**Analysis of Effects of various factors on Workday Loss**

**in Coalmine Industry Incidents**



**A PROJECT REPORT**

***Submitted By***

**Shivendra Soni (IIT2010027)**

**Kshitij Mittal (IIT2010040)**

**Shashi Prakash (IIT2010158)**

*Under the guidance of*

**Dr. Sonali Agarwal**

*in partial fulfillment for award of degree*

**BACHELOR OF TECHNOLOGY**

*in*

**INFORMATION TECHNOLOGY**

**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY**

**ALLAHABAD**

**February, 2013**

**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY**

**Certificate**

This is to certify that the dissertation entitled “Analysis of Effects of various factors on Workday Loss in Coalmine Industry Incidents” submitted by Shivendra Soni (IIT2010027) and Kshitij Mittal (IIT2010040) is approved for the award of Degree - Bachelor of Technology in Information Technology.

We hereby declare that it is an authenticated record of our original work carried out from July to November, 2012 under the guidance of Dr. Sonali Agarwal. Due acknowledgements have been made in the text to all other material used.

Place: Project Mentor

IIIT Allahabad **Dr. Sonali Agarwal**

Jhalwa Campus

Date: 29th November, 2012

**Acknowledgements**

We would like to convey our deepest gratitude to Dr. Sonali Agarwal, who guided

us through this project. Her keen inspiring personality, admirable guidance,

and constant encouragement are the motive force behind this project work.

**Shivendra Soni**

**(IIT2010027)**

**Kshitij Mittal (IIT2010040)**

**CONTENTS**

**ABSTRACT**

This work attempts to establish the relevance of risk management in mining industry, provide introduction to various methods, techniques and procedures used in risk management and then application of statistical tools for analysis of the data aquired. It further explains a general algorithm for risk management, and types of risk assessment techniques. The methodology section, explains the implementation of three methods for statistical analysis for risk management – regression, neural networks and fuzzy logic.

The aim of the research is to develop a relationship between incidents in underground coal mining industry and various factors responsible for it, so that appropriate measures are taken to prevent or minimize this.

**AIM AND SCOPE**

The aim of quantitative analysis is to statistically analyze the given data using the three methods, and predict values of various parameters related to risk. Since risk is derived from severity and probability of incidents (or a combination of both), physical quantities representing these two may be considered as the ‘result quantities’ of statistical analysis.

For instance, probability may be represented by number of incidents corresponding to a particular parameter. Similarly, severity may be represented by damage in the form of number of workdays lost, injuries, deaths or disability cases per incident, and the product of severity and frequency may be represented by cumulative results, such as overall number of injuries, deaths or disability cases throughout the history.

These all may be considered as output parameters, while the cause of incident may be linked to reason of incident, or, type of hazard. However, there are several other parameters directly or indirectly affecting the outputs, such as organ affected by incident, working shift (day/evening/night), time of the year or even age of worker. Some inputs like reason of damage and affected organ directly affect the victim, while the effects of age, time of year or work shift may be assumed to affect them indirectly. This entire set of inputs and outputs is shown in Table 1

Table 1. Inputs and Outputs for probable variables to be used for statistical analysis

|  |  |
| --- | --- |
| **Inputs**   * Direct   + Reason (Hazard)   + Affected organ of the worker * Indirect   + Age of worker   + Time of year (such as month or season)   + Work shift (day, evening or night) | **Outputs**   * Probability   + Number of Incidents * Severity   + Workday loss / incident   + Number of injuries / incident * Cumulative   + Total Number of days loss   + Total Financial Damage   + Total number of injuries, deaths or disabilities |

**MOTIVATION AND LITERATURE SURVEY**

**METHODOLOGY**

# 1. Introduction

For this project, a ten years (2002-2012) statistical data on Turkish coal mines is used for quantitative analysis. We deal with quantitative analysis using three techniques for risk analysis – (i) Ordinary Least Squares Multivariate Regression (called as OLS regression, or just ‘regression’ henceforth), (ii) Neural Networks and (iii) Fuzzy Logic. Three different methods are used for risk analysis, because each has its own sets of advantages and limitations. Methodology section explains the three methods, and discusses the algorithm required for their analysis.

