

# Workload Correlation and Visualization\*

Tom Wilson

*A great way to make your point is to “show them”. This paper demonstrates some simple comparison and correlation methods for studying workload using visualization as an aid in the analysis. First, we examine some metrics for a real system and perform correlations between a few parameters. Then, we compare workloads from various samples to determine an appropriately representative interval upon which to base a performance test. These methods provide confidence for the system development and testing.*

## 1 Introduction

As engineers in a multi-program environment, we sometimes become part of a team where the program is already well under way. Part of the learning curve involves reading established documents and reviewing existing work. Occasionally, we find opportunities for improvements and have to justify them to a decision-maker who is not necessarily familiar with the analysis details. One of the simplest communication mechanisms is a graph or chart. It is far easier to argue a point by visualizing the data rather than by showing a table of numbers, arguing emotionally, and waving your hands. We will demonstrate several visualization analyses, two of which convinced the decision-makers to make improvements to the work being done.

This particular program consists of a series of projects developing proprietary transaction systems. Each system builds upon the previous, bringing more functionality to a continuously-growing user base. The systems support logistics and maintenance for military equipment. For the purpose of anonymity, only general details about the system are provided here. One of the concerns common to all of the systems is performance.

A key component to system performance testing is the usage of a relevant workload. Since the performance test has a short duration (i.e., on the order of hours), the test workload should represent that portion of real workload that stresses the resources of the system. When an existing system provides the real workload, we are left with the decision as to how to sample that data to create the test workload. If we consider two samples, how can we compare them?

A workload characterization effort would associate resource usage with the transactions in order to determine some equivalence. For our system, associating resources with transactions was deemed difficult due to lack of support within the existing software. Carrying out a large number of isolated tests as an alternative to determining this association was not considered feasible. We need an alternative to characterization so that we can quickly compare workloads. The literature contains numerous papers on workload characterization (e.g., [EM02]), but little on simple comparison.

In this paper, we want to investigate several concerns. Is a sample workload representative? If the wrong workload is modeled, then an incorrect evaluation of performance will exist. If the test results are good, then problems may arise in production because a different workload exists. If the test results are bad, then over-engineering may occur to solve problems that may never arise and raise cost. In order to answer this question, we need to be able to compare workloads and make decisions based on the comparison. The results of the comparisons should guide us toward selecting the workload upon which to base a test. Past testing selected a sample based on experience and had no supporting analysis. This is common in a pressure-filled development environment, especially where review is absent, but does not necessarily lead to a wrong answer.

Will the workload change as load increases? If it does, we need to understand how, so that our tests are appropriate. To answer this question, we must first confirm that load is a function of users and that the user base is growing. If both are confirmed, we need to understand how the workload has changed as the user base grew in the past.

---

\*Originally published in *Proceedings of the CMG '10 International Conference*, December 2010.

We will begin by analyzing some source data. In this paper, the analyses will be constrained to those that support answering our questions. Then, we will define how we can compare workloads.

## 2 Source Data Analysis

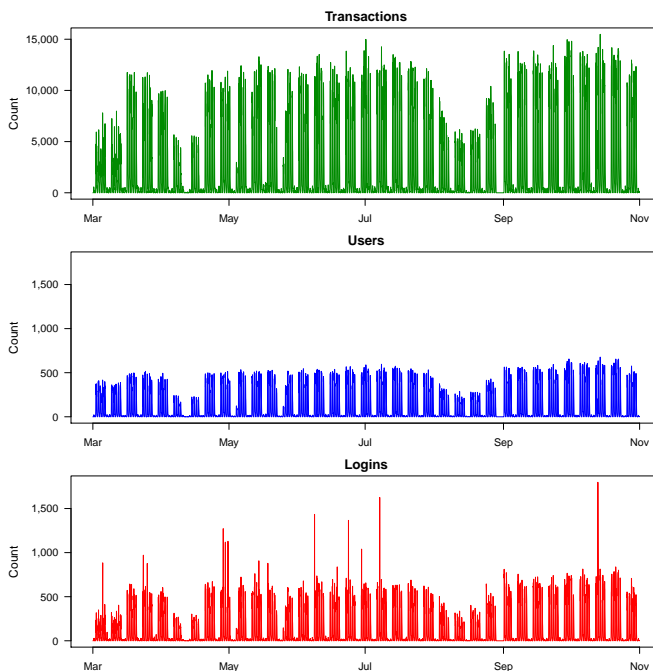
A significant amount of source data is available from the existing system against which many analyses can be performed. Some of those data are captured in transaction logs, which include the beginning and ending times of the transactions. From these data, a sampling was performed to determine the number of transactions per hour, the number of users per hour, and the number of logins per hour.<sup>1</sup> The source data cover an eight-month period during which the system was in continual use. User load varies because of many factors, such as the day-of-the-week, the time-of-day, and seasonal events (i.e., holidays). This behavior will be illustrated in Section 2.1.

The source data are filled with various problems that are mostly a small percentage in the total volume. However, one problem that was frequent concerned the naming of transactions. For reasons that we cannot explain, a large number of transactions were named “Unknown Transaction”. This many-to-one mapping masks what the real frequencies of these transactions are. At one point during the maintenance lifetime of the system, many of these transactions were distinctly renamed. We anticipate a poor result when comparing a sample from before this point to a sample after this point.

