Prediction and Evaluation of Accidents within Oilfields of Arabian Gulf Company Using Discriminant Analysis

Amer A. M. Boushaala
Faculty of Engineering, University of Benghazi
Benghazi, Libya
Amer.boushaala@uob.edu.ly

Intisar M. Elbergo
The Libyan Academy
Benghazi, Libya

Abstract— **The dynamic nature of oil and gas production is one of the major causes for various types of accidents resulting in injuries and fatalities in oil fields. The main purpose of this paper is to identify and classify accidents in Arabian Gulf Oil Company (AGOCO) through application of the Linear Discriminant Analysis (LDA) technique. Data collected were for 8 years spanning from 2005 to 2012, from four oil fields (Sarir, Nafoora, Messla, and Byda).**

The LDA is used to classify an accident into one of the accident groups; "Oil and gas Leak", "Fire", "Accident", and "Damage".

The developed discriminant functions revealed significant association between groups 60.1%, 22.75%, and 6.97% of between groups variability. However, the structure matrix revealed two significant predictors only of the first function, namely production area and camp with scores of 0.791 and 0.766 respectively. For the second function revealed also two significant predictions; the oil well and the transition station with scores of 0.806 and 0.740 respectively.

The third function has no significant predictors. The cross-validated method showed that 65.1% of the grouped cases are correctly classified.

Index Terms: **accidents prediction model, prediction model, Linear Discriminant Analysis (LDA), Discriminant Analysis (DA) methods, oilfield production, oilfield accidents.**

I. INTRODUCTION

The dynamic nature of oil production industry is one of the major causes for various type of accidents resulting in injuries and fatalities in oil production. It contains large volumes of flammable and hazardous chemicals, and it has a successive process which involves a lot of industries and professions, complicated technology and various kinds of equipments[1].

Furthermore, as the raw and assistant materials needed in the oilfield production are often flammable, explosive, poisonous and mordant, it is quite easy to have serious accidents, such as fire, explosions and poison leakage. These cause great injuries to workers, and damage to the company properties [2].

The main purpose of the paper is to study the classification or identification methods of oil fields accidents, prediction and evaluation methods using LDA. This is because each significant real world classification problem has its own properties, requirements and challenges. Oilfield accidents can be considered as a classification problem in this paper.

There are two main methods for classification: cluster analysis and discriminant analysis. In cluster analysis or unsupervised learning, the groups (or class) are unknown a prior; and the task is to determine these classes from the data. In discriminant analysis or supervised learning, we have a learning sample (or training sample) of the data to construct (or build) a classification rule to predict the outcome for unseen objects.

The classification situation is characterized by the following: one has two sets of multivariate observations. The first set, called learning sample or training sample,{(xi, yi), i=1,…, n} consists of n observations, where X Є R^d represents an attribute vector, and yi is class label in the set {1,…, J}.The second set is referred to as test sample, which consists of observations for which such prior information is not available and which has to be assigned to one of the J classes [3].

Currently, there is limited research on the urban and spatial dimensions of the prediction of accidents in oil production in Libya. Therefore, there is a need for prediction of such adverse impact of the oil accidents on the environment, workers and equipment in sites. This work lays the foundation to evaluate the effects and consequences of major accidents in oilfield. The locations selected for the study are Sarir, Messla, Naffora, and Byda oil fields.

The focus of the paper is on oil fields of AGOCO and deals with physical locations of the study areas. The locations cover accident hotspot areas in the company. AGOCO is one of the biggest oil companies in Libya and also stands as one of the largest oil company in North Africa. The data of accident records consider all aspects of oil well, production area, Transition station, and camp of oilfields.

Received 10 November 2017; revised 26 November 2017; accepted 11 November 2017.

Available online 11 December 2017.

A. *Objectives of Study*

1- Prediction and identification of oil fields accidents at four sites.
2- Minimization of misclassification error by determination of accident type that is identified with an object (number of observations).
3- Finding out the most frequent accidents affecting the sites of oil fields, using prediction models, LDA.
4- Classifying the oilfields accidents into mutually exclusive groups on basis of a priori information.
5- Studying differences among groups, using linear combination of predictor that identify the class or group of an object.

