MODELING DOMESTIC OUTPUT FROM SELECTED MACROECONOMIC VARIABLES IN NIGERIA USING CROSS-VALIDATED SCATTERPLOT SMOOTHERS

Are S.O[1], Alabi N.O[1]
[1]Department of Mathematics and Statistics, Federal Polytechnic Ilaro
e-mail: nurudene.alabi@gmail.com

Abstract

We employed generalized additive models (GAMs) with scatterplot smoothers such as cubic smoothing splines and local regression to establish non-linear relationships that exist between Nigeria's domestic output proxy by gross domestic products and four macroeconomic variables such as recurrent and capital expenditures, crude oil prices and federal government retained earnings. Four flexible generalized additive models were postulated. These are smoothing splines, local regression, local quadratic regression and composite GAM. Generalized Cross-validation and leave-one-out re-sampling techniques were employed in determining the effective degrees of freedom and the window size or span for smooth splines and local regression additive models respectively. We compared models accuracy using the residual deviance for the GAMs. Smoothing Splines was selected through a nested analysis of deviance on the four postulated models. The minimized penalized residual sum of squares optimal tuning parameters $\Omega = 1.2613x10^{-6}, 2.0153x10^{-6}, 7.9291x10^{-7}$ *and* $1.5286x10^{-7}$ *for each smoothing functions on federal government retained earnings, crude oil prices, recurrent expenditure and capital expenditure respectively which control the tradeoff between goodness-of-fit and smoothness of the smoothing spline model. Analysis shows the federal government retained earnings and volatile crude oil prices are the major macroeconomic variables that cause instability in the growth of gross domestic products in Nigeria.* ***R*** *programming language packages such as "splines","gam","ggplot2","boot"and"lattice" amongst others were employed throughout the analysis.*

Keywords: gross domestic products, generalized additive models, generalized cross-validation, smoothing splines and local regression.

Introduction

Lawal G.O., Alabi N.O., Ige S.A., Ibraheem R.A. (2016) employed *autoregressive distributed lag model (ARDL)* to establish non-linear relationships existing between domestic output and government spending with federal government retained earnings and crude oil price as static regressors. Our analysis produced a model with minimum bias in domestic output, which produced estimates very close to the true values. One advantage of this method was that in addition to reduced bias, we were able to minimize the variance of the residual error term. Furthermore, since the variables are all of the same order one (i.e. I(1)), fitting an ARDL model allowed us to establish both the long-run cointegrating relations and short-run dynamic effects amongst the endogenous and exogenous variables using the traditional ordinary least squares method of estimation. However, the estimates of the model parameters are average effects that can provide limited information about the true pattern inherent in the data. Here, we introduce an alternative method of studying the non-linear relationships existing between the same set of macroeconomic variables in a non-parametric manner. This method unlike the ARDL is a semi-parametric and non-parametric regression learning method first introduced by Hastie and Tibshirani in 1986. It is known as the *Variable Coefficient Model* or *Generalized Additive Model* (GAM, hereon). GAM belongs to the family of *Generalized Linear Models* (GLM) or the *Exponential family of distribution.* Unlike most members of GLM, GAM replaces the linear predictor of a set of predictors with a sum of unspecified smooth functions whose parameters are estimated using a scatterplot smoother such as kernel estimate or the cubic spline amongst several other methods such as running mean and running median (Hastie and Tibshirani, 1986, 1990). Although this approach is not popular amongst econometricians, its relevance in terms of bias-variance tradeoff optimization and model interpretability cannot be overlooked which has made it an important econometric analysis tool in the last decade. Our main objective in this current paper is not to duel extensively on the literature regarding domestic output and selected macroeconomic predictors (refer to Lawal *et al* (2016) for detailed literature review), instead we look at the theoretical framework of GAM vis-à-vis its relevance to the available data.

Generalized Additive Model (GAM) on Nigeria's GDP and its predictors

We begin by postulating a model in equation 1.1 on domestic output proxy by gross domestic products (***gdp***) and crude oil price (***cp***), federal government retained earnings (***fgre***), government expenditures proxy by recurrent (***re***) and capital expenditures (***ce***) using quarterly data collected from the database of Central Bank of Nigeria and Nigeria Statistical Bulletin between 2004 and 2014. The GAM is of the form;

$$gdp_i = \beta_0 + f_1(fgre_i) + f_2(cp_i) + f_3(re_i) + f_4(ce_i) + \varepsilon_i \tag{1.1}$$

Each unspecified univariate smooth function f_j in equation 1.1 is estimated using very fast non-parametric Backfitting algorithm such as additive backfitting algorithm with weights (ARBAW) and *k*-nearest neighbours[1]. The error term is assumed to be distributed normally i.e. $\varepsilon_i \sim N(0, \sigma^2)$. We assumed for simplicity that each covariate is standardized within the [0,1] interval. These two approaches use a scatterplot smoother[2] to generalize the Fisher scoring procedure for calculating MLE (Hastie and Tibshirani, 1986,1990). The non-parametric form of the functions f_j in equation 1.1 allows for relatively flexible pattern of the dependence of ***gdp*** on the covariates, but by specifying the model only in terms of smooth

---

[1] *The f_j's are standardized smooth function such that* $\mathbf{E}[f_j(X_j)]=0$*. Each smooth function is only estimable to within an additive constant.*

[2] *Scatterplot smoother is a method of estimating the smooth non-parametric functions $f_j(.)$ i.e. splines. According to Wood (2013), y_i is usually measured with noise and it is generally more useful to smooth x_i, y_i data rather than interpolating them by setting $g(x_i)$ as n free parameters of the cubic spline.*

functions. This flexibility and convenience comes at the cost of two conjectural problems. It is essential to define the smooth functions involving basis dimension and location of region boundary and to determine the degree of smoothness controlled by a tuning parameter. According to Cho H., Goude Y., Brossat X., and Yao Q. (2013), GAMs allow for implicit non-linear relationships between the response and the covariates without suffering from "*curse of dimensionality*". We estimated each of the functions in equation 1.1 separately in a forward stepwise manner using the scatterplot smoothers. We are able to fit non-linear f_j to each predictor allowing non-linear relationships missed by a linear model.

