2018 IEEE 16th Int. Conf. on Dependable, Autonomic & Secure Comp., 16th Int. Conf. on Pervasive Intelligence & Comp., 4th Int. Conf. on Big Data Intelligence & Comp., and 3rd Cyber Sci. & Tech. Cong.

Substituting Missing Values in End-to-end Internet Performance Measurements using k-Nearest Neighbors

Saqib Ali[1,2], Guojun Wang[1,*], Xiaofei Xing[1], and Roger Leslie Cottrell[3]
[1]*School of Computer Science and Technology, Guangzhou University, Guangzhou, P. R. China, 510006. Emails: {saqibali, csgjwang, xingxf}@gzhu.edu.cn*
[2]*Department of Computer Science, University of Agriculture, Faisalabad, Pakistan, 38000. Email: saqib@uaf.edu.pk*
[3]*Stanford Linear Accelerator Center, P.O. Box 4349, Palo Alto, California, 94309-4349. Email: cottrell@slac.stanford.edu*
[*]*Correspondence to: csgjwang@gzhu.edu.cn*

***Abstract*—PingER (Ping End-to-end Reporting) is a worldwide end-to-end Internet performance measurement framework running for the last 20 years and led by the SLAC National Accelerator Laboratory USA. The objective of the project is to monitor the performance of the Internet links around the world using the ubiquitous ping facility. Currently, the framework comprises of about 50 active Monitoring Agents (MAs) in 20 countries of the world. These MAs probe 700 remote sites located in 170 countries of the world. They are covering an area containing over 98% of the world's population. Currently, the size of the PingER data is about 60 GB stored in 100,000 flat files with a compression ratio of 5:1. The data is of an historical nature and very useful for fine-grained Internet performance analysis. However, the data contains missing values due to congestion, queuing overflow, faulty hardware or software and unavailability of MAs & remote sites. These missing values affect the quality of the Internet performance analysis. The objective of this paper is to substitute the missing values using the k-Nearest Neighbors algorithm (k-NN) and compare the estimation with the statistical method. Therefore, PingER historical data is first transformed into CSV format using a PingER data dimensional model. Afterward, missing values are imputed, using the statistical method and the k-NN algorithm, on data containing the different percentages of missing values. The results conclude that the k-NN algorithm is best suited for the substitution of missing values in the PingER data as compared to the method based on the statistical procedure.**

***Index Terms*—Internet performance monitoring, PingER, missing value, k-Nearest Neighbors**

1. Introduction

Internet performance measurement infrastructures periodically measure the metrics of Internet links by running different network measurement tests. The key Internet performance measurement platforms available in the literature are SamKnows, BISmark [1], Dasu [2], Netradar [3], Portolan [4], RIPE Atlas [5], and perfSONAR [6] originally partially based on the PingER architecture [7]. The detailed taxonomy of Internet performance measurement platforms is available in [8]. These platforms use `ping`, `mtr`, `cron`, `ntp`, `dig`, `netstat`, `iperf`, and `traceroute` commands to measure the performance of the Internet links. The data collected by these infrastructures are then used to understand the comprehensive view of the Internet. For example, the Federal Communications Commission (FCC)[1] which is an inter-state communication regulator in the USA, is using the Internet performance measurements to analyze the performance of broadband providers to regulate the industry in the country [8]. Further, such datasets are used to explain the impact of Powerboost [9], ISP characterization [10], [11], Broadband Mapping [12], Broadband Performance [13], Internet Congestion [14], peer-to-peer (P2P) streaming [15], and Connectivity of IPv4 & IPv6 [16]. Thus, such datasets have a significant impact in revolutionizing the performance of the Internet links around the globe.

Like other real-time performance measurement scenarios, Internet Performance measures also suffer from missing or incomplete data. For example, UC Irvine[2] provides a Machine Learning repository of 370 datasets for benchmarking Machine Learning Algorithms which contains more than 40% of missing values [17]. Similarly, the missing values percentage is quite high in Internet performance measurements as packets get lost in a network due to congestion, bottleneck links, queuing overflow, faulty network hardware or drivers and due to the measurement host or target host being unavailable due to end host outages. Further, sometimes Internet packets are deliberately dropped by the routers through efficient network management policies [7]. All these factors contribute to missing values in the Internet performance measurements. These missing values seriously affect the measurements by triggering biased in Internet performance analysis. Moreover, the effect becomes drastic when the missing values are not distributed randomly. The most common approach used to handle missing values in

1. https://www.fcc.gov/
2. http://uci.edu/

978-1-5386-7518-2/18/$31.00 ©2018 IEEE
DOI 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.000-8

Internet performance estimation is to omit such observations. Thus, the analysis is carried out only on the complete dataset. Consequently, this approach will lead to the loss of effectiveness in the Internet performance analysis.

