International Journal of Data and Network Science 7 (2023) 1033–1044

Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Global gold prices forecasting using Bayesian nonparametric quantile generalized additive model

Yudhie Andriyana[a*], Yollanda Nalita[b], Bertho Tantular[a], I Gede Nyoman Mindra Jaya[a] and Annisa Nur Falah[c]

[a]*Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia*
[b]*PT Pemeringkat Efek Indonesia (PEFINDO), Indonesia*
[c]*Doctoral Program of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia*

C H R O N I C L E

Article history:
Received: February 24, 2023
Received in revised format: April 12, 2023
Accepted: May 28, 2023
Available online: May 28, 2023

Keywords:
Gold prices
Bayesian
Quantile
Additive model

A B S T R A C T

Gold is one of the most attractive commodities and popular investments. Investment experts often recommend investing in gold because gold is one of the safest investments. It is a stable classic hedge, although the conditions of currency volatility or global markets are depreciated. However, the gold price fluctuations can be influenced by some other factors, such as the USD Index, which reflect and measure the strength of the US Dollar currency, and the Index of Dow Jones Industrial Average (DJIA) or a reflection of the political and economic conditions of the stock market. In this study, we conduct a global gold price forecast (USD) based on the USD Index, the DJIA Index, and the influence of time trends. Based on the data's characteristics, we face the fact that the data is nonlinear, contains outliers, and its pattern is not easy to specify parametrically. Due to the complexity of the model, we then propose a more flexible, robust modeling technique called the Bayesian Nonparametric Quantile Generalized Additive Model method. According to the results for the median case, the proposed method shows an accurate forecasting category due to the value of the Mean Absolute Percentage Error, MAPE <10 %.

© 2023 by the authors; licensee Growing Science, Canada.

1. Introduction

Gold is one of the most valuable precious metals in the world, which is often used as a medium of exchange in trade and financial standards in various countries. Investment experts often recommend investing in gold because gold is a classic means of hedging or a haven in conditions of volatile currency fluctuations (Gürgün & Ünalmış, 2014). The tendency of investors to buy gold is due to the rising price of gold in the long run and the ability of gold as a haven asset to fight against extreme currency movements (Reboredo, 2013). One of the important pieces of investing knowledge is predicting the future gold price. Forecasting future gold prices is a consideration before deciding to invest because this tendency is not only to make gold an investment option but also to increase profits and prevent losses. Gold price fluctuations can be influenced by several factors, including the movement of the US Dollar Index (USDX). This index number reflects and measures the strength of the US Dollar against six other major world currencies. The relationship between the United States dollar (USD) and gold is very close. Meanwhile, considering that the world gold price is traded in USD currency, if the US Dollar Index weakens, it will increase demand for commodities, including gold as an alternative investment asset to protect their wealth, which causes the price of gold to rise as well (Joy, 2011). The Dow Jones Industrial Average (DJIA), the world's largest stock exchange, reflects political and economic conditions, which also affect gold's price. The stock and gold markets are the two most important markets in the combination of assets (Zhang et al., 2011). Based on these factors, the researcher is interested in forecasting the price of gold on the USD Index, the DJIA Index, and the influence of the time trend.

* Corresponding author.
E-mail address: y.andriyana@unpad.ac.id (Y. Andriyana)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)
© 2023 by the authors; licensee Growing Science, Canada.
doi: 10.5267/j.ijdns.2023.6.002

We found that the gold price data (USD) from January 2000 to December 2019 contains a time trend component. It has a relatively high pattern of change and tends to be nonlinear. The relationship between the USD Index and the DJIA Index on the price of gold is challenging; moreover, it has some outliers. Hence, the nonparametric approach needs to be implemented. Due to the existence of several covariates, then we propose to use an additive model. Therefore, we apply the Generalized Additive Model (GAM). The GAM method is effective and flexible in nonlinear time series regression analysis (Dominici et al., 2002). However, in GAM modeling, it is known that the relationship between response variables and predictors shows that data contain outliers. Therefore, a robust regression technique is needed to overcome the outliers in the data. One of the robust techniques is quantile regression (Koenker & Bassett Jr, 1978). Based on the abovementioned factors, a precise, flexible, and efficient model is needed to predict the global gold price and build gold price classifications into several price groups.

2. Methods

2.1 Data Source

In this study, we use, as the response variable, the global gold price data, which is taken from the World Gold Council. Meanwhile, as the covariates, denoted by and respectively, the USD and DJIA Indexes are taken from Global Finacial Portal from investing.com. Although both data are taken from two different resources, the period of the data is the same, which started from January 2000 to July 2020. For the quantile levels for this study, we focus on three levels of quantiles (0.25, 0.5, and 0.75).

2.2 Generalized Additive Model

The Generalized Additive Model (GAM) is an extension of the Generalized Linear Model (GLM) (Dominici et al., 2002) (Hastie, 2017). The GAM method is modeled as an additive function of the response variable that accommodates the predictor variable's nonlinear influence without knowing the effect's shape explicitly. In contrast, the nonlinear effect can be approximated by smoothing the relationship structure between the response and predictor variables (Beck & Jackman, 1997). GAM assumes that the response variable follows an exponential family distribution and accommodates the smooth function on each predictor variable that can be estimated with various non-parametric estimators such as kernels, splines, and the Fourier series. In this study, we model univariate response variables $y_t \in \mathbb{R}$ based on the effect of several covariates $\boldsymbol{x}_t = (x_{t,1}, \cdots, x_{t,d}) \in \mathbb{R}^d$. In GAM, y_t it is assumed to be a smooth function. For $t = 1,2,\cdots,n$, $y_t = \mu(\boldsymbol{x}_t) + \varepsilon_t$, with $\mu(.)$ is an unknown smooth function and is the residual having exponential family distribution. GAM assumes that the link function of g is as follows (Gaillard et al., 2016):

$$g(\mu(\boldsymbol{x}_t)) = \beta_0 + f_1(x_{t,1}) + f_2(x_{t,2}) + f_t(x_{t,4}) + \cdots \tag{1}$$

with the smoothing function $f_j(.)$ compiled from the sum of the basis function b and the appropriate regression coefficients of β, so that it can be written as follows:

$$f_j(\boldsymbol{x}_t) = \sum_{k=1}^{q} \beta_{jk} b_{jk}(\boldsymbol{x}_t) \tag{2}$$

with q is the basis dimension chosen to avoid excessive smoothing. We use thin plate splines and cubic splines (Fasiolo et al., 2020).

