Forecasting and Event Detection in Internet Resource Dynamics Using Time Series Models

S P Meenakshi, *Member, IAENG,* and S V Raghavan

***Abstract*—At present, Internet emerges as a country's predominant and viable data communication infrastructure. Autonomous System (AS) topology occupies the top position in the Internet infrastructure hierarchy. AS resources are building blocks of this topology, and consist of AS numbers, IPv4 and IPv6 prefixes. Further, the resource requirement in each country is dynamic and driven by various technical and socio-economic factors. Hence, the organizational and national competitiveness for socio economic development is reflected in AS growth pattern. Furthermore, to assess the competitiveness, future expansion, and policy development, there is a need for both study and forecast AS growth. For Internet infrastructure development, understanding long term trends and stochastic variation behaviour are essential to detecting significant events during growth periods. In this work, we use time series based approximation for mathematical modeling, system identification, and forecasting to determine the annual AS growth. The AS data of five countries, namely India, China, Japan, South Korea, and Taiwan were extracted from the APNIC (Asia Pacific Network Information Centre) archive for this purpose. The first two countries have larger economies and the next three countries are advanced technological nations in the APNIC region. The characterization of the time series is performed by analyzing the trend and fluctuation component of the data. The model identification is carried out by testing for non stationarity and autocorrelation significance. ARIMA (Auto Regressive Integrated Moving Average) models with different Auto Regressive (AR) and Moving Average (MA) parameters are identified for forecasting the AS growth of each country. Model validation, parameter estimation, point forecast, and prediction intervals with 95 % confidence levels for the five countries are reported in the paper. The statistical analysis on long term trends and Change Point Detection (CPD) on Inter Annual Absolute Variations (IAAV) are presented. The significant level of change in variations, positive growth percentage in IAAV, and higher percentage of advertised ASes when compared to other countries indicate India's fast growth and wider global reachability of Internet infrastructure from 2007 onwards. The correlation between AS IAAV change points and GDP (gross domestic product) growth periods indicates that the service sector industry growth is the driving force behind significant annual changes.**

***Index Terms*—AS topology, statistical analysis, AS growth forecasting, long term trend, inter annual absolute variation.**

I. Introduction

INTERNET infrastructure plays a crucial role in a country's economic, educational, and social development, by expanding organizational and national competitiveness. In the rapidly developing communication infrastructure scenario, Internet converges with other communication platforms, such as public switched telecommunication networks and broadcast medias. To better understand the Internet evolution, infrastructure and performance indicators are essential. Infrastructure indicators help to apprise issues related to technological limitations and resource address portability, level of competition in the backbone market, and convergence across different communication platforms. From a policy perspective, the indicators play a significant role in regulating infrastructure.

Autonomous System (AS) topology takes up the top position in the Internet network hierarchy. In the topology, the visibility of each network and its allocated IP addresses to global data communication infrastructure is through AS Numbers (ASNs) and prefixes allocated for the network. Besides, AS resources are globally competitive because of their limited number Viz., 65535 ASNs. The AS resource requirement of each country is dynamic and driven by various factors [1] such as economic growth, industrial growth, intra-national connectivity, and backbone routing table aggregation. Hence to acquire the resources, plan for future expansions, and to develope policies there is a need for growth forecasting. Additionally, understanding about long term trend and stochastic variation behavior are essential to detecting significant events during AS resource growth as an Internet infrastructure development indicators [2]. In our work, extraction and analysis of ASNs allocated by APNIC database to five countries belonging to two different classes were performed in order to forecast AS growth. Further, long term trend and stochastic behaviour in the growth pattern were explored. We use the term ASN interchangeably with AS count in this paper.

ASNs are allocated by Internet Assigned Numbers Authority (IANA). The Regional Internet Registries (RIR) redistribute these resources directly to customers and National Internet Registries (NIR). From the global AS pool, the RIRs get a working pool of 1024 blocks in order to meet current assignment demands. ASN requirements technically depend on number of large Internet Service Providers (ISPs) deploying policy based services, multihoming users, and distributing usage of AS numbers as identifiers in multipath label switching virtual private networks [3]. When a network uses different inter domain policies, then a public AS number is used to realize each policy. In case of single policy networks, private ASNs are used between the service providers and the BGP (Border Gateway Protocol) speaking client network. The reachability of ASes through their prefixes are ensured by the global routing protocol E-BGP using routing data exchange process. In a country level, based on demand from different service providers and industries, the ASNs are assigned from the working pool by the RIRs. The ASNs are unstructured and have no direct aggregation advantage, but consecutive numbers can be used to effectively separate domestic and external traffic in firewalls implemented at

Manuscript received October 10, 2014; revised March 17, 2015.

S.P. Meenakshi was with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, TN, 600036 India e-mail: spmeena@cse.iitm.ac.in.

S.V. Raghavan is with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, TN, 600036,India email: sv.raghavan@nic.in.

Border ASes i.e., the Great Firewall of China [4].

One more technical concern in acquiring ASNs is about 16 bit and 32-bit ASN pools. As some of the existing ASes E-BGP platforms support only 16-bit ASNs, using a 32-bit ASN means a transition of the E-BGP platform. Since BGP platform transition takes place in its own temporal phase due to additional investment cost, 16-bit AS numbers are the mostly sought after ones until its predicted exhaustion in the year 2014. From the 16-bit global AS pool that consists a total of 64510 ASNs, 61438 ASNs are already allocated to RIRs [5]. There are seven thousand eight hundred and twenty eight ASNs allocated from the global 16-bit AS pool for APNIC RIR, out of which 801 ASNs are still unallocated according to statistics obtained [6] from the archive as of February 2013.

The ASN utilization of each country is dynamic and is driven by various other factors, such as economic growth, industrial growth, and government policies apart from technical factors. In this work, we employed time series based approximation for mathematical modeling, system identification and forecasting of the annual ASN growth. The AS data of five countries, Viz. India, China, Japan, South Korea, and Taiwan were extracted from APNIC archive for this study. In this, the first two countries are larger economies and the next three countries are advanced technology nations in APNIC region. Considering the dynamics of resource consumption with respect to each country, an appropriate model selection is extremely important for forecasting.

The dynamics induced by factors like economic growth, increase in policy based ISPs and multihoming users, competition for 16-bit ASN pool, and indirect summarization advantage have different impacts on the long term growth trend and yearly variations. Additionally, statistical analysis on AS data and the reason for temporal variations are sparsely addressed in the literature. So in this work, we have attempted a statistical analysis on country-wise ASN data with the following objectives:

1) To understand the country-wise AS growth that is influenced by normal and unusual occurrence of events, technology advancements and economic growth.
2) To automate the process of monitoring and forecasting of temporal behaviors (i.e., AS long term growth trend and yearly variations).
3) To understand the global reachability percentage of ASes and related growth.

We accomplished our objectives by exploring the following questions:

1) What are the characteristics of AS data associated with each country?
2) Why are ARIMA models suitable to estimate the data for forecasting?
3) How the long-term growth trend is presented for larger economies and technologically advanced countries?
4) How to detect significant changes in yearly variations and account for them?
5) Why assigned versus advertised ASN ratio is different for various countries?

The contributions of our work are time series model identification, validation, forecasting ASN growth, long-term trend analysis, event detection using inter-annual variations, event correlation with GDP, and compariosn of assigned versus advertised ASNs for the five countries.

The paper is organized as follows. The AS data analysis from APNIC data is discussed in section 2. ASN characterization, time series modeling, and forecasting are discussed in section 3. Long-term trend analysis is reported in section 4. Inter-annual variations and change point detections are discussed in section 5. AS route-view data analysis is performed in section 6 followed by related work in section 7. Our conclusion of this work is presented in section 8.

II. AS Data Analysis

Internet AS resources such as ASNs, IPv4 prefixes, and IPv6 prefixes are allocated and maintained by RIRs such as APNIC, RIPE and ARIN. The ISPs, large network users such as content providers, corporates, and universities receive AS resources on request and statically recorded in RIRs. The growth in allocated AS resources of a country indicates the growth of Internet infrastructure. We analyzed the APNIC RIR data for the countries under consideration to understand the growth behavior of ASN.

