Sample Design and Weighting for Estimating a Dose-Response Curve

Sharon Lohr[1], Pam Broene[1], Eric Jodts[1]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract
The Neighborhood Environment Survey was designed for the purpose of estimating the relationship between a household's aircraft noise exposure and the probability that the household reports being highly annoyed by aircraft noise. Unlike most surveys, in which at least some of the primary outcomes are means or totals, here the primary outcome is a logistic regression function that estimates the dose-response relationship. The first stage of the design uses balanced sampling to select airports. The household selection at the second stage of sampling takes a stratified sample of households from an address-based sampling frame. At most airports, there are relatively few households that are close to the airport and have high noise exposure levels, but there are many households that are farther away from the airport and have low noise exposure levels. We use optimal design results of Chaloner and Larntz (1989) to inform the second-stage sample allocation to noise exposure strata at each airport. The resulting sampling weights for households have high variation, and we discuss the choice of weights for estimating the dose-response curve and for nonresponse adjustments.

Key Words: balanced sampling, logistic regression, optimal design, stratification

1. The Neighborhood Environment Survey

The Neighborhood Environment Survey (NES) was conducted in 2015 and 2016 to study the relationship between exposure to aircraft noise and community annoyance at a representative sample of 20 U.S. airports in the contiguous 48 states. Exposure to high levels of environmental noise has been associated with hearing loss and with non-auditory health effects such as sleep disturbance, hypertension, and reduced cognitive performance (Basner et al. 2014), as well as interference with speech and other activities. Self-reported annoyance is used "as a summary measure of the general adverse reaction of people to living in noisy environments" (FICON, 1992, p. ES-2).

The primary goal of the NES was to calculate a national dose-response curve relating noise exposure from aircraft (dose) to the probability of being highly annoyed by aircraft noise (response). Sizov and Pickard (2011) described the background leading up to the survey. The list of sampled airports and results from the NES will be published in 2018. This paper describes the design of the survey as well as considerations for weighting and analyzing the data.

Following EPA (1974) and FICON (1992), noise exposure is measured using the day-night average A-weighted sound level (DNL), which captures both the magnitude and duration of noise events. DNL is calculated for a particular location by calculating the sound energy from each aircraft operation affecting that location over a year. Aircraft operations that occur between 10 pm and 7 am are assessed a 10 dB penalty when computing DNL. Letting $L(t)$ represent the A-weighted sound level associated with an event at time t, DNL is defined over the 86,400 seconds of the day as:

$$\mathrm{DNL} = 10\log_{10}\left[\frac{1}{86{,}400}\left(\int_{7\,am}^{10\,pm} 10^{L(t)/10}\,dt + \int_{10\,pm}^{7\,am} 10^{[L(t)+10]/10}\,dt\right)\right] \quad (1)$$

FAA (2017) defines “significant noise” as having an aircraft noise exposure of DNL 65 dB or higher.

Annoyance in the NES is measured using a 5-point scale recommended by Fields et al. (2001) with the question: “Thinking about the last 12 months or so, when you are here at home, how much does each of the following bother, disturb or annoy you?” This question is followed by a list of 13 potential irritants, of which the fifth is “Noise from aircraft.” The response options are “Not at all,” “Slightly,” “Moderately,” “Very,” or “Extremely.” Respondents checking “Very” or “Extremely” are considered to be highly annoyed (HA).

The binary response HA is used in the analysis so that the results of the NES can be compared with previous surveys. Schultz (1978), in his synthesis of early surveys, proposed using percent HA as a way to create a single response variable that could be applied to all surveys, even though the surveys used different questions and scales to measure annoyance. Subsequent aircraft noise annoyance modeling has continued this tradition to allow comparison to the Schultz (1978) curve and the similar curve given by FICON (1992), which are currently used as the basis for U.S. aircraft noise policy. Figure 1 displays these two curves.

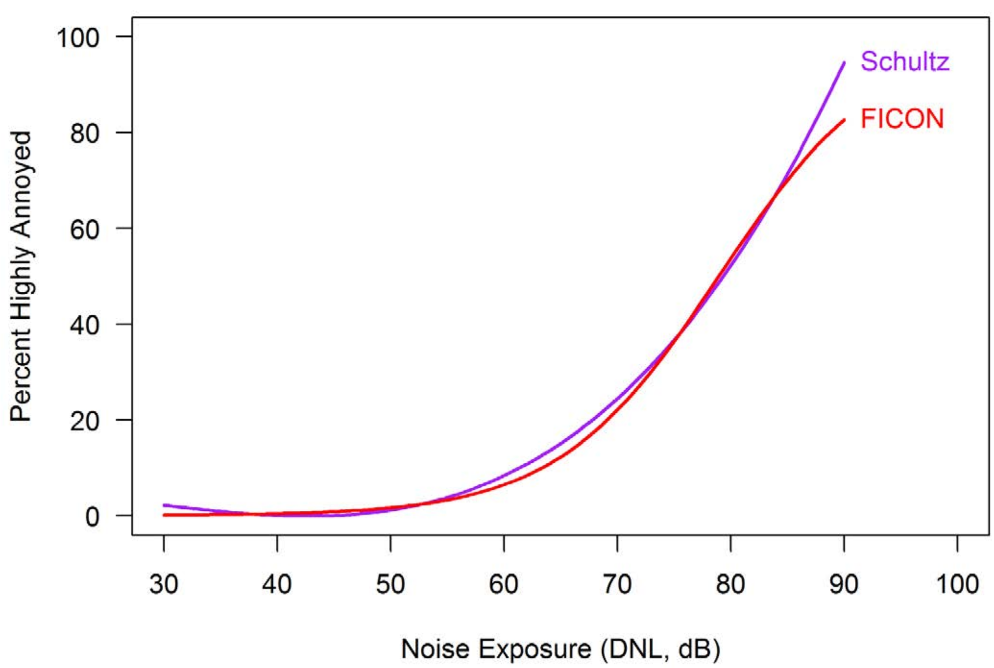


Figure 1: Dose-response curves from Schultz (1978) and FICON (1992).

