11-4-2011

# Modeling and Estimation for Transit On-time Performance Improvement

Xiaobo Wang
*Florida International University*, xwang007@fiu.edu

**DOI:** 10.25148/etd.FI11120513

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

# MODELING AND ESTIMATION FOR TRANSIT ON-TIME PERFORMANCE IMPROVEMENT

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

Xiaobo Wang

2011

To: Dean Kenneth G. Furton
College of Arts and Sciences

This thesis, written by Xiaobo Wang, and entitled Modeling and Estimation for Transit On-time Performance Improvement, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

Florence George

Zhenmin Chen

Hassan Zahedi, Major Professor

Date of Defense: November 4, 2011

The thesis of Xiaobo Wang is approved.

Dean Kenneth G. Furton
College of Arts and Sciences

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2011

ABSTRACT OF THE THESIS

MODELING AND ESTIMATION FOR

TRANSIT ON-TIME PERFORMANCE IMPROVEMENT

by

Xiaobo Wang

Florida International University, 2011

Miami, Florida

Professor Hassan Zahedi, Major Professor

Transit agencies have the opportunity to improve the delivery of services by using data from Intelligent Transportation Systems (ITS). On-time performance is an important measure. The objective of this paper is to adjust the timetables so that the probability of on-time performance is maximized. For this purpose we analyze data distributions of travel time and also consider the general case that data distribution is unknown. Statistical procedures are presented to find scheduled time for some selected distributions. Monte Carlo simulation is introduced for the purpose of finding scheduled time when data distribution is not known. Simulation studies indicate that the on-time performance would increase using the proposed methodology. The contribution of this paper is to provide transit system a procedure to set up or update their timetables based on current ITS data and its distribution, and hence increase level of service.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

# INTRODUCTON

The problem to be examined in my thesis involves how to improve transit on-time performance by modeling and estimating the distribution of transit travel time. The travel time in this paper refers to a period of time spent traveling between two adjacent stops, usually denoted by a random variable. The on-time performance is defined as percentage of the times that a bus is considered to be on-time, which used to indicate buses arriving or departing late, on-time, or early. Depending on the transit agency, on-time performance can be calculated by using arrivals, departures, or possibly a combination of both. In this research, on-time performance is calculated by using arrivals. That is, if the scheduled arrival internal is interval [*a, b*], the on-time performance is the percentage of times that a bus arrives or departures within the scheduled interval.

Transit agencies usually use the on-time performance measure as an important performance indicator for transit system. Since the distribution of arrival times for buses can be studied by using statistical methods, then one can use statistical methods to update scheduled time table to maximize the on-time performance.

To better understand this problem, a hypothetical transit route is presented in the next page. The diagram in figure 1 shows a schematic of a hypothetical transit route. The route is divided into a number of timepoints (stops) when a bus traversing along the route. When a transit bus arrives at stop i+1, the actual arrival time is recorded at that particular location. If the bus arrival time falls outside the range of the scheduled interval, which is defined by the transit agency, the bus will be regarded as either late or early. Such process is depicted in Figure 1 under Actual Arrival Time, which shows here that the bus

arrives mostly early. However, if we shift the scheduled time to a little earlier time, chances for the number of early arrivals will be reduced.

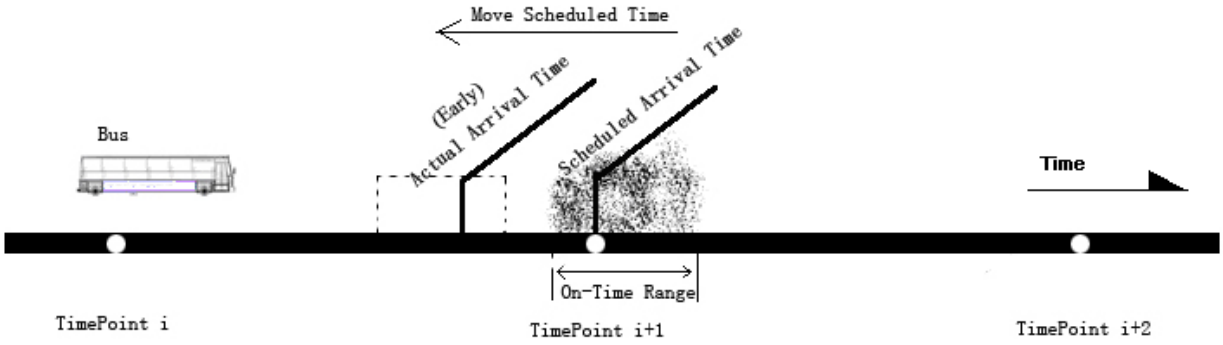


Figure 1.Moving a scheduled time in a hypothetical transit route

Table1. Definition of Variables

| Variable | Description |
|---|---|
| $A_i$ | Arrival time at stop i. |
| $S_i$ | Scheduled time at stop i before adjustment (initially set by transit agencies). |
| $S'_i$ | Optimal scheduled time at stop i after adjustment based on analysis of transit data. |
| $T_i$ | Travel time from stop i-1 to stop i. |
| $a$ | Tolerance time for early arrival, chosen by agencies. $a > 0$. |
| $b$ | Tolerance time for later arrival, chosen by agencies. $b > 0$. |
| $L$ | The width of on-time performance interval. $L = b+a$ (chosen by transit agencies). |
| $[S_i - a, S_i - a + L]$ | On-time interval before adjustment at stop i. |
| $[S'_i - a, S'_i - a + L]$ | Optimal on-time interval after adjustment at stop i. |
| $k$ | The number of stops for a given trip. |

In order to make the scheduled time to match the arrival time better, one solution is to update the timetables at each time point. Thus, such problem is reduced to the following question: What should be the new scheduled time in timetable which maximizes the on-time performance? The first step to address this question is to define the relevant variables in the problem. In the following table 1, we introduce the variables and notations used in this paper.

The goal of this work is to narrow the gap between the scheduled time and the arrival time. That is, we would like to maximize the probabilities that arrivals fall in their scheduled on-time interval as demonstrated in Figure 1. This can be achieved by adjusting the timetables so that the probability of on-time performance in each stop is maximized. Note that the probability of on-time performance is the area under the probability density function for the arrival time $A_i$ over the scheduled on-time performance interval $[S'_i - a, S'_i - a + L]$ . If the scheduled time at stop i is $x_i$, then the probability of on-time performance is equal to: $P\{x_i - a \leq A_i \leq x_i + b\}$ . Thus, our goal is to maximize $P\{x_i - a \leq A_i \leq x_i + b\}$ with respect to $x_i$ for i=1, 2,..., k. That is:

$$Goal: \max_{x_i \geq a} \sum_{i=1}^{k} P\{x_i - a \leq A_i \leq x_i - a + L\}. \quad (1.1)$$

Let $S'_i$ be the optimal value of $x_i$, then

$\sum_{i=1}^{k} P\{S'_i - a \leq A_i \leq S'_i - a + L\} \geq \sum_{i=1}^{k} P\{x_i - a \leq A_i \leq x_i - a + L\}$ for all $x_i \geq a$.

## LITERATURE REVIEW

Researchers have shown a special interest in making use of transit data, such as travel time to improve transit services. Lee et al. (2001) studied the effect of travel time in an Automatic Vehicle Location (AVL) system. Hammerle et al. (2005) pointed out that some transit agencies would like to use transit data to provide more reliable service by developing methods for extracting information from these data to compute service reliability indicators.

One main objective of these studies was to measure on-time performance. For example, New York City transit, see Nakanishi (1997), established a customer-oriented bus performance indicator program to measure the on-time performance. The on-time performance measure becomes increasingly more important since it can easily be used by supervisors and managers, so problematic issues can be detected and fixed early. Information concerning on-time performance can help to improve the delivery of services of the entire transit system, or an individual route in a particular system.

Because only measuring performance indicators is not enough, researchers need to consider how to improve service quality especially on-time performance by analyzing transit data. One approach is to make the real-time transit data available to passengers, transit dispatchers and supervisors. When the real-time information is provided, both bus operators and passengers can better arrange their schedules to be on time. Tri-Met in Portland, see Kimpel et al. (2004), shows that scheduling can be improved with monitoring transit data. Shalaby et al. (2004) have made efforts towards using transit

data to develop bus travel time models to obtain real-time information on bus arrival and departure times with the goal to improve the on-time performance.

To improve the on-time performance, one approach is to match scheduled times with arrival times by updating the timetables. Some studies have attempted to find the relationship between on-time performance and the travel time. For example, Portland State University and Tri-Met, see Strathman et al. (2002), analyzed the on-time performance from the perspective of the travel time variation. Also Cevallos et al. (2008) suggested that to improve an on-time performance, the scheduled travel time should be set at a value slightly less than the mean or the median of the travel times.

From the above reviews, it can be seen that researchers have a special interest in using travel time data to improve transit systems. They have also tried different ways to find the underlined relationships in the data. These studies could be summarized into one question which has been asked frequently: how to model and estimate the service reliability by using current transit data? For example, Ahmed and others et al. (2011) introduced four different multivariate regression models to measure different dimensions of service reliability. All these models were based on travel time under the assumption that the travel time has a normal distribution.

There are some other studies which are based on alternative statistical methodologies than regression approach. For example, Wall and Dailey (1999) developed an algorithm for predicting the arrival time of a transit vehicle using a combination of both the AVL data and the historical data. The algorithm they used in that work was based on the Kalman filter framework and statistical estimation of scheduled time were also done

under the assumption that the transit data has a normal distribution. For Kalman filter, see Kalman (1960).

However, in many practical situations, transit data may not follow a normal distribution. Presently, there are very limited results for the case that the normality assumption is not valid. In this research, we extend the work of a previous paper by Cevallos (2011) that describes a procedure for improving on-time performance at transit agencies. We will introduce several different distributions in addition to the normal distribution and develop a general method to improve on-time performance on the basis of the selected appropriate distribution. In addition, this paper also presents and compares several other new approaches for on-time performance strategies.