Scatterplots of likely association between GDP and its predictors in Nigeria from 2004 to 2014

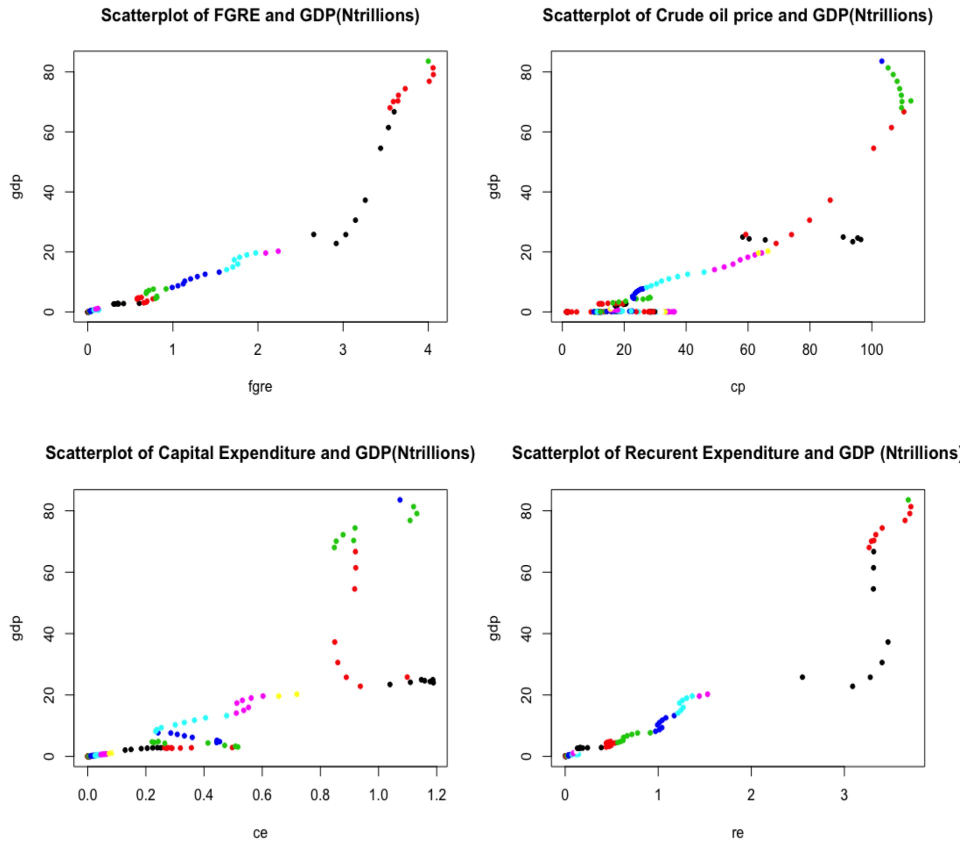


Figure 1: *Scatterplots on gross domestic products (**gdp**) and its covariates. These plots exhibit some patterns of non-linearity.*
Source*: Personal computation using **R** Studio ggplot2 package (H. Wickham, 2009).*

The Smoothing Splines Model (SSM) on Domestic Output in Nigeria

Firstly, we introduce the cubic spline scatterplot smoother. Later, we introduce the localized regression involving window size or span, which is a memory-based procedure. The former involves fitting the model in equation 1.1 by using additive regression backfitting algorithm with weights (ARBAW) in which each f_1, f_2, f_3, and f_4 are cubic smoothing splines[3]. The backfitting

[3] *Smoothing splines result from minimizing a penalized residual sum of squares criterion subject to a smoothness penalty. They are natural cubic splines with knots at every unique observation of* x_i. *The details are not covered in this research work.*

algorithm (Friedman and Stuetzle, 1981; 1982) generates a smoothing spline that minimizes the penalized residual sum of squares in equation 1.2

$$PRESS = \sum_{i=1}^{176} (y_i - \beta_0 - \sum_{j=1}^{4} f_j(x_{ij}))^2 + \sum_{j=1}^{4} \Omega_j \int g_j^{||}(t_j)^2 \, dt_j \qquad 1.2$$

$$g(x_{ij}) = \beta_0 + \sum_{j=1}^{4} f_j(x_{ij}) \qquad 1.3$$

We find the function g or smoothing spline, which minimizes the equation 1.2. This function is the shrunken natural cubic spline with region boundaries at unique values of x_i*'s* with continuous first and second derivatives at each region boundary or *knot*[4]. The level of shrinkage is controlled by the tuning parameter Ω determined by solving equation 1.6. Equation 1.2 is a function of residual sum of squares plus a smoothness penalty associated with each f_j. In order to avoid the excessive flexibility associated with smoothing splines when Ω is too low and the over-smoothness when Ω is too high, we employed the cross-validation re-sampling method to determine the tuning parameter and the corresponding effective degrees of freedom so that the estimated model is as close as possible to the true model. Specifically, the generalized cross-validation (GCV) approach, which is based on orthogonal rotation of the residual matrix such that the diagonal elements of the influence or hat matrix is as even as possible[5](Wood, 2006). It has been proven that ordinary cross-validation approach suffers from two major problems. Firstly, there is a problem of cost associated with the number of smoothing parameters involved. Secondly, there exist a problem of lack of invariance from comparing the ordinary cross-validation scores derived from equation 1.4

$$\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2 + \sum_{i=1}^{p} \boldsymbol{\Omega}_i \boldsymbol{\beta}^T \boldsymbol{S}_i \boldsymbol{\beta} \qquad 1.4$$

and its orthogonal transform

$$\|\boldsymbol{QY} - \boldsymbol{QX\beta}\|^2 + \sum_{i=1}^{p} \boldsymbol{\Omega}_i \boldsymbol{\beta}^T \boldsymbol{S}_i \boldsymbol{\beta} \qquad 1.5$$

where **Q** is an orthogonal matrix obtained from the **QR** decomposition[6]. Generalized cross-validation (GCV) addressed these problems by rotating an influence matrix **M** such that the effective degrees of freedom are the sum of the diagonal elements in the matrix M_Ω. That is

$$\text{EDF}_\Omega = \text{tr}(\mathbf{M}) \qquad 1.6a$$

Additive regression backfitting algorithm with weights (ARBAW) fits a multiple predictors model by repeatedly updating the fit for each predictor in turn holding the others fixed. This approach uses the scatterplot smoothers to generalize the Fisher scoring procedure for calculating MLE (Hastie and Tibshirani, 1986).

