Making a few talk for the many – Modeling driver behavior using synthetic populations generated from experimental data

Downloaded from: https://research.chalmers.se, 2025-12-04 23:27 UTC

Citation for the original published paper (version of record):
Schindler, R., Flannagan, C., Bálint, A. et al (2021). Making a few talk for the many – Modeling driver behavior using synthetic populations generated from experimental data. Accident Analysis and Prevention, 162. http://dx.doi.org/10.1016/j.aap.2021.106331

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Accident Analysis and Prevention 162 (2021) 106331

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Making a few talk for the many – Modeling driver behavior using synthetic populations generated from experimental data

Ron Schindler [a,*], Carol Flannagan [b], András Bálint [a], Giulio Bianchi Piccinini [a]

[a] Division of Vehicle Safety, Department of Mechanics and Maritime Sciences, Chalmers University of Technology, Hörselgången 4, 41756 Göteborg, Sweden
[b] University of Michigan Transportation Research Institute, 2901 Baxter Road, Ann Arbor, MI 48109-2150, USA

ARTICLE INFO

Keywords:
Synthetic population
Experimental data
Bayesian functional data analysis
Vulnerable road user
Active safety system design

ABSTRACT

Understanding driver behavior is the basis for the development of many advanced driver assistance systems, and experimental studies are indispensable tools for constructing appropriate driver models. However, the high cost associated with testing is a serious obstacle in collecting large amounts of experimental data. This paper presents a methodology that can improve the reliability of results from experimental studies with a limited number of participants by creating a virtual population. Specifically, a methodology based on Bayesian inference has been developed, that generates synthetic cases that adhere to various real-world constraints and represent possible variations of the observed experimental data. The application of the framework is illustrated using data collected during a test-track experiment where truck drivers performed a right turn maneuver, with and without a cyclist crossing the intersection. The results show that, based on the speed profiles of the dataset and physical constraints, the methodology can produce synthetic speed profiles during braking that mimic the original curves but extend to other realistic braking patterns that were not directly observed. The models obtained from the proposed methodology have applications for the design of active safety systems and automated driving, demonstrating thereby that the developed framework has great promise for the automotive industry.

1. Introduction

Modern science is data driven and researchers need detailed data to understand phenomena related to human behavior. However, large scale data collections are often expensive and time consuming, which limits the size of available data sets and makes it challenging to develop reliable models.

One methodology to address this issue is the creation of synthetic populations, that is already commonly used in urban planning and travel demand modelling. The detailed data needed for the modelling process is obtained through the population synthesis, where the information can be constructed based on a small collected sample (Choupani and Mamdoohi, 2016). For example, very detailed household information (e.g., age, work place, income, employment status, shopping habits) that can only be collected in specific areas can be extrapolated to larger regions where the collection with the same amount of detail is not feasible (e.g., due to time and budget constraints), but some very general data points are available (e.g. population, household size, infrastructure). The strength of this approach is that combining an initial dataset set (which defines detailed correlations between different dimensions) and information about margins can preserve the internal correlations of the initial dataset as well as the external restrictions provided by the margins (Rich and Mulalic, 2012).

Commonly used population synthesizers include sample based methods, where the synthetic reconstruction can keep the valid correlation structure from the samples during the synthesis process (Ye and Wang, 2018), and those based on iterative proportional fitting (IPF), allowing several target constraints to be represented (Rich and Mulalic, 2012). IPF has also been used in Zhu and Ferreira (2014) to estimate the joint distribution of household and individual characteristics for transportation microsimulations and in Niebuhr et al. (2013) and Kreiss et al. (2015) for approximating the EU level crash population based on crash data collected in Germany. However, the constraints in IPF can be cross-linked when sharing target variables, resulting in non-trivial consistency issues when many target constraints and dimensions are addressed (Rich and Mulalic, 2012). A more detailed overview of IPF-based population synthesis and related issues is provided in Choupani and Mamdoohi (2016).

* Corresponding author.
E-mail address: ron.schindler@chalmers.se (R. Schindler).

https://doi.org/10.1016/j.aap.2021.106331
Received 31 March 2021; Received in revised form 15 June 2021; Accepted 1 August 2021
Available online 24 September 2021
0001-4575/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Other approaches have also been used to generate synthetic populations. Wu et al. (2018) have used the maximum entropy principle to generate models for multi-dimensional categorical data. Saadi et al. (2016) have proposed a Hidden Markov Model which uses a marginal distribution as boundary condition to guide population synthesis. Additionally, the detailed information needed for agent-based simulations can be provided by a Bayesian network approach (Sun and Erath, 2015).

In fact, Bayesian approaches have been used for the generation of synthetic populations, mainly for their power in creating probability distributions based on prior beliefs. As described in Gelman et al. (2013), Bayesian approaches represent the state of knowledge about unknown quantities as a probability distribution. A distribution representing prior beliefs can then be updated once new data are available and it can be shown that under reasonable assumptions, Bayesian inference provides a mathematically optimal way of performing such an update in a quantitative way (Hoff, 2009). Bayesian methods have been successfully used to complement theoretical results and empirical analyses based on frequentist methods in image analysis, artificial intelligence research and political science, along with many other research fields; see Kruschke (2015) for some examples. Previous publications in traffic safety research (e.g. Hauer, 1983a) show the value of Bayesian methods and have implemented them for different applications: to model the visual behavior of drivers in different driving situations (e.g. Lee and Lee, 2019; Morando et al., 2020), to assess the benefits associated to the changes in road infrastructure (e.g. Gårder et al., 1998) and to the introduction of safety systems (e.g. Kovaceva et al., 2020a), to quantify the effect of treatment strategies for convicted drivers (Hauer, 1983b), to improve crash prediction models (e.g. Miaou and Lord, 2003; Mitra and Washington, 2007; Huang and Abdel-Aty, 2010), and to determine the factors contributing to crashes (Xie et al., 2018).

Overall, Bayesian approaches have shown the potential to advance research on traffic safety in different applications. However, the work to generate synthetic population data in the traffic safety field has not included Bayesian methods so far. While Leledakis et al. (2021) predict relevant crash configurations of future passenger cars through the synthesis of virtual cases, these synthetic cases result from uniform variations of relevant crash parameters such as impact speeds, position in the lane and braking level and duration. The technical report from Waymo (2020) describes the role of generating synthetic conflict situations in their product development. In this case, synthetic cases represent variations of existing scenarios observed in crash databases and own data collected by Waymo through a "fuzzing" process, as well as the creation of entirely synthetic scenarios.

Beyond crash data analysis, a further important use case for synthetic populations is the possibility to generate additional data from samples which are limited in size, due to several constraints intrinsic to the planning of empirical studies with human subjects (e.g. costs, ethical considerations), for the creation of driver behavior models. Detailed models of driver behavior have wide applications for the design of autonomous vehicles and the evaluation of active safety systems (Markkula, 2015) and rely on data collected in naturalistic driving studies (e.g. Kovaceva et al., 2020b), driving simulators (Bianchi Piccinini et al., 2020), and test track experiments (Boda et al., 2018). Since these methods are time consuming and expensive, having a methodology available that can reduce the needed sample size will benefit both quality and availability of data sets.

In this paper, we demonstrate the use of Bayesian Functional Data Analysis to create braking profiles of a virtual population, based on physical constraints and data from a limited initial sample of participants. We will demonstrate the application of the methodology using a data set that was collected during a test-track experiment, where truck drivers were performing right turn maneuvers with and without the presence of a cyclist, showing the possibilities that the methodology provides based on real data. The generated dataset provides a model of drivers' kinematic behavior in specific scenarios, which can be used for the assessment of active safety systems and the design of autonomous vehicles. Therefore, this type of utilization of Bayesian methods may have potential applications for both academia and industry.

2. Material and methods

This paper proposes a methodology, based on Bayesian Functional Data Analysis (BFDA), that synthesizes small data samples to estimate the behavior of a population of drivers in a specific scenario. To illustrate our methodology, we use traffic safety related data that were collected in a real-world experiment, as described below.

2.1. Data source

The data used to illustrate the application of the methodology were collected from 11 participants driving a tractor-semitrailer combination during an experiment conducted at the AstaZero test track in Sweden. The participants drove six laps in the city environment of the test track. After one training lap and two baseline laps, where no vulnerable road user (VRU) was present, the drivers would encounter a cyclist dummy, crossing their path during a right turn maneuver (see Fig. 1), followed by two more baseline laps for this scenario.

The purpose of the experimental study was to identify and model the differences in driver behavior between the baseline scenario (no cyclist dummy present before and during the right turn maneuver) and cyclist scenario (cyclist dummy present before and during the right turn maneuver), motivated by the high share of crashes and large number of fatalities associated with this specific cyclist scenario in the European Union (Schindler et al., 2020). During the whole experiment, the drivers were instructed to obey the speed limit of 30 km/h and followed a predefined route during every lap. The drivers were informed that there were other traffic elements such as cars and vulnerable road users with them on the test track, but not further primed to any specific interactions. More details on the experiment are provided in Schindler and Bianchi Piccinini (2021).