In the past, transit agencies typically used two methods to set up or update a transit timetable. One method is based on the transit system manual. Basically, in the manual method they use some key factors to calculate the travel time between two stations, and the travel time is used to set up or update the timetable. The calculation commonly used is usually based on a linear regression model, but this model cannot be used for nationwide transit systems. The second method uses data from running a bus for a number of times in a particular route to estimate a rough timetable. While the final time table will depend on the experience of drivers on the roads. Subsequently, when the transit agencies receive complaints from passengers, they would collect that information to update the timetables for the future. None of these two methods described are quite based on formal maximization of the on-time performance measure. Thus in this research, we attempt to find a procedure to take use of historical data and develop a

methodology with the intention of optimally updating a transit timetable which maximizes the on-time performance. If a distribution for transit data is assumed, usually a Goodness-of-Fit test can be used to validate the assumption. For goodness-of-fit test, see Hans (1967).

My research attempts to answer the problem of maximizing the on-time performance based on the selection of an appropriate statistical/mathematical model for the transit data and then study the model by estimating the related parameters, analytically if possible, or numerically by using simulations. This research also proposes and compares some other nonparametric procedures and strategies to improve the on-time performance by updating the time tables.

## METHODOLOGIES

The objective of this study is to adjust the timetables so that the probability of on-time performance is maximized. Since the probability of on-time performance depends on the distribution of the bus arrival time, therefore, it is necessary to know (or to estimate) the distribution of the arrival time. Note that the bus arrival time at stop i, $A_i$, is equal to the arrival time at stop i-1, $A_{i-1}$, plus the travel time, $T_i$ , between stop i and stop i-1. That is,

$$A_i = A_{i-1} + T_i \, , \; i = 2,3, \dots , k, \tag{3.1}$$

assuming there is no stopping time (negligible) between arrival time and departure time at stop i. Note that $A_i$ in terms of $T_i$ $s$ is given by:

$$A_i = \sum_{j=2}^{i} T_j \, , i = 2,3, \dots , k. \tag{3.2}$$

Usually the first stop is the origin of the trip, without loss of generality we assume that at time zero the bus is ready for its trip at the origin of trip (stop 1), and hence we assume: $S_1 = A_1 = 0$.

To simplify the model, we make the following initial assumptions:

**Assumption 1.** The arrival times are random variables and the scheduled times are constants. (3.3)

**Assumption 2.** The travel times between stops are independent random variables (That is, $T_1, T_2, \dots T_k$ are independent but not necessary identical random variables). (3.4)

Assumption (3.3) can be justified that drivers' habits are not expected to depend on scheduled times but rather on distances between stops, speed limits, and bus conditions. The travel time between two stops ideally should be relatively constant for a given distance at a given driving speed. However, because of many other factors such as traffic congestions, the travel time varies randomly. Therefore, $T_i$ is considered to be a random variable in the model.

To achieve the maximization goal in (1.1), first we need to know the probability distribution of arrival time $A_i$. Then based on the distribution of $A_i$, or its sample estimate, we will use some analytical approaches to find the new optimal scheduled time ,$S'_i$, which maximizes the on-time performance at the stop i.

Let $A_{i1}, A_{i2}, \dots, A_{in_i}$ be $n_i$ i.i.d observed arrival times for the stop i, with a common cdf $F_i$ and a common pdf $f_i$. Traditionally, transit agencies update their scheduled timetables for stop i by using the average of these observed arrival times. That is, they use the scheduled time $S'_i$ for the stop i given by:

$$S'_i = \frac{1}{n_i}\sum_{j=1}^{n_i} A_{ij}\,, i = 2,3, \dots, k, \tag{3.5}$$

regardless of the distribution of arrival time, $F_i$, and parameters of the selected on-time interval. Thus, a timetable set by such calculation would not necessarily be optimal since it does not consider the distribution of arrival time and the relevant transit scheduled on-time intervals.

In the following sections, we would discuss our procedures for three different cases, depending on the assumptions on the cdf $F_i$ and the pdf $f_i$.

**Case 1: Distribution of $A_i$ and its parameters are completely known**

Let $A_i \sim F_i(x; \boldsymbol{\theta})$. In this section, we assume that $F_i$ , the cumulative distribution function of arrival time, is completely known where $\boldsymbol{\theta}$ denotes a known vector of parameters. In this paper we assume arrival time distributions are all unimodal. We make this assumption to make the calculation easier, but this is not unreasonable assumption since the plotted histograms for most transit data appear to be almost unimodal.

Suppose, for stop i, the scheduled arrival time is $x_i$ and the corresponding scheduled on-time interval is $(x_i - a, x_i - a + L)$ ,where $a$ and $L$ are constants given by the transit agency. Then, the probability of on-time performance for this scheduled arrival time interval is given by:

$$P_i = P\{x_i - a \le A_i \le x_i - a + L\} = F_i(x_i - a + L) - F_i(x_i - a).$$

Therefore the goal in (1.1) is equivalent to:

$$\max_{x_i \ge a} \sum_{i=1}^{k} [\, F_i(x_i - a + L) - F_i(x_i - a)]. \qquad (3.6)$$

In order to solve the above problem, first we have the following lemma.

**Lemma 3.1** Suppose $F_i$ is a differentiable cumulative distribution function with corresponding unimodal density $f_i$ (such that $f_i$ is strictly increasing when $x_i < M_i$ and strictly decreasing when $x_i > M_i$ where $M_i$ is the unique mode of pdf $f_i$). Then the

optimal scheduled time $S'_i$ for stop i can be obtained by solving the following equation with respect to $x_i$:

$$f_i\,(x_i - a) = f_i\,(x_i - a + L).$$

*Proof:* Assume the arrival time at stop i is $A_i$ which has a cdf $F_i(x)$ with a pdf $f_i(x)$. Then the goal in (1.1) can be rewritten as follows:

$$P \equiv \max_{x_i \geq a} \sum_{i=2}^{k} [F_i(x_i - a + L) - F_i(x_i - a)]$$

$$= \sum_{i=2}^{k} \max_{x_i \geq a} [F_i(x_i - a + L) - F_i(x_i - a)]\ . \qquad (3.7)$$

Let

$$P_i = F_i(x_i - a + L) - F_i(x_i - a), \qquad (3.8)$$

then (3.7) can be written as:

$$P = \sum_{i=2}^{k} \max_{x_i \geq a} P_i\ . \qquad (3.9)$$

For *i=2,...,k ,* in order to maximize $P_i$ in (3.8), we take its derivative with respect to $x_i$ and set it equal to zero. That is:

$$\frac{\partial P_i}{\partial x_i} = 0\ . \qquad (3.10)$$

Then (3.10) reduces to the following equation:

$$f_i\,(x_i - a + L) - f_i\,(x_i - a) = 0. \qquad (3.11)$$

If $x_i = S'_i$ is the solution to the equation (3.11), then we must have:

$$f_i\,(S'_i - a) = f_i\,(S'_i - a + L). \tag{3.12}$$

Since pdf $f_i$ is assumed to be unimodal, equation (3.12) implies that for any $L > 0$,

$S'_i - a \leq M_i$ and $S'_i - a + L > M_i$ or

$$S'_i - a < M_i \text{ and } S'_i - a + L \geq M_i\ . \tag{3.13}$$

To complete the proof that $S'_i$ is the solution which maximizes $P_i$, we need to show that the second derivative of $P_i$ at point $S'_i$ is less than zero. Since $\frac{\partial^2 P_i}{\partial^2 x_i} = f_i'(x_i - a + L) - f_i'(x_i - a)$, we need to show $f_i'(S'_i - a + L) - f_i'(S'_i - a)$ is less than zero. Since $f_i$ is unimodal with mode $M_i$. We have $f_i'(x_i) < 0$ for $x_i > M_i$ and $f_i'(x_i) > 0$ for $x_i < M_i$. Since $S'_i - a + L > M_i$ and $S'_i - a < M_i$ it follows that $f'(S_i - a + L) < 0$ and $f'(S_i - a) > 0$ thus $f'(S'_i - a + L) - f'(S'_i - a) < 0$, which competes the proof.

**Optimal scheduled times for some selected unimodal families of distributions**

Assuming we know the distribution of arrival time ($A_i$), we can solve equation (3.12) to find the new scheduled time ($S'_i$) for the following selected families of distributions.

i. Normal Family of Distributions

Suppose, the arrival time $A_i$ at stop i, has a normal distribution. That is, $A_i \sim Normal\,(\mu_i, \sigma_i^2)$ ,where $\mu_i$ and $\sigma_i^2$ are known location and scale parameters, $-\infty < \mu_i < \infty$ and $\sigma_i^2 > 0$ .The pdf for $A_i$ is given by:

$$f_i(x;\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi\,\sigma_i{}^2}}\,e^{-\frac{(x-\mu)^2}{2\,\sigma_i{}^2}}.$$

It is easy to see that the solution to equation (3.12) which maximizes $P_i$ is $S'_i$ given by:

$$S'_i = \mu_i + a - \frac{L}{2}, \tag{3.14}$$

and hence the optimal on-time interval for stop i is given by:

$$[\mu_i - \frac{L}{2}, \mu_i + \frac{L}{2}],$$

where constant $L$ >0 is chosen by transit authority.

ii. Lognormal Family of Distributions

Suppose, the arrival time $A_i$ at stop i, has a lognormal distribution. That is, $A_i \sim Lognormal(\mu_i, \sigma_i^2)$ ,where $-\infty < \mu_i < \infty$ and $\sigma_i^2 > 0$, are known location and scale parameters, respectively. The pdf for $A_i$ is given by:

$$f_i(x;\mu_i,\sigma_i^2) = \frac{1}{x\sqrt{2\pi\,\sigma_i{}^2}}\,e^{-\frac{(\ln x-\mu_i)^2}{2\sigma_i{}^2}}.$$

The equation (3.12) for lognormal case is given by:

$$\frac{1}{(x-a)\sqrt{2\pi\sigma_i{}^2}}\,e^{-\frac{(\ln(x-a)-\mu_i)^2}{2\sigma_i{}^2}} = \frac{1}{(x-a+L)\sqrt{2\pi\sigma_i{}^2}}\,e^{-\frac{(\ln(x-a+L)-\mu_i)^2}{2\sigma_i{}^2}}. \tag{3.15}$$

It is easy to check that the equation (3.15) reduces to solving the following equation with respect to $x$:

$$e^{-\frac{(\ln(x-a)-\mu_i)^2}{2\sigma_i{}^2}+\frac{(\ln(x-a+L)-\mu_i)^2}{2\sigma_i{}^2}} = \frac{x-a}{(x-a+L)}, \quad (3.16)$$

taking the log of left and right sides of (3.16) we get:

$$\frac{1}{2\sigma_i{}^2}(Ln[(x-a+L)(u)] * \left[Ln\frac{x-a+L}{x-a}\right] + (2\mu_i)Ln\left(\frac{x-a}{x-a+L}\right)) - Ln\left(\frac{x-a}{x-a+L}\right) = 0\ . \quad (3.17)$$

In (3.17), if we let $t = Ln\frac{x-a+L}{x-a}$, then $x$ in terms of $t$ is given by:

$$x = \frac{L}{e^t-1} + a, \quad (3.18)$$

Now, equation (3.17) in terms of $t$ can be written as:

$$\frac{e^t-1}{\sqrt{e^t}} = \frac{L}{e^{\mu_i-\sigma_i{}^2}}\ . \quad (3.19)$$

If we set $\sqrt{e^t} \equiv z$, then (3.19) reduces to the following quadratic equation in terms of $z$:

$$z^2 - cz - 1 = 0. \quad (3.20)$$

The solutions to equation (3.20) are:

$z = \frac{c\pm\sqrt{c^2+4}}{2}$ where $c = \frac{L}{e^{\mu_i-\sigma_i{}^2}}$.

