RISK MANAGEMENT MAGAZINE

Vol. 18, Issue 3
September – December 2023

EXCERPT

https://www.aifirm.it/rivista/progetto-editoriale/

Modeling the interest rates term structure using Machine Learning: a Gaussian process regression approach

Alessio Delucchi, Pier Giuseppe Giribone

Modeling the interest rates term structure using Machine Learning: a Gaussian process regression approach

Alessio Delucchi (Avvale S.p.A) – Pier Giuseppe Giribone (University of Genoa – Department of Economics; BPER Banca – Financial Engineering)

Corresponding Author: piergiuseppe.giribone@bper.it

Article submitted to double-blind peer review, received on 2nd August 2023 and accepted on 2nd Decembre 2023

Abstract

The correct modeling of the interest rates term structure should definitely be considered an aspect of primary importance since the forward rates and the discount factors used in any financial and risk analysis are calculated from such structure. The turbulence of the markets in recent years, with negative interest rates followed by their recent substantial rise, the period of the COVID pandemic crisis, the political instabilities linked to the war between Ukraine and Russia have very often led to observe anomalies in the shape of the interest rate curve that are difficult to represent using traditional econometric models, to the point that researchers have to address this modeling problem using Machine Learning methodologies. The purpose of this study is to design a model selection heuristic which, starting from the traditional ones (Nelson-Siegel, Svensson and de Rezende-Ferreira) up to the Gaussian Process (GP) Regression, is able to define the best representation for a generic term structure. This approach has been tested over the past five years on term structures denominated in five different currencies: the Swiss Franc (CHF), the Euro (EUR), the British Pound (GBP), the Japanese Yen (JPY) and the U.S. Dollar (USD).

Key Words: Interest rates term structure, Nelson-Siegel model, Svensson model, de Rezende-Ferreira model, Gaussian process regression.

JEL code: C52-C53-C55-E43-E47

1. Introduction

The correct estimation of the interest rates term structure is of primary importance for financial analysts, risk managers, actuarial experts, and policy makers. Precisely to meet this specific need, a substantial scientific literature has developed aimed at its correct model representation.

One of the first advanced approaches is the smoothed bootstrap initially proposed by (Bliss and Fama, 1987). They proposed to derive zero rates from raw market data and then fit them to the data with a smooth and continuous curve.

To this end, numerous curve fitting spline methods have been employed: quadratic and cubic splines (McCulloch, 1971 and 1975), exponential splines (Vasicek and Fong, 1982), B-splines (Shea, 1984) and (Steeley, 1991), quartic maximum smoothness splines (Adams and Van Deventer, 1994) and penalty function-based splines (Fischer, Nychka and Zervos, 1994) and (Waggoner, 1997).

These approaches have been criticized by (Annaert et al., 2012) because they are characterized by unwanted economic properties as they are statistical techniques that do not incorporate micro-macro economic principles in their functioning.

(Seber and Wild, 2003) also highlight that these methodologies contribute to having a "black-box" type of interpretative effect and therefore should be avoided in contexts of standard financial markets free from turbulence.

(Nelson and Siegel, 1987), (Svensson 1994 and 1996) and, subsequently (de Rezende and Ferreira, 2013), approached the problem of obtaining a smooth bootstrap through nonlinear regression models able to fit reasonably well for different families of term structure shapes observed on the financial markets.

These models are parsimonious, consistent with the theoretical interpretation of the term structure suggested by (Litterman and Scheinkman, 1991) and held in high esteem by both academics and professionals.

Let us consider, for example, that the Nelson-Siegel model and the Svensson model are extensively used by central banks and monetary policy makers (Bank of International Settlements, 2005) and (European Central Bank, 2008).

It should be noted that non-linear models and in particular those which envisage the estimation of a large number of parameters can be subject to potential instability in the calibration phase, having to resort to a numerical optimization routine, which most times is constituted by an algorithm of local search for solutions, generally a quasi-Newtonian one such as L-BFGS (Nocedal, 1980) or a Direct Search one, as a simplex by (Nelder and Mead, 1965).

(Cairns and Pritchard, 2001) show that the estimates of the Nelson-Siegel model are very sensitive to the starting values used in the optimization. Moreover, time series of the estimated coefficients have been documented to be very unstable (Barrett, Gosnell and Heuson, 1995), (Fabozzi, Martellini and Priaulet, 2005), (Diebold and Li, 2006), (Gurkaynak, Sack and Wright, 2006), (de Pooter, 2007).

Finally, the standard errors on the estimated coefficients, though seldom reported, are too large (Annaert et al., 2012).

In addition to the potential technical-computational problems mentioned above, we should also consider the fact that a traditional econometric model assumes a priori the functional form according to which the data observed on the market should be explained.

This approach should be pursued whenever possible, typically during periods of stable, non-turbulent financial markets.

If we consider the anomalies that have recently characterized the financial markets, among which we mention: the issue of negative interest rates and their subsequent sharp rise, the current inflationary context and the ongoing war in Europe between Ukraine and Russia, then it could be considered as incongruous to use canonical econometric models to represent the interest rates term structure.

This aspect has already been highlighted by various studies in which a "bottom-up" approach was implemented, i.e., starting from the data, without making a priori hypotheses on the shape of the function of the interest rate curve, the problem was tackled with Machine Learning paradigms (Giribone, 2023).

Among the statistical methods pertaining to this family, the Radial Basis Function (RBF) Neural Network (Cafferata et al., 2019) and the feed-forward Artificial Neural Network (Caligaris and Giribone, 2015) are worth mentioning.
As has been reiterated in (Cafferata, Giribone and Resta, 2018) it should be emphasized that the use of Machine Learning methods and particularly those connected with Deep Learning should somehow be justified: it is unreasonable to adopt statistical methods which are more sophisticated than necessary if market conditions do not require it or if there are no available quotes.
We have chosen the Gaussian Process (GP) regression for performing this task because it is a methodology that is able to work with relatively little data available. It also inherited some theoretical principles common to diffusive processes (see third section of the paper) and it has been shown to produce good results in similar financial applications (Gonzalvez *et al.*, 2019).
The present study fits into this context and proposes a heuristic of choice between different models for the correct representation of the interest rates term structure. The proposed algorithm, starting from the simplest traditional econometric models (Nelson-Siegel and Svensson) and reaching the most complex ones (de Rezende-Ferreira), evaluates their performance in terms of goodness of fit (adjusted R^2) and estimation stability of the coefficients (analysis of confidence bands and outliers). Only if the traditional approaches are not in line with expectations, the heuristics would automatically implement a Machine Learning method: the approach proposed in this paper is a Gaussian process regression with an automatic selection of different kernels.
The selection heuristic was tested on interest rates term structures over the last five years for different currencies (USD, CHF, GBP, EUR, JPY), each characterized by different financial instruments from which the relative zero rates were derived (Deposits, Futures, FRAs and Swaps).
The following section summarizes the main features of the traditional econometric models for representing the interest rates term structure (Nelson-Siegel, Svensson and de Rezende-Ferreira). The third section illustrates the operating principles of a Gaussian Process Regression providing evidence of how an incorrect or unsatisfactory modeling reached with the previously mentioned approaches can be solved. The fourth section illustrates the operational logic of the model selection heuristics in detail and applies them to different case studies. The last section provides the statistics that emerged from the algorithm and draws the conclusions of the study.

2. Non-linear parametric models

(Nelson and Siegel, 1987) were the first to introduce a simple model for interest rates that also has a satisfactory predictive power both for short and very long maturities; these characteristics make it a still relevant approach both for scholars and professionals. Many researchers over the years further developed this model. Among the many contributions, the approaches presented here are those proposed by (Svensson, 1994) and (de Rezende and Ferreira, 2013) who, as will be explained, added additional terms to the '87 formula in order to have a better fit for the term structure under particular circumstances.

2.1 The Nelson Siegel model

A class of functions that generates the typical yield curve shapes is that associated with solutions to differential equations. The expectations theory of the term structure of interest rates provides heuristic motivation for investigating this class since, if spot rates are generated by differential equations, then forward rates, being forecasts, will be the solution to the equations discussed in (Nelson and Siegel, 1987). The researchers explored two cases:

- The instantaneous forward rate is the solution to a second order differential equation with real and unequal roots.
- The instantaneous forward rate is the solution to a second order differential equation with real and equal roots.

In the first case the instantaneous forward rate is defined as:

$$F(t) = \beta_0 + \beta_1 \cdot \exp\left(-\frac{t}{\tau_1}\right) + \beta_2 \exp\left(-\frac{t}{\tau_2}\right) \quad (1)$$

However, tests made by Nelson and Siegel showed that fitting the model considering $F(t)$ as forward rates resulted in overparameterization and they explained this in two ways: firstly, they observed the fact that changing τ_1 and τ_2 caused almost no change to the fit obtained and, secondarily, using statistical software, the model described failed to satisfactorily converge to a robust solution.

For these reasons Nelson and Siegel studied the second case. The forward rate as a solution for a differential equation with equal roots is:

$$F(t) = \beta_0 + \beta_1 \cdot \exp\left(-\frac{t}{\tau}\right) + \beta_2 \left(\frac{t}{\tau}\right) \exp\left(-\frac{t}{\tau}\right) \quad (2)$$

Integrating the formula for the spot rate they obtained:

$$R(t) = \beta_0 + (\beta_1 + \beta_2) \cdot \frac{\tau\left[1-\exp\left(-\frac{t}{\tau}\right)\right]}{t} - \beta_2 \cdot \exp\left(-\frac{t}{\tau}\right) \quad (3)$$

Or in the more canonical form:

$$R(t) = \beta_0 + \beta_1 \cdot \frac{\tau\left[1-\exp\left(-\frac{t}{\tau}\right)\right]}{t} + \beta_2 \cdot \left[\frac{\tau\left[1-\exp\left(-\frac{t}{\tau}\right)\right]}{t} - \exp\left(-\frac{t}{\tau}\right)\right] \quad (4)$$

Nelson and Siegel show that function (4) is able to capture the typical shapes assumed by interest rate term structures. Looking more closely at the function and at the meaning of the coefficients, the three βs can be seen as the strength of the different term, with β_0 measuring the weight of the long-term rates, β_1 the weight of short-term rates and β_2 the weight of the medium-term rates. This kind of interpretation depends on the τ factor, that is considered as a time decay factor which affects mostly the short-term component, only mildly the mid-term one and does not influence the long-term part at all.

An example of this model at work is shown for two cases, one in which the model can satisfactorily explain the behaviour of interest rates and another one in which instead it fails to converge at all (Figure 1).

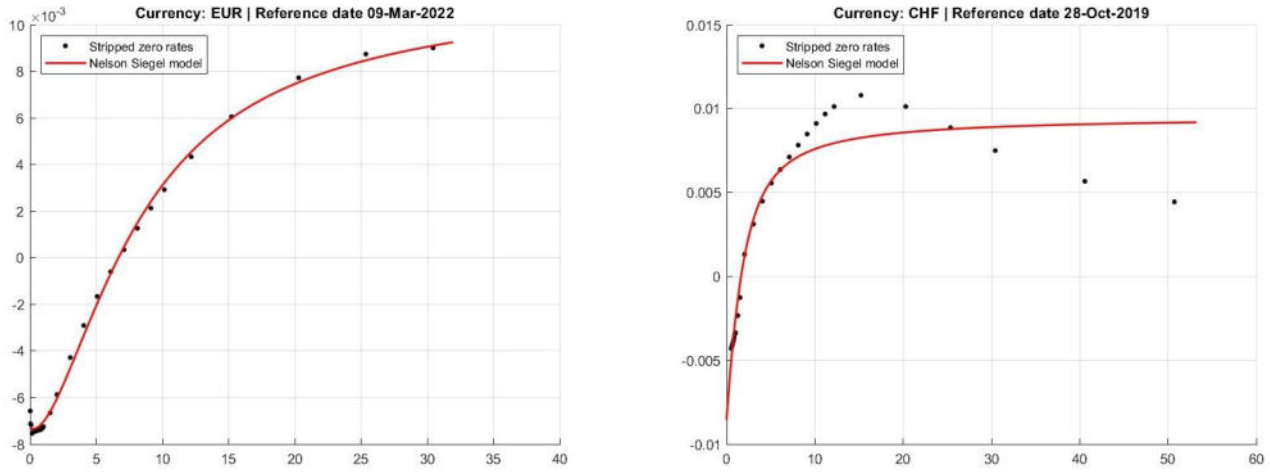


Figure 1: The Nelson Siegel model: good fitting versus poor fitting

2.2) The Svensson model

(Svensson, 1994) proposed a new version of the Nelson and Siegel model in which the author added a further term in order to catch a second hump and thus increase the flexibility of the model. The forward rate function for the Svensson model is as follows:

$$F(t) = \beta_0 + \beta_1 \cdot \exp\left(-\frac{t}{\tau_1}\right) + \beta_2 \left(\frac{t}{\tau_1}\right) \exp\left(-\frac{t}{\tau_1}\right) + \beta_3 \left(\frac{t}{\tau_2}\right) \exp\left(-\frac{t}{\tau_2}\right) \quad (5)$$

Applying the formula of the spot rate by integrating the forward rate, the new function defining the spot rate is:

$$R(t) = \beta_0 + \beta_1 \cdot \frac{\tau_1\left[1-\exp\left(-\frac{t}{\tau_1}\right)\right]}{t} + \beta_2 \cdot \left[\frac{\tau_1\left[1-\exp\left(-\frac{t}{\tau_1}\right)\right]}{t} - \exp\left(-\frac{t}{\tau_1}\right)\right] + \beta_3 \cdot \left[\frac{\tau_2\left[1-\exp\left(-\frac{t}{\tau_2}\right)\right]}{t} - \exp\left(-\frac{t}{\tau_2}\right)\right] \quad (6)$$

The fourth term, made of two parameters β_3 and τ_2 (which must be positive) is then able to capture a term structure which shape includes two humps. The interpretation of the other terms remains the same.
Svensson proves that the Nelson Siegel model goodness of fit is fulfilling in most cases, but sometimes, when the term structure shape proves to be more complex, the extended model can improve the fit in a significant way.

