Application of discriminant analysis-based model for prediction of risk of low back disorders due to workplace design in industrial jobs

Ganga, G. M. D.[a,*], Esposto, K. F.[b] and Braatz, D.[a]

[a] *Production Engineering Department, University Federal of São Carlos, Via Washington Luis, Km 235, São Carlos, São Paulo, Brazil.*
[b] *Production Engineering Department, University of São Paulo, Avenida Trabalhador São-carlense, 400, São Carlos, São Paulo, Brazil.*

Abstract. The occupational exposure limits of different risk factors for development of low back disorders (LBDs) have not yet been established. One of the main problems in setting such guidelines is the limited understanding of how different risk factors for LBDs interact in causing injury, since the nature and mechanism of these disorders are relatively unknown phenomena. Industrial ergonomists' role becomes further complicated because the potential risk factors that may contribute towards the onset of LBDs interact in a complex manner, which makes it difficult to discriminate in detail among the jobs that place workers at high or low risk of LBDs. The purpose of this paper was to develop a comparative study between predictions based on the neural network-based model proposed by Zurada, Karwowski & Marras (1997) and a linear discriminant analysis model, for making predictions about industrial jobs according to their potential risk of low back disorders due to workplace design. The results obtained through applying the discriminant analysis-based model proved that it is as effective as the neural network-based model. Moreover, the discriminant analysis-based model proved to be more advantageous regarding cost and time savings for future data gathering.

Keywords: lifting task assessment; statistical analysis; musculoskeletal injuries.

1. Introduction

The occupational exposure limits of different risk factors for development of low back disorders (LBDs) have not yet been established. The main problem in setting such guidelines is the high difficulty in understanding and correlating interactions between different risk factors for LBDs, with regard to causing injuries, given that the nature and mechanism of these disorders are relatively unknown phenomena. Within this context, industrial ergonomists' role becomes further complicated, since they cannot clearly define the complex manner in which the potential risk factors that may contribute towards the onset of LBDs interact and impact on each other. This makes it difficult to discriminate well between the jobs that place workers at high or low risk of LBDs [9].

Papers relating to the application of quantitative (mathematical and statistical) methods and artificial intelligence to predict and evaluate human, health, safety and environmental factors, etc. [3, 5, 9-11], within such contexts, are increasingly common.

The purpose of the present paper was to develop a comparative study between applying the neural network-based model proposed by Zurada, Karwowski

*Corresponding author. E-mail: ganga@dep.ufscar.br.

and Marras [9] and a linear discriminant analysis-based model, for predicting the potential risk of low back disorders due to workplace design among industrial jobs. The main hypothesis put forward was that the discriminant analysis-based model would attain a higher performance level than shown by the artificial neural network-based method. Another specific hypothesis put forward related to the capability of discriminant analysis, which is a technique based on multivariate statistical analysis, to distinguish the variables that have higher importance in determining the degree of risk of low back disorders due to workplace design [7].

This article is set out in several sections. Section 2 presents the Zurada, Karwowski & Marras model. Section 3 describes the method and explains the basic principles of discriminant analysis (DA). Section 4 described the application and testing of the discriminant analysis stepwise method in order to classify the set of jobs. Finally, section 5 presents the comparison between the discriminant model and the neural network model.

2. Zurada, Karwowski & Marras model

The Zurada, Karwowski and Marras model [9] is an artificial neural network-based diagnostic system that classifies industrial jobs according to the potential risk for low back disorders due to workplace design.

The experimental data were collected in a study involving industrial surveillance of trunk motion and quantification of workplace factors as presenting high or low risk of LBDs, in relation to repetitive tasks [15]. Data gathered on industrial lifting jobs in manufacturing companies were used. These jobs were divided into two groups, i.e. high and low risk of LBDs, based upon examination of previous injuries and medical records. Out of the 235 industrial jobs with low and high risks of LBDs, 148 jobs were randomly selected for network training. This group of 148 jobs contained 74 low-risk and 74 high-risk jobs.

