MICRO FORECASTS OF INDIA'S EXPORT SECTOR

METHODOLOGY

(NEURAL NETWORK)

February'2004

Modelling & Simulation Division
NATIONAL INFORMATICS CENTRE
Department of Information Technology
Ministry of Communication & Information Technology
A-Block, CGO Complex, Lodhi Road
New Delhi-110003

Email : analytic@hub.nic.in,
&
rkg@hub.nic.in

Part of main report
(India's Export Forecast for Financial Year 2004-05)
"Specific countries and specific commodities"

An Econometric Model

Preface:

The present report covers the micro level forecasts for specific countries and specific commodities, as a major input for planning India's export for the financial year 2004-05. The main study will use these forecasts, as a major input to develop an econometric model to derive macro level forecasts, for strategic planning of India's exports. The micro-level forecast is short-term monthly forecasts, carried out for different commodities and with principal trading partners. The forecasts obtained thus, would be reviewed mid-year subsequently, keeping in view the changed scenario with regard to fluctuations in the external and domestic market. The present report is a part of the main econometric study. The micro-level forecast study is done by the modelling division, National Informatics Centre, Ministry of Communication & Information Technology at the request of Ministry of Commerce, in close association with Research & Information System (RIS), Ministry of External Affairs.

The monthly time series behavior have been captured using Neural Network methodology. The approach integrates, captures and understands the varying demand conditions and the degree of price competitiveness (exchange rate at the discarded level). The monthly input for the present study have been provided by the Research and Information System (RIS), Ministry of External Affairs. The final forecast have been closely studied with respect to various statistical error measures and an in-build optimal selection criteria have been used in the selection of the forecasting model. The model behavior have been simulated for with-in sample, and once stabilized and coverage's, out-of-the sample forecasts have been generated.

The various level(s) of discussions have taken place over time, between Modelling Division, NIC and RIS to arrive at the final forecast(s). The total number of forecast(s) to be obtained is of the order for 322 data sets. The various internal modelling exercises have been undertaken, and the final selection is based on the forecast obtained following Neural Network methodology. Cognos 4Thought has been finally chosen for developing forecasts corresponding to all the 322 data sets. The reliability of the forecast and degree of confidence have also presented, besides major statistical measures along with the final forecasts.

A. Contents:

			Page #
1.	Introduction	….	1
2.	Objective of the Study	….	1
3.	Input Series	….	1-2
4.	Modelling Approach	….	8-10
5.	Forecasting Techniques/Selection Criterion	….	10-21
6.	Neural Network Modelling Software	….	21-22
7.	Results & Interpretation	….	23-25
8.	Limitation of the Study	….	26
9.	Conclusion	…	26

B. Tables:

		Page #
. Table – 1: Summary of country wise data sets	…	3
. Table –2: List of selected item code for five destination countries (USA, Canada, China, Japan & EU)	…	4-6
. Table – 3: Output description codes.	…	7
. Fig. 1: Graphical Representation of Actual, Fitted & Forecasts Values	…	23
. Fig 2 : Model Form	…	24
. Fig 3 : Statistical Output	…	25

C. Annexure :

. Output 1: Final Forecast model details

. Output 2: Statistical summary including Error Statistics

1. INTRODUCTION:

1.1 Department of Commerce, Ministry of Commerce and Industry have requested NIC vide their letter No.18(54)/2001-EPL dated 15th July,02 informing that the forecast model building work related to macro level forecasts of India's export has been entrusted to Research and Information System (RIS), Ministry of External Affairs. RIS in turn requested National Informatics Centre to undertake the micro-level forecast for India's export to NIC. On behalf of NIC, Modelling & Simulation division of NIC have undertaken this project of developing micro level forecasts of India's export, as a strategic planning tool to finally develop India's export policy for the financial year 2004-05. (An Econometric modelling study)

1.2 Number of interaction have taken place since then, between RIS and NIC to finally decide upon the methodology, taking into consideration the time constraints for obtaining in-depth forecasts for large number of data sets, numbering 322. The various approaches have been discussed and debated internally and preliminary analyses have been undertaken taking into consideration the availability of the software tool(s), the expertise available and over and above the limitation of the time constraints. Based on above, software availability was explored and finally decided on Cognos 4Thought. As all the data sets (time series) have a seasonality, besides other components, thus it was absolutely necessary to develop micro-level (instead of macro-level) forecasts for India's export to destination countries

2. OBJECTIVES OF THE STUDY:

As the present study is the major input to the main study, being carried out by RIS. The objective of present study is to develop the micro-level forecast for India's export, as well as the import of the major items from the chosen destination countries. The micro-level forecast thus obtained has been used as major input to derive the micro level forecasts, based on an Econometric model.

3. INPUT SERIES:

3.1 The number of commodities, variable in each commodity, the time period of the data availability for all the four variables (Import, Export, Unit Value Index of India, Unit Value Index of ROW) including rest of the world (ROW), world, European Union (EU) besides countries viz USA, Canada, China, Japan, Malaysia, Singapore, Thailand, Hong Kong and Bangladesh is given at Table-1. Thus, the monthly data set on the India's Export sector by the commodities and the destination consists of 322 variables.

3.2 The time period of each of the data set is varying and given at Table-1. The item chosen for analyses are the ones that have the major shares in the India's trading scenario, identified for each of the five destination countries and have been studied very closely.

