Estimation of Infrastructure Condition from a Biased Sample

Gautam Gupta, Jasenka Rakas, Mark Hansen
Civil and Environmental Engineering Department
University of California at Berkeley
Berkeley, California, USA
ggupta@berkeley.edu

Abstract **- In this paper, we address the problem of making inferences about a population of infrastructure facilities from a subset that is a biased sample. We consider the case in which the sample is biased towards facilities in worse condition or requiring more expensive repair. Two methods are developed that incorporate a model of the process through which the sample is selected. One of the methods is based on well-known truncated distributions, whereas the other assumes that the bias operates continuously. The methods are applied to a class of facilities under the FAA's jurisdiction known as "un-staffed facilities." These consist of structures housing radars, navigation aids, radio beacons, and other ground-based equipment, and no previous system-wide evaluation has been attempted for these facilities. We present and discuss the estimates obtained from both the methods, and examine their goodness-of-fit with the sample. Given the premise that bias exists, the continuous bias model proved more suitable. However, the continuous bias model did not surpass the truncation models in terms of goodness-of-fit.**

I. Introduction

Infrastructure maintenance and repair decisions, along with supporting budgets, are based on data about facility condition, cost factors, and budgetary constraints. In some infrastructure systems, comprehensive condition surveys are done periodically. In other systems, condition surveys over the entire population of facilities are not done. Reasons for this can include excessive cost, accessibility constraints, and a reactive "fix it when it breaks" approach to infrastructure management. But even in such cases, there may be partial data (of a subset of the population) on condition, replacement or repair needs, and associated costs. With adequate knowledge about the procedure used to gather such data, reasonable extrapolations can be made about the entire population.

In this paper, we address such a problem of making inferences about a population of infrastructure facilities from data for a sample of them. The unique aspect of the problem is that the sample is biased in a particular fashion. Specifically, we consider the case in which the sample of facilities is biased toward facilities in worse condition or requiring more expensive repair (or even replacement). Such a bias may exist for a variety of reasons. For example, if the infrastructure manager is accustomed to budgets that are insufficient to bring all facilities to "like-new" condition, it will reasonably focus its condition-monitoring resources on the more "urgent" and "expensive" facilities.

We develop a method to address such a bias in the sample. The method incorporates a model of the process by which the sample is selected. We then demonstrate this method for a case study, which is a set of Air Traffic Control (ATC) facilities operated by the Federal Aviation Administration (FAA). The contributions of this paper are two-fold. First, we develop a method for utilizing biased sample data to derive information about the entire population. The kind of bias treated here is generic and may be encountered in a wide variety of situations. Second, we apply this method to a class of facilities under the FAA's jurisdiction, known as "un-staffed facilities" and consisting of structures housing radars, navigation aids, radio beacons, and other ground-based equipment, for which no previous system-wide evaluation has been attempted. FAA has identified such a comprehensive un-staffed facility evaluation as critical to the ongoing re-structuring of its infrastructure assets [1].

The rest of this paper is organized as follows. We first motivate and state the problem. We then identify a possible method for addressing the problem derived from previous literature, and discuss its shortcomings. We next propose more innovative and appropriate methods, followed by a description of the case study. We then apply the alternative methods to several different classes of FAA un-staffed facilities, compare their results, and demonstrate the advantages of our proposed method.

II. Problem Statement and Motivation

Consider a system of diverse infrastructure facilities spread over a wide region, managed by a government or private agency. The periodic allocation of maintenance and replacement funds to facilities is based, at least in part, on information about facility condition. The specific information provided is the cost of bringing *a subset* of facilities to "like new" state. Only a subset of facilities is included because it is not feasible (or even desirable) to bring all facilities to such a state, and the cost of developing the information of any given facility in non-negligible. Finally assume that agency policy is to prioritize maintenance and repair

activities on facilities in the worst condition, and thus with the highest restoration costs, there being a strong correlation between poor condition and high restoration cost.

Given these circumstances, it is reasonable to suppose that the subset of facilities for which cost information is provided will not be representative of the entire set, but rather be skewed towards those in the poorest condition. These cost estimates provided for a given class of facility can be treated as a sample of the population, but it is a biased sample in which facilities in a poor condition (and thus with higher restoration costs) are more likely to be included.

Suppose now that the agency wishes to use restoration cost data from these biased samples to estimate properties of the populations from which they are drawn. These properties might include the cost of restoring all facilities to "like-new" condition, or the probability distribution of these costs for individual facility types. There may be various motivations for this, including internal "budget drills" or the wish to publicly document the extent to which maintenance budget is "under-funded." Whatever the reason, and irrespective of its validity, the technical problem is to use available data for a purpose it was not originally intended for, and hence is imperfectly suited. Our aim is to investigate how to do this.

Let us now formalize the above problem. Let X be a random variable that is the cost of bringing a given type of facility to "like-new" condition. Suppose there are N facilities of this type, and that we have cost information for n of these facilities. Let p_i $(i = 1, 2 \ldots N)$ be the probability that facility *i* is included in the sample, and suppose that p_i depends on x_i the value of X for facility i: $p_i = p(x_i)$. Further, assume that the function $p(\,)$ is positive monotonic; more specific assumptions about the function will be discussed below. Given our sample data $x_1,\ x_2 \ldots x_n$, our objective is to estimate the probability density function (PDF) for X, $f_X(x)$.