# 2. Introduction to Statistical Analysis Methods

This section covers introduction to the three methods of analysis, while the next section (4) covers the relative comparison of the methods.

## 

## 2.1. Regression

Regression using the following algorithm to fit the data: Find the equation of a single line in the given input parameters which fit the given data. (Lind, D. A. , Marchal, W.G. and Wathen, S.A. , 2005). This is the simplest method, since if all values of input parameters and corresponding output values are available in continuous form for the given dataset. The given equation is as follows:

The regression equation can be written, in its simplest form, as follows:

**R\*X + C = Y** where

R = Corresponding regression coefficient for that variable – this is the quantity to be found from regression, which produces the best fit for the line.

X = Value corresponding to that point (or 0/1 for dummy variable for the corresponding state)

C = Value of constant term

Y = output values for regression

Since our given dataset has more than one input parameters, the regression form to be used will be ‘multivariate regression’. Accordingly, the quantity X will be a 2-dimensional matrix, with columns representing input quantities, and R will be a vertical matrix (now replaced by β. Equation can now be modified as follows:


 \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad
 \mathbf{X} = \begin{pmatrix} \mathbf{x}^{\rm T}_1 \\ \mathbf{x}^{\rm T}_2 \\ \vdots \\ \mathbf{x}^{\rm T}_n \end{pmatrix}
 = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\
 x_{21} & \cdots & x_{2p} \\
 \vdots & \ddots & \vdots \\
 x_{n1} & \cdots & x_{np}
 \end{pmatrix}, \quad
 \boldsymbol\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad
 \boldsymbol\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.
  (Eq.1)

One of the best and most commonly used estimate method, suggested by Lind et.al. (2005) is least squares method. In this method, the line parameters are computed such that the sum of squares of distances between data-points and the line is minimized. In other words,

(δ/δx)∑{ ∑(Xj - Xji)2 + (Y-Yi)2} = 0

(δ/δx)∑{ ∑(Xj - Xji)2 + (Y-Yi)2} = 0

where i,j represent points and input variables respectively, and other symbols have usual meanings.

On solving, it comes out to the following:


  \hat{\boldsymbol\beta} = (\mathbf{X}^{\rm T}\mathbf{X})^{-1} \mathbf{X}^{\rm T}\mathbf{y}
 = \big(\, \tfrac{1}{n}{\textstyle\sum} \mathbf{x}_i \mathbf{x}^{\rm T}_i \,\big)^{-1}
 \big(\, \tfrac{1}{n}{\textstyle\sum} \mathbf{x}_i y_i \,\big).
  

This result can easily be found using some engineering or data analysis tools. For example, if matlab is used for programming, the function ‘regress’ accepts matrix forms (as given in equation 1) of input and output variables, to produce various statistical quantities, such as regression equation, correlation coefficients, errors and so on. However, as already stated, it is difficult to analyze the set of parameters in Table 7 because the set contains both continuous and discrete variables. Age is a continuous variable, while the rest others are discrete. Numbers related to such ‘discrete’ values are ordinal, and do not represent any relationship. For example, it will be wrong to state that shift two is twice of shift one, and so on. So, it is necessary that such quantities are converted to some other format which makes them look like continuous.

One such format was suggested by Hardy (1993), who explained the method of using dummy variables for regression analysis. As per the method, if variables are represented by a set of binary quantities, they can be made to behave like continuous variables. In the previous example, the behavior of ‘shift’ can be made continuous, if three shifts are represented as given in Table 8, by variables V1, V2 and V3, such that V1 represents (S1,0,0) and so on. Thus, continuous variables V1, V2 and V3 will take the required shift coefficients for states S1, S2 and S3 respectively. For instance, coefficients of (V1, V2, V3) = (1, 3, 7) represent the values 1 if state = S1, 0, 0 (or simply state = S1), 3 if state = S2, 7 if state = S3. Thus, each parameter generates the number of variables equal to the number of discrete values it may take. For our project, each discrete variable can take only one state, but this method can also be used for those situations where multiple discrete outcomes are expected. For instance, V1 = 1 and V3 = 1 can potentially take care of some observation with mixed shift S1 +S3 corresponding to an incident.