[4] *It is a natural cubic spline with knots at* $x_1,\ldots\ldots x_n$.

[5] *A hat (influence) matrix provides the estimates of vector of* ***E****(Y) when post-multiplied by the data vector y.*

[6] *We decompose X into QR where Q is the first p columns of the orthogonal reflector matrix.*

The equation 1.6a is derived from the decomposition of

$$M_\Omega = QMQ^T \qquad 1.6b$$

such that if **M**=**VV**^{T}, where V is any matrix square root, then

$$M_\Omega = QVV^TQ^T \qquad 1.6c$$

since the orthogonal matrix **Q** is such that each row of **QV** has the same mathematical length, the principal elements in the leading diagonal of the influence matrix M_Ω have the same value (see Wahba, 1990) i.e.

$$tr(M_\Omega) = tr(QMQ^T) = tr(MQ^TQ) = tr(M) \qquad 1.6d$$

Craven and Wahba (1979), Golub G.H., Heath M., and Wahba G., (1979) have shown that the generalized cross-validation score expressed in equation 1.6e derived from the rotation of the ordinary cross-validation score has predicted error unaffected by the rotation.

$$\upsilon_{gcv} = \frac{n \| y - \hat{\mu} \|^2}{[n - tr(M)]^2} \qquad 1.6e$$

The integral part of equation 1.2 measures the overall change in the slope $g^I(t)$ over its entire range measuring the roughness of the slope. We attempt to find the smoothing spline, *g* that minimizes the function in equation 1.2. As the smoothing parameter Ω approaches ∞, the EDF decreases[7] from *n* to 2, where *n* is the sample size (here, *n*=176). This controls the tradeoff between the goodness-of-fit of the model and model smoothness. ARBAW allows us to update a function using the partial residual. We exploit **R** programming language packages "*splines*" and "*gam*" to compute the smoothing parameter, the EDF and diagnostic checks for the GAM. The model is estimated as

$$E(gdp_i | fgre, cp, re, ce) = 1.0922 - 0.0265(fgre)_i + 0.4604(cp)_i + 0.4320(re)_i + 0.5610(ce)_i \qquad 1.7$$

Table 1 and **Table** 2 show the analysis of variance of semi-parametric and non-parametric effects of the covariate in the smoothing splines model of equation 1.7 respectively. Analysis of the effects of the individual smooth splines models indicate that the semi-parametric and non-parametric effects of the predictors are all statistically significant at 1 per cent level.

---

[7] *If EDF reduce to 2, then all observations are utilized and the resulting curve tallies with the least squares regression line. Of course this will result in a model with low residual variance but high bias. Hence, the EDF are determined in some ways using the cross validation method of re-sampling.*

Table 1 ANOVA of semi-parametric effects in smoothing splines model

Effect	Df	SS	MS	F-value	*p*-value
s(fgre, df=22.52)	1	1476.77	1476.77	310085.00	2.2x10^{-16}*
s(cp, df=13.06)	1	22.00	22.00	4618.00	2.2x10^{-16}*
s(re, df=25.27)	1	7.21	7.21	1513.20	2.2x10^{-16}*
s(ce, df=37.21)	1	11.83	11.83	2483.00	2.2x10^{-16}*
Residuals	76.94	0.37	0.00		

Table 2 ANOVA of non-parametric effects in smoothing splines model

Effect	Non-parametric df	Non-parametric F-value	*p*-value
s(fgre),df=22.52)	21.50	34.28	2.2x10^{-16}*
s(cp),df=13.06)	12.10	30.06	2.2x10^{-16}*
s(re),df=25.27)	24.30	94.23	2.2x10^{-16}*
s(ce),df=37.21)	36.20	60.34	2.2x10^{-16}*

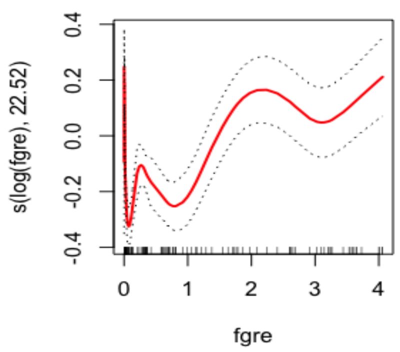
*Note: * statistically significant at 1 per cent level. Reject the null hypothesis of linearity.*

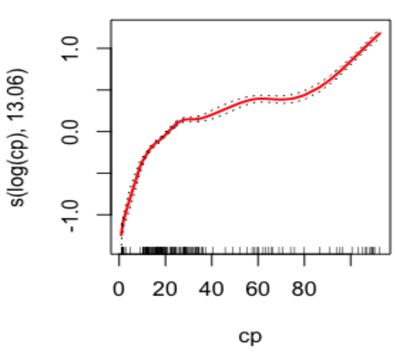
Null Deviance=1482.91 on 175 degrees of freedom, Residual Deviance= 0.37 on 76.9387 degrees of freedom, AIC= -387.11. ***Source****: Personal computation using* ***R*** *Studio "gam" package (Trevor Hastie, 2015).*

By using the "*smooth.splines*" function in the "*splines*" package of the **R** Studio programming language, we calculated the optimal tuning parameters Ω = 1.2613x10^{-6}, 2.0153x10^{-6}, 7.9291x10^{-7} and 1.5286x10^{-7} for each smoothing functions f_j, j=1,2,3,4 respectively which control the tradeoff between goodness-of-fit and smoothness of the model in equation 1.7.

The null model which estimates one parameter for the data set has a deviance of 1482.91. The deviance was reduced to 0.37 with about 76.94 degrees of freedom, one for the intercept and the remaining for the four predictors. This reduction was achieved with probability of 2.2x10^{-16}. The p-values under the non-parametric ANOVA correspond to the null hypothesis of linearity versus the alternative of non-linear relationship. Extremely low p-values for all the smoothing functions indicate non-linear functions are sufficient for all the terms in the model in equation 1.7. Also, we tested for evidence of lack of fit using the residual deviance or residual sum of squares and the degree of freedom by conducting the deviance test. This test give a maximum Chi-square *p*-value = 1.00 showing no evidence of lack of fit.

