Accident Analysis and Prevention 106 (2017) 428–436

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Development of a real-time prediction model of driver behavior at intersections using kinematic time series data

Yaoyuan V. Tan[a,*], Michael R. Elliott[a], Carol A.C. Flannagan[b]

[a] Department of Biostatistics, University of Michigan, United States
[b] University of Michigan, Transportation Research Institute, United States

ARTICLE INFO

Keywords:
Connected driverless vehicles
Connected autonomous vehicles
Bayesian additive regression trees
Naturalistic driving data
Principal components analysis
Longitudinal prediction

ABSTRACT

As connected autonomous vehicles (CAVs) enter the fleet, there will be a long period when these vehicles will have to interact with human drivers. One of the challenges for CAVs is that human drivers do not communicate their decisions well. Fortunately, the kinematic behavior of a human-driven vehicle may be a good predictor of driver intent within a short time frame. We analyzed the kinematic time series data (e.g., speed) for a set of drivers making left turns at intersections to predict whether the driver would stop before executing the turn. We used principal components analysis (PCA) to generate independent dimensions that explain the variation in vehicle speed before a turn. These dimensions remained relatively consistent throughout the maneuver, allowing us to compute independent scores on these dimensions for different time windows throughout the approach to the intersection. We then linked these PCA scores to whether a driver would stop before executing a left turn using the random intercept Bayesian additive regression trees. Five more road and observable vehicle characteristics were included to enhance prediction. Our model achieved an area under the receiver operating characteristic curve (AUC) of 0.84 at 94 m away from the center of an intersection and steadily increased to 0.90 by 46 m away from the center of an intersection.

1. Introduction

An autonomous vehicle can be loosely defined as a vehicle where no human supervision or human controlled driving is needed. The National Highway Traffic Safety Administration (NHTSA) provides a more detailed definition with five levels of classification (National Highway and Traffic Safety Administration, 2013), ranging from Level 0 – the driver completely controls the vehicle at all times (typical of a 20th century vehicle before the introduction of electronic stability control or antilock braking) – to Level 4, where vehicle performs all functions for the entire trip, with the driver not expected to control the vehicle at any time. An example of a Level 4 autonomous vehicle is a vehicle from the Google Self-Driving Car Project.

In 2009, Google started testing these self-driven vehicles on the streets of Mountain View, California and Austin, Texas. As of August 2015, Google reported that they had self-driven these vehicles for more than 1 million miles (Google, 2015), and they had been involved in a total of 14 accidents since 2009 (CNNMoney, 2015). In all these accidents, Google asserted that human error and inattention was the main cause. Google's claim is not surprising since these vehicle will have to interact with human drivers. Unfortunately, human drivers do not always communicate their decisions clearly, leading to near crashes and crashes. As such, autonomous vehicles can benefit from predicting human driver decisions using information conveyed by the human driver's vehicle.

In this paper, we hypothesized that the kinematic behavior of a human driven vehicle provides enough information to make a good prediction of driver intent within a short time frame. We envision a system whereby a driver-intent model is evaluated on the human driver's vehicle and transmitted via vehicle-to-vehicle (V2V) communication. Although current autonomous vehicles under development generally use onboard sensors to gather information, V2V communication will increasingly be available as an additional source of information, resulting in connected and automated vehicles (CAVs). Hence, using the kinematic behavior of a human driven vehicle to predict driver intent makes sense if a driver's unique tendencies are an important predictor. Thus, a human-driven vehicle can learn its driver's intent patterns and communicate these to CAVs nearby.

In particular, we studied the speed of a human driven vehicle. We focused on predicting whether a driver will stop at an intersection before executing a left turn. This is important for two reasons. First, left turns at intersections can result in injury-causing side impacts.

* Corresponding author. Present address: 1415 Washington Heights, Ann Arbor, MI 48109, Unites States.
E-mail address: vincetan@umich.edu (Y.V. Tan).

http://dx.doi.org/10.1016/j.aap.2017.07.003
Received 18 December 2016; Received in revised form 27 June 2017; Accepted 4 July 2017
Available online 20 July 2017
0001-4575/

According to the National Motor Vehicle Crash Causation Survey (NMVCCS), 22% of the tow-away crashes in the US in 2008 were due to left turn maneuvers at intersections (Choi, 2010). Second, knowing if a human driven vehicle would stop before executing the left turn maneuver would allow the driverless vehicle to make a critical decision of whether to execute its own planned maneuver or wait.

To build the prediction model, we used naturalistic driving data from about 100 licensed drivers in Michigan. We converted the time series data to a distance series and defined a new distance-varying outcome. Because we believed that recent speeds contain more information about the human driver's intention to stop compared to past speeds, we employed a moving window on the distance-varying speeds. Next, we used principal components analysis (PCA) to reduce the number of variables employed in our prediction algorithm. To link our distance-varying outcomes to our principal component (PC) variables and five other road and observable vehicle characteristics predictors, we used a model we recently developed, the random intercept Bayesian additive regression trees (riBART). riBART (Tan et al., 2016) is an extension of Bayesian additive regression trees (BART; Chipman et al., 2010) which is able to account for the repeated left turns made by the same driver. We evaluated riBART's prediction performance at every meter away from the center of an intersection by using the area under the receiver operating characteristic curve (AUC) and compared these results with standard BART and linear logistic regression. Finally, we plotted the sensitivity and false positive rate (FPR; Davis and Goadrich, 2006) profiles of riBART where sensitivity = $\frac{\text{True positives}}{\text{All left turn stops}}$ and FPR = $\frac{\text{False positives}}{\text{All left turn } \mathbf{non} - \text{stops}}$ to investigate how the predicted probability cut-off level affects unnecessary stops by the CAVs and crashes.

2. Data and methods

2.1. Naturalistic driving data

We obtained our dataset from a previous study by Sayer et al. (2011). In brief, our naturalistic driving data was collected from 108 licensed drivers in Michigan between April 2009 and April 2010. Sixteen late-model Honda Accords were fitted with cameras, recording devices, and a collision warning system – the Integrated Vehicle Based Safety System (IVBSS) – to collect visual and kinematic data from the drivers for a total of 40 days – 12 days baseline period with IVBSS switched off followed by 28 days with IVBSS activated. To avoid confounding due to the IVBSS system, we used the 12 days baseline unsupervised driving data to develop our prediction model. Because information about road types and intersections outside Michigan was not available, we restricted our analysis to driving within Michigan in order to facilitate the accurate identification of an intersection and its associated road type. Accurate identification of an intersection allows us to determine a reference time to start extracting the information necessary for our prediction model.

