**Understanding the CAD Data Generating Process**

**White Paper #1**

**By**

**Marc L. Swatt**

**2015**

This paper describes the “Data Generating Process” (DGP) behind the Los Angeles Police Department's (LAPD) Computer Aided Dispatch (CAD) data. The DGP can be understood as a model that generates the data at hand. In the paper the DGP is discussed using a set of rules that govern 1) whether valid values for response time are observed, 2) whether the observed values of response time are problematic (outliers), and 3) the expected mean for valid observed values. It is anticipated that a better understanding of the DGP will assist the LAPD in diagnosing problems with the CAD data, develop strategies for improving the quality of the CAD data, and determining appropriate metrics that can be used to evaluate response time performance.

**Missing Data Process**

Missing data is an inherent part of most data systems and often must be accounted for in descriptions of the DGP. For the purposes of this analysis we can separate the types of missing data into four separate categories: 1) missing by design, 2) conditional missing with observed mechanism, 3) conditional missing with unobserved mechanism, and 4) unconditional (or random) missing.[[1]](#endnote-1)

When missing data arise through some deterministic mechanism that either intentionally or unintentionally yields consistent missing values, this can be discussed as “missing by design.” For values that are missing by design, a key objective is to determine the rules that govern the production of missing values. For the other three types of missing data mechanisms, missingness is stochastic rather than deterministic, meaning that missing values are produced through a draw of some probability distribution and therefore contain an element of randomness. For unconditional missing data, the probability of missingness is independent of other available information in the data, the unobserved value of the variable, and any other outside mechanism not described in the data. For conditional missing data with an observed mechanism, the probability of missingness depends on the values of one or more variables with values that are observed in the data. After identifying and removing these conditional relationships the remaining missing data are unconditional missing. Finally, for conditional missing with an unobserved mechanism, missingness depends on information that is outside of what is collected in the data.

For the purposes of this examination, we examine "missingness" relative to total response time (creation to arrival) for the 2014 LAPD CAD data. Incident response times were calculated using the standard SQL query that produces these results in reports and these response times were appended to the CAD data. We first identify the deterministic rules that establish missing by design. After this we identify whether a conditional relationship exists between missingness and the values of additional provided variables using Classification and Regression Trees (CART, see Berk, 2008 for a description). It is assumed that any remaining missing values result from Bernoulli trials with a probability approximately equal to the observed frequency of missingness in the terminal nodes. We acknowledge that there may be additional unobserved mechanisms that govern the missing data process, but further exploration requires additional information beyond what is contained in the current data.

**Missing by Design**

Out of the 1,787,656 calls for service received in 2014, a total of 994,590 or 55.64% had missing data for total response time. From this, we were able to identify 5 rules that account for 938,525 or 94.36% of all missing data. Three rules were established in the SQL response time calculation and two rules were present in the data.

From Query:

Rule 1:

I\_PRIORITY > 3 => 4

Rule 2:

ITI\_TYPEID = 006 => 837,047 (also CALLFROM = “OFCR”)

Rule 3:

ITI\_TYPEID = 902 => 87,350 (always CALLFROM = “OTHER”)

From Data:

Rule 4:

ITI\_TYPEID = 200, 2001, 720, 7201, 7202, 7203, 7204, 720O, 820, 8201, 8201W, 8202

8202W, 8203, 8205, 8205W, 820H, 820O, 820OW

=> 12,328 (always CALLFROM = “OTHER”)

Rule 5:

