



Quantifying Information Loss Through Data Aggregation

Mazda A. Marvasti, Ph.D., CISSP

TECHNICAL WHITE PAPER

Table of Contents

Abstract..... 3
Introduction..... 4
Characteristic Quantification..... 4
Data Description.....7
Distribution Change Quantification..... 8
Conclusions..... 11
References..... 11

Abstract

Data aggregation is a common “feature” and practice of current IT systems management tools as a means of data growth control and minimizing storage requirements. The purpose of this study is to assess the information loss associated with IT data aggregation. The results will show that more than 50 percent of IT type data lose their characteristic distribution at 2.5 hour aggregation.

The results will also show that obtaining statistical estimation from aggregated data is possible for computations of mean, median and mode. However, range reduction due to aggregation can lead to loss of critical regions of the data where capacity information resides. The conclusion of this study is that usage of aggregated data (namely those data sets with a greater than 2.5 hour aggregation) for reporting and capacity analysis will have lost a significant amount of its original behavior such that the analyzed results cannot be appropriately counted on as a fair representation of the original data set.

Introduction

Data aggregation has been a staple of systems management tools as a means of controlling the storage growth of the monitoring database. Although data can be aggregated in many ways, averaging is the most prevalent technique used by tool vendors. Based upon policies the typical aggregation may look like the following:

- Keep raw data for 24 hours
- Data greater than 24 hours but less than 7 days aggregate to once per hour
- Data greater than 7 days aggregate to 24 hours

The actual policies for each user may be more or less complex than what is depicted above but it does show a typical aggregation strategy. The other factor that may come in is to aggregate based on the type of data. For example typical operating system (OS) metrics (CPU usage, memory usage etc.) may be aggregated more often due to volume, while transaction response metrics may have aggregation schedules that keep the raw data for a longer period.

Because of the need of the typical systems management tool to aggregate raw data (for performance and cost reasons), the reporting and capacity planning functions (or experts) within organizations have been forced to use aggregated data. It is the purpose of this study to quantify data characteristics and identify at what aggregation point the characteristic of the data set is lost (data characteristics in IT systems management is synonymous with behavior, pattern, baseline, etc.).

Characteristic Quantification

Information loss classically belongs to the realm of entropy quantification as described in Information Theory. Measures such as Kullback–Leibler divergence can quantify information loss when an estimated distribution is used to encode a message with a different actual distribution. However, these types of techniques were designed to assess information loss with respect to uniform distributions and thus cannot be used effectively in the current context. The quantification process used here involves the following techniques:

- Using Average Shifted Histogram (ASH) to determine the actual distribution of the non-aggregated (raw) and aggregated data sets
- Using the Kolmogorov-Smirnov (KS) statistics quantify how close the aggregated data distribution is to the raw data distribution
- Using the 5 percent criticality level of the KS statistics determine the aggregation point beyond which the aggregated data distribution no longer resembles the raw data distribution
- Using ASH to determine the distribution of the above critical aggregation level for a variety of IT type data

The results of the above procedure will be a distribution representation of the aggregation level of metrics beyond which the data set no longer resembles the original, non-aggregated data. The techniques described here will be summarized below and a more detail description can be found in the references.

The first step in the above process requires obtaining the actual distribution of the data. Average Shifted Histogram (ASH) technique is used to obtain the distribution. This technique allows one to very closely model the data distribution without having to rely upon typical histogram limitations of wide band data buckets (see Figure 1).