Table 8 - Representation of shift through three variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Shift** | **V 1** | **V 2** | **V3** |
| S1 | 1 | 0 | 0 |
| S2 | 0 | 1 | 0 |
| S3 | 0 | 0 | 1 |

There are inbuilt functions to convert the variables to the format described in Table 8. For instance, function ‘dummyvar’ in matlab converts a discrete variable carrying values 1-n to dummy variables. However, unless states are already provided in natural numbers format (1, 2, 3), the mapping of states with natural numbers need to be done wither through implementation of logic through a code or manually.

As will be established through comparison later, out of the three methods, regression is easiest to design, implement and get processed by machine, but it is not used too often because it is found that some other methods produce better results.

## The other two methods under consideration namely, Neural Network and Fuzzy Logic are described in brief below and would be covered in details in post-mid semester report.

## 3.2 Neural Networks

A Neural Method works in a way similar to regression, but uses past available values, or ‘learning system’ for analysis, instead of one-time method like least squares. It initially creates a hidden layer of neurons, pre-decided and unalterable parameters.

The layers input and output contains the information provided in training set, whereas the hidden layer contains information on weights of regression coefficient, which get modified after each cycle.

Once weights are obtained through training, given system uses those weight values to get actual results from its database. For regression analysis type problem, a part of the given observation set itself acts as a training set and the remaining as a testing set.

## 2.3 Fuzzy Logic

The third method is fuzzy logic. The method is frequently used to model the data when either the results are qualitative, or the number of variables is huge. The advantage of this method lies in the fact that it exploits the power of intuitions and common sense to come up with the expected results. On the other hand, it requires human intelligence and creativity for writing rules well-enough or else, the method may fail. This method is explained in the coming paragraphs.

The basic idea behind the method is that related to flexible and uncertain reasoning, that every apparently binary system may take any values between 0 and 1. For instance, a two-value logic analysis might allocate the values 0 or 1, to say, some fluid when it is hot or cold respectively. In contrast the same liquid might take either of the following values under fuzzy logic - 0.5 for moderate, 0.25 for hot, 0.75 for cold, 0.9 for freezing, and 1 for minimum possible temperature. The user might define certain actions depending on the state. (Ross, 2009) This example is elaborated in Table 3

Table 3. An example of fuzzy logic implementation for variable ‘Temperature’

|  |  |  |  |
| --- | --- | --- | --- |
| **Binary Value** | **Definition** | **Condition** | **Action** |
| 0.25 | T ~ 400 | Hot | Start the Air Cooler |
| 0.5 | T ~ 298 | Moderate | Do Nothing |
| 0.75 | T ~ 210 | Cold | Start the Heater |
| 0.9 | T ~ 70 | Freezing | Set the Heater Temperature to Maximum Value |
| 1 | T ~ 0 | Absolute Freeze | Condition Impossible – Examine the measuring devices |

From Table 3, it can be seen that some actions have been decided based on temperature values. Such a perspective enables decision making depending on whether the given value is closest to the given condition. For instance, if binary value corresponding to the temperature approaches 0.89, it might mean that the heater’s maximum capacity is about to reach, and it needs to be shut down.

**Progress Till Date**

So far we have been able to analyze the given data using OLS (Ordinary Least Square) Regression model. The Methodology applied, results obtained and Analysis of the results have been tabulated below.

**Quantitative Analysis using Regression**

The first model to be used for quantitative analysis is ordinary least squares (OLS) regression. This method deals with simply fitting a line which ‘best relates’ the given set of variables (continuous + dummy) linearly.

