Discussion Papers in Economics

Coincident Indicators and Forecasting in Economics using Ensemble Empirical Mode De-composition (EEMD) Analysis: A Study of IIP

Manoj Pant and Arun K. Attri

Discussion Paper 17-04

Centre for International Trade and Development

School of International Studies

Jawaharlal Nehru University

India

*

Coincident Indicators and Forecasting in Economics using Ensemble Empirical Mode De-composition (EEMD) Analysis: A Study of IIP[1]

-Manoj Pant and Arun K. Attri*

May, 2017

*The authors are Professor of Economics, Centre for International Trade and Development, School of International Studies, JNU and Professor of Environmental Bio-Physics, School of Environment Studies, JNU

(*preliminary draft not be quoted without permission)*

ABSTRACT

Many phenomena in natural and social sciences requires analysis of time series data to draw inferences about their possible future behavior. In Economics, time series analysis is frequently applied in the context of expected future course of economic activity, for example, movements in GDP, predicting recessionary cycles and so on. Research organisations have built huge macro models which are very data intensive and involve a lot of assumptions on elasticities and aspects of the macro economy to forecast into the long run. However, for small business organisations what is more important is short run estimation and forecasting with limited data requirements. Currently, the latter task is performed using various statistical techniques which are largely linear in approach, are dependent on the choice of the start-end period and have low statistical reliability. This study uses the EEMD (Ensemble Empirical Mode De-composition) approach which is not constrained by these defects. As an illustration, the Indian IIP series is used to develop a simple coincident indicator of movements in IIP which is simple to use, uses real time data and gives accurate forecasts.

[1]We are indebted to Prof. K.L. Krishna whose comments on an earlier version made this one possible. We would also like to like to thank Sugandha Huria for research assistance. We also received helpful comments during the seminar at the Centre for International Trade and Development, School of International Studies, JNU

I. Introduction

The history of forecasting-models in Economics is old and goes back to the famous “Harvard ABC curves” which predate the 1st World War. Very broadly forecasting models can be divided into long term analytical macro models and short term indicator models. One of the best examples of the former is the so called National Institute Global Econometric Model (NIGEM) of the National Institute of Economic and Social Research (NIESR), U.K. As with other such models, a set of equations is used to define any economy in its various blocks like consumption, investment, government expenditure etc., while trade equations are used to link these economies to the global economy. Typically such dynamic general equilibrium models are used to forecast economic activity for individual economies and globally over next four to five decades.

The second class of short term “indicator models” are individual economy based with the more limited objective of identifying “leading indicators” of cyclical behavior of important macro-economic variables like GDP. In the “Harvard ABC”model, the idea was that the best “leading indicators” of economic cycles were stock prices, volume of bank cheques and interest rates, in that order, where the upturn (downturn) in the first is followed by the others all culminating in an upswing (downswing) of an economy’s growth cycle. Formal work on such models was continued in the NBER and the Economic Cycle Research institute (ECRI) by various researchers. Another set of studies were conducted at the OECD Economics department by authors like Sedillot and Pain (2003) and Anabella (2006) . For an excellent survey of such indicator models see Banerjee and Dua (2007). The broad idea is that cyclical movements of macro economic aggregates like GDP and employment are themselves a function of cyclical behavior of micro level variables like raw material and other inputs, production of items like steel, manufactures etc. The cyclical movement in these micro variables may reinforce or oppose each other at different points of time. It is only when these movements in each is in the same direction that a persistent cyclical behavior in aggregates like GDP or employment would be observed. Hence, the best predictor of a cyclical economic behavior is one or more of these variables which then become leading indicators. These indicators can be divided into “coincident” indicators

(whose behavior mimics aggregate economic activity) or “leading” indicators (whose movements precede movements of macro- economic aggregates).

However, as also noted in Banerji and Dua(op.cit.), the data requirements of even such short term indicator models might be too daunting for many poorer emerging economies, where data collection and dissemination may be poor or time consuming, and where the results of such models may only be available to government agencies or available with considerable time lag, so that reliable short-term economic decision making is not possible.

In this paper we concentrate on developing a coincident indicator for India using a technique that is statistically sound and requires limited data inputs compared to established techniques or models. While this paper looks at developing a coincident indicator, a later work will focus on how the technique can be used to generate a leading indicator of cyclical economic behavior. This paper also focuses on a macro aggregate usually used for the analysis of the growth of the Indian economy: the Index of Industrial Production (IIP). However, the methodology can be used to extract reliable trends and their forecast for any time series data.

The next section looks at some statistical techniques frequently applied to estimate growth rates for Indian economy. This section also introduces a new technique of estimation and forecasting using the Ensemble Empirical Mode Decomposition (EEMD) algorithm. In Section III we discuss the IIP and its various limitations and suggest some proxies for IIP. Section IV then, compares the various techniques, using IIP time-series data,to show how the EEMD seems statistically better for estimating the growth rates. The EEMD methodology is then used in Section V to develop a coincident indicator for the IIP, which is then used to forecast growth. Finally, we conclude the paper in Section VI.

II. **Estimating Growth Rates.**

An issue often encountered in Economics is to analyze a given time series data on some proxy variable, considered appropriate, to yield growth rates. Obtained growth rates are then used to forecast into the near future with usual forecast errors. The techniques used depends on the appropriate data available. If this is large, an explanatory model for the variable in question can be set up. The model can then be subjected to analysis using various techniques like Multiple Linear Regression (MLR) (panel regression methods if the data available is extensive) to obtain the “best fit” of the variable and the parameters involved in the model. Alternatively, if the data-size availability is limited, the time series in question can be "fitted" using variants of the Box-Jenkin (BJ) techniques, for example, Auto-regressive in Moving Averages (ARIMA) or non-linear estimation (NLE). The BJ or ARIMA models are essential to handle an inherent property of most time series: non-stationarity. However, all these methodologies assume linearity in the estimated variable. Even the NLE method uses a linear approximation to derive the estimated model parameters from data and has some similarities to the MLR techniques. A third possibility, and a common practice, is to fit an exponential function to obtain growth rates using a simple regression estimation and then estimate forecasts. In this small note we will introduce a non-linear method of estimation which does not pre-specify the form of the estimated trend in time series (linear or non-linear), has limited data requirements, is not affected by non-stationarity and non-linearity of time series data or the choice of the time period and has lower estimation errors compared to the alternatives specified above. As an illustration, this note will use the time series on Index of industrial Production (IIP) but the same methodology can be applied to other data. The IIP is chosen as it has a wide variety of users: from political parties, to business men to academic establishments. Given the composite nature of IIP, which is a summation of multiple economic activity proxy predictors, we further explore the possible best single proxy indicator variable from which the future IIP trend can be forecast.