Smoothing Splines plots showing relationships between gross domestic products and its predictors





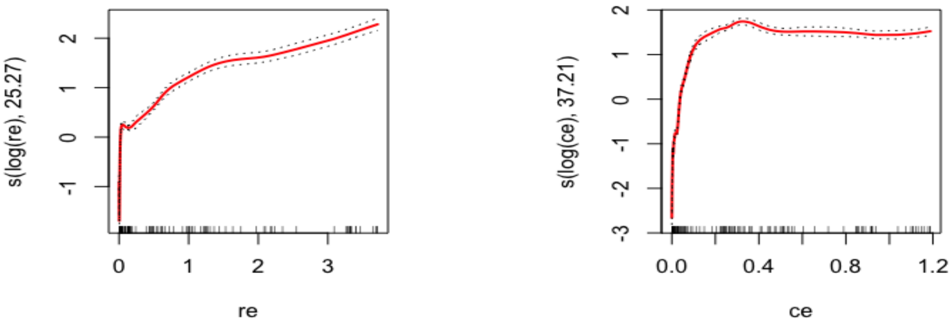


Figure 2: *Smooth spline plots of the relationships between each predictor and the gross domestic products (**gdp**) in the fitted model (equation 1.7). Each plot displays the relationship between X_j versus Y holding the remaining predictors constant in the model. Also, each plot shows the fitted function and 2 times standard errors*[8]*. All functions are smoothing splines with 22.52, 13.06, 25.27 & 37.21 EDF respectively.*
***Source**: Personal computation using **R** Studio "splines" package (**R** Core team, 2016).*

In **Figure** 2, the response variable ***gdp*** is expressed in terms of mean deviation, thus each smooth function f_j is centred and represents how ***gdp*** changes relative to its mean with unit changes in the predictors. Hence, the zero value on the response-axis is the mean of ***gdp***. The upper left panel shows that holding all other predictors fixed, gross domestic products rise with low, intermediate and high values of federal government retained earnings (***fgre***). The upper right panel showing the effect of crude oil prices on ***gdp*** indicate that ***gdp*** continue to rise for every increase in crude oil price (***cp***). However, there was a sharp rise in ***gdp*** from $80 per barrel mark. The lower left panel of **Figure** 2 shows the effects of recurrent expenditures (***re***) on ***gdp*** is also positive on ***gdp***. The lower right panel indicate that holding all other covariates fixed, ***gdp*** remained stable for every increase in the value of capital expenditures (***ce***) above the ₦400 billion mark.

The Local Regression Model (LRM) on Domestic Output in Nigeria

Since p = 4, we postulated another model in equation 1.1 referred to as the local regression based on local scoring. Its methodology involves estimating the effects in equation 1.1 at a target point using only the nearby observations. For each function of the predictors f_j in equation 1.1, we computed a span $h = k/n$ of training observations closest to $\boldsymbol{x}_o$, k is the number of observations in the neighbourhood. We then assign a weight $W_{io} = W(x_i, x_o)$ to each point in the neighbourhood. A *weighted least squares* regression of Y_i on X_i using the W_{io} was fitted by minimizing

$$\sum_{i=1}^{n} W_{io}(Y_i - \beta_o - \beta_1 X_i)^2 \qquad 1.8$$

[8] *The dotted curve is the estimated 95 per cent confidence interval for the smoothing splines.*

We assumed a linear form for the functions and calculated the span *h* using the efficient cross-validation approach. Span selection was done by employing the expensive leave-one-out cross-validation of the deviance or residual sum of squares in order to obtain an unbiased approximation of the Kullback-Liebler distance. This method selected span of ~0.29, ~1.63, ~0.30, ~0.46 for the local regression functions on federal government retained earnings, crude oil prices, recurrent expenditure and capital expenditure respectively. Therefore, *k* the number of nearest observations is 51, 287, 52 and 81 for f_1, f_2 , f_3 *and* f_4 respectively. The lower the value of *h*, the "*wigglier*" the curve. Also, smaller values of *h* imply that the curve is more flexible and "*local*" (**Figure** 3) which reduces the model bias, at the same time minimizing the variance proportionately. The local regression fit on ***gdp*** results in equation 1.9 below with the resultant local plots shown in **Figure** 3.

$$E(gdp_i / fgre,cp,re,ce) = 2.18 + 0.34\,fgre - 0.26cp + 0.09(cp)^2 + 0.48re + 0.18ce \qquad 1.9$$

Table 3 ANOVA of semi-parametric effects in local regression model

Effect	df	SS	MS	F-value	*p*-value
Lo(fgre,*h*=0.29)	1	1469.95	1469.95	525208.37	$2.2\text{x}10^{-16}$*
lo(cp,*h*=1.63,degree=2)	2	7.24	3.62	128.63	$2.2\text{x}10^{-16}$*
lo(re,*h*=0.30)	1	7.58	7.58	269.35	$2.2\text{x}10^{-16}$*
lo(ce,*h*=0.46)	1	0.86	0.86	30.51	$1.4\text{x}10^{-7}$*
Residuals	154.98	4.35	0.03		

Table 4 ANOVA of non-parametric effects in local regression model

Effect	Non-parametric df	Non-parametric F-value	*p*-value
lo(fgre, *h*=0.29)	6.8	6.25	$2.4\text{x}10^{-6}$*
lo(cp, *h*=1.63,degree=2)	0.1	4.69	0.0585***
lo(re, *h*=0.30)	5.5	20.64	$2.2\text{x}10^{-16}$*
lo(ce, *h*=0.46)	3.0	30.30	$1.3\text{x}10^{-15}$*

*Note: *,*** statistically significant at 1 per cent and 10 per cent levels, do not accept the null hypothesis of linearity. Null Deviance=1482.91 on 175 degrees of freedom, Residual Deviance= 4.35 on 154.98 degrees of freedom, AIC= -106.86.*
***Source**: Personal computation using **R** Studio "gam" package (Trevor Hastie, 2015).*

Table 4 shows that only the local function on crude oil price has a significant non-parametric and non-linear effect in the gross domestic products model of equation 1.9 only at10 level of significance. This indicates a local polynomial fit between ***cp*** and ***gdp***. A reduced deviance from 1482.91 to 4.35 with 154.98 degrees of freedom was achieved on this model. Similar to the smooth spline model, the deviance test on the local regression model give a maximum Chi-square *p*-value = 1.00 indicating no evidence of lack of fit.

