Forecasting of Traffic Load in a Live 3G Packet Switched Core Network

Philipp Svoboda, Manfred Buerger, Markus Rupp
Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology, Austria
Gusshausstrasse 25/389, A-1040 Vienna, Austria
Email: {psvoboda, mbuerger, mrupp}@nt.tuwien.ac.at

***Abstract*— In this paper we analyze different methods for long term forecasts of packet switched traffic from live 3G networks. The dataset consists of over 400 values, each representing the peak load for a separate day. Four different methods were applied to forecast the increase in traffic, two simple: linear and exponential regression, and two more sophisticated ARMA and DHR. We will show in which cases the sophisticated models deliver a better performance and discuss the question if the gain is significant to justify the increased complexity. We present numerical results for long, e.g., more than 100 day, and short time,e.g., hourly or daily, fitting for our case study based on real traces from a live network. The paper concludes with a benchmark based on the observed mean absolute error.**

I. Introduction

In this paper we provide a long-term forecast of packet switched traffic for a live 3G network. The forecasting of packet switched traffic is one of the most important tasks in the network planing of a mobile operator, as field deployments of new hardware have a long starting phase. In the recent past the introduction of flat rate billing for packet switched services led to a massive increase of packet switched traffic cf. [1]. As a normal upgrade plan is fixed nine month in advance an unpredicted increase in bandwidth may cause serious problems in form of bottlenecks, which can not be removed before the end of the planing phase. Therefore, operators are interested in long-term forecasts for capacity planing on aggregated links.

The actual planing is often performed by linear or exponential estimations based on historic data collected from different sources. In the past these methods worked quite well. However, forecasts based on these models had a strong deviation compared with actual data collected.

We evaluated the performance of more sophisticated models, namely ARMA (AutoRegressive Moving Average) and DHR (Dynamic Harmonic Regression), compared to the old methods. The goal was to benchmark the models for a long-term forecast based on the prediction error.

The document is structured in the following way. Section II explains the measurement setup and the traces. The data preprocessing is also covered in this section. In Section III we introduce the forecast models DHR and ARMA. Section IV presents the results of the forecasts and benchmarks the different methods. Finally we present related work and a summary and conclusions.

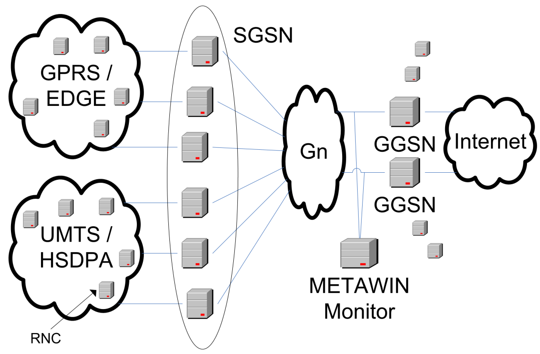

Fig. 1. Measurement Setup

II. Measurement Setup

The reference network scenario is depicted in Fig. 1. As most access networks, the 3G mobile network has a hierarchical tree-like deployment.

The mobile stations and base stations are geographically distributed. Going up in the hierarchy (first BSC/RNC, then SGSN, ultimately GGSN) the level of concentration increases, involving a progressively smaller number of equipments and physical sites. In a typical network there are relatively few SGSNs and even fewer GGSNs. Therefore it is possible to capture the whole data traffic from home subscribers on a small number of Gn/Gi links. For further details on the structure of a 3G mobile network referee to [2].

A. METAWIN

The input traces were captured on a live GPRS/UMTS network at the Gi interface by the METAWIN monitoring system [1]. It is a monitoring tool designed to record traffic in a mobile core network. Although the underlying protocols and interfaces are similar to normal core networks, the presence of user mobility introduces intermediate protocols between the transport network and the user data (see [2]). Therefore METAWIN has to accomplish two main tasks: decoding the additional protocols and tracking the individual user sessions. This preprocessing allows to perform research on live traces.