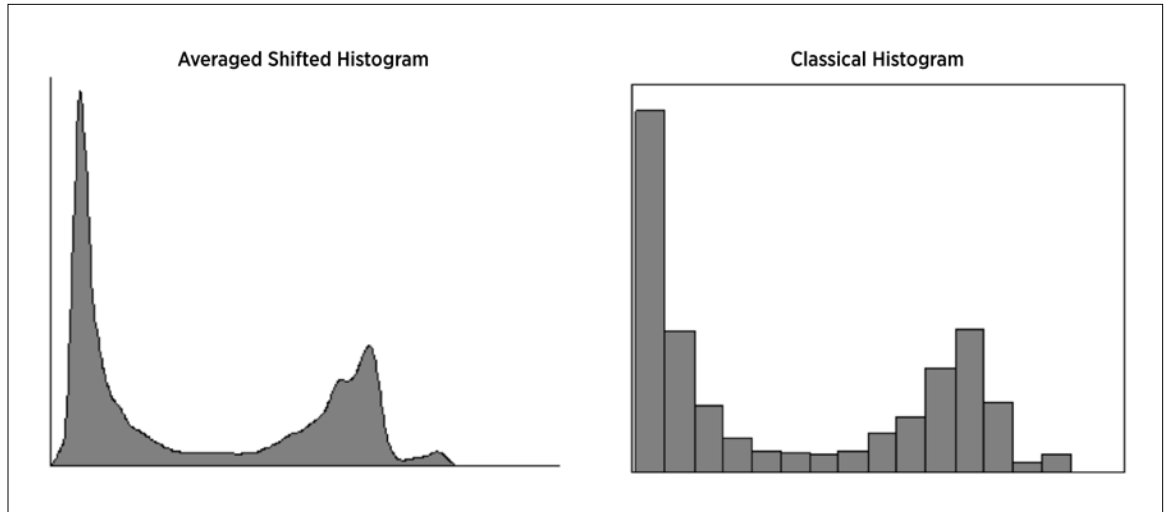


Figure 1. Resolution of ASH vs. Classical Histogram Distribution Representation

The ASH works by starting with the traditional banding of data into N buckets each of size (width) h , but then shifting each of these N buckets $M-1$ times to average the behavior across the data set using:

$$P_h(x) = \frac{1}{M} \sum_{l=0}^{M-1} \frac{1}{Nh} \sum_{i=1}^N \left\{ \sum_j I(X_i \in V_{jl}) I(x \in V_{jl}) \right\}$$

$$V_{jl} = [(j + l/M)h, (j + 1 + l/M)h], \quad l = 0, \dots, M - 1; \quad j = 0, \dots, N - 1$$

where $P_h(x)$ is the distribution of the data set and $I(X_i \in V_{jl})$ represents the count of number of x_i which are elements of the set V_{jl} .

The Kolmogorov-Smirnov technique is a non-parametric test of how “close” one distribution is to another. The KS test uses the largest distance between the cumulative distribution function of aggregated and raw data (see Figure 2).

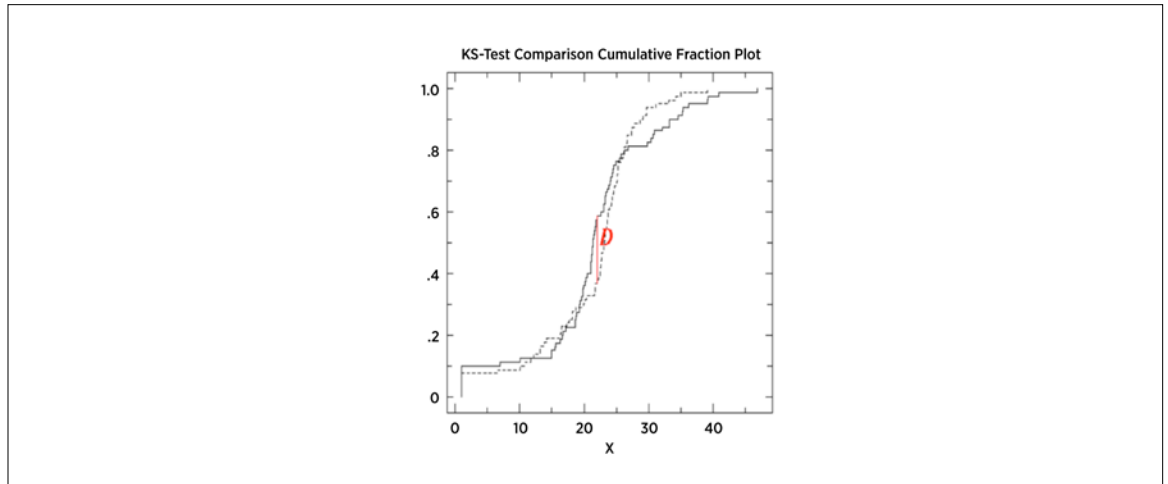


Figure 2. KS-Test Comparison Cumulative Fraction Plot

This measure is then compared to a critical value (obtained at the desired criticality level) to ascertain whether the two data sets have distributions that are acceptably close. If the value of the measure (D) is less than the critical value then the two data sets have distributions that are very similar. If D is greater than the critical value, then the two data sets have very differing distributions.

The purpose of the above procedure is to answer the question: at which aggregation point the data has lost enough information such that it cannot effectively be used for analysis? When data is aggregated two things get impacted, the data distribution and data set range. As the aggregation time increases, the range of the data set starts to decrease (since they get averaged into the data which is more prevalent in the distribution). This range reduction may be acceptable up to a certain point but beyond that lies a critical region of data that is of most interest to capacity management. Figure 3 shows an idealized distribution of data and various regions of the data set.