Note that only the solution $z_1 = \frac{c+\sqrt{c^2+4}}{2}$ is admissible. Thus, if we equate $z_1 = \sqrt{e^t}$ and solve it for $t$, we get:

$$t = Ln(\frac{c+\sqrt{c^2+4}}{2})^2 \text{ where } c = \frac{L}{e^{\mu_i-\sigma_i{}^2}}\ . \quad (3.21)$$

Now, substituting for $t$ from (3.21) in equation (3.18), we obtain the solution to the equation (3.15) which is:

$$x = \frac{L}{e^{t}-1} + a = \frac{L}{e^{Ln(\frac{c+\sqrt{c^2+4}}{2})^2}-1} + a = \frac{L}{\left(\frac{c+\sqrt{c^2+4}}{2}\right)^2-1} + a, \text{ where } c = \frac{L}{e^{\mu_i-\sigma_i{}^2}}.$$

Hence the optimal on-time scheduled time $S'_i$ for the lognormal case is given by:

$$S'_i = \frac{L}{\left(\frac{c+\sqrt{c^2+4}}{2}\right)^2-1} + a \quad \text{where } c = \frac{L}{e^{\mu_i-\sigma_i{}^2}}, \tag{3.22a}$$

and thus the optimal scheduled on-time interval for stop i for the lognormal case is given by:

$$\left[\frac{L}{\left(\frac{c+\sqrt{c^2+4}}{2}\right)^2-1} \quad , \frac{L}{\left(\frac{c+\sqrt{c^2+4}}{2}\right)^2-1} + L\right], \tag{3.22b}$$

where $c = \frac{L}{e^{\mu_i-\sigma_i{}^2}}$ and the constant $L > 0$ is given by the transit authority.

iii. Gamma Family of Distributions

Suppose, the arrival time $A_i$ at stop i, has a $Gamma(k_i, \theta_i)$ distribution, where $k_i > 0$ and $\theta_i > 0$ are known shape and scale parameters, respectively. That is, the pdf for $A_i$ is given by:

$$f_i(x; k_i, \theta_i) = x^{k_i-1} \frac{e^{-\frac{x}{\theta_i}}}{\theta_i{}^{k_i}\Gamma(k_i)} . \tag{3.23}$$

The equation (3.12) for the gamma case is given by:

$$(x-a)^{k_i-1}\frac{e^{-\frac{x-a}{\theta_i}}}{{\theta_i}^{k_i}\Gamma(k_i)} = (x-a+L)^{k_i-1}\frac{e^{-\frac{x-a+L}{\theta_i}}}{{\theta_i}^{k_i}\Gamma(k_i)}, \tag{3.24}$$

which reduces to the following equation:

$$e^{\frac{L}{\theta_i}} = \left(1+\frac{L}{x-a}\right)^{k_i-1}. \tag{3.25}$$

It is easy to check that the solution to equation (3.25) which maximizes $P_i$ is given by:

$$S'_i = \frac{L}{e^{L/(\theta_i k_i-\theta_i)}-1} + a. \tag{3.26}$$

Hence the optimal scheduled on-time interval based on the optimal scheduled time $S'_i$ for stop i is given by:

$$\left[\frac{L}{e^{L/(\theta_i k_i-\theta_i)}-1}\ ,\frac{L}{e^{L/(\theta_i k_i-\theta_i)}-1}+L\right],$$

where $L$ >0 is a known constant given by the transit authority for stop i.

Table 2. Summary of $\boldsymbol{S'_i}$Values

| $A_i$ | $S'_i$ | Optimal Interval |
|---|---|---|
| $A_i \sim Normal(\mu_i, {\sigma_i}^2)$ | $S'_i = \mu_i + a - \frac{L}{2}$ | $Lower\ Limit = \mu_i - \frac{L}{2}$<br>$Upper\ Limit = \mu_i + \frac{L}{2}$ |
| $A'_i \sim Lognormal(\mu_i, {\sigma_i}^2)$ | $S'_i = \frac{L}{\left(\frac{c+\sqrt{c^2+4}}{2}\right)^2-1} + a$<br>where $c = \frac{L}{e^{\mu_i-{\sigma_i}^2}}$ | $Lower\ Limit = \frac{L}{\left(\frac{c+\sqrt{c^2+4}}{2}\right)^2-1}$<br>$Upper\ Limit \frac{L}{\left(\frac{c+\sqrt{c^2+4}}{2}\right)^2-1} + L$<br>Where $c = \frac{L}{e^{\mu_i-{\sigma_i}^2}}$ |

| | | |
|---|---|---|
| $A_i \sim Gamma(k_i, \theta_i)$ | $S'_i = \frac{L}{e^{L/(\theta_i k_i - \theta_i)} - 1} + a$ | $Lower\ Limit = \frac{L}{e^{L/(\theta_i k_i - \theta_i)} - 1}$<br>$Upper\ Limit = \frac{L}{e^{L/(\theta_i k_i - \theta_i)} - 1} + L$ |

Table 2 shows a summary of optimal scheduled times calculated in this section. Generally speaking, an optimal scheduled time can be obtained by solving equation (3.12) if the distribution of arrival time is known.

**Case 2: Distribution of $\mathbf{A_i}$ is known but its parameters are unknown.**

Let $A_i \sim F_i(x;\ \boldsymbol{\theta})$, where the functional form of $F_i$ is assumed to be known but its parameter vector $\boldsymbol{\theta}$ may be partially or completely unknown and need to be estimated from the observed arrival times. In this section, for each family of distributions discussed earlier in case 1, we try to estimate the optimal scheduled time and the scheduled on-time interval using sample observed arrival times. Since the Maximum Likelihood Estimator (MLE) of a parameter $\theta$ is a consistent estimator which is a function of sufficient statistics for the unknown parameter, the procedure used here is to replace any unknown parameters of the model with its maximum likelihood estimators. If a MLE is not unbiased, whenever possible, we take an unbiased estimator based on the MLE. However, there may be cases that the MLEs do not have a nice closed existed form and it may be difficult to obtain them analytically. For those cases, one alternative approach is to use their moment estimators (the unbiased version if possible). In the following table we summarize the MLEs and moment estimators for the parameters of those models considered in case 1. The MLE of a parameter $\theta$ is denoted by $\hat{\theta}$ and the moment

estimator is denoted by $\tilde{\theta}$. Once unknown parameters are replaced by their sample estimates, the procedure in case 1 can be used to estimate the optimal scheduled on-time intervals.

Table 3. MLEs and Moment Estimators for the Selected Models in Case 1

| **Model (distribution)** | **Unknown parameters ($\theta$)** | **MLEs ($\hat{\theta}$)** | **Moment Estimator ($\tilde{\theta}$)** |
|---|---|---|---|
| $Normal(\mu_i, {\sigma_i}^2)$ | $\mu$<br>$\sigma^2$ | $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (*)<br>$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$<br>$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (*) | $\tilde{\mu} = m_1$<br>$\tilde{\sigma}^2 = m_2 - {m_1}^2$ |
| $Lognormal(\mu_i, {\sigma_i}^2)$ | $\mu$<br>$\sigma^2$ | $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$<br>$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2$ | $\tilde{\mu} = \ln \frac{{m_1}^2}{\sqrt{m_2}}$<br>$\hat{\sigma}^2_{mle} = \ln \frac{m_2}{{m_1}^2}$ |
| $Gamma(k_i, \theta_i)$ | $k$<br>$\theta$ | No closed form | $\tilde{k} = \frac{{m_1}^2}{m_2 - {m_1}^2}$<br>$\tilde{\theta} = \frac{m_2 - {m_1}^2}{m_1}$ |

Remark 1: The moment estimators are obtained by equating the distribution moments $\mu_k = E(X^k)$ with sample moments, $m_k = \frac{1}{n}\sum_{i=1}^{n} {X_i}^k$ and used to solve for the unknown parameters. For details, see for example VK. Rohatgi and AK, saleh (2001)

Remark 2: If MLE does not have a closed form, we use the moment estimator instead.

(*): unbiased estimator

**Case 3: Distribution of $A_i$ is completely unspecified (completely nonparametric case)**

Let $A_i \sim F_i$ where $F_i$ is completely unspecified. In this section, we discuss three approaches to address this case. The first approach is based on replacing the unknown arrival time distribution, $F_i$, in equation (3.7) with its sample empirical cdf, $\hat{F}_i$, and then proceed with maximization. The second approach is replacing the unknown pdf $f_i$, with a suitable estimator $\hat{f}_i$ .This method involves density estimation and generally the procedures are more complicated. The third approach is constructing the on-time performance of a given length based on some appropriate sample statistics such as sample mean, sample median, sample mode, sample middle point, sample geometric mean, sample harmonic mean and sample skewness.

**i. Empirical distribution approach:**

Let $X_1, X_2, \ldots, X_{n_i}$ be iid observations from arrival time $A_i \sim F_i$, where $F_i$ is unknown. The empirical cdf, $\hat{F}_i$, is defined as:

$$\hat{F}_i(x) = \frac{Number\ of\ X_i \leq x}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} I_{X_j}(x), \tag{3.27}$$

where $I_{X_j}(x) = \begin{cases} 1 & if\ X_j \leq x \\ 0 & if\ X_j > x \end{cases}$.

