HHS Public Access

Author manuscript

J Safety Res. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:

J Safety Res. 2016 September ; 58: 79–87. doi:10.1016/j.jsr.2016.07.002.

Off-road truck-related accidents in U.S. mines

Saeid R. Dindarloo[a,*], **Jonisha P. Pollard**[b], and **Elnaz Siami-Irdemoosa**[c]

[a]Department of Mining and Nuclear Engineering, Missouri University of Science and Technology, MO, USA

[b]Workplace Health Branch, Pittsburgh Mining Research Division, NIOSH, PA, USA

[c]Department of Geoscience and Geological and Petroleum Engineering, Missouri University of Science and Technology, MO, USA

Abstract

Introduction—Off-road trucks are one of the major sources of equipment-related accidents in the U.S. mining industries. A systematic analysis of all off-road truck-related accidents, injuries, and illnesses, which are reported and published by the Mine Safety and Health Administration (MSHA), is expected to provide practical insights for identifying the accident patterns and trends in the available raw database. Therefore, appropriate safety management measures can be administered and implemented based on these accident patterns/trends.

Methods—A hybrid clustering-classification methodology using K-means clustering and gene expression programming (GEP) is proposed for the analysis of severe and non-severe off-road truck-related injuries at U.S. mines. Using the GEP sub-model, a small subset of the 36 recorded attributes was found to be correlated to the severity level.

Results—Given the set of specified attributes, the clustering sub-model was able to cluster the accident records into 5 distinct groups. For instance, the first cluster contained accidents related to minerals processing mills and coal preparation plants (91%). More than two-thirds of the victims in this cluster had less than 5 years of job experience. This cluster was associated with the highest percentage of severe injuries (22 severe accidents, 3.4%). Almost 50% of all accidents in this cluster occurred at stone operations. Similarly, the other four clusters were characterized to highlight important patterns that can be used to determine areas of focus for safety initiatives.

Conclusions—The identified clusters of accidents may play a vital role in the prevention of severe injuries in mining. Further research into the cluster attributes and identified patterns will be necessary to determine how these factors can be mitigated to reduce the risk of severe injuries.

Practical application—Analyzing injury data using data mining techniques provides some insight into attributes that are associated with high accuracies for predicting injury severity.

[*]Corresponding author at: 226 McNutt Hall, 1400 N. Bishop Ave., Rolla, MO 65409, USA. srd5zb@mst.edy (S.R. Dindarloo).

Disclaimer

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health. Mention of any company or product does not constitute endorsement by NIOSH.

Keywords

Off-road mining trucks; Fatalities and injuries; K-means clustering; Genetic programming; Classification

1. Introduction

Analysis of workplace injuries has been heavily utilized as a means to determine high-risk tasks, prioritize workplace redesign, and determine areas of concern for worker safety in many industries including healthcare, construction, retail and services, and mining (Cato, Olson, & Studer, 1989; Drury, Porter, & Dempsey, 2012; Mardis & Pratt, 2003; Moore, Porter, & Dempsey, 2009; Pollard, Heberger, & Dempsey, 2014; Schoenfisch, Lipscomb, Shishlov, & Myers, 2010; Turin, Wiehagen, Jaspal, & Mayton, 2001; Wiehagen, Mayton, Jaspal, & Turin, 2001). While many industries would require injury records from individual companies or insurance providers to perform an analysis, mining is uniquely suited for a more comprehensive injury analysis. An important feature of U.S. mining is the accessibility of injury records. The Mine Safety and Health Administration requires all mine operators and contractors to file a Mine Accident, Injury and Illness Report (MSHA Form 7000-1) for all reportable accidents, injuries, and illnesses incurred at U.S. mining facilities. Reportable illnesses include any illness or disease that may have resulted from work. The database of these reports is available in the public domain and is provided by the National Institute for Occupational Safety and Health (http://www.cdc.gov/niosh/mining/data/default.html). Each entry of the database contains 36 unique attributes including: mine id, mining method, accident date, degree of injury, accident classification, mining equipment, employee's experience and activity, and a narrative briefly explaining the accident. Previous mining research has examined the injury and fatality causes associated with maintenance and repair, haulage vehicles, ingress and egress from mobile equipment, operating underground and surface mining mobile equipment, and other mining tasks (Drury et al., 2012; Moore et al., 2009; Pollard et al., 2014; Reardon, Heberger, & Dempsey, 2014; Turin et al., 2001; Wiehagen et al., 2001). Traditional injury data analysis uses counts and cross-tabulations as a means to determine trends in injuries. While this typically yields useful information, more sophisticated data mining techniques may allow for more improved classification of injuries through identification of injury patterns.

Clustering and classification are the two widely used methods of data mining for the purpose of pattern recognition. Clustering is among the unsupervised methods of pattern recognition while the classification is a supervised learning method. By an unsupervised method, one means that the data analyzer does not have any prior hypothesis or pre-specified models for the data, but wants to understand the general characteristics or the structure of the high-dimensional data. A supervised method means that the investigator wants to confirm the validity of a hypothesis/model or a set of assumptions, given the available data (Jain, 2010). Clustering and classification are also called un-labeled and labeled, respectively. In pattern recognition, data analysis is concerned with predictive modeling: given some training data, the prediction task is to find the behavior of the unseen test data. This task is also referred to as learning. Often, a clear distinction is made between learning problems that are (i)

supervised (classification) or (ii) unsupervised (clustering), the first involving only labeled data (training patterns with known category labels), while the latter involves only unlabeled data (Duda, Hart, & Stork, 2001; Jain, 2010). Clustering and classifications are performed using differing algorithms but may be used together to improve prediction accuracy.

The aim of this research was to gain a better understanding of the factors associating with severe injuries (fatalities and permanent disabilities) in U.S. mining by employing data mining techniques. Clustering and classification were employed for a comprehensive analysis of off-road truck-related accidents and injuries reported to MSHA during a 13-year period (2000–2012). Gene expression programming was used for classification, allowing all injury attributes to be considered and tested to determine which were associated with the highest prediction accuracies. The most explanatory attributes were selected among the available 36 unique attributes in the MSHA database. Then, K-means clustering was used as a means to identify similarity/dissimilarity between the accident records using the selected attributes for the purposes of pattern recognition in the raw data. It should be noted that the goal of this study was not to establish cause–effect relationships between accident attributes and outcomes, but to: (a) use data mining to systematically identify important attributes from MSHA incident reports that are highly associated with the outcomes of accidents (classification), and (b) recognize patterns in the accidents (clustering) given a set of work-related attributes.