In this study, we had data from 108 drivers who made 3795 turns. Of these 3795 turns, 1823 were left turns. We took the time at 100 m away from the center of an intersection (−100 m) as the reference point for the start of data extraction and stopped extraction at the time the vehicle was beyond the center of an intersection i.e. 0 m. We extracted both the speed of the vehicle (in m/s) and the amount of distance traveled (in m) at 10 ms intervals starting from our reference point. We also defined a vehicle as stopped when its speed was ≤1 m/s.

Because our goal is prediction of stopping before turning for future turns, we rescale the original time series predictors to measure distance from the intersection. We do this because, in a turn that is not complete, only the distance from the intersection will be known in advance; we will know the duration that the vehicle takes to reach the center of an intersection only after the vehicle has reached the center of an intersection. Fig. 1 illustrates this conversion using an example with Driver 40 Trip 34 Turn 1. Fig. 1(a) shows the speed profile of this particular turn. In this example, our target is the vehicle speed at 70 m away from the center of an intersection (−70 m). To obtain this speed, we first "draw" a line at −70 m and focus on the speed sample points closest to this −70 m line. Fig. 1(b) shows the blow up of this focal point. To set the speed at −70 m, we then compared which of the two speed sample points was closest to −70 m. In our example, because the point on the left was closest, it was set as the speed at −70 m for this turn. For the speeds of this turn from −100 m to −1 m at every 1 m interval, we employed a similar approach. In the situation where more than one speed sample point was closest to the line, we took their average as the speed at that distance.

Because vehicles can stop and restart before reaching the center of the intersection, we define "stopping" as a distance-varying outcome. Let i be the ith turn and j be the jth meter away from the center of intersection, $j = -100,\ldots,-1$. Let s_{ij} be the distance series of vehicle speed and y_{ij} be the distance-varying outcome (1 = stopped in future, 0 = will not stop in future). We defined y_{ij} as follows:

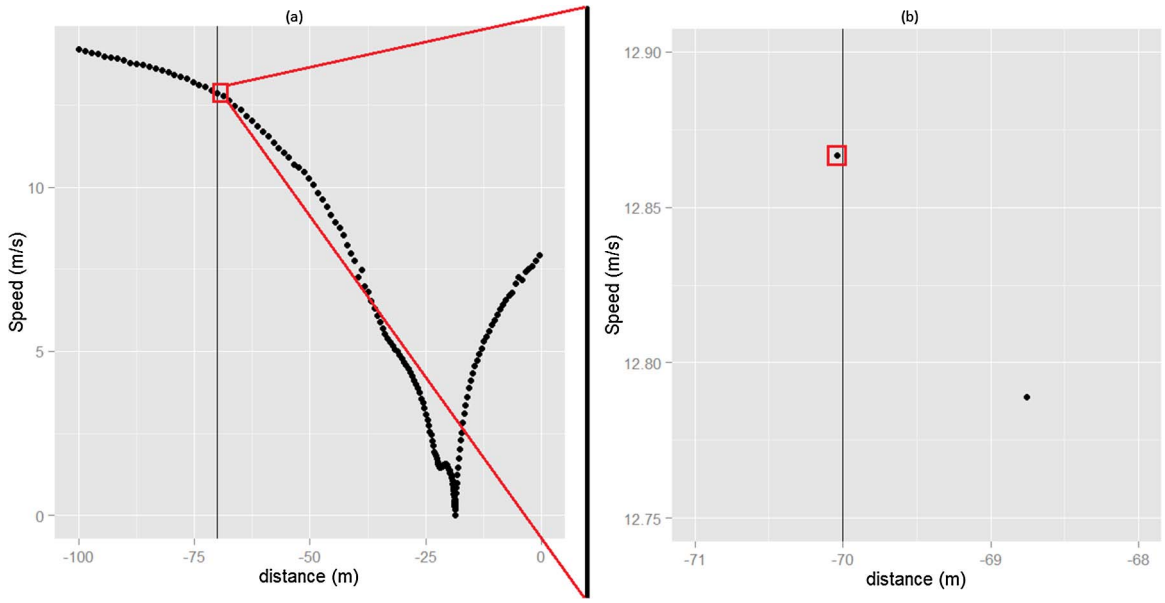

Fig. 1. Original speed profile of Driver 40 Trip 34 Turn 1.

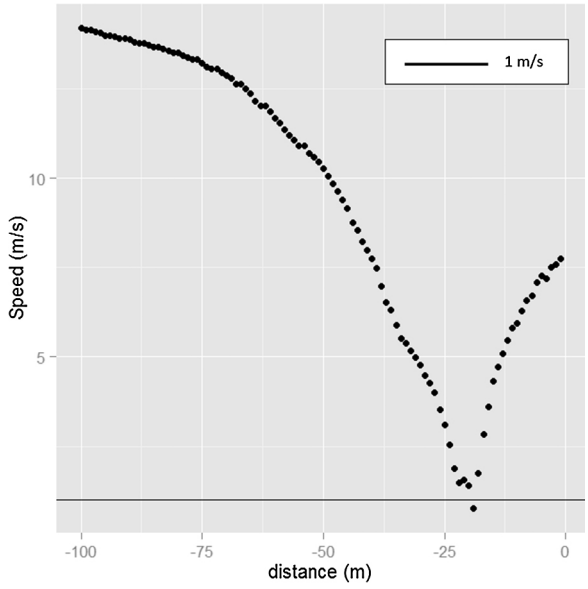

Fig. 2. New distance series speed profile of Driver 40 Trip 34 Turn 1. The horizontal line indicates 1 m/s.

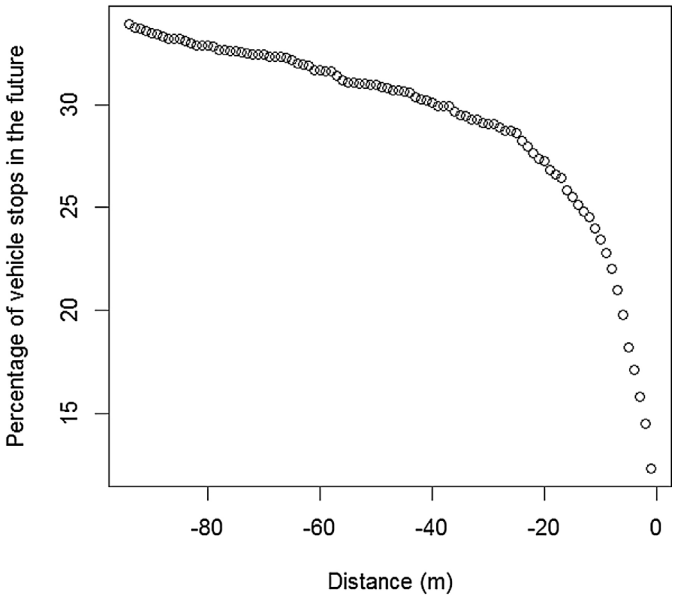

Fig. 3. Proportion of vehicles in our study that would be stopped ($\leq$1 m/s) at some future point for each meter away from the center of an intersection.

