MODELING INTERNET TRAFFIC PACKET LENGTH USING PROBDISTID: A CASE STUDY

Dragiša Miljković*,
Siniša Ilić,
Branimir Jakšić,
Petar Milić,
Stefan Pitulić

Faculty of Technical Sciences,
University of Pristina in
Kosovska Mitrovica,
Kosovska Mitrovica, Serbia

Abstract:

In this study, we apply the ProbDistID tool, a user-friendly tool based on nonlinear regression, designed for fitting probability distributions and estimating their parameters, to model internet traffic packet length using a real-world internet traffic dataset. The tool requires no a priori knowledge of input data, making it suitable for real-time fitting recognition and for data mining tasks. Our primary objectives in this case study are to identify distributions that offer the best fit for internet traffic datasets. We utilized our tool to fit and estimate parameters for eight cumulative density functions (CDFs). The fitting results are presented using utilized several model selection methods and goodness-of-fit tests to determine the most appropriate distribution model. The case study indicate that the Generalized Extreme Value (GEV) and Pareto distributions provide the most accurate fit. Our findings are presented graphically and in tabular form, demonstrating the effectiveness of ProbDistID and its potential applicability across various fields, including data mining tasks.

Keywords:

Data Mining, Internet Traffic, Nonlinear Regression, Cumulative Distribution Function, Model Selection.

INTRODUCTION

Nonlinear regression [1] (NR) is a valuable statistical method for understanding and modelling data and processes in various fields. Although there are other approaches available, such as linear regression, nonlinear regression has proven to be particularly important due to its ability to model complex relationships between variables. In this paper, we focus on the application of nonlinear regression in several domains, including fading in wireless systems [2], [3], medical signal analysis [4], [5], computer vision [6], internet traffic modelling [7], [8], and other diverse examples.

Despite the rise of numerous machine learning (ML) and artificial intelligence (AI) techniques, such as deep learning, support vector machines, and decision trees, nonlinear regression remains a relevant and valuable tool for data modelling. Its continued importance can be attributed to its flexibility and robustness in handling complex relationships and noisy data.

Correspondence:
Dragiša Miljković

e-mail:
dragisa.miljkovic@pr.ac.rs

DOI: 10.15308/Sinteza-2023-172-177

In this study, we utilize ProbDistID, a user-friendly tool with a graphical user interface (GUI), designed for fitting probability distributions and their parameters. The underlying algorithm for ProbDistID, which is based on the Levenberg-Marquardt nonlinear regression algorithm [9], [10], has already been tested and published in a previous paper. This research serves as a continuation of our efforts to apply nonlinear regression to fit data in real-world processes using ProbDistID.

The primary objectives of this study are to identify the best modelling distribution for internet traffic and to analyse the applicability of eight appropriate cumulative density functions (CDFs) for fitting experimental traffic data. Our contributions include analysing the selected CDFs for fitting the experimental traffic data (traces) and determining the best-fitting distribution to accurately represent internet traffic patterns in the given context.

2. PROBDISTID: A BRIEF OVERVIEW

ProbDistID[1] is a web-based tool specifically developed to simplify the process of identifying probability distributions and their parameters for user-selected scenarios. It streamlines the process of selecting appropriate probability distributions and estimating their parameters, making it a valuable asset for data-driven decision-making and data mining tasks across various domains. The underlying algorithm description and pseudocode for ProbDistID can be found in reference [1].

Built upon previous work [2], ProbDistID presents an enhanced version of the approach, which was validated in a prior study using a large set of 38,400 randomly generated signals with five different probability distributions commonly employed to model wireless fading, such as Gamma, Rayleigh, Rician, Nakagami-m, and Weibull.

Key upgrades in the new version include the development of a user-friendly GUI interface (Figure 1), replacing the previous CLI tool; the implementation of the entire algorithm in R as opposed to MATLAB for signal generation; the addition of more probability distributions (currently 14, with support for more in the pipeline), and the incorporation of more model selection criteria and goodness-of-fit tests (also being continually enhanced).

ProbDistID offers a range of capabilities, such as signal generation using any of the supported distributions with user-selected parameter values, signal recognition, and batch signal recognition of a set of inputs (allowing users to upload data in RData, CSV, JSON, XML, and plain text formats). It also provides a tabular presentation of fitting results, making it easier for users to understand and interpret the results. Additionally, the tool provides a tabular representation of the estimated parameters, which assist users in selecting the best-fitting model for their data.