The order of these 148 jobs in the training set was also randomized. The remaining 87 jobs were used for testing the network performance after training. This test group consisted of 50 low-risk jobs and 37 high-risk jobs. The purpose of breaking up the data into a training set and a test set was to provide a check on a 'real world' situation. The training set was used to train the neural network, a procedure that reduces the least mean square error between the cor-

rect response and the actual response until all examples from the training set are learned within an acceptable overall error. New data that the system had not been exposed to were then presented to the network, and its performance relating to these new data patterns was evaluated.

Each observation in the training data set contained the five variables that described occupational risk factors for development of LBDs. These variables were as follows (Figure 1): (1) lift rate in number of lifts per hour (LIFTR); (2) peak twist velocity average (PTVAVG); (3) peak moment (PMOMENT); (4) peak sagittal angle (PSUB); and (5) peak lateral velocity maximum (PLVMAX). To prevent network saturation, these variables were normalized to values from within the interval [0, 1]. The classification variable (RISK of LBDs) took values of 1 or 0 for high and low-risk jobs, respectively.
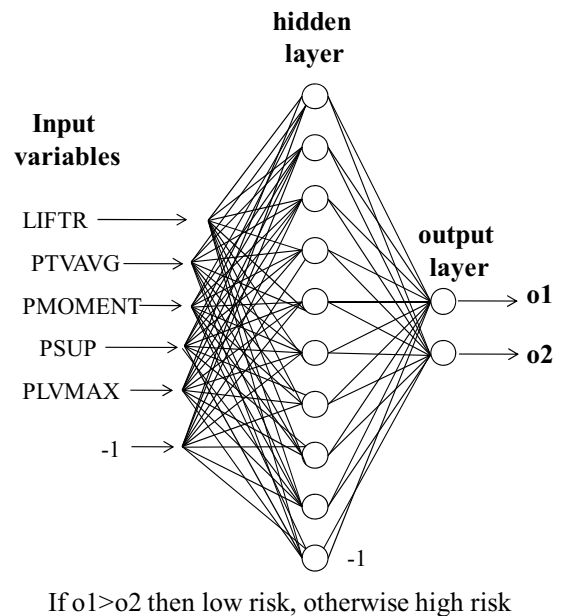

If o1>o2 then low risk, otherwise high risk

Figure 1. Architecture of the neural network.

The results from the Zurada, Karwowski and Marras model [9] showed that the diagnostic system developed could successfully classify jobs into the low and high-risk categories of LBDs based on lifting task characteristics. Figure 2 illustrates the efficiency and effectiveness of the classification based on this model.

As can be seen, the diagnostic system developed correctly classified 72% of the low-risk jobs, 78% of the high-risk jobs and 75% of the medium-risk jobs.
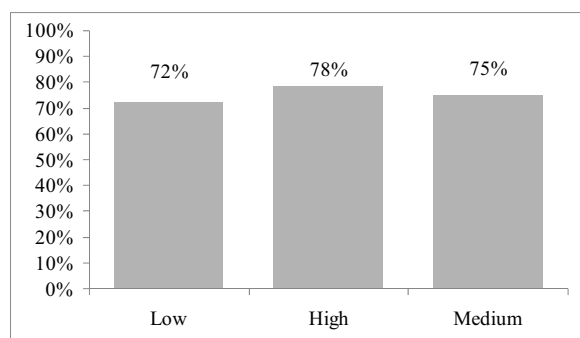
Figure 2. Classification Results – Neural Networks

3. Discriminant Analysis method

The research presented in this paper was descriptive, empirical and quantitative [8]. The primary interest in this class of model-based research is in creating a model (a prediction model in this case) that adequately classifies industrial jobs according to the potential risk for low back disorders due to workplace design. In order to achieve this aim, a multivariate statistical technique of discriminant analysis was used [1, 2, 4, 6, 12-14].

Discriminant analysis is a statistical multivariate technique used when the intention is to establish the relationship between a non-metric dependent variable and metric dependent variables. Through discriminant analysis, it is possible to identify which variables are most relevant for explaining differences among groups that are heterogeneous within a specific context, but homogeneous overall. This analysis searches for a function, i.e. a linear combination of two or more independent variables, that discriminates best among *a priori* groups [7].