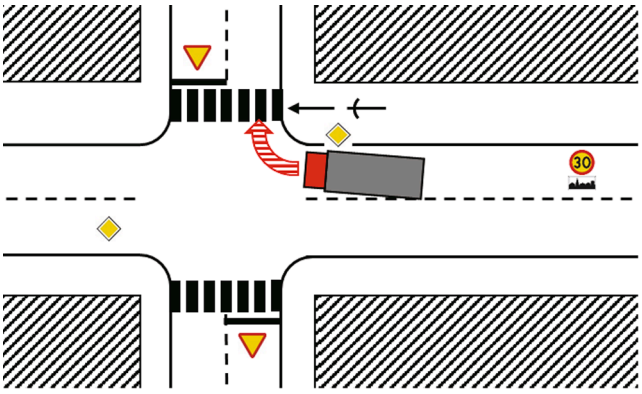

Fig. 1. Schematic representation of the interaction during the right-turn maneuver at the test-track (adapted from Schindler and Bianchi Piccinini, 2021).

2.2. Methodology

In this paper, we modeled the driver behavior by using the speed collected in the original experiment during the braking sequence. In the remainder of this paper, the expression "braking sequence" is used to describe the drivers' braking behavior during the approach to the intersection, from the start of braking until the lowest speed of the right turn maneuver is reached.

To model the data, we used BFDA. Functional Data Analysis (Ramsay, 2004) treats *functions* of data, as opposed to individual data values, as the dependent measure. In our work, entire braking sequences over time are represented by a function, and the coefficients of the function are modeled in a Bayesian framework. Bayesian models (Gelman et al., 2013) have the advantage of allowing for very complex parametric and distributional structures. Our dependent measure, the speed profile during braking, is described by a cubic curve that has six coefficients (see Section 3.1), and thus we are modeling the six-dimensional joint distribution of these coefficients as the dependent outcome using a hierarchical Bayesian model of those parameters.

As described in more detail in the next section, we first modeled each subject's speed profile during braking to obtain the six coefficients of each speed profile as a function of distance traveled. Next, we modeled the joint distribution of these coefficients and included hierarchical components to model driver-specific patterns. The use of hierarchical models enables separate representations of the distribution of driver-specific patterns across different drivers and the variation in braking produced by an individual driver across events.

3. Theory and calculations

3.1. Raw data modelling

Fig. 2 shows the raw speed profiles as a function of distance traveled from a fixed starting point at which the cyclist dummy movement was triggered (corresponding to 0 m) before the right-turn maneuver from the original experiment. Laps when the cyclist was present are indicated with solid red curves, and the gray box labeled "Cyclist appears" indicates the range of distances at which the cyclist was first visible to the driver (which varied slightly between drivers due to experimental limitations). The location of the center of the intersection, which is also the theoretical conflict point, at 66 m from the trigger point is also labeled.

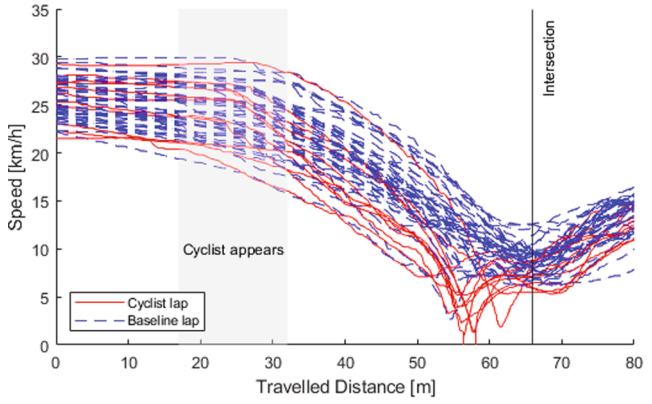

Fig. 2. Plot of speed over travelled distance from the trigger point (for cyclist (solid red) and baseline (dashed blue) laps) in the right-turn maneuver. The grey area marked as "Cyclist appears" represents the area when the cyclist dummy became visible to the truck drivers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The speed profiles were truncated at the start and end of braking, and the speeds in between were modeled by a cubic function (Eq. (1)),

$$y = d_0 + d_1 x + d_2 x^2 + d_3 x^3 \quad (1)$$

with y as the speed, x as the travelled distance from the trigger point and d_0 to d_3 as the coefficients. Functions with higher polynomial order did not provide improvements to the fit of the data, so the model with the fewest coefficients was chosen. For all laps, the start of the braking maneuver was defined as when the truck decelerated with more than 0.5 m/s^2, as this deceleration threshold was able to capture the start of the braking while excluding noise (e.g. coasting). The end of the braking maneuver for all laps was defined as when the truck had reached the lowest speed during the turning maneuver (or stopped completely). In the functional representation, speeds prior to the start of braking were assigned the start value (since the approach happened at an approximately constant speed) and speeds after the end of braking were assigned the end value (for cosmetic purposes). For the cubic function modeling, the speed profile between the start and end of braking was constrained to be monotonic non-increasing. Although the unconstrained best-fit cubic was briefly increasing at the start or end in a few cases, the raw data were never increasing, so the constraint was imposed to better match the main characteristics of the observed braking maneuver.

Table 1 shows the six coefficients that were used to describe each braking event. The first four coefficients describe the start and endpoints of the braking maneuver, and d_2 and d_3 are the quadratic and cubic terms from Eq. (1). The coefficients d_0 and d_1 in Eq. (1) are determined by the distance and speed values at the start point S and the end point E of the braking maneuver, and thus do not need to be separately modeled. The equations used for modelling speed based on the coefficients listed in Table 1 can be found in Appendix A.

Fig. 3 illustrates the raw data collected in four different laps and the modeled cubic curve that was fitted to this data. The coefficients describing the fitted curve can be found in the upper right corner of each plot. The plots also emphasize the fact that the fitted curve is only valid inside the interval between start and end of braking (delimited by red circles). Thus, as noted earlier, for graphical display, we assign the starting speed, S_y, to all distances prior to S_x and we assign the end speed, E_y, to all distances after the end of the maneuver E_x. For the analysis, we do not include modeled speeds before the start or after the end of the braking maneuver. Fig. 4 shows the fitted curves for all laps in a similar fashion to Fig. 2.

Table 1
Functional data coefficients (raw data).

Coefficient	Description of Coefficient
S_x	Travelled distance from trigger point at start of braking [m]
S_y	Speed at start of braking [km/h]
E_x	Travelled distance from trigger point at end of braking [m]
E_y	Speed at end of braking [km/h]
d_2	quadratic term [−]
d_3	cubic term [−]

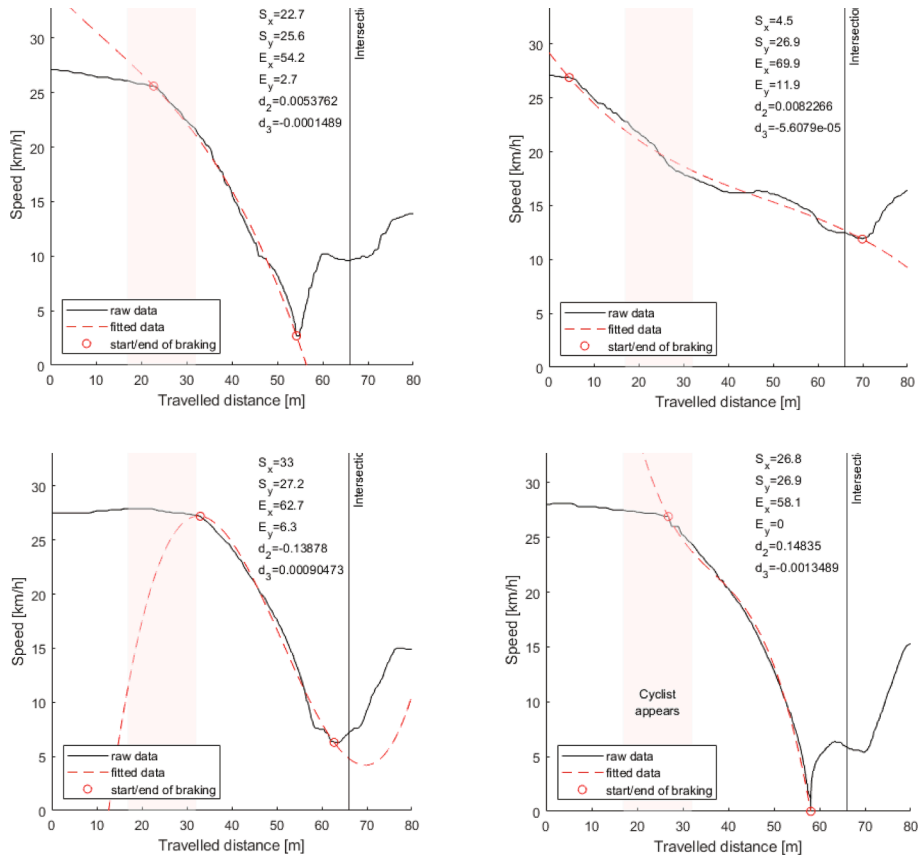

Fig. 3. Raw data and fitted curves for four different laps from four different participants. The plots show baseline laps and one cyclist lap (bottom right). These plots are directly comparable to the raw data shown in Fig. 2. Note that although speeds prior to the start point are generally constant in the raw data, speeds after the endpoint are not.