III. Alternative Approaches

We now present three approaches to this estimation problem. The main difference between them is the specific assumptions we make about the function $p(\,)$.

A. Truncation Models

The most basic approach for modeling this kind of data is to assume that the sample is a truncated sample, with truncation point being a. Thus, we assume in this case that the cost data is drawn from the set of facilities whose restoration cost is above *a*. Given this assumption, a truncated distribution function can be constructed from the probability density function (PDF) and cumulative distribution function (CDF). Thus, if $f_X(x)$ is the PDF of the un-truncated distribution and $F_X(x)$ is the cumulative distribution function (CDF), the density of the truncated random variable can be written as [2]:

$$f_X(x|x > a) = \frac{f_X(x)}{\text{Prob}(x \geq a)} = \frac{f_X(x)}{1 - F_X(a)} \qquad (1)$$

Along with the parameters of $f_X(x)$, the truncation point a will also be an unknown parameter to be estimated. The estimation of a is subject to the constraint that $a \leq x_1$, where x_1 is the lowest value in the sample. The likelihood function for this approach can be written as

$$L = \prod_{i=1,2..n} \frac{f_X(x_i)}{1 - F_X(a)} \qquad (2)$$

For estimation of this model, it can be shown that the constraint $a \leq x_1$ is binding (proof is included in appendix 1). Thus, the truncation point would be the lowest value in the sample, and estimation involves determining the parameters of $f_X(x)$ only.

The above approach employs the cost data only. In most cases, the total number of facilities, including those without cost data, is also known. This information can be used in the estimation process. In order to do so, we must make an assumption about the process that determines whether or not a given facility appears in the sample. Two assumptions may be considered. First, we can assume that all of the facilities whose cost is above *a* are included. In this case, we know that if a facility is excluded, its cost must be below *a*. This yields the likelihood function:

$$L = \prod_{i \in P \setminus S} F_X(a) \prod_{i \in S} f_X(x_i) \qquad (3)$$

where S be the set of facilities in the sample, and P is the entire set, with $S \subseteq P$. Again, the estimation of a is subject to the constraint that $a \leq x_1$, where x_1 is the lowest value in the sample. Using arguments similar to appendix 1, it can be easily seen that this constraint is binding. We call this approach truncation with complete sampling (TWCS).

Alternatively, we can assume that the data represents only a fraction of the facilities whose restoration cost is greater than the truncation value. This sampling fraction thus becomes an additional parameter to be estimated, along with the truncation point and the parameters of $f_X(x)$. In effect, we assume that the facilities are initially screened to eliminate those whose restoration cost is less than *a*, and that a fraction *p* of the remaining facilities are included in the sample. The process could also begin by choosing a sample from all the facilities, and then eliminating from that sample those with a cost below *a*. The likelihood function is the same regardless of the sequence, but is most intuitively expressed if it is assumed that the initial sample includes all the facilities. It is given by:

$$L = \prod_{i \in P \setminus S} [pF_X(a) + (1 - p)] \prod_{i \in S} p f_X(x_i) \qquad (4)$$

The first product is for facilities not in the sample, either because they were initially sampled and had a cost less than *a*, or because they were not initially chosen for the sample. The second product corresponds to facilities in the sample. Note that when $p = 1$, this model reduces to the previous one. As before, estimation of a is subject to the constraint that $a \leq x_1$, and using arguments similar to appendix 1, it can be seen that this constraint is binding. We call this approach truncation with incomplete sampling (TWIS).

Both of these models have an important limitation. The bias toward facilities with higher restoration costs takes the form of rule that simply excludes facilities with costs below a certain value. The data is treated as an unbiased sample of those facilities that pass the cost test. A more plausible assumption is that the bias operates continuously: the higher the cost, the more likely the facility will be included. This is the basis for the next set of models.

B. Continuous Bias Models

These models assume that as facility repair cost increases, the probability of the facility being included in the data increases in a continuous fashion. These may be no absolute minimum cost, but facilities with low repair costs are very unlikely to be sampled. In the previous truncation based models, we modeled this selection as a constant probability, *p*, for all facilities with repair cost above the minimum value. Now selection is modeled as a monotonically increasing function of the repair cost.

Perhaps the simplest such model is that the selection probability is a linear function of the repair cost, with the probability being zero for the lowest repair cost in the sample, and 1 for highest. Formally, if x_1 and x_n be the lowest and highest values in the sample respectively, the probability of inclusion in the sample is:

$$p(x_i) = \begin{cases} 0, & x_i \leq x_1 \\ \dfrac{x_i - x_1}{x_n - x_1}, & x_1 \leq x_i \leq x_n \\ 1, & x_n \leq x_i \end{cases} \quad (5)$$

In this case, the likelihood function can be written as:

$$L = \prod_{i \in P \setminus S} \left(1 - \int_0^\infty p(y) f_X(y) dy\right) \prod_{i \in S} p(x_i) f_X(x_i) \quad (6)$$

The first product term covers the facilities that are not included in the sample, and the second term gives the contribution of the sample in the likelihood function. The integral $\int_0^\infty p(y) f_X(y) dy$ is the probability that a facility is selected. The likelihood maximization should yield a result such that $\int_0^\infty p(y) f_X(y) dy$ nearly equal to the ratio of sample to population.