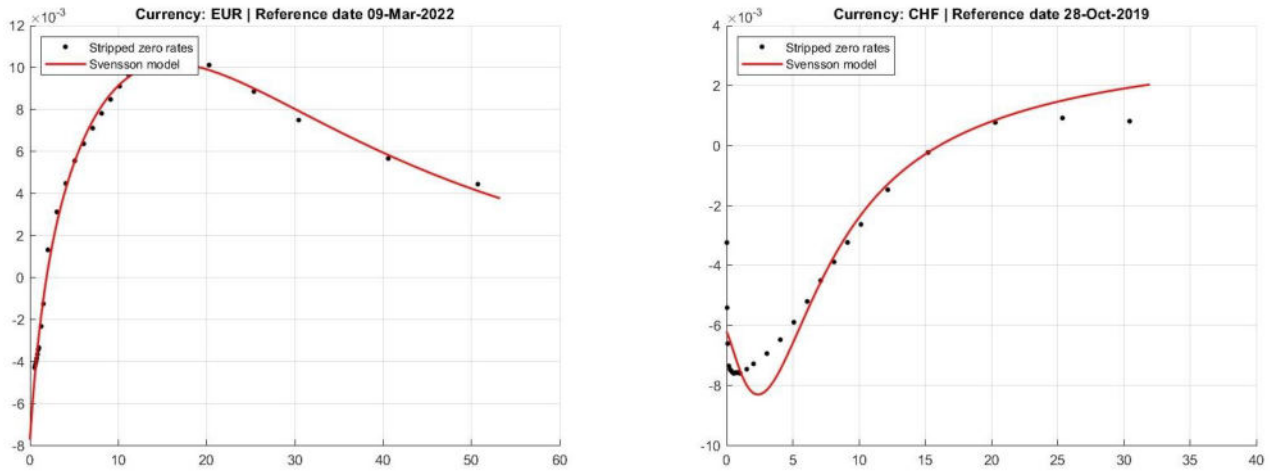


Figure 2: The Svensson model: good fitting versus poor fitting

2.3) The de Rezende and Ferreira model

(de Rezende and Ferreira, 2011) introduced a model that further develops the Nelson Siegel and the Svensson models, adding a fifth factor to address an additional need for flexibility. The new formulas for spot and forward rates are:

$$F(t) = \beta_0 + \beta_1 \cdot \exp\left(-\frac{t}{\tau_1}\right) + \beta_2 \left(\frac{t}{\tau_1}\right) \exp\left(-\frac{t}{\tau_1}\right) + \beta_3 \left(\frac{t}{\tau_2}\right) \exp\left(-\frac{t}{\tau_2}\right) + \beta_4 \left(\frac{t}{\tau_3}\right) \exp\left(-\frac{t}{\tau_3}\right) \quad (7)$$

$$R(t) = \beta_0 + \beta_1 \frac{\tau_1\left[1-\exp\left(-\frac{t}{\tau_1}\right)\right]}{t} + \beta_2 \left[\frac{\tau_1\left[1-\exp\left(-\frac{t}{\tau_1}\right)\right]}{t} - \exp\left(-\frac{t}{\tau_1}\right)\right] + \beta_3 \left[\frac{\tau_2\left[1-\exp\left(-\frac{t}{\tau_2}\right)\right]}{t} - \exp\left(-\frac{t}{\tau_2}\right)\right] + \beta_4 \left[\frac{\tau_3\left[1-\exp\left(-\frac{t}{\tau_3}\right)\right]}{t} - \exp\left(-\frac{t}{\tau_3}\right)\right] \quad (8)$$

We notice that the fifth term proposed recalls the one introduced by Svensson. The interpretation for this new term is that of a second slope of the curve (to catch a third hump), while the interpretation of the other terms remains unaltered.

This model is expected to work well in case of a very complex and twisted curve, so in those few cases in which the preceding models may fail to fit or tend to underfit.

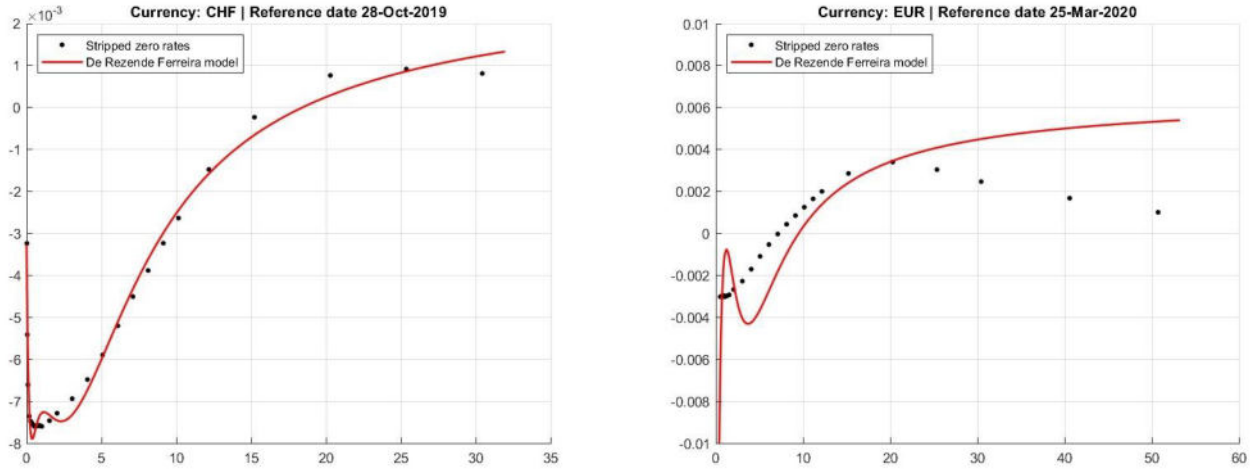


Figure 3: The de Rezende – Ferreira model: good fitting versus poor fitting

3. A Machine Learning approach through a Gaussian process regression

There are many ways to interpret GP regression models, and following the work of Rasmussen and Williams (2006) there are two main approaches to GPs:

-A weight-space view: the typical way of looking at a regression model, mostly focused on parameters.

-A function-space view: considering GP as a distribution over functions.

In both cases Bayesian Linear Regression (BLR) is involved and indeed we can think of GPs as a generalization of this peculiar type of regression.

BLR takes advantage of normal distribution properties (conditioning and marginalization) in order to analytically solve the regression problem. It is composed of three elements: Prior distribution, Likelihood and Posterior distribution.

3.1 The Weight-space view

Given the dataset: $D = \{(x_i, y_i) | i = 1, \cdots, n\}$, the linear regression model is:

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w} + \varepsilon \qquad y = f(\boldsymbol{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2) \quad (9)$$

Where $\boldsymbol{x}$ is the input vector, $\boldsymbol{w}$ is the vector of weights (i.e. parameters) of the linear model, f is the function value and y is the observed target value. We have assumed that the observed values y differ from the function values $f(\boldsymbol{x})$ by additive noise, which follows an independent, identically distributed Gaussian distribution with zero mean and variance σ_n^2.

This noise assumption, together with the model directly gives rise to the likelihood, the probability density of the observations given the parameters, which is factored over cases in the training set because of the independence assumption to give

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - x_i^T \boldsymbol{w})^2}{2\sigma_n^2}\right) = \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2} |\boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{w}|^2\right) = \mathcal{N}(\boldsymbol{X}^T \boldsymbol{w}, \sigma_n^2 I) \quad (10)$$

Where $|\boldsymbol{z}|$ denotes the Euclidean length of vector $\boldsymbol{z}$, I is the identity matrix of dimension n. In the Bayesian formalism we need to specify a prior over the parameters, expressing our belief about the parameters before we look at the observations. We put a zero mean Gaussian prior with covariance matrix Σ_p on the weights: $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_p)$.

Through Bayes theorem it is possible to find the posterior distribution of the parameter:

$$Posterior = \frac{Likelihood \times Prior}{Marginal\ Likelihood} \quad (11)$$

Considering the particular case of $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})} \quad (12)$$

The marginal likelihood as the name suggests is found through marginalization and it is independent from the parameters (the parameters are marginalized out). Analytically, it can be written as:

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})\, d\boldsymbol{w} \quad (13)$$

The posterior in Eq. (12) combines the likelihood and the prior, and captures everything we know about the parameters. Writing only the terms from the likelihood and prior which depend on the weights, we obtain:

$$P(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{w})\right) \exp\left(-\frac{1}{2}\boldsymbol{w}^T\Sigma_p^{-1}\boldsymbol{w}\right) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \bar{\boldsymbol{w}})^T\left(\frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^T + \Sigma_p^{-1}\right)(\boldsymbol{w} - \bar{\boldsymbol{w}})\right) \quad (14)$$

Where $\bar{\boldsymbol{w}} = \sigma_n^{-2}\left(\sigma_n^{-2}\boldsymbol{X}\boldsymbol{X}^T + \Sigma_p^{-1}\right)^{-1}\boldsymbol{X}\boldsymbol{y}$ and we recognize the form of the posterior distribution as Gaussian with mean $\bar{\boldsymbol{w}}$ and covariance matrix A^{-1}:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \sim \mathcal{N}\left(\bar{\boldsymbol{w}} = \frac{1}{\sigma_n^2}A^{-1}\boldsymbol{X}\boldsymbol{y}, A^{-1}\right) \quad (15)$$

Where $A = \sigma_n^{-2}\boldsymbol{X}\boldsymbol{X}^T + \Sigma_p^{-1}$ (Rasmussen and Williams, 2006). Notice that for this model (and indeed for any Gaussian posterior) the mean of the posterior distribution $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$ is also its mode, which is called the maximum a posteriori (MAP) estimate of $\boldsymbol{w}$.

To make predictions for a test case we average over all possible parameter values, weighted by their posterior probability. Thus the predictive distribution for $f_* = f(\boldsymbol{x}_*)$ at $\boldsymbol{x}_*$ is given by averaging the output of all possible linear models with reference to the Gaussian posterior

$$p(f_*|\boldsymbol{x}_*, \boldsymbol{X}, y) = \int p(f_*|\boldsymbol{x}_*, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{X}, y)d\boldsymbol{w} = \mathcal{N}\left(\frac{1}{\sigma_n^2}\boldsymbol{x}_*^T A^{-1}\boldsymbol{X}y, \boldsymbol{x}_*^T A^{-1}\boldsymbol{x}_*\right) \quad (16)$$

The predictive distribution is again Gaussian, with a mean given by the posterior mean of the weights from Eq. (15) multiplied by the test input, as one would expect from symmetry considerations. The predictive variance is a quadratic form of the test input with the posterior covariance matrix, showing that the predictive uncertainties grow with the magnitude of the test input, as one would expect for a linear model.

The Bayesian linear model suffers from limited expressiveness. A simple idea to overcome this problem is to first project the inputs into some high dimensional space using a set of basis functions and then apply the linear model in this space instead of directly on the inputs themselves. As long as the projections are fixed functions (i.e. independent of the parameters $\boldsymbol{w}$) the model is still linear in the parameters, and therefore analytically tractable.