2. Materials and methods

2.1. MSHA injury data

A dataset comprised of 13 years (2000–2012) of Mine Accident, Injury and Illness Reports was selected beginning with 1/1/2000 (MSHA, 2014). From this dataset, records of severe injuries (fatalities and permanent disabilities) and non-severe injuries associated with off-road trucks were selected. The NIOSH code “minemach-44, all accidents related to off-road mining trucks” was identified to select the records of interest in this study. A total of 5,831 records of injuries (both severe and non-severe) were filtered for further analysis. This dataset included 125 severe records that affected 140 employees. These severe injuries consisted of 88 fatalities and 52 permanent disabilities. These records were analyzed using Minitab (Minitab Inc., State College, Pennsylvania), MATLAB (MathWorks, Inc., Natick, Massachusetts), Rapidminer (RapidMiner, Inc., Cambridge, Massachusetts) and GenExprotools (Gepsoft Limited, Bristol, UK) to determine factors associated with the highest counts of severe injuries.

2.2. GEP-clustering modeling

The objective of clustering is to discriminate between dissimilar data by dataset partitioning (clustering). As an unsupervised data mining technique, the aim of clustering is to split a heterogeneous dataset into several more homogenous groups. The optimization task is to maximize the similarity between the in-cluster members and dissimilarity between the out-cluster members. K-means clustering is used to partition a large, highly variable dataset such that like data are grouped together. As an example, one is given a set of n data points in d-dimensional space (R^d) and an integer k. The goal is to determine a set of k points in R^d,

called centers, so as to minimize the mean squared distance from each data point to its nearest center (Kanungo et al., 2002). Let $X = \{x_i\}$, $i = 1, \ldots, n$,be the d-dimensional observations which are clustered into a set of k clusters, $C = \{c_k, K = 1, \ldots, K\}$. The K-means clustering finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let μ_k be the mean of the cluster c_k. The squared error between μ_k and the points in cluster c_k is defined as shown in Eq. (1). The goal of K-means clustering is to minimize the sum of the squared error over all k clusters as shown in Eq. (2). (For a detailed review of the theory and background of K-means clustering see: Halkidi et al., 2001, Fraley and Raftery, 2002, Jain, 2010.)

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (1)$$

$$J(c) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (2)$$

Genetic programming (GP) creates a functional relationship between inputs (attributes) and outputs to predict the occurrence of the output based on the properties of the attributes. Genetic programming can be represented as a hierarchically structured tree comprising functions and terminals. Fig. 1 illustrates a simple representation of a GP tree for the function $Coshx/(C_1 sinx)$. The tree reads from left to right and from bottom to top. Mimicking the Darwinian principle of survival, the fittest solutions (smallest error) are chosen to generate a population of new offspring programs for the next generation (Koza, 1992). In the next step, some genetic operations, namely mutation and crossover, will generate new offspring from the fittest programs of the previous generation. The operator selects a random node in a tree and replaces it with another node or subtree. The new offspring will be evaluated with the error or fitness function. The process continues until reaching a predefined threshold in terms of the best fit or error. In GP, thousands of solutions (computer programs) are generated and evolved consecutively based on the Darwinian principle of survival. The search for the solution starts with a population of completely randomly generated programs (solutions) from a predefined set of available functions (e.g., arithmetic functions) and terminals (independent variables). All programs are measured against a fitness function (e.g., root mean square error in a regression problem) and the best ones survive and are bred to the next generation.

Proposed by Ferreira (2001), gene expression programming (GEP) is a development of the conventional GP. As with GP, in GEP the main steps include: the function set, terminal set, fitness function, control parameters, and stop criteria. The fundamental difference between GP and GEP resides in the nature of the individuals. In GP the individuals are nonlinear entities of different sizes and shapes (parse trees). In GEP the individuals are also nonlinear entities of different sizes and shapes (expression trees), but these complex entities are encoded as simple strings of fixed length (chromosomes). The chromosomes in this study

are the attributes of each record of injuries (60 attributes and 5831 records). Unlike the parse tree representation in conventional GP, GEP uses a fixed length of character strings to represent solutions to the problems, which are afterwards expressed as parse trees of different sizes and shapes. These trees are called GEP expression trees (ETs). One advantage of the GEP technique is that the creation of genetic diversity is extremely simplified as genetic operators work at the chromosomal level. Another GEP strength is its unique, multi-genic nature that allows the evolution of more complex programs composed of several subprograms. The GEP algorithm begins with an initial population of chromosomes, which are randomly generated, linear strings of a fixed length. Then, the linear chromosomes are expressed as ETs and the fitness of each individual is evaluated based on a predefined fitness function. The individuals are then selected, according to fitness, to form a new generation—i.e., the higher the fitness value, the more chance an individual has to be selected. The selected individuals are also subjected to reproduction with modification, through genetic operators like crossover, mutation, and rotation. The individuals of this new generation are, in their turn, subjected to the same developmental process: expression of the genomes, selection, and reproduction. The process is repeated for a certain number of generations or until a solution has been found.

The GEP algorithm was used for classification of the off-road truck-related injury dataset. The objective was to identify the most explanatory (independent) variables (attributes) to be used in the subsequent clustering sub-model. One advantage of GEP over conventional classification algorithms (such as support vector machines and artificial neural networks) is that it can select the best explanatory features to achieve the highest fitness or accuracy. Hence, all available independent variables included in the incident reports, including both the categorical and numerical variables, were initially used in the GEP model building as explanatory variables. The injury severity was converted into a binary variable with: class 0 for those severe truck-related incidents that resulted in either fatalities or permanent disabilities, and class 1 for all other injuries. Two-thirds of the records (3,888) were used for training the algorithm and the remaining (1,943) for model testing. The fitness function was set as the sensitivity/specificity to achieve maximum classification accuracy. K-means clustering was used to group similar injuries therefore maximizing the similarity of records in each cluster. All data (both severe and non-severe) were analyzed together and separated into clusters.

