A Cluster-Based Taxonomy of Bus Crashes in the United States

Dooti Roy* Ved Deshpande† M. Henry Linder‡

Abstract

Accident taxonomy or classification can be used to direct the attention of policymakers to specific concerns in traffic safety, and can subsequently bring about effective regulatory change. Despite the widespread usage of accident taxonomy for general motor vehicle crashes, its use for analyzing bus crashes is limited. We apply a two-stage clustering-based approach based on self-organizing maps followed by neural gas clustering to construct a data-driven taxonomy of bus crashes. Using the 2005–2015 data from General Estimates System (GES), we identify four clusters and expose the qualitative traits that characterize four distinct types of bus crash. Our analysis suggests that cluster characteristics are largely stable over time. Consequently, we make targeted policy recommendations for each of the four subtypes of bus crash.

Key Words: bus crashes, accident classification, clustering, self-organizing maps, General Estimates System

1. Introduction

Some forms of mass transit, such as rail-based systems, exist within a closed environment, and traffic accidents are prevented by human dispatchers. Other forms, however, depend upon infrastructure used by heterogeneous populations. In particular, buses operate on municipal roadways, often in dense urban areas. Bus-automobile interactions pose unique problems to public safety, especially for public and school buses, which stop frequently. Other problems for bus safety include bus-pedestrian collisions, and an outsized impact of dangerous weather conditions. All told, in 2014, 22000 persons were injured, and 281 died, in bus accidents (Analysis Division, 2014).

As a mass transit solution, bus systems experience unique challenges. To begin with, the bus system operates within the larger system of public-access roads. This poses problems to the entire community, starting with pedestrians, who account for nearly 25% of bus accident fatalities, as well as drivers of other cars (26%), small truck drivers (18%), and bus passengers (16%). Furthermore, the problem is concentrated on buses operating in urban or suburban areas: 73% of bus fatalities occur in school buses or transit buses, as opposed to long-distance bus lines (Analysis Division, 2014). In order to promote the welfare of all individuals impacted by bus accidents, it is essential to understand the causes and characteristics of traffic accidents involving buses. Such an understanding will provide explanations for the causes of bus accidents, which has policy relevance for traffic control, transit system design, and distracted-driver regulations.

A taxonomy of bus accidents would provide insight into the variety of conditions that frequently occur simultaneously in each accident subtype. These conditions could be very specific to each subtype, allowing policymakers to focus their attention on relevant regulations that need to be changed. We believe that such taxonomies should be data-driven to avoid subjectivity, and discover novel subtypes of accident.

The construction of data-driven bus accident taxonomies is not widespread in the traffic safety literature. The leading study of Prato and Kaplan (2013) employs data collected

*Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA
†Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA
‡Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA

nearly a decade ago. By applying their method to more recent data, we can understand how the taxonomy itself has changed from the 2005–2009 period to the 2010–2015 period.

The National Highway Traffic Safety Administration (NHTSA) annually publishes detailed datasets of auto accidents, which include variables describing the events that caused the accident, demographic information about any drivers or pedestrians involved, and other categorical variables describing the incident. The latent structure of accidents in the NHTSA data can be used to identify subpopulations of accidents to guide a taxonomic study of risk factors of bus accidents.

However, the large volume of observations and high-dimensionality of variables makes it difficult to characterize bus accidents by manual exploratory data analysis. Therefore, in this paper we constructed a taxonomy of bus accidents using cluster analysis. Our analysis discriminates several subpopulations of bus accidents on the basis of patterns of shared attributes, which provide compelling qualitative profiles of these subpopulations.

We aggregated a subset of variables in the NHTSA's General Estimates System (GES), and considered crashes from two separate time periods, 2005–2009, and 2010–2015. We find clusters representing subpopulations of bus accidents that are broadly consistent across the two datasets, but we also note the differences in the clusters across the datasets, which gives us insight into how the taxonomy itself changed over the two time periods.

The paper is organized as follows. In section 2, we describe the NHTSA's national crash data sample, an ongoing data-collection initiative, which is the dataset used in our analysis. In section 3, we describe the two-stage clustering approach used to build the taxonomy. In section 4, we take a close look at the clusters and interpret them to understand the distinct subpopulations of bus accidents formed by the taxonomy. Section 5 discusses possible implications of our analysis and addresses some of its limitations.

2. Data

The GES dataset is a nationally representative sample of police-reported motor vehicle crashes, released annually. The data is collected continuously at 60 locations around the United States by the NHTSA's National Center for Statistics and Analysis. In releasing the data, the NHTSA intends to provide researchers access to traffic safety data (Highway Traffic Safety Administration, 2015). The GES dataset is a probability sample, and corresponding to each accident is a complex survey weight. The incorporation of these weights allows estimation of quantities at the national level.

We constructed two datasets, covering the time periods 2005–2009 and 2010–2015 respectively. We chose to consider two disjoint time periods in order to assess the stability of the cluster analysis. There is also a significant break between 2009 and 2010 in the format, structure, and definitions of the NHTSA's GES sample.

All of the variables we employed in our analysis were either indicator variables for a certain characteristic, or categorical, as shown in the following list. We selected variables for the two datasets along the lines of Prato and Kaplan (2013).

- **Binary**
  - Accident involved non-motorists;
  - A school bus was involved;
  - Located at an intersection;
  - Single lane road;

  - Bus driver distracted;
  - Bus driver impaired or under the influence of alcohol or drugs;
  - Bus driver speeding;
  - Bus driver was male;
  - Pickup truck, SUV, or van involved;
  - Light / heavy truck involved;
  - Other driver distracted;
  - Other driver impaired or under the influence of alcohol or drugs;
  - Other driver speeding;
  - Surface conditions adverse (wet, straight, or unlevel road);
  - Occurred during daylight hours.

- **Categorical**
  - Number of vehicles involved: one, two, three or more;
  - Speed limit: < 35MPH, $35 - -55$MPH, > 55MPH;
  - Critical event that made crash imminent: loss of control, vehicle turning, in another vehicle's lane, other vehicle encroaching into lane, non-motorist, object or animal;
  - Traffic control at crash site: no control, traffic signal, traffic sign, other;
  - Bus movement prior to accident: parking, going straight, stopping, decelerating, turning right, turning left, overtaking, reversing, negotiating a curve, other.