Specifically, we introduce the function $\boldsymbol{\phi}(\boldsymbol{x})$ which maps D-dimensional input vector $\boldsymbol{x}$ into an N dimensional feature space. Further let the matrix $\Phi(\boldsymbol{X})$ be the aggregation of columns $\boldsymbol{\phi}(\boldsymbol{x})$ for all cases in the training set. Now the model is

$$f(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{w} \quad (17)$$

Where the vector of parameters has length N. The analysis for this model is analogous to the standard linear model, except that everywhere $\Phi(\boldsymbol{X})$ is substituted for $\boldsymbol{X}$. Thus the predictive distribution becomes

$$f_*|\boldsymbol{x}_*, \boldsymbol{X}, y \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\boldsymbol{\phi}(\boldsymbol{x}_*)^T A^{-1}\Phi y, \boldsymbol{\phi}(\boldsymbol{x}_*)^T A^{-1}\boldsymbol{\phi}(\boldsymbol{x}_*)\right) \quad (18)$$

With $\Phi = \Phi(\boldsymbol{X})$ and $A = \sigma_n^{-2}\Phi\Phi^T + \Sigma_p^{-1}$. To make predictions using Eq. (18) we need to invert the A matrix of size $N \times N$ which may not be convenient if N, the dimension of the feature space, is large. However, we can rewrite the equation in the following way:

$$f_*|\boldsymbol{x}_*, \boldsymbol{X}, y \sim \mathcal{N}\left(\boldsymbol{\phi}_*^T\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}y, \boldsymbol{\phi}_*^T\Sigma_p\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^T\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\Phi^T\Sigma_p\boldsymbol{\phi}_*\right) \quad (19)$$

Where we have used the shorthand $\phi(\boldsymbol{x}_*) = \phi_*$ and defined $K = \Phi^T \Sigma_p \Phi$. To show this for the mean, first note that using the definitions of A and K we have $\sigma_n^{-2}\Phi(K + \sigma_n^2 I) = \sigma_n^{-2}\Phi\left(\Phi^T \Sigma_p \Phi + \sigma_n^2 I\right) = A\Sigma_p \Phi$.
Now multiplying through by A^{-1} from left and $(K + \sigma_n^2 I)^{-1}$ from the right gives $\sigma_n^{-2} A^{-1}\Phi = \Sigma_p \Phi (K + \sigma_n^2 I)^{-1}$, showing the equivalence of the mean expressions in Eq. (18) and Eq. (19). For the variance we use the matrix inversion lemma, setting $Z^{-1} = \Sigma_p$, $W^{-1} = \sigma_n^2 I$ and $V = U = \Phi$ therein.

3.2 The Function-space view

Gaussian Processes are defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" (Rasmussen and Williams, 2006).

A Gaussian process is completely specified by its mean function and covariance function. We define the mean function $m(\boldsymbol{x})$ and the covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ of a real process $f(\boldsymbol{x})$ as:

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})] \quad (20)$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[\left(f(\boldsymbol{x}) - m(\boldsymbol{x})\right)\left(f(\boldsymbol{x}') - m(\boldsymbol{x}')\right)\right]$$

And will write the Gaussian process as:

$$f(\boldsymbol{x}) \sim GP\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right) \quad (21)$$

A Gaussian process is defined as a collection of random variables. Thus, the definition automatically implies a consistency requirement, which is also sometimes known as the marginalization property. This property means that if the GP e.g. specifies $(y_1, y_2) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then it must also specify $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ where Σ_{11} is the relevant submatrix of Σ.

This can be seen as the prior over functions. According to this prior it is possible to define the joint distribution of f, the training outputs and f_* the test outputs:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right) \quad (22)$$

From this joint distribution, the predictive distribution can be derived, that is, as before, the conditional distribution of f given $\boldsymbol{x}_*, \boldsymbol{x}$ and y:

$$P(f_* | \boldsymbol{x}_*, \boldsymbol{x}, y) \sim N(K_* K^{-1} y, K_{**} - K_* K^{-1} K_*^T) \quad (23)$$

That is exactly the same function defined in the weight-space view. Proceeding through these steps, the reason why GP regression is considered a generalization of BLR becomes clear: GP regression uses kernels instead of basis functions to find the families of the functions for regression.
Using kernels allows to define a very broad family of functions that basis functions alone could not handle. This makes GPs more flexible as it is still possible to implement the Bayesian update and reach a good posterior predictive fit.

3.3 Kernel functions and hyperparameters

Kernel functions control the model, they determine which kind of function is more or less likely to be sampled. The kernel is a function that measures how similar two inputs are and therefore it is quite clear why such functions are used to produce covariance matrices in GPs.

Let us suppose we have $\boldsymbol{x}$ and $\boldsymbol{x}'$; the kernel function is:

$$k(\boldsymbol{x}, \boldsymbol{x}' | \boldsymbol{\tau}) \quad (24)$$

$\boldsymbol{x}$ and $\boldsymbol{x}'$ can refer to any two objects, provided that we can measure similarity between them. $\boldsymbol{\tau}$ is a vector of hyperparameters used to tune the kernel function. The output of the kernel function will be a similarity measure, large and positive if the inputs are very similar, large and negative otherwise.

There are technical restrictions on which functions can be used as kernels, since the covariance matrix must be positive definite (the reason for this restriction is based on the (Mercer, 1909) theorem). The model comes down to which prior functions are likely to be sampled and this is dictated by the kernel, meaning that this aspect is fundamental. It is essential to decide what makes two x similar or dissimilar.

The idea is that GP will sample functions with close y values for $\boldsymbol{x}$ deemed similar by the kernel. In order to clarify this concept, consider the Squared exponential kernel function (or RBF function), one of the most used for Gaussian Processes:

$$k(\boldsymbol{x}, \boldsymbol{x}' | \boldsymbol{\tau}) = \sigma_f^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right) + \sigma_n^2 \quad (25)$$

Hyperparameters in the squared exponential are as follows:

ℓ: length scale. It scales distances between the x. It means that if ℓ is a small value, most input pairs are considered different, and this implies that the sample functions can "wiggle" more rapidly. If ℓ is large on the other hand, most input pairs will be considered similar, leading to smoother sample functions. Intuitively, this function tells how far we have to go before things become virtually uncorrelated with one another.

σ_f^2: output scale (or signal variance) determines the scale of the y values. If σ_f^2 increases, the function spans a bigger part of the y axis; if it decreases, the opposite holds.

σ_n^2: noise variance. It is not a direct parameter of the kernel function, but it influences the likelihood function from which the optimal values of σ_f^2 and ℓ are found.

But how do we find the values for these parameters? As anticipated, the likelihood function plays a fundamental role. The τ vector determines how kernel measures similarity.

The idea is to select those hyperparameters that maximize the log likelihood of y after integrating out possible functions.

The idea is to optimize (i.e., find the maximum of):

$$\ln P(y|\boldsymbol{x}, \boldsymbol{\tau}, \sigma_n^2) = \ln \int p(y|f, \sigma_n^2) P(f|\boldsymbol{x}, \boldsymbol{\tau})\, df \quad (26)$$

Computing $\ln P(y|\boldsymbol{x}, \boldsymbol{\tau}, \sigma_n^2)$ essentially means finding hyperparameters that improve the fitting of the data through a function f sampled by the prior, and this is done over an infinite number of sample functions.

Intuitively we want to pick the hyperparameters where the prior functions explain the data in the best way, thus hyperparameters for which the f_s fit the data well without conditioning.

This means that:

$$\ln P(y|\boldsymbol{x}, \boldsymbol{\tau}, \sigma_n^2) \sim \ln N(y|\boldsymbol{0}, K(X, X) + \sigma_n^2 I) \quad (27)$$

The gradient of this function can be computed with respect to the hyperparameters. The probability function is differentiable, hence any algorithm based on gradient is able to obtain hyperparameters that maximize the probability. With GPs we can optimize a huge number of hyperparameters.

There are many types of kernels, and the designer can even decide to combine them by adding one to the other (sum of kernels means that two functions are sampled and then the sum of the two leads to the "final" kernel) or multiplying one to the other.

This allows to create more complicated models that can better explain data.

An application of this Gaussian Processes could be done in the fitting of the yield curves for which the parametric models in the previous chapter gave poor results. The results of the GP regression are reported in Figure 4.

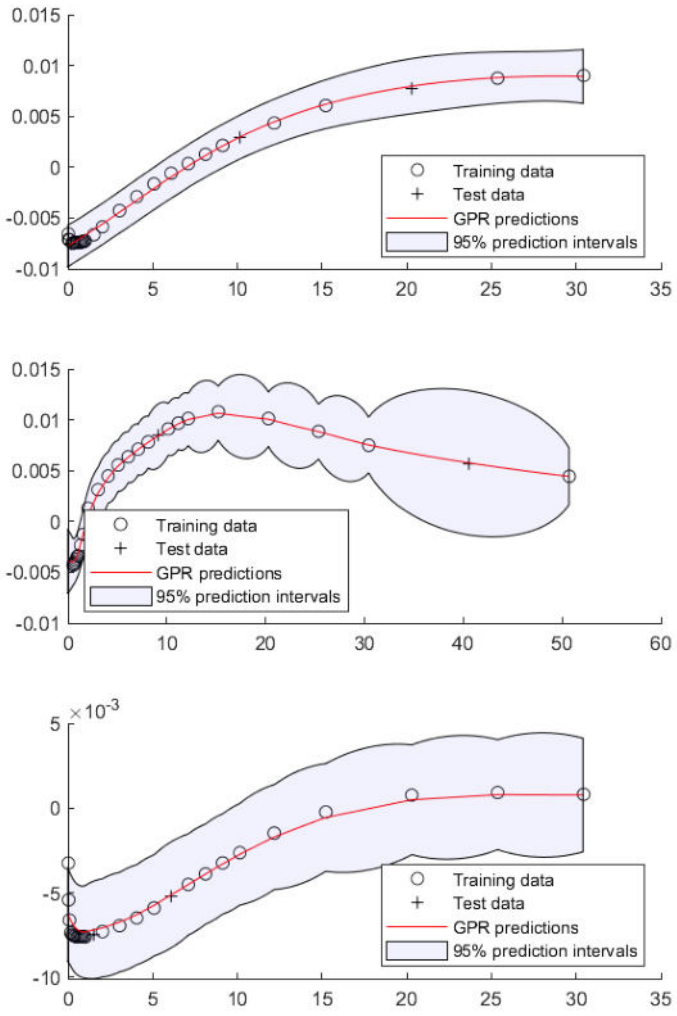


Figure 4: Gaussian Process applications for the previous poor fitting cases

The kernel functions considered in this study are:

- Squared Exponential: already discussed in this section.

- Exponential:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{\|x-x'\|}{\ell}\right) \quad (28)$$

Both the Squared Exponential and the Exponential kernel functions work quite well for smoother functions, while in case of functions with kinks or local structures, other kernel functions, such as the (Matérn, 1960), perform better.

- Matérn 3/2:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \left(1 + \frac{\sqrt{3}\|x-x'\|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\|x-x'\|}{\ell}\right) \quad (29)$$

- Matérn 5/2:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \left(1 + \frac{\sqrt{5}\|x-x'\|}{\ell} + \frac{5\|x-x'\|^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}\|x-x'\|}{\ell}\right) \quad (30)$$

In short, the Matérn kernel functions tend to be more flexible as they are derived by considering a smoothness parameter with value 3/2 and 5/2 in the cases considered here. Higher values of the smoothness parameter result in smoother and more differentiable functions.

- Rational quadratic:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \left(1 + \frac{\|x-x'\|^2}{2\alpha\ell^2}\right)^{-\alpha} \quad (31)$$

Where α is a positive scale parameter.

The Rational quadratic kernel function incorporates a balance between short-range and long-range correlations. This is possible thanks to the scale parameter α. A higher α value results in a smoother function, capturing long-range correlations. Conversely, a lower α value leads to a rougher function that emphasizes short-range correlations. The idea is that as the value of α changes, a higher weight is assigned to a different section of the curve.

An Automatic Relevance Determination (ARD) version of all the previous kernels can be applied to any kernel function that has a length scale. This method, by introducing a separate length scale parameter for each input variable in the covariance function of the GP model, is a check for the relevance of the input variable.

When the length scale for a particular input variable is small, the GP model becomes more sensitive to variations in that variable, conversely, when the length scale is large, the GP model becomes less sensitive to variations in that variable.

The ARD kernel functions are then:

- ARD Squared exponential kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\sum_{i=1}^{n} \frac{\|x_i-x_i'\|^2}{2\ell_i^2}\right) \quad (32)$$

- ARD Exponential kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\sum_{i=1}^{n} \frac{\|x_i-x_i'\|}{\ell_i}\right) \quad (33)$$

- ARD Matérn 3/2:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \left(1 + \sqrt{3}\sum_{i=1}^{n} \frac{\|x_i-x_i'\|}{\ell_i}\right) \exp\left(-\sqrt{3}\sum_{i=1}^{n} \frac{\|x_i-x_i'\|}{\ell_i}\right) \quad (34)$$

- ARD Matérn 5/2:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \left(1 + \sqrt{5}\sum_{i=1}^{n} \frac{\|x_i-x_i'\|}{\ell_i} + \frac{5}{3}\sum_{i=1}^{n} \frac{\|x_i-x_i'\|^2}{\ell_i^2}\right) \exp\left(-\sqrt{5}\sum_{i=1}^{n} \frac{\|x_i-x_i'\|}{\ell_i}\right) \quad (35)$$

- ARD Rational quadratic:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \left(1 + \frac{1}{2\alpha}\sum_{i=1}^{n} \frac{\|x_i-x_i'\|^2}{\ell_i^2}\right)^{-\alpha} \quad (36)$$

4. Case Study

The methodologies described previously will be used to model the interest rates term structure, considering different time periods, different currencies and for each currency different financial instruments (according to their market liquidity). The time range examined starts from 1st January 2018 and ends on 21st March 2023. Five currencies have been considered: the Swiss Franc (CHF), the Euro (EUR), the British Pound (GBP), the Japanese Yen (JPY) and the U.S. Dollar (USD).The instruments used to model the term structure for each currency are: swaps for CHF, GBP and JPY; deposits, forwards and swaps for EUR; deposits, futures and swaps for USD. The granulometry for each currency, with the corresponding terms and instruments, is shown in Table 1. Data used in this study can be considered in line with the best market practice given that they are retrieved from the Bloomberg® yield curves module. Zero rates and discount factors for each eligible date in the time span have been bootstrapped from market rates. The eligibility of dates depends on one criterion: the number of par rates available; if the missing rates are more than ten, the date is considered ineligible. In case the number of missing rates is less or equal to ten, the missing rates will be interpolated, and the zero rates will be computed. This selection process is further developed in the flow chart depicted in Figure 5.

