UDK 614.8:622.261.2:519.1
DOI: 10.7562/SE2017.7.01.01
Original article
www.safety.ni.ac.rs

BULENT TUTMEZ [1]
MERT G. OZDOGAN[2]

[1]Inonu University
School of Engineering, Turkey

[1]bulent.tutmez@inonu.edu.tr
[2]mertozdogan90@hotmail.com

ASSESSMENT OF ACCIDENTS INVOLVING FATALITIES AT COLLIERIES USING COUNT REGRESSION

Abstract: *Although some studies were conducted on various parameters that affect safety and health of mine workers, there has been scarce research on specific accidents involving fatalities. The most important hard coal reserves of Turkey are located only around Zonguldak province in the northwest of Anatolia. The number of fatalities that occurred in the coal-mine fields in the Zonguldak Basin from 2000 to 2014 is appraised by count regression analysis. To describe the relationships between the locations and the cases, a Poisson Regression model with count data has been developed. In this way, general attention has been drawn to this risky work area. Based on the statistical analyses and tests, the proposed model provided considerable explanation. The model and the risk scenarios showed that the results can be used by decision makers to provide a comprehensive sustainability assessment.*

Key words: Coal mining, fatality, count data, Poisson regression, loss prevention.

INTRODUCTION

As clearly discussed in the recent work [1], workers in developing countries believe their organizations will take all the necessary measures to ensure that they return home safely at the end of the work day. Yet, work-related injuries and fatalities continue to occur at an alarming rate. In particular, mining work-related injuries and diseases exhibit a serious and costly burden to some countries and a major challenge to governments and especially workers themselves. Despite increased safety measures and decreased fatality toll, coal mining has received a general attention as a dangerous industry, partly from recent highly publicized disasters [2].

Global consumption of coal is growing and is projected to increase even more as developing countries expand their energy needs [3-4]. The most important hard coal reserves of Turkey are placed around Zonguldak province in the northwest of Anatolia. The Turkish Hard Coal Enterprise (TTK) controls 100% of this bituminous coal production [5]. Bituminous coal can be used in power and heat manufacturing as a coking coal, mainly for aluminum and steel production.

The Zonguldak basin has hard underground mining conditions. Recently [6] has reported that the main causes of accidents are non-regulated working conditions in the mines. In addition to focusing on social bound of work accidents and their reasons, analyzing and modeling recorded data would have critical contributions to make provision against the future accidents. From a general structure, the collected data include both physical injuries and fatalities. In practice, the number of injuries is very high and from different sources. Luckily, the number of fatality cases is limited and the large part of their sources can be assessed and controlled.

Modelling the relationship between the fatalities and their potential sources can provide useful information to consider the problem and take some precautions. In the present study, the number of fatalities and their locations recorded in Zonguldak Basin from 2000 to 2014 are evaluated by count regression analysis. Poisson Regression (PR) is suitable when we are estimating an outcome variable representing counts from a set of categorical and/or continuous variables. The number of occurrences of behaviour in a fixed period of time can be depicted by count data [7]. In cases in which the target variable is a count with low arithmetic mean, conventional regression algorithms may result in biased outputs. In these cases, instead of the traditional methods such as Ordinary Least Squares (OLS) Regression, using Poisson Regression and Log-linear models are preferred [8]. In recent years, many works of Poisson Regression Analysis in occupational safety have been published. [9] conducted a research which addresses Poisson regression based on identification for mapping cancer mortality. The incidence of lung cancer around asbestos mine in Finland was investigated [10]. In a novel paper on occupational safety, [11] discussed occupational exposure to asbestos and mortality among asbestos removal workers via Poisson regression. Similarly, the GWR (Geographically Weighted Regression) and the PR models for country-level crash modelling have been used. An evaluation on accidents using a binominal panel count data model has also been made [13].

In a recent paper, the count regression analysis has directly been applied for transportation safety planning [14]. Although some interesting papers play a part in occupational safety literature [15], there is no distribution-based evaluation on occupational safety and health in coal mining. Starting from this, the accident involving fatalities that occurred in the underground coal mines have been handled and estimated by count regression analysis. Because the number of fatalities is a count data, instead of the conventional regression algorithms, we used relatively new and explanatory method - known as Poisson Regression Analysis - to describe the relationships of the mine sites (locations) and accident involving fatalities. The aim of this study is to provide findings for a comprehensive sustainability assessment and take some precautions for making decisions.