A. APNIC AS Data

The archived delegated resource data was obtained from APNIC RIR. The registered ASN details related to a country were extracted from it. The APNIC RIR provides the allocated/assigned Internet resource statistics for the countries in the Asia Pacific Region. The resource statistics accounted are for the following:

1) Autonomous Systems
2) IPv4 Addresses
3) IPv6 Addresses

Public Internet address spaces and ASNs allocated by IANA are redistributed by APNIC RIR to the countries located in the region. The resources are managed with well-structured policy guidelines by these organizations.

We have used the record format details provided by the RIR to interpret the retrieved data [6]. The standard record format for the data entry in all the RIRs is as follows: $<$ registry, cc, type, start, value, date, status, extensions $>$. Registry specifies the name of the RIR, and cc indicates the country code to which the resource has been delegated. The type field provides details on whether the resource is an ASN, IPv4 address, or IPv6 address. The rest of the fields give information on starting number, total count of the resource from its starting number, allocation/assigned date, whether the resource is reserved or assigned to an organization, and available count for future extensions. The aut-num attribute [7] of AS class holds a short description or a name of the organization to which an ASN is assigned. The summary report specifies the total number of ASNs, IPv4 prefixes, and IPv6 prefixes.

III. Time Series Modeling

In our work, the AS count over the years is considered as non-stationary time series data. The non-stationary property is confirmed using the Dickey Fuller hypothesis test. The data has been analyzed for modeling and forecasting, as well as to extract the following statistical properties for detecting events.

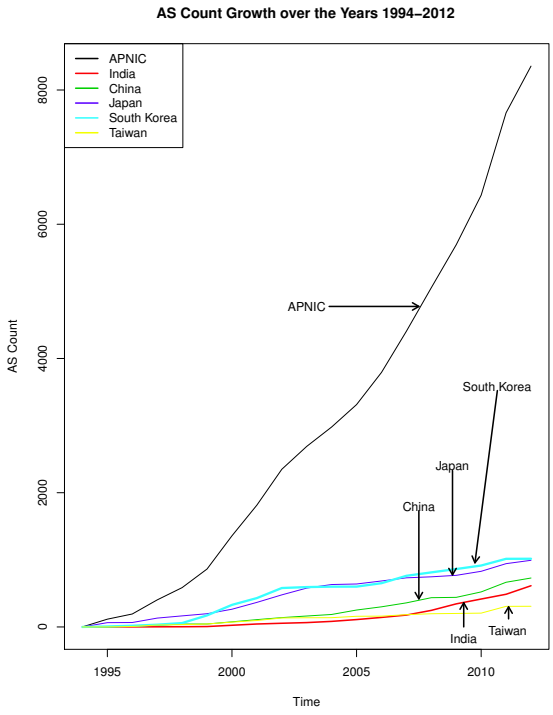

Fig. 1. ASN Growth Pattern Comparison

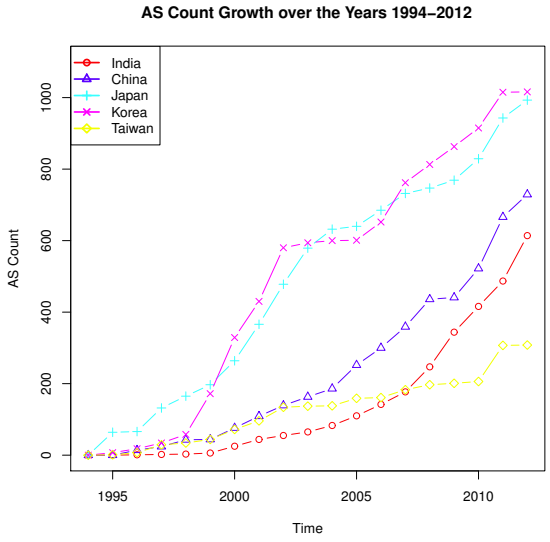

Fig. 2. ASN Growth Pattern for Five Countries

1) Long-term trend on AS growth over the years
2) Average annual growth rate and direction
3) Inter Annual Absolute Variation (IAAV) and direction

The AS count time series for APNIC region and the other five countries are given in Figs. 1 and 2. We have made the following observations on the annual AS count data. The AS number allocation in the APNIC region started in 1994. The countries such as Japan, South Korea, and Taiwan have started their registrations from 1995 onwards, which is a year ahead of India and China. The annual AS count of APNIC, India and China exhibits exponential growth trend while Japan, South Korea, and Taiwan exhibits exponential growth combined with intermittent linear trends. As the data is annual data, there is no seasonality associated with it. The variation is not constant, and different variation levels are observed in the growth trend. This property can't be captured by using a single linear or non linear equations or regression based models. This observation is also made in [8], [9]. Since the data is time series and non stationary, and additionally the growth is influenced by many factors, such as economical growth of a country and new content providers, we choose to use time series models. The time series model that *best* captures these data trends while relating the present values with past values and prediction error is the ARIMA model. This model consists of an Auto Regression (AR) component, known as lag term, and a Moving Average (MA) component, known as error term. Two of the significant criteria considered to select the best fitted forecasting model among the model candidates were based on correctness and the accuracy of forecasting with narrow prediction interval. Other criteria, such as minimum number of parameters, goodness of fit value, and residual independency with constant variance. The non seasonal general form of ARIMA model considered in our work is given in Eq. 1.

$$(1-\phi_1 B-\cdots-\phi_p B^p)(1-B)^d(y_t-\mu t^d/d!) = (1+\theta_1 B+\cdots+\theta_q B^q)e_t \quad (1)$$

where c = $\mu(1-\phi_1-\cdots-\phi_p)$ and μ is the mean of $(1-B)^d y_t$. ϕ_i is the AR coefficient and 'p' is the order of AR component. The MA coefficient is denoted by θ_i where as the order of it is represented by 'q'. The number of required differencing is denoted by the term 'd'. The lag terms of 'y' is specified by the notation 'B' while the error at time t is denoted by e_t. Specific values for the order terms, coefficients and differencing terms were computed from the respective countries' AS count data.

The Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) coefficients were computed between the series and lag terms to identify the dependencies of the response variable to past observations. The order of AR and MA components are determined by analyzing the pattern of significant coefficient values present for the lag terms. $Y_{1t},\ldots,Y_{6t}$ represent the AS annual growth of APNIC, India, China, Japan, South Korea, and Taiwan, respectively. The ACF and PACF plots of AS annual growth are given in the Figs. 3 and 4. The significant values of the ACF and PACF indicate MA(2) and AR(1) order components for the ARIMA model. The linear trend observed in Fig. 1 supports for a first order differencing term in the model. The linear dependency of the AS Count on its lag 1 values shown in Fig. 5 also suggests an auto regressive model for the data. We selected ARIMA(1,1,2), ARIMA(1,1,1) and ARIMA(2,1,3) as candidate models for estimation to overcome the sample noise issue. We presented complete details on ARIMA components estimation and validation for the time series of India. The forecasting results with prediction intervals are reported for all of the countries under consideration.

