Thierry Duchesne
Bruno Rémillard
Odile Marcotte

Comptes rendus

Septième atelier de résolution de problèmes industriels de Montréal

16 au 20 mai 2016

Proceedings

Seventh Montréal Industrial Problem Solving Workshop

May 16–20, 2016

CRM-3357

Préface

L'atelier de résolution de problèmes qui eut lieu au CRM en mai 2016 était entièrement consacré aux mathématiques financières et actuarielles. Nous sommes très reconnaissants aux organismes suivants de nous avoir fourni des problèmes : la Banque Nationale, The Co-operators, Desjardins Assurances générales et la Caisse de dépôt et placement du Québec. Nous aimerions aussi exprimer notre reconnaissance à Jean-François Quessy et Jean-François Plante, qui ont accepté de coordonner l'équipe étudiant le problème de Desjardins Assurances générales, et à Louis Doray, coordonnateur pour le problème de la Caisse de dépôt et placement. Finalement nous remercions le CRSNG d'avoir accordé aux trois instituts de mathématiques la subvention appelée Plateforme d'innovation des instituts, nous permettant ainsi d'organiser des ateliers où se nouent des collaborations entre mathématiciens et partenaires industriels. L'atelier a suscité l'enthousiasme des participants et nous espérons qu'il sera un prélude à l'inclusion de nombreux problèmes de statistique et de mathématiques financières et actuarielles dans les ateliers futurs.

The Problem Solving Workshop held at the CRM in May 2016 *was dedicated to financial and actuarial mathematics. We are very grateful to the following organizations for having submitted problems to the workshop participants: the National Bank*, *The Co-operators*, *Desjardins Assurances générales*, *and the Caisse de dépôt et placement du Québec. We are also grateful to Jean-François Quessy and Jean-François Plante*, *who were the coordinators of the team tackling the problem submitted by Desjardins Assurances générales*, *and to Louis Doray*, *who was the coordinator for the problem submitted by the Caisse de dépôt et placement. Finally we wish to express our thanks to NSERC for the Institutes Innovation Platform* (*IIP*), *a project of the three Canadian mathematics institutes funded by NSERC*: *the IIP allows us to organize workshops fostering collaborations between mathematicians and industrial partners. The May* 2016 *workshop was much appreciated by its participants and we hope that future workshops will include many problems from statistics and financial and actuarial mathematics.*

avril 2017

Thierry Duchesne
Bruno Rémillard
Odile Marcotte

Contents

1

Regimes Switching in Stock-Bond Correlations

Project Submitted by the National Bank of Canada

Rosemonde Lareau-Dussault, Helen Samara Dos Santos, Mario Palaciano, Éric Tsala, Kris Schmaltz Tziritas, Adel Benlagra, Caio De Naday Hornhardt, Farshid Zoghalchi, Manuel Morales, Bruno Rémillard, Pierre Laroche, and Alessandro Mina

1.1 Introduction

Correlations are a statistical measure of the linear dependence between two assets and are one of the building blocks of a diversified investment portfolio. Until recently the correlations between stock prices and bond yields have been assumed to be constant. It is known, however, that they fluctuated in time, although tending to be negative throughout much of the 20th century. They have even been largely positive since the late 1990s as shown in Fig. 1.1 for the 10-year US government bond yields and S&P500. Careful examination of this dynamics must play an important role in the process of portfolio construction.

The same figure shows the time evolution of inflation, as measured by the consumer price index, over the same time period. Notice how periods of high and variable inflation generally coincide with strong negative stock-bond cor-

Rosemonde Lareau-Dussault · Mario Palaciano · Farshid Zoghalchi
University of Toronto

Helen Samara Dos Santos · Caio De Naday Hornhardt
Memorial University of Newfoundland

Éric Tsala
Université de Sherbrooke

Kris Schmaltz Tziritas · Manuel Morales
Université de Montréal

Adel Benlagra
Université du Québec à Montréal

Bruno Rémillard
HEC Montréal

Pierre Laroche · Alessandro Mina
National Bank of Canada

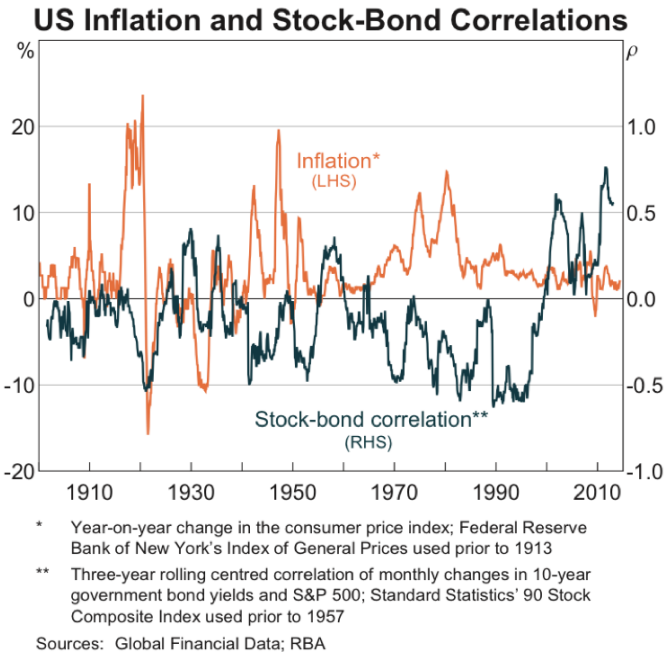

Fig. 1.1 Yearly change in the consumer price index and three-year rolling centered correlation of monthly changes in 10-year government bond yields and S&P 500.

relations. It is known that expectations for inflation, and growth, determine forecasts for dividends and interest rates. On the other hand, whether these expectations cause the stock price and the bond yields to move in the same, or opposite, directions is still poorly understood.

The aim of this project is to understand how macroeconomic factors like inflation or growth expectations affect the dynamics of stock-bond correlations and to explore whether we can define chronological regimes with specific levels for these correlations. This would help build models for their dynamics and, if possible, devise predictive techniques based on these models.

We investigated four different approaches to defining regimes in the correlations between the daily DEX bond index and S&P TSX60 index futures prices from 2011 to 2016 and their relation to proxies for the Canadian inflation and growth time evolution over the same period. We started with a naive approach (based on the variations of macroeconomic factors) to defining the aforementioned regimes. Our second approach assumes that the stock and bond log returns are described by a Hidden Markov Model (HMM) with a time-varying discrete latent variable and normal conditional distributions of the stock and bond log returns given a value for that latent variable. The distribution parameters are then estimated using maximum likelihood. Finally

we use two unsupervised machine learning techniques to define regimes in terms of hidden structures in the data.

1.2 The Naive Approach

In the so-called naive approach, economists define macroeconomic regimes based on positive or negative variations in quantities such as expected inflation, expected growth, or aggregated risk aversion. These variations are assumed to influence greatly stock and bond returns as well as their mutual dependence. Hence, depending on a given macroeconomic regime in time, it may be better to invest in stock or bond assets at that particular time.

For exemple, as shown in Fig. 1.2 below, when inflation and growth both have positive variations (regime 1), it is better to invest in stocks than in bonds. Alternatively, in regime 3, when both inflation and growth have negative variations, it is safer to invest in bonds than in stocks.

In what follows we will assess this naive picture on the basis of inflation and growth variations. We will use the daily DEX bond index and S&P TSX60 index futures prices from 2011 to 2016 together with proxies for the Canadian inflation and growth time evolution. The latter are, respectively, the Canadian breakeven inflation (BEI) rate and the Bloomberg Commodity Index (BCOM), which tracks prices of futures contracts on physical commodities on the commodity markets. The corresponding time series have $T = 1297$ data points.

According to the naive picture, we define 4 regimes based on the upward and downward daily movements in BEI and BCOM. We then associate each

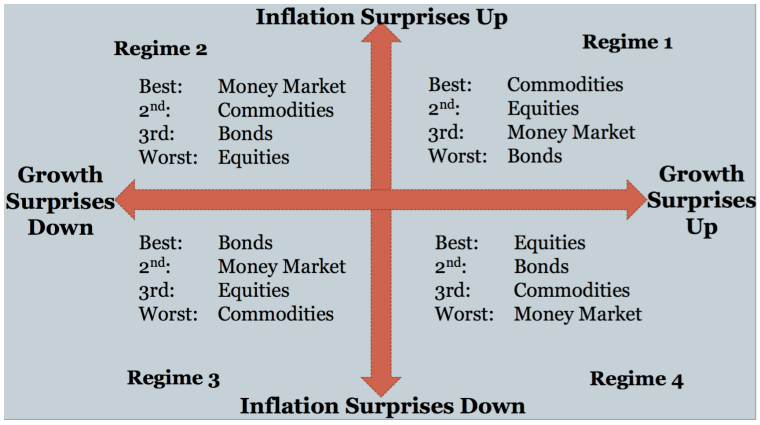

Fig. 1.2 The four regimes defined by economists with two factors: the inflation surprises and the growth surprises.

data point in the stock and bond time series to a given regime. This results in an even distribution of the data across the 4 regimes as shown in Table 1.1.

Figure 1.3 shows the stock and bond time series coloured according to the 4 macroeconomic regimes.

A first observation from this figure is that all 4 regimes are represented throughout the whole time span considered, regardless of how good or bad the economy is: these regimes do not define distinct periods of time (as one would expect). Rather they seem to correspond to different levels of stock and bond returns, regardless of the actual state of the economy. For example regime 1 seems to be associated mainly with high (positive) stock returns and low (negative) bond returns throughout the time period.

Table 1.2 displays summary statistics for the average stock and bond log returns within each regime, including their respective volatilities and correlation.

As observed earlier, regime 1 corresponds to the maximum average stock return and minimum average bond return. Regime 3, on the contrary, cor-

Table 1.1 Frequency of data points in each of the defined macroeconomic regimes. The total number of data points is $T = 1297$.

Regime	1	2	3	4
Frequency (%)	26.29	20.43	30.30	22.98

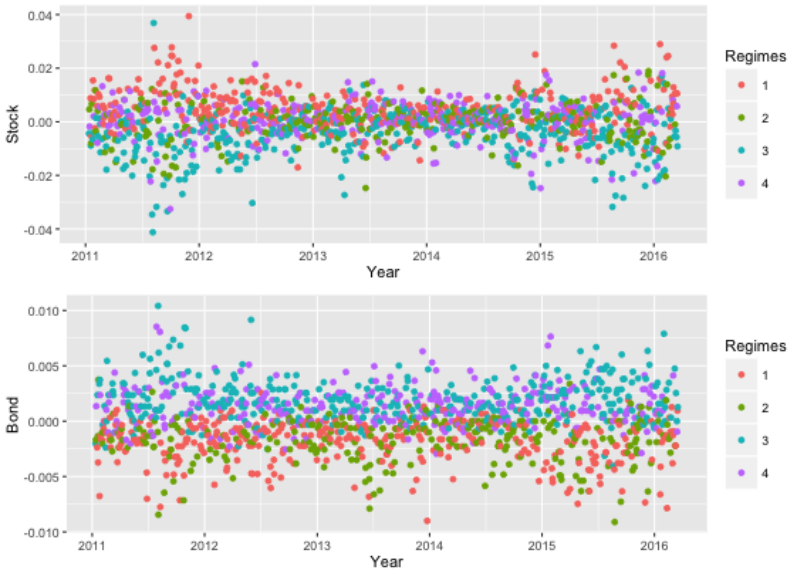

Fig. 1.3 Time series of the stock and bond log returns coloured by the expected regime given the inflation and growth variations. The time series correspond respectively to the DEX bond index and S&P TSX60 index futures from 2011 to 2016.

Table 1.2 Stock and bond averages $\mu^{S,B}$, stock and bond volatilities $\sigma^{S,B}$, and their correlations ρ within each macroeconomic regime defined by Canadian inflation and growth variations. Maximum values for the average return are displayed in red, while minimum values are displayed in blue. The data used are the DEX bond index and S&P TSX60 index futures as well as the BEI rate and BCOM index from 2011 to 2016.

Regime	1	2	3	4
μ^S	1.23	−0.02	−1.15	0.14
μ^B	−0.48	−0.43	0.48	0.34
σ^S	0.13	0.11	0.14	0.12
σ^B	0.03	0.03	0.03	0.03
ρ	−0.19	0.05	−0.30	−0.11

responds to the minimum average stock return and maximum average bond return. Both regimes have a negative stock-bond correlation. Notice that the volatilities of the stock index and the bond index do not seem to vary across the 4 macroeconomic regimes. Hence these regimes do not discriminate between normal and crisis periods, for example. In particular they are not useful as predictors, even on a short time scale, since they seem oblivious to the time dimension.

1.3 The Hidden Markov Model Approach

The HMM is a model for regime switching where the observables are assumed to follow a distribution function whose parameters depend on a **latent**, i.e., nonobservable, discrete process Z_t that switches randomly between, say, K different values (each value belonging to $i = 1, 2, \ldots, K$). Each value corresponds to a regime. The dynamics describing the switching between the different regimes is modelled by a Markov chain with a transition matrix $\boldsymbol{Q}$ defined as follows:

$$\mathbb{P}(Z_{t+1} = j \mid Z_t = i) = Q_{ij} \ . \tag{1.1}$$

In the Gaussian case the observables follow a normal distribution conditionally on Z_t. In particular, in our case, the log returns of the stock index and the bond index are assumed to be distributed according to

$$\boldsymbol{Y}_t = \begin{pmatrix} Y_t^B \\ Y_t^S \end{pmatrix} \Bigg| \, Z_t = i \sim \mathcal{N}_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \ , \quad i = 1, \ldots, K \ , \tag{1.2}$$

where $\boldsymbol{\mu}_i$ is a two-dimensional vector and $\boldsymbol{\Sigma}_i$ is a 2×2 variance-covariance matrix.

When $K = 1$ and $\Sigma_{12} = 0$ hold, we recover the special case of the Black and Scholes (BS) model for the dynamics of the stock index and bond index. The HMM retains some of the simplicity of the BS model, in particular its mathematical tractability, while it captures more accurately the stylized facts observed empirically in financial markets, in particular the stochastic persistent volatility in financial time series.

Figure 1.4 shows a simulated time series using the HMM with $K = 2$. On the left-hand side of the figure, we display a plot of the stock log return y_t^{S} dynamics together with the values of the latent variable Z_t. The stock index switches between a low return/low volatility regime when $Z_t = 2$ holds and a high return/high volatility regime when $Z_t = 1$ holds.

Note that the differences between the respective return and volatility values in the two regimes have been exaggerated for visual clarity. On the right-hand side, we show the correlation between the stock and bond log returns, calculated on a rolling window over 15 time periods, and its comparison with the correlation ρ_{12} in the HMM model.

Given a sample $\boldsymbol{y}_{1:T} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, and assuming that the underlying model is an HMM, it is possible to estimate the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{Q})$ of the model. This is usually carried out by using maximum-likelihood estimation. The likelihood function $\mathcal{L}(\boldsymbol{y}_{1:T}; \boldsymbol{\theta})$, however, may be difficult to evaluate or maximize in the presence of latent variables. Fortunately it is possible to find the maximum likelihood estimate using the expectation-maximum (EM) iterative algorithm introduced by Dempster et al. [1977]. This algorithm has two main steps. Starting from an initial guess $\boldsymbol{\theta}^{(0)}$, the value $\boldsymbol{\theta}^{(k)}$ of the parameters $\boldsymbol{\theta}$ at the kth iteration is calculated as follows:

1. Define the function

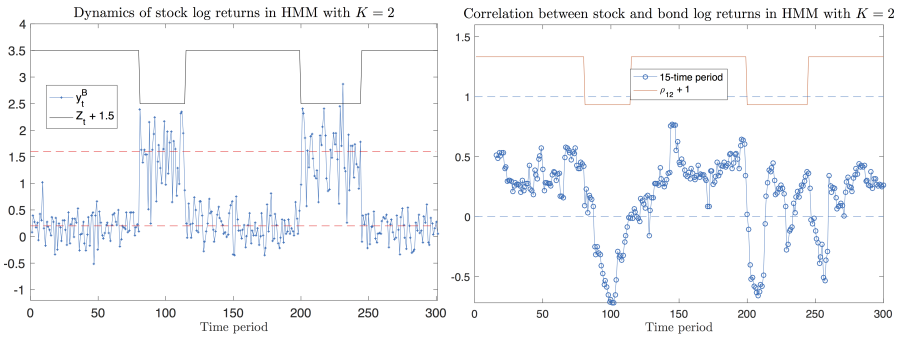

Fig. 1.4 Dynamics of the stock log return (left) and rolling window correlation, over 15 time periods, between the stock and the bond log returns (right) in the HMM with $K = 2$. The values of the latent variable as well as the correlation ρ_{12} are also shown in each plot, respectively. The parameters used to simulate $\boldsymbol{Y}_t$ are: $\boldsymbol{\mu}_1 = (1.6, -0.4)^\top$, $\boldsymbol{\mu}_2 = (0.6, -1.4)^\top$, $Q_{12} = 0.0371$, and $Q_{21} = 0.0101$.

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) \equiv \mathbb{E}\big[\ln f(\boldsymbol{y}_{1:T}, Z_{1:T}; \boldsymbol{\theta}) \,\big|\, \boldsymbol{y}_{1:T}, \boldsymbol{\theta}^{(k-1)}\big] \,, \tag{1.3}$$

where $f(\boldsymbol{y}_{1:T}, Z_{1:T}; \boldsymbol{\theta})$ is the joint density of $\boldsymbol{y}_{1:T}$ and $Z_{1:T}$.

2. Find $\boldsymbol{\theta}^{(k)}$ as

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) \,. \tag{1.4}$$

Using the chain rule, the joint density function in (1.3) can be written as

$$f(\boldsymbol{y}_{1:T}, Z_{1:T}; \boldsymbol{\theta}) = f(\boldsymbol{y}_{1:T} \mid Z_{1:T}; \boldsymbol{\theta}) f(Z_{1:T}; \boldsymbol{\theta}) \,, \tag{1.5}$$

and the function Q can be split into two terms:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) = \mathbb{E}\big[\ln f(\boldsymbol{y}_{1:T} | Z_{1:T}; \boldsymbol{\theta}) \,\big|\, \boldsymbol{y}_{1:T}, \boldsymbol{\theta}^{(k-1)}\big] \\ + \mathbb{E}\big[\ln f(Z_{1:T}; \boldsymbol{\theta}) \,\big|\, \boldsymbol{y}_{1:T}, \boldsymbol{\theta}^{(k-1)}\big] \,. \end{aligned} \tag{1.6}$$

The first term depends only on the parameters of the conditional distribution of $\boldsymbol{Y}_t$ given $Z_t = i$, i.e., $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$, while the second term depends only on the distribution of the latent variable Z_t, i.e., only on the transition matrix $\boldsymbol{Q}$. Hence the problem of maximizing the likelihood function $\mathcal{L}(\boldsymbol{y}_{1:T}; \boldsymbol{\theta})$ is split into two maximization subproblems of reduced dimensionality. We will not go into further details but we can find an explicit expression of $\boldsymbol{\theta}^{(k)}$ in the Gaussian case. Dempster et al. [1977] showed that the likelihood function $\mathcal{L}(\boldsymbol{y}_{1:T}; \boldsymbol{\theta})$ is increasing at each iteration:

$$\mathcal{L}(\boldsymbol{y}_{1:T}; \boldsymbol{\theta}^{(k)}) \geq \mathcal{L}(\boldsymbol{y}_{1:T}; \boldsymbol{\theta}^{(k-1)}) \,. \tag{1.7}$$

The algorithm, however, may converge towards a local maximum of the likelihood function; one should use different initial guesses or a global optimization routine in order to find the global maximum.

Until now, the number of regimes K has been fixed. To choose an appropriate value for K, Rémillard et al. [2010] and Rémillard [2011] proposed a test of goodness-of-fit using the Cramér–von Mises test based on the Rosenblatt transform. They proposed to choose the smallest K for which the P-value of the goodness-of-fit test for K regimes is greater than 5%.

Using simulated data and the real market data introduced in the last section, the estimation of the model and the test of goodness-of-fit were carried out using the C and Matlab codes kindly provided by Bruno Rémillard.

We began with the simulated data. The aim here is to show that the EM algorithm is able to recover the parameters of the model from data simulated with that same model and to recover the correct regime at a given point in time. We choose a model with $K = 2$ in which the only differences between the two regimes lie in the stock and bond standard deviations and the correlation. This means that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ holds but $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ also holds. The simulated data have $T = 1280$ data points. Table 1.3 below shows the estimation of the

parameters of the model together with their true values (used to produce simulated data).

Table 1.3 Estimation of the parameters of the model from the simulated data. Values between parentheses are the true values of the HMM used to produce the simulated data.

Regime	μ^{S}	μ^{B}	σ^{S}	σ^{B}	ρ_{12}
1	0.6140 (0.6)	−0.3750 (−0.4)	0.5949 (0.6)	0.6309 (0.6)	0.1063 (0.1)
2	0.5996 (0.6)	−0.4112 (−0.4)	0.2482 (0.25)	0.2482 (0.25)	−0.4020 (−0.4)

As we can see from the table, the algorithm is able to estimate the parameters of the model and discriminate between the two regimes (although they only differ in the value of a single parameter). Using the same code, we can estimate $\mathbb{P}(Z_t = i)$ for all values of t and i. Then we could *filter* the value of Z_t by choosing i so that $\mathbb{P}(Z_t = i)$ is maximal. The confusion matrix (Table 1.4) shows the number of true values of Z_t that we recover with this method.

The accuracy of the filtering is of the order of 92.65%. The method used to recover the regime is thus fairly accurate.

We now turn to the results using the DEX bond index and S&P TSX60 index futures from 2011 to 2016. With $N = 1000$ bootstrap samples generated to estimate the P-value, we find a P-value of 0% for 4 (or fewer) regimes and a P-value of 6.6% for 5 regimes. Hence we choose a Gaussian model with 5 regimes. The parameters of the conditional distribution in each regime are shown in Table 1.5.