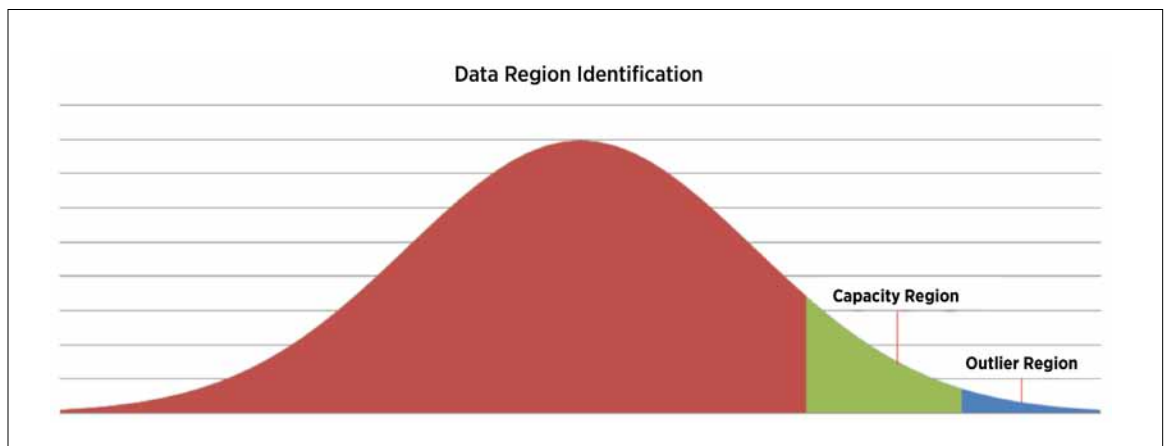


Figure 3. Data Region Identification

The outlier region typically represents a range of data that is considered to be anomalies and not part of regions of interest. Immediately adjacent to that is the capacity region which represents the upper behavior of data set that dominates the peak performance of the application. From this figure, it is evident that too much range loss will severely impact the capacity region and thus impact capacity analysis. One of the goals of this study is also the quantification of acceptable range loss.

Data Description

The data sets used in this study were gathered from a variety of IT systems and applications including four different verticals. The data sets ranged from Operating System metrics (CPU, memory, disk IO, bandwidth usage, etc.) to usage metrics (number of hits, sessions, connections, etc.) to response time metrics (transaction time, server time, etc.) to environmental metrics (power supply temperature, processor temperature, etc.). The total number of metrics ranged in the thousands to ensure a proper sampling of the metrics. Figure 4 shows a User Hits metric at zero aggregation, 6-hour aggregation, and 24-hour aggregation.

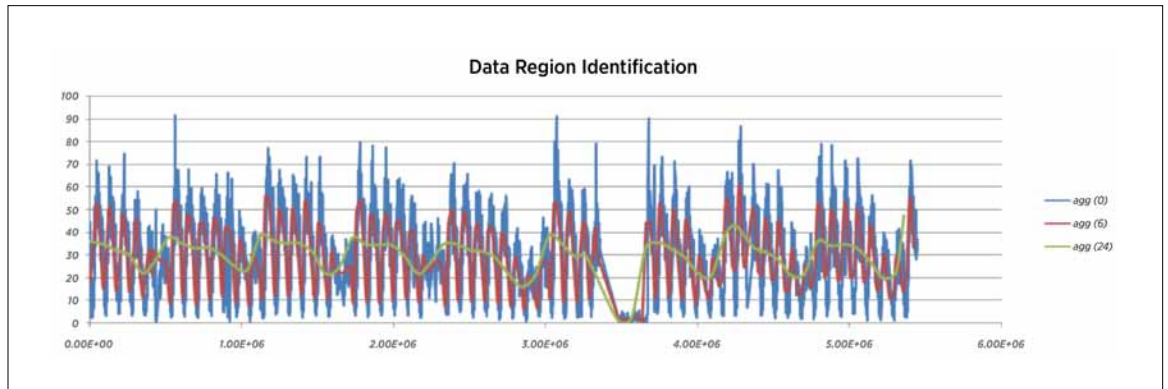


Figure 4. "User Hits Data" Metrics

The Probability Density Function (PDF) of this data set at various aggregation points is shown in Figure 5.