III. The IIP

As economic growth assumes greater importance, indicators of this have become extremely important not only for the government, but also for industry and private demand forecasts. In addition, as growth in the developed world grinds to a halt, there is an increasing attention given to growth potential in both India and China. Here, as Chinese growth seems to be leveling off in recent years, there is even more focus on the growth potential in India.

In India, there are well known infirmities in the data bases. For one, insufficient computerization of regional data bases implies that the time lag in data's availability is too large. The best estimates of GDP growth, for example, are only available at 3-monthly intervals and that too with a time lag of a couple of months. This is normally insufficient for most industry/financial planning purposes. A quicker estimate of growth in economic activity is the Index of industrial Production (IIP) available on a monthly basis but even here the time lag is about two months.

The IIP index also suffers from a number of limitations. First, the IIP data is based on production in three final consumption sectors– mining, electricity and manufacturing and three intermediate production sectors, basic goods, capital goods intermediate goods. However, it only represents the organized sector firms and ignores the service sectors altogether. As services now account for almost 50 percent of GDP, it is debatable that the movements in IIP trend give a true picture of changes in the level of economic activity unless one assumes a perfect correlation between economic activity in the manufacturing and service sector. In addition, one has to assume a one to one correspondence between the organized and unorganized sectors. India's experience since independence shows that neither of these assumptions may be valid (see, for example, Kundu, 1993; Harris and Sinha (2007); Chakrabarti and Kundu (2009); Kalyani (2016))

Second, the methodology for computing the IIP does not allow for a change in product mix till the index is redefined at periodic intervals when the list of companies/products is changed to reflect the new economic reality. This periodicity is also not consistent as can be seen in Table 1 below.

Table 1: IIP Index: Years of revision and coverage

Base year	**Number of Items Covered**
1946	35
1951	88
1956	201
1960	312
1970	352
1980	352
1993-94	543
2004-05	682

As is clear from Table 1, even a five year periodicity has not been consistently followed. This makes comparison over time difficult.

Third, since companies produce different products, it is often necessary to convert physical outputs into values (for example, for estimating manufacturing output) where there is the added problem of the price series used in conversion.

In other words, the IIP as an index of economic activity seems at best only a limited indicator for very short periods (no structural change). Even more important, it is an extremely insufficient (and possibly misleading) indicator of economic activity over longer periods. This is particularly important for those trying to forecast future movements in economic activity. What is probably most important is the time lag in the availability of the IIP series. However, it is the best indicator currently available. The issue then is how best to obtain a true description of the underlying IIP series? Second, to avoid all the problems of ignored sectors and valuation of outputs, why not look at the country wide data on inputs which go into the production function of the economy? In that case, do the IIP figures actually reflect accurately the level of economic activity? Finally, can one do an accurate forecast with this data? These are some of the issues we will take up in the next section.

III.1. Forecasting Economic Activity using IIP.

However imperfect a measure, it is necessary to see if IIP does have any relation to the level of activity in the economy. It is clear that economic activity can be measured directly by looking at changes in the output or indirectly in terms of the inputs that go into the production process. To that extent, changes in the production/consumption of critical inputs like cement, steel, transportation, electricity etc. also indicates the level of economic activity. This input approach has the advantage that one does not have to reduce the inputs to a common value unit. Second, data on such inputs is normally available on a real time basis. Hence, an indicator based on these inputs is better suited to real time forecasting. Third, the input approach does not require disaggregation by a sector of use and avoids the criticism that it is based only on manufacturing sectors and a sample of firms. The input approach, however, does face the criticism that it does not reflect productivity changes-even with constant inputs, output can increase if productivity increases. Thus, lack of demand for inputs does not imply that economic production is stagnant. Hence, an input approach estimate of activity is a lower bound for the actual level of activity in an economy.

The other issue for users of indicators of economic activity is some forecasting of trends in the near future. How well the IIP series lends itself to forecasting depends itself on the method of “ line fitting” used to approximate the series-trend. Here a number of approaches are possible. The simplest approach is to fit a regression line to the given data. The fitted equation then become the basis of simple forecasts. Further, depending on the explanatory hypothesis, the set of statistically significant regression coefficients can be the basis of defining an “instrumental variable” which can be used for forecasting and thus avoids the problem of the infirmities in IIP data already indicated earlier. This approach, however, has two problems. For one, a basic assumption here is that the movements of the IIP are assumed to be linear. Second, since the matching data on all inputs is necessary the data requirements are considerable. Third, linear methods in particular do not account for the presence of multi-cyclical influences over time in most time series data. Given leads and lags in links between inputs and outputs, it can be argued that the point to point data do not indicate any long term trend after multi-cyclical behavior has been netted out.

An alternative, less data intensive, approach is to use time series techniques of estimating a best fit line for IIP. An example is the ARIMA approach which also lends itself to forecasting. However, here it is still assumed that the “de-seasonalised and adjusted” trend in IIP is linear.

The third possibility is to use non-linear estimation for fitting a trend line to the IIP data. The usual approach is to assume that an exponential trend fits the IIP series best. Typically, a semi-log estimation process is used to generate compound growth rates of IIP. This again is amenable to simple forecasts. However, this methodology does not take any account of the presence of multi-cyclical behavior in the data.

A final approach is to allow the form of the inherent non-linearity in the IIP time series data to emerge endogenously, without any *a priori* assumption, using the Ensemble Empirical Mode Decomposition (EEMD) approach[2]. Here the same method of estimating the non-linear function can be used for the corresponding input series to see which one best mirrors the behavior of the estimated IIP series. This input (one or more) can then be the best proxy for estimating economic activity. The EEMD approach has two advantages over the above three methodologies: it is not constrained by the presence of non-linearity and non-

[2] For a detailed description of the EEMD approach see Annexure.

stationarity in a time series data and accounts for all cyclical behavior (see, Huang et.al.,1998; Wu et.al., 2011).