3.3 Although, item wise destination is not made for countries such as Malaysia, Thailand, Singapore, Hong Kong, Bangladesh, Rest of the World and the world for the present study. Instead an additional exercise have been carried out from 26 items imported by European Union (EU) and unit index of EU imports from rest of the world for the same 26 items on yearly basis.

***Table 1**: SUMMARY OF COUNTRY WISE DATA-SETS*

(Time Series Forecasting Carried for the listed number of data sets)

Number of				Time periods of data availability			
Country List	**Com-Codes**	**Var. for each Code**	**Total Vbls**	*Var 1*	*Var 2*	*Var 3*	*Var 4*
Canada	13	4	52+2(rest) =54	Apr 1996 to June 2003	Jan 1995 to Nov 2003	Jan 1995 to Nov 2003	Jan 95 to Nov 2003
USA	17	4	68	Apr 1996 to May 2003	Jan 1993 to Oct 2003	Jan 1993 to Oct 2003	Jan93 to Oct 2003
China	10	4	40	Apr 1996 to May 2003	Jan 1995 to Nov 2003	Jan 1995 to Nov 2003	Jan 1995 to Nov 2003
Japan	11	4	44	Apr 1996 to June 2003	Jan 1994 to Nov. 2003	Jan 1994 to Nov. 2003	Jan 1994 to Nov 2003
Malaysia*	1	1	1	Apr 1996 to Aug 2002	NA	NA	NA
Singapore*	1	1	1	-do-	NA	NA	NA
Thailand*	1	1	1	-do-	NA	NA	NA
Hong Kong*	1	1	1	-do-	NA	NA	NA
Bangladesh*	1	1	1	-do-	NA	NA	NA
Rest of World*	1	1	1	-do-	NA	NA	NA
World*	1	1	1	-do-	NA	NA	NA
EU✡	26	4	104+2(rest) =106	Apr 1996 to June 2003	Jan 1996 to June 2003	Jan 1996 to June 2003	Jan 1996 to June 2003
TOTAL			**322**				

* Only single variable total export of "all commodities" from India is considered.

✡ Includes both the series- monthly as well as annual - with 26 items in each series.

Where,

- Var1 = India's **Export** to destination country
- Var2 = Total **Imports** of the destination country
- Var3 = Unit value **Index** of India
- Var4 = Unit value **Index** of row

Table 2, presents the item code along with their complete description as given by **WTA.**

Table 2: List of Selected Item Codes for 5 Destination Countries

Notation	Item Code	Description
United States of America		
1	30613	Shrimps and prawns frozen
2	420310	Articles of apparel
3	500720	Other fabrics, containing 85% or more by weight of silk or of silk waste other than oil silk
4	570110	Carpets and other textile coverings of wool or fine animal hair
5	570231	Carpets and other textile coverings of wool or fine animal hair
6	610510	Men's or boys' shirts of cotton, knitted or crocheted
7	610910	T-shirts, singlets & other vests, of cotton, knitted or crocheted
8	620442	Women's or Girls' Suits, Ensembles etc. of cotton
9	620443	Women's or Girls' Suits, Ensembles etc. of synthetic fibers
10	620520	Men's or Boys Shirts of cotton
11	620630	Women's or Girls' shirts blouses or shirt blouses of cotton
12	630492	Other furnishing articles, not knitted or crocheted, of cotton
13	640351	Other footwear with outer soles of leather covering the ankle
14	680223	Simply cut or sawn granite with a flat/even surface
15	710239	Other non-industrial diamonds
16	732599	Other cast articles of malleable cast iron, nes
17	Rest	Rest of the codes
Canada		
CAN 1	30613	Shrimps and prawns frozen
CAN 2	100630	Semi-milled or wholly milled rice, whether or not polished or glazed.
CAN 3	520511	Measuring 714.29 decitex or more (not exceeding 14 metric number)
CAN 4	610510	Men's or boys' shirts of cotton, knitted or crocheted
CAN 5	610610	Women's or girls, blouses, shirts and shirt-blouses of cotton, knitted or crocheted
CAN 6	610831	Women's or girls' nightdresses & pyjamas of cotton, knitted or crocheted
CAN 7	610910	T-shirts, singlets & other vests, of cotton, knitted or crocheted
CAN 8	620442	Womens or Girls Suits,Ensembles etc. of cotton
CAN 9	620443	Women's or Girls Suits, Ensembles etc. Of synthetic fibers
CAN 10	620520	Men's or Boys Shirts Of cotton
CAN 11	620630	Women's or Girls shirts blouses or shirt blouses of cotton
CAN 12	630492	Other furnishing articles not knitted crocheted of cotton
CAN 13	710239	Other non-industrial diamonds