First, we will look at metrics for a few parameters during several intervals. Then, we will perform correlations between those parameters. Finally, we will perform a trend analysis. These analyses are foundational to answering our workload questions.

### 2.1 Metrics

Figure 1 shows metrics for the three parameters for the eight-month interval. The data consist of hourly counts for each metric. A weekly pattern is fairly apparent; each “column” consists of data for the weekdays, while the gaps in between them are the weekends. The “valleys” that occur in April and August are due to holidays. The users and logins graphs use the same y-axis scale, so as to better appreciate their relationship.

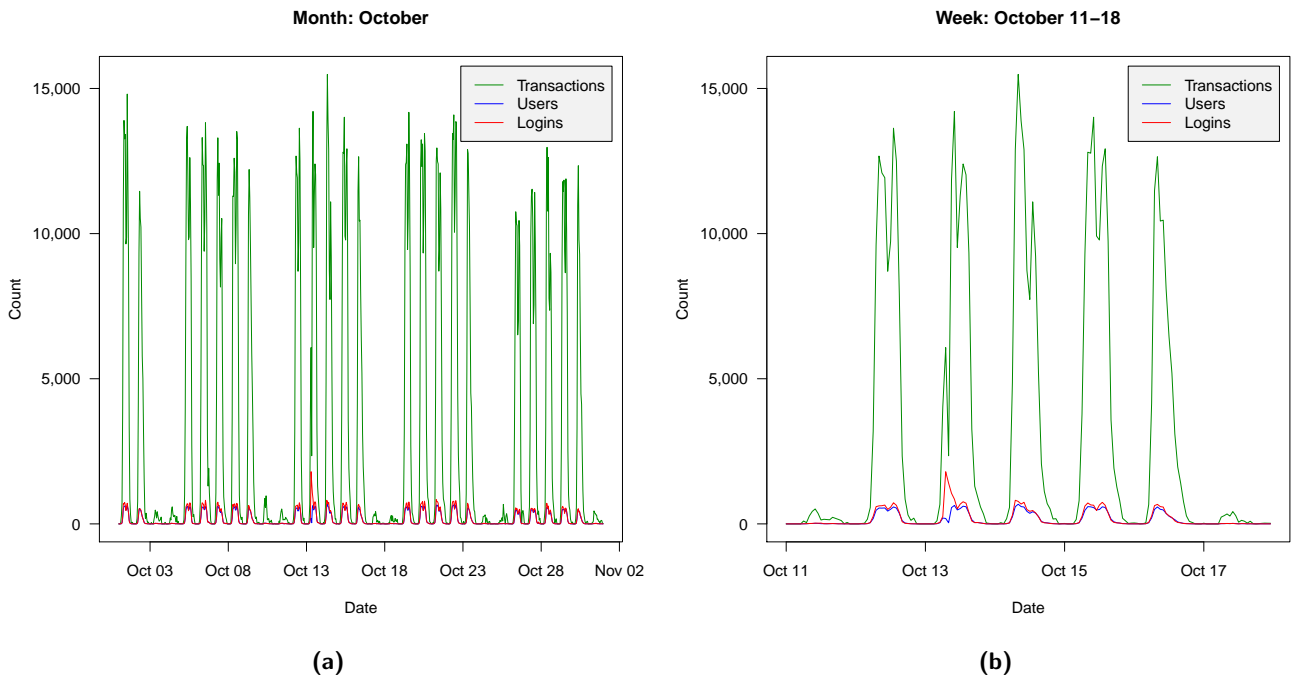


**Figure 1:** These charts show metrics for the three parameters over an eight-month period: transactions, users, and logins. The users and logins charts use the same y-axis scale; the transactions chart does not.

Figure 2 shows a few charts of the same three parameters for specific one-month and one-week time intervals. In Figure 2a, the weekly pattern is now more obvious: five work days followed by two non-work days (i.e., a weekend). While there are some minor differences in the data for each day, there is nothing that easily explains it. Figure 2b shows data for

<sup>1</sup>We have provided login data purely for contrast with the other parameters.

