J Netw Syst Manage (2016) 24:189–221
DOI 10.1007/s10922-015-9348-6

A Multivariate Approach to Predicting Quantity of Failures in Broadband Networks Based on a Recurrent Neural Network

Željko Deljac[1] · Mirko Randić[2] · Gordan Krčelić[1]

Received: 21 March 2014 / Revised: 10 April 2015 / Accepted: 25 April 2015 /
Published online: 5 May 2015

Abstract In this paper, we present a multivariate recurrent neural network model for short-time prediction of the number of failures that are expected to be reported by users of a broadband telecommunication network. An accurate prediction of the expected number of reported failures is becoming increasingly important to service providers. It enables proactive actions and improves the decision-making process, operational network maintenance, and workforce allocation. Our previous studies have shown that the recursive neural network is flexible enough to approximate the dynamics of the failure reporting process. Development of the model is based on long-term monitoring of failure-reporting processes and experience gained through fault management related to the network of one of the leading Croatian telecom providers (T-HT). Many factors, both in the network and outside the network, influence the time series representing failure reporting. The model encompasses the most important predictor variables and their logical and temporal dependencies. Predictor variables represent internal factors such as profiles of past and current quantities of failures as well as external factors like weather forecasts or announced activities (scheduled maintenance) in the network. External factors have a strong effect on fault occurrence, which finally results in failures reported by users. These factors are quantified and included as input variables to our model. The model is fitted to the data from different sources like an error-logging database, a trouble-ticket archive, announced settings logs and a meteo-data archive. The accuracy of

✉ Željko Deljac
zeljko.deljac@t.ht.hr

Mirko Randić
mirko.randic@fer.hr

Gordan Krčelić
gordan.krcelic@t.ht.hr

[1] Technical Functions, T-Hrvatski Telekom (T-HT), Zagreb, Croatia

[2] Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

the model is examined on simulation tests varying the prediction horizons. Assessment of the model's accuracy is made by comparing results obtained by prediction and the actual data. This research represents a real-world case study from telecom operations. The developed prediction model is scalable and adaptable so that other relevant input factors can be added as needed. Hence, the proposed prediction approach based on the model can be efficiently implemented as a functionality in real fault-management processes where a variety of available input data of different volumes exist.

Keywords Proactive fault management · Failure reporting · Failure prediction · Predictor variables · Multivariate model · NARX · Telecommunication network

1 Introduction

Plain Old Telephone Service (POTS) networks were developed over a long period of time during the last century. Due to their long-term development and fault-tolerant design, these systems are reliable and work in a stable manner. Fault tolerance is achieved by duplication of the essential parts of access and transit switches as well as transmission and management systems. Even though an IP (Internet Protocol)-based broadband telecommunication network in Croatia was introduced to the mass market more than 10 years ago, and the quality of the network is continuously improving, it still hasn't reached the level of reliability that the POTS networks have. Therefore, fault detection, diagnosis, and correction are still major concerns for a telecom operator. The parameter that best reflects the quality of a network regarding fault occurrence is the mean time between failures (MTBF). The service complexity, longer average service usage time, many more instances of terminal equipment compared to the POTS network, and a higher bandwidth demand on the access network all have an impact, so the MTBF of broadband services is 2–6 times lower than the MTBF of narrowband services.[1]

A broadband network, as opposed to a POTS network, includes a multitude of active and passive elements that can be subject to fault. Elements most susceptible to faults are as follows: an Asymmetric Digital Subscriber Line (ADSL) modem, customer's equipment, Internet Protocol TeleVision (IPTV) set-top box, ADSL splitter, home installation, copper twisted pair, distribution point, main distribution frame, fibre optic cable, and ADSL Digital Subscriber Line Access Multiplexer (DSLAM) port. Faults, as commonly defined, are problems that can be detected and handled directly. A consequence of a fault, i.e., its manifestation is a discrepancy between some observed value or condition and a true, specified, or theoretically correct value or condition. Faults are usually reported by a surveillance system in the form of alarms. Generally, faults can be caused by the following:

[1] The range of values of this MTBF reduction factor has been published as a result of an internal technical analysis encompassing networks of 16 telecom operators in Western and Central Europe. The range (2–6) is quite large because of considerable differences in the equipment that is installed in the analyzed national networks and because of different efficiencies of the fault-repair systems and processes implemented.

(a) *intrinsic factors* such as age of equipment, life expectancy of equipment, manufacturing defects, or software bugs;

(b) *environmental factors* such as wind, animals, humidity, lightning (electric discharge), ice, or power supply outages;

(c) *human factors* such as misconfigurations, incorrect maintenance (upgrades, replacement of equipment, or addition of new functions), accidents caused by an operator, contractor, or a third party work crew, other effects caused by human activity, etc.

Some faults result in the service delivered deviating from the agreed or specified service that is visible to the outside world. The term failure is used to denote this situation. Failures are commonly defined as follows: a system failure occurs when the service delivered deviates from the specified service, where the service specification is an agreed description of the expected service [1]. Similar definitions can be found in papers of Melliar-Smith and Randell [2], Laprie and Kanoun [3], and Salfner [2–4]. The main point here is that a failure refers to misbehavior that can be observed by the user, which can either be a human or another software or hardware component. Failures can be reported either by a surveillance system or by users. For example, the most common failures that can be reported by users are complete interruption of a service, low downstream bandwidth, inability to access web sites, noise during Voice Over Internet Protocol (VoIP) phone calls, inability to establish a phone call, and problems with IPTV service like error blocks or jerkiness.

There are many other problems that could affect customer service, so for IPTV the following issues are listed Tiling, Ringing, Quantization Noise, Aliasing Effects, Artifacts, Object Retention, Slice Losses, Blurring, and Color Pixelation [5, 6].

In some cases, faults are not recognized immediately from the systemic alarms, but later they become apparent due to failures reported by users. More often, occurrence of a fault is accompanied by one or more alarms while users report failures afterward. Operators can minimize failure occurrence with proper design and preventive maintenance of the network. In order to resolve a failure, the fault that caused it has to be detected and fixed. Failures should be eliminated as soon as possible for the sake of the customers' satisfaction as well as respecting signed Service-Level Agreements (SLA) and rules laid down by the regulatory agencies. They need to be resolved reactively after a user complaint, but it would be better to act preventively and proactively. The enabler for this is failure prediction. Generally, proactivity based on failure prediction increases overall quality of service (QoS) and a customer's perception of the QoS has a major impact on her satisfaction and loyalty.

Two types of failure predictions are considered in the literature: online failure-occurrence prediction and quantity-of-failures prediction. The aim of online prediction is to predict the occurrence of failures during runtime based on system-state monitoring, [4]. This type of prediction can enable proactive action and thus, directly increase customer satisfaction and loyalty at the individual level. On the other hand, accurate prediction of the expected number of failures (quantity of failures) that will be reported by customers of a broadband network is becoming increasingly important to service providers. Information about the number of failures that could appear in the future gives enough time for preparation and proactive action. Long-term predictions allow providers to anticipate future operating expenses enabling more confident strategic decisions for planning investments and increasing business efficiency. Short-term predictions improve the decision-making process, operational network maintenance and workforce allocation. Temporal tracking and predicting the number of failures that will be reported can be achieved by observing reported failures as a time series. A time series that describe failure-reporting processes are influenced by a large number of random factors, which make them a complex, nonlinear, time-series. Moreover, imperfections of surveillance systems used for monitoring have an impact. Errors during data collection, ambiguous interpretation of alarms, or complete absence of alarms in specific cases introduce noise in a time series. Also, there are many relevant but unobserved variables whose values are not known at all.

The contributions of this paper are as follows:

On the basis of a sample that covers fault-management data collected during a four-year period, we have done analysis of the main causes of faults related to broadband networks. We present locations in the network, i.e. network elements where faults appear statistically more frequently, as well as causes of these faults. Among all causes, we identify those whose occurrences can be predicted for a period of time. Through the analysis, we gain insight into the variables that could be selected as candidates for predictor variables for our model.

By using real operational data, we explain the characteristics of a time series that represent the number of reported failures. Also, we explain how fault–occurrence processes, the users' average daily usage of services, and the users' expected behavior in the moment when they become aware of service failure, impact the dynamics of failure reporting. By applying a formal statistical method, elements such as outages, lightning, rainfall, and announced work on the network are identified as the most significant predictable causes of faults. To the best of our knowledge, temporal analysis of the interdependence between meteorological factors like humidity (rain) and failures noticed and reported by customers as presented in this paper has not yet been studied. Furthermore, the significance of the historical data on the number of reported failures (e.g., weekly averages or reports in the last few days) for prediction accuracy is noticed by examination of the time series. These data yield information about seasonality and information about the current buffer of the unresolved failures and their impact on the future sequence of reporting failures.

Considering the complexity of the dynamics of failure reporting and the results of our previously-published comparative analysis of various predicting methods, we propose the nonlinear autoregressive network with exogenous input (NARX) as the most promising one for predicting quantities of reported failures in the context of complex systems, such as broadband networks. We believe that due to the inherent adaptability of the proposed multivariate predicting approach based on NARX, it can be used and implemented in fault-management processes by other telecom operators. A system that is able to predict the expected number of reported failures several days in advance enables operators to plan and allocate necessary resources and can considerably decrease operational costs.

This paper is organized as follows. In the introductory Section we state the objectives and motivation for the work and provide an adequate background with minimum details. Related work on failure predictions and multivariate predictions that has generally influenced this research is presented in Sect. 2. Locations and causes of faults and failures in a broadband network are explained in Sect. 3. Characteristics of the time series representing the dynamics of the failure reporting process are also presented. A multivariate model for short-time prediction of reported failure quantities based on the recursive neural network, together with an approach to determination of significant predictor variables, is described in Sect. 4. Accuracy of the model and predictions are presented in relation to variable prediction horizons and the number of input variables. Finally, conclusions are drawn at the end of the paper.

2 Related Work

There are various systems developed for real-world networks, SHRINK [7], NetworkMD [8], Draco [9] and so forth, whose purpose is to enable proactive action based on network analysis and network diagnosis. Improvement of performance management and network reliability in similar types of networks are analyzed in papers [10–12].