In this paper, the missing values in the Internet performance measurements are substituted using k-Nearest Neighbors (k-NN). This is because the Internet performance measurements contain historical data covering several years. For example, the dataset used in this paper is from PingER (Ping End-to-end Reporting)[3] which is an Internet End-to-end Performance Measurement (IEPM)[4] framework led by SLAC National Accelerator Laboratory [5] [7]. It is running for the last 20 years and contains a multi-domain historical dataset (e.g., bandwidth, delay, jitter and loss) of nearly 20 years from 700 nodes in 170 countries of the world [18]. The analysis performed by removing the missing values or replacing the missing values by other statistical methods from the PingER dataset introduces bias in the estimation. On the other hand, k-NN algorithm provides a precise estimation of the missing values. The reason is that the k-NN algorithm uses a weighted average feature to provide a better local estimation of the missing values as compared to the other statistical methods. Thus, making k-NN algorithm best suits for substituting the missing values in the PingER dataset.

The paper comprises three major contributions. First, the PingER historical flat files containing the missing values are converted into Comma Separated Value (CSV) file format using PingER dimensional model. Secondly, the dimensional model containing different percentages of missing values i.e., 5, 10, 15 and 20% of missing values are replaced by using row average and k-NN algorithm. Finally, the performance analysis is carried out to conclude that the k-NN algorithm substitutes the missing values with more realistic ones as compared to row average. Thus, improving the overall estimation of the Internet performance analysis in the PingER dataset.

The remaining paper is organized as follows. Related work is discussed in Section 2. Section 3 describes the PingER framework. Section 4 formulates the problem. The proposed approach for substituting missing values is explained in Section 5. Performance evaluation is outlined in Section 6, and finally, Section 7 concludes the paper.

2. Related Work

Internet performance measurement frameworks use dedicated probes that periodically run network measurement tests to mine the performance of Internet links on wired and mobile networks [8], [19]. Currently, the key Internet performance measurement frameworks are SamKnows[7], BISmark [1], Dasu [2], Netradar [3], Portolan [4], RIPE Atlas [5], and perfSONAR [6] originally partially based on PingER architecture [7]. The detail discussion on these platforms based on their deployment strategy, probing methodology, features, and research impacts is summarized by Bajpai & Schonwalder in [8]. The data collected by these platforms is used to analyze the end-to-end performance of the links, quantifying the digital divide among the regions, detecting congested routes, identifying last mile problems and evaluating the impact of major events i.e., cable cuts, tsunamis, earthquakes, and social upheavals.

3. www-iepm.slac.stanford.edu/pinger/
4. http://www-iepm.slac.stanford.edu/
5. https://www6.slac.stanford.edu/
6. https://confluence.slac.stanford.edu/display/IEPM/PingER+Regions
7. https://www.samknows.com/

TABLE 1: PingER monitored countries and populations by region[6]

Region	No. of countries	Population of the region (Millions)	% of world population
Africa	50	988	14.57
Balkans	10	69	1.02
Central Asia	9	80	1.18
East Asia	4	1534	22.62
Europe	31	527	7.76
Latin America	21	557	8.21
Middle East	13	226	3.33
North America	3	342	5.05
Oceania	4	33	0.49
Russia	1	142	2.09
S.E. Asia	11	578	8.52
South Asia	8	1585	23.37
Total	165	6660	98.21

Among above mentioned Internet performance measurement frameworks, PingER is led by SLAC and is running for the last 20 years. It covers a geographical area containing over 98% of the world's population as shown in Table 1. The detail discussion on the PingER framework is available in Section 3. It has gathered interesting multi-domain historical performance data of the Internet links worldwide. However, the data suffer from the common missing value problem, like other real-time measurements, which affects the accuracy of the estimation [17], [20], [21]. Normally, the analysis is performed on such data by removing the missing values completely. However, this approach introduces bias in the estimated values.

Another approach in substituting the missing values without losing the information available in the missing observation is the imputation method [17], [22]. In this technique, missing values are estimated using the known association among the complete set of values in the dataset. There are many standard statistical procedures (i.e., List-wise Deletion (LD), Row Average or Person Mean Substitution, Hot Deck, Multiple Imputation (MI), and Regression Imputation) that are available for this purpose [23], [24], [25]. Similarly, many machine learning algorithms are also available for handling missing data imputation (e.g., k-NN, Self-organizing Maps (SOM), Multi-layer Perceptron (MLP), and Decision Tree (DT) construction algorithms) in different problem domains [26], [27], [28].

The machine learning algorithms are more suitable for substituting missing values because of their flexibility and power of capturing higher order interaction among the

