DOI: 10.61782/fa.2025.0690

FORUM ACUSTICUM
EURONOISE 2025

KEY ASPECTS FOR SURVEY DATA PROCESSING AND STATISTICAL ANALYSIS WHEN MODELLING EXPOSURE-ANNOYANCE RELATIONSHIPS

Mark Brink[1*] Beat Schäffer[1] Dirk Schreckenberg[2]
[1] Swiss Federal Office for the Environment, CH-3003 Bern, Switzerland
[2] Empa, CH-8600 Dübendorf, Switzerland
[3] ZEUS GmbH, D-58095 Hagen, Germany

ABSTRACT

In the context of (classical) noise annoyance surveys, the main product of interest is usually exposure-response relationships to inform health risk assessments or policy makers. Surveys conducted for this purpose are becoming increasingly complex, raising the question of what the key aspects in terms of data processing and statistical analysis are, that allow valid and representative exposure-response functions to be derived from the data collected. In this talk we will look at relevant decisions at different stages of analysis and illustrate their impact, in parts using data from the NORAH and SiRENE noise annoyance surveys as examples. The focus is on the necessary data processing and statistical considerations when aiming at statistically sound predictions of the percentage of 'highly annoyed' (%HA) as a function of noise exposure, alongside potential further predictors. The paper touches on the issues of (a) the explanatory or predictive approach related to variable selection and modelling, (b) the use of survey weights or not, (c) the proper handling of cases with low or missing exposure values in survey samples, (d) population-averaged vs. subject-specific responses and (e) the treatment of non-focal predictors when plotting exposure-response curves.

Keywords: noise annoyance, exposure-response relationship, statistical modelling

[*] ***Corresponding author***: *mark.brink @bafu.admin.ch.*

1. STARTING POINT

Annoyance is one of the most widespread immediate and long-term effects of transportation noise exposure, and this outcome has always played a pivotal role in informing policy, carrying out environmental impact assessments, or setting noise limits. All these goals require solid and valid exposure-response functions, which are derived from the data of so-called socioacoustic surveys. Most common statistical techniques to derive these functions are generalized linear (mixed) models. Corresponding model parameters are used to predict the outcome (mostly, the percentage highly annoyed, %HA) for a given level of noise exposure, which can be visualized in exposure-response curves (i.e., graphical representations of the exposure-effect relationship). Even though statistical analyses are *the* crucial part in the processing of data from socio-acoustic studies to derive exposure-response relationships, the steps to arrive there are usually hardly questioned out of habit – although there are various important decisions to be made regarding how to proceed in each particular case. While the things that are usually decided during the planning/conceptualization phase of a survey, such as sampling strategy, stratification, the way the questionnaires are presented, survey mode, response scales used etc., are of course crucial factors for the final result (cf. for example [1, 2]), the focus of this conference talk is basically on what happens *after* data collection. The talk covers the following points:

(1) The decision for an either explanatory or predictive approach related to variable selection and modelling

(2) Whether to use survey weights in the data prior to inferential statistical modelling
(3) The proper handling of low and missing exposure values in survey samples (censoring vs. truncation)
(4) The choice of modelling population-averaged or subject-specific exposure-response relationships (in the case of longitudinal surveys)
(5) The treatment of non-focal predictor variables when plotting exposure-response curves.

The present short paper is a brief preview of the conference talk. Our thoughts on the above topics will be consolidated in a journal paper containing a more extensive discussion and detailed recommendations.

2. BRIEF ROUNDUP OF COVERED POINTS

2.1 Aspect #1: Purpose of an annoyance survey: Explanation versus prediction

We argue that the distinction between predictive and explanatory approach is key when analysing data from socioacoustic surveys. Is the goal of the study to obtain, from amongst a large collection of variables, a set which best predicts the outcome variable? Or is the goal to estimate the association between a risk factor (i.e., exposure or treatment variable) and the outcome? Indeed, this principal decision should be taken even before designing a survey as it affects every aspect of model construction and evaluation thereafter. Just look back at your own career: you will find introductory texts from term papers to theses to maybe even the first peer-reviewed articles later on, where your research goals were rather vaguely described, perhaps in the sense of "by this research we aim at a *better understanding* of the association between x and y" or "this research wants to *examine* the factors that influence x"... things of that sort (by the way, we are no exception). A very basic question that should be answered before the set-up of any noise study (and even more so before the analysis begins) is whether the research pursues a *descriptive* goal (summarizing characteristics), a *predictive* goal (forecast effects [of noise, and maybe additional variables]), or, whether the goal is *explanatory* (unravel causal mechanisms). Various statisticians have pointed out the importance of this distinction [e.g. 3]. When it comes to environmental impact assessments, the explanatory ("risk factor") approach is perhaps the least cumbersome because it leads to more easily interpretable exposure-effect curves.