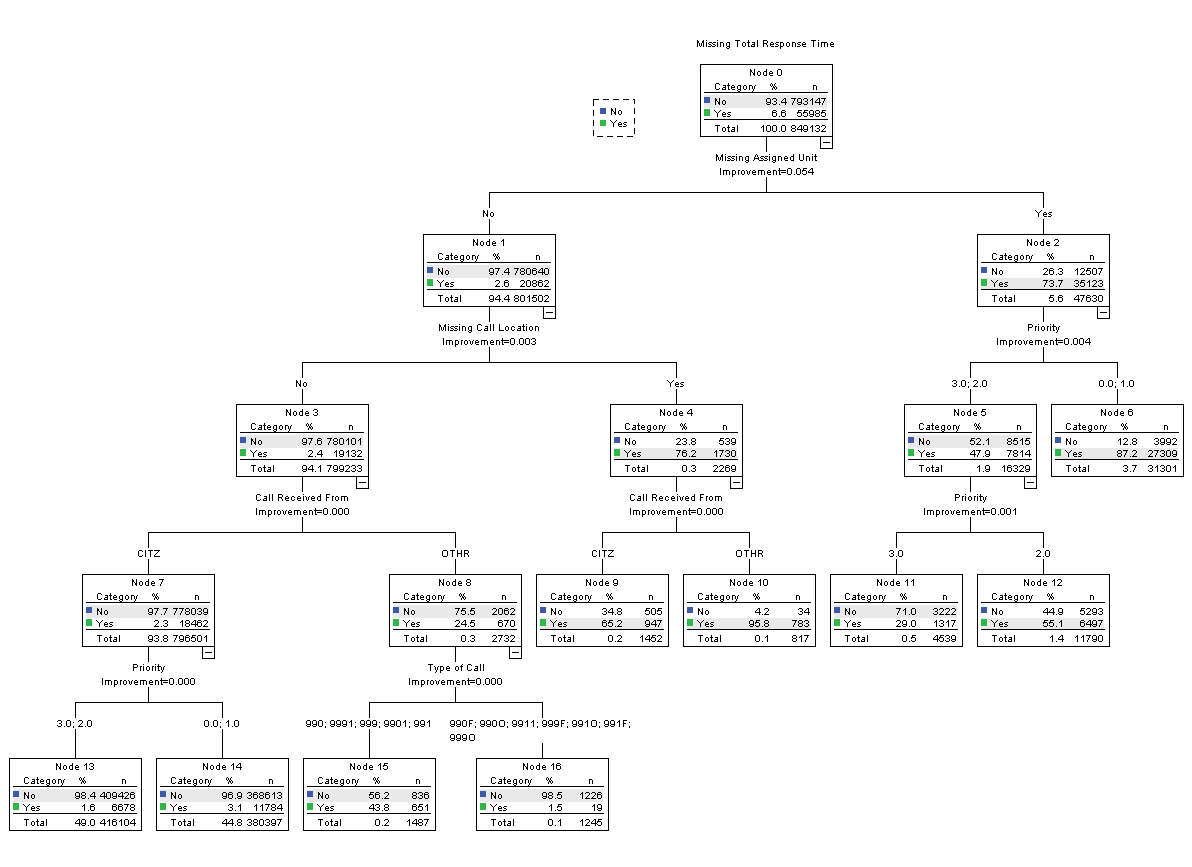
ITI\_TYPEID = 100 => 1,796 (always CALLFROM = “CITZ”)

**Conditional Missing with Known Mechanism**

For the remaining 56,065 calls with missing data (approximately 3.1% of total calls), we were unable to identify deterministic rules that accounted for missingness based on the data provided. As an alternative, we used the information provided in other fields to generate a conditional model that captures the interrelationship between these variables and whether the value for total response time is missing. Specifically, we used the CART algorithm available in SPSS 22 to generate a classification tree that models missingness on the basis of the available fields. While this model is being used for descriptive purposes, similar CART models can form the basis for multiple imputation models for missing data (see Burgette & Reiter, 2010).

