Optimization of Two-Phase Sampling Designs with Application to Naturalistic Driving Studies

Citation for the original published paper (version of record):
Imberg, H., Lisovskaja, V., Selpi, S. et al (2022). Optimization of Two-Phase Sampling Designs with Application to Naturalistic Driving Studies. IEEE Transactions on Intelligent Transportation Systems, 23(4): 3575-3588. http://dx.doi.org/10.1109/TITS.2020.3038180

(article starts on next page)

Optimization of Two-Phase Sampling Designs With Application to Naturalistic Driving Studies

Henrik Imberg, Vera Lisovskaja, Selpi, *Member, IEEE*, and Olle Nerman

***Abstract*— Naturalistic driving studies (NDS) generate tremendous amounts of traffic data and constitute an important component of modern traffic safety research. However, analysis of the entire NDS database is rarely feasible, as it often requires expensive and time-consuming annotations of video sequences. We describe how automatic measurements, readily available in an NDS database, may be utilized for selection of time segments for annotation that are most informative with regards to detection of potential associations between driving behavior and a consecutive safety critical event. The methodology is illustrated and evaluated on data from a large naturalistic driving study, showing that the use of optimized instance selection may reduce the number of segments that need to be annotated by as much as 50%, compared to simple random sampling.**

***Index Terms*— Case-control studies, naturalistic driving studies, optimal design, pseudo-likelihood, safety critical event, unequal probability sampling.**

Nomenclature

ND	Naturalistic driving
NDS	Naturalistic driving studies
SCE	Safety critical event
SD	Standard deviation
SRS	Simple random sampling
WMLE	Weighted maximum likelihood estimator

I. Introduction

IN RECENT years, naturalistic driving studies (NDS), including naturalistic field operational tests, have been employed all around the globe, providing an important source of data for analysis and enabling a better understanding of driver behavior and traffic safety, for example 100-car [1], [2] and SHRP2 in the U.S.A. [3], [4], euroFOT [5], PROLOGUE [6], and UDRIVE [7] in Europe, as well as NDS in Australia [8] and in Japan [9]. In NDS, data is collected automatically for all driving sessions in a large fleet of vehicles for several months. These automatic recordings include vehicle data such as speed and direction; environmental conditions, lane position, location and surrounding traffic recorded by radar, video and other external instrumentation; and video recordings of the drivers face, pedal, and eye movements. The data provided by the NDS design thus offer many opportunities for analysis of both normal driving and safety critical events, and is richer than more traditional data sources such as crash databases [10], [11].

Despite recent advancements and investments into naturalistic data sources, there are many challenges remaining, largely related to the huge amount of heterogeneous and sometimes noisy data generated by NDS. For instance, the SHRP2 project collected more than a million hours of driving data, including both video and recordings of vehicle kinematics [12]. Thus, the sheer volume of data poses a major challenge in analysis of naturalistic driving (ND) data. On top of this comes issues with data quality, including data losses and errors in recorded vehicle kinematics [13] and challenges in the annotation of video recordings [14]. To address the data quality issue, the SHRP2 study employed a rigorous procedure for quality assurance and quality control [12]. Others have proposed using the Geographic Information System for quality control in NDS, for example to understand missing data due to existence of tunnels or to understand speed profile in relation to the road profile [13], [15]. Thus, there is a need for rigorous and efficient procedures to ensure high-quality data to be extracted from NDS.

In order to handle the large amounts of data produced by NDS, data thinning or subsampling is commonly employed. For example, [16] proposed a matched case-crossover approach to extract event and control information from the video part of ND data, while [10] used random baseline sampling method. Sampling based approaches become of even greater relevance when the analyses rely on information derived from the video data: the great cost associated with video annotation implies that statistical analyses based on video sequences must be restricted to only a limited subset of the original database. Thus, choosing this subset in a manner that captures as much of the available information as possible is essential.

In this paper, we address the issue of appropriate subset selection: we present an inferential framework that enables a flexible selection of video fragments for annotation from an NDS database, and show how this selection may be optimized

Manuscript received October 22, 2019; revised July 14, 2020 and September 30, 2020; accepted October 30, 2020. This work was supported by the Swedish Research Council under Grant VR 340-2012-5995. The work of Selpi was supported by the Chalmers Area of Advance Transport for the later stage of this study. The Associate Editor for this article was P. Edara. *(Corresponding author: Henrik Imberg.)*

Henrik Imberg and Olle Nerman are with the Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Gothenburg, Sweden (e-mail: imbergh@chalmers.se).

Vera Lisovskaja was with the Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Gothenburg, Sweden.

Selpi is with the Division of Vehicle Safety, Department of Mechanics and Maritime Sciences, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.

Digital Object Identifier 10.1109/TITS.2020.3038180

using information readily available in the database through automatic recordings of vehicle maneuver data. The methodology is illustrated using data collected in Sweden as part of the European large scale field operational test (euroFOT) study [5]. We demonstrate that a variance reduction of up to 50% compared to simple random sampling can be achieved. In other words, optimal sampling can lead to a performance on par with that of ordinary methods with up to 50% less annotation demand.

In the next section, we start by presenting a motivating example. We then review a common procedure for analysis of complex, two-phase, samples in Section III, and show in Section IV how the sample selection may be optimized. The application of the methodology to the euroFOT data, collected using Volvo cars in Sweden, can be found in Sections V and VI.

II. Motivating Example

A. An Embedded Experiment

Consider a traffic situation involving two vehicles, the vehicle taking part in the NDS study (the index car) and a front car. The two are driving at similar speeds, when the front car brakes. This scenario describes a situation where a potential safety critical event (SCE) can occur, namely a rear-end collision. Of interest is the question of whether the glancing behavior of the driver of the index car, namely whether he/she looks at the car in front when braking is initiated, the speed of the vehicles and time gap between the two cars at this initiation, have an impact on the likelihood that a safety critical event will occur. Mathematically, we could explore this relationship through e.g. an application of a logistic regression model, with presence, or absence, of an SCE being the dichotomous response, and speed at brake light, time gap at brake light and glancing away at brake light as explanatory variables.

B. Definition of Events

Explicitly, we define an instance to be a time segment initiated by the turning on of the brake lights of the front car and ending with the driver of the index car returning to normal driving after braking. It is also the turning on of the brake lights that define the timepoint at which braking is initiated. As collisions are rarely observed in naturalistic driving data, other safety critical events are often used as proxy endpoints; these SCEs are typically defined by a combination of kinematic triggers that identifies SCE candidates, and a visual review of videos, classifying the SCE candidates by whether they are relevant for traffic safety or not [2], [17]. Specifically, we are interested in the safety critical event characterized by the presence of a surprised reaction of the driver of the index car, commonly referred to as an "oops reaction" [18], [19]. Thus, the following information requires video annotation in order to be obtained: the timepoint at which the brake light turns on, whether the driver looks on or off road at this timepoint, and whether the event is safety critical with the driver displaying a subsequent surprised reaction after looking back on road.

C. Poisson Sampling

Due to financial constraints, only part of the relevant instances in an NDS database can be annotated; typically all the identified SCEs (cases) and some of the instances with no SCE associated (controls) [10]. A simple way of choosing controls to be annotated would be to toss a hypothetical weighted coin for each of the available non-SCE instances, a process referred to as Bernoulli random sampling in [20]. A somewhat more complex alternative would be a hypothetical sequence of tosses of different weighted coins, the so called Poisson random sampling. In this paper we will describe how the weights in such a sampling procedure can be chosen in a way that maximizes the information that could potentially be provided by this smaller sub-sample, ideally approaching the precision of estimation that would have been present were the whole data set (i.e. all the instances) analyzed.

D. Data

In the examples that follow, we use data from the euroFOT study, containing data from 100 Volvo cars collected during one year. All vehicles were supplied with specialized equipment, including video cameras and external radars. Thus, driver actions, environmental conditions, vehicle data and vehicle maneuvers were continuously recorded and stored. Additional details can be found from [21].

For the purpose of demonstrating the sampling approach described in this paper, 49 instances with an SCE and 500 randomly selected instances without an SCE were identified in the database. The subset from which these were selected constituted more than 1,000 driving hours of suitable filtered instances for the rear-end conflict described above.

Video review revealed data quality issues in 65 of the 500 control candidates, including no video ($n = 13$), poor video quality ($n = 12$), external factors hindering video annotation (e.g. poor light conditions or driver wearing glasses, $n = 26$) or the control candidate being judged as irrelevant for the event of interest (e.g. due to lane change or lead vehicle not braking, $n = 14$). The remaining 435 controls and 49 SCEs were fully annotated.

In brief, the length of the annotated events ranged from 20 to 30 seconds. The mean (SD) vehicle speed was 53.0 km/h (16.0) and time gap was 1.8 (0.8) seconds. Glances off road at brake light were present in 21 (42.9%) of the cases and 63 (14.5%) of the controls.

III. Weighted Estimation From Complex Samples

With the motivating example above in mind, consider a statistical model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ relating a response variable Y to a set of explanatory variables X, indexed by a parameter vector $\boldsymbol{\theta}$ (e.g. a logistic regression modeling the probability of an SCE). Consider also a collection $\mathcal{D}$ of instances (e.g. time segments started by the frontal car initiating a brake), for which the responses y_i and the explanatory variables $\boldsymbol{x}_i$ are registered.

For each instance in $\mathcal{D}$, two types of variables can be available: the ones that are measured automatically (e.g. acceleration) and the ones that require video annotation in order to obtain them (e.g. glancing behavior of the driver). We will denote the former by Z and the latter by W. In the model $f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, both Y and X can, at least partly, belong to this latter class of measurements that require annotations.

Suppose that each instance $i \in \mathcal{D}$ is assigned a positive probability π_i of being sampled, and that a subset $\mathcal{S}$ of $\mathcal{D}$ has been sampled and annotated; consequently, complete records $(\boldsymbol{x}_i, y_i)$ are observed for this subset only. Since different instances can have different sampling probabilities, as is the case for Poisson sampling, the ordinary maximum likelihood estimation, which assumes an independent and identically distributed sample, is generally not applicable. Instead, one may consider a weighted maximum likelihood estimator (WMLE), defined by:

$$\hat{\boldsymbol{\theta}}_\pi := \arg\max_{\boldsymbol{\theta}} \ \hat{\ell}_\pi(\boldsymbol{\theta}), \tag{1}$$

$$\hat{\ell}_\pi(\boldsymbol{\theta}) := \sum_{i\in\mathcal{S}} w_i \log f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i), \tag{2}$$

where the weights w_i may be taken as $w_i = 1/\pi_i$.