3. Results

3.1. MSHA injury data overview

A brief analysis of the MSHA injury data was performed to give an understanding of the types of data available in the injury records. A temporal illustration of off-road truck-related fatalities and disabilities (Fig. 2) shows fluctuations in the fatality and disability counts over the years with a clear decreasing trend. The distribution of accident types is shown in Fig. 3. An analysis of the type of mine operation found that most severe injuries occurred at surface mines that use strips, quarries, or open pits as shown in Fig. 4. Employee job experience was also examined as shown in Fig. 5a. Frequencies of severe versus other injuries among victims of different job experiences are illustrated in Fig. 5b. Nearly half of the affected

employees had less than 5 years of job experience. Work activity at the time of injury was also examined as shown in Fig. 6.

Most of the severe injuries were sustained while the employee was operating the off-road truck. In the 56 severe accidents (57 victims) that occurred while operating a truck, 46 employees sustained fatal injuries and 11 employees were permanently disabled. The four main causes of these accidents (as identified in the report of the MSHA investigation) were: (i) losing control of the truck, (ii) berm/dump failure, (iii) unsafe/careless actions, and (iv) truck/component mechanical failure. In the category of “unsafe/careless actions,” MSHA identified that the underlying root causes were either associated with poor safety training, management and enforcement with managerial causes, or failure of the employees to obey safety regulations (personal issues). In nearly all of the 26 accidents (affecting 27 victims) that occurred during the “maintenance/repair” activity, either the employees' failure to regard safety regulations or a component/equipment malfunction/failure caused the injuries. A major cause of fatalities in this category was failure to block off/lockout the truck/bed. This was either due to the employees' unsafe actions or the mechanical malfunction of the equipment. In the “other” category, 17 fatal accidents resulted in 15 fatalities and 7 permanent disabilities. The main causes of this category were: (i) the operator(s) exited the cab while operating the equipment and (ii) unsafe/careless minor repairing attempts.

3.2. Results of gene expression programming

Gene expression programming utilized all available variables in the incident records to determine those that best predicted injury severity. The important GEP parameters (number of chromosomes, head size, and number of genes) were determined through a grid search algorithm to achieve the maximum classification accuracy. A grid search is an exhaustive search in the pre-specified domains of the parameters of interests. In this case, using the grid search, the best results (highest accuracies) were achieved when the three parameters were set to 34, 11, and 5, respectively. The resulting best model is shown in the GEP expression tree in Fig. 7 with the key in Table 1. The two classes are “severe” and “other injuries.” Results of this programming elicited five attributes that best explained the injury severity. These attributes were: the operation subunit (NIOSH code “subunit”), the month of the year in which the accident occurred (NIOSH code “month”), the victim's job experience at the time of the accident (NIOSH code “expjob”), the employee's activity at the time of the accident (NIOSH code “mwactiv”), and the type of operation (NIOSH code “commod,” including: coal, metal, non-metal, stone, sand & gravel operators, as well as, coal and non-coal contractors). The overall classification accuracies of 64.55% and 65.65% were obtained for the training and testing datasets, respectively. The classification accuracies for the severe injury class and the all other injuries class in the testing dataset were 63.04% and 64.69%, respectively (see Table 2).

In order to examine the accuracy of the GEP classification algorithm, the whole database was modeled with the widely-used traditional method of logistic regression for binary classification (Cox, 1958). In a binary logistic regression, the dependent variable is a two-class categorical variable. Here, the two classes were severe and non-severe accidents. The objective of the regression was to classify the dataset using the available attributes. Similar

to GEP, the logistic regression builds a relationship between the (categorical) dependent variable and (a combination of categorical, ordinal, and numerical) independent variables (attributes). A logistic regression model assigns to each set of attributes a number between 0 and 1. This output can be interpreted as the probability of a set of attributes belonging to each class. For instance, an output value equal to 0.3 means that the accident (given its attributes) is a non-severe one (class 0) with a probability of 70% and a severe accident (class 1) with a probability of 30%. Using the same training, testing, and attribute sets as the GEP, the results of logistic regressions were obtained for the whole dataset and are compared with the GEP results in Table 2. GEP had a superior performance compared to logistic regression resulting in a more accurate classification of both severe and non-severe accidents.

3.3. Results of K-means clustering

Using the GEP sub-model's output, the resulting clusters from the K-means clustering were defined based on their most dominant attribute, which discriminated them from the other clusters while maintaining the highest similarity inside the cluster. Therefore, the clusters were defined based on their attributes as specified in Table 3a. The optimal number of clusters was identified using the GAP Statistics method (Tibshirani, Walther, & Hastie, 2001). Five clusters were created as defined in Table 4. The number of incidents in each cluster and the average job experience of employees at the time of the accident are provided in Tables 3a, 3b, 3c, and 3d along with the dominant subunit and activity, operation, and month codes for each cluster. The activity, operation type, and subunit codes in Table 3a are defined in Tables 3b–3d. The last column in Table 3a shows the number (percentage) of severe accidents (Codes 1 and 2) inside each of the five proposed clusters. In summary, Tables 3a, 3b, 3c, and 3d show the most discriminating attribute(s) for each cluster. For instance, in cluster 0, the operation code 4 (stone operator) is dominant (48%). This means almost half of the severe injuries were occurred in the “stone operator” classification. Over two-thirds (72%) of the victims had less than 5 years of “job experience.” Also, the dominant subunit code was 9 (i.e., mill or preparation plants, see Table 3d), which includes over 90% of this cluster's accidents. Similarly, the major activity codes, in cluster 0, were 1 and 5 (see Table 3b) which implied that the two activities of “getting on/off truck” and “driving off-road truck” were associated with 72% of the accidents during the period 2000–2012. In this cluster, a total of 22 fatalities and permanent disabilities have been recorded that account for 3.4% of the total reported accidents (including non-severe ones).

