Journal of Statistical and Econometric Methods, vol. x, no. x, 2017, x-x
ISSN: xxxx-xxxx(print), xxxx-xxxx(online)
Scienpress Ltd, 2017

Forecasting a Composite Indicator of Economic Activity in Ghana: A Comparison of Data Science Methods

Emmanuel Thompson[1], **Ahmad M. Talafha**[2]

Abstract

Of recent, data science methods have been used to study and forecast financial and economic problems. This paper uses historical data to build a more parsimonious predictive model for making short term forecasts of the future values for the Composite Indicator of Economic Activity (CIEA) in Ghana. Based on our studies of a variety of shrinkage methods and a dimension reduction technique, we show empirically that the estimated model based on the Adaptive Elastic Net (Adaptive ENET) algorithm offers the greatest forecasting potential for the CIEA. A major finding in this paper was that, the Adaptive ENET model outperformed the benchmark model: Principal Component Regression (PCR) according to the cross validation root mean square error difference Statistic.

Mathematics Subject Classification : G12, C15, G22

Keywords: Composite Index of Economic Activity, Least Absolute Shrinkage and Selection Operator, Elastic Net, Principal Component, Artificial Neural Network.

[1] PhD, Department of Mathematics, Southeast Missouri State University. e-mail: ethompson@semo.edu

[2] PhD Student, Department of Mathematics, Western Michigan University. e-mail: a.talafha@wmich.edu

Article Info: *Received* : July 31, 2017. *Published online* : Month 00, 2017

1 Introduction

The aim of this paper is to use historical data to build a more parsimonious predictive model that can be used to make short term forecasts of the future values for the Composite Indicator of Economic Activity (CIEA) in Ghana. The CIEA symbolizes a point estimate that tracks the current state of an economy by providing useful picture of where the economy is headed. Policy makers in developing countries rely heavily on it as a way of judging how best an economy is performing [1]. CIEA is a measure that correlates with the current level of economic activity such as the real Gross Domestic Product (GDP) [2]. One of the biggest challenges in applied economic and financial research is finding a dependable model for forecasting macroeconomic and financial variables due to the simultaneous and comovements in most of these variables. Also, the use of only one indicator for short term forecasting is not reliable for the simple reason that most leading indicators produce erroneous signals [3]. To overcome some of these challenges is the emergence of a composite index that reflects a broader spectrum of the entire economy (real, monetary, fiscal, and external sector data) and [4] reviews the benefit of composite indexes. In 2003, economists at the Bank of Ghana (BoG) produced a working paper aimed at constructing a CIEA for Ghana and explored the likely use of the index to explain short run economic fluctuations. Using the conference board methodology synonymous to the Moore-Shiskin methodology, their findings indicate that economic activity for Ghana is still on the rising trend and that economic activity stagnated between February and April 2003. They concluded that, the composite indicator can be used to predict the overall direction of economic activity in Ghana[1].

In macroeconomic studies and monetary policy analysis, forecasting variables plays a crucial role. Authors of [5, 6] have established the importance of accurate prediction in having a better understanding of economic movements and implementation of effective monetary policies respectively. In the recent past, econometricians and statisticians have used Dynamic Factor Model (DFM) and its modifications and extensions to obtain valuable insights from a large panel of time series [7, 8, 9]. In applying DPMs, a comprise between loss of information and curse of dimensionality has to be made. The use of a limited number of principal components to summarize most of the information in time series results in a loss of information because the remaining principal components also contain some about of information. Also, in an attempt to use a large number of principal components enlarges the dimensionality of the model causing the degrees of freedom problem. For these reasons we put forward a different method of forecasting rooted in data science. Data science is a scientific discipline that combines statistics and computer science with foundations in mathematics [10]. The most common applications of data science techniques by companies are for tracking business processes and building a wide variety of fancy predictive models. For example, A credit card company uses predictive models to monitor its underwriting, pricing, and marketing activities. Healthcare insurers adjust payments and quality measures based on risk factors which are estimated from predictive models [11].

The proposed forecasting data science methods use the concept of shrinkage as opposed to dimension reduction to estimate model parameters. To explain in simple terms, the shrinkage methods operate by imposing a penalty on the least squares estimation method. Various assumptions have been made in the literature where the $l_1 - norm$ and $l_2 - norm$ or both which represent the penalty term were used to influence the parameter estimates in order to minimize the impact of collinearity. This paper focuses on two of the shrinkage methods - the least absolute shrinkage and selection operator (LASSO) and the elastic net (ENET). To improve upon the LASSO and the ENET results, we included the Adaptive LASSO and the Adaptive ENET. It is important to add that, with the advent of Big Data, low dimensional models suffer massive computational and statistical challenges such as scalability and storage problems, noise accumulation, spurious correlation, incidental endogeneity, and measurement error in capturing complex, dynamic patterns underlying large panels of time series [12]. In the context of an approximate factor model with dynamics, it has been shown that forecasts of a single time series based on principal components of a large number of predictors are first-order asymptotically efficient [7]. For this reason, we will use the principal component regression (PCR) as a benchmark to compare the proposed data science models. This paper is not intended to repeat what has already been well covered in many standard data science textbooks, for detailed review on PCR, we direct readers to [13, 7]. The rest of the paper is partitioned into six sections. Section 2 provides a short conceptual overview of the shrinkage estimation of the forecasting models. Section 3 describes the time series data used. Section 4 presents the statistical procedures for evaluating out-of-sample predictive accuracies of the forecasting models. Section 5 shows detailed results and section 6 concludes.

2 Data Science Methods

This section reviews the data science methods used in constructing the forecasting models. These methods can be classified as either shrinkage or dimension reduction methods. For this paper, the emphasis is on shrinkage methods.

2.1 Shrinkage Methods

Consider a stationary autoregressive model with endogenous and exogenous regressors as follows:

$$y_t = \alpha + \sum_{l=1}^{p} \phi_l y_{t-l} + \sum_{i=1}^{k} \sum_{j=1}^{q_i} \beta_{ij} x_{it-j} + \epsilon_t \tag{1}$$

where $y_{t-1}, ..., y_{t-p}$ and $x_{1t}, ..., x_{kt}$ are the endogenous and exogenous regressors respectively. ϵ_t is a random sequence of independent Gaussian with $E\epsilon = 0$ and $E\epsilon^2 = \sigma^2$. Let $\theta = (\phi, \beta)$, then the ordinary least squares (OLS) estimate $\hat{\theta}^{OLS} = \left(\hat{\phi}, \hat{\beta}\right)$ minimizes

$$RSS = \frac{1}{2} \sum_{t=1}^{T} \left(y_t - \alpha - \sum_{l=1}^{p} \phi_l \, y_{t-l} - \sum_{i=1}^{k} \sum_{j=1}^{q_i} \beta_{ij} \, x_{it-j} \right)^2 \tag{2}$$

The OLS estimates usually perform poorly in prediction and interpretation particularly when the size of the sample T is small relative to the number of predictors (both exogenous and endogenous variables). Shrinking or regularizing the coefficient estimates towards zero improves fit and significantly reduce their variances. In this paper we consider important shrinkage techniques, the LASSO and the ENET and their variants.

In LASSO regression [14], the penalty term has the form of the sum of absolute values. The solution can be obtained solving (3):

$$\hat{\theta}^{LASSO}(\lambda) = \underset{\theta}{argmin} \, RSS + \lambda \left(\sum_{l=1}^{p} |\phi_l| + \sum_{i=1}^{k} \sum_{j=1}^{q_i} |\beta_{ij}| \right) \tag{3}$$

where $\lambda \geq 0$. The LASSO simultaneous does continuous shrinkage and automatic variable selection with the imposition of an l_1-penalty.

ENET regression [15] is basically a combination of RIDGE and LASSO procedures. The estimates from ENET method are defined by

$$\hat{\theta}^{ENET}(\lambda) = \underset{\theta}{argmin}\, RSS + \lambda_2 \left(\sum_{l=1}^{p} \phi_l^2 + \sum_{i=1}^{k} \sum_{j=1}^{q_i} \beta_{ij}^2 \right) + \lambda_1 \left(\sum_{l=1}^{p} |\phi_l| + \sum_{i=1}^{k} \sum_{j=1}^{q_i} |\beta_{ij}| \right) \tag{4}$$

The Adaptive LASSO seeks to minimize (5) [16]:

$$\hat{\theta}^{AdpLASSO}(\lambda) = \underset{\theta}{argmin}\, RSS + \lambda \left(\sum_{l=1}^{p} \hat{w}_l |\phi_l| + \sum_{i=1}^{k} \sum_{j=1}^{q_i} \hat{w}_{ij} |\beta_{ij}| \right) \tag{5}$$

where $\hat{w}_l$ is the adaptive data-riven weight and can be computed by $\hat{w}_l = \left(\left| \hat{\phi}_l(ridge) \right| \right)^{-\gamma}$. γ is a positive constant and $\hat{\phi}^{ini}$ is a consistent estimate of ϕ. Also, $\hat{w}_{ij}$ is computed by $\hat{w}_{ij} = \left(\left| \hat{\beta}_{ij}(ridge) \right| \right)^{-\gamma}$ where $\hat{\beta}^{ini}$ is a consistent estimate of β.

The adaptive ENET can be viewed as a combination of the ENET and the adaptive LASSO. Thus, the Adaptive ENET solves the problem [15]

$$\hat{\theta}^{AdpENET}(\lambda) = \underset{\theta}{argmin}\, RSS + \lambda_2 \left(\sum_{l=1}^{p} \phi_l^2 + \sum_{i=1}^{k} \sum_{j=1}^{q_i} \beta_{ij}^2 \right) + \lambda_1 \left(\sum_{l=1}^{p} \hat{w}_l |\phi_l| + \sum_{i=1}^{k} \sum_{j=1}^{q_i} \hat{w}_{ij} |\beta_{ij}| \right) \tag{6}$$

where the adaptive weights $\hat{w}_l$ and $\hat{w}_{ij}$ are obtained by $\hat{w}_l = \left(\left| \hat{\phi}_l(enet) \right| \right)^{-\gamma}$ and $\hat{w}_{ij} = \left(\left| \hat{\beta}_{ij}(enet) \right| \right)^{-\gamma}$ respectively. Note that $\hat{w}_l$ and $\hat{w}_{ij}$ in (6) are estimators associated with the ENET algorithm.

3 Data

This section describes the data series and their associated transformations.The data set can be downloaded from the BoG website: `https://www.bog.gov.gh/`. The data set consists of 33 variables (Table 1 in Appendix A) spanning the period February 2000 − March 2016.

Table 2 displays the descriptive statistics of the original variables. These variables were all positive in terms of skewness and kurtosis, however, the JB test for normality indicates that all the original variables were not normally distributed ($p - values < 0.01$).

