International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 4, Issue 7, July – 2020, Pages: 70-81

On the efficiency of the weighted Generalized Cross-Validation and Unbiased Risk Smoothing Method for Time Series Observations with Autocorrelated Error

Samuel Olorunfemi Adams[1], Rueben Adeyemi Ipinyomi[2]

[1]Department of Statistics, University of Abuja, Nigeria.
[2]Department of Statistics, University of Ilorin, Nigeria.
Corresponding Author Email: samuel.adams@uniabuja.edu.ng

Abstract: *Spline Smoothing is a technique under the Non-parametric regression used to filter out noise in observations, it is one of the most popular methods used for the prediction of non-parametric regression models and its performance depends on the choice of smoothing parameters. Most of the past works applied to smooth methods to time series data, this method over fits data in the presence of Autocorrelation error. There are many methods of estimating smoothing parameters; most popular among them are; Generalized Maximum Likelihood (GML), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR), these methods tend to overfit smoothing parameters in the presence of autocorrelation error. An efficient new Spline Smoothing estimation method is proposed and compared with three classical methods to eliminate the problem of overfitting associated with the presence of Autocorrelation in the error term. It is demonstrated through a simulation study performed by using a program written in R based on the predictive Mean Score Error criteria. The result indicated that the predictive mean square error (PMSE) of the four smoothing methods decreases as the smoothing parameters increases and decrease as the sample sizes increases. This study discovered that the proposed smoothing method is the best for time-series observations with Autocorrelated error because it doesn't overfit and works well for large sample sizes. This study will help researchers overcome the problem of overfitting associated with applying Smoothing spline method time series observation.*

Keywords: Autocorrelation; generalized maximum likelihood; generalized cross-validation; splines smoothing; time series and unbiased risks.

1. Introduction

In non-parametric regression, smoothing is of great importance because it is used to filter out noise or disturbance in observation; it is commonly used to estimate the mean function in a nonparametric regression model, it is also the most popular methods used for prediction in non-parametric regression models, the general spline smoothing model is given as:

$$y_i = f(X_i) + \varepsilon_i \tag{1}$$

Where; y_i is the response variable, f is an unknown smoothing function, X_i is the observation values of the predictor variable and ε_i is zero mean Autocorrelated stationary process. The main objective of this research is to estimate f (.) when $x_i = t_i$ but not necessarily equally spaced, with $t_1 < . . . < t_n$ (time) and ε_i are assumed to be correlated. Therefore, this research shall consider the spline smoothing for non-parametric estimation of a regression function in a time-series context with the model;

$$y_i = f(t_i) + \varepsilon_{ti} \tag{2}$$

Where; Y_i = observation values of the response variable, f = an unknown smoothing function, t_i is the time for i = 1 . . . n, and e_{ti} = zero mean Autocorrelated stationary process.

Smoothing spline arises is the solution to a nonparametric regression problem having the function f(x) with two continuous derivatives that minimize the penalized sum of squares;

$$S(f) = \sum_{i-1}^{n}\left(y_i - f(x_i)\right)^2 + \lambda \int_0^1 \left(f^!(x_i)\right)^2 dx \tag{3}$$

Where; λ denotes a smoothing parameter, that is, the rate of exchange between residual error and roughness of the curve f, the parameter λ controls the trade-off between goodness-of-fit and the smoothness of the estimate. If λ is 0 then $f^!(x)$ simply interpolates the data, if λ is very large, then $f^!(x)$ will be selected so that $f^!(x)$ is 0, which implies a globally linear least-squares fit all data. Wahba et.al (1995). There are some many literatures on Spline Smoothing modeling of time series data in the presence autocorrelated error; Diggle and Hutchinson (1989), Yuedong (1998), Yuedong et. al. (2000), Opsomer, Yuedong and Yang

(2001), Wahba et. al. (1995), Carew et. al (2002), Hall and Keilegom (2003), Francisco-Fernandez and Opsomer (2005), Hart and Lee (2005), Krivobokova and Kauermann (2007), Shen (2008), Morton et.al. (2009), Wang, Meyer, and Opsomer (2013), Adams, Ipinyomi and Yahaya (2017).

This study aims to propose a new Smoothing method (PSM) by modifying two of the existing spline smoothing methods (i.e. the Generalized Cross-Validation (GCV) and Unbiased Risk (UBR)) and compare this modified smoothing methods with three existing estimation methods namely; Generalized Maximum Likelihood (GML), Generalized Cross-Validation (GCV) and Unbiased Risk (UBR) for time series observations in the presence of Autocorrelated error to eliminate the problem of overfitting associated with the presence of Autocorrelation in the error term. Section one presents the introduction to the study. Section two reviews the existing spline smoothing method and the proposed selection method, Section 3 presents the Materials and Methods, Monte Carlo simulation study, the equation used for generating values in simulation experimental design and data generation, section four compares the four methods via a simulation study, and finally, the result discussion and conclusion were presented in the last section.

2 Parameter Estimation

2.1 Generalized cross-validation (GCV) with autocorrelation structure

The term Generalized Cross-Validation (GCV) was coined by Wahba (1977) and applied by Hastie and Tibshirani, (1999), Aydin and Memmedli (2011) then, Diggle and Hutchinson (1989) and Wahba (1983) introduced the Autocorrelation structure in GCV, this is given as;

$$GCV(\lambda) = \frac{(y - \hat{g})^T V^{-1} (y - g)}{[trace(I - S_\lambda)]^2} \tag{4}$$

Where; (Sλ) = the ith diagonal element of smoother matrix, V = the correlation structure, y = $(y_1, \ldots, yn)^T$ and f = $(f(t1), \ldots, f(tn))^T$

2.2 Generalized maximum likelihood (GML) estimation method with Autocorrelation structure

The Generalized Maximum Likelihood (GML) estimation method is an empirical Bayes type criteria developed by Wecker and Ansley (1983) and Wahba (1985) while Yuedong (1998) proposed the GML methods for correlated observations with one smoothing parameter given by;

$$GML(\lambda) = \frac{\lambda' W(I - S_\lambda)}{[det^+ W(I - S_\lambda)]^{\frac{1}{n-m}}} \tag{5}$$

Where; $det^+ (I - S_\lambda)$ is the product of the n – m nonzero eigenvalues of $(I - S_\lambda)$, λ is Smoothing parameter, W is the correlation structure, S_λ is the diagonal element of the smoother matrix, n is $n_1 + n_2$, Pairs of measurement/observations and m is the number of zero eigenvalues.