B. *The Significance of Study*

1- Based on this study, scientific and efficient countermeasures can be put forward so as to provide a base for reducing the accident rate and severity.
2- It is of important theoretical and practical significance to improve oilfield production safety in AGOCO.
3- It establishes the roles of the company management and workers with regard to vigilance in ensuring their safety and health in the workplace.

II. RESEARCH METHODOLOGY

One of the methods available to estimate test error or (generalization error), is the expected prediction error of future observations (drawn independently from the same distribution), including; re-substitution, cross-validation and bootstrap [4].

LDA is to be applied to develop accidents prediction models using regression methods. The performance of the approach is evaluated using statistical modeling metrics and incident prediction ability of each model. The data sets to be used in the analysis, and the statistical methods used to develop and evaluate the performance of the LDA based models are to be described. The classification and the creation of accidents type and the type of the sites are to be presented.

This research is based on reports of accident data collected from AGOCO which occurred between 2005 and 2012. Thus oilfields which had changed in terms of conditions and after 2012 are not included. The available sources for quantitative data collection and techniques of accidents analysis such as SPSS are used. The data, could be of various types accidents severity such as road accident, fire, oil and gas leak, damage.

The discriminant function analysis is to be introduced, and the discriminant coefficient discussed. Discriminant analysis derives an equation as linear combination of the independent variables that will discriminate best between the two or more groups of the dependent variables.

The discriminant equation:

$$D = \nu 0 + \nu 1\mathrm{X} + \nu 2\mathrm{X} + \cdots + \nu p\mathrm{X}p$$

where:

D= discriminant function.

ν0= a constant.

νp= the discriminant coefficient or weight for that variable.

Xp= respondent score for that variable.

p= the number of predictor variables.

The major underlying assumptions of LDA are:
1- The observations are random sample;
2- Predictor variable is normally distributed;
3- The allocations for the dependent categories in the initial classification are correctly classified;
4- There should be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group) [5].

There are several purposes of LDA:
1- To investigate differences between groups on the basis of the attributes of the cases, indicating which attributes contribute most to group separation.
2- To identify the linear combination of attributes known as canonical discriminant functions (equations) which contribute maximally to group separation.
3- To assign new cases to groups. The DA function uses a person’s scores on the predictor variables to predict the category to which the individual belongs.
4- To determine the most parsimonious way to distinguish between groups.
5- To test theory whether cases are classified as predicted.

III. ANALYSIS AND RESULTS

LDA predicts a group membership. Firstly, it examines whether there are any significant differences between groups on each of the independent variables using group means and ANOVA. The group statistics and Tests of Equality of Group mean tables provide this information. For example, mean differences between Transition Station and Production area depicted in Table I suggest that both may be good discriminators as the separations are large.

Figure 1 presents the data of Table 1 graphically, where the severity of the four classifications of accidents in each of the sections considered in the oil fields are presented. However, Figure 2 depicts ratings of the four types of accidents considered, in each of the oil fields included in this work.

Table 2 provides evidence of significant differences between means of the four types of accidents for all independent variables, with production area and camp indicating high values of F.

Table 1. Group Statistics

Category		Mean	Std. Deviation	Un-weighted	Weighted
Oil & Gas Leak	Transition Station	5.96	8.291	24	24.00
	Production area	0.08	0.408	24	24.00
	Camp	0.42	0.776	24	24.00
	Oil well	2.21	2.536	24	24.00
Fire	Transition Station	0.53	0.640	15	15.00
	Production area	0.87	0.834	15	15.00
	Camp	0.40	0.632	15	15.00
	Oil well	0.20	0.414	15	15.00
Accident	Transition Station	1.25	1.481	28	28.00
	Production area	2.89	2.061	28	28.00
	Camp	6.64	4.847	28	28.00
	Oil well	1.11	1.370	28	28.00
Damage	Transition Station	0.74	0.653	19	19.00
	Production area	0.21	0.419	19	19.00
	Camp	1.16	1.642	19	19.00
	Oil well	0.53	0.964	19	19.00