It is well known that empirical cdf $\hat{F}_i(x)$ is an unbiased and consistent point estimator of $F_i(x)$, at each $x \in R$.

Note that for any fixed $x$, the indicator $I_j \equiv I_{X_j}(x)$ is a Bernoulli random variable with parameter $p = F_i(x)$, and $n\hat{F}_i(x)$ has a binomial ($n = n_i$ , $p = F_i(x)$) distribution. Hence,

$E(\hat{F}_i(x)) = F_i(x)$ and $V\left(\hat{F}_i(x)\right) = \frac{F_i(x)\left(1-F_i(x)\right)}{n}, x \in R$. That is, $\hat{F}_i(x)$ is an unbiased estimator for $F_i(x)$ and by the strong law of large numbers, the estimator $\hat{F}_i(x)$ converges to $F_i(x)$ as $n \to \infty$, almost surely at each $x \in R$. Thus the unbiased estimator $\hat{F}_i(x)$ is strongly consistent. If in equation (3.8), we replace $F_i$ with the empirical distribution function, $\hat{F}_i$ ,we get:

$$P_i = \hat{F}_i(x_i + L) - \hat{F}_i(x_i) . \quad (3.28)$$

Hence we propose the following numerical procedure to find the value of $x_i$ which maximize $P_i$.

1. Sort sample data, in ascending order, mark them as $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

$$\hat{F}_i(x) = \frac{Number\ of\ X_i \le x}{n} = \frac{j}{n}\ if\ X_{(j)} \le x < X_{(j+1)} .$$

2. In (3.28), initialize variable $x_i = c = X_{(1)}$ and define a gap variable $\Delta = 1$.
3. Calculate $P_i(1) \equiv \hat{F}_i(c-a) - \hat{F}_i(c-a+L)$.
4. Shift $c$ by $\Delta$ value, that is, $new\ c = old\ c + \Delta$ and then repeat step 3 to obtain $P_i(2)$.
5. Exit loop when $new\ c = X_{(n)}$ . Then sort $P_i(j)$ values and obtain the maximum value with the corresponding $c = c^*$ value. Finally we get the optimal scheduled time interval which is $[c^* - a, c^* - a + L]$, where $c^*$ is the obtained value of $x_i$ which maximized $P_i$.

Note that the $c^*$ obtained in step 5 may not be unique. In that case, we suggest to take the smallest $c^*$ if the histogram of the data is almost right skewed, take the largest $c^*$ if the

histogram of the data is almost left skewed and take the midpoint between the largest $c^*$ and the smallest $c^*$ if the histogram of the data is almost symmetric.

### ii. Density function estimation approach

This approach involves density estimation which requires more advanced theoretical statistical procedures beyond the scope of this thesis, but it could be a separate research project for the future, we put less effort on it and only would introduce it in this section. Presently, there are several nonparametric density estimator methods where most of them are based on Kernel-smoothing, using a normal kernel.

The density at *x* by the kernel smoothing method is given by:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \qquad (3.29)$$

where *K* is the kernel function, *h* is the window size and $X_i$ is the observed data. Once unknown density function is estimated, then one can follow the procedure described in case 1 to estimate the optimal scheduled on-time interval.

### iii. Sample statistics approach

The statistics considered here are sample mean, sample geometric mean, sample harmonic mean, sample median, sample mode, sample midpoint, and sample interquartile midpoint. If a statistics *T* is used then the scheduled time is given by:

$$S'_i = T + a - \frac{L}{2},$$

Where $a$ and $L$ are given by transit agencies. Hence the proposed scheduled on-time interval based statistics T is:

$$[S'_i - a, S'_i - a + L] = \left[T - \frac{L}{2}, T + \frac{L}{2}\right]. \quad (3.30)$$

**iv. Guess model**

Here we suggest another easy procedure to approximate scheduled time which we call it guess model. In lemma 3.1, we showed that if $S'_i$ is the optimal scheduled time for a given distribution, then $S'_i \in (M_i + a - L, M_i + a)$ , where $M_i$ denotes the mode of the distribution of arrival time $A_i$. Since in this section we assume distribution is completely unknown, we suggest the following on-time arrival time based on the sample mode:

$$S'_i = \widehat{M}_i + a - \frac{L}{2} + \hat{\eta}_i * \frac{L}{2} \ , \quad (3.31)$$

where $\widehat{M}_i$ and $\hat{\eta}_i$ denote the sample mode and the sample skewness, respectively. The sample skewness is defined by:

$$\hat{\eta}_i = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}(x_j - \bar{x})^3}{\left(\frac{1}{n_i}\sum_{j=1}^{n_i}(x_j - \bar{x})^2\right)^{3/2}} . \quad (3.32)$$

Since the data have multiple modes, we suggest replacing $\widehat{M}_i$ with the midpoint of the interval with the highest frequency in the histogram for the observed arrival time data.

**Model Comparison**

In earlier sections, we have discussed the procedures for the case that the distribution of arrival time is known, or can be estimated and the completely nonparametric case. In this

section we use computer simulation to compare the performance of the procedures introduced earlier.

We used SAS program for the Monte Carlo simulation to compare performance of different scheduled on-time intervals. The source codes for the SAS programs are given in the appendix. Figures 2, 3, 4, 5, 6 and 7 illustrate that the on-time performance changes when updating the scheduled time using different methods discussed in previous sections. In Figures 2, 3 and 4, the arrival time is randomly generated from a Gamma distribution with different values of the scale and shape parameters. In Figures 5, 6 and 7, the arrival time is randomly generated from a skewed normal distribution for different values of scale and shape (hence skewness) parameters. In figure 8, the arrival time is randomly generated from a lognormal distribution for different values of scale parameter. Since transit agencies currently consider an arrival to be on-time if it is at most five minutes earlier or at most two minutes later than the scheduled arrival time, for the purpose of simulation, in equation (3.7) we take $a = 5$ and $b = 2$ (and hence $L$=7). Since the skewnormal does not have a closed form, we did not include this method in skewnormal figures. The optimal on-time performance based on various sample statistics, guessing method, and statistical models discussed in previous sections are given in the table 4. The following symbols are used in table 4:

$\bar{X}_a = \frac{1}{n}\sum_{i=1}^{n} X_i$ represents the sample arithmetic mean of arrival times.

$\bar{X}_h = \frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}}$ represents the sample harmonic mean of arrival times.

$\bar{X}_g = \sqrt[n]{X_1 X_i \ldots X_n}$ represents the sample geometric mean of arrival times.

$Q_2$ represents the sample median (or 2nd quartile ) of arrival times.

$M$ represents the sample mode of arrival times.

$\hat{X}_{mid} = \frac{X_{(1)}+X_{(n)}}{2}$ represents the sample midpoint of arrival times.

$Q_m = \frac{Q_1+Q_3}{2}$ represents the midpoint between $Q_1$ and $Q_3$,where $Q_1$ and $Q_3$ are the 1st and 3rd sample quartiles, respectively.

For example, *otp_geomean* represents the on-time performance determined from geometric mean. That is, the new scheduled time based on geometric mean is $S'_i = \bar{X}_g - \frac{L}{2} + a$ .

Table 4. Calculation formula for each method

| *Methods* | $S'_i$ (Scheduled Time) | Scheduled time interval |
|---|---|---|
| otp_median | $S'_i = Q_2 - L/2 + a$ | $[Q_2 - \frac{L}{2}, Q_2 + \frac{L}{2}]$ |
| otp_mean | $S'_i = \bar{X}_a - L/2 + a$ | $[\bar{X}_a - \frac{L}{2}, \bar{X}_a + \frac{L}{2}]$ |
| otp_harmean | $S'_i = \bar{X}_h - L/2 + a$ | $[\bar{X}_h - \frac{L}{2}, \bar{X}_h + \frac{L}{2}]$ |
| otp_mid | $S'_i = Q_m - L/2 + a$ | $[Q_m - \frac{L}{2}, Q_m + \frac{L}{2}]$ |
| otp_mode | $S'_i = M - L/2 + a$ | $[M - \frac{L}{2}, M + \frac{L}{2}]$ |
| otp_transit * | $S'_i = \bar{X}_a$ | $[\bar{X}_a - a, \bar{X}_a - a + L]$ |
| otp_geomean | $S'_i = \bar{X}_g - L/2 + a$ | $[\bar{X}_g - \frac{L}{2}, \bar{X}_g + \frac{L}{2}]$ |

| | | |
|---|---|---|
| otp_guess | $S_i = M + a - \frac{L}{2} + \hat{\eta}_i * \frac{L}{2}$ | $[M - \frac{L}{2} + \hat{\eta}_i * \frac{L}{2}, M + \frac{L}{2} + \hat{\eta}_i * \frac{L}{2}]$ |
| otp_nqmid | $S'_i = \hat{X}_{mid} - L/2 + a$ | $[\hat{X}_{mid} - \frac{L}{2}, \hat{X}_{mid} + \frac{L}{2}]$ |
| otp_equation | Solve equation (3.12). See table 3 | Solve equation (3.12). See table 3 |

*Methods currently used by transit agency.

The findings from these simulations are as follows. If we don't know the distribution of arrival time, we have the following conclusions:

1. The on-time performance using these methods would be very close to each other when the density of arrival time is almost symmetric.
2. The scheduled time defined by $S'_i = Q_2 - \frac{L}{2} + a$ as the optimal scheduled time overall performs well when we have no sufficient information about the distribution of arrival times.
3. As expected, the Monte Carlo simulations illustrate that the scheduled on-time intervals obtained using analytical method introduced in this research is better than all other procedures when the distribution of arrival time is known or can be effectively estimated.