Table 2: Descriptive statistics for the Original Variables

						JB Normality Test	
Variable	N	**Mean**	**SD**	**Skewness**	**Kurtosis**	**Statistic**	p**-value**
CIEANom	194	511.32	400.1743	0.7972	2.419	23.278	0.0000
CIEAReal	194	224.74	126.5762	2.1609	15.704	1455.6	0.0000
GSE-ASI	194	3693.10	2776.8922	0.8077	2.6749	21.948	0.0000
INF-YOY	194	16.93	7.8581	1.4372	4.5592	86.435	0.0000
TBR-91	194	21.49	9.5100	0.7663	3.1644	19.207	0.0001
BCROIL	194	66.33	32.9065	0.2483	1.7412	14.803	0.0006
BNCG	194	2064.65	2636.3761	1.4457	3.7718	72.391	0.0000
CIC	194	2383.39	2406.0205	1.1117	3.0696	40.001	0.0000
CITOB	194	224.91	242.6918	1.0462	3.0679	35.426	0.0000
COB	194	2154.13	2164.5157	1.1317	3.1126	41.515	0.0000
CocoaP	194	2152.17	710.5970	0.0633	1.6579	14.689	0.0006
CPI- F	194	186.35	83.9133	0.3446	1.7483	16.505	0.0003
CPI- NF	194	225.45	127.9649	0.7529	2.45	20.773	0.0000
CPI- O	194	214.12	136.8050	3.6004	30.2735	6431.9	0.0000
DD	194	3101.02	3483.5252	1.2646	3.4464	53.318	0.0000
FCD	194	2857.99	3335.3403	1.4398	4.0926	21.948	0.0000
GIR	194	2443.80	1553.5424	0.2214	1.7737	13.741	0.0010
IBKEXRENDMUSD	194	1.45	0.8959	1.566	4.5682	99.171	0.0000
IBKEXRMAVEUSD	194	1.44	0.8813	1.558	4.5348	97.521	0.0000
IBKXEMAVEGBP	194	2.34	1.3286	1.4085	4.1928	75.644	0.0000
IBKXRAVEEURO	194	1.80	1.0817	0.9732	3.0501	30.645	0.0000
IBKXRENDMEURO	194	1.82	1.0921	1.0001	3.0808	32.392	0.0000
IBKXRENDMGBP	194	2.35	1.3346	1.408	4.1659	75.089	0.0000
INF-F	194	12.66	7.6997	1.0718	3.5357	39.465	0.0000
INF-NF	194	20.26	9.9586	1.699	5.7405	154.04	0.0000
IntBkWAve	194	19.62	8.5443	1.184	3.9632	52.825	0.0000
M1	194	5261.35	5658.9232	1.2071	3.2764	47.729	0.0000
M2	194	8647.06	9178.0168	1.1727	3.2887	45.137	0.0000
M2+	194	11496.16	12458.4691	1.2332	3.444	50.769	0.0000
MPR	194	18.88	5.0663	0.4807	1.7904	19.3	0.0001
RM	194	3509.35	3903.0914	1.2613	3.4063	52.771	0.0000
TBR-91 day	194	21.49	9.5093	0.7687	3.1676	19.332	0.0001
TotDep	194	11730.07	9087.4002	1.0912	3.3666	39.587	0.0000

In accordance with standard practice in macroeconomic literature, we used two widely known methods: the Augmented Dickey Fuller (ADF) [17] and Phillip Perron (PP) [18] tests to check for the existence or otherwise of unit root in the original variables. The results of these tests are shown in Table 3. The ADF test indicates that apart from CPI-O ($p-value = 0.010$) where the null hypothesis of a unit root can be rejected, the null hypothesis of a unit root for the rest of the variables cannot be rejected at the 0.05 significance level. Of all the variables tested, the PP test shows that the null hypothesis of a unit root can be rejected at significance level of 0.05 for CIEAReal ($p-value = 0.010$), CPI-O ($p-value = 0.010$), and INF-F ($p-value = 0.022$). The time plot of the individual variables are shown in Figures 1 to 12 in Appendix B. It is evident that majority of the variables have unit root non-stationarity property.

Table 3: ADF and PP Tests for the Original Variables.

Variable	**ADF Test**		**PP Test**	
	Test Statistic	***p*-value**	**Test Statistic**	***p*-value**
CIEANom	1.985	0.990	-1.113	0.918
CIEAReal	-2.765	0.069	-12.394	0.010
GSE-ASI	-1.727	0.415	-1.737	0.686
INF-YOY	-2.106	0.274	-2.685	0.290
TBR-91	-2.053	0.294	-1.824	0.650
BCROIL	-2.122	0.268	-1.488	0.791
BNCG	-0.036	0.952	-1.869	0.631
CITOB	1.762	0.990	-0.746	0.965
CNBP	0.091	0.962	-3.065	0.130
COB	1.820	0.990	-0.570	0.978
CocoaP	-1.625	0.452	-2.831	0.228
CPI- F	-1.804	0.386	-1.369	0.840
CPI- NF	-2.270	0.213	-2.729	0.271
CPI- O	-3.620	0.010	-6.410	0.010
DD	4.099	0.990	1.018	0.990
FCD	3.252	0.990	0.086	0.990
GIR	-1.305	0.572	-2.839	0.225
IBKEXRENDMUSD	2.982	0.990	-0.012	0.990
IBKEXRMAVEUSD	2.967	0.990	0.312	0.990
IBKXEMAVEGBP	1.186	0.990	-0.874	0.953
IBKXRAVEEURO	0.593	0.989	-1.842	0.642
IBKXRENDMEURO	1.304	0.990	-1.441	0.810
IBKXRENDMGBP	1.008	0.990	-1.154	0.912
INF-F	-2.143	0.260	-3.767	0.022
INF-NF	-1.974	0.323	-2.414	0.403
IntBkWAve	-1.789	0.391	-1.510	0.781
M1	3.309	0.990	0.272	0.990
M2	3.056	0.990	0.911	0.990
M2+	5.356	0.990	1.496	0.990
MPR	-1.047	0.667	0.294	0.990
RM	2.400	0.990	0.095	0.990
TBR-91 day	-2.046	0.296	-1.823	0.650
TotDep	-0.272	0.920	-1.286	0.875

To guarantee stationarity of all variables before modeling, we transformed all variables by taking natural log and first difference. Descriptive statistics and the unit root test results on the transformed variables are shown in Tables 4 and 5 respectively. The transformed variables exhibited positive kurtosis and the Jarque-Bera (JB) test for normality shows that none of the transformed variable was normal since $p - values$ were all less than 0.01. While some of the transformed variables were skewed right, majority were left skewed. The null hypothesis of a unit root for the all transformed variables can be rejected at the 0.05 level of significance for both the ADF and the PP tests (Table 5). It can therefore be concluded that all the transformed variables have achieved stationarity.

Table 4: Descriptive statistics for the Transformed Variables

						JB Normality Test	
Variable	N	**Mean**	**SD**	**Skewness**	**Kurtosis**	**Statistic**	p**-value**
CIEANom	193	0.01454	0.09119	−1.18734	69.36817	35467.00	0.0000
CIEAReal	193	0.00936	0.32143	−0.35764	72.81053	39195.00	0.0000
GSE-ASI	193	0.00492	0.15246	−10.87958	139.37037	153360.00	0.0000
INF-YOY	193	0.00083	0.09296	1.03531	16.02756	1399.30	0.0000
BCROIL	193	0.00189	0.09211	−0.84348	4.14395	33.41	0.0000
BNCG	193	0.02105	0.22824	−0.95372	7.70036	206.93	0.0000
CIC	193	0.02163	0.07714	1.38692	9.35858	387.01	0.0000
CITOB	193	0.02707	0.21117	1.85345	16.32924	1539.30	0.0000
COB	193	0.02091	0.06213	0.79845	4.14414	31.03	0.0000
CocoaP	193	0.00595	0.06895	−0.37683	4.44314	21.32	0.0000
CPI- F	193	0.00414	0.08761	−11.85668	157.40598	196250.00	0.0000
CPI- O	193	0.00597	0.16327	−3.24883	63.82491	30091.00	0.0000
DD	193	0.02442	0.06747	0.10831	4.46293	17.59	0.0002
FCD	193	0.02406	0.07703	−0.76251	16.48692	1481.50	0.0000
GIR	193	0.01267	0.09933	0.87737	4.55741	44.27	0.0000
IBKEXRENDMUSD	193	0.01183	0.03023	−1.14607	25.54137	4128.30	0.0000
IBKEXRMAVEUSD	193	0.01208	0.06471	2.37768	61.14852	27373.00	0.0000
IBKXEMAVEGBP	193	0.01148	0.04033	0.20849	11.49324	581.48	0.0000
IBKXRAVEEURO	193	0.01324	0.18846	−0.26362	86.58825	56189.00	0.0000
IBKXRENDMEURO	193	0.01270	0.04699	−0.55862	16.97745	1581.10	0.0000
IBKXRENDMGBP	193	0.01140	0.04624	0.51966	13.24765	853.18	0.0000
INF-F	193	0.00009	0.16575	0.66202	11.62322	612.08	0.0000
INF-NF	193	−0.00022	0.13351	−0.25997	23.29625	3314.80	0.0000
IntBkWAve	193	−0.00086	0.09291	−0.01772	9.67280	358.08	0.0000
M1	193	0.02253	0.05870	−0.02386	15.44010	1244.50	0.0000
M2	193	0.02300	0.03323	0.47490	3.44751	8.87	0.0119
M2+	193	0.02350	0.03187	0.51824	7.68764	185.35	0.0000
RM	193	0.02233	0.09058	−0.04048	12.26741	690.71	0.0000
TotDep	193	0.01217	0.28729	−2.30957	63.13548	29252.00	0.0000

Table 5: ADF and PP Tests for the Transformed Variables.

Variable	**ADF Test**		**PP Test**	
	Test Statistic	***p*-value**	**Test Statistic**	***p*-value**
CIEANom	-16.479	0.01	-30.271	0.01
CIEAReal	-16.610	0.01	-33.300	0.01
GSE-ASI	-9.239	0.01	-12.957	0.01
INF-YOY	-8.386	0.01	-12.387	0.01
BCROIL	-8.560	0.01	-10.507	0.01
BNCG	-11.956	0.01	-19.527	0.01
CIC	-10.133	0.01	-15.610	0.01
CITOB	-14.436	0.01	-24.426	0.01
COB	-9.560	0.01	-11.261	0.01
CocoaP	-11.287	0.01	-14.200	0.01
CPI- F	-9.341	0.01	-13.818	0.01
CPI- O	-12.317	0.01	-19.851	0.01
DD	-10.693	0.01	-17.413	0.01
FCD	-13.266	0.01	-20.002	0.01
GIR	-9.199	0.01	-16.435	0.01
IBKEXRENDMUSD	-6.910	0.01	-14.602	0.01
IBKEXRMAVEUSD	-12.345	0.01	-21.077	0.01
IBKXEMAVEGBP	-9.073	0.01	-14.695	0.01
IBKXRAVEEURO	-16.027	0.01	-29.680	0.01
IBKXRENDMEURO	-9.674	0.01	-16.754	0.01
IBKXRENDMGBP	-9.535	0.01	-16.538	0.01
INF-F	-10.878	0.01	-14.665	0.01
INF-NF	-10.102	0.01	-14.835	0.01
IntBkWAve	-9.629	0.01	-12.337	0.01
M1	-10.436	0.01	-15.263	0.01
M2	-9.324	0.01	-10.848	0.01
M2+	-9.566	0.01	-15.598	0.01
RM	-12.038	0.01	-19.839	0.01
TotDep	-9.694	0.01	-13.984	0.01

4 Forecasting Accuracy Metrics

In this section, we describe the statistical metrics for comparing the out-of-sample predictive accuracies of alternative models. In particular, we focus on two quantitative methods. One, the Root Mean Square Error difference ($\Delta RMSE$) statistic similar to the one used by [19, 20] for assessing the predictive abilities of asset pricing models. In this paper, we prefer to call the metric Cross Validation Root Mean Square Error difference ($\Delta CV - RMSE$) statistic because we hold-out a sample of observations before fitting begins and once fitting is completed, the held-out sample is used to evaluate the predictive performance of the fitted models. Two, the Diebold-Mariano test [21] to assess the forecast accuracy of two predictive models.

4.1 The Cross Validation Root Mean Square Error Difference Statistic

The $\Delta CV - RMSE$ statistic is given by:

$$\Delta CV - RMSE = \sqrt{\frac{\sum_{t=T+1}^{N}(y_t - \bar{y_t})^2}{N-T}} - \sqrt{\frac{\sum_{t=T+1}^{N}(y_t - \hat{y_t})^2}{N-T}} \tag{7}$$

where $N - T$ represents the number of held-out sample forecasts, y_t is the observed value of variable being predicted at time t, $t = T+1, ..., N$, $\bar{y_t}$ and $\hat{y_t}$ are the forecasts from a benchmark model and a proposed model respectively. A positive $\Delta CV - RMSE$ is an indication that the proposed model has a better predictive power compared to the benchmark model.