4. Discussion and conclusions

Truck-related fatalities have been previously examined in the literature. An analysis of fatal truck-related accidents during the period 1995–2006 revealed the three most frequent causes of the haul truck-related fatalities as: (i) failure of victims to respect haul truck working area, (ii) failure to provide adequate berms, and (iii) failure of mechanical components (Md-Nor, Kecojevic, Komijenivic, & Groves, 2008). Another study within the period of 1995–2002 (Kecojevic & Radomsky, 2004) categorized the major causes of fatalities as follows: (i) failure of mechanical components (22%); (ii) lack of and/or failure to obey warning signals (20%); (iii) failure to maintain adequate berm (13%); (iv) inadequate hazard training (10%);

and (v) failure to recognize adverse geological conditions (10%). Kecojevic and Radomsky (2004) associated 70% of all fatalities to the above five categories, which are consistent with the results of studies for periods 1995–2006 and the current study. Ruff, Coleman, and Martini (2011) studied the equipment-related fatalities for the period 2000–2007 and found that, for mobile equipment, the most frequent fatalities were related to loss of control or visibility issues during the operation of the equipment. A more recent analysis examined 133 fatality reports for the period 1995–2010 using a previously developed coding scheme to determine repeating patterns of accidents (Drury et al., 2012). In this work, the authors were able to more fully develop the classification patterns previously reported by Wenner and Drury (2000). This scheme was broken into driving and non-driving accidents. Under driving, the factors included: loss of control, failure of ground, two-vehicle collisions, mechanical failure (sudden and inadequate performance), and leaving the driving track. Under non-driving, the factors included: unexpected movement (of the vehicle or part of the vehicle or vehicle's load), falls from vehicle, and hit by other vehicle. Comparison of the results of these previous studies with the current study reveals that the root causes of truck-related fatalities have not changed in the past two decades. A new method to determine factors associated with these injuries is needed.

Gene expression programming was found to successfully predict severe accidents in mining based on available injury data. The accuracies obtained using GEP for both the training and testing data are comparable or superior to those reported in car and traffic accident studies (Abdelwahab & Abdel-Aty, 2001; De Oña, López, Mujalli, & Calvo, 2013; De Oña, Mujalli, & Calvo, 2011; Mujalli & De Oña, 2011). Therefore, the GEP model's inputs (variables) had the highest explanatory characteristics, among 36 attributes, for severity estimation. The break-down of injuries within the clusters was quite interesting.

Minerals processing mills and coal preparation plants essentially formed their own cluster (cluster 0). Therefore, there appears to be a unique element to these plants that created different truck-related injury scenarios from the other subunits. Typically, this subunit includes mills, coal preparation plants, breaker operations, shops, and yards associated with one specific mine. These are non-production locations and likely utilize differing types of equipment and have different geographies and layouts from the pits or underground mining locations. Also of the 653 incidents in this cluster, in the past 13 years about 31% of all accidents (203 instances) happened during June–August. The main reason(s) behind the high accident rates during these months should be examined in a case by case manner for all of the mining sites in this cluster. Therefore the root cause(s) for the higher accident rates cannot be identified for this particular cluster of mining sites. Furthermore, per Table 3a over 70% of the victims had job experiences in the range of 0–5 years. Therefore, the inexperienced crews in all of the mining sites in cluster 0 experienced more accidents than other four clusters. The distribution of these inexperienced workers within this cluster as compared to the other clusters is not known so determining the relative rates of accidents was not possible. However, the high number of accidents in the inexperienced worker can justify implementing further training and safety measures. Additionally, over 70% of all off-road truck-related accidents happened when the truck operators were either getting on/off the trucks or when they were operating the trucks. Previous research has identified the safety issues associated with getting on and off of mobile equipment (Moore et al., 2009). Thus

future safety improvement plans for this cluster of mining operations should focus on the root causes of ingress and egress injuries as well as those injuries sustained during operation of these trucks.

The interesting pattern in Cluster 1 is that in 55% of all cases, the victim had been inspecting the truck for maintenance/repair with non-powered hand tools. Moreover, over half of the accidents were related to either the coal or stone operators. Although this cluster did not include any severe injuries, it is clear that the use of hand tools for vehicle inspection creates a hazard at mines. Also similar to cluster 0, in this cluster the highest numbers of accident were recorded in June–August period. Unlike cluster 0, in cluster 1 the share of the inexperienced employees for all of the accidents is not very high with respect to more experienced employees. Also, a majority of the accidents (78%) took place in either strip or open pit mines.

With 2,352 accidents, cluster 2 was the largest cluster in this study. Interestingly all of the accidents in the past 13 years in this cluster happened in the second half of the year. Although it is not within the scope of this study to analyze the root causes or contributing factors for this unique pattern, it provides a well-justified guideline for conducting future, more in-depth analyses to determine activities or tasks associated with high numbers of accidents. Also, similar to cluster 1, in the past 13 years most of the accidents in this cluster happened in surface (strip or open pit) mines. In summary, cluster 2 showed that ingress and egress from off-road trucks is still a problem and requires immediate attention.

With 187 accidents, cluster 3 was the smallest cluster in this study. Similar to cluster 1, no severe accidents had been reported since 2000. Over 60% of the accidents were associated with a non-specified activity and coded as code 6 (other). Therefore, the activities preceding accidents in this cluster were not within the more frequent accident causing activities (e.g., getting on/off truck, handling supplies). In terms of the victim job experience this cluster did not differ from clusters 1, 2, and 4 (see Table 3a).

Similar to cluster 2, one interesting pattern in cluster 4 is that no accident (severe or non-severe) had been recorded in one half of the year. Unlike cluster 2, all of the accidents in cluster 4 occurred in the first half of the year since 2000. Again, this pattern (along with the pattern in cluster 2) is a reasonable justification for conducting seasonal studies for improving safety in all of the mining sites in this cluster. Thus, the main reasons for zero accidents in the first half of each year and a total of 2,192 in the second half should be identified. Also, nearly 90% of these accidents occurred at surface mines. Cluster 4 was also found to have the highest number of severe accidents. More in-depth analysis would need to be conducted to determine the key differences between the accidents in clusters 2 and cluster 4 to determine what appears to be a seasonal division in accidents contributing to increased accident severity. More training may be needed to address the injuries in this cluster.

This study was limited to data associated with off-road trucks in mining. Thus all other injuries that were not associated with off-road trucks were not included in this study. It is unclear whether this hybrid methodology will have sufficient accuracies when applied to other types of injuries. Furthermore, the study analyzed only the accident reports within the

period January 2000 to December 2012. Data included in this analysis were limited to that provided in the MSHA incident reports which have been collected, organized, and released to the public by NIOSH. Increasing the level of detail provided in incident reports will likely result in improved injury clustering and classifications. It is also important to note that correlation does not equal causation and no causal relationships are implied based on the results of this analysis. A more in-depth analysis will be necessary to determine causal factors for injuries sustained in each cluster.