Generally, proactivity assumes existence of data and knowledge about processes as well as efficient intelligent methods for data analysis, learning, and predictions. Selection of the optimal prediction method depends on the nature of the processes being modeled, data availability, and the duration of the monitoring period, as well as on adaptability of involved operational support systems. In order to improve the accuracy of prediction models, research is conducted in two directions. First, there are efforts to improve performance of the existing prediction methods, e.g., to develop a new training method for a neural network or to propose a new network topology, and second, researchers are developing their own predictive models customized to a specific application [13, 14]. An interesting example of multivariate forecasting is presented in [13], where the authors develop their own predictive model to forecast the overall sales of retail products. Their model consists of three modules: Data Preparation and Pre-processing (DPP), Harmony search Wrapper-based Variable Selection (HWVS), which prunes redundant and irrelevant variables and selects out the optimal input variable subset, and Multivariate Intelligent Forecaster (MIF) used to establish the relationship among variables and forecast the sales' volumes. The proposed model has proved to be effective in handling multivariate forecasting problems. Similar principles for using three-stage predictive models (preprocessing-selecting-forecasting) were used extensively in the field of forecasting; this idea is used in our work as well. Variation and evaluation of different configurations of neural networks are frequently encountered for predictive purposes. In [14], the authors compare different types of neural networks with their own method, Multiple Forward Stepwise Logistic Regression (MFSLR), to predict customer purchase behavior. Their conclusion is that the MFSLR method shows superior performance.

An interesting application of a recurrent neural network on real world telecommunications data is shown in paper [15] where the authors describe a forecaster for telecommunications call-volume forecasting. The forecaster is a recursive neural network called the Block-diagonal Recurrent Neural Network (BDRNN) with a simpler structure than fully recurrent networks usually have. BDRNN forecasting results are compared with a series of well-established forecasting models such as the traditional forecasting methods—the Naive Forecast, Linear Extrapolation with Seasonal Adjustment (LESA), Simple Exponential Smoothing (SES), and the Seasonal Auto Regressive Integrated Moving Average method (SARIMA), methods that use computational intelligence like the Orthogonal Least Squares-based Fuzzy Forecasting Model (OLS-FFM), Locally Recurrent Neurofuzzy Forecasting System (LR-NFFS), and Recurrent Neurofuzzy Forecaster (ReNNFOR). Results show that the BDRNN gives much more precise forecasts in comparison with the other models. The data set covers a period of 10 years and the model can be applied for telecom-bill optimization and unnecessary-cost reduction in large organizations.

The impact of currently existing element outages and announced work on the network that can potentially cause new outages on failure occurrences is the subject of a lot of research, e.g., [16–18]. Impact of lightning on faults in the network and suggestions for mitigation of their effects were also a subject of research [19]. Similarly, a research on how severe weather (e.g., thunderstorms) can cause loss of Internet connection for residential users is presented in [20]. To determine the effect of weather conditions on connectivity, the authors designed and implemented a measurement tool called ThunderPing. This tool is able to detect the loss of connectivity by pinging residential Internet hosts. An example of an online failure-occurrence prediction system is NEVERMIND, [21]. The system is composed of two components: a trouble-ticket predictor and a trouble locator. Both components infer future tickets and trouble locations based on existing sparse line measurements, and the inference models are constructed automatically using supervised machine learning techniques. The ticket predictor utilizes existing measurements on individual DSL lines to proactively identify lines on which customers are likely to register problem tickets in the near future.

3 Failures in a Broadband Network

Failure reporting is a complex process influenced by numerous environmental and technical factors. Service failures arise due to the occurrence of faults in a network and the lack of redundancy in its design. In this Section, we explain the structure of a broadband network, locations in the network where faults appear, types of faults, and root causes of these faults. Understanding the root causes helps in selecting variables that have significant impact on the failure reporting process.

3.1 The Process of Failures

We distinguish two random processes related to faults and failures: a process representing the occurrence of faults caused by the many factors previously mentioned and a process that represents failure reporting influenced by fault occurrences and users' habits, i.e. their behavior while perceiving and reporting failures. Users' failure-reporting behavior is complex. The moment of reporting the failure depends primarily on the following:

> The users' average daily usage of services—whether the user is using the service at the time or shortly after the fault occurs, (e.g., after midnight service usage is minimal so almost no reporting exists),
> The users' expected actions/behavior (active or passive/indifference) in the moment when they become aware of a service failure (if the user knows a reporting procedure, whether the user is trying to fix the problem alone, has a habit of calling the call center or a habit of passively waiting until the service starts to work again, etc.).

Both processes directly drive the execution of the failure-handling process implemented by the operator. Processes together with the factors that influence them are depicted on Fig. 1.

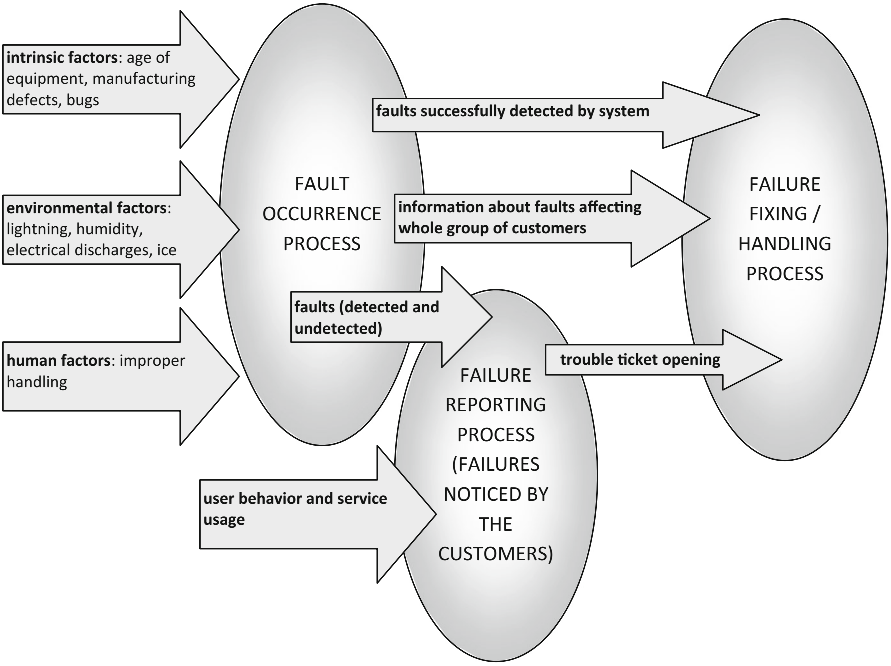

intrinsic factors: age of equipment, manufacturing defects, bugs
environmental factors: lightning, humidity, electrical discharges, ice
human factors: improper handling
FAULT OCCURRENCE PROCESS
faults successfully detected by system
information about faults affecting whole group of customers
faults (detected and undetected)
FAILURE FIXING / HANDLING PROCESS
trouble ticket opening
FAILURE REPORTING PROCESS (FAILURES NOTICED BY THE CUSTOMERS)
user behavior and service usage

Fig. 1 Fault occurrence and failure reporting processes

In this paper, we only consider failures caused by detected faults or failures reported by users. Other phenomena such as transient failures, unreported failures, or undetected faults, which were observed neither by the user nor by a fault-management system, represent additional problems and are not the subject of this paper.

3.2 Locations and Types of Faults

A broadband network contains a multitude of hardware and software components in different locations that can be subject to fault more or less frequently depending on their MTBF. A typical broadband network shown in Fig. 2 consists of three main parts:

1. The Internet Protocol/Multiprotocol Label Switching (IP/MPLS) core part (1) is based on MPLS technology for switching labels. There are two types of routers in the core part: the Label Switch Router (LSR) and Label Edge Router (LER). The core part also includes head-end servers that provide services to users such as: Internet access, access to video services, IPTV and Video On Demand (VOD), and telephone (Voice) over IP (VoIP).
2. The second part of the network is the access part (2) where DSLAM architecture is used as the link to the Ethernet aggregation. Subscriber traffic is directed over an ADSL port to the DSLAM, which does the conversion of Asynchronous Transfer Mode based DSL traffic into Ethernet frames. Ethernet

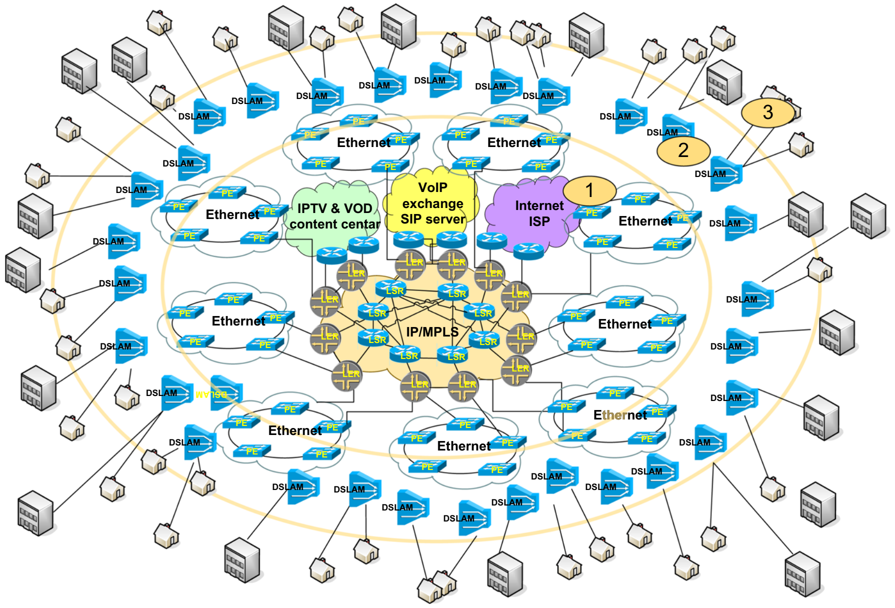


Fig. 2 Broadband network—main parts

traffic flows into an aggregation card and is transported to the network through Gigabit Ethernet rings (Metro Ethernet transport). Cables with twisted copper pairs that form a part of the access network were inherited from the Public Switched Telephone Network (PSTN). Broadband technology has imposed much higher requirements on this part because the speed, and therefore the spectrum, have dramatically increased. These have initiated new problems. The physical link between user and the DSLAM port is a twisted copper pair in a subscriber cable. Access via the copper pair is, at the moment, the most common kind of access to the network. In each area of central or remote subscriber access there is a main distribution frame (MDF), which marks the beginning of the subscriber lines. The end point of the access part is a Distribution point behind which the customer installation begins.

3. The third, user part (3) includes network termination equipment (ADSL modem, Splitter), other customer premises equipment (IPTV set-top box, television set, handset and other devices) and in-house customer installations. This part of the network is spatially the most abundant.

A variety of elements in all three parts are possible locations of faults. The fault-management system is designed to record all faults detected and failures resolved.

The data about alarms are entered into the database automatically while other data are entered by technicians during the resolving process. The result is that the database gives an accurate insight into the faults, causes of faults, and failures that have been reported. By analyzing operational data on faults and their locations we get the distribution of faults displayed in Table 1.

The majority of faults, 70.86 %, occur in the customer part of the network. Of these, 34.55 % relate to the user equipment, 14.36 % to the ADSL modem (router) and 12.36 % to the in-house customer installation. In the access part we find 26.53 % of the faults, while the rest, or 2.61 %, occur in the core part of the network. Over the years, causes have been recorded for each fault detected. Their frequencies are shown in Table 2. The majority of failures that occur in the users part of the network (34.89 %) are caused by users themselves (by improper handling and wrong initial settings) or because of errors in CPE software.