The EEMD considers that a time series data in question is composed of both cyclic and acyclic components. The method enables the extraction of Internal Model Functions (IMFs), representing multi-cycles, each having distinct periodicity, embedded in a time-series. The sum of all IMFs, representing each cycle, would represent total cyclical activity in the series data, which is also evident from the observed time dependent oscillations in the data. Extraction of different IMFs is initiated by determining the maxima and minima points present in the series. The points representing the various maxima and minima are then fitted using a cubic spline. The fitted spline functions would designate respective maxima and minima envelopes. The average values of the these two envelopes is subtracted from the original data to gives us the first IMF, say, IMF(1). The same procedure is then applied to the original data minus IMF(1).The extracted IMFs are in ordered sequence of high to low frequency (period). Significance of each extracted IMF is ascertained on the basis of a statistical test algorithm. The time series data, subsequent to the extraction of all embedded IMFs, represent the trend where no cyclical variations are observed. The obtained trend may display linear or non-linear attributes. The trend estimated on a given time series, in a given time span, is unaffected by future addition of new data in the time series. Determination of all IMFs and trend using EEMD (Empirical Ensemble Mode De-composition) represent their respective ensemble averages, which are obtained after the addition of multiple different proportion of random white noise of the time series variance of the time series data under analysis. Addition of known proportion of white noise does not affect the embedded signal (IMF), at the same time it allows the determination of each IMF and trend as ensemble averages from different noise added sets of the original time-series data. Consequently, the uncertainty region at 95% confidence interval can be determined for respective IMFs and trend.

In this study we will use all these methodologies to arrive at the best indicator for the IIP and forecast of future economic activity. The methods used are compared in out of sample forecasts.

III.2. Data base.

As already noted, the IIP is a composite indicator of production of firms belonging to the Mining, Electricity and Manufacturing sectors with maximum weightage given to the latter. For the index, production values of a sample of ASI firms in these sectors is calculated by allotting different weights to each sector and then indexing the final production series to a base, which varies when the set of firms/products changes. The monthly data for IIP values was available from November, 1988 to February,2017 from the Ministry of Statistics and Programme Implementation (MOSPI) data base.

By definition, the chosen firms all belong to the set of organized industries. We also collected separately monthly data on major inputs in the economy. These inputs were Rail Freight (in Tonne Kilometers), Cement (Million Tonnes), Finished Steel ('000 Tonnes), Coal ('000 Tonnes), Thermal Power (Million KwH), Road Trucks (Number of vehicles) and Production of Truck and Bus Tyres (numbers). But due to too many gaps in the data for some variables, we dropped Road Truck production and Production of Truck and Bus tyres from the final set of inputs . Further, uninterrupted data for a matching time span for IIP and all the inputs was only available for April, 1999 to March, 2012 (a total of 156 observations). These time series data-sets were used in all our estimation after reducing all numbers to the common base 1993-94 using the CSO's conversion formula. We have chosen 2012 as one watershed point as, for GDP (and IIP) calculations, the basis of calculations has changed after 2012. Out of sample forecasts were made for five periods beyond March 2012.

The data for Rail Freight and Coal Production is taken from CSO, whereas Department of Industrial Policy and Promotion furnishes the data for Cement consumption. Finished steel production data was taken from the Ministry of Commerce and Thermal power generation from Ministry of Power.

IV. The Estimation.

As the introduction indicated, we thus have four methods of estimation: the full MLR model, single series ARIMA method, fitting an exponential function and EEMD analysis.

IV.1 Multiple Linear Regression (MLR)

MLR was used to estimate the parameters of the multi-variable regression model followed by the prediction of future values. We assume there is an economy wide production function where the indicator of production (IIP) is a function of inputs used. The generalized form of regression equation is given by –

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u_i$$

Where

Y_i= Index of Industrial Production (IIP), X_1= Freight Rail (in tonnes), X_2= Cement Consumption (in tonnes), X_3= Finished Steel Production (in tonnes), X_4= Coal Production (in tonnes), X_5= = Power (Thermal, in million units) and u_j the usual error term.

The linearity implies that there is no substitutability between inputs. The results of the MLR estimation are given in Table 1.

Table 1. MLR estimation of IIP,

VARIABLES	**IIP**
Freight Rail	.554**
Cement Consumption	-1.26**
Finished Steel Production	0.00012**
Coal Production	0.026**
Power (Thermal)	.039**
Constant	-9.79**
Observations	156
Adj R-squared	0.983

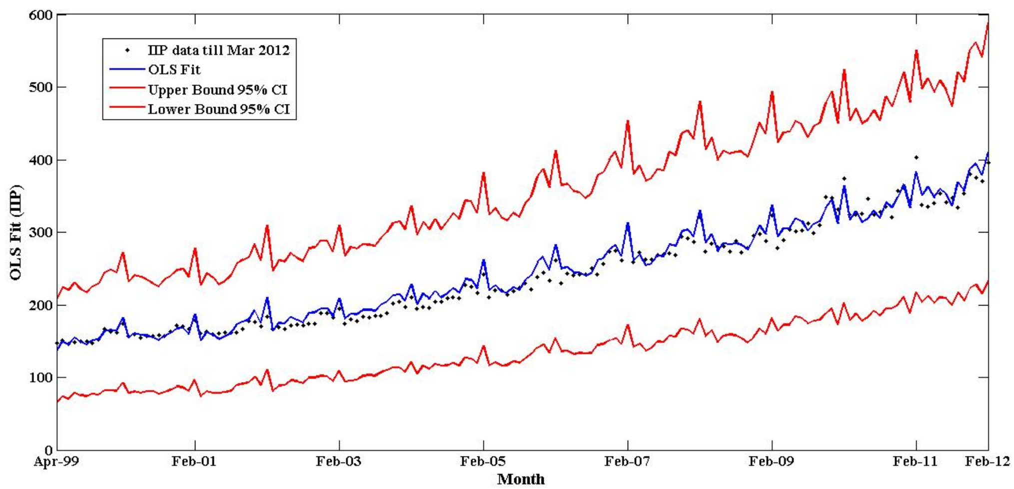
** significant at 5 percent level

The correlation matrix indicated a high degree of multi-collinearity between all the variables so that the individual coefficients are not true estimates. However, since our purpose is to obtain forecasts of IIP, the multicollinearity problem can be ignored. What the estimation shows is that a linear combination of the explanatory variables explains about 98 percent of the total variation in IIP. In other words, use of IIP as an indicator of economic activity is not entirely unjustified despite all the infirmities in the data base. However, the non-stationarity of all the variables raises questions on the statistical validity of this approach.