The rest of the paper is structured as follows: Section 2 describes the problem and methodology, Section 3 presents the application and detailed discussion, whereas Section 4 concludes the paper.

MATERIALS AND METHODS

2.1. Field and problem statement

Coal is one of the world's largest sources of electricity. 40% of global electricity production is provided from coal sources. It has been the world second largest primary energy source, and is projected to replace oil as the world's largest source within a few years [16].

Zonguldak Coal Basin, placed in the Black Sea coast, is the only bituminous coal area of Turkey. It is predicted that the coal reserves in the Zonguldak coal basin are around 1.1 billion tons and their areal extent is 13.000 km^2. The proven reserves include the seams which are present to a depth of 1200 m, with the thickness varying between 1 and 10 m. [17]. The study area is illustrated in Figure 1. Zonguldak Basin consists of five collieries and General Directorate unit: Armutcuk (AR), Amasra (AM), Uzulmez (UZ), Karadon (KA) and Kozlu (KO). In addition, some workers are working in General Directorate (GR) units. The Turkish Hard Coal Enterprise manages the collieries that produced approximately 1.6 million tons of saleable coal as of 2011 [18].

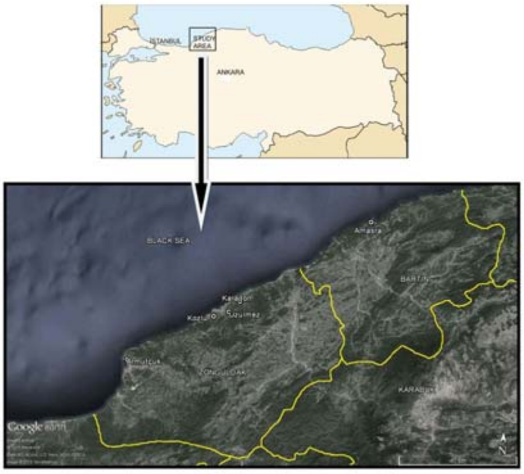


Figure 1. *Study area and locations of collieries*

There are still serious problems as consequences of natural difficulties of the basin as well as the practices. All the collieries in Zonguldak basin have rough underground mining conditions. As in Table 1, due to hard working conditions and ill-advised practices, in five collieries (excluding GD) 2147 injuries were recorded in the period 2000-2014 [17]. A range of 20-33% for injury rate per number of workers was calculated from the public records. This range addresses a general consistency for all sites. In addition to these injuries, 74 fatalities were reported by the authorities.

Table 1. *Production sites and some quality parameters*

Colliery	Product (Metric tons)	Number of Workers	Injuries	Injury Rate per Number of Workers
Armutcuk	214479	1136	242	0.213
Amasra	222349	667	136	0.203
Uzulmez	482050	1664	546	0.328
Karadon	680215	3222	917	0.285
Kozlu	484550	1832	306	0.167

2.2 Regression analysis of count data

In a count data set, the measurements can take only the non-negative integer values {0, 1, 2, ...}. One of the most practicable methods for modelling count data is Poisson regression method. As discussed in [8] using count variables in OLS regression may potentially pose problems. For example, when the mean of output is low (e.g. computed as smaller than 10), the conventional OLS algorithm exhibits undesirable outcomes including biased standard errors. Similarly, [19] expressed that if the variation of the explanatory is very small, it may easily appear in count variables with small range; the regression coefficient for that explanatory variable will be changeable and will have a great standard error.

Poisson regression (PR) analysis is a special type of regression analysis utilized to analyse count data. The conventional Poisson approach presumes the response variable Y has a Poisson distribution, and supposes the logarithm of its expected value can be identified via a linear combination of unknown parameters. If Y is the number of occurrences, its probability distribution can be expressed as

$$f(y) = \frac{\mu^y e^{-\mu}}{y!},\ y = 0,1,2,\ ... \qquad (1)$$

where μ is the average number of occurrences. It can be presented that $E(Y) = \mu$ and $var(Y) = \mu$. The influence of predictor variables on the target Y is modelled through the parameter μ. The average number of occurrences μ determines both the mean and variance of the distribution; both the mean and variance equal μ [7].