2.1 Data Extraction

One of the major challenges of this analysis was data extraction and processing. For the benefit of researchers who may wish to replicate our findings, we briefly describe some of the important aspects of our data processing. The NHTSA undertook a massive effort to standardize the Fatality Analysis Reporting System (FARS) and National Automotive Sampling System General Estimates System (NASS GES) data with the goal of simplifying crash data coding and analysis, as well as to reduce costs and errors. Major changes to the coding scheme were implemented in datasets for the years starting from 2010. The resulting inconsistencies in variables is one of the reasons why the two datasets were considered separately in our analysis. Variables whose coding underwent substantial changes included those relating to driver impairment (drugs, alcohol), light conditions, and manner of collision, among others.

For each year, the GES data encompasses several tables, each containing data measured at several levels—vehicle-, person-, and accident-level, among others. Table 1 lists and describes the purpose of these data files. Since we performed our analysis at the accident level, we collapsed data from the more granular levels to the accident level, a procedure that involves some subjective assessment.

3. Methods

Our goal was to build a data-driven taxonomy of bus crashes which classifies accidents into cohesive groups that share recurrent characteristics. The broad class of clustering techniques naturally lends itself to this task, because they are intended to identify emergent group structures from data in an unsupervised way (Friedman et al., 2001).

We adopted the two-stage clustering method of Prato and Kaplan (2013), who performed the primary study of bus accident taxonomy. In the first stage, we reduced the dimensionality of the data using self-organizing maps (SOM) (Kohonen, 1982). An SOM is an artificial neural network that reduces a high-dimensional input space to a low-dimensional, topographic map of the input space. Frequently, a two-dimensional map is used which consists of a grid of neurons. The SOM algorithm maps high-dimensional observations to the lower-dimensional plane, and organizes the neurons spatially. The method is structured so each neuron represents a unique region of the input space, and the overall grid of neurons covers the entire input space.

Figure 1 provides an illustration of the training process for the SOM algorithm. The SOM algorithm can thus be viewed as a clustering algorithm is its own right, with each neuron representing a cluster of observations.

To organize the neurons, the SOM algorithm utilizes competitive learning, in which neurons compete among each other to represent input data vectors. Associated with each neuron is a weight vector with the same dimension as the input vectors. When the SOM considers a randomly-selected input vector, the neuron whose weight vector is "closest" to the input vector is deemed the winner. This nearest neuron is called the best matching unit (BMU). Then, the weights of the BMU and the neurons close to it in the SOM grid are adjusted towards the input vector. This procedure is repeated until convergence, at which point the neurons cease to "move"—that is, their weight vectors do not change (Kohonen, 1990).

In the second stage of clustering, we applied the neural gas algorithm of Martinetz and Schulten (1991) to the self-organizing map from the first stage, thereby producing the final clustering. The neural-gas clustering algorithm was historically inspired by the SOM algorithm and can be viewed a robust version of the sequential K-means algorithm (MacQueen et al., 1967). In sequential K-means, cluster centers are updated by processing individual input vector sequentially. This contrasts with the approach of the traditional K-means algorithm, which considers all input vectors simultaneously. In the sequential method, only the cluster center closest to the input vector is adjusted towards it. The neural gas algorithm deviates from this by updating *all* the cluster centers at each new input vector, with centers closer to the input vector adjusting more than those far away. This modification provides robustness to noisy input vectors and stabilizes the algorithm's convergence.

As mentioned previously, the SOM algorithm is a clustering algorithm in its own right, and hence could be used to form the taxonomy of bus crashes by itself. However, as mentioned in Prato and Kaplan (2013), the neural-gas algorithm has advantages over methods like SOM in terms of convergence speed and accuracy (Vesanto and Alhoniemi, 2000). The high computational complexity of the neural-gas algorithm, its main disadvantage, is mitigated by its application to the lower-dimensional SOM neuron grid. This allows more flexibility in dealing with high-dimensional output, without sacrificing the practicality of applying the technique.

4. Results

A total of 4386 bus accidents were reported in the GES data from 2005–2015, representative of an approximate total of 530797 bus accidents. We performed basic exploratory data analysis to identify relevant clustering variables, and to compare the data composition across the two time periods.

Figure 2 compares the proportion of "yes" responses for the binary variables. Generally, the proportions are consistent across the two time periods. We see a decrease in driver distraction in the later dataset, and a proportional increase in the accidents in the presence of adverse surface conditions. Other-driver distraction fell by almost 50% in the later dataset. Incidents where the other driver was impaired or under the influence of alcohol or drugs also fell in the later years by a comparatively small proportion. There were relatively fewer incidents involving trucks from 2010–2015 than 2005–2009. We also observe that in both time periods very few incidents involved bus drivers driving while impaired. But, these variables reflect only a small portion of the explanatory variables, so that for the most part, the two datasets are similar in terms of their binary responses.

Figure 3 gives radar plots for the categorical variables considered in the analysis. Within each subplot, the axes correspond to different levels of the categorical variable, the heights sum to one, and the distance of the point from the center represents the magnitude of that factor level's proportional constitution. As these plots indicate, the composition of the categorical variables is quite similar in the two datasets. There are noticeable differences in the distribution of the variables for traffic control, critical event that made the crash imminent, and bus movement prior to the event to avoid the accident. These variables seem to be distributed differently in the two time periods, information which is used in the clustering.

We chose a 20×20 grid of neurons for the SOM clustering method, similar to Prato and Kaplan (2013). The neural gas algorithm partitioned the data into four clusters, which we chose to optimally balance cluster homogeneity and differentiation. We used the R programming language (R Core Team, 2016) to implement the two-stage clustering. We performed SOM using the package "kohonen" (Wehrens and Buydens, 2007) and we used "cclust" for the neural gas algorithm (Dimitriadou, 2015).

Based on the results of the analysis, the four clusters in the **2005–2009 dataset** can be described as follows:

Cluster 1: Multi-vehicle crashes (86.4% involving two vehicles, see Figure 4 (a).) which happened largely due to the bus stopping (49.8%), going straight (25.7%) or decelerating (10.4%) (Figure 7 (a)). The cause of the crash was typically due to the bus being in another vehicle's lane (92.3%) (Figure 6 (a)). 27.4% of the time, the bus driver was distracted. 51.8% of the time, a school bus was involved. Most of the crashes happened during the daytime (90.1%). 40% of the time, the crash occurred at an intersection. In 30.6% of the cases, the police found at least one driver of the other vehicles involved in the crash distracted and in 22.1% of the cases, at least one of the other drivers was under the influence of alcohol, or drugs, or was impaired in some way. In 21.6% of the cases, one of the other vehicles was a truck and in another 28.4%, crashes involved an SUV, pickup truck, or van. See Figure 5 (a). 65.6% of the crashes happened on roads with speed limit between 35–55 MPH, implying that these happened within a city or town and not on the highways (Figure 9 (a)). 50.3% of the crashes happened where the road had no traffic control, and in 27.2% of the cases, there was a traffic signal (Figure 8 (a)).