While the volatility of the bond log return is of the same order of magnitude in all 5 regimes, the volatility of the stock log return clearly varies across the regimes. Regime 5 has the highest volatilities in both the stock and the bond log returns while it has the second and third lowest values (respectively) in the stock and bond mean log returns. The correlation between the stock and bond indices is also the largest in magnitude, though with a

Table 1.4 Confusion matrix for the filtered values of Z_t using the EM algorithm on the simulated data.

		True regime	
		1	2
Filtered regime	1	319	43
	2	51	867

Table 1.5 Parameters of the conditional normal distribution in each HMM regime. Maximum values for the average return are shown in red while minimum values are shown in blue. The data used for estimation are the daily DEX bond index and S&P TSX60 index futures from 2011 to 2016.

Regime	1	2	3	4	5
μ^{S}	−0.0397	−0.7980	0.6491	0.2788	−0.2045
μ^{B}	0.0199	0.0635	−0.3429	0.0926	0.0394
σ^{S}	0.1294	0.0957	0.0537	0.0664	0.2258
σ^{B}	0.0356	0.0439	0.0476	0.0242	0.0585
ρ_{12}	−0.3879	0.1281	0.1744	−0.1710	−0.4298

negative sign, in this regime. Regime 1 has the smallest bond and stock mean log returns in magnitude, the second largest stock volatility and, comparatively, a strong negative stock-bond correlation. Only regimes 2 and 3 have a positive stock-bond correlation. These correspond, respectively, to the lowest stock and bond mean log returns.

Figure 1.5 shows the time series of the stock and bond log returns coloured according to the 5 different regimes, filtered using the method described earlier and used on the simulated data.

Over the time period considered, the stock and bond indices are in regime 1 almost half of the time (619 data points amongst a total of 1297 data points)

Fig. 1.5 Time series of the stock and bond log returns coloured by the filtered values of Z_t. The time series correspond to the daily DEX bond index and S&P TSX60 index futures from 2011 to 2016.

while regimes 2 and 3 are barely represented(respectively 58 and 65 data points). It can be seen that regime 5, i.e., the regime with the highest volatilities, can be associated with two major events: the US debt-ceiling crisis in mid 2011 and the emergent markets crisis at the end of 2015. It is not clear how to interpret regimes 1 and 4. It is tempting to associate the former with regime 5, the two regimes corresponding to heating up and cooling down periods around high volatility times. Regime 4 could be associated with more quiet periods.

As can be seen in Fig. 1.6, these 5 regimes cannot be associated with positive or negative variations of the BEI rate (inflation proxy) and the BCOM index (growth proxy). Hence they are not clearly related to the macroeconomic regimes defined in the naive approach. They prove, however, to be better indicators of the state of the economy.

Finally, based on the estimation of the transition matrix $\boldsymbol{Q}$ and given the assumed conditional distribution within each regime, we may have some predictive power: if we are currently in a specific regime, the transition matrix $\boldsymbol{Q}$ provides us with the probability of being in another regime at any time in the future. For any realized regime in the future, we could sample a data point from the conditional distribution and obtain a stock and bond return prediction. We did not investigate this possibility during the short time we spent at the CRM. But in principle we could have split our data into a training data set and a testing data set; we would have been able to check whether future regimes can be predicted accurately or not.

Fig. 1.6 Time series of the BEI rate and the BCOM index coloured by the filtered values of Z_t. The time series correspond to the daily DEX bond index and S&P TSX60 index futures from 2011 to 2016.

1.4 The Unsupervised Machine Learning Approach

So far we have defined regimes somewhat arbitrarily, either by linking them to the positive and negative variations of some macroeconomic factors or by assuming an underlying model describing the stock and bond data. In this section we will explore whether the data contain some structure without any prior assumption. This will be done using unsupervised machine learning techniques, in particular clustering and neural networks. These are sets of algorithms that can describe hidden structures or patterns in unlabelled data. We hope to uncover patterns related to chronological regimes with specific stock-bond correlations. If we could uncover these patterns, the distribution of the data within each regime and the transitions from one regime to another would enable us to predict the future levels of stock-bond correlations.

1.4.1 *Hierarchical Clustering* (*HC*)

Clustering is a set of techniques used to group data into subsets of *similar* points called clusters. There are many ways to define the notion of "cluster" and hence many different clustering algorithms. We will use the HC algorithm, where the similarity between observations is based on the notion of distance in a multidimensional space.

HC is a bottom-up approach where each observation is initially in its own cluster; clusters are then merged gradually using a notion of relative distance. The hierarchical arrangement of the clusters is nicely illustrated by a so-called dendrogram, displayed in Fig. 1.7.

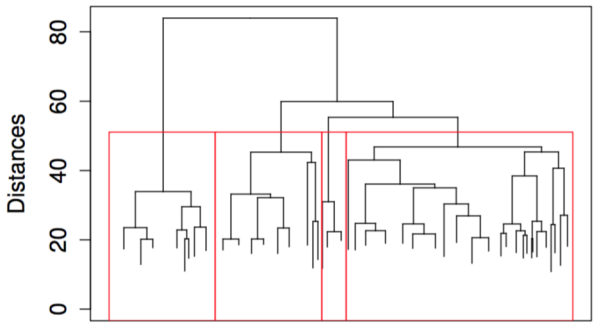

Fig. 1.7 Illustration of a dendrogram and the manner of selecting the number of clusters. Here 4 clusters have been selected.

In a dendrogram, each node corresponds to a cluster linked to its children clusters through branches. The vertical axis represents the height of the

cluster, i.e., the intergroup dissimilarity between its children clusters. The measure of dissimilarity is the distance between the clusters: the more distant the clusters, the more dissimilar they are. This distance depends on the choice of an appropriate metric (Euclidean, Manhattan, etc.), as well as a so-called linkage criterion. The former determines the distance between pairs of observations considered as points in a multidimensional space. The latter determines how to define the distance between two sets of observations. We may for instance take the distance between two sets to be the maximum distance between any two points not belonging to the same set, or the distance between their centroids. We chose to stick with the standard Euclidean metric and chose Ward's method (see Ward [1963]) as a linkage criterion; it is based on the minimization of the total within-cluster variance. We must keep in mind that the choice of a metric and a linkage criterion influences the shape and definition of the clusters.

Note that the HC algorithm does not choose the adequate number of clusters to consider. We can cut the dendrogram tree arbitrarily at any level, as shown in red in Fig. 1.7, to obtain a given number of clusters. Various methods can be used to determine the *best* cutting section. One of the simplest and most visual methods is to cut at a point where the branches from a cluster to its children clusters are the longest (since this maximizes the dissimilarity between the children clusters). In Fig. 1.7 this cutting criterion would result in two clusters.

We begin by applying the HC algorithm on the simulated data used in Sect. 1.3. These were data generated using an HMM model and hence the regimes and the conditional distribution within each regime are known.

In the first step we will only use the simulated stock and bond log returns time series as the input to the HC algorithm. With two variables, this means that the multidimensional space has dimension $d = 2$. Figure 1.8 below shows the resulting dendrogram in the left panel.

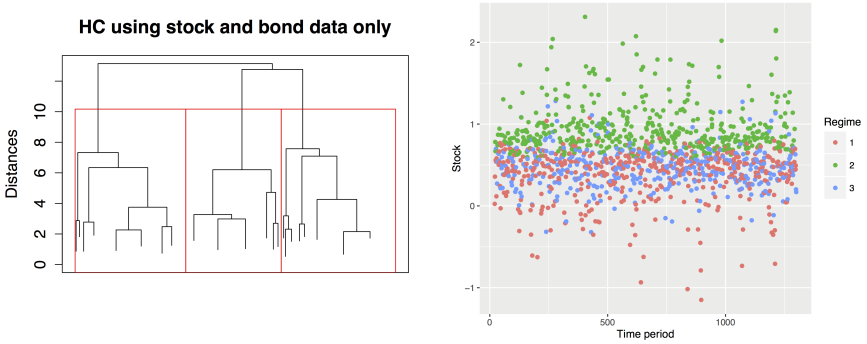

Fig. 1.8 Dendrogram structure (left) and simulated stock log return time series coloured by clustered regimes (right). Only the stock and bond log returns have been used in the clustering algorithm.

Given the cutting criterion chosen above, we would obtain 3 clusters (while the data have been simulated under a two regimes hypothesis). The right panel in Fig. 1.8 shows the stock log return time series coloured by the 3 HC clusters. It is clear that these clusters have been chosen according to the level of the stock (and bond) log returns and not the stock-bond correlation. Remember that the simulated regimes differed in the stock and bond volatilities and the correlation but not in the average stock and bond log returns.

Table 1.6 Parameters of the conditional distribution in each HC regime. We used the HC algorithm with the simulated stock and bond data of Sect. 1.3. The true value of the parameters are given in Table 1.3

Regime	μ^{S}	μ^{B}	σ^{S}	σ^{B}	ρ_{12}	Number of data points
1	0.3775	−0.6474	0.3100	0.3305	0.1221	457
2	0.9543	−0.4661	0.2973	0.3075	0.0488	442
3	0.4649	−0.0274	0.2385	0.2932	0.0387	381

Table 1.6 displays some summary statistics within each of the 3 HC clusters. The cardinalities of these clusters are roughly equal and we confirm that the distinction between them is mainly due to the level of the stock and bond log returns. The volatilities are of the same order of magnitude and the stock-bond correlation is positive in all clusters.

How can we explain that the HC algorithm does not find the 2 regimes in our simulated data? And how can we explain that the 3 HC regimes extend over the whole time span instead of being tied to specific time periods? The answer to both questions, in particular the second one, is that the HC algorithm seems to be completely blind to the time dimension of our data (which is, of course, a crucial aspect of a time series).

To remedy this blindness and help the HC algorithm, we derived 4 new variables from the stock and bond log returns. These are the 20-time-period rolling-window stock and bond standard deviations and correlation as well as the 10-time-period rolling-window stock-bond correlation. Hence the multi-dimensional space is now of dimension $d = 6$. Figure 1.9 shows the resulting dendrogram on the left and the coloured stock log return time series on the right.

The cutting criterion now yields two clusters. It is clear from the stock log return time series that the algorithm is now aware of the time nature of our data thanks to the additional input from the derived data. The two HC clusters now defines periods of time instead of log return levels. We can visually associate the first cluster to a high volatility regime and the second cluster to a low volatility regime.

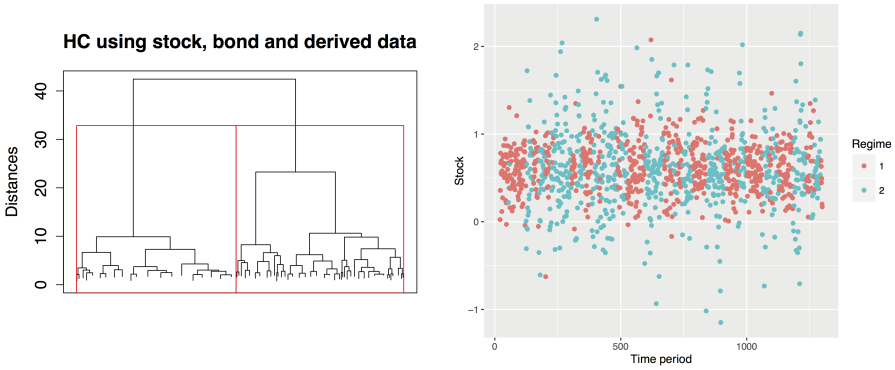

Fig. 1.9 Dendrogram structure (left) and simulated stock log return times series coloured by clustered regimes (right). The stock and bond log returns as well as the derived data have been used in the clustering algorithm. The derived data are the 20-time period rolling-window stock and bond standard deviation and correlation as well as the 10-time period rolling window stock-bond correlation.

Table 1.7 Parameters of the conditional distribution in each HC regime. We used the HC algorithm with the simulated stock and bond data of Sect. 1.3 as well as the derived data defined in the text. The values of the HMM regime parameters are given within parentheses.

Regime	μ^{S}	μ^{B}	σ^{S}	σ^{B}	ρ_{12}
1	0.5931 (0.6)	−0.3434 (−0.4)	0.4614 (0.6)	0.4851 (0.6)	0.0723 (0.1)
2	0.6128 (0.6)	−0.4600 (−0.4)	0.2830 (0.25)	0.2789 (0.25)	−0.3449 (−0.4)

Table 1.7 shows some summary statistics within each of the two HC clusters. These parameters values are rather close to the parameters used to simulate the two regimes with the HMM model. The confusion matrix (Table 1.8) shows the concordance between the HC clusters and the HMM regimes used to simulate the data. The accuracy of the HC algorithm is 67.5%. It is far less accurate than the EM algorithm. The EM algorithm, however, is tailored to data generated using the HMM while the HC algorithm is more general in its use.

Using the simulated data, we showed that the HC algorithm can be sensitive to the structure of the data if we provide it with the appropriate data. Using only the stock and bond log return time series kept the algorithm blind to the time nature of our data. With the appropriate derived data, however, we were able to help it discover this aspect of the data and find the correct regimes with an acceptable accuracy.

Keeping this in mind, we turn to the market data considered previously. Together with the DEX bond index, the S&P TSX60 index futures, the BEI

Table 1.8 Confusion matrix for the HC clusters compared to the HMM regimes.

		HMM regime	
		1	2
HC cluster	1	559	65
	2	351	305

rate, and the BCOM index, we used the following variables when applying the HC algorithm: the Chicago Board Options Exchange Volatility Index or VIX (measuring the implied volatility of the S&P 500 index, and appropriate for the S&P TSX60 index as well); the ratio between the bond and stock volumes; the stock and bond momenta; and the lagged time series of all the variables up to 5 time periods. All these additional variables are assumed to provide enough information to the HC algorithm for it to capture the time dimension of the problem.

The HC algorithm and the resulting dendrogram point towards the presence of 4 regimes. Figure 1.10 shows the clusters-coloured stock and bond time series as well as the empirical density functions within each regime. By comparing with Fig. 1.5, we first observe that there is a rather large overlap between the regimes defined by the two algorithms. In particular we observe that the fourth HC regime is closely related to the fifth HMM regime with the largest stock and bond volatilities and the third HC regime closely follows the fourth HMM regime. The first HC and HMM regimes seem also to be related.

When we look at the empirical density functions, however, it is clear that the stock and bond data are not normally distributed within each HC regime. This is particularly the case in the third and fourth regimes. Hence the assumption of normality used in the HMM is clearly rejected. We may still think of a transition from a regime to another driven by a hidden Markov chain. We can even empirically estimate a transition matrix by monitoring the historical transition from one regime to another.

1.4.2 *SOM*

1.4.2.1 Introduction

Recall that in financial and economic literature the notion of a "state of the economy (or regime)" is often taken for granted. Regimes are often defined in a crude fashion based on intuition (see naive approach in Sect. 1.2). The goal of this section is to define the notion of "regime" in a concrete manner using actual market data, i.e. by using real data we would like the answer

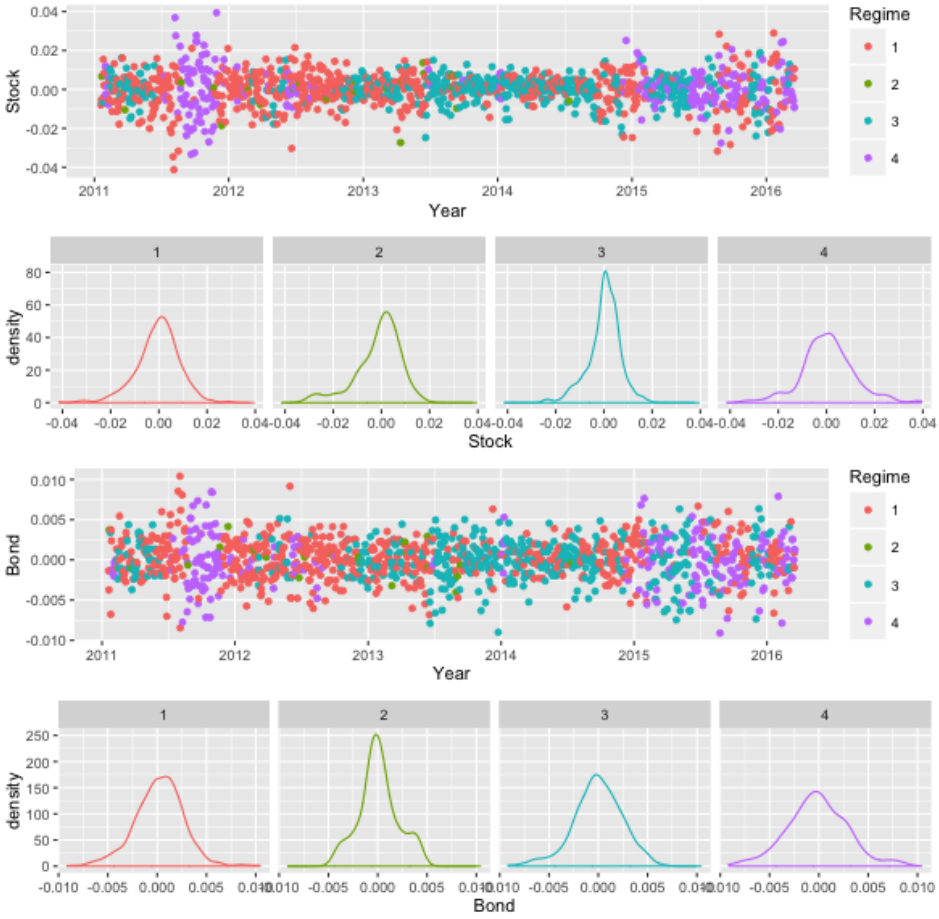

Fig. 1.10 Stock (two upper panels) and bond (two lower panels) coloured times series and empirical density functions within each HC cluster. The data used are the daily DEX bond index and S&P TSX60 index futures, the BEI rate, the BCOM index, as well as the derived data described in the main text.

the following question: what is a "regime"? We will try to shed some light on this question by using the machine learning algorithm SOM (self-organizing maps). This SOM algorithm will be described in the following section.

The data set that we are using can be thought of as a toy market. In our data set we have a collection of time-dependent vectors that are of the following form:

$$(t, B_t, V_t^{\mathrm{B}}, S_t, V_t^{\mathrm{S}}, \Sigma_t, I_t, G_t), \tag{1.8}$$

where

Table 1.9 Parameters of the conditional distribution in each HC regime. Maximum values for the average return are shown in red, while minimum values are shown in blue. The data used by the HC algorithm are the daily DEX bond index and S&P TSX60 index futures, the BEI rate, the BCOM index, as well as the derived data described in the main text.

	1	2	3	4
μ^{S}	−0.09	−0.28	0.04	0.16
μ^{B}	0.05	0.02	−0.01	−0.04
σ^{S}	0.14	0.14	0.11	0.17
σ^{B}	0.04	0.03	0.04	0.05
ρ	−0.39	−0.25	−0.12	−0.43

t is the date,
B_t is the log return on the bond index at date t,
V_t^{B} is the volume of the bond index at date t,
S_t is the log return on the stock index at date t,
V_t^{S} is the volume of the bond index at date t,
Σ_t is a stock market volatility index,
I_t is a proxy for inflation at date t, and
G_t is a proxy for GDP at date t.

In order to define a regime we must first check whether there is any clustering in the data. If there are clusters then, naturally, we can define a regime as a cluster. The problem is that we are dealing with high-dimensional data vectors; therefore looking for clusters is not trivial. In order to look for clusters in the data we will use the machine learning algorithm SOM.

The goal of this section, which is to define what is meant by regime, can thus be reframed as the problem of determining whether there are clusters in our high-dimensional time-dependent data set.

1.4.2.2 Self-Organizing Maps (SOM)

Self-organizing maps (SOM) are one the most widely used neural network models, which can be used to classify, organize, and visualize large data sets. SOM belongs to a large class of methods based on competitive learning networks. SOM is an unsupervised machine learning algorithm, which means that no human intervention is required during the learning process and little knowledge of the input data is needed. It can be used to discover elaborate structures and patterns in high-dimensional data.

Here are two advantages of using SOM.

- It is a non-parametric method. This is useful in finance because estimating model parameters using market data can be challenging.
- There is no need to make any a priori assumptions about the distribution of the data. This is also useful in finance because erroneous distributional assumptions are common.

SOM takes high-dimensional data and assigns each data vector to a node. This creates the type of two-dimensional grid illustrated in Fig. 1.11 (all figures were made with the R package kohonen). Note that this is a "counts" grid, one of many grid types. Each node (circle) on this grid contains a certain number of data vectors from the original data set, represented by the colour of the node. One may think of each node as a bin containing a certain number of vectors from the data set. Before we move on it is important to state that SOM provides a topology-preserving mapping from the data set to the grid. In other words, data vectors in the same node are "close" to one another (in terms of Euclidean distance); also if two nodes are close the data vectors in those nodes are close as well.

In Fig. 1.12 we have what is called a heatmap. These are used when we want to identify the distribution of a specific variable (i.e. an individual component of the data vectors) across the entire grid. The bond index in our data set is the DEX universe bond index and Fig. 1.12 provides a picture of the average value of the DEX in each node.

Before we move on to clustering with SOM we provide an intuitive description of the SOM algorithm. Each node on the grid contains a single "weight vector." These weight vectors must be of the same dimension as the data vectors in the data set.

STEP 1: Initialize all weight vectors randomly.
STEP 2: Choose a data vector at random from the data set.