It is useful to look at the graph of the fitted function (Graph 1).

Graph 1.

MLR Estimation of IIP as function of various inputs, 1999-2012



The coloured band around the MLR fitted line indicates the 95 percent confidence interval for the fitted values of IIP. On the vertical axis we plot the values of IIP based on the estimated coefficients of Table 1. The blue line is the MLR fit while the red lines indicate the 95 percent confidence interval (CI). From Graph 1 it is clear that the movements in estimated IIP mirror quite closely the actual data.

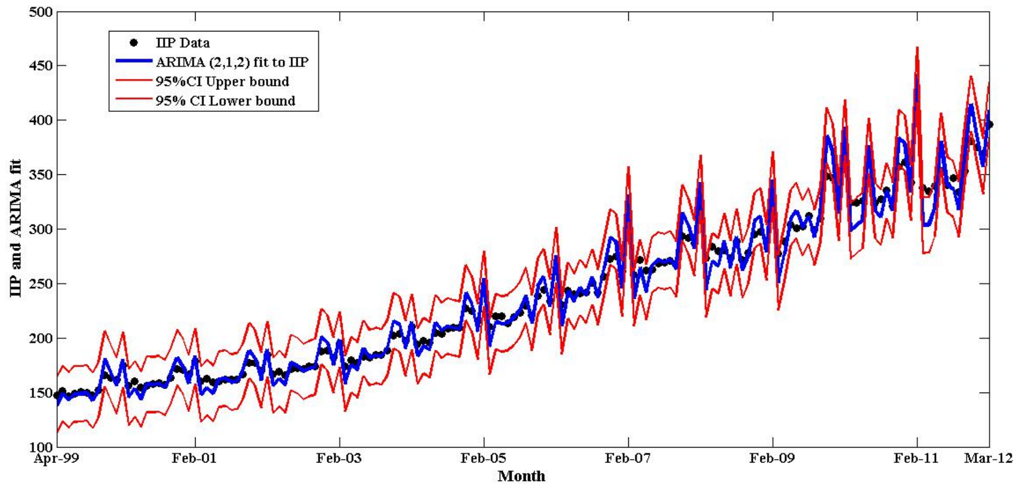
The compound growth rate (CAGR) for the sample period April 1999 to March 2102 based on the actual values of IIP is 0.629 percent per month as compared to 0.627 with the estimated values (see, Table 2 below)[3]. While this may seem close, the wide CI band indicates that these estimates are not very reliable. In graph 1, the CI bands are actually diverging over time. In addition, looking at the blue line, the high variability in the estimated values indicates that forecasting values beyond the sample period would be extremely hazardous. The forecast error is minimized– but it is larger than the errors in residuals– only when the values of the X's in the forecast period equal the means of the sample X's (see, for example, Pindyck and Rubinfeld, 1991, Ch. 8). This is unlikely in a time series models especially when cyclical influences are not accounted for. Finally, since the IIP time series is clearly non-stationary, MLR estimation will not yield unbiased estimates and this would further add to the forecast errors. Hence, we have chosen not to report the forecast errors in the case of MLR.

IV.2. ARIMA Methodology.

We also conducted an ARIMA fit of the IIP series. This seems warranted since the IIP series is clearly non-stationary. Usual tests of non-stationarity confirmed this. The graph of the fitted IIP data and a 95 percent confidence band is shown in Graph 2 below.

Graph 2. ARIMA fit of IIP data, 1999-2012

[3] All growth rates are calculated by fitting a semi-log function log Y= log a +bt, where Y represents the actual/estimated values of IIP and b is the estimated CAGR.



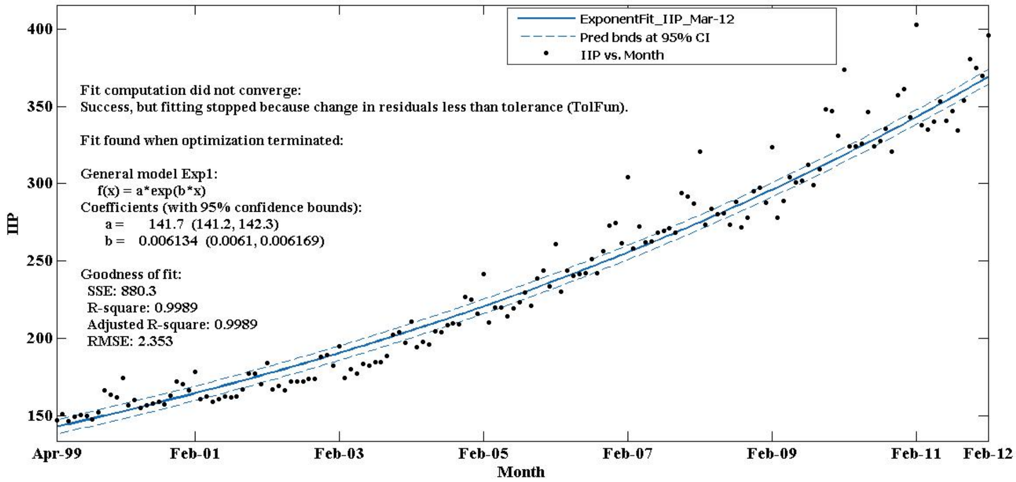
Adjusted R^2 = .968

The CAGR of IIP based on the fitted values is 0.636 percent per month which is higher than the value obtained from the MLR procedure .Inspection of Graphs 1 and 2 indicates that both fits are very similar but the ARIMA estimated values show less cyclical fluctuations and the estimated values are statistically more reliable as the CI bands are much smaller than in the case of the MLR estimation. In addition, the CI bands include all the actual values of IIP indicating that the ARIMA process explains the generation of IIP values quite well.

IV.3. Non-linear Regression, Exponential Model

As we have noted, it is often assumed the IIP series follow an exponential path of the form $Y_t = a.\ e^{bt.}$ and b is the growth rate over time. A semi-log linear regression equation fit gives the fitted line as shown in Graph 3.