In the survey sampling literature, the WMLE (1) is known as a pseudo maximum likelihood estimator [22], and the sum (2), with weights taken as $w_i = 1/\pi_i$, as a Horvitz-Thompson estimator [23] of the log-likelihood

$$\ell_0(\boldsymbol{\theta}) := \sum_{i\in\mathcal{D}} \log f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i), \tag{3}$$

i.e. the log-likelihood we would have obtained if all data had been annotated. In particular, $\hat{\ell}_\pi(\boldsymbol{\theta})$ is an unbiased estimator of $\ell_0(\boldsymbol{\theta})$ provided that all sampling probabilities are strictly positive. Furthermore, it holds under general regularity conditions, as the size of the sample $\mathcal{S}$ gets large, that the distribution of $\hat{\boldsymbol{\theta}}_\pi$ under repeated subsampling from $\mathcal{D}$ converges to a normal distribution with mean $\boldsymbol{\theta}_0$ and covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$ [22], [24], where $\boldsymbol{\theta}_0$ is the maximizer of the log-likelihood (3) and

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})^{-1}\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})^{-1}, \tag{4}$$

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ell_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}, \tag{5}$$

$$\boldsymbol{V}(\boldsymbol{\theta}) = \left[\sum_{i\in\mathcal{D}} \frac{1-\pi_i}{\pi_i}\boldsymbol{s}_i\boldsymbol{s}_i^T + \sum_{i,j\in\mathcal{D}} \frac{\pi_{i,j}-\pi_i\pi_j}{\pi_i\pi_j}\boldsymbol{s}_i\boldsymbol{s}_j^T\right],$$

where $\pi_{i,j}$ is the probability of selecting both instances i and j, $\boldsymbol{s}_i = \boldsymbol{s}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})$ is the column vector defined by $\boldsymbol{s}_i = \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)$ (i.e. the score), and $\boldsymbol{H}(\boldsymbol{\theta})$ is the Hessian matrix of the log-likelihood $\ell_0(\boldsymbol{\theta})$ given in (3). Hence, the WMLE $\hat{\boldsymbol{\theta}}_\pi$ may be regarded as an estimator of the finite population parameter $\boldsymbol{\theta}_0$, i.e. the maximum likelihood estimator we would have obtained if the entire database $\mathcal{D}$ had been annotated.

The WMLE (1) may be obtained by standard software routines by supplying the sampling weights to the estimation procedure, e.g. using the `weights` option in the `glm` function in the R language for statistical computing [25]. Obtaining appropriate standard errors of the estimates does, however, require software routines specialized for inference from complex samples. This is available e.g. through the `svyglm` function in the `survey` package in R [26]–[28]. Formulas for variance estimation may also be found in e.g. [29, Chapter 6.5].

We point out that the properties of $\hat{\boldsymbol{\theta}}_\pi$ given above are stated with respect to the sampling mechanism, taking the database $\mathcal{D}$ as fixed. The additional uncertainty arising from the random process generating the initial database may be accounted for by adding a term $-\boldsymbol{H}(\boldsymbol{\theta})^{-1}$ to the covariance matrix (4), which is the usual covariance matrix of the maximum likelihood estimator $\boldsymbol{\theta}_0$ [29]. Since we are considering the problem of sample selection from a specific database we will ignore this term in the remaining part of the paper, as it is unaffected by the subsampling procedure.

For the special case of Poisson sampling, each instance $i \in \mathcal{D}$ is sampled independently, leading to $\pi_{i,j} = \pi_i\pi_j$ and a simplification of the covariance matrix (4) of the WMLE to

$$\boldsymbol{H}^{-1}\left(\sum_{i\in\mathcal{D}} \frac{1-\pi_i}{\pi_i}\boldsymbol{s}_i\boldsymbol{s}_i^T\right)\boldsymbol{H}^{-1}.$$

This simplification allows obtaining a closed form solution to the optimal choices of π_i for certain optimality criteria, as is detailed in the next section. Note that we, from now on, write the Hessian (5) of the log-likelihood (3) as $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\theta})$, leaving the dependence on the parameter $\boldsymbol{\theta}$ implicit to simplify the notation.

IV. Optimal Sampling Schemes

We will now describe how sample selection in NDS with the use of Poisson sampling may be optimized for a class of optimality criteria known as linear optimality criteria, which aims to minimize the average variance of a collection of linear combinations of the parameter $\boldsymbol{\theta}$. The motivation for this particular choice of optimality criterion is threefold: first, it is a natural optimization criterion in many studies where the individual or simultaneous effect(s) of one or multiple covariates are of primary interest; second, it leads, when considered together with Poisson sampling, to an optimization problem that is numerically tractable with a simple closed-form solution for the optimal choice of sampling probabilities; third, as we will show, it may be used as a building-block for more complex non-linear optimization criteria.

We start with a single linear combination (c-optimality) and continue with the general case with multiple linear combinations (L-optimality). This includes, as a special case, minimizing the average variance of the parameters (A-optimality). We then show how this may be extended to optimization with respect to non-linear optimality criteria, such as to minimize prediction variance (V-optimality) [30].

A. Linear Optimality Criteria

Consider first a linear combination of the model parameters of a regression model $\boldsymbol{a}^T\boldsymbol{\theta} = a_1\theta_1 + a_2\theta_2 + \ldots a_p\theta_p$, where $\boldsymbol{\theta}$ is the parameter vector and $\boldsymbol{a}$ is a column vector of linear coefficients. Such a linear combination may represent the effect of a single covariate, or the effect associated with a simultaneous change in multiple covariates. Conditionally on $\mathcal{D}$, i.e. considering the variation due to subsampling from the database $\mathcal{D}$, the variance of the WMLE of such a linear combination is given by

$$\operatorname{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_\pi|\mathcal{D}) = \boldsymbol{a}^T\operatorname{Var}(\hat{\boldsymbol{\theta}}_\pi|\mathcal{D})\boldsymbol{a}\,, \tag{6}$$

which for Poisson sampling becomes

$$\begin{aligned}\operatorname{Var}(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}_\pi|\mathcal{D}) &= \boldsymbol{a}^T\boldsymbol{H}^{-1}\left(\sum_{i\in\mathcal{D}}\frac{1-\pi_i}{\pi_i}\boldsymbol{s}_i\boldsymbol{s}_i^T\right)\boldsymbol{H}^{-1}\boldsymbol{a}\\ &= \sum_{i\in\mathcal{D}}\frac{c_i}{\pi_i}+k\,,\end{aligned}$$

where

$$c_i = (\boldsymbol{a}^T\boldsymbol{H}^{-1}\boldsymbol{s}_i)^2 \tag{7}$$

and k is a constant not depending on the π_i's. Thus, the optimal sampling scheme in terms of minimizing the variance (6) is obtained by choosing

$$\pi_i \propto \sqrt{c_i}\,, \tag{8}$$

normalized so that $\sum_{i\in\mathcal{D}}\pi_i$ equals the desired sample size (Proposition 1, Appendix B). As this may result in sampling probabilities greater than one, a simple adjustment described in Algorithm 1 in Appendix A may be necessary. The optimality of the sampling scheme after this modification is governed by Proposition 2 in Appendix B.

More generally, we may consider a collection of parameter combinations captured by an $r \times p$ matrix $\boldsymbol{L}$, where each row $\boldsymbol{a}_k^T$ of $\boldsymbol{L}$ defines a linear combination as described above. Thus, the matrix $\boldsymbol{L}$ may be defined to capture several relevant evaluations and comparisons of interest. Using the total variance of the linear combinations specified by the matrix $\boldsymbol{L}$ as optimality criterion, the result in Equation (7) generalizes to:

$$\begin{aligned}c_i &= \boldsymbol{v}_i^T\boldsymbol{v}_i\,,\\ \boldsymbol{v}_i &= \boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{s}_i\,.\end{aligned} \tag{9}$$

The special case where $\boldsymbol{L}$ is the $p \times p$ identity matrix corresponds to minimizing the average variance of the parameters in the vector $\hat{\boldsymbol{\theta}}_\pi$, commonly referred to as A-optimality [30].

B. Non-Linear Optimality Criteria

The results of the previous section can also be applied to optimization with respect to certain classes of non-linear optimality criteria where the optimization criterion can be expressed in terms of a differentiable function $h(\boldsymbol{\theta})$. For instance, considering a logistic regression model, we may optimize the sample selection with respect to the variance of estimators of absolute risks and smooth functions of those, rather than the estimators of log-odds ratios, as would otherwise commonly be the case. To see this, we have, by the use of the Delta method [31], that the variance of $h(\hat{\boldsymbol{\theta}}_\pi)$ may be approximated by

$$\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta})^T\operatorname{Var}(\hat{\boldsymbol{\theta}}_\pi)\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta})\,,$$

provided that $\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \neq \boldsymbol{0}$. Hence, minimizing the average variance of r such functions $h_1(\hat{\boldsymbol{\theta}}_\pi), \ldots, h_r(\hat{\boldsymbol{\theta}}_\pi)$ translates into a linear optimality criterion discussed above with $\boldsymbol{L}$ to be a matrix with rows equal to $\nabla_{\boldsymbol{\theta}}h_k(\boldsymbol{\theta})^T$.

C. Maximizing the Expected Log-Likelihood

Another important example of a non-linear optimality criterion is obtained when the linear coefficient matrix $\boldsymbol{L}$ is taken as $\boldsymbol{L} = \boldsymbol{H}^{1/2}$, where $\boldsymbol{H}^{1/2}$ is a square root of the matrix $\boldsymbol{H}$ such that $\boldsymbol{H}^{1/2}\boldsymbol{H}^{1/2} = \boldsymbol{H}$, leading to a simplification of (9) to

$$c_i = \boldsymbol{s}_i^T\boldsymbol{H}^{-1}\boldsymbol{s}_i\,. \tag{10}$$

As we show in [32], the resulting sampling scheme satisfies the optimality criterion

$$\max_{\boldsymbol{\pi}}\operatorname{E}\left[\ell_0(\hat{\boldsymbol{\theta}}_\pi)\right], \tag{11}$$

with expectation taken with respect to the sampling mechanism. In words, this means that the sampling scheme derived from (10) (using Algorithm 1 in Appendix A) optimizes the generalization performance of the estimator $\hat{\boldsymbol{\theta}}_\pi$ in the sense of maximizing, in expectation, the total log-likelihood (3).[1] Compared to the other optimization criteria discussed above, the criterion (11) has the advantage of not requiring explicit specification of the linear coefficient matrix $\boldsymbol{L}$; instead, it is specified implicitly with respect to the geometry of the model space. The corresponding optimal sampling scheme is also invariant to linear transformations and non-singular re-codings of the design matrix, and implicitly accounts for the relevance of the variables in terms of their anticipated contribution to the log-likelihood.