A. ARIMA Model and Residual Analysis

The AS annual growth data is modeled using the candidate ARIMA models. Regression is performed with observed predictor values on the response variables using minimization

TABLE I
ARIMA MODEL ESTIMATION FOR AS COUNT ANNUAL GROWTH DATA - INDIA

Model	AR Coeff.ϕ_i, Std. error	z Statistic	MA Coeff.θ_i, Std error	z Statistic	AICc	Sample Variance
ARIMA(1,1,1)	(0.96,0.07)	13.71	(0.85,0.16)	5.3	163	246
ARIMA(1,1,2)	(0.76,0.18)	4.22	(1.74,0.28)	6.21	164	176
			(1.0,0.31)	3.2		
ARIMA(2,1,3)	(0.87,0.40)	2.18	(1.26,0.37)	3.4	169	151
	(0.099,0.388)	0.25	(0.057,0.454)	0.125		
			(-0.53,0.28)	-1.89		

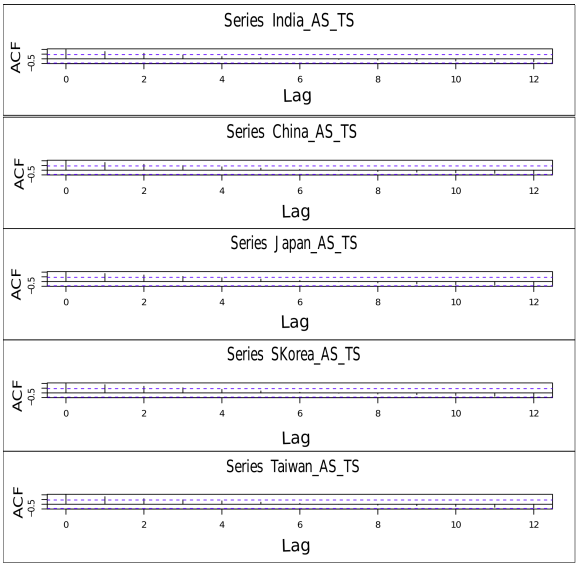

Fig. 3. AS Count Annual Growth ACF Plots

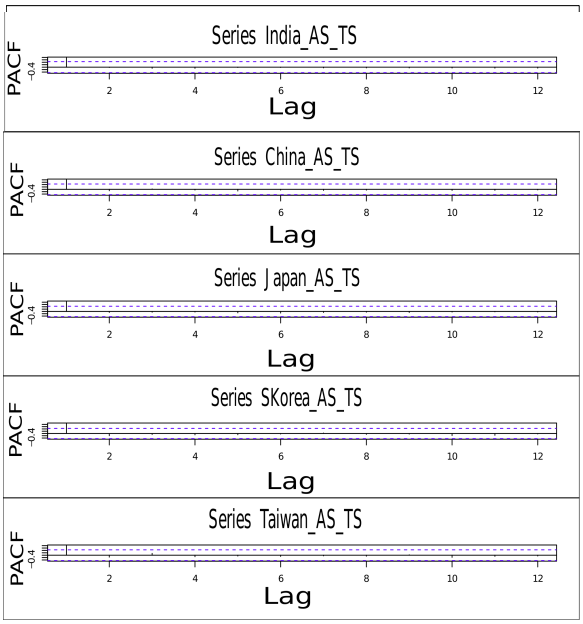

Fig. 4. AS Count Annual Growth PACF Plots

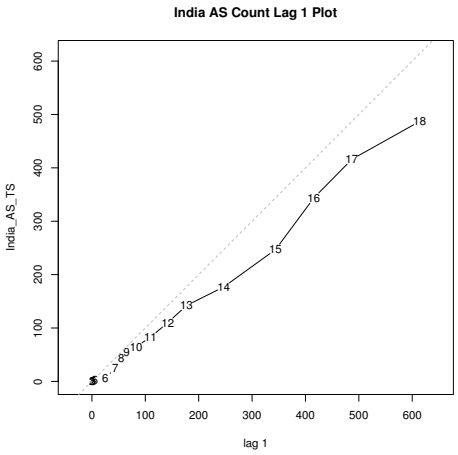

Fig. 5. Lag 1 Plot - India

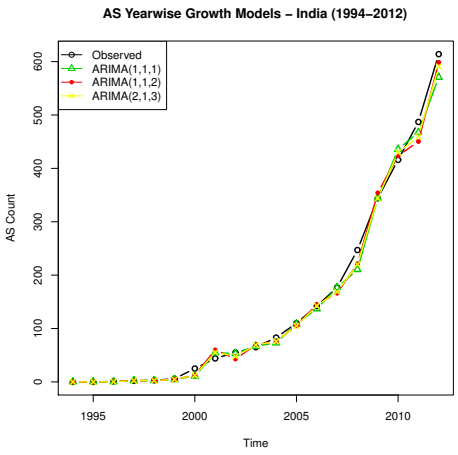

Fig. 6. AS Count Annual Growth Models - India

of the Conditional Sum of Squares (CSS) method combined with the Maximum Likelihood (ML) method. The non seasonal ordering is used in the model. ARIMA models do internal mathematical transformations on the series to detrend and stabilize the variance of the data. The transformations on data for detrending include differencing. For variance stabilization, functions such as log and square roots are used. The predictions are performed on the transformed data and been converted to original series by reversing the transformation process. We used the R statistical package for modeling [10]. The fitted models to the Indian AS annual data are shown in Fig. 6. The ARIMA(1,1,2) model visual fitness is good for the data. The estimated model component coefficients (ϕ_i,θ_i), standard error, z-statistic, and AICc (Akaike Information criterion with correction) are given in Table I. The z-statistic

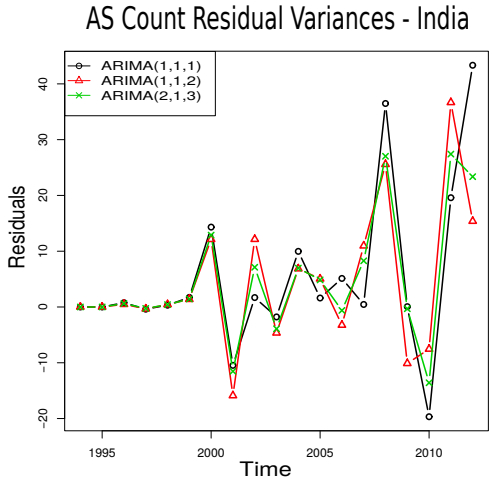

Fig. 7. Model Residual Variances - India

TABLE II
AS COUNT MODEL RESIDUAL NORMALITY TEST RESULTS

Model	Jarque Bera p-value	Shapiro-Wilk p-value
ARIMA(1,1,1)	0.04362	0.002736
ARIMA(1,1,2)	0.2215	0.2248
ARIMA(2,1,3)	0.4522	0.03706

is computed as a ratio between the component coefficient and standard error. The parameter is statistically significant when z-statistic is greater than 1.96. AICc is a goodness of fit measure computed based on number of parameters and information loss. The ARIMA(2,1,3) model coefficients, namely AR2, MA2, and MA3 are not statistically significant since the z-statistic is less than 1.96. Also, AICc increases and variance decreases along with the number of parameters. This may be attributed to over fitting of the observed values and more information loss caused by the increase in number of parameters of the models.

Residuals were analyzed to validate the fitness of the model for the observed data. The variances of the residuals for all three models are shown in Fig. 7. The time series plot of the residuals does not exhibit any abnormalities, and the variances are roughly constant over a period of time. The residual ACF given in Fig. 8 for all three models has no significant evidence for autocorrelation upto lag 12. Hypothesis tests for checking white noise property of the residuals were done using the Jarque Bera test and the Shapiro-Wilk normality test. The null hypothesis for both of the tests is that the residual series follows normal distribution. The p-values for each model are given in Table II. ARIMA(1,1,2) model has a p-value greater than 0.05 for both the tests, and hence the null hypothesis is accepted. The other two models fail in both or either one of the tests.