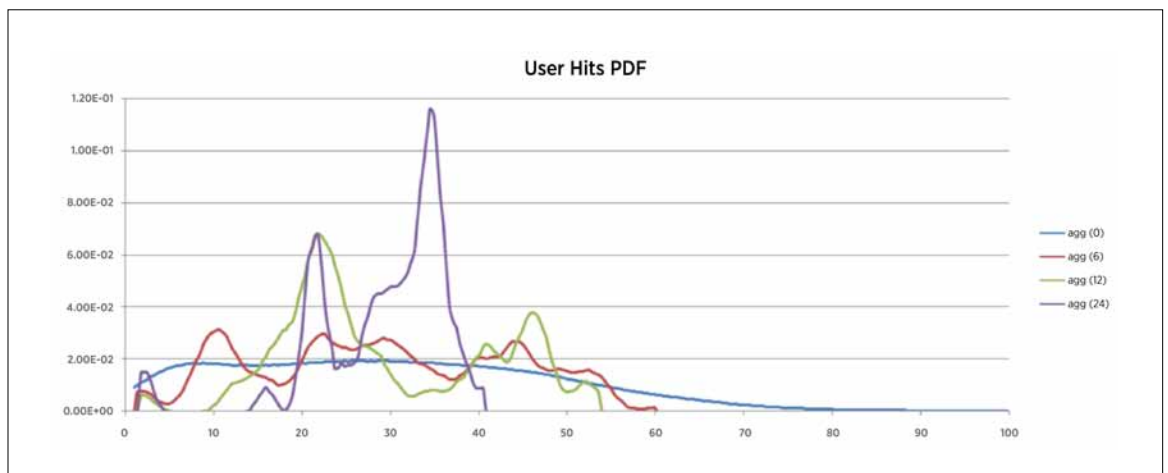


Figure 5. "User Hits PDF" Metrics

Qualitatively, it can be observed that the shape of the distribution function changes as the degree of aggregation is increased. The quantification of this change and the changes in the statistical properties of the data will be explored next.

Distribution Change Quantification

The first attempt at identifying the impact of aggregation on a given data set is the quantification of the degree to which the aggregated data distribution deviates from the raw data. As mentioned above, the KS technique is used for this quantification process. Figure 6 shows the Statistics D as a function of degree of aggregation (in hours) for four metrics.

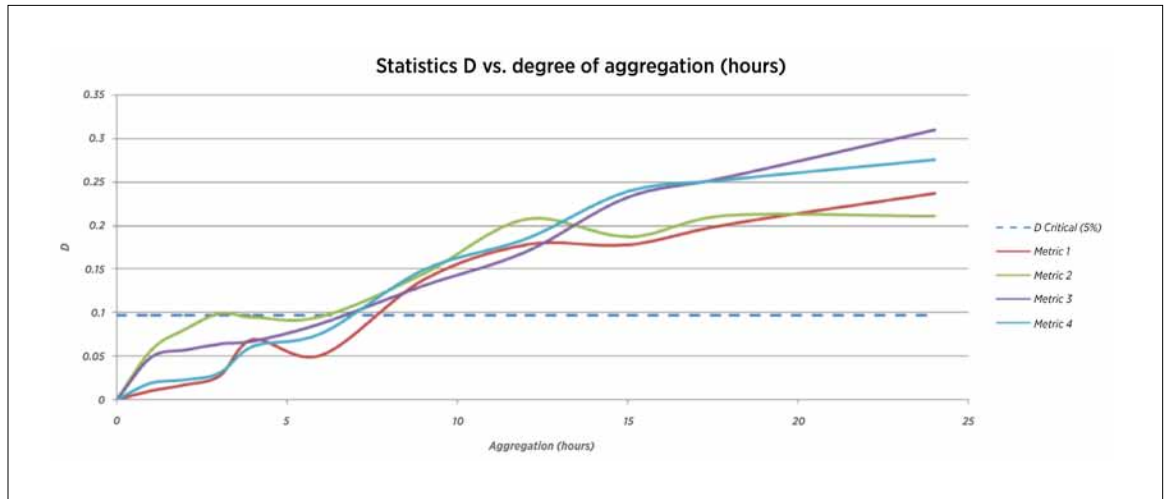


Figure 6. Statistics D vs. Degree of Aggregation

The dashed line represents the Critical D value at the 5 percent criticality level (i.e. 95 percent confidence level) below which the metric is considered to be correlated to the raw data and above which the data is considered to be deviated from the original distribution. This figure shows that different metrics start deviating from the raw data at different aggregation levels (from now referred to as critical aggregation). Another observation that can be made here is that the larger the aggregation level the further away the data characteristics deviate from the raw data set. The next task is to determine the distribution of the critical aggregation level of IT type metrics.

Figure 7 shows the PDF of all IT metrics examined vs. the critical aggregation level.