The advantage of the above method of linearly increasing probability is the ease of estimation. Equation (5) gives a pre-determined probability for each data point in the sample, and there are no added parameters to be estimated besides the underlying distribution. This stems from the linear relationship and assigning probability values to the largest and smallest data-points. However, these assumptions themselves are a shortcoming of the above approach. This is because firstly, the relationship need not be linear, and secondly, even if the relationship is linear, determining the parameters of the linear relationship by assigning probabilities to smallest and largest values is restrictive. A more refined approach is to estimate the parameters of the sample selection model along with those of the cost distribution.

As stated earlier, the functional form used to model sample selection should be positive monotonic, since the selection is skewed towards higher values. Further, for the ease of estimation, the function should preferably be smooth for $x > 0$, where x is the repair cost. Equation (5) shows that estimation involves evaluating a definite integral that varies with the parameters being estimated, and a non-differentiable selection function would add to the complexity of an involved estimation. This criterion questions the logical extension to the functional form in (5), where x_1 and x_n would be parameters to be estimated rather than lowest and highest values in the sample. The function $p(x_i)$ in (5) is not differentiable at x_1 and x_n, and this would lead to a complicated estimation procedure.

Due to the above reasons, we adopt a binary logit model for sample selection. The probability of a facility with a particular repair cost, *x*, being present in the sample is:

$$p(x_i) = \frac{e^{\beta + \alpha f(x_i)}}{1 + e^{\beta + \alpha f(x_i)}} \quad (7)$$

where α and β are parameters to be estimated and $f(\cdot)$ is a positive monotonic function. For $\alpha > 0$, $p(x_i)$ approaches 1 as x_i becomes very large. Moreover, in this model sample selection can be represented as a utility maximization process, in which the expression $\beta + \alpha f(x_i)$ can be interpreted as the deterministic utility of including facility *i* in the sample as compared to the alternative of not including it, whose deterministic utility is assumed to be 0. The likelihood function remains the same as in (6). Estimation, however, is considerably more difficult as compared to the earlier methods because the definite integral includes unknown parameters of both the probability function and the repair cost distribution. We call this approach the continuous bias approach (CB).

IV. Case Study: Un-staffed Facilities in the National Airspace System

A. Introduction and Objective

The Federal Aviation Administration, as a part of its ongoing re-structuring of its infrastructure assets, plans to comprehensively evaluate the *un-staffed* facilities in the National Airspace System (NAS). Un-staffed facilities are structures that house communication, navigation, and surveillance equipment. The number of physical assets in the NAS is significant. The

NAS contains about 5,000 unstaffed facilities and 9,000 structural towers. Although the number and importance of such facilities is significant, it appears that the state of the art in assessing facility conditions and its performance is not very advanced [1,3,4].

NAS facilities are very diverse in terms of their type, construction and size, geographic location, environment and the traffic area they serve [9]. This diversity represents one of the major challenges in assessing facility condition and performance at an aggregate level. The entire NAS has been divided into nine different regions, based on different climactic and local conditions [9]. Because of the large number of diverse unstaffed facilities distributed across the entire NAS it would be extremely expensive to systematically assess and evaluate each facility and to establish a comprehensive data-base within a short time-frame. Instead, it is proposed to assess the condition of a representative sample of facilities from different regions.

The ultimate objective of our work is to develop a sampling methodology for this assessment. In order to do this, it is desirable to develop preliminary estimates of the mean and variance of repair cost for different classes of facilities so that accuracy of estimates yielded by different sample sizes can be predicted. These preliminary estimates are based on samples of facilities for which cost data are available. As noted above, these existing samples are skewed toward facilities with high repair costs. The preliminary estimates can then be used to design a sampling methodology based on stratified sampling [5].

B. Data and Methodology

As stated before, the entire NAS is divided into 9 regions, with each region having its own sub-divisions. There are 12 different types, and a table with abbreviations for different facilities is reproduced from [9] in appendix 2. For each facility, the respective sub-division assigns a subjective measure (Facility Condition Index, FCI) of the condition of the facility, with the assignment being updated at least once every year. This FCI is on a scale of 1 to 5, with 1 denoting a new facility, and 5 denoting need for replacement.

The data provided included the following:

- The population of a particular type of facility in any one of the 9 regions.
- For a certain subset of facilities, the *deferred maintenance cost* (DMC) was provided. DMC represents the cost of the repair that has been deferred for that financial year. This estimate is based on subjective judgment and periodic cursory inspection of facilities suspected to be in need of repair. Further, most of the data is for facilities which are judged to have an FCI value of 4 or 5.

In [9], the authors give the sizes of the total population and the sample for which deferred maintenance cost is available for each facility type. These data are not the result of a thorough inspection but based on quick appraisals, to be used for budget allocation. Nevertheless, some FAA managers state that these estimates are a good representation of the real cost. In the following sections, we assume these data are accurate and use them to estimate, for individual facility types, the underlying distributions of deferred maintenance cost. We do this for all types except TDWR's and ASDE's, whose sample sizes of 8 and 6 are too small to yield meaningful results.