Statistically, Poisson regression model has a generalized linear structure with Poisson distribution error form and the natural *log (ln)* link function. The general structure of the model can be given as [8]

$$\ln(\hat{\mu}) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p, \qquad (2)$$

where $\hat{\mu}$ is the predicted count on the target variable presented the definite values on the explanatory variables $X_1, X_2, \dots, X_p$. When an equation has count outcomes, use of eq. (2) instead of standard OLS structure provides some inputs such as non-constant variance of the errors and non-normal conditional distribution errors.

The generalized linear model for the Poisson model above can be expressed as follows:

$$E(Y_i) = \mu_i = n_i e^{x_i^T \beta}. \qquad (3)$$

where Y_i denotes the number of events observed from exposure n_i. The natural link function given in eq. (2) can be presented as the logarithmic function

$$\log \mu_i = \log n_i + x_i^T \beta. \qquad (4)$$

In a regression analysis of count data, residuals can be utilized to identify model misspecification; to identify outliers, or observations with poor fit; and to identify influential measurement, or measurements with a big impact on the fitted model. To analyse the errors, the standard error of Y_i is predicted by $\sqrt{e_i}$. Thus, the Pearson residuals can be computed as follows [20]

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}} \qquad (5)$$

Where, o_i represents the measured value of Y_i. Starting from this, the chi-squared goodness of fit statistic and deviance residuals can be given as follows:

$$X^2 = \sum r_i^2 = \sum \frac{(o_i - e_i)^2}{e_i} \qquad (6)$$

$$d_i = \mathrm{sign}(o_i - e_i)\sqrt{2\left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i)\right]} \qquad (7)$$

The deviance residuals produced by eq. (7) are the elements of deviance D, $D = \sum d_i^2$. As presented in [7], X^2 and D can be utilized directly as measures of goodness of fit, as they can be computed from the data and the fitted model.

RESULTS AND DISCUSSION

3.1. Data and descriptive statistics

The data considered in this study concern the accidents involving fatalities at collieries in Zonguldak Basin, Turkey [18]. Table 2 summarizes the data set which includes the fatalities resourced from the mine-related works. The frequency graph for the total fatality is illustrated in Figure 2. An observable variation is recorded in Figure 2. In Figure 3, the number of fatalities and corresponding collieries are presented.

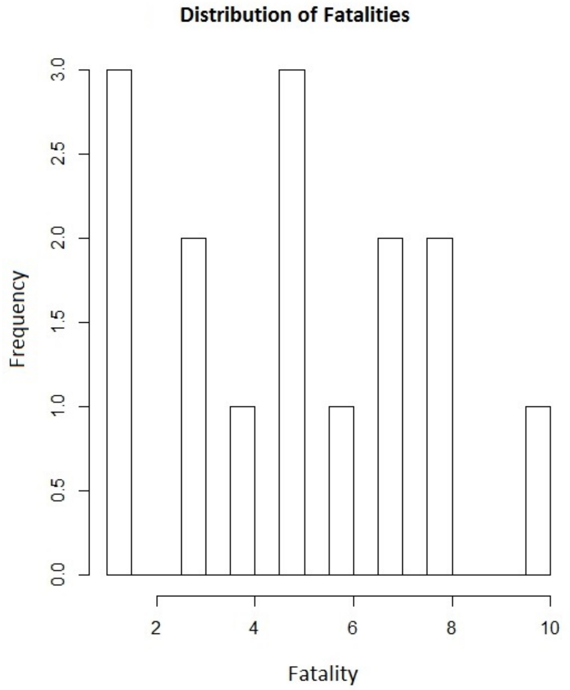


Figure 2. *Frequency distribution of Zonguldak Basin for total fatalities*

Table 2. *Recorded fatalities*

Year	General Directorate	Armutcuk	Amasra	Uzulmez	Karadon	Kozlu	Total
	Accidents involving fatalities						
2000	0	0	0	1	0	0	1
2001	0	0	0	1	0	0	1
2002	0	0	1	4	3	0	8
2003	1	0	0	2	3	2	8
2004	0	0	2	0	2	1	5
2005	0	1	0	0	6	3	10
2006	0	0	0	2	1	0	3
2007	0	0	0	1	2	2	5
2008	0	1	0	3	2	1	7
2009	1	0	0	0	4	2	7
2010	0	0	0	2	2	1	5
2011	0	0	0	3	1	0	4
2012	0	1	1	1	1	2	6
2013	0	0	1	1	1	0	3
2014	0	1	0	0	0	0	1

3.2. Model development

As seen in Figure 2, target variable (number of fatalities) has a skewed nature. However, there is no finding to show the possible presence of outliers. The fitting parameters of the Poisson regression model of Zonguldak Basin data set are summarized in Table 3. It can be recorded from the coefficients; Kozlu (KO) is one of the effective determinants.