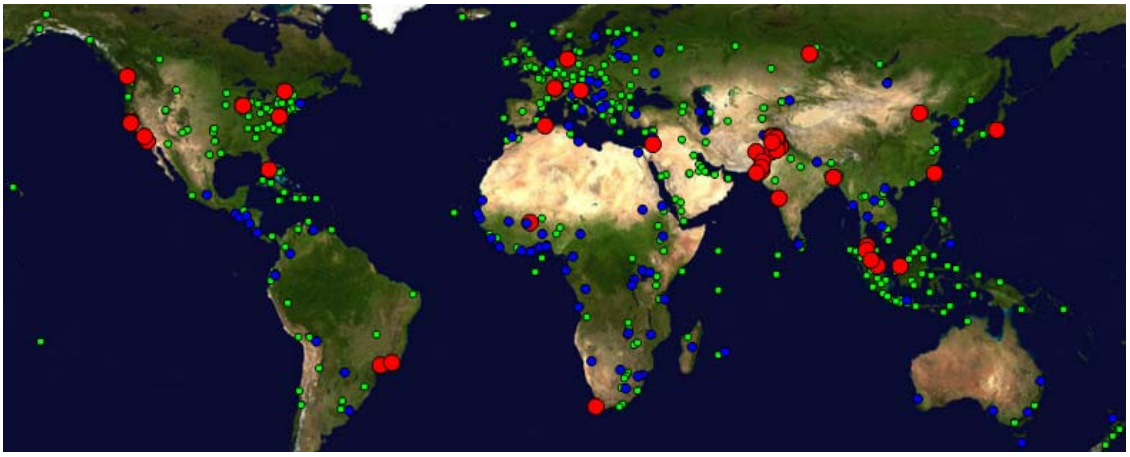

Figure 1: Worldwide geographical locations of PingER MAs, beacons and remote sites

completed values as compared to standard statistical techniques [20]. Since PingER dataset is of historical nature, therefore, k-NN which is a machine learning algorithm provides a precise estimation of the missing values. This is because k-NN uses a weighted average feature to provide a better local estimation of the missing values as compared to the statistical methods.

3. The PingER Framework

PingER is a framework to monitor end-to-end performance of the Internet links worldwide. It was specially designed in 1995 by SLAC to facilitate modern High Energy Nuclear and Particle (HENP) physics experiments taking place among sites such as SLAC, the Brookhaven National Laboratory (BNL)[8] and the European Center for Particle Physics (CERN)[9]. However, for the last decade, the objective of the project is to monitor the performance of the Internet links around the world. Currently, PingER comprises of about 50 active Monitoring Agents (MAs) in 20 countries of the world [29]. These MAs send `ping` probes to 700 remote sites located in 170 countries of the world. As a result, 10,000 MA-remote sites are developed covering an area containing over 98% of the world's population as shown in Table 1. The North Korea, Central African Republic, Chad and Guinea-Bissau, each more than one Million population, are the only countries which do not have any MAs or remote sites. Similarly, Figure 1 indicates the geographical locations of PingER MAs (colored red), Beacons (monitored by most MAs are colored blue) and Remote sites (colored green) which cover nearly 98% of the of Internet users in the world.

3.1. PingER Methodology

In PingER, each sample measurement set is sent every 30 minutes. The MA goes through its list of remote sites and for each sends an initial 100-byte ping that is used to prime the routing caches and is discarded. This is followed by sending up to thirty 100-byte pings at one second intervals until ten responses are received. This is then repeated for 1000-byte pings. Thus, each MA-remote site pair only produces a little extra traffic of 100 bits/s on average making PingER a lightweight Internet active performance measuring framework [7]. The data collected for each set of pings consists of an MA name, list of target remote sites, IP addresses of MA & target remote sites, payload in ping request, minimum Round Trip Time (RTT), maximum RTT and average RTT [30]. Afterward, the data archived by each MAs is pulled daily by the SLAC to a centralized repository of text archives. Sixteen different network performance metrics are extracted including packet loss, jitter, unreachability, throughput, directivity, unpredictability, and quiescence from the collected data.

3.2. Significance of PingER Dataset

Currently, the size of PingER data repository consists of 100,000 flat files of 60 Gigabytes which is growing at the rate of 800 Megabytes per month. The historical compressed data files can be downloaded from the Pingtable web interface [10] or by anonymous FTP[11]. The data is in the form of tab-separated-value (.tsv) on an hourly, monthly and yearly basis. The data is valuable for fine-grained analysis to predict current and future end-to-end performance, bottleneck links, queuing effect, and congested routes. For example, historical throughput trendline for SLAC to world region is shown in Figure 2. Although it is clear from the graph that Internet performance of Africa is improving, however, it still lags the rest of the world. The major event that caused this upgrade was 2010 FIFA World Cup which not only brought three million football fans to Johannesburg but was also accompanied by the landing of new submarine

8. https://www.bnl.gov/world/
9. http://home.cern/
10. http://www-wanmon.slac.stanford.edu/cgi-wrap/pingtable.pl
11. ftp://ftp.slac.stanford.edu/users/cottrell/