The distribution of faults by locations together with the preponderance of faults in the user part of the network have an influence on the shape of the time series of reported failures. In fact, the dynamics of service usage by customers introduces periodicity and seasonality in the time series. Slight discrepancies between percentages related to total quantities of faults per parts of the network (Tables 1, 2) arise from the measurement noise already mentioned in the Introduction. Table 3 shows an overview of the most frequent causes of faults recorded during the period 2010–2012.

Table 1 Locations of faults

Fault location	Fault equipment	Frequency (%)	Total (%)
Customer Premises Equipment (CPE)	ADSL modem	14.36	70.86
	Customer equipment	34.55	
	Set top box	6.16	
	ADSL splitter	3.43	
	Customer house installation	12.36	
Access network	Cooper twisted pair	3.22	26.53
	Network termination point	6.53	
	Main distribution frame	4.24	
	Optical cable	2.82	
	ADSL DSLAM port	6.44	
	DSLAM	3.28	
Core network	Internet service provider	0.76	2.61
	Core network	0.19	
	Ethernet aggregation	0.73	
	IPTV content center	0.93	

Source: T-HT internal report, 2012

Table 2 Causes of faults

Fault location	Fault reason	Frequency (%)	Total (%)
Customer Premises Equipment	Misconfiguration	8.31	71.26
	Improper handling	34.89	
	In-house installation fault	11.93	
	Electrical discharge	7.32	
	Worn-out equipment	8.81	
Access network	Corrosion	1.22	26.25
	Breakdown	6.53	
	Hardware defect	11.24	
	Electrical discharge	3.82	
	Over-threshold attenuation	3.44	
Core network	Misconfiguration	0.33	2.49
	Incorrect wiring	0.07	
	Hardware defect	0.57	
	Failed upgrade	0.59	
	Low-grade content	0.93	

Source: T-HT internal report, 2012

3.3 Number of Reported Failures: Time Series

The diagram in Fig. 3 shows annual quantities of reported failures by services in the last 10 years. During this period the transition from the POTS network to the

Table 3 Most frequent causes of faults

Fault causes	2010	2011	2012	%
Fault due to deterioration and oxidation	173,010	151,699	148,564	26.95
Defective equipment	90,610	103,035	106,728	17.11
Fault on the twisted pair (shunt, short circuit, break)	61,519	56,867	61,183	10.23
Blockade	41,937	44,828	52,635	7.94
Fault caused by improper handling (user)	50,698	44,038	42,870	7.84
Fault caused by severe weather conditions	35,658	40,104	38,399	6.50
Fault in configuration	17,812	30,905	42,261	5.18
Faults on a system	27,736	28,485	33,006	5.08
Occasional faults with unknown cause	20,868	23,767	25,968	4.02
Damage caused by outside contractors	16,891	21,295	19,932	3.31
Faulty cabling	12,224	13,306	17,502	2.45
Damage due to internal works	7186	7944	8538	1.35
Large attenuation on a twisted pair	3938	4002	4442	0.71
Great distance from the switch	2559	4001	3576	0.58
Wrong account	2122	2249	2024	0.36
Splitter is connected incorrectly	955	1138	906	0.17
Faults occurred under the influence of humidity	933	642	453	0.12
Crosspoint	664	538	351	0.09
Blown fuse—equipment damage	25	53	164	0.01
Power outage	4	6	40	0.00
Blown fuse—cause unknown	4	17	24	0.00

Source: T-HT internal report, 2012

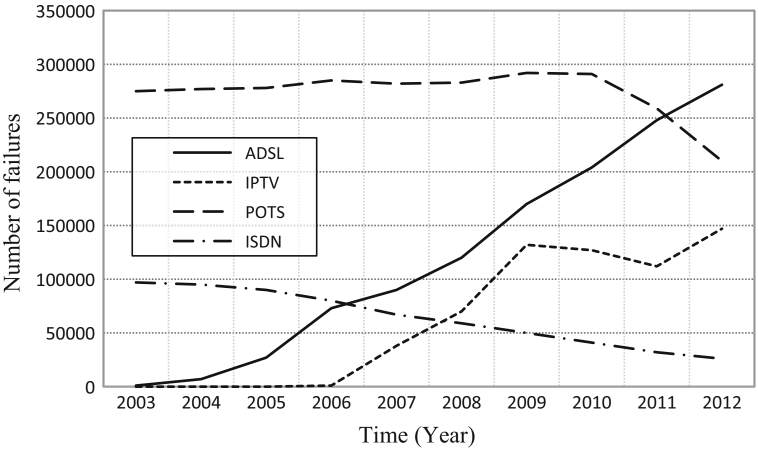


Fig. 3 Number of failures reported by customers in period 2003–2012

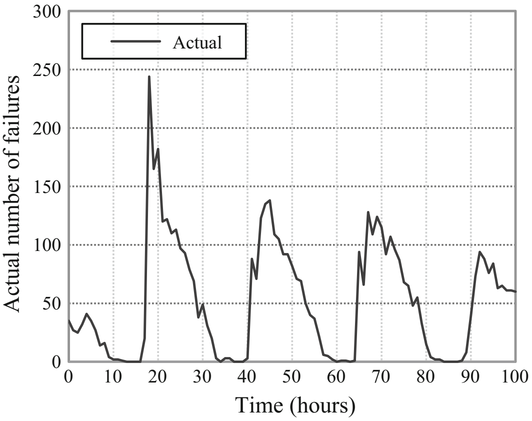
broadband network took place. The number of failures related to the traditional services such as POTS voice telephony and Integrated Services for Digital Network (ISDN) services is in decline because these services are being replaced by alternative services on broadband platforms. Another reason for the decline of POTS failures is the migration to the other shared models of services among multiple service providers. On the other hand, the number of failures on broadband services such as ADSL and IPTV grows along with the total number of users of these services. Note that the total number of failures in the whole observed period increases, and this trend is expected to continue in the upcoming years.

The failure-reporting process can be represented by a time series. These are stochastic series whose future development in time can be estimated based on previous values. Thanks to the fault-management system we have precise data about failure reporting, i.e., daily, weekly, monthly and annual series on the number of reported failures are available. Sampling was carried out in equal intervals so that interval sequences are cumulative in nature and can be considered as a series of a discrete time parameter.

As we said before, the failure-reporting process is strongly driven by customers. Failure reporting has daily dynamics that depend on the users' habits of using services and on their actions when they notice the presence of failure. Therefore, the time series representing the number of reported failures displayed in hourly and daily intervals reveals periodicity in time (Figs. 4, 5).

Daily or weekly patterns simply reflect work activities (usage of services) of residential and business customers during a day or week (Monday–Saturday). In the daily pattern a notable drop happens during the night while in the weekly pattern a reduction in the usage of the services on Sundays results in a lower number of reported failures. In a time series representing reported failures displayed in weekly and monthly intervals (monthly and yearly patterns) no periodicity in time is notable (Figs. 6, 7), because the cumulative number of reported failures in a week or in a month is more under the influence of random factors, such as bad weather or unexpected breakdowns in the network. In regular circumstances, time series have a

Fig. 4 Number of failures in hourly intervals



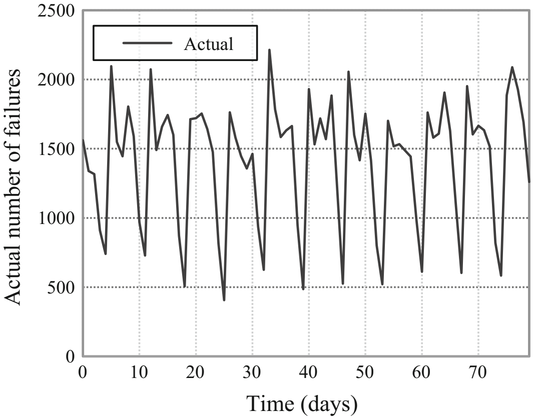


Fig. 5 Number of failures in daily intervals

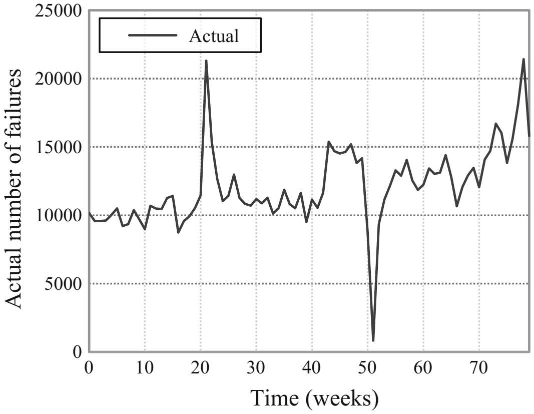
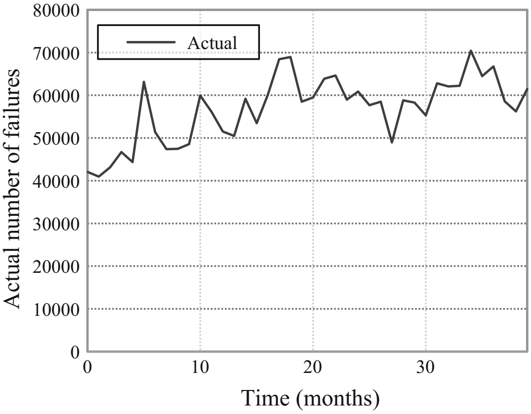


Fig. 6 Number of failures in weekly intervals

common and recognizable shape. However, unexpected events such as core network element malfunctions or thunderstorms can significantly affect the shape of the curve. One such anomaly that occurred in the 21st week can be seen in Fig. 6. Furthermore, the sharp decline that reaches a minimal value in the 52nd week is caused by characteristics of the calendar, i.e., by an incomplete last week in the year. Prediction of such anomalies on the curve is important in order to reserve the human resources that can prevent or resolve additional failures. A common daily distribution of reported failures (working day) is shown in Fig. 8 (gray line).