A case study demonstrates the effectiveness and utility of ProbDistID in describing internet traffic, highlighting its potential for application in a wide array of fields and scenarios.

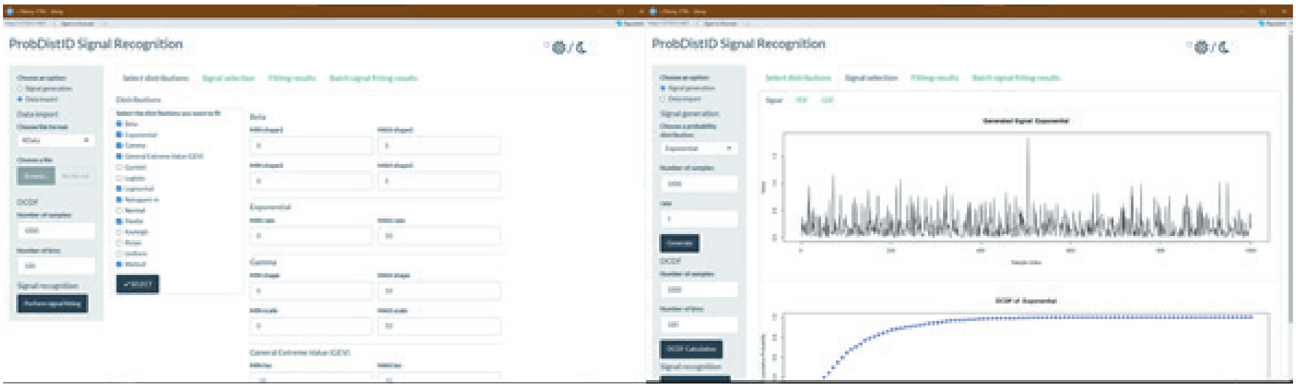
ProbDistID Signal Recognition
Select distributions
Signal selection
Fitting results
Batch signal fitting results
ProbDistID Signal Recognition
Select distributions
Signal selection
Fitting results
Batch signal fitting results

Figure 1 - ProbDistID Graphical User Interface.

[1] Available online: https://x9u9lx-despot.shinyapps.io/ProbDistID/

3. DATASET: MAWI WORKING GROUP TRAFFIC ARCHIVE

In this study, we utilized the MAWI Working Group Traffic Archive [11], which is a collaborative project between Japanese network research and academic i nstitutions. Their goal is to analyse the performance of networks and networking protocols within Japan's wide area networks. The dataset is made up of daily traces taken at the transit link of WIDE to the upstream ISP, as a component of the "Day in the Life of the Internet" project. This link functions at a capacity of 1 Gbps.

We selected the DumpFile: 202304301400.pcap [12], generated between Sun Apr 30 14:00:00 2023 and Sun Apr 30 14:15:00 2023. The file contains 53,857,184 packets captured during a 15-minute period, with a size of 4 GB and an average rate of 306.45 Mbps. We used the tshark network protocol analyser for data pre-processing and analysis. Tshark is a versatile tool that allows users to capture packet data from a live network or read packets from a previously saved capture file, decoding and printing the packet information or writing the packets to a file. In our study, we employed tshark to filter out Ethernet packages, limiting the packet sizes in our dataset to a maximum of 1500 Bytes, which corresponds to the Maximum Transmission Unit (MTU) size.

This case study focuses on modelling internet traffic packet length using the MAWI Working Group Traffic Archive dataset. However, the methodology can be extended to live packet capturing and fitting distributions with real-time data. This information is valuable for various applications, including:

- Network congestion prediction [13]: By understanding the distribution of packet lengths, network administrators can anticipate congestion and take appropriate measures to prevent bottlenecks.
- Network performance optimization [14]: Analyzing the characteristics of the traffic can help in adjusting parameters such as buffer sizes and transmission rates to optimize network performance.
- Anomaly detection and network security [7]: Comparing real-time traffic with expected patterns based on historical data can aid in detecting anomalies or potential attacks on the network, thereby enhancing its security.

4. METHODOLOGY

Our approach is based on the rationale that ProbDistID provides an efficient and user-friendly way to identify the best-fitting probability distribution for a given dataset.