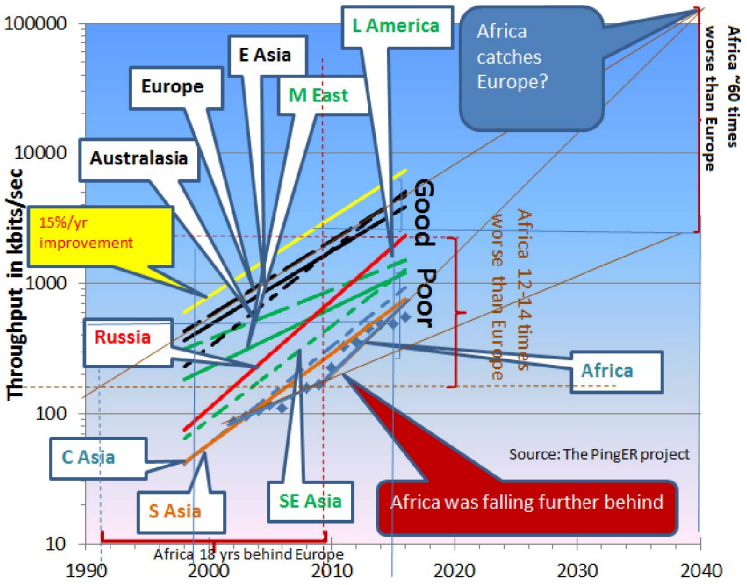

Figure 2: Throughput trendline from SLAC to the world regions

cables in the region. As a result, Africa's Internet went from 18 years behind Europe in 2010 to 16 years behind in 2015. It may even catch up Europe by 2030 [31]. However, this uptick in growth may be a temporary after effect of the 2010 World Cup. The current PingER throughput trendline shows that Africa is falling further behind Europe i.e., nearly 60 times worse than Europe by 2040. PingER data can also reveal information about major events like fiber cuts, tsunamis, and social upheavals. Further, PingER monitoring data is also used to develop case studies regarding the Internet performance in the different regions of the world i.e., Africa, Latin America, East Asia, Middle East, South Asia & South East Asia[12]. This indicates the importance of the PingER historical data in Internet performance analysis of the world.

3.3. Issue: Handling Missing Values

The PingER dataset contains missing values which can affect the quality of the analysis. The reason for missing data in PingER is because of packet loss triggered by the congestion, bottleneck links, queuing overflow, faulty network hardware or drivers, and due to the MAs or remote site being unavailable owing to end host outages. All these factors contribute to missing values in the PingER dataset which is Missing Completely at Random (MCAR) [22], thus affecting the quality of the dataset for critical analysis. The quality of PingER dataset can be improved by substituting the missing values. Several standard statistical and machine learning imputation methods are available as discussed in Section 1 & 2. However, this paper focus on the use of k-NN which is a machine learning algorithm to handle missing values in PingER dataset.

4. Problem Formulation Using k-NN

k-NN is an instant-based machine learning algorithm where the objective function is only approximated among the neighbors such that they minimize some distance measure. Further, the computation takes place only at the time of classification or regression. The known values of the instants act as a training set for the algorithm. However, no prior training is required by the algorithm to generate an explicit model or classifier and works fine with both qualitative and quantitative type of datasets [17], [32]. Another important feature of k-NN is that it can successfully predict accurate results even with the increasing percentage of the missing values in the data. Further, k-NN imputation is also robust for noisy datasets and is less sensitive to the selection of the number of nearest neighbors [21]. This makes k-NN the best choice for handling missing values in the PingER dataset.

Consider a feature vector x of average Round Trip Time (RTT)[13] values of PingER dataset containing different percentages of the missing values. Suppose that the j^{th} instant of the feature vector x is missing. Euclidean distances from x to all training instance are calculated such that they minimize some distance measure and are arranged in the ascending order while excluding the missing instances in the feature vector x [21], [32]. Let V be the set of k nearest neighbor of x feature vector arranged in the increasing order of the distances and is define by the Equation 1.

$$V = \{\nu_k\}_{k=1}^{k} \tag{1}$$

12. https://confluence.slac.stanford.edu/display/IEPM/PingER+Case+Studies

13. http://www-wanmon.slac.stanford.edu/cgi-wrap/pingtable.pl

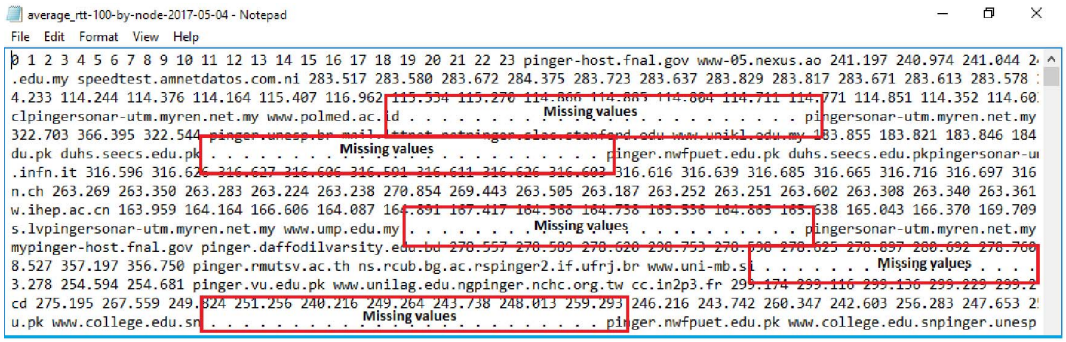

Figure 3: Average RTT file with missing values

The optimal value of k (nearest neighbor) is selected using cross-validation technique. Afterward, the substituted value for the j^{th} instant of the feature vector x is calculated using mean estimation of k nearest neighbors (when all neighboring instances are considered with the same level of significance) as defined in Equation 2.