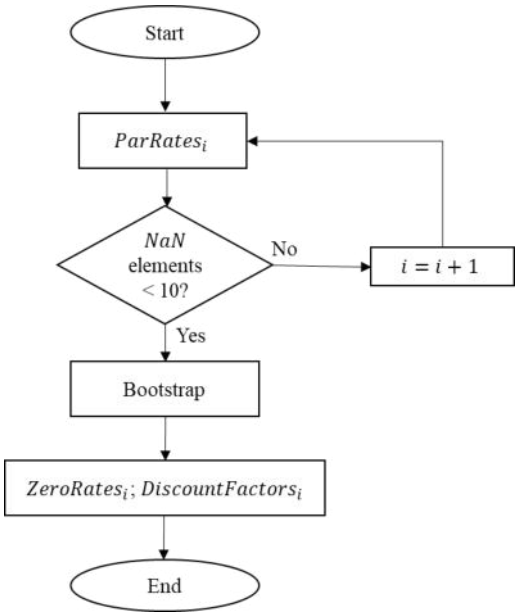


Figure 5: Selection process for the Zero rates stripping algorithm

Instruments per currency

Starting date: 01/01/2018						End date: 03/21/2023			
CHF		EUR		GBP		JPY		USD	
Term	Instrument	Term	Instrument	Term	Instrument	Term	Instrument	Term	Instrument
1W	Swap	6M	Deposit	1W	Swap	1W	Swap	3MO	Deposit
2W	Swap	FRA1X7	FRA	2W	Swap	2W	Swap	FUT_1	Futures
1MO	Swap	FRA2X8	FRA	1MO	Swap	3W	Swap	FUT_2	Futures
2MO	Swap	FRA3X9	FRA	2MO	Swap	1MO	Swap	FUT_3	Futures
3MO	Swap	FRA4X10	FRA	3MO	Swap	2MO	Swap	FUT_4	Futures
4MO	Swap	FRA5X11	FRA	4MO	Swap	3MO	Swap	FUT_5	Futures
5MO	Swap	FRA6X12	FRA	5MO	Swap	4MO	Swap	2Y	Swap
6MO	Swap	FRA9X15	FRA	6MO	Swap	5MO	Swap	3Y	Swap
7MO	Swap	FRA12X18	FRA	7MO	Swap	6MO	Swap	4Y	Swap
8MO	Swap	2Y	Swap	8MO	Swap	7MO	Swap	5Y	Swap
9MO	Swap	3Y	Swap	9MO	Swap	8MO	Swap	6Y	Swap
10MO	Swap	4Y	Swap	10MO	Swap	9MO	Swap	7Y	Swap
11MO	Swap	5Y	Swap	11MO	Swap	10MO	Swap	8Y	Swap
12MO	Swap	6Y	Swap	12MO	Swap	11MO	Swap	9Y	Swap
18MO	Swap	7Y	Swap	18MO	Swap	12MO	Swap	10Y	Swap
2Y	Swap	8Y	Swap	2Y	Swap	15MO	Swap	11Y	Swap
3Y	Swap	9Y	Swap	3Y	Swap	18MO	Swap	12Y	Swap
4Y	Swap	10Y	Swap	4Y	Swap	2Y	Swap	15Y	Swap
5Y	Swap	11Y	Swap	5Y	Swap	3Y	Swap	20Y	Swap
6Y	Swap	12Y	Swap	6Y	Swap	4Y	Swap	25Y	Swap
7Y	Swap	15Y	Swap	7Y	Swap	5Y	Swap	30Y	Swap
8Y	Swap	20Y	Swap	8Y	Swap	6Y	Swap	40Y	Swap
9Y	Swap	25Y	Swap	9Y	Swap	7Y	Swap	50Y	Swap
10Y	Swap	30Y	Swap	10Y	Swap	8Y	Swap		
12Y	Swap	40Y	Swap	12Y	Swap	9Y	Swap		
15Y	Swap	50Y	Swap	15Y	Swap	10Y	Swap		
20Y	Swap			20Y	Swap	11Y	Swap		
25Y	Swap			25Y	Swap	12Y	Swap		
30Y	Swap			30Y	Swap	15Y	Swap		
				40Y	Swap	20Y	Swap		
				50Y	Swap	25Y	Swap		
						30Y	Swap		
						35Y	Swap		
						40Y	Swap		

Table 1: Financial instruments used for bootstrap and Interest rates term structures granulometry

After this preliminary filter, the initial 1362 dates for each of the five currencies become 1349 for CHF, 1342 for EUR, 1326 for GBP, 1293 for JPY, and 1347 for USD.
The surfaces of all the stripped zero rates and discount factors, divided by currency, are reported from Figure 6 to Figure 10.