In the initial phase, we examined packet capture files using GUI tools such as Wireshark. However, we found that data analysis tools like the R programming language offered more flexibility and efficiency. To process the data more rapidly, we employed the tshark command-line tool. We used tshark commands to extract the lengths of all Ethernet frames. The resulting file is very large (200MB), but a significant amount of data in it is redundant, there are several million rows, but only 1500 possible lengths. To improve efficiency, we summarized the data as frequencies of occurrences of packet data lengths.

Based on the literature [7], [8], [14], we selected the following probability distributions for fitting: Beta, Exponential, Gamma, Generalized extreme value (GEV), Log-normal, Nakagami, Pareto, and Weibull. We then used the data obtained in the previous step as input in ProbDistID tool. Subsequently, we calculated the discrete cumulative distribution function (DCDF) to obtain the cumulative distribution of packet sizes, and fitted the distribution models accordingly.

In order to determine the best-fitting model for our data, we employed several model selection methods and goodness-of-fit tests. These include: Akaike Information Criterion (AIC) [15], [16], a that balances the goodness-of-fit with the complexity of the model, Bayesian Information Criterion (BIC), similar to AIC, but with a stronger penalty for model complexity, Residual Sum of Squares (RSS), Root Mean Square Error (RMSE), R-squared, and Adjusted R-squared (a modified version of R-squared that adjusts for the number of predictors in the model). By applying these methods and tests, we were able to identify the most appropriate distribution model for the internet traffic packet length dataset.

5. RESULTS AND DISCUSSION

In this section, we present a comprehensive analysis of the results obtained from our application of ProbDistID to model internet traffic packet length. We provide visual representations in the form of plots for the fitted distributions, as well as detailed tabular summaries of the values for model selection and goodness-of-fit tests, along with the fitted parameters.

Figure 2 illustrates the cumulative distribution of packet lengths and bytes in the original dataset, with a focus on packets up to 1500 bytes (Ethernet packages). This figure offers a clear visualization of the distribution of packet lengths and bytes, revealing the presence of very large and very small packets as the spike points. This is a typical characteristic of Internet traffic, referred to as Internet Mix (IMIX) [17], [18].

To ensure readability and avoid cluttering the plots, we present the fitting results in two separate figures. Figure 3 showcases the results of fitting the GEV, Nakagami, Gamma, and Beta distributions, while Figure 4 offers a comparison between the Weibull and Lognormal distributions. These two figures enable a clear evaluation of the eight distributions' performance in fitting the data.

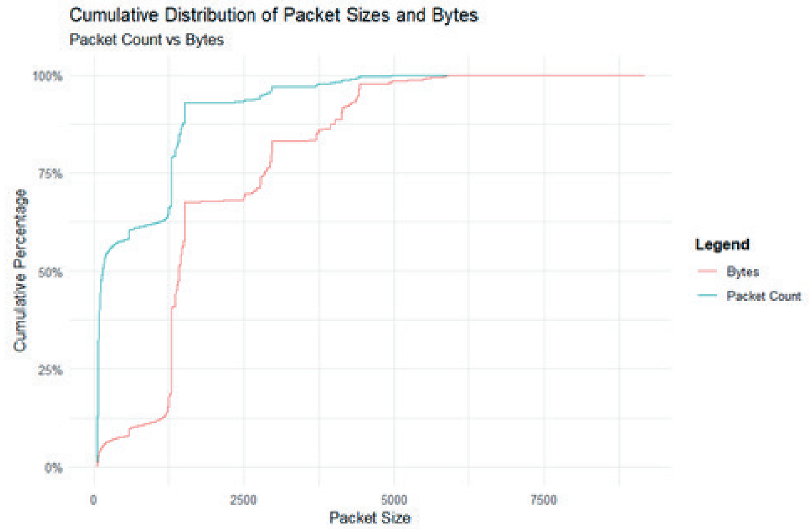

Figure 2 - Cumulative Distribution of Packet Sizes and Bytes.