Discriminant analysis was applied in two groups, because the dependent variable involved in this analysis (risk) had two possible categories, i.e. low and high. When two classifications are involved, the technique refers to both groups of discriminant analysis, and it is appropriate to test the hypothesis that the averages of the independent variable groups located for two or more groups are equal. The averages of the discriminant scores of the groups are related to their centroids. There are as many centroids as the number of groups; comparison among them shows the extent to which the groups are separated during the testing of the discriminant function.

Application of discriminant analysis makes it possible:
- To establish discriminant functions that better discriminate between the dependent variable categories (for risk: low or high);
- To check whether there are any significant differences among the groups regarding the independent variables;
- To identify the independent variables that contribute most to the differences among the groups;
- To classify the cases in one of the groups regarding the values of the independent variables.

Two computing methods are usually used to determine the discriminant function: simultaneous estimation (concurrent estimation), when the analysis involves all independent variables; and stepwise estimation, as used in the present study, when the intention is to identify the variables that have the greatest discriminant strength. In this latter method, the independent variables are analyzed one by one, starting from the one that discriminates best among all of them [12].

3.1 Model

Discriminant analysis implies searching for a theoretical value that is the linear combination of the independent variables that discriminates best between the previously defined groups, as shown in Equation 1 [1, 2, 4, 6, 12-14]:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + W_3 X_{3k} + \ldots + W_n X_{nk} \quad \text{Eq. (1)}$$

where:
Z_{jk}: value of the discriminant function j for the object k;
a: constant;
W_I = the discriminant coefficient or weight for variable i.
X_{ik} = independent variable i for the object k.

The data used in developing the model in this study was based on the Zurada, Karwowski and Marras model [9]. The proposed construct for developing the discriminant analysis model is illustrated in Figure 3.

Figure 3. Construct of discriminant analysis model

4. Results

The first test in developing the model was Wilk's Lambda, to ascertain the equality of the group averages, thereby identifying most discriminant variables. The lower the result from this test is, the higher the strongest discriminatory function of the groups involved is. Table 1 illustrates the result from this test applied to the variables involved.

Table 1

Equality tests on group means

Variables	Wilks Lambda	F	Sigma
LIFTR (1)	0,979	3,133	0,079
PTVAVG (2)	0,889	18,295	**0,000**
PMOMENT (3)	**0,817**	32,601	**0,000**
PSUP (4)	0,938	9,640	**0,002**
PLVMAX (5)	0,891	17,817	**0,000**

Analysis on Table 1 showed that the variable with the highest discriminant strength was peak moment (PMOMENT), which had a Wilk's Lambda value of 0.817.

Table 1 also shows the results from the F-ANOVA test, which indicated that there was a significant difference among the averages for the group, at the significance level < 0.05. The results support the variables presented and show another three possible candidates for discriminative variables: PTVAVG, PSUP and PLVMAX, respectively.

The correlation matrix illustrated in Table 2 makes it possible to identify probable cases of multicollinearity and to anticipate the variables that may affect the stepwise process and, consequently, cease to be part of the discriminant function.

Table 2

Correlation matrix

	(1)	(2)	(3)	(4)	(5)
(1)	1,000	0,159	-0,050	0,031	-0,002
(2)	0,159	1,000	0,243	0,053	0,456
(3)	-0,050	0,243	1,000	0,180	0,081
(4)	-0,031	0,053	0,180	1,000	0,219
(5)	-0,002	0,456	0,081	0,219	1,000

No significant correlation was identified among the variables in the model. If any correlation had occurred, it would have been advisable take such variables out of the stepwise process, because they could have an influence on identifying variables with higher discriminant strength.

Box's M test can confirm whether a premise of broken equality has occurred among the covariance matrixes. This test is based on an F transformation. It tests H_0 regarding covariance equality matrixes through the resultant significance level. Table 3 presents the results from Box's M tests.