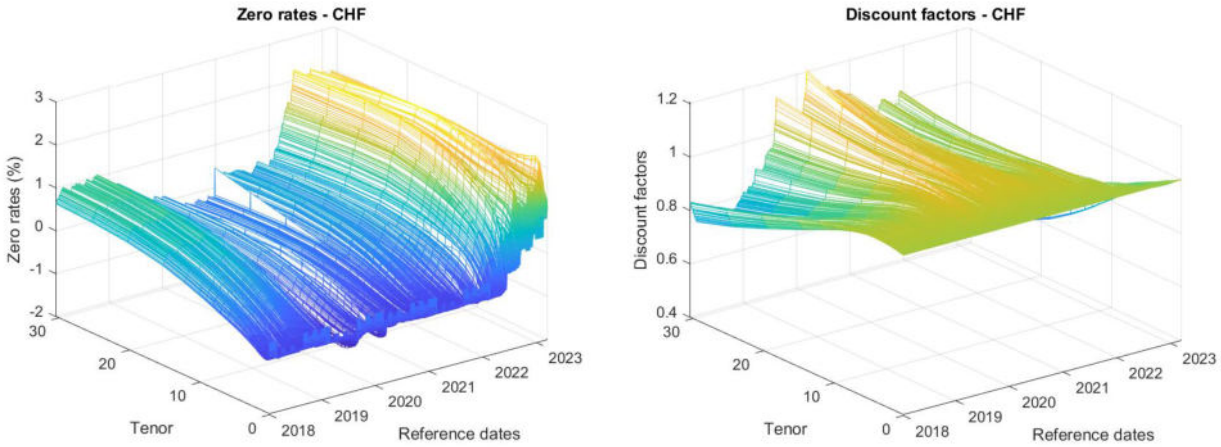


Figure 6: Term structure and Discount factors surface - CHF

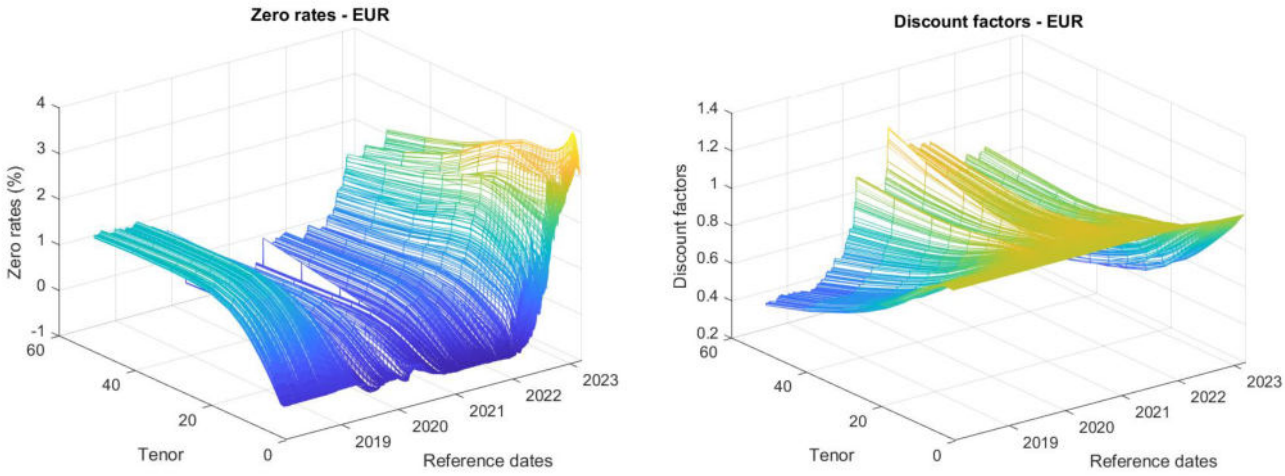


Figure 7: Term structure and Discount factors surface - EUR

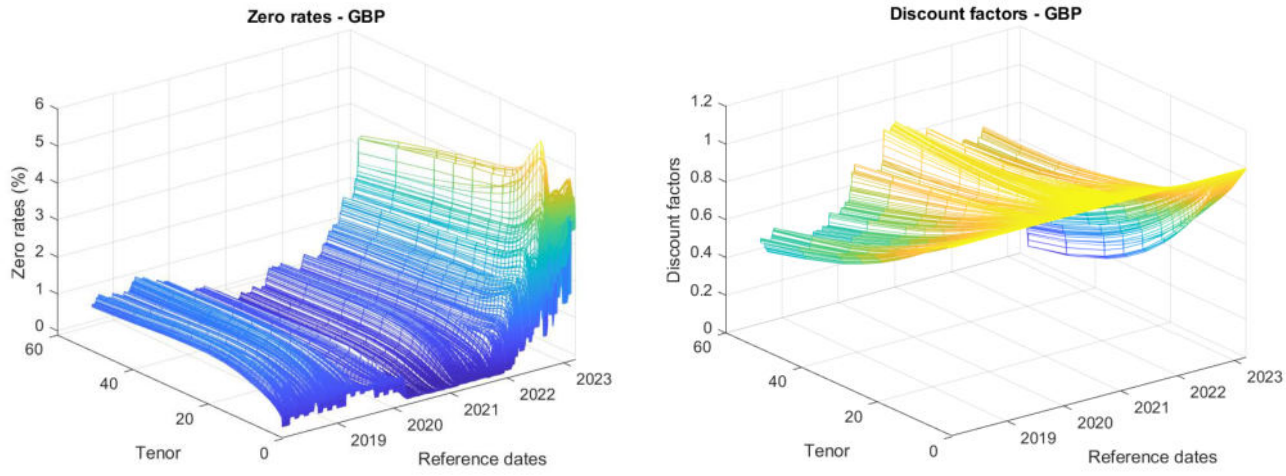


Figure 8: Term structure and Discount factors surface - GBP

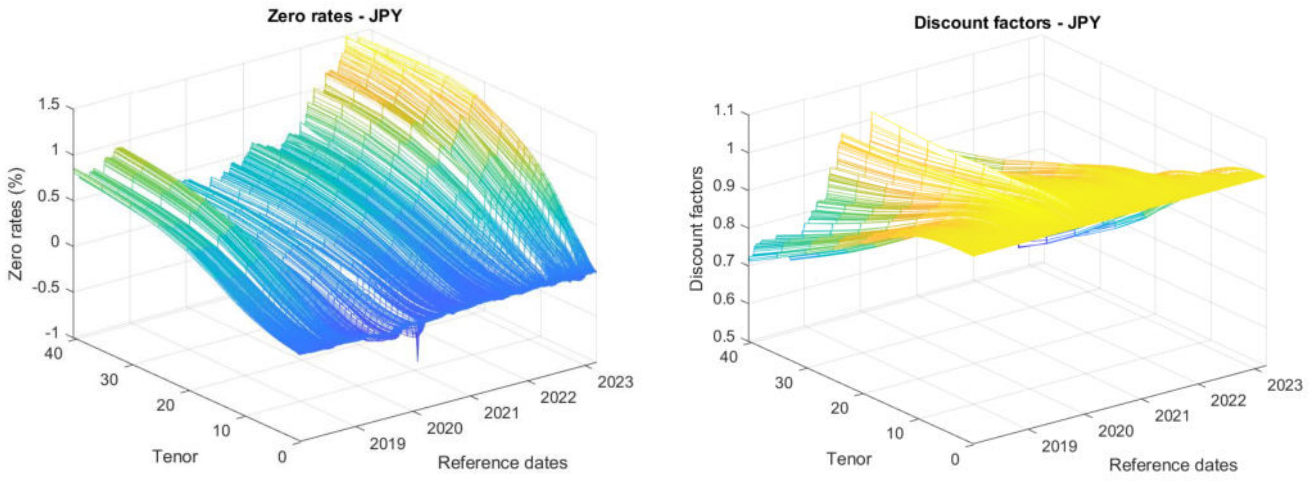


Figure 9: Term structure and Discount factors surface - JPY

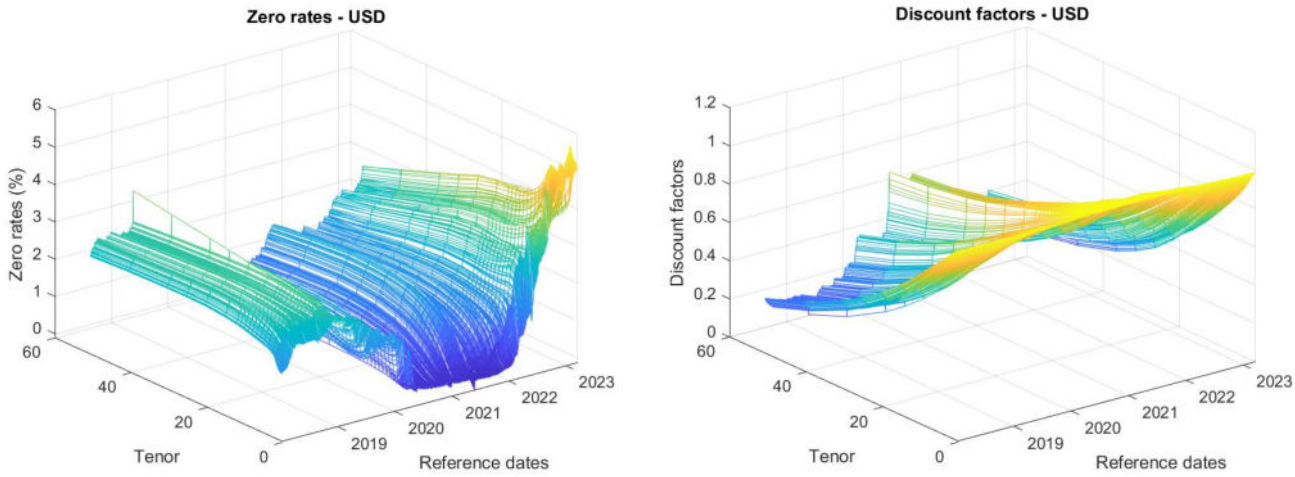


Figure 10: Term structure and Discount factors surface – USD

The term structure for each eligible date is then modelled following parsimonious criteria. The idea is to start from the simplest model and use more complex models only if certain criteria on the goodness of fit or stability of the results are not fulfilled.

This means that the first model that will be used is Nelson Siegel, then Svensson, then de Rezende and if none of the parametric models works satisfactorily, then the Gaussian Process Regression will be implemented.

The first discriminating criterion is the goodness of fit, measured through the adjusted R-square. The threshold is set to a value of 0.95, and below such value the model is rejected in favour of more complex (and flexible) models.

After checking the goodness of fit, the focus is on the stability of results. The selection for stability is conducted following two steps:

- Detection of unrealistically unstable results,

- Detection of outliers.

In the first step, the coefficients of the parametric models are analyzed observing their mean value and using a 95% confidence level. If a model's coefficient has an upper (lower) bound of the confidence band which is ten times higher (lower) than the mean value, then it is considered unrealistically unstable, and the model is discarded.

The second step, on the other hand, implies a more "classical" procedure of outlier detection. If the model's coefficient has an upper (lower) bound that is higher than the mean value plus two times the standard deviation, then the model is discarded. It is worth to highlight that we have implemented a robust check on the starting guesses related to the nonlinear least squares solver.

If statistical performances are not aligned with the previous criteria using the first random initial values, the algorithm automatically produces two other sets of values for the solver.

If none of these attempts works, then the heuristic will take into consideration a more complex parametric model. The process of model selection is further developed in the flow chart below (Figure 11).

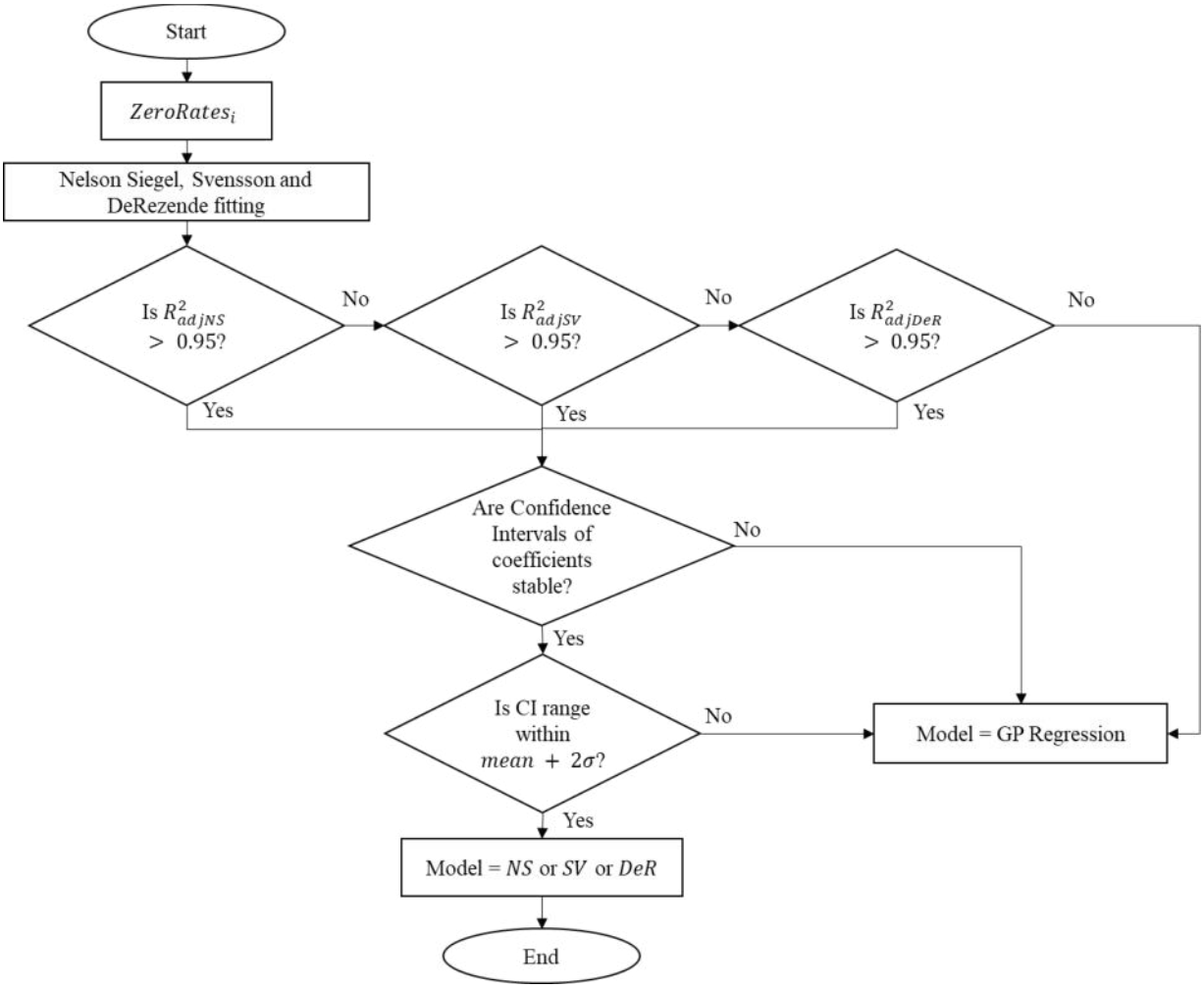
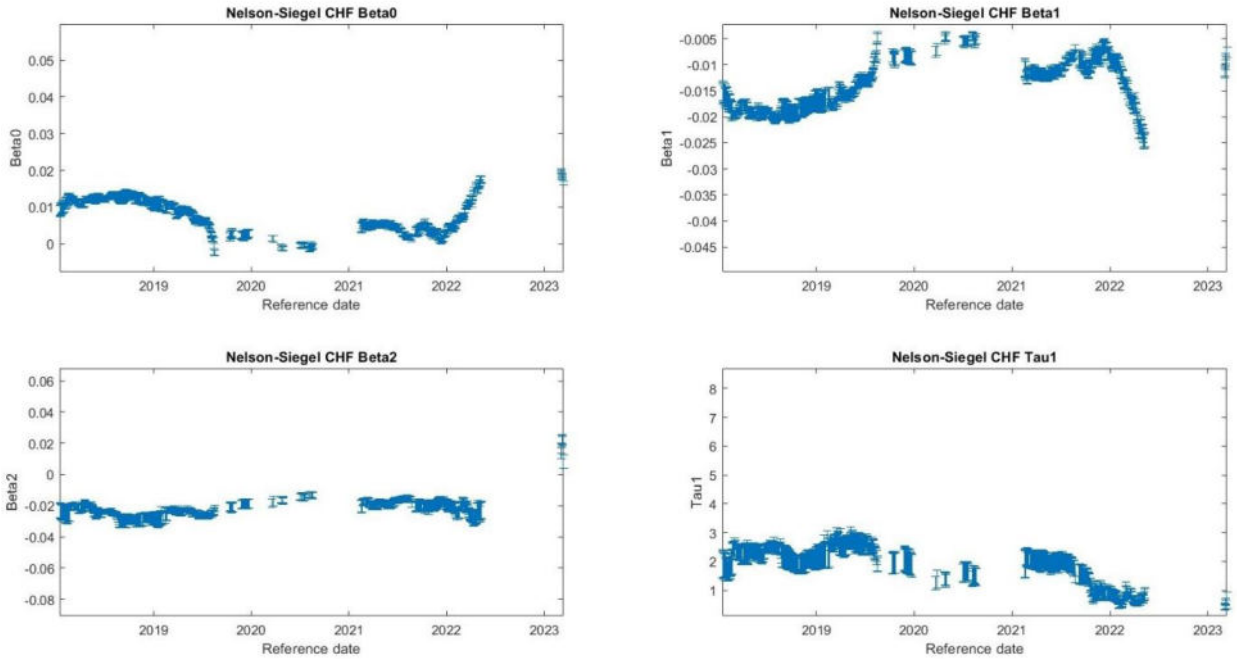


Figure 11: Heuristic for the model selection

The results of this selection are reported in the next Figures (12-16).



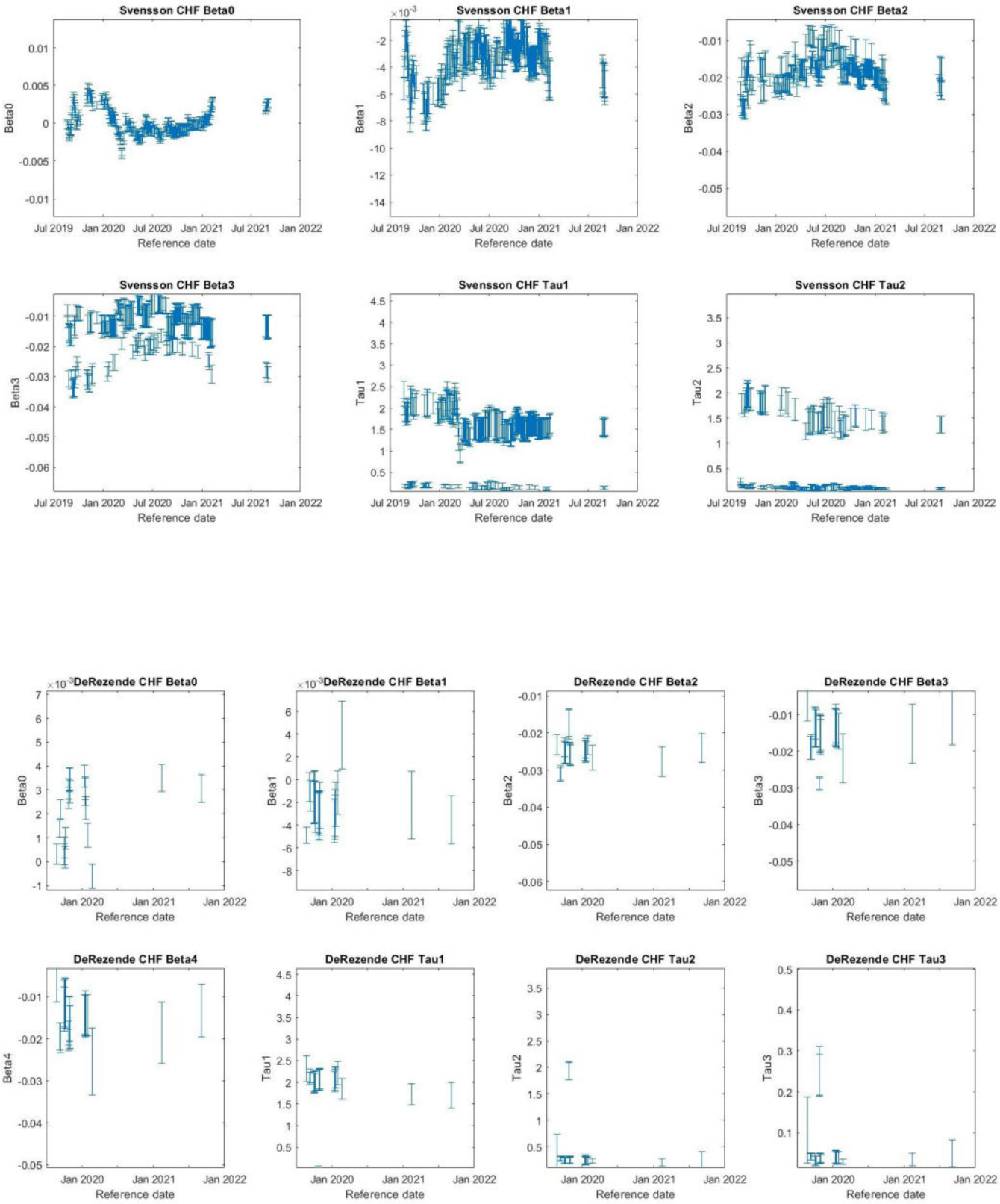


Figure 12: Nelson-Siegel, Svensson and de Rezende coefficients for the CHF Interest rates term structures