2.3 Quantile Regression

Quantile regression is a generalization of the median regression, which is not affected by the presence of outliers (Koenker & Bassett Jr, 1978) . This method describes the τ - quantile ($\tau \in [0,1]$) of the response variable, conditionally on vector covariate, $\boldsymbol{x}$. The τ-th quantile of $\boldsymbol{x}$ is defined by $\mu_\tau(\boldsymbol{x})$ where the quantile itself is inverse from the cumulative distribution function $\mu_\tau(\boldsymbol{x}) = F^{-1}(y|\boldsymbol{x})$. Given a sample of size n, an approximation of $dF(y)$ with an empirical function $dF_n(y)$ leading to a quantile estimate is obtained by specifying the following:

$$\hat{\mu}_\tau(\boldsymbol{x}) = \underset{\mu}{argmin} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\{y_i - \mu(\boldsymbol{x}_i)\}, \tag{3}$$

where,

$$\rho_\tau(z) = (\tau - 1)z\mathbb{I}(z < 0) + \tau z\mathbb{I}(z \geq 0) \tag{4}$$

$\rho_\tau(.)$ known as "pinball loss", while $\mathbb{I}(.)$ is an indicator function. Pinball loss has the disadvantage that it is less efficient in computation and less optimal for smoothed generalization of loss, so pinball loss is replaced with Extended Log-F (ELF) loss which is a generalization of scaled pinball loss (Fasiolo et al., 2021):

$$\tilde{\rho}_\tau(y - \mu) = (\tau - 1)\frac{(y - \mu)}{\sigma} + \lambda \log\left(1 + e^{\frac{(y-\mu)}{\lambda\sigma}}\right) \tag{5}$$

where $\lambda > 0$ is the smoothness loss and $1/\sigma$ called the learning rate, which shows the relative weight of the ELF loss. The form of normalization $\exp(-\tilde{\rho})$ becomes:

$$\tilde{p}_F(y - \mu) = \frac{e^{-\tilde{\rho}_\tau(y-\mu)}}{\int e^{-\tilde{\rho}_\tau(y-\mu)}dy} = \frac{e^{(1-\tau)\frac{y-\mu}{\sigma}}\left(1 + e^{\frac{y-\mu}{\lambda\sigma}}\right)^{-\lambda}}{\lambda\sigma\text{Beta}[\lambda(1-\tau), \lambda\tau]}, \tag{6}$$

with $[.,.]$ is a Beta function. Eq. (6) is known as the density of the ELF distribution, which is an extension of the *log-f* distribution (Jones, 2008). The complexity of the model based on the ELF loss function can be controlled with an empirical Bayesian approach which can specifically represent the smooth regression between the regressors and the number of quantiles.

2.4 Bayesian Quantile Regression

In the quantile regression with the Bayesian framework, the complexity of the smooth and random effects are controlled using the prior distribution of the regression coefficients, which are described by $p(\boldsymbol{\beta})$. We define the mechanism for updating the corresponding posterior $p(\boldsymbol{\beta}|\boldsymbol{y})$. The prior distribution used is Gaussian smoothing prior $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \mathbf{S}^-)$, where $\mathbf{S}^-$ is the inverse of the matrix $\mathbf{S}^\gamma = \sum_{i=1}^m \gamma_i \mathbf{S}_j$, while $\mathbf{S}_j$ is a semi-definite positive matrix using the penalty the wiggliness from $\mu(\boldsymbol{x})$ which is scaled with positive parameters $\boldsymbol{\gamma} = \{\gamma_1, \dots, \gamma_m\}$. $\mathbf{S}_j$ is known while the vector $\boldsymbol{\gamma}$ is selected (Fasiolo et al., 2021). $\mathbf{S}_j$ is related to the smooth effect which has a more complex structure and aims as a smoothing penalty, while $\boldsymbol{\gamma}$ as a smoothing parameter vector with several elements that control the prior accuracy of the smooth effect. However, the direct application of Bayes' rule is hindered by the fact that we are basing on quantile regression based on ELF loss rather than on the probability model for the observed density $p(\boldsymbol{y}|\boldsymbol{\beta})$, so there is a missing in the likelihood function. These constraints were overcome by adopting the Belief Updating (BU) framework and also combining it with a calibration method that explicitly aims to achieve good frequency properties (Fasiolo et al., 2021) (Bissiri et al., 2016). The BU framework is based on updating the prior to posterior distribution using a loss function, not the likelihood function, minimizing the expected loss:

$$\text{E}\{L(\boldsymbol{\beta})\} = \int L(y; \boldsymbol{\beta})\, f(y)dy, \tag{7}$$

with, $L(.;.)$ is a general loss function $f(y)$ and is the Probability Density Function (p.d.f) of y. For example, belief prior $\boldsymbol{\beta}$ with prior density $p(\boldsymbol{\beta})$, with some data y, then the coherent approach to update $p(\boldsymbol{\beta})$ is posterior with the following formula (Bissiri et al., 2016).

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto e^{-\frac{1}{\sigma}\sum_{i=1}^n L(y_i, \beta)} p(\boldsymbol{\beta}) \tag{8}$$

where, $\sum_{i=1}^n L(y_i, \beta)$ is an estimate of Eq. (7). Eq. (8) is called the Posterior Gibbs as a constant negative normalization of marginal loss (Syring & Martin, 2019).

The quantile regression based on ELF loss fits perfectly into this Belief Updating coherent framework because the Gibbs Posterior corresponds to a quantile regression loss function based on the ELF loss.

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto \prod_{i=1}^n \tilde{p}_F\{y - \mu(\boldsymbol{x}_i)\} p(\boldsymbol{\beta}) \tag{9}$$

where, $\tilde{p}_F$ is the ELF density implicitly dependent on τ , λ , σ and β.

2.5 Bayesian Non-parametric Quantile Generalized Additive Model

The analytical method used in this research is the Quantile Generalized Additive Model method using the Bayesian approach, so which is also known as the Bayesian Nonparametric Quantile Generalized Additive Model. Based on Eq. (1) in the quantile context, the τ-th Quantile Generalized Additive Model for global gold prices is as follows:

$$\mu_\tau(\boldsymbol{x}_t) = \beta_{0,\tau} + f_1(x_{t,1}) + f_2(x_{t,2}) + f_3(t) + f_4(s_t) \tag{10}$$

where $\beta_{0,\tau}$ is the intercept in the th-quantile model $\tau, f_1(x_{t,1})$ and $f_2(x_{t,2})$ are the smooth effects for the predictor variable $x_{t,1}$ and $x_{t,2}$ which is constructed using the thin plate splines function, $f_3(t)$is the subtle effect for the trend factor, which in this case uses the time *t* which is also constructed using the thin plate splines function. The trend factor is also seen using the time of year $f_4(s_t)$ to see the movement of gold prices from year to year. $f_4(s_t)$is a cyclic effect constructed using the cubic spline function. The cyclic effect on the year factor in order for the smoothing effect has adjusted from year to year (Fasiolo et al., 2020).

In Bayesian Non-parametric Quantile Generalized Additive Model modeling, three iteration processes that are nested and interrelated, generally explained (Fasiolo et al., 2020):

Iteration in: maximum estimation of a posterior MAP regression coefficient

In general, quantile regression, for example, $\mu(\boldsymbol{x}_t) = \mathbf{x}_t^T\boldsymbol{\beta}$ with $\mathbf{x}_t$ is the *t*-th row matrix $\mathbf{X}_{(nxd)}$ containing the basis splines function evaluated on the *t*-th $\boldsymbol{x}_t$ covariate vector. In this process, there is a work process of maximizing the posterior Gibbs logarithm of the ELF loss in Eq. (9) when $\boldsymbol{\gamma}$ and σ_0 is given. Regression coefficient estimates can be obtained by minimizing the following criteria:

$$\tilde{V}_D\{\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_0\} = \sum_{t=1}^{n} \mathrm{Dev}_t\,\{\boldsymbol{\beta}, \sigma(\boldsymbol{x}_t)\} + \sum_{j=1}^{m} \gamma_j\, \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{S}_j\boldsymbol{\beta} \tag{11}$$

where $\mathrm{Dev}_t\{\boldsymbol{\beta},\ \sigma(\boldsymbol{x}_t)\}$ is the *t*-th deviation component based on the ELF density in Eq. (5). Furthermore, the regression coefficient can be estimated using the Penalized Iteratively Reweighted Least Square (PIRLS) algorithm by minimizing:

$$\sum_{t=1}^{n} w_t\{z_t - \mu_t\}^2 + \sum_{j=1}^{m} \gamma_j\, \boldsymbol{\beta}^T\mathbf{S}_j\boldsymbol{\beta} \tag{12}$$

where $z_t = \mu_t - \frac{1}{2w_t}\frac{\partial \mathrm{Dev}_t}{\partial \mu_t}$; $w_t = \frac{1}{2}\frac{\partial^2 \mathrm{Dev}_t}{\partial \mu_t^2}$; $\mu_t = \mathbf{x}_t^T\boldsymbol{\beta}$, and $\mathrm{Dev}_t = \mathrm{Dev}_t\{\boldsymbol{\beta}, \sigma(\boldsymbol{x}_t)\}$.