These are "regular failures" that occur mainly in the access network, and are caused by problems in the customer equipment; their reporting can be expected in similar daily distributions. For these quantities of failures, telecom operators have reserved resources to deal with their removal. However, in some situations anomalies occur, i.e., incidents that lead to an increased number of failures. One such anomaly recorded by the management system is shown in Fig. 8. The black-colored addition to the common shaped baseline curve represents the increased number of reported failures caused by a serious fault in equipment. Two things are important to define a prediction model of good quality: knowing the characteristics

Fig. 7 Number of failures in monthly intervals



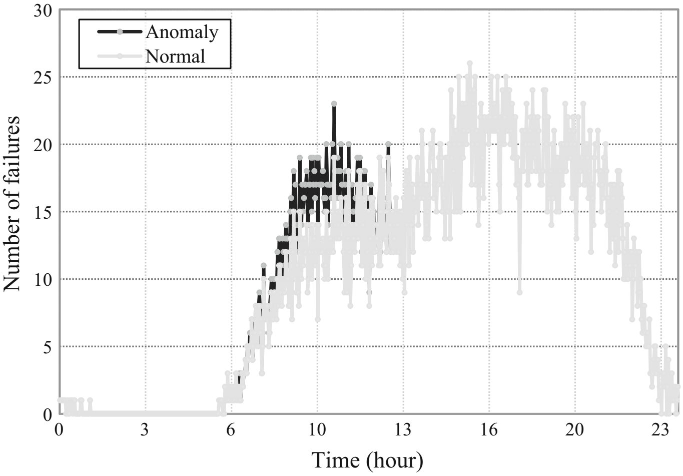


Fig. 8 Daily distribution of reported failures with anomaly

of the common series (trend, periodicity) and detecting the main external factors that cause substantial increases in the number of reported faults.

Estimates of future failures only on the basis of past information concerning quantities, without recognizing the external factors (environmental and human factors) and their influences can be optimistic and insecure. For example, the access and customer parts of the network are susceptible to environmental factors. Particular weather conditions (lightning or high humidity) can result in negative influences on lines and equipment, leading to sharp increases in the failure rate. In the next Section we show that the additional quantity of failures that appear under the influence of external factors (like adverse weather conditions) can be successfully predicted by the recursive neural net model. The multivariate modeling concept is introduced to reflect the effect of continuously varying influences of internal and external factors. It is very important that the model is scalable in a way that allows inclusion of additional factors that will be subsequently detected as relevant. Selection of important factors can be done by determining correlations between quantified factors and the number of reported faults. An example is presented in Sect 4.3.

4 Multivariate Model for Short-Term Prediction of Failure Quantities

A large variety of data from network management systems or data about external conditions are now available to the service providers. In this Section, we discuss the influence of the data and other characteristics of the broadband network environment and its processes on the choice of prediction method, predictor type, topology of predictor, and learning method. On the other hand, comprehensive data analysis and evaluation of the significance of input variables represent a precondition for development of a multivariate prediction model of good quality that encompasses the most relevant predictor variables. With such an approach it is possible to eliminate redundancy, enhance processing efficiency, and improve prediction accuracy.

4.1 Data Set

The data sets that have been used in this study were obtained from three different sources.

The first source—the Trouble Tickets database—contains information related to trouble reporting and troubleshooting. Three fields from the database were used, see Table 4. The second data source—the Error Logging database—is a component of the Network Management System. It includes information about network-element outages (alarm logs). Relevant data extracted from the Error Logging database is shown in Table 5.

Finally, the third data source—the Meteorological database—contains data from external sources. These data represent daily readings of meteorological measurements from 3 main regional centers in Croatia (Zagreb, Split, and Rijeka) that cover the most populated areas in the country. Seven relevant fields were extracted from the Meteorological database, Table 6.

Table 4 Relevant data extracted from the Trouble Tickets database (Trouble Tickets table)

ID	Field name	Field description
1	Faulty_Service	Affected service, identified according to customer's reports. For the purpose of this study only ADSL and IPTV related services have been selected
2	Reporting_Time	The time at which the customer who reported the failure called the contact center
3	General_Description	General description of a failure and possibly additional text about noticed causes

Table 5 Relevant data extracted from the Error Logging database (Alarm Logs table)

ID	Field name	Field description
1	Element_Name	DSLAM identification. Unique ID for the entire network This field is used as a link to the trouble-tickets table
2	Fault_Type	Fault type. Possible types are: breakdown, service degradation or occasionally occurring fault, and announced work
3	Fault_Cause	Causes of problems are grouped as software errors, hardware failure, transmission, and power supply
4	Alarm_Start_Time	The time at which the alarm first appeared
5	Alarm_End_Time	Alarm ceasing time, after repair
6	Affected_Customers	The number of customers affected by the network-element outage

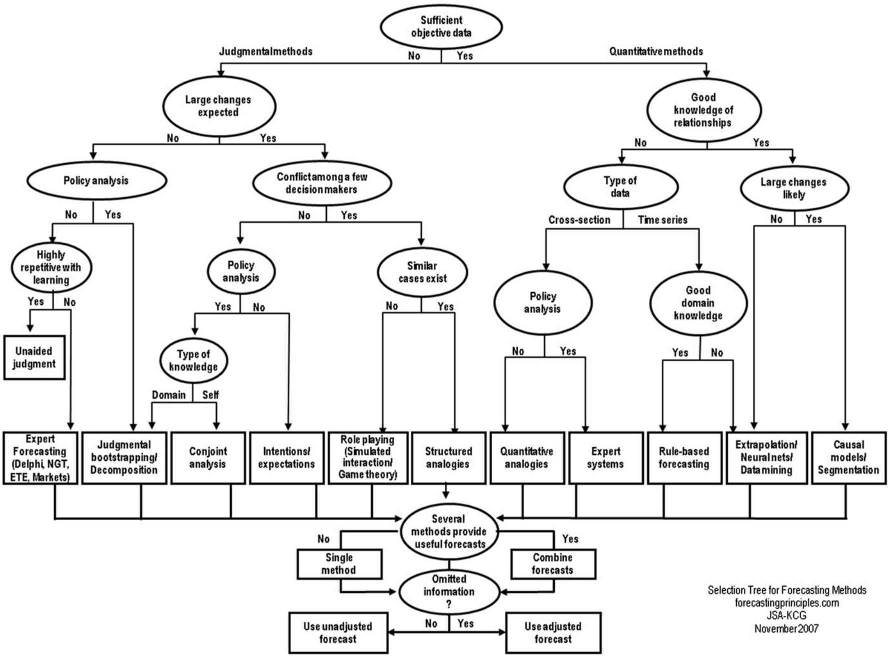
Table 6 Relevant data extracted from the Meteorological database

ID	Field name	Field description
1	Meteo_Station	Location of the meteorological station
2	Meteo_Date	Date of reading
3	Var_Rain	Binary variable indicating whether on the Meteo_Date rain is observed or not in the area of Meteo_Station
4	Var_Lightning	Binary variable indicating whether on the Meteo_Date lightning is observed or not in the area of Meteo_Station
5	Var_Snow	Binary variable indicating whether on the Meteo_Date snow is observed or not in the area of Meteo_Station
6	Var_Hail	Binary variable indicating whether on the Meteo_Date hail (hailstorm) is observed or not in the area of Meteo_Station
7	Var_Fog	Binary variable indicating whether on the Meteo_Date fog is observed or not in the area of Meteo_Station

We analyzed data collected during the period from January 2012 to August 2012. The total number of failure reports recorded in the Service Management Center in this period was 585,000, while the number of network-element outages in the same period was 591. There were a total of 103 rainy days, 53 days with lightning, 23 days with snow, 2 days with fog, and a day with hail observed in all 3 meteo-stations in the period of observation.

4.2 Description of the Model

As we already said in the Introduction, both long-term and short-term predictions of the number of reported failures are valuable for telecom providers. The long-term prediction procedure based on a time series is not as complex as the short-term prediction because of exogenous factors that can be largely ignored. Actually, effects of exogenous factors are reduced and diluted over long periods (months and years). Exogenous phenomena are relatively rare so accurate long-term predictions can be made on the basis of the past values alone. We can say that the impact of exogenous factors is already incorporated in the time series, and the time series themselves embody all information necessary for modeling. On the other hand, a



Selection Tree for Forecasting Methods
forecastingprinciples.com
JSA-KCG
November2007

Fig. 9 Armstrong's decision tree that helps in the selection of an appropriate prediction method

short-time prediction should be based on multivariate models supported with input data of adequate quality that represent external influences. Correlation analysis helps in determining the significance of input factors. That input has no significant impact that can be excluded from the model, thereby reducing its complexity. In order to ensure a more accurate short-term prediction of the number of reported faults, a prediction model based on a neural network with exogenous input has been developed. All available and relevant input data collected from various sources inside and outside the network, as well as historical data, are included in the model. Also, the model has the ability to deal with temporal cross-correlations, i.e., the input variables can lag behind the dependent variables. The model meets the following criteria:

Users shouldn't be able to know variable dependencies or enter initial presumptions about dependencies in the system;

Relationships among the variables ought to be calculated automatically, not manually;

The model is nonlinear, due to the nature of the process being described;

The model allows an arbitrary number of input variables (new input can be added subsequently);

The model has sufficient tolerance to noise;

The possibility of generalization, prediction of new situations;

The model is tolerant of data incompleteness;

There is a possibility to predict an arbitrary number of steps ahead;
The model isn't too slow during the learning and prediction.

In [22], Armstrong developed a decision tree (Fig. 9) that helps in the selection of an appropriate method. Bearing in mind the characteristics of the broadband network environment and its processes, the following facts, which are relevant to the selection of the method, can be stated:

Data sets of sufficient size and accuracy are available—input data for prediction can be obtained from Operations Support Systems (OSS), Business Support Systems (BSS), and external sources. In the research described in this paper, we used actual data that exceeds 1.5 million items collected by performance and fault management systems in the period of 2009–2012, data about users' habits, data warehousing, external sources—meteo-data logs and data about relevant announced events;
Good knowledge about the relationship between relevant variables is missing.
Data type—discrete time series;
Inner nature of the system is not known well;
The system is massive with inertia, i.e., there is a low probability of changing conditions in the system during a period of predictions.

Considering these facts in relation to the Armstrong decision tree reveals that the most appropriate prediction approaches should be: extrapolation, neural networks, and data mining. These prediction approaches were reached by passing through the decision tree according to the responses to the following questions:

1. Do we have a data set with sufficient size and accuracy?… Yes
2. Is there good knowledge of the relationship between the variables of the system?… No
3. Type of data—time series?… Yes
4. Is there good knowledge of the nature of the system?… No

We used the decision tree as guidance for selection of an appropriate method. Besides this, we carried out a comparison test of several promising candidate methods [23, 24]. We compared and analyzed the prediction quality of the following prediction methods applied to the time series representing the numbers of reported failures: Autoregressive Integrated Moving Average (ARIMA), Multi Layer Perceptron with time-delayed inputs (MLP), Focused Time-Delay Neural Network (FTDNN)—also known as Input Delayed Neural Network (IDNN)-, Layer Recurrent Network (LRN) using back connection from a hidden layer to the input layer and Nonlinear Autoregressive Network with Exogenous Inputs network where output data is preserved in a delayed memory line (NARX). The recursive models (LRN and NARX) showed dominant results, while the NARX model achieved the best overall prediction results during the comparison test [25], so it is used for the development described in this paper.