1. If $s_{ij} > 1$ m/s $\forall\, j = -100,\ldots, -1$, then set $y_{ij} = 0$ for every j.
2. If $s_{ij} \leq 1$ m/s for some $j \in \{-100,\ldots, -1\}$, let $c \in \{-100,\ldots, -1\}$ be the index such that for every $j > c$, $s_{ij} > 1$ m/s. We set $y_{i,-100} = y_{i,-99} = \cdots = y_{i,c} = 1$ and $y_{i,c+1} = y_{i,c+2} = \cdots = y_{i,-1} = 0$.

Point 1 in the above definition means that if the new distance series speed profile of a particular turn was more than 1 m/s throughout, the distance-varying outcome would be set to 0 throughout. Fig. 2 clarifies point 2 in the above definition. Fig. 2 corresponds to the new distance series of Driver 40 Trip 34 Turn 1. We can see that for $j > -19$, the speed of the vehicle was more than 1 m/s. Hence, $c = -19$ and the distance-varying outcome y_{ij} is set to 1 for $j = -100,\ldots, -19$, and to 0 for $j = -18,\ldots, -1$. In general, c can be viewed as the final distance from the center of an intersection such that the speed of the vehicle is < 1 m/s. In the event that there are multiple stops, c will be the closest distance to the center of an intersection where the vehicle "stopped".

Using our new definition of stops, we show the proportion of vehicles in our study that had a future stop as the vehicles approached the center of an intersection meter-by-meter in Fig. 3. About 30% of the vehicles would be stopped at some point at around 40 m away form the center of an intersection, with the proportion rapidly dropping to less than 10% less than 5 meters away from the center of an intersection.

2.1.1. Summary measures of speed trajectories

With the conversion and definition of the distance-varying outcome in place, we developed our prediction model by first employing a moving window of speeds. This was done because as the vehicle approaches the center of intersection, we believed that recent vehicle speeds contain more information on whether a human driver will decide to stop. The full profile of a vehicle's past speeds may include this information as well, but past speeds may contain irrelevant information, making the full profile of a vehicle's past speeds noisier compared to a window of recent speeds. For every jth meter $j = -100,\ldots,-1$, we defined the moving window of speeds as,

$$M_{ij} = \{s_{i,j-w+1},\ s_{i,j-w+2},\ \ldots, s_{ij}\}$$

where w is the size of the moving window.

Next, we used PCA on these M_{ij}s to reduce the number of covariates in our prediction model. Before reduction, the covariates are $s_{.,j-w+1}$, $s_{.,j-w+2}$, ..., $s_{.j}$. We let

$$M_j = \begin{bmatrix} s_{1,j-w+1} & s_{1,j-w+2} & \cdots & s_{1j} \\ s_{2,j-w+1} & s_{2,j-w+2} & \cdots & s_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n,j-w+1} & s_{n,j-w+2} & \cdots & s_{nj} \end{bmatrix} \quad \text{and} \quad u(j) = \begin{bmatrix} u_{j-w+1} \\ u_{j-w+2} \\ \vdots \\ u_j \end{bmatrix}$$

where M_j is the matrix of moving windows with the first row being M_{1j}, second row being M_{2j},..., and nth row being M_{nj}. There are w (number of columns in M_j) orthogonal vectors $u(j)$ that decompose the variance of M_j into w parts under the condition that for each $u(j)$, $||u(j)|| = 1$. To obtain the w decomposed variances, we used the formula: $PC_j = Var[M_j u(j)]$. If we let $PC_{j(q)}$ be the ordered statistic where $q = 1,\ \ldots, w$ and $u(j)_{(q)}$ be the ordered vector corresponding to $PC_{j(q)}$, then the first PC is $X_{j1} = M_j u(j)_{(w)}$, the second PC is $X_{j2} = M_j u(j)_{(w-1)}$, and so on. In our study, we found that the first two PCs explained at least 99% of the variation in M_j for all $j = -100,\ldots,-1$. Fig. 4 shows selected PC loadings from -95 m to -15 m. Based on visual inspection, the first PC approximates the mean velocity and the second PC approximates acceleration/deceleration (first derivative). Although the first PC explains the vast majority of the variability in vehicle speed, the second PC added non-trivial information for prediction, as we show in Section 3. In addition to the two PC predictors, we added five other road and observable vehicle characteristics predictors to adjust and enhance our prediction model. Details of these predictors are provided in Section 3.

2.2. Modeling stopping behavior using riBART

There are many alternative prediction models that are able to predict, or equivalently, classify a binary outcome such as stopping behavior. The impact of our two predictors on the risk of stopping may be highly non-linear, as well as multiplicative (interactions). We conducted preliminary analyses (not reported here), starting with logistic regression that explored linear and non-linear main effects and interactions. We then tried the "Super Learner" ensemble method that uses a weighted combination of various machine-learning and statistical models. The machine-learning and statistical models we included in our "Super Learner" were elastic net (Friedman et al., 2010), linear logistic regression, K-Nearest Neighbor, Generalized Additive Models (Hastie and Tibshirani, 1990), mean of the outcomes, and BART. Although the prediction performance of Super Learner was better than logistic