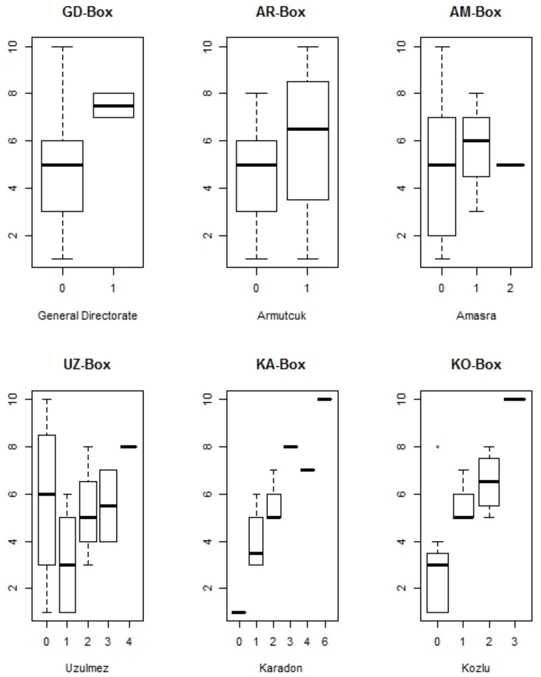


Figure 3. *Box plots for input data.*

Table 3. *Fitting statistics*

Model Parameter	Estimate	Std. Error	Z value	Pr(>[z])
Intercept	0.217	0.392	0.554	0.579
GD	0.158	0.405	0.392	0.695
AR	0.029	0.368	0.080	0.936
AM	0.323	0.215	1.502	0.133
UZ	0.289	0.118	2.453	0.014
KA	0.158	0.100	1.579	0.114
KO	0.389	0.247	1.574	0.116

The estimated values given in Table 3 are the outcomes of the GLM model described in (3). In a Poisson regression analysis, the target variable being identified is the log of the conditional mean $log_e(\mu)$. For example, the regression parameter 0.158 for General Directorate (GD) shows that one new fatality in GD site is associated with a 0.16 increase in the log mean number of total fatalities. It's much easier to translate the regression coefficients in the original scale of the target variable. In this case, one new fatality increase in GD site multiplies the total number of fatality by 4.86, holding the other variables constant.

Table 4 summarizes the observed and predicted numbers of fatalities in Zonguldak Basin and the residuals for the Poisson model implemented. As seen in Table 4, X^2 and D were calculated as 1.361 and 1.435, respectively. From a statistical view, deviance is a measure of goodness of fit. Therefore, the residual deviance describes the difference between the deviance of the current model and the maximum deviance of the ideal model. In addition, the residual deviance is about normally distributed if the model is defined correctly [21]. In this application, because median is not quite zero, a little bit of skewness (0.025) is recorded.

Table 4. *Actual and predicted fatalities with residuals*

Years	Observed fatalities	Predicted fatalities	Pearson residual	Deviance residual
2000	1	1.659	-0.512	-0.553
2001	1	1.659	-0.512	-0.553
2002	8	8.741	-0.251	-0.254
2003	8	9.054	-0.350	-0.357
2004	5	4.786	0.098	0.097
2005	10	10.562	-0.173	-0.174
2006	3	2.593	0.253	0.247
2007	5	4.944	0.025	0.025
2008	7	6.153	0.341	0.334
2009	7	5.946	0.432	0.420
2010	5	4.476	0.248	0.243
2011	4	3.461	0.290	0.283
2012	6	6.005	-0.002	-0.002
2013	3	2.681	0.195	0.191
2014	1	1.280	-0.247	-0.257
Sum of squares			1.361	1.435

Statistically, a large divergence shows unsatisfactory fit relative to a perfect model. Monitoring the residuals is one of the best ways to investigate the large deviance [22]. Existence of an outlier can display this problem. The half-normal plot of the residuals given in Figure 4 indicates no outlier.