C. Analysis

In this analysis, we assume that the repair costs are lognormal, or the log of the costs is normally distributed. This is clearly more plausible than the normal, because cost must be non-negative, but includes only two parameters, making estimation tractable. We also conducted experiments with other distributions including the folded normal and the exponential distribution (see table 4), but found the lognormal to have the best results. More complicated distributions, such as the Gamma might also be tried, but they would make estimation more difficult. Moreover, as discussed below, goodness-of-fit tests performed after estimation yielded acceptable results for most of the facility types. The functional forms for the lognormal distribution are given below in (8).

$$f_Y(y) = \frac{1}{By\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln y - A}{B}\right)^2}$$
$$F_Y(y) = \Phi\left(\frac{\ln y - A}{B}\right) \tag{8}$$

1) Truncation: We first present results from maximizing the likelihood function in (2), which considers only the sample cost values, not the fraction of the population sampled. As mentioned above, the maximum likelihood estimate for the truncation point is the lowest value in the sample, leaving just the two log-normal distribution parameters for numerical estimation. The results of this estimation are given in table 1. The parameters A and B are the same as defined in (8).

The Kolmogorov-Smirnov (KS) test was performed on the truncated probability function defined in (2). As shown in table 1, the estimated distribution passes the KS test for the .05 significance level in almost all cases.

TABLE 1. ESTIMATION RESULTS FROM SIMPLE TRUNCATION

Facility Type	A	B	KS Test Value
ALS	*8.768* (0.426)	*1.945* (0.327)	0.083#
ARSR	*10.483* (0.162)	*1.476* (0.105)	0.063#
ASR	*9.47* (0.209)	*1.718* (0.183)	0.057#
AWOS / ASOS	*9.27* (0.181)	*1.035* (0.123)	0.236
GS	*9.295* (0.115)	*1.471* (0.078)	0.111
LOC	*9.3* (0.134)	*1.717* (0.098)	0.118
MALS / SSALS	*9.422* (0.855)	*2.221* (0.87)	0.073#
RCAG	*9.6* (0.101)	*1.45* (0.055)	0.076#
RTR	*9.673* (0.121)	*1.529* (0.097)	0.079#
VOR	*9.588* (0.062)	*1.355* (0.038)	0.049#

(Values in brackets are standard errors for estimates)
Underlined and italicized estimates are significant at 0.05 level
Estimated distribution passes KS test at 0.05 level

As discussed before, the likelihood function used in the above estimation ignores potentially important information concerning the proportion of the population included in the sample. This information is particularly relevant if we assume that the sample includes all facilities whose DMC is greater or equal to the truncation value. In this case the likelihood function is given in (3). Estimation results, which appear in table 2, are much different than those in table 1, with lower A values and higher B values. Moreover, in virtually all cases, the fitted distribution fails the KS test. This suggests that the assumption that the data is a complete sample of values exceeding the truncation value is wrong.

TABLE 2. ESTIMATION RESULTS FROM MODIFIED TRUNCATION WITH COMPLETE SAMPLING (TWCS)

Facility Type	A	B	KS Test Value	Sample Size	Population
ALS	*4.874* (0.572)	*3.826* (0.702)	0.177#	42	126
ARSR	*7.769* (0.409)	*3.89* (0.603)	0.361	85	136
ASR	*3.824* (0.594)	*4.804* (0.839)	0.234#	77	249
AWOS / ASOS	-0.851 (2.73)	*5.02* (1.721)	0.370	36	600
GS	0.445 (0.839)	*6.136* (0.953)	0.316	187	914
LOC	-0.909 (0.96)	*6.652* (0.941)	0.273	194	1150
MALS / SSALS	-7.494 (5.383)	*7.382* (2.661)	0.156#	17	711
RCAG	*0.983* (0.599)	*7.447* (1.349)	0.412	217	633
RTR	-0.172 (0.997)	*6.451* (1.007)	0.293	179	1030
VOR	*5.9* (0.148)	*4.203* (0.314)	0.335	487	967

(Values in brackets are standard errors for estimates)
Underlined and italicized estimates are significant at 0.05 level
Estimated distribution passes KS test at 0.05 level

The third truncation model, described by (4), relaxes the assumption that the entire population above the cutoff point is included in the sample, but still exploits information about the proportion of the facilities included in the sample. The parameters for estimation are the parameters of the lognormal distribution (A and B) and the sampling fraction (p). Further, it should be noted that as a consequence of the estimation, the expression $1 - p\big(1 - F_X(a)\big)$ should be equal to the ratio of the sample to the population. The results of the estimation, along with the values of this expression and the ratio of sample to population are given below in table 3.

As expected, the sample-to-population ratio predicted matches to observed ratio. Moreover, the estimated values for (p) are very close to this ratio as well. This means that virtually all the exclusions of facilities excluded from the sample are the result of incomplete sampling rather than truncation. This also explains why the estimated values for A and B in Table 3 are so close to the estimates for the Simple Truncation model (Table 1). If exclusions from the sample are nearly always a consequence of random sampling rather than truncation, then accounting for the excluded facilities in the log likelihood function has very little effect.

2) Continuous Bias Models: Estimation results from the truncation models reveal that, if the samples in our data are indeed biased toward more facilities with higher DMC values, then this bias is not well represented using models based on truncation. It appears from those results that if a bias in fact exists, it results not from categorically excluding facilities whose DMC is below a certain value, but from a tendency to include more facilities with high DMCs. This suggests the use of a continuous bias model. Since we are using the lognormal distribution, we used natural logarithm for the function $f(\,)$ in (7).