Iteration between: Smoothing Parameter selection and ELF Smoothness Loss

Selection of smoothing parameters $\boldsymbol{\gamma}$ with σ_0 fixed, namely maximizing Laplace's approximation based on ELF loss (Fasiolo et al., 2020):

$$\log p\{\mathbf{y}|\boldsymbol{\gamma}, \sigma_0\} = \int \tilde{p}_\tau\,\{y_i - \mu(\boldsymbol{x}_i)\}p(\boldsymbol{\beta}|\boldsymbol{\gamma})d\boldsymbol{\beta} \tag{13}$$

In particular, Eq. (13) is an integral that is difficult to solve so it can be approached using the Laplace Approximate Marginal Loss (LAML) criteria:

$$G_L\{\boldsymbol{\gamma}, \sigma_0\} = -\frac{1}{2}\tilde{V}_D\{\widehat{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \sigma_0\} + \tilde{\Pi} - \frac{1}{2}[\log|\mathbf{X^TWX} + \mathbf{S}^\gamma| - \log|\mathbf{S}^\gamma|_+] + \frac{M_p}{2}\log(2\pi) \tag{14}$$

where $\tilde{\Pi}$ is the saturated loss $\tilde{\Pi} = -(1-\tau)\lambda\log(1-\tau) - \lambda\tau\log(\tau)$, $\mathbf{W}$ is the diagonal matrix thus $\mathrm{W}_{tt} = w_t$, M_p is the dimension of the null space of $\mathbf{S}^\gamma$ and $|\mathbf{S}^\gamma|_+$ Product of the non-zero positive eigenvalues of the matrix $\mathbf{S}^\gamma$. LAML can be maximized efficiently in smoothing parameter selection $\boldsymbol{\gamma}$ using the outer Newton algorithm (Müller, 2013).

Before the outer iteration, we define the ELF loss smoothness λ and the x dependent component of the learning rate $\tilde{\sigma}(\boldsymbol{x})$ when σ_0 fixed. Consideration of heteroscedasticity data treatment is required, assuming the response variable follows the scale-location model $y|\boldsymbol{x} \sim \alpha(\boldsymbol{x}) + k(\boldsymbol{x})z$, where z_t is i.i.d with E $(z|\boldsymbol{x}) = 0$; $var(z|\boldsymbol{x}) = 1$. Suppose $\tilde{h}_z^*$ that is the optimal bandwidth for regression z on x, so that the appropriate optimal bandwidth for y is $\tilde{h}^*(\boldsymbol{x}) = \tilde{h}_z^* k(\boldsymbol{x})$. In a context $\tilde{h}_z^* k(\boldsymbol{x}) = \lambda\sigma$ where one of the right sides must depend on $\boldsymbol{x}$. By choosing $\sigma(\boldsymbol{x}) = \sigma_0 \tilde{\sigma}(\boldsymbol{x})$, by σ_0 is the learning rate and $\tilde{\sigma}(\boldsymbol{x})$ the component that it depends on $\boldsymbol{x}$. Based on the scale-location model, the Asymptotic MSE (AMSE) of the regression coefficient is minimized by (Fasiolo et al., 2021):

$$\tilde{h}^*(\boldsymbol{x}) = \left[\frac{d\ 9 f_z\{F_z^{-1}(\tau)\}}{n\pi^4 f_z'\{F_z^{-1}(\tau)\}^2}\right]^{\frac{1}{3}} k(\boldsymbol{x}) \tag{15}$$

where $\tilde{h}^*(\boldsymbol{x})$ is the degree of loss smoothness, d is the Dimension $\boldsymbol{\beta}$, f_z is the Probability density function (p.d.f) of z, f_z' is the first derivative p.d.f of z, F_z is the cumulative distribution function (c.d.f) z, $F_z^{-1}(\tau)$ and is the τ-th quantile of z. Based on Equation (15) and it is known $n^{-1}\sum_{t=1}^{n} \tilde{\sigma}(\boldsymbol{x}_t) = 1$ that the smoothness loss λ can be formulated as follows:

$$\hat{\lambda} = n^{-1}\sum_{t=1}^{n} \tilde{h}^*(\boldsymbol{x}_t)\ /\sigma_0. \tag{16}$$

Smoothness loss $\hat{\lambda}$ is a smoothed generalization of loss.

Outer iteration: Learning rate selection

Calibration σ_0 is a σ_0 selection process that aims to obtain credible, calibrated for quantile functions. Calibration σ_0 with a velocity $1/\sigma_0$ that determines the loss and prior relative weights of the posterior Gibbs in Eq. (9). Learning rate selection with a calibration procedure based on the Integrated Kullback-Leibler (IKL) estimation, which is smooth and convex ensures that the quantile interval obtained is reliable and includes the correct frequency, which is achieved by minimizing (Fasiolo et al., 2021):

$$\widehat{\mathrm{IKL}}(\sigma_0) = n^{-1}\sum_{t=1}^{n}\left[\frac{\hat{\tilde{v}}(\boldsymbol{x}_t)}{v(\boldsymbol{x}_t)} + \log\frac{v(\boldsymbol{x}_t)}{\hat{\tilde{v}}(\boldsymbol{x}_t)}\right]^{\zeta} \tag{17}$$

Suppose that the posterior covariance matrix for $\widehat{\boldsymbol{\beta}}$ is $\mathbf{V} = (\boldsymbol{\mathcal{J}} + \mathbf{S}^{\gamma})^{-1}$ and $\mathbf{V} = (\boldsymbol{\mathcal{J}} + \mathbf{S}^{\gamma})^{-1}$, with $v(\boldsymbol{x}) = \mathbf{x}^{\mathrm{T}}\,\mathbf{V}\,\mathbf{x}$ and $\tilde{v}(\boldsymbol{x}) = \mathbf{x}^{\mathrm{T}}\,\widetilde{\mathbf{V}}\,\mathbf{x}$ is the posterior variance $\mu(\boldsymbol{x})$of the two posterior covariance matrices for $\widehat{\boldsymbol{\beta}}$. Equation (17) $\widetilde{\mathbf{V}}$ is replaced with $\hat{\widetilde{\mathbf{V}}} = \left(\boldsymbol{\mathcal{J}}\widehat{\boldsymbol{\Sigma}}_{\nabla}^{-1}\boldsymbol{\mathcal{J}} + \boldsymbol{S}^{\gamma}\right)^{-1}$ that is deterministic because it can be minimized efficiently using the bisection method so that the ups and downs of $\sigma_0$lead to smoothness and fluctuation of $\tilde{v}(\boldsymbol{x})/v(\boldsymbol{x})$.