The FICON (1992) curve was computed from a logistic regression model fit to 400 data points from different surveys, relating the percent HA to noise exposure from aircraft, road traffic, and railroads. The equation of the FICON curve is

$$\text{Percent HA} = 100\ \frac{\exp(-11.13+0.141\ DNL)}{1+\exp(-11.13+0.141\ DNL)} \qquad (2)$$

The FICON (1992) curve in Equation (2) relied on data from historical surveys collected in North America, Europe, and Australia. Bassarab et al. (2009) catalogued 628 U.S. and international social surveys that had been conducted between 1943 and 2008 to investigate residents' reactions to noise from aircraft, railways, road traffic, and other environmental noise sources. These surveys used different questionnaires, survey procedures, interviewing modes, languages, and methods for measuring noise exposure. Some surveys were conducted over time periods of a few days, while in others the data collection spanned several months. Most importantly, some of the surveys included transportation noise from road traffic and railroads. In addition, the majority of the more recent surveys catalogued in Bassarab et al. (2009) have been conducted outside of the U.S. and may not describe current dose-response relationships in the U.S.

The NES, by using uniform procedures and collecting data over the same time period (late October 2015 to early November 2016) from each of the 20 sampled airports, is designed to provide data for estimating an updated, U.S.-based, dose-response relationship between annoyance and noise exposure from aircraft. The goals of the NES are to develop logistic regression curves relating percent HA and DNL: (1) separately for each of the 20 airports in the study and (2) for the set of 20 airports, with the goal of estimating a curve that describes the average of individual dose-response curves for the population of airports.

Most surveys have a primary goal of estimating means (for example, unemployment rate) or totals (for example, total out-of-pocket expenditures for all Medicare recipients). The primary goal of the NES is to estimate logistic regression equations that give the dose-response relationships between noise exposure and annoyance. This results in a different design than would be used if the primary interest were in population means, because the variances of estimated logistic regression parameters depend on the sampled values of DNL in the survey as well as on the sample size and variability of the response variable.

Section 2 describes the balanced sampling design used to give a representative sample of 20 airports. Section 3 describes the sampling design used to select households within the sampling region for each sampled airport. Issues of using weights for estimation and nonresponse adjustment are discussed in Section 4, and Section 5 describes other surveys where the design innovations of this paper may be useful.

2. Selection of Airports

The Federal Aviation Administration (FAA) determined that 95 airports in the contiguous U.S. met the requirements for inclusion in the study, and determined the sample size of 20 airports. FAA also designated that three airports would be in the sample because of their large number of operations, with a fourth airport to be selected from the three major New York City-area airports. With these restrictions, it was desired to select 20 airports that were representative of the population of 95 airports on a wide range of characteristics. In a large sample, random selection methods provide assurance that the sample represents the population with respect to most characteristics. With a small sample, however, the larger sampling variability results in greater average absolute differences between sample means and corresponding population values of characteristics, and thus a smaller sample has a higher chance of being unrepresentative with respect to an important characteristic. A stratified sample can provide more control over the sample selection, but a sample of size 20 limits the number of strata that can be formed. It was desired to ensure that the NES sample of airports mirror population characteristics for six factors, and a stratified design could not guarantee that the sample would be representative on all six factors.

Therefore, to ensure that the sample was representative with respect to factors thought to be related to annoyance, and to obtain a sample that was geographically distributed across the U.S., a balanced sampling design (Valliant et al., 2000; Tillé, 2011) was used. This design allowed the sample to be balanced around the set of predetermined airports and yet still mirror the characteristics of the population of 95 airports. Hansen and Madow (1978), among the strongest advocates in the statistical literature for probability sampling designs, argued for model-based designs in small samples, which were considered to be samples of size less than 25 by Hansen et al. (1983).

The sample of airports was selected so that the proportion of airports in the sample for each category of each factor in Table 1 matched the corresponding population proportion. This was done by generating a large number of samples that were stratified on FAA region (thus ensuring balance on that factor in the initial set of potential samples that were generated), rejecting those that were not balanced on all of the other factors as well, and finally selecting one sample at random from the set of possible samples that met the balancing requirements.

Table 1: Balancing Factors Used in Design

Factor	*Categories*
FAA Region	Central, Eastern, Great Lakes, New England, Northwest Mountain, Southern, Southwest, Western Pacific
Average Daily Temperature	Below 55 degrees F, between 55 and 70 degrees F, above 70 degrees F
Percentage of Nighttime Operations	Less than 20 percent; more than 20 percent[1]
Average Number of Daily Flight Operations	Less than 300, more than 300
Fleet Mix Ratio of Commuter to Large Jet Aircraft	Less than 1, greater than 1
Population within 5 Mile Radius of Airport	Less than 230,000, greater than 230,000

Because the sample of airports was selected using balanced sampling, the national airport curve will be estimated using a model-based analysis. A random coefficients logistic regression model, where the curve from airport i is described by its own logistic regression

[1] After the sample of airports was drawn and after questionnaires were mailed, it was discovered that the value of percentage of nighttime operations was incorrect for some of the airports in the sampling frame. However, this error did not affect the validity of the sample, which was drawn so as to be balanced with respect to the values recorded in the frame. After the correct values for percentage of nighttime operations were determined, it was found that in fact the distribution of percentage nighttime operations in the sample was very similar to that in the population.

equation, accounts for the clustering by airport. The dose-response relationship for airport i is modeled by:

$$\text{Percent HA} = 100\ \frac{\exp(\beta_{0i}+\beta_{1i}\, DNL)}{1+\exp(\beta_{0i}+\beta_{1i}\, DNL)} \tag{3}$$

where β_{0i} and β_{1i} are assumed to follow a bivariate normal distribution with means β_0 and β_1 and unstructured positive definite covariate matrix. In this formulation, both the airport-specific slope and the airport-specific intercept are assumed to be random effects; Lohr (2014) found that a random intercept model, in which the intercept is allowed to vary for different airports but the slope is assumed to equal the same value β_1 for every airport, could substantially understate the variability in the data.