In conclusion, clustering, classification, and a hybrid methodology of both—using K-means clustering and gene expression programming—was shown to be effective when analyzing mining injury data. In particular, the GEP classification sub-model had a better performance than the traditional method of logistic regression in accident type classification. Furthermore, the identified patterns in the accident database using the clustering method are not achievable with the traditional injury data analysis which uses counts and cross-tabulations. Therefore, determining the most dominant attributes for specific types of injuries (classification) and separating (clustering) the data based on these attributes may allow researchers to better understand the nature and causal factors of mining injuries. Clustering could likely improve traditional injury analysis methods. Injuries sustained in minerals processing mills and coal preparation plants, and injuries sustained during the operation of trucks, should be analyzed separately from larger datasets. The use of non-powered hand tools for off-road truck inspections, ingress and egress from off-road trucks, and newer employees operating dump trucks should be investigated further to determine ways to prevent these severe and non-severe accidents. While this analysis was limited to those injuries associated with off-road trucks, it is expected that larger, broader injury data analyses will also benefit from this hybrid clustering-classification methodology.

5. Practical applications

Analyzing the injury data using data mining techniques provides some insight into attributes that are associated with high accuracies for predicting injury severity. Off-road truck-related injuries continue to plague the mining industry resulting in fatalities, permanent disabilities, and other less severe injuries. Many factors contribute to these injuries, and many of these are likely preventable. This analysis revealed that injuries associated with the use of non-powered hand tools for off-road truck inspections, ingress and egress from off-road trucks, and newer employees operating dump trucks are areas deserving of attention to determine ways to prevent future injuries.

References

Abdelwahab HT, Abdel-Aty MA. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record. 2001; 1746:6–13.

Cato C, Olson DK, Studer M. Incidence, prevalence and variables associated with low back pain in staff nurses. American Association of Occupational Health Nurses Journal. 1989; 37:321–327.

Cox DR. The regression analysis of binary sequences (with discussion). Journal of the Royal Statistical Society B. 1958; 20:215–242.

De Oña J, López G, Mujalli R, Calvo FJ. Analysis of traffic accidents on rural highways using latent class clustering and Bayesian networks. Accident Analysis and Prevention. 2013; 51:1–10. [PubMed: 23182777]

De Oña J, Mujalli RO, Calvo FJ. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention. 2011; 43(1):402–411. [PubMed: 21094338]

Drury, CG.; Porter, WL.; Dempsey, PG. Proceedings of the human factors and ergonomics society 56th annual meeting. Santa Monica, CA: Human Factors and Ergonomics Society; 2012. Patterns in mining haul-truck accidents; p. 2011-2015.

Duda, R.; Hart, P.; Stork, D. Pattern classification. 2nd. New York: John Wiley and Sons; 2001.

Ferreira C. Gene expression programming: A new adaptive algorithm for solving problems. Complex Systems. 2001; 13(2):87–129.

Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002; 97(458):611–631.

Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. Journal of Intelligent Information Systems. 2001; 17(2-3):107–145.

Jain AK. Data clustering: 50 yrs beyond K-means. Pattern Recognition Letters. 2010; 31(8):651–666.

Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithms: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002; 24(7):881–892.

Kecojevic V, Radomsky M. The causes and control of loader- and truck-related fatalities in surface mining operations. Injury Control and Safety Promotion. 2004; 11(4):239–251. [PubMed: 15903158]

Koza, JR. Genetic programming, on the programming of computers by means of natural selection. Cambridge, MA: MIT Press; 1992.

Mardis AL, Pratt SG. Nonfatal injuries to young workers in the retail trades and services industries in 1998. Journal of Occupational and Environmental Medicine. 2003; 45(3):316–323. [PubMed: 12661189]

Md-Nor ZA, Kecojevic V, Komijenivic D, Groves W. Risk assessment for haul truck-related fatalities in mining. Mining Engineering. 2008; 60(3):43–49.

Mine Safety and Health Administration (MSHA). Summary of selected accidents/injuries/illnesses reported to MSHA under 30 CFR part 50, mine injury and worktime quarterly and self extracting files, 2000–2014. 2014. (Retrieved Dec 25, 2014 from) http://www.msha.gov

Moore SM, Porter WL, Dempsey PG. Fall from equipment injuries in U.S. mining: Identification of specific research areas for future investigation. Journal of Safety Research. 2009; 40:455–460. [PubMed: 19945559]

Mujalli RO, De Oña J. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research. 2011; 42(5):317–326. [PubMed: 22093565]

Pollard JP, Heberger J, Dempsey PG. Maintenance and repair injuries in US mining. Journal of Quality in Maintenance Engineering. 2014; 20(1):20–31.

Reardon LM, Heberger JR, Dempsey PG. Analysis of fatalities during maintenance and repair operations in the U.S. mining sector. IIE Transactions on Occupational Ergonomics & Human Factors. 2014; 2(1):27–38.

Ruff T, Coleman P, Martini L. Machine-related injuries in the US mining industry and priorities for safety research. International Journal of Injury Control and Safety Promotion. 2011; 18(1):11–20. [PubMed: 20496188]

Schoenfisch AL, Lipscomb HJ, Shishlov K, Myers DJ. Nonfatal construction industry-related injuries treated in hospital emergency departments in the United States, 1998–2005. American Journal of Industrial Medicine. 2010; 53(6):570–580. [PubMed: 20506460]

Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society, Series B (Statistical Methodology). 2001; 63(2):411–423.

Turin, FC.; Wiehagen, WJ.; Jaspal, JS.; Mayton, AG. Haulage truck dump site safety: An examination of reported injuries (DHHS (NIOSH) publication no 2001–124, information circular 9454).

Pittsburgh, PA: U.S. Department of Health and Human Services, Public Health Services, CDC-NIOSH; 2001.

Wenner C, Drury CG. Analyzing human error in aircraft ground damage incidents. International Journal of Industrial Ergonomics. 2000; 26(2):177–199.

Wiehagen, WJ.; Mayton, AG.; Jaspal, JS.; Turin, FC. An analysis of serious injuries to dozer operators in the U S Mining industry (DHHS (NIOSH) publication no 2001–126, information circular 9455). Pittsburgh, PA: U.S Department of Health and Health Services, Public Health Services, CDC-NIOSH; 2001.