B. Forecasting

Forecasting of AS annual growth data was examined for correctness and accuracy using the three models. In the 19 annual ASN data, 14 values were considered for estimating the parameters of each model using which predictions were

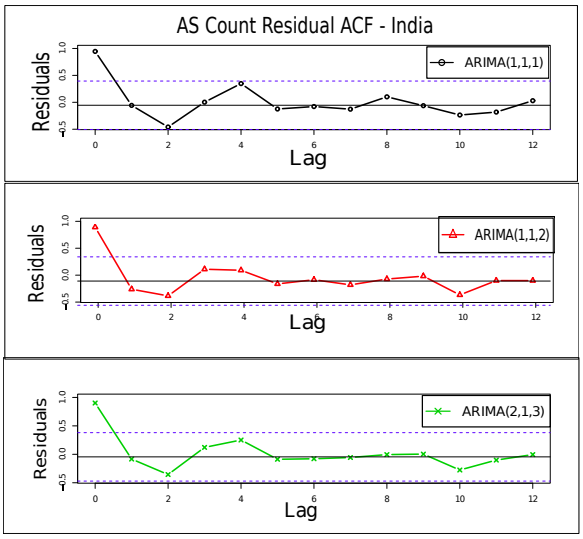

Fig. 8. Model Residual ACF - India

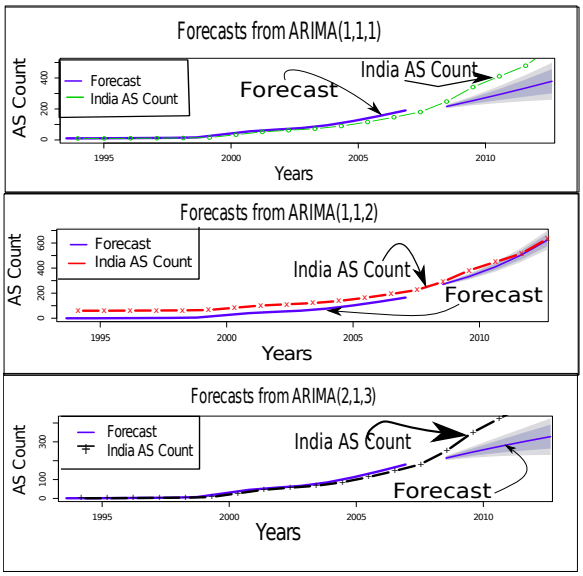

Fig. 9. Model Forecasts With Prediction Intervals - India

made for the next five years. Point forecasts as well as prediction intervals were computed. The prediction interval was computed using the standard error of the forecasts. The 95 percent confidence interval was computed by assuming the model error follows normal distribution. Forecast standard error was computed from Psi weights (ψ_i) that were obtained by converting the ARIMA model to the infinite-order MA model. The prediction interval (PI) was computed using current predicted value (CPV) and standard forecast error (SFE) from the Eq. 2. The forecast series along with PI for the three models are shown in Fig. 9.

$$PI = CPV \pm 1.96 * SFE \tag{2}$$

From the graphs, it is observed that ARIMA(1,1, 2) model fits and forecasts the annual AS data within the PI. For the other two models, the forecast values are well beyond the PI, except for the first value. This prooves the correctness of the ARIMA(1,1,2) model. Also, the PI of the model is narrow, which indicates that the variability of the data is stabilized

TABLE III
AS COUNT ANNUAL FORECASTING MODELS -INDIA

Model	Data Predictions	Prediction Interval	Accuracy RMSE
Observed	(247,344,416,487,614)		
ARIMA(1,1,1)	(214,253,294,337,383)	(203-225),(222-284), (238-351),(250-425), (259-507)	141.36
ARIMA(1,1,2)	(226,292,374,476,603)	(215-236),(266-317), (333-414),(417-534), (522-683)	32.32
ARIMA(2,1,3)	(211,242,271,298,324)	(200-222),(213-271), (221-320),(226-370), (228-420)	174.9

and establishes the accuracy of the model. We have given the point forecasts, 95 % confidence level prediction interval, and Root Mean Square Error (RMSE) accuracy in the Table III. The accuracy is computed as the square root of mean square errors, which occur as difference between the predicted and the observed values. ARIMA(1,1,2) model exhibits good point forecast, prediction interval, and RMSE accuracy when compared to other models for the AS count data. The ACF and PACF analysis, visual fitness of the model, number of parameter significance, residual analysis, and prediction properties confirms the adequacy of ARIMA(1,1,2) model for the Indian AS annual growth data.

Similar procedure was followed for model estimation, validation, forecasting, and selection when the other four countries and the APNIC region were considered. For the APNIC region, when compared to other models, ARIMA(1,1,1), estimation was visually good. The MA1 coefficient was less than the z-statistic threshold. Variances were roughly constant and no evidence of autocorrelation was found upto lag 12. Both the normality tests confirmed the normal distribution of the residuals. RMSE was also less and most of the forecast values have fallen within a 95 % confidence interval. Similar observations were made with this model for Japan, South Korea, and Taiwan. For China, when compared to other models, ARIMA(2,1,3) was relatively good with the expected residual and forecasting properties. The parameters and forecasting observations for the selected models are given in Tables IV and V.

IV. LONG-TERM TREND ANALYSIS

The long-term trend, which is a component of AS count time series, was estimated and analyzed in this section. As we observed before, AS count data has an additive long-term trend. This trend can be estimated from the observations using a statistical model that explicitly includes a local or global trend parameter. We identified a random walk model with drift parameter (ARIMA(0,1,0)) and a linear model to estimate the trend. The drift parameter of ARIMA(0,1, 0) and slope of the linear model represent the average annual growth rate. In addition, the slope of the linear model provides the direction of the growth. The estimated growth trend using both of the models to India and Japan are shown in Figs. 10 and 11. From the long-term trend graph, we observed that the

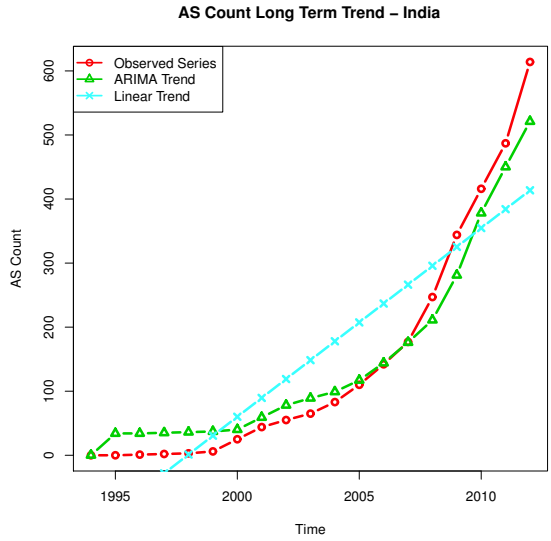

Fig. 10. AS Count Long Term Trend (1996-2012) - India

actual AS count cycles around the linear trend line. The half-cycle period is approximately twelve years for APNIC, India, and China. But for technologically advanced countries, the half-cycle period is approximately five years. Table VI gives the trend values for all of the countries under consideration.

Since the average growth rate computed by both methods are almost similar, we took the average from these two methods for further analysis. Technologically advanced countries Japan, Korea, and Taiwan have average annual growth rates of 56, 62 and 18 AS counts, respectively. Among larger economies, India and China have average annual growth rate of 36 and 44 AS counts, respectively. The APNIC region has an annual average growth rate of 459 AS counts. Relative average growth rate with respect to the APNIC region is computed for the purpose of comparison. India had 1.8 % less in relative average growth than China and approximately 5 % less average growth than Japan and South Korea. Out of 56 countries in the APNIC region, the average annual growth rate contributed by these five countries is 47 %.