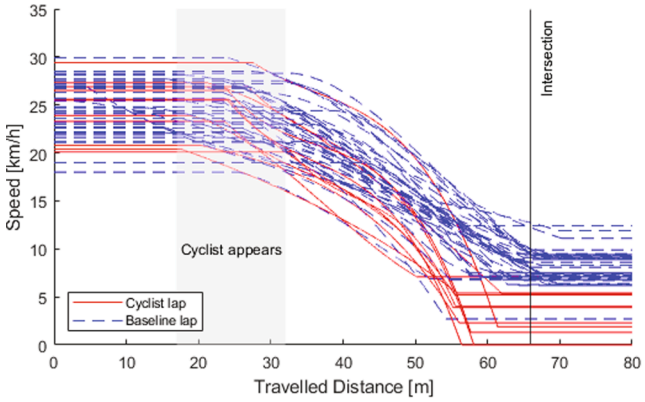

Fig. 4. Plot of modeled speed over travelled distance (for cyclist (solid red) and baseline (dashed blue) laps) in the right-turn maneuver. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Parameterization for modeling

The six coefficients of the function describing the braking profiles in the previous section 3.1 were reparameterized for the Bayesian model to simplify the implementation, especially of constraints on the coefficients. The constraints on the coefficients are easier to implement if the end distance E_x and end speed E_y are replaced by the change in distance E_x-S_x and change of speed E_y-S_y during the braking maneuver. Therefore, the four coefficients describing the start and end of the driver's braking profile were reparameterized as:

1) starting distance: S_x,
2) starting speed: S_y,
3) total distance traveled during braking maneuver: T_x, and
4) total speed reduction during braking maneuver: T_y.

This reparameterization allowed a straightforward implementation of boundary constraints. For example, total speed reduction was constrained to be non-negative, but no greater than start speed. Total distance was also constrained to be non-negative but was not constrained at the upper end in our implementation (because the braking maneuver could continue as the truck passed through the intersection). However, with this parameterization, it would be straightforward to impose an upper limit on total distance for the maneuver in future studies.

In order to implement the constraint of monotonically decreasing speed during the braking maneuver, the two coefficients of the cubic function between endpoints, i.e., d_2 and d_3 in Table 1, were reparameterized following Fenimore et al. (2000). We transformed the distance and speed dimensions to lie in a unit space, such that the start and endpoints were located at (0,1) and (1,0), respectively. The quadratic and cubic coefficients for approximating the unit-space speed curve as a cubic function of the unit-space distance (analogously to Eq. (1)) were labeled $\widehat{d_2}$ and $\widehat{d_3}$. Fenimore et al. (2000) derive the bounds on unit-space quadratic and cubic terms ($\widehat{d_2}$ and $\widehat{d_3}$) that produce monotonically decreasing functions as follows:

1) $-3 \leqslant \widehat{d_2} \leqslant 6.465$, and
2) $\widehat{d_3}$ is constrained by the most restrictive (on either end) of the following:
   a. $\left(1-\widehat{d_2}\right)/2$
   b. $\left(1-\widehat{d_2}\right)$
   c. $\frac{1}{6}\left(-3-3\widehat{d_2}-\sqrt{3}\sqrt{3+6\widehat{d_2}-\widehat{d_2}^2}\right)$
   d. $\frac{1}{6}\left(-3-3\widehat{d_2}+\sqrt{3}\sqrt{3+6\widehat{d_2}-\widehat{d_2}^2}\right)$

In practice, this means that for $\widehat{d_2} \leq 0$, $\widehat{d_3}$ must be between $-1-\widehat{d_2}$ and $\left(1-\widehat{d_2}\right)/2$. For $0 \leq \widehat{d_2} \leq 3$, $\widehat{d_3}$ must lie between $\frac{1}{6}\left(-3-3\widehat{d_2}-\sqrt{3}\sqrt{3+6\widehat{d_2}-\widehat{d_2}^2}\right)$ and $\left(1-\widehat{d_2}\right)/2$, and for $\widehat{d_2}$ greater than 3, $\widehat{d_3}$ must lie between $\frac{1}{6}\left(-3-3\widehat{d_2}-\sqrt{3}\sqrt{3+6\widehat{d_2}-\widehat{d_2}^2}\right)$ and $\frac{1}{6}\left(-3-3\widehat{d_2}+\sqrt{3}\sqrt{3+6\widehat{d_2}-\widehat{d_2}^2}\right)$. Fig. 5 shows the allowable coefficient combinations.

Because of the complex shape (and discontinuities in the edge) of this region, we further transformed the coefficients to another (0,1)-(0,1) space. In this space, which we call "beta space" because the transformed parameters will be modeled by a pair of draws from beta distributions in the BFDA setting described below, $\widehat{d_2}$ is transformed to $\tilde{d}_2$ and $\widehat{d_3}$ is transformed to $\tilde{d}_3$ as follows:

$$\tilde{d}_2 = \frac{\left(\widehat{d_2}+3\right)}{(6.465+3)}$$

$$\tilde{d}_3 = \frac{\left(\widehat{d_3}-ll_{d3}\right)}{(ul_{d3}-ll_{d3})}$$

where ll_{d3} and ul_{d3} are the lower and upper limits of $\widehat{d_3}$, given the choice of $\widehat{d_2}$. In other words, the beta space codifies each parameter as its proportion of the available range from lowest to highest value.

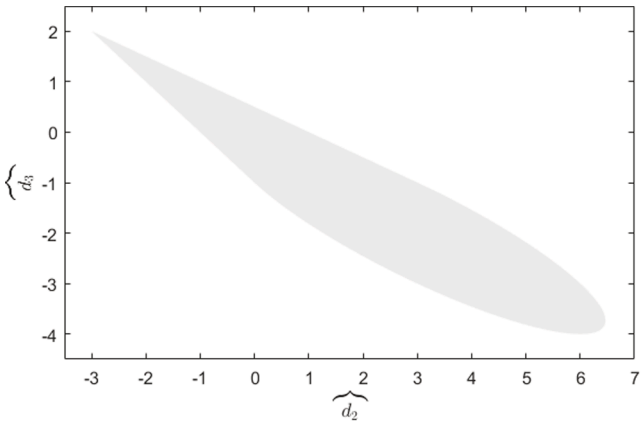

Fig. 5. Region of allowable combinations of $\widehat{d_2}$ and $\widehat{d_3}$ coefficients that permit monotonically decreasing approximation of standardized speed by a cubic polynomial of standardized distance.

Areas of the beta space can be thought of as representing "styles" of deceleration. This is illustrated in Fig. 6, which shows the general deceleration patterns from different parts of the beta space. The figure is organized to map to beta space with axes from 0 to 1 left to right and bottom to top. The curves were selected from deciles of each of the parameters. Thus, curves near the upper left corner of this figure correspond to values in the upper left corner of a graph of beta space, i.e., low values of $\tilde{d}_2$ and high values of $\tilde{d}_3$. We observe that as $\tilde{d}_2$ increases, the curves have an inflection point in the middle, which would reflect initial braking, followed by less braking and then a second episode of braking. In contrast, curves with $\tilde{d}_2$ in the range from ~

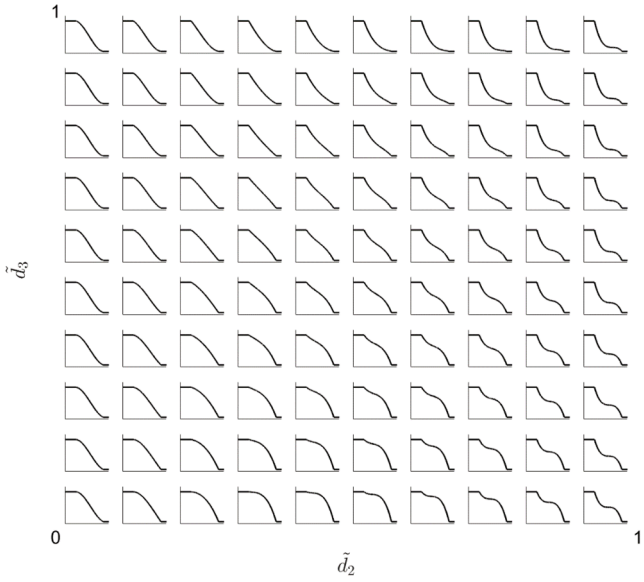

Fig. 6. Illustration of curve shapes for different coefficient pairs in beta space.

Table 2
Functional data parameters (modeled data).

Parameter	Description of Parameter
S_x	Travelled distance from trigger point at start of braking [m]
S_y	Speed at start of braking [km/h]
T_x	Total distance travelled during braking maneuver [m]
T_y	Total speed reduction during braking maneuver [km/h]
$\tilde{d}_2$	quadratic term (as proportion of the available range) [−]
$\tilde{d}_3$	cubic term (as proportion of the available range) [−]

0.2–0.5 and $\widetilde{d_3}$ below 0.5 tend to involve later braking, while curves with higher values of $\widetilde{d_3}$ and middle values of $\widetilde{d_2}$ involve early braking.