[1]More Information on the METAWIN project can be found here: `http://www.ftw.at/ftw/research/projects/`

Window Size	Mean Error	Max Error
2 min	0.8%	2.9%
4 min	1.5%	4.9%
6 min	2.2%	5.6%
15 min	3.1%	8.0%

TABLE I
ERROR INTRODUCED BY AVERAGING

To meet privacy requirements traces are anonymized by hashing all fields related to user identity at the lower 3G layers (e.g. IMSI, MSISDN), while the user payload above the TCP/IP layer is removed.

B. Nexus

The Nexus Reporting Suite (NRS) provided by Nexus Telecom is a monitoring tool for GSM, GPRS and UMTS network components. It is designed to monitor usage trends and traffic volumes within a multi-vendor core network. In order to collect this data internal counters of the networks MSCs and other entities are collected and stored in a data base. Based on this we extracted a time series for the 2G packet switched traffic for 485 days. A draw back to this method is the low granularity of the provided data sets. In the 2G case only the peak load based on a 15 minutes average is stored in the data base. In contrast to this the METAWIN data is stored with a granularity of one minute.

C. Collected Traces and Preprocessing

The data samples presented in this paper were taken throughout the years 2006 and 2007. The time series contains 485 consecutive days. Due to our NDA we are not allowed to publish absolute values nor to disclose the exact time the traces were recorded. Therefore, we divided all traces by an arbitrary factor, which will change the total values but leave the shape unchanged.

Each point of the time series represents the peak load of one day averaged in a 15 minutes time-bin. The averaging leads to an underestimation of the real peak in bandwidth. Therefore, we extracted a time series for 45 days on a 1sec time-bin and compared the average value for different time windows. Table II-C presents the numbers for different time bins. The smoothing introduces an underestimation of 3.1%, averaged for all days, to the actual peak value. The maximum error occurred was 8.0%.

As our traces were recorded in a live system they included several time-bins where either wrong or no data was recorded. We filled this missing information by the similar day method [3]. In this approach an average day is generated using historical data. Missing data points are replaced by the corresponding data points taken from the average day weighted with an actual trend.

III. FORECASTING METHODS

Four different methods were used to forecast the traffic, namely: linear, exponential, ARMA and, DHR. In the first two cases we fitted the parameter using a least squares fitting. We included these simple functions because we were interested in the question how much we can gain if we apply more sophisticated models.

A. ARMA

The time series of data throughput within a cellular core network exhibits strong periodic patterns. We observed daily fluctuations due to the fact that there is low traffic in the early morning and high traffic with peaks on the evenings. Furthermore, this daily cycle is superimposed by a weekly cycle where the highest throughput values tend to be in the middle of the week whilst the weekends are showing lower peak values.

Classical seasonal decomposition is one of the modeling approaches for cyclic time series [4]. There, a time series x_t is assumed to be the sum of separable components,

$$x_t = m_t + c_t + Z_t, \tag{1}$$

where m_t is a slowly changing trend, c_t is a cyclic[2] component and Z_t is a remaining stochastic component which can be further modeled. In particular, for decomposition a method proposed in [3] was used, namely a combination of differencing and moving average-type smoothing. For daily recorded values (and therefore 7 day periodicity) the stochastic component Z_t is given by

$$Z_t = x_t - \left(\frac{1}{N}\sum_{i=1}^{N} x_{t-i\cdot 7} + \frac{1}{7}\sum_{j=1}^{7} x_{t-j} - \frac{1}{7N}\sum_{i=1}^{N}\sum_{j=1}^{7} x_{t-i\cdot 7-j} \right), \tag{2}$$

and similarly for hourly recorded values the daily periodicity have to be considered,

$$Z_t = x_t - \left(\frac{1}{N}\sum_{i=1}^{N} x_{t-i\cdot 168} + \frac{1}{7}\sum_{j=1}^{7} x_{t-j\cdot 24} - \frac{1}{7N}\sum_{i=1}^{N}\sum_{j=1}^{7} x_{t-i\cdot 168-j\cdot 24} \right). \tag{3}$$