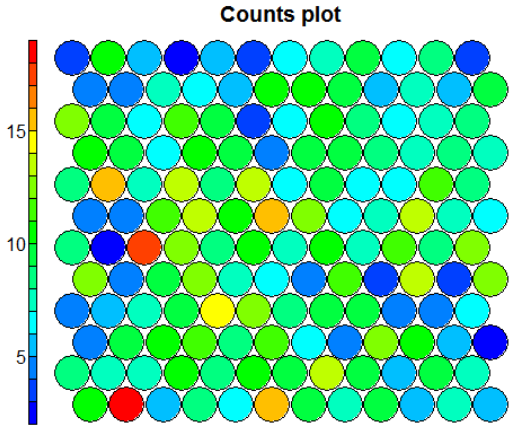

Fig. 1.11 Number of observations within each node

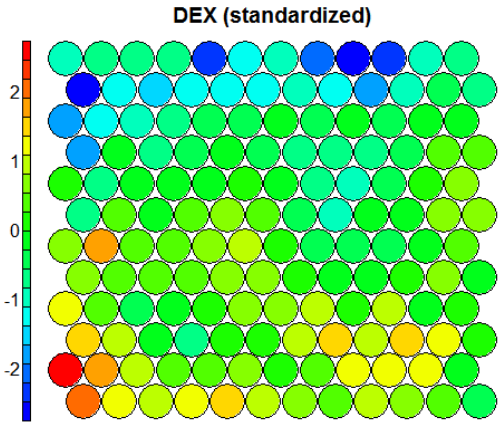

Fig. 1.12 Average standardized value of the bond index (DEX) in each node

STEP 3: Find the "winning" node. This is the node whose weight vector is closest to the data vector in terms of Euclidean distance.

STEP 4: Determine the "neighbourhood" of the winning node. The neighbourhood is defined by a circle centered at the winning node. This circle has a radius determined by a decreasing "neighbourhood function" of the form

$$R(n) = a\mathrm{e}^{-n/b} ,$$

where n denotes the iteration number. Note that the size of the neighbourhood decreases from one iteration to the next.

STEP 5: Adjust every weight vector in the neighbourhood by applying the following equation:

$$W(n+1) = W(n) + L\big(n\big)\big(V(n) - W(n)\big) ,$$

where

n is the iteration number,
$L(n) = c\mathrm{e}^{-n/b}$ is the "learning rate function" for some fixed c and b,
$W(n)$ is the weight of the node considered in the neighbourhood,
$V(n)$ is the data vector from STEP 2.

Note that the constant b is the same in both the neighbourhood function and the learning rate function.

STEP 6: Return to STEP 2.

1.4.2.3 Clustering with SOM

SOM gathers data vectors that are "similar" and includes them into a grid node; therefore each node can be thought of as some sort of cluster. In general, however, we would like to have only a handful of clusters; these clusters will be considered as regimes. In order to do this we start with a "neighbour distance plot", shown in Fig. 1.13. This plot can tell us whether there are clusters of nodes on the grid. Nodes with low neighbour distances are deemed to be similar. If there are clusters of nodes on the grid, this plot can be used to identify how many clusters there are.

Another way to visualize clustering in the data is to use standard statistical techniques to determine the number of clusters and then examine a cluster plot (see Fig.1.14). Note that in order to create a cluster plot one must input the number of clusters in advance.

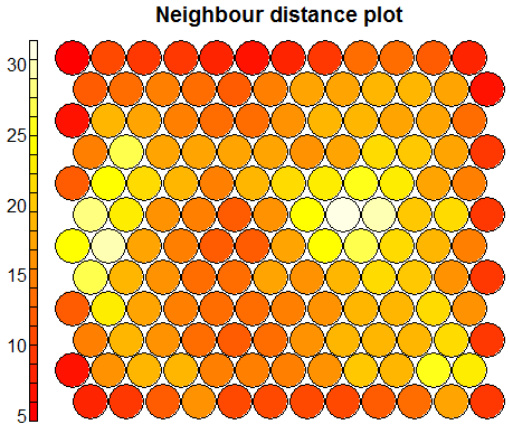

Fig. 1.13

1.4.2.4 Results and Conclusion

While working on our project we recognized that SOM works very well to determine elaborate structures in high-dimensional data when the data are NOT time-dependent, for example, if one has a large number of data vectors of the form

$$(\text{height}, \text{weight}, \text{age}, \text{salary}) \ .$$

In our case, however, the data vectors are time-dependent and SOM did not yield useful results. This is because there is nothing in the SOM algorithm that would allow SOM to "understand" the difference between time-

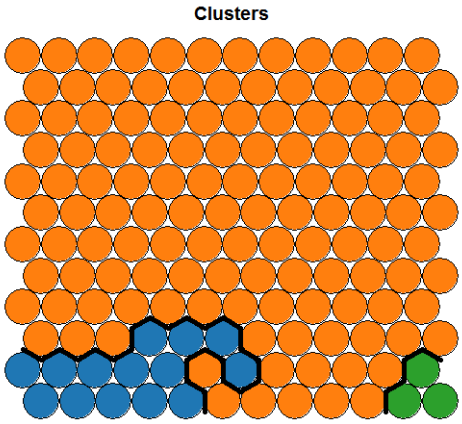

Fig. 1.14 Clustering of nodes created by SOM

dependent and time-independent data vectors. In order to deal with this problem, we inserted time-dependent components into our data vectors in order to get SOM to "understand" that our data vectors are time-dependent. For example we added data such as running volatilities to each data vector; so instead of (1.8) we obtained data vectors of the form

$$(t, B_t, \mathrm{RV}(5), V_t^{\mathrm{B}}, S_t, V_t^{\mathrm{S}}, \Sigma_t, I_t, G_t)\ , \tag{1.9}$$

where RV(5) denotes the running volatility over the past 5 days. The goal was to do this in such a way that SOM learn to find elaborate structures in the data while taking time-dependence into account. We tried to implement this idea in many different ways but were unable to achieve our goal. We did, however, discover some promising literature at the end of the workshop; therefore we are confident that SOM can be used to discover clusters in time-dependent high-dimensional data.

References

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

B. Rémillard. Validity of the parametric bootstrap for goodness-of-fit testing in dynamic models. Working Paper 1966476, SSRN, 2011.

B. Rémillard, A. Hocquard, and N. A. Papageorgiou. Option pricing and dynamic discrete time hedging for regime-switching geometric random walks models. Working Paper 1591146, SSRN, 2010.

J. H. Ward, Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.

2
Event Variables in Client Analytics
Project Submitted by The Co-operators

Alberto Alinas, Thierry Duchesne, Émilie Lavoie-Charland, Mernoosh Malekiha, James McVittie, Idir Saïdani, Arusharka Sen, Joey Wang, and Meng Zhao

2.1 Introduction

This report is a summary of the work completed during the Seventh Industrial Problem Solving Workshop held by the CRM at the Université de Montréal. Our problem was posed by the R&D team of the company, The Co-operators, which has approximately 5000 employees and is considered one of the largest general insurance firms in Canada. The Co-operators is involved with P&C insurance covering vehicles, houses, businesses and farms, life insurance, travel insurance, and investment insurance, corresponding in total to over 3 million policies. The R&D team wants to utilize better their vast amounts of data in the improvement of marketing practices for client

Alberto Alinas
Department of Modelling and Computational Science, University of Ontario Institute of Technology, e-mail: alberto.alinas@uoit.ca

Thierry Duchesne
Département de mathématiques et de statistique, Université Laval, e-mail: thierry.duchesne@mat.ulaval.ca

Émilie Lavoie-Charland · Idir Saïdani
Research and Innovation, The Co-operators, Insurance and Financial Services, e-mail: emilie.lavoie-charland@cooperators.ca, e-mail: idir.saidani@cooperators.ca

Merhnoosh Malekiha · Arusharka Sen
Department of Mathematics and Statistics, Concordia University, e-mail: merhnoosh.malekiha@gmail.com, e-mail: arusharka.sen@concordia.ca

Joey Wang
Department of Mathematics, University of British Colombia Okanagan, e-mail: wjoey2015@gmail.com

James McVittie · Meng Zhao
Department of Mathematics and Statistics, McGill University, e-mail: james.mcvittie@mail.mcgill.ca, e-mail: meng.zhao3@mail.mcgill.ca

retention and cross-sale. Their specific goal for the workshop was to develop models to answer two questions.

- When will an existing client leave the company (client retention)?
- When will an existing client add a new product – in particular, life insurance (cross-sale)?

For each of these questions, four objectives were set: formulation of the response variable; development of statistical methodology for incorporating effects of time-varying covariates; and determining suitable methods of validation and measures of performance. This report is organized in such a way as to address the two questions sequentially, outlining our attempts to meet these objectives.

2.2 Overview of the data

Two data sets were available during the workshop. The training set consisted of data on approximately 43,000 households with approximately 99,000 rows, where each row records the policies owned by a household for a particular interval of time. At any time, a household can own any combination of 5 types of insurance products: car, home, life, farm, and business. Also recorded for each time interval are other covariates about the household such as tenure with the company (from the first entry of the household in the data set); marital status; presence of a male and presence of a female in the household; auto premium amount; home premium amount; number of vehicles owned by the household; highest level of education; age of oldest household member, etc. Each event, meaning the purchase of at least one policy by the household or the filing of at least one claim by the household, corresponds to an entry (a row) belonging to that household's ID. The variable "event_start" is the calendar date of this change and the variable "event_stop" is the calendar date of the next change. The time-fixed "tenure" covariate giving the household's "age" in the company (from its first entry) is expressed as a whole number of years. Each household's entries terminate at the earliest date when it no longer owns any insurance products from The Co-operators. The binary variable "is_leaving" indicates whether a household ID has left the company at the end of the interval.

The testing set has exactly the same structure as the training set, except that it contains fewer rows; its data on approximately 34,000 households (different from the ones in the training set) were set aside for use in the model validation and prediction steps.

2.3 Client retention: When will an existing client leave the company?

This initial question motivated the more precise and practical question: at a given point in time, which clients have the highest risk of leaving the company? The ideal solution to such a question would be actionable; The Co-operators wants to be able to rank clients in order of their risk of leaving the company. For example ranking the clients in the middle of a given policy year would give marketing agents guidance about which clients to prioritize.

2.3.1 *Hazard-based modelling*

In studying the problem of predicting the risk of leaving the company, one must first define the response variable. We defined it as the length of time between the first moment when a household joins the company and the first moment when it has no longer insurance products from the company. This means that time 0 for each household is the time when the household first becomes a client of the company, and that the risk we model is a function of household "age" in the company; we want a household's age and covariates to differentiate its risk from that of another household. Because of this, the "event_start" and "event_stop" variables have to be carefully modified for every household in the training and testing sets, in order to reflect not calendar time but how long the household has been a client.

In survival analysis terms, the (instantaneous) risk of an individual leaving the company is represented by the *hazard*. The most widespread approach to modelling the hazard function is the Cox proportional hazards (Cox PH) model, which is largely the approach that The Co-operaters had already been using (they were incorporating only time-fixed covariates, however). We gave them potential directions to pursue in the modelling of time-varying covariates. We present here a mathematical explanation for estimating the hazard under the assumption that all variables are time-independent as well as numerical alternatives through R for modelling time-dependent covariates.

2.3.1.1 Semiparametric Cox PH model with time-varying covariates

The semiparametric Cox PH model is a well-known model in survival analysis. In brief, it assumes that for a subject with a vector of (possibly time-dependent) covariates $\boldsymbol{z}(t)$, the *hazard* function is of the form:

$$\lambda(t|\boldsymbol{z}(t)) = \lambda_0(t)e^{\boldsymbol{z}(t)^\top \boldsymbol{\beta}}, \tag{2.1}$$

where there are no assumptions on the form of $\lambda_0(t)$, the baseline hazard common to all subjects, and $\boldsymbol{\beta}$ is the vector of parameters for the effects corresponding to $\boldsymbol{z}(t)$. If the covariates are time-independent, then it is straightforward to obtain the *cumulative hazard* function via the following integration.

$$\Lambda(t|\boldsymbol{z}) = \int_0^t \lambda_0(u) e^{\boldsymbol{z}^\top \boldsymbol{\beta}} du = e^{\boldsymbol{z}^\top \boldsymbol{\beta}} \int_0^t \lambda_0(u) du = e^{\boldsymbol{z}^\top \boldsymbol{\beta}} \Lambda_0(t) \tag{2.2}$$

The cumulative hazard at t is not a probability itself but has a nice relationship with the probability of survival past time t: individuals ranked in *increasing order of survival past time* t are ranked in *decreasing order of cumulative hazard at time* t.

Standard survival software readily yields the nonparametric (Anderson-Gill) estimate of $\Lambda_0(t)$, and hence estimating the cumulative hazard for a subject with time-fixed covariates $\boldsymbol{z}$ is straightforward. On the other hand, if the covariates are time-varying, then the integration becomes more complicated, though it is still relatively easy if $\boldsymbol{z}(t)$ is piecewise constant because the interval of integration can be subdivided as necessary.

The cumulative hazard, however, is always an increasing function and is thus generally higher for subjects who have been with the company longer; we cannot say that the subjects at highest risk of leaving the company at a given time are the ones with the highest cumulative hazard at that time. Rather the highest-risk subjects are the ones whose cumulative hazards have the *steepest increase* at that time. Indeed they are the subjects with highest values of fitted hazard $\lambda(t|\boldsymbol{z})$ (or, equivalently, the derivative of the fitted $\Lambda(t|\boldsymbol{z})$):

$$\lambda(t|\boldsymbol{z}) = \frac{d\Lambda(t|\boldsymbol{z})}{dt} = e^{\boldsymbol{z}^\top \boldsymbol{\beta}} \cdot \frac{d\Lambda_0(t)}{dt}.$$

The nonparametric estimate of $\Lambda_0(t)$, however, is piecewise constant (as is the nonparametric estimate of $\Lambda(t|\boldsymbol{z})$); to obtain the baseline hazard $\lambda_0(t)$ we must either kernel-smooth the estimate of $\Lambda_0(t)$ or impose some parametric assumptions. We opted for the latter for various reasons, including sensitivity of results to choice of kernel and smoothing bandwidth, and convenience of computation.

2.3.1.2 Cox PH model with parametric baseline hazard

The R package flexsurv is highly developed and allows for flexible fitting of survival models. One has many options for the parametric form of baseline hazard function: exponential, Weibull, Gompertz, gamma, generalized gamma, to name a few. We used the flexsurvreg function to fit our Cox PH model, allowing $\lambda_0(t)$ to be the hazard of the generalized gamma distribution whose probability density function is

$$f(x|\alpha,\beta,\kappa) = \frac{\beta}{\Gamma(\kappa)\cdot\alpha}\left(\frac{x}{\alpha}\right)^{k\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}}. \tag{2.3}$$

Note that the hazard function of the generalized gamma is unavailable in closed form. Given the large sample size at our disposal, the numerical optimizer often had trouble converging for other parametric choices of baseline; even though the model fit procedure involving generalized gamma baseline is successful, the process still took much longer than when fitting with the same covariates in the semiparametric Cox PH model.

By fitting the Cox PH model with generalized gamma baseline hazard and no covariates to the training set, we found that the fitted generalized gamma baseline hazard approximates quite well the nonparametric estimate of the baseline hazard (see Figure 2.1). Moreover, when some covariates were added to the model, the overall agreement continued to hold (see Figure 2.2). The non-parametric and parametric estimates, however, become less and less similar as values on the time axis increase; this is to be expected, as there are much fewer households with cumulative tenures as high as 50-60 years than there are households with tenure between 0 and 30 years.

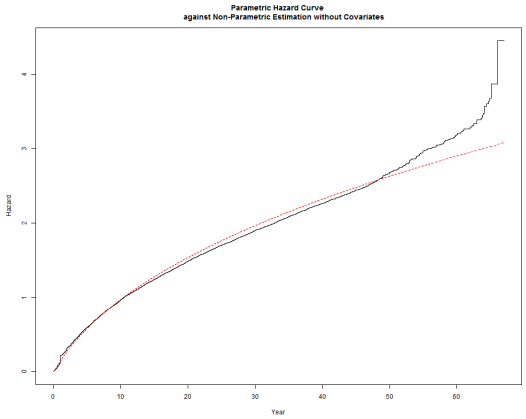

Fig. 2.1 Parametric estimate of baseline cumulative hazard function $\Lambda_0(t)$ (red) compared with the nonparametric approximation of $\Lambda_0(t)$ (black), for intercept-only model.

We must emphasize, however, that since the timescale of the graphs in Figures 2.1 and 2.2 spans approximately 70 years, the discrepancy between the curves at the local level was initially unclear.

Examining Figure 2.3 reveals that, once we zoom in on Figure 2.2 for times near 0, the discrepancy between the fitted parametric baseline and the fitted nonparametric baseline is actually non-negligible. Indeed, upon closer inspection, it is clear that there is a sudden increase in baseline risk at approximately the end of every year; this is to be expected because most

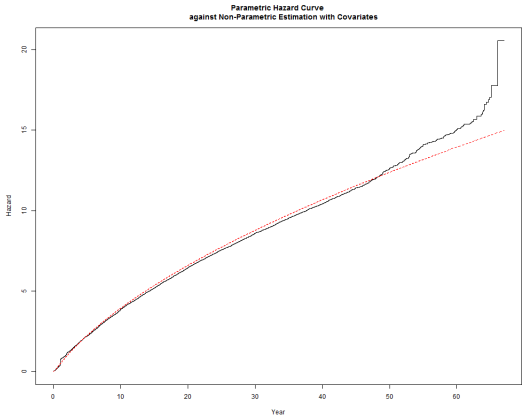

Fig. 2.2 Parametric estimate of baseline cumulative hazard function $\Lambda_0(t)$ (red) compared with the nonparametric approximation of $\Lambda_0(t)$ (black), for model with two covariates.

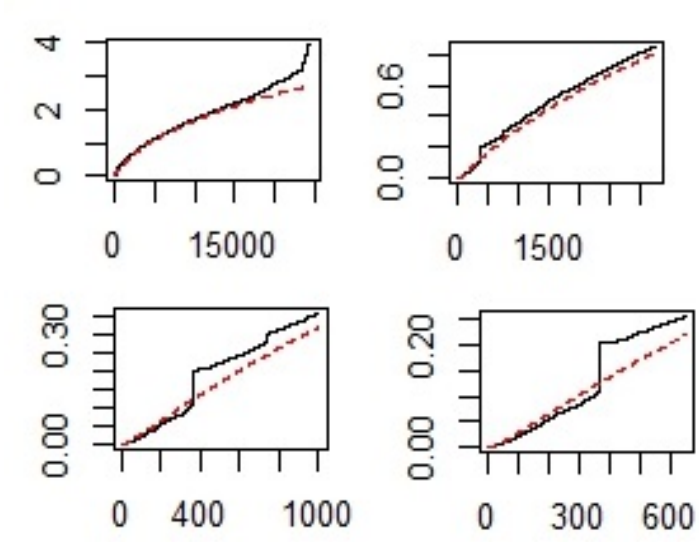

Fig. 2.3 Parametric estimate of baseline cumulative hazard function $\Lambda_0(t)$ (red) compared with the nonparametric approximation of $\Lambda_0(t)$ (black), for different lengths of time axis starting at 0.

clients would decide to leave the company when it is time to renew their policy for the coming year. Another key observation upon closer inspection is that the baseline cumulative hazard seems to be piecewise linear by year; this indicates that the baseline hazard is close to piecewise constant (i.e., takes a – possibly different – constant value each year). In light of this, our recommendation to The Co-operators is that it may even be reasonable to fit the model using only data of the year preceding the date on which they want to rank households' risks of leaving the company.

Nevertheless we will continue using the generalized gamma baseline hazard Cox PH formulation and complete the calendar timescale to illustrate the next (potential) steps.

2.3.2 *Selection of covariates*

The next major issue to address in our model fitting was selecting covariates to be included in the Cox PH model. Our intercept-only model is specified by only three parameters because of the generalized gamma baseline hazard but the possibilities quickly multiply when covariates enter the picture; the training data available during the workshop contained only a small amount of potential covariates compared to the database of The Co-operators, and it was already a time-consuming challenge to try and find a good subset to incorporate into the model. This section will outline preliminary approaches to covariate selection that weren't feasible during the workshop but would be feasible with the company's computational resources.

2.3.2.1 Model selection criteria

One of the classical model selection criteria is the Akaike Information Criterion (AIC), which is defined, for a given dataset and a fitted model, as

$$-2\,l(\theta|\boldsymbol{z}) + 2k, \tag{2.4}$$

where $l(\theta|\boldsymbol{z})$ is the log-likelihood maximized at the parameter estimates of the model and k is the number of parameters determining the model. On the same dataset, if one model has a smaller AIC value than another model, the one with the lowest AIC is preferred; the term $2k$ acts as a penalty (imposed when more parameters are selected), so that parsimonious models with best fit are favoured. Another classical model selection criterion is the Bayesian Information Criterion (BIC), which differs from the AIC only in that the $2k$ term in equation (2.4) is replaced by $k \log n$, where k is the sample size. Other criteria to consider are the Watanabe-Akaike Information Criterion (WAIC)

and the Deviance Information Criterion (DIC). Finally let us mention the Leave One Out Cross Validation (LOOCV) method, which is a standard method in machine learning to select among possible prediction models but is computationally expensive.

In situations where there are many covariates that can be chosen and added to the linear model, automation through R or other software is necessary. If the set of useful covariates can be narrowed down using domain knowledge and prior understanding in such a way that the total number of possibly informative covariates is K (a constant that is not too large), then perhaps one can afford an all-possible-subsets exhaustive search for the best model, where 2^K models could be fit and evaluated according to one of the aforementioned criteria. The Co-operators database, however, records approximately 100 covariates; therefore an exhaustive search is clearly inefficient and even infeasible.

Thus the approach we recommend (and attempted ourselves during the workshop) is backward stepwise and forward stepwise selection.

2.3.2.2 Backward stepwise and forward stepwise model selection

In backward selection, we first fit a full model with all the candidate covariates. By comparing the p-value (from the drop-one t-tests) for each coefficient with a chosen significance level, the least significant covariate is dropped from the full model. This algorithm then repeats the step just described and produces reduced models until the only covariates remaining are all statistically significant. The final model, which cannot be further reduced without compromising the goodness-of-fit (given a threshold of significance), is the model we select.