TABLE 3. ESTIMATION RESULTS FROM MODIFIED TRUNCATION WITH INCOMPLETE SAMPLING (TWIS)

Facility Type	A	B	p	$1 - p + pF_X(a)$	Sample Fraction	Sample Size	KS Test Value
ALS	*8.764* (0.427)	*1.947* (0.327)	*0.375* (0.056)	0.333	0.333	42	0.083#
ARSR	*10.478* (0.162)	1.476 (0.105)	*0.625* (0.042)	0.624	0.625	85	0.062#
ASR	*9.468* (0.208)	*1.716* (0.183)	*0.316* (0.03)	0.309	0.309	77	0.058#
AWOS / ASOS	*9.266* (0.181)	*1.035* (0.123)	*0.061* (0.01)	0.060	0.060	36	0.234
GS	*9.292* (0.114)	*1.47* (0.078)	*0.205* (0.013)	0.205	0.205	187	0.112
LOC	*9.297* (0.134)	*1.718* (0.099)	*0.17* (0.011)	0.169	0.169	194	0.119
MALS / SSALS	*9.419* (0.857)	*2.222* (0.871)	*0.028* (0.009)	0.024	0.024	17	0.073#
RCAG	*9.597* (0.101)	*1.449* (0.055)	*0.342* (0.019)	0.342	0.343	217	0.077#
RTR	*9.671* (0.121)	*1.529* (0.097)	*0.174* (0.012)	0.174	0.174	179	0.080#
VOR	*9.584* (0.062)	*1.356* (0.038)	*0.504* (0.016)	0.503	0.504	487	0.050#

(Values in brackets are standard errors for estimates)
Underlined and italicized estimates are significant at 0.05 level
Estimated distribution passes KS test at 0.05 level

As stated before, this approach involves evaluating a large definite integral that varies with the parameters to be estimated. One way to do this is to use maximum simulated likelihood, where simulated probabilities are used instead of actual probabilities [6, 7]. However, our model definition involves only a one-dimensional definite integral, and hence we used the Newton-Cote's quadrature rules (trapezoidal rule) to approximate the integral [8]. Newton-Cote's formulas work by using interpolating functions to evaluate the integral, and in our case, we use the linear interpolation. The results from the estimation are given in table 4.

TABLE 4. ESTIMATION RESULTS FROM THE CONTINUOUS BIAS APPROACH

Facility Type	A	B	α	β	$\int_a^\infty p(y)\cdot f_X(y)dy$	Sample Fraction	Sample Size	KS Test Value
ALS	*7.025* (1.088)	*2.33* (0.668)	*1.206* (0.659)	*-9.891* (1.362)	0.334	0.333	42	0.091#
ARSR	*9.415* (0.269)	*1.973* (0.232)	*1.846* (0.691)	*-16.136* (2.612)	0.620	0.625	85	0.055#
ASR	*8.94* (1.521)	*1.676* (0.313)	0.339 (0.82)	-3.898 (3.332)	0.309	0.309	77	0.066#
AWOS / ASOS	*9.221* (0.757)	*0.998* (0.118)	0.086 (0.789)	-3.544 (3.641)	0.060	0.060	36	0.242§
GS	*7.86* (0.41)	*1.682* (0.166)	*0.84* (0.201)	*-8.443* (0.628)	0.203	0.205	187	0.100§
LOC	*7.474* (0.456)	*1.95* (0.184)	*0.782* (0.155)	*-8.073* (0.451)	0.167	0.169	194	0.100§
MALS / SSALS	7.779* (1.971)	*1.767* (0.483)	0.66* (0.428)	*-9.566* (1.35)	0.022	0.024	17	0.104#
RCAG	*8.291* (0.381)	*1.764* (0.194)	*0.941* (0.261)	*-8.783* (0.849)	0.340	0.343	217	0.067#
RTR	*8.196* (0.566)	*1.71* (0.197)	*0.774* (0.249)	*-8.39* (0.842)	0.173	0.174	179	0.068#
VOR	*8.247* (0.17)	*1.908* (0.132)	*1.81* (0.331)	*-14.909* (0.948)	0.502	0.504	487	0.077

(Values in brackets are standard errors for estimates)
Underlined and italicized estimates are significant at 0 .05 level
* Estimates significant at 0.1 level
# Estimated distribution passes KS test at 0.05 level
§ Estimated distribution passes KS test at 0.01 level, fails at 0.05 level

The parameter that captures bias in this model is α. Estimates for this parameter are positive in every case, implying that, as expected, higher cost facilities are more likely to be included in the sample. Moreover, based on a one-tailed test, α is significant at the .10 level in eight cases and at the .05 level in seven. Thus in most cases we can be fairly sure that a bias toward higher DMC facilities exists. Furthermore, based on the earlier estimation results, the truncation models are essentially equivalent to the continuous model with $\alpha = 0$. Thus rejection of this hypothesis implies that the continuous bias model is the most valid of those considered.

Table 4 also summarizes the KS test results comparing the predicted distribution for the observations in the sample to the observed data. In six of the 10 cases, the null hypothesis that the data came from the fitted distribution cannot be rejected at the .05 level. Of the remaining four cases, in three the null hypothesis cannot be rejected at the .01 level, while in one it is rejected at both .05 and .01 levels.