2.3 Unbiased risk (UBR) estimation method with autocorrelation structure

The UBR method or CP criterion was suggested by C.L. Mallows' (1973) and had been applied successfully by Craven and Wahba (1979), Gu (1992); Wahba, Wang, Gu (1995); Klein, and Klein (1995) and (Wang, 1998), but Yuedong (1998) provides UBR method with a known Autocorrelation structure for selecting smoothing parameters for spline estimates with non-Gaussian data. It is written as;

$$UBR(\lambda) = \frac{\frac{1}{n}\left\| W^{\frac{k}{2}}(I - S_\lambda)\lambda \right\|^2}{\left[\frac{1}{n} trace\left(W^{k-1}(I - S_\lambda)\right)\right]^2} \quad k = 0,1,2 \tag{6}$$

Where; n is the pairs measurement/observations$\{x_i, y_i\}$, W is the correlation structure, λ is Smoothing parameters, S_λ = is the ith diagonal element of the smoother matrix.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 4, Issue 7, July – 2020, Pages: 70-81

2.4 The weighted GCV and UBR estimation method with autocorrelation structure

A Spline Smoothing model is defined as

$$y_i = f(x_i) + \varepsilon_i \tag{7}$$

Where; y is the variable of interest, x is a vector of the predictor variable, f is Regression function and $\varepsilon_i \sim N\left(N(0, \sigma W^{-1})\right)$.
There is some option to consider when the model (7) above is to be used to take care of non-linearity, they include; Data transformation, additive terms e.g. quadratic or cubic term, and Spline smoothing. This study is interested in Spline Smoothing because it considers non-linearity based on the regression curve by introducing a kink or bends in $\hat{Y}$, these kinks are produced by hinge function and the point of bend on the fit is called knots.
The main purpose of the conventional regression analysis is to minimize the residual sum of Square (RSS), the model with the minimum RSS is the best model. It is worthy to note that in Spline Smoothing, a method of selection known as Cross-Validation (CV) was proposed by Wahba (1979). In place of RSS in the conventional simple regression analysis, the error term is defined as;

$$\varepsilon = y_i - \hat{y}_i \tag{8}$$

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{9}$$

$Y_i = f(x_i)$, for the observed and $\hat{Y}_i = f_\lambda(x_i)$ the fitted value when several knots are introduced.
The weighted smoothing method is the minimizer of the Penalized Residual Sum of Square of the Generalized Cross-Validation (GCV) and Unbiased risk (UBR) as proposed by Wang (1996).

$$P_k = \frac{1}{n}\left(y - \hat{f}\right)^T W^k (y - f) + \lambda \int_0^1 \left(f''(x_i)\right)^2 dx \tag{10}$$

The first term in the equation is the residual sum of the square for the goodness of fit of the data while the second term is a roughness penalty, which is large when the integrated second derivative, λ is a smoothing parameter, of the regression function, $f''(x)$ is large when f(x) is rough (i.e. with rapidly changing slope). If λ approaches 0 then $f''(x)$ simply interpolates the data and when λ is very large, then f(x) will be selected so that $f''(x)$ is everywhere, which implies a globally linear least-squares fit all data.
Taking the Euclidean norm of (10) as established by Wang (1996) and setting $\lambda = 0$, we have;

$$P_k = \frac{1}{n}\left\| W^{\frac{k}{2}}\left(y - \hat{f}\right) \right\| \quad k = 0{,}1 \tag{11}$$

Note that; if $f = (f(x_1), \ldots, f(x_n))$ be the vector of values of function f at the knot points $x_1, \ldots, x_n$. The smoothing spline estimate $\hat{f}$ of this vector or the fitted values for data $y = (y_1, \ldots, y_n)^T$ are projected by;

$$\hat{f} = \begin{bmatrix} \hat{f}(x_1) \\ \hat{f}(x_2) \\ \cdot \\ \cdot \\ \cdot \\ \hat{f}(x_n) \end{bmatrix}_{(n\times 1)} = S_\lambda y \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}_{(n\times 1)} \quad \text{or} \quad \hat{f} = S_\lambda y \tag{12}$$

where $\hat{f}_\lambda$ is a natural cubic spline with knots at $x_1, \ldots, x_n$ for a fixed $\lambda > 0$, and S_λ is a well-known positive-definite (symmetrical) smoother matrix which depends on λ and the knot points $x_1, \ldots, x_n$, but not on y. Function f_λ, the estimation of function f, is obtained by cubic spline interpolation that rests on the condition $\hat{f}_\lambda(xi) = (\hat{f})i$, $i = 1, 2, \ldots, n$.

$$P_k = \frac{1}{n^2}\left\| W^{\frac{k}{2}}(y - S\lambda y) \right\| \quad k = 0{,}1 \tag{13}$$

The GCV method is well known for its optimal properties in smoothing estimation method (Wahba 1990) while the Mallow's CP criterion or UBR method has been successfully applied to estimate smoothing parameters for spline estimates with non-Gaussian data, fit data appropriately and do not overfit data, Gu (1992), Wahba, Wang, Gu, Klein, and Klein (1995), Morris and Lieberman (2008). The proposed Smoothing method combines the optimal properties of GCV and UBR.
The minimizer of GCV is given as;

$$GCV(\lambda) = \frac{1}{n} \frac{\sum_{k=1}^{n}\{y - f_k(x_1)\}^2}{\{1 - n^{-1} trace(S\lambda)\}^2}$$

$$GCV(\lambda) = \frac{n^{-1}\|(y - S\lambda)y\|^2}{[n^{-1} trace(I - S\lambda)]^2} = \frac{n^{-1}\|(I - S\lambda)y\|^2}{[n^{-1} trace(I - S\lambda)]^2} \tag{14}$$