In a similar fashion, forward selection applies the same rule, but starts from the other end by adding covariates into the model one at a time until all the included covariates are statistically significant.

Backward stepwise and forward stepwise model selection is streamlined with the use of the R function `step`, which supports objects of class `coxph`. Note that backward selection and forward selection are not always guaranteed to end up with the same model.

This brings us to the next objective, which is to determine how best to assess model performance. What would it mean for the company to find a well-fitting model if it cannot make useful predictions about future client behaviour?

2.3.3 *Predictive capability*

Assuming a generalized gamma for the baseline hazard function, we fitted Cox PH models for several different combinations of covariates. It is worth noting that typically in survival analysis, it is essential to verify that the assumptions required by a model hold. For instance, the Cox PH model fitted to our client data assumes that, if we compare the hazard of a household associated with covariates $\boldsymbol{z}_1(t)$ at t days of being a company client with the hazard of another household associated with covariates $\boldsymbol{z}_2(t)$ at t days of being a company client, then the ratio of their hazards at the same age in their (not necessarily aligned) timelines is a value equal to $\exp\left\{(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top \beta\right\}$. This is sometimes termed the assumption of time-independent *coefficients*, since we demand that the hazard ratio be of the form $\exp\left\{(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top \beta\right\}$, not $\exp\left\{(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top \beta(t)\right\}$.

The measures of model validity that are truly relevant to The Co-operators must center, however, on the model's predictive capability. It is reasonable to assume that having an agent call an at-risk-of-leaving household before the end of a policy year decreases the chances of the household leaving the following year, but this costs time and money for the company. Therefore, given that resources are limited, the main use of a statistical model is to maximize the proportion that is spent on highest-risk-of-leaving households.

Unfortunately, verifying the assumption of time-independent coefficients is often performed visually or using p-values and significance levels, and after this work has been carried out one still has to evaluate the model's performance using the test set. Moreover, even if the assumption holds imperfectly, we may still use the model to make useful predictions. Therefore we suppress the habits inculcated into us through our classical survival analysis courses and forego further discussion about how precisely to check the assumption of time-independent coefficients in the Cox PH model. Instead we propose that The Co-operators directly use prediction-oriented measures such as lift curves. Other well-established measures in the industry such as gain curves and variants of ROC curves are not discussed in this report.

In this section the test set finally comes into play, and though it can be used in many different ways, we decided to use it as follows. Let us define a date of interest, say July 1, 2011, at which there is a true "ranking" of households based on the order in which they left afterwards. A model without covariates would predict the instantaneous risk of leaving the company for each household in the test using only the length of time for which the household has been a company client as of July 1, 2011; this would give a ranking of the theoretical urgency of calling each household in the test set at that point in time. Each of the two simple models we use to demonstrate our lift curves has two covariates: model 1 incorporates the effects of the age of the oldest household member and of whether the household has auto insurance on July 1, 2011, and model 2 incorporates the effects of marital status and of whether the household has home insurance on July 1, 2011.

2.3.3.1 Lift curves

We briefly explain here how we arrived at our lift curve. In general a lift curve illustrates how well a model predicts an outcome, over varying numbers of subjects, relative to random guessing.

Let the horizontal axis of the plot be the number of years (x) elapsed since July 1, 2011. On July 1, 2011, N households are clients of the company and thus eligible to leave the company. Then by each time point x, we have a number of households that had actually left the company, say J. Now let us take the top J households predicted by our model to be most at risk of leaving on July 1, 2011. If we spent the resources to call all J of these households predicted by the model, then there would be some number $W \leq J$ of overlap between the J actually-leaving households and the model's J predicted-leaving households. W is the number the model predicted "correctly". On the other hand, if we had called J households in a purely random fashion, then the *expected* number of households chosen "correctly" would be J^2/N (given by the formula for the expected value of a hypergeometric distribution). The "lift" value on the vertical axis plotted against each value on the horizontal axis is given by $\frac{W}{J^2/N}$, the ratio of how many households the model would choose "correctly" to how many households a random selection would choose "correctly", if the same number of households were called.

Figure 2.4 shows that both model 1 and model 2 yield better predictions than random ones for a sample size that is at most the number of households that left within 1.25 years after July 1, 2011; model 2 seems to be a slightly better choice than model 1.

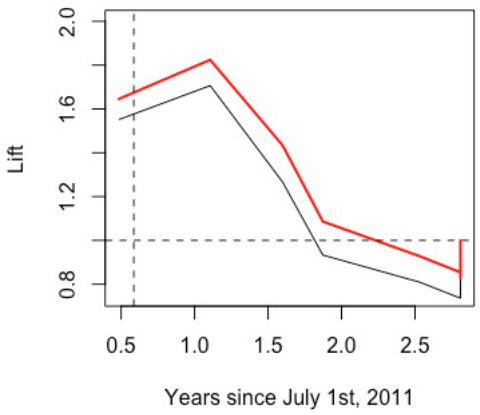

Fig. 2.4 Lift curves for model 1, with covariates "age of oldest household member" and "whether household has auto insurance" (black line), and model 2, with covariates "whether household has home insurance" and "marital status" (red line).

We conclude this section on lift curves by providing an alternative formulation of lift curve. This time let the horizontal axis of the plot be p, a fraction that increases from 0 to 1. Again, on July 1, 2011, N households are clients of the company and thus eligible to leave the company. Then for each value of p, we have the first $L = p \cdot N$ households who actually leave after July 1, 2011. Now let us take the top L households predicted by our model to be most at risk of leaving on July 1, 2011. If we spent the resources to call all L of households predicted by the model, then, as before, there would be some number $W \leq L$ representing the overlap between the L actually-leaving households and the model's L predicted-leaving households. If we had called L households in a purely random fashion, then the *expected* number of households chosen "correctly" would be L^2/N. The "lift" on the vertical axis is $\dfrac{W}{L^2/N}$, the ratio of how many households the model would choose "correctly" to how many households a random method would choose "correctly", if the same amount of resources (for calling L households) were spent in both cases.

The difference here is that whereas Figure 2.4 does not directly show the number of households that would have to be called to hit all the ones predicted to leave by time x after July 1, 2011, Figure 2.4 gives an idea of how well the model would predict if we were willing to call a proportion p of the N clients existing on July 1, 2011.

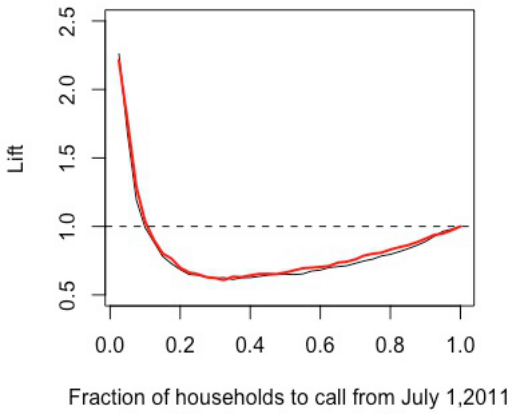

Fig. 2.5 Lift curves for model 1, with covariates "age of oldest household member" and "whether household has auto insurance" (black line), and model 2, with covariates "whether household has home insurance" and "marital status" (red line).

We note that Figure 2.4 and Figure 2.5 (in particular) are purely illustrative and in an actual setting, the test set, the model predictions, and lift

curve computations must be set up carefully so as to avoid errors. An error in the coding for Figure 2.5 is strongly suspected.

2.3.3.2 Model assessment accounting for net profit

Lastly we suggest a more precise measure of performance for candidate models that involves attaching an amount lost for every household called and an amount gained for every household whose risk of leaving was adequately estimated. For each model, we could have a more discerning lift curve by constructing a contingency table (involving true and false positives and negatives) for each value on the horizontal axis and computing the hypothetical net profit of the model's predictions versus that of random calling. We did not have time to pursue this course of action but this is a general idea that would apply to prediction models for many different kinds of problems, not just the current one.

2.4 Cross-sale: When will an existing client add life insurance?

The second problem presented by The Co-operators is more complicated than the first. In the previous client retention problem we merely modelled the *marginal* hazard of a household leaving the company as a function only of its age in the company and its covariates, regardless of its client activity. In this cross-sale problem, we face a challenge that is actually a particular case of the more general problem of modelling *transition intensities* from one state to another, where each state is defined by a certain combination of insurance products. Transition intensity functions are the multi-state analogue of the time-to-single-event hazard function.

We first narrowed the scope of possible states by considering only 3 insurance products: auto, home, and life. Getting a client to commit to buying a life insurance product is much more difficult than getting a client to buy auto or home insurance; The Co-operators would like to have a model that can estimate, for instance, the probability that a household that has had auto and home insurance for X years will purchase life insurance.

2.4.1 *Multi-state modelling*

At any time a client can hypothetically have any combination of insurance products. Setting up a multi-state framework where each state is associated with a particular combination of our 3 particular insurance products entails

considering 8 possible states, where we have defined the state associated with no insurance products as an absorbing state. If we allow a possible transition between every pair of states, however, then this multi-state model immediately becomes too complicated for our purposes; we wanted to avoid simultaneously modelling over 20 transition intensities. Note that in principle this is not actually an issue if there is enough data representing the transitions that are relatively rare (e.g., a transition from having all three products to having none). Hence we allow a simpler set-up: state 1 is associated with product combination A ("having at least one of auto & home insurance"); state 2 is associated with product combination B ("having life insurance AND at least one of auto & home insurance"); and state 3 is "no longer with the company". The allowed states and transitions are illustrated in Figure 2.6.

This simplifies the problem because we need only estimate four different hazard transition probabilities. Implementation can be completed using the mstate library in R, where the trickiest part is setting up the data frame so as to define clearly which state a household is in and which households are *at risk of* (eligible for) particular transitions, in each time interval. In certain cases even quantities such as the probability of transitioning from one product combination to another, given that the customer has been in another state at some time t_0, can be computed.

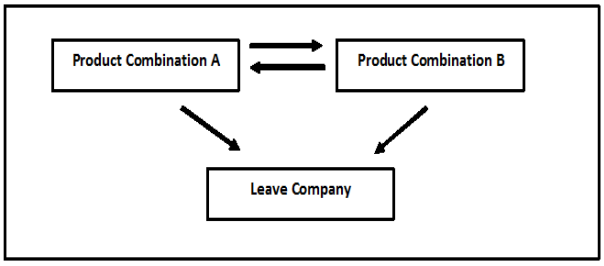

Fig. 2.6 Graphical illustration of a 3-state multi-state model and the allowed transitions.

For demonstration we carried out this procedure, ignoring covariate effects for simplicity, on the training set using flexsurv (which contains multistate). Figure 2.7 illustrates the results: it features on the left the estimate of $\Lambda_{AB}(t)$ (the cumulative hazard of going from product combination A to product combination B) and on the right the estimate of $\Lambda_{AL}(t)$ (the cumulative hazard of going from product combination A to leaving the company). For each graph the nonparametric estimate (black) was plotted with the generalized gamma fit (red). For each transition we could take the derivative of the generalized gamma cumulative hazard function to find the instantaneous risk of undergoing that transition at a particular time.

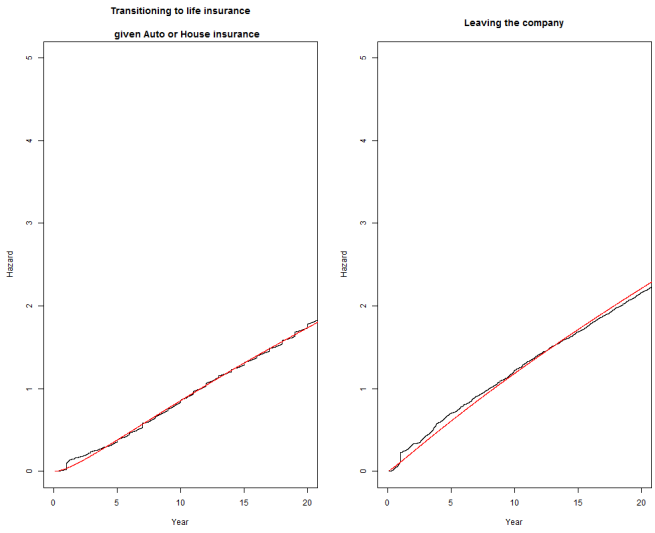

Fig. 2.7 Left: Nonparametric estimate of cumulative hazard function plotted with generalized gamma fitted cumulative hazard function, for the transition from product combination A to product combination B (left) and the transition from product combination A to leaving the company (right).

We conclude this section by noting that it is possible to determine the probability at time $t_0 + x$ of transitioning from product combination A to product combination B *before* or transitioning to leaving the company, given that at a particular time t_0 the individual was in product combination A. More specifically, under the Markov assumption for the possible transition probabilities in Figure 2.6, we have the following:

$$\begin{aligned} &P(\text{ buy life before leaving } | \text{ have auto or home at time } t_0) \\ &= \int_{t_0}^{t_0+x} \exp\left\{ - \int_{t_0}^{t_0+x} \lambda_{AB}(s) + \lambda_{AL}(s) ds \right\} \lambda_{AB}(u)\, du. \end{aligned}$$

We can even incorporate time-fixed covariates such that a household associated with covariates $\boldsymbol{z}$ has a cumulative hazard of transitioning from A to B given by

$$\lambda_{AB}(t) = \lambda_0^{AB}(t) e^{\boldsymbol{z}^\top \beta_{AB}}$$

and cumulative hazard of transitioning from A to leaving given by

$$\lambda_{AL}(t) = \lambda_0^{AL}(t) e^{\boldsymbol{z}^\top \beta_{AL}},$$

where $\lambda_0^{AB}(t)$ and $\lambda_0^{AL}(t)$ are the respective baseline hazards for the two transitions in question and coefficient vectors β_{AB} and β_{AL} respectively capture the effect of $\boldsymbol{z}$ on each transition intensity.

Then, assuming a Cox PH model, we can write the following:

$$\begin{aligned} &P(\text{ buy life before leaving } | \text{ have auto or home at time } t_0) \\ &= \int_{t_0}^{t_0+x} \exp\left\{-e^{\boldsymbol{z}^\top \beta_{AB}} \left[\Lambda_0^{AB}(t_0+x) - \Lambda_0^{AB}(t_0)\right]\right\} \\ &\qquad \cdot \exp\left\{-e^{\boldsymbol{z}^\top \beta_{AL}} \left[\Lambda_0^{AL}(t_0+x) - \Lambda_0^{AL}(t_0)\right]\right\} \\ &\qquad \cdot e^{\boldsymbol{z}^\top \beta_{AB}} \lambda_0^{AB}(u)\, du. \end{aligned}$$

All these quantities are available from R once the multi-state model has been fitted, allowing the above quantity to be computed manually. Alternatively, once an object of class flexsurvreg has been fitted, careful use of the pmatrix.fs function will also us to compute the above quantity for various values of x as well as other transition probabilities.

2.4.2 *Alternative approach: self-exciting marked point processes*

The final candidate that would be eligible for both problems presented by The Co-operators is the class of self-exciting marked point processes. We give a very brief outline here.

If we think of a household's history with the company as a "customer pathway," the pathway is marked by time points each of which represents a purchase, cancellation, claim, or change of covariate as well as the quantity or type of those events. A marked point process lets time points T_i represent, for instance, the purchase times of an insurance product by a customer, while each T_i is associated with a mark, say X_i, representing the type of insurance. A marked point process is given by

$$N(x,t) = \sum_{i \geq 1} I(Xi \leq x, Ti \leq t) = \#\{i : Xi \leq x, Ti \leq t\}, \tag{2.5}$$

where $I(A)$ denotes the indicator function of event A, $t \geq 0$ represents time, and x is a point in the mark space. The distribution of $N(x,t)$ can be determined by its intensity process $\lambda(x,t)$, which represents the rate at which the purchase activities corresponding to mark x occur over time and may depend on the internal history of the point process over the period prior to t (as well as covariates such as age or number of claims and external stimuli such as advertisements and promotional offers, etc.). We choose a suitable semiparametric model for the intensity measure, of the form

$$\lambda(x,t) = \lambda(x, t \mid \theta, \lambda_0(t), N(.,s), 0 \leq s \leq t), \tag{2.6}$$

incorporating a finite-dimensional parameter θ connecting the different components of $\lambda(x,t)$, a specified baseline intensity λ_0, and the internal history $N(\cdot,s)$. By this method, once we find the intensity function for each event, we will have a general model for predicting the probability of events such as leaving the company (thereby addressing the client retention problem) or transitioning from one state to another (thereby addressing the cross-sale problem).

2.5 Future Work and Discussion

We found the problems proposed by The Co-operators to be exciting and challenging; a more thorough review of the literature on extensions of classical survival models might reveal more answers but it is certainly the case that there is current ongoing research on the challenges we tried to tackle during the 4-day workshop.

We note that even though we only applied multi-state modelling to the latter problem (that of cross-sale), it is also applicable to the first problem (general client retention). The other major concern that we did not have time to address is related to the time-varying nature of covariates into our models, or even the number of the covariates. Predicting future client behaviour would involve also predicting the path of values covariates take over time. In this vein the most plausible idea we have is the use of self-exciting marked point processes.

Even though our work during the workshop included only outlines and prototypes of potentially useful models, we hope that the representatives from the R&D team at The Co-operators will have gained insight that they didn't have before and that some of our proposed ideas and methods will amount to fully implemented, actionable models in the future.

References

F. Gao, A. Manatunga, and S. Chen. Non-parametric estimation for baseline hazards function and covariate effects with time-dependent covariates. *Statistics in Medicine*, 2007.

C. Jackson. flexsurv: A platform for parametric survival modelling in r. *Journal of Statistical Software*, 2016.

K. Koppenschmidt and W. Stute. The statistical analysis of self-exciting point processes. *Statistica Sinica*, 2013.

F. Larocque and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 2011.

T. Petersen. Fitting parametric survival models with time-dependent covariates. *Journal of the Royal Statistical Society*, 1986.

J. Yan and J. Huang. Model selection for Cox models with time-varying coefficients. *Biometrics*, 2012.

3
Simulation d'évènements extrêmes en présence de dépendance spatiale
Projet proposé par Desjardins

Nicholas Beck, Bouchra Nasri, Fateh Chebana, Marie-Pier Côté, Juliana Schulz, Jean-François Plante, Martin Durocher, Marie-Hélène Toupin, Jean-François Quessy, Jonathan Jalbert, Véronique Tremblay et Nouredine Daili

3.1 Introduction

Au Canada, les inondations ont des répercussions sociales et économiques importantes pour de nombreux citoyens. Des catastrophes de grande ampleur comme les inondations de Saguenay (en mai 1996) ou de Calgary (en juin 2013) ont causé des dégâts de plusieurs milliards de dollars et ont eu pour effet de sensibiliser la société à l'importance de se protéger du risque de telles pertes économiques. À titre d'illustration, la Figure 3.1 présente un sommaire des plus importantes inondations survenues au Canada. On y voit que les pertes assurées sont parfois bien en dessous des pertes économiques réelles. Cet écart indique qu'il est possible d'améliorer la couverture des biens assurés en cas d'inondation. Desjardins Assurances désire donc proposer des polices d'assurance adaptées qui serviront à protéger davantage les victimes d'8inondations.

Les inondations peuvent être causées par différents phénomènes météorologiques distincts. Par conséquent, il existe des types particuliers d'inondations qui surviennent à des fréquences propres et entraînent des dommages

Nicholas Beck · Marie-Pier Côté · Juliana Schulz · Jonathan Jalbert
McGill University

Bouchra Nasri · Fateh Chebana
INRS-ETE

Jean-François Plante
HEC Montréal

Martin Durocher · Jean-François Quessy
UQTR

Marie-Hélène Toupin · Véronique Tremblay
Université Laval

Nouredine Daili
Université Sétif-1

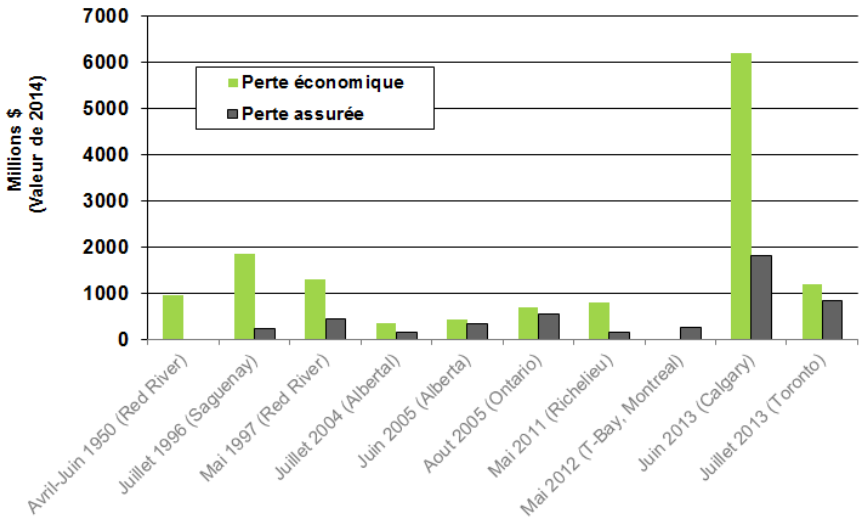

Fig. 3.1 Coût des inondations au Canada

d'ampleurs inégales. Il est ainsi nécessaire d'étudier séparément les risques encourus lors de chacun de ces types d'inondations afin de tenir compte de leurs particularités. À l'heure actuelle, Desjardins Assurances ne propose aucune couverture pour les inondations côtières. Celles-ci peuvent être provoquées par des séismes ou explosions volcaniques sous-marins produisant des séries d'ondes pouvant se transformer en vagues imposantes à l'approche du rivage. Les ondes de tempêtes peuvent également occasionner des niveaux anormalement élevés de la mer. En effet, au centre d'un cyclone, une zone de basse pression se forme et crée ainsi un effet de succion. Lorsqu'une onde de tempête s'échoue sur les côtes à marée haute, elle tire avec elle une masse d'eau suffisante pour entraîner une inondation. Ce mécanisme est illustré à la Figure 3.2.