Local regression plots showing relationships between gross domestic products and its predictors

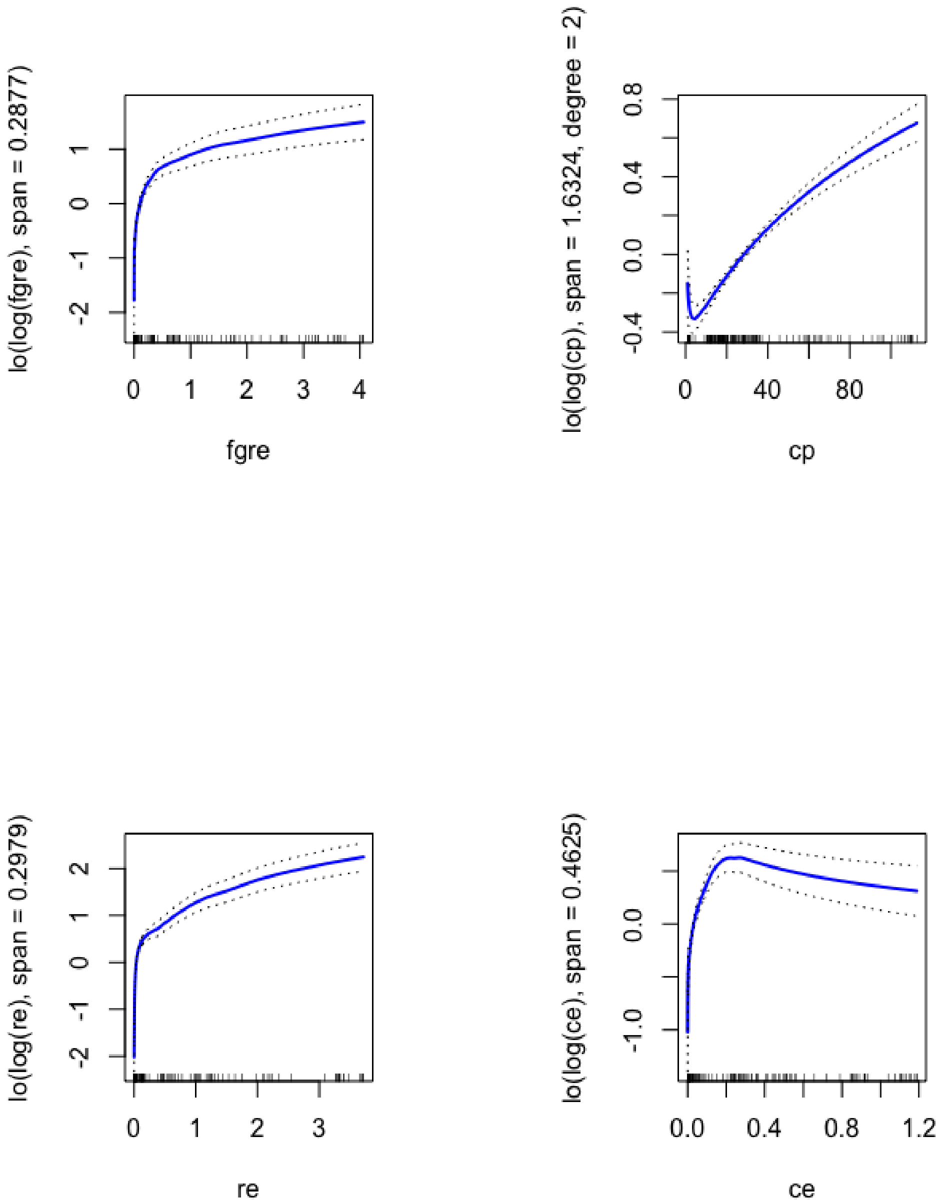


Figure 3: *Plots of the relationships between each predictor and the gross domestic products (**gdp**) in the fitted model (equation 1.9). Each plot displays the fitted function and two times standard errors. All functions are local regression with h=0.29, 1.63, 0.30 and 0.46 for **fgre**, **cp**, **re** and **ce** respectively. **Source**: Personal computation using **R** Studio "gam" package (Trevor Hastie, 2015).*

Generally, the curves in **Figure** 3 are smoother than **Figure** 2, which is especially obvious in the cases of functions on federal government retained earnings, crude oil price and capital expenditures. However, the acceptance of the null hypothesis of linearity for the function crude oil prices is surprising. We suggest a further postulation whereby the function on crude oil prices will be local polynomial of degree 2.

Local Quadratic Regression Model and Composite GAM on Domestic Output in Nigeria

Next we introduced a polynomial of degree 2 (i.e. Local Quadratic Regression) in the model on local regression (equation 2.0) and a composite GAM in which some of the functions in equation 1.1 are local regression and the remaining are smooth splines (equation 2.1). These are to ensure we achieve models with reduced deviances and stronger predictability. Specifically, for the composite GAM, we specify a smooth spline function for the additive model on crude oil price due largely to its high window size h=1.63 and k=287 nearby observations in equation 1.9. We summarize the results in **Table** 5 below;