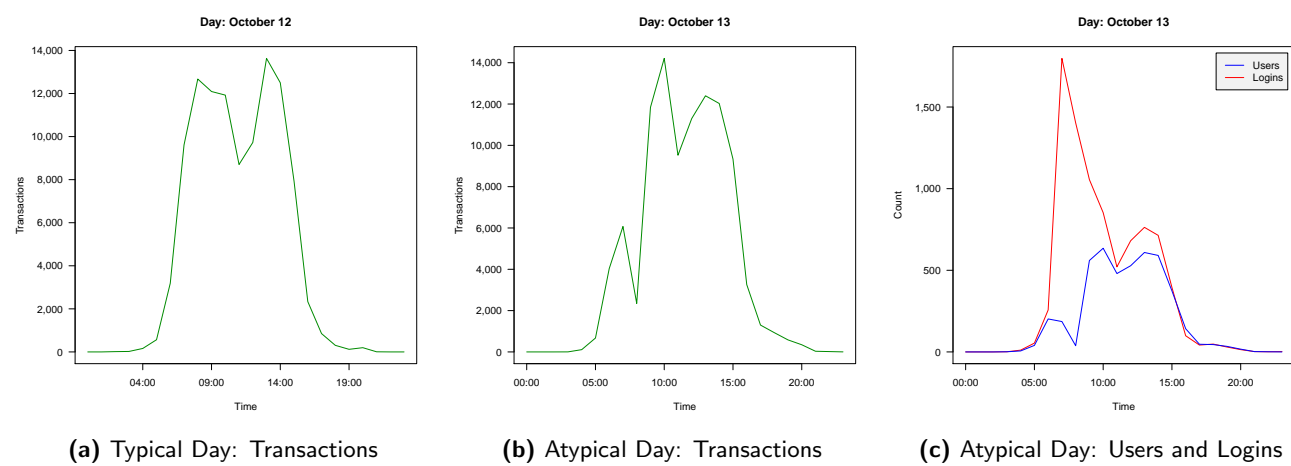
one week. Most weekdays show a common double-hump feature that results from peaks before and after a lunch break, although Friday lacks this feature. This double-hump artifact is a property of people, not the system.



**Figure 2:** These charts show the three metrics over (a) a one-month interval and (b) a one-week interval.

Most other months are similar to the example, although the peaks might be lower. Months with holidays are certainly different, but we are not expecting to base a test on a holiday load.

Figure 3 shows two different one-day intervals. Figure 3a shows transaction counts during a typical work day. The corresponding graphs of the other metrics would look similar and are not shown. Figure 3b shows transaction counts during an atypical work day, where some event has caused a reduction in transaction counts around 8:00. Figure 3c shows the user and login counts for the same time period. The user counts are similar in shape to transaction counts, while the login counts reveal a spike reflecting users logging in numerous times. The event probably brought the system down and required the users to log in again.



**(a) Typical Day: Transactions      (b) Atypical Day: Transactions      (c) Atypical Day: Users and Logins**

**Figure 3:** These charts show the three metrics during a day interval: (a) Transactions during a typical work day, (b) transactions during a work day with an anomaly, and (c) logins and users during that same atypical day.

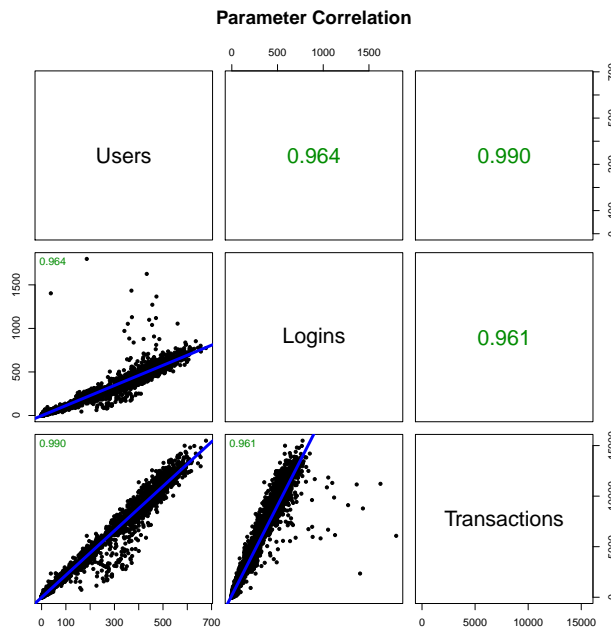
These metrics give some insight to system usage and are the basis for several subsequent analyses. What we did not

study here is what the users did on the system (i.e., the activities).

## 2.2 Parameter Correlation

Correlation expresses a (linear) relationship between two parameters. If the relationship is strong, then one parameter may be a function of the other. Correlation can be computed by a function and expressed as a quantity (called the *correlation coefficient*), and it can be visualized with a scatterplot and a regression line. We will look at the correlation between pairs of our three parameters. One expectation is that the number of users dictates the number of transactions.

Figure 4 plots each parameter against the other as a matrix of graphs. The cells of the diagonal contain the parameter name. Any other cell is a comparison of the two parameters specified in the cell's row and column. The cells below the diagonal show a scatterplot with a blue regression line and the correlation coefficient (in green). Cells above the diagonal show the same text as below the diagonal for readability purposes. The correlation coefficient for users and transactions is very high, so we should conclude that the number of users drives the number of transactions. This confirms our expectation that the number of users drives the number of transactions.



**Figure 4:** This chart shows the correlation between three metrics: users, logins, and transactions. Note that axes have different scales.

For contrast, we also compare the number of logins to the number of users. The correlation coefficient is still fairly high, but not as high as the previous comparison. It is easy to see many outlying values that reduce the coefficient. One of them corresponds to the feature in Figure 3c where the number of logins was much higher than the number of users. Many others correspond to the spikes in Figure 1. The anomalies are not numerous and can probably be explained (e.g., a system crash). A similar correlation occurs between logins and transactions. Many outlying values are visible. Note that the scatterplot tells us more than the correlation coefficient does (because it expresses more data).