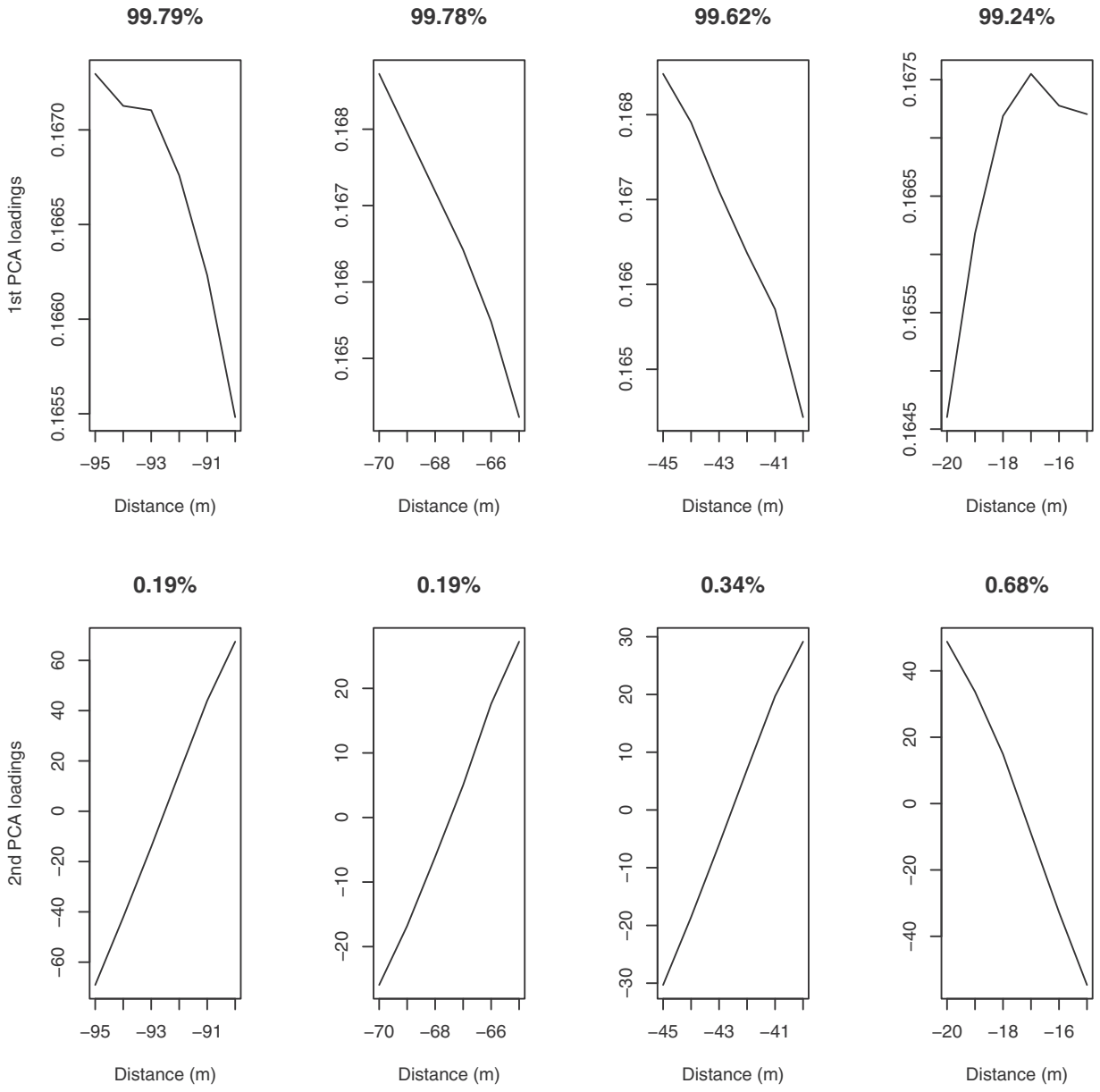

Fig. 4. Principal component loadings for the first and second PC from −95 m to −90 m, −70 m to −65 m, −45 m to −40 m, and −20 m to −15 m (left to right). The percentages indicate the proportion of variation explained by each PC.

regression, non-linear logistic regression, and BART as measured by the AUC criterion, it was highly unstable. In contrast, BART produced estimates that performed almost as well as the Super Learner but far more stable.

In essence BART models the mean outcome of Y_{ij} given covariates by a sum of regression trees and incorporates the additive effects of predictors. An important advantage of BART is that it can stably estimate non-linear effects of predictors in main effects and interactions in relatively high dimensions. Because BART in its current formulation assumes independent observations; for this analysis, we extend BART to account for the presence of repeated measurements in the dataset by using a random intercept. In the next few paragraphs, we briefly present the riBART model for correlated binary outcomes. Readers interested in a more complete overview of BART or random intercept BART (riBART) can refer to Chipman et al. (2010) and Tan et al. (2016) respectively.

Let $X_{kij} = \{X_{kij1}, \ldots, X_{kijp}\}$ be the p predictors in our model where $k = 1, \ldots, K$ indexes the drivers, $K = 108$, $i = 1, \ldots, I_k$ the turns, $j = -100,\ldots,-1$, the distance from the intersection, and l the regression tree. Formally, we write binary outcomes riBART as

$$P(y_{\text{kij}} = 1|X_{\text{kij}}) = \Phi[G_a(X_{\text{kij}})], \tag{1}$$

$$G_a(X_{\text{kij}}) = \sum_{l=1}^{m} g(X_{\text{kij}}, T_l, V_l) + a_k, \quad a_k \sim N(0, \tau^2), \tag{2}$$

where T_l is the lth tree structure and $V_l = \{\mu_{1l}, \ldots,\mu_{b_l l}\}$ is the set of b_l terminal node parameters associated with tree structure T_l. To elaborate, each binary tree T_l is made up of both internal nodes and terminal nodes. At each internal node, there is a decision rule that splits estimation of the mean of $G_a(X_{kij})$ depending on the covariates X_{kij}. Using a hypothetical or toy example (Fig. 5), the first internal node at the top of the tree drops the mean to the left if the corresponding covariate $X_{kij2} < 100$ or to the right if $X_{kij2} \geq 100$. At a terminal node (a node with no decision rules to split an outcome), the sample mean of the outcomes allocated to the terminal node can be calculated to obtain the

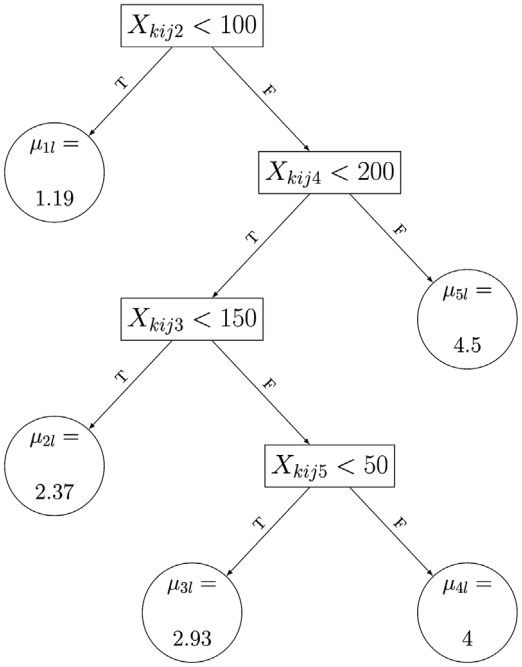

Fig. 5. An example of a regression tree where μ_{pl} is the mean of the pth node for the lth regression tree.

parameter μ_{pl} at the terminal node, $p = 1, \ldots, b_l$. Thus, $g(X_{kij}, T_l, V_l)$ can be viewed as the lth function that assigns the mean μ_{pl} to the outcome, $G_a(X_{kij})$.