4.2 The Diebold-Mariano test

The Diebold-Mariano test statistic is given by:

$$DM = \left[\frac{1}{N-T}\left(\hat{\gamma_0} + 2\sum_{j=1}^{r-1}\hat{\gamma_j}\right)\right]^{-\frac{1}{2}} \frac{1}{N-t}\sum_{t=T+1}^{N} d_t \tag{8}$$

where d_t is the difference of squared forecast errors which is defined as:

$$d_t = e_{j,t}^2 - e_{k,t}^2 \tag{9}$$

$\hat{\gamma}_r$ is the estimated j^{th} autocovariance of the time series d_t, and r denotes the forecast horizon. The test statistic has an asymptotic standard normal distribution under the null hypothesis of no difference between the expected forecast performance of models j and k.

5 Results

The data set used for empirical analysis comprises of the variables that were transformed to ensure stationarity. Specifically, 29 stationary time series variables where CIEAReal was the predictor variable were used. The original time series variables started from February 2000 to March 2016 and they were transformed to include up to lags of 4 in both the response (natural log transformation) variable and the predictor (first difference of natural log) variables. In total there were 128 predictor variables and the modeling period commenced from July 2000. We define a training set or an in-sample period starting from July 2000 to January 2013 (151 months in total) and a testing set or an out-of-sample period from February 2013 to March 2016 (38 months in total). The training set was used to estimate alternative predictive regression models including lags up to a maximum of 4. The estimated regression models were validated on the testing set to ascertain their predictive capabilities. The proposed predictive regression models were estimated using shrinkage methods discussed under Section 2. However, for purposes of benchmarking, the PCR was estimated.

Table 6 shows the estimation results using the LASSO, ENET, Adaptive LASSO, and Adaptive ENET algorithms and PCR on the training set. The MSEs shown on the table were computed based on the testing set. The LASSO and ENET algorithms at their optimal values of λ estimated 75 ($MSE = 0.16647$) and 13 ($MSE = 0.06255$) non-zero coefficients out of 128 respectively. When $\gamma = 1$, the Adaptive LASSO and Adaptive ENET algorithms at their optimal λ values gave respectively 8 ($MSE = 0.14628$) and 4 ($MSE = 0.04434$) non-zero coefficients out of 128. For the PCR, the minimum mean square error ($MSE = 0.04481$) was achieved with the first principal component in the regression. From a comparison of the MSEs of the various predictive regression models, it is obvious that Adaptive ENET model has the least MSE, followed by PCR, ENET, Adaptive LASSO, and LASSO in that order. Clearly, the predictive power of Adaptive ENET model supersedes the rest of the models. Figures 12 to 16 (see Appendix C) display the graphs of out-of-sample comparison between the predicted values from each of the LASSO, ENET, Adaptive LASSO, Adaptive ENET, and PCR models to the $ln(CIEAReal)$. A close observation of these figures reveals the superiority of the Adaptive ENET model relative to the other proposed models in terms of predictive accuracy.

A rigorous comparison between the proposed models (LASSO, ENET, Adaptive LASSO, and Adaptive ENET) and the benchmark model (PCR) using formal tests are displayed in Table 7. The $\Delta CV - MSE$ statistic for each comparison of LASSO, ENET, or Adaptive LASSO model to the PCR was negative. However, the sign of the $\Delta CV - MSE$ statistic for the comparison between Adaptive ENET and PCR was positive. These test results show that, the Adaptive ENET model has a better predictive power relative to the benchmark model. The LASSO, ENET, and Adaptive LASSO models underperformed relative to the benchmark model in terms of the ability to forecast.

Table 6: Estimated Non-Zero Coefficients

Model	$\boldsymbol{\alpha}$	$\boldsymbol{\gamma}$	$\boldsymbol{\lambda}_{optimal}$	**Coefficients**	**MSE**
LASSO	1.0	-	0.21227	75	0.16647
ENET	0.2	-	0.00038	13	0.06255
Adaptive LASSO	1.0	1.0	0.17027	8	0.14628
Adaptive ENET	0.7	1.0	0.16993	4	0.04434
PCR	-	-	-	-	0.04481

The DM tests show that there were statistically significant differences $(p - value < 0.01)$ between the expected forecast performances of each of the LASSO, ENET, and Adaptive LASSO model to the benchmark model since the $p - values$ associated with these tests were less than 0.01 and all the DM statistics were negative. However, no statistically significant difference was observed between the expected forecast performance of the Adaptive ENET model and that of the PCR because the $p - value$ of the test was greater than 0.1 and the DM statistic was positive.

Table 7: Formal Tests for Model Comparisons

Benchmark - Proposed Model	**ΔCV-RMSE Statistic**	**DM-Test**	
	ΔCV-RMSE	**DM**	***p*-value**
PCR - LASSO	-0.29555	-8.1867	0.00000
PCR - ENET	-0.05355	-4.7108	0.00003
PCR - Adaptive LASSO	-0.17079	-2.4785	0.01802
PCR - Adaptive ENET	0.00059	0.0901	0.92870

6 Conclusions

Recent literature has amplified the significance of large information for forecasting and has recommended approaches based on factor models to handle problems with huge dimensionality. However, the fact remains that in statistics and econometrics, forecasting time series with a massive dimensional predictor space is an essential and stimulating problem.

This paper has studied a variety of shrinkage methods (LASSO, ENET, Adaptive LASSO, and Adaptive ENET algorithms) with a high dimensional predictor space to propose a more parsimonious predictive model for short term forecast of the future values for the CIEA in Ghana. In the modeling process, we have estimated the LASSO, ENET, Adaptive LASSO, Adaptive ENET, and the PCR models, and evaluated their out-of-sample predictive accuracies using the MSE metric. Results based on the MSE seem to suggest that the Adaptive ENET model has the greatest potential in accurately forecasting CIEA in Ghana, followed by PCR, ENET, Adaptive LASSO, and LASSO in that order.

A major finding in this paper was that, among the four proposed models, the Adaptive ENET model was the only one that outperformed the benchmark model (PCR) according to the $\Delta CV - MSE$ statistic. Using data science approaches for fashioning out forecasting models offer great tools for practicing statisticians and economists when dealing with intricate and high-dimensional economic and financial time series variables.

References

[1] B. Amoah, P. Abradu-Otoo, F. F. Blankson, M. Bawumia, Estimating a Composite Leading Indicator of Economic Activity for Ghana, *Bank of Ghana Working Paper No. 2003/04*, (2003).

[2] Francis Leni Anguyo, A Model to Estimate a Composite Inductor Of Economic Activity (CIEA) for Uganda, *Research Department Bank of Uganda*, (2011).

[3] Atabek et al, E. E. Cosar, S. S. Sahinoz *A New Composite Leading Indicator for Turkish economic activity,* Emerging Markets Finance and Trade, **41**(1), (2005), 45-64.

[4] Ronny Nilsson, *Confidence Indicators and Composite Indicators,* OECD, paper presented at the CIRET conference, Paris, (2000).

[5] J. Bai, S. Ng, Large Dimensional Factor Analysis, *Foundations and Trends in Econometrics*, **3**(2), (2008), 89-163.

[6] David E. Rapach, Jack K. Strauss, Guofu Zhou, Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy, *Review of Financial Studies*, **23**(2), (2010), 821-862.

[7] J. H. Stock, M. W. Watson, Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association*, **97**, (2002), 1167-1179.

[8] J. H. Stock, M. W. Watson, Macroeconomic Forecasting Using Diffusion Indexes, *Journal of Business and Economic Statistics*, **20**, (2002), 147-162.

[9] Jiahan Li, Weiye Chen, Forecasting Macroeconomic Time Series: LASSO-Based Approaches and Their Forecast Combinations with Dynamic Factor Models, *International Journal of Forecasting*, **30**(4), (2014), 996-1015.

[10] Benjamin S. Baumer, Daniel T. Kaplan, Nicholas J. Horton, *Modern Data Science with R*, Chapman and Hall/CRC, United Kingdom, 2017.

[11] Liran Einav, Jonathan Levin, The Data Revolution and Economic Analysis, *Innovation Policy and the Economy*, **14**(4), (2014).

[12] Jianqing Fan, Fang Han, Han Liu, Challenges of Big Data analysis, *National Science Review*, **1**(2), (2014), 293-314.

[13] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, New York City, 2013.

[14] Ryan J. Tibshirani, Regression Shrinkage and Selection via the Lasso, *J R Stat Soc.*, **58**(1), (1996), 267-288.

[15] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J R Stat Soc. Series B (Statistical Methodology)*, **67**(2), (2005), 301320.

[16] Hui Zou, The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**(476), (2006), 1418-1429.

[17] Said E. Said, David A. Dickey, Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order, *Biometrika*, **71**(3), (1984), 599-607.

[18] Peter C. B. Phillips, Pierre Perron, Testing for Unit Root in Time Series Regression, *Biometrika*, **75**(2), (1988), 335-346.

[19] Ivo Welch, Amit Goyal, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, *Rev Financ Stud*, **21**(4), (2008), 1455-1508.

[20] Corte P Della, L. Sarno, I. Tsiakas, Spot and forward volatility in foreign exchange, *Journal of Financial Financial Economics*, **100**, (2011), 496-513.

[21] Francis X Diebold, Robert S Mariano, Comparing Predictive Accuracy, *Journal of Business & Economic Statistics*, **13**, (1995), 253-265.

APPENDIX A

Table 1: Definition of Variables

NAME	**DEFINTATION**
CIEANom	Composite Index of Economic Activities (nominal)
CIEAReal	Composite Index of Economic Activities (real)
GSE-ASI	Ghana Stock Exchange All Share Index
INF-YOY	Overall Inflation
TBR-91	91 Day Treasury Bill Rate (%)
BCROIL	Brent Crude Oil
BNCG	Bank of Ghana Net Claims on Govt (GHC'm)
CIC	Currency in Circulation (GHC'm)
CITOB	Cash in the Tills of the Commercial Banks (GHC'm)
COB	Currency Outside Banks
CocoaP	Indicative Cocoa Prices: dollars/tonne
CPI- F	CPI-Food Index
CPI- NF	CPI-Non Food Index
CPI- O	CPI-Overall Index
DD	Demand Deposit (GHC'm)
FCD	Foreign Currency Deposit
GIR	Gross International Reserves ($ million)
IBKEXRENDMUSD	Inter-Bank Exchange Rate End Month (GHC/US$)
IBKEXRMAVEUSD	Inter-Bank Exchange Rates Monthly Average GHC/US$
IBKXEMAVEGBP	Inter-Bank Exchange Rates Monthly Average GHC/GBP
IBKXRAVEEURO	Inter-Bank Exchange Rates Monthly Aerage GHC/EURO
IBKXRENDMEURO	Inter-Bank Exchange Rates End Month GHC/EURO
IBKXRENDMGBP	Inter-Bank Exchange Rates End Month GHC/GBP
INF-F	Food Inflation
INF-NF	Non-Food Inflation
IntBkWAve	Inter-Bank Weighted Average
M1	Narrow Money (GHC'm)
M2	Broad Money (M2) (GHC'm)
M2+	Total Liquidity (M2+)(GHC'm)
MPR	Monetary Policy Rate (%)
RM	Reserve Money (GHC'm)
TBR-91 day	91 Day Treasury Bill Rate (%)
TotDep	Total Deposits