Minimizes the difference between the posterior marginal for $\mu(\boldsymbol{x})$ based on $\mathbf{V}$ and $\widetilde{\mathbf{V}}$ has good asymptotic frequency properties. However, $\widetilde{\mathbf{V}}$ does not cause the posterior to reach the lowest asymptotic risk (Wood et al., 2016). Therefore, intervals based on the true marginal variance of $\hat{\mu}(\boldsymbol{x}) = \mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ in objective odds of $\mathbb{P}$ provide better coverage even in a small sample. $\mathbb{P}[\mu^0(\boldsymbol{x}) \in C_\alpha\{\sigma_0, \boldsymbol{y}\}] \approx \alpha$ is an objective probability measurement based on the data generalization process, where $C_\alpha\{\sigma_0, \boldsymbol{y}\}$ is the credible interval $\mu(\boldsymbol{x})$ for the level $\alpha \in (0,1)$ and $\mu^0(\boldsymbol{x})$ is the true quantile. The assumption that τ is fixed, λ and $\sigma(\boldsymbol{x})$ are functions of σ_0 so that IKL loss can be estimated through the bootstrap algorithm and obtained by IKL loss estimation:

$$\widehat{\mathrm{IKL}}_{\mathbb{P}}(\sigma_0) = n^{-1}\sum_{t=1}^{n}\left[\frac{\widehat{\mathrm{var}}\{\hat{\mu}(\boldsymbol{x}_t)\}}{v(\boldsymbol{x}_t)} + \log\frac{v(\boldsymbol{x}_t)}{\widehat{\mathrm{var}}\{\hat{\mu}(\boldsymbol{x}_t)\}} + \frac{1}{v(\boldsymbol{x}_t)}\{\hat{\mu}_t^0 - \bar{\mu}(\boldsymbol{x}_t)\}^2\right]^{\gamma} \tag{18}$$

where $\bar{\mu}(\boldsymbol{x}_t)$ and $\widehat{\mathrm{var}}\{\hat{\mu}(\boldsymbol{x}_t)\}$ are the mean and variance of the bootstrap quantile vector $\hat{\mu}_t^1, \ldots, \hat{\mu}_t^k$.

3. Result and Discussion

3.1 Model Specification

Before specifying the model, we divided global gold price data into 2 parts, training data, taken from January 2000 up to December 2019, and testing data from January 2020 to July 2020. Firstly, we check the pattern of the data. We plot the data to get the pattern of the relationship between the predictors and the response variable, which are presented in Fig. 1 and Fig. 2.

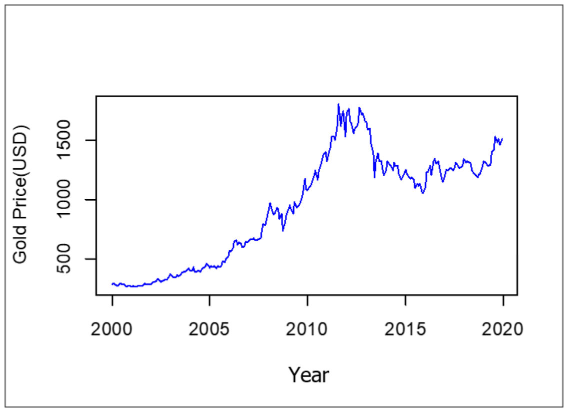
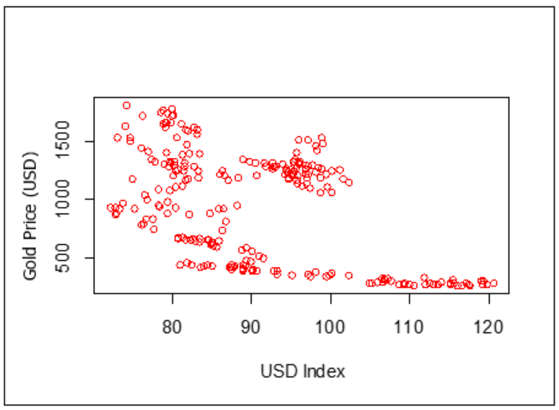
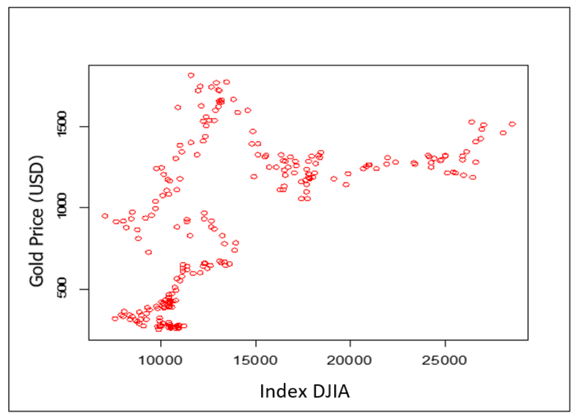


Fig 1. Gold Price Monthly Plot (USD) from 2000 to 2019





(a) (b)

Fig. 1. Scatterplot of Gold Price and USD Index (a) and Gold Price and DJIA Index (b)

Based on the data plot of the gold price (USD) per ounce in Fig. 1, the data contains a time trend component and tends to be non-linear. It is proven based on the linearity test of terracotta with a significant p-value = 0.009517 on $\alpha = 0.05$, so that the smoothing function is included for the time variable which is explained by the time variable *t* and the time variable of year s_t. Fig. 2 shows that the relationship pattern of the USD Index and DJIA Index variables with the global gold price is difficult to form parametrically. The nonparametric approach is used due to the difficulty of data patterns prespesification. Hence, we use a nonparametric approach which is more flexible. In this study, using the GAM approach. In the GAM approach, the residual plot of the mean GAM approach is used to see outliers from the distribution of data. The residuals from GAM are also useful in seeing whether the model can adequately explain the distribution of the data. Based on the boxplot in Fig. 3, it can be seen that the relationship between the response variables and predictors shows the existence of outliers.

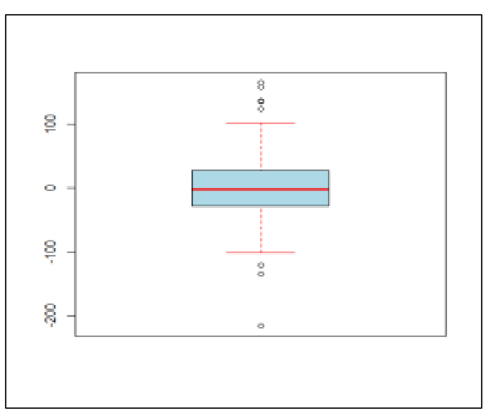


Fig. 3. Boxplot Residual Modeling with Generalized Additive Model

Fig. 3 shows some outliers. We then need to use a more robust technique where in this case, we implement quantile regression for Generalized Additive Model (GAM). The complexity of the quantile GAM based on ELF loss controlled by the Bayesian approach (Fasiolo et al., 2021).