2.2.1. Prior distribution

Assuming that $(T_l, V_l) \perp \tau^2$ for all l and the terminal node parameters μ_{pl} are independent of each other within each lth tree, we then have

$$\begin{aligned} P[(T_1, V_1), \ldots, (T_m, V_m), \tau] &= [\textstyle\prod_{l=1}^{m} P(T_l, V_l)]P(\tau) \\ &= [\textstyle\prod_{l=1}^{m} P(V_l|T_l)P(T_l)]P(\tau) \\ &= \left[\textstyle\prod_{l=1}^{m} \left\{\prod_{p=1}^{b_l} P(\mu_{\mathrm{pl}}|T_l)\right\} P(T_l)\right] P(\tau). \end{aligned}$$

This decomposition implies that we need to impose priors on T_l, μ_{pl}, and τ. According to Chipman et al. (2010), the prior distribution $p(T_l)$ can be specified using three aspects: (i) the probability that a node at depth $d = 0, 1, 2, \ldots$ is an internal node given by

$$\alpha(1 + d)^{-\beta}, \quad \alpha \in (0, 1), \beta \in [0, \infty).$$

where α controls how likely a terminal node in the tree would split and β controls the number of terminal nodes, (ii) a uniform distribution on how to choose the covariate to be allocated as the decision rule in a internal node of the regression tree T_l, and (iii) a uniform distribution on how to choose the value of the covariate to be allocated as the decision rule in a internal node. For the priors on μ_{pl} and τ, we set them as

$$\mu_{\mathrm{pl}}|T_l \sim N(\mu_\mu, \sigma_\mu^2),$$
$$\tau^2 \mathrm{IG}(1, 1)$$

where $IG(\alpha, \beta)$ is the inverse gamma distribution with shape parameter α and rate parameter β.

2.2.2. Posterior distribution

The posterior draws of T_l, μ_{pl}, a_k and τ can be obtained by employing the technique of data augmentation (Albert and Chib, 1993, 1996; Tanner and Wong, 1987) where we first draw a latent variable Z_{kij} according to

$$Z_{\mathrm{kij}}|y_{\mathrm{kij}} = 1 \sim N_{(0,\infty)}(G_a(X_{\mathrm{kij}}), 1)$$
$$Z_{\mathrm{kij}}|y_{\mathrm{kij}} = 0 \sim N_{(-\infty,0)}(G_a(X_{\mathrm{kij}}), 1).$$

where $N_{(a,b)}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2 truncated at (a, b). Next, we remove a_k from Z_{kij} to obtain $\widetilde{Z}_{\mathrm{kij}} = Z_{\mathrm{kij}} - a_k$. Then, using BART with the $\widetilde{Z}_{\mathrm{kij}}$ as the outcome, X_{kij} as the independent predictors, and the error in the model set as 1, we sample from the posterior distribution of T_l and V_l. After which, we sample the posterior draws for a_k and τ^2 from $N\left(\frac{\tau^2 \sum_{k,i}(Z_{\mathrm{kij}} - \hat{Z}_{\mathrm{kij}})}{n_k\tau^2 + 1}, \frac{\tau^2}{n_k\tau^2 + 1}\right)$ and $\mathrm{IG}\left(\frac{K}{2} + 1, \frac{\sum_{k=1}^{K} a_k^2 + 2}{2}\right)$ respectively, where $\hat{Z}_{\mathrm{kij}}$ is a posterior draw of $\sum_{l=1}^{m} g(X_{\mathrm{kij}}, T_l, V_l)$ and n_k is the number of turns each driver k made in the dataset. After obtaining the posterior draws of T_l, M_l, and a_k, we compute $G_a(X_{kij})$ and then sample a new latent variable Z_{kij} for the next Monte Carlo Markov Chain (MCMC) iteration. Explicit details of how the posterior draws are done can be found in Tan et al. (2016).

2.2.3. Prediction evaluation

To evaluate riBART, we plotted the AUC value and its 95% confidence interval (CI) at every jth meter before the center of the intersection. We also plotted the AUC value of BART and linear logistic regression together with their corresponding 95% CI as a comparison. AUC calculates the proportion of observed outcomes that were ranked higher in terms of their predicted probability compared to the observed non-outcomes. Thus, a value close to 1 indicates that the prediction model is performing much better than chance while a value close to 0.5 indicates that the prediction model performs no better than chance. We computed the CI of the AUC using the method of Hanley and McNeil (1982), which uses a linear approximation of the AUC to the Somer's D statistic to obtain an estimate of the variance of AUC.

In addition to the AUC, we plotted the sensitivity (the y-axis of the Receiver Operating Characteristic [ROC] curve) profile and the FPR (the x-axis of the ROC curve) profile for riBART. We plotted both profiles at nine different predicted probability cut-offs. Plotting the sensitivity and FPR profile allows us to investigate how each different predicted probability cut-off level will affect the probability of an unnecessary stop by the CAV and the probability of a crash between the CAV and a human driven vehicle.

All analyses were done in R 3.3.1.

3. Results

Our dataset contained 1823 left turns, with 894 of these turns started on arterials, 613 started on collectors, and 316 were started on local roads. We also found 812 eventual stops defined as $s_{ij} \leq 1$ m/s for some $j \in \{-100, \ldots, -1\}$. The average speed in all turns was 10.5 m/s with a standard deviation of 4.2 m/s and each driver took about 17 left turns (mean 16.9, standard deviation 10.8).

We determined the length of our moving window w by using a 10-fold cross validation AUC (cvAUC) with the first 3 PCs as the variables and BART as the prediction model. We compared the cvAUC profiles with w from 3 to 50. Fig. 6 shows the results of w from 3 to 10. (We did not present the cvAUC profiles of w from 11 to 50 because they were all below the cvAUC profiles of w from 3 to 10.) We chose a window length of 6 because the 10-fold cvAUC profile was higher compared to window lengths of 3–5 from −95 m to −30 m. Similarly, for distances more than −30 m, the cvAUC of window length 6 was more than that of window lengths 7–10.