Graph 3.Exponential fit of IIP data, 1999-2012



Adj R^2 = .9989

Inspection of Graph 3 indicates that the exponential function (as in the case of ARIMA estimation) fits the data better than MLR estimation since the CI bands are clearly smaller. The CI band is smaller than the ARIMA model (see Table 2 below). However, the exponential fit does not track the actual IIP values well as most of these values seem to lie outside the CI levels especially in later years. This is probably because cyclical variations in data are not accounted for. The CAGR per month for fitted values is about .613 per month. Inspection of the three graphs does indicate that the ARIMA estimates are statistically the best but this does not account for the possible non-linearity in the data.

One of the objectives of our estimation is to compute the out of sample forecasts for IIP. We have considered the five periods from April 2012 to August 2012. The out of sample forecasts for the MLR estimation is obtained by running the regression for the whole period including the forecast period. This has already been noted in the section on data collection. Our results are summarised in Table 2 below.

Table 2 : Five period forecasts based on MLR, ARIMA and Exponential Models

7.9	**IIP values**	**Forecast Values#**		
		MLR*	**ARIMA(dynamic)**	**Exponential**
Apr-12	346.4	353.4	379 (+- 25.3)	379.9(+-21)

May-12	359.5	370.5	385.6 (+-27.1)	380.0 (+-21)
Jun-12	354.6	357.5	393.9 (+31.9)	382.4(+-21)
July-12	352.7	360.8	386.3 (+-37.9)	384.8 (+-21)
Aug-12	347.7	357.9	390.9 (+-40.1)	387.3 (+-21)
CAGR (% p.m.), sample period	0.630	0.630	0.636	0.613

*The MLR forecasts were obtained since in this case all the values of the X's were available in the out of sample period except for Cement. For cement we fitted a linear function over time and forecast the values in the forecast period. These were used in the regression equation to get out of sample forecasts for IIP. This makes estimation of errors in forecasts even more difficult. Hence, they are not reported here.
the figures in parenthesis indicate the 95 percent upper and lower CIs.

Inspection of table 2 indicates that the estimated CAGRs seem to be about the same with all the methodologies and fairly close to the CAGR of the raw data on IIP. However, while the MLR out of sample forecasts are quite close to the actual values of IIP they are statistically not very reliable especially as the IIP series is non-stationary. The broad trends observed from the Graphs is confirmed from Table 2: the ARIMA model seems to give the most reliable estimates yet it assumes the underlying movements in the time series is linear.

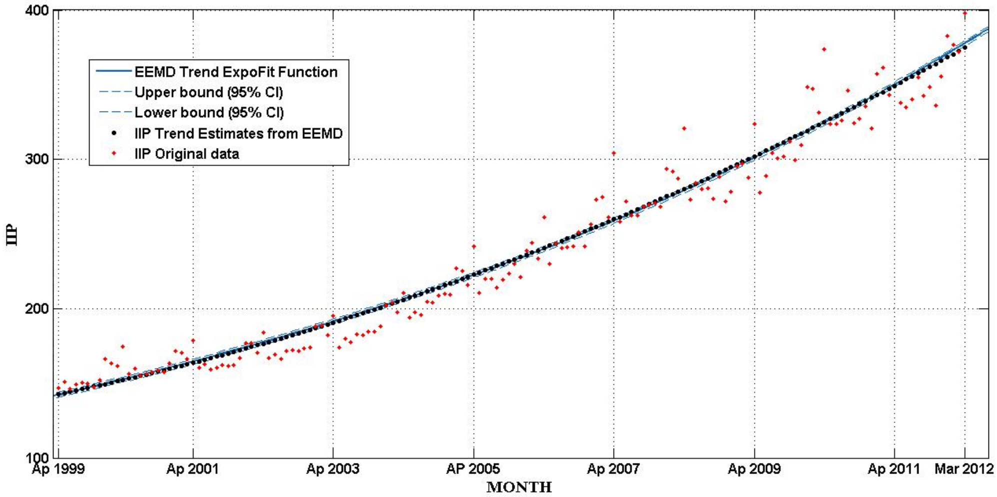
Thus, the problem is to balance the need to incorporate problems of non-stationarity, statistical reliability, non-linearity and cyclical variations in the underlying series. This is taken care off in the EEMD estimation of the next section.

IV.4.Non-linear Estimation, EEMD fit[4]

We have already noted that the EEMD estimation technique removes all cyclic components (IMFs) embedded in the data, while also endogenously determining the form of the estimated trend line. This is shown below in Graph 4.

Graph 4. EEMD fit of IIP data, 1999-2012

[4]The EEMD analysis was conducted using the MATLAB software and an algorithm developed by the authors.



Adj. R^2 = .9998.

The estimated growth rate of 0.630 percent per month from EEMD trend for the sample period is close to the growth rates obtained from other three methods as shown in Table 2 and very close to the actual growth rate of IIP . However, inspection of Graph 4 indicates two things. One, that the EEMD method also indicates that the best endogenously estimated trend function is non-linear. Second, the 95 percent confidence interval (CI) indicates that the EEMD technique yields the " best fit" of all four estimation methods on a probabilistic basis. Finally, the fitted line also tracks the actual IIP data fairly well.

We next obtained forecast values with the EEMD method. The five period forecasts are given in Table 3 below.

Table 3. Forecast Using the EEMD technique

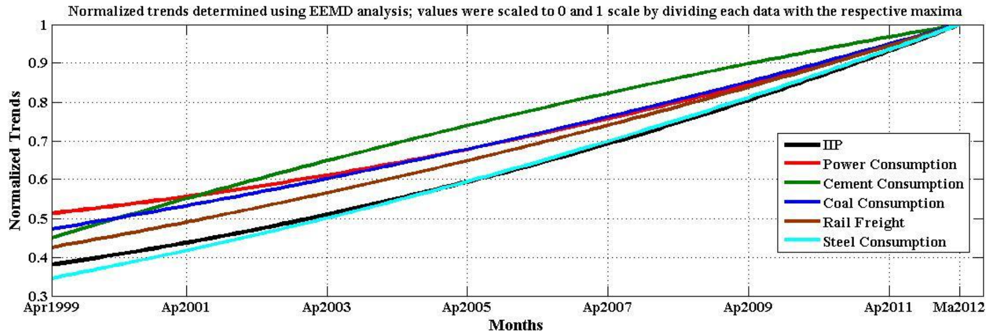
Month	Extrapolated Values of IIP
April-12	379.9 (-1.3 +1.3)
May-12	382.3 (-1.3 +1.3)
June-12	384.7 (-1.3 +1.3)
Jul-12	387.7 (-0.8 +1.9)
August-12	389.6 (-1.3 +3.4)