D. Using Auxiliary Information

A practical complication in optimal design theory is the fact that the optimal design typically depends on unknown quantities, such as the actual value of the parameter $\boldsymbol{\theta}$. In particular, the optimal design does in our case depend on the Hessian $\boldsymbol{H}$ and score vectors $\boldsymbol{s}_i$, which in turn depend on the parameter $\boldsymbol{\theta}$, outcomes y_i and explanatory variables $\boldsymbol{x}_i$, some of which are unknown. Consequently, the optimal sampling scheme can not be evaluated, and we must resort to approximations. In NDS, the availability of auxiliary information in the form of automatically measured variables provides an opportunity to derive such an approximation by minimizing the expected variance under an assisting auxiliary model for the distribution of the unknowns.

Formally, let Z denote a collection of auxiliary variables that are automatically measured and thus readily available for all instances in the database, and $g(y, \boldsymbol{x}|z)$ denote an auxiliary model for the conditional distribution of the response Y and explanatory variables X given the auxiliary variables Z. Considering, as before, a linear coefficient matrix $\boldsymbol{L}$, the expected variance of the linear combinations specified by $\boldsymbol{L\theta}$ is minimized by sampling with probability proportional to

[1] This follows from Proposition 2 in [32] by taking the negative log-density $-\log f_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ as loss function.

$\sqrt{\mathrm{E}[c_i]}$, i.e. replacing c_i in (7) - (10) and Algorithm 1 by its expectation under the auxiliary model $g(y, \boldsymbol{x}|z)$.

In general, this expectation can not be obtained analytically and numerical methods, such as Monte Carlo integration [33], will have to be employed. An algorithmic description of such a procedure is provided by Algorithm 2 in Appendix A. We also present a simplified and computationally less demanding version in Algorithm 3 in Appendix A, which additionally requires a parameter guess $\boldsymbol{\theta}^*$ and Hessian matrix $\boldsymbol{H}^*$ as input. The auxiliary model $g(y, \boldsymbol{x}|z)$, parameter guess $\boldsymbol{\theta}^*$ and Hessian $\boldsymbol{H}^*$ may be obtained using e.g. a small pilot study, prior knowledge, existing data and simulations.

V. Motivating Example, Continued

A. Optimization Criteria

Re-visiting the example introduced in Section II, we consider a logistic regression model for the risk of an SCE given by

$$\begin{aligned}\text{logit } P(Y=1|X) = \; & \theta_0 + \theta_1 \textit{Time gap} + \theta_2 \textit{Speed} \\ & + \theta_3 \textit{Glance} + \theta_4 \textit{Glance} \times \textit{Time gap}\,, \end{aligned} \tag{12}$$

where Y is a binary indicator of the SCE, *Time gap* is the distance between the vehicles measured in seconds, *Speed* is the speed of the index car and *Glance* is a binary indicator whether the driver is having eyes-off-road at brake light. To illustrate the proposed sampling procedure, we will consider four linear optimization criteria directed towards estimating the effects of time gap, vehicle speed and glancing, and a high-low risk contrast involving all parameters, as further described below.

i) *Time gap.* Say that we are primarily interested in the regression coefficient corresponding to time gap when having the eyes on road. Explicitly, we are interested in minimizing the variance of an estimator of θ_1. In this case, $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)^T$ and the linear combination of interest consists of one single parameter, implying that $\boldsymbol{a} = (0, 1, 0, 0, 0)^T$.

ii) *Vehicle speed.* Alternatively, we might be interested in the effect of speed, i.e. in the parameter θ_2, corresponding to a linear combination determined by the coefficient vector $\boldsymbol{a} = (0, 0, 1, 0, 0)^T$.

iii) *Glancing.* We may also be interested in the effect of glancing at a certain time gap to the front vehicle, say at 1, 2 and 3 s. time gap. This is described by the parameter combinations $\theta_3 + \theta_4$, $\theta_3 + 2\theta_4$ and $\theta_3 + 3\theta_4$, corresponding to the coefficient vectors $\boldsymbol{a}_1 = (0, 0, 0, 1, 1)^T$, $\boldsymbol{a}_2 = (0, 0, 0, 1, 2)^T$, $\boldsymbol{a}_3 = (0, 0, 0, 1, 3)^T$ and the coefficient matrix

$$\boldsymbol{L} = \begin{pmatrix} 0, 0, 0, 1, 1 \\ 0, 0, 0, 1, 2 \\ 0, 0, 0, 1, 3 \end{pmatrix}.$$

iv) *High-low risk contrast.* As a final example of a linear optimality criterion, we consider a contrast between a hypothetical high risk and low risk scenario, defining the high risk scenario as glancing off road when driving at 70 km/h and 1 s. time gap, and the low risk scenario as having eyes-on-road when driving at 30 km/h and 3 s. time gap. The parameter combination corresponding to the high risk scenario is given by $\theta_0 + \theta_1 + 70\theta_2 + \theta_3 + \theta_4$ and a coefficient vector $\boldsymbol{a}_{high} = (1, 1, 70, 1, 1)^T$. Similarly, the low risk scenario may be described by $\boldsymbol{a}_{low} = (1, 3, 30, 0, 0)^T$. The contrast between the two is thus described by the linear combination $-2\theta_1 + 40\theta_2 + \theta_3 + \theta_4$, and we may take $\boldsymbol{a}$ as $\boldsymbol{a}_{high} - \boldsymbol{a}_{low} = (0, -2, 40, 1, 1)^T$.

As an example of a non-linear optimality criterion we also consider the optimality criterion (11) introduced in Section IV-C:

v) *Maximizing the expected log-likelihood.* This does not require an explicit specification of the linear coefficient matrix $\boldsymbol{L}$, but simply amounts to replacing c_i in (9) by (10) in the optimization.

B. Auxiliary Information

Recall that there are three variables present in the example model that require annotation: time gap, vehicle speed and driver glancing behavior (eyes on/off road), at brake light. Information about the first two can be obtained by automatic measurements of vehicle data. Information of the latter would ideally be obtained by automatic extraction of relevant signals from the video sequences. Lacking such information, we proceeded using automatic measurements of vehicle data also to predict glancing. We used data from the 49 a priori annotated SCEs included in this study to derive auxiliary models, pretending, in order to mimic a real-world scenario, that the corresponding information for the controls was unavailable at this stage.

Proxies for vehicle speed and time gap at brake light were obtained as follows. Based on the deceleration profiles of the annotated SCEs, a proxy for time of brake light onset was first identified (Figure 1). The speed and time gap at the predicted timepoint for brake light were consequently used as proxies for the corresponding variables at brake light. To identify auxiliary variables for glancing, we employed a logistic regression model with eyes on/off road as a response and used a stepwise search for predictors among the following automatically measured variables: vehicle speed (km/h), time gap (s), acceleration of vehicle ahead (m/s^2), a binary indicator whether the driver of the index car is braking, and time to collision (s), defined as the expected time for the index car to collide with the front car if they remain on the same path and at the same speeds. This procedure resulted in deceleration of vehicle ahead as the sole predictor of glancing behavior.

After auxiliary variables had been identified, we used data from the annotated SCEs to estimate auxiliary models, using univariate linear regression for time gap and vehicle speed, yielding a coefficient of determination (R^2) of 0.83 and 0.88, and univariate logistic regression for glancing away from road (Figure 2). The predictive performance of the latter was, however, rather weak. Indeed, the auxiliary model predicted a higher probability of having eyes-on-road with increasing deceleration of the lead vehicle, but no such trend was actually observed among the controls (Figure 2).

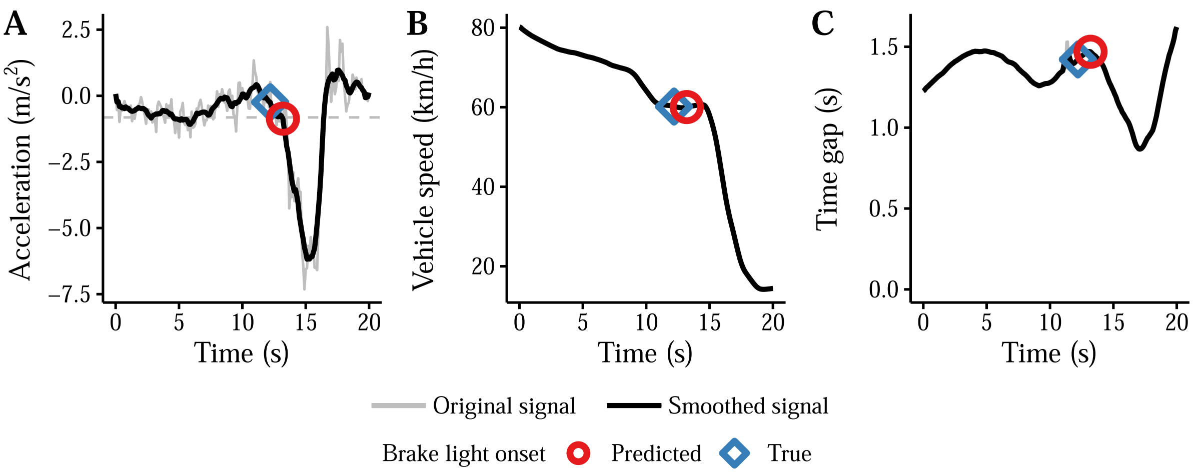


Fig. 1. Extracting auxiliary variables from continuous measurements of the acceleration of the front vehicle, vehicle speed and time gap between vehicles. The last time point prior to the point of maximal deceleration where the deceleration exceeded 0.8 m/s^2 was used to predict brake light onset, as derived from 49 annotated cases. Vehicle speed and time gap at the predicted time point for brake light were consequently used as proxies for the corresponding variables at brake light. To reduce noise, the signals were smoothed using a moving average.