$$\tilde{x}_j = \frac{1}{k}\sum_{k=1}^{k} v_{kj} \tag{2}$$

However, the weighted mean (i.e., by assigning greater weight to nearest neighbors as explained by [33]) of k nearest neighbors which is a refinement to the mean estimation, is calculated by using Equation 3.

$$\tilde{x}_j = \frac{1}{kW}\sum_{k=1}^{k} w_k v_{kj} \tag{3}$$

and

$$W = \sum_{k=1}^{k} w_k \tag{4}$$

where w_k in Equation 3 & 4 indicates the corresponding weight to the k^{th} nearest neighbor of the j^{th} instant of the feature vector x. Further, w_k in case of Euclidean distance based metrics is calculated by using Equation 5 [34].

$$w_k = \frac{1}{d(x, v_k)^2} \tag{5}$$

Finally, the missing j^{th} instant of the feature vector x of PingER dataset is substituted using the weighted mean as given by Equation 3 using the k-NN algorithm.

5. Proposed Approach for Substituting Missing Values in PingER Dataset

The proposed approach of substituting the missing values in PingER dataset consists of the following steps.

A. Extraction

The PingER server at SLAC fetches the zipped raw data from the remote monitoring nodes daily. It compresses and stores them into flat files with `txt.gz` extension. Each file name consists of the name of the performance metric, packet size (100 or 1000 bytes), node, and date i.e., `average_rtt-100-by-node-2017-05-05.txt.gz`. The data is retrieved via anonymous FTP [14] server at SLAC.

B. Missing values

The extraction of tar file contains raw flat files for each day with the hourly data for all the days in the PingER archive. Each flat file is appeared as shown in Figure 3. The format of the file consists of a first line "`0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23`" followed by lines of the form "`pinger-host.fnal.gov www-05.nexus.ao 241.197 240.974 241.044 242.636 240.974 240.954 240.990 249.199 241.093 241.464 241.091 241.079 241.091 241.170 241.088 241.046 241.039 241.037 240.999 240.984 241.236 240.988 241.039 240.967 pinger-host.fnal.gov www-05.nexus.ao`" per day/host pair where 0 to 23 are 24 tokens (one for each hour) between the initial and final `src_name` and `tgt_name` tokens. In Figure 3, the missing data is shown by a period (.) followed by a space i.e., "`icfamon.dl.ac.uk lns62.lns.cornell.edu 108.871 . . 107.671 . 109.657 . . . . . . . . . 108.892 . . . . . 109.620 icfamon.dl.ac.uklns62.lns.cornell .edu`". These missing values in the data are due to the ping request or reply gets lost in the network. It may happen due to congestion, bottleneck links, queuing overflow, faulty network hardware or drivers. In some cases, ICMP packets are deliberately dropped by the routers as a part of efficient network management policies. Further, sometimes the measurement host may not have been working or target host may not have replayed to the pings due to outages. In all cases, there is no value for the performance metric being measured and a dot is recorded in the system. Total missing values or number of dots per year in PingER data files are shown in Figure 4.

14. ftp://ftp.slac.stanford.edu/users/cottrell

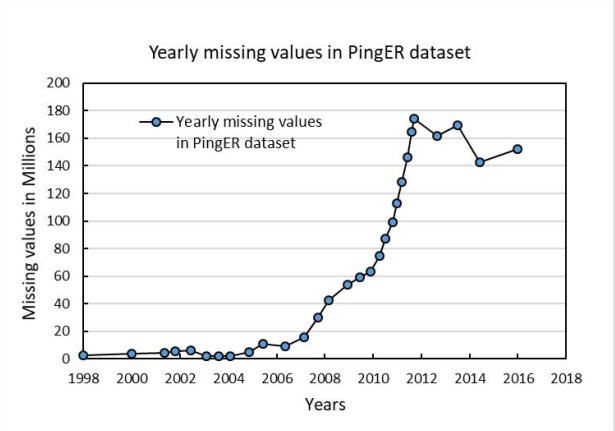

Figure 4: Missing values in PingER dataset

C. Transformation

In this step, a MapReduce [35] data flow is executed using the Cloudera Distribution of Hadoop (CDH) [36] and SciCumulus [37] to transform the flat files of PingER data into CSV format. The MapReduce data flow consists of two steps. In the first step, mapper reads each raw PingER text file for a given metric and day of the year. During the second step, it transforms the raw text files of data into a CSV format according to the PingER dimensional data model [38].

6. Performance Evaluation

A. Evaluation Setup

The evaluation setup consists of a cluster of 4 virtual machines running on Red Hat Linux 7.2[15]. Each virtual machine has four cores, 16 Gigabytes of Random Access Memory (RAM), and 220 Gigabytes of storage. After transforming the raw text files into a PingER dimensional data model, missing values are imputed using the k-NN algorithm as discussed in Section 4 and row average values. The monitoring node is `pinger.slac.stanford.edu` located in SLAC, California, USA. The target node is `www.startel.ao` located in Luanda, Angola, Africa. The monthly average RTT metric is used in the analysis. The dataset ranges from December 2003 to April 2017. In order to compare the effectiveness of the substitution method, a reference dataset of average RTT is estimated by removing all the missing values from the data; the estimation process is described as listwise or case deletion [20]. Later, 5 to 20% values are deleted from the data at random to generate 5, 10, 15 and 20% test dataset of missing values. Each missing value dataset is substituted with k-NN algorithm and row average to recover the missing values in the test dataset. Afterward, the substituted values are correlated with the original data using Normalized Root Mean Square Error (NRMSE) to verify the accuracy of the estimation process between k-NN algorithm and row average [21]. The results are discussed in the next section.