Biographies

Saeid R. Dindarloo holds Ph.D., M.Sc. and B.Sc. degrees, all in Mining Engineering, from Missouri University of Science and Technology, USA, and Amirkabir University of Technology (Tehran Polytechnic), Iran. Dr. Dindarloo has 10 years of research and professional experience. He has extensive experience in mine design, planning, and computer application. His current research interests include: mini safety and health, mining machinery, and mining equipment management.

Jonisha P. Pollard M.S., CPE is a Research Engineer at the National Institute for Occupational Safety and Health in the Office of Mine Safety and Health Research. She joined NIOSH while pursuing her Master's degree in Bioengineering from the University of Pittsburgh in 2007. During her time at NIOSH, she has been involved in numerous research studies and has published research on knee injuries in mining, the effect of cap lamp lighting on balance, the effect of kneepads on balance, injuries due to maintenance and repair work in mining, locomotion in restricted workspaces and a variety of other topics. Ms. Pollard enjoys conducting field and laboratory research to positively impact the health and safety of mine workers.

Elnaz Siami Irdemoosa obtained her B.Sc. in Mining Engineering and her M. Sc. in Rock Mechanics from Amirkabir University of Technology (Tehran Polytechnic). She is currently a Ph. D. candidate at Missouri University of Science and Technology in the field of Geological Engineering. Her research interests include underground design and construction, tunneling, construction management, and geophysical method.

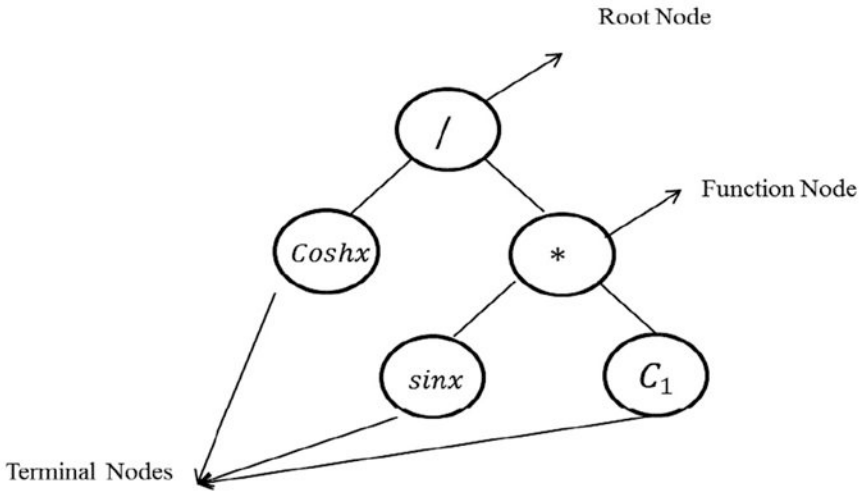

Fig. 1.
GP tree representation of $Coshx/(C_1 sinx.)$.

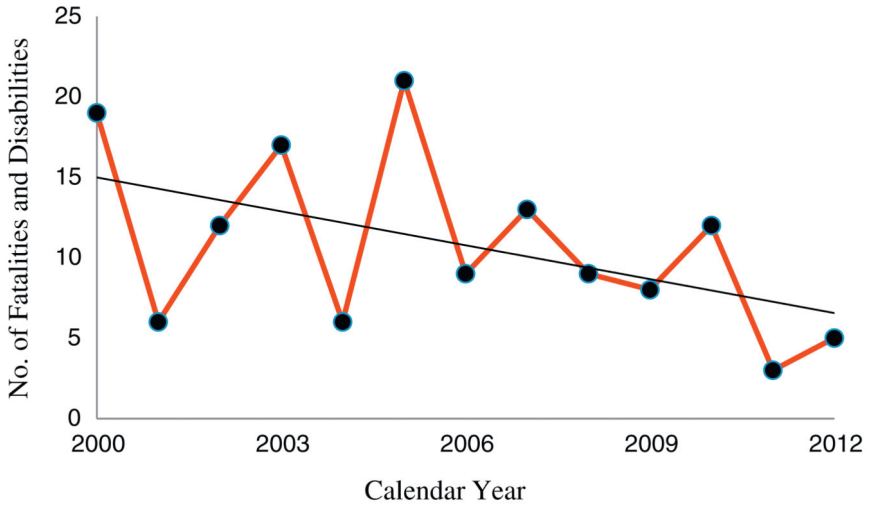

Fig. 2.
Time series of off-road truck-related severe injuries at US mines (2000–2012).

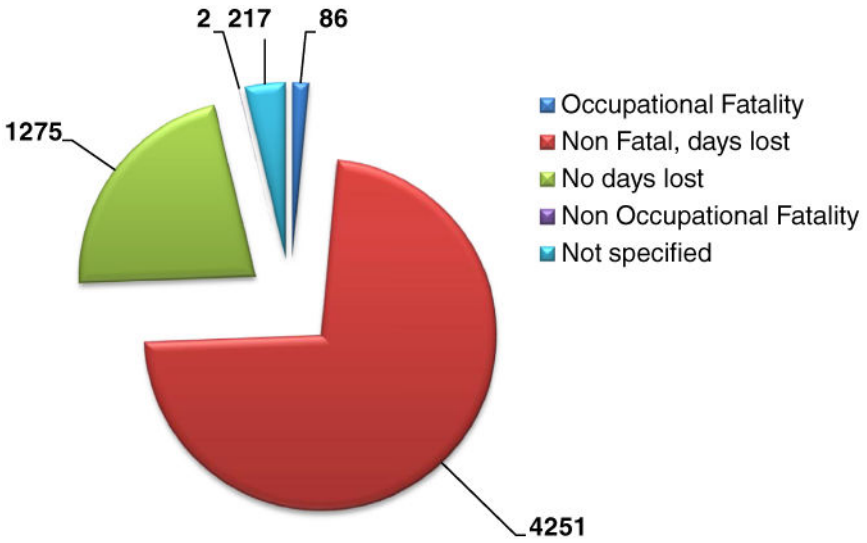

Fig. 3.

Distribution of off-road truck-related accident types at US mines (2000–2012). Only, two Non-Occupational Fatalities occurred.