Desjardins Assurances désire se doter d'un modèle permettant d'évaluer les risques attachés aux inondations côtières, afin d'établir de nouveaux produits d'assurance. À l'heure actuelle Desjardins Assurances possède des modèles hydrodynamiques capables de déterminer les zones inondables (étant donné le niveau de la mer). Ces modèles permettent d'identifier les zones qui seront inondées suite à l'occurrence d'un évènement météorologique extrême. Le problème proposé par Desjardins Assurances est de créer un modèle stochastique permettant de simuler conjointement et de façon réaliste les niveaux extrêmes de la mer en plusieurs endroits. Ces simulations fourniront des données aux modèles hydrodynamiques qui, à leur tour, permettront d'évaluer concrètement les risques d'inondations côtières.

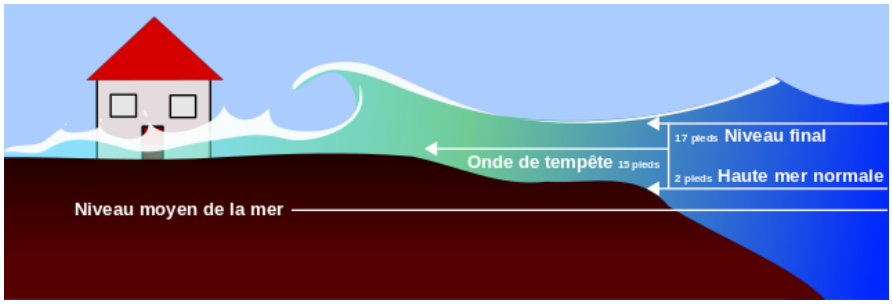

Fig. 3.2 Schéma démontrant le mécanisme des inondations côtières provoquées par les ondes de tempêtes
.

3.2 Données et hypothèses

L'ensemble des données fournies par Desjardins Assurances provient de marégraphes qui mesurent sur une base horaire les niveaux de la mer à différents sites du Canada. Ces sites sont situés principalement sur la côte pacifique, la vallée du Saint-Laurent et la région atlantique. La région est du pays est présentée à la Figure 3.3. À cause des besoins spécifiques de Desjardins Assurances concernant les inondations côtières, l'équipe décide de restreindre son analyse à un groupe de 21 sites situés à l'est de la ville de Québec (y compris Québec elle-même).

Les données horaires représentent un important volume d'information qui peut s'avérer difficile à gérer, d'autant plus que la majorité de l'information contenue dans ces données correspond à des niveaux de la mer normaux. Les phénomènes physiques sous-jacents à ces niveaux normaux ne sont pas les mêmes que les évènements extrêmes donnant lieu à des inondations côtières. Pour la région étudiée, on peut s'attendre à ce qu'un certain nombre de tempêtes se produisent chaque année. Par conséquent, il est raisonnable de présumer que le maximum annuel du niveau de la mer est en général causé par les évènements extrêmes qui nous intéressent. L'équipe s'entend ainsi pour adopter l'approche de la modélisation des maximums annuels du niveau de la mer, qui permettra de filtrer l'information pertinente des marégraphes. Ainsi, à moins d'indication contraire, le terme maximum annuel sera réservé à la variable du niveau de la mer.

La période d'opération des marégraphes va de 1966 à 2015, mais le nombre réel d'années où des observations ont été faites varie d'un site à l'autre. Afin de s'assurer que chaque maximum annuel représente une mesure valide, l'équipe s'est assurée qu'un maximum annuel a été calculé à partir d'enregistrements couvrant au moins 90% de l'année. Les sites de la région d'étude ont ainsi en moyenne 28.5 maximums annuels. Certains marégraphes ayant une série

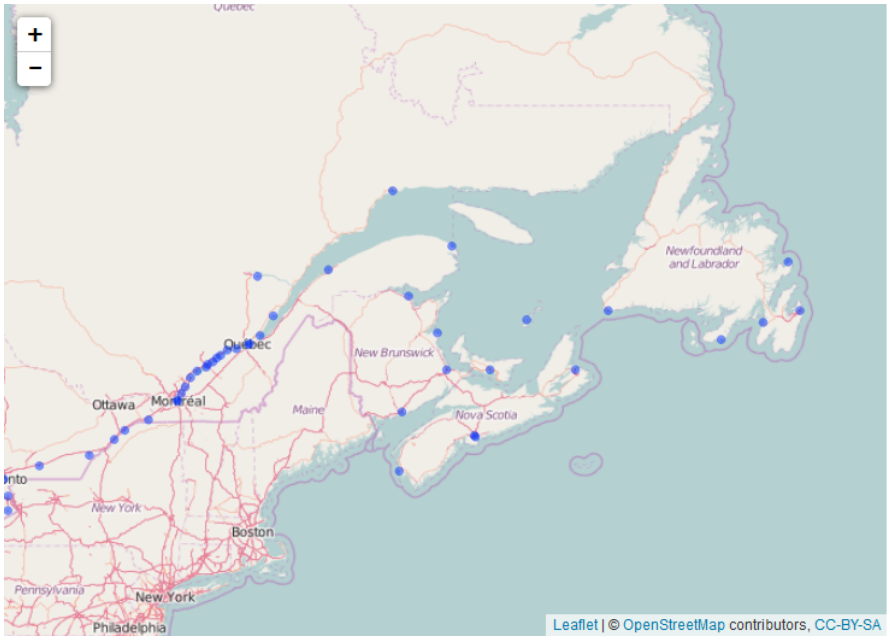

Fig. 3.3 Cartes des marégraphes pour la partie est du Canada
.

complète de mesures permettent d'obtenir des séries chronologiques de 50 années. Toutefois 7 des sites retenus ont moins de 20 années.

L'équipe s'accorde également à faire l'hypothèse que les maximums annuels d'un site surviennent de manière indépendante dans le temps. Cette hypothèse est dans un premier temps vérifiée visuellement par l'examen individuel des séries chronologiques. Un exemple est fourni à la Figure 3.4 pour différents sites. Dans ces graphiques on remarque que les droites de régression ont des pentes faibles et ne semblent pas significatives. De plus, le test de Mann-Kendall est appliqué aux sites ayant plus de 20 ans de données afin de détecter des tendances monotones [McLeod et al., 1991]. Afin de contrôler la probabilité d'obtenir de faux résultats positifs (à cause du fait que plusieurs sites sont testés pour une même région), la procédure FDR (*false discovery rate*) servant à corriger les p-values obtenus par les tests individuels est adoptée [Benjamini and Hochberg, 1995]. Cette procédure, conçue au départ dans un contexte iid, a été aussi validée dans un contexte de dépendance spatiale [Wilks, 2006]. À un niveau de confiance de 95%, l'hypothèse d'une tendance monotone est ainsi rejetée pour l'ensemble des sites.

En plus de l'information fournie par les marégraphes, à partir de la bathymétrie et de la gravité il est également possible d'obtenir pour un site quelconque le niveau normal de la mer à marée haute. On appellera cette dernière marée haute normale. Le panneau de gauche de la Figure 3.5 montre

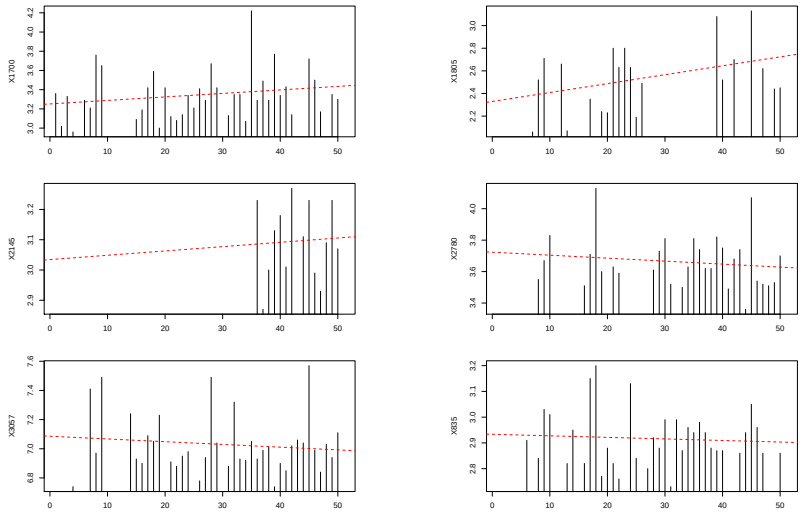

Fig. 3.4 Série de maximums annuels pour certains sites.

les maximums annuels pour la région étudiée, en fonction de la longitude. On constate une tendance décroissante qui se stabilise à la longitude de -65°Ouest. De plus, on observe qu'un site possède des maximums annuels relativement élevés par rapport aux autres sites. Ce site correspond au marégraphe installé à Saint-Jean (Nouveau-Brunswick). Son comportement peut être expliqué par la localisation du site dans la baie de Fundy, reconnue pour ses grandes marées. Dans le panneau de droite de la Figure 3.5, les maximums annuels ont été corrigés en soustrayant la marée haute normale. On remarque alors que tous les maximums annuels corrigés possèdent un centre et une variabilité semblables. En particulier, le site de Saint-Jean (Nouveau-Brunswick) possède un comportement beaucoup moins aberrant.

3.3 Modèle

L'objectif est de pouvoir simuler un champ aléatoire $Z(s)$ de maximums annuels à des sites $s = s_1, \ldots, s_d$. Pour simplifier la notation, on utilisera par la suite $Z_j = Z(s_j)$ pour représenter la variable aléatoire d'un site donné. La loi conjointe du champ aléatoire est donnée par

$$H(z_1, \ldots, z_d) = Pr\{Z_1 < z_1, \ldots, Z_d < z_d\}. \tag{3.1}$$

On cherche alors un modèle pour H que l'on saura simuler et qui sera cohérent avec le comportement des maximums annuels. L'approche par copules

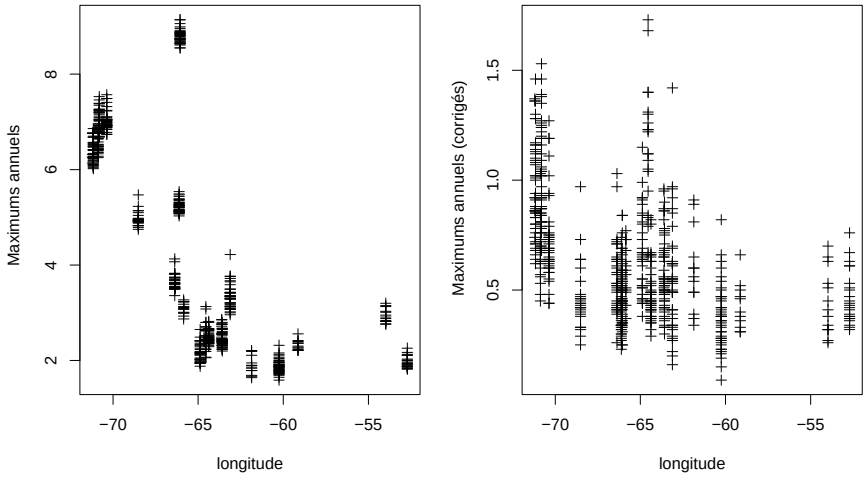

Fig. 3.5 Maximums annuels corrigés pour la marée haute normale en fonction de la longitude.

est attrayante dans le cas présent, puisqu'elle permet de décomposer la loi de H en deux composantes : les lois marginales et la copule. En raison de l'hypothèse d'indépendance temporelle, l'approche par copule permet d'écrire la probabilité jointe du champ aléatoire Z pour une année donnée de la façon suivante :

$$H\left(z_1,\ldots,z_d\right) = C\left\{F_1(z_1),\ldots,F_d(z_d)\right\}, \tag{3.2}$$

où C est la copule et F_j la fonction de répartition (cdf) de la loi marginale au site j.

La théorie des valeurs extrêmes possède un théorème fondamental qui spécifie la loi marginale asymptotique d'un maximum empirique [e.g. Coles, 2001]. Soit $Y_1,\ldots,Y_n$ une suite de variables aléatoires iid et définissons M_n comme $\max\{Y_1,\ldots,Y_n\}$. Le théorème de Fisher-Tippett énonce qu'il existe alors une suite $b_n > 0$ et une suite $a_n \in \mathbb{R}$ telles que

$$Pr\left\{\frac{(M_n - a_n)}{b_n} < z\right\} \to G(z) \quad \text{lorsque } n \to \infty, \tag{3.3}$$

où G est une loi de probabilité non dégénérée. Plus spécifiquement, la loi de probabilité G, dite loi des valeurs extrêmes ou GEV (de l'anglais *generalized extreme values*), a la forme

$$F(z \mid \mu,\sigma,\xi) = \begin{cases} \exp\left\{-\left[1+\xi\frac{(z-\mu)}{\sigma}\right]^{-1/\xi}\right\} & \xi \neq 0,\ 1+\xi\frac{(z-\mu)}{\sigma} > 0 \\ \exp\left\{-\exp\left[\frac{(z-\mu)}{\sigma}\right]\right\} & \xi = 0, z \in \mathbb{R} \end{cases}, \tag{3.4}$$

où μ est un paramètre de position, σ un paramètre d'échelle et ξ un paramètre de forme.

Si l'on fait l'hypothèse que n est suffisamment grand, la loi GEV peut être utilisée comme loi approchée pour des blocs de maximums de taille comparable. En pratique, l'hypothèse que les variables Y_i sont iid est restrictive, parce que les phénomènes météorologiques, comme les tempêtes et les marées, évoluent dans le temps selon des processus physiques précis. Néanmoins, une hypothèse plus réaliste est de supposer que les variables Y_i sont stationnaires, c'est-à-dire que la loi de probabilité jointe ne change pas lorsqu'on se déplace dans le temps. Dans ces conditions il est démontré que le choix de la loi GEV demeure valide, malgré le risque d'une convergence plus lente vers la loi asymptotique [Coles, 2001]. L'équipe décide donc d'utiliser une loi GEV pour les marginales F_j.

La dépendance spatiale est prise en compte en choisissant une structure qui est uniquement fonction de la distance entre deux sites. L'équipe se met d'accord pour utiliser la distance sphérique, qui assimile la planète terre à une sphère de rayon $R = 6378.388$ km. L'équipe fait de plus l'hypothèse que Z est un champ aléatoire isotropique, ce qui signifie que la dépendance entre deux sites ne dépend ni de la position, ni de l'orientation des sites en question.

Idéalement, la loi multivariée de H devrait être une loi multivariée extrême, c'est-à-dire une généralisation du théorème de Fisher-Tippett aux espaces de dimension $d > 1$ [de Haan and Ferreira, 2006]. Les avantages d'utiliser des lois de type extrême pour la modélisation de valeurs extrêmes spatiales sont examinés par Davison et al. [2012]. Dans le cadre du présent travail, pour obtenir une loi multivariée extrême il est nécessaire que les lois marginales soient de type GEV et que la copule C soit de type extrême, c'est-à-dire que pour un vecteur $(u_1, \ldots, u_d) \in [0,1]^d$ la copule respecte

$$C\left(u_1^t, \ldots, u_d^t\right) = C^t\left(u_1, \ldots, u_d\right), \quad \forall t > 0. \tag{3.5}$$

La simulation efficace d'une copule extrême de grande dimension est un problème ouvert et l'équipe propose de se tourner vers une copule non extrême qui puisse offrir également une approximation réaliste.

Une propriété importante pour la modélisation des maximums annuels est celle de la dépendance codale, qui sert à décrire le comportement d'une copule pour les faibles et grandes valeurs de manière spécifique. Ce comportement est important, puisque l'objectif principal est de simuler l'occurrence conjointe de deux évènements proches et d'ampleur suffisante pour provoquer des inondations côtières. Soit (X_1, X_2) un couple de variables aléatoires ayant les cdf F_1 et F_2, respectivement. On définit la dépendance codale à droite comme

$$\lambda_u = \lim_{u \to 1^-} Pr\left(X_1 > F_1^{-1}(u) \mid X_2 > F_2^{-1}(u)\right). \tag{3.6}$$

Une définition semblable existe pour la dépendance codale à gauche.

$$\lambda_l = \lim_{u \to 1^-} Pr\left(X_1 < F_1^{-1}(u) \mid X_2 < F_2^{-1}(u)\right). \tag{3.7}$$

Dans la théorie classique de la géostatistique, la copule normale joue un rôle central. La dépendance codale de celle-ci est toutefois égale à zéro, ce qui s'interprète comme une probabilité nulle qu'un évènement extrême survienne (étant donné qu'un autre évènement extrême s'est produit à proximité du premier). Cette propriété est contraire à ce que l'équipe cherche à reproduire pour les inondations côtières. Par conséquent, à défaut d'utiliser une copule extrême, l'équipe propose de considérer la t-copule, qui a une dépendance codale non nulle.

Nous présentons ici brièvement certains résultats importants sur la t-copule [Demarta and McNeil, 2005]. Soit $\mathbf{X} = (X_1, \ldots, X_d)'$ un vecteur de dimension d, de moyenne μ_X et de matrice de covariance Σ_X. Alors $\mathbf{X} \sim t_d(\nu, \mu, \Lambda)$ suit une loi de Student de densité (pdf) :

$$f_t(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^d|\Lambda|}}\left(1 + \frac{(\mathbf{x}-\mu)'\Lambda^{-1}(\mathbf{x}-\mu)}{\nu}\right)^{-\frac{\nu+d}{2}}, \tag{3.8}$$

où $\nu > 2$ est le degré de liberté de la loi. La t-copule correspond à la structure de dépendance de la loi de Student et appartient à la famille des copules elliptiques. De plus la copule normale correspond au cas limite d'une t-copule lorsque $\nu \to \infty$. La t-copule elle-même peut être évaluée grâce à la formule

$$C_t(u_1, \ldots, u_d) = F_t\left[F_{t,1}^{-1}(u_1), \ldots, F_{t,d}^{-1}(u_d)\right], \tag{3.9}$$

où F_t est la cdf de $t_d(\nu, 0, \Sigma)$, Σ une matrice de corrélation, et $F_{t,j}$ la cdf d'une loi de Student $t_1(\nu, 0, 1)$. La pdf de la t-copule est

$$c_t(u_1, \ldots, u_d) = \frac{f_t\left[F_{t,1}^{-1}(u_1), \ldots, F_{t,d}^{-1}(u_d)\right]}{\prod_{i=1}^{d} f_t\left[F_{t,1}^{-1}(u_i)\right]} \tag{3.10}$$

et peut être utilisée pour l'estimation des paramètres. À des fins de simulation, un vecteur de la loi de Student peut être obtenu à partir de la transformation d'un vecteur de loi normale. En effet, un résultat bien connu montre que

$$\mathbf{X} \overset{d}{=} \mu + \sqrt{W}Y, \tag{3.11}$$

où $Y \sim N_d(0, \Sigma)$ et $\nu/W \sim \chi^2_\nu$ suit une loi du Khi-deux.

Grâce à sa matrice de corrélation, la t-copule peut modéliser chaque paire de sites selon la distance qui les sépare. On dénote $C = C_{\Sigma,\nu}$ la t-copule servant à modéliser les maximums annuels des niveaux de la mer et pour laquelle la matrice de corrélation est définie selon une fonction de lien g_γ telle que

$$(\Sigma)_{j,k} = g_\gamma \left(D(s_j, s_k)/\theta \right). \tag{3.12}$$

Dans les articles scientifiques, il existe une grande variété de fonctions de lien pour lesquelles la matrice de corrélation Σ est correctement définie. Par manque de temps, l'équipe décide de se limiter à la famille des fonctions de puissance exponentielle

$$g_\gamma(\delta/\theta) = \exp\left[-3 \left(\frac{\delta}{\theta} \right)^\gamma \right], \tag{3.13}$$

qui est fréquemment utilisée en pratique [Cressie, 1993]. Le paramètre γ est un paramètre de lissage qui détermine la vitesse à laquelle la dépendance décroît en fonction de la distance δ. Le paramètre de portée θ contrôle pour sa part la distance à partir de laquelle la dépendance entre deux points devient négligeable. Remarquons que lorsque δ est égal à θ, la relation $g_\gamma(1) \approx 0.05$ est satisfaite. Notons que ce résultat ne dépend pas de γ.

La dépendance codale à droite pour la t-copule bivariée est donnée par la formule

$$\lambda_u = 2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{1-\rho}/\sqrt{1+\rho} \right) \tag{3.14}$$

et dépend de ν (son degré de liberté) et ρ (son coefficient de corrélation). De plus, comme la t-copule a une symétrie radiale (c'est-à-dire que la relation $c_t(u_1, \ldots, u_d) = c_t(1-u_1, \ldots, 1-u_d)$ est vérifiée), la dépendance codale à droite et la dépendance codale à gauche sont identiques. Dans un contexte multivarié, le paramètre ρ est l'un des coefficients de la matrice de corrélation Σ et dépend donc uniquement de la distance. Le degré de liberté sert alors à ajuster la dépendance codale. L'effet de la dépendance codale est illustré à la Figure 3.6, qui compare sur une grille régulière les réalisations d'une copule normale (haut) et d'une t-copule (bas) ayant toutes deux la même matrice de corrélation, alors que la t-copule vérifie $\nu = 8$. Nous pouvons constater que dans le cas de la t-copule, les pixels rouges (les points les plus élevés) apparaissent régulièrement en grappe. Ce phénomène crée alors des taches qui sont plus évidentes que dans le cas de la copule normale. Ce comportement décrit alors de manière plus réaliste l'occurrence d'un même évènement pouvant inonder l'ensemble d'une région.

3.4 Solution 1 : l'approche classique

Cette section présente une première méthodologie permettant d'estimer le modèle présenté à la section précédente pour les maximums annuels du niveau de la mer en utilisant des lois GEV et une t-copule. Cette méthodologie suit l'approche classique de la statistique et sera suivie dans la section suivante d'une approche bayésienne.