Our previous results in the area of predicting quantities of reported failures related to broadband networks are presented in [24, 25]. The objective was to compare different prediction methods and assess their applicability for short-term prediction. The paper [25] presents detailed comparative analysis related to the following traditional prediction method and methods based on artificial intelligence: ARIMA, MLP—Multi Layer Perceptron, IDNN (FTDNN)—Input Delayed Neural net, LRN—Elman network, and NARX—Jordan network. Accuracy of predictions obtained by these methods was estimated by using the Root Mean Square Error (RMSE) and Model Efficiency Index or Coefficient of determination (R2) parameters. These parameters are calculated by comparing the actual and predicted data. In short and long-term predictions, the NARX and LRN models achieve the best overall results. The results obtained show that dynamic recurrent neural networks outperform static neural networks. Also, the traditional prediction method (ARIMA) couldn't achieve the accuracy of the recurrent neural networks. This confirmed the hypothesis about the non-linear character of most fault causes in telecommunications' networks. Detailed simulations have been made with variable prediction horizons from 1 to 8 days ahead. Comparisons of the accuracy of prediction results are shown in Table 7.

Models based on recursive networks (LRN, NARX) give better results, i.e., smaller errors in comparison with two other models. Therefore, we have selected dynamic memory models, such as the "NARX" and "LRN" neural networks [26], due to their ability to deal with logical and temporal correlations among variables, as the most promising candidates [27] for prediction in complex systems, which broadband telecommunication networks indeed are. In the text that follows, we show how further improvements in prediction accuracy can be achieved by applying additional fine tuning of network parameters and, of course, by recognition and introduction of relevant input variables that would bring additional knowledge to the model. Exogenous input allows feeding the NARX network with an arbitrary number of data values related to input (predictor) variables. Optimal choice of input variables is critical for accurately predicting the output variable. In the domain of

Table 7 Prediction accuracy—prediction horizon is varied

Prediction hor. (days)	ARIMA		LRN		NARX		TDNN	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
1	362.435	0.38207	328.901	0.51459	284.795	0.59163	333.689	0.42811
2	364.259	0.35403	294.386	0.57808	292.749	0.58276	296.25	0.55525
3	368.452	0.32625	283.769	0.5952	278.002	0.61644	289.743	0.58336
4	326.821	0.44902	300.255	0.53495	287.594	0.57334	363.788	0.31733
5	301.387	0.53001	283.052	0.58546	294.247	0.55202	404.855	0.15192
6	300.79	0.50199	284.25	0.55525	289.452	0.53883	337.187	0.37417
7	290.587	0.52492	268.475	0.59447	304.043	0.47991	300.573	0.49171
8	417.017	0.29588	396.452	0.31756	371.276	0.44187	430.625	0.24918
Mean	341.468	0.42052	304.942	0.53445	300.27	0.5471	344.589	0.39388

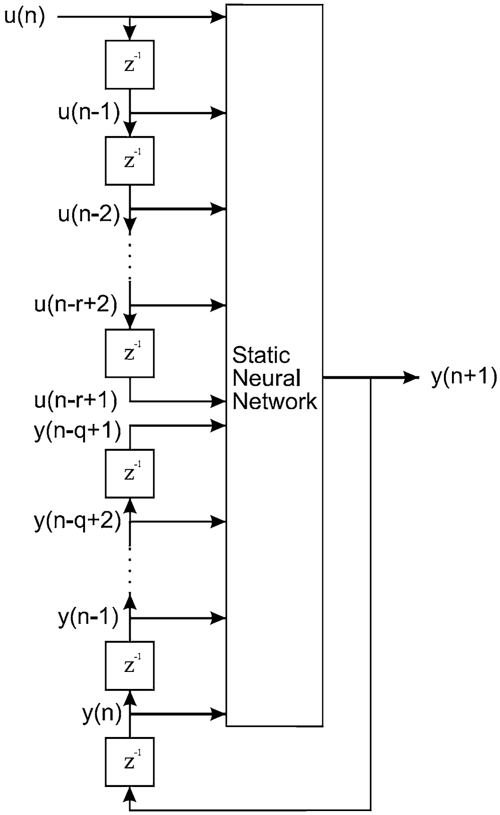
broadband networks, the following data can be used as input for predicting failure quantities:

Data from the very recent past—this relates to data about reported failures collected in a period from the past few minutes up to the past few hours;
Archived data (data warehousing (DWH))—short-term and long-term historical data about failures (from the past few days up to several years); these data allow time series analysis and identification of trends and seasonal patterns;
Data about network loads, performances and the operational statuses of network elements—provide information about the operating regime of network elements or some kind of observed irregularity (peak loads, overloads, traffic rejections, …);
Fault-management data—degradation of services, outages of individual network elements, faults logged and stored in the fault-management system that are correlated with failures reported by users;
Service-usage information—data about users' habits, i.e., about average daily service usage time, average service usage time per session, preferences according to types of service, the distribution of daily traffic volumes, etc.;
Equipment reliability—reliability of network elements can be calculated from error logging and error statistic databases, providing a basis for calculating Time To Failure (TTF) and Time To Repair (TTR) parameters; by using these parameters, we can predict the dynamics of equipment breakdowns,
Data about external influences on network—nformation about the events that cause external influences on the system; there are many different influences that can cause faults in the network; for example, faults in power supply can disrupt the operational status of customer premises equipment;
Meteorological data—due to the large impact of humidity and electrical discharge on the network, particularly the access and transport part of the network, these data represent important input to the model;
Data about social events—information on social events that may affect the network, for example large gatherings of customers in a given area, a variety of migrations, seasonal loads, "nomadism", scheduled or announced events that can be taken as input variables for prediction.

Certainly, data of good quality from all the above-mentioned sources were not available during the research described in this paper. Some data were not available at all and it would take significant, additional effort to make them available.

Therefore, the model has been developed on the basis of available data sets. As we emphasized before, the model is scalable, i.e. subsequent addition of input variables are possible. The neural network used in the model belongs to the group of nonlinear dynamic networks, which is known in the literature as NARX, NARMA, NARMAX [28]. This is a nonlinear autoregressive neural network with exogenous inputs, also referred to as an input–output recurrent model; the principle scheme is shown in Fig. 10. A special feature of this network configuration is two delay lines. The first line, known as a recurrent delay line, connects output with input and allows the dynamics of the signal to be captured. The second line, known as a tapped delay

Fig. 10 NARX network—principle scheme



line, accepts an input vector with a time delay. Both nonlinear and linear functions can be employed in one hidden and one output layer; input and output can be multidimensional.

An additional advantage of the model, as opposed to some other recurrent models, is a standard multilayer perceptron located in the center of the network and enabling learning by a standard algorithm. This ensures simplicity and reduces learning time. The network dynamics is described by Eq. (1):

$$y(n+1) = F(y(n), \ldots, y(n-q+1), u(n), \ldots, u(n-r+1)) \tag{1}$$

where u(n) is the current observed value, u(n − 1), …, u(n − r + 1) are past observations of the variables memorized up to r − 1 lags, and y(n), …, y(n − q + 1) are q past output fed into a recurrent delay line. In this way, past information can be preserved, which means that information from the initial moment up to the current moment affects the calculation of the new output value.

We used the Matlab [26] implementation of the neural network with configuration parameters shown in Table 8. The input and output vectors consist of a total of 250 records that represent the observed period (approximately 8 months).

Table 8 Neural network configuration parameters

Parameter	Parameter mark	Value
Input vector	P	6 × 250
Output vector	T	1 × 250
Length of input tapped delay lines	ID	4–8
Length of output tapped delay lines	OD	4–8
Hidden layer size	Si	2–6
Transfer function of hidden layer	TFi	tansig
Transfer function of output layer	TFi	purelin
Training function	BTF	trainlm
Network performance function	PF	mse
Function for normalizing input vector	IPF	removeconstantrows, mapminmax
Function for normalizing output vector	OPF	removeconstantrows, mapminmax
Data division function	DDF	dividerand
Epochs between displays	trainParam.show	10
Maximum number of epochs to train	trainParam.epochs	150–500

The network training function updates weight and bias values according to the Levenberg–Marquardt optimization (function trainlm) [26], because this is much faster than training by a basic error back-propagation algorithm (function train). Another advantage of the Levenberg–Marquardt method is the ability to find a solution in situations when training starts far away from a global minimum. The only problem related to this method is its memory consumption. Memory exhaustion can be avoided by adjusting the parameters to reduce memory usage, or by using other methods such as the quasi-Newton back-propagation method, which is slower, but uses memory more efficiently. The choice of transfer functions, like tansig, logsig, and purelin, depends on the characteristic of the modeled system. In this specific case, a hyperbolic tangent sigmoid transfer function (tansig) was used in the hidden layer while a linear transfer function (purelin) was used in the output layer. The configuration of the NARX network with 6 inputs, 1 recurrent feedback loop, and the above-listed parameter values is shown in Fig. 11. The network consists of three main parts. The input section ensures that the values of the variables are passed into the 6 inputs through the delay lines. Central place is occupied by the classic multilayer perceptron. Its function is to determine the significance and functional dependencies using regression in the hidden and output layers of the network. The third part is the output layer, which makes input/output links through the delay lines. Input delay lines ensure that the input data related to past events impact the output value (links for temporal cross-correlations); for example, how a thunderstorm that occurred 2 days ago affects the number of faults today.