While the GCV method for estimating spline smoothing function (f) in the presence of Autocorrelation structure as developed by Diggle and Hutchinson (1989) was given by,

$$GCV(\lambda) = \frac{(y - \hat{f})^T W(y - f)}{[trace\{1 - S\lambda\}]^2} \quad (15)$$

Therefore; equation (13) becomes;

$$P_k = \frac{1}{n}\left\|W^{\frac{k}{2}}(y - S\lambda y)\right\| = \frac{1}{n}\left\|W^{\frac{k}{2}}(I - S\lambda)y\right\| \quad (16)$$

The expected value (U_k) of (16) as established by Wang (1996) is given as;

$$U_k = \frac{1}{n}(I - S\lambda)yW^k(I - S\lambda)y - \frac{\sigma^2}{n}tr(S\lambda W^k S\lambda W^{-1})y \qquad k = 0,1,2 \quad (17)$$

While the unbiased estimate of (16) is given as;

$$U_k = \frac{1}{n}y(I - S\lambda)W^k(I - S\lambda)y - \frac{\sigma^2}{n}tr(W^{k-1})S\lambda y - 2\frac{\sigma^2}{n}trW^{k-1}S_\lambda \quad (18)$$

The minimization of $U_k$in (18) is referred to as the Unbiased Risk (UBR), by making σ^2 the subject of the formula, UBR estimator is given as;

$$\text{UBR}(\lambda) = \frac{\frac{1}{n}\left\|W^{\frac{k}{2}}(I - S\lambda)y\right\|^2}{\left[\frac{1}{n}trace\left(W^{k-1}(I - S\lambda)\right)\right]^2} \quad (19)$$

Therefore, a new smoothing method is proposed by introducing an additional weighted parameter ***g*** and combining properties of the generalized cross-validation (GCV) and Mallow CP criterion (UBR). The combination, measurement, and expression of the quantities of the two methods will yield an optimal performance and smoothing model that does not overfit data. The minimizer of the combination of (15) and (19) is the weighted smoothing method given as;

$$WSM\ (\lambda) = gf_1(x) + (1 - g)f_2(x), \qquad where;\ 0 < g < 1 \quad (20)$$

$$WSM\ (\lambda) = g\frac{(y - \hat{f})^T W(y - f)}{[trace(I - S_\lambda)]^2} + (1 - g)\frac{\frac{1}{n}\left\|W^{\frac{K}{2}}(I - S_\lambda)\right\|^2}{\left[\frac{1}{n}trace\{W^{k-1}(I - S_\lambda)\}\right]^2} \quad (21)$$

Now the behavior of the minimize λ in GCV and UBR methods under the substituted value of k = 1 as the optimum value of WSM yields;

$$WSM\ (\lambda) = g\frac{(y - \hat{f})^T W(y - f)}{[trace(I - S_\lambda)]^2} + (1 - g)\frac{\frac{1}{n}\left\|W^{\frac{1}{2}}(I - S_\lambda)\right\|^2}{\left[\frac{1}{n}trace\{W(I - S_\lambda)\}\right]^2} \quad (22)$$

Where;
n = number of observations, $0 < g < 1$, g = weighted values, W = V^{-1} = Correlation Matrix for the error term
$y = (y_1, \ldots, y_n)^T$ = Smoothing function, $\hat{f} = (f(t_1) \ldots f(t_n))y_n)^T$ = $S_\lambda y$, Sλ= is the diagonal element of the smoother matrix, $\left\|W^{\frac{1}{2}}(I - S\lambda)y\right\|$= norm of the Euclidean vector $W^{\frac{1}{2}}(y - \hat{f})$

3.0 Material and Method

3.1 Equation used for generating values in the simulation

A simulation study is conducted to evaluate and compare the performance of the four estimation methods presented in previous sections. The model considered is

$$y_t = \frac{Sin\pi_i}{t} + \varepsilon_t\ i = 1, 2, \ldots n, t = \varepsilon[1, 100] \quad (23)$$

Where; ε's are generated by a first-order autoregressive process AR (1) with mean 0, standard deviations 0.8 and 1.0 and first-order correlations (i.e. ρ = 0.0. 0.2, 0.5 and 0.8) and its 95% Bayesian confidence interval, Wahba, (1983) and Diggle, (1989).

3.2 Experimental design and data generation

The experimental plan applied in this research work was designed to have three (3) time series sample Sizes (T) of 20, 60 and 100, four (4) Autocorrelation levels, (ρ = 0.0, 0.2, 0.5 and 0.8), and two (2) standard deviations were considered (σ = 0.8 and 1.0). The data were generated for 1000 replications for each of the $3 \times 4 \times 2 = 24$ combinations of cases T, ρ, and σ. The criterion used is the PMSE values to evaluate $\hat{f}_{\lambda}$ computed according to each of the estimation given as;

$$PMSE(\lambda) = E\left[\sum_{i=1}^{n}\left(f(x_i) - \hat{f}(x_i)\right)^2\right] \tag{24}$$

The Predictive Mean Square Error can be divided into two terms, the first term is the sum of square biases of the fitted values while the second is the sum of variances of the fitted values.

Where; $f(x_i)$ is the observed value and $\hat{f}(x_i)$ = fitted/predicted/estimated value. Aydin, Memmedli, and Omay (2013). The simulation study was performed by using a program written in R, it was used to estimate all the model parameters, the criterion, the effect of autocorrelation on the estimated parameters and the performances of the four estimation methods i.e. Generalized Maximum Likelihood (GML), Generalized Crossed Validation (GCV), Unbiased Risk (UBR) and the Weighted Smoothing Method (WSM).

4 Results

In this study, the results of the proposed Spline smoothing estimation method were compared with three existing estimation methods namely; the Generalized Cross-Validation, Generalized Maximum Likelihood, and Unbiased Risks, the Predictive mean square errors criterion was used to measure their efficiency.

