Transportation Research Part F 58 (2018) 25–38

Contents lists available at ScienceDirect

Transportation Research Part F

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

Using topic modeling to develop multi-level descriptions of naturalistic driving data from drivers with and without sleep apnea

Elease J. McLaurin [a,*], John D. Lee [a], Anthony D. McDonald [a,1], Nazan Aksan [b], Jeffrey Dawson [c], Jon Tippin [b], Matthew Rizzo [d]

[a] *University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, United States*
[b] *University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, United States*
[c] *University of Iowa, 145 N. Riverside Drive, Iowa City, IA 52242, United States*
[d] *University of Nebraska Medical Center, 42nd and Emile, Omaha, NE 68198, United States*

ARTICLE INFO

Article history:
Received 11 April 2017
Received in revised form 10 March 2018
Accepted 16 May 2018

Keywords:
Machine learning
Topic modeling
Driver behavior
Drowsiness
Sleep apnea
Naturalistic driving data

ABSTRACT

One challenge in using naturalistic driving data is producing a holistic analysis of these highly variable datasets. Typical analyses focus on isolated events, such as large g-force accelerations indicating a possible near-crash. Examining isolated events is ill-suited for identifying patterns in continuous activities such as maintaining vehicle control. We present an alternative approach that converts driving data into a text representation and uses topic modeling to identify patterns across the dataset. This approach enables the discovery of non-linear patterns, reduces the dimensionality of the data, and captures subtle variations in driver behavior. In this study topic models were used to concisely described patterns in trips from drivers with and without untreated obstructive sleep apnea (OSA). The analysis included 5000 trips (50 trips from 100 drivers; 66 drivers with OSA; 34 comparison drivers). Trips were treated as documents, and speed and acceleration data from the trips were converted to "driving words." The identified patterns, called topics, were determined based on regularities in the co-occurrence of the driving words within the trips. This representation was used in random forest models to predict the driver condition (i.e., OSA or comparison) for each trip. Models with 10, 15 and 20 topics had better accuracy in predicting the driver condition, with a maximum AUC of 0.73 for a model with 20 topics. Trips from drivers with OSA were more likely to be defined by topics for smaller lateral accelerations at low speeds. The results demonstrate topic modeling as a useful tool for extracting meaningful information from naturalistic driving datasets.

1. Introduction

Crash statistics and self-reported data show that drivers with untreated obstructive sleep apnea (OSA) are at an increased risk of involvement in a motor vehicle crash (Tregear, Reston, Schoelles, & Phillips, 2009; Williamson et al., 2011). These data sources provide details on the safety risk of OSA in naturalistic settings, however, they do not capture the subtle changes in

[*] Corresponding author.
E-mail address: emclaurin@wisc.edu (E.J. McLaurin).
[1] Present address: Texas A&M University, 3131 TAMU, College Station, TX 77845, United States.

https://doi.org/10.1016/j.trf.2018.05.019
1369-8478/© 2018 Elsevier Ltd. All rights reserved.

driver behavior that may be responsible for the increased crash risk. As a result, assumptions must be made about the connection between the OSA condition and crash risk. The prevailing understanding is that the increased crash risk is due to increased driver drowsiness (Tippin, Aksan, Dawson, & Rizzo, 2013; Williamson et al., 2011). The chronic narrowing or closure of the upper airway and associated disordered breathing causes recurrent partial awakenings during the sleep of OSA patients (American Academy of Sleep Medicine (AASM), 2014). As a result, drivers with OSA are prone to higher incidences of daytime drowsiness.

1.1. Drowsiness effects in controlled environments

Drowsiness has been shown to undermine driver performance in controlled settings, such as a driving simulator (Contardi, Pizza, Sancisi, Mondini, & Cirignotta, 2004; Thiffault & Bergeron, 2003; Turkington, Sircar, Allgar, & Elliott, 2001; Yuan, Du, Qu, Zhao, & Zhang, 2016). Drowsy drivers have delayed sensory processing ability and perception, need longer periods to react to external stimuli, and have substantially degraded ability to control their vehicle (Eskandarian, Mortazavi, & Sayed, 2012). In understanding the effects of drowsiness on vehicle control, steering wheel movements have emerged as one of the more sensitive measures of the effect of drowsiness on driving behavior (Krajewski & Sommer, 2009; McDonald, Lee, Schwarz, & Brown, 2018; McDonald, Lee, Schwarz, & Brown, 2013). The effects of drowsiness on steering behavior include: "fewer small, smooth steering adjustments (micro-corrections), more zigzag and slow oscillation, greater steering entropy (measure of steering variability), larger erratic steering movements (indicating e.g. overcorrecting for unanticipated road changes), lateral drift outside the driver's comfort zone, and more large and fast steering corrections" (Krajewski & Sommer, 2009). These steering behaviors are further reflected in the lateral movement of the vehicle, resulting in the common use of lateral movement measures in drowsiness detection simulator studies (Forsman, Vila, Short, Mott, & Van Dongen, 2013).

1.2. Drowsiness effects in naturalistic environments

Although they lack the precise control of simulator studies, naturalistic driving studies can reveal a link between driver behavior and safety outcomes. In naturalistic driving studies, video and data recorders in participants' vehicles unobtrusively record behavior and environments during normal driving (Bärgman, Lisovskaja, Victor, Flannagan, & Dozza, 2015; Klauer, Dingus, Neale, Sudweeks, & Ramsey, 2006). These studies generate large amounts of heterogeneous data, which make it difficult to identify comparable scenarios and extract a holistic understanding of driver behavior. A typical approach to address this challenge is to identify and compare the relative frequency of data segments which have similar kinematic or physiological signatures, for example extracting the five to ten seconds of data surrounding a large vehicle acceleration value or the prolonged closure of a participant's eyes (Barr, Yang, Hanowski, & Olson, 2005; Dozza, Bärgman, & Lee, 2013; Guo & Hankey, 2009; Hanowski, Wierwille, & Dingus, 2003; Rau, 2005; Wu & Jovanis, 2013). This approach has been particularly useful in yielding insights regarding common activities that precede crashes and near-crashes (Carney, Harland, & McGehee, 2016; Dingus et al., 2016). Using this approach, naturalistic driving studies investigating the effects of drowsiness on driving found drowsy drivers have a higher crash risk and tend to develop "tunnel vision", where they neglect to scan areas outside of the forward roadway (Barr et al., 2005; Klauer et al., 2006). A naturalistic study investigating the effects of sleep fragmentation (the amount and duration of awakenings after sleep onset) on driving found that for days following high levels of fragmented sleep, drivers with untreated OSA displayed more severe physical symptoms of drowsiness while driving, even while navigating intersections, and made more safety errors during drowsy episodes as compared to drivers without OSA (Aksan, Dawson, Tippin, Lee, & Rizzo, 2015). Another study using the same dataset found that short periods (3–30 s) of speed and acceleration data captured driving patterns that distinguished drivers with different levels of OSA treatment adherence (McDonald et al., 2017).

1.3. Limitation of typical naturalistic driving data analysis approach

The results from prior work suggest that drowsiness affects driving performance in both simulated and naturalistic environments. However, in contrast to the simulator studies, the naturalistic studies have tended to treat drowsiness as a transient event by focusing on identifying isolated periods where drivers exhibited symptoms of drowsiness. In the case of drivers with untreated OSA, drowsiness is often a chronic condition (American Academy of Sleep Medicine (AASM), 2014). As a result, these drivers may continuously exhibit driving behavior that differs from the behavior of drivers without the condition. To determine the presence of more pervasive, and perhaps more subtle differences in driving behavior requires a different approach than the critical event method typically used to analyze naturalistic driving data. The objective of this study is to explore the use of an alternative approach for analyzing naturalistic driving data to identify trip-level patterns in the behavior of drivers with OSA. The approach used for the analyses described in this paper differs from the critical events method in that it describes all the data from a driver's trip instead of isolated periods of the trip. The approach also enables the discovery of unknown driving patterns by not requiring the analyses to have pre-determined the features that will be extracted from the raw data (e.g., pre-determine to extract all instances where g-force exceeded a threshold).