Table 2 shows the final six parameters that were used in the Bayesian model. Each set of six parameters describes a specific deceleration curve, and thus the joint distribution of these parameters describes the probability of each curve being observed in the context from which the data arose.

3.3. Bayesian model

Fig. 7 shows the model for the first four parameters: S_x, S_y, T_x, T_y. Each parameter was modeled as a truncated normal at the lowest level, with truncation at 0 on the left (since all parameters must be non-negative). Truncation is denoted by "Trunc[lower,upper]" in the figure. Speed reduction (T_y) was also truncated on the right at start speed (S_y) for each draw. The parameters of the four normal distributions are represented by a linear model for the mean and a single parameter for the variance. The linear model included an intercept plus a slope for the cyclist-present dummy variable, resulting in three parameters per truncated normal at the base level. The priors on the eight mean parameters (including intercept and cyclist slope for each parameter) were hierarchical normal distributions with mean and variance parameters for each. Each subject's priors were drawn from this distribution. Hyperpriors on the hierarchical parameters were normal for means with values indicated in Fig. 7. The prior on S_x was selected to be centered on the location where the cyclist could appear, and the prior on S_y was selected to be slightly lower than the speed limit set for the experiment. The prior on T_x was selected such that when paired with the mean prior on S_x, the maneuver would end shortly after the center of the intersection. This was partially based on observing this to occur in some of the braking events. Finally, the prior on T_y was set such that when paired with the mean of S_y, the mean end speed would be 5 km/h. The prior on the cycle parameter was a regularizing prior, centered at 0. All variance parameters had half-normal (hN) priors with values indicated in Fig. 7 as recommended by Gelman et al. (2013). We initially tested half-Cauchy priors on variances, but these produced unreasonable values in our prior predictive checks because the Cauchy tails are too fat for this application. As described later, priors were evaluated using prior predictive checks, and we also conducted a sensitivity analysis on the variances of the priors to determine whether the results were being overly influenced by the choice of prior.

Fig. 8 shows the model structure for the two shape parameters after transformations, $\widetilde{d_2}$ and $\widetilde{d_3}$. The two parameters are modeled as independent Beta mixtures with three components. The mixture components are different for the cycle and no-cycle conditions, and individual drivers have their own mixing proportions (k_{ij}), indicating their style preferences for each of the shape parameters. The Dirichlet distribution represents the driver-level distribution of mixing proportions. Priors on the Beta distribution parameters and the Dirichlet are indicated in Fig. 8. Parameterization of the Beta as lambda and phi (rather than alpha and beta), as well as the Gamma and Pareto priors are recommended by Gelman et al. (2013). Because each point in the two-dimensional beta space represents a different braking style, we wanted the priors to be as flat as possible across this space. The specific values of these prior distributions were selected using prior predictive checks (described in the

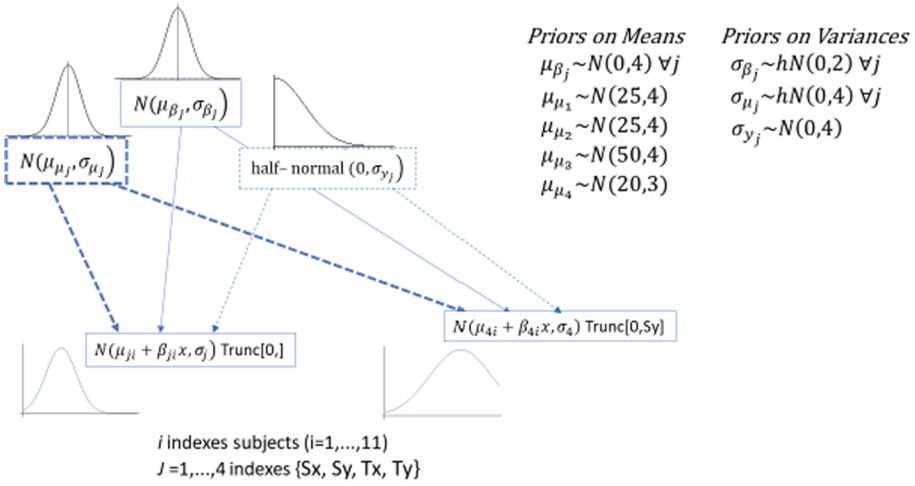

Fig. 7. Visualization of Bayesian hierarchical model structure for the four endpoint-related parameters of the deceleration curve.

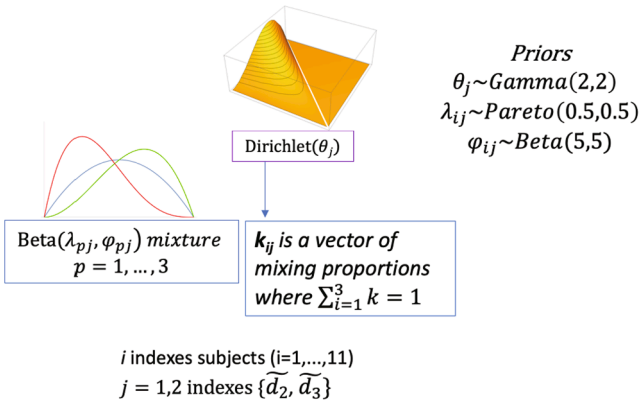

Fig. 8. Visualization of the Bayesian hierarchical model structure for the two shape parameters of the deceleration curve.

results). From this, we selected the combination of priors that produced the most even spread of draws in beta space.

The model was run in Stan using R Version 4. To obtain prior predictive draws and to generate the full posterior, we ran four chains to 200,000 draws after a warmup of 100,000 draws. The draws were thinned by selecting every other draw because of computing limits. Graphics in the results were based on a random subsample of 2,500 draws, though numerical results were based on the full set of posterior draws (after thinning). For the sensitivity analyses, we obtained 100,000 posterior draws and thinned to 50,000.

Diagnostics indicated that the chains had converged after the warmup draws and throughout the posterior draws. The largest R-hat was 1.021, well under the common cutoff of 1.1 to indicate lack of convergence (Gelman and Rubin, 1992).

4. Results

4.1. Prior predictive draws

The first task in setting up the model was to select and test the priors for each parameter. This was done through prior predictive checks, in which values (i.e., 6-tuples, each representing a braking curve) are generated using only the prior distributions. The goal is to ensure that the selected values are reasonable, which for some parameters is straightforward (i.e., checking that total speed reduction is not more than starting speed and checking that total distance is not unreasonably large relative to the location of the intersection). For others (e.g., selection of deceleration style from beta space), there is no source of prior information, so relatively uninformative priors are preferred.

Fig. 9 shows the distribution of each of the four endpoint-related coefficients from the prior predictive check. Cycle and no-cycle conditions are separated. Values are in reasonable ranges, especially total distance, which without constraint on the upper end could have produced unreasonably large values. Since the priors for the cycle parameters are centered at zero, the cycle prior draws have greater variability (because of the extra slope parameter) but are centered in the same location.

Fig. 10 shows the shape parameters in beta space. The distribution across the space is generally (but not perfectly) flat and represents an uninformative prior.

Fig. 11 shows the distribution of maximum deceleration from the prior predictive draws. One of the challenges of working with complex multi-parameter functions is that it is not always clear what parameter values are reasonable. Particularly for the purpose of prior and posterior

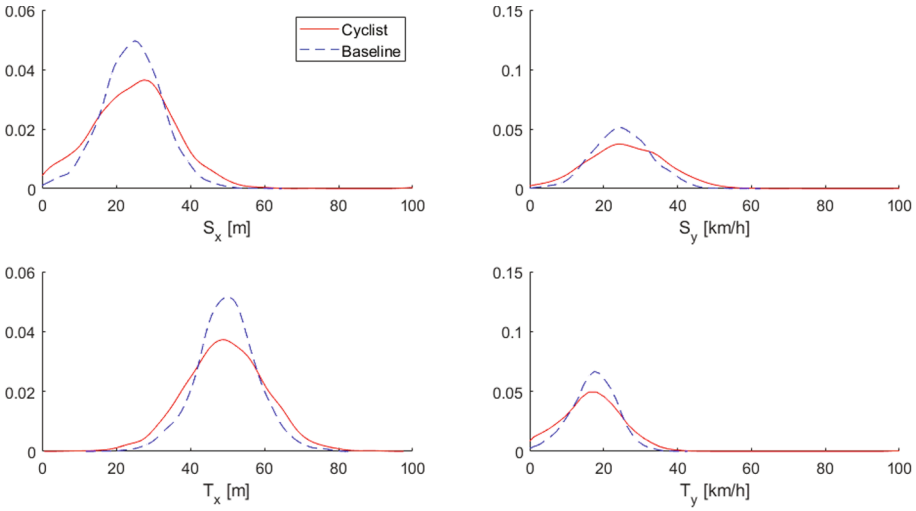

Fig. 9. Prior predictive draws for values of the four endpoint-related coefficients. Note that the speed reduction is constrained to be less than the starting speed. (S_x: travelled distance from trigger point, at start of braking; S_y: speed at start of braking; T_x: total distance travelled during braking maneuver; T_y: total speed reduction during braking maneuver).