If this remaining component Z_t is weakly stationary, it can be modeled using AutoRegressive Moving Average (ARMA) models. An ARMA model of order (p,q) can be written as (cf. [3] p.82-86):

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}, \tag{4}$$

with ϕ_i are the AR coefficients and θ_i are the MA coefficients, $\epsilon_t$are random noise. The order of the model was determined by using Akaike's Information Criterion and Bayesian Information Criterion. For details about information criteria and the parameter estimation we refer to the standard time series analysis literature (e.g. [4] or [5])

B. DHR

Dynamic Harmonic Regression (DHR) is a modeling approach designed from scratch to cope with nonstationary time series. In particular adaptive seasonal adjustment, signal extraction [6] and gap interpolation are inherent features of this Unobserved Components (UC) type model. Detailed insights into the fundamentals of this model class is provided by [7], [8], [9] and in [10] the relation to ARMA models is discussed. Moreover DHR has been applied in forecasting telephone call demand [11] and other related areas (cf. [12]).

[2] In time series literature also called seasonal component, especially in context of annual cycles.

As in Equation (1) the time series is assumed to be the sum of a trend, a cyclical and a remaining stochastic term. The important difference is now that the cyclical component is given by,

$$c_t = \sum_{i=1}^{R_s} a_{i,t} \cos(\omega_i t) + b_{i,t} \sin(\omega_i t), \qquad (5)$$

with $a_{i,t}$ and $b_{i,t}$ are so called stochastic Time Variable Parameters (TVP) and ω_i are the frequencies contained in the time series, revealed by an Autoregressive (AR) Spectrum (cf. [11]). In the case of daily values the fundamental period is 7. The hourly recorded series shows a fundamental period of 24. For the modeling such fundamental frequencies and several harmonics had been used.

The estimation problem for $a_{i,t}$ and $b_{i,t}$ is formulated in a stochastic state space context and then solved by Kalman Filter and Fixed Interval Smoothing algorithms. Detailed information about the estimation process is given in the references cited above.

The trend is extracted by using an Integrated Random Walk (IRW) plus noise modeling approach. The full state space description is given by the state equation (cf. [13] p.17-20),

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \eta_{t-1}, \qquad (6)$$

and the observation equation,

$$y_t = (10) \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} + e_t, \qquad (7)$$

where x_{1t} x_{2t} are state variables. Here, η_t is a zero mean, serially uncorrelated white noise process with a certain variance. The variance and the missing states are estimated by the algorithms stated above. Finally, e_t is zero mean white noise and contains the remaining uncertainty.

IV. Forecasting Results

In this section we present the results of the forecasts archived with ARMA and DHR. All DHR related modeling and forecasting have been performed using the CAPTAIN Toolbox [14] for MATLAB. We used the MAE (Mean Absolute Error) and the MSE (Mean Squared Error). The MAE is the mean of the absolute differences between the real data and the forecast, this metric is often used in electrical load forecasts. The MSE is the mean of the absolute squared differences between the real data and the forecast, this metric puts more weight to larger differences than to smaller. We normalized both metrics to the mean (by mean squared for MSE) of the data series to keep the data confidential. The normalization is different for the two time series, as in the 2G throughput case we only consider peak values the resulting value is significant larger than for the HTTP time series where all data is considered.

A. Results for Forecasting 1 Month

This paragraph presents results for a step-wise prediction of the dataset. We predicted the 2G throughput and HTTP separated for a period of four weeks. Figure 2 depicts a four week forecast of the 2G throughput data with DHR. The frequency components applied in the DHR 2G forecast were: $0; 0.7; 3.5; 2.33; 1.75; 1.4$, the IRW parameters were identified with a maximum likelihood method. The same settings were implemented for the step wise prediction with DHR. The step wise DHR forecast for HTTP was set to the following frequency components: $0; 24; 12; 8.03$, the IRW parameters were estimated in the same way. The four week forecast for HTTP was modeled each day separately to improve the poor results.