15. https://www.redhat.com

B. Performance Results

The performance of k-NN and row average is evaluated on average RTT values of PingER data with 5, 10, 15 and 20% of the missing values. In row average, missing values are simply replaced with the row mean whereas in k-NN, missing values are replaced with estimated values of RTT using 1, 3, 5, 7, 9, 11, 13 and 15 nearest neighbors. The optimal value of k is 7-11 selected by cross-validation. Consider the results as shown in Figure 5, k-NN surpasses row average in estimating the missing values accurately. At low percentage of missing values i.e., 5 or 10% the average deviation from the true value is only 3 to 6%. This is because the missing values are few, consequently, row average precisely captures the hidden pattern information in the average RTT values and correctly recovered the missing values in PingER dataset. However, as the percentage of missing values increases from 10 to 20%, row average failed to capture the hidden pattern information in the data. As a result, the NRMSE is significantly higher when compared with reference dataset. Thus, row average leads to the unsatisfactory estimation of the missing values in the PingER data.

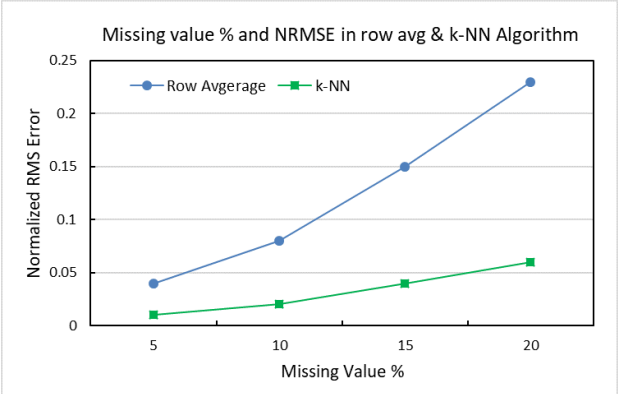

Figure 5: NRMSE in row average and k-NN algorithm

On the other hand, the performance of k-NN is quite satisfactory even with the high percentage of missing values i.e., 15 and 20%. The results are shown in Figure 5 & 6. At 20% of missing values, the average derivation in the estimated values is less than 6% from the original values. This, indicates the accuracy of the k-NN algorithm as compared to row average. In k-NN, missing values are estimated based on the local region (i.e., nearby neighbors) whereas in row average the neighborhood comprises of the entire row which makes it highly irrelevant to the estimation problem. Further, in k-NN, nearest neighbors are assigned with greater weight as compared to far neighbors based on the Euclidean distance and an optimal value of k. This weighted average procedure provides a better local estimation of the missing