3.2 Bayesian Non-parametric Quantile Generalized Additive Model

The movements of global gold price fluctuations that are influenced by the USD Index, the DJIA Index, and the effect of time trends are shown in three quantile regression models ($\tau = 0.25, 0.50,$ and 0.75). Based on Eq. (2) and Eq. (10) by entering the splines base smoothing function for each predictor variable, the following model is obtained:

$$\mu_\tau(\boldsymbol{x}_t) = \beta_{0,\tau} + \sum_{k=1}^{q_1} \beta_{1,k} b_{1,k}(x_{t,1}) + \sum_{k=1}^{q_2} \beta_{2,k} b_{2,k}(x_{t,2}) + \sum_{k=1}^{q_3} \beta_{3,k} b_{3,k}(t) + \sum_{k=1}^{q_4} \beta_{4,k} b_{4,k}(s_t) \tag{19}$$

with $b_{1,k}(x_{t,1})$, $b_{2,k}(x_{t,2})$, and $b_{3,k}(t)$ are the basis for thin plate splines with $b_{j,k}(\boldsymbol{x}_t) = \left\|\boldsymbol{x}_t - \boldsymbol{x}_{t_k}\right\|^2 \log\left\|\boldsymbol{x}_t - \boldsymbol{x}_{t_k}\right\|$, whereas $b_{4k}(s_t)$ is the basis for cubic splines with $b_{4,k}(s_t) = \left(s_t - s_{t_u}^*\right)_+^3$. The basic dimensions for each variable are obtained, namely $q_1 = q_2 = q_3 = 9$ and $q_4 = 8$. We model the variance $k(\boldsymbol{x})$ using a cyclic effect for the year variable s_t on the Gaussian GAM location-scale with the basis of cubic splines.

3.3 Estimated Parameters

The parameter estimations are optimized simultaneously with a nested iteration process in R-package "qgam". Firstly, the learning rate parameter can be selected by calibration. The learning rate at each quantile value is shown in Fig. 4:

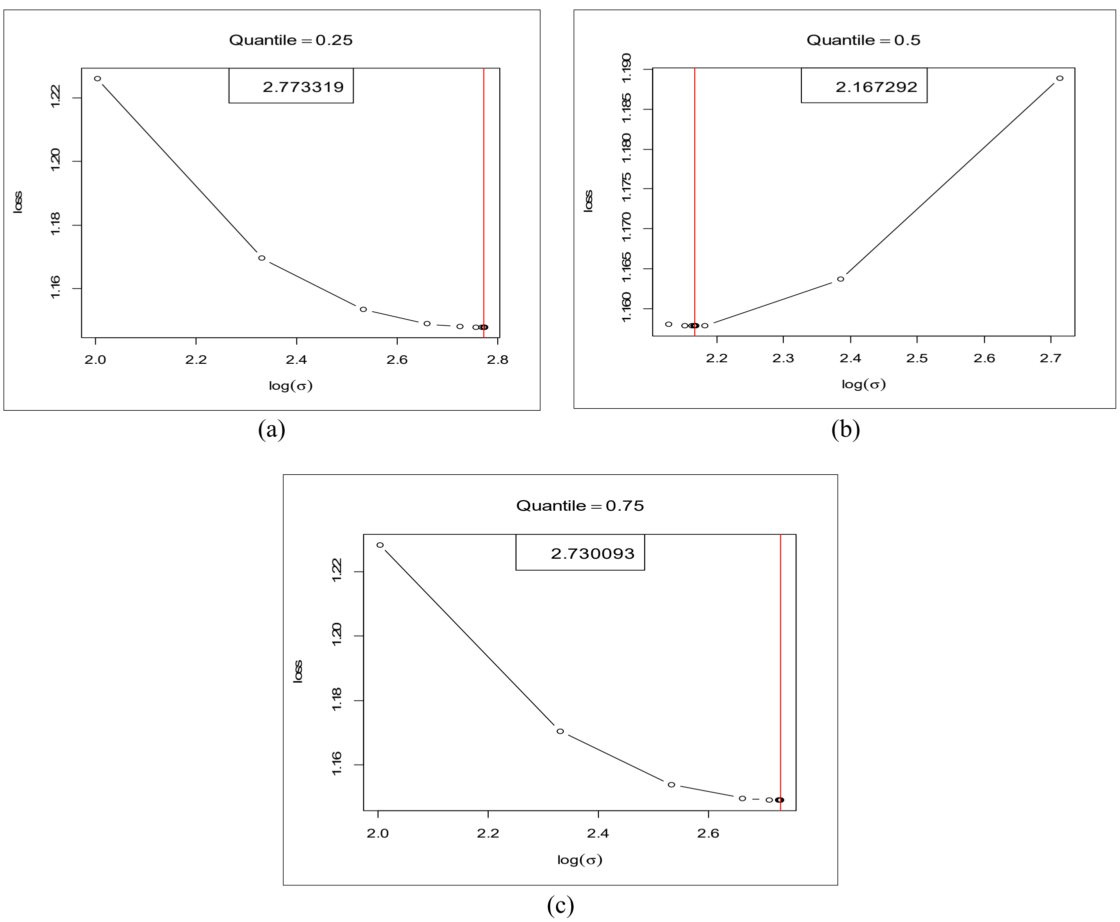


Fig. 4. Learning Rate in 0.25th (a), $0.25th$ (b), and $0.75th$ (c) quantiles

Based on Fig. 4, it is shown that in $\tau = 0.25$ the optimum learning rate is 2.773319, in $\tau = 0.50$ the optimum learning is 2.167292, and in $\tau = 0.75$ we obtain the optimum learning rate 2.730093. In the intermediate iteration process, the smooth generalization of the ELF loss function is defined by $\tilde{h}^*(\boldsymbol{x}_t)$. The optimum parameters $\hat{\lambda}$, are obtained based on Eq. (16), and the range of smoothing parameters $\boldsymbol{\gamma}$ are obtained based on the outer Newton algorithm by minimizing LAML which can be summarized in Table 1.

Table 1

Smoothness Loss $\hat{\lambda}$ and Range $\hat{\gamma}$ Optimum in 0.25,0.50, and 0.75 quantiles

Quantile ($\boldsymbol{\tau}$)	$\hat{\boldsymbol{\lambda}}$ Optimum	Range $\hat{\boldsymbol{\gamma}}$ Optimum
$\mathbf{0.25}$	1.23831	$[-6.38319 \times 10^{-8};\ 1.014285 \times 10^{-7}]$
$\mathbf{0.50}$	6.63386	$[-7.946099 \times 10^{-6};\ 0.000282722]$
$\mathbf{0.75}$	1.28482	$[9.689726 \times 10^{-7};\ 1.611026 \times 10^{-5}]$

The optimum regression coefficients are then obtained by minimizing the penalized ELF loss, which is minimized through PIRLS. The results are presented in Table 2.

Table 2

Estimated Coefficients for each quantile levels (0.25, 0.50, and 0.75)