The CART model for missingness is presented in Figure 1. This model produced a total of 16 nodes, 8 of which were terminal nodes, suggesting that a series of eight rules (each generated by following the branches to the terminal node) can be used to explain the observed patterns of missingness. The precision of the predicted model is presented in the classification table (also called confusion table, see Berk, 2008) below the tree with an overall precision of 96.4%. The model was better at predicting non-missingness (98.8%) than missingness (63.5%), but still offers a considerable improvement in the identification mechanism generating missingness.[[2]](#endnote-2)

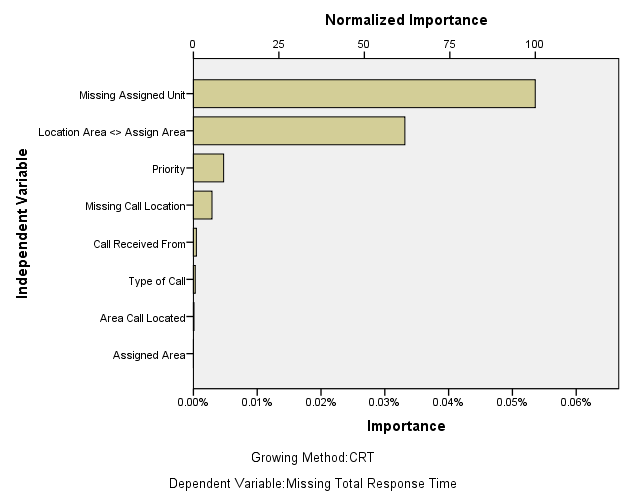
Figure 1. CART Model for Missing Response Time Values.



|  |  |  |  |
| --- | --- | --- | --- |
| **Classification** | | | |
| Observed | Predicted | | |
| No | Yes | Percent Correct |
| No | 783323 | 9824 | 98.8% |
| Yes | 20449 | 35536 | 63.5% |
| Overall Percentage | 94.7% | 5.3% | 96.4% |
| Growing Method: CRT  Dependent Variable: Missing Total Response Time | | | |

In Figure 2, we present a visualization of the “normalized importance” of each of the independent variables that were used in the CART fitting algorithm. This provides a measure to gauge the relative impact of variables in the CART model as the strongest predictor is set at 100% and all other predictors are gauged relative to the strongest predictor. The variable “missing primary unit assignment” had the greatest impact in the model as cases where unit assignment was missing (47,630) were missing response time 73.7% of the time, whereas cases where unit assignment wasn’t missing (799,233) were missing response time only 2.4% of the time. The variable measuring whether the area where the call occurred was the same as the area assigned also had considerable importance, but this was predominately because it shared a substantial amount of information with whether the unit was missing (further discussed below). Beyond these two variables, only priority (0,1 vs. 2,3) and missing location of event had any additional appreciable impact.

Figure 2.



Missing unit assignments were all associated with cases where the area where the crime was located did not match the area that was assigned the call. Specifically, whenever there was a missing value for assigned unit, the call location area did not match the assigned area. However, the converse was not always true, when the call location area did not match the assigned area, 60% of the time primary assigned unit was missing and 40% of the time it was not. This relationship appeared conditioned by Priority as the proportion of time that a mismatch between call location area and assigned area was high for Priority 0, 1 (76.3%), moderate for Priority 2 (52.7%) and low for Priority 3 (28.3%).

**Response Time Distribution and Outliers**

Of the total 1,787,656 calls for service received in 2014, 793,066 or 44.36% had observed values for total response time. Outliers, or observations with values that are: 1) outside of the bounds of logically observable values (such as negative or zero response times), 2) sufficiently outside the bounds of the mass of observed data that they are likely to have arisen through error (user error, equipment failure, etc.), or 3) represent isolated or rare events that are qualitatively different than the majority of observed events (ex. FD/EMT response to a plane crash).

Of these three different types of outliers, only the first category can be directly observed without reference to the underlying distribution of the data. In the current data, only 194 (0.01%) have negative or zero values. Most of the inappropriate values are zero (167 or 86.1%) with the remainder being negative values ranging from -3385 to -71 seconds. All of the negative time calls were the result of calls from citizens as were 71.3% of the calls with zeroes. Only 48 (or 28.7%) of the zero calls were calls from others (non-citizen, non-officer). Out of the total 194 cases with discrepant values, 96 (49.5%) had a Priority = 0, 73 (37.6%) had a Priority = 2, and 25 (12.9%) had a Priority = 3. The cases occur too infrequently for the CART algorithm to provide any additional detail regarding the mechanism that generates these cases. It is recommended that these values be treated as missing or an individual case review of these 194 cases be conducted to determine specifically how these values are being generated.