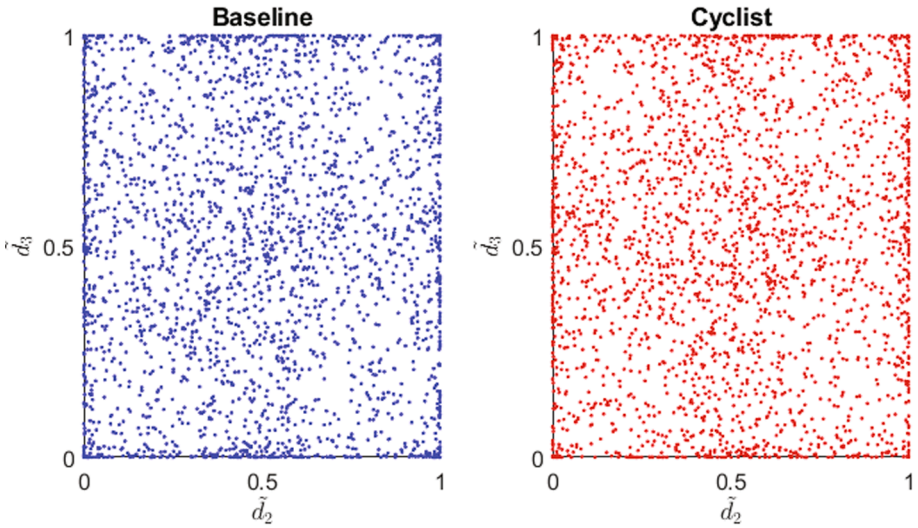

Fig. 10. Distribution of observations of $\tilde{d}_2$ and $\tilde{d}_3$ parameters in beta space from prior predictive draws.

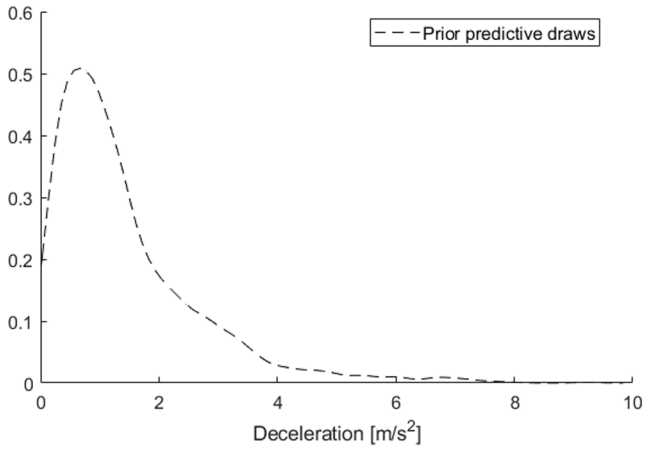

Fig. 11. Distribution of maximum deceleration from curves generated from prior predictive draws.

predictive checks, it is useful to have a means of "sense checking" the resulting models. To provide a sense-check on the models, we focused on maximum deceleration in the resulting curves because the braking constraints of heavy trucks are known (see e.g. Economic Commission for Europe, 2016). Thus, it is possible to identify curves that are implausible based on maximum deceleration. While decelerations of 8 m/s^2 are unlikely, the distribution covers a range of very plausible values and, at the high end, possible values of deceleration, suggesting that the priors are appropriate.

4.2. Posterior predictive draws

For each observation, posterior predictive draws were obtained for each model in the posterior. To evaluate model fit, we visualized each original speed curve against a random sample of the curves generated by the posterior predictive draws for that observation (i.e., using the values of the subject ID and cycle condition for that observation). Fig. 12 shows selected examples of this comparison; for a figure including all curves, see Appendix B. The bold black lines show the observed speed curves for the particular subject and either baseline or cyclist-present condition. For the baseline condition, each participant had multiple observations that are graphed together, and the blue lines represent posterior predictive draws of the baseline condition for that participant. For the cycle condition, each participant had one observation and the red lines are a sample of posterior predictive draws for the cycle condition for that participant. The posterior draws mimic the original curves in both the endpoints and curve shapes, giving confidence in the parameterization of the model.

Table 3 shows descriptive statistics for the four cycle parameters associated with the hierarchical distributions. That is, these four parameters, labeled μ_{β_j} in Fig. 7, control the mean of the hierarchical distribution from which subject-specific mean cycle parameters are drawn. There is one for each of the first four parameters, as noted in the table. For the initial speed and distance, the cycle parameter posterior distribution contains 0. However, the 95% credible interval (crI) for the posterior distribution of total distance traveled during the maneuver is entirely negative, and for the speed reduction, the crI is entirely positive. This means that when the cyclist is present, maneuvers cover less distance and the total speed reduction is greater than when the cyclist is not present, in line with the expectations from what is seen in Fig. 2.

Fig. 13 shows the distribution of the four endpoint-related coefficients from the posterior predictive draws with observations when the cyclist was present and absent separated. Consistent with the results of the hierarchical mean cyclist parameters, the posterior predictive draws for distance and speed at the start of the maneuver are nearly identical, but the total distance is less, and speed reduction is greater when the cyclist is present.

The shift of the T_x-curves to the left (i.e. reduction of total distance travelled during braking maneuver) and T_y-curves to the right (i.e. higher speed reduction during the braking maneuver) in the cyclist condition, results in a higher likelihood of larger decelerations of the

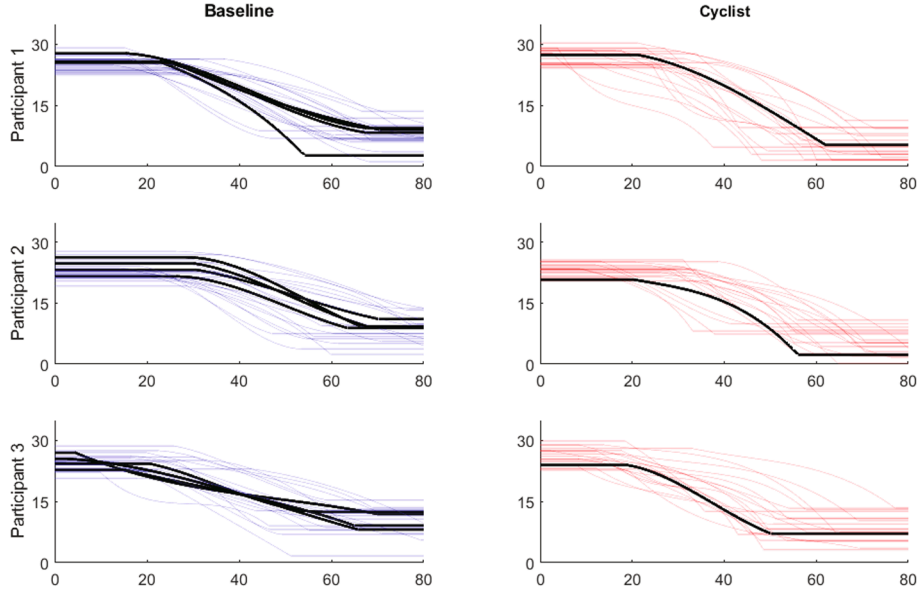

Fig. 12. Posterior predictive checks against the original curves (solid black lines) for 3 participants. Each plot shows a sample of 20 randomly selected blue/red curves produced by posterior draws associated with the specific observation (based on the subject and cycle condition). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Descriptive statistics for four μ_{β_j} parameters.

Parameter	Mean	95% Credible interval (2.5th, 97.5th)
μ_{β_1} (related to S_x [m])	−0.818	(−4.283, 2.648)
μ_{β_2}, (related to S_y [km/h])	0.046	(−1.249, 1.351)
μ_{β_3} (related to T_x [m])	−5.926	(−9.708, −2.071)
μ_{β_4} (related to T_y [km/h])	4.752	(2.658, 6.933)

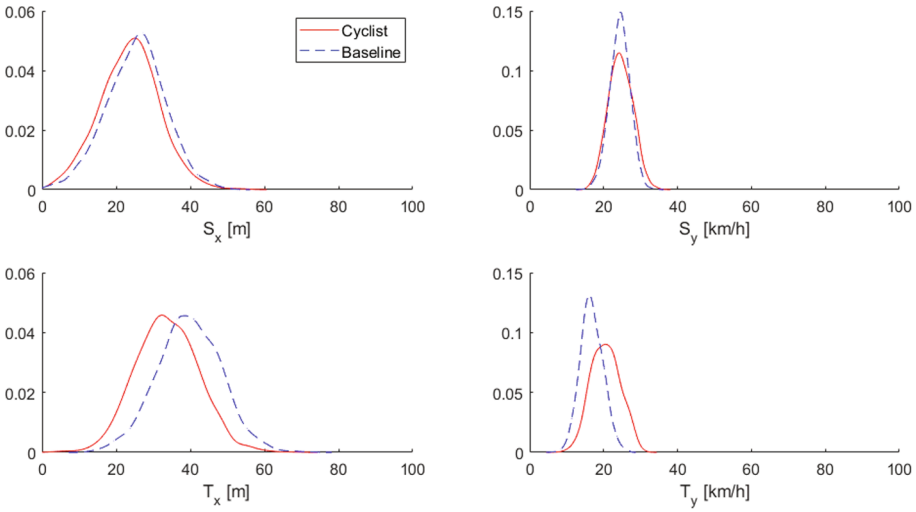

Fig. 13. Density of endpoint-related coefficient values in posterior predictive draws with cyclist and baseline conditions overlaid. (S_x: travelled distance from trigger point, at start of braking; S_y: speed at start of braking; T_x: total distance travelled during braking maneuver; T_y: total speed reduction during braking maneuver).