4.1 Performance of the four smoothing methods based on predictive mean square error criterion when σ = 0.8.

Table 1 presents the predictive mean square error for the four smoothing methods, three (3) time series periods, and four (4) correlation error levels at 0.8 sigma level. It was discovered that for GCV; the predictive mean square error increases as the level of Autocorrelation increases from 4.938284 when ρ = 0.2 to 5.70041 when ρ = 0.5 and to 5.735483 when ρ = 0.8 and time series size = 20. It was also discovered that the predictive mean square error decreases as and time series size increases; when T = 20 the PMSE decreased from 4.938284 to 1.353605 when T = 60 and further decreases from 1.353605 to 1.334855 when T = 100.
The predictive mean square error (PMSE) of GML increases from 3.788134 to 3.902353 as the Autocorrelation error level increases of 0.2 to 0.5 and also increased 4.557857 as the Autocorrelation level increases from 0.5 to 0.8. It was also discovered that the predictive mean square error decreases as the time series size increases; when T = 20 the PMSE decreased from 3.788134 to 2.328352 when T = 60 and further decreases series from 2.328352 to 2.314015 when T = 100.
For the Weighted Smoothing Method (WSM), it was discovered that the predictive mean square error decreases as the time series size and Autocorrelation level increases. WSM decreased from 0.046551 at time series size = 20 to 0.023102 at a time series size of 60 and further decreased to 0.022368 at a time series size of 100. The predictive mean square error of WSM increased from 0.046551 to 3.49336 as the Autocorrelation error level increases of 0.0 to 0.2.
The predictive mean square error for UBR decreases as the time series size increase at ρ =0.5; UBR decreases from 3.777261 to 2.101405 when the time series size is 60 and further decreased to 1.913073 when time series sizes are 100. Furthermore, the predictive mean square error of UBR increased from 2.1014051 to 2.3170046 as the Autocorrelation error level increases of 0.2 to 0.5 for time series sizes of 60.

Table 2 presents the predictive mean square error for the four smoothing methods, three (3) time series sizes, and three Autocorrelation error levels at 1.0 sigma level. It was discovered that for GCV; the predictive mean square error increases as the level of Autocorrelation increases from 2.217985 when ρ = 0.2 to 4.652218 when ρ = 0.5 and to 5.219991 when ρ = 0.8 and time series sizes = 20. It was also discovered that the predictive mean square error decreases as the time series sizes increase; at T = 20 the PMSE decreased from 2.217985 to 1.5079261 at T = 60 and further decreases from 1.5079261 to 0.109678 at T = 100.
The predictive mean square error (PMSE) of GML increases from 1.402249 to 2.213838 as the Autocorrelation error level increases from 0.2 to 0.5 and also increased to 2.854191 as the Autocorrelation level increases from 0.5 to 0.8. The predictive mean square error decreases as the time series sizes increase; at T = 20 the PMSE decreased from 1.402249 to 1.285324 at T = 60 and further decreases from 1.285324 to 0.917754 at T = 100.
For the Weighted Smoothing Method (WSM), it was discovered that the predictive mean square error decreases as the time series sizes increase. WSM decreased from 0.046236 at time period = 20 to 0.044161 at time period = 60 and further decreased to 0.035276 at time period=100. The predictive mean square error of PSM also decreased from 1.181876 to 1.86648 as the Autocorrelation error level increases from 0.2 to 0.8 at time series size of 60.

The predictive mean square error for UBR decreases as the time series sizes increase, UBR decreases from 3.946115 at time series sizes = 20 to 3.477279 at time series sizes = 60 and further decreased to 0.715411 at time series sizes =100. Furthermore, the predictive mean square error of UBR increased from 2.170123 to 2.854018 as the Autocorrelation error level increases from 0.5 to 0.8 for time series sizes of 20.

Figures 1 to 8 presents the observed and estimated values of GCV, GML, WSM, and 1000 replications. From these plots, it is observed that WSM and GCV estimates have small PSMEs compare with UBR and GML. It is concluded that all four methods estimate the smoothing parameters and the functions well but the WSM and GCV provide better estimates than UBR and GML in terms of predictive mean-square error. The WSM method is more stable when the sample size is small (T = 20) while the GCV method performs slightly better when T = 60. This behavior of the GCV method was investigated in Wahba and Wang (1993) and Wang (1998).

Table 1: The PMSE result of the simulated study for GML, GCV, WSM and UBR in the presence of autocorrelation (ρ) = 0.0, 0.2, 0.5 and 0.8 for T = 20, 60 and 100 when standard deviation (σ) = 0.8

Smoothing parameters	***Smoothing Methods***	***Autocorrelation levels*** $\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
T = 20, σ = 0.8	GCV	0.053273	4.938284	5.700411	5.735483
	GML	0.081617	3.788134	3.902353	4.557857
	WSM(k=0.1)	0.046551	3.49336	2.09983	2.272232
	UBR	0.04802	3.777261	2.810875	2.449087
T = 60, σ = 0.8	GCV	0.027264	1.353605	3.179886	5.817303
	GML	0.074695	2.328352	2.429546	2.625861
	WSM(k=0.2)	0.023102	1.25185	1.28961	1.05828
	UBR	0.034561	2.101405	2.317046	1.118518
T = 100, σ = 0.8	GCV	0.025094	1.334855	4.190077	4.753061
	GML	0.068976	2.314015	2.836043	2.438085
	WSM(k=0.3)	0.022368	1.05761	1.01288	1.01515
	UBR	0.025485	1.913073	2.079789	2.841755

Table 2: PMSE result of the simulation study for GML, GCV, WSM and UBR in the presence of autocorrelation levels ρ = 0.0, 0.2, 0.5 and 0.8, λ = 1 and σ = 1.0