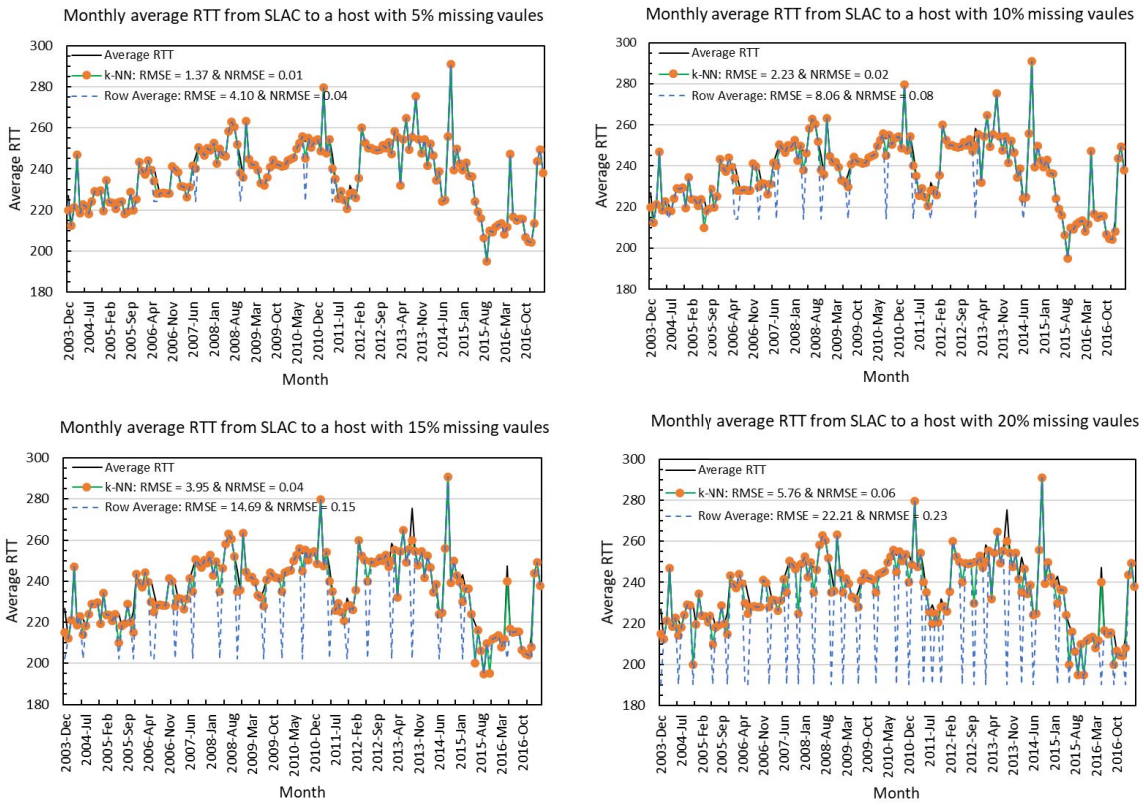

Figure 6: Monthly average RTT from SLAC to African host with substituted values

values. Figure 6 indicates that k-NN NRMSE in PingER average RTT with 15 and 20% of missing values only ranges from 0.04 to 0.06 as compared to row average where error vary from 0.15 to 0.23 with respect to original values. Thus, making k-NN the best choice for the estimation of the missing values in the PingER dataset.

7. Conclusion

The missing values in Internet performance metrics captured through PingER framework is a generic problem. This is due to the congestion in links, queuing overflow, faulty hardware or software and unavailability of MAs & remote sites all of which are unavoidable. These missing values directly affect the quality of the fine-grained Internet performance analysis. Therefore, in this work, PingER historical flat files are first converted into CSV format using a PingER data dimensional model. Afterward, missing values are imputed using row average and k-NN algorithm on a dataset of average RTT between SLAC-USA and Luanda-Angola pair. The data contain a different percentage of missing values i.e., 5, 10, 15 and 20% of missing values with respect to a reference dataset. At low percentages of missing values, both methods provide estimated values with low values of NRMSE. However, as the percentage of missing values is raised from 10 to 20%, k-NN algorithm outperforms the row-average method in PingER dataset. Thus, it concludes that k-NN is the best approach to estimate the missing values in PingER historical dataset to improve the quality of the Internet performance analysis worldwide. However, the results cannot be generalized to different types of dataset.

Acknowledgments

This material is partially based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Contract DE-AC02-76SF00515. Further, this work is supported in part by the National Natural Science Foundation of China (Grant No. 61632009 & 61472451), in part by the CERNET Innovation Project NGII20170102, in part by the Guangdong Provincial Natural Science Foundation (Grant No. 2017A030308006) and High Level Talents Program of Higher Education in Guangdong Province under Grant No. 2016ZJ01.

References

[1] S. Sundaresan, S. Burnett, N. Feamster, and W. De Donato, “Bismark: A testbed for deploying measurements and applications in broadband access networks.” in *USENIX Annual Technical Conference*, 2014, pp. 383–394.

[2] M. A. Sánchez, J. S. Otto, Z. S. Bischof, D. R. Choffnes, F. E. Bustamante, B. Krishnamurthy, and W. Willinger, "Dasu: Pushing experiments to the internet's edge." in *NSDI*, 2013, pp. 487–499.

[3] S. Sonntag, J. Manner, and L. Schulte, "Netradar - measuring the wireless world," in *2013 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2013, pp. 29–34.

[4] A. Faggiani, E. Gregori, L. Lenzini, V. Luconi, and A. Vecchio, "Smartphone-based crowdsourcing for network monitoring: Opportunities, challenges, and a case study," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 106–113, January 2014.

[5] V. Bajpai, S. J. Eravuchira, and J. Schönwälder, "Lessons learned from using the ripe atlas platform for measurement research," *SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 3, pp. 35–42, Jul. 2015.

[6] A. Hanemann, J. W. Boote, E. L. Boyd, J. Durand, L. Kudarimoti, R. Łapacz, D. M. Swany, S. Trocha, and J. Zurawski, "Perfsonar: A service oriented architecture for multi-domain network monitoring," in *Proceedings of Third International Conference Service-Oriented Computing - ICSOC 2005*, Amsterdam, Netherlands, 2005, pp. 241–254.

[7] W. Matthews and L. Cottrell, "The pinger project: active internet performance monitoring for the henp community," *IEEE Communications Magazine*, vol. 38, no. 5, pp. 130–136, May 2000.

[8] V. Bajpai and J. Schonwalder, "A survey on internet performance measurement platforms and related standardization efforts," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 3, pp. 1313–1341, 2015.

[9] S. Bauer, D. Clark, and W. Lehr, "Powerboost," in *Proceedings of the 2Nd ACM SIGCOMM Workshop on Home Networks*, ser. HomeNets '11, 2011, pp. 7–12.