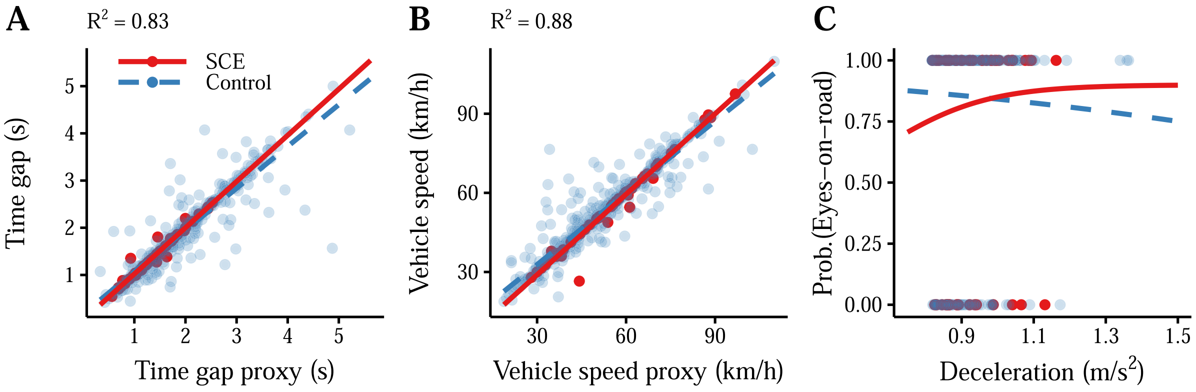


Fig. 2. Fitted mean trends from the auxiliary models for time gap (A), vehicle speed (B) and glancing (C) using univariate linear regression (A, B) and univariate logistic regression (C). The auxiliary models were derived from 49 a priori annotated SCEs (red solid lines). The blue dashed lines show the actual association observed among the controls. Since we anticipated a higher proportion of off-road glances among the SCEs than among the controls, the intercept of the logistic regression model for glancing was modified to predict 20% off-road glances among the controls, based on estimates of off-road glancing during normal driving in [34], [35].

Finally, a guess of the value of the parameter $\boldsymbol{\theta}$ was obtained by generating 100 complete datasets by stochastic simulation of covariate vectors $\boldsymbol{x}_i^*$ for the 500 control candidates included in this study, using the auxiliary models for vehicle speed, time gap and glancing. For each of the simulated data sets, the parameters of the logistic regression model (12) were estimated and the mean of these estimates was used as a guess $\boldsymbol{\theta}^*$ of the value of the target parameter $\boldsymbol{\theta}$. Similarly, we used the mean of the corresponding covariance matrices to estimate $-\boldsymbol{H}^{-1}$.

C. Optimal Sampling Schemes

We used, next, the auxiliary models to compute optimal control sampling schemes with respect to the effect of time gap when having eyes-on-road, glancing away at 1, 2, and 3 s. time gap, vehicle speed, and a high vs. low risk contrast, as detailed in Section V-A. We also implemented the optimal sampling scheme with respect to the optimality criterion (11), i.e. maximizing the expected log-likelihood. The optimization was implemented according to Algorithm 3 in Appendix A for selection of 100 out of 500 controls.

Optimal control sampling schemes for the four linear optimization criteria discussed above are illustrated in Figure 3. As observed in this figure, there are substantial variations between the sampling schemes, depending on the linear combination of interest. The sampling probabilities depend on the expected values of the covariates of interest, and, for linear combinations involving multiple parameters, also on their anticipated correlations, and further on the anticipated risk of SCE. Generally, controls at high anticipated risk of SCE should be oversampled, i.e driving at high speed, small time gap, and with a high predicted tendency of glancing. Additionally, relatively large sampling probabilities are assigned to controls at moderate to mild risk, constituting a subset to which the characteristics of the cases and high risk controls may be contrasted. Controls with low anticipated risk of an SCE tend to be selected with low probability, as these contribute with little information with regards to safety.

The sampling scheme optimized to maximize the expected log-likelihood is illustrated in Figure 4. Without explicitly specifying the linear coefficient matrix $\boldsymbol{L}$, this sampling scheme assigns sampling probabilities proportional to the

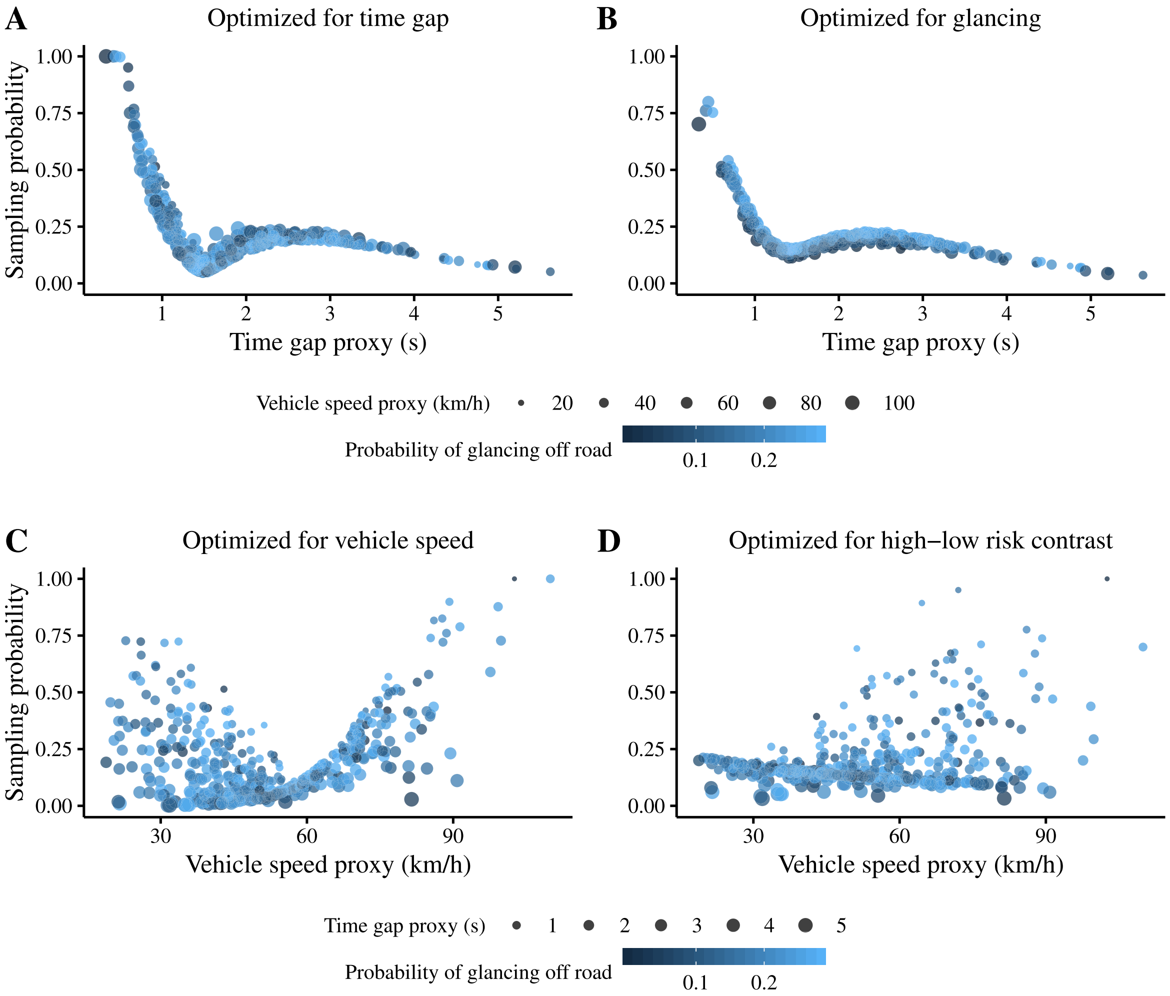


Fig. 3. Optimal control sampling schemes for selection of 100 out of 500 controls, optimized for estimating the effect of time gap when having eyes-on-road (A), glancing off road at 1, 2, and 3 s. time gap (B), vehicle speed (C), and a contrast between a hypothetical high risk and low risk scenario (D). The high risk scenario is defined as glancing off road when driving at 70 km/h and 1 s. time gap, and the low risk scenario as having eyes-on-road when driving at 30 km/h and 3 s. time gap.

relative importance of the instances in terms of the anticipated contribution to the log-likelihood. The sampling scheme is to a greater extent determined by time gap than vehicle speed, demonstrating a greater importance of the former in explaining the risk of an SCE, according to the auxiliary models and parameter guess $\boldsymbol{\theta}^*$.

VI. Empirical Evaluation: Efficiency of the Sampling Schemes

To evaluate the performance of the presented optimization and analysis procedure, using auxiliary information for instance selection followed by a correspondingly weighted analysis, we conducted an empirical evaluation by repeated subsampling from the cohort of 49 cases and 500 control candidates, as further described below.

A. Methods

Sample scheme optimization was performed according to the optimality criteria described in Section V-A and V-C, using the auxiliary models and parameter guess from Section V-B. For each of the optimization criteria, optimal control sampling schemes were computed according to Algorithm 3 in Appendix A, and a sample of controls was selected accordingly, using Poisson sampling. Thus, cases were selected with probability 1. A weighted analysis was then performed, and the estimated parameter vector was stored. The procedure was repeated 10^4 times for control samples of expected size $n = 50, 100, 150$ and 200. For each sample size, the standard deviations (SD) of the estimated parameters and linear combinations of interest were calculated and stored. For targets including multiple parameter combinations, the square root of the average variance was instead computed. Two analyses were conducted for comparison: a complete information analysis using the entire study cohort, and an analysis based on a sub-sample chosen using simple random sampling of $n = 50, 100, 150$ and 200 controls. Ordinary non-weighted logistic regression was used for both of these approaches, as commonly is done in logistic regression analysis of case-control studies [36], [37].

B. Results

1) Complete Information Analysis: The result of a logistic regression analysis of the full study cohort, after exclusion of non-relevant controls (n = 14) and controls with missing data

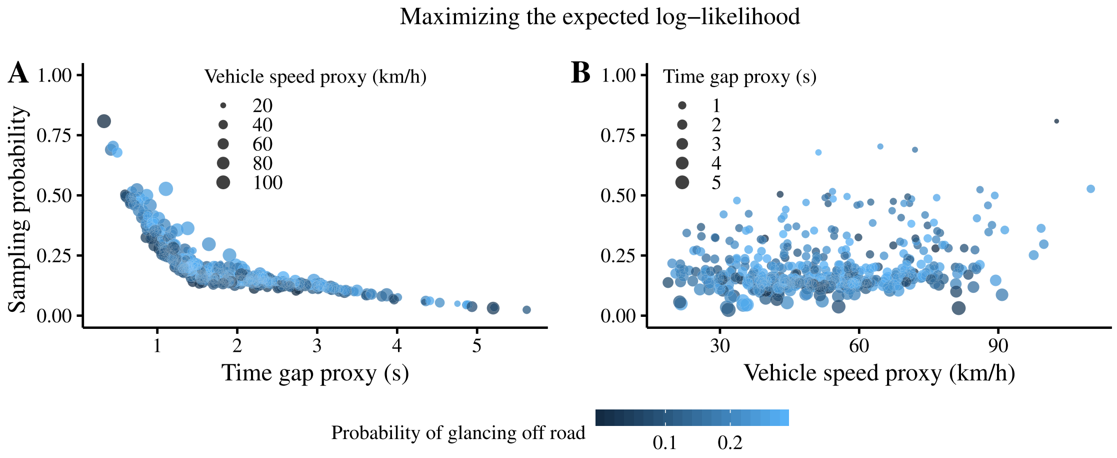


Fig. 4. Optimal control sampling scheme for selection of 100 out of 500 controls vs time gap (A) and vehicle speed (B), optimized according to the optimality criterion (11), i.e. to maximize the expected log-likelihood.