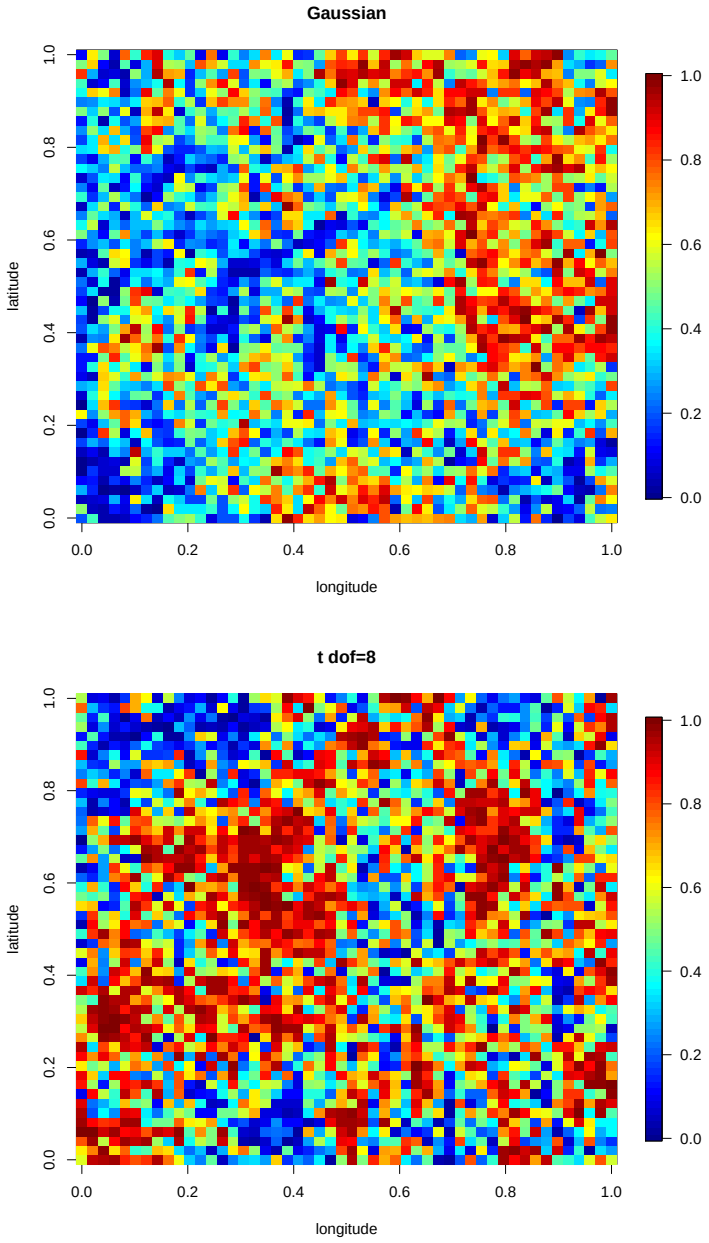

Fig. 3.6 Comparaison de la copule normale (haut) et de la t-copula (bas).

L'équipe décide d'utiliser une approche en deux étapes. Dans une première étape, un modèle pour les lois marginales GEV est proposé et estimé. Ce modèle devra permettre de prédire la loi marginale d'un site quelconque à partir de ses coordonnées et de la marée haute normale. De plus, cette estimation des lois marginales est réalisée sans la connaissance des paramètres de la t-copule. Dans une seconde étape, les paramètres de la t-copule sont estimés conditionnellement à la connaissance des lois marginales.

La classe de modèles adoptée par l'équipe s'écrit sous la forme générale suivante pour chaque site s_j ayant une marée haute normale x_j.

$$\begin{aligned} \mu_j &= \mu_\beta(s_j, x_j) \\ \sigma_j &= \sigma \\ \xi_j &= \xi \end{aligned} \tag{3.15}$$

Les β_j représentent les paramètres spécifiques à la fonction μ_β déterminant le paramètre de position μ_j. L'utilisation de paramètres régionaux constants pour σ et ξ est courante et vise à améliorer l'estimation parfois difficile de ces parmètres à partir des sites individuels, tout en évitant une surparamétrisation. En particulier, le paramètre de forme est crucial puisque'il affecte le comportement de la queue d'une loi et est sensible aux observations les plus extrêmes (peu nombreuses). Le choix d'un paramètre régional permet alors de tenir compte de l'information des sites voisins pour améliorer son estimation.

Afin d'estimer le modèle décrit par (3.15), l'équipe utilise l'approche de vraisemblance composée [Varin et al., 2011]. De manière générale, soit $f(\mathbf{y} \mid \phi)$ une fonction de densité, où $\mathbf{y}$ est un vecteur de dimension d et ϕ un ensemble de paramètres de dimension inconnue. Si $\{A_1, \ldots, A_m\}$ représente un ensemble de sous-évènements d'un modèle, alors la forme générale de la vraisemblance composée est la suivante :

$$L_c(\phi \mid y) = \prod_{k=1}^{m} f(y \in A_k \mid \phi)^{w_k}, \tag{3.16}$$

où les $w_k > 0$ sont des poids. Sous certaines hypothèses généralement vérifiées en pratique, l'estimateur du maximum de vraisemblance composée $\hat{\phi}$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\phi} - \phi) \sim N_d(0, \Lambda_\phi), \tag{3.17}$$

où Λ_ϕ est la matrice d'information de Godambe.

Pour le problème de modélisation des lois GEV, un cas particulier de la vraisemblance composée est la vraisemblance indépendante. Celle-ci correspond à la vraisemblance complète si toutes les observations étaient effectivement indépendantes. Soit $z_{i,j}$ le maximum annuel de l'année $i \in \{1, \ldots, n\}$ au site $j \in \{1, \ldots, d\}$. La vraisemblance indépendante est alors donnée par

$$L_{ind} = \prod_{i=1}^{n} \prod_{j=1}^{d} f(z_{i,j} \mid x_j, \beta, \sigma, \xi), \tag{3.18}$$

où f est la densité d'une loi GEV. Remarquons par ailleurs que la vraisemblance indépendante ne dépend pas du choix de la copule.

Dans un premier temps, un modèle est estimé en considérant une relation quadratique entre le paramètre de position μ_j et la longitude s_j^* :

$$\begin{aligned} \mu_j &= x_j + 4.74 + 0.16 s_j^* + 0.002 (s_j^*)^2 \\ \sigma_j &= 0.20 \\ \xi_j &= 0.07. \end{aligned} \tag{3.19}$$

La Figure 3.7 présente les boxplots des lois GEV estimées et ceux des maximums annuels. Bien que les boxplots suggèrent que le modèle puisse reproduire avec une certaine fidélité le niveau moyen des maximums annuels, ce modèle n'est pas satisfaisant. La relation quadratique du modèle pour le paramètre de position est trop restrictive. Celui-ci fournit une solution où le paramètre d'échelle régionale σ est hautement surestimé et le paramètre de forme ξ est aussi trop élevé.

À des fins de comparaison, des lois GEV ont été ajustées individuellement pour tous les sites ayant plus de 20 années d'observations. La valeur moyenne du paramètre d'échelle est alors 0.14 et celle du paramètre de forme est -0.02. Pour obtenir une solution suffisamment flexible pour le paramètre de position, il est possible d'utiliser un paramètre β_j tel que μ_j égale $\beta_j + x_j$ pour chacun des sites. L'estimation de ce modèle conduit à un paramètre d'échelle régional de 0.14 et un paramètre de forme régional de -0.01, correspondant aux moyennes des paramètres GEV estimés individuellement.

Un problème soulevé par ce dernier modèle est qu'il ne permet pas de prédire les paramètres de la loi marginale pour des sites sans marégraphe. Faute de temps, l'équipe n'a pas été en mesure de proposer un modèle prédictif pour les lois marginales qui soit meilleur que le modèle quadratique. Pour des travaux futurs, l'équipe suggère d'utiliser des bases de fonctions flexibles, comme les fonctions splines [Green and Silverman, 1993], afin de prédire les évènements aux sites sans marégraphe. Une autre option (simple) serait d'interpoler les β_j obtenus par des techniques de krigeage [Cressie, 1993].

La seconde étape de l'approche proposée consiste à estimer la t-copule. Cette estimation se fait conditionnellement à la connaissance des lois marginales. Les cdf $\hat{F}_j$ des lois GEV estimées permettent alors de calculer des pseudo-observations $\hat{u}_{i,j} = \hat{F}_j^{-1}(z_{i,j})$ qui seront modélisées par la t-copule. On rappelle que la t-copule permet de contrôler la dépendance codale par son degré de liberté ν et que sa matrice de corrélation Σ dépend d'une fonction de lien (3.13) avec un paramètre de portée θ et un paramètre de lissage γ.

En vertu de l'indépendance temporelle, la vraisemblance de la t-copule (en supposant que les lois marginales sont connues) peut être écrite comme le produit des pdf de toutes les années. Par contre, en raison des différentes

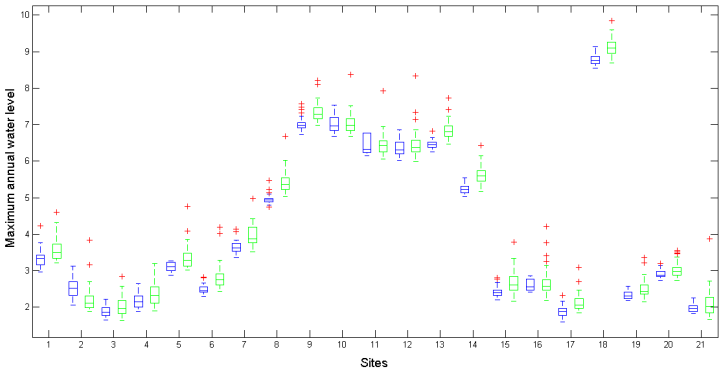

Fig. 3.7 Comparaison des boxplots des lois GEV estimées (vert) et des maximums annuels pour la t-copule.

périodes d'opération des marégraphes, le nombre de maximums annuels disponible n'est pas le même pour chaque année. Nous notons alors d_i le nombre de sites disponibles pour l'année i et la fonction de vraisemblance est donnée par la formule

$$L_{cop}(\nu, \gamma, \theta \mid u_1, \ldots, u_d) = \prod_{i=1}^{n} c_t(u_{i,1}, \ldots, u_{i,d_i} \mid \nu, \gamma, \theta). \tag{3.20}$$

Nous pouvons calculer l'optimum de cette fonction afin d'obtenir une estimation des paramètres de la t-copule.

À des fins illustratives, la t-copule a été estimée à partir des pseudo-observations fournies par le modèle (3.19) (utilisant une relation quadratique pour le paramètre de position). Cette estimation ne peut toutefois pas être considérée comme satisfaisante, car elle dépend de la qualité des lois marginales. À partir de ces résultats préliminaires, on trouve une t-copule de paramètres $\hat{\theta} = 2850$, $\gamma = .37$ et $\nu = 21$. Afin de visualiser l'évolution de la dépendance spatiale, le tau de Kendall empirique a été calculé entre les membres de chaque paire de sites ayant au moins 20 maximums annuels disponibles. La Figure 3.8 présente ainsi l'évolution du tau de Kendall en fonction de la distance. Le tau de Kendall théorique déduit de la t-copule pour un coefficient de corrélation ρ est donné par la relation

$$\tau = \frac{2}{\pi} \arcsin(\rho). \tag{3.21}$$

Dans la Figure 3.8, ces valeurs sont représentées par une ligne passant au milieu du nuage des points, suggérant que l'évolution de la dépendance spatiale pour le modèle estimé correspond aux observations. Toutefois cette valida-

tion par un graphique ne prend pas en compte l'ajustement de la dépendance codale.

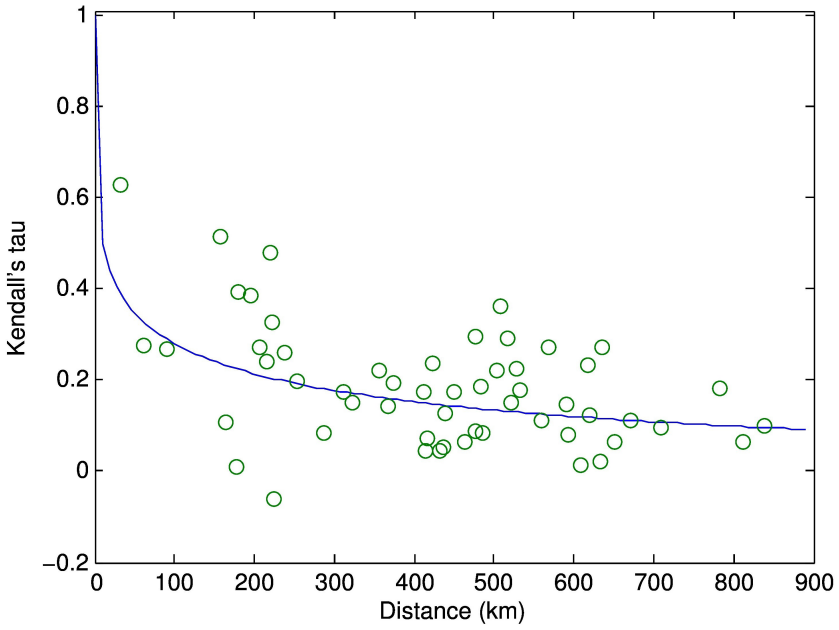

Fig. 3.8 Évolution du tau de Kendall en fonction de la distance pour les paires de sites ayant au moins 20 maximums annuels.

3.5 Solution 2 : l'approche bayésienne

Au lieu d'utiliser une estimation en deux étapes, la deuxième solution estime conjointement le modèle marginal et la copule de dépendance en utilisant un modèle hiérarchique bayésien. L'estimation conjointe permettra d'améliorer l'estimation des paramètres et de caractériser adéquatement l'incertitude. Cependant, un tel modèle requiert des calculs plus difficiles et sa formalisation est plus complexe. Il est à noter que le modèle pour les lois marginales est différent de ce que la solution 1 proposait.

Soit le domaine spatial $\mathcal{D} \subset \mathbb{R}^2$ et soit $s = (s_1, s_2, \ldots, s_d)$ le vecteur contenant la localisation des d points de mesure (tous dans $\mathcal{D}$). Dénotons $M(s_i, t)$ le maximum annuel de l'année t enregistré à la station s_i. Selon la

théorie des valeurs extrêmes, nous pouvons supposer que la loi marginale du maximum annuel du site s_i de l'année t peut être approximée par la loi GEV :

$$M(s_i,t) \approx GEV\left\{\mu(s_i),\sigma(s_i),\xi\right\}.$$

La modélisation bayésienne peut se subdiviser en trois niveaux conditionnels : un niveau pour les maximums annuels, un niveau pour les lois marginales et un niveau pour les lois *a priori*. Tous ces niveaux sont reliés de façon conditionnelle. Le premier niveau modélise la cohérence spatiale des maximums annuels pour une année donnée. L'ajustement des paramètres se fera conjointement grâce à un algorithme Monte-Carlo par chaîne de Markov.

Comme nous l'avons mentionné dans les sections précédentes, les maximums annuels des sites $(s_1,\ldots,s_d)$ pendant l'année t sont cohérents spatialement. Cette dépendance spatiale est modélisée par une t-copule, aussi dénotée $C_{\nu,\theta}$, voir (3.9) et (3.12). La vraisemblance pour une année t peut donc s'écrire de la façon suivante :

$$f_{[M(s,t)|\boldsymbol{\mu}(s),\boldsymbol{\sigma}(s),\boldsymbol{\xi}(s)]} = \frac{\partial^d}{\partial m(s_1,t)\ldots m(s_d,t)} C_{\nu,\theta} \left[G\left\{m(s_1,t)|\mu(s_1),\sigma(s_1),\xi(s_1)\right\},\ldots,G\left\{m(s_d,t)|\mu(s_d),\sigma(s_d),\xi(s_d)\right\}\right], \tag{3.22}$$

où $G(x|\mu,\sigma,\xi)$ dénote la fonction de répartition de la loi GEV (de paramètres (μ,σ,ξ)) évaluée en x. Jusqu'à maintenant, la matrice de corrélation de la copule n'a pas été estimée par manque de temps. Les estimations données par la solution 1 ont été utilisées.

En considérant la suite des maximums annuels indépendants dans le temps, la *pseudo-vraisemblance* peut s'écrire de la façon suivante :

$$f_{[M(s)|\boldsymbol{\mu}(s),\boldsymbol{\sigma}(s),\boldsymbol{\xi}(s)]} = \prod_t \frac{\partial^d}{\partial m(s_1,t)\ldots m(s_d,t)} C_{\nu,\theta} \left[G\left\{m(s_1,t)|\mu(s_1),\sigma(s_1),\xi(s_1)\right\},\ldots,G\left\{m(s_d,t)|\mu(s_d),\sigma(s_d),\xi(s_d)\right\}\right]. \tag{3.23}$$

Pour le deuxième niveau, nous supposons que le paramètre de position de la GEV, $\boldsymbol{\mu}(\cdot)$, est un champ aléatoire lisse sur le domaine $\mathcal{D}$. Nous modélisons la variation spatiale du paramètre de position en utilisant un champ aléatoire gaussien :

$$\boldsymbol{\mu}(s) \sim N_d\left\{X(s)\beta\,,\ \Sigma(s,s)\right\}, \tag{3.24}$$

où $X(s)$ est une matrice de variables explicatives de taille $(m \times d)$ (où m dénote le nombre de variables explicatives), β regroupe les m vecteurs de régression et $\Sigma(s,s)$ est une matrice de covariance de taille $(d \times d)$ qui dépend de la distance. Dans notre cas, la variable explicative utilisée est la marée haute normale. La matrice de covariance est modélisée de la même façon que la matrice de corrélation de la t-copule :

$$\Sigma(s,s) = \tau^2 \Lambda_\rho(s,s), \tag{3.25}$$

où τ^2 est la variance marginale et Λ_ρ une matrice de corrélation définie de la façon suivante :

$$\Lambda_\rho = \{g\left[D(s_i,s_j)/\rho\right] : 1 \leq i \leq d, 1 \leq j \leq d\}\,. \tag{3.26}$$

Plusieurs fonctions de corrélation g peuvent être utilisées : nous avons choisi la fonction exponentielle en (3.13) avec la valeur 1 pour γ. Il serait intéressant de tester d'autres fonctions (dans le cadre de travaux futurs).

Jusqu'à maintenant, nous n'avons pas identifié de variable explicative permettant de modéliser la variation spatiale du paramètre d'échelle de la GEV. Ce paramètre est donc estimé de façon marginale pour chacun des d sites.

Quant au paramètre de forme de la GEV, il est courant dans les articles scientifiques de supposer sa valeur pour un domaine spatial de ce type. Nous utilisons cette simplification dans le cadre de ce travail.

Le dernier niveau concerne les lois *a priori* des paramètres. Nous avons utilisé des lois non informatives pour tous les paramètres :

$$f_{\boldsymbol{\xi}}(\xi) \propto 1; \tag{3.27}$$

$$f_{\boldsymbol{\sigma}}(\sigma) \propto \prod_{i=1}^{d} \frac{1}{\sigma(s_i)}; \tag{3.28}$$

$$f_{\boldsymbol{\beta}}(\beta) \propto 1; \tag{3.29}$$

$$f_{\boldsymbol{\tau^2}}(\tau^2) \propto \frac{1}{\tau^2}; \tag{3.30}$$

$$f_{\boldsymbol{\rho}}(\rho) = U\left\{0, \max_{s_i,s_j \in \mathcal{D}} D(s_i,s_j)\right\}, \tag{3.31}$$

où U est une loi uniforme.

Nous présentons maintenant des résultats sur l'ajustement de la loi GEV aux marégraphes. La figure 3.9 illustre l'ajustement de la loi GEV aux maximums annuels du marégraphe de l'Île d'Orléans. L'intervalle de crédibilité à 95% du modèle hiérarchique bayésien est tracé en tirets tandis que l'intervalle de crédibilité d'un modèle non spatial est tracé en pointillés. Nous constatons que l'utilisation de la cohérence spatiale des maximums permet de réduire la variance d'estimation des niveaux. La Figure 3.10 illustre l'ajustement de la loi GEV au marégraphe du Vieux-Québec. Dans ce cas, le nombre de maximums annuels (5) était insuffisant pour ajuster un modèle marginal : les algorithmes ne convergeaient pas. Toutefois, en considérant la cohérence, il a été possible d'obtenir un ajustement de la loi GEV pour ce site. L'information, bien que limitée, a pu être prise en compte dans le modèle global.