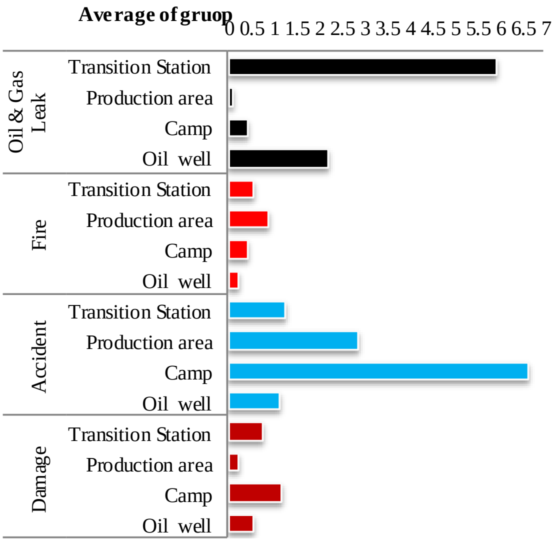

Figure 1. Graph of Four Categories

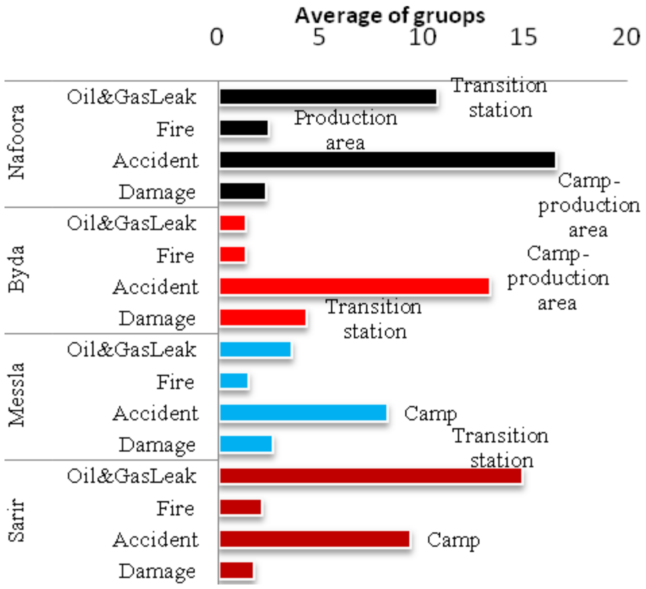

Figure 2. Graph of Four Fields

The Pooled Within-Group Matrices (Table III) also supports use of these independent variables as inter-correlations are low.

Table 2. Test of equality of group means

Independent variables	Wilks's Lambda	F	df$_1$	df$_2$	p-value
Transition Station	0.788	7.360	3	82	0.000
Production area	0.502	27.108	3	82	0.000
Camp	0.507	26.624	3	82	0.000
Oil well	0.820	6.002	3	82	0.001

Table 3. Pooled Within-Groups Matrices

Correlation matrix	Transition Station	Production area	Camp	Oil well
Transition Station	1.000	0.073	0.152	0.289
Production area	0.073	1.000	0.597	0.208
Camp	0.152	0.597	1.000	0.302
Oil well	0.289	0.208	0.302	1.000

A. *Eigen values result*

This provides information on each of the discriminate functions (equations) devolped. In our problem (Table IV) the three canonical correlations of 0.775, 0.477 and 0.264 suggest that the models explain 60.1%, 22.75% and 6.97% respectively, of the variation in the grouping variables.

Table 4. Eigen values

Function (Predictio n Model)	**Eigen value**	**% of Variance**	**Cumulative %**	**Canonical Correlation**
1	1.503^a	80.3	80.3	0.775
2	$.295^a$	15.7	96.0	0.477
3	$.075^a$	4.0	100.0	0.264

B. *Wilks's lambda*

Wilks's lambda indicates the significance of the discriminant function. Table V indicates a highly significant function (p-value < 0.05).

Table 5. Wilks's Lambda

Test of Function(s)	**Wilks's Lambda**	**Chi-square**	**df**	**p-value**
1 through 3	0.287	101.084	12	0.000
2 through 3	0.719	26.773	6	0.000
3	0.930	5.847	2	0.049

C. *The standardized canonical discriminant function coefficients*

Table VI provides an index of the importance of each predictor like the standardized regression coefficients (β's).