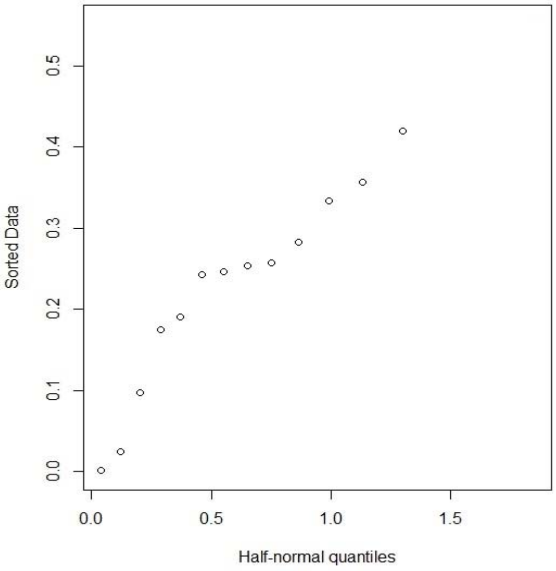


Figure 4. *Half-normal plot of residuals*

Model performance is also illustrated in Figure 5 and it can be evaluated by various indicators. From OLS perspective, Coefficient of Determination (r^2) addresses a measure of the proportion of variation in the results. It exhibits the extent to which the model accounts for the measured data. For this model r^2 was computed as 0.95. However, as stressed in [23], in count regression the total deviation in the result cannot be exactly partitioned into explained and unexplained partitions.

Deviance is decreased by adding exploratory variables to the intercept-only model if the exploratory variables have some precision in accounting for the result [8]. Therefore, an alternative index, the variation of the model has been adapted to compute a pseudo-R^2 measure as follows:

$$R^2_{deviance} = 1 - \frac{deviance(fitted\ model)}{deviance(intercept\ only)} \tag{8}$$

From eq. (8), $R^2_{deviance} = \left(\frac{4.99}{23.493}\right) = 0.707$. Thus, six explanatory variables decreased the deviance by 78.7% compared to using no predictors.

[20] suggested practising on standard errors for the parameter estimates to control mild violation of the distribution assumption that the mean equals the variance. Determining the p-values and 95% confidence interval, the parameter estimates and their robust standard errors were utilized. Table 5 summarizes Robust Standard Errors (RSE) and the confidence intervals for the estimated coefficients. Besides GD, the algorithm presents relatively narrow intervals. The wide range given for GM may be resourced from the limited number of occurrence (fatality).

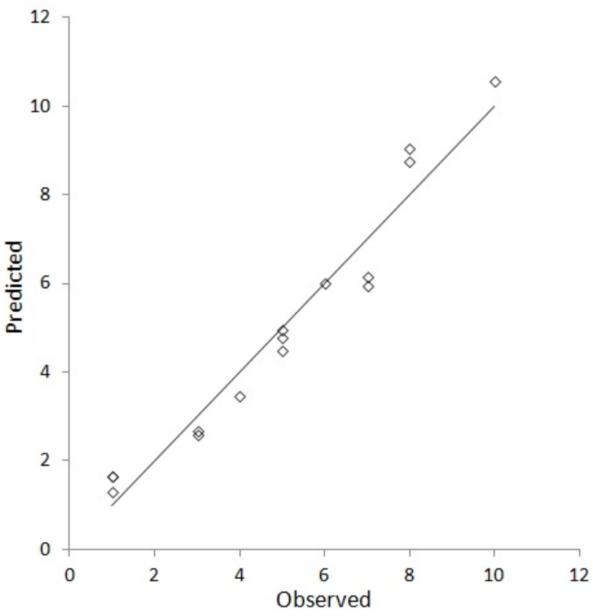


Figure 5. *Model performance*

3.3. Over-dispersion analysis

In some situations, the individual counts may demonstrate more variation than it is predicted from the count (Poisson) model. Therefore, one of the main risks of such analysis is high variability. In a standard Poisson distribution, the variance and mean should be equal. However, over-dispersion exists in Poisson regression when the measured variance of the target variable is larger than it would be estimated by the Poisson distribution.