**Distribution of Total Response Time**

The minimum observed value for total response time after removing negative and zero response times is 0.02 minutes (1 second) and the maximum observed value is 1456.3 minutes (24.3 hours). Not surprisingly, the distribution has considerable skew (Sk = 4.474) with the mean (36.53) considerably higher than the median (20.48). The distribution is highly leptokurtotic as well, suggesting that a considerable amount of values cluster around a high peak (Kt = 33.139). Order statistics provide a value of which the specified percentage of cases fall below. The order statistics for every 5% increment are presented in Figure 3 below.

These order statistics indicate that 90% of observed response times fall between 3.62 minutes and 123.52 minutes (or just over 2 hours). A histogram for total response time is presented in Figure 4. A very obvious peak can be observed and although the bars are difficult to see, they extend nearly to the 1500 minute mark (25 hours), which is well beyond the 95 percentile.

Figure 4. Histogram of Total Response Time.



A hanging rootogram makes the skewness and kurtosis of the distribution easier to see. The hangroot plots the normal curve against the square root of the frequency. The hanging rootogram for the untransformed distribution is presented in Figure 5. The zero reference line presents the expected value in relation to the normal distribution. If the bars do not reach the zero line, the bar has fewer cases than expected under a normal distribution. If the bars extend beyond the zero line, there are more cases than expected under a normal distribution. To assist visualization, this rootogram truncated the data at 800 minutes (far outside the 95 percentile).

Figure 5. Hanging Rootogram for Total Response Time



Based on the skewness and kurtosis of the untransformed distribution, it is clear that the DGP does not result from a normal distribution. It is possible, however, that a logarithmic transformation of the distribution is approximately normal. This would indicate that total response time follows a lognormal distribution (which is often encountered in queuing theory as well as many natural processes, see Limpert, Stahel, & Abbt, 2001).

A histogram for the natural logarithmic transformation of total response time is presented in Figure 6 below. It is fairly apparent that the log transform has resulted in a distribution that is approximately normally distributed.[[3]](#endnote-3) While there still appear to be outliers (particularly at the low end of this distribution, many of the seemingly large numbers of outliers now lie under the normal curve. To reinforce this point, a hanging rootogram of the transformed data is presented in Figure 7 below. From this graph, the remaining outlier issues after transformation appear to be predominately at the low tails of this distribution beyond 0, which corresponds to less than 1 minute response time. There are 3,175 observations with an observed response time less than 1 minute. This represents less than 0.5% of the total records with valid response times.

The implication of using the lognormal distribution is that the geometric mean and multiplicative standard deviation (corresponding to the mean and standard deviation of the transformed distribution) are the appropriate measures of central tendency and dispersion. The geometric mean (20.4428) is substantially smaller than the arithmetic mean (36.5268) for these data. The median (20.4833) is often a common choice for a measure of central tendency in a lognormal distribution as the exponentiated median of the transformed distribution is equal to the median of the untransformed distribution.

Figure 6. Histogram of Log Transformed Total Response Time



Figure 7. Hanging Rootogram of Log Transformed Total Response Time



While the median is often recommended as a robust measure of central tendency due to its high breakdown point, it is a very inefficient measure when the data are approximately normal (see Andersen, 2008). In this instance, the logarithmic transformation of response time shows a remarkably normal distribution save the lower tail. A compromise measure would be a trimmed mean, where the highest and lowest percentile of the distribution would be eliminated prior to the computation of the mean. This introduces some slight inefficiency, but provides some protection against outliers given the current data. Table 1 presents the values of a number of difference options for a trimmed mean. The percentage of trim refers to the percentage of data discarded at each tail, so the Trim 1% Mean eliminates the top and bottom 1% of cases. No trim corresponds to the mean, Trim 25% Mean is often called the Interquartile Mean, and the 50% Trim Mean is the Median. Confidence intervals are presented for all options other than the median, as these confidence intervals typically require bootstrapping to generate confidence intervals.[[4]](#endnote-4)



The lower trim percentages are larger than the mean as they remove the outliers at the lower end of the distribution. Of all the estimators, the Trim 1% has the narrowest confidence intervals and is the most efficient estimator of central tendency. For the current data, the Trim 1% appears to provide sufficient protection against outliers and will be estimator used for the remainder of this report. However, for applications using less than a full year, it may be necessary to consider using a larger trim depending on the average number of records across the time period.