Table 6. Standardized Canonical Discriminant Function Coefficients

Independent variables	**Function**		
	1	**2**	**3**
Transition Station	-0.296	0.534	0.261
Production area	0.525	0.020	1.124
Camp	0.601	0.265	-1.071
Oil well	-0.341	0.568	-0.004

D. *The structure matrix table*

1- Table VII provides another way of indicating the relative importance of the predictors and it can be seen that the same pattern holds.

2- Many researches use the structure matrix correlations because they are considered more accurate than the standardized canonical Discriminant Function coefficients.

3- The structure matrix, TableVII shows the correlations of each variable with each discriminate function. By identifying the largest loadings for each discriminate function the researcher gains insight into how to name each function .

Table 7. structure matrix

Independent variables	Function(Prediction Models)		
	1	2	3
Production area	0.791*	0.335	0.503
Camp	0.766*	0.529	-0.362
Oil well	-0.136	0.806*	-0.018
Transition Station	-0.265	0.740*	0.179

Figure 3 (Radar graph) illustrates the largest loading for each discriminate function. Through this, the researcher gets another way to look at the structure of each function. From this graph for example, it can be seen that, production area and camp are of high scores which suggest a label of function 1 and Transition Station and Oil well have low scores in this function blue line.

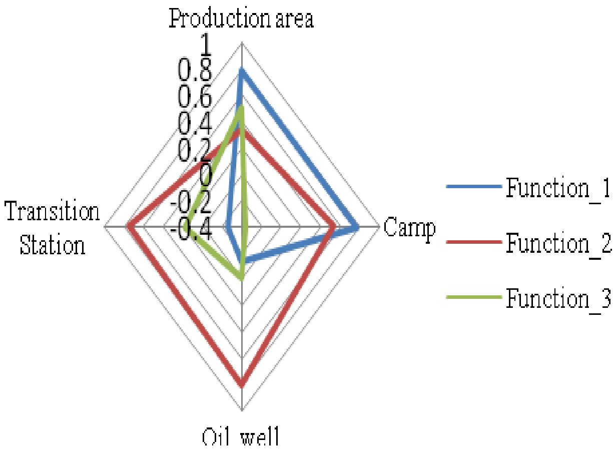

Figure 3. Discriminant Loadings

E. *The canonical discriminant function coefficient*

These un-standardized coefficients (β) are used to create the discriminant function (equation).In this case, Table VIII ilustrates the Canonical Discriminant Function Coefficients.

Table 8. Canonical Discriminant Function Coefficients

Independent variables	Function ((Prediction Models)		
	1	2	3
Transition Station	- 0.066	- 0.119	- 0.058
Production area	0.415	0.016	0.888
Camp	0.205	0.090	- 0.366
Oil well	- 0.210	0.348	- 0.002
(Constant)	- 0.627	- 0.924	- 0.212

Un-standardized coefficients

$$D_1 = -0.627 - 0.066(\mathit{Transition\ Station}) + 0.415(\mathit{Production\ area}) + 0.205(\mathit{Camp}) - 0.210(\mathit{Oil\ well})$$

$$D_2 = -0.924 - 0.119(\mathit{Transition\ Station}) + 0.016(\mathit{Production\ area}) + 0.090(\mathit{Camp}) + 0.348(\mathit{Oil\ well})$$

$$D_3 = -0.212 - 0.058(\mathit{Transition\ Station}) + 0.888(\mathit{Production\ area}) - 0.366(\mathit{Camp}) - 0.002(\mathit{Oil\ well})$$

F. *Group centroids*

A further way of interpreting discriminant analysis results is to describe each group in terms of its profile, using the group means of the predictor variables. These group means are called centroids. These are displayed in the Group Centroids as presented in Table IX. Figure 4 shows the canonical discriminant function, another way of looking at Groups Centroids.