3. Address Selection in Sampled Airports

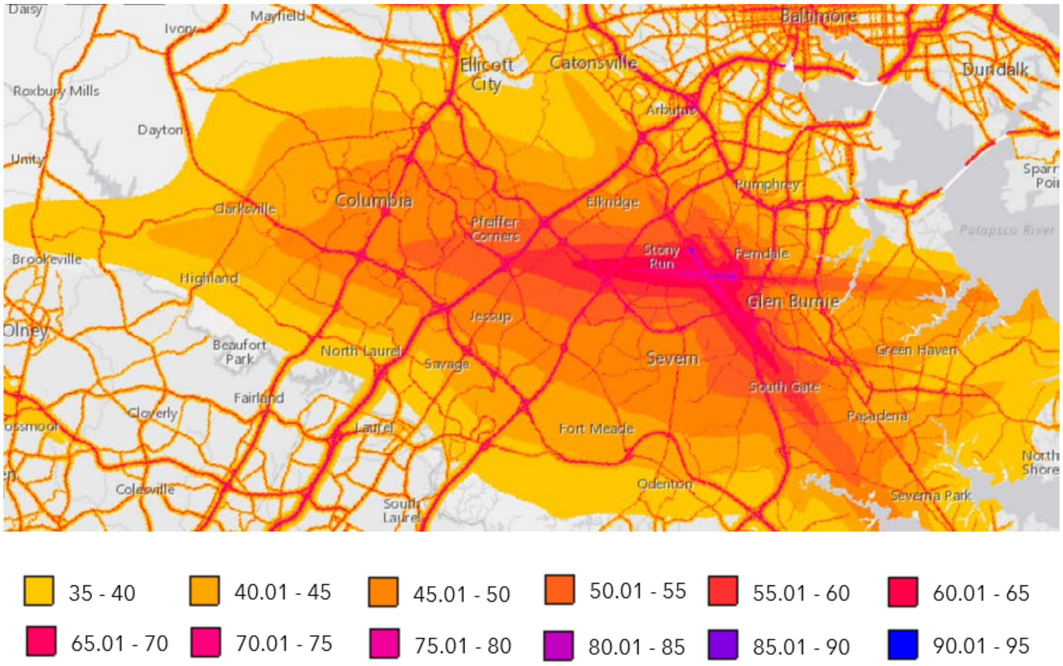
The second stage of sample selection involved selecting addresses at each sampled airport to receive the mail questionnaire. The primary goal of the design was to achieve high precision for estimating the logistic regression curve at each airport, as well as high precision for estimating the predicted percent HA at specific values of DNL between 50 and 75 dB. High precision for the individual airport curves would also contribute to the precision of the estimated national airport curve by reducing the within-airport variability. In addition, it was desired that the design would allow assessment of the fit of the model.

3.1 Features of Sampling Regions

Figure 2 displays noise contours for Baltimore/Washington International Thurgood Marshall Airport (BWI), chosen for illustration because the JSM 2017 conference was held in Baltimore. The figure shows noise from road traffic as well as aircraft, but the aviation noise from BWI can be seen in the ellipsoidal contours around the two runways, which are close to the purple regions of the figure. Figure 2 displays contours for the A-weighted equivalent sound level, which is similar to DNL but does not have the 10 dB nighttime penalty: $\text{LEQ} = 10\log_{10}\left[\frac{1}{86{,}400}\int_{7\ am}^{7\ am} 10^{L(t)/10}\, dt\right]$. BTS (2017) gave the following comparable sounds for LEQ levels: 40 dB corresponds to a humming refrigerator, 50 dB to a quiet office, 60 dB to conversational speech, 70 dB to a vacuum cleaner, and 80 dB to a garbage disposal.

Two features from Figure 2 have implications for the design and analysis of the NES. First, there are many addresses with low noise exposure but relatively few with high noise exposure. This implies that a random or self-weighting sample of addresses from the region would have few (if any) addresses with noise exposure above the “significant noise” level of DNL 65 dB. The second feature evident from Figure 2 is the irregular shape of the noise contours. The contours roughly follow the runways and do not accord with census blocks. Thus, known population counts such as the number of owner-occupied units in a census block do not correspond to the strata in the sampling design.

The sampling region for each airport was defined to be addresses that are inside the DNL 50 dB aircraft noise contour. EPA (1974) stated that “interference with activity and annoyance will not occur if outdoor levels are maintained at an energy equivalent of 55 dB,” and the sampling region was extended downwards to 50 dB in accordance with earlier studies. It was thought that households with noise exposure below 50 dB would report low annoyance and, therefore, would contribute little information for estimating the parameters of the logistic regression curves. It was expected that few households would have noise exposures above 75 dB.



A-weighted 24-hour equivalent sound level (dB)

Figure 2: Aviation and road traffic noise contours for Baltimore/Washington International Thurgood Marshall Airport in 2014. Source: BTS (2017).

3.2 Optimal Design for Logistic Regression

The goal of the survey is to fit a logistic regression curve relating noise exposure (DNL) and the binary response of high annoyance. A D-optimal design allocates the observations to x values so as to maximize the log determinant of the information matrix. Since the information matrix is proportional to the inverse of the asymptotic covariance matrix of the parameter estimates, the D-optimal design minimizes the volume of the confidence ellipsoid of the estimated regression parameters. In linear regression, the D-optimal design places half of the data points at each endpoint of the study region, and the same design is optimal for any possible values of the parameters. For logistic regression, however, the optimal design depends on the parameters of the curve.