TABLE I

EFFECT OF VEHICLE SPEED, TIME GAP AND GLANCING BEHAVIOR ON THE PROBABILITY OF OOPS-REACTION, USING LOGISTIC REGRESSION ANALYSIS ON THE FULL STUDY COHORT

Variable	At	Log-odds ratio (SE)	OR (95% CI)
Vehicle speed, per 10 km/h		-0.026 (0.100)	0.97 (0.80 to 1.19)
Time gap, per 1 s decrease	Eyes-on-road	0.771 (0.324)	2.16 (1.14 to 4.08)
Time gap, per 1 s decrease	Glancing off road	1.495 (0.531)	4.46 (1.58 to 12.62)
Glancing	Time gap 1 s	2.096 (0.469)	8.13 (3.25 to 20.39)
Glancing	Time gap 2 s	1.372 (0.446)	3.94 (1.64 to 9.45)
Glancing	Time gap 3 s	0.648 (0.967)	1.91 (0.29 to 12.73)

CI, confidence interval; OR, odds ratio; SE, standard error.

TABLE II

EFFICIENCY OF OPTIMIZED CONTROL SAMPLING SCHEMES USING POISSON SAMPLING, COMPARED TO SIMPLE RANDOM SAMPLING (SRS) AND COMPLETE INFORMATION ANALYSIS OF THE FULL STUDY COHORT OF 49 CASES AND 500 CONTROLS

			Ratio of standard deviations vs. simple random sampling				Ratio of standard deviations vs. complete information analysis[f]			
n[a]	Control sample	Optimized for parameter	Vehicle speed	Time gap[b]	Glancing[c,e]	High-Low risk[d]	Vehicle speed	Time gap[b]	Glancing[c,e]	High-Low risk[d]
50	SRS		1.00	1.00	1.00	1.00	2.16	1.84	1.92	3.00
	Poisson	Vehicle speed	0.97	1.54	1.78	1.17	2.12	2.30	2.64	3.34
		Time gap[b]	1.34	0.86	1.77	1.60	2.56	1.73	2.63	4.19
		Glancing[c]	1.29	0.89	1.66	1.53	2.50	1.75	2.53	4.05
		High-Low risk[d]	1.19	1.02	1.93	1.46	2.39	1.86	2.78	3.92
		Maximize expected log-likelihood	1.11	0.97	1.77	1.35	2.29	1.82	2.63	3.70
100	SRS		1.00	1.00	1.00	1.00	1.74	1.54	1.45	2.02
	Poisson	Vehicle speed	0.89	1.51	1.93	1.17	1.65	1.82	1.86	2.19
		Time gap[b]	1.32	0.83	1.34	1.29	1.97	1.45	1.60	2.31
		Glancing[c]	1.28	0.87	1.39	1.35	1.94	1.47	1.62	2.37
		High-Low risk[d]	1.14	0.94	1.33	0.99	1.84	1.51	1.59	2.01
		Maximize expected log-likelihood	1.05	0.93	1.27	1.02	1.77	1.50	1.57	2.03
150	SRS		1.00	1.00	1.00	1.00	1.57	1.41	1.33	1.73
	Poisson	Vehicle speed	0.80	1.52	1.93	1.23	1.46	1.63	1.63	1.89
		Time gap[b]	1.26	0.79	1.21	1.18	1.72	1.33	1.40	1.86
		Glancing[c]	1.23	0.82	1.22	1.20	1.70	1.34	1.40	1.87
		High-Low risk[d]	1.11	0.92	1.19	0.94	1.63	1.38	1.39	1.68
		Maximize expected log-likelihood	0.96	0.88	1.14	0.93	1.55	1.36	1.38	1.68
200	SRS		1.00	1.00	1.00	1.00	1.47	1.33	1.26	1.58
	Poisson	Vehicle speed	0.72	1.56	1.88	1.23	1.34	1.51	1.48	1.71
		Time gap[b]	1.20	0.76	1.16	1.12	1.56	1.25	1.30	1.64
		Glancing[c]	1.20	0.80	1.17	1.13	1.56	1.26	1.30	1.65
		High-Low risk[d]	1.03	0.91	1.13	0.87	1.48	1.30	1.29	1.50
		Maximize expected log-likelihood	0.90	0.87	1.10	0.89	1.42	1.28	1.28	1.51

[a] n is the expected size of the control sample.

[b] Effect of time gap when having eyes-on-road.

[c] Effect of glancing off road at 1, 2 and 3 s. time gap.

[d] Contrast between glancing off road at 70 km/h and 1 s. time gap vs. having eyes-on-road at 30 km/h and 3 s. time gap.

[e] The standard deviation is calculated as the square root of the average variance of the specified linear combinations.

[f] 65 controls were excluded due to missing data (no video, poor video quality or external factors hindering video annotation) or considered irrelevant for the event of interest after video review.

due to no video or poor video quality (n = 51), is presented in Table I. There was a significant increase in the risk of an oops-reaction when glancing off road, more severely so at small time gap to the front car. Reduced time gap to the front vehicle was also associated with increased risk of SCE, and the risk increased at faster rate when having eyes-off-road. On the other hand, increased speed alone was not significantly associated with an increased risk of an SCE.

2) Subsampling Study: A comparison of the sampling variability in the estimated parameters between different control sampling procedures is presented in Table II. Poisson sampling optimized for a specific linear combination of parameters generally resulted in increased precision of the corresponding parameter estimates, as compared to simple random sampling, the gain in precision increasing with the size of the control sample. With $n = 50$ controls, the standard deviation (SD) of the estimator for the effect of time gap was reduced by 14% using instance selection optimized for this particular parameter, compared to simple random sampling. The corresponding precision loss, measured as increase in SD compared to analyzing the entire database, was 73%. At $n = 200$, the results were improved further to an SD reduction of 24% compared to simple random sampling (SRS). In this case, using 40% of the database resulted in only 25% loss of precision. Similar results were observed for optimization with respect to the effect of vehicle speed.

For the high-low risk contrast, improvement compared to SRS was observed first at $n = 150$ controls, yielding an SD reduction of 6%, which was further improved to 13% reduction when $n = 200$. The effect of glancing, on the hand, was poorly estimated with all sampling schemes, particularly at small sample sizes. In a sensitivity analysis where we artificially created a new proxy for glancing, we found that explaining only 10% of the variability in glancing behavior would suffice to achieve a performance equal to simple random sampling. With a further increase to explaining 20% of the variability in glancing, an SD reduction of more than 20% was observed (data not shown).

Optimization with respect to one parameter combination sometimes resulted in loss of precision for the other parameters: at n = 100 controls, the greatest loss was an SD increase of 93% compared to SRS, as was observed for the estimating the effect of glancing when the sampling scheme was optimized with respect to the effect of vehicle speed (Table II). In contrast to the linear optimality criteria, optimization with respect to the expected log-likelihood generally performed well with respect to all parameters, producing a simultaneous SD reduction of approximately 10% for vehicle speed, time gap and the high-low risk contrast when $n = 200$.

VII. Conclusion

A. Summary of Main Results

We have presented an inferential framework for analysis of large databases in which complete data annotation is costly, and shown how instance selection in naturalistic driving data may be optimized by use of auxiliary information readily available for all instances in an NDS database. We have furthermore illustrated through a case study how such sampling designs may be implemented in practice, and demonstrated that a substantial gain in statistical efficiency may be achieved. Specifically, we were able to achieve almost 50% variance reduction in estimating the effect of vehicle speed and time gap when optimizing for the corresponding parameters, and up to 20% simultaneous variance reduction in all parameters except glancing when optimizing with respect to the expected log-likelihood.

B. Explanations and Interpretations

For a successful implementation of the analysis pipeline, the availability of auxiliary information and proxies for the study variables on interest that require annotation is crucial. In our case study, vehicle speed and time gap at brake light were well approximated by the corresponding automatically recorded signals at the predicted timepoint for brake initiation. Consequently, optimization with respect to the corresponding model parameters resulted in substantial increase in precision. Driver glancing behavior, on the other hand, was poorly predicted by the automatically measured variables available in this study. In this case, non-uniform instance selection actually resulted in loss of precision. This may partially be explained by the loss of optimality in the optimization when no or little auxiliary information is available, partially by increased variability when using a random size design such as Poisson sampling, and partially by increased variability when using a weighted estimator, as compared to ordinary non-weighted logistic regression. Nevertheless, a sensitivity analysis revealed that an auxiliary model explaining only 10-20% of the variability in the variables of interest may be sufficient to guarantee performance on par with that of simple random sampling, when the model also included some variables for which good auxiliary information was available.

Although the focus of this paper is on estimation uncertainty in terms of variance, all results could equivalently have been stated in terms of mean squared error since the bias of our estimator, seen as an estimator of the maximum likelihood estimate we would have obtained if the entire database had been annotated, vanishes at a faster rate than the variance as the size of the annotated subsample increases [24].

C. Limitations and Directions of Future Research

Due to the need of good auxiliary information when implementing the optimization procedure, effort should be made to find good proxies for the variables of interest. Finding good auxiliary variables that normally require video annotation remains a challenging task. Attempts to develop algorithms for automated detection of driver glancing behavior from video sequences have been made [19], [38]–[42]. Further development and application of such algorithms could increase the benefit of optimized sample selection with respect to analysis of driver behavior, as predictions of such algorithms could be used as auxiliary information for driving tasks. Similar results to those obtained for time gap and vehicle speed could then potentially be achieved also for estimating the effect of glancing and other driver tasks.

Another direction of improvement could be to collect data sequentially, as outlined in [43]. The auxiliary models may then be derived from a pilot sample and updated as more data is collected, thus reducing the need of prior knowledge.

For case-control analyses using logistic regression, further variance reduction of weighted estimators may also be achieved by re-scaling the weights in the control sample, see e.g. [44]–[48]. Implementation of such procedures for NDS is a possible direction of future research.