Cluster 2: Single vehicle crashes involving non-motorists (99.5%) that happened when the bus was going straight (39.2%), turning left (19%) or trying to park (14.7%). There were three primary reasons for the crash: the bus turned (48.1%), or the non-motorist was encroaching into the bus's lane (30.7%), or due to an object or animal (17.1%). In 33.9%

of the cases, the bus driver admitted to have been distracted. 39.8% of the time, a school bus was involved. For almost all of the crashes, the non-motorist involved was distracted, but only 4.7% of the cases involved the non-motorist being impaired or under the influence. 44.3% of the crashes happened at an intersection, and 35.3% of the accidents occurred on single lane roads. 69.6% of the crashes happened where the road had no traffic control, and in 16.3% of the cases there was a traffic signal. Only 0.1% of the cases involved a light or a heavy truck, and 24.8% of the crashes involved an SUV, pickup truck, or van.

Cluster 3: Multi-vehicle crashes (93.4% involving two vehicles) that occurred when the bus was going straight (64%) or stopping (15.9%). The crash was primarily due to the bus trying to encroach into another vehicle's lane (90.5%), implying that the crash was mainly the bus driver's fault. In 20.2% of the cases, the bus driver admitted to have been distracted, and 39.9% of the time, a school bus was involved. Most of the crashes happened on roadways with moderate to high speed limits (60.1% on roads with 35–55 MPH and 30.3% on roads with greater than 55 MPH). For 24.9% of the crashes, at least one of the other drivers involved was distracted, and 20.5% of the cases involved at least one of the other drivers being impaired or under the influence. 49.6% of the crashes happened at an intersection, and 31.1% of the accidents occurred on single lane roads. 49.9% of the crashes happened where the road had no traffic control, for 26.3% of the cases there was a traffic signal, and in 18.3% of the cases there was a traffic sign. 22.5% of the cases involved a light or a heavy truck, and 27.1% of the crashes involved an SUV, pickup truck, or van.

Cluster 4: Multi-vehicle crashes (98.8% involving two vehicles) that happened when the bus was going straight (18.6%) or turning (48.5%). The crash was primarily due to the bus trying to turn (97%). In 31.8% of the cases, the bus driver admitted to being distracted, and 32.2% of the time, a school bus was involved. Most of the crashes happened on roadways with moderate to high speed limits (61.6% on roads with 35–55 MPH and 32.3% on roads with greater than 55 MPH). In only 4.9% of the crashes was at least one of the other drivers involved distracted, and only 10.5% of the cases involved at least one of the other drivers being impaired or under the influence. 66.9% of the crashes happened at an intersection, and 31.1% of the accidents occurred on single lane roads. 38.9% of the crashes happened where the road had no traffic control, and in 35.9% of the cases there was a traffic signal. 19.9% of the cases involved a light or a heavy truck, and almost no crashes involved an SUV, pickup truck, or van.

Table 2 summarizes the characteristics of the 2005–2009 bus crash clusters.

The four clusters from the **2010–2015 dataset** can be described as follows. Note that we have rearranged the arbitrary ordering, so that the cluster number matches the corresponding cluster from the 2005–2009 results.

Cluster 1: Multi-vehicle crashes (90.7% involving two vehicles, see Figure 4 (b).) which happened largely due to the bus stopping (52.9%) or going straight (26.2%) (Figure 9 (b)). The reason of the crash was mostly due to the bus being in another vehicle's lane (96.4%), as observed in in the earlier dataset (Figure 6 (b)). 10.8% of the time, the bus driver was distracted. 53.2% of the time, a school bus was involved. Most of the crashes happened during the daytime (88.4%). 48.7% of the time, the crash happened when the bus was at an intersection. Although in most of the cases, the other driver was not distracted, in 18.4% of the cases, the other driver was impaired or under the influence of alcohol or drugs. Almost none of the crashes involved a truck, but another 26.8% of the crashes involved an SUV, pickup truck, or van (see Figure 5 (b).) 62.7% of the crashes happened on roads with speed limit between 35 and 55 MPH, and another 29.2% happened on roads with speed limits greater than 55 MPH (Figure 9 (b)). 54.4% of the crashes happened where the road had no traffic control, and in 28.4% of the cases there was a traffic signal (Figure 8 (b)).

Cluster 2: Single vehicle crashes involving non-motorists (100%) which mostly happened

when the bus was going straight (37.9%) or overtaking (27.1%). There were two primary reasons for the crash: the bus was turning (23.5%), or a non-motorist was encroaching into the bus's lane (72%). In 14.3% of the cases, the bus driver admitted to have been distracted, and 24.6% of the time, a school bus was involved. For only 5.5% of the crashes, the non-motorist involved was distracted, but 25.7% of the incidents involved the non-motorist being impaired or under the influence. 59.9% of the crashes happened at an intersection, and 3.2% of the accidents occurred on single lane roads. 69.6% of the crashes happened where the road had no traffic control, and in 16.3% of the cases, there was a traffic signal. 15.2% of the cases involved a light or a heavy truck, and only 0.2% of the crashes involved an SUV, pickup truck, or van. This was the cluster with the highest percentage of male drivers (79,6%). 65.9% of the crashes happened during the daytime, which implies that approximately 35% of the accidents involving non-motorists happen after dark.