APPENDIX B

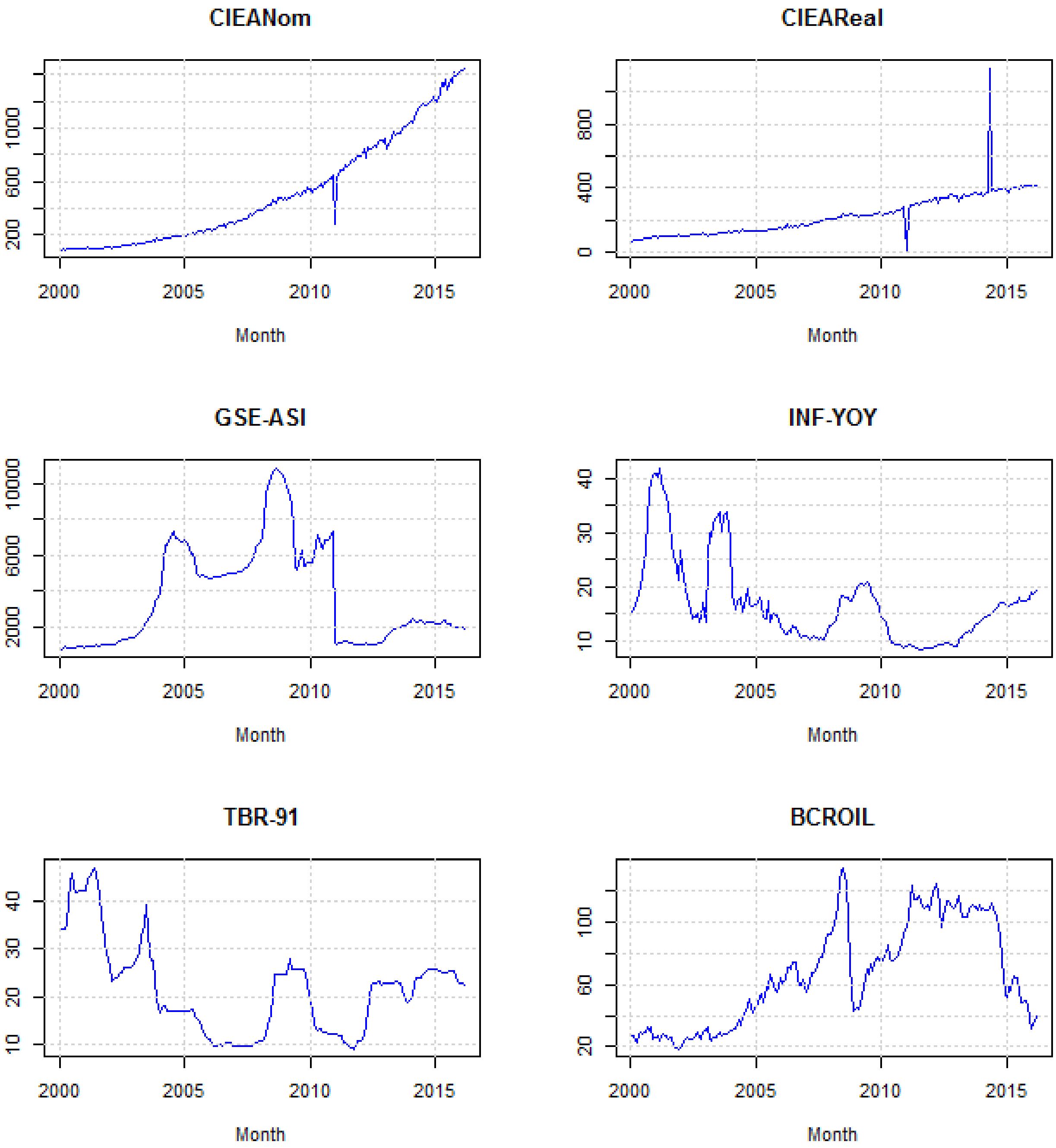

Figure 1: Time plot of the Original Variables

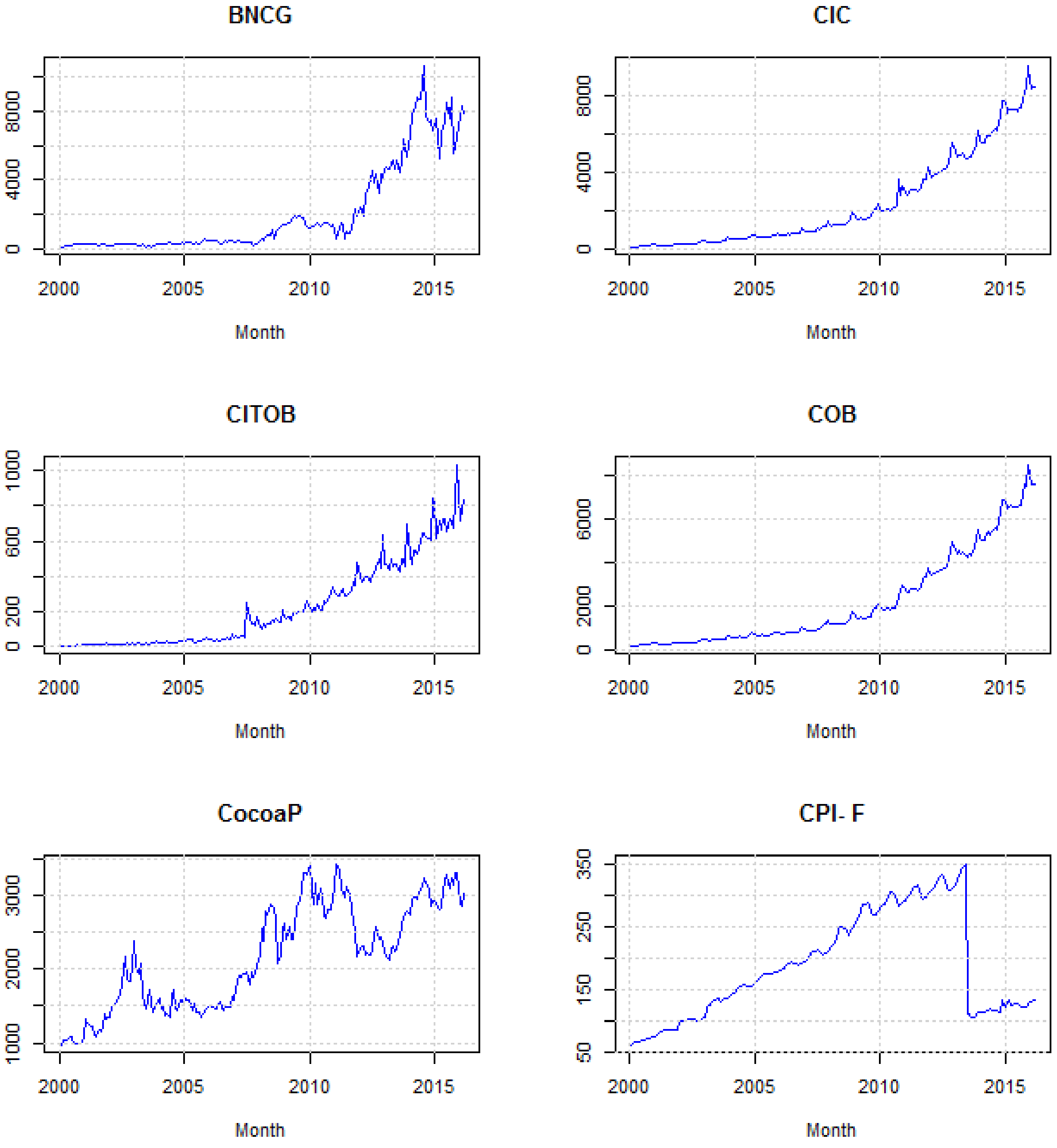

Figure 2: Time plot of the Original Variables

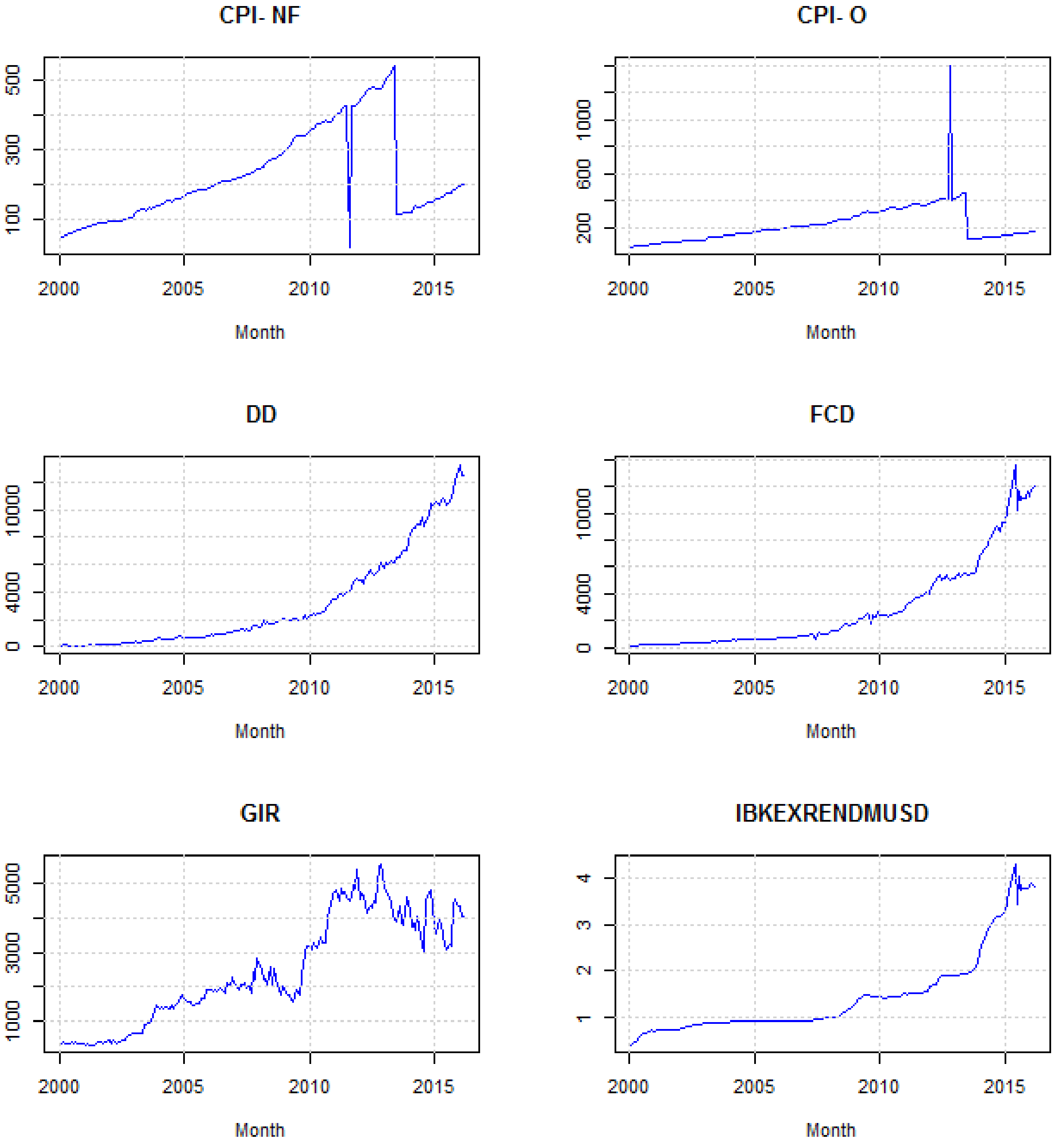

Figure 3: Time plot of the Original Variables

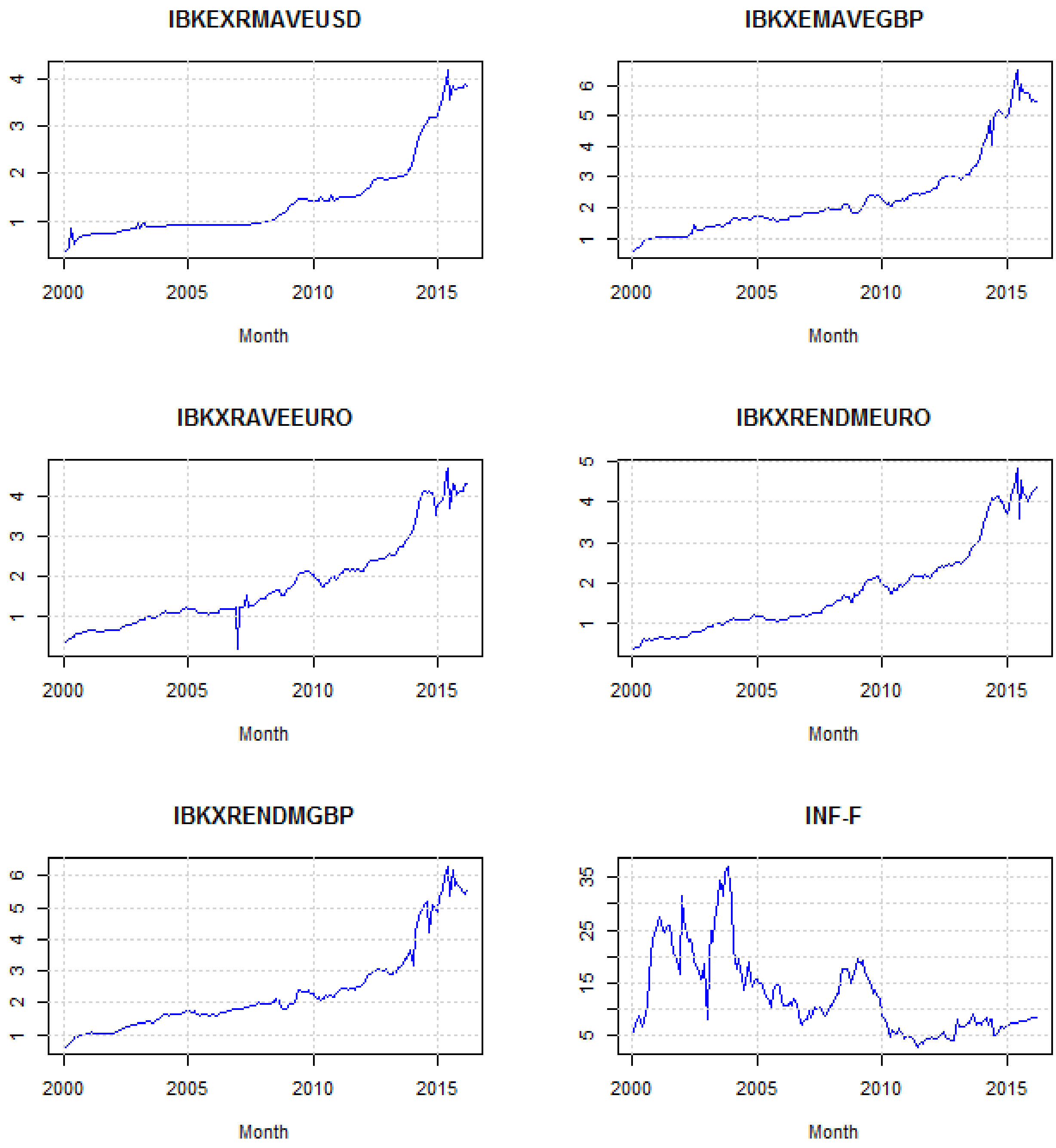