3) Model Comparison: We now compare results of the various models, in particular the CB and the TWIS, which we found to be the most satisfactory of the truncation models. With regard to goodness-of-fit, KS test results from the continuous bias and TWIS models closely resemble one another. The three facilities with the poorest distribution fits are the same in both models, and the magnitudes of the KS statistics in these cases—and most others--are quite similar. Thus, while estimation results, as well as the perception of FAA facility managers, comport better with the continuous bias model, the goodness-of-fit results do not support this conclusion.

To further explore the differences between the CB and TWIS models, we plotted the fitted PDFs and sample selection probabilities, and compared predicted CDFs and observed data. The plots for three facility types are given in figure 1. The PDF derived from the TWIS model is almost always shifted to the right of the one obtained from the CB model. The shift is greater when the probability of selection derived from the CB model changes more (in a proportional sense) over the central region of the PDF—this is the situation in which the bias will have the greatest effect. An instructive counterexample is the AWOS/ASOS case, where the PDFs are nearly identical and the selection probability is nearly constant in the central region.

While the PDFs and sample selection probabilities generally look very different for the two models, the resulting CDFs for sampled observations are strikingly similar. As the KS tests also revealed, for most facilities modeled CDFs also fit the empirical distributions quite well. Of the three cases with the poorest fits, in two the modeled distributions appear to have thicker right tails than the observed data, while in the other (which has just 17 observations), it appears that the central part of the distribution is more complicated than suggested by either model.

The ultimate aim of these models is to estimate the average DMC for each type of facility. Estimates from the four models appear in Table 5. Estimates from the CB model are less than those from any of the others. Comparing CB and TWIS estimates, the difference ranges from around 9% for ARSR's, to over 300% for localizers. It is also notable that only the CB model yields estimates of the population mean that are consistently below those of the sample mean. This again demonstrates that the CB model was the only one that actually demonstrated the bias believed to exist by FAA subject matter experts.

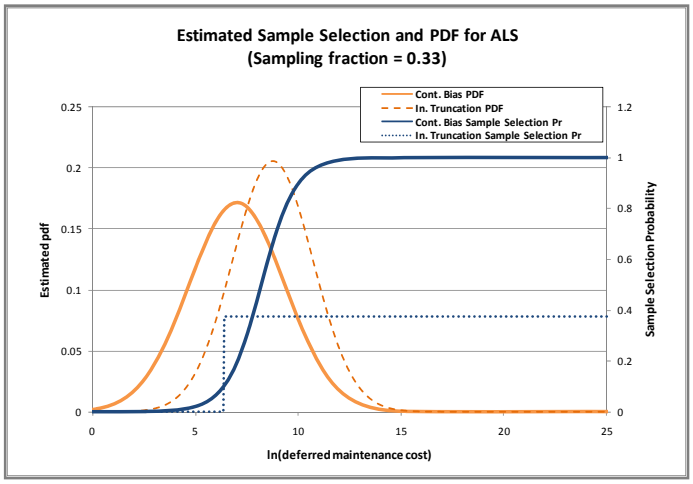
TABLE 5. COMPARING SAMPLE MEAN TO ESTIMATED POPULATION MEAN

Facility Type	Mean of Data or Sample Mean ($\times 10^3$)	Estimated Population Mean ($\times 10^3$)			
		Simple Truncation	Truncation with complete sampling	Truncation with incomplete sampling	Continuous Bias
ALS	42	43	198	43	17
ARSR	95	106	4,571	106	86
ASR	46	57	4,699	57	31
AWOS / ASOS	17	18	127	18	17
GS	30	32	233,800	32	11
LOC	44	48	1,636,000	48	12
MALS / SSALS	86	146	378,900	146	11
RCAG	40	42	2,948,000,000	42	19
RTR	40	51	916,400	51	16
VOR	42	37	2,501	37	24

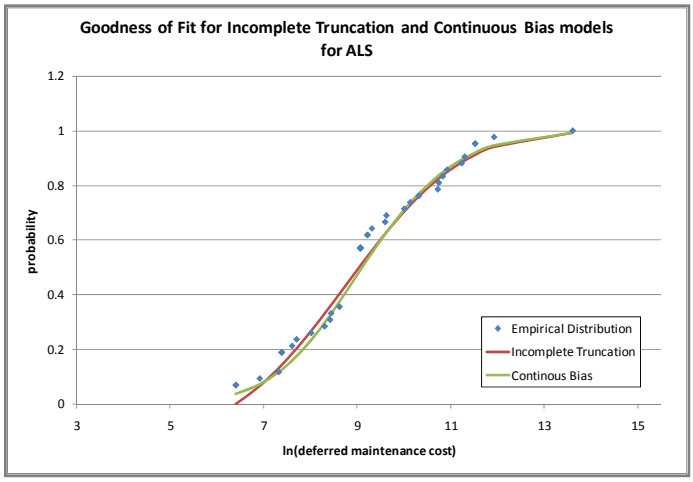
V. CONCLUSION

We have investigated the problem of estimating the parameters of a distribution from a sample of data in the face of known, or assumed, biases in the sampling process. In our application, the data are costs of restoring unstaffed facilities maintained by the FAA to support flight operations. Cost estimates are available for some facilities, but the samples are believed to be biased toward high cost instances. We have sought practical methods of inferring the cost distribution for the entire population that take this bias into account. There are many other settings, in infrastructure management and beyond, in which such a situation may exist, and to which our methods may also apply.