A comparison of Tables 2 and 3 clearly indicates that the EEMD forecasts are close to the ARIMA fit but the reliability is substantially larger in that the 95 percent CI levels from trend forecasts lie within 2 points while the CI limits for exponential fitting forecasts have a variability of around 21 points and over 25 points for the ARIMA forecasts (see Table 2). In addition, as is well known, in ARIMA estimation the errors in forecasts increases rapidly for long term forecasts so that the statistical reliability is limited. This is seen in Table 2 where, even in a 5-period space, the forecast errors increase dramatically. Hence, the EEMD methodology seems to overcome the shortcomings of the MLR, ARIMA and Exponential fitting methodologies noted earlier.

V. What is a good Coincident Indicator of Economic Activity?

We have already noted in Section 2 that the IIP has serious infirmities in terms of coverage and measurement and time lag in availability. However, we also noted that one way of getting round these issues is to assume that output growth can be well represented by input growth. In Section III.1. we saw that the MLR model indicates that variations in inputs like freight, cement, steel, coal and power explains nearly all the variations in IIP so that these can be used as good instruments for predicting movements in the IIP. However, this still creates the issue of aggregation and measurement to reduce all the physical values for inputs into one monetary value. An alternative possibility is to find which of these inputs best tracks the movements in IIP and could be an efficient coincident indicator of economic activity. We do this by applying the EEMD technique to IIP and each input separately. The results of our estimation are shown in Graph 5 below, where the normalized values, between 0 and 1, of EEMD trends of each proxy variable and IIP are plotted.

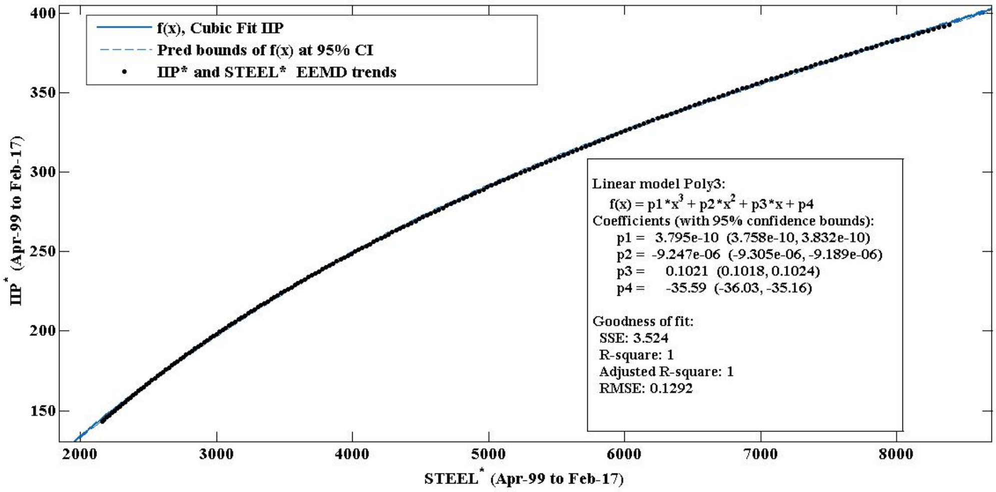
Graph 5.EEMD estimation of IIP and Inputs



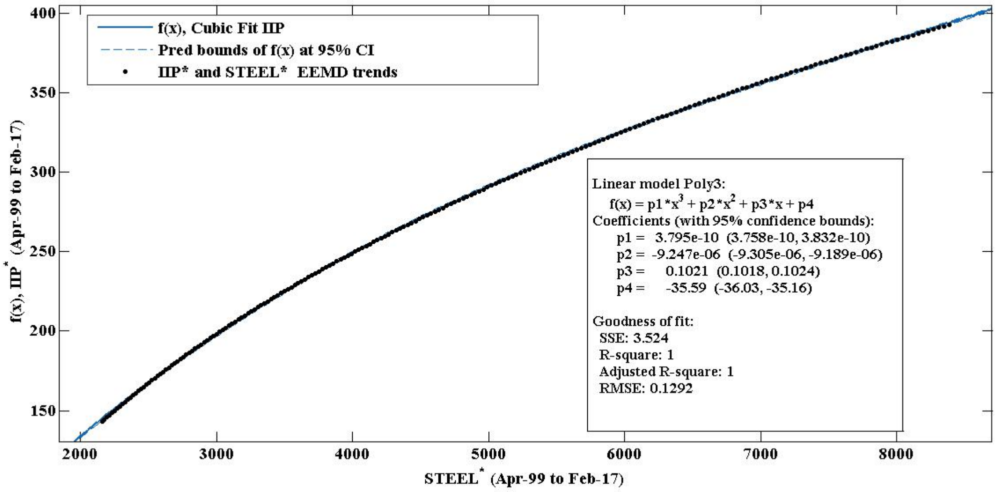
Inspection of Graph 5 clearly indicates that the movement in IIP is best tracked by Steel production. Hence we can argue that one coincident indicator of economic activity (one indicator of this is IIP) could be obtained using data on Steel production which should be readily available in real time.

We thus tried to estimate values of IIP for the period till February 2017 for which actual values of steel production and IIP were available at the time of writing. Since data on STEEL was available till February 2017, some estimate for IIP could be obtained based on the relationship between IIP and STEEL. In obtaining this relationship we used the following procedure. First, the EEMD fit to data for IIP (IIP*) and Steel (STEEL*) till Feb 2017 was performed . We then investigated the relationship between IIP* and STEEL* taking the latter as the explanatory variable. The best fit function (moving from linear to non-linear) was obtained as one which minimizes the sum of squares of the residuals. The function so obtained was found to be a cubic polynomial, f(x). IIP*, STEEL* and f(x) are plotted below in Graphs 6(a) and 6(b).

GRAPH 6(a).