While a variety of data are collected during naturalistic driving studies, the analyses in this paper focus on the use of vehicle kinematics data. The vehicle kinematics data collected during naturalistic driving studies provide a measure of driver behavior in terms of vehicle control (Fuller, 2005; Reymond, Kemeny, Droulez, & Berthoz, 2001). Because these data are often collected continuously, they offer an opportunity to analyzed normal driving behavior over an extended period. The data provide information on what driving activities occurred, when they occurred, and where they occurred. With a driver sample that travels in similar roadway environments at similar times (e.g., a sample of rural drivers with regular work schedules), the information on driving activity timing and location can be removed and the analysis of driver behavior can focus on what activities occurred in terms of how the vehicle was handled. In removing the time and spatial order of the data, the data are effectively converted from continuous spatiotemporal series to variables with discrete values. These variables can then be compared to identify patterns of interest.

1.4. Changing from critical events to topics

To discover long-range patterns in large volumes of highly variable driving data, the data needs to be transformed in a way that reduces the volume and provides structure, without subsetting the dataset (Wang et al., 2013). One common approach for such a data transformation is to summarize the distribution of the data using measures of central tendency. This approach provides simple metrics for comparison; however, it masks the subtle variability that likely constitutes the patterns of interest. An alternative approach is to use linear or non-linear dimensional reduction methods to represent the data using fewer parameters (Van Der Maaten, Postma, & Van Den Herik, 2009). Many of the driving patterns identified in controlled settings that were associated with drowsiness represent non-linear changes in vehicle control (Krajewski & Sommer, 2009). As a result, non-linear dimensional reduction methods may be well suited to represent the driving data in a way that preserves the patterns of interest.

The interpretability of the identified patterns is a key factor in using the results of a naturalistic driving data analysis to better understand driver behavior. While a number of non-linear dimensional reduction methods exist, the topic modeling method stands out because it is intended to produce interpretable results (Blei, 2012). Topic modeling is an unsupervised dimensional reduction method designed to identify word co-occurrence patterns, called latent topics (Steyvers & Griffiths, 2007). In text applications, topics provide a meaningful overall description of a corpus of documents as well as a description of the individual documents. For example, topic modeling has been applied to the thousands of papers published in the journal *Science* to describe the papers according to their topic areas and to identify the domains of investigation that have defined science over the last 100 years (Blei & Lafferty, 2007). Fruitful topic modeling applications beyond text analyses include population genetics, computer vision, and human activity recognition (Blei, 2012; Farrahi & Gatica-Perez, 2011; Huynh, Fritz, & Schiele, 2008).

Topic models require a text representation of data, so continuous driving data need to be transformed to a discrete "textual" representation. The bag-of-words text representation is particularly advantageous because it preserves much of the local variation in the data while also regularly outperforming other representations when used in subsequent analyses such as classification or clustering (Baydogan, Runger, & Tuv, 2013; Lin, Keogh, Wei, & Lonardi, 2007; Lin & Li, 2009). Driving data are represented as a bag-of-words by segmenting the data and encoding the segments as symbols or "words". The application of topic modeling to driving data requires a shift in perspective from documents and words to trips and driving-related variables encoded as driving words. In this context, topic models identify driving topics representing common trip elements analogous to how the models identified themes across a set of articles from Science (Blei & Lafferty, 2007). Common elements across trips could include, for example, the vehicle kinematics associated with driving on a highway or having to stop. By using topic modeling, it is possible to shift from a focus on isolated data segments to describing entire trips without relying on measures of central tendency that mask the variability in behavior. Instead, the topic models describe trips as an aggregate of units of behavior that occur at the level of seconds. Such a perspective can reveal previously unrecognized patterns of behavior.

This paper describes the use of topic modeling to compare how drivers with and without OSA manage the control of their vehicle. Two different analyses were conducted using speed, lateral acceleration, and longitudinal acceleration data from a naturalistic study of drivers with OSA and comparison drivers (Aksan et al., 2015). The first analysis examined how the number of topics included in the topic models affects what patterns are identified in the trips. The second analysis used the topics in a random forest prediction model to assess to what degree the topics distinguish drivers with OSA from comparison drivers. Together, these analyses show the utility of moving beyond considering only critical events when using naturalistic data to understand driver behavior.

2. Methods

2.1. Data description

The study involved naturalistic observations of OSA patients driving their own vehicles equipped with an in-vehicle data acquisition system (IV-DAS) for a two-week period before and three-months after beginning CPAP (Continuous Positive Airway Pressure)-therapy. Comparison drivers were observed for the same three and one-half month duration as the OSA drivers. OSA drivers met ICSD-2 clinical criteria for OSA and had a Respiratory Distress Index (RDI) > 15, while comparison drivers had no sleep complaints and an RDI < 5 as confirmed by an overnight sleep study (American Academy of Sleep

Medicine (AASM), 2014). They were matched with OSA drivers at the group level on age within five years, education within two years, distribution of gender, and county of residence for balancing rural vs. urban driving.

The IV-DAS contained an internal camera cluster, GPS, OBD-II (for vehicle speed), and three-axis accelerometers. Video data collection was triggered intermittently based on accelerometer exceedances and a baseline data collection schedule. The GPS, OBD-II, and accelerometer data were continuously recorded at 10 Hz. These data were partitioned into individual drive files defined by ignition start and stop (Aksan et al., 2015).

For this analysis, the driver sample included 100 drivers; 66 OSA drivers (22 female, age M = 47.4 years, SD = 7.6) and 34 comparison drivers (16 female, age M = 44.0 years, SD = 8.5). The ratio for comparison drivers to OSA drivers reflected the distribution of participants in the broader study. The study was conducted in the United States; participants drove mostly in Iowa and the surrounding states including Minnesota, Wisconsin, Illinois, Missouri, Nebraska, and South Dakota. Data from the two-week period before CPAP-therapy were used from the OSA drivers along with the first two-week period of data collected from the comparison drivers. This pre-CPAP sample provided data corresponding to the time when performance differences between OSA and comparison drivers would be the greatest, as CPAP therapy was expected to mitigate adverse effects of OSA.

2.2. Processing of driving data

A minimum trip length threshold was used to reduce the inclusion of partial trips in the analysis. The threshold of 5 minutes was decided as it is a common minimum trip length reported by travelers (Bricka & Bhat, 2006). To prevent skewing the analysis towards any one driver due to the varying numbers of trips from different drivers, only the first 50 trips from each driver that met the trip length requirement were included in the analysis, resulting in a total of 5000 trips (trip length: M = 19.2 minutes, SD = 18.6 minutes, minimum = 5 minutes, maximum = 2.9 hours). The sample included the first trips after the installation of the IV-DAS and as a result captured any driver adjustments to the system. Prior naturalistic studies using similar, unobtrusive IV-DAS found that drivers quickly forget about the system and return to their normal driving behaviors (Hanowski et al., 2003; Klauer et al., 2006). As a result, the effects of the system novelty on driver behavior were expected to be small. Driving variables used for the present analysis included speed, lateral acceleration, and longitudinal acceleration. These variables were selected because they are a direct output of vehicle control behaviors (Fuller, 2005; Reymond et al., 2001). Table 1 describes the variables used in the analysis.