In practice, both the implementation and performance of the optimization procedure may be affected by data quality issues such as data losses in kinematic measurements and video derived features. Nonetheless, we were able to demonstrate substantial improvements to standard sampling procedures in a small but realistic case study.

D. Choosing an Optimization Criterion

We have focused in this paper on linear optimality criteria, and shown how these results may be extended also to smooth non-linear functions of the parameter of interest. Comparing the results obtained by the linear optimization criteria and optimization with respect to the expected log-likelihood, the former achieved a better precision in the particular linear combination of interest, but the latter performed better with respect to multiple parameters and thus serves as a better general-purpose criterion. While the restriction to linear optimality criteria offers a numerically tractable solution to the optimal sample selection problem, it would be an interesting topic of further research to investigate whether similar procedures could be developed also for other classes of optimality criteria such as D and E-optimality [30].

E. Implications

In our empirical study, we found an SD reduction of 10-30% compared to simple random sampling when optimizing the sampling procedure for a particular parameter combination, and a simultaneous reduction of 10% in multiple parameters when optimizing for the expected log-likelihood. Translating the gain in efficiency in terms of SD reduction into power to detect possible associations between the variables of interest, a reduction in the standard deviation of an estimator by 10% roughly corresponds to a sample size reduction of $1 - 0.9^2 \approx 20\%$ [49, Chapter 9.2.4]. Similarly, with 30% SD reduction, the sample size could be reduced by $\approx 50\%$ without loss of power. Thus, the use of optimized instance selection implies that fewer instances need to be annotated, as compared to simple random sampling, potentially reducing the annotation demands by as much as 50%. Considering the high cost associated with manual video annotation, and the loss of information induced by having to restrict the analysis to a subset of the collected data, our proposed inferential framework provides a viable approach to reduce the cost of the analysis of naturalistic driving data.

Appendix A
Computation of Optimal Poisson Sampling Schemes

We present in Algorithm 1 a procedure for computation of Poisson sampling schemes that ensures that valid probabilities $0 < \pi_i \leq 1$ are obtained. A proof of its optimality is provided by Proposition 2 in Appendix B. A Monte Carlo procedure to approximate the optimal sampling scheme of Algorithm 1 using auxiliary information is presented in Algorithm 2, and a simplified version, replacing step 5 and 6 in Algorithm 2 by pre-computed estimates or guesses of the parameter $\boldsymbol{\theta}$ and Hessian $\boldsymbol{H}$, is provided in Algorithm 3.

Algorithm 1 Optimal Poisson Sampling Scheme

```
Input: Index set $\mathcal{D}$, coefficients $\{c_i\}_{i\in\mathcal{D}}$, sample size $n$.
Initialization: Let $\mathcal{M}$ be the empty set. Let $m := |\mathcal{M}| = 0$.
for $i \in \mathcal{D}$ do
    Compute $\pi_i^* = n\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}}\sqrt{c_j}}$.
end for
while any $\pi_i^* > 1$ do
    Let $\mathcal{M}$ be the collection of elements in $\mathcal{D}$ with $\pi_i^* \geq 1$.
    Update $m = |\mathcal{M}|$, the number of elements in $\mathcal{M}$.
    Set $\pi_i^* = 1$ for all $i \in \mathcal{M}$.
    For $i \in \mathcal{D} \setminus \mathcal{M}$, update $\pi_i^*$ according to
        $\pi_i^* = (n - m)\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}\setminus\mathcal{M}}\sqrt{c_j}}$ .    (A.1)
end while
Output: Optimal sampling probabilities $\{\pi_i^*\}_{i\in\mathcal{D}}$.
```

Algorithm 2 Optimal Auxiliary Variable Assisted Poisson Sampling Scheme

```
Input: Index set $\mathcal{D}$, auxiliary variables $\{z_i\}_{i\in\mathcal{D}}$, auxiliary
model $g(y, \boldsymbol{x}|\boldsymbol{z})$, linear coefficient matrix $\boldsymbol{L}$, number of
Monte Carlo simulations $M$, sample size $n$.
1:  for $m = 1, \ldots, M$ do
2:      for $i \in \mathcal{D}$ do
3:          Simulate $(y_i^*, \boldsymbol{x}_i^*)$ from $g(y_i, \boldsymbol{x}_i|\boldsymbol{z}_i)$.
4:      end for
5:      Compute parameter guess $\hat{\boldsymbol{\theta}}^*$ as
            $\hat{\boldsymbol{\theta}}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i\in\mathcal{D}} \log f_{\boldsymbol{\theta}}(y_i^*|\boldsymbol{x}_i^*)$ .
6:      Compute Hessian matrix $\boldsymbol{H}^*$ as
            $\boldsymbol{H}^* = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \sum_{i\in\mathcal{D}} \log f_{\boldsymbol{\theta}}(y_i^*|\boldsymbol{x}_i^*)$ ,
        evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^*$.
7:      for $i \in \mathcal{D}$ do
8:          Compute $\boldsymbol{s}_i^* = \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(y_i^*|\boldsymbol{x}_i^*)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^*}$ .
9:          Compute $c_{i,m}^* = \boldsymbol{s}_i^{*T} \boldsymbol{H}^{*-1} \boldsymbol{L}^T \boldsymbol{L} \boldsymbol{H}^{*-1} \boldsymbol{s}_i^*$ .†
10:     end for
11: end for
12: for $i \in \mathcal{D}$ do
13:     Compute $\hat{c}_i = \frac{1}{M}\sum_{m=1}^{M} c_{i,m}^*$.
14: end for
15: for $i \in \mathcal{D}$ do
16:     Compute $\pi_i^*$ using Algorithm 1,
        taking $\mathcal{D}$, $\{\hat{c}_i\}_{i\in\mathcal{D}}$ and $n$ as input.
17: end for
Output: optimal sampling probabilities $\{\pi_i^*\}_{i\in\mathcal{D}}$.
```

† For the optimality criterion (11) in Section IV-C, i.e. maximizing the expected log-likelihood, this expression should be replaced by $c_{i,m}^* = \boldsymbol{s}_i^{*T} \boldsymbol{H}^{*-1} \boldsymbol{s}_i^*$.

Algorithm 3 Optimal Auxiliary Variable Assisted Poisson Sampling Scheme, Simplified

Input: Index set $\mathcal{D}$, auxiliary variables $\{z_i\}_{i\in\mathcal{D}}$, auxiliary model $g(y, \boldsymbol{x}|z)$, parameter guess $\boldsymbol{\theta}^*$, Hessian matrix $\boldsymbol{H}^*$, linear coefficient matrix $\boldsymbol{L}$, number of Monte Carlo simulations M, sample size n.

```
for $i \in \mathcal{D}$ do
  for $m = 1, \ldots, M$ do
    Simulate $(y_i^*, \boldsymbol{x}_i^*)$ from $g(y_i, \boldsymbol{x}_i|z_i)$.
    Compute $\boldsymbol{s}_i^* = \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(y_i^*|\boldsymbol{x}_i^*)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$.
    Compute $c_{i,m}^* = \boldsymbol{s}_i^{*T}\boldsymbol{H}^{*-1}\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{H}^{*-1}\boldsymbol{s}_i^*$.†
  end for
  Compute $\hat{c}_i = \frac{1}{M}\sum_{m=1}^{M} c_{i,m}^*$.
end for
for $i \in \mathcal{D}$ do
  Compute $\pi_i^*$ using Algorithm 1,
  taking $\mathcal{D}$, $\{\hat{c}_i\}_{i\in\mathcal{D}}$ and $n$ as input.
end for
```

Output: optimal sampling probabilities $\{\pi_i^*\}_{i\in\mathcal{D}}$.

† For the optimality criterion (11) in Section IV-C, i.e. maximizing the expected log-likelihood, this expression should be replaced by $c_{i,m}^* = \boldsymbol{s}_i^{*T}\boldsymbol{H}^{*-1}\boldsymbol{s}_i^*$.

APPENDIX B
THEOREMS AND PROOFS

We provide below two propositions with proofs of the optimality of the sampling schemes proposed in Equation (8) (Proposition 1) and Algorithm 1 (Proposition 2).

Proposition 1: Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ *and consider the function*

$$f(\boldsymbol{\pi}) = \sum_{i=1}^{N} \frac{c_i}{\pi_i}, \quad c_i > 0$$

subject to the constraints

$$\sum_{i=1}^{N} \pi_i = n, \qquad \pi_i > 0, \quad i = 1, \ldots, N,$$

for some $n > 0$. *Then,* $f(\boldsymbol{\pi})$ *is minimized by choosing* π_i *according to*

$$\pi_i^* = n\frac{\sqrt{c_i}}{\sum_{j=1}^{N}\sqrt{c_j}}, \quad i = 1, \ldots, N.$$

Proof of Proposition 1: Using the method of Lagrange multipliers [50, Chapter 5], we introduce the auxiliary function

$$\Lambda(\boldsymbol{\pi}, \lambda) = f(\boldsymbol{\pi}) + \lambda h(\boldsymbol{\pi}), \quad h(\boldsymbol{\pi}) = \sum_{i=1}^{N} \pi_i - n.$$

Critical points of the Lagrangian are found by solving the equation system

$$\nabla\Lambda(\boldsymbol{\pi}, \lambda) = \boldsymbol{0} \quad \Leftrightarrow \quad \begin{cases} h(\boldsymbol{\pi}) = 0 \\ -\nabla_{\boldsymbol{\pi}} f(\boldsymbol{\pi}) = \lambda\nabla_{\boldsymbol{\pi}} h(\boldsymbol{\pi}) \end{cases}.$$

Since $\frac{\partial f(\boldsymbol{\pi})}{\partial \pi_i} = -c_i/\pi_i^2$ and $\frac{\partial h(\boldsymbol{\pi}_t)}{\partial \pi_i} = 1$, this implies that $\lambda = c_1/\pi_1^2 = \ldots = c_N/\pi_N^2$, and further that

$$\pi_i \propto \sqrt{c_i}.$$

By the constraints $\pi_i > 0$ and $\sum_{i=1}^{N}\pi_i = n$, we obtain

$$\pi_i = n\frac{\sqrt{c_i}}{\sum_{j=1}^{N}\sqrt{c_j}}. \tag{A.2}$$

Thus, the point $(\boldsymbol{\pi}^*, \lambda^*)$ with entries π_i^* defined according to (A.2) and $\lambda^* = c_1/\pi_1^{*2}$ is a stationary point of $\Lambda(\boldsymbol{\pi}, \lambda)$. Hence, $\boldsymbol{\pi}^*$ is a stationary point of $f(\boldsymbol{\pi})$ under the specified constraints. Furthermore, the Hessian of $f(\boldsymbol{\pi})$ is positive definite on the domain specified by $\pi_i > 0$, so $\boldsymbol{\pi}^*$ is a local minimum. By convexity, this implies that $\boldsymbol{\pi}^*$ is the global minimum of $f(\boldsymbol{\pi})$ under the specified constraints.