It was observed that the long-term trend curve computed using the ARIMA model has structural difference between

TABLE IV
SELECTED MODEL ESTIMATIONS FOR AS ANNUAL DATA

Region / Country	Model	AR (Parameter, Std. error)	MA (Parameter, Std error)
APNIC	ARIMA(1,1,1)	(0.975,0.038)	(-0.33,0.24)
India	ARIMA(1,1,2)	(0.76,0.18)	(1.74,0.28) (9.00,0.31)
China	ARIMA(2,1,3)	(0.91,0.27) (0.061,0.269)	(-0.62,0.47) (-0.62,0.67) (1.00,0.48)
Japan	ARIMA(1,1,1)	(0.91,0.14)	(-0.33,0.44)
SKorea	ARIMA(1,1,1)	(0.75,0.20)	(-0.10,0.32)
Taiwan	ARIMA(1,1,1)	(1.0,.005)	(-0.98,0.12)

TABLE V
SELECTED MODEL FORECASTING FOR AS ANNUAL DATA

Region / Country	Model	Observation	Point Forecast	Prediction Interval	Accuracy RMSE
APNIC	ARIMA(1,1,1)	(5055,5695, 6433,7661 8356)	(5064,5765, 6513,7311, 8161)	(4839-5288,5243-6287, 5613-7413,5955-8666 6275-10048)	185.52
India	ARIMA(1,1,2)	(247, 344 416, 487 614)	(226,292, 374,476, 603)	(215-236,266-317, 333-414,417-534, 522-683)	32.32
China	ARIMA(2,1,3)	(436,441, 522,666) 729)	(391,434, 469,506, 539)	(365-417,381-488, 377-560,382-629, 381-698)	115.48
Japan	ARIMA(1,1,1)	(747,769 829,943 993)	(771,805 835,861 883)	(701-841,666-944, 620-1049,567-1154 507-1259)	64.42
SKorea	ARIMA(1,1,1)	(813,863 915,1015 1016)	(851,924 982,1029 1067)	(742-961,696-1151, 628-1335,545-1512, 454-1679)	49.64
Taiwan	ARIMA(1,1,1)	(197,201, 206,307 308)	(197,211, 226,240, 254)	(170-225,166-257 163-289,158-321 153-355)	39.89

TABLE VI
AS COUNT LONG TERM TREND

Region / Country	Period	Model	Average Annual Growth	Standard Error	Relative (APNIC) Growth Percentage
APNIC	1994-2012	ARIMA(0,1,0)	464	64	100
		Linear Model	453	25	100
India	1996-2012	ARIMA(0,1,0)	38	9	8.1
		Linear Model	34	4	7.5
China	1996-2012	ARIMA(0,1,0)	45	9	9.7
		Linear Model	43	3	9.5
Japan	1995-2012	ARIMA(0,1,0)	55	8	11.8
		Linear Model	57	2	12.6
SKorea	1995-2012	ARIMA(0,1,0)	59	12	12.7
		Linear Model	64	3	14.1
Taiwan	1995-2012	ARIMA(0,1,0)	18	6	3.8
		Linear Model	17	1	3.7

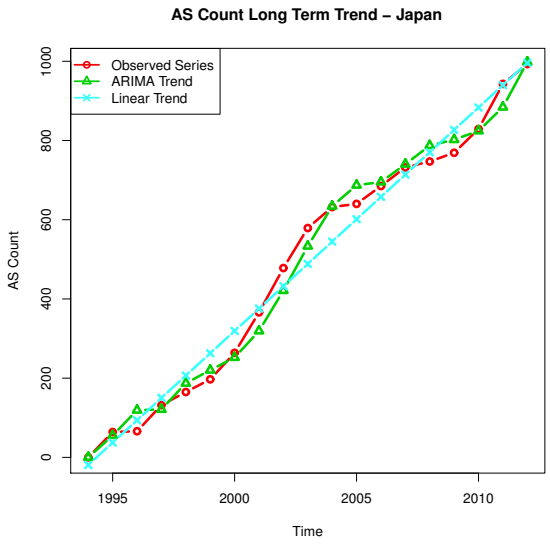

Fig. 11. AS Count Long Term Trend (1995 -2012) - Japan

larger economies and technologically advanced countries. We considered countries belonging to larger economies as Group I and countries belonging to technologically advanced countries as Group II. The structural deviation significance in the long-term trends between the two groups has been established statistically through hypothesis testing. To accomplish this, correlation coefficients between the series were used. Correlation coefficients r_1 between long-term trends of Group I countries was computed and compared with similarly computed correlation coefficents r_2 between long-term trends of Group II countries. Hypothesis testing on difference of sample correlation coefficients r_1 and r_2 was used to determine the statistical significance of structural deviations.

The population variables X and Y are AS-count annual growth rates of Group I and Group II countries. Bivariate normal distribution is assumed for the variables. The samples drawn for X and Y assign values of long term annual growth rate fitted from the ARIMA(0,1,0) model for a country. The correlation coefficient r was computed between two samples drawn from the same group or across groups. Fisher's transformation [11] is applied on *r* for variance stabilization. The procedure for hypothesis testing on significant statistical difference between correlation coefficients is as follows:

1) r1 is correlation coefficient computed between samples of size n1 within Group I.
2) r2 is correlation coefficient computed between samples of size n2 across Groups I and II .
3) z1 and z2 are computed using Fisher's transformation applied on r1 and r2 for variance stabilization.
4) z = $\frac{1}{2*ln[\frac{(1+r)}{(1-r)}]}$
5) zd = $\frac{(z1-z2)}{\sqrt{(\frac{1}{(n1-3)}+\frac{1}{(n2-3)})}}$
6) The difference zd is assumed to be standard normal
7) H0: z1 and z2 are equal, H1: z1 and z2 are different
8) If the absolute value of zd is less than 1.96, then accept null hypothesis H0
9) Otherwise, reject null hypothesis with a 95 % confidence level
10) p-value is computed as pnorm(zd)*2 when zd is negative
11) Otherwise p-value is computed as (1-pnorm(zd))*2

The r and z values computed with and across groups are given in Table VII. We can observe that for India, the r value is high within the group and less across the groups. This can be interpreted as that India has strong structural similarities in long-term trend pattern within Group I countries and significant structural difference with Group II countries. The statistical significance of difference in z1 and z2 and p-values for India are given in Table VIII. The zd values confirm that India has significanty different long-term trend when compared to the Group II countries.

V. Inter Annual Variations

The Inter Annual Variation (IAV) is another component of the time series. These values are random and do not contribute to the long-term trend. Each country has a different number of AS registrations annually, depending upon the domestic market. The demand for new ASes is influenced by the factors such as increase in the number of new ISPs, content providers, application providers, increase in intra-national connectivity and future reservations. All of these factors can be represented using the macro-economic variable service sector industries. If there is a significant variation in yearly AS count, it should have occurred due to the domino effect of fluctuations occurring in some of the aforementioned factors.

The IAV values are either positive or negative. We consider the absolute values for our analysis, which is denoted by the term, Inter Annual Absolute Variation (IAAV). The computation of it is carried out as $Y_t - Y_{t-1}$, which is the first differenced data of the time series. The annual AS count data has an additive trend, but IAAV is computed on the first difference of the annual data. In order to detect the presence of trend in the IAAV, ACF, and PACF values were examined. In addition, normality hypothesis test was performed on the values to identify the distribution. The ACF and PACF values are non significant for all of the lags of the computed IAAV data to the considered countries, excluding India. In order to remove the trend component present in the data for India, we applied the second differencing. The computed series is free of trend and data are considered to be independent. Furthermore, we applied the statistical tests Shapiro-Wilk and Jarque Bera to find the presence of normal distribution in the IAAV data. Both the tests assume normal distribution for the data as null hypothesis. The p-values of the tests are greater than 0.05 only for Japan, which confirms the acceptance of null hypothesis. For other countries, the p-values are non significant. Hence, normality assumption to the data is rejected.

To detect significant events Viz. change point in variances, we assume the data are independent and the distribution of the data is unknown. The IAAV distributions for the countries are shown in Fig. 12.