4.3 Determination of Significant Input Variables

In this subsection, we demonstrate how correlation-based procedures can be used to detect factors that significantly affect the number of reported failures, and as such

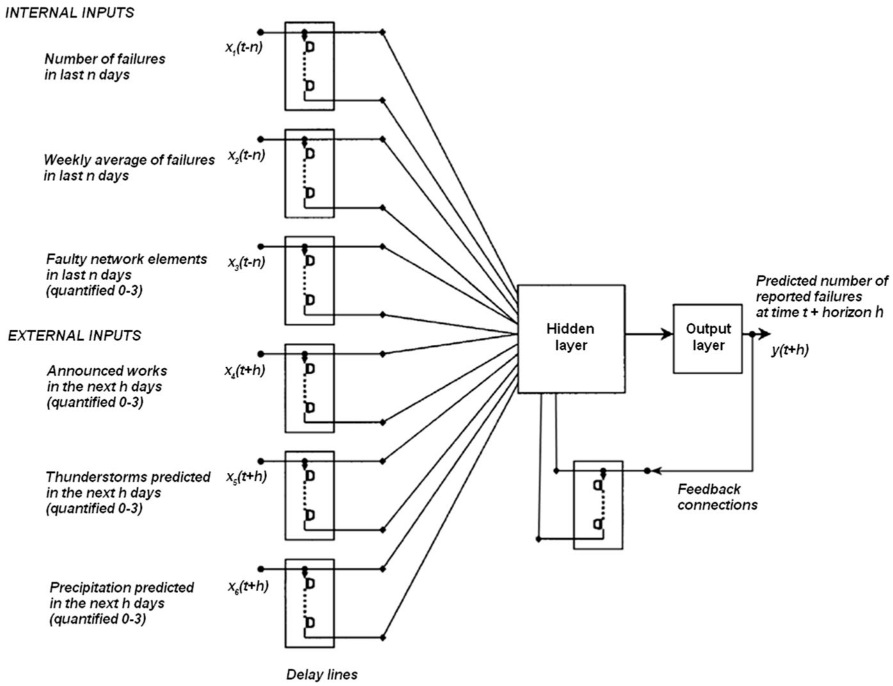

INTERNAL INPUTS
Number of failures in last n days
x1(t-n)
Weekly average of failures in last n days
x2(t-n)
Faulty network elements in last n days (quantified 0-3)
x3(t-n)
EXTERNAL INPUTS
Announced works in the next h days (quantified 0-3)
x4(t+h)
Thunderstorms predicted in the next h days (quantified 0-3)
x5(t+h)
Precipitation predicted in the next h days (quantified 0-3)
x6(t+h)
D
Delay lines
Hidden layer
Output layer
Predicted number of reported failures at time t + horizon h
y(t+h)
Feedback connections

Fig. 11 Network configuration

they are candidates to be quantified and included as input variables in the prediction model. We can simply use scatter diagrams or mathematical methods (e.g., calculating Spearman's or Kendall's rank correlation coefficient, or coefficient of linear correlation) to compare two phenomena. The scatter diagram in Fig. 12 shows the influence of electrical discharges on the number of failures. We can see a significant, positive impact. The diagram in Fig. 13 shows the effect of temperature. In this case, the impact is negligible—the trend line is horizontal.

An additional possibility to find relations between data is detecting temporal cross-correlations, as shown in the following example. In a large network that is built up over years, users' access lines do not have the same characteristics, and thus are not all equally resistant to the effects of humidity. The reasons for this are manifold:

Distance to the users varies (in practice from 0 to 7 km).
All wires in the cable do not have the same characteristics (some were factory defective or deteriorated more rapidly).
Some line units are sensitive to the influence of humidity, others less so.
Some underground joints are susceptible to the influence of water, others are not, etc.

Even though pure water is an insulator, the water that reaches the cable usually is not clean. It picks up dust and pollutants from the air and minerals that come from

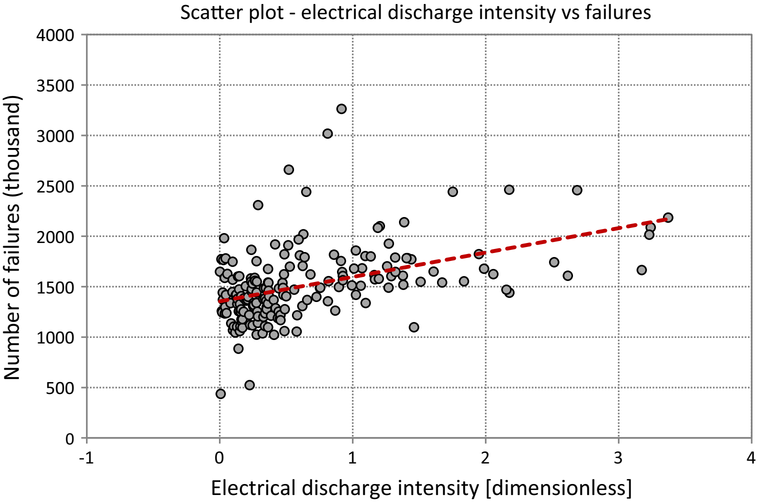


Fig. 12 Impact of electrical discharges

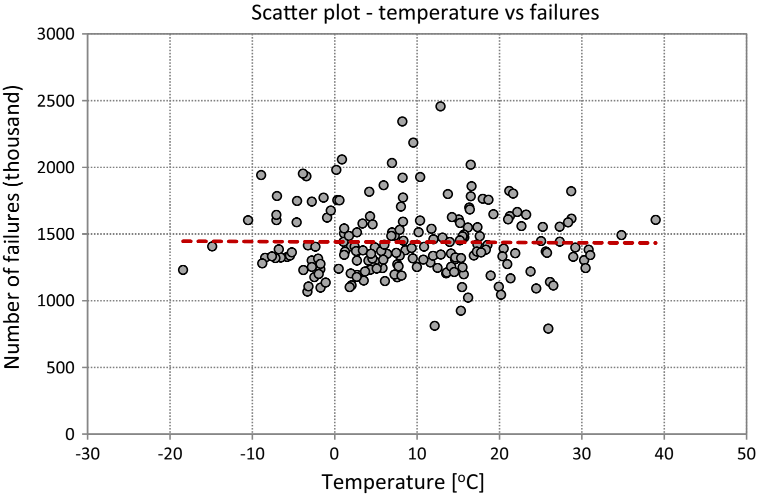


Fig. 13 Impact of temperature

the soil. Manufacturing defects and wire insulation deteriorate over time (aging), allowing moisture to penetrate into cable and therefore into the twisted pair. The results of this are changed electrical parameters, such as the capacity and impedance of the cable, causing a stronger attenuation of the higher frequency spectrum of the ADSL signal [29, 30]. To analyze the influence of humidity on the occurrence of failures, a number of measurement data were collected on a large number of lines. The measurement data were collected from the DSLAM measurement system and correlated with data on humidity. Our previous research [31] showed that the higher values of humidity caused by humid weather have significant influence on a specific

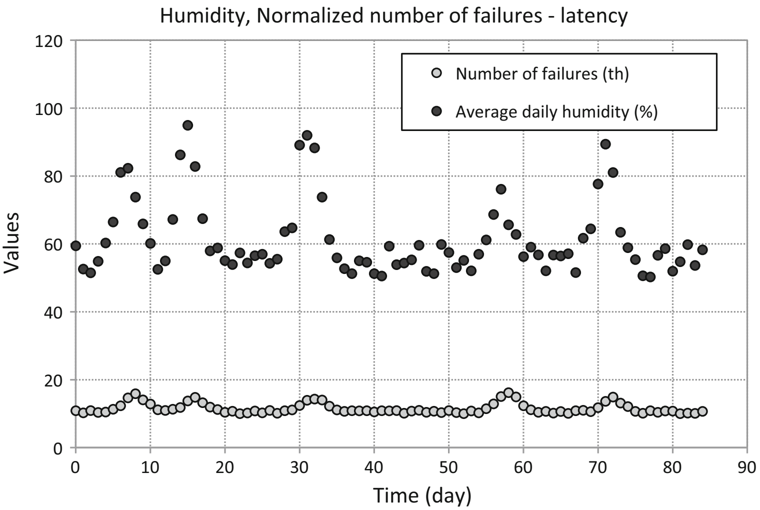


Fig. 14 Humidity versus number of reported failures—time series

number of the twisted pairs that are poorly protected from these external influences. Based on long-term monitoring, it was found empirically that the appearance of rain and humidity increases the total number of failures. So the idea was to discover the degree of correlation between the humidity and the number of reported failures. The measured data and the data obtained as a result of the failure fixing/handling process allowed us to show the dependency between two variables: relative humidity in the air and number of faults (see Fig. 14). The diagram depicts two time series. A time delay between peaks of humidity and the number of existing reported failures indicates the correlation between them.

If we display the humidity and the number of failures in a scatter diagram, Fig. 15, correlation is evident. Moreover, one can recognize a circular shape of the scattered points in the diagram that suggests the existence of temporal cross-correlation with the period of delay *d*.

For two series x(i) and y(i), with the mean values *mx* and *my* that are shifted by the delay *d*, the expression (2) defines their cross-correlation.

$$r = \frac{\sum_i [(x(i) - mx) * (y(i-d) - my)]}{\sqrt{\sum_i (x(i) - mx)^2} * \sqrt{\sum_i (y(i-d) - my)^2}} \tag{2}$$

Peak values of reported failures are delayed with regard to the peaks of measured humidity by approximately 24 h. Figure 16 shows the temporal cross-correlation and its dependence on the variable of time delay (lags) d (the highest correlation is around the 1 day mark.). This delay is explained by the time necessary for the moisture to enter the cable, combined with the time it takes the user to notice and report a failure in the service.

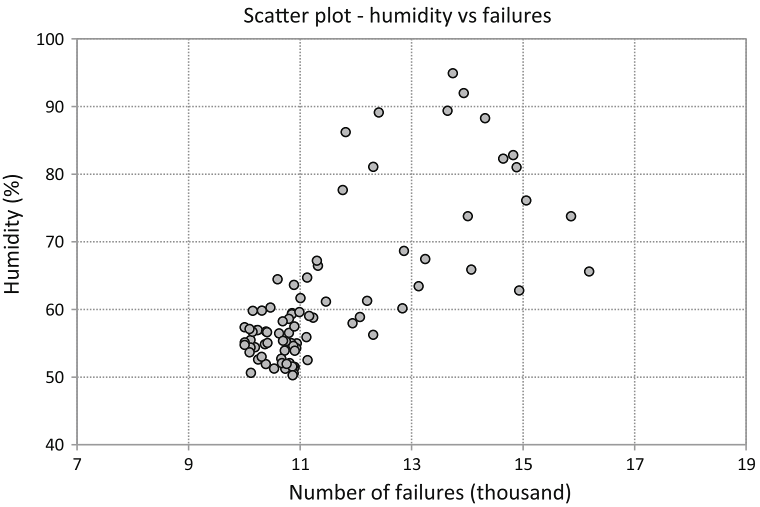


Fig. 15 Humidity versus number of reported failures—scatter diagram

The accuracy of predictions by the NARX neural network model strongly depends on the choice of input variables. Factors with a strong impact should be quantified and introduced into the model as input variables, while factors with small and insignificant correlations are rejected. We divided input variables into two groups. The first group includes internal variables whose values represent the state of the network equipment and quantities of reported failures in the past (listed in the upper part of Table 9). The second group includes external variables whose values represent the environmental and human factors with external influences on the network (lower part of Table 9).

Significant input is represented by the following data:

Data about lightning and rainfall are downloaded from the three main meteorological stations in Croatia [32];

Data about outages of network elements, data on the number of failures in the last 4 days, data about announced work on the network, and historical data on weekly averages of the number of failures are all taken from the OSS and DWH systems in the T-HT.

Network element outages and external factors do not equally affect all parts of the network. Electrical discharge mainly causes problems with equipment in the user and the access part of the network, rainfall affects the operation of equipment in the access part, while network element outages or announced works significantly impact the access and the core part of the network.