We trained the model with 362 days and predicted 28 days. After that we compared the result with the actual real data recorded. The solid blue line represents the forecast, the dashed red lines the 95% convidence intervals. The real data for this timespan, represented by the small circles, lies well inside these intervals. Overall the MAE_{rel} (Mean Absolute Error) is below 2% in this case. We consider this to be a good performance for a 4 week forecast.

Table II shows the results for the 2G throughput data series. We found the DHR model to fit best in the step ahead case. In fact the DHR also outperforms the ARMA in the four week prediction, in this scenario we only applied the differencing smoothing part of the ARMA algorithm.

However, DHR provides better performance in the short and the long term prediction. All results for the 2G throughput have a small prediction error.

Step Ahead	MAE_{rel} %	MSE_{rel} %
ARMA $(2, 1)$	2.21	0.081
ARMA $(1, 0)$	2.10	0.076
DHR (stepwise)	1.79	0.052
4 Weeks	MAE_{rel} %	MSE_{rel} %
diff.-smoo.	2.16	0.068
DHR	1.95	0.057

TABLE II

ARMA AND DHR RESULTS FOR 2G THROUGHPUT

In the HTTP case, which is shown in Table III the error for the long term prediction is larger than in the 2G throughput case. In this scenario the time series has a much higher granularity, one hour, compared to the first scenario, therefore this time series contains a higher number of frequency components, e.g., the daily and the weekly periodicity, than the 2G throughput, which only contains a weekly periodicity. Due to the high errors we conclude that these methods can not be directly applied to predict service mix data on a long time base.

We assume that it is necessary to find a model for each day of the time series instead of one model for the whole data set. The best model for step ahead prediction was an ARMA(2,0) with an error of 4.3% and the differencing smoothing for a four week forecast with an error of 12.25%.

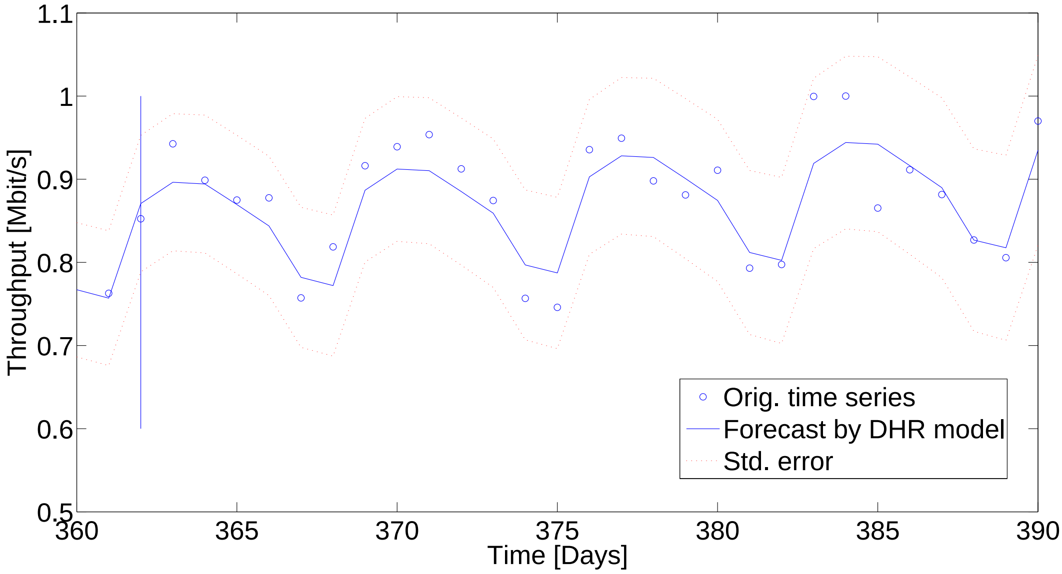

Fig. 2. A DHR four weeks ahead forecast of daily throughput maxima time series.