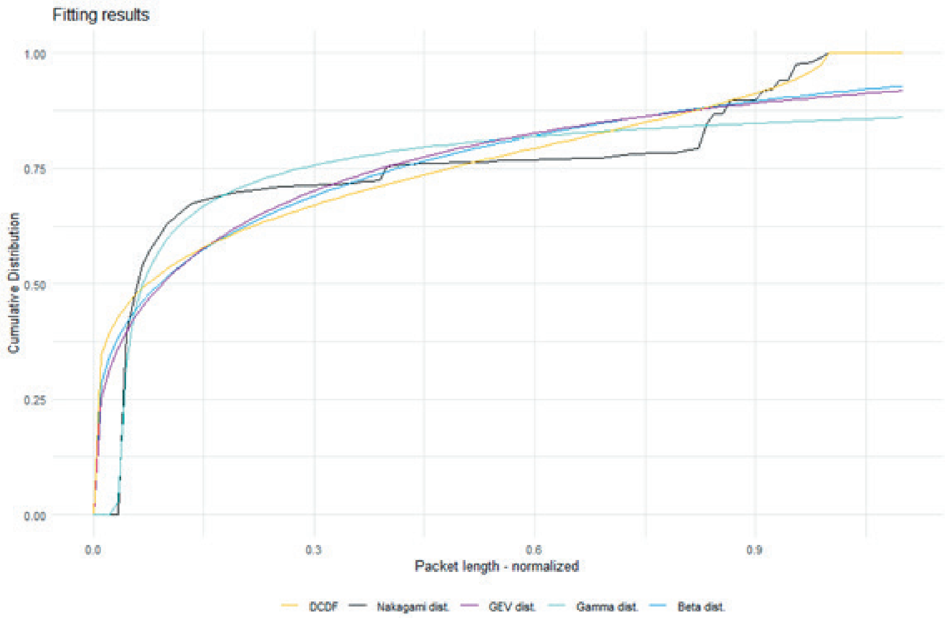

Figure 3 - Fitting of GEV, Nakagami, Gamma, and Beta distributions.

Complementing the visual representations, we provide detailed tables that encapsulate our findings. Table 1 lists the parameter values of the fitted distributions.

Table 2 presents the values of model selection methods and goodness-of-fit tests, such as AIC and BIC criteria, where lower values indicate better fit. From these results, it becomes evident that the GEV and Pareto distributions provide the best fit.

Table 1 - Fitted distribution parameters.

Distribution	Parameter 1	Parameter 2	Parameter 3
Pareto	shape = 0.0215	scale = 0.5265	
Weibull	shape = 0.4997	scale 0.1927	
Exponential	rate = 4.057		
Log-normal	meanlog = 0.023	sdlog = 1.966	
Nakagami	shape = 0.2914	scale = 0.1352	
GEV	shape = 2.23463	scale = 0.03438	location = 0.0478
Gamma	shape = 0.3319	scale = 1.0104	
Beta	shape1 = 0.1941	shape2 = 0.5430	

Table 2 - Criteria for model selection.

Distribution	AIC	BIC	RSS	RMSE	R squared	Adj. R squared
Pareto	-242.7	-234.9	0.48	0.069	0.87	0.867
Weibull	-212.54	-204.73	0.65	0.081	0.824	0.82
Exponential	-111.8	-106.6	1.83	0.135	0.51	0.5
Log-normal	-221.9	-214	0.59	0.077	0.84	0.836
Nakagami	-203	-195	0.72	0.085	0.807	0.803
GEV	-250.7	-240.2	0.44	0.066	0.882	0.878
Gamma	-206.4	-198.6	0.69	0.083	0.813	0.809
Beta	-205.9	-198.0	0.7	0.085	0.811	0.808

The IMIX phenomenon also explains why the Generalized Extreme Value (GEV) and Pareto distributions offer the most accurate fit. The Pareto distribution, often referred to as the 80:20 rule, suggests that 80% of outcomes result from 20% of causes. In the context of Internet traffic, this concept is particularly relevant. The GEV distribution is commonly employed to model extreme events, such as the largest or smallest values in a dataset. One reason why the GEV distribution fits well with the cumulative distribution of internet packet sizes is the heavy-tailed behaviour exhibited by the packet size distribution. This means there is a relatively high probability of observing numerous small and large packet sizes, which can be considered extreme events.

Our study offers valuable insights into the statistical modelling of internet traffic, paving the way for future experimentation and analysis. Nonetheless, further research is necessary to explore this topic in greater depth.

6. CONCLUSION

In this paper, we have presented a show case of using the ProbDistID tool to fit cumulative probability distributions to a real-world internet traffic dataset. In our experiment, eight probability distributions were used to model a large dataset of internet traffic data, with a focus on the packet lengths. The obtained results indicate that the Generalized Extreme Value (GEV) and Pareto distributions offer the best fit for the data.