CAN 14	Rest	Rest of the codes
China		
CH 1	30379	Other frozen fish, excluding livers and roes
CH 2	251611	Granite Crude or Roughly Trimmed
CH 3	260111	Iron ores and concentrates Non-agglomerated
CH 4	260112	Iron ores and concentrates Agglomerated
CH 5	261000	Chromium ores and concentrates
CH 6	520511	Cotton yarn Measuring 714.29 decitex or more (not exceeding 14 metric number)
CH 7	520521	Cotton yarn Measuring 714.29 decitex or more (not exceeding 14 metric number)
CH 8	520710	Cotton Yarn containing Cotton>=85% by Wt Put Up for Retail Sale
CH 9	670300	Human Hair, dressed, thinned, bleached or otherwise worked; wool or other animal hair or other textile materials, prepared for use in making wigs or the like.
CH 10	Rest	Rest of the codes
Japan		
J 1	30379	Other frozen fish, excluding livers and roes
J 2	30613	Shrimps and prawns frozen
J 3	50610	Ossein and bones treated with acid.
J 4	151530	Castor Oil and Its Fractions
J 5	230400	Oil-cake and other solid residues, whether or not ground or in the form of pellets, resulting from the extraction of soya-bean oil.
J 6	260111	Iron ores and concentrates Non-agglomerated
J 7	620442	Women’s or Girls Suits, Ensembles etc. of cotton
J 8	620520	Men's or Boys Shirts of cotton
J 9	620630	Women’s or Girls shirts blouses or shirt blouses of cotton
J 10	680223	Simply cut or sawn granite with a flat/even surface
European Union		
EU 1	30613	Shrimps and prawns frozen
EU 2	80132	Shelled cashew nuts
EU 3	90111	Coffee neither roasted nor decaffeinated
EU 4	90240	O black tea,fermntd
EU 5	151530	Castor oil and its fractions
EU 6	251611	Granite,crude/rough
EU 7	294200	Other organic comp ounds.
EU 8	300490	Other medicaments put in doses for therapeutic use
EU 9	380810	Insecticides
EU 10	410619	Goat or kid skin leather tanned/retanned
EU 11	420221	Leather handbags
EU 12	500720	Woven fabrics of silk >85%silk, nt n oil silk

EU 13	520811	Woven fabrics of cotton unbl, plain weave=<100g/m2
EU 14	570231	Carpets and other textile coverings of wool or fine animal hair
EU 15	610510	Men's or boys' shirts of cotton, knitted or crocheted
EU 16	610910	T-shirts, singlets & other vests, of cotton, knitted or crocheted
EU 17	620442	Women’s or Girls Suits, Ensembles etc. of cotton
EU 18	620520	Men's or Boys Shirts of cotton
EU 19	620630	Women’s or Girls shirts blouses or shirt blouses of cotton
EU 20	630419	Bedspread, n kn/cr tex
EU 21	630790	Other made-up articles including dress patterns
EU 22	640351	Other footwear with outer soles of leather covering the ankle
EU 23	640610	Uppers and parts thereof other than stiffeners
EU 24	680223	Simply cut or sawn granite with a flat/even surface
EU 25	710239	Other non-industrial diamonds
EU 26	711319	Articles of jewelry and parts thereof of other precious metal, whether or not plated or clad with precious metal.

Table-3 : Output Description* - Codes

Country	**Code**	**Variable Description**	**Code**
. United State	1,2,3…. (Numeric)	Export	EXP
. Canada	CAN1, CAN2…	Import	IMP
. China	CHI, CH2 ...	Index-India	IND
. Japan	J1,J2 J3 ...	Index-Rest of World	ROW
. European Union	EU1, EU2 …		
. Rest of World	ROW		

+ Note : This to be link with Table 1 describing the item code.
- The final code consists of country code & corresponding variable code

4. MODELLING APPROACH:

4.1 The monthly time series for each of the variable is available as an input series. The raw plot of data indicates that there is definite seasonal factor in the time series behavior besides others. Different methodologies have been attempted to capture the past behavior to predict the immediate future. The brief description of main approaches explore as follows:

Trend (T_t), Seasonality (S_t) , Cycling (C_t) and Irregularity (I) components. The first three components are deterministic called as signals while "I" is a random variable called "Noise" thus S_t is equal to $T_t * C_t * I$

4.2 In order to get the reliable forecast, it is necessary to determine the extent of each component in a time series. Hence, to understand and measure these components the procedures involved initially removing the component from the input series (the composition). After the facts are measured, making a forecast involve to be packed the components on forecast estimates (the composition).

4.3 The Auto-Regressive Integrated Moving Average (ARIMA) methodology is a generalized time series modelling approach for forecasting, and after long debate, considering the magnitude of the work involved in the analyses of large number of time series as well as the degree of the reliability of the forecast required, besides the experience already held in using ARIMA models for time series analysis considerable efforts have already be put in this area. Thus, ARIMA methodology was given more wheightage as compared to other time series techniques. Besides, state-of-the-art forecasting system viz. exploring neural network approach for developing forecasts. As Time Series is non-linear in behavior in almost all the cases, the interpretation and the availability of the latest version of the neural network based software viz. 4Thought was explored and procurement of the software was explored .

4.4 **Selection Criterion:** Neural network was finally chosen for developing the above forecasts for all the variables. The corresponding software viz. 4Thought was finally followed for the above analysis.

4.5 Theoretical Background:

4.5.1 Time Series Analyses

Since a large number of factors are likely to influence time series, it is thus very important that influences or components be separated out, at the 'raw' data levels. In general, as there are four types of components in any time series (X_t): Seasonality (S_t), Trend (T_t), Cycling (C_t) and Irregularity (I). The first three components are deterministic which are called "Signals", while the last component is a random variable, called

"Noise".

$$X_t = S_t * T_t * C_t * I$$

To be able to make a proper forecast, it is necessary to know the extent of presence of each component in the series. It is thus essential to understand and measure these components.