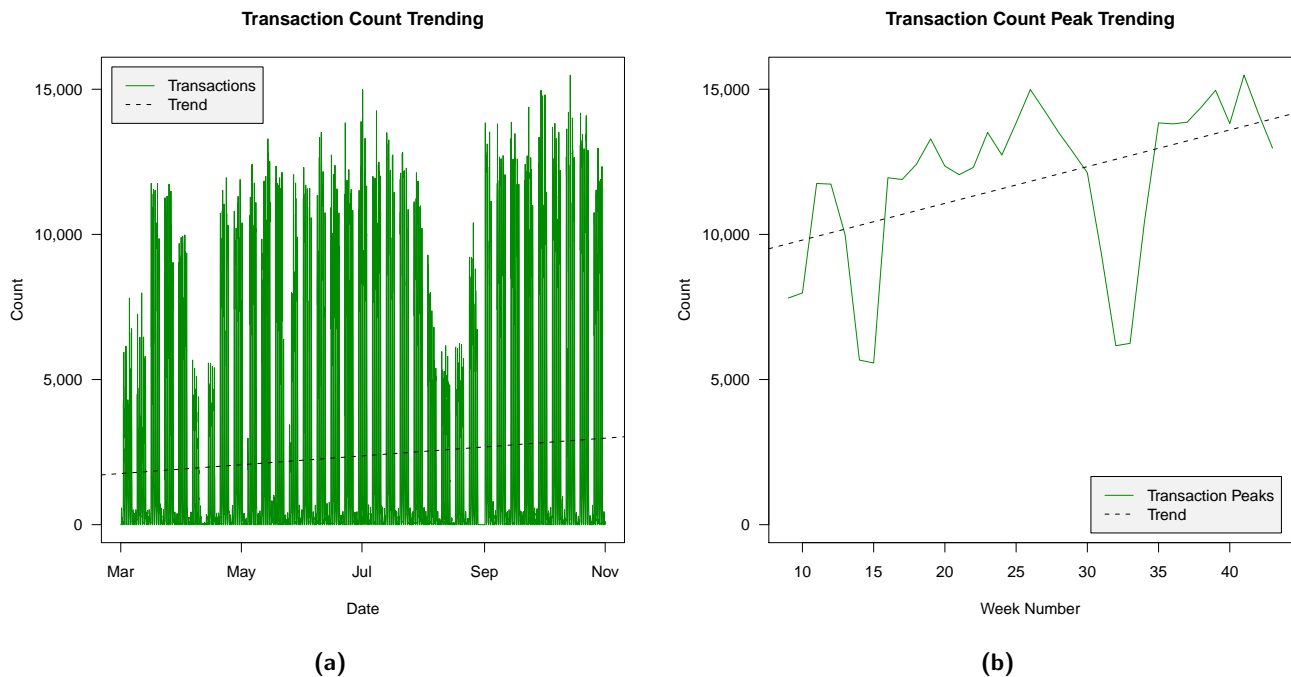
We have performed the correlation to establish a strong relationship between users and transactions. This is necessary as we consider a growing user base. This will imply a growing transaction load. Our later analysis will attempt to determine if the workload changes with the growth.

Another reason for doing this analysis is because some testing results were providing contradictory evidence against a relationship. The results here say that we should question our tests if the tests do not have a number of transactions that relates to the number of users. It turns out that some simulated users were failing during tests, resulting in a varying number of transactions for the tests. In reality, the number of users was sometimes lower than was reported. Therefore, the relationship was still probably very strong. Thorough analysis will often uncover problems and eventually explain them.

## 2.3 Trending

Trending analysis provides us with an understanding of a general parameter behavior over time. More specifically, we are interested in whether or not the load on the system is growing. The rate of growth is not an issue here, but could be for a capacity planning effort. What we need to know is whether or not the workload changes as the load grows. We will limit ourselves to the “growth” aspect here and will consider the impact of the growth on workload in Section 3.2.

A trend is computed via linear regression. Figure 5a shows the growth trend for transaction counts across the entire measurement interval. Even though the trend shows an increase, its location is deceptive (it is well below the peaks). This is because the data vary so much.



**Figure 5:** These charts show the transaction growth trend for the measurement interval. (a) This chart shows the trend computed from all of the data. (b) This chart shows the trend computed from weekly maximums.

We would prefer to understand how the peaks are trending. Peaks can be computed in a few ways, but we will compute a maximum transaction count for each week. This will eliminate a lot of data from the regression. Figure 5b shows the growth trend for transactions by looking only at the peaks. Again, an increase is shown. The elimination of detail is very helpful. The seasonal events are more obvious with this view of the data. Their impact on the trend location is worth noting.

Both figures in Figure 5 tell us that there is an increase in load over time. This is expected and the explanation is simple: The customer has been increasing its user base over time as more and more organizations have transitioned to using the system. This explanation implies a user growth rather than a transaction growth, but we already know there is a strong relationship between users and transactions. While the first chart is sufficient to establish the growth trend, it may not leave a lasting impression due to the location of the trend line. The peak trend is certainly more demonstrative.

## 3 Workload Correlation

Workload is the collection of activities being executed on the system. It is a function of the number of users, the types of activities, and their timings. The timing of activities will be effectively eliminated by just counting the activities over an interval. These counts are a function of the number of users. For simplicity, we will normalize the counts to frequencies. So, a workload is summarized as a set of frequencies. Different user counts and different intervals can be compared if we can compare sets of frequencies.