The Value of Estimated Parameters			
$\boldsymbol{\tau = 0.25}$			
Intercept	896.200		
$\hat{\boldsymbol{\beta}}_{1,k}$	$\hat{\boldsymbol{\beta}}_{2,k}$	$\hat{\boldsymbol{\beta}}_{3,k}$	$\hat{\boldsymbol{\beta}}_{4,k}$
$\hat{\boldsymbol{\beta}}_{1,1}$ = 53.445	$\hat{\beta}_{2,1}$ = -113.871	$\hat{\beta}_{3,1}$ = -229.587	$\hat{\beta}_{4,1}$ = 43.922
$\hat{\boldsymbol{\beta}}_{1,2}$ = -101.665	$\hat{\beta}_{2,2}$ = -331.817	$\hat{\beta}_{3,2}$ = -939.635	$\hat{\beta}_{4,2}$ = -31.958
$\hat{\boldsymbol{\beta}}_{1,3}$ = 28.817	$\hat{\beta}_{2,3}$ = 34.740	$\hat{\beta}_{3,3}$ = 645.675	$\hat{\beta}_{4,3}$ = -6.710
$\hat{\boldsymbol{\beta}}_{1,4}$ = 36.639	$\hat{\beta}_{2,4}$ = 337.441	$\hat{\beta}_{3,4}$ = -660.608	$\hat{\beta}_{4,4}$ = -4.346
$\hat{\boldsymbol{\beta}}_{1,5}$ = 2.397	$\hat{\beta}_{2,5}$ = 98.879	$\hat{\beta}_{3,5}$ = 85.323	$\hat{\beta}_{4,5}$ = 9.053
$\hat{\boldsymbol{\beta}}_{1,6}$ = 26.327	$\hat{\beta}_{2,6}$ = -143.074	$\hat{\beta}_{3,6}$ = -117.296	$\hat{\beta}_{4,6}$ = -139.080
$\hat{\boldsymbol{\beta}}_{1,7}$ = -15.702	$\hat{\beta}_{2,7}$ = 46.904	$\hat{\beta}_{3,7}$ = 83.296	$\hat{\beta}_{4,7}$ = -70.316
$\hat{\boldsymbol{\beta}}_{1,8}$ = -132.286	$\hat{\beta}_{2,8}$ = -589.130	$\hat{\beta}_{3,8}$ = -1016.665	$\hat{\beta}_{4,8}$ = 116.449
$\hat{\boldsymbol{\beta}}_{1,9}$ = -114.025	$\hat{\beta}_{2,9}$ = 83.190	$\hat{\beta}_{3,9}$ = 600.360	
$\boldsymbol{\tau = 0.50}$			
Intercept	935.519		
$\hat{\boldsymbol{\beta}}_{1,k}$	$\hat{\boldsymbol{\beta}}_{2,k}$	$\hat{\boldsymbol{\beta}}_{3,k}$	$\hat{\boldsymbol{\beta}}_{4,k}$
$\hat{\boldsymbol{\beta}}_{1,1}$ = 60.946	$\hat{\beta}_{2,1}$ = -124.001	$\hat{\beta}_{3,1}$ = -181.791	$\hat{\beta}_{4,1}$ = 5.355
$\hat{\boldsymbol{\beta}}_{1,2}$ = -81.857	$\hat{\beta}_{2,2}$ = -196.234	$\hat{\beta}_{3,2}$ = -978.577	$\hat{\beta}_{4,2}$ = -49.162
$\hat{\boldsymbol{\beta}}_{1,3}$ = 29.987	$\hat{\beta}_{2,3}$ = -6.892	$\hat{\beta}_{3,3}$ = 687.782	$\hat{\beta}_{4,3}$ = -8.198
$\hat{\boldsymbol{\beta}}_{1,4}$ = 23.273	$\hat{\beta}_{2,4}$ = 213.651	$\hat{\beta}_{3,4}$ = -760.066	$\hat{\beta}_{4,4}$ = 20.788
$\hat{\boldsymbol{\beta}}_{1,5}$ = -2.045	$\hat{\beta}_{2,5}$ = 106.397	$\hat{\beta}_{3,5}$ = 103.877	$\hat{\beta}_{4,5}$ = 61.898
$\hat{\boldsymbol{\beta}}_{1,6}$ = 6.126	$\hat{\beta}_{2,6}$ = -25.835	$\hat{\beta}_{3,6}$ = -53.465	$\hat{\beta}_{4,6}$ = -100.648
$\hat{\boldsymbol{\beta}}_{1,7}$ = -8.894	$\hat{\beta}_{2,7}$ = 13.263	$\hat{\beta}_{3,7}$ = 57.957	$\hat{\beta}_{4,7}$ = -66.467
$\hat{\boldsymbol{\beta}}_{1,8}$ = -91.300	$\hat{\beta}_{2,8}$ = -311.100	$\hat{\beta}_{3,8}$ = -1126.526	$\hat{\beta}_{4,8}$ = 163.758
$\hat{\boldsymbol{\beta}}_{1,9}$ = -127.794	$\hat{\beta}_{2,9}$ = 12.773	$\hat{\beta}_{3,9}$ = 665.055	
	$\boldsymbol{\tau = 0.75}$		
Intercept	974.494		
$\hat{\boldsymbol{\beta}}_{1,k}$	$\hat{\boldsymbol{\beta}}_{2,k}$	$\hat{\boldsymbol{\beta}}_{3,k}$	$\hat{\boldsymbol{\beta}}_{4,k}$
$\hat{\boldsymbol{\beta}}_{1,1}$ = 62.432	$\hat{\beta}_{2,1}$ = -133.620	$\hat{\beta}_{3,1}$ = -226.175	$\hat{\beta}_{4,1}$ = -30.733
$\hat{\boldsymbol{\beta}}_{1,2}$ = -70.883	$\hat{\beta}_{2,2}$ = -113.730	$\hat{\beta}_{3,2}$ = -849.066	$\hat{\beta}_{4,2}$ = -61.231
$\hat{\boldsymbol{\beta}}_{1,3}$ = 23.863	$\hat{\beta}_{2,3}$ = -30.715	$\hat{\beta}_{3,3}$ = 666.645	$\hat{\beta}_{4,3}$ = -11.844
$\hat{\boldsymbol{\beta}}_{1,4}$ = 11.754	$\hat{\beta}_{2,4}$ = 124.648	$\hat{\beta}_{3,4}$ = -667.276	$\hat{\beta}_{4,4}$ = 33.437
$\hat{\boldsymbol{\beta}}_{1,5}$ = -7.084	$\hat{\beta}_{2,5}$ = 106.068	$\hat{\beta}_{3,5}$ = 38.802	$\hat{\beta}_{4,5}$ = 82.869
$\hat{\boldsymbol{\beta}}_{1,6}$ = -2.297	$\hat{\beta}_{2,6}$ = 42.071	$\hat{\beta}_{3,6}$ = 125.382	$\hat{\beta}_{4,6}$ = -46.668
$\hat{\boldsymbol{\beta}}_{1,7}$ = -0.279	$\hat{\beta}_{2,7}$ = -8.451	$\hat{\beta}_{3,7}$ = -5.142	$\hat{\beta}_{4,7}$ = -42.330
$\hat{\boldsymbol{\beta}}_{1,8}$ = -67.100	$\hat{\boldsymbol{\beta}}_{2,8}$ **= -157.615**	$\hat{\boldsymbol{\beta}}_{3,8}$ **= -892.125**	$\hat{\boldsymbol{\beta}}_{4,8}$ **= 193.570**
$\hat{\boldsymbol{\beta}}_{1,9}$ = -125.200	$\hat{\beta}_{2,9}$ = -15.232	$\hat{\beta}_{3,9}$ = 582.744	

Based on the estimated model, the contribution of each predictor for each quantile level can be seen in Fig. 5. As shown in Fig. 5, in general, both predictors have negative effects. This means that the bigger the USD Index, the smaller the global gold price. For the DJIA Index, in some parts, goes down and goes up at specific intervals. This means that the global gold price implications will differ at each interval. For the time trend, the time predictor t tends to have positive effects, which means that the global gold price will rise relatively over time. In addition to the time predictor *t*, the effect of the time trend has a more constant but wiggly effect. The cyclic effect becomes wider for the more significant time *t*.