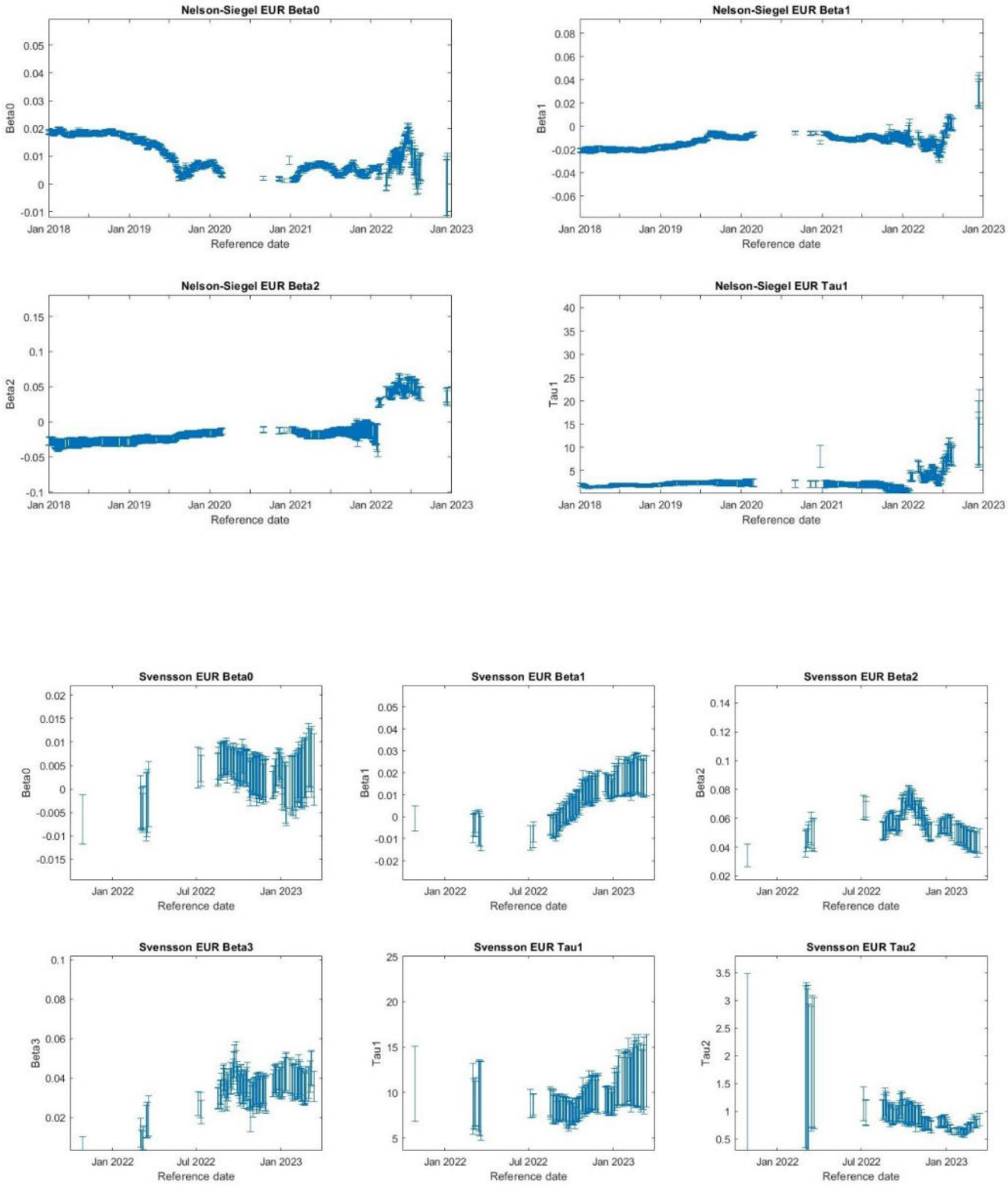
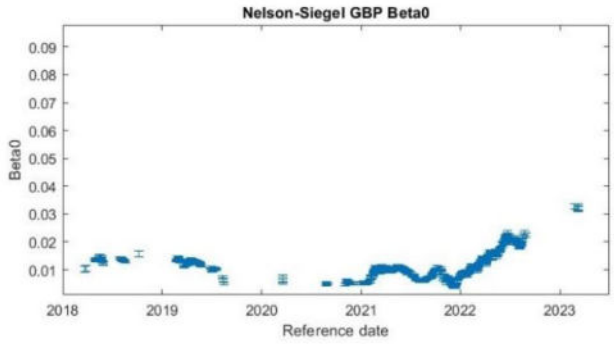
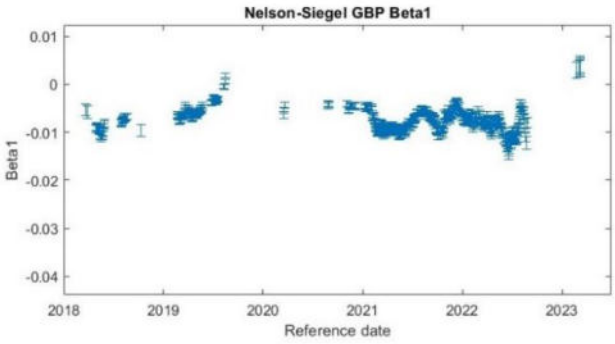
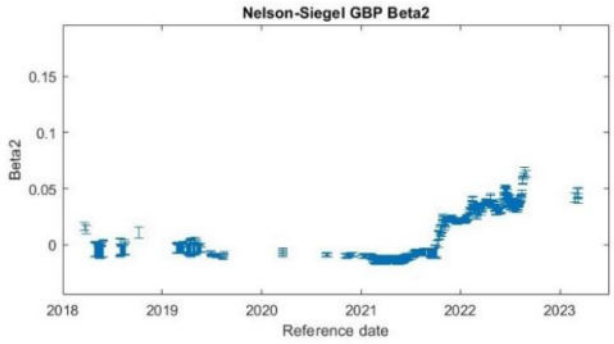


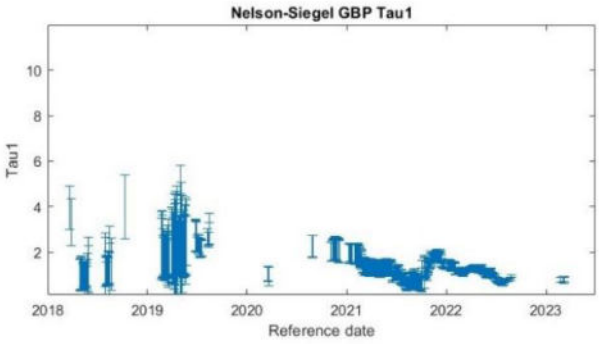
Figure 13: Nelson-Siegel and Svensson coefficients for the EUR Interest rates term structures

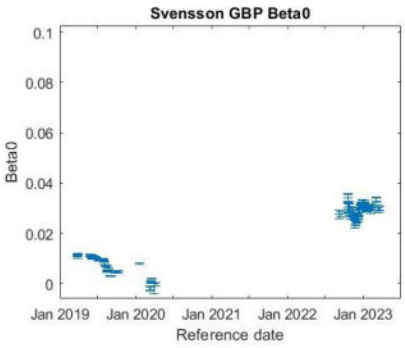
The results for the de Rezende model for the EUR currency are quite peculiar: for every estimation made with this model, the coefficients are deemed as unstable.