We will define workload correlation and then compare various workloads in order to answer our remaining questions. Our analyses in the previous section support the analyses here.

### 3.1 Example Correlation

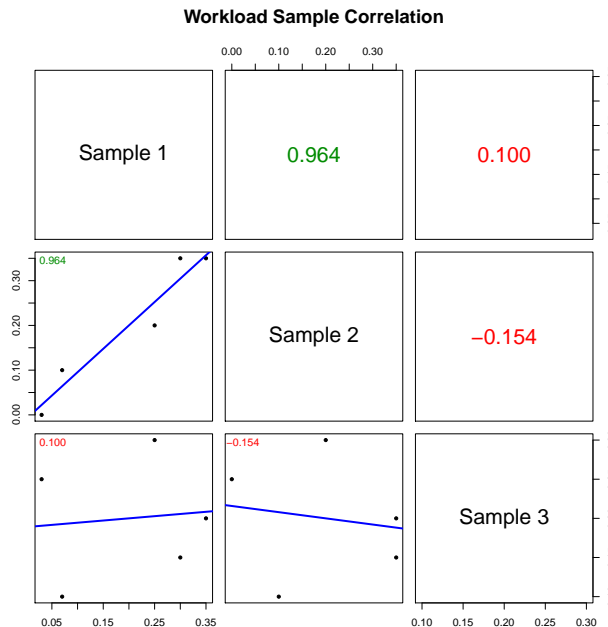
We will adjust the concept of correlation slightly to apply it to workload. Instead of comparing data for two parameters, we will be comparing two sets of frequencies. In the case of the parameter correlation presented earlier, the data are paired according to the time at which they were measured. Such a pairing can also exist based upon a subject with which the measurements are associated (e.g., a subject’s height and weight). Here, the pairing is formed by the transaction name (e.g., a search transaction).

We will demonstrate the workload correlation concept with three trivial workload samples shown in Table 1. There are five transactions. Each sample consists of a collection of transactions from which the frequencies of the transactions are derived. The first two samples “look” similar to each other and should have a high correlation. The third sample is significantly different.

**Table 1:** Sample Workloads

Transaction	Sample		
	1	2	3
Transaction 1	35%	35%	20%
Transaction 2	30%	35%	15%
Transaction 3	25%	20%	30%
Transaction 4	7%	10%	10%
Transaction 5	3%	0%	25%

Figure 6 shows a scatterplot, the regression line, and the correlation coefficient for each pair of samples. Here, strong correlation is shown in green and poor correlation is shown in red. Sets 1 and 2 are similar; set 3 is significantly different when compared to either of the other two. Other examples can be created where the correlation coefficient can take on any value between -1 and 1.<sup>2</sup> What values are considered to be “good” is subjective.



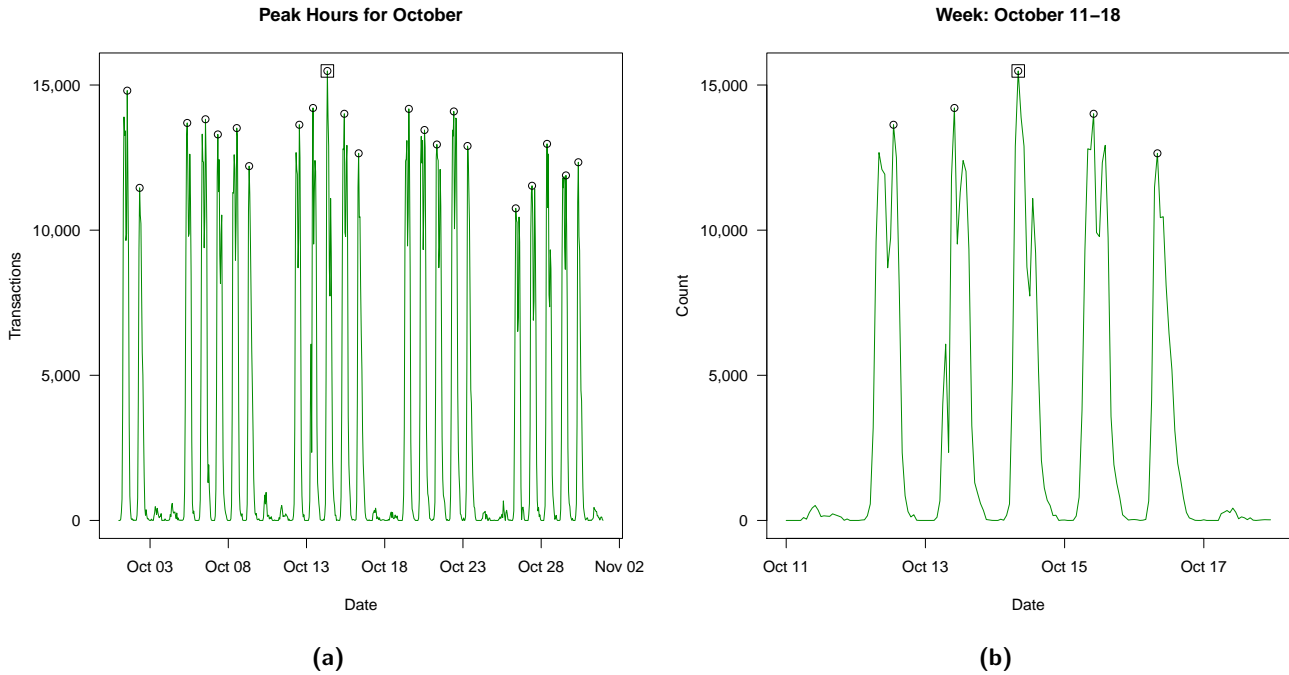
**Figure 6:** This notional example illustrates how correlation is performed for workload samples. A good correlation coefficient is in green; a bad one is in red.