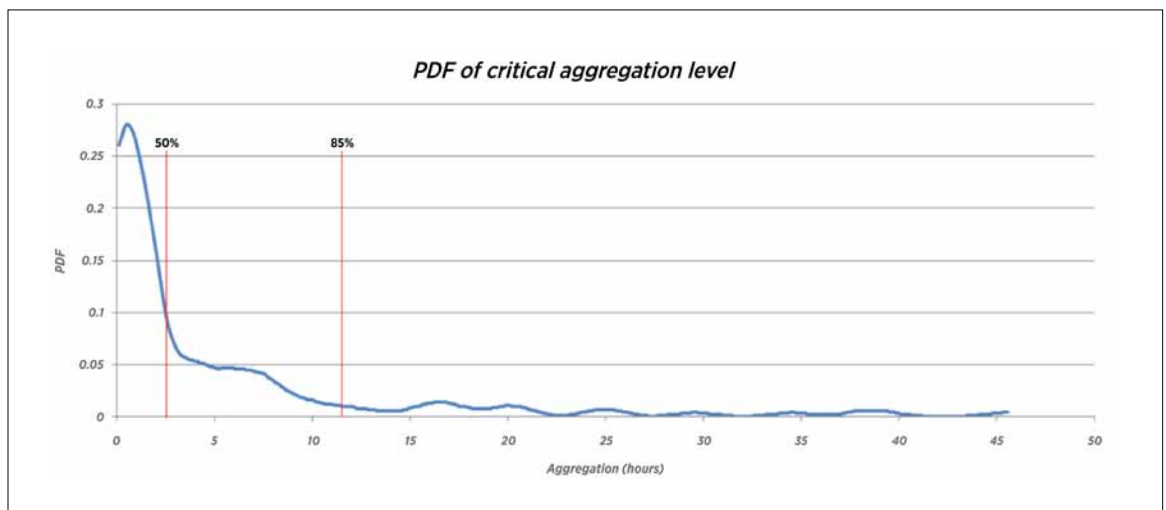


Figure 7. PDF of Critical Aggregation Level

Essentially this figure shows that 50 percent of metrics exceed their critical aggregation level at about 2.5 hours of aggregation and 85 percent of the metrics exceed their critical aggregation level at about 12 hours of aggregation. This surprising result indicates that any analysis performed on metrics which have been aggregated to greater than 2.5 hours may not be representative of what the raw data may have shown. In other words one can expect the analysis of half the metrics using an aggregation of 2.5 hours (or more) to be in error.

The next item to be studied is the range reduction of the data set being analyzed. The question then becomes how does aggregation impact those metrics? To answer this question the first step is to find a way to compare the various properties of the aggregated data to that of the raw data in a way that can compare one metric with another. This requires scaling of the various properties with respect to the range of the raw data. In other words:

(D_i) = The set of data at aggregation level i (zero indicates raw data)

$P(D_i)$ = Property of data set D_i (such as mean, median, mode, standard deviation, range, etc.)

$\bar{P}(D_i)$ = Scaled property of data set (scaled for comparison purposes)

$R(D_i)$ = Range of data set $D_i = \max(D_i) - \min(D_i)$

$$\bar{P}(D_i) = \frac{P(D_i) - P(D_0)}{R(D_0)}$$

This method allows for comparison of different types of metrics (such as cpu, memory, I/O) directly with each other for characterization of range reduction. Appendix A shows how this procedure was used to characterize the mean, median, mode, and standard deviation of the aggregated metrics with respect to the raw metrics.

Figure 8 shows the cumulative probability of the scaled range of aggregated data with respect to the raw data.

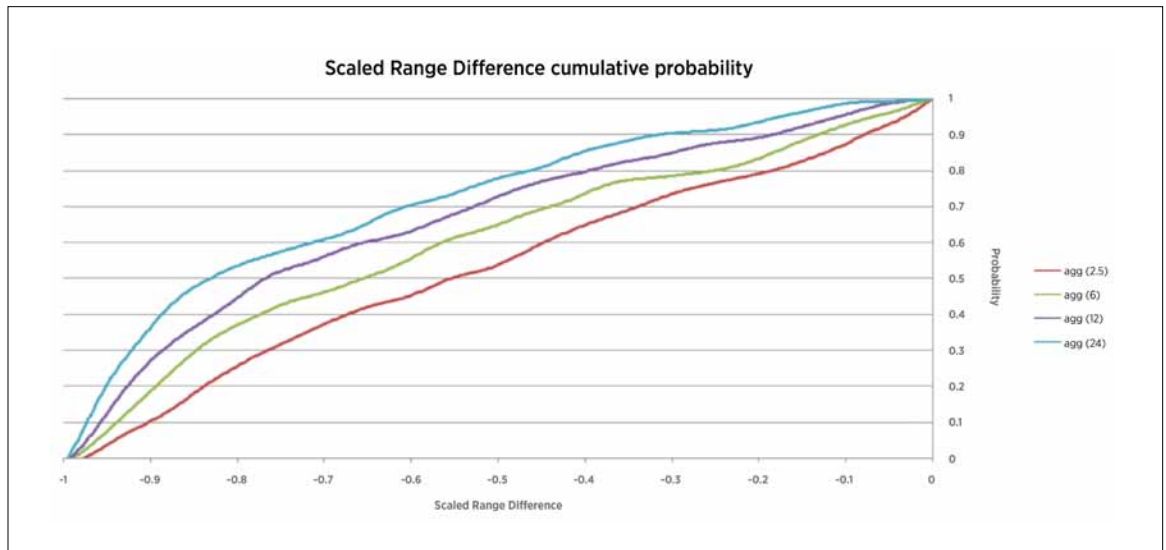


Figure 8. Scaled Range Difference Cumulative Probability

This figure shows the cumulative probability of range difference which can be obtained by integrating the probability density function at the appropriate limits of integration (see cited references). This figure can be read as the probability that the scaled range is at least a certain value or less. For example, at a probability of 0.5 the 6 hour aggregated data shows a value of approximately -0.65. What this says is that 50 percent of the metrics in that aggregation group showed a range reduction of 65 percent or more compared to the raw data. However, comparison to the raw data is not the relevant comparison as it was previously determined that the characteristics of the distribution are preserved for aggregations up to 2.5 hours. Figure 9 shows a different perspective of the data in Figure 8.