There were some practical limitations on the data that generally reduce the accuracy of predictions. These limitations could be avoided by improving the data-collection process. Variables representing thunderstorms and rainy weather are

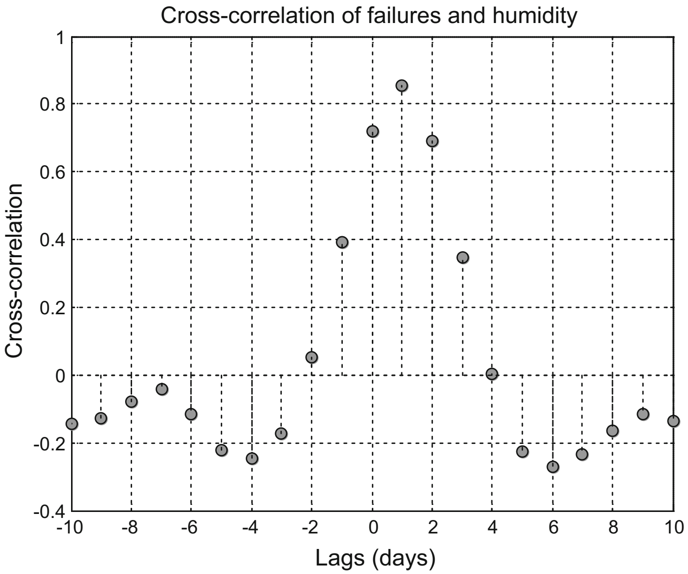


Fig. 16 Humidity versus number of reported failures—cross-correlation diagram

Table 9 Significance of variables

Variable A → Number of reported failures (NRF)	Spearman's correlation coefficient	Variable A is significant
Network element outages → NRF	0.238	Yes
Number of failures (last 4 days) → NRF	0.19	Yes
Number of failures (weekly averages) → NRF	0.082	Yes
Number of failures (monthly averages) → NRF	−0.097	No
Electrical discharges (lightning) → NRF	0.293	Yes
Announced works in network → NRF	0.107	Yes
Rain → NRF	0.241	Yes
Fog → NRF	−0.03	No
Snow → NRF	−0.032	No

quantized by four values (weighting factors) on the set {0, 1, 2, 3} where 3 denotes the greatest impact. This resolution is too low to provide the precise expression of impact. The same quantization is applied to the variables representing outages of network elements and announced work on the network. Data on daily amounts of failures are quite accurate with the exception of a few cases where data are not collected due to errors in the OSS and DWH systems. Also, there were some minor errors in the classification of failures in the DWH system. These limitations on the data, to some extent, affect the accuracy of predictions. Due to the aforementioned constraints, the significances of input variables are not estimated solely on the basis of correlation coefficients but also by using special knowledge of experts.

4.4 Prediction Results and the Accuracy of the Model

Model learning and testing were conducted on data collected during the first 8 months of 2012. Data collected during the first 5 months were used for learning while the rest of the data were used for prediction and model testing. This learning/predicting ratio (5/3) was chosen because of relatively rare influences of external factors on the network. A longer learning period is necessary to ensure proper inclusion of the external influences into the model parameters. But once the learning is completed, predictions can be performed for months ahead with periodic updates, assuming a stable, unchangeable system. It is known that large systems like broadband networks are slowly changing systems.

Figure 17 shows the actual, daily reported number of failures and values obtained by prediction (gray circles) for a 4-day prediction horizon. The daily average failure values, mean values of all Mondays, Tuesdays, etc. (black pluses in the figure) were taken as reference values. The prediction approach based on the daily average values was probably the simplest method of rough prediction and was used in the real process. Values of all input variables were fed into the NARX model. Relative to the reference value, the gain of using NARX model is shown (with a gray circles) in the figure.

Certain prediction deviations are visible at some points. For example, on the 105th day the network suffered a quite unexpected outage of a large number of devices caused by a sudden fault in an aggregation of DSLAMs. This event was not predictable on the basis of available input information. The large number of devices affected by the fault caused a significant increase in the number of failure reports. Therefore, the deviation between the predicted and actual number of reports was significant.

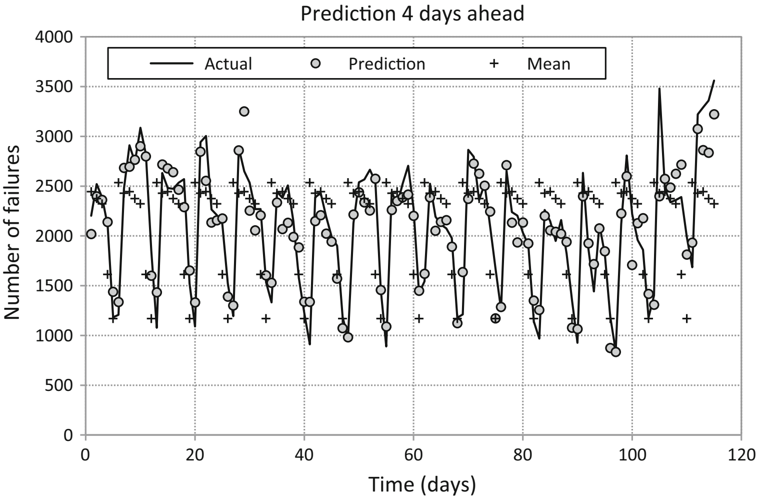


Fig. 17 Prediction for 4 days ahead

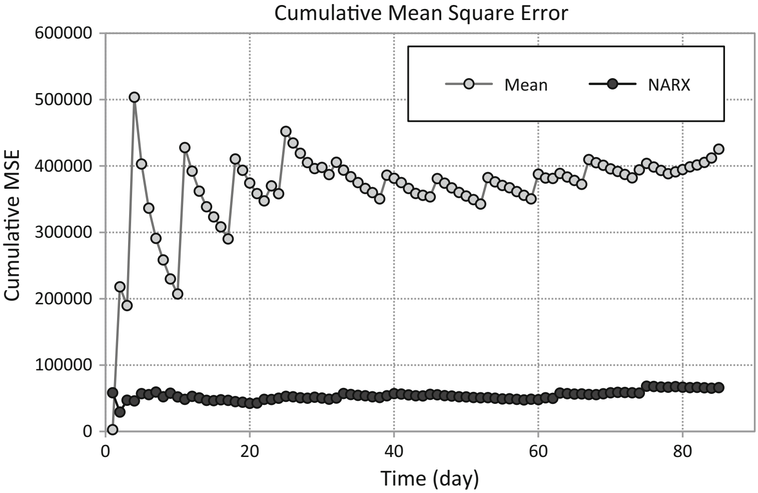


Fig. 18 CMSE—NARX versus simple prediction based on daily averaged values

Usually, accuracy of prediction is estimated by using the Root Mean Square Error (RMSE) as a measure of the differences between values predicted by the model and actual values. To denote changes of prediction error over the time, we calculate the Cumulative Mean Square Error (CMSE). The CMSE provides insight into the accuracy of the predictions at all stages and facilitates finding possible inconsistencies in the model. The gain of using the NARX model, compared to the simple method of using the daily average failure values, is evident, Fig. 18. The mean square prediction error of the NARX model is approximately 5 times lower.

An interesting view on the accuracy of predictions is obtained by varying the prediction horizon from 1 to 8 days ahead. The model was tested with prediction horizons from 1 up to 8 days (Table 10). Mean Square Errors (MSE) and R^2 are used as a measure of accuracy. R^2 is often used in statistics to assess the accuracy of

Table 10 Prediction accuracy as the prediction horizon is varied

Prediction horizon (day)	Day averages		NARX		Gain ratio
	MSE	R^2	MSE	R^2	
1	401,491.7	−0.212	81,883.3	0.753	4.9
2	410,255.8	−0.169	87,954.6	0.749	4.7
3	416,814.2	−0.120	93,448.9	0.749	4.5
4	420,657.0	−0.090	84,743.9	0.780	5.0
5	431,894.4	−0.117	97,787.6	0.747	4.4
6	434,448.6	−0.119	97,928.6	0.748	4.4
7	455,691.4	−0.141	168,591.1	0.578	2.7
8	454,374.2	−0.137	375,599.3	0.060	1.2
Mean	428,203.4	−0.138	135,992.2	0.646	4.0

Table 11 Prediction accuracy as the number of variables involved is varied

Prediction horizon (day)	Day averages		NARX		Gain ratio
	MSE	R^2	MSE	R^2	
x1 + x2 + x3 + x4 + x5 + x6	425,303.6	−0.097	63,134.1	0.837	6.7
x1 + x2 + x3 + x4 + x5	425,303.6	−0.097	87,558.5	0.774	4.9
x1 + x2 + x3 + x4	425,303.6	−0.097	255,600.1	0.341	1.7
x1 + x2 + x3	425,303.6	−0.097	262,776.2	0.329	1.6
x1 + x2	425,303.6	−0.097	273,027.2	0.316	1.6
x1	425,303.6	−0.097	312,312.8	0.194	1.4
–	425,303.6	−0.097	415,466.3	−0.072	1.0
Mean	425,303.6	−0.097	238,553.6	0.389	2.7

a model that quantifies the correspondence between actual and modeled data. $R^2 = 1$ means complete correspondence between the model and reality. In real models, the values are always less than 1; the lower the value of the coefficient, the less correspondence.

We have also measured the quality of the multi-variable prediction model by varying the number of input variables included. By varying the number of input variables, it is possible to observe the effect that each individual input has on the accuracy. Table 11 shows mean square errors of the model when the number of variables involved in the model is varied.

The model is designed based on the assumption that the input variables have the same accuracy during the whole prediction horizon. In reality, it is very difficult to maintain the same level of variable accuracy over a longer period. For the input variables that represent environmental factors, such as rain and electrical discharges, it is realistic to expect that the accuracy of prediction declines considerably when the prediction period exceeds 3 or 4 days (accuracy of weather forecasts). Usually, for the announced work there are plans for more than 10 days in advance, while the outages of the network elements are very unpredictable. However, there are various methods to anticipate outages of elements; the use of any particular method depends on the nature and type of analyzed systems. These methods have already been described in Salfner's study [4]. Estimated prediction horizon lengths in which the input variables have satisfactory accuracy are shown in Table 12. Each variable has its own characteristics, but it is reasonable to assume that the multivariable model developed here will provide prediction results of considerable accuracy for a 3–4 day horizon. This would be sufficient for the needs of the telecom business, i.e., operational network maintenance and workforce allocation.

Notwithstanding the described positive characteristics of the model, it should be stated that the model is susceptible to erroneous input data. Generally, erroneous data that enter into neural network models with long-term memory cause considerable reduction in the accuracy. Therefore, it is important to estimate the likelihood of errors in data or even better, if possible, to provide some procedure for identifying and correcting erroneous data before they enter the model. Because we were aware of the problem, we have been using a semi-automatic procedure for

Table 12 Estimated prediction horizon lengths with satisfactory accuracy for input variables

Variable ID	Variable description	Estimated horizon length (days)
x1	Number of failures—weekly averages	–
x2	Data about upcoming rainfall	2–6
x3	Data about upcoming thunder and lightning	2–4
x4	Data about announced work on the network	10–30
x5	Number of failures in last 4 days	–
x6	Data about network element outages	0–2

recognition and correction of erroneous data. According to our experience, most errors in data appear due to malfunctions in fault-management or data-warehousing systems, and in some cases because of errors caused by human factors. Our semi-automatic procedure for recognition and error correction includes log analysis, outlier detection, and methods for reconstruction of the actual data.