This example had only 5 transactions. Our data have almost 350. That observation does not change the approach—the more data the better.

<sup>2</sup>Whether or not a result of -1 has a logical meaning has not been investigated. Such a result would mean that one workload is a reflection of the other.

## 3.2 Correlation of Various Workloads

If we are going to compare real workloads, we need to define some. Figure 7 shows the transaction counts for a one-month interval where circles highlight the peak transaction hour for each work day and a square highlights the peak transaction hour for the month. We will define several periods for each month: *All hours* (AH) is the period (i.e., the entire month) that contains all data; *work-day maximum hours* (WMH) is the collection of one-hour periods containing the most transactions for each work day (i.e., a week day that is not a holiday); and *monthly maximum hour* (MMH) is the one hour period containing the most transactions for the month. So, in Figure 7, the entire interval represents AH, all of the circles represent WMH, and the square represents MMH. The AH data set contains intervals where the number of transactions varies greatly. The WMH data set is a small subset of the AH data set, where the number of transactions in the intervals does not vary greatly. The MMH data set is a subset of the WMH data set.



**Figure 7:** This chart highlights peak values for the transaction count metric for each weekday (i.e., weekends are not considered) using circles. The maximum value is highlighted with a square. (a) This chart shows the entire month. (b) This chart shows one week in order to highlight that the circles and square correspond to only one hour.

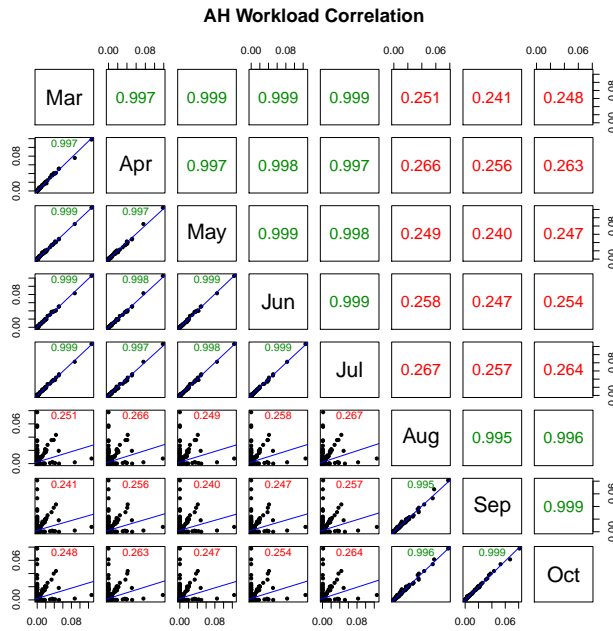
Figure 8 shows a matrix of month-by-month comparisons of the AH workloads. The text color is green if the correlation coefficient is greater than or equal to 0.990, purple if it is greater than or equal to 0.900 but less than 0.990 (there are none in this figure), and red otherwise. The coloring scheme and the associated ranges are arbitrary. A lot of information can be quickly consumed with this visualization approach.

For months March through July, there is high correlation between each pair of months—similarly, for months August through October. When comparing a month from the former range to a month from the latter range, the correlation is very poor. The bad correlation is a result of the transaction renaming modifications as previously described (refer to Section 2). The unique naming of many transactions had a large impact on the correlation because the pairings are affected. However, we can probably safely conclude that the workloads are consistent over the eight-month interval.