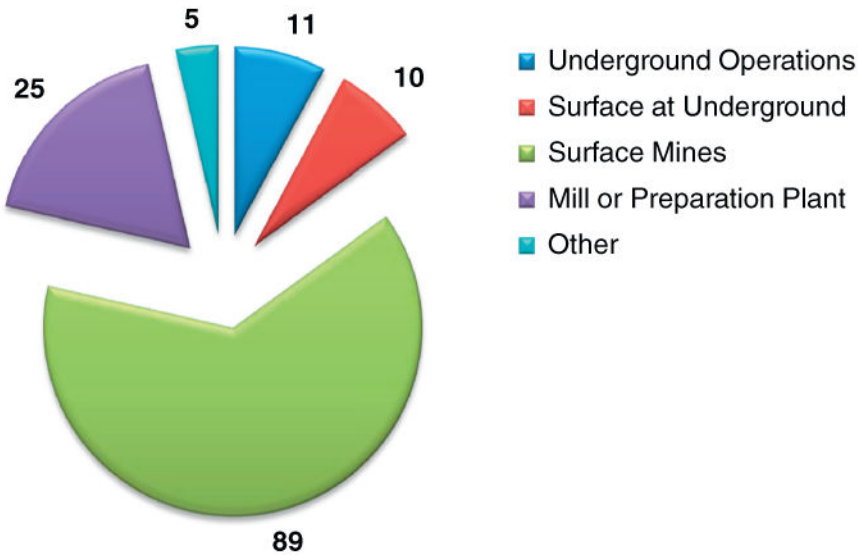

Fig. 4.

Distribution of mine operation type for severe injuries related to off-road trucks at US mines (2000–2012).

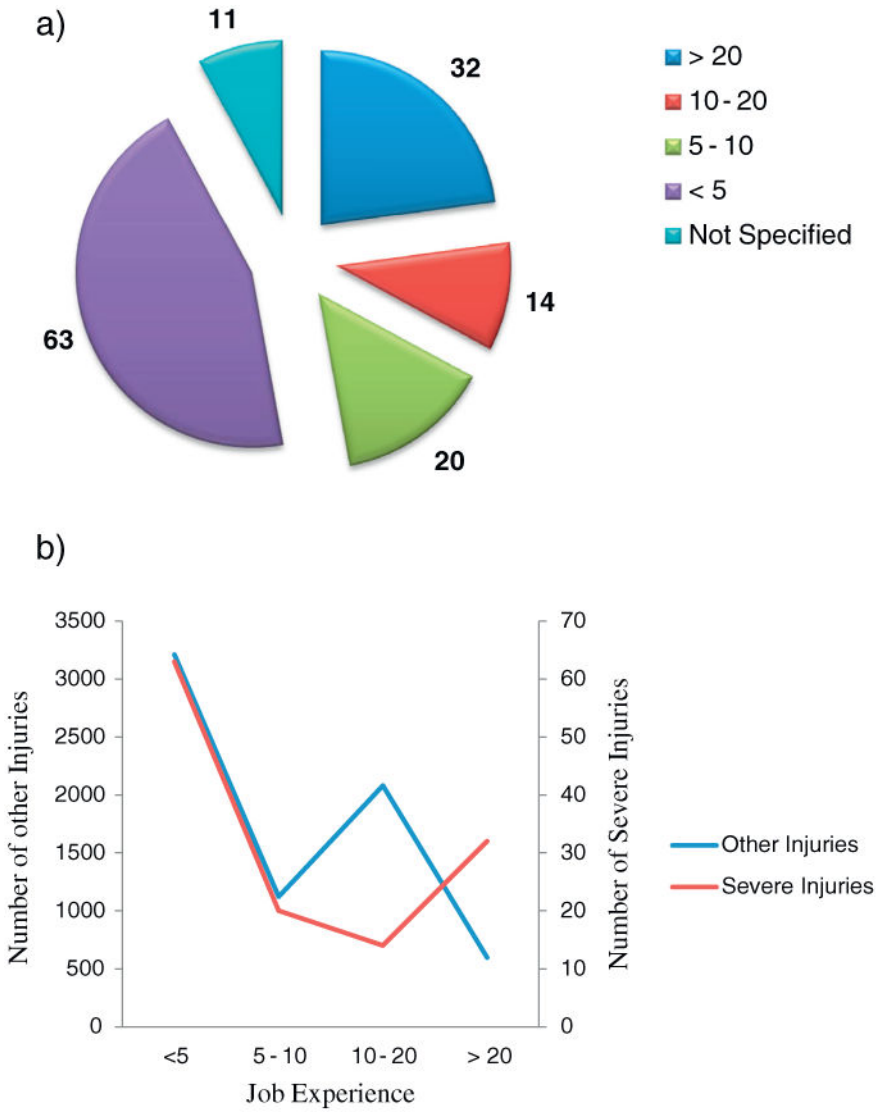

Fig. 5.

a Distribution of job experience (years) for severe injuries related to off-road trucks at US mines (2000–2012). b Distribution of job experience (years) for severe injuries vs. other injuries, related to off-road trucks at US mines (2000–2012).

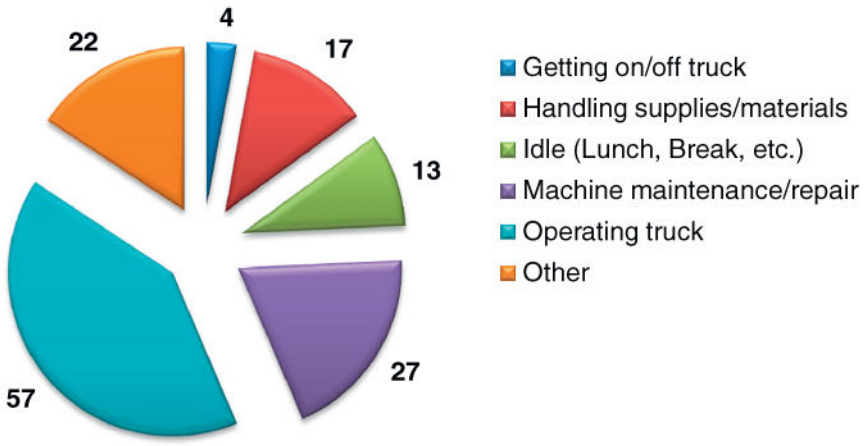

Fig. 6.
Distribution of worker activity at the time of off-road truck-related severe injury in US mines (2000–2012).