Step Ahead	MAE$_{rel}$ %	MSE$_{rel}$ %
ARMA $(0,0)$	10.08	2.378
ARMA $(2,0)$	6.03	0.725
DHR	6.79	0.755
4 Weeks	**MAE$_{rel}$ %**	**MSE$_{rel}$ %**
diff.-smo. 4 weeks	17.08	5.610
DHR	19.41	6.455

TABLE III
ARMA AND DHR RESULTS FOR HTTP

B. Results for Forecasting 6 Month

In this paragraph we discuss the results for a longterm trend prediction for different methods. We compare linear, exponential, and, DHR models to predict the traffic increase for 200 days. In case of the DHR model we only apply the trend prediction. Figure 3 shows the result for the long term prediction, x axis is the time in days and y axis represents the renormalized bandwidth and the MAE for the different models. The measured data is drawn in solid blue.

The vertical line at day 200 shows the start of the prediction. The red, magenta and, green line represent the MAE for linear, exponential and DHR (IRW) models starting to predict at day x. First thing we see is the fact that for short time, below 100 days, a forecast of the trend preforms equally for all three methods with an MAE of about 5%.

In the long term case, more than 150 days, the DHR model performs better than the two others. Even more the MAE for the linear and the exponential model start to increase strong in this time span. However, the DHR model does have problems too. In the time span around day 300 the MAE for DHR rises in an abrupt way to 15%. Analyzing the measured data we observe that the local trend of the time series changed twice from day 280 to day 320. In this scenario the DHR followed the first change and produced a very high error.

From these results we conclude that the DHR model is the best to predict the trend on a long term time base. However, it is sensitive to local trend changes and can therefore, produce considerable worse results as seen in our traces. In a time span of one to two month the trend can be equally well described by all three models.

V. Related Work

Forecasting a time-series is often done in engineering. We found related literature for the forecast of the demand for electrical power. Such time-series have a similar profile with daily and weekly periodicity [3]. However, as bandwidth is cheap forecasting in networking often relies on simple methods and/or simple over provisioning. Therefore, to our knowledge, up to now, this is the first paper dealing with the forecast of packet switched traffic in a mobile network.

VI. Summary and Conclusions

This paper presents different methods to predict the load increase in a 3G live network. A precise forecast of traffic loads in the network is an essential task for a network operator, as the upgrade path for network equipment needs two to four month time to be implemented. We used two data sets the first consisted of 485 daily peak throughput values for 2G traffic, the second time series recorded the HTTP traffic on a one hour granularity.

In the modeling phase we applied two different methods, namely DHR and ARMA. The ARMA model was implemented using MATLAB, while in the DHR case the CAPTAIN toolbox was used. We processed three different forecast scenarios: one step, four weeks, six month.

The DHR results for the 2G throughput data series outperformed the ARMA model for step wise (daily) and 4 weeks, the MAE was below 2% in both cases. In the HTTP time