2.3. Data reduction and conversion

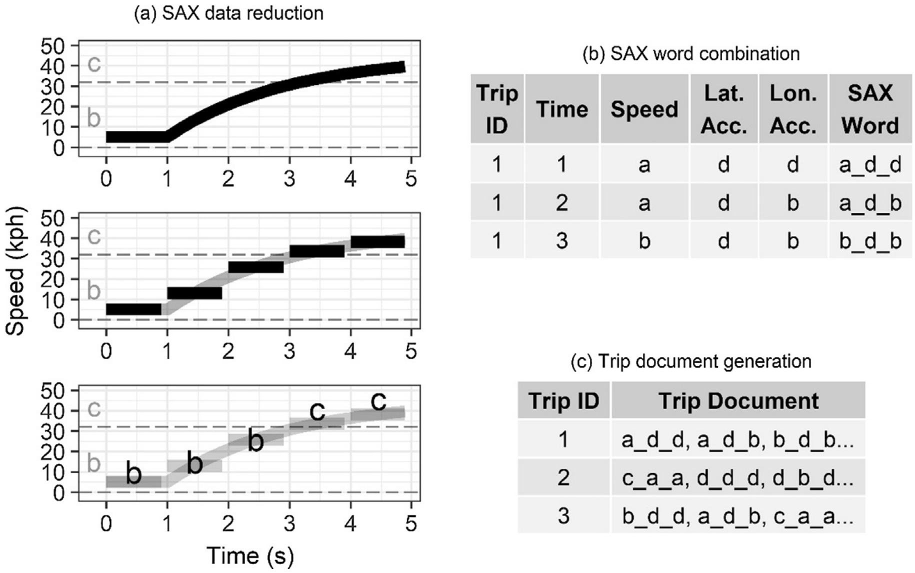
The large size of naturalistic driving datasets demands some type of data reduction. Symbolic Aggregate Approximation (SAX) was used to reduce the data for the analyses presented in this paper. SAX was developed as a generic method for converting time series data to a lower dimensional, less noisy format that retains the general shape of the original data (Lin et al., 2007). While a number of methods exist for reducing the dimensionality of time series, such as Fourier and wavelet analysis, SAX was preferred because the relative distance between values in the original data is preserved, it does not rely on assumptions of stationarity (i.e., that the mean, variance and autocorrelation structure do not change over time) and is robust to outliers (Lin et al., 2007). These properties of SAX made it possible to compare trips based on the details of drivers' behavior rather than simple summary statistics (Lin et al., 2007; McDonald et al., 2013). SAX was used to reduce the dataset to ten percent of its original size (i.e., from one point per 0.1 s to one point per 1 s) and convert the numeric speed and acceleration data into the symbolic form required for fitting the topic models.

Fig. 1a shows the steps of applying SAX to a 10-second segment of speed data. The top plot shows the raw numeric data at 10 Hz. The middle plot shows the aggregation step that reduces the number of data points, in this case from 100 data points to 10, and reduces the number of unique values by grouping them into bins, which are indicated on the y axis. The bottom plot shows the step of converting the numeric values to alphabetic symbols, which in the figure are overlaid on the raw data. The letter sequence, "abbbcccccc" is the final output of SAX for this segment of speed data. To enable multivariate analyses, letter sequences from different data variables are combined to form word sequences. Fig. 1b shows letter sequences from speed, lateral acceleration, and longitudinal acceleration data being combined to form words, each of which indicate the approximate values of the three variables for one second of data. For the present analysis the combination was: (speed)_(lateral acceleration)_(longitudinal acceleration). The order in which the letters are combined does not influence the topic model results so long as it is consistent. The word sequences created from the driving data for each trip comprise the documents that are analyzed using topic modeling as shown in Fig. 1c.

Table 1
Driving variables.

Variable name	Units	Precision
Speed	Kilometers per hour (kph)	1 kph
Lateral acceleration	Gravitational force (g)	0.001 g
Longitudinal acceleration	Gravitational force (g)	0.001 g
Time	Greenwich mean time zone (GMT)	1 s

Note. All variables recorded at 10 Hz.


(a) SAX data reduction

(b) SAX word combination

Trip ID	Time	Speed	Lat. Acc.	Lon. Acc.	SAX Word
1	1	a	d	d	a_d_d
1	2	a	d	b	a_d_b
1	3	b	d	b	b_d_b

(c) Trip document generation

Trip ID	Trip Document
1	a_d_d, a_d_b, b_d_b...
2	c_a_a, d_d_d, d_b_d...
3	b_d_d, a_d_b, c_a_a...

Fig. 1. Illustration of using SAX to reduce time-series data for use in the topic models.

2.3.1. SAX data aggregation process

The speed data were aggregated into one second, non-overlapping segments. The collection rate was 10 Hz, so each segment consisted of ten points. The arithmetic mean was used for the aggregation. Next, the sequence of segment summary values was checked for the presence of consecutive zero values. These zero speed periods were generated by non-driving events, such as a participant sitting in their vehicle for a while before turning off the ignition or idling while waiting for a traffic signal change. Any sections of consecutive zero speed summary values were reduced to a single zero value to retain information on the occurrence of a stop while avoiding the generation of meaningless topics based on the inclusion of many identical driving words associated with a single long period of zero speed.

Like the speed data, the acceleration data were aggregated into one second, non-overlapping segments. To avoid masking high acceleration events, which would hide potentially safety-relevant information, the mean was not used to aggregate the acceleration data. Additionally, in reviewing the acceleration values it was observed that using the maximum absolute values would result in the inclusion of many large spurious values which were not indicative of the driver's behavior. To avoid the inclusion of these values a threshold slightly lower that the maximum acceleration was used for the segment summary. For each segment the value at the 85-percentile (i.e., high positive accelerations) and the 15-percentile (i.e., high negative accelerations) was determined and the larger absolute value was selected as the segment summary. The sign of the selected value was retained.

2.3.2. Bin thresholds

The values used for the bin thresholds are shown in Table 2. The speed bin thresholds approximate typical speed limits for different road functional classes (National Highway Traffic Safety Administration (NHTSA), 2012). The acceleration bin thresholds align with normal, comfortable accelerations for more subtle maneuvers such as coasting or gradually increasing speed to more deliberate maneuvers such as making a turn or coming to a complete stop (Bosetti, Da Lio, & Saroldi, 2014; Rice, 1973; Seppelt & Lee, 2007).

Table 2
Bins for SAX.

Latitude & longitude acceleration		Speed	
Symbol	Range	Symbol	Range
a	<(−0.30) g	a	=0 kph
b	(−0.30) to (−0.20) g	b	<32 kph
c	(−0.20) to (−0.06) g	c	32–73 kph
d	(−0.06) to (−0.03) g	d	73–96 kph
e	(−0.03) to (0.03) g	e	>96 kph
f	0.03–0.06 g		
g	0.06–0.20 g		
h	0.20–0.30 g		
i	>0.30 g		

2.3.3. Driving word meanings

The order of the letters in each driving word corresponds to: (speed)_(lateral acceleration)_(longitudinal acceleration). Based on the possible combination of speed and acceleration letters, the words describe one of four vehicle states: during the one second period described by the word, the word represents either (1) maintaining a constant speed and lateral acceleration, (2) having changes in the vehicle speed only, (3) having changes in the vehicle lateral acceleration only, or (4) having changes in both vehicle speed and lateral acceleration. Words that are in the same vehicle state can have differences in the magnitude of the speed and/or acceleration values. Table 3 shows the word structures that correspond to the different vehicle states. Note the letter "e" corresponds to ~0 g for both lateral and longitudinal acceleration. These driving words provide a direct way to interpret the patterns identified using the topic modeling dimensional reduction method.

2.4. Identifying patterns using topic modeling

Topic modeling was originally developed to summarize the contents of large document sets. The method uses the frequency of co-occurring words to identify document topics. The topics capture patterns in word usage across sub-groups of documents. To apply topic modeling to naturalistic driving data, we use an analogy between the text documents and the numeric data. A driving document consists of all the speed and acceleration data generated during one trip. Therefore, each participant generated many driving documents. The words that comprise a driving document are the summary values for the one second data segments. These values were converted from numbers to strings using SAX. The topics identified by the model describe driver behaviors that tend to co-occur during a trip, as conveyed by vehicle state combinations. Given this analogy, topic modeling was used to identify patterns in the data and develop concise descriptions of each trip.

The topic models were generated using the Latent Dirichlet Allocation (LDA) method (Blei, Ng, & Jordan, 2003). The LDA algorithm is the simplest in the family of algorithms used to generate topic models. In an LDA topic model, the documents are treated as "bags of words" where the order of the words is ignored. Topics are identified using groups of words that co-occur frequently in the documents (Blei, 2012). These topics are formally specified as a probability distribution over the combined vocabulary of words found in all of the documents, where larger probability values are assigned to the vocabulary words that define the topic and near-zero probability values are assigned to the vocabulary words that are not part of the given topic (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009). In addition to identifying which words define the topics, the topic model identifies which topics describe each document. An LDA topic model is a mixed-membership model, so that each document can have multiple topics in different proportions (Blei et al., 2003). The model formally describes each document using a probability distribution over the combined set of topics, where larger probability values are assigned to the topics that describe the document and near-zero probability values are assigned to the topics that do not describe the document. In this way the LDA topic model differs from other common unsupervised machine learning algorithms such as k-means, which assigns each element in a dataset to only one group or cluster (Jain, 2010).