Figure 3 shows the dependence of the D-optimal design on the form of the curve. If the line in Figure 3 represents the true logistic regression curve, then the D-optimal design places half of the data points at the dose corresponding to a predicted probability of 0.824, and the other half at the dose corresponding to a predicted probability of 0.176 (Abdelbasit and Plackett, 1983). These doses are indicated by the arrows in Figure 3. To be able to estimate the curve, at least some design points are needed in the shaded region of the graph, because that is where the curve is changing. If all observations were taken outside of the shaded region, there would be little information for estimating the logistic regression parameters: for example, if all observations were taken in the area to the left of the shaded region, it would be expected that most of the responses would be 0 and there would be no data points in the "action area" of the curve. Thus, using a design that is D-optimal for the wrong regression function could lead to an inability to estimate the curve.

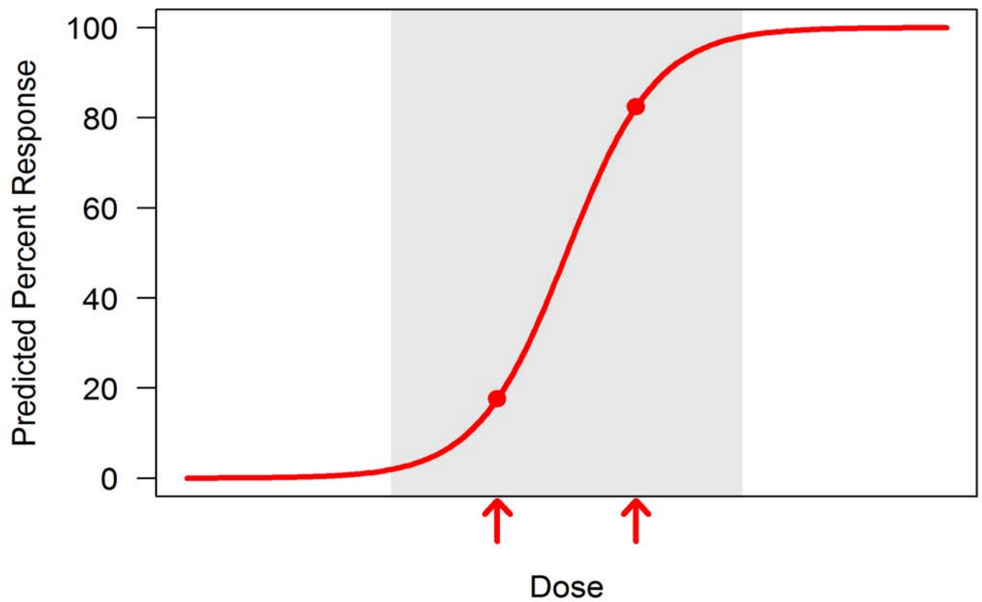


Figure 3: D-optimal design points, assuming that the logistic regression curve parameters are known. The D-optimal design specifies taking half of the observations at the values of x by each arrow.

In reality, however, the D-optimal design cannot be implemented because the true curve parameters are unknown—indeed, if the parameters were already known, there would be no need for the survey. For the FICON (1992) curve given in Equation (2), the optimal design places half the observations at DNL 68 dB and the other half at DNL 89.9 dB. If the new curve has different parameters, however, the design that is optimal for estimating the FICON curve may miss the "action region" of the new curve entirely. In addition, almost all residential addresses have DNL values less than 80 dB, so that the upper D-optimal design point of 89.9 dB cannot be sampled.

Chaloner and Larntz (1989) developed Bayesian optimal design theory for logistic regression, which incorporated uncertainty about the parameters as a prior distribution on the parameters and then computed the design that maximizes the preposterior expected utility. They showed that the number of support points in the Bayesian optimal design increases with increasingly vague prior information; moreover, uncertainty about the intercept has more effect on the support points of the Bayesian optimal design than uncertainty about the slope. A different intercept shifts the entire curve horizontally, so uncertainty about the intercept requires more design support points to ensure that the design includes the region where the logistic curve is changing.

3.3 Prior Information

Previous social surveys relating annoyance to aircraft noise exposure were used as prior information to inform the sampling design. Fidell and Silvati (2004) and Fidell et al. (2011) summarized data from some of the surveys that had been taken before those dates. The summaries provided the percentage highly annoyed at different sites in the survey, where all of the addresses for a particular site were assumed to have the same value of DNL.

When available, we used the sample size for each site from Fidell and Silvati (2004); otherwise, we assumed that the total sample size for the study, from Bassarab et al. (2009), was divided equally among the sites. We did not verify the accuracy of the data with the original sources, but we corrected inconsistencies between the data values given in the two papers. For example, Fidell et al. (2011) excluded data points from Fidell and Silvati (2004) that had values of zero for percent highly annoyed. After verifying from some of the original studies and other sources (Fidell et al., 1989; Fidell et al., 2002) that the data points with zero percent HA for that study were valid, we included those zeroes in the data.

Figure 4 shows the estimated logistic regression curves from the 42 studies that reported percent HA for at least two distinct values of DNL. These are grouped by decade, with darker colors representing more recent studies. There is substantial variability from study to study; some of the previous studies even exhibit a negative slope. In addition, the more recent studies appear to have an increased level of annoyance (see also Janssen et al., 2011); however, those differences might also be attributed to different survey methodologies or differences among countries rather than a time trend.