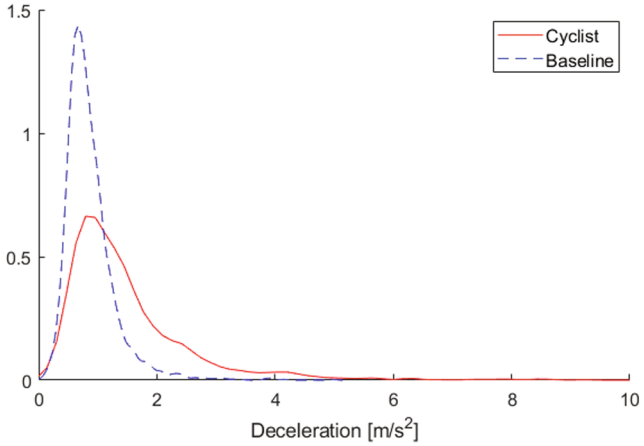

Fig. 14. Distribution of maximum deceleration for baseline and cyclist laps.

truck compared to the baseline condition (see Fig. 14).

Fig. 15 shows the bivariate plot of the two transformed curve-shape parameters, $\widetilde{d_2}$ and $\widetilde{d_3}$ with cyclist conditions separated. When the cyclist is not present, the posterior values of $\widetilde{d_2}$ most commonly fall in narrow ranges from 0.08 to 0.11. Reviewing Fig. 6, curves with this $\widetilde{d_2}$ parameter range tend to have fairly constant deceleration throughout the braking maneuver. When the cyclist is present, the curves cluster loosely around a value of $\widetilde{d_2}$ of 0.5 and $\widetilde{d_3}$ of 0.32. Shapes in this area of the space tend to have greater deceleration in the middle of the curve and in some cases, two periods of increased deceleration.

Fig. 16 shows the distribution of maximum deceleration for the posterior predictive draws, with the prior curve from Fig. 11 shown for reference. The posterior is less variable than the prior and tail values are lower.

When the resulting parameters are fed back into Eq. (1), the curves represented in Fig. 17 can be modeled. The dashed blue line represents the average speed profile of a truck driver when there is no cyclist present, and the red solid line when there is a cyclist present, for the given scenario. The area around the curves represents the values between the 1st and 3rd quartile for each situation.

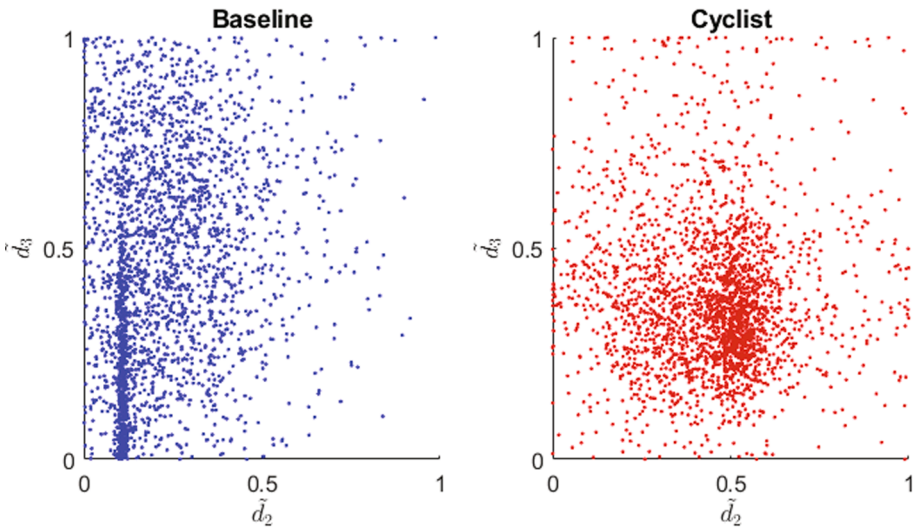

Fig. 15. Posterior predictive draws of $\tilde{d}_2$ and $\tilde{d}_3$ curve-shape parameters.

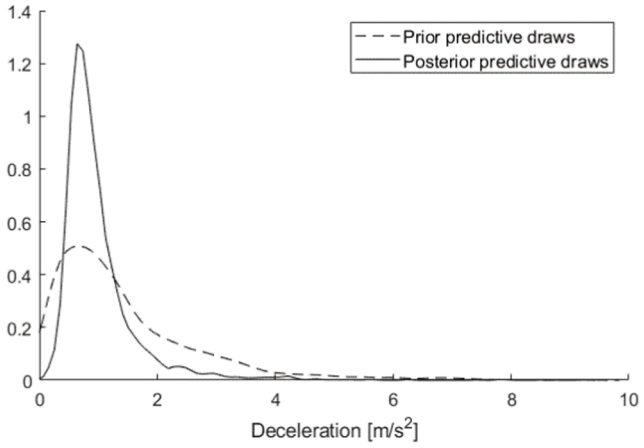

Fig. 16. Distribution of maximum deceleration for posterior and prior predictive distributions.

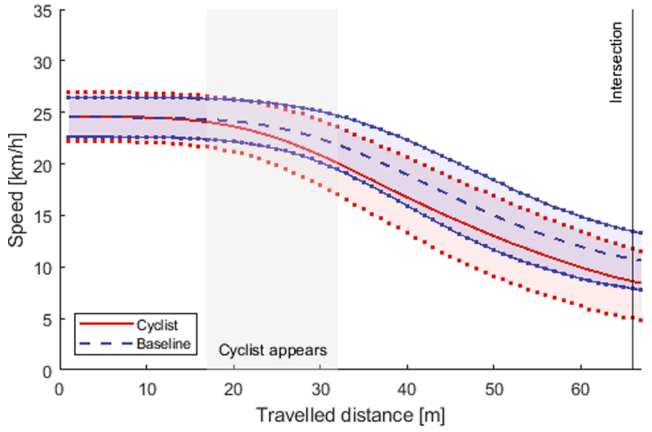

Fig. 17. Modeled speed curve with 25th and 75th percentile for baseline and cyclist maneuvers.

4.3. Sensitivity to priors

While the prior predictive checks give confidence that the selected priors are reasonable, we also conducted a sensitivity analysis to understand how stable the results are with respect to the choice of priors. Because we used a very flat prior on the style-related parameters (that result in draws of $\tilde{d}_2$ and $\tilde{d}_3$) and have no particular justification for any specific more informative prior, we did not further investigate the sensitivity of the model to those priors. However, the priors on the components of the endpoint-related coefficients (i.e., those shown in Fig. 8) can influence draws of those parameters. In particular, the variances at both hierarchical levels can influence the variance in the distributions of posterior predictive draws. Thus, the sensitivity analysis focused on the variance components of the endpoint-related priors.

Rather than run all possible combinations of 12 variance parameters, we constructed two groups of such parameters: 1) "high-variance" priors in which all 12 variance priors were increased by 2 standard deviation units relative to the baseline model, and 2) "low-variance" priors in which all 12 variance priors were decreased by 2 standard deviation units relative to the baseline model. This enables observation of the extent to which the data are driving model results, even in the context of fairly strong differences in the priors. The results, provided in Figs. 18 and 19, show relatively small differences between the cases with different variance values and imply that the main conclusions of smaller total distance, larger speed reduction and larger maximum deceleration for the cyclist condition do not change for either considered prior.

5. Discussion

In this paper, we have shown how Bayesian methods can be used to

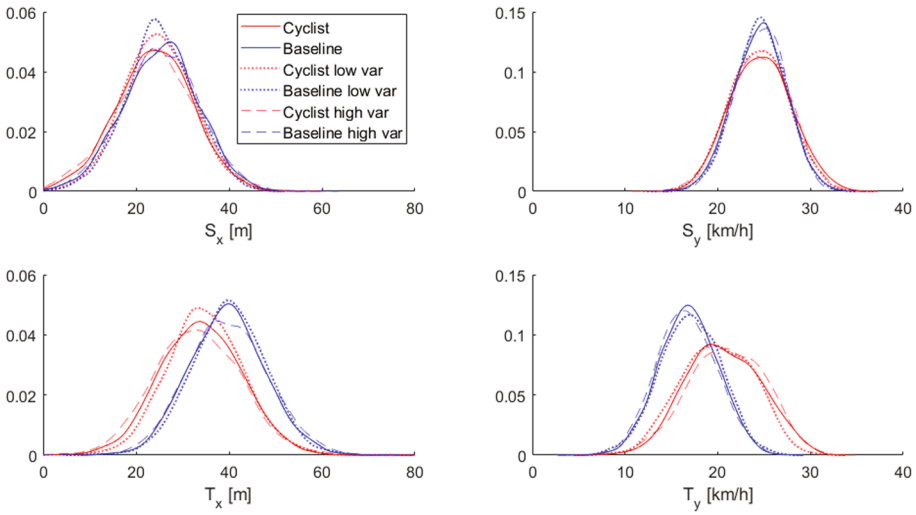

Fig. 18. Density of endpoint-related coefficient values in posterior predictive draws with high-variance and low-variance prior conditions overlaid. (S_x: travelled distance from trigger point, at start of braking; S_y: speed at start of braking; T_x: total distance travelled during braking maneuver; T_y: total speed reduction during braking maneuver).