To decide which PCs to include in our model, we considered the proportion of variance of vehicle speed explained by the PCs and the gain in prediction obtained by adding each PC into the model. We measured gain in prediction by looking at the AUC produced by BART as each PC is included one by one starting from the PC1. From Fig. 7, we can see that when we include PC2 into our model, the AUC profile increases substantially. Although PC3 seemed to reflect the change in

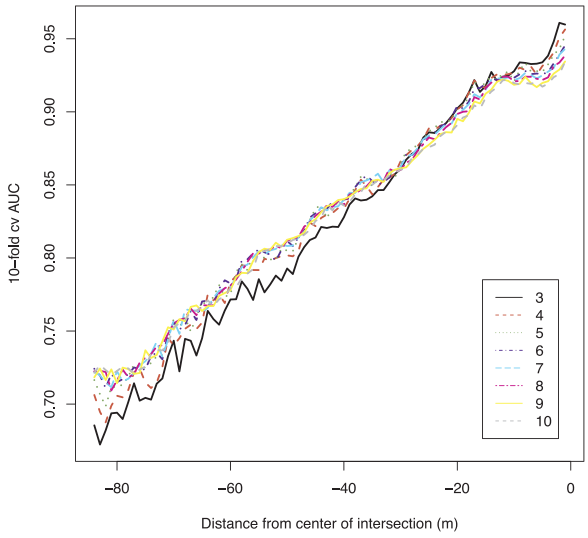

Fig. 6. Predictive value of model using 10-fold cross validation AUC profiles for window lengths 3–10.

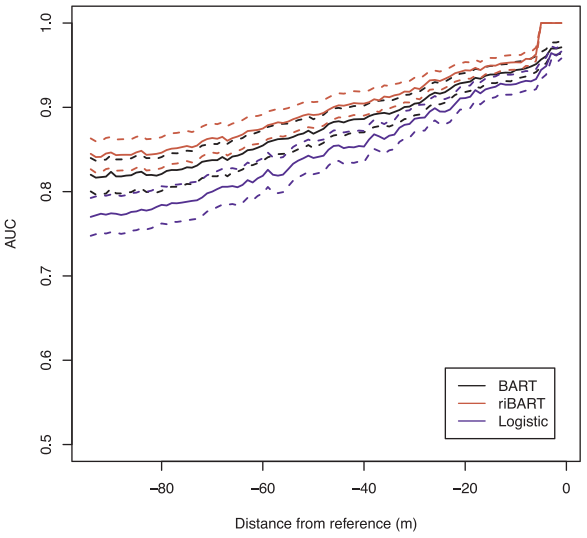

Fig. 8. Area under the receiver operating characteristic curve (AUC) profile of riBART under $\tau^2 \sim IG(1, 1)$, BART, and linear logistic regression. Dotted lines indicate 95% point wise credible interval.

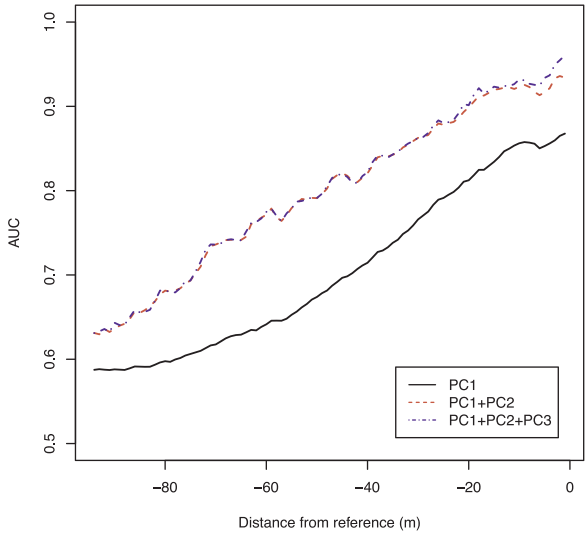

Fig. 7. Comparing the area under the receiver operating characteristic curve (AUC) profile gains of including each principal component (PC) in the logistic regression model.

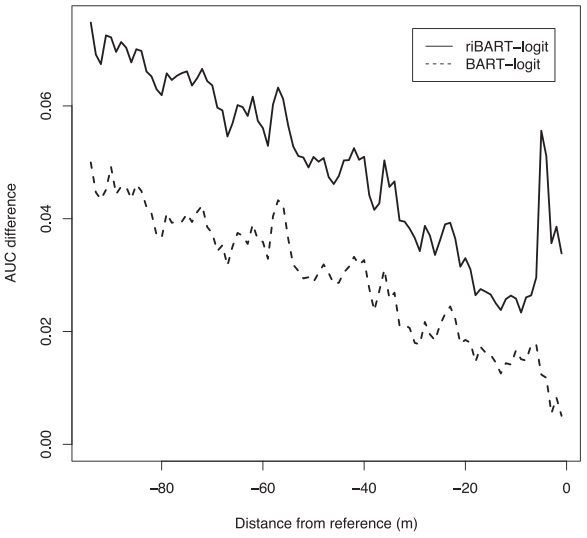

Fig. 9. Comparing the AUC difference profile of riBART under $\tau^2 \sim IG(1, 1)$ and linear logistic regression (logit) versus BART and linear logistic regression.

acceleration of a vehicle (results not shown), addition of PC3 into our model did not seem to increase the prediction performance of our model. Hence, we used the first two PCs as the predictors in the model. We added five more predictors to adjust for the road characteristics and observable conditions within the vehicle. These predictors are:

- The number of times the vehicle has stopped up to the current location to adjust for the likely correlation of stopping behavior within each turn;
- The type of road leading to the center of an intersection;
- Turn signal use;
- Presence of lead vehicle; and
- Brake use.

Other than the type of road leading to the center of an intersection, the rest of the predictors were distance-varying.

To compare the predictive ability of riBART with existing BART and linear logistic regression approaches, Fig. 8 gives the AUC profiles of riBART under $\tau^2 \sim IG(1, 1)$, BART, linear logistic regression, while Fig. 9 shows the AUC profile difference between riBART and linear logistic regression versus BART and linear logistic regression. For the AUC profile, we see that riBART performs slightly better than BART, and both models in turn perform better than the linear logistic regression model for our dataset. BART produced an AUC estimate of about 0.82 with an estimated 95% CI of (0.80, 0.84) at 94 m away from the center of an intersection (−94 m). For riBART, the AUC was 0.84 [95% CI (0.83, 0.86)] at −94 m. The difference in AUC profile between riBART and linear logistic regression was higher compared to the AUC difference profile of BART and linear logistic throughout the left turn maneuver suggesting that riBART performed better than BART.