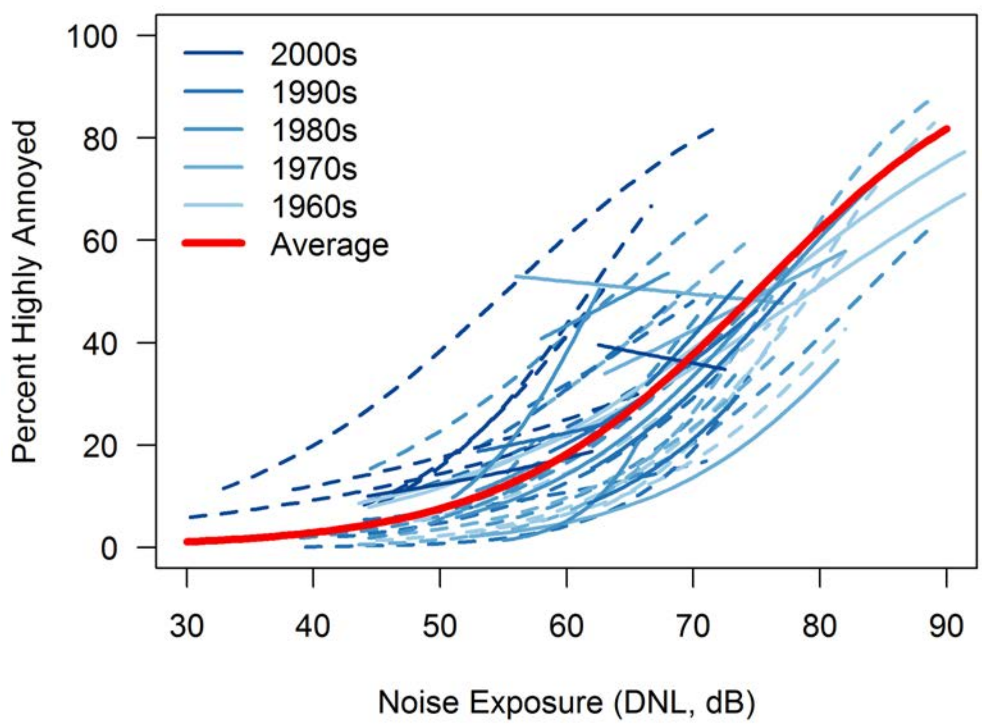


Figure 4: Logistic regression curves fit to 42 previous aircraft noise surveys, together with the average of the curves. The shading corresponds to the decade in which the study data were collected. Solid lines represent surveys conducted in the U.S.; dashed lines are from surveys conducted in other countries.

For the studies shown in Figure 4, the slopes have mean 0.10 and standard deviation 0.053; the intercepts have mean –7.5 and standard deviation 3.7; the values of μ (the value of DNL for which the predicted percent HA equals 50) have mean 75 and standard deviation 11. These standard deviations likely understate the prior uncertainty about the parameters because the data contain few recent U.S. surveys and many of the previous studies were conducted using different methods. If independent uniform (50, 100) and uniform (0, 0.2) distributions are used for μ and β_1, the Bayesian D-optimal design for estimating the regression parameters places half of the design mass at the extremes of the sampling region; this deviation from the usual placement of points in the middle of the region occurs because the mean logistic curve from the prior information is relatively flat and reaches a maximum percent HA of 50 percent at DNL 75 dB instead of having the full range of possible values of percent HA. However, that design is optimal under the assumption that the two-parameter logistic model is appropriate, and it does not allow assessment of the fit of the model. The optimal design moves toward a uniform distribution in the sampling region if the prior information also incorporates uncertainty about the form of the model.

3.4 Design

Optimal design theory provides guidance for the design; it does not prescribe it. There may be multiple designs that produce precisions close to that for the optimal design. In addition, the designs are optimal under the specific assumptions used, and may perform poorly if those assumptions are not met. For the NES, however, the optimal and near-optimal designs all specified taking at least 20 percent of the observations in the upper 20 percent of the noise exposure range for an airport. The Bayesian D-optimal design, assuming a two-parameter logistic regression function with prior information from previous studies, specifies taking observations at the extremes of the sampling region. If the variability of the prior information is increased, however, or if different curves (such as a logistic regression with quadratic and/or cubic terms in DNL) are allowed, the Bayesian D-optimal design adds points to the middle of the design region and approaches a uniform spread of points.

In this section we compare the Bayesian D-optimal and uniform designs with a design that uses proportional allocation. The sample was to be selected using address-based sampling, with the probability of selection depending on the anticipated noise exposure at the address. Contours of DNL values for each sampled airport were calculated by Harris Miller Miller and Hanson Inc. using flight information from 2012 and 2013. The contours were used to define the sampling region for each airport as all addresses inside of the 50 dB contour, and also to define up to five noise strata for each airport. The noise strata consisted of DNL values of 50-55 dB, 55-60 dB, 60-65 dB, 65-70 dB, and 70+ dB. The population count of addresses was then determined for each noise stratum within each sampled airport.

For most airports, as illustrated in Figure 2 for BWI, there is a greater area and thus many more residential addresses in low noise strata than in high noise strata. Table 2 displays a typical distribution of population counts by noise stratum, for a hypothetical airport that has all noise strata present. Table 2 shows that there are many more residential addresses in the low noise stratum than in the highest noise stratum. If proportional allocation were used with a sample of size 500, the resulting sample would have only ten households that are exposed to noise above 65 dB. The Bayesian D-optimal allocation would put half of the observations close to the lower boundary and the other half close to the upper boundary of the sampling region. This would give the most precision for estimating the parameters of the curve, but would not have any points in the middle of the data that could be used to check that the two-parameter logistic model accurately summarizes the true curve in that region. An “equal allocation” places 100 observations in each noise stratum.