Table 5. *Robust standard errors and intervals for coefficients*

Model Parameters	Estimate	Lower	Upper	Robust SE
Intercept	0.217	0.059	0.493	0.141
GD	0.158	-0.069	0.386	0.116
AR	0.029	-0.129	0.187	0.081
AM	0.323	0.228	0.416	0.048
UZ	0.289	0.216	0.362	0.037
KA	0.158	0.119	0.196	0.020
KO	0.389	0.273	0.504	0.059

If a decision maker does not account for over-dispersion in a model, too small standard errors and confidence intervals could be received. In the present application, the ratio of the residual deviance to the residual degrees of freedom has been found as follows:

$$\frac{Residual\ Deviance}{Residual\ df} = \frac{1.4345}{8} = 0.179$$

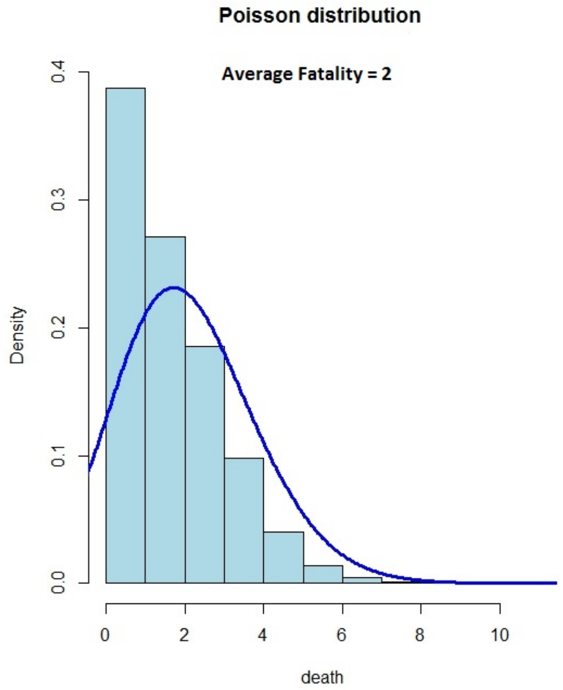
The proportion value is clearly smaller than 1. Thus, an over-dispersion has not been recorded. To control the analysis, a significance test for over-dispersion in the Poisson case has been also performed. As a result of this test, *p*-value has been provided as 0.077. The test which has a *p*-value more than 0.05 does not addresses the presence of over-dispersion.

To check the amount of the variance, another way of dealing with over-dispersion -Quasi-Poisson model - has been developed. The model follows the predicting function view of the Poisson model and does not correspond to model with specified likelihood. The parameter estimates in the Quasi-Poisson model are equal to those provided by the Poisson model. Although smaller standard errors have been obtained, the outcomes showed no over-dispersion appeared in the model.

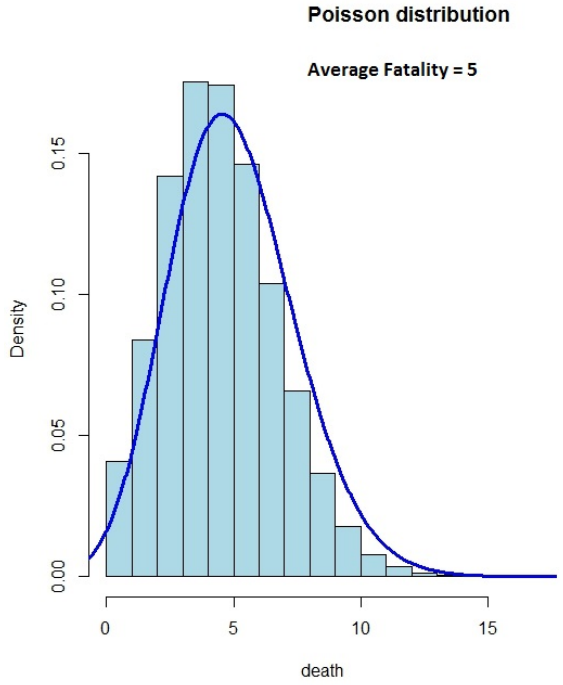
3.4. Risk scenarios

Typical risk assessments address the assessment of probabilistic measures that estimate the average expected value for the situation being considered across a range of potential outcomes [23]. In this problem, the observed and estimated numbers of fatality values have been computed as close values which are 5 and 4.93, respectively. Based on the average value 5 and the standard deviation 2.8, some simulations have been carried out to describe the probabilistic structure and variability. Considering a simulated random set (10000 observations), probability density functions have been obtained for the optimistic $\mu - \sigma$, known μ and pessimistic $\mu + \sigma$ occurrences (fatality). These scenarios are illustrated in Figure 6.