Five topic models were constructed using the R package *topicmodels* (Hornik & Grün, 2011). The models differed in the number of topics; the five models included 2, 5, 10, 15, and 20 topics. As with other machine learning methods that are initialized with random seeds, different runs on the data produced different topic models. Solutions were considered stable when similar sets of word distributions were identified as topics across different runs. To further improve solution stability, the models were fit using ten-fold cross validation and the log-likelihoods were used to select the final models, which are common approaches for fitting topic models (Hornik & Grün, 2011).

2.5. Comparing trip descriptions from different drivers using random forests

Random Forest models were then used to assess whether trips described by topics can differentiate between OSA drivers and comparison drivers. Random Forest models consists of many decision trees trained on random bootstrapped samples of training data, and randomly selected predictor subsets (Breiman, 2001). Classification is performed by a majority vote of the predictions of the trees.

The random forest algorithm was selected because it is robust to noisy predictors and it does not rely on assumptions of linearity (Breiman, 2001). Like beta weights with linear regression models, Random forests produce a measure of variable importance. This measure is based upon the number of times a given predictor appears in the algorithm and its predictive power in each appearance in the decision trees that make up the random forest. The variable importance compensates for the loss of interpretability of a multi-tree algorithm by indicating which variables contribute the most to successful classification.

Table 3
Vehicle states described by the driving words.

Vehicle state	Word structure	Examples
1. Constant speed and no lateral acceleration	(any letter)_(e)_(e)	a_e_e/b_e_e/c_e_e/d_e_e
2. Changed speed only	(any letter)_(e)_(not "e")	d_e_f/d_e_g/d_e_h/d_e_i
3. Experience lateral acceleration only	(any letter)_(not "e")_(e)	c_a_e/c_b_e/c_c_e/c_d_e
4. Changed speed and have lateral acceleration	(any letter)_(not "e")_(not "e")	e_d_f/e_c_g/e_b_h/e_a_i

Ten-fold cross validation was used to assess the expected performance of the models. Due to the repeated measures for each participant in the dataset (i.e., multiple trips for each participant), the data for each participant was grouped within the folds to avoid introducing bias in the performance predictions that would have occurred if the trips from a given participant were included in both the training and testing data.

Random Forest models were fit using R packages *caret* and *randomForest* (Kuhn, 2008; Liaw & Wiener, 2002) with the trip descriptions from the five topic models as features to identify whether the driver on a particular trip was suffering from OSA. If the trip descriptions based on topic models are good predictors, (i.e., the Random Forest models perform well) the topic models would seem to capture important aspects of driver behavior related to the effects of OSA.

3. Results

3.1. Differences in topic models—topic definitions

Topic models with 2, 5, 10, 15, and 20 topics were constructed and compared. The model computation time increased non-linearly with the number of topics, with the two-topic model taking approximately five minutes to run while the twenty-topic model took approximately four days to run using a dual-core, 64-bit Windows PC with 16 GB of RAM. The topic models specified the identified topics as a probability distribution over the combined vocabulary of words found in all the documents (i.e., trips), where larger probability values were assigned to the vocabulary words that define the topic and near-zero probability values were assigned to the vocabulary words that did not define the topic.

The two-topic model was the simplest model and identified the most outstanding characteristics in the trip compositions. In this model, Topic 1 represented the proportion of a trip driven at low to medium speeds (i.e., < 73kph), both when the driver had constant speed and no lateral acceleration as well as when they changed their speed and/or lateral acceleration at medium to low speeds. Topic 2 represented the proportion of a trip driven at high speeds (i.e., 73–96 + kph), both when the driver had constant speed and no lateral acceleration and when they changed their speed and/or lateral acceleration. Fig. 2a shows the five highest probability words associated with each topic in the two-topic model. As shown in the figure, the probability of the driving words differed within each topic. By comparison, a "null" topic would assign equal probability to each of the words. The probability of the driving words also differed across topics, with words that had high probability in one topic having little to no probability in the other topic. In Topic 1 the word with the highest probability is "c_e_e" from vehicle state 1 (constant speed and no lateral acceleration), where the speed is between 32–73 kph, with ~0 g lateral and longitudinal acceleration. The other four words have a similar probability as the highest probability word, are all from vehicle state 2 (changes in speed only), and have speed letters "b" or "c" indicating driving at low to mid speeds, similar to speeds in urban environments. In contrast, of the five words for Topic 2, the probability is concentrated on two words that are both from vehicle state 1 (constant speed and no lateral acceleration). The words "e_e_e" and "d_e_e" reflect driving at mid to high speeds (73–96 kph) similar to speeds on highways, with ~0 g lateral and longitudinal acceleration. The differences in the topics are also reflected in the trips most associated with each topic as shown in Fig. 2b, which compares a trip highly associated with Topic 1 (Topic 1 probability = 0.99) versus one highly associated with Topic 2 (Topic 2 probability = 0.99).

Increasing the number of topics in the models led to more specific topics. The topic definitions shifted from describing general differences in speed (i.e., high speed vs. low speed) regardless of longitudinal and lateral acceleration, as in the two-topic model, to more precisely describing speed ranges as well as describing differences in the lateral and longitudinal accelerations. This change can be seen in the highest probability word for each topic in the different models. These words correspond to one of four vehicle states as indicated in Table 3. In the two-topic model, the highest probability word for both topics corresponded to the vehicle state when the vehicle speed is constant and there is no lateral acceleration. However, as

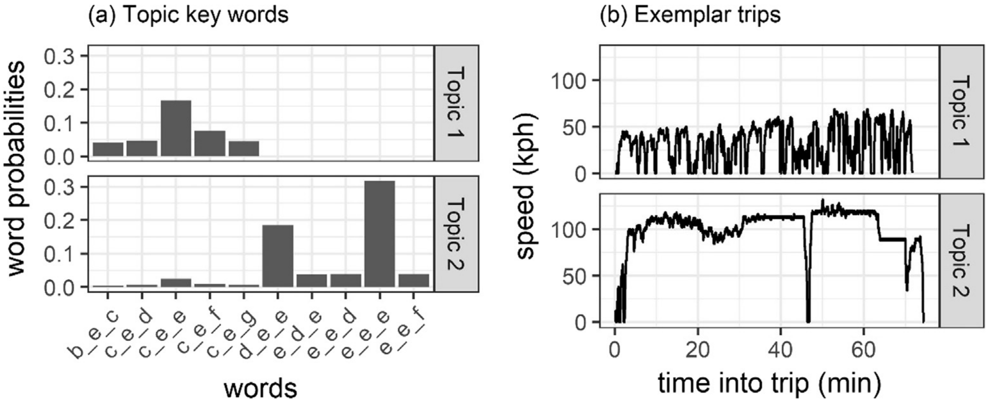


Fig. 2. Comparison of the key words and exemplar trips for each topic in the two-topic model.

shown in Fig. 3, with an increase in the number of topics in the model there was an increase in the diversity of vehicle states represented by the highest probability words.

Increasing the number of topics also led to increasing overlap in the topic definitions. Examining the highest probability words for each topic illustrates this trend. All the highest probability words were unique in the two- and five-topic models. However, the ten-topic model had only eight topics with unique highest probability words, the fifteen-topic model had only twelve topics with unique highest probability words, and the twenty-topic model had only nine topics with unique highest probability words. The increasing overlap in the topic key words suggests the models were identifying smaller variations in the driving word composition of the trips. The diminishing return of adding more topics to improve the model fit can be measured by perplexity, a measure of the accuracy of the predicted word distribution in a held-out sample, which decreases with improved model fit (Blei & Lafferty, 2007; Chang et al., 2009). Fig. 4 shows perplexity declining with the increasing number of topics; the improvement to model fit significantly diminished after the ten-topic model.