Table 2: Hypothetical Airport: Population Counts and Three Allocations for a Sample of Size 500

	Noise Stratum				
	50-55 dB	*55-60 dB*	*60-65 dB*	*65-70 dB*	*70+ dB*
Population Count	65,000	25,000	8,000	1,600	400
Proportional Allocation	325	125	40	8	2
Optimal Allocation	250	0	0	0	250
Equal Allocation	100	100	100	100	100

Figure 5 displays the anticipated precision from each of these allocations if the true dose-response curve is the average of the logistic regression curves in Figure 4, with intercept −7.5 and slope 0.1. The dose-response curve is shown in black, along with the anticipated 95 percent confidence bands for each allocation. For illustration, Figure 5 shows the anticipated precision for only one specific dose-response curve; however, the same patterns appeared when these calculations were repeated with other possible values of the parameters. For all values of parameters tested, the proportional allocation had low precision for estimating the parameters because it placed most of the observations at the low end of the noise exposure range. The equal allocation had almost as much precision as the optimal allocation, and had additional advantages of allowing assessment of the model form and being simple to administer. The equal allocation also guaranteed that any dose-response curve that has an “action region” somewhere between DNL 50 dB and 75 dB will be able to be estimated from the data; the optimal allocation and proportional allocations both might have few observations in the “action region” for the curve and thus might have low precision for estimating some possible curves.

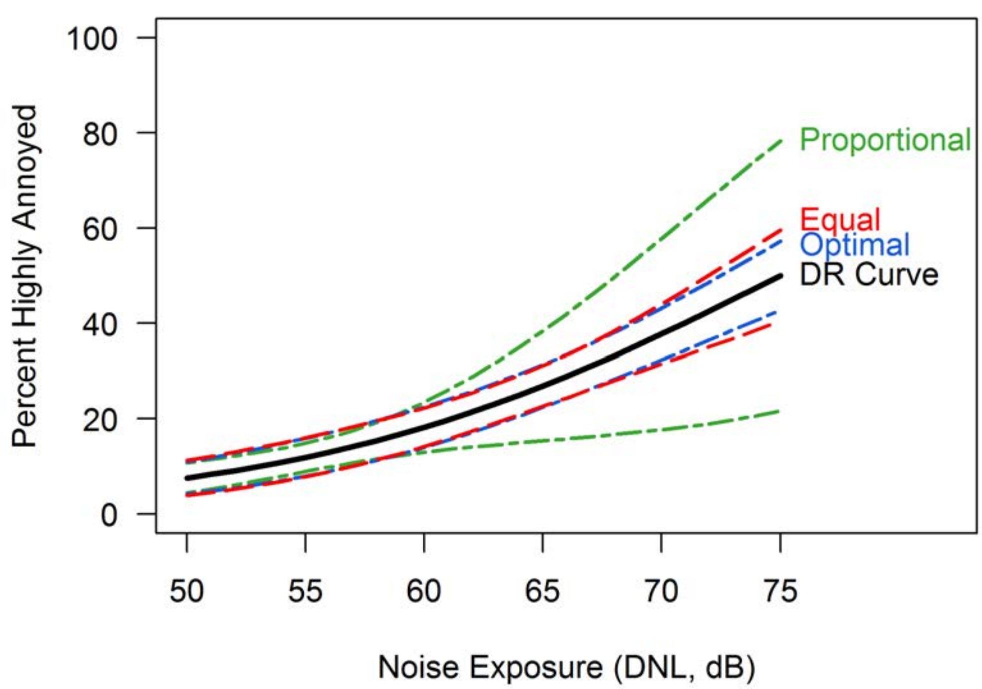


Figure 5: Anticipated 95 percent confidence bands for proportional, optimal, and equal allocation if the true dose-response (DR) curve is a logistic regression function with intercept −7.5 and slope 0.1.

The equal allocation is thus a good choice for the design. However, the equal allocation leads to a highly disproportionate stratified sample of addresses from the sampling region, and this issue will be discussed in the next section.

4. Weighting

4.1 Sampling Weights

Section 2 described the balanced sampling design used to select the airports for the study. This design was intended to obtain a sample of 20 airports that represented the 95 airports in the population, in that the characteristics in Table 1 from the sample match those for the population. Because the sample was designed to be representative of the population as a whole, a model-based analysis was planned to fit the national airport curve. The sample was selected with the purpose of using a random coefficients regression model for the analysis. This implies that each airport is weighted equally in the analysis.

As described in Section 3, however, the sample of addresses at each airport with the equal allocation is highly disproportionate. If the within-airport sampling weights were used to estimate the individual airport logistic regression relationships, the variability of the weights would result in low precision for the estimated parameters. The design weights are high for addresses with low noise exposure, and very low for addresses with high noise exposure. Thus, using the weights would mean that the regression relationship would be determined almost exclusively by low-noise-exposure households. In essence, using the sampling weights from the stratified sample to estimate the dose-response curve would negate the efficiency gains from the equal allocation design. Thus, it was decided to weight each sampled address equally when fitting the dose-response curves.

The decision to weight each observation from an airport equally when fitting the airport-specific dose-response curve can be justified on several grounds. From a model-based perspective, the logistic regression model for each airport is assumed to describe the relationship between dose and response. The regression relationship models annoyance conditionally on the noise exposure; therefore, if the model holds, the sample selection mechanism does not affect the relationship.

Because regression analyses are performed conditionally on the independent (x) variable (here, DNL), if there were full response weights would not be needed in the analysis. Pfeffermann and Sverchkov (1999) provided a theoretical justification for using equal weights for each sampled address. They proposed estimators for regression models that account for the effect of nonignorable sample schemes. If the sample selection is related to the dependent variable, then ignoring the sampling weights can create bias. For the NES, however, the sample selection is related to the independent variable, DNL. Pfeffermann and Sverchkov (1999) calculated modified “q weights” that divide the design weight (the inverse of the inclusion probability) by the conditional expected value of the design weight given x. Pfeffermann et al. (2006) extended this method to multilevel models. Because the inclusion probabilities in the NES are functions of x, the q-weight for each unit can be treated as 1.