Cluster 3: Multi-vehicle crashes (92.9% involving two vehicles) which mostly happened when the bus was going straight (66.4%) or stopping (16%). The crash was primarily due to the bus trying to encroach into another vehicle's lane (95.6%), implying that the crash was mainly the bus driver's fault. In only 3.2% of the cases did the bus driver admit to have been distracted, and 38.8% of the time, a school bus was involved. Most of the crashes happened on roadways with moderate to high speed limits (65.6% on roads with 35–55 MPH and 27.5% on roads with greater than 55 MPH). For 14.2% of the crashes, at least one of the other drivers involved was distracted, and 13.9% of the cases involved at least one of the other drivers being impaired or under the influence. 53.3% of the crashes happened at an intersection, and 1.6% of the accidents occurred on single lane roads. 57.3% of the crashes happened where the road had no traffic control, in 30% of the cases there was a traffic signal and in 6.3% of the cases there was a traffic sign. 19.4% of the cases involved a light or a heavy truck, and 26.7% crashes involved an SUV, pickup truck, or van.

Cluster 4: Multi-vehicle crashes (98.5% involving two vehicles) which mostly happened when the bus was going straight (22.4%) or turning (43.7%). The crash was primarily due to the bus trying to turn (87.5%), or due to the object or animal on road (10%). In 19.6% of the cases, the bus driver admitted to have been distracted, and 44.6% of the time, a school bus was involved. Most of the crashes happened on roadways with moderate to high speed limits (57.3% on roads with 35–55 MPH and 38.6% on roads with greater than 55 MPH). In 12.5% of the crashes, at least one of the other drivers involved was distracted, and only 5.4% of the cases involved at least one of the other drivers being impaired or under the influence. 54.9% of the crashes happened at an intersection, and only 1.7% of the accidents occurred on single lane roads. 52.6% crashes happened where the road had no traffic control, and in 26.3% of the cases, there was a traffic signal. 13.4% cases involved a light or a heavy truck, and 23.7% of the crashes involved an SUV, pickup truck, or van.

Table 3 summarizes the characteristics of the 2010–2015 bus crash clusters.

5. Discussion

The cluster compositions described above are strikingly stable across the two datasets. The clusters represent distinct subpopulations of bus accidents, with clear interpretations and policy implications.

Three of the clusters represent distinct types of bus accident. First, cluster 2 depicts accidents involving non-motorists, such as pedestrians. This cluster has a high proportion of accidents at intersections (60%), which is a reasonable result. Most pedestrian-automobile interactions occur at intersections, due often to pedestrians crossing at cross-walks. To reduce collisions with non-motorists, then, bus driver training programs can emphasize particular focus and attention to pedestrians at intersections. Furthermore, boundaries be-

tween automobiles and sidewalks, like lane lines and curbs, can be accentuated in order to reduce accidental bus-pedestrian interaction.

Second, cluster 3 features buses that are changing lanes. Most of these accidents (57.3%) occurred in the absence of traffic controls, and nearly half involved larger vehicles. This may mean that these accidents occur on larger roads, and points to accidents that involve other large vehicles. There are also many school buses in this cluster, which may be transporting school-children on highways. Also, 28% of the other drivers were distracted. Taken together, this cluster indicates the need for increased penalties for distracted driving, better local awareness by bus drivers, and perhaps mechanical upgrades to the buses. Proximity sensors that notify drivers of other cars in their blind spot, or generally adjacent to the vehicle, are becoming increasingly common in consumer automobiles, and could prevent lane-changing accidents in this cluster. A less costly, similar method would be improving visibility of bus turn signals, and additional mirrors to increase the driver's field of vision.

Third, cluster 4 contains accidents that occur when buses are turning. This type of accident is unsurprising, given the large physical profile of buses, which can limit movement. This type of accident can be discouraged by improving bus driver precision and driving skill. Also, 20% of these accidents occurred while the bus driver was distracted, suggesting that driver attentiveness is important in preventing accidents.

The signal in cluster 1 is difficult to identify, and we interpret this cluster as the category for "miscellaneous" accidents not accounted for by the other clusters. It is difficult to gain any understanding from the cluster composition, although we note the large proportion of school buses, and higher speeds. We also notice that 96.4% of the crashes are due to the bus being "in another vehicle's lane." This level of the "critical event" factor certainly merits further attention in terms of investigating the cause of this critical event.

Other than cluster 2, the proportion of accidents in the absence of traffic controls is roughly 50%. This initially suggests that bus accidents might be reduced simply by installing additional traffic controls. However, this number may simply represent those accidents that occur on straight roads, and not at a location where more traffic signals would be useful. A simple categorical variable cannot account for this nuance, and we hope to clarify these subtypes by considering conditional combinations of variables—e.g., accidents at intersections, classified by the presence of traffic controls.

We note, too, that most (70%) of the accidents in cluster 2 occur in the absence of traffic controls. We interpret this to indicate that non-motorists regularly stray from sidewalks into the street. A ready way to address this may be to increase penalties for jay-walking, and further data analysis and research can provide insight into the effectiveness of such policies.

The characteristics of the clusters which change across the two time periods are highlighted in Table 4. One of the major changes for cluster 2—which broadly represents single vehicle crashes involving non-motorists—is the critical event that caused the accident. The percentage of crashes due to "other vehicle/ non-motorist encroaching" increased from 31% to 72%. At the same time, the proportion of crashes due to "other driver distracted" decreased massively from 100% to 5%, and the proportion of crashes due to "other driver under influence" increased from 5% to 26%. This suggests that non-motorists, possibly pedestrians, have become more mindful about their surroundings, except when intoxicated.

For cluster 4—which broadly represents multi-vehicle crashes that occurred when the bus was turning—the proportion of crashes which occurred at intersections decreased from 67% to 55%. At the same time, the proportion of crashes which occurred in the absence of traffic controls increased from 39% to 53%, which suggests that traffic safety officials should examine the presence of traffic controls at intersections on common bus routes. One encouraging change that occurred in all clusters is the decrease in proportion of crashes due to the bus driver being distracted.

A primary limitation of our analysis is the lack of model-based use of the sampling weights. Summary statistics and exploratory data analysis indicate that for many of our variables, the composition remains consistent between the weighted and unweighted datasets. Although our exploratory data analysis and visualization of the clusters account for the sampling weights, our literature search did not indicate methods for incorporating survey weights in complex clustering methods such as SOM or neural gas. Consequently, in order to improve the validity of our assumptions, one avenue for further work is to develop methods to directly incorporate the weights into both stages of the clustering approach.