3.2. Differences in topic models—driver condition prediction

The topic models described each trip in the naturalistic driving data using a probability distribution over the combined set of topics, where larger probability values were assigned to the topics that described the trip and near-zero probability values were assigned to the topics that did not describe the trip. The set of topic probabilities for each trip were used as predictors in a series of Random Forest (RF) models to identify the condition of the drivers (i.e., comparison or OSA). The performance of the RF models in distinguishing OSA drivers from comparison drivers was assessed using receiver operating characteristic (ROC) curves and the associated area under the curve (AUC) values for each model, which serve as a summary for the ROC curves (Bradley, 1997). As shown in Fig. 5, the RF model performance improved as the number of topics in the topic model increased.

The RF model failed to differentiate the drivers based on trip descriptions from the two-topic model, which only distinguished the amount of driving at low to medium speeds and the amount of driving at high speeds (AUC: .51 ≈ random classification). Using the predictors from the five-, ten-, fifteen-, and twenty-topic models, the RF models identified differences in the driving data for the two groups. As the number of topics increased for the models, they became more specific about the driving patterns they described. As the topic definitions became more specific, the performance of the RF models improved. The rate of performance improvement for the RF models plateaus after the ten-topic model, with the AUC for the twenty-

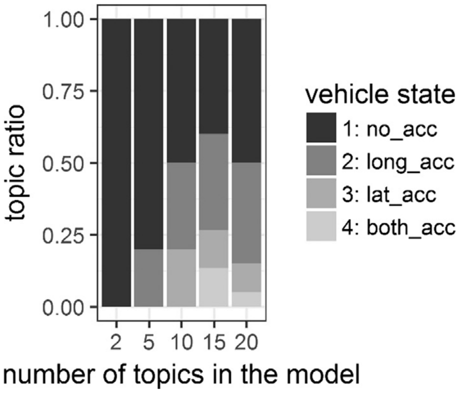


Fig. 3. Ratio of vehicle states represented in each model by the highest probability word in each topic.

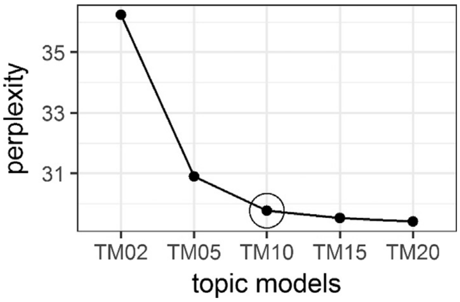


Fig. 4. Scree plot of perplexity for each topic model. Lower values of perplexity indicate better fit.

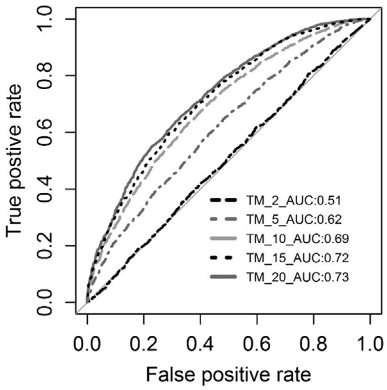


Fig. 5. ROC curves and AUC for the Random Forest models.

topic model being nearly the same as for the ten-topic model (AUC for 10-topics: 0.69; 15-topics: 0.72; 20-topics: 0.73) despite doubling the number of topics.

3.3. In-depth investigation of the ten-topic model

The five topic models were compared in terms of their descriptiveness, based on the topic key words, the fit of the models in terms of perplexity, and the performance of the models in distinguishing drivers with OSA from comparison drivers. Of the five models, the ten-, fifteen-, and twenty-topic models were the best performing in terms of AUC and perplexity. The fifteen- and twenty-topic models were less interpretable than the ten-topic model due to the significant overlap in their topic key words. Given the goal of identifying interpretable patterns, the model interpretability was a deciding factor in the model selection. The ten-topic model provided a detailed description of the data that revealed differences in the data from drivers with and without OSA while limiting the production of extraneous topics with overlapping definitions. Based on these results the model was selected for further investigation.

3.3.1. Ten-topic model topic descriptions

The key words used to define each topic in the ten-topic model are listed in Table 4. To identify these key words, the vocabulary words were ordered by decreasing probability and a cumulative probability distribution was calculated for each topic. The words that accounted for a larger fraction of the total probability for a given topic were selected as the topic's key words. As shown in the table, some topics were defined by only a few words while others are defined by many words. Topics with fewer key words were largely defined by words that indicated a constant speed with little to no lateral or longitudinal acceleration (e.g., "d_e_e", "e_e_e"—state 1 words). These words had higher overall prevalence in the data. Topics with more key words were defined by words that indicated some level of lateral and/or longitudinal acceleration (e.g., "b_c_g", "c_f_c"—state 2, 3 and 4 words). These words had lower overall prevalence in the data.

Fig. 6 shows the overall prevalence in the data of the topics from the ten-topic model. The topic prevalence was calculated by totaling the probability for each topic across all the trips. The topics varied in prevalence, with the mid and low speed topics having higher prevalence than the high-speed topics. Fig. 6 also shows a numeric summary of the topics in terms of the associated speed, lateral and longitudinal acceleration values. This summary was generated by relating the topic key words back to the speed and acceleration combinations they represent. Table 5 shows an example of how the numeric summary was determined for Topic 3. First, each letter in the topic key words was replaced by the mean value of the data represented by the letter (see Table 2 for the speed and acceleration ranges represented by each letter). Next, a weighted average of these replacement values was calculated for each data variable. The weight used for the replacement values was the key word probability divided by the total probability of all the topic key words. Using the numeric summary, the topics could be grouped by speed, with Topics 3, 5, 6, 8, and 10 referring to driving at high speeds, Topics 1, 2, and 9 referring to driving at mid speeds, and Topics 4 and 7 referring to driving at low speeds.

3.3.2. Random forest results for the ten-topic model

The variable importance measure for the ten-topic model was investigated to understand the differences in trip descriptions identified by the RF models between the two driver conditions. The measure, Gini importance, is a univariate variable selection measure that scores the variables used in the RF models independently on their importance in differentiating the data (Archer & Kimes, 2008). The topics in order of their importance are shown in Fig. 7 along with the prevalence of each topic and the numeric summary of the topics.

As shown in Fig. 7, five of the topics had particularly high importance in the Random Forest model. These topics had higher prevalence in the data and had key words referring to mid speeds (Topics 1, 2, and 9) and low speeds (Topics 4

Table 4
Topic descriptions for the ten-topic model.