4.2 Weighting for Nonresponse Adjustments

As argued in the previous section, the analysis is conditional upon the x values, and in the absence of nonresponse the national curve is estimated using a logistic regression with each observation having a weight of one. In reality, though, there is nonresponse to the survey and it is possible that the nonresponse is related to the outcome variable (whether the household reports being highly annoyed) in addition to DNL. Part of a nonresponse bias assessment involves evaluating differences between respondents and nonrespondents for variables that are known from the sampling frame. In addition, it may also be desired to develop nonresponse-adjusted weights and explore whether the estimated dose-response curves differ when those weights are used.

Many surveys adjust the weights so that estimates calculated using the adjusted weights agree with independent control totals which are often derived from the decennial census or the American Community Survey (ACS). Estimates of the characteristics used in the control totals that are calculated using the adjusted weights are thus unbiased because the census or ACS quantities are considered to be a gold standard. It is thought in general that these adjustments reduce nonresponse bias for other variables as well (Brick, 2013). For the NES, however, the sampled area does not correspond with a set of census blocks because of the irregular shapes of the strata. Therefore, adjusting to census block or block group population characteristics could actually increase rather than decrease bias. The bias could increase if the population characteristics for the parts of the census block groups outside of the sampled area differ from those of the block groups inside the sampled area. Additionally, even if census characteristics were available for units of geography that coincide with the sampling region, the highly disproportionate sampling design would lead to instability in the weight adjustments.

Instead, the initial weight for every observation in the selected sample from an airport is set to one, and the weights can be adjusted using characteristics known from the sample frame for the selected sample. Ratio adjustments or raking could be used for these adjustments. For the NES, it is proposed to estimate response propensities separately for each airport using tree-based models (Lohr et al., 2015). The construction of nonresponse-adjusted weights allows a sensitivity analysis to be performed, to assess whether the curve differs when these weights are used. If the nonresponse adjustments have an effect on the curve, it may be desired to include some of the variables used in the nonresponse adjustments as covariates in the dose-response model.

5. Conclusions

This paper summarized the methods and considerations used for the design of the NES. The design and weighting considerations may also be useful for other surveys in which the primary goal is to estimate a dose-response relationship. In some environmental surveys, the dose might be considered to be the distance from the source of a toxin, and responses might include various health outcomes. In other surveys, dose contours could be constructed using data from air pollution monitors. For other situations, there may be a database of persons listing the number of years each person worked at a chemical plant, where years can be considered as a proxy for dose.

For dose-response relationships where the dose is related to the geographical distance from a location, as with the NES, there are many people who are exposed to low doses but relatively few who are exposed to high doses. In these situations, sampling addresses with allocation proportional to population size yields a sample with poor precision for estimating the dose-response relationship, and a highly disproportionate allocation is needed to be able to have adequate representation of high-dose individuals in the sample.

For surveys in which a dose-response relationship is of primary interest, we recommend using a stratified sampling allocation in which different exposures are represented approximately equally in the sample. In this design, the selection probabilities depend only on the dose variable, which allows the weight for each observation to be set equal to one when estimating the dose-response curve. If the population to be sampled has many individuals at low dose levels but few individuals at high dose levels, using equal weights gives high precision for estimating the curve; an analysis using the highly disproportionate sampling rates will rely almost entirely on the low-dose individuals in the sample and will have precision closer to that from a proportional allocation.

Note that most of the other aircraft noise surveys in the literature reviewed by Fidell et al. (2011) and Gelderblom et al. (2017) have not used sampling weights. Nguyen et al. (2016), for example, selected 13 sites near Hanoi Noi Bai International Airport to be able to study differences between sites under the landing route and sites under the take-off route, and interviewed approximately 100 households at each site. This resulted in a cluster sample of households, and the logistic regression function was fit conditionally on the noise exposures for each site so that weights were not needed. An exception is CAA (2017), which slightly oversampled high-noise exposure addresses and used weights to account for small differences in probabilities of selection of addresses for different airports and noise bands. The survey had two noise strata and most of the sampled addresses in the upper noise stratum were at the lower end of the band, so that few addresses with high noise exposure were selected. However, the goal of the CAA survey was to obtain a representative sample of about 2,000 adults in proximity to nine airports in England, and the survey estimated characteristics of the entire population. The CAA survey obtained 1,369 responses for Heathrow and 204 for Gatwick but the sample sizes for some of the other airports (less than 30 for four of the airports) were too small to allow estimation of airport-specific dose-response curves.

Many of the face-to-face surveys have selected clusters of addresses to reduce interviewer travel costs. For the NES, however, the mail mode allowed selection of a simple random sample of addresses in each noise stratum with no increase in cost over a cluster sample. This increases precision for the estimated airport-specific dose-response curves because the observations at each airport can be treated as independent.

The goal of the NES was to fit separate logistic regression relationships for each airport in the sample, and to estimate a national dose-response curve. In other applications, there may be interest in estimating the characteristics of the exposed population as well as the dose-response relationship. In such a situation, it may be desired to also compute the survey weights to estimate population means and totals, and it may be desired to use a design that is between the highly disproportional design that is efficient for estimating the dose-response curve, and a proportional allocation design that has lower variance for estimating means and totals.

Acknowledgments

This research was supported by the Federal Aviation Administration under contract FAA DTFACT-15-D-00008. The views presented in this paper are those of the authors and do not represent the official views of any Federal Agency/Department or Westat. The authors are grateful to Natalia Sizov for her helpful comments on this paper.

References

Abdelbasit, K. M., and Plackett, R. L. (1983). "Experimental Design for Binary Data." *Journal of the American Statistical Association*, 78, 90-98.

Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S., and Stansfeld, S. (2014). "Auditory and Non-Auditory Effects of Noise on Health." *The Lancet*, 383, 1325-1332

Bassarab, R., Sharp, B., and Robinette, B. (2009). "An Updated Catalog of 628 Social Surveys of Residents' Reactions to Environmental Noise (1943-2008)." Report DOT/FAA/AEE/2009-01. Available from https://www.faa.gov/about/office_org/headquarters_offices/apl/research/science_integrated_modeling/media/An%20Updated%20Catalog%20of%20628%20Social%20Surveys.pdf.