**Predicting Total Response Time**

A final objective of this report is to explore the relationship between total response time and other variables available in the data. These relationships may either help to identify types of calls that should be examined separately or to diagnose potential problems in the CAD data. Exploratory analyses often form the basis of later predictive models.

An obvious relationships worth examining is the relationship between call priority and response time. Priority 3 calls have the lowest expected response time (Trim 1% geometric mean) at 6.95 minutes. Priority 2 calls have an expected response time over twice as long at 16.87 minutes. Finally, Priority 0 calls have an expected response time twice as large as Priority 2 calls with an expected response time of 34.16 minutes. This relationship is illustrated below using boxplots in Figure 8. The boxes represent the interquartile distance (25% and 75% order statistics); the center line represents the median (50% order statistic); the whiskers extend 1.5 times the interquartile distance, and the dots are points outside this range. Importantly, these box plots use the log transformed response time with no trimming.

Figure 8. Box Plots of the Transformed Total Response Time over Call Priority



An exploratory CART model was used to examine the relationship between a number of variables and the log transformed response time. For this model, the top and bottom 1% of total response time were trimmed. Since this model is merely exploratory, no data was withheld for cross validation purposes.Variables considered were priority (0, 2, 3,), caller (citizen v. other), assigned area, location where call occurred, whether primary assigned unit was missing, and whether there was a mismatch between location of call and the area where the call was assigned. Unfortunately, the finer grained-call type variable could not be included into the model without substantial reduction in the number of categories.

The CART model is presented below in Figure 9 (a high zoom is necessary to navigate). The resulting model explaned only 29.90% of the variance in response time, so additional information outside of the provided data would be needed to improve the predictions. In Figure 10, we present the normed variable importance. Not surprisingly, call priority was the most important predictor of response time by far. The assigned area of the call and location of the call had small impacts as well.