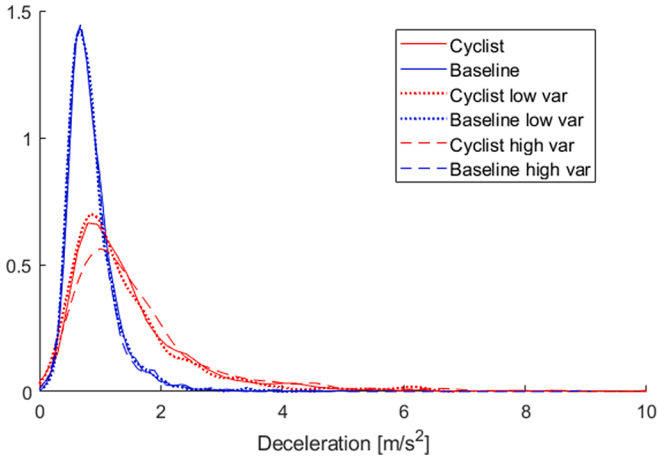

Fig. 19. Distribution of maximum deceleration for posterior predictive distributions with high and low variance prior conditions overlaid.

estimate driver behavior for a larger population of drivers, based on an initial small data sample. Our research focused on the modelling of speed during a right-turn maneuver, but the same methodology can be applied to other driving behavior measurements and different driving scenarios. What has been common procedure in urban planning and travel demand modelling – the creation of synthetic populations – has been combined with recent steps taken in the traffic safety research by implementing these procedures based on Bayesian methods.

When it comes to validating the results, Ma and Srinivasan (2015) addressed the inherent issue of validation of a synthetic population (i.e., if we knew the true population against which to validate, we would not need a synthetic population) in the context of transportation (in their case, sociodemographics of households for travel data synthesis). They used two methods: 1) comparing to population statistics with fewer control variables, and 2) comparing to artificial true data based on a small sample of census data. Analogous to the first approach, we generated a distribution of maximum deceleration that could be compared to other sources of maximum deceleration for heavy trucks. This is the strongest indication that our method produces reasonable braking curves. We do not have an external source of sample data analogous to the second method, but the posterior predictive checks against the original sample values at least assures us that the model is fitting reasonably well (granting that showing fit is not the same as validation).

Beyond these tests of model validity, the sensitivity analysis indicated that the data themselves were the strongest driver of results, although the tails of the posterior predictive distributions show a slightly higher influence by the priors. In general, while the Bayesian approach enables draws of a wide range of credible braking patterns based on a small-sample experiment, draws from the extremes of the posterior should be viewed with caution. Like all forms of extrapolation from small samples, extremes are much more dependent on model assumptions than are values closer to the center of the distribution. Thus, if this approach were used to generate unusual braking patterns (e.g. for crash simulations), sensitivity analysis, some external source of crash-related braking maneuvers for validation and general caution in reliance on extreme draws would be crucial.

A benefit of the implemented hierarchical models is that they can be used to generate synthetic braking profiles that, if demographic variables are included, could be tuned to different populations of drivers. In our case, we did not include demographics, but the hierarchical components allow us to draw examples of drivers who exhibit different (but plausible) braking behavior than the ones tested. If a larger data set is available, there is also the possibility to include more parameters in the modelling, extending the capabilities of the methodology.

The applications of the methodology for research on traffic safety are multiple. For the scenario described in this paper, the model of deceleration can be used in the design of future active systems that aim to prevent a collision with a cyclist, during the right-turn maneuver (e.g. Right turn assist with emergency braking or Blind Spot Information System combined with Automatic Emergency Braking). The results of our analysis (e.g., Fig. 17) can be fed into active safety systems as an input and be used by the systems to assess whether a driver has adapted their speed to the presence of a cyclist who is approaching the same intersection. The modeled speed profiles generated by the Bayesian methods allow to distinguish between two different strategies when approaching an intersection: the speed profile of the first "prototypical" driver type (blue curves in Fig. 17) is purely imposed by vehicle dynamics constraints, while the speed profile of the second "prototypical" driver type (red curves in Fig. 17) is also dictated by the possible risk to enter into a conflict with the crossing cyclist. Future generations of active safety systems for trucks could use the speed signal and perform a real-time comparison with the speed profiles of the "prototypical" drivers, to determine the type of deceleration initiated by the driver (e.g., deceleration that appears to be driven only by vehicle dynamics constraints or deceleration that is likely to be driven by both vehicle dynamics constraints and possible risk of collision with cyclist). Together with other information acquired from sensors (e.g. presence of cyclist), the real-time comparison of speed profiles can be included in the algorithm of the active safety systems, to determine the most appropriate timing for a warning or an intervention.

The models of speed profiles resulting from this paper also have an application for automated vehicles, by describing the comfortable deceleration patterns of the two "prototypical" drivers for the specific right-turn scenario. Previous research showed that the speed during automated driving is considered appropriate by drivers when it mimics the speed maintained by humans in the same maneuver (Abe et al., 2018). Future automated vehicles will have access to detailed information from the driving environment, including the presence of a cyclist and the distance to the intersection. They could therefore choose the driving speed using this information and the models of speed profiles presented in this paper and in future research for additional scenarios.

A methodological benefit of the approach presented in this paper is that additional data collected from other sources (e.g. other experiments on test tracks or Naturalistic Driving Studies) can be used to update the model (see also Kovaceva et al., 2020a) as long as the driving scenario is comparable to the one modeled.

In conclusion, this paper proposes a new methodology based on a Bayesian approach to make use of small data samples. Following the approach of synthetic populations, the sample data is used to estimate population behavior through Bayesian Functional Data Analysis, making better data available at less cost. The methodology and its specific results presented in this paper have wide applications, both in the design of active safety systems as well as the design of autonomous vehicles. Future research will focus on linking up behavior in different contexts such as naturalistic driving data or simulator studies and extending application of the methodology to other scenarios and contexts, e.g. to EuroNCAP use cases. With EuroNCAPs strategy to include virtual assessments (EuroNCAP, 2017), the methodology itself as well as the results can provide valuable input to the driver behavior models and assessment tools used.

6. Funding


Parts of this work were supported by the European Union's Horizon 2020 project "AEROFLEX" [grant number 769658].

7. Statement of Contribution/Potential Impact

Modern science is data driven and researchers need detailed data to understand phenomena related to human behavior. In urban planning, synthetic populations are used to provide the needed data, but this concept has not been applied in traffic safety related areas. Our proposed methodology enhances the use of data that is already available, but too small in size to perform robust analysis with traditional statistical approaches, e.g. classical null hypothesis testing. Using Bayesian methods, we developed a model to predict the behavior of a larger population of truck drivers, where the starting point for the generation of the large synthetic population was data collected in a test-track experiment with a limited number of participants. This methodology has practical applications for the design of future active safety systems and automated driving. In addition, the models can take further advantage of the Bayesian methods by being easily updated if new information becomes available.

CRediT authorship contribution statement

Ron Schindler: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft. **Carol Flannagan:** Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft. **András Bálint:** Formal analysis, Supervision, Validation, Writing - review & editing. **Giulio Bianchi Piccinini:** Data curation, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

$$y = E_y - E_x\left(\frac{S_y - E_y - d_2\left(S_x{}^2 - E_x{}^2\right) - d_3\left(S_x{}^3 - E_x{}^3\right)}{S_x - E_x}\right) - d_2E_x{}^2 - d_3E_x{}^3 + x\left(\frac{S_y - E_y - d_2\left(S_x{}^2 - E_x{}^2\right) - d_3\left(S_x{}^3 - E_x{}^3\right)}{S_x - E_x}\right) + d_2x^2 + d_3x^3 \quad \text{(A1)}$$

Function used for modelling speed (y), based on start of braking (S_x,S_y), end of braking (E_x,E_y) and shape of the braking curve (d_2,d_3)