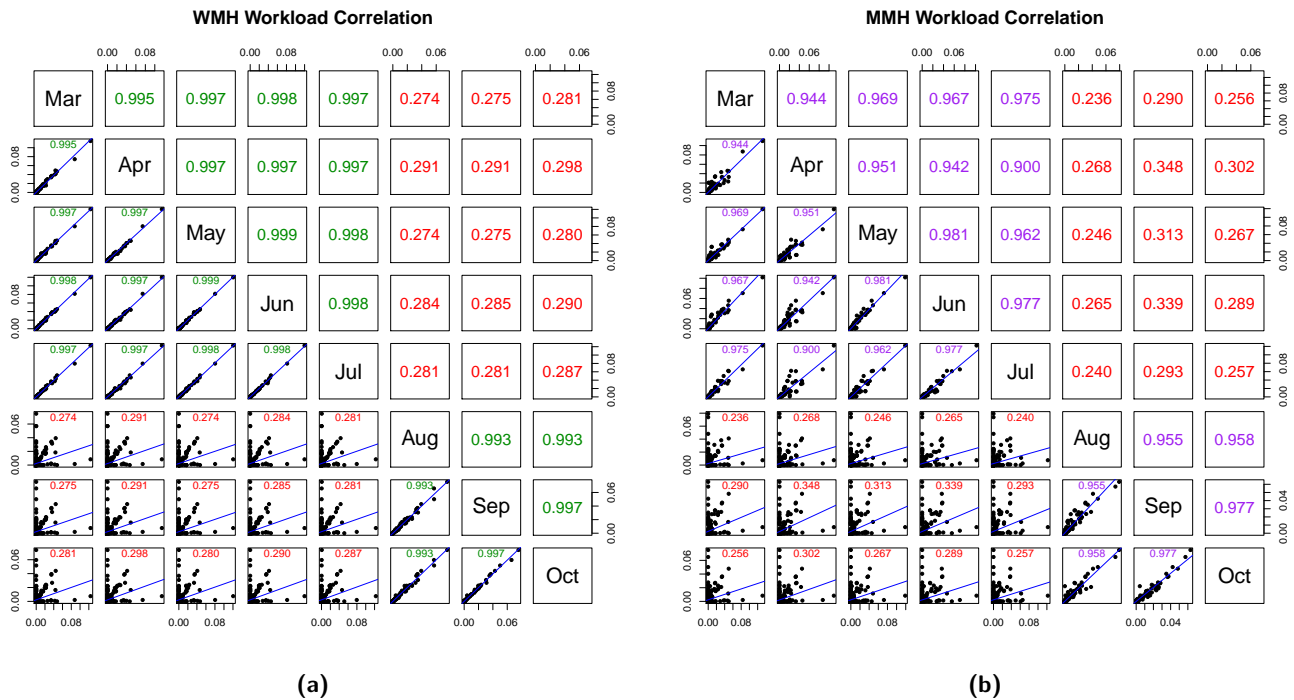
Figure 9a shows a matrix of month-by-month comparisons of the WMH workloads. The analysis is similar to that for the AH workloads. Again, we can conclude that the workload is not changing appreciably for different WMH intervals.

Figure 9b shows a matrix of month-by-month comparisons of the MMH workloads. In this case, all of the correlation coefficients are purple, reflecting less correlation than the other workloads. While the workloads have high correlation, there is more variation across the workloads in comparison to the AH and WMH workload correlations. While the risk is probably low that a bad test would be created, it would be advisable not to base a test on such a sample if using a larger sample is not costly to analyze.

All of the matrices of graphs provide a quick glance at the comparisons. The conclusion that we can draw is this: Constructing a workload from an interval that is one hour out of a month carries some risk, even if the risk is low. However,



**Figure 8:** This matrix of charts shows a month-by-month comparison of the AH workloads. Color is used as a quick indicator of high and low correlation.



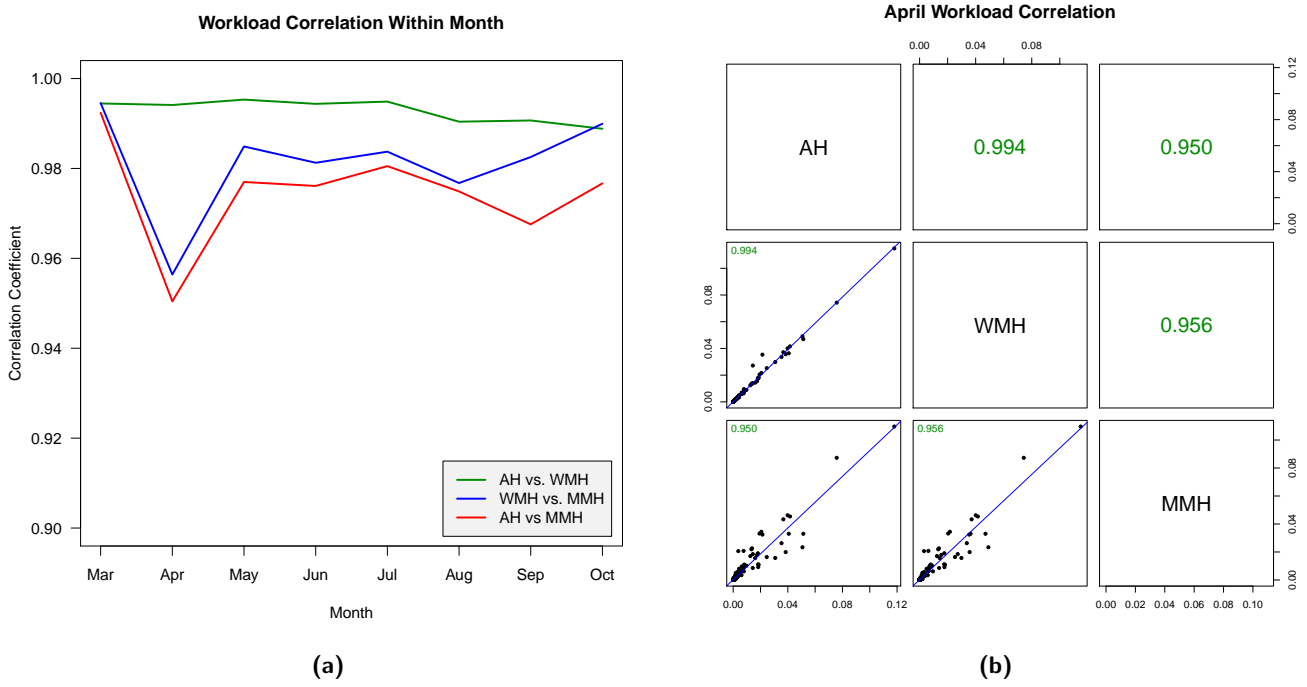
**Figure 9:** These two matrices of charts show a month-by-month comparison of the WMH and MMH workloads. The correlation of the MMH workloads is slightly less than the correlation of the AH workloads or the WMH workloads.