Smoothing parameter	***Smoothing Method***	***Autocorrelation levels*** $\rho = 0.0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
T = 20, σ = 1.0	GCV	0.1268264	2.217985	4.652218	5.219991
	GML	0.182682	1.402249	2.213838	2.854191
	WSM(k=0.4)	0.046236	1.704177	2.01762	2.18525
	UBR	0.056372	3.946115	2.170123	2.854018
T = 60, σ = 1.0	GCV	0.0424942	1.5079261	3.032906	3.355379
	GML	0.0902741	1.285324	2.424851	2.860878
	WSM(k=0.5)	0.040161	1.181876	1.86648	1.12187
	UBR	0.0444551	3.477279	1.895938	1.904192
T = 100, σ = 1.0	GCV	0.0242396	0.109678	0.205153	2.068174
	GML	0.0598298	0.917754	1.498209	1.460676
	WSM(k=0.6)	0.032276	0.618494	1.21775	1.81974
	UBR	0.0373787	0.715411	1.410622	1.391461

4.2 Smoothing Curves of Time Series Observations with Autocorrelation in the Error Term

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 4, Issue 7, July – 2020, Pages: 70-81

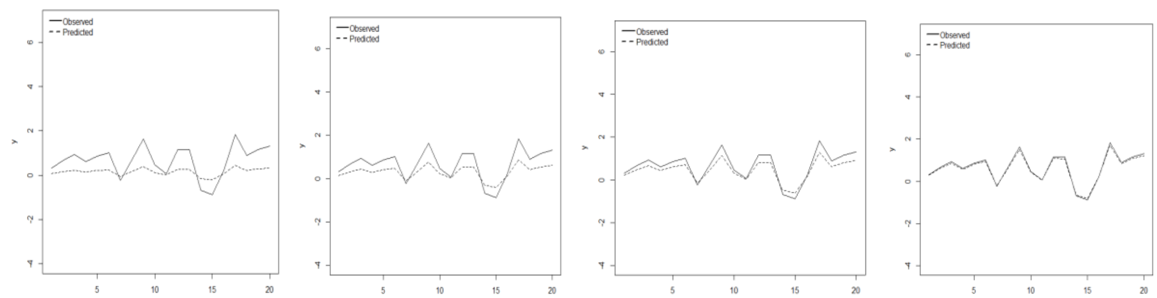

Figure 1: Plots of the observations (. . .) *and Estimates* (---) *With Smoothing Parameters Chosen by GCV (a), GML (b), WSM (c), and UBR (d) σ = 0.8, ρ = 0.0 and T = 20*

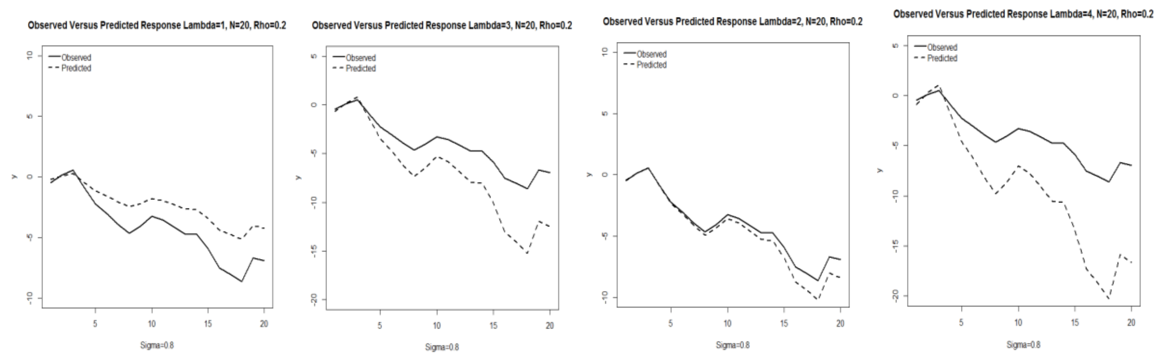

Figure 2: Plots of the observations (. . .) *and Estimates* (---) *With Smoothing Parameters Chosen by GCV (a), GML (b), WSM (c), and UBR (d) for, σ = 0.8, ρ = 0.2 and T = 20*

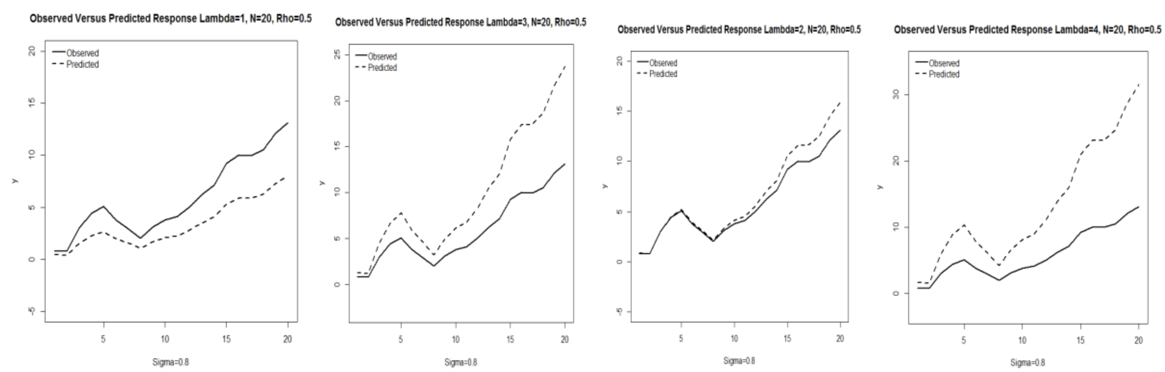

Figure 3: Plots of the observations (. . .) *and Estimates* (---) *With Smoothing Parameters Chosen by GCV (a), GML (b), WSM (c), and UBR (d) for, σ = 0.8, ρ = 0.5 and T = 20*