4.5.1.1 TREND (T_t): A time series may be stationary or exhibit trend over time. Long-term forecast procedure involves initially removing the component effects from the data (decomposition). After the effects are measured, making a forecast involves putting back the components on forecast estimates (recomposition).

4.5.1.2 SEASONAL VARIATION (S_t) : When a repetitive pattern is observed over some time horizon, the series is said to have seasonal behavior. Seasonal effects are usually associated with calendar or climatic changes. Seasonal variation is frequently tied to yearly cycles, Trend is typically modeled as a linear, quadratic or exponential function.

4.5.1.3 Cyclical variation (C_t): An upturn or downturn not tied to seasonal variation. Usually results from changes in economic conditions. Besides,

- **Seasonalities** are regular fluctuations, which are repeated from year to year with about the same timing and level of intensity. The first step in a times series decomposition, is to remove seasonal effects in the series. Without deseasonalizing the data, we may, for example, incorrectly infer that recent increase patterns will continue indefinitely (i.e., a growth trend is present) when actually the increase is 'just because it is that time of the year' (i.e., due to regular seasonal peaks). To measure seasonal effects, a series of seasonal indexes are calculated. A practical and widely used method to compute these indexes is the ratio-to-moving-average approach. From such indexes, it may be possible to quantitatively measure how far above or below a given period stands in comparison to the expected data period (the expected data are represented by a seasonal index of 100%, or 1.0).

- **Trend,** is growth or decay that are the tendencies for data to increase or decrease fairly steadily over time. Fitting a line or any other function on the deseasonalized data does measurement of the trend component. This fitted function is calculated by the method of least squares represents the overall trend of the data over time.

- **Cyclic** oscillations, are general up-and-down data changes due to changes e.g., in the overall economic environment (not caused by seasonal effects) such as recession-and-expansion. Again a series of cyclic indexes are calculated to measure how the general cycle affects data levels. Theoretically, the de-seasonalized data still contains trend, cyclic, and irregular components. Also, the predicted data levels using the trend equation do represent pure trend

effects. Thus, it stands to reason that the ratio of these respective data values should provide an index which reflects cyclic and irregular components only. As the business cycle is usually longer than the seasonal cycle, it should be understood that cyclic analysis is not expected to be as accurate as a seasonal analysis.

Due to the tremendous complexity of general economic factors on long-term behavior, a general approximation of the cyclic factor is the more realistic aim. Thus, the specific sharp upturns and downturns are not so much the primary interest as the general tendency of the cyclic effect to gradually move in either direction. To study the general cyclic movement rather than precise cyclic changes (which may falsely indicate more accurately than is present under this situation), the data is 'smoothed' off the cyclic plot by replacing each index calculation often with a centered 3-period moving average. Note that as the number of periods in the moving average increases, the smoother or flatter the data become. The choice of 3 periods perhaps viewed as slightly subjective may be justified as an attempt to dampen out the many up-and-down minor actions of the cycle index plot so that only the major changes remain.

- **Irregularities** (I) are any fluctuations not classified as one of the above. This component of the time series is unexplainable therefore it is unpredictable. Estimation of I can be expected only when its variance is not too large. Otherwise, it is not possible to decompose the series. If the magnitude of variation is large the projection for the future values will be inaccurate. The best one can do is to give a probabilistic interval for the future value give the probability of "I" is know

5. FORECASTING TECHNIQUES & SELECTION CRITERION

5.1 The selection and implementation of the proper forecast methodology has always been an important planning and control issue. There are two main approaches to forecasting. Either the estimate of future value is based on an analysis of factors which are believed to influence future values (the explanatory method) or else the prediction is based on an inferred study of past general data behavior over time (the extrapolation method). variables, automatic and user specified forecasting models for predictor variables.

5.2 Output Statistics

The major output statistics produced are:

- t-statistic;
- Squared t-statistic;
- Slope error (or inverse of t-statistic) expressed as a percentage;
- Model R^2 (model fit), test R^2 (test fit), Overall (model + test) R^2 (overall fit)

and Adjusted overall R^2; (R : Multiple Correlation Co-efficient)

- Root Mean Square Error (residual);
- Standard Deviation of model;
- 95% Confidence Forecasting Error;
- Mean Absolute Error (residual) and mean absolute percentage error (residual);
- F-statistic;
- Durbin-Watson (DW) statistic.

Classical definitions of some of these statistics should be used with caution since they are defined for use with linear models such as those obtained by linear regression. In case a given procedures produces non-linear model (relationship in the series is non-linear?), thus producing more accurate models than linear models. However, this means that some statistics become meaningless without being adapted to the non-linear case, which is what some software tools produces, including 4Thought Each statistic is explained below, both in the classical sense and how it is used in 4Thought Suggested use of each statistic is also given.

5.2.1 t-statistic

Definition:

In classical linear regression, the t-statistic is used to determine, if an input variable affects or is correlated with the output variable. In other words, it is used to test whether the slope of the cross section of the input is significantly different from zero. Significant in this sense means, "given the distribution of the data, what is the probability that the observed slope is in fact zero and only appears non-zero because of random error?" In statistical literature the significance level, usually adopted is 5% or 1%, meaning, "if the t-statistic is greater than X at the 5% significance level then we may conclude, with a certainty greater than 95%, that the slope does actually differ from zero," where X is a value obtained from a table of the t-distribution given in most statistics literature.