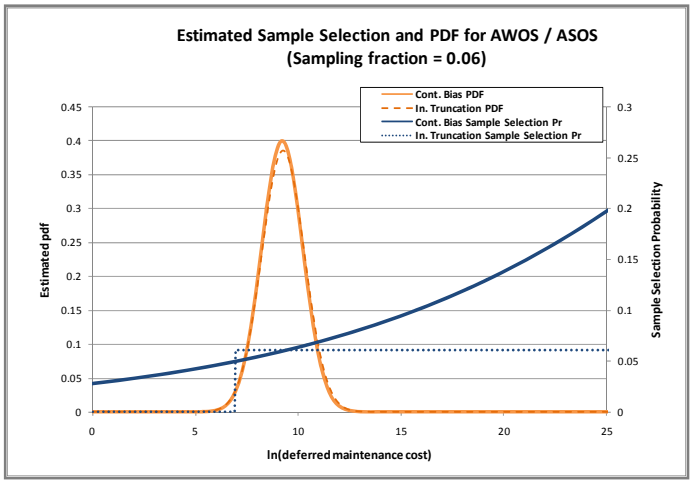
We have experimented with two ways of modeling the bias: (1) truncation, which assumes that facilities are systematically excluded if their restoration cost falls below a certain value, and (2) continuous bias, which allows the sampling probability to increase gradually as cost increases. Estimation results for the truncation models suggest that truncation is not a major source of sample bias, while results from the continuous model do suggest bias. Thus, if we accept the premise that such bias exists, then the continuous bias model proved more suitable. On the other hand, the latter model did not surpass the truncation models in terms of goodness-of-fit. In other words, the data do not, in and of themselves, favor the continuous bias model.



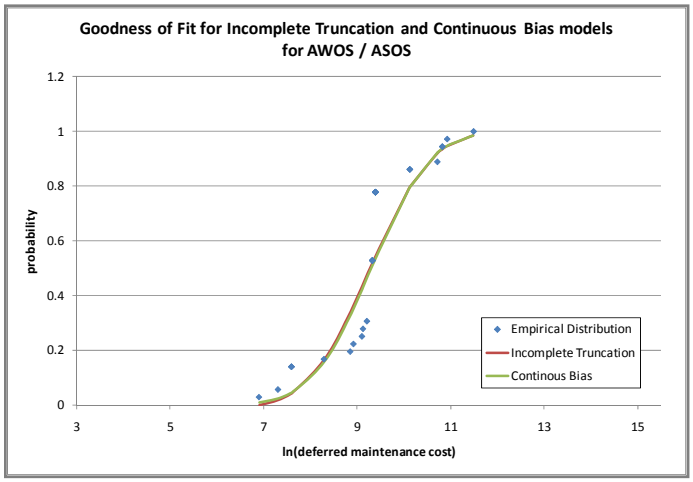
(a)



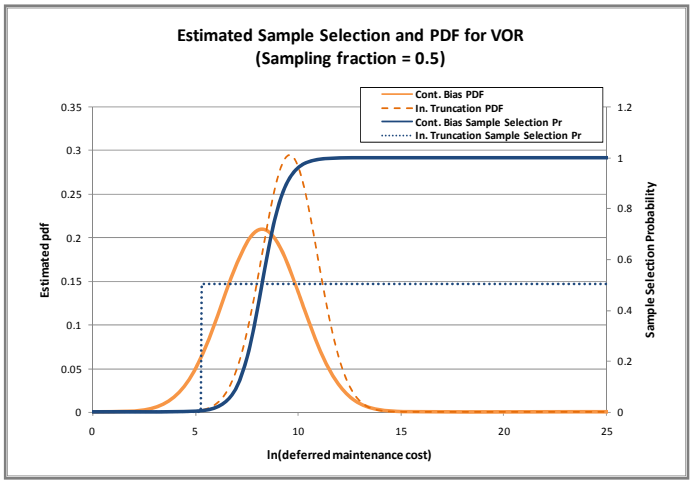
(b)



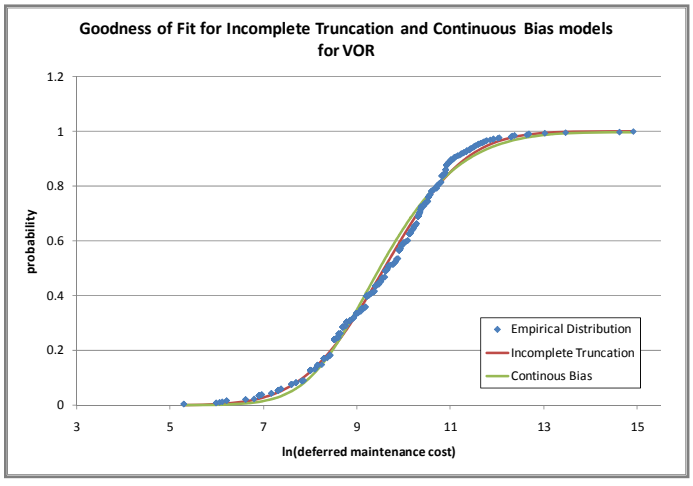
(c)



(d)



(e)



(f)

Figure 1. Plots of estimated PDF, CDF, sample selection and empirical distribution for TWIS and CB approaches

Throughout our analysis, we have maintained the assumption that the restoration cost has a log normal distribution. This was chosen as the most plausible two-parameter distribution for our setting, and results show that it works quite well for most facilities. It would certainly be desirable to extend these methods to more general distributions, but this may prove difficult, particularly for the continuous bias model. Estimation proved challenging even with a two-parameter distribution. As the number of parameters increases, bias effects and properties of the actual distribution will become harder to disentangle. Ultimately, the better approach is almost certainly to collect an unbiased sample. The methods presented here are best used as means of developing preliminary estimates from which efficient sampling strategies can be devised.