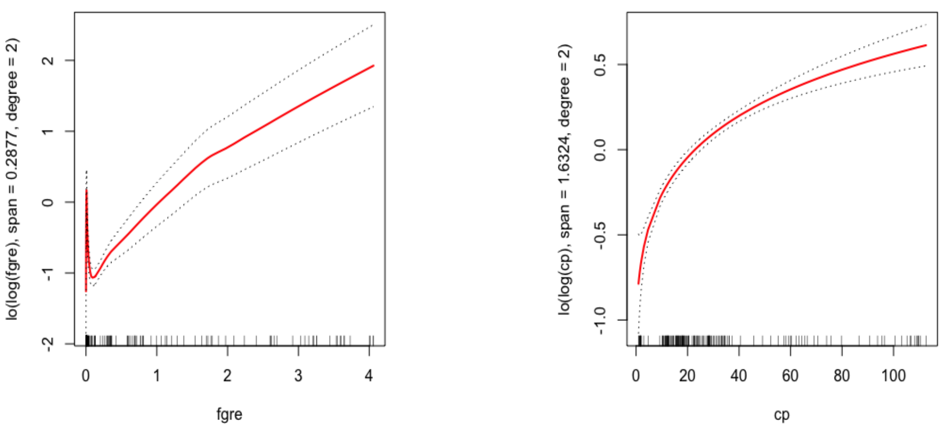
$$E(gdp_i / fgre,cp,re,ce) = 1.636 + 0.4242 fgre_i + 0.520(fgre_i)^2 + 0.1592 cp_i + 0.288(cp_i)^2 + 0.3357 re_i - 0.0203(re_i)^2 + 0.3632 ce_i - 0.0055(ce_i)^2 \quad 2.0$$

$$E(gdp_i / fgre,cp,re,ce) = 1.3268 + 0.3173 fgre_i + 0.3180 cp_i + 0.4243 re_i + 0.2429 ce_i \quad 2.1$$

Table 5: Summary of analysis of Local Quadratic Regression and Composite GAM

Model Type	# Effects	#Significant effects	Null Deviance	Residual Deviance	AIC
Local Quadratic Regression	8	7	1482.91	2.69	-171.10
Local Composite GAM	8	8	1482.91	2.72	-167.99

Local Quadratic regression plots showing relationships between gross domestic products and its predictors



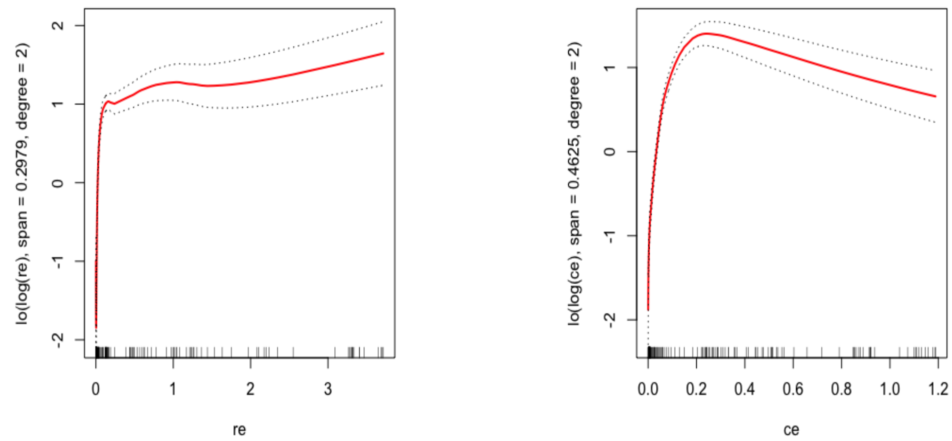


Figure 4:*Plots of the local polynomial regression of degree = 2 of gross domestic products (**gdp**) on the predictors in the fitted model (equation 1.9). Each plot displays the fitted function and two times standard errors.*
***Source**: Personal computation using **R** Studio "gam" package (Trevor Hastie, 2015).*

Composite GAM plots showing relationships between Domestic Output and its predictors

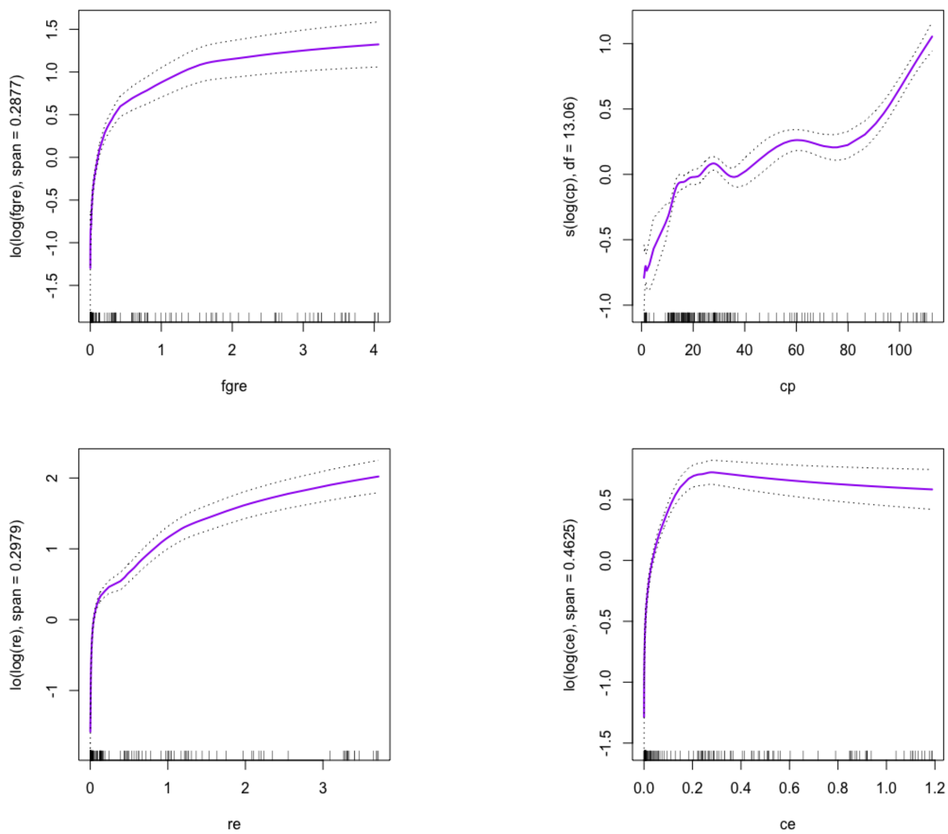


Figure 5: *Composite GAM with smoothing spline function on crude oil price and local regression functions on federal government retained earnings, recurrent and capital expenditures. **Source**: Personal computation using **R** Studio "gam" package (Trevor Hastie, 2015).*

In **Figure** 5, the upper right panel shows the smooth spline on crude oil price indicating a rather unstable effect on gross domestic products. This plot is wigglier than the smooth function in **Figure** 2. These two models show no evidence of lack of fit with high Chi-square *p*-value = 1.00 and 1.00 respectively. The main aim of this current work is find an optimal model for predicting the gross domestic products. Hence, we conducted a nested analysis of deviance on the four postulated models as follows.