Brick, J. M. (2013). "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics*, 29, 329-353.

Bureau of Transportation Statistics (BTS, 2017). "National Transportation Noise Map." Available from https://www.bts.gov/newsroom/national-transportation-noise-map.

Chaloner, K., and Larntz, K. (1989). "Optimal Bayesian Design Applied to Logistic Regression Experiments." *Journal of Statistical Planning and Inference*, 21, 191-208.

Civil Aviation Authority (CAA, 2017). "Survey of Noise Attitudes 2014: Aircraft." Civil Aviation Authority: West Sussex, UK. Available from http://publicapps.caa.co.uk/docs/33/CAP%201506%20FEB17.pdf.

Environmental Protection Agency (EPA, 1974). *Information on Levels of Environmental Noise Requisite to Protect Public Health and Welfare with an Adequate Margin of Safety*. 550/9- 74-004. Washington, D.C.: Office of Noise Abatement and Control. March. Available from https://www.rosemonteis.us/documents/usepa-1974.

Federal Aviation Administration (FAA, 2017). "Aircraft Noise Issues." Available from https://www.faa.gov/about/office_org/headquarters_offices/apl/noise_emissions/airport_aircraft_noise_issues/.

Federal Interagency Committee on Noise (FICON, 1992). "Federal Agency Review of Selected Airport Noise Analysis Issues, Final Report: Airport Noise Assessment Methodologies and Metrics." Report for the Department of Defense, Washington, D.C. Available from ftp://ftp.agl.faa.gov/Materials%20Released%20Related%20to%20the%20OM%20EIS/3-31-2005%20World%20Gateway%20Related%20Documents/787_122.pdf.

Fidell, S., Barber, D., and Schultz, T. J. (1989). "Updating a Dosage-effect Relationship for the Prevalence of Annoyance due to General Transportation Noise." Report HSD_TR-89-009, Brooks Air Force Base, TX: Noise and Sonic Boom Impact Technology, Air Force Systems Command.

Fidell, S., Mestre, V., Schomer, P., Berry, B., Gjestland, T., Vallet, M., and Reid, T. (2011). "A First-Principles Model for Estimating the Prevalence of Annoyance with Aircraft Noise Exposure." *Journal of the Acoustical Society of America*, 130, 791-806.

Fidell, S., and Silvati, L. (2004). "Parsimonious Alternatives to Regression Analysis for Characterizing Prevalence Rates of Aircraft Noise Annoyance," *Noise Control Engineering Journal*, 52, 56-68.

Fidell, S., Silvati, L., and Haboly, E. (2002). "Social Survey of Community Response to a Step Change in Aircraft Noise Exposure." *Journal of the Acoustical Society of America*, 111, 200-209.

Fields, J., De Jong, R. G., Gjestland, T., Flindell, I. H., Job, R. F. S., Kurra, S., ... and Guski, R. (2001). "Standardized General-Purpose Noise Reaction Questions for Community Noise Surveys: Research and a Recommendation." *Journal of Sound and Vibration*, 242, 641-679.

Gelderblom, F. B., Gjestland, T., Fidell, S., and Berry, B. (2017). "On the Stability of Community Tolerance for Aircraft Noise," *Acta Acustica United with Acustica*, 103, 17-27.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." *Journal of the American Statistical Association*, 78, 776-793.

Hansen, M. H. and Madow, W. G. (1978). "Estimation and Inferences from Sample Surveys: Some Comments on Recent Developments." In *Survey Sampling and Measurement*, ed. N. Namboodiri, New York: Academic Press, 341-357.

Janssen, S. A., Vos, H., van Kempen, E. E., Breugelmans, O. R., and Miedema, H. M. (2011). "Trends in Aircraft Noise Annoyance: The Role of Study and Sample Characteristics." *The Journal of the Acoustical Society of America*, 129, 1953-1962.

Lohr, S. (2014). "Design Effects for a Regression Slope in a Cluster Sample." *Journal of Survey Statistics and Methodology*, 2, 97-125.

Lohr, S., Hsu, V., and Montaquila, J. (2015). "Using Classification and Regression Trees to Model Survey Nonresponse." In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2071-2085.

Nguyen, T. L., Yano, T., Nishimura, T., Sato, T., Morinaga, M., and Yamada, I. (2016). "Social Surveys on Community Response to a Change in Aircraft Noise Exposure Before and After the Operation of the New Terminal Building in Hanoi Noi Bai International Airport." In INTER-NOISE and NOISE-CON Congress and Conference Proceedings, pp. 1487-1495.

Pfeffermann, D., Moura, F. A. S., and Nascimento-Silva, P. L. (2006). "Multilevel Modeling Under Informative Sampling." *Biometrika*, 93, 943-959.

Pfeffermann, D. and Sverchkov, M. (1999). "Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data." *Sankhyā: The Indian Journal of Statistics, Series B*, 61, 166-186.

Schultz, T. J. (1978). "Synthesis of Social Surveys on Noise Annoyance." *Journal of the Acoustical Society of America*, 64, 377-405.

Sizov, N. and Pickard, L. (2011). "An Update on Research to Guide United States Policy on Aircraft Noise Impact." *Proceedings of the 11th International Congress on Noise as a Public Health Problem*. Available from https://www.faa.gov/about/office_org/headquarters_offices/apl/research/science_integrated_modeling/noise_impacts/media/ICBEN_11_SIZOV.pdf.

Tillé, Y. (2009). "Ten Years of Balanced Sampling with the Cube Method: An Appraisal." *Survey Methodology*, 37, 215-226.

Valliant, R. Dorfman, A. H., and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach.* New York: Wiley.