Another limitation is the uncertainty introduced by the GES data revision in 2009. The differences we observe above between the 2010–2015 and 2005–2009 datasets may reflect a true change in the taxonomy of bus accidents, but it may also be due to changes in variable definitions or non-determinism in the clustering algorithm. However, because the cluster results are so compatible across the two populations, we are confident that our findings reflect a stable taxonomy of bus accidents. Furthermore, the stability of the cluster composition indicates real, latent structure in the data: bus accidents can be typified by a cluster-based taxonomy, and the differences between the types are distinct and informative.

Finally, these cluster results are preliminary, and as discussed above, the simplicity of the clustering features limited the extent to which we can interpret the clusters for practical use. This indicates that more detailed and fine-grained analysis of bus taxonomy will provide yet further understanding of the causes of bus accidents.

In this paper, we constructed a data-driven taxonomy of bus accidents in the United States using a two-stage clustering method. We investigated the stability of the cluster composition by assembling independent taxonomies for two datasets from different time periods. As anticipated, clearly distinguished accident subtypes are evident in the data. Furthermore, accompanying these subtypes is a better picture of the nature and causes of bus accidents. These results can be used by policy-makers to increase safety regulations targeted to specific accidents, and allocate funds for improved traffic controls. These results can also be used to improve training for bus drivers, inform pedestrians, and influence bus design and manufacturing. Finally, these results can be used by the researcher as a base for understanding and taxonomizing bus accident types.

References

Analysis Division, Federal Motor Carrier Safety Administration, U. D. o. T. (2014). Large Truck and Bus Crash Facts 2014. [Online; accessed 2016-09-01].

Dimitriadou, E. (2015). *cclust: Convex Clustering Methods and Clustering Indexes*. R package version 0.6-20.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Highway Traffic Safety Administration, N. (2015). Ges analytical user's manual. http://www.nhtsa.gov/NASS, accessed: September 1, 2016.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Martinetz, T. and Schulten, K. (1991). *A" neural-gas" network learns topologies*. University of Illinois at Urbana-Champaign.

Prato, C. G. and Kaplan, S. (2013). Bus crash patterns in the united states: a clustering approach based on self-organizing maps. In *WCTR 2013: 13th World Conference on Transportation Research*.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE transactions on neural networks*, 11(3):586–600.

Wehrens, R. and Buydens, L. (2007). Self- and super-organising maps in r: the kohonen package. *J. Stat. Softw.*, 21(5).

Wikipedia (2016). Self-organizing map — Wikipedia, the free encyclopedia. [Online; accessed 25-September-2016].

List of Figures

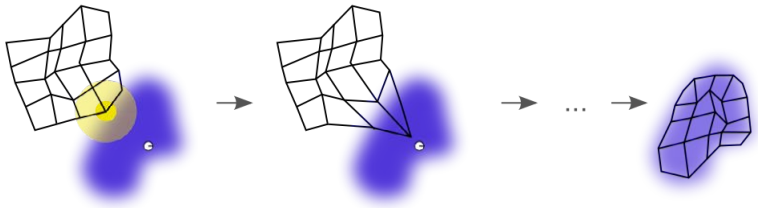


Figure 1: An illustration of the training of a self-organizing map. The blue blob is the distribution of the training data, and the small white disc is the current training datum drawn from that distribution. At first (left) the SOM nodes are arbitrarily positioned in the data space. The node (highlighted in yellow) which is nearest to the training datum is selected. It is moved towards the training datum, as (to a lesser extent) are its neighbors on the grid. After many iterations the grid tends to approximate the data distribution (right). (Wikipedia, 2016)

Figure 2: Mosaic plots comparing the binary response variables across the 2005–2009 and 2010–2015 datasets

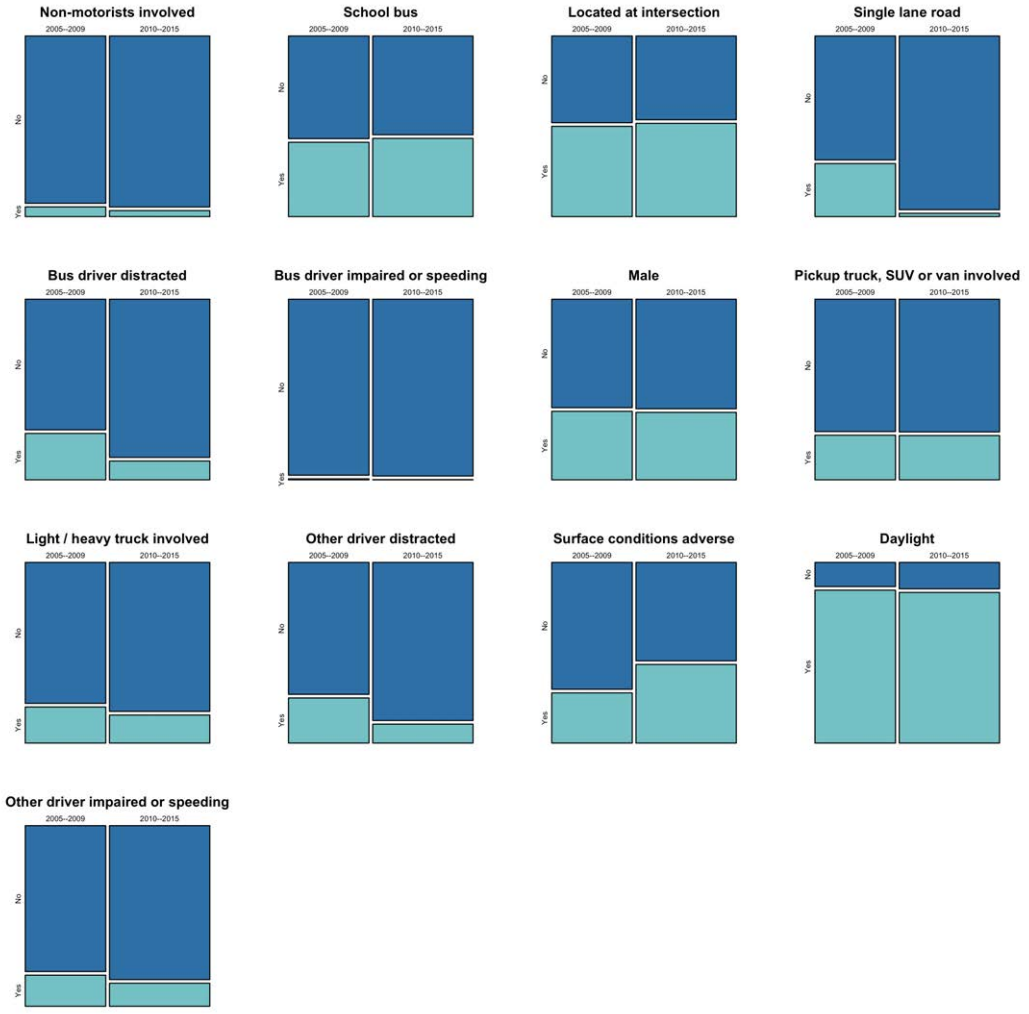


Figure 3: Radar plots comparing the categorical response variables across the 2005–2009 and 2010–2015 datasets