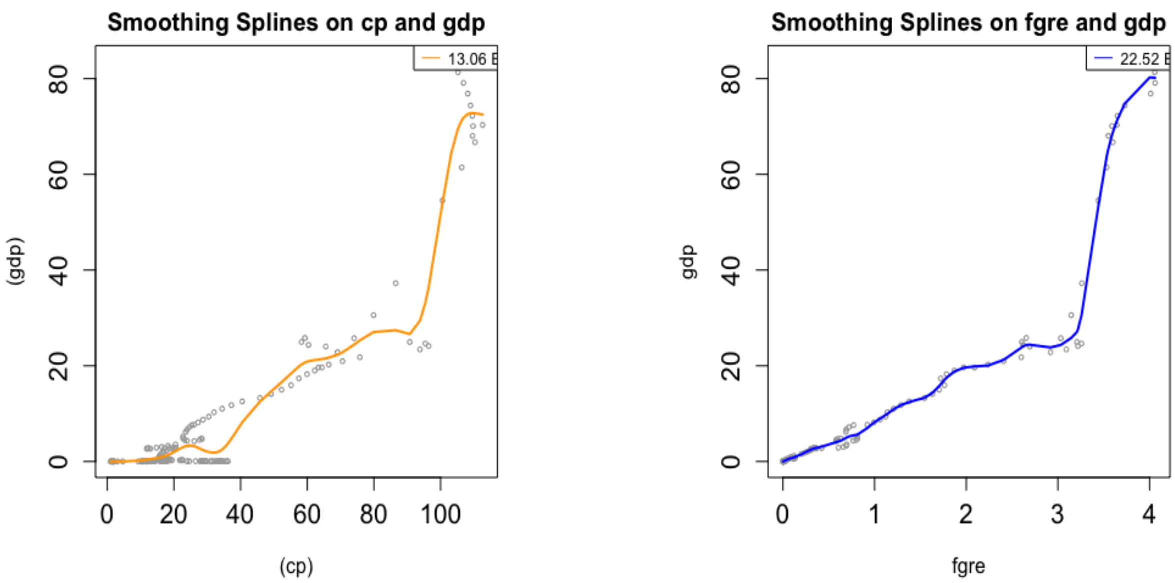
Table 6 Analysis of deviance on four postulated models

Model	Residual df	Residual Deviance	Deviance	Pr(>Chi-square)
Local Regression	154.98	4.35		
Composite GAM	143.65	2.72	1.67	2.2x10^{-16} *
Local Quadratic Regression	144.13	2.68	-0.03	
Smoothing Splines	76.94	0.37	2.35	2.2x10^{-16} *

**Indicate model is superior to the previous model at 1 per cent level of significance.*
***Source**: Personal computation using **R** Studio "gam" package (Trevor Hastie, 2015).*

Analysis of deviance in **Table** 6 above shows the residual degrees of freedom, residual deviance, model deviance and the chi-square *p*-value of achieving a lower deviance for each of the postulated models. This analysis indicates that whilst composite GAM is better than the local regression, smoothing splines is the best out of the four models. However, the test is inconclusive for local quadratic regression over composite GAM because the degree of freedom was less than one. Specifically, a lower deviance of 1.67 and 2.35 was recorded for Composite GAM and smoothing splines respectively with a probability of 2.2x10^{-16} each. Hence, the smoothing splines provide the best estimates for the actual gross domestic products values during the period under review as shown in the **Figure** 6. Furthermore, we employed a simple graphical assessment technique to compare the residuals of the four models using a boxplot (**Figure** 6).

Smoothing splines superimposed on scatterplots showing long-term trend between gross domestic products and its predictors



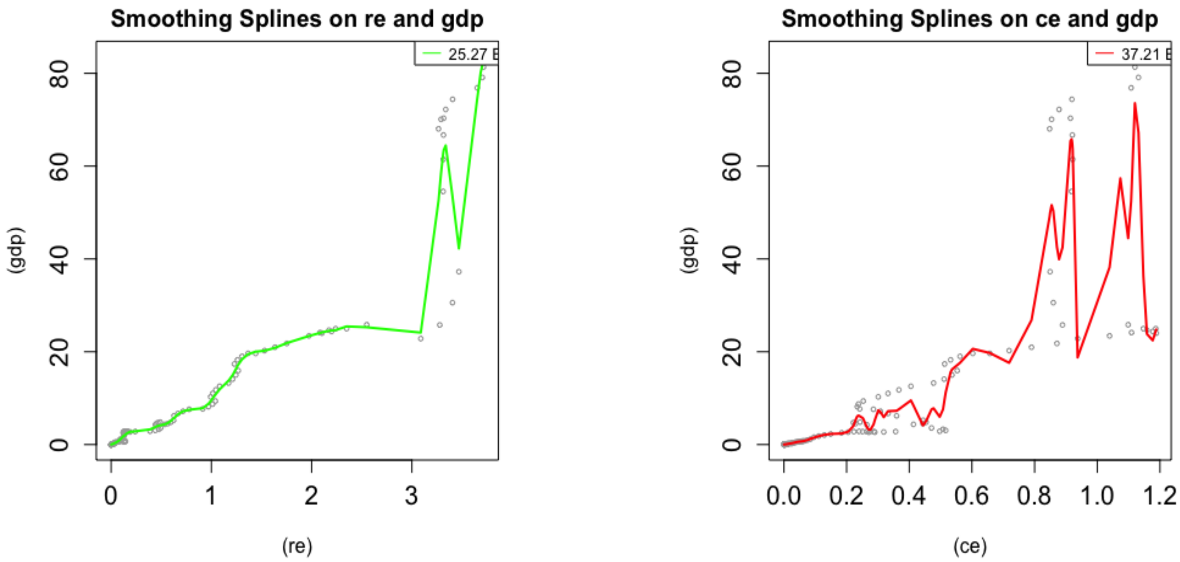


Figure 6: *Scatterplots with superimposed smoothing splines of the relationships between each predictor and the gross domestic products (**gdp**) in the fitted model (equation 1.7). Each plot shows various flexibilities and existence of non-linear long-term trend between **gdp** and the predictors.*
***Source**: Personal computation using **R** Studio "splines" package (**R** Core Team, 2016).*