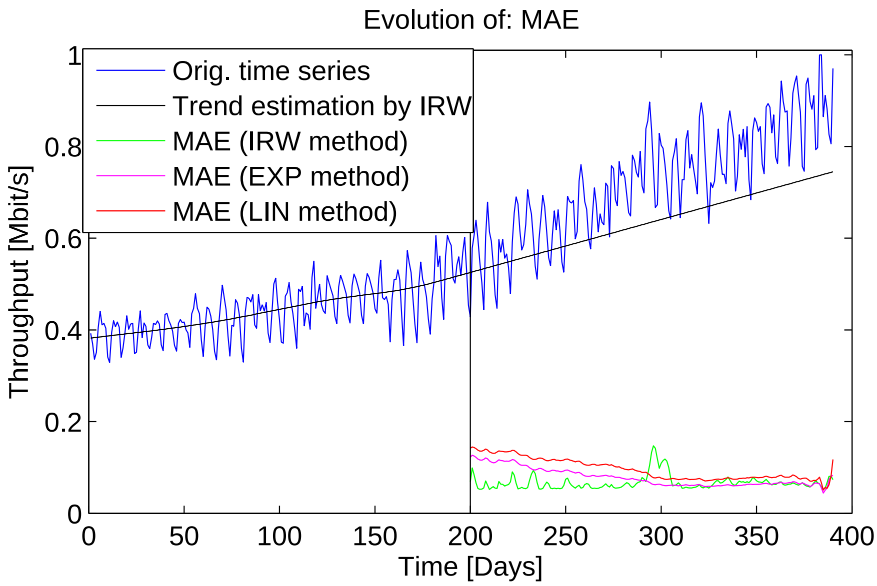

Fig. 3. A trend estimation and extrapolation for up to 200 days with the corresponding MAE.

series the ARMA outperformed the DHR. It had an error of more than 6% in the step wise forecast and even 17% for the 4 week forecast. We assume this high error is caused due to the higher number of periodicities occurring in this data set.

In the long term prediction we benchmarked the DHR against linear and exponential regression. We obtained that the performance of the three methods is nearly equal for forecasts for less than 100 days. However for longer forecasts, including trend changes, the DHR delivers a better MAE performance. In case of two trend changes the DHR produced a significant error in the forecast.

In this paper we showed that the prediction of traffic loads on a highly aggregated link, like the Gi, up to six month in advance is feasible. The DHR performs best for the 2G daily peak throughput time series. However, it has problems with the high variability of the hourly HTTP time series.

Acknowledgements

Special thanks to the ftw, hosting the METAWIN project, which provided us with data sets for this paper.

References

[1] Rundfunk und Telekom Regulierungs-GmbH. Rtr telekom monitor 4. quartal 2007, 2007.
[2] H. Holma and A. Toskala. *WCDMA for UMTS, Radio Access For Third Generation Mobile Communications, Third Edition*. Wiley, 2004.
[3] Rafal Weron. *Modeling and forecasting electricity loads and prices: a statistical approach*. John Wiley & Sons, Ltd., 2006.
[4] P.J. Brockwell and R.A. Davis. *Introduction to time series and forecasting*. Springer-Verlag New York, Inc., 2 edition, 2002.
[5] Rainer Schlittgen. *Zeitreihenanalyse*. Oldenbourg, Muenchen, 1995.
[6] P.C. Young, W. Tych, and D.J. Pedregal. Stochastic unobserved component models for adaptive signal extraction and forecasting. *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, pages 234–243, 31 Aug-2 Sep 1998.
[7] P.C. Young, D.J. Pedregal, and W. Tych. *Dynamic Harmonic Regression*. Journal of Forecasting, Vol. 18 (1999), pp. 369-394, 1999.
[8] C.N. Ng and P.C. Young. Recursive estimation and forecasting of non-stationary time series. *Journal of Forecasting, Vol. 9, 173-204*, 1990.
[9] P.C. Young. Time-variable parameter and trend estimation in non-stationary economic time series. *Journal of Forecasting, Vol. 13, 179-210*, 1994.
[10] M. Bujosa, Antonio Garca Ferrer, and Peter Young. An arma representation of unobserved component models under generalized random walk specifications: New algorithms and examples. 2002.
[11] W. Tych, D. J. Pedregal, P. C. Young, and J. Davies. *An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system*. International Journal of Forecasting, Vol. 18 (2002), pp. 673-695, 2002.
[12] D. J. Pedregal and J. R. Trapero. *Electricity prices forecasting by automatic dynamic harmonic regression models*. Energy Conversion & Management, Vol. 48 (2007), pp. 1710-1719, accepted November 2006.
[13] D.J. Pedregal, C.J. Taylor, and P.C. Young. *System Identification, Time Series Analysis and Forecasting. The Captain Toolbox Handbook v2.0*. Centre for Research on Environmental Systems and Statistics (CRES), 2007.
[14] P.C. Young, C.J. Taylor, W. Tych, and D.J. Pedregal. *The Captain Toolbox*. Centre for Research on Environmental Systems and Statistics (CRES), Lancaster University, UK, 2007.