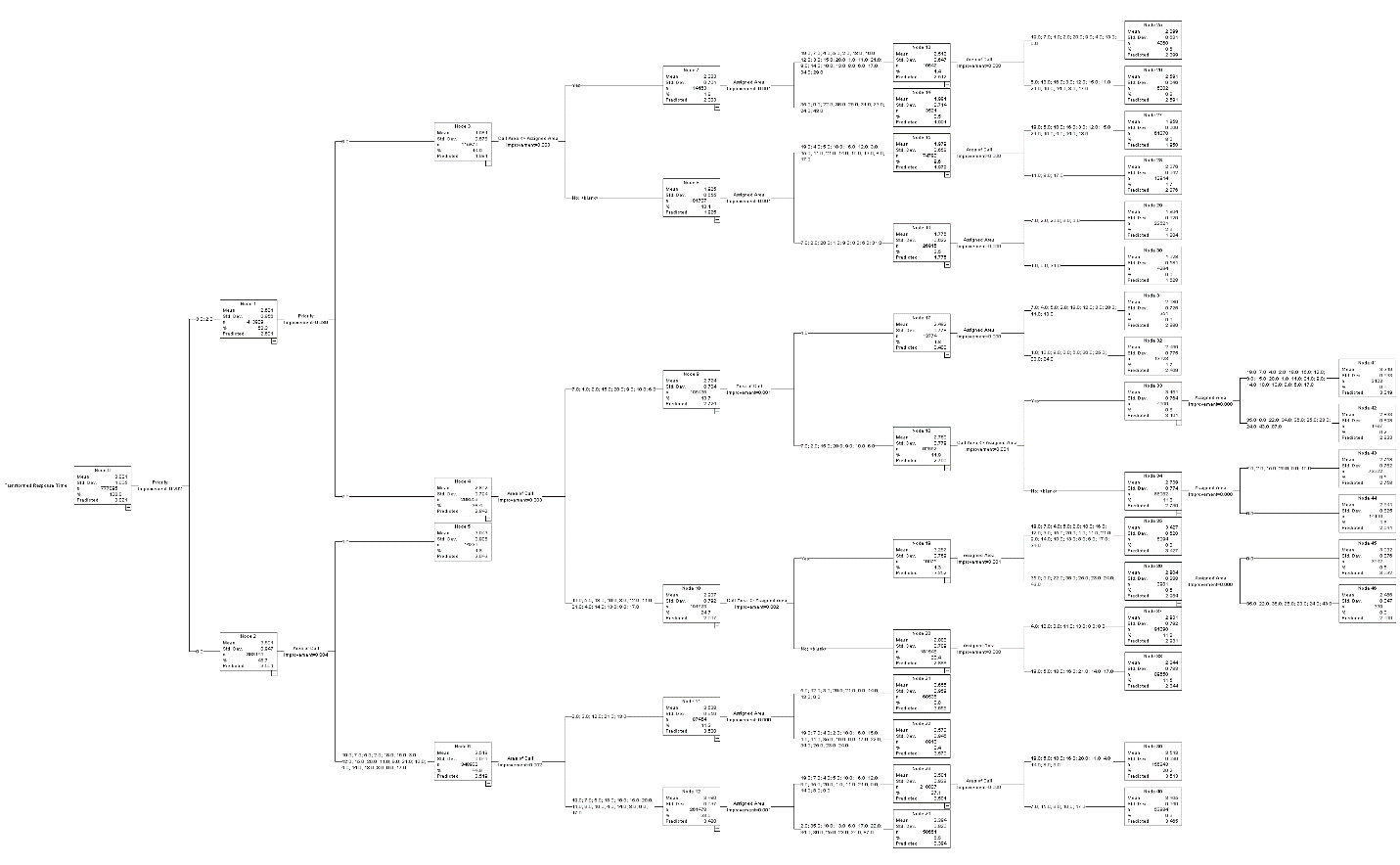
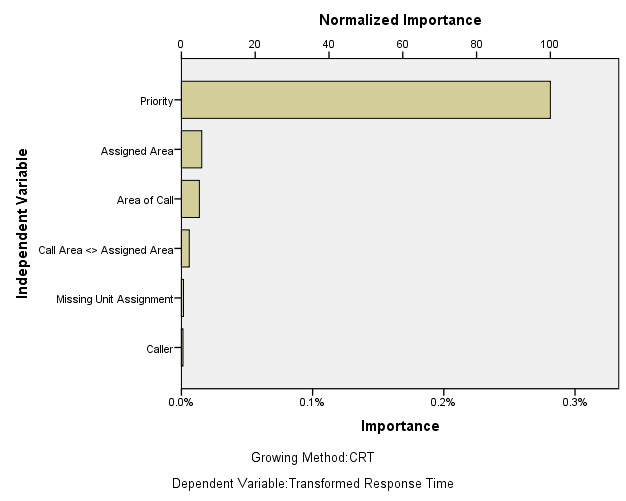
Figure 9. CART Model for Total Response Time

Figure 10. Normalized Importance of Variables in CART Model for Total Response Time



**Conclusions**

This project examined CAD data from 2014 to better identify the DGP for response time by examining: 1) The nature and extent of missing data; 2) The frequency and seriousness of outliers; and 3) What factors are associated with the observed values of response time. From this exercise, we can identified what we have learned and the recommendations that we would make.

**1. Missing Data is Not Very Problematic.**

In these data, there is a substantial amount of missing data, 55.64% of the total records were missing calculated values for total response time after running the standard query. However, of these, 94.36% of all missing data appeared to be missing by design. The rules that were identified as missing by design were a combination of intentional omissions from the query (inappropriate priorities, call type 006, and call type 902) and a number of call types that do not appear to provide valid response times (a number of types identified as a call from “other” and type 100 from “citizen”).