Proposition 2: Let $\mathcal{D} = \{1, \ldots, N\}$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ *and consider the function*

$$f(\boldsymbol{\pi}) = \sum_{i=1}^{N} \frac{c_i}{\pi_i}, \quad c_i > 0$$

subject to the constraints

$$\sum_{i=1}^{N} \pi_i = n, \qquad 0 < \pi_i \leq 1, \quad i = 1, \ldots, N,$$

for $0 < n < N$. *Then,* $f(\boldsymbol{\pi})$ *is minimized by choosing* π_i *according to Algorithm 1.*

Proof of Proposition 2: The claim follows immediately from Proposition 1 if

$$n\frac{\sqrt{c_i}}{\sum_{j=1}^{N}\sqrt{c_j}} < 1$$

for all $i = 1, \ldots, N$. Hence, we assume that there exists some index i for which this is not fulfilled.

We note first that Algorithm 1 terminates within a finite number of iterations, since $n < N$. Indeed, a feasible solution is obtained within at most $n-1$ iterations. We show below that the achieved solution satisfies the Karush-Kuhn-Tucker (KKT) conditions [50, Chapter 5.5.3], and that this is sufficient for global optimality in this setting.

Introducing the constraint functions

$$g_i(\boldsymbol{\pi}) = \pi_i - 1, \quad i = 1, \ldots, N$$
$$h(\boldsymbol{\pi}) = \sum_{i=1}^{N} \pi_i - n,$$

we may formulate the constrained optimization problem of Proposition 2 as

$$\begin{aligned} &\min_{\boldsymbol{\pi}} \ f(\boldsymbol{\pi}) \\ &\text{where} \ \ f(\boldsymbol{\pi}) = \sum_{i=1}^{N} \frac{c_i}{\pi_i}, \quad c_i > 0 \\ &\text{subject to} \ \ \pi_i > 0, \qquad i = 1, \ldots, N \\ &\qquad\qquad\quad g_i(\boldsymbol{\pi}) \leq 0, \quad i = 1, \ldots, N \\ &\qquad\qquad\quad h(\boldsymbol{\pi}) = 0 \end{aligned}$$

and introduce the Lagrangian

$$\Lambda(\boldsymbol{\pi}, \boldsymbol{\mu}, \lambda) = f(\boldsymbol{\pi}) + \sum_{i=1}^{N} \mu_i g_i(\boldsymbol{\pi}) + \lambda h(\boldsymbol{\pi}) , \tag{A.3}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ and λ are the Lagrange multipliers.

Let us assume, without loss of generality, that the elements are ordered so that $c_1 \geq c_2 \geq \ldots \geq c_N$, and let $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_N^*)$ denote the sampling scheme obtained by Algorithm 1. Thus, we have for the first m instances that $\pi_1^* = \ldots \pi_m^* = 1$, and for the remaining $N - m$ instances that $\pi_i^* < 1$. Taking

$$\begin{aligned} \lambda^* &= \frac{c_N}{\pi_N^{*2}} , \\ \mu_i^* &= \begin{cases} c_i - \lambda^* & i = 1, \ldots, m , \\ 0 & i = m+1, \ldots, N , \end{cases} \end{aligned} \tag{A.4}$$

we show that the point $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \lambda^*)$ satisfies the KKT conditions:

- *Stationarity*
  Noting that $c_j/\pi_j^{*2} = c_N/\pi_N^{*2} = \lambda^*$ for all $j > m$, we see that

$$\begin{aligned} -\nabla f(\boldsymbol{\pi})\big|_{\boldsymbol{\pi}=\boldsymbol{\pi}^*} &= -\left(-\frac{c_1}{\pi_i^{*2}}, \ldots, -\frac{c_N}{\pi_N^{*2}}\right) \\ &= \left(c_1, \ldots, c_m, \frac{c_{m+1}}{\pi_{m+1}^{*2}} \ldots, \frac{c_N}{\pi_N^{*2}}\right) \\ &= \left(\mu_1^* + \lambda^*, \ldots, \mu_N^* + \lambda^*\right) \\ &= \boldsymbol{\mu}^* + (\lambda^*, \ldots, \lambda^*) \\ &= \sum_{i=1}^{N} \mu_i^* \nabla g_i(\boldsymbol{\pi}^*) + \lambda^* \nabla h(\boldsymbol{\pi}^*) . \end{aligned}$$

  Thus, $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \lambda^*)$ is a stationary point of (A.3).
- *Primal feasibility*
  By inspection of the algorithm, we see that it produces a solution with
  i) $\sum_{i=1}^{N} \pi_i^* = n$.
  ii) $0 < \pi_i^* \leq 1$ for all $i = 1, \ldots, N$.
  Thus, all equality and inequality constraints are fulfilled and a feasible solution is obtained.
- *Dual feasibility*
  To prove dual feasibility, we must show that $\mu_i^* \geq 0$ for all $i = 1, \ldots, N$. This is trivially fulfilled for $\mu_{m+1}^*, \ldots, \mu_N^*$, since these are all zero. For $i \leq m$ we have that $\mu_i^* = c_i - \lambda^*$, so it remains to show that this is positive for all $i \leq m$. To show this, let

$$\tilde{\pi}_i = (n - m) \frac{\sqrt{c_i}}{\sum_{j=m+1}^{N} \sqrt{c_j}} \quad \text{for all } i = 1, \ldots, N, \tag{A.5}$$

  and note that

$$\begin{aligned} c_i/\tilde{\pi}_i^2 &= c_j/\tilde{\pi}_j^2 \quad \text{for all } i, j, & \text{(A.6)} \\ 1 = \pi_i^* &\leq \tilde{\pi}_i \quad \text{for } i = 1, \ldots, m, & \text{(A.7)} \\ \pi_i^* &= \tilde{\pi}_i \quad \text{for } i = m+1, \ldots, N. & \text{(A.8)} \end{aligned}$$

  Here, (A.6) follows immediately from (A.5), and (A.8) from (A.1) at the final iteration of the algorithm. To show (A.7), note that Algorithm 1 iteratively increases the sampling probabilities assigned to non-certainty selections, meaning that the factor $\frac{(n-m)}{\sum_{j \in \mathcal{D} \setminus \mathcal{M}} \sqrt{c_j}}$ in (A.1) gradually increases as more certainty selections are added to the index set $\mathcal{M}$. Thus, any instance that according to (A.1) achieved a sampling probability exceeding 1 in any iteration of the algorithm will also have $\tilde{\pi}_i \geq 1$ in (A.5). Next, (A.7) implies that $c_i \geq c_i/\tilde{\pi}_i^2$ for $i \leq m$, and (A.8) that $c_N/\tilde{\pi}_N^2 = c_N/\pi_N^{*2}$, which for $i = 1, \ldots, m$ gives

$$c_i - \lambda^* = c_i - \frac{c_N}{\pi_N^{*2}} \geq \frac{c_i}{\tilde{\pi}_i^2} - \frac{c_N}{\tilde{\pi}_N^2} = 0 .$$

  were the first equality follows from (A.4), and the last equality from (A.6). This gives the desired result.
- *Complementary slackness*
  We finally note that complementary slackness

$$\mu_i^* g_i(\boldsymbol{\pi}^*) = 0 \quad \text{for all } i = 1, \ldots, N$$

  is fulfilled, since
  i) $g_i(\boldsymbol{\pi}^*) := \pi_i^* - 1 = 0$ for $i = 1, \ldots, m$.
  ii) $\mu_i^* := 0$ for $i = m+1, \ldots, N$.

Thus, we conclude that the point $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \lambda^*)$ with $\boldsymbol{\pi}^*$ computed according to Algorithm 1 and $(\boldsymbol{\mu}^*, \lambda^*)$ taken as in (A.4) satisfies the KKT conditions. Furthermore, since we consider a convex optimization problem with convex inequality constraint functions and affine equality constraints, the KKT conditions are sufficient for global optimality [50, Chapter 5.5.3], which completes the proof.

Acknowledgment

The study was conducted at SAFER Vehicle and Traffic Safety Centre at Chalmers. The authors thank all participants in the study, and Daniel Gellerman for assistance with video annotation.

Financial Support and Sponsorship

The funding sources had no involvement in the design and conduct of the study, nor in the writing or decision to submit the article for publication.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman, "An overview of the 100-car naturalistic study and findings," in *Proc. 19th Int. Tech. Conf. Enhanced Saf. Vehicles*, 2005, pp. 1–10.

[2] T. A. Dingus *et al.*, "The 100-car naturalistic driving study: Phase II—Results of the 100-car field experiment," Virginia Tech Transp. Inst., Blacksburg, VA, USA, Tech. Rep. DOT HS 810 593, 2006. [Online]. Available: https://trid.trb.org/view/783477

[3] T. Victor *et al.*, "Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk," Transp. Res. Board Nat. Academies, Washington, DC, USA, SHRP 2 Rep. S2-S08A-RW-1, 2014. [Online]. Available: http://onlinepubs.trb.org/onlinepubs/shrp2/SHRP2prepubS08AReport.pdf

[4] T. A. Dingus *et al.*, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 10, pp. 2636–2641, Mar. 2016.

[5] C. Kessler *et al.*, "EuroFOT deliverable D11.3 final report," euroFOT Consortium, Aachen, Germany, euroFOT Deliverable D11.3, 2012. [Online]. Available: https://research.chalmers.se/publication/171349/file/171349_Fulltext.pdf

[6] I. van Schagen *et al.*, "Towards a large-scale European naturalistic driving study: Final report of PROLOGUE: Deliverable D4.2," SWOV Inst. Road Saf. Res., Leidschendam, The Netherlands, PROLOGUE Deliverable D4.2, 2011. [Online]. Available: https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/9341/5/D4.2.pdf

[7] J. Bärgman *et al.*, "The UDRIVE dataset and key analysis results," UDRIVE Consortium, UDRIVE Deliverable 41.1, 2017. [Online]. Available: https://erticonetwork.com/wp-content/uploads/2017/12/UDRIVE-D41.1-UDrive-dataset-and-key-analysis-results-with-annotation-codebook.pdf

[8] M. A. Regan, A. Williamson, R. Grzebieta, and L. Tao, "Naturalistic driving studies: Literature review and planning for the Australian naturalistic driving study," in *Proc. Australas. College Road Saf. Conf. Safe Syst., Expanding Reach*, 2012, pp. 1–13.