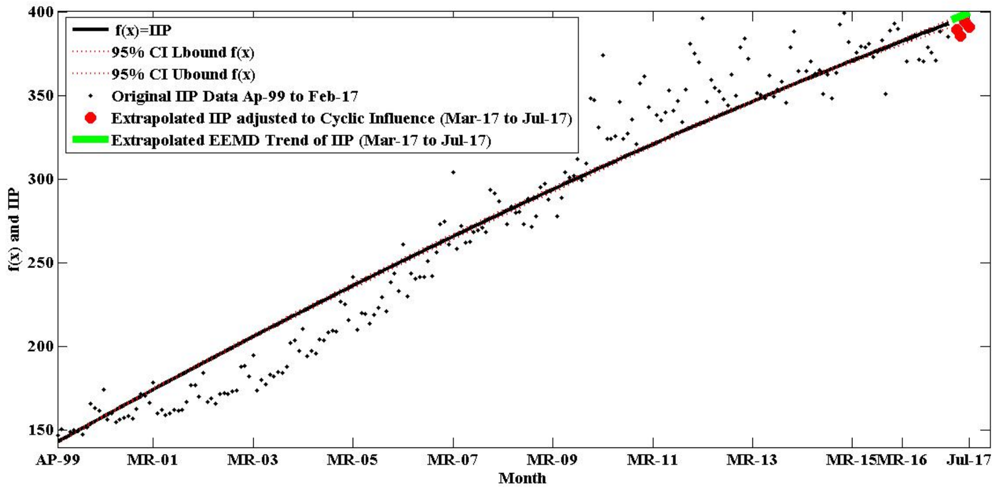
Graph 6(b)



In Graph 6(a), the relationship between IIP* and STEEL* is shown as the series of dots. The relationship is clearly non-linear. The best fit non-linear function which explains this relationship is derived as the third degree polynomial function, f(x). As a demonstration, plotting f(x) against STEEL* gives the blue line in Graph 6 (b) which replicates the IIP* vs STEEL* plot.

The relationship f(x) is then used to obtain a 5-period ahead forecast (March 2017 to July 2017) for IIP. This is shown in Graph 7 below.

Graph 7



In graph 7 we have plotted the function f(x) as an indicator for IIP, from April 1999 to February, 2017. Each month also represents a corresponding value for STEEL*. Beyond February 2017 we have first obtained extrapolated values for STEEL* for the months March 2017 to July 2017 to obtain corresponding extrapolated values for f(x) which represent our predictions for IIP. These predictions are shown as the green line in Graph 7. However, from these extrapolated values cyclical influences must again be removed using EEMD technique. The corrected forecast values are indicated by the red dots in Graph 7. As is clear, the corrected forecast values show a clear decline in IIP from the figures till February 2017.

As some validation we have calculated the correlation coefficient between the original values of IIP from April 1999 to February 2017 and our predicted values for IIP, f(x). The Spearman's rank correlation is about 0.98 so that our predicted IIP values track the original values quite closely.

Using our coincident indicator, f(x), we calculated three growth rates of IIP: till March 2012, g1; from March 2012 till February 2017, g2,(for which data is available) and March 2017 till July,2017, g3. A comparison of g2 and g1 indicates how the Indian economy was impacted by the slowdown in global growth after the initial impetus of tax and other cuts in 2008 died out. The rate g3 indicates what our forecast is for the coming months. It turns out that g1=.533 percent per month, g2=.280 and g3=.214.

it has often been argued that the world recession had a limited impact on the Indian economy and the Indian economy is "decoupled" from the global economy . Our figures show that this is not true. Till 2012, the various fiscal measures taken by the Government of

India (excise cuts, largesse of the 6th Pay Commission etc.) did postpone the impact of the recession. However, this impact seems to have worn off by 2012 as the CAGR falls to almost half at .280 per month in the period April 2012 to February, 2017. Our calculations show that in the coming months till July 2017, the growth is expected to fall further to about .214 percent per month.

VI. Conclusion.

Our main objective has been to demonstrate that trends in time series data where cyclical behavior (in the presence of multi-cycle influences) is likely can be best analysed using EEMD analysis. Analysing IIP time series data we argue that, despite all the infirmities, the IIP is a reasonable indicator of economic trend activity but only after all cyclical influences are accounted for. However, the IIP *per se* is an index which is data intensive. We suggest that, using an input approach to predicting economic activity indicates that a good proxy for IIP is the production of steel. Our analysis indicates that using data for steel (which should be available on a real time basis) future downsings or upswings in the IIP can be predicted quick. We use this relationship to develop a co-incident indicator for IIP which uses only data on STEEL production. Using available data, our analysis of the indicator shows that, in the five years after 2012, there was a significant downturn in economic activity. Our indicator also suggests that this slowdown is likely to continue till at least the second quarter of this year.

What is necessary is to use this analysis to develop a good predictor of booms and busts in economic growth. This will constitute the next phase of our study.

References

1. Banerji, Anirvan, and Pami Dua. "Predicting recessions and slowdowns: a robust approach." Working paper, *Centre for Development Economics, Delhi School of Economics, Delhi University*(2011).

2. Chakrabarti, Saumya, and Anirban Kundu. "Formal-informal sectors' conflict: A structuralist framework for India." *Journal of Economic Development* 34.2 (2009): 27.

3. Huang, Norden E., et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 454. No. 1971. The Royal Society, 1998.

4. Kalyani, Muna. "Indian Informal Sector: An Analysis."*International Journal of Managerial Studies and Research* (2016): 78-85.

5. Kundu, Amitabh (1993). "Urban Informal Sector: An Overview of Macro Dimensions, Socio Economic Correlates and Policy Perspectives in India ", in H. Nagamine (ed.) Urban Informal Sector Employment: A Cross National Comparative Study, Research Institute for Regional Planning and Development, Nagoya.

6. Klein, L. R. "Simulated Reality-A Contradiction in Terms?." *Eastern Economic Journal* 3.1 (1976): 28-29.

7. Pindyck, R.S. and Rubinfeld,D.L. Econometric Models and Economic Forecasts, McGraw-Hill, 3rd edn., 1991.

Sambamurty, R., Barbara Harris White, and Anushree Sinha. Trade Liberalisation and India's Informal Economy, 2009, pp. 147-149.

Wu, Zhaohua, et al. "On the time-varying trend in global-mean surface temperature." *Climate dynamics* 37.3-4 (2011): 759-773.