[10] Z. S. Bischof, J. S. Otto, M. A. Sánchez, J. P. Rula, D. R. Choffnes, and F. E. Bustamante, "Crowdsourcing isp characterization to the network edge," in *Proceedings of the First ACM SIGCOMM Workshop on Measurements Up the Stack*, ser. W-MUST '11, 2011, pp. 61–66.

[11] Z. S. Bischof, J. S. Otto, and F. E. Bustamante, "Up, down and around the stack: Isp characterization from network intensive applications," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 515–520, Sep. 2012.

[12] G. Bernardi, D. Fenacci, and M. K. Marina, "Bsense: A flexible and open-source broadband mapping framework," *Mobile Networks and Applications*, vol. 19, no. 6, pp. 772–789, 2014.

[13] I. Canadi, P. Barford, and J. Sommers, "Revisiting broadband performance," in *Proceedings of the 2012 Internet Measurement Conference*, ser. IMC '12, New York, NY, USA, 2012, pp. 273–286.

[14] D. Genin and J. Splett, "Where in the internet is congestion?" *CoRR - Computing Research Repository*, vol. abs/1307.3696, 2013.

[15] T. Wang, Y. Cai, W. Jia, S. Wen, G. Wang, H. Tian, Y. Chen, and B. Zhong, "Maximizing real-time streaming services based on a multi-servers networking framework," *Computer Networks*, vol. 93, pp. 199 – 212, 2015.

[16] V. Bajpai and J. Schonwalder, "Ipv4 versus ipv6 - who connects faster?" in *2015 IFIP Networking Conference (IFIP Networking)*, May 2015, pp. 1–9.

[17] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," *Neurocomputing*, vol. 72, no. 7-9, pp. 1483–1493, 2009.

[18] P. Calyam, M. Dhanapalan, M. Sridharan, A. Krishnamurthy, and R. Ramnath, "Topology-aware correlated network anomaly event detection and diagnosis," *Journal of Network and Systems Management*, vol. 22, no. 2, pp. 208–234, 2014.

[19] U. Goel, M. P. Wittie, K. C. Claffy, and A. Le, "Survey of end-to-end mobile network measurement testbeds, tools, and services," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 105–123, 2016.

[20] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105 – 115, 2010.

[21] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, p. 520, 2001.

[22] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[23] G. Hawthorne, G. Hawthorne, and P. Elliott, "Imputing cross-sectional missing data: comparison of common techniques," *Australian and New Zealand Journal of Psychiatry*, vol. 39, no. 7, pp. 583–590, 2005.

[24] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring, "Missing-data methods for generalized linear models," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 332–346, 2005.

[25] T. Wang, Y. Li, G. Wang, J. Cao, M. Z. A. Bhuiyan, and W. Jia, "Sustainable and efficient data collection from wsns to cloud," *IEEE Transactions on Sustainable Computing*, pp. 1–1, 2017.

[26] S. Peng, G. Wang, Y. Zhou, C. Wan, C. Wang, and S. Yu, "An immunization framework for social networks through big data based influence modeling," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2017.

[27] W. Yang, G. Wang, K.-K. R. Choo, and S. Chen, "Hepart: A balanced hypergraph partitioning algorithm for big data applications," *Future Generation Computer Systems*, vol. 83, pp. 250 – 268, 2018.

[28] Y. Dai and G. Wang, "Analyzing tongue images using a conceptual alignment deep autoencoder," *IEEE Access*, vol. 6, pp. 5962–5972, 2018.

[29] S. Ali, G. Wang, B. White, and R. L. Cottrell, "A blockchain-based decentralized data storage and access framework for pinger," in *IEEE Trustcom/BigDataSE/SPTIoT/HSAT*, New York, USA, jul 2018, p. In press.

[30] R. Les Cottrell, T. Barbosa, B. White, J. Abdullah, and T. White, "Worldwide internet performance measurements using lightweight measurement platforms," in *2015 NETAPPS 4th Int. Conf. on Internet Applications, Protocols and Services*, 2015, pp. 25–30.

[31] R. L. Cottrell, "How bad is africa's internet," *IEEE Spectrum*, p. 2, 2013.

[32] G. E. Batista, M. C. Monard *et al.*, "A study of k-nearest neighbour as an imputation method." *HIS*, vol. 87, no. 251-260, p. 48, 2002.

[33] T. Cover, "Estimation by the nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 50–55, January 1968.

[34] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 325–327, April 1976.

[35] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[36] M. Kornacker and J. Erickson, "Cloudera impala: Real time queries in apache hadoop, for real," *ht tp://blog. cloudera. com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real*, 2012.

[37] D. de Oliveira, E. Ogasawara, F. Baião, and M. Mattoso, "Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows," in *2010 IEEE 3rd International Conference on Cloud Computing*. Miami, FL, USA: IEEE, July 2010, pp. 378–385.

[38] T. Barbosa, R. Souza, S. Cruz, M. Campos, and R. Les Cottrell, "Applying data warehousing and big data techniques to analyze internet performance," in *2015 NETAPPS 4th Int. Conf. on Internet Applications, Protocols and Services*, 2015, pp. 31–36.