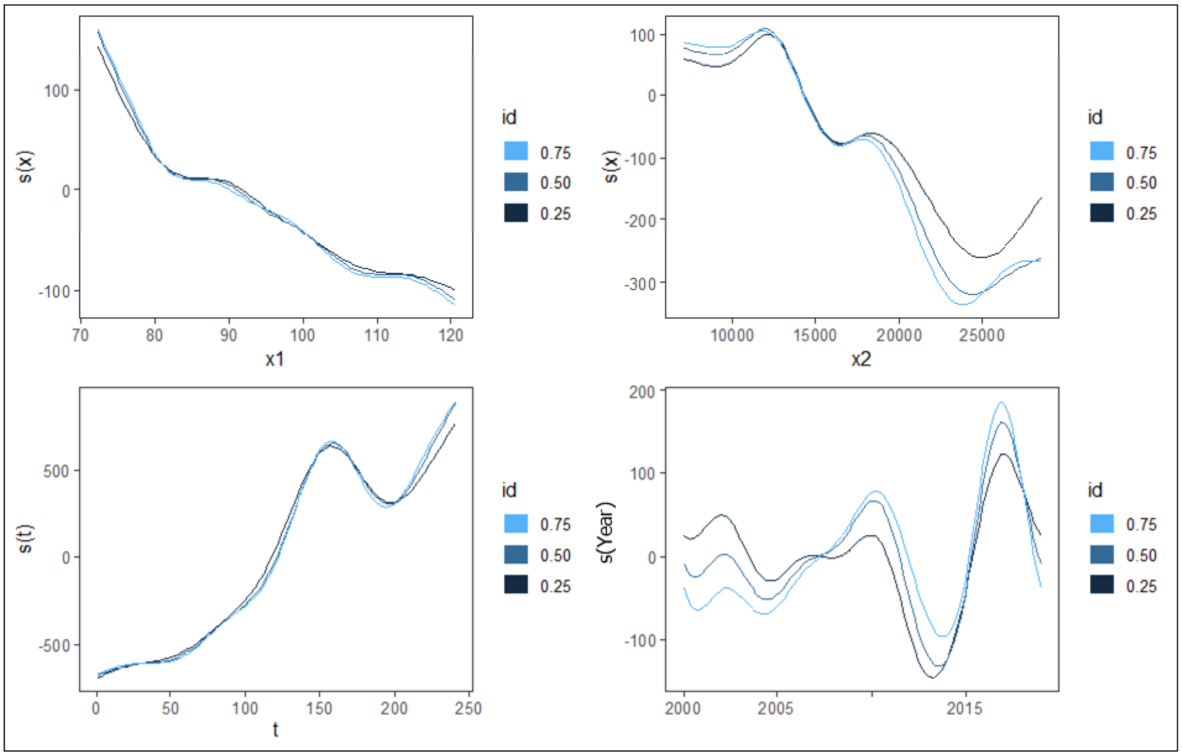


Fig. 5. Plot of the predictors for quantile levels

3.4 Model Analysis

We are now classifying the global gold price over time *t* at levels 0.25, 0.50, and 0.75 quantiles based on the influence of the USD Index, the DJIA Index, and the influence of the time trend. The estimated values of each quantile with the actual global gold price plot can be displayed in graphical form as presented in Fig. 6 below:

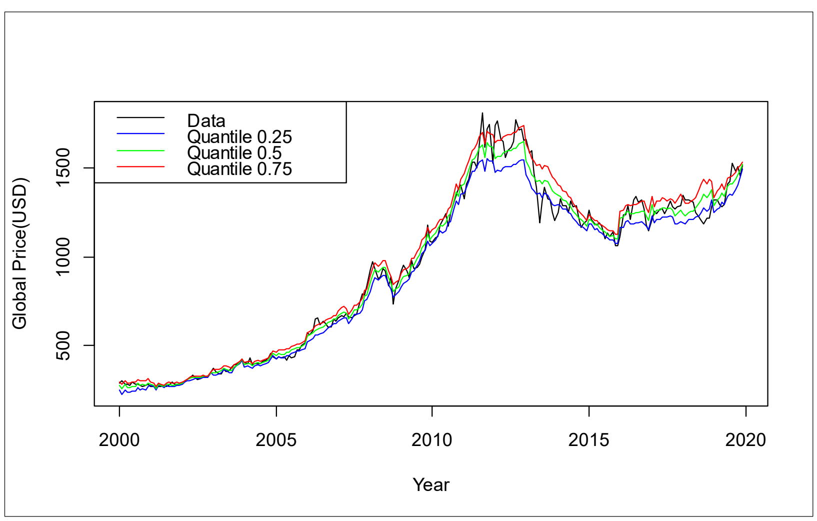
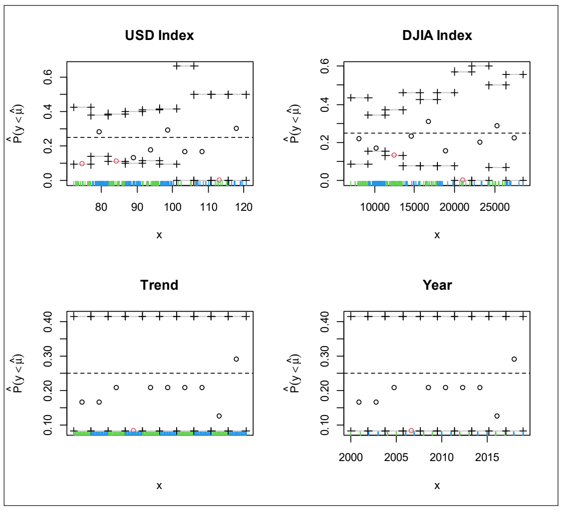


Fig. 6. Plot of Actual Data and Estimated Quantile Curve of Global Gold Prices

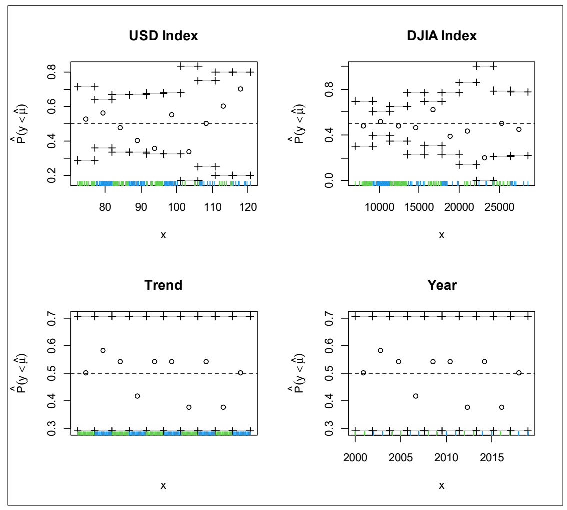
In this study, we assume that the global gold price can be grouped into low gold prices, which are located below the 0.25 quantile. The medium gold price is between the 0.25-0.50 quantile, the high gold price is between the quantile 0.50 - 0.75 quantile, and the price of gold is very high, located above the 0.75 quantiles. Based on Fig. 6, it can be seen that the global gold price, indicated by the solid black line, moves between the four categories. At the end of the year, it seems that the data presented in the solid black line are very high, indicated by its location above the 0.75*th* quantile.