The classical t-statistic is given by:

$$t = \frac{\bar{a}_j \sqrt{\sum_{i=1}^{n} \left(x_i^j - \bar{x}^j \right)^2}}{s_e} \quad ,$$

where $\bar{a}_j$ is the estimated cross-section slope of the j^{th} input, x_i^j is the value of the j^{th} input for the i^{th} data point, $\bar{x}^j$ is the mean value of the j^{th} input, s_e is the sample estimate of residual standard deviation and n is the number of data points. Also,

$$s_e^{\,2} = \frac{\sum_{i=1}^{n} e_i^{\,2}}{n-k} \quad ,$$

where e_i is the residual error (actual value A_i – model value M_i) for the i^{th} data point and

k is the number of coefficients in the regression equation (n - k is the number of degrees of freedom).
The expected value $E(s_e)$ is s_e, the true residual standard deviation of the whole population that the data was obtained on.
Sometimes, t-statistic is defined slightly differently as,

$$t = \left[\text{Average slope on cross - section}\right] \times \frac{s_i}{s_e} \times \sqrt{n-k} \quad ,$$

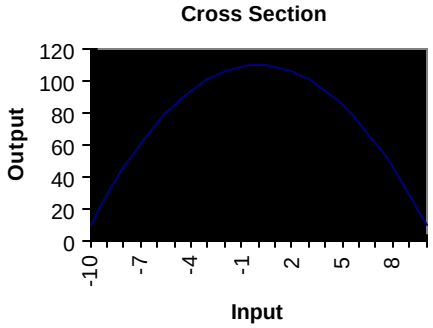
where s_i is the standard deviation of the input and s_e is the standard deviation of the residual error.
The average slope is taken because a Freefor model does not approximate the relationship between the input and output variables by a straight line as in linear model.

5.2.1.1 What does it mean?

The Freefor t-statistic may be interpreted in a similar way to the classical one, i.e. as a measure of how significant an input is to the output. If a variable 'x` has a high positive t-statistic then one may infer that 'x` has a significant positive influence on the output, and if the t-statistic is highly negative then 'x` has a significant negative influence. A very low t-statistic means that the input is insignificantly correlated with the output and thus should be removed from the model.

5.2.1.2 Rigorously speaking, to determine what constitutes a 'minimum' (or threshold) t-statistic under which value the input is considered insignificant, it (referred to as 'X` above) may be looked up in a table of the t-distribution (single tailed) in any statistics book. The number of degrees of freedom d.f., required for the table, is given by $n$- $k$ where, $k$ is the number of inputs to the model. Thus, if there are 30 data points and 5 inputs, 'X` is found to be equal to 1.03 at a 5% significance level (95% confidence level). So a reported t-statistic of 1.03 would mean that there is a 5% chance that the input variable is actually insignificant and that the observed cross-section is a product of mere random noise due to sampling error. Conversely, there is a 95% chance that the variable does positively influence the output.

5.2.1.3 Practically speaking, however, for a model where n- k > 20, a t-statistic greater than 1 may be considered as significant. Also, the t-statistic is only a good measure as long as the cross-section does not become too non-linear. As an example, consider a highly significant input which produces a cross-section like:



In this case the average slope is zero and the t-statistic will be zero, even though there is a significant correlation.

5.2.1.4 Important note:

The t-statistic assumes that any unexplained variation in the output is due to random noise. If the model is built with too few inputs or the wrong inputs then most of the unexplained variation is not noise. In this case the t-statistic will be 'artificially' low and it's absolute value becomes meaningless; inputs may only be compared relative to each other rather than using the t distribution tables as described above.

As a rule-of-thumb, the absolute t-statistic will only be valid when there is a corresponding high R^2, or when the modeller is sure that the residual error comes from random sources. Otherwise, we recommend that an input variable be judged as significant if it's t-statistic is at least 5% the magnitude of the largest t-statistic.

5.2.2 Squared t-statistic

Definition :

The squared t-statistic is defined as

$$t_{\text{squared}} = \sqrt{\left[\text{average squared slope on cross - section}\right]} \times \frac{s_i}{s_r} \times \sqrt{n-k} \quad ,$$

where s_i and s_r are the standard deviations of the input and output respectively, n is the number of data points and k is the number of inputs to the model.

5.6.2.1 What does it mean?

This is a specific statistic with no classical parallel and it should be used instead of the t-statistic when the input has a strong non-linear effect. It may be used in the same way as a t-statistic for determining if an input is significant and if it should be kept or removed. The squared t-statistic is always positive.

5.2.3. Slope Error

Definition:

The slope error is expressed as a percentage and is given by,

$$E_{Slope} = \frac{1}{t_{\text{squared}}} \quad .$$

5.6.3.1 What does it mean?

When a linearised equation of a cross section is generated, the slope error represents the random error of the coefficient of an input. It defines the single standard deviation of the coefficient as a percentage of the coefficient's value. So when looking at a term in a linearised equation, the slope error may be used to determine the 95% confidence interval for the coefficient of the particular input variable, given by $\pm 2E_{Slope}$.

5.2.4 R^2-statistics

Also known as the *Coefficient of Determination*, the R^2-statistic is classically given by,

$$R^2 = \frac{s^2_{O_i - \overline{O}}}{s^2_O} \equiv \frac{\text{variance of output explained by model}}{\text{total variance of output variable}} \quad ,$$

where O_i is the model value for a particular set of inputs, O is the corresponding required output, and $\overline{O}$ is the mean output. See figure A.