The presented approach may have practical usage in computer networks, as network administrators can utilize similar tools to get better insights in system throughput, load, and security threats.

However, it is important to note that this was a limited experiment, and the best-fitting distribution models we identified might not necessarily be the optimal models for other datasets. Further research is needed, preferably using even larger datasets, to validate and extend these findings.

The results add to the previously successful application of the tool for fading in telecommunications, and shows that the presented tool can be used in numerous fields. In the pipeline, as ProbDistID tool continues to evolve, the authors plan to apply the tool to real-time internet traffic analysis and to explore its application in other domains.

7. ACKNOWLEDGEMENTS

The authors of this paper would like to express their gratitude to the organizers of the **Sinteza 2023** conference for covering the costs of registration fees.

8. REFERENCES

[1] D. M. Bates and D. G. Watts, *Nonlinear regression analysis and its applications*. Wiley, 1988.

[2] D. Miljković, S. Ilić, D. Radosavljević, and S. Pitulić, “Application of nonlinear regression in recognizing distribution of signals in wireless channels,” *Proc. Est. Acad. Sci.*, vol. 72, no. 2, pp. 105–114, 2023.

[3] S. Panic, M. Stefanovic, J. Anastasov, and P. Spalevic, *Fading and Interference Mitigation in Wireless Communications*. Boca Raton: CRC Press, 2015. doi: 10.1201/b16275.

[4] A. Bhattacharjee, S. Saha, S. A. Fattah, W.-P. Zhu, and M. O. Ahmad, “Sleep apnea detection based on rician modeling of feature variation in multiband EEG signal,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1066–1074, 2018.

[5] J. L. Peacock and P. J. Peacock, “Probability and distributions,” in *Oxford Handbook of Medical Statistics 2e*, J. L. Peacock and P. J. Peacock, Eds., Oxford University Press, 2020, p. 0. doi: 10.1093/med/9780198743583.003.0007.

[6] L. Zhang *et al.*, “Nonlinear Regression via Deep Negative Correlation Learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 982–998, Mar. 2021, doi: 10.1109/TPAMI.2019.2943860.

[7] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” presented at the Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, 2010, pp. 267–280.

[8] K. M. Rezaul and A. Pakštas, “Web traffic analysis based on EDF statistics,” *transformation*, vol. 9, no. 1, p. 14, 2006.

[9] T. V. Elzhov, K. M. Mullen, A.-N. Spiess, and B. Bolker, “minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds.” Apr. 13, 2022. Accessed: Nov. 13, 2022. [Online]. Available: https://CRAN.R-project.org/package=minpack.lm

[10] H. P. Gavin, “The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems,” *Dep. Civ. Environ. Eng. Duke Univ.*, vol. 19, 2019.

[11] “MAWI Working Group Traffic Archive.” http://mawi.wide.ad.jp/mawi/ (accessed May 06, 2023).

[12] “Traffic Trace Info.” http://mawi.wide.ad.jp/mawi/samplepoint-F/2023/202304301400.html (accessed May 06, 2023).

[13] Y. Yang, Y. Fan, and J. O. Royset, “Estimating probability distributions of travel demand on a congested network,” *Transp. Res. Part B Methodol.*, vol. 122, pp. 265–286, 2019.

[14] E. R. Castro, M. S. Alencar, and I. E. Fonseca, “Probability density functions of the packet length for computer networks with bimodal traffic,” *Int. J. Comput. Netw. Commun.*, vol. 5, no. 3, p. 17, 2013.

[15] J. Kuha, “AIC and BIC: Comparisons of Assumptions and Performance,” *Sociol. Methods Res.*, vol. 33, no. 2, pp. 188–229, Nov. 2004, doi: 10.1177/0049124103262065.

[16] K. P. Burnham and D. R. Anderson, “Multimodel Inference: Understanding AIC and BIC in Model Selection,” Sociol. Methods Res., vol. 33, no. 2, pp. 261–304, Nov. 2004, doi: 10.1177/0049124104268644.

[17] W. John and S. Tafvelin, “Analysis of internet backbone traffic and header anomalies observed,” presented at the Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007, pp. 111–116.

[18] K. Pentikousis and H. Badr, “Quantifying the deployment of TCP options-a comparative study,” *IEEE Commun. Lett.*, vol. 8, no. 10, pp. 647–649, 2004.