Further, of the remaining 5.64% of missing cases (which corresponds to approximately 3% of the total records), the strongest predictor of missingness appeared to be whether or not a primary unit was assigned to the call and these all occurred when there was a mismatch between the location code for the area the call was received and the location code for the assigned area for the call. This hints at the possibility of an additional undetected rule for the observed pattern of missing data. If this is the case, then the fraction of missing data not by design drops further.

Regardless, the amount of conditional missing data is small and it would be reasonable to simply omit these data points from an analysis. Allison (2001) suggests that casewise deletion does not introduce bias when the data are assumed missing completely at random (although it is inefficient). Even in situations where the missingness is associated with one or more observed variables, casewise deletion is a reasonable strategy where the proportion of missing is less than 5%. If missingness remains a concern, a model similar to the CART can be used to compute weights to correct for the differential probability of missingness.

**2. Total Response Time Follows a Lognormal Distribution.**

At first glance it appears as though there are a substantial amount of outliers in the response time data – this is a consequence of assuming that response time should be normally distributed. However, on further examination after transforming the original data using the natural logarithm, most of the large outlying points disappear and the variable appears to follow the normal distribution closely for most of the curve. This suggests that the lognormal distribution is appropriate for total response time and the appropriate measures of central tendency and dispersion are the geometric mean and the multiplicative standard deviation respectively.

The lognormal distribution is fairly common, especially in applications of queuing theory. Queuing theory is often used to describe processes similar to a CAD environment, such as the distribution of telephone calls to a call center, process requests to a central computer server, and wait times in a queue at a bank or other business. Often, service delivery time is assumed to follow a lognormal distribution – which matches the finding here.

One of the key advantages of identifying the appropriate distribution is that it is possible to generate control charts which are involved in statistical process control. Since the log transform of the distribution is approximately normal, it should be fairly straightforward to develop and implement control charts that lend themselves to monitoring response time metrics on a yearly, deployment period, or even weekly basis given the volume of data obtained by the CAD system. Further, it should be possible to chart and monitor response time by priority and by division. This strategy can be used for quality assurance purposes to ensure that response times continue to remain within designated benchmarks on a rolling basis.

**3. A 1% Trimmed Mean Will Eliminate Most Outliers.**

Only approximately 0.01% of all records have inappropriate values for total response time (zero or negative response times). These can be treated as missing values with little impact on response time metrics. After eliminating these cases, there are very few remaining outliers according to the lognormal distribution. The majority of the remaining outliers appear to be records that have a response time less than one minute. Trimming the top 1% off the high and low values of the distribution provides sufficient protection against outliers while retaining the most data.

Previously the median response time was calculated as the median is strongly resistant against outliers. Further, the median of the untransformed distribution is equal to the exponentiated median of the transformed distribution. This ensured that the median response time was consistent regardless of the shape of the distribution, a property not shared by the simple mean. However, the median is an inefficient estimator and offers very little in regard to performance management. A trimmed mean from the appropriate statistical distribution provides protection against outliers but is substantially more responsive to policy changes or operational practices. As such we recommend the 1% trimmed geometric mean (or the exponentiated 1% trimmed mean of the log transformed distribution) as the appropriate metric of total response time.

**4. The Current Data Do not allow for Predictive Modeling.**

There is a clear relationship between call priority and response time and response time should be monitored across Priority 0, 2, and 3 calls to ensure that response times remain within established benchmarks. Beyond this, however, very few available variables have a substantial relationship with response time. There are some indications that the area where the call was located and the area where the call was assigned is related to total response time, but their overall impact was small and the predictive power of the model was minimal.