Appendix B

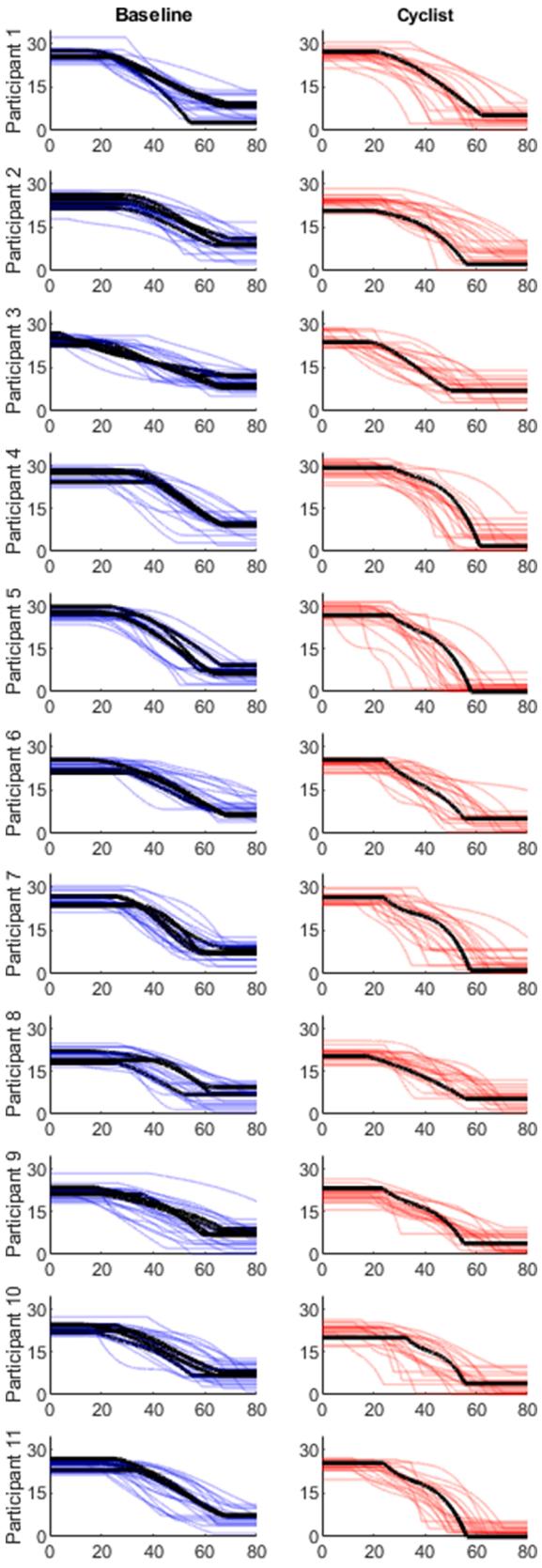

Fig. B1. Posterior predictive checks against the black original curves for all participants. Each plot shows a sample of 20 randomly selected blue/red curves produced by posterior draws associated with the specific observation (based on the subject and cycle condition).

References

Abe, G., Sato, K., Itoh, M., 2018. Driver trust in automated driving systems: the case of overtaking and passing. IEEE Trans. Hum.-Mach. Syst. 48 (1), 85–94.

Bianchi Piccinini, G., Lehtonen, E., Forcolin, F., Engström, J., Albers, D., Markkula, G., Lodin, J., Sandin, J., 2020. How do drivers respond to silent automation failures? Driving simulator study and comparison of computational driver braking models. Hum. Factors 62 (7), 1212–1229.

Boda, C.-N., Dozza, M., Bohman, K., Thalya, P., Larsson, A., Lubbe, N., 2018. Modelling how drivers respond to a bicyclist crossing their path at an intersection: How do test track and driving simulator compare? Accident Anal. Prevent. 111, 238–250.

Choupani, A.-A., Mamdoohi, A.R., 2016. Population synthesis using iterative proportional fitting (IPF): a review and future research. Transp. Res. Procedia 17, 223–233.

Economic Commission for Europe. (2016). Regulation No 13 - Uniform provisions concerning the approval of vehicles of categories M, N and O with regard to braking. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/0a43f880-d612-11e5-a4b5-01aa75ed71a1/language-en on 28 January 2021.

EuroNCAP. (2017). EuroNCAP 2025 Roadmap. Retrieved from https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf on 09 March 2021.

Fenimore, C. D., Libert, J. M., & Brill, M. (2000). Algebraic constraints implying monotonicity for cubics (NIST Interagency/Internal Report (NISTIR)-6453).

Gårder, P., Leden, L., Pulkkinen, U., 1998. Measuring the safety effect of Raised Bicycle crossings using a new research methodology. Transp. Res. Rec. 1636 (1), 64–70.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statist. Sci. 7, 457–511.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. CRC Press.

Hauer, Ezra, 1983a. Reflections on methods of statistical inference in research on the effect of safety countermeasures. Accid. Anal. Prev. 15 (4), 275–285.

Hauer, Ezra, 1983b. An application of the likelihood/Bayes approach to the estimation of safety countermeasure effectiveness. Accid. Anal. Prev. 15 (4), 287–298.

Hoff, P., 2009. A first course in Bayesian statistical methods, Vol. 580. Springer.

Huang, Helai, Abdel-Aty, Mohamed, 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42 (6), 1556–1565.

Kovaceva, Jordanka, Bálint, András, Schindler, Ron, Schneider, Anja, 2020a. Safety benefit assessment of autonomous emergency braking and steering systems for the protection of cyclists and pedestrians based on a combination of computer simulation and real-world test results. Accid. Anal. Prev. 136, 105352. http://dx.doi.org/10.1016/j.aap.2019.105352.

Kovaceva, Jordanka, Bärgman, Jonas, Dozza, Marco, 2020b. A comparison of computational driver models using naturalistic and test-track data from cyclist-overtaking manoeuvres. Transport. Res. F Traffic Psychol. Behav. 75, 87–105.

Kreiss, J.-P., Pastor, C., Dobberstein, J., Feng, G., Krampe, J., Meyer, M., Niebuhr, T., 2015. Extrapolation of GIDAS accident data to Europe no. 15–0372-O.

Kruschke, John K., 2015. In: Doing Bayesian data analysis. Elsevier, pp. 193–219. http://dx.doi.org/10.1016/B978-0-12-405888-0.00008-8.

Lee, Ja Young, Lee, John D., 2019. Modeling microstructure of drivers' task switching behavior. Int. J. Hum Comput Stud. 125, 104–117.

Leledakis, A., Lindman, M., Östh, J., Wågström, L., Davidsson, J., Jakobsson, L., 2021. A method for predicting crash configurations using counterfactual simulations and real-world data. Accid. Anal. Prev. 150, 105932.

Ma, Lu, Srinivasan, Sivaramakrishnan, 2015. Synthetic population generation with multilevel controls: a fitness-based synthesis approach and validations. Comput.-Aided Civ. Infrastruct. Eng. 30 (2), 135–150.

Markkula, G., 2015. Driver Behavior Models for Evaluating Automotive Active Safety: From Neural Dynamics to Vehicle Dynamics. Chalmers University of Technology.

Miaou, Shaw-Pin, Lord, Dominique, 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transp. Res. Rec. 1840 (1), 31–40.

Mitra, Sudeshna, Washington, Simon, 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accid. Anal. Prev. 39 (3), 459–468.

Morando, Alberto, Victor, Trent, Dozza, Marco, 2020. A Bayesian reference model for visual time-sharing behaviour in manual and automated naturalistic driving. IEEE Trans. Intell. Transp. Syst. 21 (2), 803–814.

Niebuhr, T., Kreiss, J.-P., Achmus, S., 2013. GIDAS-Aided quantification of the effectiveness of traffic safety measures in EU 27. In: Proceedings of the 18th Conference on the Enhanced Safety of Vehicles-Nagoya, p. 541.

Ramsay, J. O. (2004). Functional data analysis. Encyclopedia of Statistical Sciences, 4.

Rich, Jeppe, Mulalic, Ismir, 2012. Generating synthetic baseline populations from register data. Transport. Res. A Policy Pract. 46 (3), 467–479.

Saadi, Ismaïl, Mustafa, Ahmed, Teller, Jacques, Farooq, Bilal, Cools, Mario, 2016. Hidden markov model-based population synthesis. Transport. Res. B Methodol. 90, 1–21.

Schindler, R., Bianchi Piccinini, G., 2021. Truck drivers' behavior in interactions with vulnerable road users at intersections: results from a test-track experiment. Accident Anal. Prevent. 159, 106289.

Schindler, R., Jänsch, M., Johannsen, H., Bálint, A., 2020. An analysis of European crash data and scenario specification for heavy truck safety system development within the

AEROFLEX project. Transport Research Arena 2020, Helsinki, Finland. (Conference cancelled). Available at https://arxiv.org/abs/2103.05325.
Sun, Lijun, Erath, Alexander, 2015. A Bayesian network approach for population synthesis. Transport. Res. C Emerg. Technol. 61, 49–62.
Waymo (2020) Waymo safety report – September 2020.
Wu, Hao, Ning, Yue, Chakraborty, Prithwish, Vreeken, Jilles, Tatti, Nikolaj, Ramakrishnan, Naren, 2018. Generating realistic synthetic population datasets. ACM Trans. Knowl. Discov. Data (TKDD) 12 (4), 1–22.
Xie, S.Q., Dong, Ni, Wong, S.C., Huang, Helai, Xu, Pengpeng, 2018. Bayesian approach to model pedestrian crashes at signalized intersections with measurement errors in exposure. Accid. Anal. Prev. 121, 285–294.
Ye, P., Wang, X., 2018. In: Population synthesis using discrete copulas. IEEE, pp. 479–484.
Zhu, Y. & Ferreira, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. Journal of the Transportation Research Board 2429, 1 (2014), 168–177.