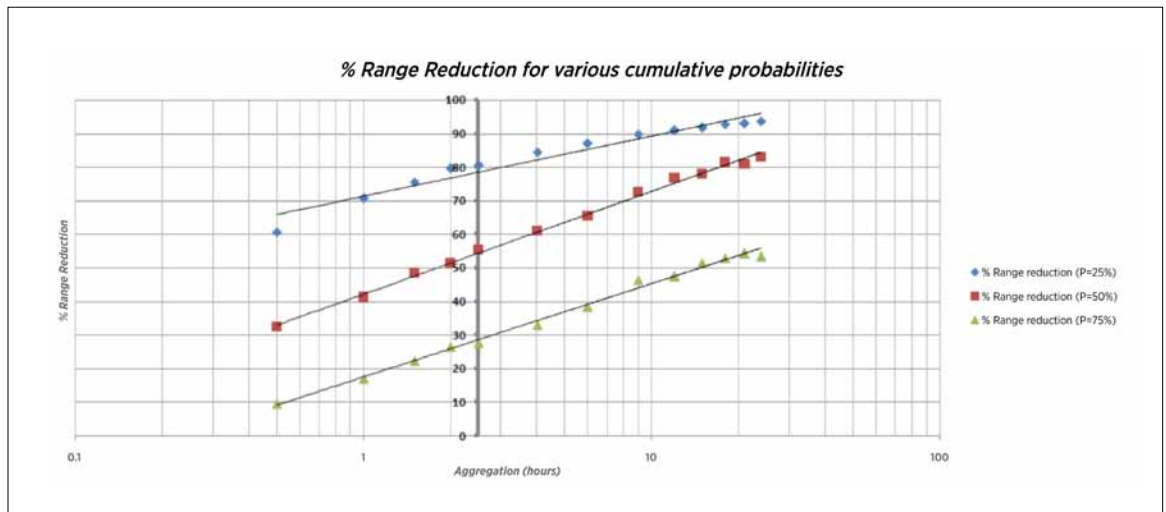


Figure 9. Range Reduction for Various Cumulative Probabilities

This figure shows percent range reduction versus aggregation level for 25 percent, 50 percent, and 75 percent cumulative probabilities. Within the aggregation range of 0.5 hours to 24 hours this data exhibits logarithmic behavior (shown linearly on a log plot). The vertical axis has been placed at the 2.5 hour aggregation point and thus the values on the vertical axis represent the maximal percent range reduction for various cumulative probabilities before data distribution divergence from that of raw data. For example at the 50 percent cumulative probability (red squares in Figure 9) the data set can go through 55 percent range reduction before distribution divergence becomes evident. What this means is that 50 percent of metrics can go through up to 55 percent range reduction before distribution divergence occurs. This 55 percent range reduction thus represents the limits of the outliers of the data set and can thus be eliminated without impacting the analysis. However, range reductions beyond 55 percent lead to information loss in a region where most interest lies for capacity analysis (see Figure 3). Essentially what Figures 8 and 9 show is that aggregation in general blends the outer tails of the data distribution into the midrange of the data. The net effects of these are that beyond a certain point the tail data becomes obscured which is where the most interesting data resides as far as capacity management is concerned. The reasons for this are:

- Typical high load periods account for only a small fraction of the total daily or weekly activity
- Ensuring systems availability requires ensuring capacity availability during peak loads
- Business cycles associated with peak activity may mean that peak periods get spread across the weeks on an uneven basis (occurring on calendar days for instance instead of the weekdays thus leading to different days exhibiting peak loads throughout the year).

This loss of tail data clearly poses a significant challenge to the capacity management practitioners. If an aggregation of greater than 2.5 hours is used for their predictions not only do they have to contend with range loss but they also have to account for loss of distribution characteristics.

Conclusions

The quantification of aggregation information loss showed that the distribution of the data set can be severely impacted after 2.5 hours of aggregation. Specifically 50 percent of IT type metrics lose their distribution coherence (relative to the raw data) after 2.5 hours of aggregation and 85 percent lose their distribution coherence after 12 hours of aggregation. Thus for capacity studies that rely upon a faithful representation of the raw data distribution, any aggregation level above 2.5 hours may severely impact the final results. The range differences highlighted the loss of critical information at the tails of the data set. The distribution of the data set shows at what values the data set “resides” most of the time. It is during the few peak periods where determination of capacity becomes essential. These peak periods are often at the tail ends of the distribution and thus most vulnerable to aggregation information loss. The results showed that a range loss of less than 55% can be considered outlier removal, however, values greater than 55% start removing and blending of important capacity regions of the data set.