A. Event Detection

We define significant variation in IAAV as an event. The impact of the factors on the significant events of the data are ill understood or unpredictable in reality, so we use the

TABLE VII
CORRELATIONS AND FISHER TRANSFORMED VALUES

Region / Country	Region/ Country	Corr. Coefficient r	z-value
India	APNIC	.97	2.1
India	China	.98	2.1
India	Japan	.86	1.3
India	SKorea	.87	1.3
India	Taiwan	.90	1.5

TABLE VIII
CORRELATION DIFFERENCE AND P-VALUE FOR INDIA

With in Group I	z1 - value	Across Group I and II	z2-value	zd (difference)	p-value
India and China	2.1	India and Japan	1.3	2.11	.03
India and China	2.1	India and Taiwan	1.5	1.66	0.09

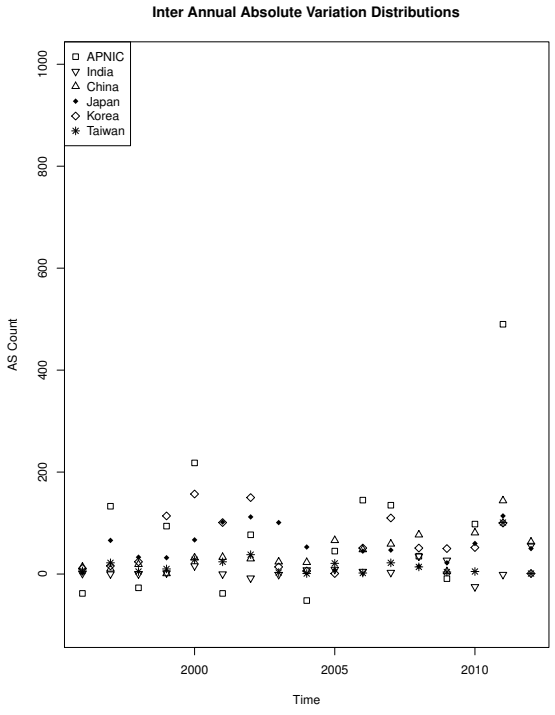

Fig. 12. AS Count IAAV Distribution

statistical properties of the sample data for detecting events [12], [13]. This is considered as a Change Point Detection (CPD) problem in statistics.

We used batch mode CPD for identifying the significant changes in IAAV in which, all the change points are detected at once. Since the distribution of the data is assumed to be unknown, CUSUM [14] is used as test statistic for the CPD. The CUSUM method followed in [14] was used in our work and given in algorithm 1. Using the computed CUSUM test statistic, the Binary Segmentation (BS) and Segment Neighborhood (SN) methods were adapted to search for multiple change points. The BS algorithm [15] uses single CPD method initially on the entire series to detect t (time) satisfying the equation 3 in which C is the cost function of the time series segment $Y_{1:t}$ and β is the penalty function. If the condition is true, then segmentation is performed on the identified t. On the two new segments, the procedure is iterated until no other change point is detected. It is an approximate method with computational cost as $\mathrm{O}(nlogn)$ where n is number of elements present in the time series.

$$C(Y_{1:t}) + C(Y_{t+1:n}) + \beta < C(Y_{1:n}) \tag{3}$$

Algorithm 1 CUSUM Algorithm

$DATA \Leftarrow IAAV$
$CUSUM \Leftarrow 0$
$m = mean(DATA)$
$len = length(DATA)$
$con = 0$
repeat
  $CUSUM_i$ = $CUSUM_i$ + ($DATA_i$ - m), i=1 ...len
  $con = con + 1$
until $con \leq len$

The SN algorithm [16] uses exact search method based on dynamic programming. Number of change points that we like to search for can be specified with parameter Q. It is assumed as the upper limit on the number of segments. The cost function is computed for all possible segments in the entire series. Change points between 0 to Q are considered from all the possible segments. In addition to the cost function, a penalty value can also be incorporated in the search method. As a consequence of exhaustive search, the computational cost is $\mathrm{O}(Qn^2)$ for the algorithm. When the number of change points increases linearly with time, there will be a cubical computational cost increase in the size of the series. The search method also employ a similar approach based on minimizing cost and a penalty function using Eq. 4 to

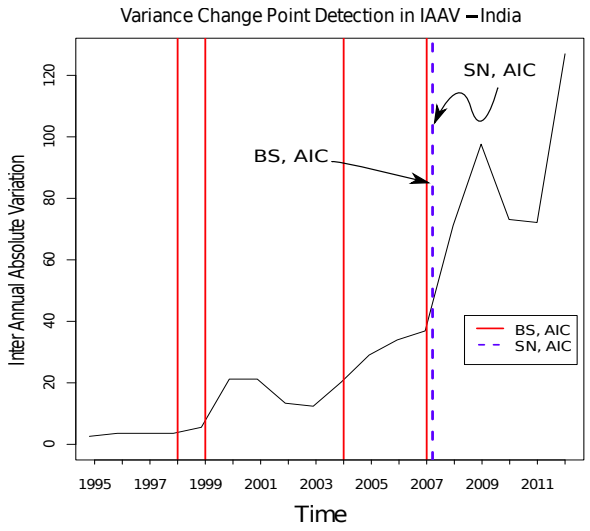

Fig. 13. Change Point Detection Using AIC Penalty

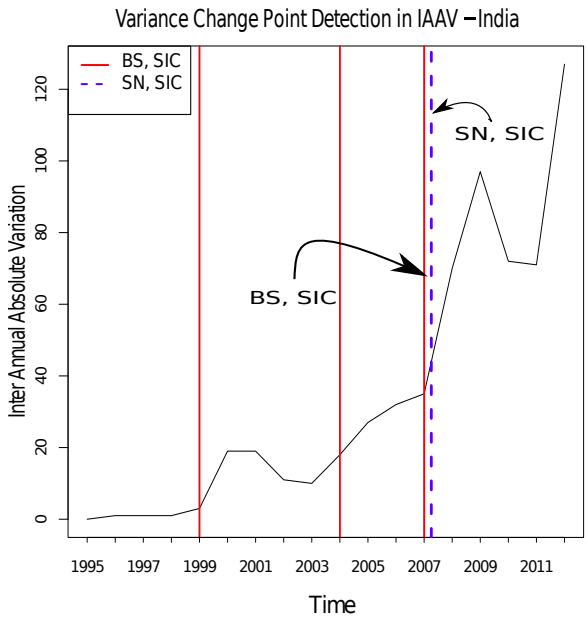

Fig. 14. Change Point Detection Using SIC Penalty

perform the segmentation [17].

$$\sum_{i=1}^{m+1}[C(Y(t_{i-1}+1):t_i)]+\beta f(m) \quad (4)$$

Here C is the cost function of the time series segment $Y(t_{i-1}+1) : t_i$ and $\beta f(m)$ is the penalty function to protect against over fitting. In the change point detection literature [18], twice the negative log likelihood is used commonly as cost function. The penalty function is generally used as linear in the number of change points m (i.e., $\beta f(m) = \beta m$). Akaike Information Criteria (AIC) [19] and Schwarz Information Criterion (S IC) [20] are widely used penalty functions. AIC uses $\beta = 2p$ and SIC uses $\beta = plog(n)$ as penalty values. In this p is the number of additional parameters introduced by adding a change point and n is number of elements in the series. The SN algorithm guarantees the global minimum of Eq. 4, but BS does not provide the guarantee.

We used AIC and SIC as penalty functions and found the variance change point in IAAV. During estimation of CPD, we observed multiple smaller segments due to small penalty value of AIC and local minima of BS. In such conditions, we consider the change point commonly detected by both the methods with SIC as penalty function. CPD in IAAV level change for India is given in Figs. 13 and 14.

We have observed that the BS method detects four change points in the variance level and SN detects one change point when AIC is used as penalty function. When SIC is used as a penalty function, BS detects three change points and SN detects one change point. The change point detected in the year 2007 for the level variation is common to both methods. The graphical analysis of the IAAV graph also confirms this level change in variation. Hence, we can conclude that there is a significant level change in variation occurred during 2007 in India. We observed significant level change in IAAV variation during 2004 in China with both search methods. But for the IAAV series of Japan, there is no level change observed in variance with either method. In the case of South Korea, both the methods detected three change points when AIC is used as a penalty function, but no change point is detected when SIC is used as a penalty function. Since the segment size performed by both the search methods is small when AIC is used, we consider the result of SIC as a penalty function. In the case of Taiwan, SN and BS detect the change point in 2010 for both of the penalty functions. The whole APNIC region has a change point for the IAAV variations level in 2006. This change point for the APNIC region is detected by using SIC as a penalty function. The change point year, and the estimated yearly growth rate of IAAV before and after the change point are given in Table IX.