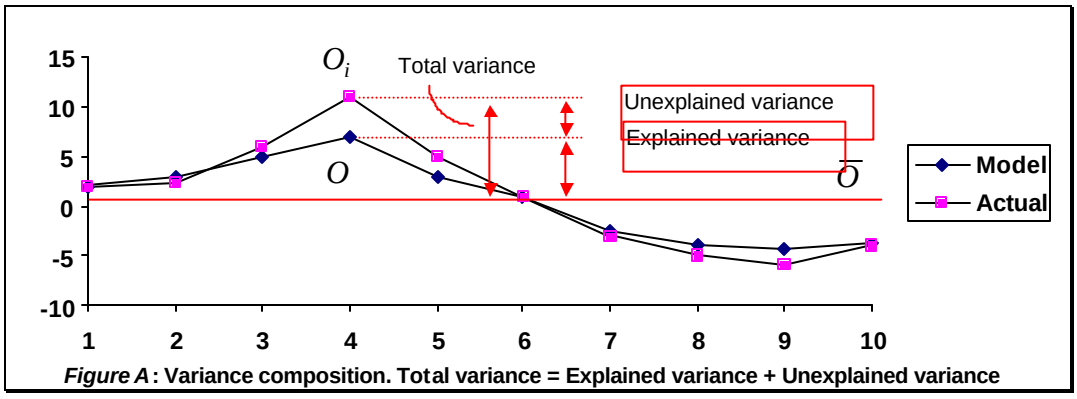


***Figure A*: Variance composition. Total variance = Explained variance + Unexplained variance**

Sometimes, a modification of the classic R^2 given by,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (A_i - M_i)^2}{\sum_{i=1}^{n} (A_i - \overline{A})^2} \quad ,$$

where M_i is the model value for a particular set of inputs, A_i is the corresponding actual

output, $\overline{A}$ is the mean of the actual values and *n* is the number of data points.

In R^2 values are given for the model data, test data and all data, and are called *model fit*, *test fit* and *overall fit* respectively. Also, R^2 has an interesting property of being zero in the naïve model case; i.e. when there is no discernible pattern in the data and the model value M_i reverts to $\overline{A}$ for all rows. It is 100% when the model exactly fits the actual values, and can go negative for the test set when the model is over fitting to the training set.

However, for a more realistic value, the **adjusted R^2** should be used because the formula above is derived using only a sample of the true population of A and not the whole population.

5.2.5 Adjusted R^2-statistic

Definition:

The number of *degrees of freedom* is classically given by the number of data points minus the number of coefficients being determined in a model. This is because each coefficient requires at least one data point to determine its value, leaving the rest of the points 'free'. The number of coefficients is the total number of weights. However, when simulation with fixed starting points for the weights and because the stopping rule will prevent the model from becoming unnecessarily non-linear, the number of coefficients is effectively lower and may safely be approximated as the number of inputs to the model. This is because, under the stated circumstances, the model will become only non-linear enough to connect the data points which, for a perfect model, is of order number of inputs. The number of inputs is the actual number of inputs plus one, the constant (or bias) coefficient.

Thus the adjusted R^2 is given by,

$$R_A^2 = R_O^2 \times \frac{(n-k)}{n} \quad ,$$

where *n* is the number of data points, k is the number of inputs and R_O^2 is the unadjusted overall R^2 (i.e. all data, model plus test).

5.2.5.1 What does it mean?

The adjusted R^2 tells us what percentage of the variance of the data is accounted for by the model. It is not equivalent to the classical coefficient of determination; this is a tool specific statistics. As the number of inputs k increases, the complexity of the model increases and so it becomes less likely that the n data points will be sufficient to describe the model, so the adjusted R^2 decreases.

5.2.6 Root Mean Square (RMS) error

Definition:

This is the square root of the mean of the squared errors (or residuals) of the model and is given by,

$$E_{rms} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \quad ,$$

where e_i is the residual error (actual value A_i – model value M_i) for each data point and n is the number of data points.

5.2.6.1 What does it mean?

RMS error gives a measure of the average magnitude of the residual errors over all data points. It should be compared to the RMS value of the data. The smaller the error, the more closely the model follows the actual data.

5.2.7 Standard Deviation of model

Definition:

Standard deviation is given by,

$$s = \sqrt{s} \quad , \quad \text{and} \quad s = \frac{\sum_{i=1}^{n} \left(M_i - \overline{M} \right)^2}{n-1} \quad ,$$

where, s is the variance, M_i is the model value for data point i, $\overline{M}$ is model mean and n is the number of data points. The -1 in the denominator comes about because variance has one derived value, $\overline{M}$, which is calculated from the data thus effectively removing one data point and giving n–1 degrees of freedom.

5.2.7.1 What does it mean?

Standard deviation is a very common measure of the dispersion of values about their mean, i.e. how much variation there is in the data. It is the square-root of variance and so has the same dimension as the data.

5.2.8 95% confidence forecasting error

Definition:

For normally distributed data, the area covered by 2σ either side of the mean defines the 95% confidence interval. Since Freefor, when left in automatic mode, always attempts to map the distribution of each variable into a normal one, this is a good measure of the confidence interval.