References

- Porkess, R.: *Web-linked Dictionary Statistics*, HarperCollins, 2006
- Jaynes, E.T.: *Probability Theory The Logic Of Science*, Cambridge, 2003
- Spiegel, M., Schiller J., Srinivasan R., *Probability and Statistics*, McGraw-Hill, 2000
- Kullback, S., *Information Theory and Statistics*, Dover Publications, 1997

Appendix A

In this section the aggregate behavior (mean, median, mode, and standard deviation) of data sets is examined to better understand how aggregation changes these values. Figure 10 shows the distribution of the scaled mean for aggregations of 6, 12 and 24 hours.

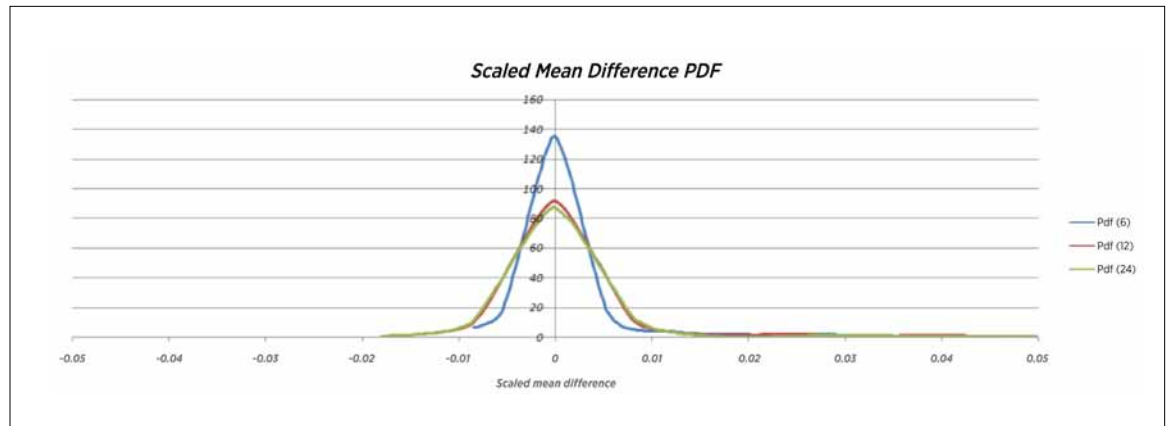


Figure A1. Scaled Mean Difference PDF

The distribution of the difference between the means of the aggregated and raw data shows a very nearly normal distribution with an increase in standard deviation with increasing aggregation level. Essentially this is nothing more than the validation of the Central Limit Theorem of statistics (which states that estimation of the mean using a sample of the population leads to a normal distribution of the estimated mean about the actual mean value). Using this, one can provide suitable confidence intervals on the estimation of the mean given the aggregation level. Thus if the mean of the data set is of interest aggregation does not provide an obstacle to its estimation. Figure 11 shows the same graph for data set median.

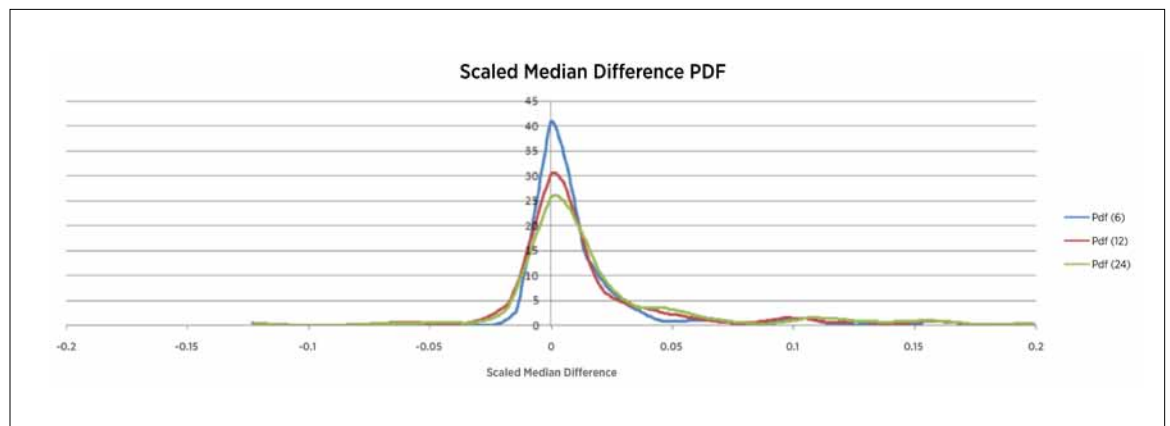


Figure A2. Scaled Median Differen PDF

This figure shows a slightly different behavior than Figure 10 in that there is very little variation due to aggregation level, however there is a clear shift to the positive side of the axis. What this says is that aggregation in general will have a tendency to overestimate the value of the median. However, having the distribution (such as the above figure) allows one to compensate appropriately for median estimation using aggregated data. Figure 11 shows the scaled mode difference between aggregated and raw data.