We can infer that the yearly change in variation has a positive rate of 2.3 % for India after 2007. The variation is constant for countries in technologically advanced countries from 1995 to 2012, except for an outlier detected in Taiwan during 2010. After 2004, the variation is also constant for China. India and the APNIC region have a significant percentage of change in IAAV after the CPD. We analyzed the macro economic variable service sector growth, which is an aggregate measure of factors that influence the IAAV to understand the significant change in variations for India. The GDP was used as a proxy variable for service sector growth.

B. IAAV Events and GDP

We observed significant changes in IAAV after 2007. To understand the driving force for these variations we analyzed the GDP of India. GDP is a measure of yearly output of products and services of a country. Indian GDP is a composition of outputs from agriculture, industry, and services sectors. After the post economic reform period (from 1991 onward) GDP is mainly driven by the services sector. The Table X shows the contribution of the services sector to the GDP during these years. The data is taken from an economic survey [21] and various publications of the Reserve Bank of India (RBI). The percentage share of the services sector is steadily increasing from 1991 onwards and more than 64 % during 2008-09. Hence, we use GDP as a proxy variable for services sector growth.

The annual growth rate of GDP is taken from an economic survey and given in Table XI. During the years of 2006-

TABLE IX
CHANGE POINT DETECTION DETAILS FOR IAAV

Country/Region	Change Point	IAAV Growth rate (%) (before, after) cpd
APNIC	2006	0.33,1.54
India	2007	0.47,2.3
China	2004	0.34,-0.05
Japan	No change	-0.08
SKorea	No change	-.03
Taiwan	2010	.02,-32.4

TABLE X
THE COMPONENTS PERCENTAGE SHARE TO GDP

Year	Agriculture	Industry	Services
1990-91	31.4	19.8	48.8
1995-96	27.3	21.2	51.4
2000-01	23.9	20.4	56.1
2005-06	19.5	19.4	61.1
2008-09	17.0	18.5	64.5
2012-2013	14.1	21.1	64.8

TABLE XI
GDP DURING 1992 -2010

Year	Annual GDP Growth Rate
1992-93	5.4
1993-94	5.7
1994-95	6.4
1995-96	7.3
1996-97	8.0
1997-98	4.3
1998-99	6.7
1999-00	6.4
2000-01	4.4
2001-02	5.8
2002-03	3.8
2003-04	8.5
2004-05	7.5
2005-06	9.5
2006-07	9.7
2007-08	9.0
2008-09	6.7
2009-10	7.2

2008, the GDP witnessed a growth rate of more than 9 % driven by contributions from services sectors, such as telecommunication, computer software, railways, and education. IAAV change point significantly correlates with this time period. This correlation establishes that services sector industry growth significantly influences variations in IAAV.

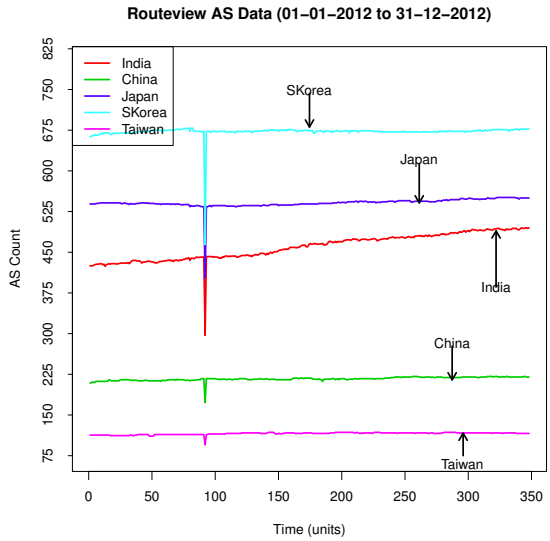

Fig. 15. AS Time Series Data Obtained from Routeview

VI. AS ROUTEVIEW DATA ANALYSIS

To understand the temporal presence of assigned ASes of each country in the global routing table, we have analyzed the routeview data [22] from 01-01-2012 to 01-01-2013. The sample snapshot was taken at fixed intervals (0000 hours) of the day. We consider the per-day AS count data for the analysis. This AS count was obtained by searching for registered AS numbers in APNIC registry of each country in the routeview daily snapshot. The presence of ASNs in the routing table can be interpreted as currently advertising ASes from the assigned ASes of a country. These ASes can be reached from one or more geographic regions using the E-BGP paths. At least one prefix of the AS is being advertised by the AS. The AS may be originating AS, transit AS or both. Events such as link failures, node failure, and attacks will cause reachability problems to the ASes. The impact of such events would be manifested in the form of stochastic variations in AS counts measured from the global routing table. The obtained AS data for the countries under consideration are given in Fig. 15.

We can observe a clear linear growth of AS count in the routeview table during 2012-2013 for India. The increase is estimated as 16 %. for this period. For the same period, the rest of the countries have growth below 6 %. A significant stochastic variation is also observed on April 17^{th}, 2012.

TABLE XII
ASSIGNED VS ADVERTISED AS COUNT 2013

Country/Region	Registered	Assigned	Advertised	Ratio	Increase % (2012-13)
APNIC	9876	8420	5285	.6	-
India	614	607	495	.8	16.4
China	729	551	220	0.4	5.2
Japan	993	800	550	0.7	2.04
SKorea	1016	857	677	0.8	2.11
Taiwan	308	196	116	0.6	2.65

During this time period, 35 % of ASes are not reachable for India, Japan, and South Korea whereas 25 % of ASes are not reachable for China and Taiwan. This may be due to a failure in the common link that is used to reach ASes of these countries, and reachability is restored within a day. The advertised versus assigned ratio was computed using the routeview and APNIC assigned AS data. When we consider the whole APNIC region, only 60 % of the assigned ASes are advertising in the global routing table. This can be interpreted as only 60 % of the assigned ASes are reachable from different geographic locations. For China, 40 % of the ASes can only be reachable from global sources. The ratios indicate that 20 to 60 % of the assigned ASes are not reachable from outside world to the countries in the APNIC region. This may be due to the reasons that the AS may not be operational or used only in I-BGP. The advertised versus assigned AS details are given in Table XII.

The analysis of AS data in the routeview table shows the highest increase of 16 % in the AS global reachability for India during the period of 2012-13. In the assigned ASes, 80 % can be reached from various global locations. The stochastic variations of short durations in AS count time series are more likely in the routeview data due to the occurrence of various external events, such as attacks and under sea cable cuts. The study on these stochastic variations and inferring events that caused variations are our future work.

VII. RELATED WORK

So far in the Internet AS topology research domain, structural analysis on the AS topology [23], topology generation methods [24] and impact of routing dynamics on AS topology [25], [26] are studied. Different topological features, such as AS node count, average node degree, the node degree relationship with node count, and different centrality measures are reported in AS structural studies. Prefix counts, path distributions, average path lengths, and peer counts are reported in AS routing dynamic studies. Web sites like Potaroo.net [5] and hurricane electric [27] provides daily reports on ASes, prefixes, peers, routing table size, withdrawn, newly announced, and bogus routes. Weekly, monthly, and yearly summaries are also provided by these sites on the aforesaid features for each country studied. Still, understanding on the temporal occurrence of events specific to a country, impact of the events on the overall trend, and yearly variances of the AS topology are to be explored further. In this work, we have chosen the AS node count to explore its growth and variations due to events internal and external to a country.

The growth trends in the number of ASes seen in the global routing system and the RIRs are studied by Dhamdhere et al.[8], [9]. The key observation they have made in their studies is that ARIN and RIPE RIRs have shown distinctly different growth trends since 2001 in terms of the number of advertised ASes. Until mid 2001, both RIRs showed exponential growth trend. After that, ARIN has grown linearly and RIPE changed to a slower but exponential increase. The number of advertised ASes is larger in RIPE than in ARIN.