Figure 4: Time plot of the Original Variables

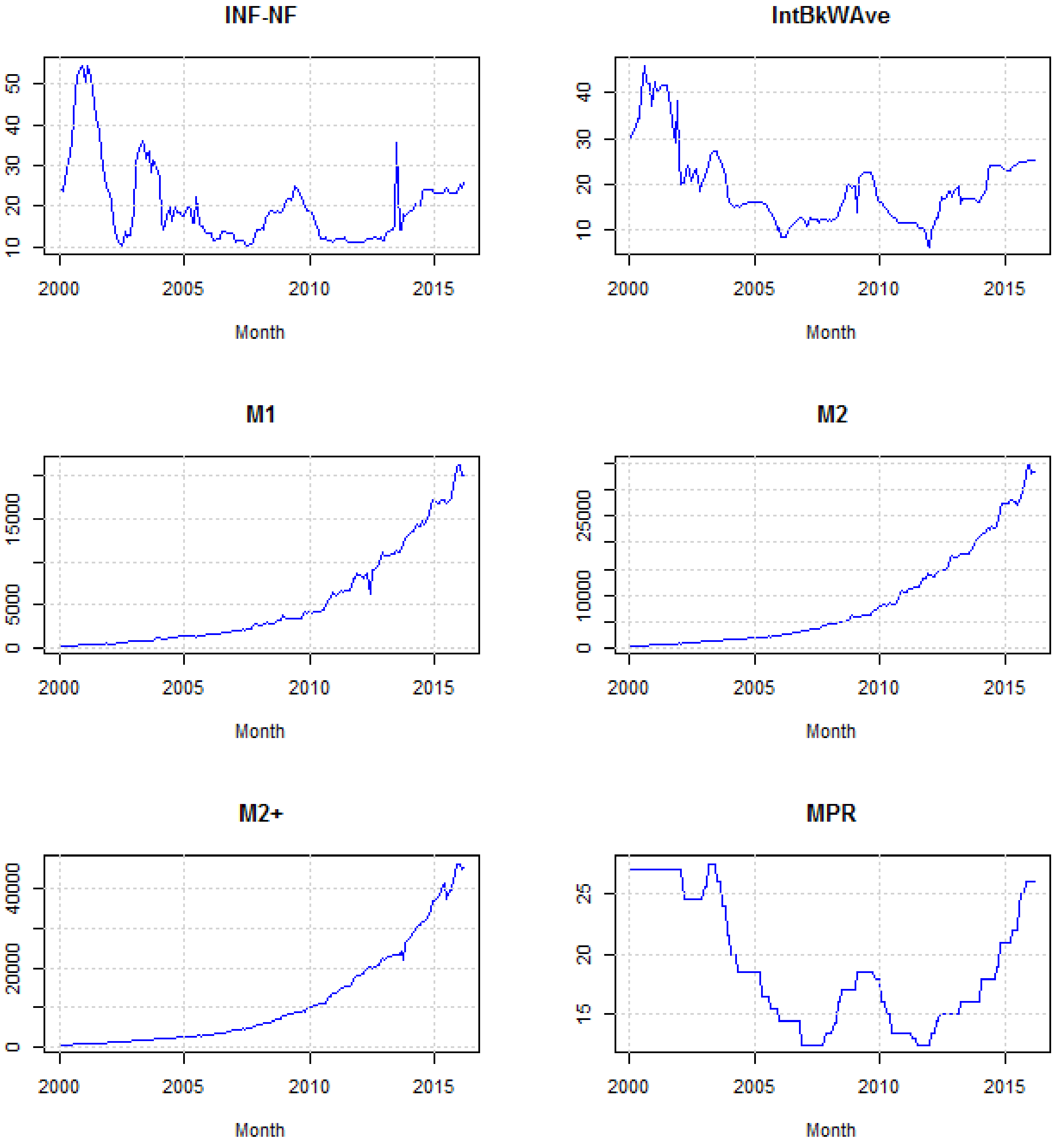

Figure 5: Time plot of the Original Variables

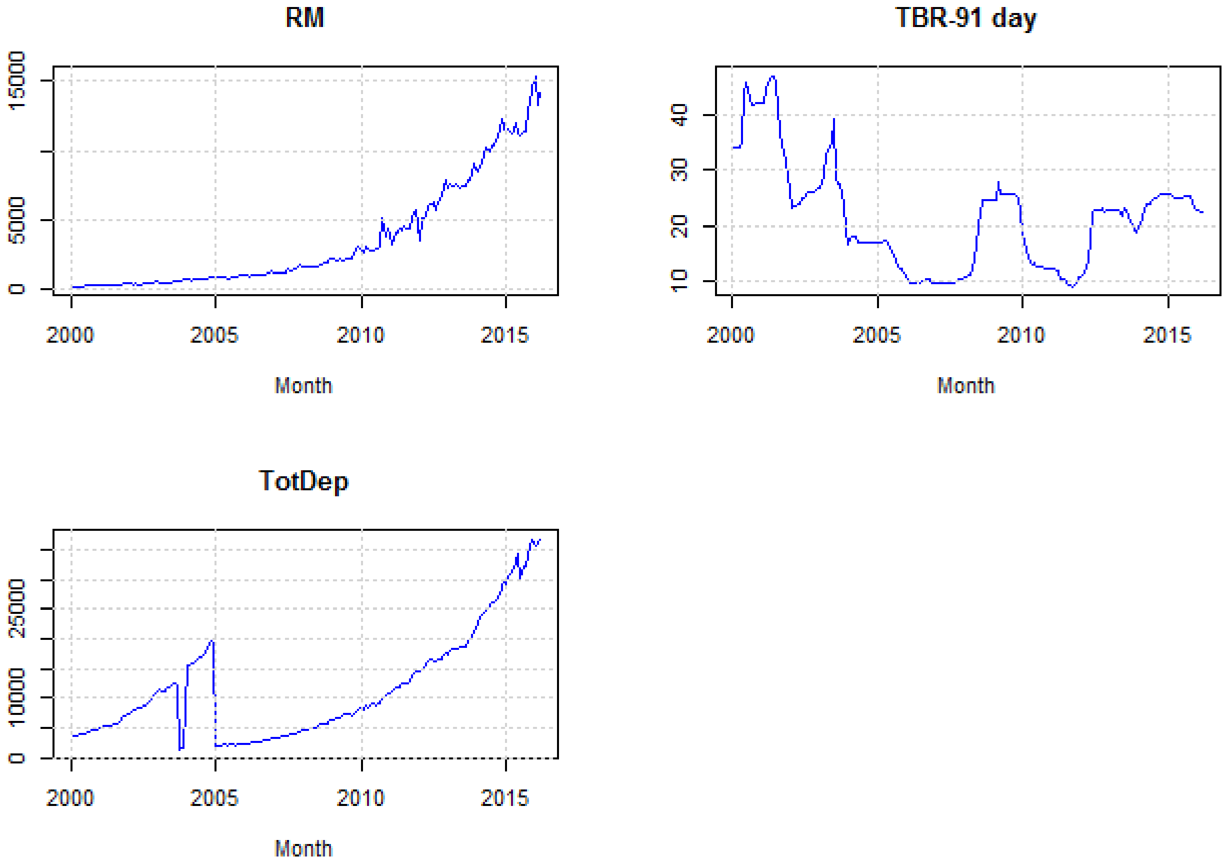

Figure 6: Time plot of the Original Variables

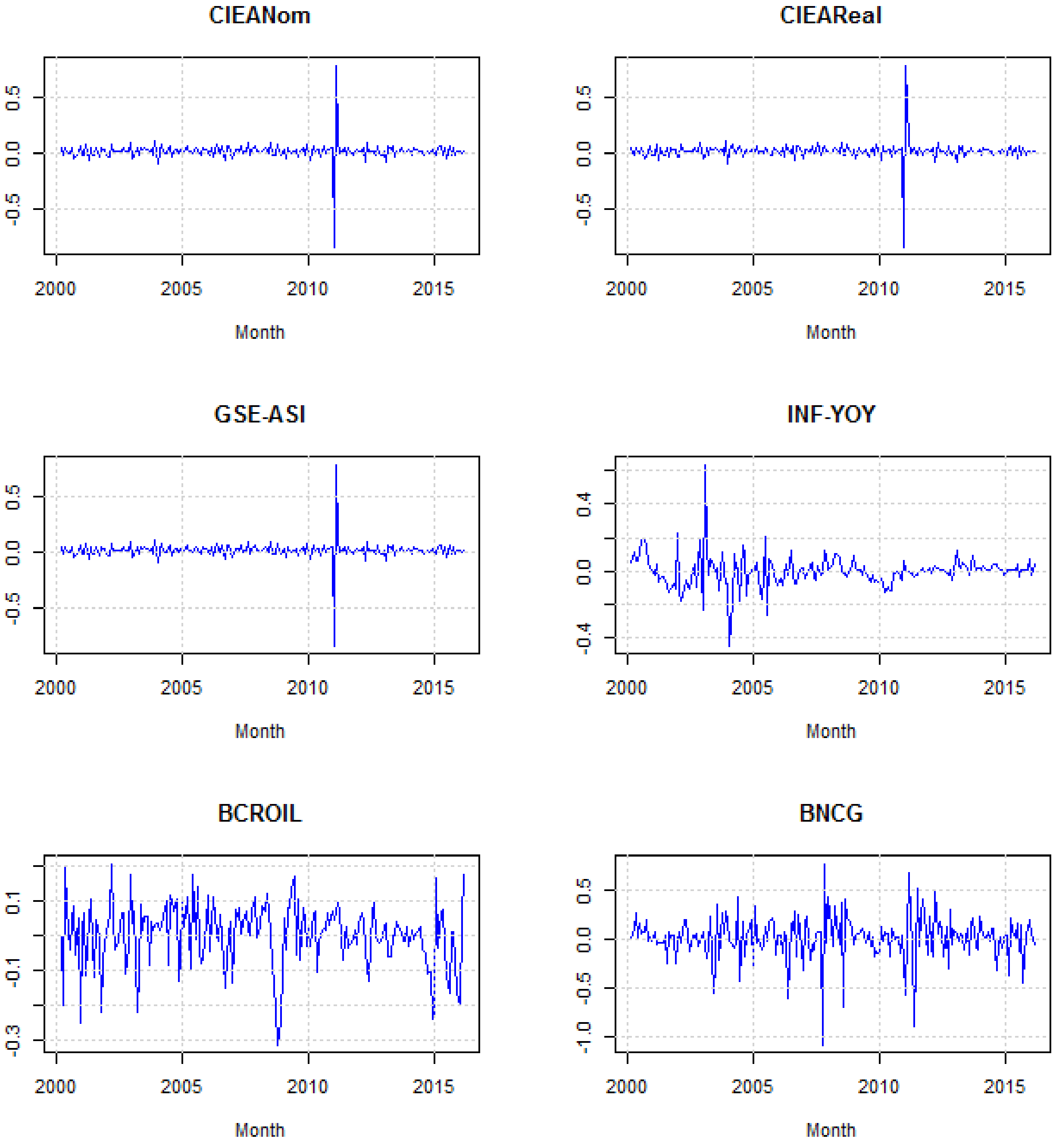

Figure 7: Time plot of the Transformed Variables