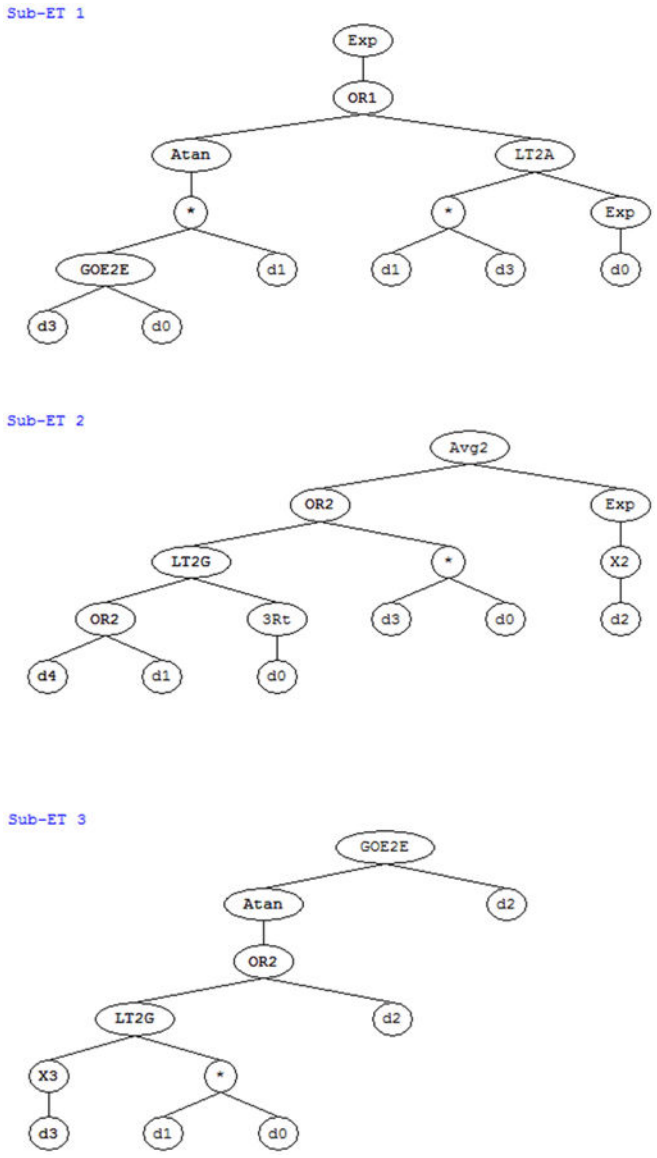

Fig. 7.

The three GEP expression trees (Sub-ET 1, Sub-ET 2, and Sub-ET 3) for classification of the injury severity.

Table 1

GEP tree key.

	Symbol	Definition
Functions	LT2A	if x<y, then x, else y
	3Rt	$x^{1/3}$
	LT2G	if x<y, then (x+y), else arctan(x*y)
	GOE2E	if x>=y, then (x+y), else (x*y)
	OR1	if x<0 OR y<0, then 1, else 0
	OR2	if x>0 OR y>0, then 1, else 0
	Avg2	avg(x,y)
	Exp	Exponential
	Atan	Arctan
	X2	X^2
Exlpanatory Variables	d0	Subunit
	d1	Month of the year
	d2	Activity
	d3	Job Experience
	d4	Operation

Table 2

GEP classification accuracies.

	Injury classification	Accuracy (%)	
		GEP	Logistic regression
Training	Severe	74.68	61.24
	Other	64.33	53.62
Testing	Severe	63.04	49.59
	Other	64.69	51.13

Table 3a

Cluster specifications.

Cluster	Number of incidents	Month of year	Operation Code	Average experience (years)	Subunit code	Activity code	Number of Severe Injuries
0	653	6, 7, 8 (31%)	4 (48%)	0–5 (72%)	9 (91%)	1, 5 (72%)	22 (3.4%)
1	371	6, 7, 8 (34%)	1, 4 (54%)	0–5 (51%)	3 (78%)	4 (55%)	0 (0%)
2	2,352	1–6 (0%)	1, 4 (55%)	0–5 (51%)	3 (78%)	1, 5 (79%)	58 (2.5%)
3	187	1, 4, 10 (33%)	1, 4 (59%)	0–5 (49%)	3 (68%)	6 (62%)	0 (0%)
4	2,191	7–12 (0%)	1, 4 (45%)	0–5 (57%)	3 (89%)	5 (52%)	45 (2.1%)
Total	5,754						125 (2.17%)

Table 3b

Activity code definitions.

Code	Description
1	Getting on/off truck
2	Handling supplies/materials
3	Idle (Lunch, break, etc.)
4	Machine maintenance/repair
5	Operating truck
6	Other

Table 3c

Operation code definitions.

Code	Description
1	Coal operator
2	Metal operator
3	Non-metal operator
4	Stone operator
5	Sand and gravel operator
6	Coal contractor
7	Non-coal contractor

Table 3d

Subunit code definitions.

Code	Description
1	Underground operations
2	Surface at underground
3	Surface: strip or open pit mining
4	Auger
5	Culm banks
6	Dredge
7	Other surface
8	Independent shop and yards
9	Mill or preparation plant
10	Office

Table 4

Cluster definitions.

Cluster	Label
0	Accidents related to minerals processing mills and coal preparation plants (91%). More than two-thirds of the victims in this cluster had job experiences of less than 5 years. This cluster is associated with the highest percentage of severe injuries (22 severe accidents, 3.4%). Almost 50% of all accidents in this cluster occurred at stone operations.
1	In 55% of the cases in this cluster, the victim had been inspecting the truck for maintenance/repair and using non-powered hand tools. Most of these accidents occurred at coal or stone operations. This cluster was not associated with any severe accidents.
2	No accidents occurred during the first half (January through June) of the year. Half of these victims were injured while getting on or off equipment, machines, etc. The average job experience for this cluster is less than 5 years. Also, the percentage of fatalities in this cluster is considerable. The cluster contains the highest absolute number of severe injuries, 58 victims.
3	In this cluster, 62% of all incidents occurred when the victim was either running or walking. No severe injuries were recorded in this cluster.
4	The second highest number of severe accidents occurred in this cluster (45 accidents). Over half of the accidents occurred when equipment operators with less than 5 years of experience (57%) were driving dump trucks (activity code 5).