The AS number resource consumption is extensively studied, and prediction for the pool exhaustion has been performed in the work [3]. In the unstructured 16-bit ASNs, from 1 to 64511 ASNs are available for global Internet routing excluding the reserved numbers such as 0 and 65,535 as well as private pool from 64512 through 65534. The ASN consumption per year in the global pool is reported as 3500 from 2002 onwards which is roughly 5.4 % per year. The RIR pool size during 2006 along with advertised and assigned ASes are also reported in [3]. The prediction models used are exponential and linear. In the exponential model, recent past three year values are considered for future predictions.

In our work, we have considered ARIMA models that use past values and errors in the forecasting. The statistical properties (i.e., long-term trend and IAAV), for technologically advanced countries and larger economies were analyzed. We also used CPD methods on IAAV to detect significant changes in ASN growth within a country. Our work is focussed towards AS resource planning, policy making, and growth anomaly detection with respect to a country.

VIII. CONCLUSION

In this work, we have analyzed the AS resource data available in the APNIC repository for five Asian countries and the APNIC region. The countries were chosen based on two categories, namely, fast growing economies and technologically advanced countries in the APNIC region. The characterization on the AS count time series data was performed to identify the appropriate time series model. The estimated autocorrelation properties up to lag 3 and partial autocorrelation properties upto lag 1 indicate the presence of AR and MA components. The location change in the data indicates a linear trend. Based on these characterizations, ARIMA models were chosen to forecast the data. The model validation was performed by analyzing the

residuals for randomness, constant variations, and normal distribution. From the model candidates, forecasting accuracy and correctness were used as important criteria in final model selection. An out of sample forecasting with point forecasts, prediction intervals, and prediction accuracy are reported for the selected model of each country. The long term-trend is analyzed using ARIMA(0,1,0) and linear trend models. The average yearly growth rate and the trend direction are reported. The hypothesis test establishes significant structural deviations in the long-term trend of India and the technologically advanced countries. We analyzed IAAV data for change in variations with the CUSUM test statistic using BS and SN search methods. Two penalty functions, AIC and BIC, were used, along with the search methods. The level change within variations and annual growth rates are reported. The change in IAAV at a rate of 2.3 % after 2007 is hypothesized as driven by Indian services sector growth. This is established using evidence from GDP data assumed as proxy variable for the services sector. The technologically advanced countries witnessed a constant IAAV value during the observation period. The global reachability percentage during 2013 with respect to assigned ASes and the annual growth percentages are reported. The significant level change in variations, positive growth percentage in IAAV, and higher percentage of advertised ASes when compared to other countries indicate India's fast growth and wider global reachability of Internet infrastructure from 2007 onwards. Our modeling effort reveal new insights and patterns in the country level Internet infrastructure indicator:AS count, for the countries that fall under two groups. The stochastic variation analysis in the global AS reachability data and event detection are considered for our future work.

ACKNOWLEDGMENT

The authors would like to thank UmaDevi, Jeanie and Professor Vijayalakshmi for their review inputs and suggestions.

REFERENCES

[1] K. Brown, "Internet society global internet report 2014," Internet Society, 1775 Wiehle Avenue, USA, Tech. Rep., 2014.
[2] S. Paltridge, "Internet infrastructure indicators," Directorate for Science, Technology and Industry, Committee for Information, Computer and Communications Policy, OECD, Working Party on Telecommunication and Information Services Policies, 1998.
[3] G. Huston, "Exploring autonomous system numbers," *The Internet Protocol Journal*, vol. 9, pp. 2–23, 2006.
[4] X. Xu, Z. M. Mao, and J. A. Halderman, "Internet censorship in china: Where does the filtering occur?" in *PAM*, 2011, pp. 133–142.
[5] G. Huston, "The 32-bit as number report," 2011. [Online]. Available: http://www.potaroo.net/tools/asn32/index.html
[6] APNIC, "Asia pacific network information center," 1992. [Online]. Available: ftp://ftp.apnic.net/pub/stats/apnic/delegated-apnic-extended-latest
[7] X. Dimitropoulos, D. Krioukov, G. Riley, and k. claffy, "Classifying the types of autonomous systems in the internet," in *SIGCOMM 2005 Poster*. Philadelphia, Pennsylvania: ACM, Aug 2005.
[8] A. Dhamdhere and C. Dovrolis, "Ten years in the evolution of the internet ecosystem," in *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, ser. IMC '08, 2008, pp. 183–196.
[9] C. R. Resources, "An analysis of advertised autonomous system (as) growth trends in different regional registries," 2010. [Online]. Available: http://www.caida.org/research/routing/as_growth/
[10] J. J. Faraway, *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, 2006.
[11] C. Ghate, R. Pandey, and I. Patnaik, "Has india emerged? business cycle stylized facts from a transitioning economy," *Structural Change and Economic Dynamics*, vol. 24, no. C, pp. 157–172, 2013.
[12] V. Guralnik and J. Srivastava, "Event detection from time series data," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '99, 1999, pp. 33–42.
[13] C. Beaulieu, J. Chen, and J. L. Sarmiento, "Change-point analysis as a tool to detect abrupt climate variations," *Philosophical Transactions of the Royal Society of London*, vol. 370, no. 1962, pp. 1228–1249, 2012.
[14] W. A. Taylor, "Change-point analysis: A powerful new tool for detecting changes," 2000.
[15] A. J. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics*, vol. 30, no. 3, pp. 507–512, 1974.
[16] I. E. Auger and C. E. Lawrence, "Algorithms for the optimal identification of segment neighborhoods," *Bulletin of Mathematical Biology*, vol. 51, no. 1, pp. 39–54, 1989.
[17] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.
[18] J. Chen and A. K. Gupta, "Testing and locating variance changepoints with application to stock prices," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 739 – 747, 1997.
[19] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on automatic control*, vol. 19, pp. 716–723, 1974.
[20] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.
[21] U. Budget, "Economic survey 2012-2013," March 2012. [Online]. Available: http://indiabudget.nic.in/survey.asp
[22] Routeviews, "University of oregon route views project," 2003. [Online]. Available: http://archive.routeviews.org/oix-route-views/
[23] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 4, pp. 251–262, Aug. 1999.
[24] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Network topology generators: Degreebased vs. structural," in *ACM SIGCOMM 2002*, 2002.
[25] D. G. Andersen, N. Feamster, and H. Balakrishnan, "Topology Inference from BGP Routing Dynamics," in *2nd ACM SIGCOMM Internet Measurement Workshop*, Boston, MA, November 2002.
[26] Y. Zhang, Z. M. Mao, and J. Wang, "A framework for measuring and predicting the impact of routing changes," in *INFOCOM*, 2007, pp. 339–347.
[27] H. Electric, "Internet services," 2013. [Online]. Available: http://bgp.he.net/

S P Meenakshi is pursuing her Ph.D. in the Deparetment of Computer Science and Engineering at IIT Madras, Chennai, India. She earned her MS in Computer Science from IIT Madras in July 2007 and received her Bachelor of Engineering degree in Computer Science from Madurai Kamaraj University, India in 1995. She has been working as Senior Assistant Professor in the School of Computing Science and Engineering at VIT University, Vellore, India, since December 2013. Her area of research includes Network Management and Security, Pattern Recognition, and Machine Learning. She is a Member of IAENG from 2014.

S V Raghavan received his B.Sc. in Physics from the University of Madras in 1971, D.M.I.T. in Electronic Engineering from Madras Institute of Technology in 1974, M.E. in Automation from the Indian Institute of Science in 1976, and Ph.D. from Indian Institute of Technology Madras, Chennai, India in 1985. He is a professor in the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai. He is currently serving as the Scientific Secretary in the Office of the Principal Scientific Advisor to Government of India and also an Adjunct professor in the Department of Computer Science in IIT Delhi. He has been recognized and awarded at National and Global level for his excellence and contribution towards the Country's growth and security. His research and development interests are in Large Scale Network Design and Cyber Security.