Nous n'avons pas eu le temps de produire des simulations des niveaux de la mer comme nous l'avons fait pour la solution 1. Afin d'y arriver, il

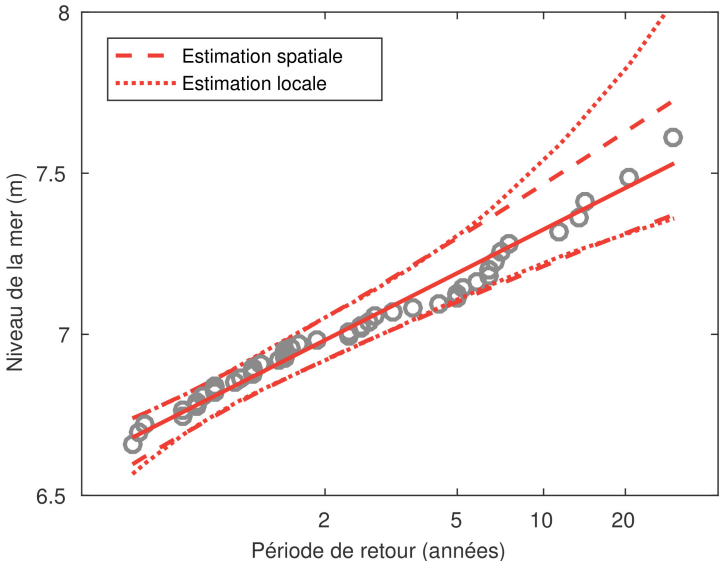

Fig. 3.9 Marégraphe de l'Île d'Orléans

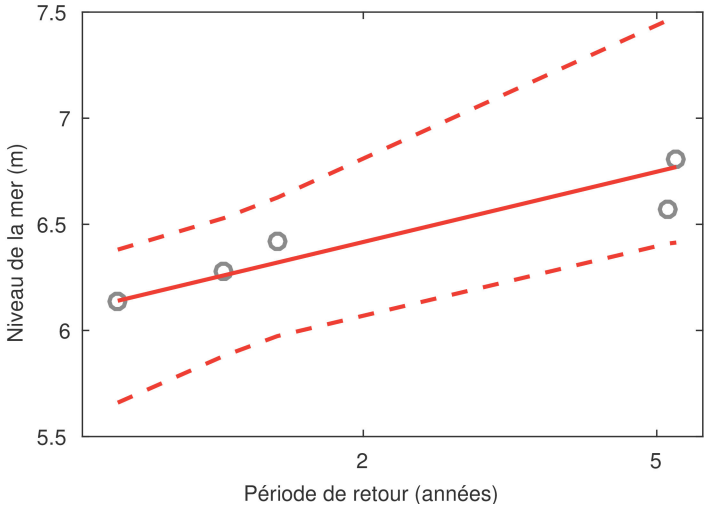

Fig. 3.10 Marégraphe du Vieux-Québec

faudrait tout d'abord introduire l'ajustement de la matrice de corrélation de la t-copule dans le modèle hiérarchique bayésien.

3.6 Conclusion

Un modèle statistique a été proposé pour simuler des maximums annuels des niveaux de la mer pour la région atlantique du Canada. Les simulations de ce modèle serviront d'entrée à un modèle hydrodynamique permettant d'identifier des zones inondables réalistes et ainsi évaluer les risques d'inondation côtière pour l'est du pays. Le modèle proposé par l'équipe suit l'approche par copule utilisant des lois marginales GEV et une t-copule. La t-copule n'est pas une copule extrême mais représente une option intéressante qui est facile à simuler et peut reproduire la dépendance codale dans une certaine mesure. Faute de temps, l'équipe n'a pas été en mesure de produire des résultats finaux : elle a quand même illustré la méthodologie proposée à partir de résultats préliminaires.

Dans de futurs travaux, plus d'efforts pourraient être consacrés à la modélisation des lois marginales. En pratique, l'étape de modélisation des lois marginales est souvent précédée d'une étape de délinéation qui vise à créer des régions homogènes d'un point de vue hydrologique (voir par exemple [Hosking and Wallis, 1997]). Dans le cadre d'une étude pancanadienne, il est raisonnable de supposer que la zone pacifique et la zone atlantique sont des régions homogènes distinctes. L'équipe recommande toutefois de vérifier si ces zones pourraient être subdivisées en sous-régions pertinentes dans le but d'étudier les inondations côtières. Cette étape de délinéation aurait des conséquences directes sur la modélisation des lois marginales GEV, puisque un paramètre d'échelle et un paramètre de forme seraient estimés pour chacune de ces régions homogènes. Par ailleurs, l'équipe a utilisé la ville de Québec comme critère pour définir la région d'étude. Ce choix a été fait en accord avec les besoins de Desjardins Assurance, mais est peut-être arbitraire d'un point de vue hydrologique. Il faudrait asseoir ce choix sur des bases hydrologiques plus solides.

Bien que les analyses préliminaires effectuées par l'équipe n'aient pas conduit à la détection de tendances dans les séries chronologiques, l'étude de la stationnarité demeure un sujet important, en particulier lorsque l'on pense à l'augmentation prévue du niveau de la mer en raison des changements climatiques.

Finalement, mentionnons qu'une autre option que l'approche par copules (telle que proposée par l'équipe) est celle des processus max-stables, correspondant à la généralisation directe de la théorie des valeurs extrêmes pour les données spatiales [Davison et al., 2012]. La simulation de processus max-stables est un sujet en plein développement. Les avancées dans ce domaine ont le potentiel de conduire à des méthodes de simulation de champs aléatoires

extrêmes qui, en théorie, sauront reproduire plus fidèlement les évènements extrêmes qui causent les inondations côtières.

Références

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) :289–300, 1995. ISSN 0035-9246. URL http://www.jstor.org/stable/2346101.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Ser. Statist. Springer, London, 2001.

Noel Cressie. *Statistics for Spatial Data*. Wiley Ser. Probab. Math. Statist. Appl. Probab. Statist. Wiley, New York, 1993.

A. C. Davison, S. A. Padoan, and M. Ribatet. Statistical modeling of spatial extremes. *Statist. Sci.*, 27(2) :161–186, 2012. ISSN 0883-4237. doi : 10.1214/11-STS376.

L. de Haan and A. Ferreira. *Extreme Value Theory : An Introduction*. Springer Ser. Oper. Res. Financ. Eng. Springer, New York, 2006.

Stefano Demarta and Alexander J. McNeil. The t copula and related copulas. *Int. Stat. Rev.*, 73(1) :111–129, 2005. ISSN 1751-5823. doi : 10.1111/j.1751-5823.2005.tb00254.x.

Peter J. Green and Bernard W. Silverman. *Nonparametric Regression and Generalized Linear Models*, volume 58 of *Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, 1993.

J. R. M. Hosking and J. R. Wallis. *Regional Frequency Analysis*. Cambridge University Press, Cambridge, 1997.

A. Ian McLeod, Keith W. Hipel, and Byron A. Bodo. Trend analysis methodology for water quality time series. *Environmetrics*, 2(2) :169–200, 1991. ISSN 1099-095X. doi : 10.1002/env.3770020205.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statist. Sinica*, 21(1) :5–42, 2011. URL http://www.jstor.org/stable/24309261.

D. S. Wilks. On "field significance" and the false discovery rate. *Journal of Applied Meteorology and Climatology*, 45(9) :1181–1189, 2006. ISSN 1558-8424. doi : 10.1175/JAM2404.1.

4
VaR and Low Interest Rates
Project Submitted by the Caisse de dépôt et placement du Québec

Zichun Ye and Louis G. Doray

4.1 Introduction

Value at Risk (VaR) is a measure of the risk of investments. It estimates how much a set of investments might lose, given normal market conditions, in a set time period such as a day. The VaR is typically used by firms and regulators in the financial industry to gauge the amount of assets needed to cover possible losses. For a portfolio of financial assets, the VaR of the portfolio is defined to be the minimal potential loss in value of the portfolio, for a given level of confidence over a certain investment horizon, namely

$$\mathrm{VaR}_{t,p}(P) = \inf\{\ l\ |\mathbb{P}(v_t(P) \geq l) \leq 1 - p\}. \tag{4.1}$$

Here t is the time horizon and p is the confidence interval; P stands for the portfolio and $v_t(P)$ denotes the value of the portfolio at time t. We refer the reader to Wipplinger [2007] for a discussion of the VaR.

Historical simulation is a popular way of estimating the VaR. It involves using past data as a guide to what will happen in the future. Suppose that we want to calculate the VaR for a portfolio using a one-day time horizon, a 99% confidence level, and 501 days of data. (The time horizon and confidence level are those typically used for a market risk the VaR calculation; 501 is a popular choice for the number of days of data used, because, as we shall see, it leads to 500 scenarios being created.) The first step is to identify the market variables affecting the portfolio. These will typically be interest rates, equity prices, commodity prices, and so on. All prices are measured in the

Zichun Ye
The University of British Columbia

Louis G. Doray
Université de Montréal

domestic currency. For example, one market variable for a German bank is likely to be the S&P 500 measured in euros.

Data are collected on movements in the market variables over the most recent 501 days. This provides 500 alternative scenarios for what can happen between today and tomorrow. Denote the first day for which we have data as Day 0, the second day as Day 1, and so on. Scenario 1 is where the percentage changes in the values of all variables are the same as they were between Day 0 and Day 1, Scenario 2 is where they are the same as between Day 1 and Day 2, and so on. For each scenario, the dollar change in the value of the portfolio between today and tomorrow is calculated. This defines a probability distribution for daily loss (gains are negative losses) in the value of our portfolio. The 99th percentile of the distribution can be estimated as the fifth highest loss. The estimate of the VaR is the loss when we are at this 99th percentile point. We are 99% certain that we will not take a loss greater than the VaR estimate if the changes in market variables in the last 501 days are representative of what will happen between today and tomorrow. We refer the reader to Hull [2006] for more details about the historical simulation method. To express the approach algebraically, define v_i as the value of a market variable on Day i and suppose that today is Day n. The ith scenario in the historical simulation approach assumes that the value taken by the market variable tomorrow will be

$$\text{Value under } i\text{th scenario} = v_n + v_i - v_{i-1}. \tag{4.2}$$

The principal parameters defining the (historical) VaR are:

- the size of the historical period used for the estimation;
- the measurement frequency;
- the confidence level;
- the investment horizon.

The choice of parameters generally depends on the desired use for the VaR. As far as the size of the historical period is concerned, if the historical period is long, there will be many more distinct events, but this will also lead to a much more stable measure over time. By contrast, if we use a very short time horizon, the VaR will be a much better reflection of current market conditions. Another important parameter is the measurement frequency. Using a high measurement frequency puts a lot of emphasis on short-term variations, which may not be in line with the objectives of the investor. Ideally the measurement frequency should coincide with the investment horizon. The last parameter is the confidence level. The higher the confidence level, the more emphasis we put on the tail of the distribution and extreme events.

In practice, we always address the question of scaling the measurement frequency. For most of trading assets in the market, like stocks and bonds, the prices are changing all the time. As a result high-frequency measurement of data is possible and it is easy to calculate the VaR with a short time horizon. On the other hand, if we try to calculate the long-time horizon VaR

directly from data with low measurement frequency, the short-term fluctuations of the price are ignored. As a result, to calculate the VaR with long time horizon, such as one year or more, it is better and more widely accepted to calculate the VaR with short time horizon, such as one week or less, and then scale it to the required time horizon. In practice, the square-root-of-time rule is widely used in the scaling problem. The square-root-of-time rule is commonly assumed when financial risk is time-aggregated: high-frequency risk estimates are scaled to a lower frequency T by the multiplication of $\sqrt{T}$. This is usually correct if the mean of daily returns equals zero and daily returns are normally and independently distributed. Explicitly, if the square-root-of-time rule holds for the VaR of the portfolio P with respect to confidence interval l, we have the following relation:

$$\mathrm{VaR}_{aT,l}(P) = \sqrt{a}\,\mathrm{VaR}_{T,l}(P). \tag{4.3}$$

We refer the reader to Danielsson and Zigrand [2006] for a discussion of the square-root-of-time rule.

In this report, we describe the results obtained at the Seventh Montreal Industrial Problem Solving Workshop and improve two aspects of the estimation of the VaR for bond portfolios. The first aspect is the scenario generation in Section 4.4. Instead of traditional historical simulation (which applies historical variation directly to the recent price), we rescale the variation by taking the interest rate level into account. The second aspect is the annualization in Section 4.5. Starting from the square-root-of-time rule, we develop a more general rule for annualization by estimating the Hurst coefficient. The data are the values of the interest rate from January 3, 2000 to April 29, 2016 for 20 maturities. Both the problem and the data were provided by *Caisse de Dépôt et Placement du Québec*, hereafter called the Caisse.

4.2 Review of the Problem

Now we introduce the method used by the Caisse for calculating the VaR for bond portfolios. The Caisse uses the historical simulation method in order to estimate the VaR. We now discuss the parameter settings. The size of the historical period is set to be 10 years and the measurement frequency is weekly. We use a historical period consisting of the last 10 years of weekly data with daily overlaps for all risk factors present in the portfolio. This gives us enough data with which to measure the VaR and only introduces a light measurement bias. Using the current level of each risk factor, we apply the historical variations to create the historical scenarios. In the case of bonds, the interest rate is the only risk factor we take into account. For interest rates with maturity m at time t_0, we create historical scenarios in the following way:

$$\tilde{r}_{m,t_0} = r_{m,t_0} + r_{m,t} - r_{m,t-w}. \tag{4.4}$$

Here t is a time point randomly selected from the past and w is the length of one week. That is, we apply the historical variation $r_{m,t} - r_{m,t-w}$ to the current level of the interest rate r_{m,t_0} in order to simulate $\tilde{r}_{m,t_0}$ one week later in the scenarios. The assumption underlying this approach is that variations are independent of the level of interest rates. Then we apply these scenarios to our current portfolio holdings in order to simulate potential variations of the portfolio. In order to simulate the value of a bond, we recompute the value of the bond under the simulated interest rates, using the formula below.

$$\tilde{V}_{t_0,t} = \sum_{k=1}^{n} \frac{C}{(1+\tilde{r}_{k,t_0,t})^k} + \frac{P}{(1+\tilde{r}_{n,t_0,t})^n} \tag{4.5}$$

In this formula C denotes the coupon and P the principal. We then calculate a weekly VaR, and subsequently annualize this number using the square-root-of-time rule.

To illustrate, we consider a short position in a Canadian bond, with a coupon of 4% maturing in June 2017. Its weekly VaR equals -0.31% and its annualized VaR -2.2%. Bond managers are very familiar with the interest rate sensitivities of their portfolios, and so they like to express the VaR in terms of sensitivities. Here is the formula.

$$\text{Duration} \times \text{Interest rate movement} \approx \text{VaR} \tag{4.6}$$

As the duration of this bond equals one, using this approximation we see that the implicit interest rate movement equals -2.2%. Combined with the fact that current one-year rates are at 0.5%, this implies that we would need the one-year rate to drop to -1.7% for this VaR to be realized!

There are two aspects of the VaR calculation that cause this problem. The first one is scenario generation. We simulate interest movements by a simple method: adding the historical variation to the current value. The distribution of interest rate variations, however, seems to depend on the interest rate level. We refer the reader to Figure 4.1 for the distribution of interest rate variations.[1] The graph shows that as the level of interest rate decreases, the variation of the interest rate becomes smaller. The formula for generating a scenario (4.4), however, violates this observation.

The second aspect is the annualization. In order to transform the VaR from the time scale of the measurement frequency to that of the investment horizon, we need our observations to be i.i.d. and to follow a normal distribution. These assumptions, however, are not met in our bond case. We refer the reader to Figure 4.2 for the distribution of simulated weekly profits and losses

[1] Source: Presentation at the Seventh Montreal Industrial Problem Solving Workshop by the Caisse.

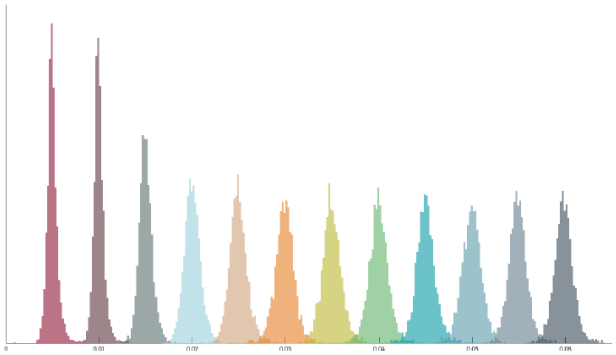

Fig. 4.1 The distribution of interest rate variations.

for the bonds[2]. The distribution shows a negative skewness and a positive excess kurtosis[3]. As a left-skewed and leptokurtic distribution, it does not fit the normal distribution (dash line) well.

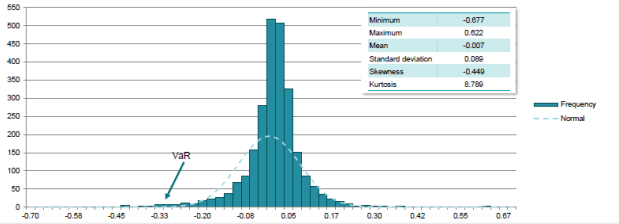

Fig. 4.2 The distribution of simulated weekly P&L.

4.3 Exploratory Analysis of Data

The Caisse provided us with the data of the interest rates from January 3, 2000 to April 29, 2016. There are 20 maturities and 4260 observations for each maturity. Selected maturities and corresponding time horizons are displayed in Table 1.

To study the interest rates as a time series, we take the rates with 1-year maturity as an example. Figure 4.3 is the plot of the rates with 1-year maturity from January 3, 2000 to April 29, 2016.

[2] Source: Presentation at the Seventh Montreal Industrial Problem Solving Workshop.

[3] Excess kurtosis = Kurtosis - 3.

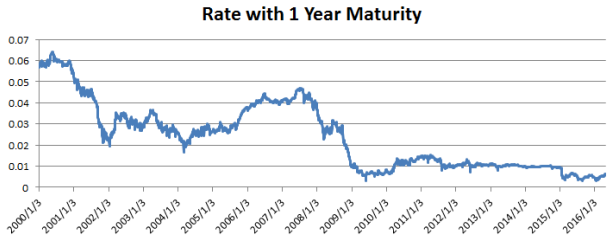

Fig. 4.3 The distribution of simulated weekly P&L.

As we can observe on the graph, the process of the rate with one-year maturity is

1. a non-stationary process: the mean, variance, and covariance change over time (the first difference of the process, however, is stationary);
2. a nonlinear process, because of the existence of different states of the world (or regimes);
3. a process with volatility clustering: "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes."

These observations motivate us to find some process other than Brownian motion or Geometric Brownian motion to model the process of interest rates.

4.4 Scenario Generation

4.4.1 *Model: Short-Term Variation*

As the level of interest rates decreases, the variation of the interest rate becomes smaller (see Figure 4.1). We now introduce a model having this property, the CEV model.

Table 4.1 Selected maturities

number of days	Maturity
1	Day (d)
7	Week (w)
30	Month (m)
90	Quarter (q)
180	Semi-annual (s)
365	Annual (a)
700	2 Years (2y)
...	...
10955	30 Years (30y)

In mathematical finance, the CEV, or constant-elasticity-of-variance model, is a stochastic volatility model attempting to capture stochastic volatility and the leverage effect Cox [1975]. The model is widely used by practitioners in the financial industry, especially for modelling equities and commodities. It was developed by John Cox in 1975. The CEV model describes a process that evolves according to the following stochastic differential equation:

$$dS_t = \mu S_t dt + \sigma S_t^{\gamma} dW_t. \tag{4.7}$$

Here W_t is the standard Brownian motion. The constant parameters σ and γ satisfy the conditions $\sigma \geq 0$ and $\gamma \geq 0$, respectively. In the CEV process the local volatility is a deterministic function of the underlying asset given by $\sigma(S,t) = \sigma S^{\gamma-1}$. The parameter γ controls the relationship between volatility and price and is the central feature of the model.

Now we assume the interest rate r_t follows the CEV process without a drift term, namely $\mu = 0$. Then the SDE (4.7) becomes

$$dr_t = \sigma r_t^{\gamma} dW_t. \tag{4.8}$$

In a small time interval $[t, t+\Delta t]$, the variation of the interest rate dr_t is small compared with the interest rate r_t. Assuming the diffusion coefficient σr_t^{γ} is constant through the time interval, we obtain

$$\begin{aligned} r_{t+\Delta t} - r_t &= \int_t^{t+\Delta t} \sigma r_t^{\gamma} dW_t \\ &= \sigma r_t^{\gamma} \int_t^{t+\Delta t} dW_t \\ &= \sigma r_t^{\gamma} (W_{t+\Delta t} - W_t). \end{aligned} \tag{4.9}$$

By the property of the standard Brownian motion and scaling of σ, we have the following discrete version of the CEV model:

$$\Delta r_t = \sigma r_t^{\gamma} Z, \tag{4.10}$$

where Z denotes the standard normal distribution. Therefore, given the interest rate level r_t, the change in interest rate Δr_t follows a normal distribution with variance depending on the current level, that is

$$\Delta r_t \sim N(0, \sigma^2 r_t^{2\gamma}). \tag{4.11}$$

Notice that when r_t decreases, the variance of Δr_t (i.e., $\sigma^2 r_t^{2\gamma}$) also decreases, which fits the data displayed in Figure 4.1.

Remark 1. Actually the CEV model without a drift term, namely the SDE (4.8), is a special case of the SABR model (see Hagan et al. [2004]). In mathematical finance, the SABR model is a stochastic volatility model attempting to capture the volatility smile in derivatives markets. The name SABR stands

for "stochastic alpha, beta, rho," referring to the parameters of the model. The SABR model is widely used by practitioners in the financial industry, especially in the interest rate derivative markets. The SABR model describes a single forward F, such as a LIBOR forward rate, a forward swap rate, or a forward stock price. The volatility of the forward F is described by a parameter σ. SABR is a dynamic model in which both F and σ are represented by stochastic state variables whose time evolution is given by the following system of stochastic differential equations:

$$dF_t = \sigma_t F_t^{\beta} dW_t, \tag{4.12}$$

$$d\sigma_t = \alpha \sigma_t \, dZ_t, \tag{4.13}$$

with the prescribed time-zero (currently observed) values F_0 and σ_0. In the above system W_t and Z_t are two correlated Wiener processes with correlation coefficient $-1 < \rho < 1$:

$$dW_t dZ_t = \rho dt. \tag{4.14}$$

The constant parameters β and α satisfy the conditions $0 \leq \beta \leq 1$ and $\alpha \geq 0$, respectively.

The above dynamics is a stochastic version of the CEV model with the skewness parameter β: in fact, it reduces to the CEV model if $\alpha = 0$ holds. The parameter α is often referred to as the vol of vol, and its meaning is that of the log-normal volatility of the volatility parameter σ.

4.4.2 *Scenario Generation*

Starting with the discrete version of the CEV model (4.10), we now derive the formula of scenario generation. The basic idea is to scale the historical variation by the corresponding interest rate level. In the following derivation, we use two samples r_t and r_{t-w} chosen among historical interest rates to generate a scenario. Here r_t denotes the interest rate at time point t and r_{t-w} the interest rate at time point $t-w$, that is, one week before t. Also we denote r_{t_0} the current interest rate level and $\tilde{r}_{t_0}$ the interest rate level generated from the scenario for the time point $t_0 + w$. The following derivation holds for all maturities: therefore we omit the parameter m.