The regression equation can be written as follows:

**∑Cvi\*Vi  + C0 = Y** where

i – element index (varies from the first to the last element of the column of that variable)

C – Corresponding Correlation Coefficient for that variable

V = Quantity corresponding to the variable (or 0/1 for dummy variable for the corresponding state)

C0 = Value of constant term

Y = Result of regression

The method involves the following assumptions:

1. The dependent variables have no mutual dependence against one another, and errors are random.

2. The two sites provide similar conditions for experimentation. This means the difference in coefficients due to the two sites would be random.

The regression equation was obtained through matlab program, which was done by first reading the matrix from file, converted the required discrete variables to dummy, and then using the ‘regress’ function to get the equation. Constant term was obtained by appending a unit column to the final data.

Regression equation for site 1:

Y = -0.0802\*A *Age*

- 2.637\*S2 – 2.935\*S3 *Season*

+ 1.103\*T1+0.145\*T2 *Time*

+ 5.671\*R3 + 9.628\*R4 + 0.3872\*R5 + 8.407 \*R(7) + 12.18\*R(8) + 11.64 \* R9 + 3.091 \* R11 + 7.664\*R12 + 5.553\*R13 + 8.141\*R14 *Reason*

- 20.95\*O1 - 14.08 \* O2 - 13.45 \*O3 - 15.54\*O4 -16.01\*O5 - 18.21\*O6+30.36

*Organ (Affected)*

Similarly, regression equation for site 2:

Y = 0.1050\*A

- 3.655\*S1 - 4.960\*S2 - 4.064 \* S3

-1.655\*T1 + 2.228 \* T2

+ 8.156\*R3 + 10.978 \* R4 + 11.459\*R5 + 6.227\*R7 + 10.166\*R8 + 8.419\*R9 + 10.633\*R11 + 10.408\*R12 + 2.75\*R13 + 8.158\*R14

- 0.6671\*O1 + 8.152\*O2 + 12.248\*O3 + 4.782\*O4 + 5.355\*O5 + 1.636\*O6 + 16.481 \*O7 where,

A = Age

Ti = ith shift of time

Ri = ith reason

Oi = ith organ affected

Y = workday loss per incident

Variables states with no incidents associated with them, or those having negligibly small coefficients, are both marked as zero. So, it cannot be assumed that absence from the equation is due to zero coefficient.

The value of regression (R^2) was found to be 0.031 for site 1 and 0.054 for site 2, which shows that the obtained equations are not strongly correlated against the given dataset. This is expected, owing to fact that the number of variables in the equation is large. However, if compared to number of values in dataset, it can be said that on an average, there are more than 30 values for each variable. This proves that the data accuracy can be enhanced using some other methods, as discussed in neural network and fuzzy logic sections.

In spite of its poor accuracy, a few inferences may be drawn from the equation generated by the method. The analysis of the equation is performed as follows:

1. For continuous variable (Age), the effect can be determined by the sign of variable – negative for decreasing and positive for increasing effect. It was found that the coefficient for age had different signs for different sites. Since this difference could not be due to the site effect, it can be inferred that age does not have a significant effect on workday loss per incident.

2. For seasons, it can be said that the coefficient for season 1 (spring) is certainly higher than the other two seasons for both sites. It has zero coefficient for site 1, because it shows 0 value, and has a reasonably large number of observations. For site 2 also, its value is higher than others. This means that workday loss is higher (per incidence) in spring season than other seasons.

3. The results for ‘reason’, were not strong enough, but some reasons like Manual Handling (7) were found to have high values for both the tables, and can therefore, be claimed to be enhancing workday loss per incidence. Other results for reason are not too commendable.

4. ‘Affected Organ’ has provided some good results. It is clearly evident that workday loss gets minimized for head injuries (1) since this column has lowest values for both the sites. Similarly, foot injury (3) has highest & arm injury (4) has the second highest values for both the sites, which proves that workday loss is higher for foot and arm injury, as compared to other injuries.