[9] N. Uchida, M. Kawakoshi, T. Tagawa, and T. Mochida, "An investigation of factors contributing to major crash types in Japan based on naturalistic driving data," *IATSS Res.*, vol. 34, no. 1, pp. 22–30, Jul. 2010.

[10] F. Guo and J. Hankey, "Modeling 100-car safety events: A case-based approach for analyzing naturalistic driving data: Final report," Virginia Tech Transp. Inst., Blacksburg, VA, USA, Tech. Rep. 09-UT-006, 2009. [Online]. Available: https://vtechworks.lib.vt.edu/bitstream/handle/10919/7406/Modeling100_CarSafety_Final.pdf?sequence=1&isAllowed=y

[11] I. van Schagen and F. Sagberg, "The potential benefits of naturalistic driving for road safety research: Theoretical and empirical considerations and challenges for the future," *Proc. Social Behav. Sci.*, vol. 48, pp. 692–701, Jul. 2012.

[12] T. A. Dingus *et al.*, "Naturalistic driving study: Technical coordination and quality control," Transp. Res. Board Nat. Academies, Washington, DC, USA, SHRP 2 Rep. S2-S06-RW-1, 2015. [Online]. Available: https://vtechworks.lib.vt.edu/handle/10919/52942

[13] J. Balsa-Barreiro, P. M. Valero-Mora, I. Pareja-Montoro, and M. Sánchez-García, "Proposal of geographic information systems methodology for quality control procedures of data obtained in naturalistic driving studies," *IET Intell. Transp. Syst.*, vol. 9, no. 7, pp. 673–682, Sep. 2015.

[14] R. J. Jansen, S. T. van der Kint, and F. Hermens, "Does agreement mean accuracy? Evaluating glance annotation in naturalistic driving data," *Behav. Res. Methods*, early access, Jul. 2020, doi: 10.3758/s13428-020-01446-9.

[15] J. Balsa-Barreiro, P. M. Valero-Mora, J. L. Berné-Valero, and F.-A. Varela-García, "GIS mapping of driving behavior based on naturalistic driving data," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 5, p. 226, May 2019.

[16] F. Guo, I. Kim, and S. G. Klauer, "Semiparametric Bayesian models for evaluating time-variant driving risk factors using naturalistic driving data and case-crossover approach," *Statist. Med.*, vol. 38, no. 2, pp. 160–174, Jan. 2019.

[17] *Naturalistic Driving Studies—Vocabulary—Part 1: Safety Critical Events*, Standard ISO/TR 21974-1:2018, International Organization for Standardization, 2018. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:tr:21974:-1:ed-1:v1:en

[18] T. Victor *et al.*, "Sweden-Michigan naturalistic field operational test (SeMiFOT) phase 1: Final report," SAFER Vehicle Traffic Saf. Centre Chalmers, Gothenburg, Sweden, SAFER Rep. 2010:02, Project C3 SeMiFOT, 2010. [Online]. Available: https://document.chalmers.se/download?docid=1773834060

[19] M. Dozza and N. P. González, "Recognising safety critical events: Can automatic video processing improve naturalistic data analyses?" *Accident Anal. Prevention*, vol. 60, pp. 298–304, Nov. 2013.

[20] C. E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*. New York, NY, USA: Springer, 2003.

[21] S. Selpi *et al.*, "Deliverable D3.3: Data management in euroFOT," euroFOT consortium, Aachen, Germany, euroFOT Deliverable D3.3, 2011. [Online]. Available: https://research.chalmers.se/publication/171344/file/171344_Fulltext.pdf

[22] C. J. Skinner, "Domain means, regression and multivariate analysis," in *Analysis of Complex Surveys*, C. J. Skinner, D. Holt, and T. M. F. Smith, Eds. New York, NY, USA: Wiley, 1989, pp. 80–87.

[23] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, pp. 663–685, Dec. 1952.

[24] D. A. Binder, "On the variances of asymptotically normal estimators from complex surveys," *Int. Stat. Rev.*, vol. 51, no. 3, pp. 279–292, Dec. 1983.

[25] R Core Team, "R: A language and environment for statistical computing," R Found. Stat. Comput., Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[26] T. Lumley, "Analysis of complex survey samples," *J. Stat. Softw.*, vol. 9, no. 8, pp. 1–19, Apr. 2004.

[27] T. Lumley, *Complex Surveys: A Guide to Analysis Using R*. Hoboken, NJ, USA: Wiley, 2010.

[28] T. Lumley. (2019). *Survey: Analysis of Complex Survey Samples*. [Online]. Available: https://cran.r-project.org/web/packages/survey

[29] W. A. Fuller, *Sampling Statistics*. Hoboken, NJ, USA: Wiley, 2009.

[30] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. Oxford, U.K.: Clarendon Press, 1992.

[31] A. DasGupta, *Asymptotic Theory of Statistics and Probability*. New York, NY, USA: Springer, 2008.

[32] H. Imberg, J. Jonasson, and M. Axelson-Fisk, "Optimal sampling in unbiased active learning," in *Proc. 23rd Int. Conf. Artif. Intell. Statist. (Proc. Mach. Learn. Res.)*, vol. 108, 2020, pp. 559–569.

[33] G. S. Fishman, *Monte Carlo*. New York, NY, USA: Springer, 1996.

[34] T. Victor, "Keeping eye and mind on the road," Ph.D. dissertation, Dept. Psychol., Uppsala Univ., Uppsala, Sweden, 2005. [Online]. Available: https://uu.diva-portal.org/smash/get/diva2:167500/FULLTEXT01.pdf

[35] J. Bärgman, V. Lisovskaja, T. Victor, C. Flannagan, and M. Dozza, "How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from SHRP2," *Transp. Res. Part F, Traffic Psychol. Behav.*, vol. 35, pp. 152–169, Nov. 2015.

[36] N. E. Breslow, "Statistics in epidemiology: The case-control study," *J. Amer. Stat. Assoc.*, vol. 91, no. 433, pp. 14–28, Mar. 1996.

[37] N. E. Breslow, "Case-control studies," in *Handbook of Epidemiology*, W. Ahrens and I. Pigeot, Eds. New York, NY, USA: Springer, 2005, pp. 287–319.

[38] T. Victor, O. Blomberg, and A. Zelinsky, "Automating driver visual behavior measurement," in *Vision in Vehicles IX*, A. G. Gale, J. Bloomfield, G. Underwood, and J. Wood, Eds. Loughborough, U.K.: Loughborough Univ., 2012, pp. 181–188.

[39] M. Dozza and N. P. Gonzalez, "Recognizing safety-critical events from naturalistic driving data," *Proc. Social Behav. Sci.*, vol. 48, pp. 505–515, Jul. 2012.

[40] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 344–349.

[41] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 49–56, May 2016.

[42] S. Vora, A. Rangesh, and M. M. Trivedi, "On generalizing driver gaze zone estimation using convolutional neural networks," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2017, pp. 849–854.

[43] H. Imberg, "Unequal probability sampling in active learning and traffic safety," Licentiate thesis, Dept. Math. Sci., Chalmers Univ. Technol., Gothenburg, Sweden, 2019. [Online]. Available: https://research.chalmers.se/publication/512677/file/512677_Fulltext.pdf

[44] A. Scott and C. Wild, "Case-control studies with complex sampling," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 50, no. 3, pp. 389–401, 2001.

[45] A. Scott and C. Wild, "On the robustness of weighted methods for fitting models to case-control data," *J. Roy. Stat. Soc., Ser. B (Stat. Methodol.)*, vol. 64, no. 2, pp. 207–219, May 2002.

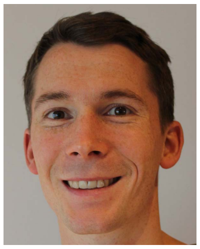
[46] A. Scott, "Population-based case control studies," *Surv. Methodol.*, vol. 32, no. 2, pp. 123–132, Dec. 2006.

[47] A. Scott and C. Wild, "Population-based case-control studies," in *Handbook of Statistics: Sample Surveys: Inference and Analysis*, vol. 29B, C. R. Rao, Ed. Amsterdam, The Netherlands: Elsevier, 2009, pp. 431–453.

[48] V. Landsman and B. I. Graubard, "Efficient analysis of case-control studies with sample weights," *Statist. Med.*, vol. 32, no. 2, pp. 347–360, Jan. 2013.

[49] D. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. New York, NY, USA: Wiley, 2003.

[50] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.



Henrik Imberg received the M.Sc. degree in mathematical statistics from the University of Gothenburg, Sweden, in 2016. He is currently pursuing the Ph.D. degree in applied mathematics and mathematical statistics with the Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg. His research is about development and optimization of sampling strategies for problems in statistics and machine learning.

Vera Lisovskaja received the Ph.D. degree in mathematical statistics from the Chalmers University of Technology, in 2013. After the defense, she proceeded to participate in a collaboration with SAFER Vehicle and Traffic Safety Centre at the Chalmers University of Technology, Gothenburg, as part of a Post-Doctoral position, this collaboration lasting over the span of two years. She then left the university to accept a position in applied statistics in an industry.

Selpi (Member, IEEE) received the B.Sc. degree in computer science from the University of Indonesia, in 2000, the M.Sc. degree in bioinformatics from the Chalmers University of Technology, Sweden, in 2004, and the Ph.D. degree in computing from Robert Gordon University, U.K., in 2008. She currently works with the Chalmers University of Technology. Her current research interests include applications of machine learning and data science for transport-related domain (e.g., understanding driving styles/driver behavior from naturalistic driving data, travel time and traffic volume predictions, and text-mining for text data in transport). She is also interested in understanding how mixed traffic, with vehicles with different driving styles and automation levels sharing the same roads, affects traffic safety, and efficiency. Beside academic work, she has several years of experiences in software industry. She is a member of the IEEE Intelligent Transportation Systems Society's technical committee on Naturalistic Driving Data Analytics. She has served as a reviewer and an associate editor for several IEEE conferences.

Olle Nerman is currently a Professor emeritus in Mathematical Statistics with the Chalmers University of Technology, Gothenburg, Sweden. During his career, he has been working broadly with research. He started with probabilistic modeling of population dynamics and continued with survey sampling, genetic epidemiology, and bioinformatics. In the last years before retiring, he worked with applied epidemiology and to some extent as a Supervisor and a Consultant to researchers with the SAFER Vehicle and Traffic Safety Centre, Chalmers University of Technology.