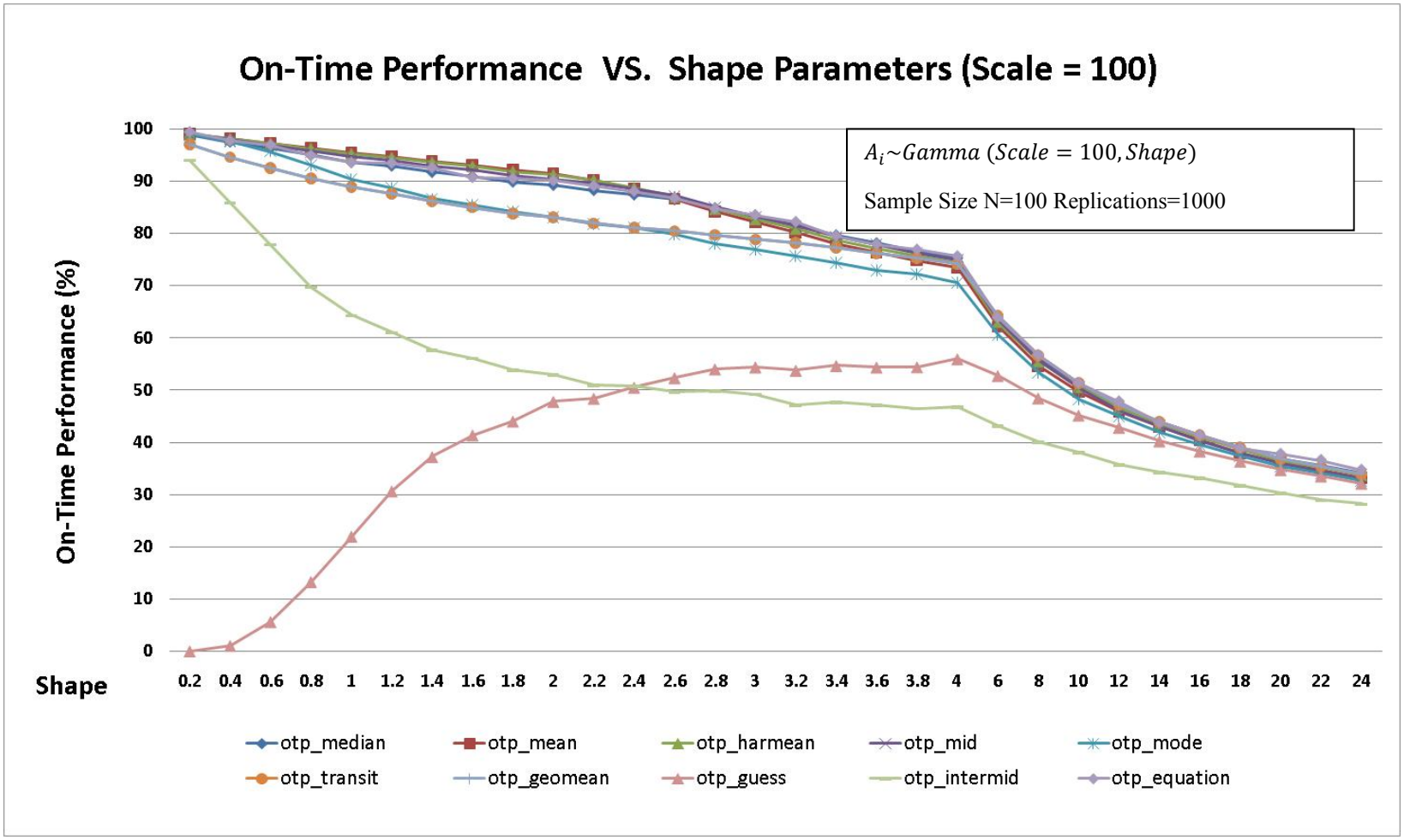


Figure 2 On-time performance measure vs. values of shape parameters (Gamma distribution): value of scale =100

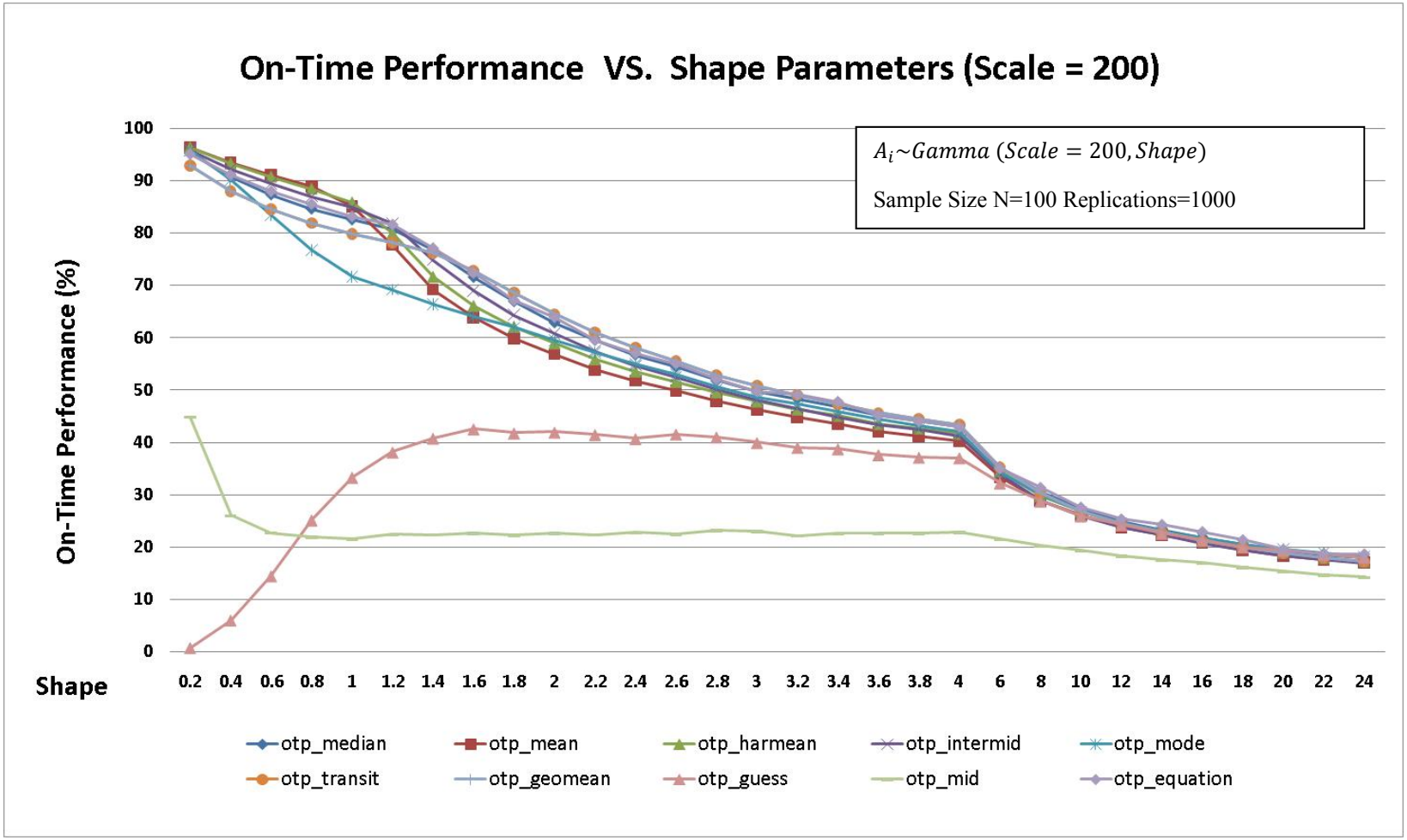


Figure 3. On-time performance measure vs. values of shape parameters (Gamma distribution): value of scale =200

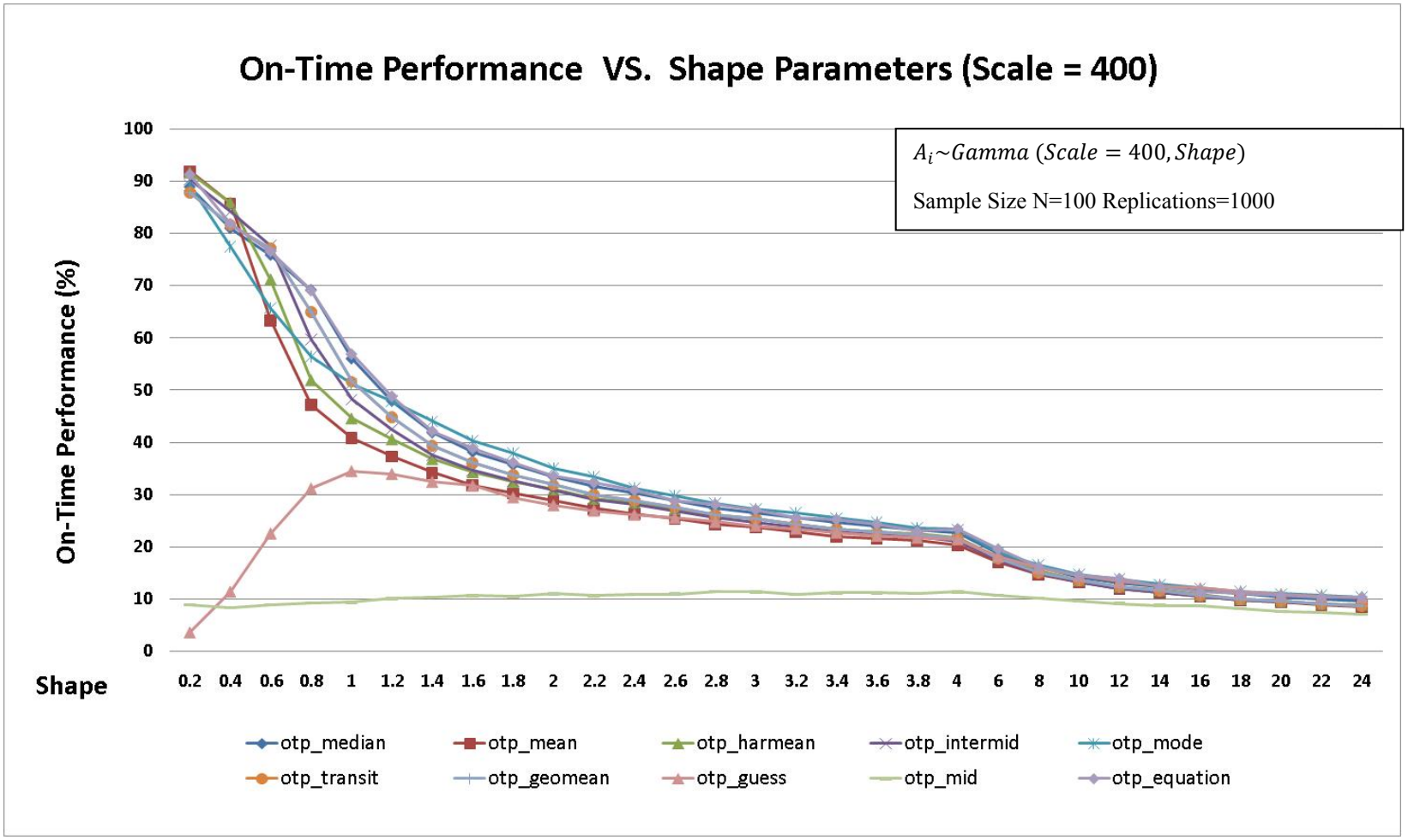


Figure 4. On-time performance measure vs. values of shape parameters (Gamma distribution): value of scale =400