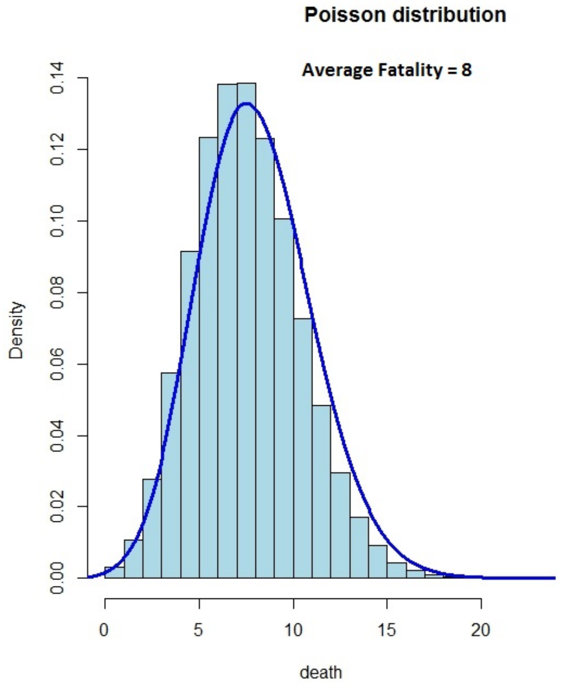
For a non-skewed Poisson distribution, mean and variance values should be the same. As it can be seen in Figure 6, the distribution progressively favours normal distribution as the projected occurrence becomes larger. Because the Poisson regression model provides good approximation for small success probabilities and large totals, new estimations and scenarios based on new accident records can be used to identify the mine sites. On the other hand, one of the underlying assumptions in the use of the Poisson arrival model is that the underlying process (causes) of accidents does not change in time. The mean and the variance, as determined by the data, are different which indicates that some change or difference in the underlying causes of accidents may have occurred.



(a)



(b)



(c)

Figure 6. *Poisson distributions with different values of expected fatalities*

CONCLUSIONS

The most important bituminous coal reserve in Turkey, the Zonguldak basin, is characterized by hard coal underground mining conditions and insufficient operational practices. In recent years, many work-related injuries and accidents involving fatalities have been recorded in the sites. In the present study, the cases recorded in the mine sites and the General Directorate units in Zonguldak have been modeled by a count regression.

The Poisson model structure indicated that Kozlu (KO) and Amasra (AM) sites have dominant effects on the model. As a result of the tests, no over-dispersion effect has been recorded. The numerical results showed that the developed model has big explanation and fitting capacity. The use of a Poisson model to represent the expected occurrences of accidents in time is also appropriate. As recorded by the simulations, there is a close connection between the number of fatalities and symmetry level of the occurrences. Finally, different simulations have been conducted and three risk scenarios have been presented for decision makers. The outputs of this study can be supervised by the decision makers to provide a comprehensive sustainability assessment and control the risks encountered in the hard coal sites.

REFERENCES

[1]Amponsah-Tawiah K, Ntow MAO, Mensah J. “Occupational health and safety management and turnover intention in the Ghanaian mining sector”, Safety and Health at Work 2016; in press, http://dx.doi.org/10.1016/j.shaw.2015.08.002

[2] Margolis KA. “Underground coal mining injury: A look at how age and experience relate to days lost from work following an injury”, Safety Science 2010; 48(4): 417-421.

[3] Lorenz U, Grudzinski Z. “Hard coal for energetic purpose: price-quality relationship; international coal market observations and Polish practice”, Applied Energy 2003; 74: 271–279.

[4] Zaman EM. Zonguldak Kömür Havzasının İki Yüzyılı (Bicentennial of Zonguldak Coal Basin), Ankara: TMMOB Publication (in Turkish); 2004.

[5] WEC Strategic Publication. World Energy Issues Monitor, London: World Energy Council; 2015.

[6] Sarikaya İ. Work accidents as the most afflictive face of neoliberalism: the Zonguldak hard coal basin case, 2nd International Conference on Public Policy, Milan; 2015.

[7] Dobson AJ, Barnett AG. “*An introduction to generalized linear models*”, Boca Raton: CRC Press; 2008.