it is just as easy to construct a workload from a larger sample that has even lower risk. We prefer the WMH workload to the AH workload, even though the two workloads do not seem to be noticeably different.

Next we consider comparing the different workloads within a month: AH vs. WMH vs. MMH. Rather than creating a collection of matrices of graphs, we have opted to create a graph of correlation coefficients. Figure 10a shows these comparisons in a chart. While we have plotted the results with lines, we should note that the x-axis is categorical. This



presentation format is simpler to visually consume than a collection of bars in a bar chart or a collection of matrices. What this says is that there is some variation in workload if we limit ourselves to sampling one hour out of the month. If we limited ourselves to sampling the work day hours of one month (typically including 20-23 hours), there would be less risk.



**Figure 10:** (a) This chart shows how the different workloads compare to each other within the same month. The WMH workload consistently correlates against the AH workload. The MMH also has good correlation against AH and WMH, but variation is higher. We should highlight that all values are above 0.9 and are still good. Note that the y-axis does not extend to 0. (b) This chart compares correlation results for various workloads within the month of April. The results define the points in the April column of (a).

Figure 10b compares the three data sets within April to illustrate how the data in Figure 10a are computed. April was selected because it had the lowest correlation coefficients in Figure 10a. What the comparisons say is this: When WMH is compared to AH, a very high correlation exists. When MMH is compared to AH or WMH, the correlation is a little lower.

Considering AH alone, there does not seem to be very much workload variation. This is further supported by analysis of WMH. We can also conclude that workload is not changing with growth since our trending analysis shows that the user base has grown during the eight-month interval.

## 4 Use R!

We have intentionally not made this paper about R, but want to highlight its merits. Originally, most of the analysis was done in Microsoft Excel. The major exceptions are the matrix figures. Excel's advantages are (1) familiarity to others, (2) an interactive nature, and (3) the ability to see a lot of numbers quickly. The interactive nature comes from using forms/controls to select data and to turn features on and off. Excel's disadvantages are (1) slowness in computing large spreadsheets, (2) the difficulty in creating some chart types (e.g., creating a boxplot), and (3) the changes that occur across releases and/or inconsistency across platforms (e.g., PC vs. Macintosh). It is certainly an acceptable analysis tool, especially considering advantage #1.

The usage of R has been written about previously (e.g., [Hol04] and [Hol05]) and there are numerous books (e.g., [ZIM09] and [Sar08]). Once the learning curve is overcome, it is possible to process data more quickly than with Excel. An expert in Excel requires time to build a system of formulas to present the data. Significantly less work is required in R. R allows us to graph our data quickly and control its presentation. It allows us to put many parameters in the same figure or examine the data according to various groups. The graphs in this paper were created with the standard graphics package ([R D09]) and the lattice package ([Sar09]).

## 5 Conclusions

One of our primary concerns was the selection of a sample of real workload on which to base a test workload. The existing choice was to select the one hour period within a month that had the largest transaction count. Such samples from several months were compared to each other as well as to larger samples. A tremendous amount of data can be quickly consumed when presented as a matrix of graphs. We concluded that choosing such a sample is not very risky, although some of that risk is unnecessary. Choosing a larger sample resulted in very little variation in content. Larger samples included peak-transaction-hours from many work days of the month. The visualization techniques were not necessary, but were helpful in reaching this conclusion.

We also looked at growth trends using visualization. This confirmed our expectation of how the system was being used. Visualization was also not necessary here. However, visualization simplifies finding irregularities in the usage patterns (like the login spike of Figure 3c).

We computed some correlations between metrics for the system. A key relationship was the one between users and transactions. We confirmed this relationship and suggested corrections to our testing model. Again, the visualization is not necessary, but is helpful in communicating the results to a decision maker.

## References

- [EM02] Said Elnaffar and Pat Martin. "Characterizing Computer Systems' Workloads". Technical report, 2002.
- [Hol04] James Holtman. "Using R for System Performance Analysis". *Proceedings of the CMG 2004 International Conference*, 2004.
- [Hol05] James Holtman. "Visualization Techniques for Analyzing Patterns in System Performance Data". *Proceedings of the CMG 2005 International Conference*, 2005.
- [R D09] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009.
- [Sar08] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Use R! series. Springer Science+Business Media, 2008.
- [Sar09] Deepayan Sarkar. *Lattice Graphics*, 2009. R package version 0.17-25.
- [ZIM09] Alain Zuur, Elena Ieno, and Erik Meesters. *A Beginner's Guide to R*. Use R! series. Springer Science+Business Media, 2009.