Table 9. Functions at Group

Category	Function(Prediction Models)		
	1	2	3
Oil & Gas Leak	- 1.362	0.593	0.051
Fire	- 0.263	- 0.741	0.442
Accident	1.621	0.258	- 0.004
Damage	- 0.461	- 0.545	- 0.407

Un-standardized canonical discriminant functions evaluated at group means

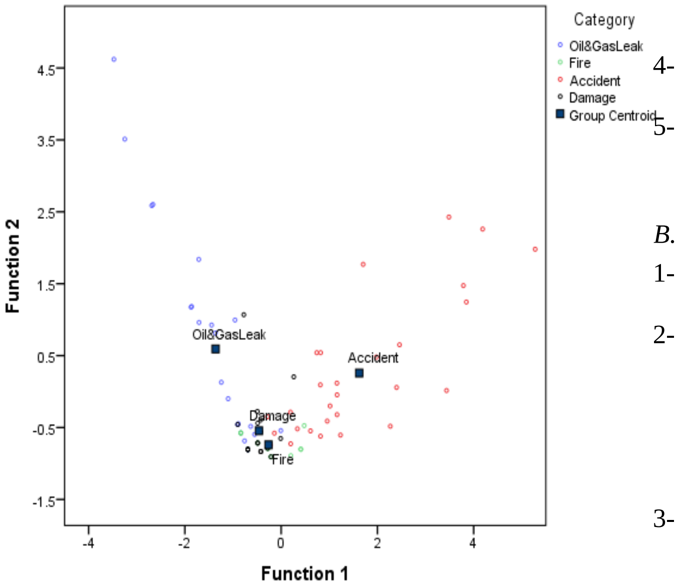

Figure 4. Canonical Discriminant Functions

G. *Confusion matrix or Classification table*

The classification results, Table X reveal that 66.3% of original categories were classified correctly with about 54.2% 'Oil & Gas Leak', 60.0% 'Fire', 75.0% 'Accident' and 73.7% 'Damage'. This overall predictive accuracy of the discriminant function is called the 'hit ratio'. The Cross-validated method has classified groups with slightly less accuracy (65.1%) than re-substitution method (66.3%).

Table 10. Classification Results a,b

Method	Category	Predicted Group Membership				Total
		Oil & Gas Leak	Fire	Accident	Damage	
Original (a) %	Oil & Gas Leak	54.2	4.2	.0	41.7	100.0
	Fire	.0	60.0	.0	40.0	100.0
	Accident	.0	7.1	75.0	17.9	100.0
	Damage	5.3	21.1	.0	73.7	100.0
Cross-validated (b) %	Oil & Gas Leak	54.2	4.2	.0	41.7	100.0
	Fire	.0	60.0	.0	40.0	100.0
	Accident	.0	7.1	71.4	21.4	100.0
	Damage	5.3	21.1	.0	73.7	100.0

A. 66.3% of original grouped cases correctly classified.
B. 65.1% of cross-validated grouped cases correctly classified

IV. CONCLUSIONs AND RECOMMENDATIONS

A. *Conclusions*

1- The study has comprehensively investigated the oilfield production accidents on the sites.

2- This study has taken a holistic view to describe the scenario of oilfield production accidents in light of the classification of accidents that increased the effectiveness of the models.

3- The DA is conducted to predict whether an accident would be classified in one of the accident groups 'Oil & Gas Leak', 'Fire', 'Accident' and 'Damage'.

4- Significant mean differences were observed for all the predictors on the dependent variable.

5- The independent variables selected are appropriate since their inter correlations are low.

B. *Recommendations*

1- Similar, further studies should be carried out, wherever appropriate.

2- Further research could investigate the effectiveness of accidents with some control variables of the companies (for example, size, industry... *etc.*) in the regression model. The study can also be extended in some other local companies to ascertain and compare conditions at the oil fields relative to accidents.

3- Accidents data in oil fields of developing countries need to collected to create benchmark values for critical variables

REFERENCES

[1] J. I. Chang, and C. C. Lin, “A study of storage tank accidents,” *Journal of Loss Prevention in the Process Industries,* vol. 19, pp 51–59, 2006.

[2] W. Wei, and L. Mao, “Analysis of accident in production in oilfield of China,” Center for Urban Public Safety, Nankai University, Tianjin 300071, China, 2003.

[3] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning - data mining, inference, and prediction,” Springer Series in Statistics, New York, 2001.

[4] B. Efron, and R. Tibshirani, “Cross-validation and the bootstrap: estimating the error rate of a prediction rule,” Technical Report, University of Stanford, 1995.

[5] Multivariate Data Analysis Using SPSS Discriminant Analysis: Lesson2. (2008), Available: http://www.researchgate.net