The lessons of “big data” suggest that there is some practical value in extracting the most information possible out of extensive datasets like the CAD data. Specifically, non-obvious relationships between input variables and our target variable, total response time, may point to strategies that the LAPD can employ to decrease response time, increase citizen satisfaction with service delivery, and possibly minimize the injury or loss associated with criminal events. To this end, it is clear that the current data need to be augmented with additional information from the CAD system and possibly external sources of data to yield consistent predictive models. This may be a productive line of inquiry in future projects.

**5. A Finer-Grained Examination of Response Time May Be Useful.**

While the current report considers total response time by call for service, there are many aspects of the dispatch system that is not considered. We have not examined how units are assigned calls nor situations where more than one unit may be responding to a single call. We have also not considered dispatches across units to identify possible inefficiencies or determining metrics that enable monitoring of unit performance. Further work also remains to adequately specify control charts and corresponding monitoring protocol for weeks, deployment periods, or years for shifts, divisions, or bureaus. In short, a substantial amount of possible work remains to be done to better understand the CAD system and this report should be seen as an initial attempt towards that end.

**Next Steps**

**1. Examine Differences across the LAPD**. We have not yet examined whether there are important differences in average response time by priority across bureaus or stations in the LAPD. This information can be very helpful in setting performance targets for these administrative areas.

**2.** **Examine Trends in Response Time over Time**. To date, we have only examined response time over calendar year 2014. It may be insightful to consider how response time performance has changed over a longer period of time (ex. 2010 – 2014). This requires additional CAD data.

**3. Evaluate the Performance of the 1% Trimmed Geometric Mean**. While the 1% Trimmed Geometric Mean was demonstrated to be effective for measuring performance across the entire LAPD, it remains to be seen whether this measure will remain appropriate with a smaller amount of data.

**4. Separate Response Time into Components and Develop Metrics**. Although we have examined total response time, it may be important to assess performance over the distinct steps that constitute the response to a call for service. From this, we hope to develop metrics and a method to benchmark performance goals for each component.

**REFERENCES**

Allison, Paul D. 2001. Missing Data. Thousand Oaks, CA: Sage.

Andersen, Robert. 2008. Modern Methods for Robust Regression. Thousand Oaks, CA: Sage.

Berk, Richard A. 2008. Statistical Learning from a Regression Perspective. New York, Springer.

Burgette, Lana F. & Jerome P. Reiter. 2010. Multiple imputation for missing data using sequential regression trees. American Journal of Epidemiology,

Limpert, Eckhard, Werner A. Stahel, & Markus Abbt. 2001. Log-normal distributions across the sciences: Keys and clues. Bioscience, 51, 341-352.

StataCorp. 2013. Stata User’s Guide. Release 13. College Station, TX: Stata Press.

1. **NOTES**

   In statistics, it is necessary to make assumptions about missing data when it is present. The three assumptions are that missing data is missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). A different terminology was created because these terms were discussing assumptions about the data whereas we are discussing the data themselves. Practically, when missing data are unconditional missing, the assumption of MCAR is appropriate. When data are conditional missing with an observed mechanism, the MAR assumption is appropriate. And when the data are conditional missing with an unobserved mechanism, the NMAR assumption is appropriate. [↑](#endnote-ref-1)
2. Caution is necessary as these numbers to not imply how accurate the model would be predicting either missing data from previous or subsequent time periods. Because the purpose of this analysis is to describe the observed patterns of missing data rather than to predict missingness in new data, we did not “prune” the tree to remove superfluous branches or generate a hold-out sample to gauge the degree of “overfit” of the model (see Berk, 2008). As such, we cannot conclude to what degree the performance of our model is generalizable to data that was not included in the fitting process. [↑](#endnote-ref-2)
3. While explicit tests for normality exist (ex. Kolmogorov-Smirnov, Shapiro-Wilks, Shapiro-Francia, and the D’Agostino et al. tests, see StataCorp, 2013), these are not recommended as the large sample size will result in the detection of nearly any departure from normality.

   [↑](#endnote-ref-3)
4. An exact closed form solution for the standard deviation of the median does exist, but it is computationally difficult. Because of this bootstrapping is a common strategy for estimating the standard deviation of the median. [↑](#endnote-ref-4)