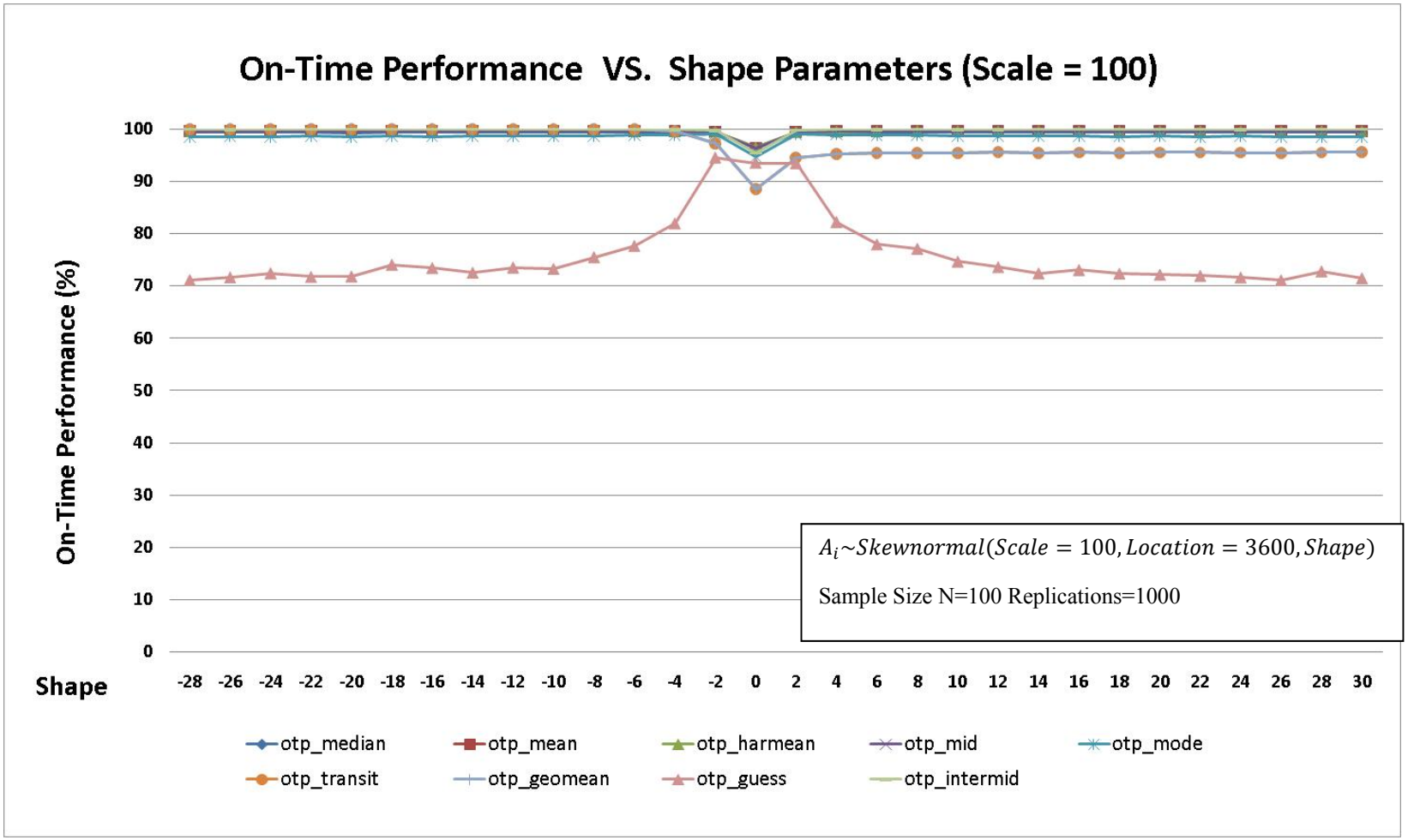


Figure 5.On-time performance measure vs. values of shape parameters (Skewnormal distribution): value of scale =100

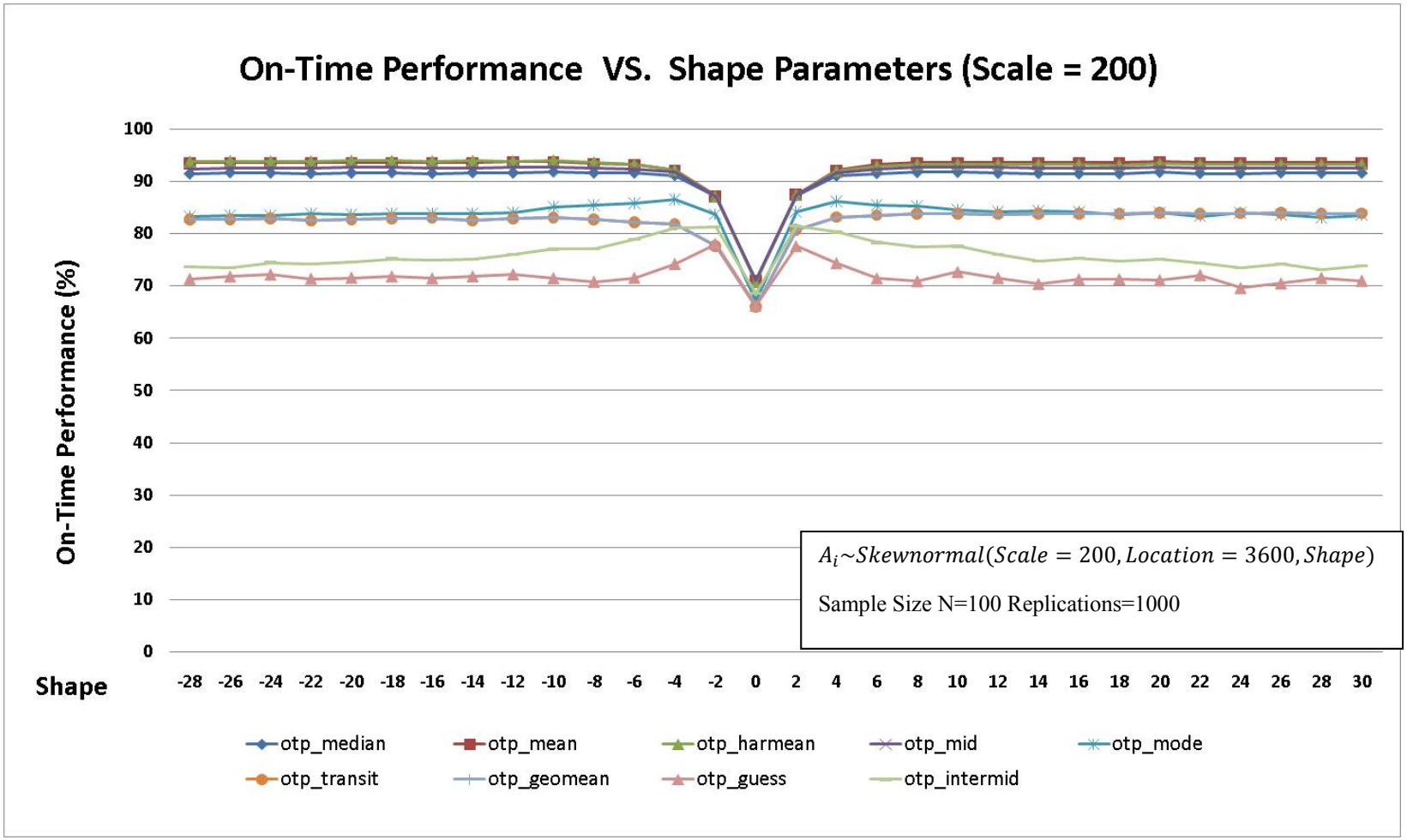


Figure 6. On-time performance measure vs. values of shape parameters (Skewnormal distribution): value of scale =200