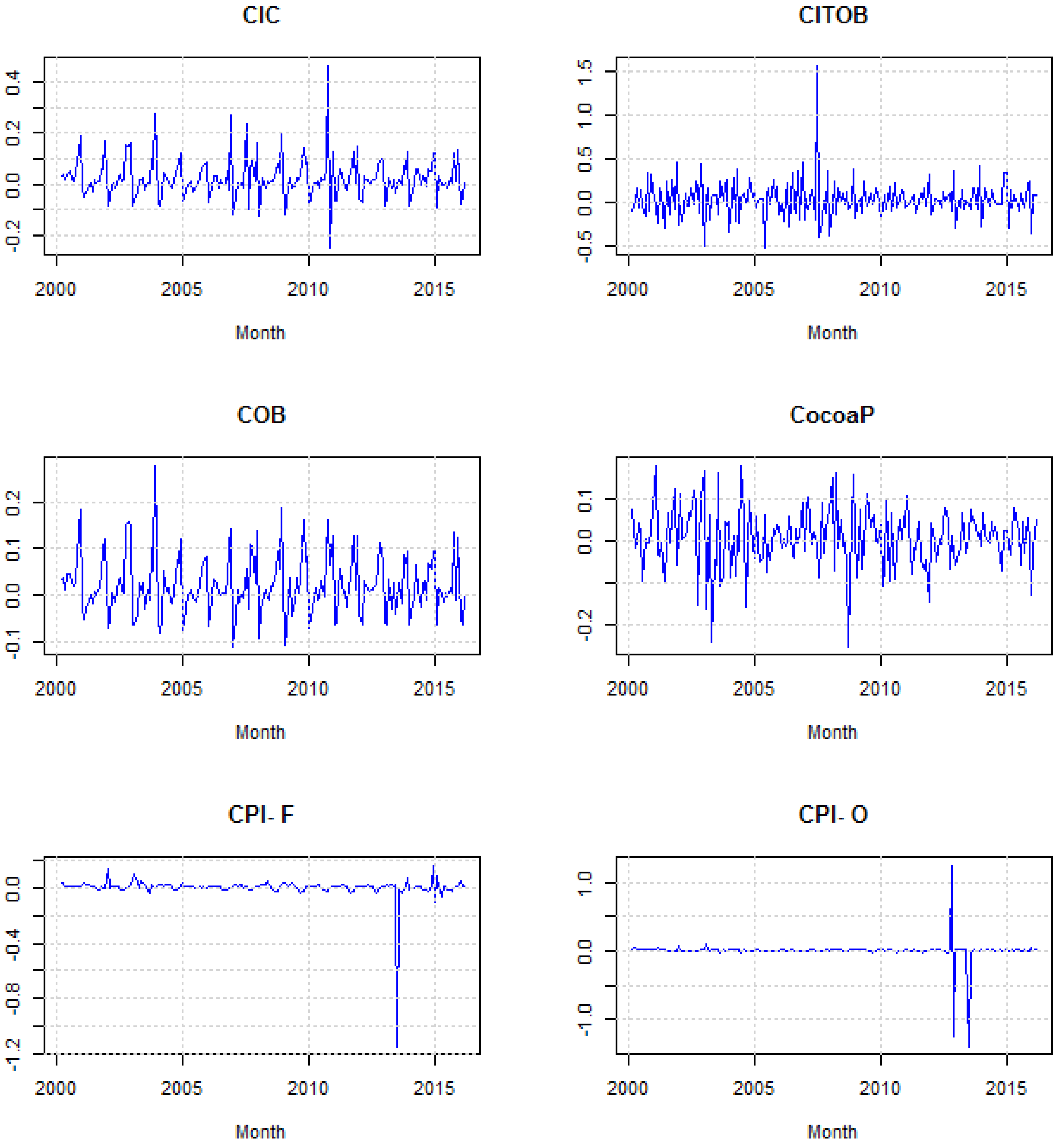

Figure 8: Time plot of the Transformed Variables

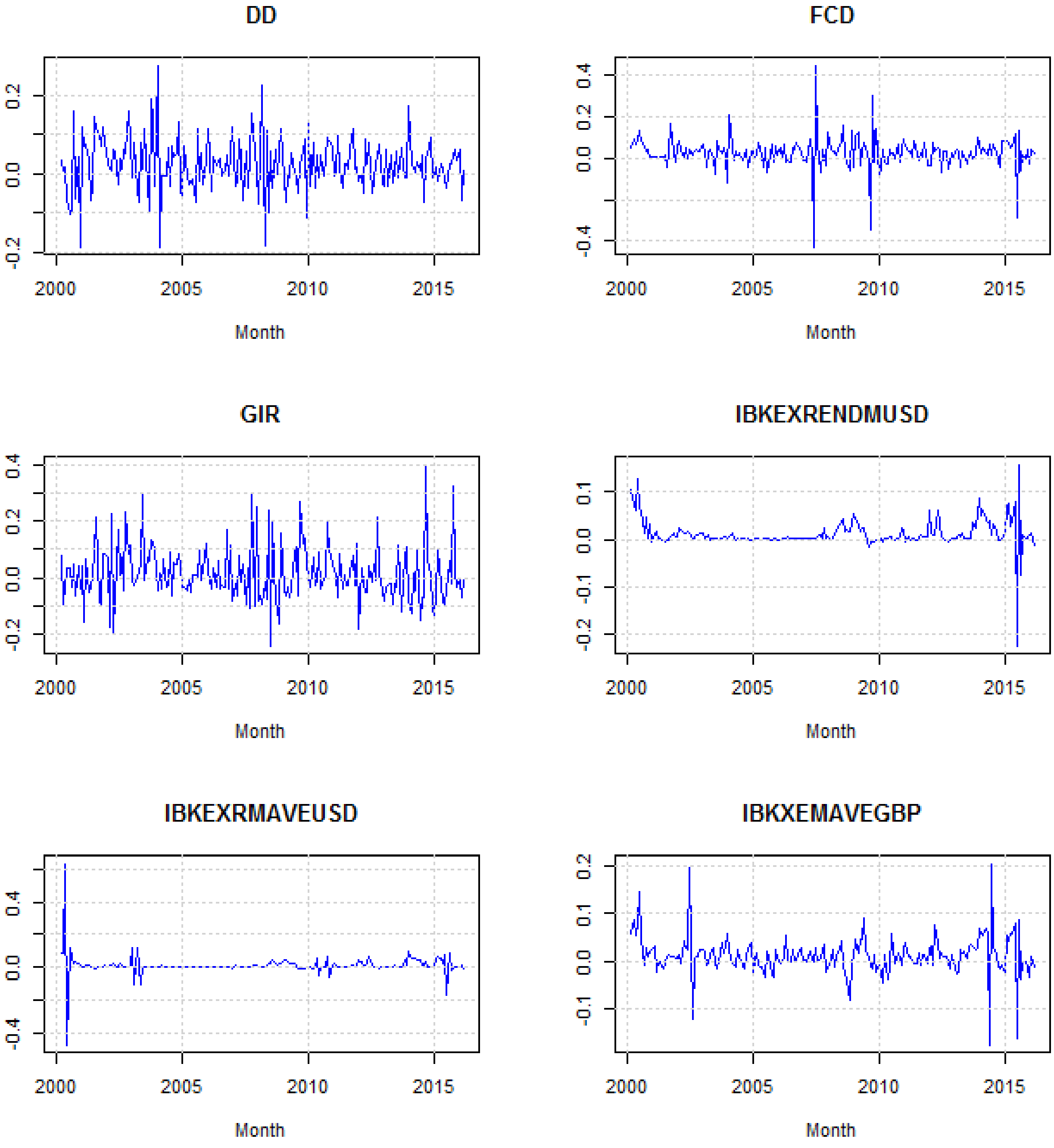

Figure 9: Time plot of the Transformed Variables

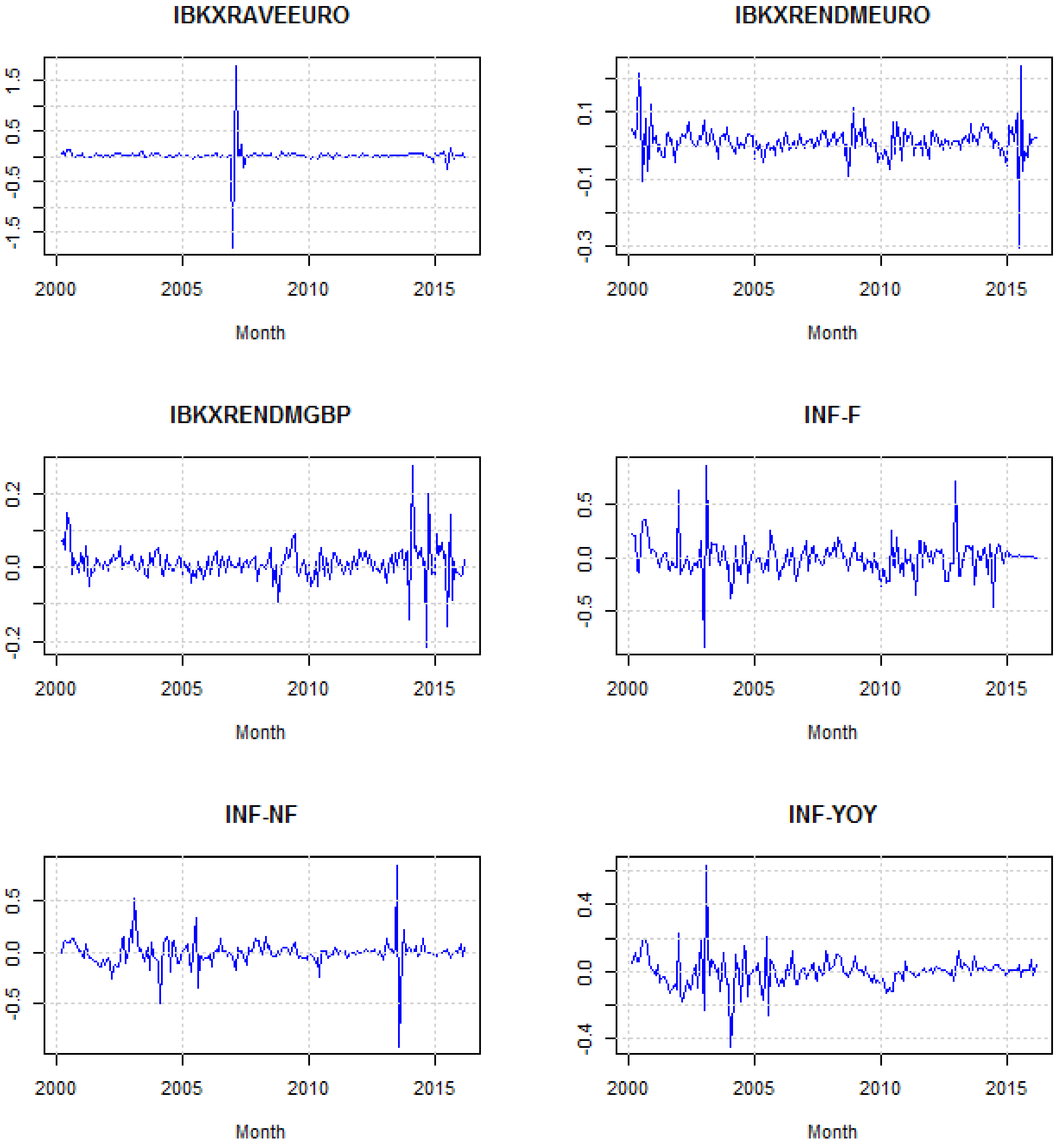

Figure 10: Time plot of the Transformed Variables

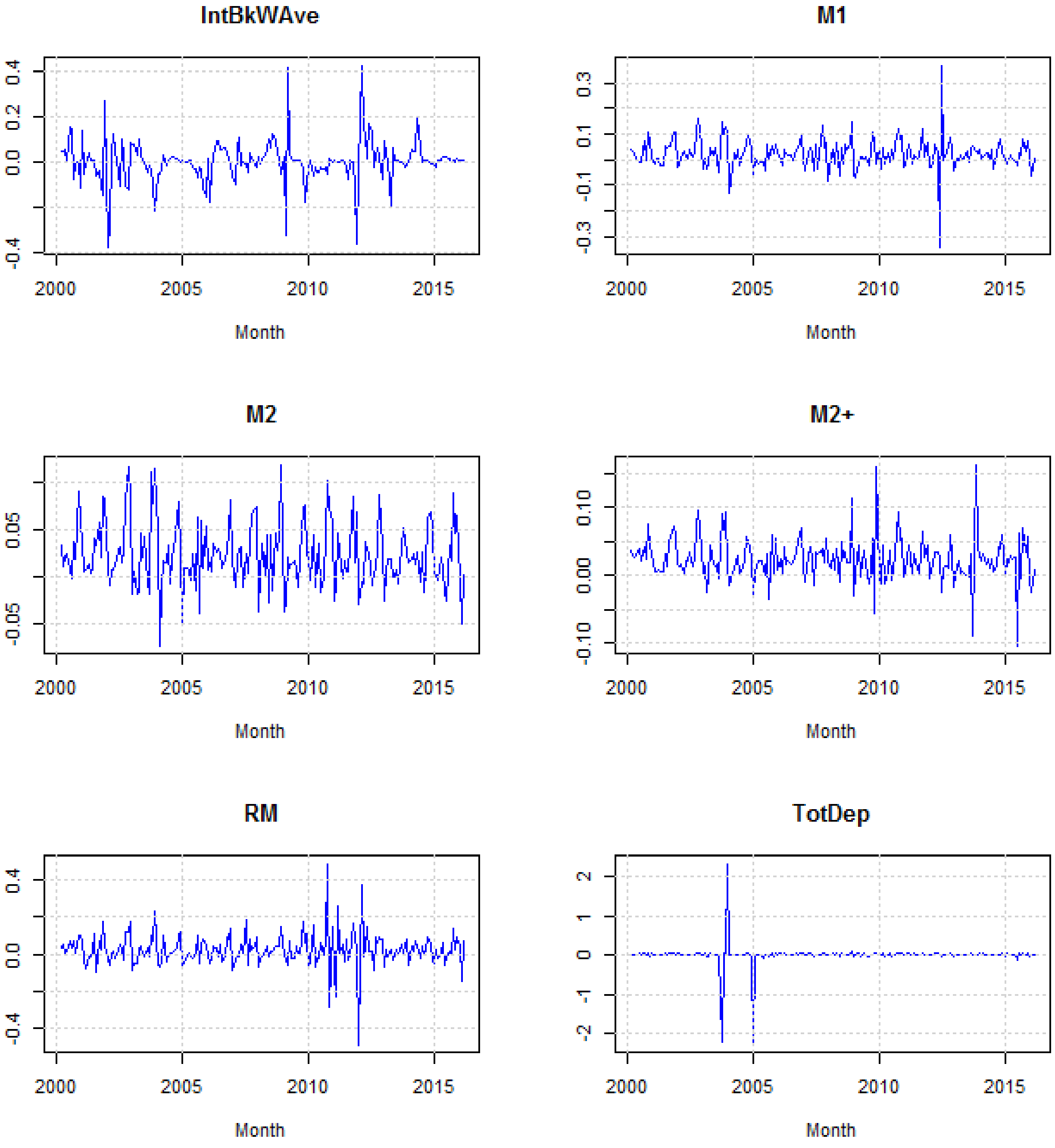