Wu, Zhaohua, Norden E. Huang, and Xianyao Chen. "Some considerations on physical analysis of data." *Advances in Adaptive Data Analysis* 3.01, n02 (2011): 95-113.

Annexure:

Ensemble Empirical Mode Decomposition (EEMD) explained in brief

Mathematical rationale and steps governing Empirical Mode De-composition (EMD) algorithm was put forth by Huang et al. (1998), the method was developed to isolate the presence of multiple cyclic factors from the trends embedded in a time series data. Frequently encountered shortcomings– non-linearity and non-stationarity– associated with time dependent variation in a time series data were either overlooked or compromised with *a priori* assumptions based massaging (smoothing, transformation, moving averaging) of the data in question. This allowed the application of statistical methods amenable to analyze oscillatory data over a time domain which are more suited for composite linear data response involving many variables. Almost all data representing natural phenomenon (global warming, wind speed, solar flux, flow of river water etc.), economic development indices based on multiple factors having different periodicity and magnitude, or time dependent variability in the number of vehicles passing a given location. Time series data of all these manifestations can be represented as:

Time Series Data (t) = Total Cyclic(t) + Acyclic Trend(t) or

$$Time\ Series\ Data(t) = \sum_{i=1}^{n} IMF_i + \mathrm{Re}\,siduals(t)_n \quad \text{...........1}$$

$$\mathrm{Re}\,siduals(t)_n = Trend(t)_n + random\,error(t)_n \quad \text{...........2}$$

In equation 1, above, ***IMF**$_i$* represent internal mode function representing ***i**$_{th}$* cycle***n*** cycles embedded in the time series data under analysis and ***t*** stands for time; each ***IMF*** in the data is tested for its significance weather the attributes of the cycle differ from equivalent cycle obtained from white random noise. Second term in equation 1 represent residuals, trend which, more often than not, is linear. Residuals can be conceived to be sum of the trend in the time series data and random error. Estimation of error in the extracted respective ***IMFs*** is not trivial, particularly when only one data point is available at a given time ***t*** in the time series, use of improved EMD algorithm EEMD allows the determination of ensemble average values for respective IMFs and trend with their respective uncertainty region at 95% Confidence Interval (CI). For clarity the following text describes the steps involved in the analysis of a time series data using EEMD algorithm to obtain embedded IMFs and trend (linear or non-linear) without having to invoke any *a priori* assumptions as the analysis is not constrained by the presence of non-stationarity or non-linearity in the data. To ensure clarity in understanding the steps involved a representative synthetic time series was created and analyzed.

EEMD analysis:

Time series data is plotted as a function of time for visual appraisal. Time dependent oscillations present in the data are indicative of the multiple cyclic influences having differing periodicity (frequency) and magnitude. The plotted data displays the presence of maxima and minima in the time series, the program identifies the coordinates of maxima and minima.

Cubic spline is fitted to maxima and the fit is designated as **upper-envelope**, similarly minima fit of cubic spline represents **lower-envelope**. Step 1 and 2 are shown in figure A1

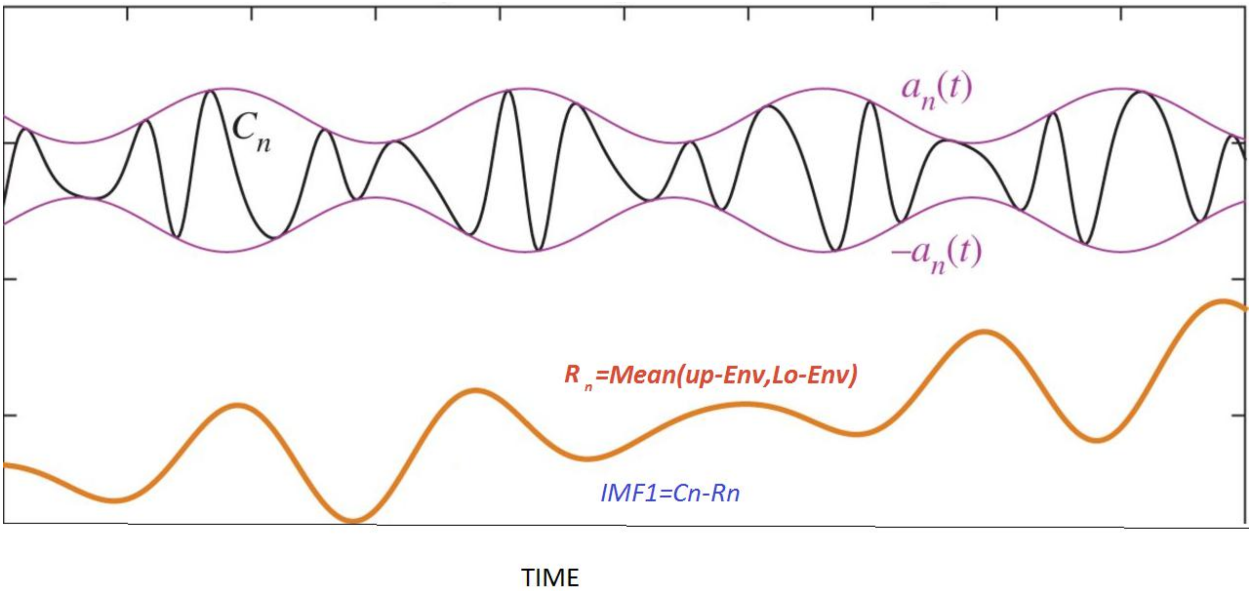


Figure A1. Upper panel displays original time series plot Cn and cubic spline fit to maxima points (upper-envelop) and minima (Lo-envelop).

The mean Rn is subtracted from Cn to yield IMF1 (highest frequency mode) through iterations to satisfy the criteria.

The spread in ensemble average IMF1 is obtained by adding random white noise in proportion to the standard deviation of the time series Cn. Six sets of noise assisted IMF_i

after addition of different proportion of white random noise (0.1xSD, 0.15xSD…0.3xSD). This facilitates in determining each respective IMF at 95% CI.

Subsequent steps are repeated on Cn-<IMF1>, where <IMF1> is the highest frequency IMF in the data, to obtain IMF2…..IMFn in increasing frequency. Finally, subsequent to the extraction of all IMFs the embedded ensemble average trend is obtained with spread (95% CI).