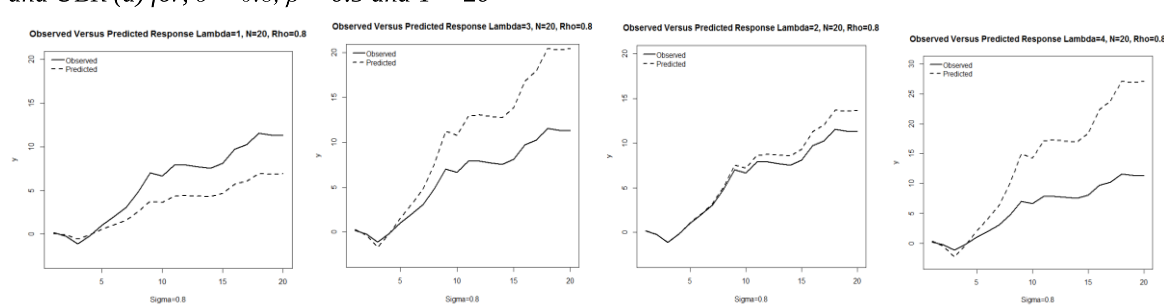

Figure 4: Plots of the observations (. . .) *and Estimates* (---) *With Smoothing Parameters Chosen by GCV (a), GML (b), WSM (c), and UBR (d) for σ = 0.8, ρ = 0.8 and T = 20*

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 4, Issue 7, July – 2020, Pages: 70-81

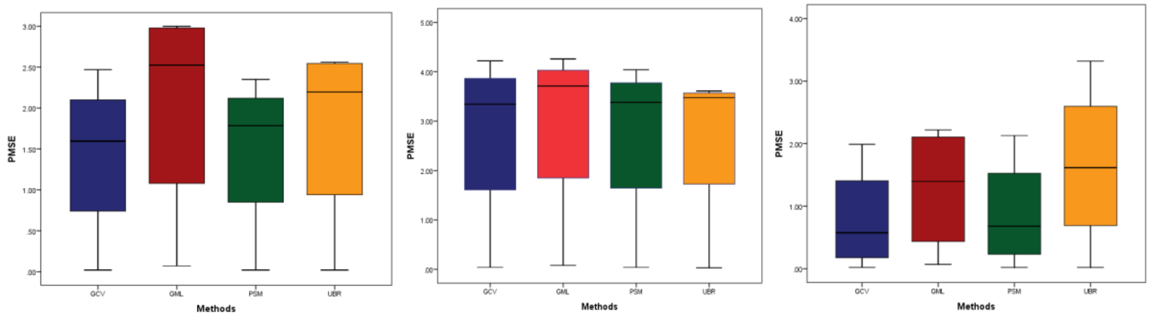

Figure 5: Box plot of GML, GCV, WSM, and UBR of the PMSE of the simulated study in the absence of autocorrelation when σ = 0.8, ρ = 0.0 and T = 20, 60 and 100

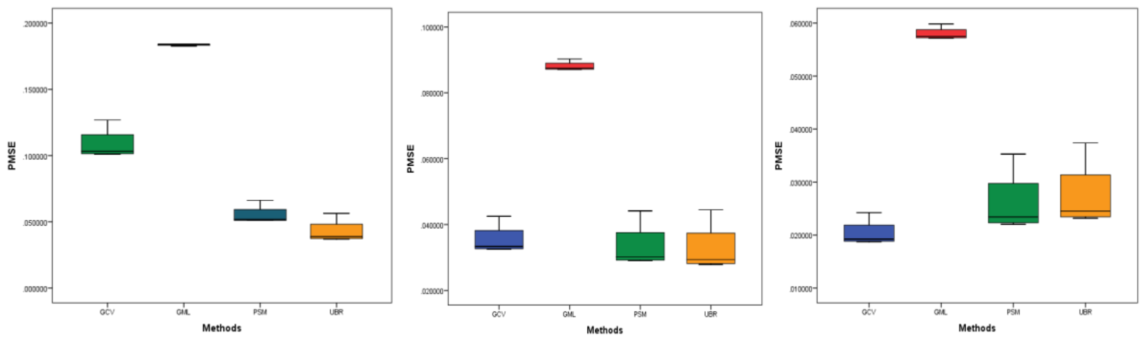

Figure 6: Box plot of GML, GCV, WSM, and UBR of the PMSE of the simulated study in the presence of autocorrelation when σ = 0.8, ρ = 0.2 and T = 20, 60 and 100

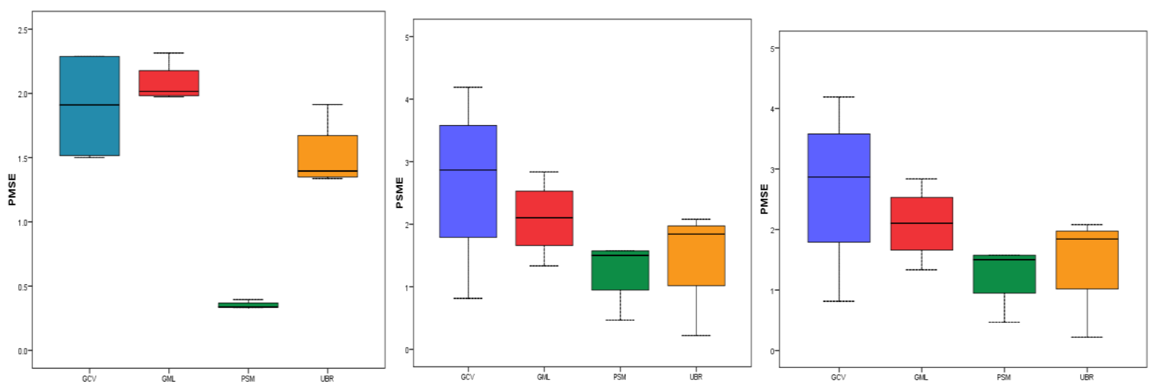

Figure 7: Box plot of GML, GCV, WSM, and UBR of the PMSE of the simulated study in the presence of autocorrelation when σ = 0.8, ρ = 0.5 and T = 20, 60 and 100