The *95% confidence forecasting error* is the confidence interval of any forecast made by Freefor. It is also given by 2σ either side of the mean, where the standard deviation σ is the same as that obtained from the model. Use of the same σ is justified because Freefor uses part of the real world data for testing — as soon as the test variance begins to diverge from the model variance Freefor stops — so that the model σ , test σ, and therefore forecast σ may safely be assumed to be equivalent.

5.2.8.1 What does it mean?

The 95% confidence forecasting error determines the distance between two limits, one above and one below a forecast, within which the true value is expected to lie 95% of the time. So the narrower these limits, the more precise are the forecast, i.e. the more confidants the forecast is.

5.2.9 Mean Absolute error

Definition:

This is the mean of the absolute values of the residual errors and is given by,

$$\overline{E}_a = \frac{\sum_{i=1}^{n} |e_i|}{n} \quad ,$$

where e_i is the residual error (actual value A_i – model value M_i) for each data point and n is the number of data points.

5.2.9.1 What does it mean?

Very similar to the RMS error. Practically, in the modelling context, they are both measures of the same thing but the RMS error gives a higher value. Generally though, the RMS value is more physically meaningful.

5.2.10 Mean Absolute Percentage Error (MAPE)

Definition:

This is the mean of the ratios of errors to their corresponding actual values, expressed as a percentage, and given by,

$$MAPE = \sum_{i=1}^{n} \left| \frac{e_i}{A_i} \right| \quad ,$$

where A_i and e_i are the actual value and residual error (A_i – model value M_i) respectively for each data point and n is the number of data points.

5.6.10.1 What does it mean?

MAPE is a relative measure of how large the error is compared to the magnitude of the values being modeled. A low MAPE means that the model follows the actual values very

well. It is conceivable that a badly built model might have an R^2 close to 100% (very good) where the variation in the model is almost equal to that in the data but is in different places to that in the data. In this case the MAPE will be very high, indicating that there is something wrong with the model.

5.2.11 F-statistic

Definition:

The F-statistic is given by,

$$F = \frac{R_O^{\,2}/k}{\left(1-R_O^{\,2}\right)/\left(n-k-1\right)} \equiv \frac{\text{Explained variation}}{\text{Unexplained variation}}$$

$$\equiv \frac{\sum_{i=1}^{n}\left(M_i-\overline{M}\right)^2}{\sum_{i=1}^{n} e_i^{\,2}} \times \text{Adjustment by number of degrees of freedom},$$

where $R_O^{\,2}$ is the overall R^2-statistic (as defined above), k is the number of inputs, n is the number of data points and M_i and e_i are the model value and residual error respectively for each data point.

This differs slightly from the classical definition since the R^2-statistic differs from the classical.

5.2.11.1 What does it mean?

The F-statistic measures the ratio of explained to unexplained variation in the data. So if the F-statistic has the value 2, there is twice as much variation in the model than in the residual errors. In the limit, as the error decreases to zero, the F-statistic increases to infinity.

Basically, it may be looked at in this way: the higher the F-statistic, the better the model. Rigorously speaking, one may use the F-statistic to decide whether the hypothesis, "The model is purely random and the correlation we are seeing is actually non-existent," is to be accepted or rejected. Here for a modeller up, based on the numbers of degrees of freedom in the numerator and denominator, the borderline value of the F-statistic for a particular significance level.

For instance, if n = 400 and k = 7, we find in the tables that the threshold value of F is 2.03 at the 5% significance level (95% confidence level). This means that if F from our model has the value 2.03, there is a 5% chance that the model is meaningless (95% chance that it is valid). If F is higher, the confidence level increases. Again, for a 1% significance level (99% confidence), the value of F would have to be 2.69.
Practically speaking, for a typical modelling task with several hundred data points and 5 to 10 inputs, an F-statistic of at least 2 may be considered to indicate a reliable model (at

the 5% significance level accepted as the standard default by many statisticians). For specific cases of little data and/or many inputs the F tables should be consulted.

5.2.12 The Durbin-Watson statistic

Definition:

The Durbin-Watson statistic is used when making time-series forecasts to determine if any periodic patterns remain unaccounted for by the model. It is given by,

$$\text{D-W} = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^{\,2}} \quad ,$$

where e_i is the residual error on data point i and n is the number of data points.

5.2.12.1 What does it mean?

The D-W statistic is the ratio of successive error differences to the error values. Thus, if there is a fast changing residual pattern where e_i varies successively from positive to negative and vice-versa, $|e_i - e_{i-1}|$ will on average be twice the value of $|e_i|$ (see figure (2a)). Thus in this limit, D-W = 4. In the opposite limit, when there is an extremely slow moving residual pattern, $|e_i - e_{i-1}|$ will generally be much smaller than $|e_i|$, thus in this limit D-W = 0 (see figure (2b)). And if there is no systematic residual pattern — i.e. only random errors — then D-W averages to 2.