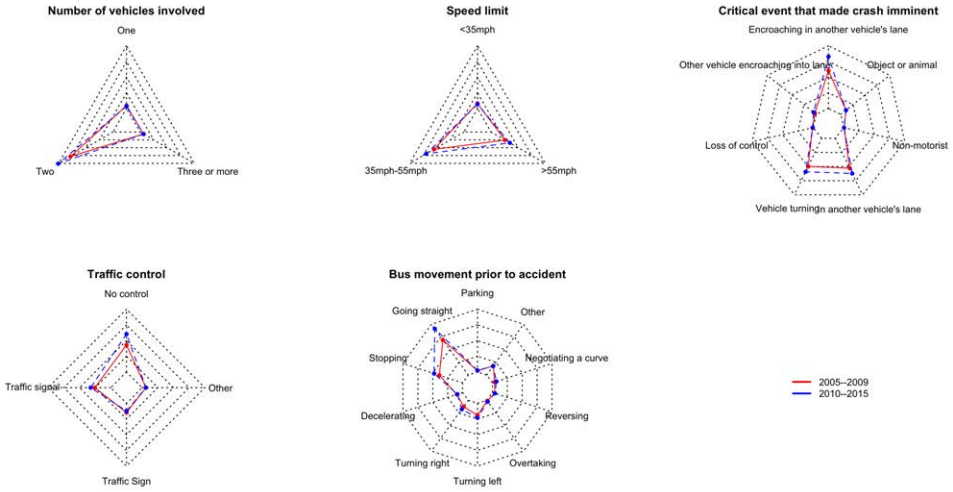
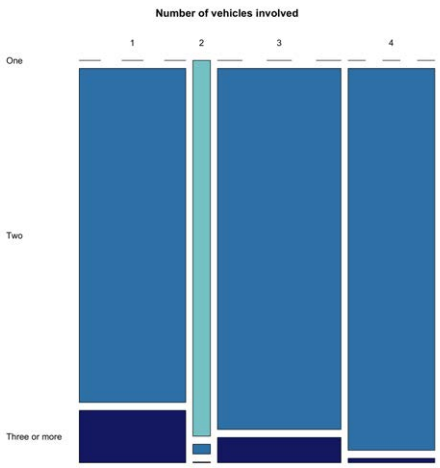


Figure 4: Mosaic plots giving the breakdown of the variable "number of vehicles involved in the accident" for each of the four clusters.

(a) 2005–2009



(b) 2010–2015

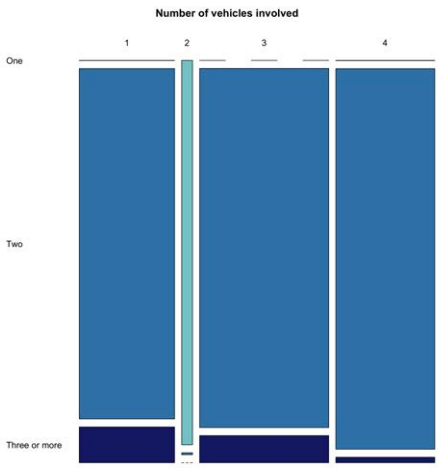


Figure 5: Observed proportions for binary response variables for each of the four clusters.

(a) 2005–2009 (b) 2010–2015

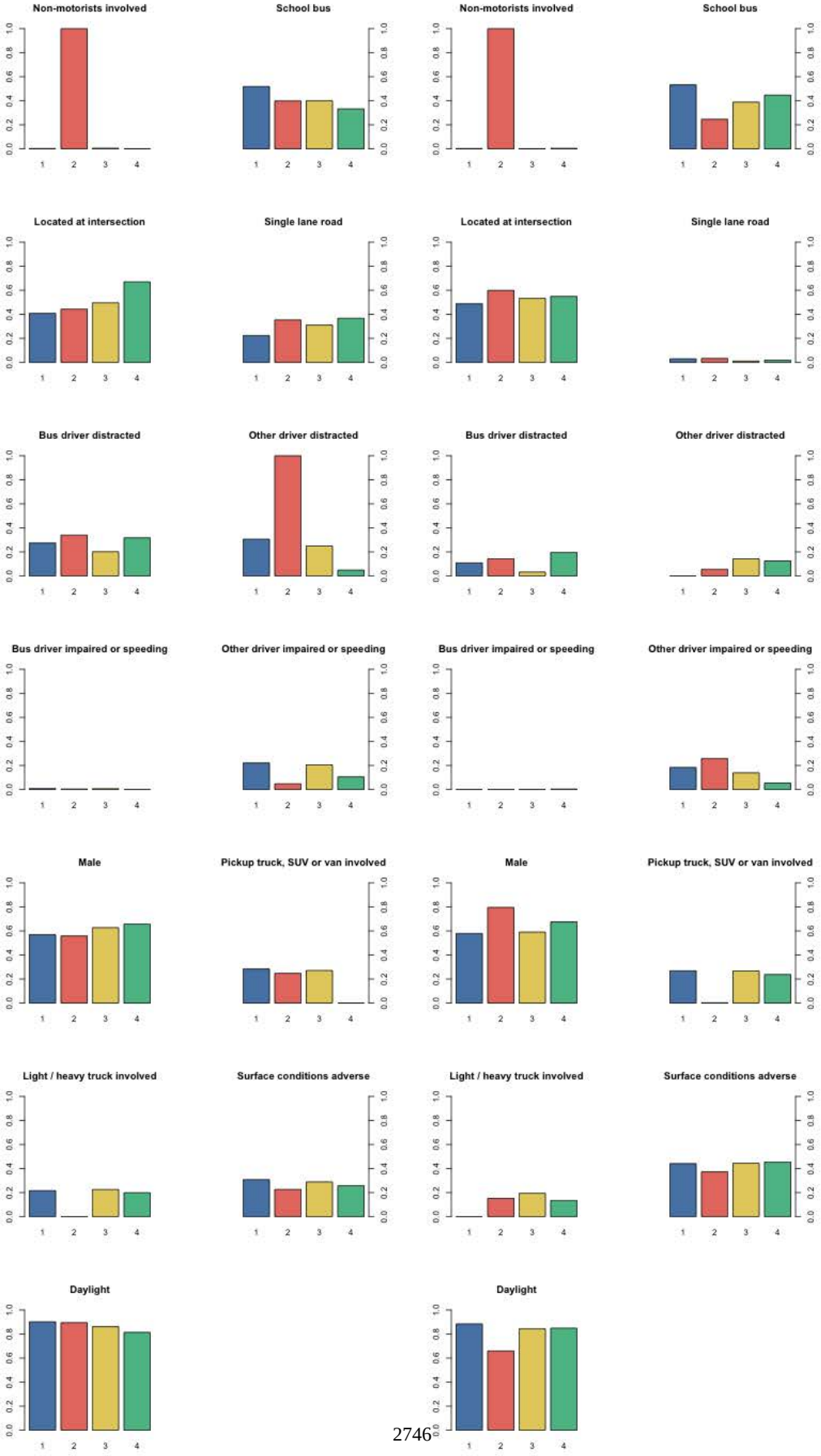
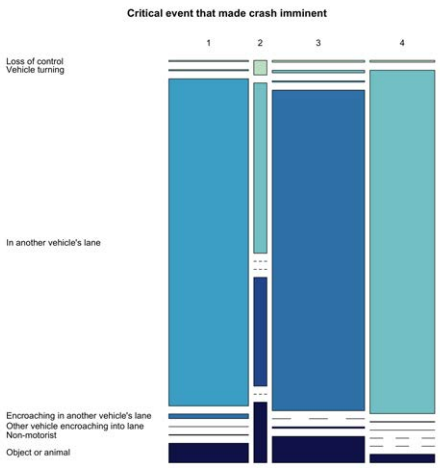