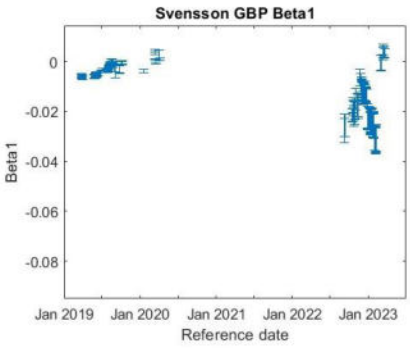

Nelson-Siegel GBP Beta0
Beta0
Reference date

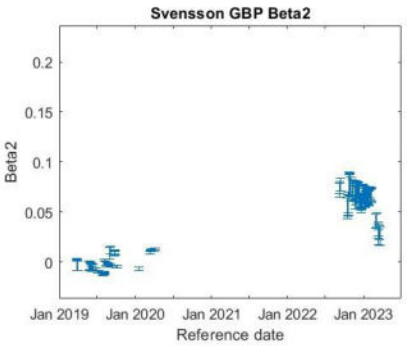

Nelson-Siegel GBP Beta1
Beta1
Reference date

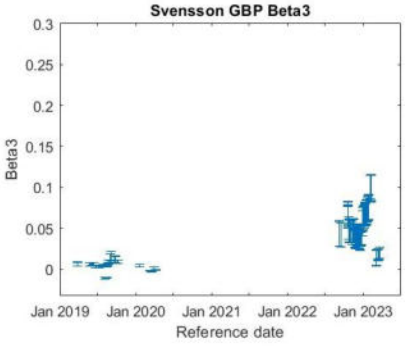

Nelson-Siegel GBP Beta2
Beta2
Reference date

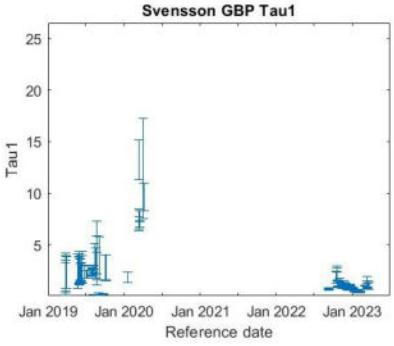

Nelson-Siegel GBP Tau1
Tau1
Reference date

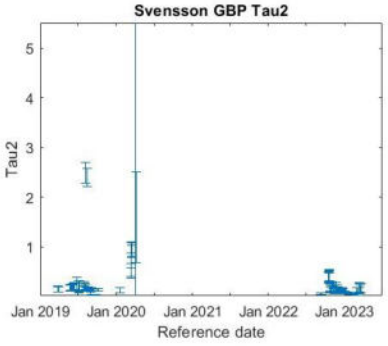

Svensson GBP Beta0
Beta0
Reference date


Svensson GBP Beta1
Beta1
Reference date


Svensson GBP Beta2
Beta2
Reference date


Svensson GBP Beta3
Beta3
Reference date


Svensson GBP Tau1
Tau1
Reference date


Svensson GBP Tau2
Tau2
Reference date

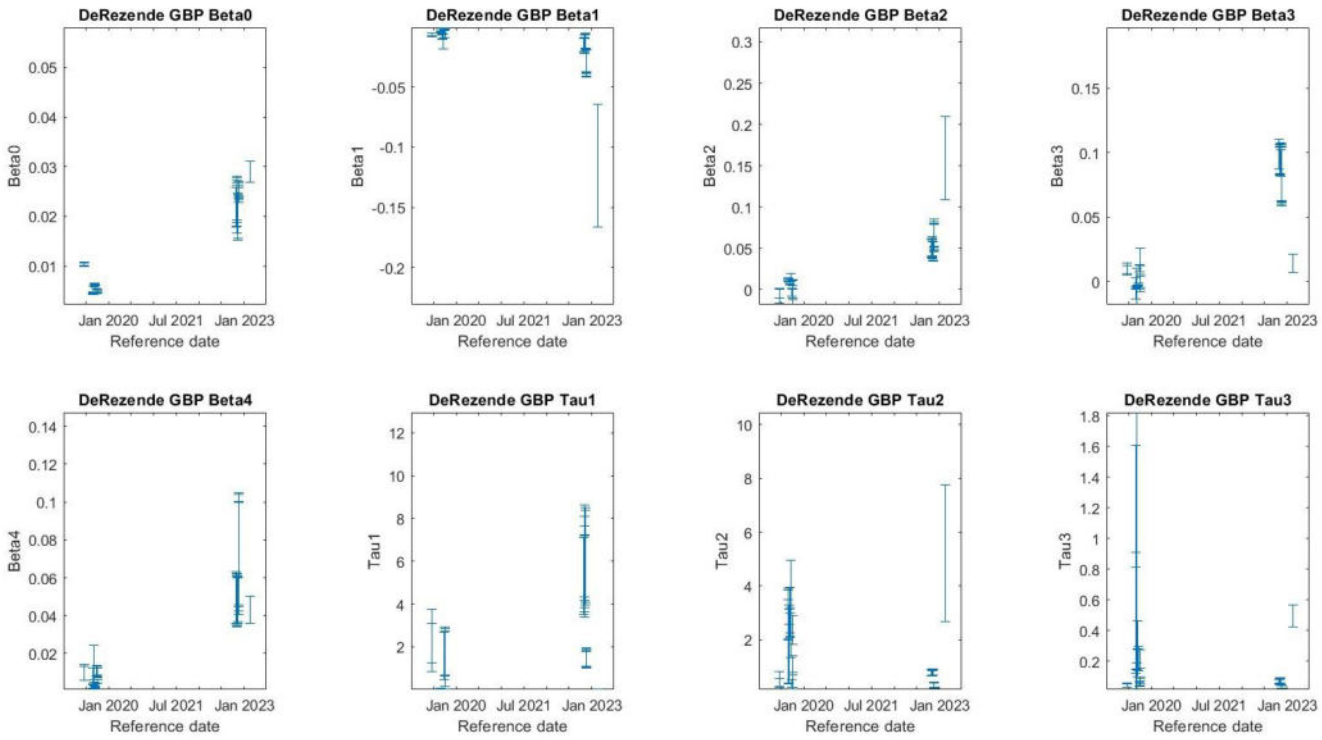


Figure 14: Nelson-Siegel, Svensson and de Rezende coefficients for the GBP Interest rates term structures

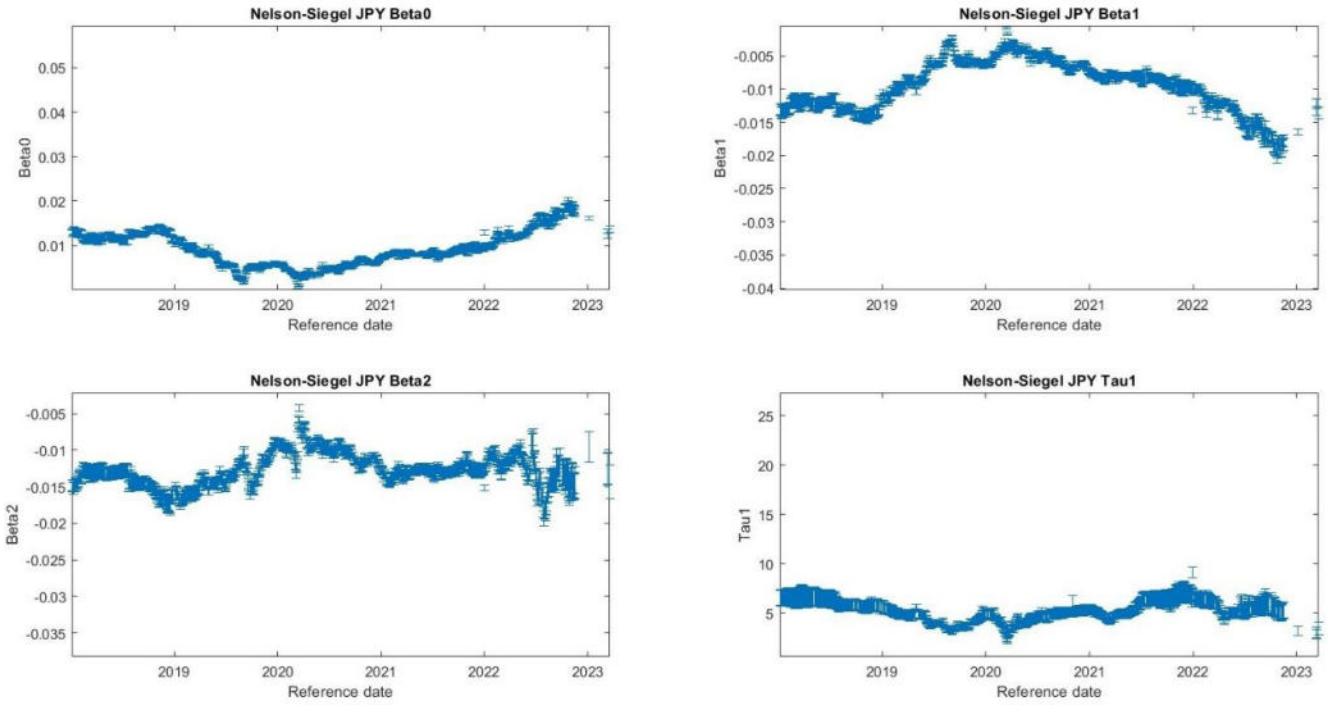


Figure 15: Nelson-Siegel coefficients for the JPY Interest rates term structures

The JPY is another interesting case, as after the procedures described above, basically only the Nelson Siegel model is used, with the GP regression used for the few points deemed as outliers.

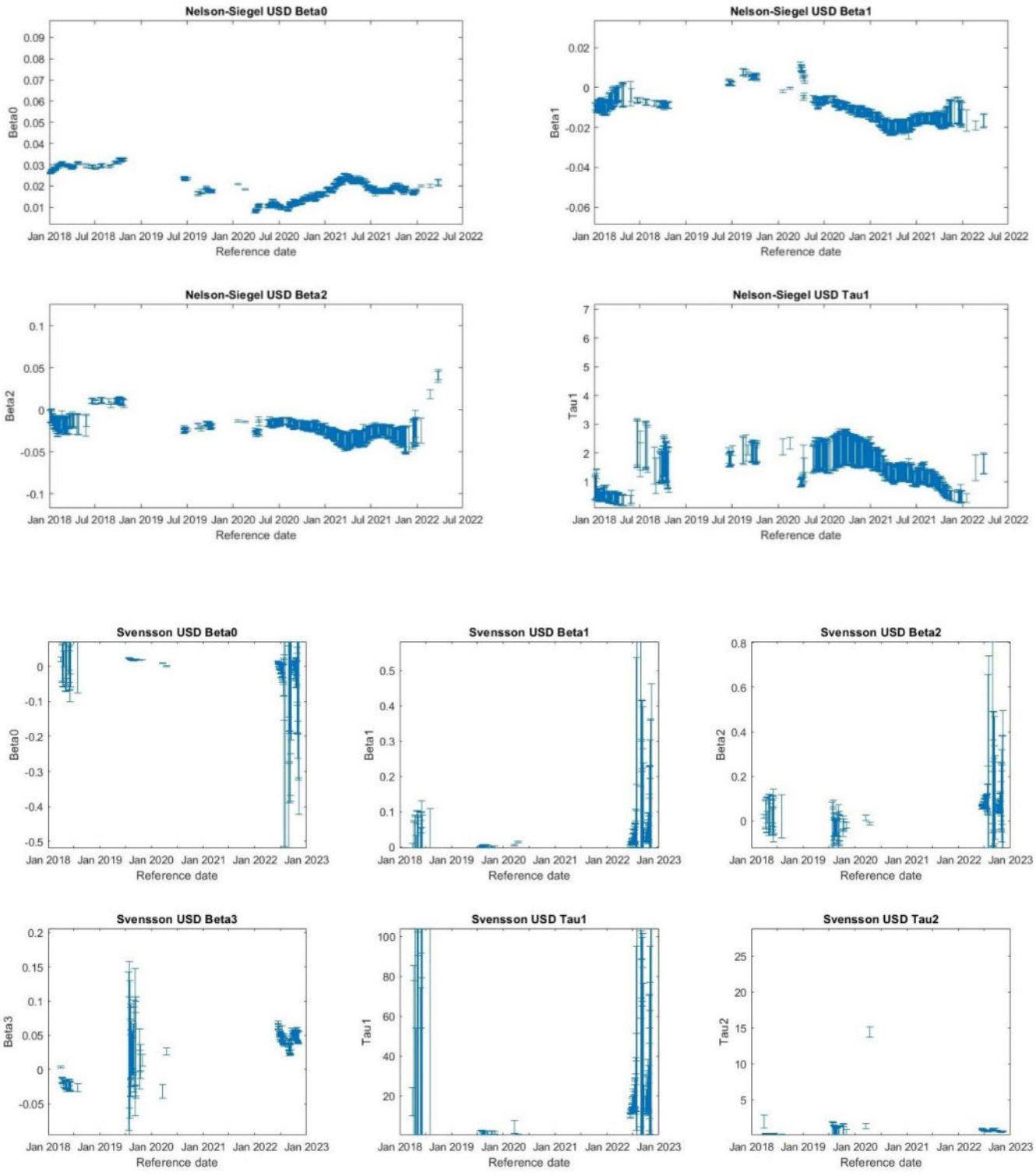


Figure 16: Nelson-Siegel and Svensson coefficients for the USD Interest rates term structures

As in the EUR case, for the USD currency, again, the results with the de Rezende model are too unstable.

After the selection process, the reference dates for which the modelling through parametric models did not give a satisfying result are represented through the Gaussian Process regression.
Given that the yield curves that failed to be modelled with the parametric approaches can assume very different shapes, the kernel choice is made automatically among a list of kernel functions, each one with peculiar characteristics that can be useful for modelling the different characteristics of the term structures. For a list of the kernel functions, see paragraph 3.

After the kernel choice, the model has been run and a 10-fold cross validation procedure has been implemented in order to check the performance of the model. There are no overfitting problems given that the MSE of the training and test sets almost match in all cases. As an example of the procedure conducted, the (10-fold) MSE of the term structure with reference date 8th February 2019, modelled through GP regression, has been plotted with respect to the iterations of the procedure (Figure 17).

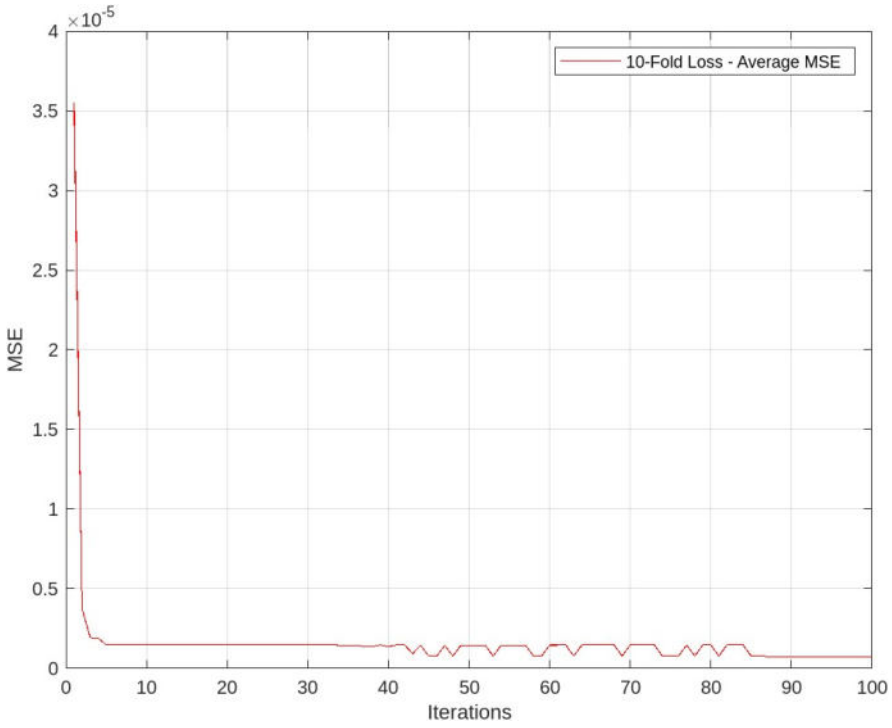


Figure 17: Coefficients for GBP Interest rates term structure

5. Main Results and Conclusions

The results computed on the parametric models are in line with evidence from other works. For example, (Svensson, 1996) and (de Rezende, 2013) in their cited works found that the (Nelson Siegel, 1987) model is the most applied, as is the case here.

Besides, the role played by Machine Learning should be highlighted. The number of models applied per currency and the overall results are displayed in Figure 18.