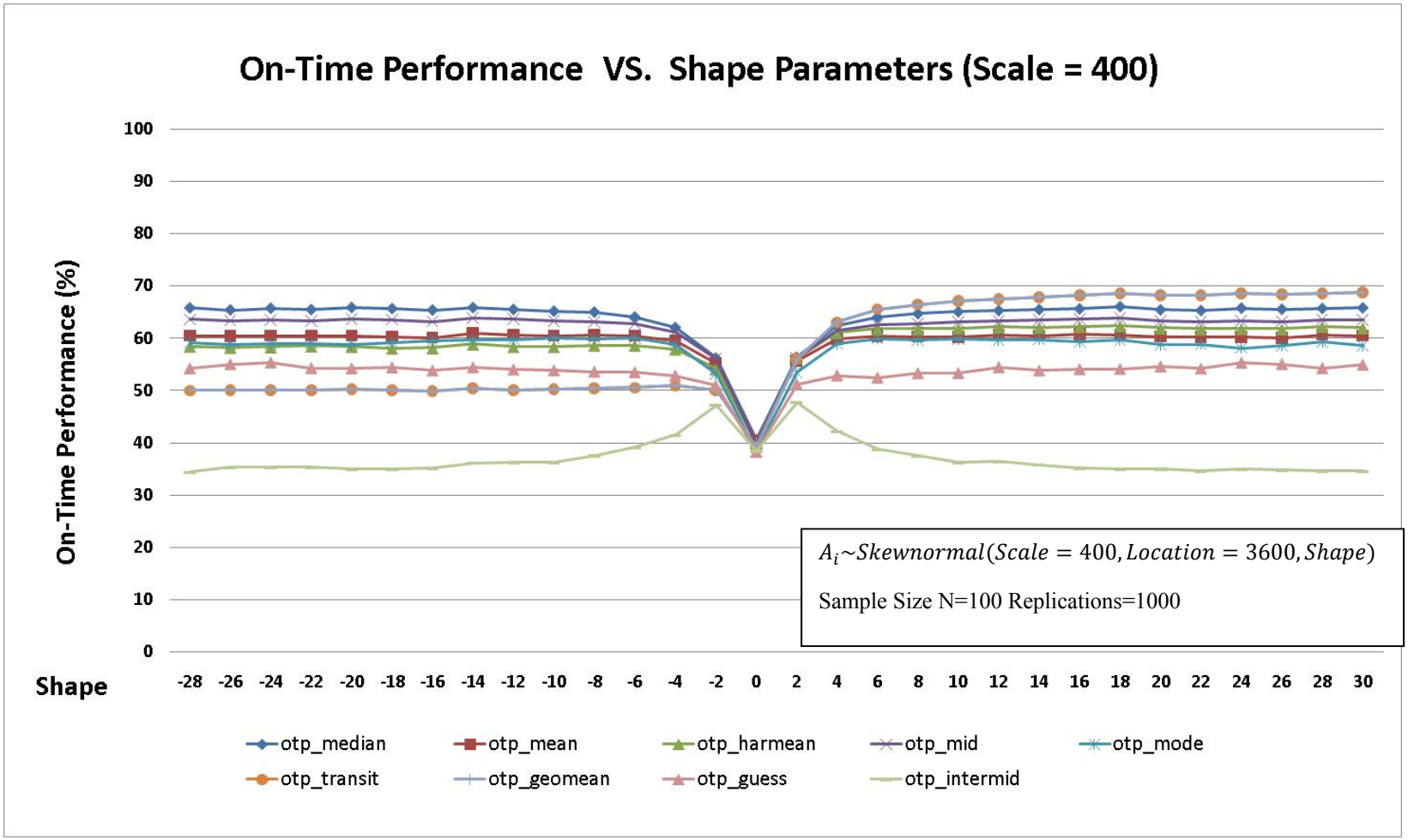


Figure 7. On-time performance measure vs. values of shape parameters (Skewnormal distribution): value of scale =400

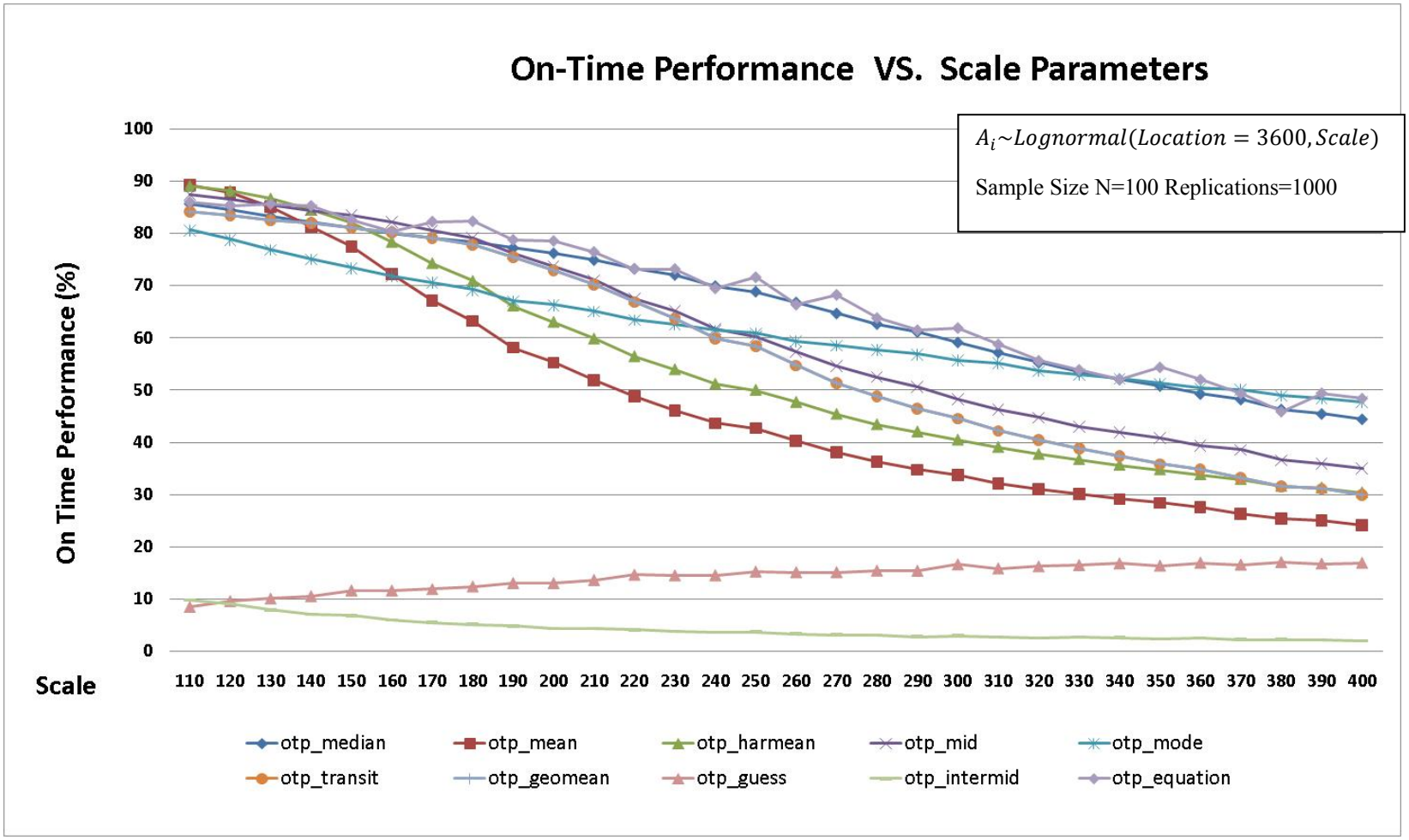


Figure 8. On-time performance measure vs. values of scale parameters (Lognormal distribution)

**Remark on arrival time $A_i$**

By (3.2), the arrival time ( $A_i$) at stop i is a summation of previous travel time. Let $T_{i-1}$ and $T_i$ be two independent continuous random variables with density functions $f(x)$ and $f(y)$. The sum $Z = T_{i-1} + T_i$ is a random variable with density function $h(z)$, where $h(z)$ is the convolution of $f(x)$ and $g(y)$. That is:

$$h(z) = \int_{-\infty}^{+\infty} f(z-y)g(y)dy = \int_{-\infty}^{+\infty} g(z-x)f(x)dx \ . \quad (3.33)$$

The above formula is to obtain density function of Z for only two independent random variables. In general, the density function of $A_i = T_1 + T_2 + \cdots + T_i$, where $T_1, T_2, \ldots, T_i$ are independent random variables with corresponding density function $f_{T_1}, f_{T_2}, \ldots, f_{T_i}$,respectively, is given by :

$$f_{A_i}(x) = \left(f_{T_1} * f_{T_2} * \ldots * f_{T_i}\right)(x) \quad , \quad (3.34)$$

where the right-hand side of (3.34) is an i-fold convolution of $f_{T_1}, f_{T_2}, \ldots, f_{T_i}$.

In some special cases, the density function of sum can easily be found based on the density function of its components. Table 5 gives some examples of these cases. As listed in the table 5, when $T_i$ has a distribution such as *Exponential* distribution, *Normal* distribution, *Cauchy* distribution or *Gamma* distribution (with the same values of scale parameter), the distribution of $A_i = T_1 + T_2 + \cdots + T_i$ , can easily be obtained and are well-known in the literature. For example, if all travel time distributions in each segment follow normal distributions, then the arrival time, $A_i$, will also have a normal distribution, though its parameters will be different.

This table only covers some simple examples, for more comprehensive list see any advance book in distribution theory such as Johnson and Kotz (1994).

Table 5. Special cases of $\mathbf{T_i}$ and $\mathbf{A_i}$

| $T_j\ (j \leq i)$ | $A_i(=\sum_{j=1}^{i} T_j\ )$ |
|---|---|
| $T_j \sim Exponential(\lambda)$ | $A_i \sim Erlang(i, \lambda)$ |
| $T_j \sim Normal(\mu_j, {\sigma_j}^2)$ | $A_i \sim Normal\left(\sum \mu_j, \sum {\sigma_j}^2\right)$ |
| $T_j \sim Cauchy(x_j, \gamma_j)$ | $A_i \sim Cauchy(\sum x_j, \sum \gamma_j)$ |
| $T_j \sim Gamma(k_j, \theta)$ | $A_i \sim Gamma(\sum k_j, \theta)$ |

## IDEAS FOR POSSIBLE FUTURE RESEARCH

The proposed methodologies can further be improved and extended: First, in the proposed models, it is assumed that drivers would not take any adjustment action when they are late or early at a particular stop. If drivers arrive at a stop i-1 earlier or later than the scheduled time, they may try to adjust their speed to meet the scheduled time for the next stop. Second, usually in transit agencies, the total amount of delays for a given road may have some constraint. That is, $\sum_{i=1}^{k}(S'_i - S_i) < C$ where $C$ is set by transit agencies. Lastly, though transit agencies like to use on-time performance measure to evaluate their services, mean squared errors of scheduled adherence can also be used as an alternative way to measure the performance of their services.

To address the first possible improvement, one can introduce a parameter β as an adjustment factor, where β is used to adjust travel time between the stops based on changes on the driver's driving habit. For example, when a driver is late in arriving at stop i-1, he may speed up in trip continuation to stop i. One way to adjust the reduction in the travel time may be modeled by $\beta(S'_i - A'_i)$, where in this case the equation (3.1) can be rewritten as follows:

$$A_i = A_{i-1} + \beta(S'_{i-1} - A'_{i-1}) + T_i \,. \qquad (4.1)$$

If $\beta = 0$, this reduces to the earlier model which does not consider driver's adjustment factor. Using (4.1), the problem will become more complicated and it needs further investigation.

To address the second and third possible improvements one may use procedure based on linear programming formulation which we would like to investigate in the future.

## SUMMARY CONCLUSION

This paper presents a perspective on how to improve transit on-time performance by using arrival time data. Several methodologies are proposed for updating the bus timetables by using arrival time data. The goal of this procedure is to find a new scheduled time ($S'_i$) which maximizes the on-time performance measure that currently transit agencies use to evaluate their services. Three different cases are discussed in this paper. In the first case, the distribution of arrival times is assumed to be completely known. In the second case, the functional form of the distribution of arrival times is assumed to be known but parameters of the model could be unknown. In the third case, the distribution of the arrival time is considered to be completely unspecified. Monte Carlo simulations are used to compare the effectiveness of these methods.