Figure 11: Time plot of the Transformed Variables

APPENDIX C

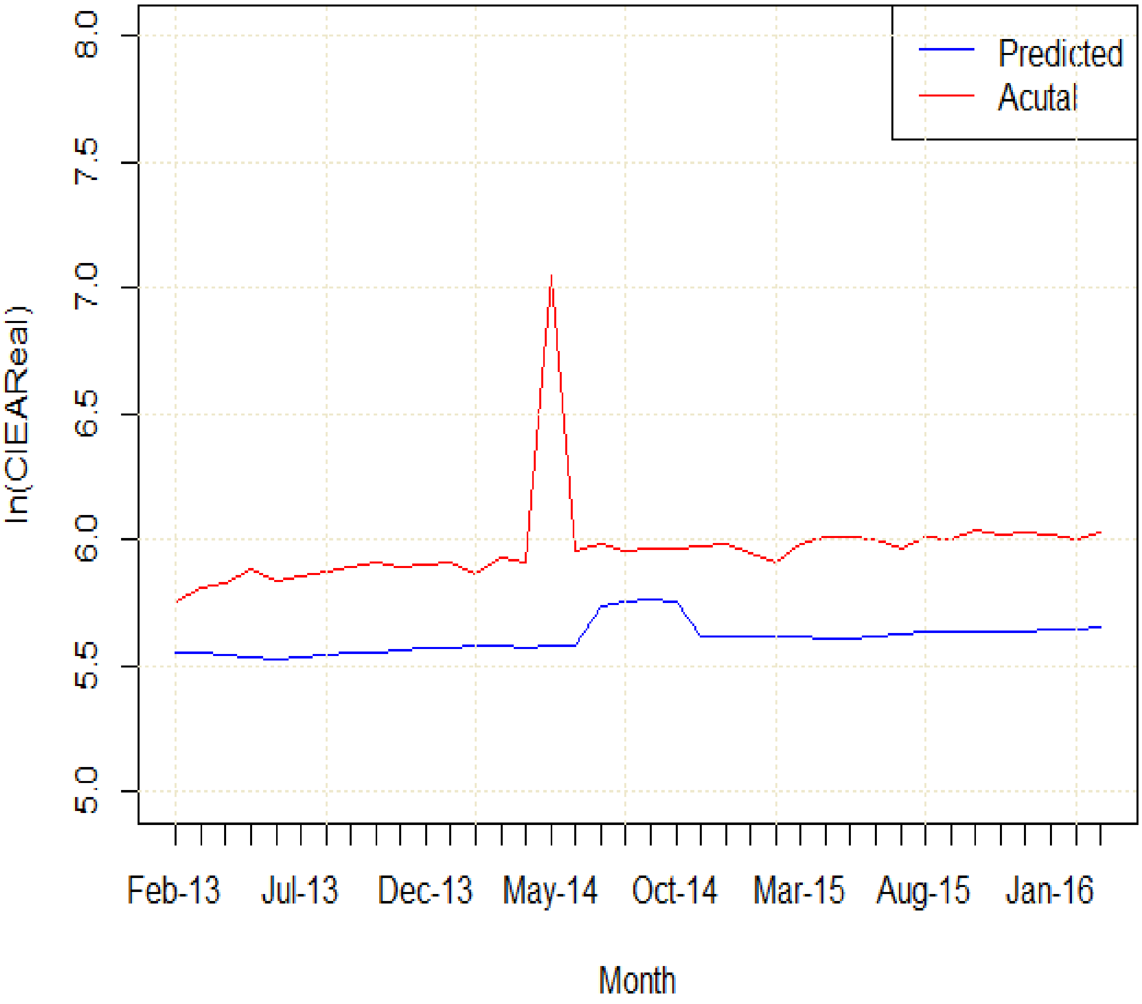

Figure 12: Out-of-sample comparison of predicted values from LASSO and $\ln(CIEAReal)$

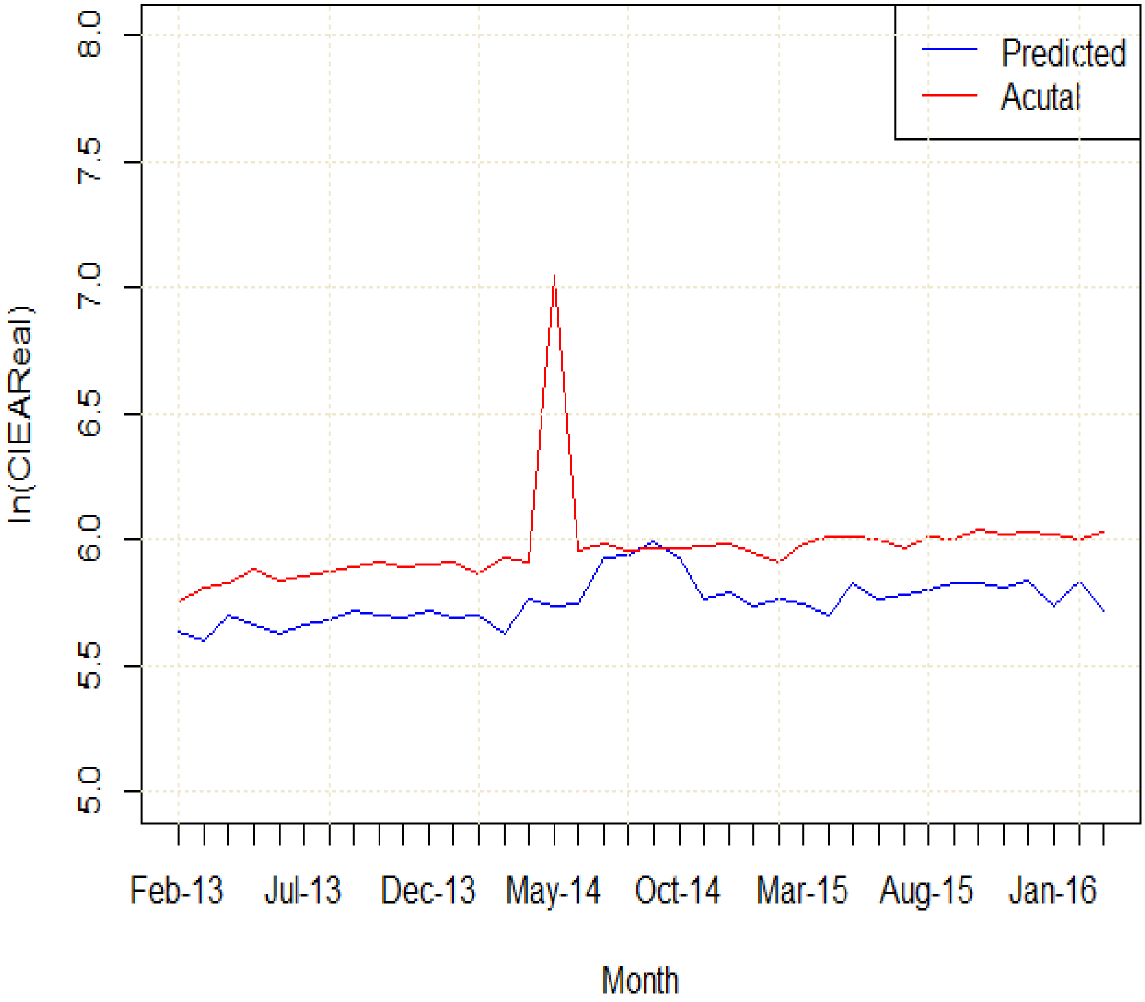

Figure 13: Out-of-sample comparison of predicted values from ENET and $\ln(CIEAReal)$