Appendix 1: Binding Nature of Constraint on Truncation Point for Simple Truncation

Let $f_X(x)$ be the probability distribution function (pdf) of the underlying population distribution that we are trying to estimate, with μ and σ being the parameters of the distribution. Let a be the truncation point for the estimation, with a being a parameter to be estimated too. Let $x_1, x_2 \ldots x_n$ be the sample values, sorted in increasing order. Thus, x_1 is the smallest value in the sample, and the estimation of the truncation point a is done subject to the constraint that $a \leq x_1$. Now, the truncated distribution that we are trying to estimate can be written as

$$f'_X(x) = \frac{f_X(x)}{1 - F_X(a)} \tag{9}$$

where

$$F_X(a) = \int_0^a f_X(x)dx \tag{10}$$

Thus, the likelihood function can be written as

$$\mathcal{L} = \prod_{i=1}^{n} \frac{f_X(x_i)}{1 - F_X(a)} \tag{11}$$

And the log-likelihood function becomes

$$\log \mathcal{L} = \sum_{i=1}^{n} \log f_X(x_i) - n \log\big(1 - F_X(a)\big) \tag{12}$$

If μ and σ are the parameters of $f_X(x)$, then both the numerator and denominator of the likelihood function are dependent on μ and σ. However, only the denominator $\big(1 - F_X(a)\big)^n$ depends on the truncation point a. Consider the cumulative distribution function $F_X(a)$. For any parameters μ and σ, the value of $F_X(a)$ increases monotonically with a, since it is the area under the pdf for $x \leq a$. Thus, the function $\frac{1}{\big(1-F_X(a)\big)^n}$ also increases monotonically with a for a given μ and σ. Hence, the constraint $a \leq x_1$ becomes binding while maximizing the log-likelihood function in terms of μ, σ and a.

Appendix 2: List of abbreviations related to different types of facilities

Abbreviation	Facility Type
TDWR	Terminal Doppler Weather Radar
ASR	Airport Surveillance Radar
ASDE	Airport Surface Detection Equipment
ARSR	Airport Route Surveillance Radar
RTR	Remote Transmitter Receiver
RCL	Radio Communication Link
RML	Remote Microwave Link
TML	Television Microwave Link
VOR	VHF Omni-directional Range
VORTAC	VOR collected with TACAN
TACAN	Tactical Aircraft Control and Navigation
LOC	Localizer
ALS	Approach Light System
MALS	Medium Intensity Approach Lighting System
SSALS	Simplified Short Approach Lighting System
AWOS	Automated Weather Observation System
ASOS	Automatic Surface Observing System
NEXRAD	Next Generation Weather Radar
LLWAS	Low Level Wind Shear Alert System
RCAG	Remote Communication Air / Ground
GS	Glide Slope

REFERENCES

[1] FAA, National Airspace System Capital Investment Plan 2003-2007, http://www.faa.gov, 2002

[2] R. J. Larsen and M.L. Marx, An introduction to mathematical statistics and its applications. 3rd ed. 2001, Upper Saddle River, NJ: Prentice Hall. x, 790.

[3] T. Brantley, "FAA's Aging ATC Facilities: Investigating the Need to Improve Facilities and Worker Conditions". Statement before the House Committee on Transportation and Infrastructure - Subcommittee on Aviation, July 24, 2007, 2007

[4] P. Forrey and P. Gilbert, "FAA's Aging ATC Facilities: Investigating the Need to Improve Facilities and Worker Conditions". Testimony before the House Committee on Transportation and Infrastructure - Subcommittee on Aviation, July 24, 2007. http://www.natca.org/legislationcenter

[5] P.S.R.S. Rao, Sampling methodologies with applications. Texts in statistical science. 2000, Boca Raton, Fla.: Chapman & Hall/CRC. 311.

[6] C. Arias and T. L. Cox, "Maximum simulated likelihood : a brief introduction for practitioners". Agricultural & applied economics staff paper series, 1999(no 421): p. 15p.

[7] K. Train, Discrete choice methods with simulation. 2003, New York: Cambridge University Press. vii, 334.

[8] C. W. Ueberhuber, Numerical computation : methods, software, and analysis. 1997, Berlin ; New York: Springer.

[9] G. Gupta, J. Rakas, and M. Hansen, “Cost-effective sampling to evaluate condition of un-staffed facilities in National Airspace System”. in Collection of Technical Papers - AIAA 5th ATIO and the AIAA 16th Lighter-than-Air Systems Technology Conference and Balloon Systems Conference. 2005. Arlington, VA, United States: American Inst. Aeronautics and Astronautics Inc., Reston, VA 20191-4344, United States.