[8] Coxe S, West SG, Aiken LS. “The analysis of count data: a gentle introduction to Poisson Regression and its alternatives”, Journal of Personality Assessment 2009; 91(2): 121-136.

[9] Vacchino MN. “Poisson regression in mapping cancer mortality”, Environmental Research 1999; 81(1): 1-17.

[10] Kokki E, Penttinen A. “Poisson regression with change-point prior in the modeling of disease risk around a point source”, Biometrical Journal *2003;* 45(6): 689-703.

[11] Frost G, Harding A-H, Darnton A. “Occupational exposure to asbestos and mortality among asbestos removal workers: a Poisson regression analysis”, British Journal of Cancer 2008; 99(5): 822-829.

[12] Li Z, Wang W, Liu P, Bigham JM, Ragland DR. “Using geographically weighted Poisson regression for county-level crash modeling in California”, Safety Science 2013; 58: 89-97.

[13] Coruh E, Bilgic A, Tortu A. “Accident analysis with aggregated data: the random parameters negative binominal panel count data model”, Analytic Methods in Accident Research 2015; 7: 37-49.

[14] Imprialou M-I M, Quddus M, Pitfield DE. “Predicting the safety impact of a speed limit increase using condition-based multivariate Poisson lognormal regression”, Transportation and Technology 2016; 39(1): 3-23.

[15] Paul PS. “Investigation of the role of personal factors on work injury in underground mines using structural equation modeling”, Int. J. Mining Science and Technology 2013; 23: 815-819.

[16] WEC Coal Report. Coal World Energy Resources, London: World Energy Council; 2013.

[17] Karacan CO, Okandan E. “Fracture/cleat analysis of coals from Zonguldak Basin (northwestern Turkey) relative to the potential of coal bed methane production”, International Journal of Coal Geology 2000; 44(2): 109–125.

[18] TTK (the Turkish Hard Coal Enterprise). “*Annual Statistics*”. Strategy Development Division, Ankara: TTK Publication; 2014.

[19] Cohen J, Cohen P, West SG, Aiken LS. “*Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.)”, NJ: Lawrence Erlbaum Associates; 2003.

[20] Cameron AC, Trivedi PK. “*Regression analysis of count data*”, 2nd edition, Cambridge University Press; 2013.

[21] Kabacoff RI. *“R in Action”*, Shelter Island: Manning Publication; 2011.

[22] Faraway JJ, “*Extending the linear model with R*”, Boca Raton: CRC Press; 2006.

[23] Cameron AC, Windmeijer FAG. “An R-squared measure of goodness of fit for some common nonlinear regression models”, Journal of Econometrics 1997; 77: 329-342.

BIOGRAPHY

Bulent Tutmez was born in Pertek, Turkey, in 1974. He received the diploma in Mining Engineering from Çukurova University, and the Ph.D. degree from Hacettepe University. His main areas of research include: evaluation of uncertainties in engineering, soft computing, geostatistics and occupational safety. He is currently working as a professor at the Faculty of Engineering at Inonu University.

PROCENA NEZGODA SA SMRTNIM ISHODOM NA KOPOVIMA UGLJA POMOĆU REGRESIONE ANALIZE

Bulent Tutmez, Mert G. Ozdogan

Rezime: *Veliki broj studija se bavi različitim parametrima koji utiču na bezbednost i zdravlje radnika u rudnicima, ali i dalje ne postoji dovoljno istraživanja o proceni nesreća sa smrtnim ishodom na površinskim kopovima uglja. Najvažnije rezerve tvrdog uglja u Turskoj nalaze se isključivo u provinciji Zonguldak na severozapadu Anatolije. U radu je izvršena procena broja smrtnih slučajeva koji su se desili na kopovima uglja u basenu Zonguldak u periodu od 2000. do 2014. pomoću regresione analize. Za objašnjenje povezanosti između lokacije i smrtnih slučajeva, razvijen je model Poasonove regresije za prebrojive podatke. Na ovaj način je ukazano na rizičnu radnu oblast. Na osnovu statističkih analiza i testova predložen je model i mogućnost njegove upotrebe. Model i scenariji rizika pokazali su da donosioci odluka mogu iskoristiti dobijene rezultate u cilju sveobuhvatne ocene održivosti.*

Ključne reči: eksploatacija uglja, smrtnost, prebrojivi podaci, Poasonova regresija, prevencija gubitaka.