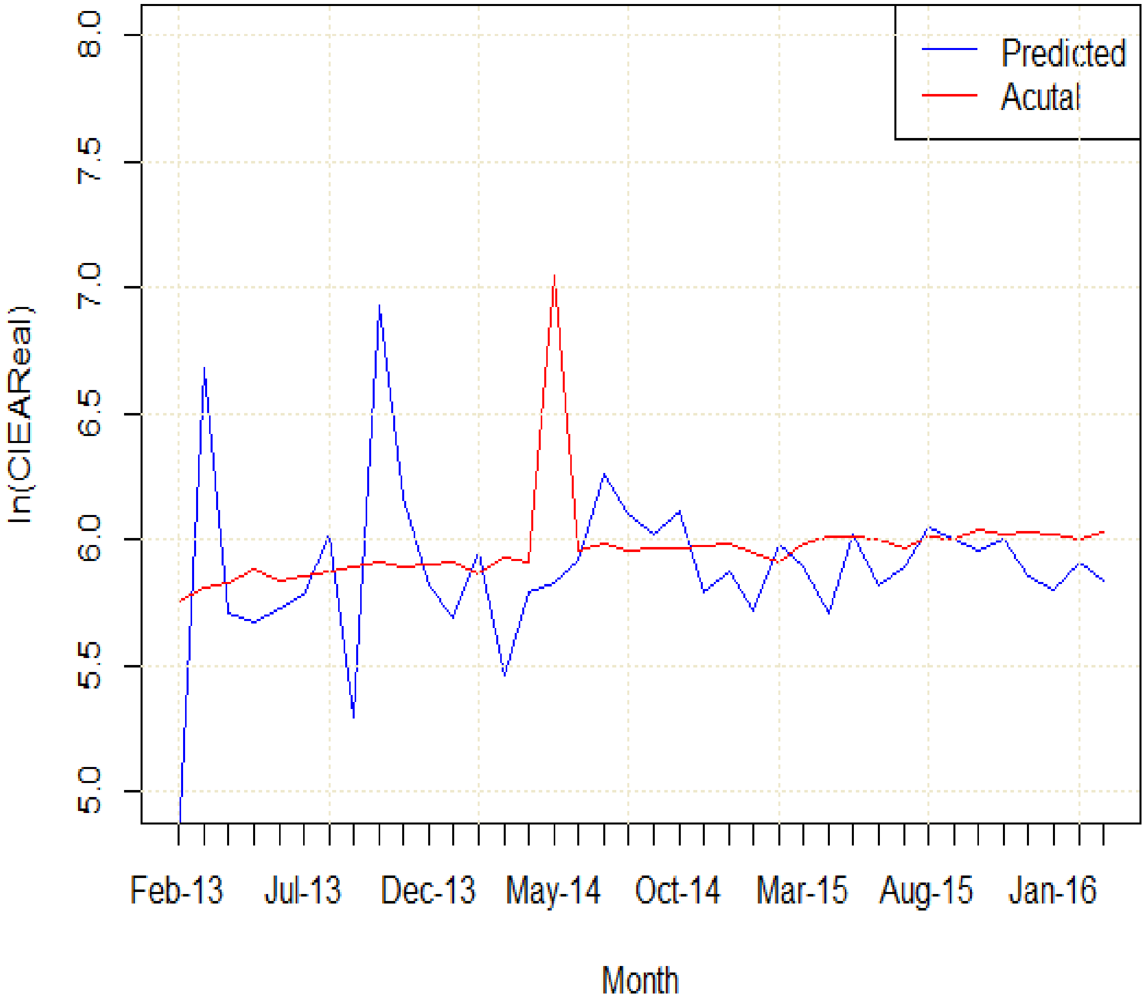

Figure 14: Out-of-sample comparison of predicted values from Adaptive LASSO and ln($CIEAReal$)

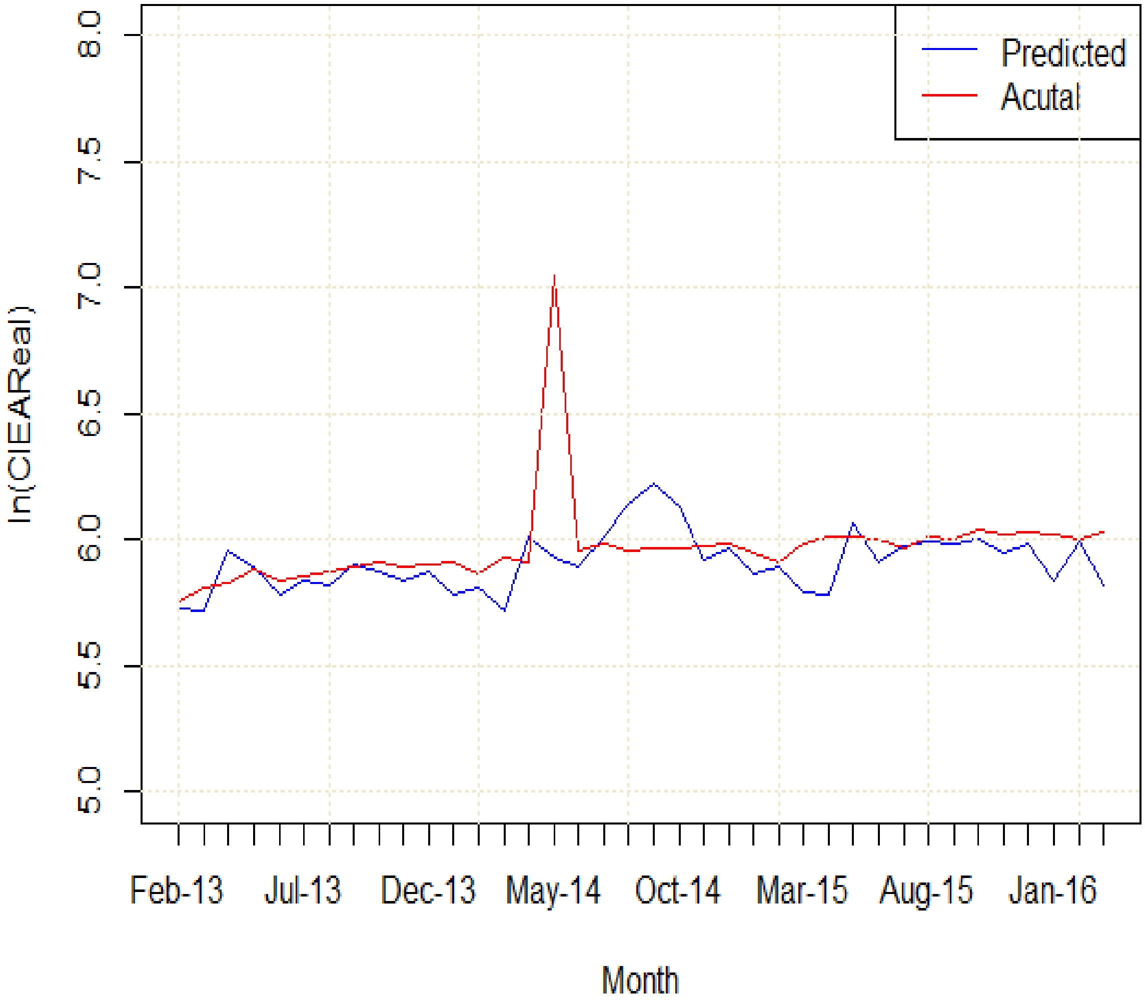

Figure 15: Out-of-sample comparison of predicted values from Adaptive ENET and $\ln(CIEAReal)$

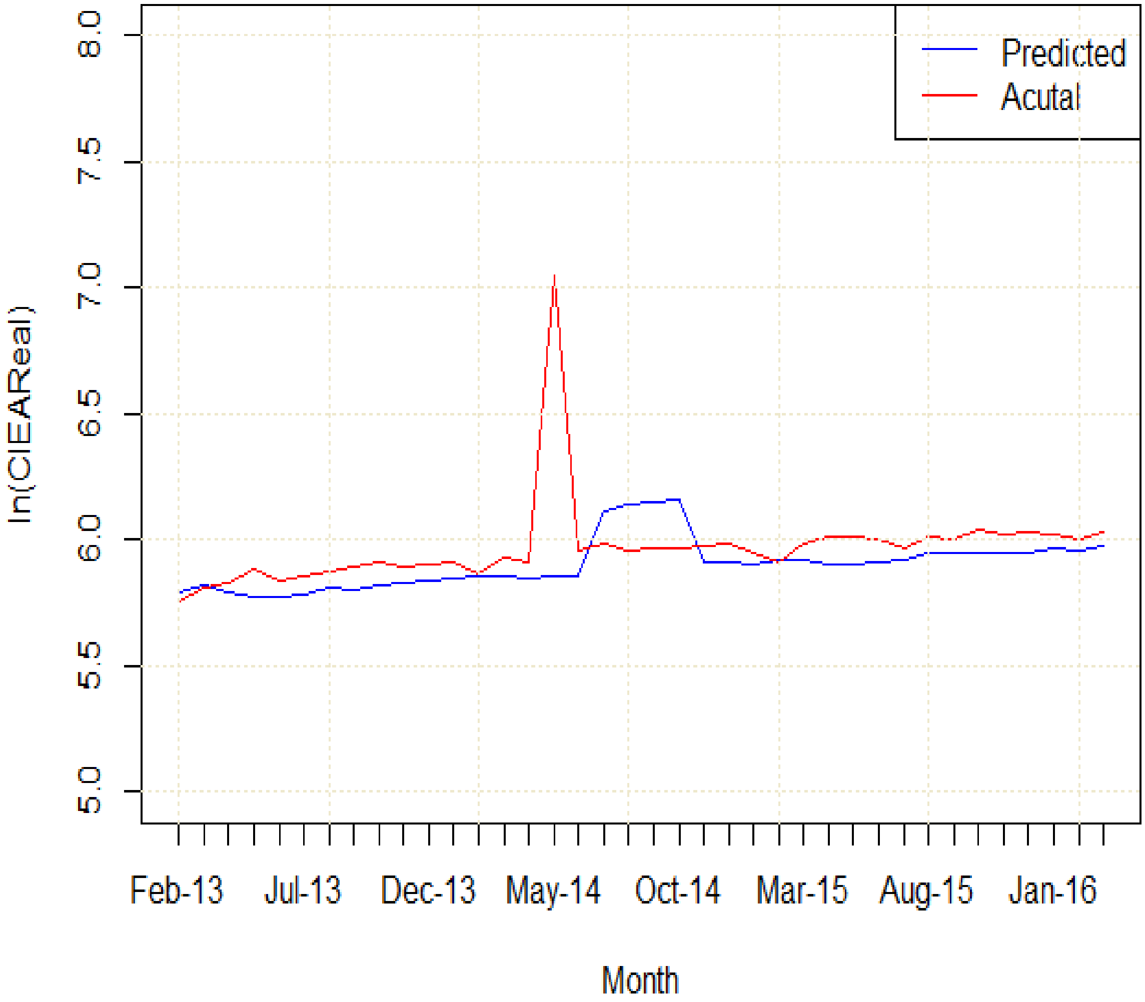

Figure 16: Out-of-sample comparison of predicted values from PCR and $\ln(CIEAReal)$