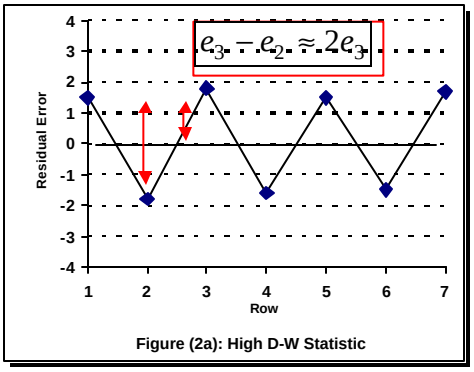


Figure (2a): High D-W Statistic

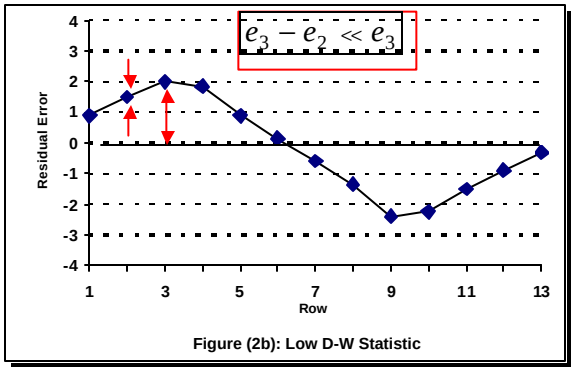


Figure (2b): Low D-W Statistic

So if D-W is significantly higher than 2, there is a fast moving unexplained pattern left, and if it is significantly less than 2 then there is a slow moving unexplained pattern left. The definition of 'significantly' is again given by a probability function tabulated in any statistical table.

Practically speaking, for a typical modelling application with several hundred data points and ten inputs, a D-W value of less than 1.2 or greater than 2.8 may be considered as significant.

6. ***4Thought* :** In earlier forcasts Arima technology was used for forecasting but this time we have chosen 4thought neural network technology for forecasting . We can use 4thought to analyze business situations, discover the variables that have greatest impact on your business, perform what-if analysis, predict future behaviour etc..

By performing a seven stage process 4thought can help to get the most out of your data

- Identify the question
- Determine the data requirements
- Find and prepare the data
- Build the model
- Build the forecast
- Explore the results
- Summarize the results

6.1 **Identify the question :** First stage in building a model is to identify the questions which 4thought is supposed to answer. These questions describe the type of performance we want to measure or explain. These questions should be very specific as they will identify the data which we need to gather

6.2 **Determine the data requirements** : in this stage we determine the data which is required to build the model. First we need to decide on a target that is the performance measure which best represents the question we formed in stage one. Next identify the variables which will affect our target and these variables become the factors in our model.

6.3 **Find and prepare the data** : this stage includes selecting data sources, preprocessing the data as necessary prior to bringing it into 4 Thought, importing the data into 4thought

6.4 **Building the model :** This stage involves building the model , evaluating its predictive ability and refining it as required. 4thought models can fall in 2 broad categories

- Time series models
- Profile model

Since our data was time series based we have used time series models. When we begin to build the model, 4thought makes repeated passes through the data to analyze the relationships between each factor and target. With each iteration of the build 4thought sets aside part of data to test the accuracy of the model. The model's performance is measured by how well te predicts the values of the test set data. 4Thought modelling process is unique in that is identifies the iteration of the model with highest predictive ability and returns to it after it finishes.

Once the model is finished , we can check its predictive ability by comparing the actual target values to the values specified by the model. The modeled values will never be identical to actual values . The correspondence between actual and predicted values are encapsulated in the model fit and Test fit statistics.

6.5 **Building the forecast :** Forecasts are time series models that predict future values of a target based on historical data. There are 2 types of forecasting

- **Univariate ,** in which we forecast values based only on historical target values and time or seasonal factors.
- **Multivariate,** in which we forecast target values based on one or more factors.When we build a multivariate forecast, 4Thought predicts future values for each factor by building univariate forecasts and then using these forecasts generates a forecasts for target.

6.6 **Explore the results :** After we finalize our model or forecast, we can explore the results in greater depth, we can identify periodicity and time trends in our data. A forecast enables to take the analytical and predictive powers of modelling and apply them to future . This provides us with unparalleled support for decisions that can be made to impact the future of our business

6.7 Summarize the results : After we analyze the model or forecast , we can summarize the results and prepare them for distribution. We can make our business information available to others who don't have access to 4Thought by presenting the results graphically and preparing and printing the reports .

7.1 Final Forecasts:

The detailed output giving the final estimated model coefficients, the error statistics as well as the graphical representation of actual, fitted and forecast value for each of the 322 variables are given at Appendix…

8. LIMITATIONS IN THE STUDY

8.1 Limitation of 4Thought was that it doesn't choose the best model…we had to generate the models based on different factors and choose the best one so it so every variable we have to regenerate the model till we get the best model.

8.2 The analyses of all series have certain limitations with regard to the model fitting. It is observed that in cases where the data variation is within a small range, the fit is only a simple average straight-line model

8.3 In cases where there is one or two extreme observation(s) but within the 3sigma limit, i.e. not identified as outliers, the fit is modified to adjust these extremities causing increase in the actual-fit deviation of remaining observations.

9. CONCLUSION

9.1 The medium range forecasts have been generated with all the corresponding details for 322 variables. The details of each is given in the following order.

Var 1: India's Export to destination country (Exp)
Var 2: Total Imports of the destination country (Imp)
Var 3: Unit value Index of India (Ind)
Var 4:Unit value Index of Rest of World (Row).

9.2 An attempt, to generate forecast based on Neural Network based approach, may have been better, in lieu of time constraints and to in-corporate in-consistence behavior of the large number of time series.

9.3 There were some ambiguities in the data itself i.e. having some –ve values foe excport-import data but it will be taken care by RIS while doing Macro level analysis.