Topic	Keywords[a,b]	Topic definition
Topic 1	c_e_f/c_e_g/c_e_c/b_e_g/c_e_d/b_e_c	• Speed: mid (32–73 kph), ~suburban speed limits • Lateral acceleration: minimal (<±0.03 g) • Longitudinal acceleration: large (±0.03 g to ±0.2 g)
Topic 2	c_f_e/c_d_e/c_c_e/c_g_e/c_e_e/c_d_f/c_f_f/c_g_f/ c_c_f/c_c_d/c_f_d	• Speed: mid (32–73 kph), ~suburban speed limits • Lateral acceleration: large (±0.03 g to ±0.2 g) • Longitudinal acceleration: moderate (>±0.03 g to ±0.06 g)
Topic 3	d_e_e/d_e_f/d_e_d	• Speed: high (73–96 kph), ~highway/arterial speed limits • Lateral acceleration: minimal (<±0.03 g) • Longitudinal acceleration: minimal (<±0.03 g)
Topic 4	b_e_e/b_e_d/b_e_f/b_e_c/b_e_g/b_d_e/b_f_e	• Speed: low (1–32 kph), ~residential speed limits • Lateral acceleration: minimal (<±0.03 g) • Longitudinal acceleration: moderate (±0.03 g to ±0.06 g)
Topic 5	d_f_e/d_d_e/d_c_e/d_g_e/d_e_e/c_e_e	• Speed: high (73–96 kph), ~highway/arterial speed limits • Lateral acceleration: large (±0.03 g to ±0.2 g) • Longitudinal acceleration: minimal (<±0.03 g)
Topic 6	e_e_e/e_e_d/e_e_f	• Speed: very high (+96 kph), ~freeway/controlled-access highway speed limits • Lateral acceleration: minimal (<±0.03 g) • Longitudinal acceleration: minimal (<±0.03 g)
Topic 7	b_e_g/b_e_c/b_d_g/b_g_c/c_e_f/c_e_e/b_c_c/b_f_c/ b_c_g/c_e_g/b_g_g	• Speed: low (1–32 kph), ~residential speed limits • Lateral acceleration: large (±0.03 g to ±0.2 g) • Longitudinal acceleration: large (±0.03 g to ±0.2 g)
Topic 8	d_e_f/d_e_d/d_e_g/c_e_c/e_e_f/c_e_g/d_d_f/c_f_c/ d_e_c/d_d_g/c_e_d/c_g_c/d_f_f/e_e_g	• Speed: ranges from mid to very high (32–96 + kph), ~suburban to highway/arterial speed limits • Lateral acceleration: moderate (±0.03 g to ±0.06 g) • Longitudinal acceleration: large (±0.03 g to ±0.2 g)
Topic 9	c_e_e/c_e_f/c_e_d/c_d_e/c_f_e	• Speed: mid (32–73 kph), ~suburban speed limits • Lateral acceleration: minimal (<±0.03 g) • Longitudinal acceleration: minimal (<±0.03 g)
Topic 10	e_e_e/e_d_e/e_f_e/e_e_d/e_e_f	• Speed: high (+96 kph), ~freeway/controlled-access highway speed limits • Lateral acceleration: moderate (±0.03 g to ±0.06 g) • Longitudinal acceleration: minimal (<±0.03 g)

[a] Letter Order: (speed letter)_(lateral acceleration letter)_(longitudinal acceleration letter).
[b] See Table 2 for the numeric ranges corresponding to the letters.

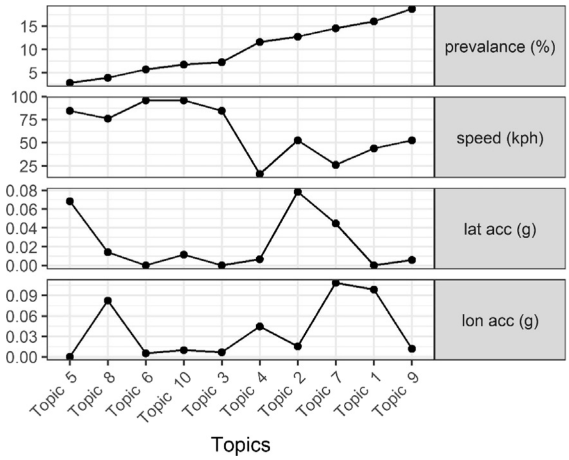


Fig. 6. Topic prevalence and description summary.

and 7). Fig. 7 also shows the difference in topic prevalence for the drivers with and without OSA. As shown in the figure, of the five topics with high importance in the Random Forest model, the drivers with OSA had a lower probability for the presence of Topics 2 and 7 and a higher probability for the presence of Topics 1 and 4 in their trip descriptions. Topics 1 and 4 are

Table 5
Example of numeric summary of key words for Topic 3.

Keywords	Speed (kph)[a]	Lateral acc. (g)[b]	Longitudinal acc. (g)[b]	Word probability ÷ total probability of key words
d_e_e	84.5	0	0	0.71/0.84
d_e_f	84.5	0	0.045	0.07/0.84
d_e_d	84.5	0	0.045	0.06/0.84
Weighted mean values for Topic 3	84.5	0	0.007	–

[a] Mean speed values for each letter (kph): “a” = 0, “b” = 16, “c” = 52.5, “d” = 84.5, “e” = 96.
[b] Mean absolute acceleration values for each letter (g): “a” = 0.3, “b” = 0.25, “c” = 0.13, “d” = 0.045, “e” = 0, “f” = 0.045, “g” = 0.13, “h” = 0.25, “i” = 0.3.

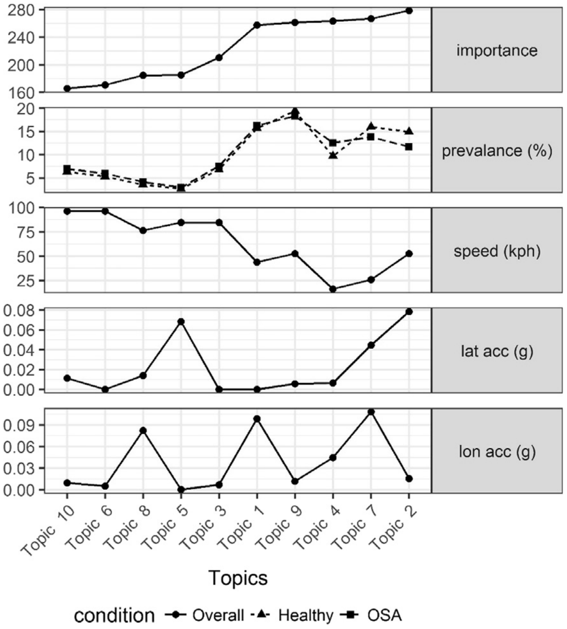


Fig. 7. Topic importance, prevalence and description summary.

related in that they are both defined by little to no lateral acceleration activity. Topics 2 and 7 are related in that they are both defined by medium to high amounts of lateral acceleration activity. These results suggest lateral acceleration is a differentiating factor for the two driver groups.

4. Discussion

4.1. Overview

We had two goals in conducting the analyses discussed in this paper: (1) to develop trip-level summaries of a naturalistic driving dataset, without oversimplifying the summary by using measures of central tendency (i.e., means, medians, standard deviations, etc.) in describing the continuous activity of maintaining vehicle control and (2) to use this advanced description to compare the behavior of drivers with and without OSA. These goals were accomplished by using SAX and topic modeling to identify patterns in speed and acceleration data from the drivers' trips. Topic models with 2, 5, 10, 15 and 20 topics were developed and compared. Three approaches were used to compare the models. First, the descriptiveness of the topics was compared. Increasing the number of topics from two to ten changed the topic definitions from only describing the general differences in the speeds selected by drivers during the trips to a more specific description of the changes in lateral and longitudinal acceleration while driving at certain speeds. Further increases in the number of topics from ten to twenty led to many topics with overlapping definitions. These additional topics were undesirable as they reduced the model interpretability. Second, the models were compared using perplexity, a common metric for examining the fit of language models. Based on this metric the fit improved with an increase in the number of topics. However, the improvement plateaued after the ten-topic model. Finally, the models were compared in terms of their ability to distinguish trips made by drivers with OSA from the trips made by the comparison drivers. For this analysis, the set of topic probabilities for each trip were used as the input to a Random Forest classification model. The twenty-topic model had the best RF model performance with an AUC = 0.73.

However, the performance of the twenty-topic model was not substantially different from the performance of the ten-topic model (AUC = 0.69) despite doubling the number of input variables.

Based on the descriptiveness of the topics, the model's fit based on perplexity, and the RF model performance, the ten-topic model was selected as the one that balanced specificity with potential for generalization. The ten-topic model was further investigated to better understand the systematic differences in the topic probabilities identified by the RF model for the trips from the two driver groups. In measuring the importance of the topics, the topics describing the amount of lateral acceleration while driving at mid to low speeds differentiated the two groups. The comparison drivers had higher probabilities for topics indicating significant amounts of lateral acceleration and the OSA drivers had higher probabilities for topics indicating little to no amount of lateral acceleration. As part of an exploratory analysis, the results raise new questions. (1) Would the same differences be found if the model was applied to a new sample? Using cross-validation when assessing the model performance suggests the findings will generalize; however, true generalization requires that the models be applied to a new sample of drivers. (2) Assuming the results do generalize, why did drivers with OSA have less lateral acceleration activity? Were the groups driving in different contexts or were they in similar contexts with different behaviors? As context was not directly observed, but inferred from speed, this remains an open question. Further analyses taking into consideration the spatial and time information of the naturalistic driving data could shed light on this distinction. (3) How do these group level differences translate into effects at the level of individual drivers? Prior work has found a correlation between the severity of an individual's OSA condition and the amount of increased crash risk (Williamson et al., 2011). Considering these results, it is of interest to explore if there are variations in the driving behavior between drivers with different severity levels of OSA and within the results for a single driver who experiences nights with varying levels of sleep restriction.