3.5 Diagnostic and Model Forecasting

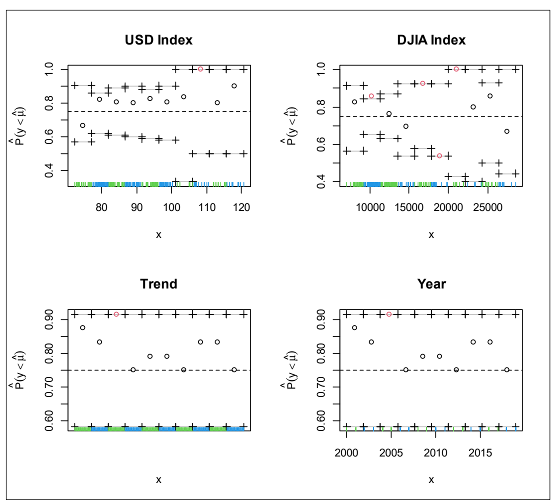
Around 18.75% of observations are located below the 0.25 quantile, 49.16% of observations below the 0.50 quantile, and 81.25% of observations under the 0.75 quantile level. The diagnostic plot for each quantile is presented based on the proportion of negative residuals, as can be seen in Fig. 7 below.



(a)



(b)



(c)

Fig. 7. Diagnostics Model for 0.25th Quantile (a), 0.50th Quantile (b) and 0.75th Quantile (c)

The model diagnostic in Fig. 7 shows the proportion of negative residuals in a sequence of bins for each predictor in each quantile model. The boundaries, denoted by (+++), describes the confidence intervals for $\mu_\tau(\boldsymbol{x}_t)$ based on the binomial probability. In the 0.50th quantile model, it can be seen that using confidence interval of 95%, all the observed points are within the curve boundaries, meaning there is no deviation. While in the 0.25*th* and the 0.75*th* quantile models, some observations are located outside the boundaries (indicated by some red dots). It indicates a deviation but not so big since the locations are still around the boundaries. The accuracy of the models is indicated by the number of observations located inside or outside the boundaries. The more data located outside or deviating away from the curve indicates the model's inaccuracy, where the predictors could be better enough in explaining the response variable. Based on Fig. 7, it can be seen that we have a good model.

The next step is to look at the estimated performance using the Mean Absolute Percentage Error (MAPE) value in the median case ($\tau = 0.50$) shown in Table 3.

Table 3
MAPE Values for Data Training and Data Testing for the Median case

	Data Training	Data Testing
MAPE	3.715568	9.820569

Based on Table 3, the MAPE value for training data and testing data is below 10%. MAPE on the training data is categorized as very low (3.72%), while on the testing data, we have 9.82%. It shows that we have an accurate estimated model since our forecasting is located at a very good level of accuracy for predicting the global gold price.

We are now presenting the forecasting of the global gold prices for the next seven months, from January to July 2020. Visually the comparison of the actual value and the predicted value of the gold price can be illustrated in Fig. 8 as follows (median case):

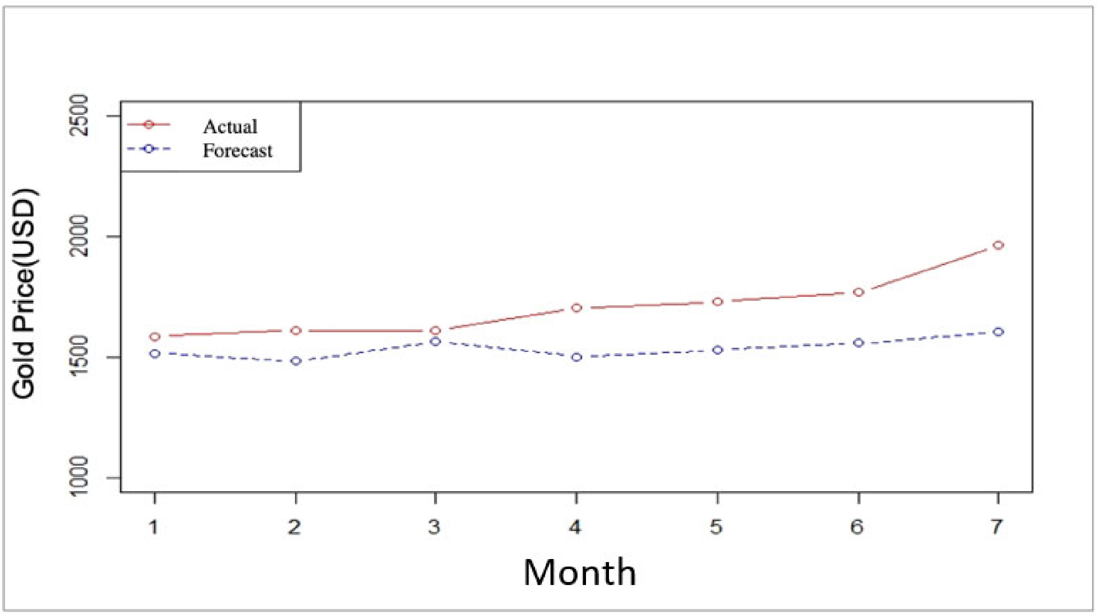


Fig. 8. Actual Data Plot and Global Gold Price Forecast for the Period of January-July 2020

Based on Fig. 8, it can be seen that the estimated model has a good performance, especially if it is implemented in a short term, for example, only for the first three months. The longer the distance of forecasting, the lower the performance, as can be seen from the figures that the forecasts after the first three months tend to under-forecast. However, although their distances become wider, the performances are still acceptable due to the MAPE values.

4. Conclusion

Based on the previous discussion, the Bayesian Nonparametric Quantile Generalized Additive Model fits nicely to describe the relationship between the USD Index and the DJIA Index between global gold prices. The influence of time trends can not be ignored since it has an effect on the estimated model. Forecasting global gold prices for the next seven months (January - July 2020) has a high forecasting accuracy or is in the very good forecasting category (<10%). This is indicated by the small MAPE value of 9.82% (Median case, $\tau = 0.50$).

Acknowledgments

The authors gratefully thank the Universitas Padjadjaran for supporting the research which is funded by RPLK scheme with the contract No: 1549/UN6.3.1/PT.00/2023. We also thank the reviewers for the valuable review for this paper.

References

Beck, N., & Jackman, S. (1997). Getting the mean right is a good thing: generalized additive models. *Society for Political Methodology Working Papers*.

Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series b, Statistical Methodology*, *78*(5), 1103.

Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, *156*(3), 193–203.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., & Goude, Y. (2020). qgam: Bayesian non-parametric quantile regression modelling in R. *ArXiv Preprint ArXiv:2007.03303*.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., & Goude, Y. (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, *116*(535), 1402–1412.

Gaillard, P., Goude, Y., & Nedellec, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, *32*(3), 1038–1050.

Gürgün, G., & Ünalmış, İ. (2014). Is gold a safe haven against equity market investment in emerging and developing countries? *Finance Research Letters*, *11*(4), 341–348.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249–307). Routledge.

Jones, M. C. (2008). On a class of distributions with simple exponential tails. *Statistica Sinica*, 1101–1110.

Joy, M. (2011). Gold and the US dollar: Hedge or haven? *Finance Research Letters*, *8*(3), 120–131.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Müller, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, *81*(5), 1805–1849.

Reboredo, J. C. (2013). Is gold a safe haven or a hedge for the US dollar? Implications for risk management. *Journal of Banking & Finance*, *37*(8), 2665–2676.

Syring, N., & Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, *106*(2), 479–486.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563.

Zhang, Q., Ma, J. H., & Wang, Y. (2011). Study on forecasting of gold price based on varying-coefficient regression model. *Key Engineering Materials*, *467*, 1398–1403.

© 2023 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).