Fig. 10 shows the sensitivity and FPR profiles of riBART under nine different predicted probability cut-offs, 10%, 20%, …, 90%. By a x%

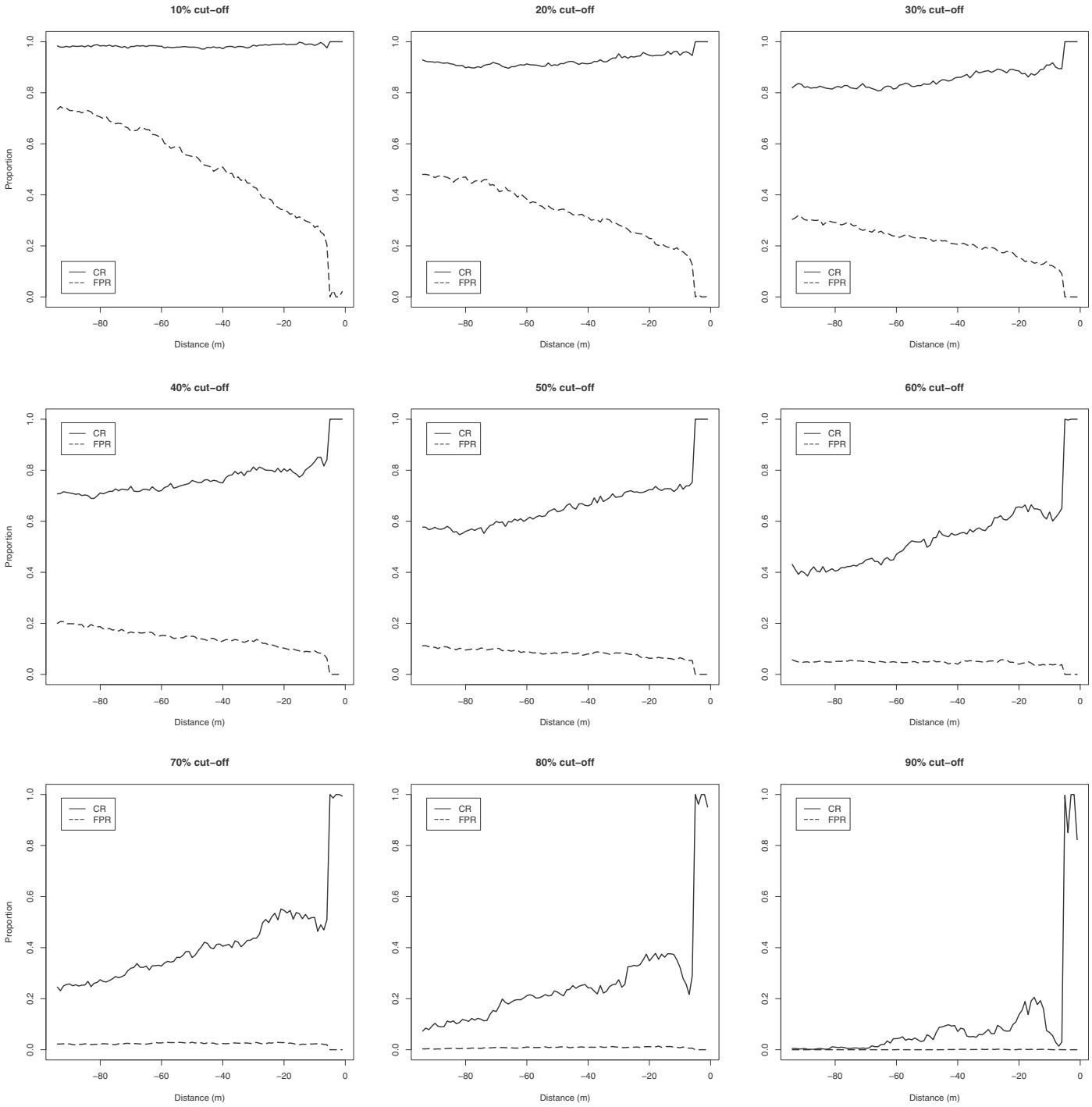

Fig. 10. Sensitivity and false positive ratio (FPR) profiles under nine different predicted probability cut-offs for riBART with PC1 and PC2 as the independent predictors.

predicted probability cut-off we mean that for any predicted probability produced by the riBART prediction model, those that were more than x % were labeled as stops and those that were less than or equal to x% were labeled as non-stops. The sensitivity thus looks at the proportion of actual stops that were labeled correctly as stops and the FPR looks at the proportion of non-stops that were labeled incorrectly as stops. The solid lines in Fig. 10 represent the sensitivity and is equivalent to the CAV correctly predicting that the human driven vehicle would stop using our riBART model with the particular predicted probability cut-off. The dotted lines represent the FPR and is equivalent to the CAV incorrectly predicting that the human driven vehicle would stop and hence a crash with the human driven vehicle would occur. Hence, the ideal prediction would have sensitivity values close to 1 (so that the driverless vehicle does not stop unnecessarily) and FPR values close to 0 (so that the driverless vehicle will not fail to stop, potentially avoiding a crash). We can see that as the predicted probability cut-off increases, both sensitivity and FPR decreases and when the vehicle reaches the center of an intersection, sensitivity approaches 1 and FPR approaches 0 regardless of the probability cut-offs.

4. Discussion and conclusion

In this study, we showed how we could use the kinematic behavior of speed from a human driven vehicle to predict the human driver's decision of stopping before executing a left turn. We also demonstrated the importance of the driver random effect, thus indicating that a driver-specific intent model (at least with a driver-specific intercept) performs best. To effectively implement such a model, it would need to be evaluated on the human driver's vehicle and communicated to nearby CAVs via V2V communication. We employed a moving window of vehicle speeds to capture relevant information for prediction and used PCA to reduce the number of variables in our model. We then employed a model we recently proposed, riBART, to link our distance-varying outcome to the PC variables. Finally we evaluated our prediction model by plotting the AUC, sensitivity, and FPR profiles.