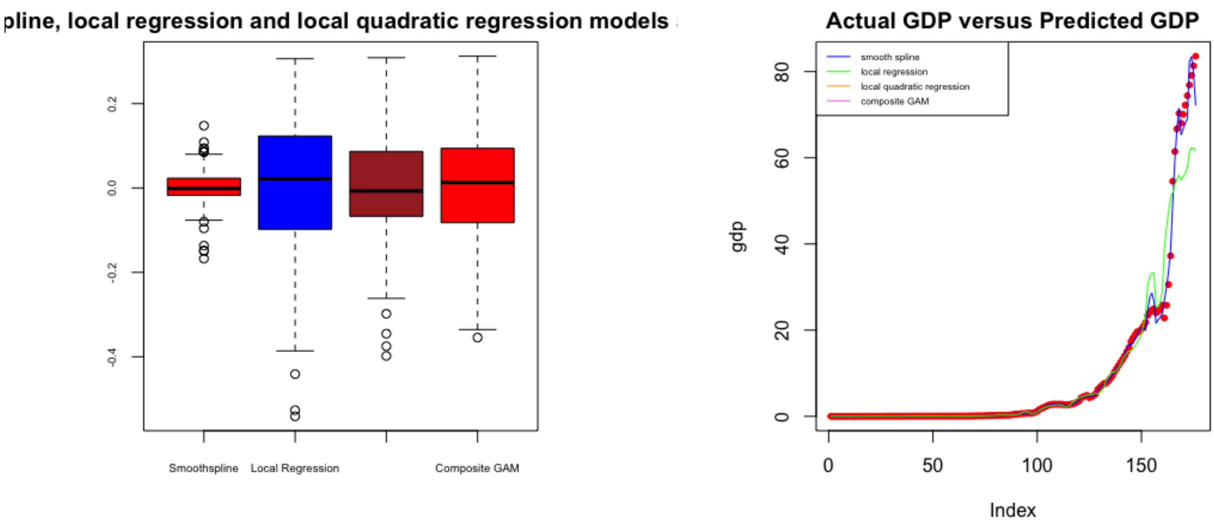


Figure 7: *Boxplot on residuals of smoothing splines, local regression, local quadratic regression and composite GAM models, which shows the upper, mid and lower quartiles are quite different for all the GAMs. Local regression, local quadratic regression and Composite GAM have extreme residuals concentrated outside the lower whisker outside. On the plots of predicted values of the gross domestic products in Nigeria from 2004 to 2014, it is obvious that the prediction power of the smoothing splines model is strongest especially from periods after 2009, which confirms the superiority of the model over the others.*

The two plots in **Figure** 7 alongside the analysis of deviance in **Table** 6 favor the model in equation 1.7 on smoothing splines for studying and predicting gross domestic products in Nigeria between 2004 to 2014. **Figure** 6 depicts the fitted lines of each of the smoothing splines function in equation 1.7 against the scatterplots in **Figure** 1. The graphs show that the long-term trend line in each function is non-linear, the residuals and the biases are small at each point. These lead to the smoothing of “noisy” gross domestic products data in Nigeria during the period under review.

Conclusion

In this research work, we have attempted to look at the relationships between four major macroeconomic variables and the Nigeria gross domestic products between 2004 and 2014. The purpose is to determine the nature of the relationships and make predictions, which are comparable to conventional methods. The analysis shows that rise in federal government retained earnings which comprise federation account levies, company income tax, custom and excise duties, value added tax (VAT) pool, allocation from excess crude account and independent sources such as government ministries/departments/agencies (MDAs) surpluses and volatile crude oil prices remain the major variables causing instability in the growth of the country’s gross domestic products, which is in line with conclusions drawn by Lawal *et al* (2016).

Reference

1. Craven, P. and G. Wahba (1979), “Smoothing noisy data with spline function”, *Numerische Mathematik* 31:377-403.
2. Friedman J.H and Stuetzle W. (1981), “Projection Pursuit Regression”, *Journal of American Statistical Association*, 76 817-823.
3. Friedman J.H and Stuetzle W. (1982), “Smoothing of Scatterplots”, Technical Report, Orion 3, Department of Statistics, Stanford University.
4. Golub, G.H., M.Heath and G. Wahba (1979), “ Generalized Cross-validation as a method for choosing a good ridge parameter”, *Technometrics* 21(2):215-223.
5. Haeran Cho, Yannig Goude, Xavier Brossat & Qiwei Yao (2013), “Modeling and Forecasting Daily Electricity Load Curves: A Hybrid Approach”, *Journal of the American Statistical Association*, 108:501,7-21.
6. Hastie, T. and R. Tibshirani (1986), “Generalized additive models (with discussion)”. *Statistical Science 1*, 297–318.
7. Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall/CRC.
8. Lawal Omoniyi Ganiyu, Alabi Nurudeen Olawale, Ige Sikiru Ajibade, Ibraheem Rahmat Abisola (2016) “The Nexus between Nigerian Government Spending and Domestic Output in the Presence of Long-Term Crude Oil Price Shock: A Conditional Unrestricted Equilibrium Correction Model Approach”, *Open Journal of Statistics*, (6): 412:425.
9. Wahba, G. (1990), Spline models for observational data, Philadelphia:SIAM.
10. Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*. CRC Press.
11. Trevor Hastie (2015). Gam: Generalized Additive Models. R package version 1.12. http://CRAN-project.org/package=gam.
12. R Core Team (2016), R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. URL http://www.R-project.org/.

13. H.Wickham (2009), ggplot: Elegant graphs for data Analysis, Springer-Verlag New York.