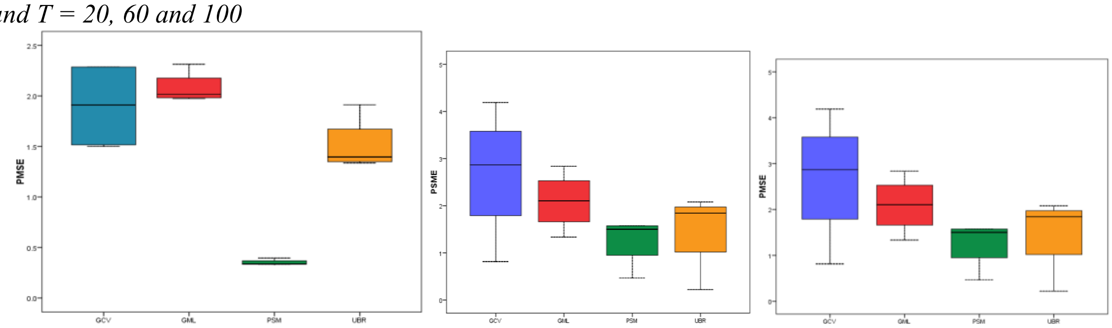

Figure 8: Box plot of GML, GCV, WSM, and UBR of the PMSE of the simulated study in the presence of autocorrelation when σ = 0.8, ρ = 0.8 and T = 20, 60 and 100

4.3 Discussion of findings

Table 3, presented the preferred smoothing methods in the presence of four Autocorrelation levels (ρ = 0.0, 0.2, 05 and 0.8) at three time series sizes (T = 20, 60 and 100) for two standard deviations (σ = 0.8 and 1.0). It was discovered that WSM was the most preferred smoothing method at all-time series size small (T = 20, 60, and 100). Similarly, Table 4, showed the preferred Spline Smoothing Method based on four Autocorrelation levels (ρ = 0.0, 0.2, 05 and 0.8) for two standard deviations (σ = 0.8 and 1.0). It was discovered that WSM was the most preferred spline smoothing method at four Autocorrelation levels (ρ =0.0, 0.2, 0.5, and 0.8)

Table 3: Preferred Smoothing method in the presence of Autocorrelated error based on time series sizes

T	σ = 0.8				σ = 1.0				MOST PREFERRED
	ρ = 0.0	ρ = 0.2	ρ = 0.5	ρ = 0.8	ρ = 0.0	ρ = 0.2	ρ = 0.5	ρ = 0.8	
20	WSM	WSM	WSM	WSM	UBR	GML	WSM	WSM	**WSM**
60	WSM	WSM	WSM	WSM	GCV	WSM	WSM	WSM	**WSM**
100	UBR	WSM	WSM	WSM	GCV	GCV	GCV	GCV	**WSM**

Table 4: Preferred smoothing method based on the presence of Autocorrelation

T	σ = 0.8				σ = 1.0			
	ρ = 0.0	ρ = 0.2	ρ = 0.5	ρ = 0.8	ρ = 0.0	ρ = 0.2	ρ = 0.5	ρ = 0.8
20	WSM	WSM	WSM	WSM	UBR	GML	WSM	WSM
60	WSM	WSM	WSM	WSM	GCV	GCV	WSM	WSM
100	UBR	WSM	WSM	WSM	GCV	GCV	GCV	GCV
MP	**WSM**	**WSM**	**WSM**	**WSM**	**GCV**	**GCV**	**WSM**	**WSM**

***MP = Most preferred**

5 Discussion and Conclusion

In this study, the spline smoothing estimation method for time series observations in the presence of Autocorrelated errors was compared based on three sample sizes. The simulation result under the finite sampling properties of the PMSE criterion shows that all smoothing methods were consistent but adversely affected by the presence of Autocorrelation in the error term, the smoothing methods rank as follows, WSM, GCV, GML, and UBR. The result suggested that WSM should be preferred at all levels of the Autocorrelation level. It was also observed that UBR and GML were mostly affected by the presence of Autocorrelation and therefore had an asymptotically similar behavioral pattern. It was also discovered that WSM does not overfit when the time series size is medium and large (T = 60 and 100) **(see Appendix A).** This result is slightly similar for GML with; Yuedong (1998) and Yuedong et.al (2000) but differs from; Hart and Wehrly (1986), Diggle and Hutchinson (1989), Altman (1990), Herrman, Gasser, and Kniep (1992) and Krivobokova and Kauermann (2007). The study also discovered that the Weighted Smoothing method is preferred at all the time series sizes studied in this research.

.

References

Adams, S.O., Ipinyomi R.A., and Yahaya H.U. (2017). Smoothing Spline of ARMA Observations in the Presence Of Autocorrelation Error. *European Journal of Statistics and Probability*, 5(1): 1 – 8.

Aydin, D., M. Memmedli, and R. E. Omay. (2013). Smoothing parameter selection for nonparametric regression using a smoothing spline. *European Journal of Pure and Applied Mathematics* 6:222–38

Barry, D. (1983). Nonparametric Bayesian regression, Ph.D. thesis, Yale University, New Haven, Connecticut.

Carew, J. D., Wahba, G., Xie X, Nordheim, E.V. and Meyerand M. E. (2003), Optimal Spline Smoothing of FMRI Time Series by Generalized Cross-Validation, *NeuroImage*, 18(4): 950 – 961.

Chen, C.S., and Huang H.C. (2011). An improved Cp criterion for spline smoothing. *Journal of Statistical Planning and Inference, 144(1): 445 – 471.*

Diggle, P.J., and Hutchinson, M.F. (1989). On spline smoothing with Autocorrelated errors. *Australian Journal of Statistics*, 31: 166 –182.

Francisco-Fernandez, M., and Opsomer, J.D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics, 33(2):* 279–295.

Hall, P., and Keilegom, I. (2003). Using Difference-Based Methods for Inference in Nonparametric Regression with Time Series Errors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65 (2): 443 – 456.

Hart, J. D., and Lee, C. (2005). Robustness of one-sided cross-validation to autocorrelation. *Journal of Multivariate Analysis*, 92:77 – 96.

Kim, T., Park, B., Moon, M., and Kim C. (2009). Using bimodal kernel for inference in nonparametric regression with correlated errors. *Journal of Multivariate Analysis.* 100 (7), 1487 – 1497.

Krivobokova T. and Kauermann G. (2007). A note on Penalized Spline Smoothing with Correlated Errors. *Journal of the American Statistical Association*, 102: 1328 – 1337.