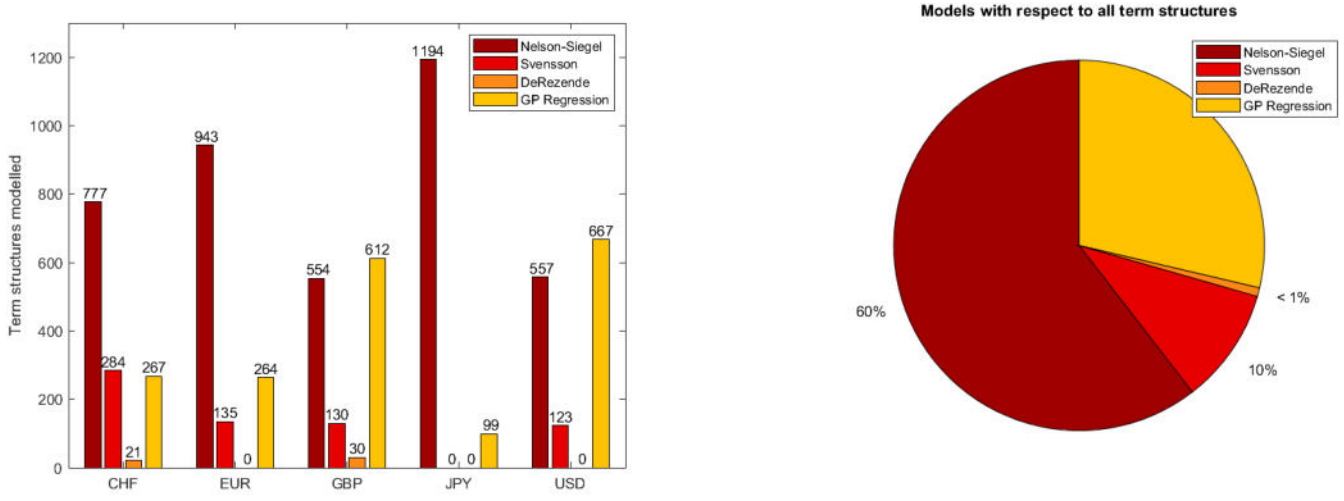


Figure 18: Models used per currency

As shown, most of the term structures have been modelled through the Nelson Siegel and the GP regression. The case of Japan is quite curious, as almost the entirety of the term structures over the reference period have been modelled through Nelson Siegel. This may also be explained observing the smoothness of the surface of the zero rates, due to a greater stability in interest rates compared to the other currencies. Another observation, deduced from Figure 18, is the similarity between USD and GBP and between EUR and CHF.

Overall, 29% of the models used are GP Regression models: it is a large figure. This may be because the reference period considers sub periods of strong turbulence in financial markets in general, and in particular related to monetary policies. This assessment seems to be confirmed if we analyse both the graphs of the coefficients and the zero rates surfaces in section 4: the period of the COVID-19 outbreak and the recent surge in inflation have led to high volatility, and it seems to have affected the capability of parametric

models to effectively model the term structure. This can be seen in the graphs (12-16) related to models' coefficients, in the white gaps (i.e. the time periods without the confidence intervals) in the first period of the pandemic and in the very last part of the reference period that coincided with high inflation and stricter monetary policies. As highlighted above, observing the graphs of coefficients (12-16), again, the case of Japan looks quite striking.
A less "qualitative" result concerns the kernel functions used in the GP regressions. As for the term structure modelling problem, Automatic Relevance Determination kernel functions appear to be better than their "standard" counterparts. Surprisingly, the Squared Exponential kernel function proves to be the less used kernel function for this kind of regression problem. These results are shown in the bar chart in Figure 19.

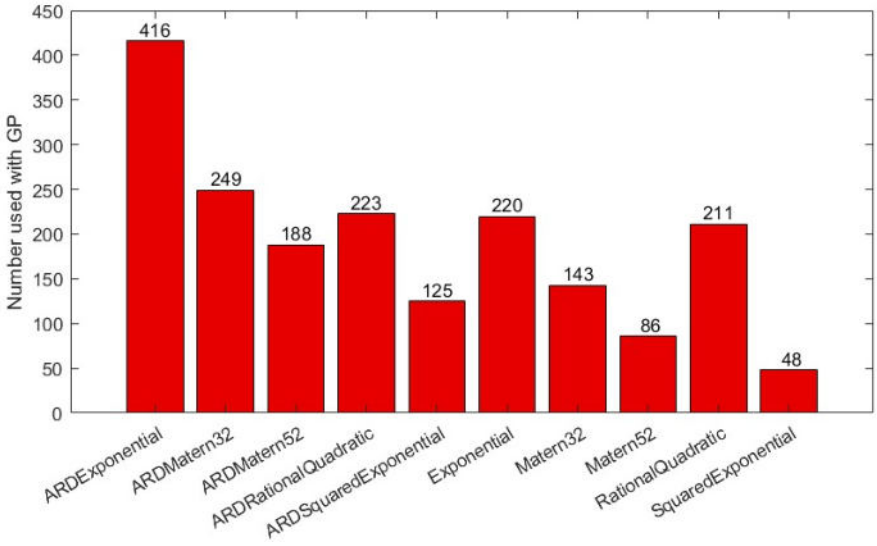


Figure 19: Kernels used with respect to GP models

In order to increase the level of clarity and explainability of the Gaussian Process output, the implementation of statistical methods similar to those applied in (Giudici and Raffinetti, 2023) and in (Giudici, Centurelli and Turchetta, 2024) should be considered for further research developments.
At the current state of this study, the gap between the prediction obtained from the Machine Learning method and the output computed by a spline interpolation of the zero rates points has been used as a potential control for the responses. If this gap were higher than a prefixed threshold, the anomaly would be reported to the analyst who can thus control and intervene in the choice of the method to be applied to efficiently solve the regression.
Another potential improvement to the model could be implemented combining different kernels by adding (and/or multiplying) them together to model the term structure, starting with those kernels that appeared as the most promising: a linear combination of ARD kernels. The reason for this suggestion is to increase the fitting potential of the regressive methodology as it would entail a higher level of adaptability compared to the one obtained from the implementation of a single kernel. The better the fitting is, the better the estimation of the discount factors and of the implied forwards rates are.

References

[1] Adams, K. J., Van Deventer D. R. (1994). "Fitting Yield Curves and Forward Rate Curves with Maximum Smoothness", Journal of Fixed Income, Vol. 4, No. 1, 52-62.
[2] Annaert, J., Claes, A. G. P., de Ceuster, M. J. K., Zhang, H. (2013). "Estimating the Yield Curve Using the Nelson-Siegel Model: A Ridge Regression Approach". International Review of Economics & Finance, Vol. 27, 482-496.
[3] Bank of International Settlements (2005). "Zero-Coupon Yield Curves – Technical Documentation", BIS Paper No. 25.
[4] Barrett, W. R., Gosnell, T. F. Jr., Heuson, A. J. (1995). "Yield Curve Shifts and the Selection of Immunization Strategies", The Journal of Fixed Income, Vol. 5, No. 2, 53-64.
[5] Bliss, R., Fama, E. (1987). "The Information in Long-Maturity Forward Rates". American Economic Review, Vol. 77, 680-692.
[6] Cafferata A., Giribone P. G., Neffelli M., Resta M. (2019). "Yield curve estimation under extreme conditions: do RBF networks perform better?" – Chapter 22 in book: "Neural Advances in Processing Nonlinear Dynamic Signals" – Springer.
[7] Cafferata A., Giribone P. G., Resta M. (2018). "Interest rates term structure models and their impact on actuarial forecasting" – QFW18: Quantitative Finance Workshop 2018 (UniRoma3 – Rome).
[8] Cairns, A. J. G., Pritchard D. J. (2001). "Stability of Descriptive Models for the Term Structure of Interest Rates with Applications to German Market Data", British Actuarial Journal Vol. 7, Issue 3, 467-507.
[9] Caligaris O., Giribone P. G. (2015). "Modellizzare la curva dei rendimenti mediante metodologie di apprendimento artificiale: analisi e confronto prestazionale tra le tecniche regressive tradizionali e le reti neurali", AIFIRM Magazine Vol. 10, N. 3.
[10] de Pooter, M. (2007). "Examining the Nelson-Siegel Class of Term Structure Models", Tinbergen Institute Discussion Paper, IT 2007-043/4.
[11] De Rezende R.B., Ferreira M.S. (2013). "Modeling and Forecasting the Yield Curve by an Extended Nelson-Siegel Class of Models: A Quantile Autoregression Approach", Journal of Forecasting Vol. 32, Issue 2, p. 111–123.
[12] Diebold, F. X., Li, C. (2006). "Forecasting the Term Structure of Government Bond Yields", Journal of Econometrics, Vol. 130, Issue 2, 337-364.
[13] European Central Bank (2008). "The New Euro Area Yield Curves", Monthly Bulletin, (February 2008), 95-103.
[14] Fabozzi, F. J., Martellini, L., Priaulet, P. (2005). "Predictability in the Shape of the Term Structure of Interest Rates", The Journal of Fixed Income, Vol. 15, No. 1, 40-53.
[15] Fisher, M., Nychka, D., Zervos, D. (1994). "Fitting the Term Structure of Interest Rates with Smoothing Splines", Finance and Economics Discussion Series, Federal Reserve Board.
[16] Giribone P. G. (2023). "Notes on Quantitative Financial Analysis". AIFIRM Edizioni – Educational Book Series. ISBN: 979-12-80245-19-9.
[17] Giudici P., Centurelli M., Turchetta S. (2024). "Artificial Intelligence risk management". Expert Systems with Applications, Vol. 235, ISSN 0957-4174.
[18] Giudici P., Raffinetti E. (2023). "SAFE Artificial Intelligence in finance". Finance Research Letters, Vol. 56, ISSN 1544-6123.
[19] Gonzalvez J., Lezmi E., Roncalli T., Xu J. (2019). "Financial Applications of Gaussian Processes and Bayesian Optimization". Capital Markets and Asset pricing eJournal. url: https://api.semanticscholar.org/CorpusID:159451782.
[20] Gurkaynak, R. S., Sack, B., Wright, J. H. (2007). "The U.S. Treasury Yield Curve: 1961 to the Present", Journal of Monetary Economics Vol. 54, Issue 8, 2291-2304.
[21] Litterman, R. B., Scheinkman J. (1991). "Common Factors Affecting Bond Returns". Journal of Fixed Income, Vol. 1, No. 1, 54-61.
[22] Matérn B. (1960). "Spatial Variation". Meddelanden från Statens Skogsforskningsinstitut, Vol. 49, No. 5. Almänna Förlaget, Stockholm. Second Edition (1986), Springer-Verlag, Berlin.
[23] McCulloch (1971). "Measuring the Term Structure of Interest Rates". The Journal of Business, Vol. 44, Issue 1, 19-31.
[24] McCulloch (1975). "The tax-adjusted yield curve". The Journal of Finance, Vol. 30, Issue 3, 811-830.
[25] Mercer, J. (1909). "Functions of Positive and Negative Type, and Their Connection with the Theory of Integral Equations". Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 209, 415-446.
[26] Nelder J. A., Mead R. (1965). "A Simplex Method for Function Minimization". The Computer Journal Vol. 7, Issue 4, 308 – 313.
[27] Nelson, C., Siegel, A. F. (1987). "Parsimonious Modeling of Yield Curves", Journal of Business, Vol. 60, 473-489.
[28] Nocedal, J. (1980). "Updating Quasi-Newton Matrices with Limited Storage". Mathematics of Computation Vol. 35, Issue 151, 773-782.
[29] Rasmussen C. E., Williams C. K. I. (2006). "Gaussian Processes for Machine Learning". The MIT Press.
[30] Seber, G. A. F., Wild, C. J. (2003). "Nonlinear Regression", Wiley Series in Probability and Statistics.
[31] Shea G. (1984). "Pitfalls in Smoothing Interest Rate Term Structure Data: Equilibrium Models and Spline Approximations". Journal of Financial and Quantitative Analysis, Vol. 19, Issue 3, 253-269.
[32] Steeley, J. M. (1991). "Estimating the Gilt-Edged Term Structure: Basis Splines and Confidence Intervals", Journal of Business Finance & Accounting, Vol. 18, No. 4, 513-529.
[33] Svensson, L. E. O. (1994). "Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994". IMF Working Paper, WP/94/114, 1-49.
[34] Svensson, L. E. O. (1996). "Estimating the Term Structure of Interest Rates for Monetary Policy Analysis". Scandinavian Journal of Economics, Vol. 98 (1996), 163-183.
[35] Vasicek O. A., Fong H. G. (1982). "Term Structure Modeling Using Exponential Splines". The Journal of Finance, Vol. 37, Issue 2, 339-348.
[36] Waggoner, D. F. (1997). "Spline methods for extracting interest rate curves from coupon bond prices". Working Paper, No. 97-10, Federal Reserve Bank of Atlanta.