Six meters of speed data at each jth meter away from the center of intersection gave us good cvAUC performance both near and far from the center of intersection. We used the first two PCs as the covariates in our prediction model because they explained at least 99% of the variation in the speed data at each jth meter. We also included the number of times the vehicle has stopped up to the current location as a predictor to adjust for the likely within turn correlation, as well as road type, turn signal use, presence of a lead vehicle, and brake use as additional predictors of future stopping. Our riBART model produced an AUC of 0.84 at -94 m away from the center of intersection and this value increased steadily to 1 as the vehicle approaches the center of intersection.

Because our proposed riBART prediction model performed substantially better compared to BART without a random intercept, and linear logistic regression, this implies that knowing the propensity to stop of a driver is important when trying to decide whether a human driver would stop before executing a left turn. In addition, there is significant impact of non-linear and complex interactions on the likelihood of a vehicle stopping at an intersection before executing a left turn. This conclusion follows from our results because riBART takes into account non-linear and complex interaction effects as well as clustering in our dataset. The random intercept of our model can be viewed as the estimate for a driver's propensity to stop before executing a left turn. To practically implement our results, our algorithm could first be used to estimate the propensity to stop before executing a left turn for a given driver apriori and have this stored in the driver's vehicle. When the vehicle is being driven, they could then emit their propensity to stop before executing a left turn to the CAV. The CAV can then include this information into their prediction of whether the vehicle would stop before executing a left turn.

When we converted the time series of vehicle speed to a distance series, we could have used a more sophisticated method to determine the vehicle speed. Some examples are linear interpolation, non-linear interpolation, and smoothing splines. However, we chose not to employ any of these methods because the likely loss in precision of estimating the speed would be at most 0.01–0.02 m/s. Such a small loss in the precision of the vehicle speed is unlikely to influence the final results.

We also considered many alternative statistical methods which we have not presented. For the use of a moving window of speeds, we originally employed a long window of vehicle speeds where at each meter, we kept increasing the window of speeds we considered. This corresponded to the definition of M_j as

$$M_j = \begin{bmatrix} s_{1,-100} & \cdots & s_{1j} \\ s_{2,-100} & \cdots & s_{2j} \\ \vdots & \vdots & \vdots \\ s_{n,-100} & \cdots & s_{\mathrm{nj}} \end{bmatrix}.$$

We found that the predictive ability of this model as measured by AUC was not as good as the prediction obtained by using a moving window. We also tried to see if taking a weighted combination of moving window and long window definition would improve the AUC profile but it did not. Hence, we concluded that after accounting for recent speeds, past speeds essentially amounted to “noise”, reducing the model's prediction performance.

In addition, we considered using speed and acceleration in place of the first and second PC, given their resemblance to these quantities. However, we found that the resulting AUC profile was substantially lower compared to the AUC profile from using the first and second PC. Another alternative to using the first two PCs was the direct use of the 6 m of speed as variables in the BART model. The rationale behind this is we can view the first two PCs as linear combinations of the 6 meters of speed since $X_{j(q)} = M_j u(j)_{(q)}$. So the use of the first two PCs and the 6 meters of speed data would be “similar”. In addition, PCA involves matrix multiplications which could slow down computation when the number of observations increase. Unfortunately, this alternative method does not produce an AUC profile better or comparable to the AUC profile produced using the first two PCs. We suspect the reason is PCA further extracts useful information from the 6 meters of speed data compared to simple summary statistics of average speed and acceleration. Moreover, using all the information from the 6 meters of speed data may add noise instead of signal.

Inspection of the sensitivity and FPR profile plots suggests that different decision rules should be defined at different distances, depending on the cost we decide to allocate to either correctly predicting a driver stop and hence avoid unnecessary stops in the CAV, or incorrectly predicting a driver stop and hence a crash with the human driven vehicle would occur. To obtain the different optimal cut-offs that would balance the sensitivity and FPR at each distance, we suggest attaching different costs to the sensitivity and FPR at each j and then employ numerical methods to solve for the optimal cut-off.

Finally, while we have focused on prediction of left turns as a particularly important and complex decision for CAVs to make about human drivers, this methodology could be adapted and employed to other driver behaviors, including right turns, stopping at stop lights or stop signs, and lane changes.

Acknowledgments

This work was supported jointly by Dr. Michael Elliott and an ATLAS Research Excellence Program project awarded to Dr. Carol Flannagan. This work was also funded in part by the Toyota Class Action Settlement Safety Research and Education Program. The conclusions are those of the authors and have not been sponsored, approved, or endorsed by Toyota or plaintiffs' class counsel. We would like to thank Dr. Jian Kang and Dr. Brisa Sánchez for their valuable suggestions. The authors would also like to thank the Editor-in-Chief and two referees for their many helpful comments that significantly improved this article.

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. 88, 669–679.

Albert, J., Chib, S., 1996. Bayesian modeling of binary repeated measures data with application to crossover trials. In: Berry, D.A., Stangl, D.K. (Eds.), Bayesian Biostatistics. Marcel Dekker, New York.

Chipman, H., George, E., McCulloch, R., 2010. BART: Bayesian Additive Regression Trees. Ann. Appl. Stat. 4, 266–298.

Choi, E., 2010. Crash Factors in Intersection-Related Crashes: An On-Scene Perspective DOT HS 811 366. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811366.pdf (17.11.16).

CNNMoney, 2015. Injuries in Google Self-Driving Car Accident. Retrieved from: http://money.cnn.com/2015/07/17/autos/google-self-driving-car-injury-accident/ (26.08.15).

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 233–240.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.

Google, 2015. What We're Up To. Retrieved from: http://www.google.com/selfdrivingcar/ (26.08.15).

Hanley, J., McNeil, B., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.
National Highway and Traffic Safety Administration, 2013. U.S. Department of Transportation Releases Policy on Automated Vehicle Development. Retrieved from: http://www.nhtsa.gov/About+NHTSA/Press+Releases U.S.+Department+of +Transportation+Releases+Policy+on+Automated+Vehicle+Development (26.08.15).
Sayer, J., Bogard, S., Buonarosa, M., LeBlanc, D., Funkhouser, D., Bao, S., Blankespoor, A., Winkler, C., 2011. Integrated Vehicle-Based Safety Systems Light-Vehicle Field Operational Test Key Findings Report DOT HS 811 416. . Retrieved from: http://www.nhtsa.gov/DOT/NHTSA/NVS/Crash%20Avoidance/Technical%20Publications/2011/811416.pdf (26.08.15).
Tan, Y., Flannagan, C., Elliott, M., 2016. Random Intercept Bayesian Additive Regression Trees: Application to Longitudinal Prediction. arXiv page 1609.07464.
Tanner, M., Wong, W., 1987. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 82, 528–540.