Morton, R., Kang, E.L. and Henderson, B.L. (2009), Smoothing Spline for trend estimation and prediction in time series. *Environmetrics*, 20(3): 249 – 259.

Opsomer J., Yuedong W., and Yang Y. (2001). Nonparametric Regression with correlated Error *Statistical Sciences,* 6: (2) 134 – 153.

Shen, H. (2008). On Modeling and Forecasting Time Series on Smooth Curve, *Technometrics,* 51(3): 227 – 238.

Wahba, G. (1975). Optimal convergence properties of variable knot kernel and orthogonal series methods for density estimation. *Annals of Statistics*, 3: 15 – 29.

Wahba, G. (1979). Convergence rates of thin plate smoothing splines when the data are Noisy in T. Gasser and M. Rosenblatt (eds), Smoothing techniques for curve estimation, *Springer-Verlag*.

Wahba, G. (1983), Bayesian Confidence intervals for the cross-validated smoothing Spline, *Journal of Royal Statistical Society Service*. B. 45:133 – 150.

Wahba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameters in the Generalized Spline Smoothing Problem. *The Annals of Statistics*, 4:1378 – 1402.

Wahba, G., Wang, Y., Gu, C., Klein, R., and KIein, B. (1995). Smoothing Spline ANOVA for Exponential Families, With Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *The Annals of Statistics*, 23:1865 – 1895.

Wang, H., Meyer M. C., and Opsomer J.D. (2013). Constrained spline regression in the presence of AR(p) errors. *Journal of Nonparametric Statistics*, 25:809 – 827.

Yuedong, W., and Brown M.B. (1996). A Flexible Model for Human Circadian Rhythms, *Biometric*, 52: 588 – 596

Yuedong, W. (1998), Smoothing Spline Models with Correlated Random Errors. *Journal of American Statistical Association*, (93) 441: 341 – 348.

Yuedong, W., Wensheng G., and Brown M.B. (2000). Spline Smoothing for Bivariate data with application to an association between hormones, *Statistica Sinica,* 10: 377 – 397.

APPENDIX A: Test for overfitting

Table 1: Test of Over-fitting (GCV)
GCV: n = 60, rho = 0.8, Sigma = 0.8
The regression equation is
y = 0.855 + 0.122 x

Predictor	Coef	SE Coef	T	P
Constant	0.8552	0.4509	1.90	0.074
x	0.1216	0.4741	0.26	0.800

S = 1.45436   R-Sq = 14.4%   R-Sq(adj) = 0.0%
PRESS = 43.6008   R-Sq(pred) = 0.00%

Table 2: Test of Over-fitting (GML)
GML: n = 60, rho = 0.8, Sigma = 0.8
The regression equation is
Y = 1.15 + 0.0239 X

Predictor	Coef	SE Coef	T	P
Constant	1.1507	0.1511	7.62	0.000
X	0.02387	0.02778	0.86	0.401

S = 0.454909   R-Sq = 13.9%   R-Sq(adj) = 0.0%
PRESS = 4.49669   R-Sq(pred) = 0.00%

Table 3: Test of Over-fitting (WSM)
WSM: n = 60, rho = 0.8, Sigma = 0.8
The regression equation is
Y = 0.300 + 0.134 X

Predictor	Coef	SE Coef	T	P
Constant	0.30041	0.07217	4.16	0.001
X	0.13436	0.03793	3.54	0.002

S = 0.207129   R-Sq = 61.1%   R-Sq(adj) = 57.8%
PRESS = 0.905132   R-Sq(pred) = 50.94%

Table 4: Test of Over-fitting (UBR)
UBR: n = 60, rho = 0.8, Sigma = 0.8
Y = 0.377 - 0.0222 X

Predictor	Coef	SE Coef	T	P
Constant	0.37664	0.06330	5.95	0.000
X	-0.02217	0.01918	-1.16	0.263

S = 0.172663   R-Sq = 16.9%   R-Sq(adj) = 1.7%
PRESS = 0.649678   R-Sq(pred) = 0.00%

Table 5: Test of Over-fitting (GCV2)
GCV: n = 100, rho = 0.8, Sigma = 0.8
The regression equation is
Y = 0.795 + 0.0077 X

Predictor	Coef	SE Coef	T	P
Constant	0.79504	0.02881	27.60	0.000
X	0.00772	0.01237	0.62	0.540

S = 0.108278   R-Sq = 20.1%   R-Sq(adj) = 0.0%
PRESS = 0.271302   R-Sq(pred) = 0.00%

Table 6: Test of Over-fitting (GML2)
GML: n = 100, rho = 0.8, Sigma = 0.8
The regression equation is
Y = 2.25 + 0.235 X

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 4, Issue 7, July – 2020, Pages: 70-81

Predictor	Coef	SE Coef	T	P
Constant	2.2533	0.3798	5.93	0.000
X	0.23546	0.08131	2.90	0.010

S = 1.03928   R-Sq = 31.8%   R-Sq(adj) = 28.0%
PRESS = 28.4788   R-Sq(pred) = 0.07%

Table 7: Test of Over-fitting (WSM2)
WSM: n = 100, rho = 0.8, Sigma = 0.8
The regression equation is
Y = 0.494 - 0.0965 X

Predictor	Coef	SE Coef	T	P
Constant	0.49423	0.02105	23.48	0.000
X	-0.096522	0.008244	-11.71	0.000

S = 0.0929687   R-Sq = 88.4%   R-Sq(adj) = 87.7%
PRESS = 0.184664   R-Sq(pred) = 86.22%

Table 8: Test of Over-fitting (UBR2)
UBR: n = 100, rho = 0.8, Sigma = 0.8
The regression equation is
Y = 0.816 + 0.0204 X

Predictor	Coef	SE Coef	T	P
Constant	0.8161	0.1135	7.19	0.000
X	0.02039	0.07643	0.27	0.793

S = 0.113964   R-Sq = 4.4%   R-Sq(adj) = 0.0%
PRESS = 0.289895   R-Sq(pred) = 0.00%