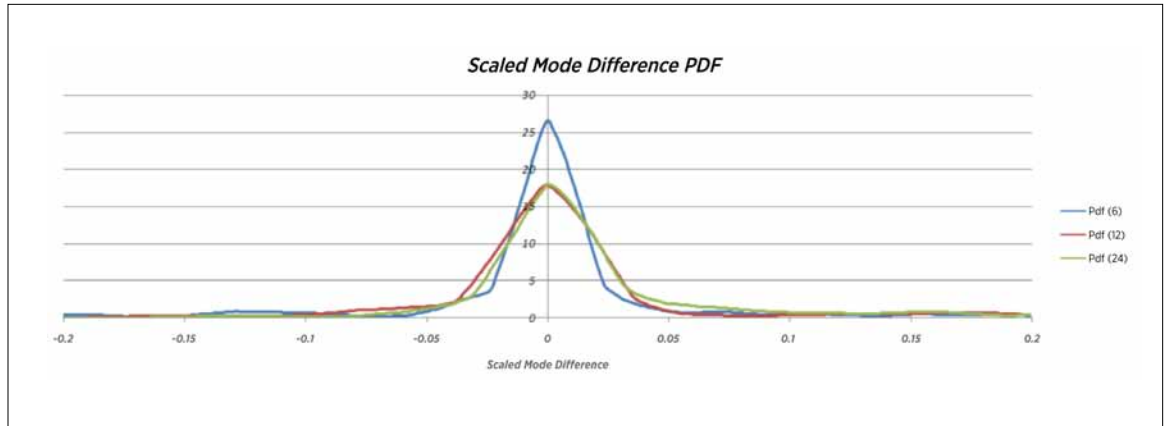


Figure A3. Scaled Mode Difference PDF

The mode of a distribution represents the data value with the largest probability density. This figure shows that in general the estimation of mode very much follows that of the mean in terms of obtaining a confidence band based on a normal distribution of the estimated mode value. Figure 12 shows the same type graph for the standard deviation of the data sets.

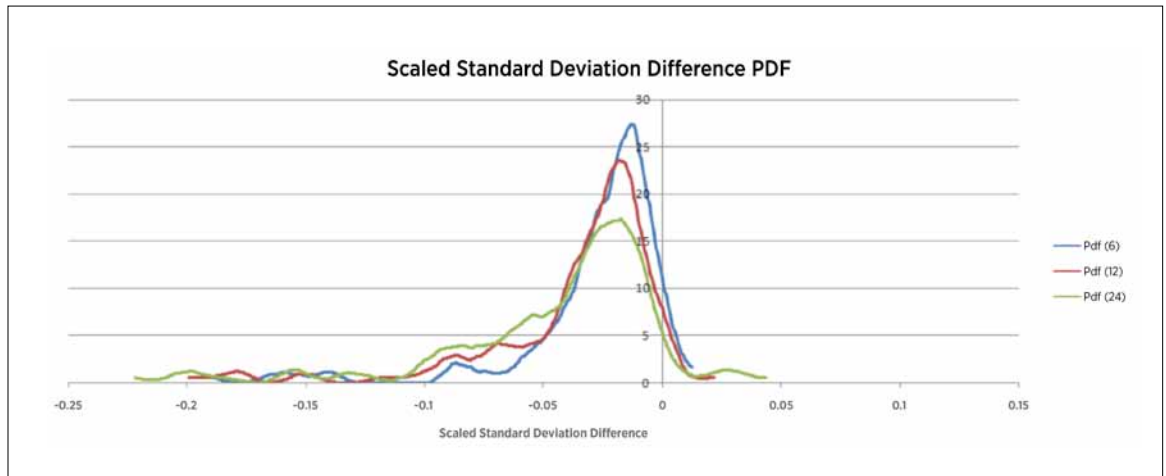


Figure A3. Scaled Standard Deviation Differences PDF

This statistics is markedly different than the others presented. What this shows is a clear bias to reducing of the standard deviation with an increase in aggregation level. This should be intuitive as with increased aggregation the outliers tend to be more blended into the mainstream data set and thus leading to a reduced range as shown in Figures 8 and 9.



VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Copyright © 2011 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at <http://www.vmware.com/go/patents>. VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies. Item No: VMW-TECH-WP-VC-OPS-IT-DATA-INFO-USLET-101