2.2 Aspect #2: Should one use survey weights?

In response to a request from a colleague, we deal with the question of whether case weighting may have an advantage in socio-acoustic surveys or not. For purely descriptive purposes, where the aim is to achieve the highest possible representativeness for the population, the weighting of sample data is widely accepted. For analytical conclusions about model parameters, however, there is a wide range of opinions about the role of sample weights among statisticians. We concluded that it is better to avoid weighting and stick to the original data when inferential statistical questions are in the foreground. The disadvantage of weighting is that it moves the data away from what has actually been empirically observed. And experience shows that the more data have been processed, the more difficult it becomes to understand what they represent. Also, results based on a weighted data set can be more difficult to interpret or communicate to stakeholders, especially if the weighting factors are complex or not well known.

2.3 Aspect #3: How to handle cases with low and missing exposure values?

Standard uncertainties in noise exposure assessments are larger at lower exposure levels than in the vicinity of infrastructure (roads, railways, airports) where the respective noise levels are higher. For some of the receiver points in a sample, the noise levels may even fall outside the calculation grid and thus are completely missing. This is often the case in socio-acoustic surveys, in which several noise sources are examined simultaneously in the same sample, but where not all noise sources are present at each receiver. Further, modelled exposure values may be so low that they may not actually be perceptible in "real life" with other ambient sound sources [4]. This raises the question of how low the bottom of an exposure-response function should be for a particular noise source and whether cases with very low or missing exposure values should be a) included in the analysis, b) replaced by a specified value (censoring), or c) excluded from the analysis altogether (truncation). In the talk, we discuss these alternatives and come to a fairly clear judgment, namely, that the inclusion of essentially unexposed cases has the potential to bias the resulting exposure-response curve. We therefore recommend that, given the weaknesses in exposure estimation at low exposure levels, cases with exposure levels below some 30 to 40 dB (L_{den} or L_{Day}) should be excluded from the analysis. This may limit the number of cases available, but will lead to better, i.e. more representative %HA predictions, especially at higher exposure levels.

2.4 Aspect #4: GEE vs GLMM modelling

The primary product one usually has in mind when carrying out socio-acoustic surveys are noise exposure-response relationships which are important for predicting the effects of noise on individuals and populations. In the case of %HA-research, such relationships can be derived from independent (this means: non-nested and non-hierarchical) or, – admittedly – less often, repeated binary observations ("repeated measures"; e.g. in laboratory listening studies/experiments). As long as the binary observations are independent, a "standard" logistic regression analysis is applied, which yields so-called population-averaged exposure-response relationships that are directly applicable in risk assessment on the population level. However, if data from repeated binary observations are collected, one needs to account for the correlation of the data within individuals. Possible approaches to do so are generalized linear mixed models (GLMM) [5] or generalized estimating equations (GEE), [6]. We thoroughly discuss the two options of deriving exposure-response relationships in longitudinal or repeated measures studies, namely, the population average versus subject-specific prediction, that can be achieved with the GEE and GLMM modelling approaches, respectively.

2.5 Aspect # 5: Treatment of non-focal predictor variables when plotting exposure-response curves

For most use cases, socioacoustic surveys are carried out to describe or indeed, predict the association between some amount of noise exposure and a corresponding health effect. The minimum set of variables that is needed for this is an exposure variable, here L_{den}, and an effect variable, here HA. However, most survey designs are more complex, and modelling involves additional variables such as level-2 variables in hierarchical (multilevel) designs and/or additional covariates and/or confounders. Regardless of the number of predictors in the model, one can just use two dimensions to describe the relationship between exposure and effect in a graphically simple (two-dimensional) exposure-response plot with the so-called *focal term* (e.g. L_{den}) on the x-axis and the outcome on the y-axis. This requires all *non-focal terms* (non-focal predictors in the model) to be fixed (kept constant) at a certain value. For continuous predictors, this is often (but not necessarily) the sample mean. If categorical variables with multiple levels are involved, e.g. "gender", or "school", it is not immediately evident how to "average" over them, especially when the group sizes are different. We show in detail which methods of marginalization of non-focal predictors are possible when it comes to draw exposure-response curves for %HA (or similar outcomes). We will particularly shed light on *empirical average* and *estimated marginal means* prediction.

Of course, we cannot – for all the aspects mentioned above – give an unequivocal recommendation in every case for or against a specific way of action. However, we believe it is important that our peers from the field of noise effects research are aware of the issues and pitfalls we raise in the talk, and that they can overcome them by making well-informed decisions. We hope that our soon to be published full paper on the topic will prove to be a useful resource for this decision-making process.

3. REFERENCES

[1] Brink M, Schreckenberg D, Vienneau D, et al. Effects of Scale, Question Location, Order of Response Alternatives, and Season on Self-Reported Noise Annoyance Using ICBEN Scales: A Field Experiment. *Int J Environ Res Public Health*2016 Nov 23;13(11).

[2] Gjestland T. Measuring Community Response to Noise—Factors Affecting the Results of Annoyance Surveys. *International Journal of Environmental Research and Public Health*2024;21(4):420.

[3] Shmueli G. To Explain or to Predict. *Statistical Science*2010;25(3):289-310.

[4] Mestre VE. Describing the uncertainty associated with low aircraft noise level data derived from noise models: implications for noise effects studies. *Internoise 2010*; Lisbon2010.

[5] Stiratelli R, Laird N, Ware JH. Random-effects model for serial observations with binary response. *Biometrics* 1984;40:961–71.

[6] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*1986;73(1):13-22.