5 Conclusion

In this paper, we give an overview of locations and types of faults in a common broadband telecommunications network. The total number of reported failures is increasing, and this trend is expected to continue in the coming years. Failures need to be eliminated as soon as possible, reactively, after a user's complaint, but it would be better to act preventively and proactively. It is long and short-term failure predictions that make this possible. Therefore, telecom operators are compelled to include prediction procedures in their operational and strategic network managements. An accurate short-term prediction is a challenge due to a number of relevant factors. It is shown that a prediction model based on a NARX neural network, well trained and fed with relevant input data, is accurate and effective enough to support operational maintenance and workflow allocation processes. The proposed model takes into account the impact of 6 factors (network-element outages, number of failures in the last 4 days, number of failures—weekly averages, electrical discharges, announced work on the network, and upcoming rainfall). But the model is also scalable in the sense that new input can be simply added. Also, it is demonstrated how correlation-based procedures can be used for determination of factors that significantly affect the number of reported failures in the future, and as such, are candidates to be quantified and included as input variables to the prediction model. As explained in the paper, predictions of expected failures are based on large quantities of data originating from multiple sources. Consequently, erroneous data are inevitable, but such data when entering a neural network model with long-term memory, cause considerable reduction in the accuracy. Therefore, as a continuation of this study, it would be necessary to improve procedures for identifying and correcting erroneous data before they enter the model.

References

1. Avizienis, A., Laprie, J.-C.: Dependable computing: from concepts to design diversity. In: Proceedings of the IEEE (1986). doi:10.1109/PROC.1986.13527
2. Melliar-Smith, P.M., Randell, B.: Software reliability: the role of programmed exception handling. In: Proceedings of the ACM Conference on Language Design for Reliable Software (1977). doi:10.1145/800022.808315
3. Laprie, J.-C., Kanoun, K.: Software reliability and system reliability. In: Lyu, M.R. (ed.) Handbook of Software Reliability Engineering, chapter 2, pp. 27–69. McGraw-Hill, New York (1996)
4. Salfner, F., Lenk, M., Malek. M.: A survey of online failure prediction methods. J. ACM Comput. Surv. (2010). doi:10.1145/1670679.1670680
5. A guide to standard and high-definition digital video measurements. Publishing Tektronix. http://www.tek.com/regional-page/guide-standard-hd-digital-video-measurements (2007). Accessed 15 August 2013
6. Held, G.: Understanding IPTV (Informa Telecoms and Media). Auerbach publications, Boston (2006)
7. Kandula, S., Katabi, D., Vasseur, J.P.: Shrink: A tool for failure diagnosis in IP networks. In: Proceedings of the ACM SIGCOMM Workshop on Mining Network Data (MineNet 2005), pp. 173–178. ACM (2005). doi:10.1145/1080173.1080178
8. Mao, Y., Jamjoom, H., Tao, S., Smith, J.M.: Networkmd: topology inference and failure diagnosis in the last mile. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement (IMC 2007), pp. 189–202. ACM (2007), doi: 10.1145/1298306.1298333
9. Kavulya, S.P., Joshi, K., Hiltunen, M., Daniels, S., Gandhi, R., Narasimhan, P.: Draco: Top-down statistical diagnosis of large-scale VoIP networks. Carnegie Mellon University, AT&T Labs-Research. http://www.pdl.cmu.edu/PDL-FTP/ProblemDiagnosis/CMU-PDL-11-109.pdf (2011)
10. Mahimkar, A., Ge, Z., Shaikh, A., Wang, J., Yates, J., Zhang, Y., Zhao, Q.: Towards automated performance diagnosis in a large IPTV network. In: Proceedings of the ACM SIGCOMM 2009 conference on Data communication, pp. 231–242. ACM (2009). doi:10.1145/1592568.1592596
11. Song, H.H., Ge, Z., Mahimkar, A., Wang, J., Yates, J., Zhang, Y.: Analyzing IPTV set-top box crashes. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Home Networks (HomeNets 2011), pp. 31–36. ACM (2011). doi:10.1145/2018567.2018575
12. Mahimkar, A., Song, H.H., Ge, Z., Shaikh, A., Wang, J., Yates, J., Zhang, Y., Emmons, J.: Detecting the performance impact of upgrades in large operational networks. In: Proceedings of the ACM SIGCOMM 2010 Conference, pp. 303–314. ACM (2010). doi:10.1145/1851182.1851219
13. Guo, Z.X., Wong, W.K., Li, M.: A multivariate intelligent decision-making model for retail sales forecasting. Decis. Support Syst. **55**, 247–255 (2013). doi:10.1016/j.dss.2013.01.026
14. Soroush, A., Bahreininejad A., van den Berg, J.: A hybrid customer prediction system based on multiple forward stepwise logistic regression mode. Intell. Data Anal. **16**, 265–278 (2012). doi:10.3233/IDA-2012-0523
15. Mastorocostas, P., Hilas, C., Varsamis, D., Dova, S.: A recurrent neural network-based forecasting system for telecommunications call volume. Appl. Math. Inf. Sci. **7**(5), 1643–1650 (2013). doi:10.12785/amis/070501
16. Oduro-Gyimah, F.K., Azasoo, J.Q., Boateng, K.O.: Statistical analysis of outage time of commercial telecommunication networks in Ghana. In: Proceedings of the International Conference on Adaptive Science and Technology, pp 1–8 ICAST (2013). doi:10.1109/ICASTech.2013.6707520
17. Jaudet, M., lqbal, N., Hussain, A., Sharif, K.: Temporal classification for fault-prediction in a real-world telecommunications network. In: Proceedings of the International Symposium on Emerging Technologies, pp. 209–214 IEEE (2005). doi:10.1109/ICET.2005.1558882
18. Zhang, X., Sugiyama, A., Kitabayashi, H.: Estimating telecommunication equipment failures due to lightning surges by using population density. In: Proceedings of the International Conference on Quality and Reliability, pp. 182–185, ICQR (2011). doi:10.1109/ICQR.2011.6031705
19. Barbosa, C., Ying, X., Day, P., Zeddam, A.: Recent progress of ITU-T recommendations on lightning protection. In: Proceedings of the 7th Asia-Pacific International Conference on Lighting (APL 2011), pp. 258–262. IEEE (2011). doi:10.1109/APL.2011.6110120
20. Schulman, A., Spring. N.: Pingin' in the rain. In: Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC 2011), pp. 19–28. ACM (2011). doi:10.1145/2068816.2068819
21. Jin, Y., Duffield, N., Gerber, A., Haffner, P., Sen, S., Zhang Z.: NEVERMIND, the problem is already fixed: Proactively detecting and troubleshooting customer DSL problems. In: Proceedings of

the 6th International Conference (Co-NEXT '10), Artic. 7. ACM (2010). doi:10.1145/1921168.1921178

22. Armstrong, J.S.: Principles of forecasting: a handbook for researchers and practitioners. Stanford University, Kluwer. http://www.gwern.net/docs/predictions/2001-principlesforecasting.pdf (2002)
23. Deljac, Ž., Kunštić, M.: A comparison of methods for fault prediction in the broadband networks. In: Proceedings of the 18th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2010), pp. 42–46. IEEE (2010)
24. Deljac, Ž., Kunštić, M., Spahija, B.: A comparison of traditional forecasting methods for short-term and long-term prediction of faults in the broadband networks. In: Proceedings of 34th international convention on information and communication technology, electronics and microelectronics (MIPRO 2011), pp. 517–522. IEEE (2011)
25. Deljac, Ž., Kunštić, M., Spahija, B.: Using temporal neural networks to forecasting of broadband network faults. In: Proceedings of the 19th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2011), pp. 1–5. IEEE (2011)
26. Demuth, H., Beale, M., Hagan, M.: Neural Network Toolbox$^{\mathrm{TM}}$ 5. The MathWorks Inc., Natick (2006)
27. Coulibaly, P., Anctil, F., Bobee, B.: Multivariate reservoir inflow forecasting using temporal neural networks, pp. 367–376. J. Hydrol. Eng., ASCE (2001)
28. Dijk, O.E.: Analysis of recurrent neural networks with application to speaker independent phoneme recognition. University of Twente, Department of Electrical Engineering, pp. 21–24. http://www.eskodijk.nl/doc/Dijk99_Recurrent_Neural_Networks.pdf (1999)
29. Dodds, E.D., Celaya, B.: Locating Water Ingress in Telephone Cables Using Frequency Domain Reflectometry. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering, pp. 324–327. IEEE (2005). doi:10.1109/CCECE.2005.1556938
30. Celaya, B., Dodds, E.D.: Single-ended DSL line tester. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering, pp. 2155–2158. IEEE (2004). doi:10.1109/CCECE.2004.1347670
31. Spahija, B., Deljac, Ž.: Proactive copper pair troubleshooting utilizing principal component analysis. In: Proceedings of the 18th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2010). IEEE (2010)
32. Free Meteo Forecast Archive (2012) (Stations: Zagreb/Grič, Rijeka/Kozala and Split/Marjan). http://freemeteo.com.hr

Željko Deljac has been working as a Senior Service Management and Quality Assurance Expert at the T-Croatian Telecom. He is currently pursuing his Ph.D. in the department of Telecommunications at the Faculty of electrical engineering and computing, University of Zagreb. His doctoral research focuses on the development and application of data mining techniques in fault management and failure prediction. His research interests include research on usage of artificial intelligence in network and services management.

Mirko Randić is an Assistant Professor at the Faculty of Electrical Engineering and Computing, University of Zagreb where he received his Ph.D. His research interests include systems, networks and service management, software systems modeling and service performance modeling. His work has been published in several peer reviewed journals, such as Software, practice & experience, Journal for Control, Measurement, Electronics, Computing and Communications, Journal of Computing and Information Technology, and other journals and book chapters.

Gordan Krčelić received his B.S. in Mechanical Engineering from the Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb and M.S. in Electrical Engineering from the Faculty of Electrical Engineering and Computing, University of Zagreb. He is working as Quality Assurance Expert at the T-Croatian Telecom. He is certified Project Management Professional and Six Sigma Black Belt. His projects were mostly process improvement focused using Six Sigma + LEAN methodology. He started his postgraduate doctoral study at the Faculty of Organization and Informatics, University of Zagreb. His research interests include processes management, service management and high-level planning.