This research can provide transit agencies with procedures to help them to find optimal scheduled on-time intervals using historical ITS data. The methods discussed in this paper are expected to improve transit agencies' on-time performance percentage at a very low cost (Just by updating the timetable).

In this research, we have simplified the model by reducing variables and making some appropriate assumptions that may not be valid for some cases. For example, the travel time could be modeled by more variables like bus travel time plus passenger boarding and unloading time. In big cities, those data could play a very significant role. In this work we implicitly counted the passenger boarding and unloading time as a part of the travel time. The boarding and unloading time could be considered separately but this needs to be investigated in the future.

LIST OF REFERENCES

Ahmed M. El-Geneidy, Jessica Horning, and Kevin J. Krizek. 2011. Analyzing transit service reliability using detailed data from automatic vehicular locator systems. In *Journal of Advanced Transportation* Volume 45, Issue 1, pages 66–79.

Cevallos F., X. Wang, Z. Chen, and A. Gan. 2011. Using AVL data for Improving Transit On-Time Performance. *Journal of Public Transportation*,P21-40.

Cevallos, F., K. Kirwin, and R. Pearsall. 2008. Using CAD/AVL Data for Performance Management. Proceedings of the 10th International Conference on Applications of Advanced Technologies in Transportation, ASCE, Athens, Greece, May 27- 31.

Hammerle M., M. Haynes, and S. McNeil. 2005. Use of Automatic Vehicle Location and Passenger Count Data to Evaluate Bus Operations. In Transportation Research Record: *Journal of the Transportation Research Board*, No.1903, Transportation Research Board of the National Academies, Washington, D.C. pp. 27–34.

Hans Riedwyl. 1967. "Goodness of Fit", *Journal of the American Statistical Association,* Vol 62, No.318,P390-398.

Kalman, Rudolph Emil. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME, Journal of Basic Engineering* V82, P35-45.

Kimpel T. and J. Strathman. 2004. Improving Scheduling Through Performance Monitoring Using AVL and APC Data. Submitted to University of Wisconsin-Milwaukee as a Local Innovations in Transit project report under the Great Cities University Consortium.

Lee Y., K. Chon, D. Hill, and N. Desai. 2001. Effect of Automatic Vehicle Location on Schedule Adherence for Mass Transit Administration Bus System. In Transportation Research Record: *Journal of the Transportation Research Board*, No.1760, Transportation Research Board of the National Academies, Washington, D.C. P 81-90.

Nakanishi Y. 1997. Bus Performance Indicators-On-Time Performance and Service Regularity. In Transportation Research Record: *Journal of the Transportation Research Board*, No.1571, Transportation Research Board of the National Academies, Washington, D.C. pp. 1–13.

N. L. Johnson, S. Kotz, Continuous Univariate Distributions, vol.1, 2nd edition, ISBN: 978-0-471-58495-7.

Shalaby A. and A. Farhan. 2004. Prediction Model of Bus Arrival and Departure Times Using AVL and APC Data. *Journal of Public Transportation*, Vol. 7, No. 1.

Strathman, J. G., Kimpel, T. J., Dueker, K. J., Gerhart, R., & Callas, S. (2002). Evaluation of transit operations: Data applications of Tri-Met's automated Bus Dispatch System. *Transportation*, V29, P321-345.

Thevor Hastie, Robert Tibshirani, Jerome Friedman. 2001. The elements of statistical learning: Data mining, Inference, and prediction. Page 18-19.

Z. Wall, D.J. Dailey. 1999. An Algorithm for Predicting the Arrival Time of Mass Transit Vehicles Using AutomaticVehicle Location Data. In Transportation Research Record: *Journal of the Transportation Research Board*, No.0870, Transportation Research Board of the National Academies, Washington, D.C.

# APPENDICES

Appendix 1- Find the relationship between shape and on-time performance (SAS)

```
%let llm= -5;
%let ulm= 2;


data SimData;
array x x1-x100;
array a a1-a100;
array b b1-b100;
array c c1-c100;
      array d d1-d100;
      array e e1-e100;
      array f f1-f100;
      array g g1-g100;

      array xmean xmean1-xmean1000;
      array xmedian xmedian1-xmedian1000;
      array xmiddlepoint xmiddlepoint1- xmiddlepoint1000;
      array xharmean xharmean1- xharmean1000;
      array xskewness xskewness1-xskewness1000;
      array xmode xmode1-xmode1000;
      array xgeomean xgeomean1-xgeomean1000;


      array otp_median otp_median1-otp_median1000;
      array otp_mean otp_mean1-otp_mean1000;
      array otp_mid otp_mid1-otp_mid1000;
      array otp_mode otp_mode1-otp_mode1000;
      array otp_harmean otp_harmean1-otp_harmean1000;
      array otp_old otp_old1-otp_old1000;
      array otp_geomean otp_geomean1-otp_geomean1000;
      array otp_transit otp_transit1-otp_transit1000;
      array counts {100} counts1-counts100 ;

      shape =-10;
      scale =300;
      location =3600;


      do n=1 to 40;
      shape = shape +0.5;

        do r=1 to 1000;
        harmean = 0;
        geomean =1;
        do i=1 to 100;


                sigma = shape/sqrt(1+(shape)**2);
                  normal1= rand('NORMAL',0,1);
                  normal2= rand('NORMAL',0,1);
                 u1= sigma*normal1+sqrt(1-sigma*sigma)  * normal2;
                  if(normal1>=0) then t=u1; else t=-u1;
        x{i}= (t* scale +location);

```

```
55.       /*
56.        seed= ranpoi(123,87);
57.
58.        x{i} = rangam(seed,shape);  x{i} =
   scale*rangam(seed,shape);           /* gamma with shape a */
59.            /* gamma with shape a & scale b */
60.
61.        harmean + 1/x{i};
62.        geomean =(abs(x{i}) **0.01) * geomean;
63.        end;
64.
65.       do i = 1 to 100;
66.        counts{i} = 0;
67.       end;
68.
69.        do ii = 1 to 100;
70.          compare = x{ii};
71.          do jj = 1 to 100;
72.            if x{jj}-20 <= compare and x{jj}+20 >= compare  then
   counts{ii} + 1;
73.          end;
74.        end;
75.
76.        biggest = max(of counts1-counts100);
77.
78.        do k = 1 to 100;
79.          if counts{k} = biggest then leave;
80.        end;
81.
82.        xmode{r} = x{k};
83.        xmedian{r}= median( of x1-x100);
84.        xmean{r} = mean( of x1-x100);
85.        lower_quartile=PCTL(25,of x1-x100);
86.        uper_quartile=PCTL(75,of x1-x100);
87.        xmiddlepoint{r}= (lower_quartile +uper_quartile)/2;
88.        xharmean{r}= 100/harmean;
89.        xskewness{r}= skewness( of x1-x100);
90.        xgeomean{r}= geomean;
91.        otp_median{r} =0;
92.        otp_mean{r} =0;
93.        otp_mid{r} =0;
94.        otp_harmean{r} =0;
95.        otp_old{r} =0;
96.        otp_mode{r} =0;
97.        otp_transit{r}=0;
98.        otp_geomean{r}=0;
99.
100.       do i=1 to 100;
101.         mid = (&llm+&ulm)/2*60;
102.           a{i}=x{i}+ mid -xmedian{r};
103.           b{i}=x{i}+mid-xmean{r};
104.           c{i}=x{i}+mid-xmiddlepoint{r};
105.           d{i}=x{i}+mid-xharmean{r};
106.           f{i}=x{i}+mid-xmode{r};
107.           e{i}= x{i}-xmean{r};
```

```
108.            g{i}= x{i}-xgeomean{r};
109.
110.            lowlimit= &llm *60;
111.            uplimit= &ulm *60;
112.            if(lowlimit<a{i}<uplimit) then otp_median{r}+1;
113.            if(lowlimit<b{i}<uplimit) then otp_mean{r}+1;
114.            if(lowlimit<c{i}<uplimit) then otp_mid{r} +1 ;
115.            if(lowlimit<d{i}<uplimit) then otp_harmean{r}+1;
116.            if(lowlimit<f{i}<uplimit) then otp_mode{r}+1;
117.            if(lowlimit<x{i}-3600<uplimit) then otp_old{r} +1 ;
118.            if(lowlimit<e{i}<uplimit) then  otp_transit{r} +1 ;
119.            if(lowlimit<g{i}<uplimit) then  otp_geomean{r} +1 ;
120.
121.        end;
122.
123.        end;
124.
125.        median_out = mean (of xmean1-xmean1000);
126.        mean_out = mean (of xmedian1-xmedian1000);
127.        mid_out = mean (of xmiddlepoint1- xmiddlepoint1000);
128.        harmean_out = mean (of xharmean1- xharmean1000);
129.        skewness_out = mean (of xskewness1-xskewness1000);
130.        mode_out =mean( of xmode1-xmode1000);
131.        geomean_out =mean (of xgeomean1-xgeomean1000);
132.
133.        otp_median_out = mean(of otp_median1-otp_median1000);
134.        otp_mean_out = mean(of otp_mean1-otp_mean1000);
135.        otp_harmean_out = mean(of otp_harmean1-otp_harmean1000);
136.        otp_mid_out = mean(of otp_mid1-otp_mid1000);
137.        otp_old_out = mean(of otp_old1-otp_old1000);
138.        otp_mode_out = mean(of otp_mode1-otp_mode1000);
139.        otp_transit_out = mean(of otp_transit1-otp_transit1000);
140.        otp_geomean_out = mean(of otp_transit1-otp_transit1000);
141.        output;
142.
143.        keep shape scale location mode_out skewness_out
    median_out mean_out mid_out harmean_out otp_old_out
    otp_median_out  otp_mean_out otp_harmean_out otp_mid_out
    otp_mode_out otp_transit_out otp_geomean_out geomean_out;
144.
145.     end;
146.     run;
```

Appendix 2- Test distribution of sample data (SAS)

```
1. proc univariate data= Data  noprint;
2.
3. histogram travel_1 travel_2 travel_3 travel_4/
4. kernel ( k = normal
5. c = MISE
6. w = 3.0
7. color = green )
8. lognormal
9. weibull
10.      gamma
11.
12.      midpoints = 0 to 500 by 50; run;
```