4.2. Advantages of topic models

The driving patterns described by the topics in the ten-topic model cover a wide range of driving contexts and behaviors. The overall prevalence of the topics in the data reflects typical travel patterns on different roadway types, with most travel occurring on mid to low speed roads that provide access and less travel occurring on high speed roads that provide mobility (Meyer & Miller, 2001). The topics also reflect typical behaviors while driving on these roads, with more limited acceleration changes occurring on high-speed roads and significant lateral and longitudinal changes occurring on mid and low- speed roads due to driving maneuvers such as turns and stops initiated in response to encountering at-grade intersections and traffic control devices (Roess, Prassas, & McShane, 2011). The ability of topic models to identify these patterns in an unsupervised manner makes it a promising method for exploratory analyses. Furthermore, it enables a comparison of drivers in terms of how they responded to more general roadway demands.

Most investigations using naturalistic driving data to study the effect of OSA or drowsiness on drivers focus on events of interest, such as the number of eye blinks or the occurrence of high g-force accelerations (Barr et al., 2005; Hanowski et al., 2003; Rau, 2005). In this paper the question is approached differently, by not focusing on specific events that occur during a trip and instead using data from the entire trip as the unit of analysis. The ability of a topic model to leverage, in an unsupervised manner, small units of data to generate a picture of the larger dataset made it an ideal tool for this analysis. Unlike some other unsupervised machine learning methods, particularly the more commonly used methods such as k-means or principal component analysis, topic modeling does not make linear assumptions about the structure of the patterns in the data. Topic models also recognize the association of data elements with multiple groups or topics, unlike the single group membership approach used in the typical categorization methods such as k-means clustering (Jain, 2010). These features make topic modeling a method with substantial promise as a general tool to explore the vast, unknown contents of the datasets generated by naturalistic driving studies. The "big picture" view of the data generated by topic models can provide context for interpreting analyses that rely on events and other forms of data segmentation. In addition to its value for exploratory analyses, topic modeling is a more sophisticated data summary method than other methods that use means, standard deviations, or counts. Topic models can capture subtle elements of behavior that might otherwise be obscured by these measures of central tendency. Topic modeling can provide additional benefits when used in combination with traditional analyses. The analyses described in this paper used the topic modeling results as features in a random forest classification model. Using the topics as features instead of using the raw or summarized data observations as the random forest input increased the interpretability of the classification results (Deng, Runger, Tuv, & Vladimir, 2013; Farrahi & Gatica-Perez, 2011; Huynh et al., 2008). Adding topics as a covariate to a regression model of high g acceleration events may significantly reduce unexplained variance. Other advantages of topic models may be obtained by using algorithms different from the simple LDA topic model used in this paper. For example, structural topic models make it possible to include covariates as metadata to help develop the topics (Roberts et al., 2014).

4.3. Topic interpretation

As an unsupervised method, topic models identify data topics without input from the analyst, other than setting the number of topics. However, an analyst is needed to define and interpret the meaningfulness of the topics generated. Examination of the topic key words and visualization of the raw data to observe what items the model gives similar topic definitions can be helpful for validation (Chuang, Manning, & Heer, 2012). However, topic interpretation is recognized as a challenge even for literary analyses, which, unlike the artificially constructed words in the present analysis, involve real words that connect to established concepts (Chang et al., 2009). The task of interpretation is further complicated by the ability of the method to generate results regardless of the number of topics selected. In this sense, topic modeling is similar to clustering, another unsupervised method that requires domain knowledge to determine the significance and interpretation of the output (Jain, 2010).

In this study, speed, lateral and longitudinal acceleration data were the variables selected for analysis and SAX was used to prepare the numeric driving data for the topic model analysis. A similar procedure could be used to prepare other driving data collected, including vehicle-based variables such as vehicle headway or driver-based variables such as eye gaze location. This study also demonstrated topic models can be used for time-synced, multivariate analyses, an analysis that is desirable to leverage the synchronous data collection process used in naturalistic driving studies.

5. Conclusion

Analyses of naturalistic driving data offer an opportunity to learn about drivers, the strategies and behaviors they use to navigate in the world and how impaired states such as drowsiness contribute to crashes. However, the size and complexity of the datasets make it challenging to obtain these insights. Considering these challenges, typical analysis methods focus on identifying critical events in the data while summarizing the remainder of the data using measures of central tendency. In this exploratory paper, topic modeling was used to shift the focus from critical events to continuous activities when describing naturalistic driving data and identifying characteristics that distinguished drivers with and without OSA. The method provided a holistic description of the data that might have been lost by using methods that simply determine the central tendency across the trips. Topic modeling is a useful complement to the more traditional event-focused analysis methods because it provides an understanding of the context surrounding the local influences observed in critical events.

Funding source

Funding: This work was supported by the National Institutes of Health [grant number RO1 HL091917].

References

Aksan, N., Dawson, J. D., Tippin, J., Lee, J. D., & Rizzo, M. (2015). Effects of fatigue on real-world driving in diseased and control participants. *Proceedings of the Ellipsis International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, 2015*, 268–274. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26618204.

American Academy of Sleep Medicine (AASM) (2014). *International classification of sleep disorders: Diagnostic and coding manual* (3rd ed.). Darien, IL: American Academy of Sleep Medicine. Retrieved from <http://www.aasmnet.org/library/default.aspx?id=9>.

Archer, K., & Kimes, R. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249–2260. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167947307003076.

Bärgman, J., Lisovskaja, V., Victor, T., Flannagan, C., & Dozza, M. (2015). How does glance behavior influence crash and injury risk? A "what-if" counterfactual simulation using crashes and near-crashes from SHRP2. *Transportation Research Part F: Traffic Psychology and Behaviour, 35*, 152–169. https://doi.org/10.1016/j.trf.2015.10.011.

Barr, L., Yang, C. Y. D., Hanowski, R. J., & Olson, R. (2005). Assessment of driver fatigue, distraction, and performance in a naturalistic setting. *Transportation Research Record: Journal of the Transportation Research Board* (Vol. 1937). http://doi.org/10.3141/1937-08.

Baydogan, M. G., Runger, G., & Tuv, E. (2013). A bag-of-features framework to classify time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(11), 2796–2802. https://doi.org/10.1109/TPAMI.2013.72.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM.* Retrieved from <http://dl.acm.org/citation.cfm?id=2133826>.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics, 1*(1), 17–35. https://doi.org/10.1214/07-AOAS114.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bosetti, P., Da Lio, M., & Saroldi, A. (2014). On the human control of vehicles: An experimental study of acceleration. *European Transport Research Review, 6*(2), 157–170. https://doi.org/10.1007/s12544-013-0120-2.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Bricka, S., & Bhat, C. (2006). Comparative analysis of global positioning system-based and travel survey-based data. *Transportation Research Record: Journal of the Transportation Research Board, 1972*, 9–20. https://doi.org/10.3141/1972-04.

Carney, C., Harland, K. K., & McGehee, D. V. (2016). Using event-triggered naturalistic data to examine the prevalence of teen driver distractions in rear-end crashes. *Journal of Safety Research, 57*, 47–52. https://doi.org/10.1016/j.jsr.2016.03.010.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296.

Chuang, J., Manning, C., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the International Working Conference on Advanced Visual Interfaces, ACM*, 74–77. Retrieved from http://dl.acm.org/citation.cfm?id=2254572.

Contardi, S., Pizza, F., Sancisi, E., Mondini, S., & Cirignotta, F. (2004). Reliability of a driving simulation task for evaluation of sleepiness. *Brain Research Bulletin, 63*(5), 427–431. https://doi.org/10.1016/j.brainresbull.2003.12.016.

Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. *Information Sciences, 239*, 142–153. https://doi.org/10.1016/J.INS.2013.02.030.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the United States of America, 113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113.

Dozza, M., Bärgman, J., & Lee, J. D. (2013). Chunking: A procedure to improve naturalistic data analysis. *Accident Analysis & Prevention, 58*, 309–317. https://doi.org/10.1016/j.aap.2012.03.020.

Eskandarian, A., Mortazavi, A., & Sayed, R. A. (2012). Drowsy and fatigued driving problem significance and detection based on driver control functions. In *Handbook of intelligent vehicles* (pp. 941–974). London: Springer, London. http://doi.org/10.1007/978-0-85729-085-4_36.

Farrahi, K., & Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(1), 3. Retrieved from http://dl.acm.org/citation.cfm?id=1889684.

Forsman, P. M., Vila, B. J., Short, R. A., Mott, C. G., & Van Dongen, H. P. A. (2013). Efficient driver drowsiness detection at moderate levels of drowsiness. *Accident Analysis and Prevention, 50*, 341–350. https://doi.org/10.1016/j.aap.2012.05.005.

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention, 37*(3), 461–472. https://doi.org/10.1016/j.aap.2004.11.003.

Guo, F., & Hankey, J. (2009). Modeling 100-car safety events: A case-based approach for analyzing naturalistic driving data. *Final report*. Retrieved from <https://www.researchgate.net/profile/Feng_Guo16/publication/265045998_Modeling_100-Car_Safety_Events_A_Case-Based_Approach_for_Analyzing_Naturalistic_Driving_Data_Final_Report/links/5630c49208aef3349c29f526.pdf>.

Hanowski, R. J., Wierwille, W. W., & Dingus, T. A. (2003). An on-road study to investigate fatigue in local/short haul trucking. *Accident Analysis & Prevention, 35*(2), 153–160. https://doi.org/10.1016/S0001-4575(01)00098-7.

Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30. Retrieved from http://epub.wu.ac.at/3987/.

Huynh, T., Fritz, M., & Schiele, B. (2008). Discovery of activity patterns using topic models. *Proceedings of the 10th international conference on ubiquitous computing, ACM*, pp. 10–19. Retrieved from <http://dl.acm.org/citation.cfm?id=1409638>.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

Klauer, S., Dingus, T. A., Neale, V., Sudweeks, J., & Ramsey, D. J. (2006). *The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data*. Blacksburg, VA. Retrieved from <http://trid.trb.org/view.aspx?id=786825>.

Krajewski, J., & Sommer, D. (2009). Steering wheel behavior based estimation of fatigue. In *5th International driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 118–124). Retrieved from <http://www.wirtschaftspsychologie.uni-wuppertal.de/fileadmin/wieland/lehrstuhl/Krajewski_Draw/017_KrajewskiSommer.pdf>.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26. https://doi.org/10.1053/j.sodo.2009.03.002.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22. Retrieved from http://ai2-s2-pdfs.s3.amazonaws.com/6e63/3b41d93051375ef9135102d54fa097dc8cf8.pdf.

Lin, J., & Li, Y. (2009). Finding structural similarity in time series data using bag-of-patterns representation. In *Proceedings of the 21st international conference on scientific and statistical database management* (pp. 461–477). Springer-Verlag. http://doi.org/10.1007/978-3-642-02279-1_33.

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery, 15*(2), 107–144. Retrieved from http://www.springerlink.com/index/pdf/10.1007/s10618-007-0064-z.

McDonald, A. D., Lee, J. D., Aksan, N., Dawson, J. D., Tippin, J., & Rizzo, M. (2013). The language of driving: Advantages and applications of symbolic data reduction for analysis of naturalistic driving data. *Transportation Research Record: Journal of the Transportation Research Board, 2392*, 22–30. https://doi.org/10.3141/2392-03.

McDonald, A. D., Lee, J. D., Aksan, N. S., Dawson, J. D., Tippin, J., & Rizzo, M. (2017). Using kinematic driving data to detect sleep apnea treatment adherence. *Journal of Intelligent Transportation Systems, 1–13*. https://doi.org/10.1080/15472450.2017.1369060.

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2013). Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 56*(5), 986–998. https://doi.org/10.1177/0018720813515272.

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2018). A contextual and temporal algorithm for driver drowsiness detection. *Accident Analysis and Prevention, 113*, 25–37. https://doi.org/10.1016/j.aap.2018.01.005.

Meyer, M., & Miller, E. (2001). *Urban transportation planning: A decision-oriented approach* (2nd ed.). Boston: McGraw-Hill.

National Highway Traffic Safety Administration (NHTSA). (2012). *Summary of state speed laws* (12th ed.). Washington, DC.

Rau, P. (2005). *Drowsy driver detection and warning system for commercial vehicle drivers: Field operational test design, data analyses, and progress. National highway traffic safety administration*. Retrieved from <https://www-nrd.nhtsa.dot.gov/Pdf/nrd-01/ESV/esv19/Other/Print 20.pdf>.

Reymond, G., Kemeny, A., Droulez, J., & Berthoz, A. (2001). Role of lateral acceleration in curve driving: Driver model and experiments on a real vehicle and a driving simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43*(3), 483–495. Retrieved from http://hfs.sagepub.com/content/43/3/483.short.

Rice, R. (1973). Measuring car-driver interaction with the gg diagram. *SAE technical paper*. Retrieved from <http://papers.sae.org/730018/>.

Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., ... Rand, D. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. Retrieved from http://www.jstor.org/stable/24363543.

Roess, R. P., Prassas, E. S., & McShane, W. R. (2011). *Traffic engineering* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible. *International Journal of Human-Computer Studies, 65*(3), 192–205. https://doi.org/10.1016/j.ijhcs.2006.10.001.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis: A road to meaning*. Laurence Erlbaum. Retrieved from <https://books.google.com/books?hl=en&lr=&id=JbzCzPvzpmQC&oi=fnd&pg=PA427&dq=Probabilistic+topic+models+steyvers&ots=aMN5O4Q_LI&sig=DpixZiOg2aOXqwgSVt3NMedDkm4>.

Thiffault, P., & Bergeron, J. (2003). Monotony of road environment and driver fatigue: A simulator study. *Accident Analysis & Prevention, 35*(3), 381–391. https://doi.org/10.1016/S0001-4575(02)00014-3.

Tippin, J., Aksan, N., Dawson, J. D., & Rizzo, M. (2013). Neuroergonomics of sleep and alertness. In *Neuroergonomics* (pp. 110–128). London: Palgrave Macmillan UK. http://doi.org/10.1057/9781137316523_6.

Tregear, S., Reston, J., Schoelles, K., & Phillips, B. (2009). Obstructive sleep apnea and risk of motor vehicle crash: Systematic review and meta-analysis. *Journal of Clinical Sleep Medicine, 5*(6), 573–581. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=medl&AN=20465027.

Turkington, P. M., Sircar, M., Allgar, V., & Elliott, M. W. (2001). Relationship between obstructive sleep apnoea, driving simulator performance, and risk of road traffic accidents. *Thorax, 56*(10), 800–805. https://doi.org/10.1136/THORAX.56.10.800.

Van Der Maaten, L., Postma, E., & Van Den Herik, J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research, 10*, 66–71. Retrieved from http://www.uvt.nl/ticc.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery, 26*(2), 275–309. https://doi.org/10.1007/s10618-012-0250-5.

Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. (2011). The link between fatigue and safety. *Accident Analysis and Prevention, 43*(2), 498–515. https://doi.org/10.1016/j.aap.2009.11.011.

Wu, K.-F., & Jovanis, P. P. (2013). Defining and screening crash surrogate events using naturalistic driving data. *Accident Analysis & Prevention, 61*, 10–22. https://doi.org/10.1016/j.aap.2012.10.004.

Yuan, Y., Du, F., Qu, W., Zhao, W., & Zhang, K. (2016). Identifying risky drivers with simulated driving. *Traffic Injury Prevention, 17*(1), 44–50. https://doi.org/10.1080/15389588.2015.1033056.