Figure 6: Mosaic plots giving the breakdown of the variable "critical event that made the crash imminent" for each of the four clusters.

(a) 2005–2009



(b) 2010–2015

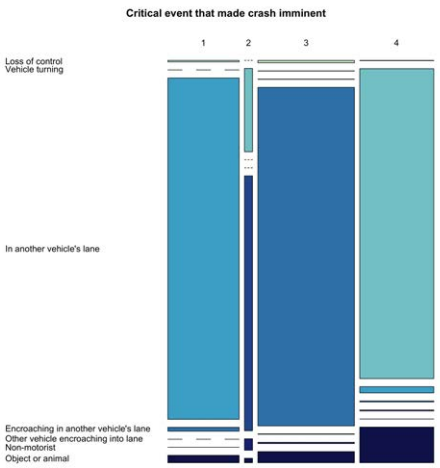
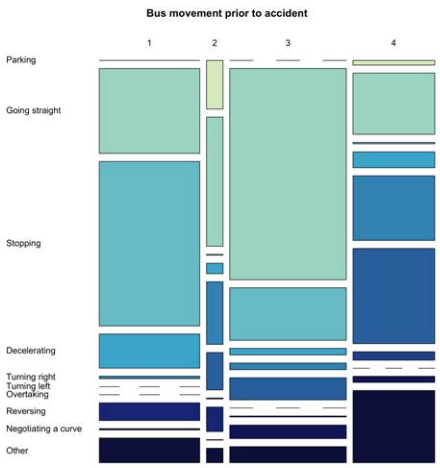


Figure 7: Mosaic plots giving the breakdown of the variable "bus movement prior to the crash" for each of the four clusters.

(a) 2005–2009



(b) 2010–2015

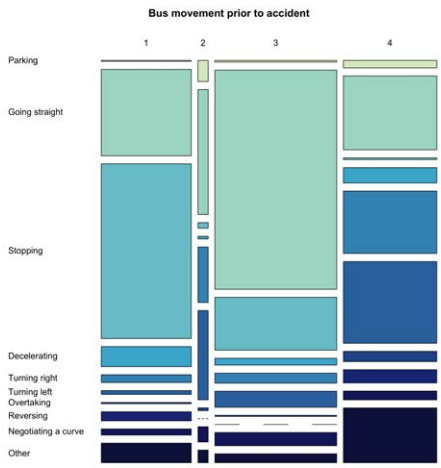
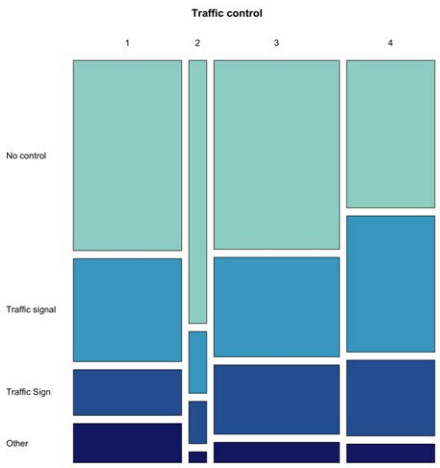


Figure 8: Mosaic plots giving the breakdown of the variable "traffic control devices applicable to the bus at the time of the accident" for each of the four clusters.