5. The results for ‘time’ are either not obtained properly through this method. However, it is difficult to claim that shift does not affect workday loss per incidence, because one reason of not getting good results could be attributed to smaller number of discrete states possible for shift.

The results ( 1 – 5 ) can be summarized as shown in Table 4.

Table 4. Effects of parameters on workday loss – results from regression

|  |  |  |
| --- | --- | --- |
| Sr. No. | Parameter | Effect on workday loss |
| 1 | Age | - |
| 2 | Shift | - |
| 3 | Season | (1)Spring (+) |
| 4 | Reason | (7) Manual handling (+) |
| 5 | Affected Organ | (1) Head (--), (3) Foot (++), (4) Arm (+) |

Another type of result which could be obtained from this method is the ‘intensity of replacement within variables’. For example, the coefficient values corresponding to seasons (2, 3) for site 1 were found to be (-2.63,-2.93) respectively. Replacing one by another modifies the final result (workday loss per incident) by (+/-) 0.3. This is because dummy variables take the value 1 or 0 for availability or unavailability, which means coefficients simply add up for all such existing states. Such a difference for Site 2 was found to be 0.9. Differences in corresponding values for shifts (1, 2) were found to be 0.95 and 3.8 respectively. So, it can be said that the effect of shift replacement is stronger than the effect of time replacement. This logic can be quantified by taking the median values of the obtained regression coefficients, and then comparing them against highest and lowest coefficient values for the same parameter. By doing this, it could be established that reason is the most significant parameter as per the current method, as it has the highest replacement effect (Max-median = 4.3, median – min = 7.5). This was further verified by neural networks analysis as well, as will be discussed later. For fuzzy logic, it was found that affected organ factor was slightly stronger but still, it can be said without any doubt that reason is a parameter which strongly affects severity of incidents.

**Final Deliverable**

Upon the completion of the project we will be able to :

* Three successfully prepared models/software codes for the given data – regression, neural networks and fuzzy logic.
* Identify those factors, which enhance or reduce the severity of such incidents in the considered scenario.
* Do analysis of the effect of changing various parameters, strong and weak parameters were identified.
* Compare the three methods, and suggest the best one for the considered dataset.

**CONCLUSION**

By the analysis of the given data and its comparative study on three different statistical models we were able to analyze the various reasons responsible for the accidents in coalmines.

This could prove as a valuable insight to the Risk/Hazard assessment and Management techniques. Since severity and probability completely define a risk, the goal of analysis is to estimate the values of these two quantities associated with risks, which we successfully accomplished.

And finally this knowledge could be applied to Risk Treatment. Once risks are identified, analyzed and evaluated, they are finally required to be treated, meaning some suitable remedial action is suggested for them. This might be a lifesaving decision in the real world scenario.

That being said, the most important goal of the project was to utilize a given data by the use of technology and convert it into information, information that could really be of utility in real world scenarios.

**REFERENCES**

1. Sari, M. (2002). *Risk Assessment Approach on Underground Coal Mine Safety Analysis*. Ankara: Middle East Technical University. Department of Mining Engineering
2. Lind, D. A., Marchal, W.G. and Wathen S.A. (2005). *Statistical techniques in business & economics*. New Delhi: Tata McGraw Hill Education Private Limited
3. Hardy, M. A. (1993) .*Regression with Dummy Variables, Issue 93*. London: SAGE
4. Gurney, K. and Gurney, N. K. (1997). *Neural Network.* London: CRC Press
5. Ross, T. J. (2009). *Fuzzy Logic with Engineering Applications*. Chichester: John Wiley & Sons
6. Bagot, A. (2009). *Accidents in Mines: Their Causes and Prevention.* Retrieved from books.google.com
7. McDonald L.B., Pomroy W.H. and United States Bureau of Mines (1980). *A statistical analysis of coal mine fire incidents in the United States from 1950 to 1977*.US Department of the Interior, Bureau of Mines,8830,1-42