By (4.11), given the current interest rate level r_{t_0}, the variation in the next one-week period (i.e., $\Delta r_{t_0} = \tilde{r}_{t_0} - r_{t_0}$) follows a normal distribution,

$$\tilde{r}_{t_0} - r_{t_0} \sim N(0, \sigma^2 r_{t_0}^{2\gamma}). \tag{4.15}$$

Thus by the property of a normal distribution, we have

$$\frac{\tilde{r}_{t_0} - r_{t_0}}{r_{t_0}^{\gamma}} \sim N(0, \sigma^2). \tag{4.16}$$

Similarly, for the historical data, we have

$$\frac{r_t - r_{t-w}}{r_{t-w}^{\gamma}} \sim N(0, \sigma^2). \tag{4.17}$$

From (4.16) and (4.17) we conclude, after scaling by the corresponding interest rate level, that the variations have the same distribution (d denoting equality in distribution).

$$\frac{\tilde{r}_{t_0} - r_{t_0}}{r_{t_0}^{\gamma}} \overset{d}{=} \frac{r_t - r_{t-w}}{r_{t-w}^{\gamma}} \tag{4.18}$$

Furthermore the variation has the same distribution as the rescaled historical variation:

$$\tilde{r}_{t_0} - r_{t_0} \overset{d}{=} \frac{r_{t_0}^{\gamma}}{r_{t-w}^{\gamma}}(r_t - r_{t-w}). \tag{4.19}$$

Thus we can scale the historical variation $r_t - r_{t-w}$ by $r_{t_0}^{\gamma}/r_{t-w}^{\gamma}$ and apply it to the current level of the interest rate r_{t_0}, so that our scenario generation formula based on the CEV model becomes

$$\tilde{r}_{t_0} = r_{t_0} + \frac{r_{t_0}^{\gamma}}{r_{t-w}^{\gamma}}(r_t - r_{t-w}). \tag{4.20}$$

Before moving to the next step, we will first have a look at two special cases of the equation (4.20). When $\gamma = 0$ holds, then (4.20) becomes

$$\tilde{r}_{t_0} = r_{t_0} + (r_t - r_{t-w}), \tag{4.21}$$

which is identical to the original scenario-generating function (4.4) (or the additive model). Under this assumption SDE (4.8) becomes

$$dr_t = \sigma dW_t. \tag{4.22}$$

Thus the corresponding CEV process for the interest rate r_t is the Brownian motion. When $\gamma = 1$ holds, then (4.20) becomes

$$\tilde{r}_{t_0} = \frac{r_{t_0}}{r_{t-w}} r_{t-w}, \tag{4.23}$$

which is the multiplicative model. Under this assumption SDE (4.8) becomes

$$dr_t = \sigma r_t dW_t. \tag{4.24}$$

Thus the corresponding CEV process for the interest rate r_t is the Geometric Brownian motion. Hence our model is a generalization of two well-known processes with an extra degree of freedom.

4.4.3 *Parameter Estimation*

In the scenario-generating function (4.20), the only unknown parameter is γ, which determines the power of scaling. Now we introduce the method of maximum likelihood estimation for estimating γ.

From (4.11) we conclude that the following holds, given the current interest rate level r_t.

$$\Delta r_t \sim N(0, \sigma^2 r_t^{2\gamma}) \tag{4.25}$$

The contribution to the likelihood function for one observation is given by

$$L(\Delta r_t | r_t, \sigma^2, \gamma) = \left(2\pi\sigma^2 r_t^{2\gamma}\right)^{-1/2} \exp\left(-\frac{\Delta r_t^2}{2\sigma^2 r_t^{2\gamma}}\right). \tag{4.26}$$

Thus for a sequence of independent observations $\{(r_{t_i}, \Delta r_{t_i})\}_{i=1,2,\ldots,N}$, the total log-likelihood function is

$$\begin{aligned} l(\gamma, \sigma^2) &:= \ln \prod_{i=1}^{N} L(\Delta r_{t_i} | r_{t_i}, \sigma^2, \gamma) \\ &= -\sum_{i=1}^{N} \left(\frac{1}{2} \ln\left(2\pi\sigma^2 r_{t_i}^{2\gamma}\right) + \frac{\Delta r_{t_i}^2}{2\sigma^2 r_{t_i}^{2\gamma}}\right) \\ &= -\sum_{i=1}^{N} \left(\frac{1}{2} \ln(2\pi) + \frac{1}{2} \ln\sigma^2 + \gamma \ln r_{t_i} + \frac{\Delta r_{t_i}^2}{2\sigma^2 r_{t_i}^{2\gamma}}\right). \end{aligned} \tag{4.27}$$

Then the maximum likelihood estimators of σ^2 and γ are respectively $\hat{\sigma^2}$ and $\hat{\gamma}$, i.e., those values that maximize the total log-likelihood function $l(\gamma, \sigma^2)$.

Remark 2. Here the total log-likelihood function $l(\gamma, \sigma^2)$ is a function of σ^2 (the variance). The standard method is to obtain the maximum likelihood estimator of σ^2 directly instead of considering $l(\gamma, \sigma^2)$ as a function of σ (the standard deviation).

To find the maximum likelihood estimators, we take the partial derivatives of $l(\gamma, \sigma)$ with respect to γ and σ^2:

$$\frac{\partial l(\gamma,\sigma^2)}{\partial \gamma} = -\sum_{i=1}^{N}\left(\ln r_{t_i} - \frac{\Delta r_{t_i}^2}{\sigma^2 r_{t_i}^{2\gamma}}\ln r_t\right) \tag{4.28}$$

$$\frac{\partial l(\gamma,\sigma^2)}{\partial \sigma^2} = -\sum_{i=1}^{N}\left(\frac{1}{2\sigma^2} - \frac{\Delta r_{t_i}^2}{2\sigma^4 r_{t_i}^{2\gamma}}\right). \tag{4.29}$$

To compute the critical point, we let $\frac{\partial l(\gamma,\sigma^2)}{\partial \gamma} = 0$ and $\frac{\partial l(\gamma,\sigma^2)}{\partial \sigma^2} = 0$ hold, and thus with time points $t_i = 1,\ldots,T$, we need to solve the system

$$\sum_{t=1}^{T}\ln r_t = \sum_{t=1}^{T}\frac{\Delta r_t^2}{\sigma^2 r_t^{2\gamma}}\ln r_t, \tag{4.30}$$

$$T = \sum_{t=1}^{T}\frac{\Delta r_t^2}{\sigma^2 r_t^{2\gamma}}. \tag{4.31}$$

Computing σ^2 from (4.31) and substituting its value into (4.30), we conclude that the MLE solution $\hat{\gamma}$ satisfies the equation

$$\sum_{t=1}^{T}\ln r_t = \left(T\sum_{t=1}^{T}\frac{\Delta r_t^2}{r_t^{2\gamma}}\ln r_t\right) \Big/ \left(\sum_{t=1}^{T}\frac{\Delta r_t^2}{r_t^{2\gamma}}\right). \tag{4.32}$$

This equation has no analytic solution but for a given data set, we can obtain a numerical solution of the equation. We will introduce the method for doing so in the following section.

4.4.4 *Numerical Results*

To get the total log-likelihood function $l(\gamma,\sigma^2)$ (4.27) from the likelihood function $L(\gamma,\sigma^2)$, we require the observations to be independent. More precisely $L(\Delta r_t|r_t,\sigma^2,\gamma)$ is the likelihood function for Δr_t given r_t. Thus the sequence of interest rate variations $(\Delta r_{t_i})_{i=1,\ldots,N}$ needs to be a sequence of independent values given the sequence of interest rate levels $(r_t)_{t=1,\ldots,T}$. Here we consider the interest rate variations from one week to the next, namely

$$\Delta r_t = r_{t+w} - r_t. \tag{4.33}$$

If we consider a sequence of times $\{t_i\}_{i=1,\ldots,N}$ with $t_{i+1} - t_i = w$, then the sequence of $\{\Delta r_{t_i}\}$ represents the interest rate variations in non-overlapping time intervals; hence we regard this sequence as a sequence of independent observations of interest rate variations. In practice we choose a specific weekday and gather all the interest rate data for this weekday. The data given

by the Caisse consist of 852 weeks and we choose Friday as the weekday for data gathering.

To get the MLE $\tilde{\gamma}$, we solve (4.32) through a numerical method. Here we use the bisection method to find the root of the equation. This method is outlined below.

1. Define $f(\gamma)$ as

$$f(\gamma) := \sum_{t=1}^{T} \ln r_t - \left(T \sum_{t=1}^{T} \frac{\Delta r_t^2}{r_t^{2\gamma}} \ln r_t \right) \Bigg/ \left(\sum_{t=1}^{T} \frac{\Delta r_t^2}{r_t^{2\gamma}} \right). \tag{4.34}$$

2. For initialization, take $a = 0$ and $b = 1$. It can be checked whether $f(a)f(b) < 0$ holds.
3. For each iteration, take $c = \frac{a+b}{2}$. If $f(a)f(c) > 0$, then let a take the value of c (leaving b unchanged); otherwise, let b take the value of c (leaving a unchanged).
4. When $|f(c)|$ is small enough (we use the condition $|f(c)| < 10^{-6}$), terminate the algorithm and set $\tilde{\gamma}$ equal to c.

Here are the values obtained for two maturities.

Table 4.2 Parameter estimation

Maturity	180 (Semi annual)	365 (Annual)
$\tilde{\gamma}$	0.4981587	0.43337

To illustrate the estimation, we compared empirical and theoretical densities of the interest rate variation at two interest rate levels. We plot the empirical density in black based on the dataset and the theoretical density in red based on the equation (4.11). The graphs are displayed in Figure 4.4 and 4.5.

As a test, we calculate the VaR of a short position in the following two kinds of bonds:

1. Zero coupon bond with one year maturity;
2. 4% coupon bond with one-year maturity and semi-annual coupon payment.

We consider the case of a one-week time horizon and 99% confidence interval. Then the VaR of the above bonds are given in Table 4.3. The results are quite reasonable. As a 4% coupon bond has a shorter duration than a zero coupon bond, it has a smaller VaR (in absolute value).

Table 4.3 VAR with a one-week time horizon and 99% confidence interval

Bond	4% Coupon	Zero Coupon
VaR	-0.18%	-0.22%

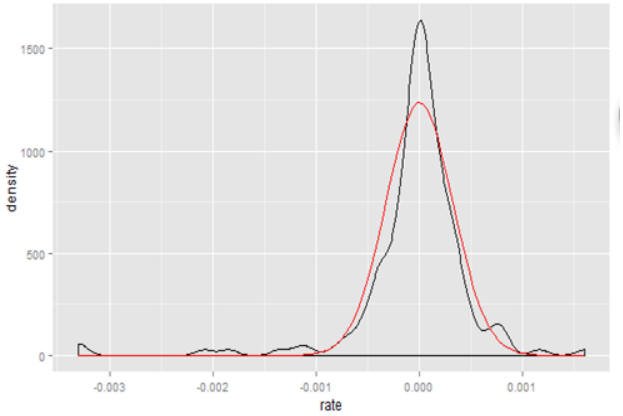

Fig. 4.4 $r_t = 0.01$

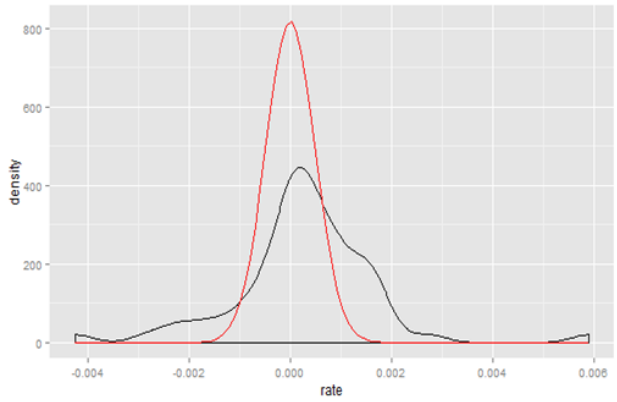

Fig. 4.5 $r_t = 0.26$

4.5 Annualization

4.5.1 *Model: Long Term Aggregation*

In this section we discuss the second question: annualization. Instead of considering short-time variations as in the previous section, we consider how interest rates correlate over a relatively long time horizon. Also, as showed in Figure 4.1, we need to find a process with a positive excess kurtosis (heavy tail). We want our process to be a generalization of Brownian motion so that we have a generalization of the square-root-of-time rule. We will use self-similar processes.

Self-similar processes are stochastic processes that behave in the same fashion when viewed at different levels of magnification, or at different scales in a specific dimension (for instance space or time). Self-similar processes can sometimes be described using heavy-tailed distributions, also known as long-tailed distributions. We now give the mathematical definition of these processes.

Definition 1. A stochastic process $\{X(t)\}_{t\geq 0}$ is said to be self-similar if there exists $H > 0$ (the *Hurst coefficient*) such that the following condition holds for any $k > 0$.

$$X(kt) \stackrel{d}{=} k^H X(t) \tag{4.35}$$

The following are two instances of self-similar processes.

1. If $\{B_t\}_{t\geq 0}$ is the standard Brownian motion, then for $k > 0$,

$$B(kt) = k^{1/2} B(t) \tag{4.36}$$

holds. Thus $\{B_t\}_{t\geq 0}$ is a self-similar process with Hurst coefficient equal to 0.5.
2. If $\{X_t\}_{t\geq 0}$ is an α-stable process, then for $k > 0$,

$$X(kt) = k^{1/\alpha} X(t) \tag{4.37}$$

holds. Thus $\{X_t\}_{t\geq 0}$ is a self-similar process with Hurst coefficient $1/\alpha$.

From Theorem 4.1 in Embrechts and Maejima [2000], a self-similar process is heavy-tailed when $H < 0.5$. This corresponds to a long time-dependence of the process. Moreover the following theorem from (13.2.3) in Miller et al. [2007] shows that a generalized square-root-of-time rule holds for self-similar processes.

Theorem 4.5.1 *Assume that the value of the portfolio P is a self-similar process with Hurst coefficient H. Then the following relation holds.*

$$\mathrm{VaR}_{aT,l}(P) = a^H \,\mathrm{VaR}_{T,l}(P) \tag{4.38}$$

4.5.2 *Estimation and Numerical Results*

The estimation of the Hurst coefficient H is based on the Rescaled Range (R/S) Calculation. We will follow the notation and methods in Kaplan [2013].

For a process $\{X(t)\}_{t\geq 0}$, the range, $R(\tau)$, is defined as the difference between the maximum value $X(t_b)$ and the minimum value $X(t_a)$ of the process over the time period τ.

$$R(\tau) = \max_{t\in[0,\tau]} X(t) - \min_{t\in[0,\tau]} X(t). \tag{4.39}$$

The standard deviation over the range from 1 to τ, $S(\tau)$, is defined as

$$S(\tau) = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}\left(X(t) - \bar{X}_\tau\right)^2}, \tag{4.40}$$

where $\bar{X}_\tau$ is the mean of the process over the range from 1 to τ. Then the rescaled range is calculated by dividing the range $R(\tau)$ by the standard deviation $S(\tau)$.

$$R/S(\tau) = \frac{R(\tau)}{S(\tau)} \tag{4.41}$$

Then Equation 10 in Kaplan [2013] shows that as n tends to ∞, we have

$$\mathbb{E}\left(R/S(n)\right) = Cn^H. \tag{4.42}$$

To get an estimator for H, we run a linear regression line through a set of points, each of which having two coordinates: the log of n (the size of the area on which the average rescaled range is calculated) and the log of the average rescaled range over a set of regions of size n. The slope of the regression line is the estimate of the Hurst exponent. This method for estimating the Hurst exponent was developed and analyzed by Benoît Mandelbrot and his co-authors in articles published around 1968 Mandelbrot and Van Ness [1968]. As shown in Figure 4.6, the rescaled range is calculated for the entire data set (here R/S Ave_0 = RS_0). Then the rescaled range is calculated for the two halves of the data set, resulting in RS_0 and RS_1. These two values are averaged, resulting in R/S Ave_1. In this case the algorithm continues by subdividing each of the previous sections into two parts and calculating the rescaled range for each new section. The rescaled range values for sections of a given size are then averaged. At some point the algorithm stops subdividing the sections, since the regions become too small.

As an example we run the regression for the interest rate data corresponding to a one-year maturity, over the period 2000-2016. The independent variable is log(Scale) and the response variable log(R/S). Figure 4.7 is the Scatter plot of log(R/S) vs log(Scale), with the regression line, and Figure 4.8 is the result of the regression.

The regression shows that an estimate of the Hurst coefficient is 0.05599. Combining this result with the weekly VaR for one-year zero coupon bonds (-0.22%) obtained in the last section, we conclude that the annualized VaR is -0.22%*$52^{0.056}$ = -0.274%.

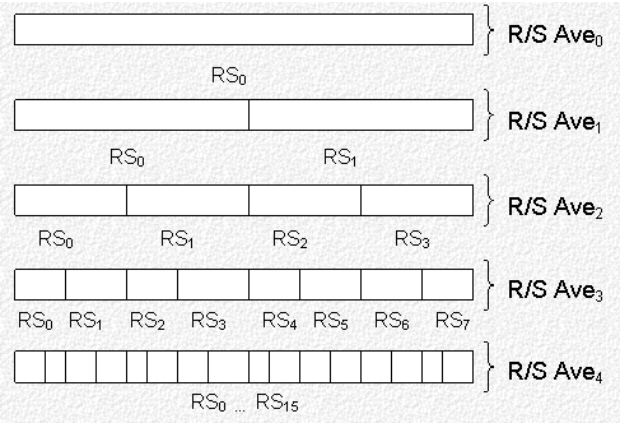

Fig. 4.6 Estimating the Hurst Exponent

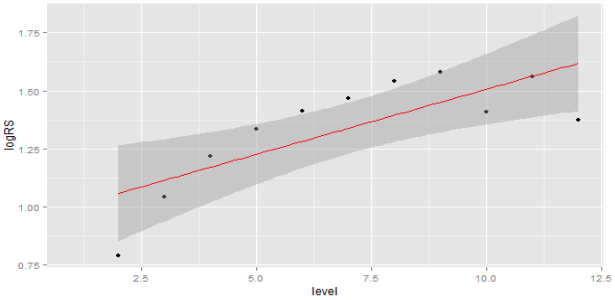

Fig. 4.7 log(R/S) vs log(Scale)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.94757    0.11810   8.024 2.16e-05 ***
level        0.05599    0.01537   3.642  0.00539 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 4.8 log(R/S) vs log(Scale)

4.6 Further Work

We could study many more models for the interest rate process. The following are two instances of such models.

GARCH

GARCH (Generalized Auto-Regressive Conditional Heteroskedasticity) models volatility clustering. In this model the volatility process is time-varying

and dependent upon both the past volatility and past innovations.

$$r_t = \mu + \delta h_t + \epsilon_t \tag{4.43}$$

$$h_t = \gamma_1 + \gamma_2 h_{t-1} + \gamma_3 \epsilon_{t-1}^2 \tag{4.44}$$

MS-GARCH

Markov-switching GARCH model (MS-GARCH) is an extension of GARCH. The conditional mean and variance switch in time from one GARCH process to another. The conditional variance may originate from structural changes in the variance process which are not accounted for by standard GARCH models. Let $\{s_t\}$ be an ergodic Markov chain on a finite set $S = \{1, \cdots, n\}$, with transition probabilities $\{\eta_{ij} = P(s_t = i | s_{t-1} = j)\}$. Then the MS-GARCH model is given by

$$y_t = \mu_{s_t} + \sigma_t \mu_t, \tag{4.45}$$

$$\sigma_t^2 = \omega_{s_t} + \alpha_{s_t} \epsilon_{t-1}^2 + \beta_{s_t} \sigma_{t-1}^2. \tag{4.46}$$

Estimating such a model might be difficult because it allows regime switching in the parameters.

Acknowledgements

We wish to thank the CRM (which organized the Seventh Montreal Industrial Problem Solving Workshop), NSERC and CANSSI (the sponsors of the workshop), and Têkogan Hemazro and Yannis Papageorgiou, from the Caisse de dépôt et placement du Québec. Zichun Ye also wishes to thank the Université de Montréal for their hospitality while this work was in progress, and Dr. Odile Marcotte and Dr. Stéphane Rouillon for their help during the application process and the workshop.

References

John Cox. Notes on option pricing i: Constant elasticity of variance diffusions. Stanford University, Graduate School of Business, 1975.

Jon Danielsson and Jean-Pierre Zigrand. On time-scaling of risk and the square-root-of-time rule. *Journal of Banking & Finance*, 30(10):2701–2713, 2006.

Paul Embrechts and Makoto Maejima. An introduction to the theory of self-similar stochastic processes. *International Journal of Modern Physics B*, 14(12-13):1399–1420, 2000.

Patrick S Hagan, Deep Kumar, Andrew S Lesniewski, and Diana E Woodward. Managing smile risk. In Paul Wilmott, editor, *The Best of Wilmott 1: Incorporating the Quantitative Finance Review*, pages 249–296, New York, 2004. Wiley.

John C Hull. *Options, Futures, and Other Derivatives*. Pearson, 2006.

Ian Kaplan. Estimating the Hurst exponent. Web resource available at http://www.bearcave.com/misl/misl_tech/wavelets/hurst/, 2013.

Benoit B Mandelbrot and John W Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437, 1968.

John Miller, David Edelman, and John Appleby. *Numerical Methods for Finance*. Chapman and Hall/CRC Financial Mathematics Series. Chapman and Hall/CRC, Boca Raton, FL, 2007.

Evert Wipplinger. Philippe Jorion: Value at Risk – The New Benchmark for Managing Financial Risk. *Fin Mkts Portfolio Mgmt*, 21(3):397–398, 2007.