(a) 2005–2009



(b) 2010–2015

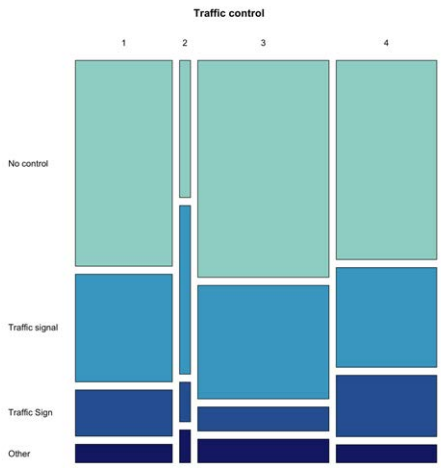
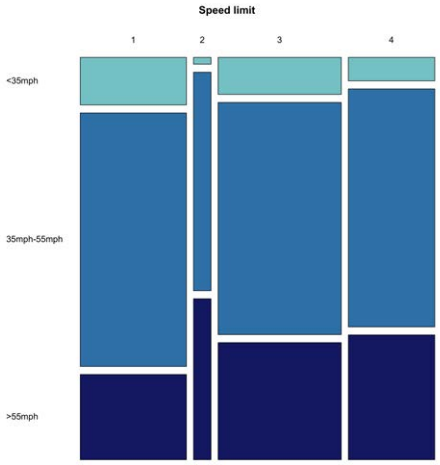
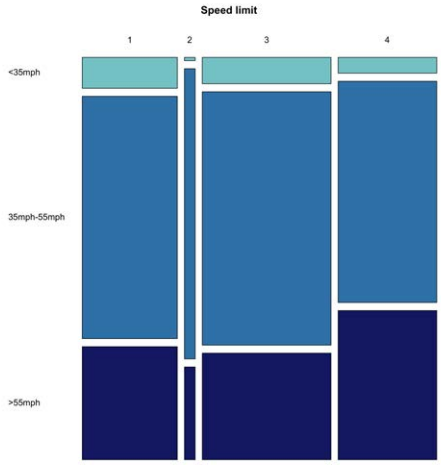


Figure 9: Mosaic plots giving the breakdown of the variable "speed limit of the traffic way where the accident occurred" for each of the four clusters.

(a) 2005–2009



(b) 2010–2015



List of Tables

Table 1: Structure of the various data files that form the GES data

Data file name	Multiple records per accident	Description of variables in file	Unique record identifier
Accident	No	Accident characteristics	Case number
Vehicle	Yes	Characteristics of all vehicles involved	Case number, vehicle number
Person	Yes	Characteristics of all persons involved	Case number, vehicle number, person number
Distract	Yes	Lists person-level distractions	same as Person
Visual	Yes	Lists person-level visual obstructions	same as Person
Impair	Yes	Lists person-level impairments	same as Person

Table 2: Taxonomy of crashes, 2005–2009

Characteristic	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Single vs. Multiple vehicles (>85%)	Multiple	Single	Multiple	Multiple
Non-motorist involvement	0.1%	99.5%	0.4%	0%
Bus movement prior to crash	stopping (49.8%)	going straight (39.2%), turning left (19%), parking (14.7%)	going straight (64%), stopping (15.9%)	overtaking (28.9%), other (21.9%), turning left (19.6%)
Critical event that made the crash imminent	in another vehicle's lane (92.4%)	vehicle turning (48.1%), other vehicle encroaching into lane (30.7%)	encroaching in another vehicle's lane (90.5%)	vehicle turning (96.9%)
Bus driver distracted	27.4%	33.9%	20.2%	31.8%
Other vehicle's driver or the non-motorist charged with alcohol/drug/impairment related offense	22.1%	4.7%	20.5%	10.5%
Was a school bus involved?	51.8%	39.8%	39.9%	32.2%
Daylight condition - was there sufficient light?	90.1%	89.5%	86.1%	81.4%
Did the accident happen at an intersection?	40.8%	44.3%	49.6%	66.9%
Bus driver gender (male?)	56.9%	55.8%	62.7%	65.7%
Light/ heavy truck involvement	21.6%	0.1%	22.5%	19.9%
Was the other driver/ non motorist distracted?	30.6%	100%	24.9%	4.9%
Was there a pick-up truck/ van/ SUV involved?	28.4%	24.8%	27.1%	0%
Single lane vs. multiple lanes	22.4%	35.3%	31.1%	36.7%

Table 3: Taxonomy of crashes, 2010–2015

Characteristic	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Single vs. Multiple vehicles (>85%)	Multiple	Single	Multiple	Multiple
Non-motorist involvement	0.1%	100%	0.02%	0.3%
Bus movement prior to crash	going straight (66.4%), stopping (16%)	going straight (37.9%), overtaking (27.1%)	overtaking (24.8%), going straight (22.4%)	stopping (52.9%), going straight (26.2%)
Critical event that made the crash imminent	in another vehicle's lane (96.4%)	other vehicle encroaching into lane (72%), vehicle turning (23.5%)	encroaching in another vehicle's lane (95.6%)	vehicle turning (87.5%)
Bus driver distracted	10.8%	14.3%	3.2%	19.6%
Other vehicle's driver or the non-motorist charged with alcohol/drug/impairment related offense	18.4%	25.7%	13.9%	5.4%
Was a school bus involved?	53.2%	24.6%	38.8%	44.6%
Daylight condition - was there sufficient light?	88.4%	65.9%	84.3%	84.9%
Did the accident happen at an intersection?	48.7%	59.9%	53.3%	54.9%
Bus driver gender (male?)	57.9%	79.6%	59.1%	67.5%
Light/ heavy truck involvement	0%	15.2%	19.4%	13.4%
Was the other driver/ non motorist distracted?	0%	5.5%	14.2%	12.5%
Was there a pick-up truck/ van/ SUV involved?	26.8%	0.2%	26.7%	23.7%
Single lane vs. multiple lanes	2.9%	3.2%	1.6%	1.7%

Table 4: Key differences between 2005-2009 and 2010-2015 clusters. Figures are rounded to nearest percentage point.

Characteristic	Cluster	2005-2009	2010-2015
Bus movement prior to crash	2	Parking (15%) Turning left (11%)	Parking (6%) Turning left (27%)
Critical event that made the crash imminent	2	Vehicle turning (48%) Other vehicle/ non motorist encroaching (31%)	Vehicle turning (23%) Other vehicle/ non motorist encroaching (72%)
	4	Object/animal (2%)	Object/animal (10%)
Speed limit of the road	2	35–55 MPH (57%) >55 MPH (42%)	35–55 MPH (75%) >55 MPH (24%)
Traffic control devices	2	No control (69%) Traffic signal (16%)	No control (36%) Traffic signal (45%)
	3	No control (50%) Traffic sign (18%)	No control (57%) Traffic sign (6%)
	4	No control (39%) Traffic signal (36%)	No control (53%) Traffic signal (26%)
School bus involved?	2 4	40% 33%	24% 45%
Located at intersection?	2 4	44% 67%	60% 55%
Single lane?	all	Decreased proportion for all from 2005-2009	
Bus driver distracted?	all	Decreased proportion for all from 2005-2009	
Other driver distracted?	2 4	100% 5%	5% 12%
Other driver under influence?	2	5%	26%
Daylight?	2	89%	66%
Pickup truck/SUV/ van involved?	4	0%	24%