Table 3

Results from Box's M tests

Box's M		15,727
F	Approx.	5,165
	df1	3
	df2	3836880,0
	Sig.	**0,001**

The analysis showed that the equality premise among the covariance matrixes was violated, given that the result from the test was lower than 0.05.

Therefore, in order to evaluate the capacity of the model to separate and classify the groups, another Wilk's Lambda test was applied at the end of the process. From these two steps in developing the model, the best variables for the discriminant function were found. These were selected based on the significance level that they reached with the predefined confidence interval (95%). These variables were PMOMENT and PLVMAX, both with significance < 0.05, as illustrated in Figure 4. The other variables were not selected.

<div style="page-columns: 2">

Table 4
Variables in the Analysis[a,b,c,d]

	Step	1	2
	Entered	PMOMENT	PLVMAX
Wilks' Lambda	Statistic	0,817	0,857
	df1	1	2
	df2	1	1
	df3	146	146
Exact F	Statistic	32,601	23,253
	df1	1	2
	df2	146	145
	Sig.	**0,0000**	**0,0000**

At each step, the variable that minimized the overall Wilk's Lambda was entered.
a - Maximum number of steps was 10.
b - Maximum significance of F to enter was 0.05.
c - Minimum significance of F to remove was 0.10.
d - F level, tolerance or VIN insufficient for further computation.

In this manner, the resultant canonical discriminant function is illustrated in Table 5.

Table 5
Canonical Discriminant Function Coefficients

Variables	Function
	1
PMOMENT	0,015
PLVMAX	0,033
(Constant)	-2,101

This function can be interpreted such that the variable PLVMAX has greater strength than the variable PMOMENT in the composition of the discriminant function.

To classify each job in a group, two classification functions were generated, which were named Fischer's linear functions and are illustrated in Equation 2 and Equation 3, respectively.

$$\text{Low} = 0.006 \text{ PMOMENT} + 0.122 \text{ PLVMAX} -2.885 \quad \text{Eq. (2)}$$

$$\text{High} = 0.024 \text{ PMOMENT} + 0.159 \text{ PLVMAX} -5.248 \quad \text{Eq. (3)}$$

Analysis on the coefficients of the two equations (2 and 3) made it possible to affirm that the jobs that presented higher values for PLVMAX and PMO-

MENT would be classified as presenting a high level of risk; otherwise, they would be considered to have a low level of risk.

The results from the classification of the discriminants functions are illustrated in Table 6.

Table 6
Classification Results

			Class of risk	Predicted group membership		Σ
				low	high	
Jobs Selected	Original	Freq.	low	60	14	74
			high	26	48	74
		%	low	81,1	18,9	100
			high	35,1	64,9	100
	Cross-validated (a)	Freq.	low	60	14	74
			high	28	46	74
		%	low	81,1	18,9	100
			high	37,8	62,2	100

Cross-validation was done only for the cases in the analysis. In cross-validation, each case is classified according to the functions derived from all cases other than that case. Thus, the final results regarding the efficiency of the discriminant functions generated were:

– 73.0% of the originally grouped cases were correctly classified;
– 71.6% of the cross-validated grouped cases were correctly classified.

Figure 4 also illustrates the efficiency of classification for the two risk groups.
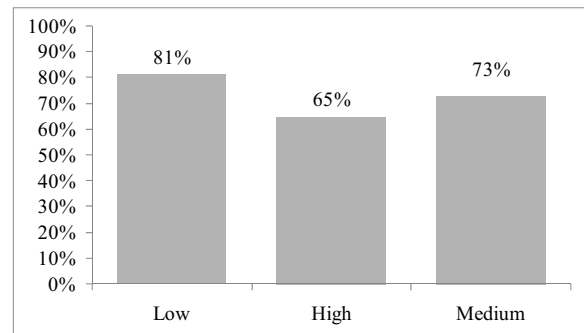

Figure 4
Classification Results – Discriminant Analysis

</div>

The discriminant analysis model classified 73% of the cases correctly. For low-risk jobs, 81% were correctly classified. For high-risk jobs, 65% were correctly classified.

4.1. Comparative analysis

The comparison between the two models considered in this study on the classification efficiency is illustrated in Figure 5.
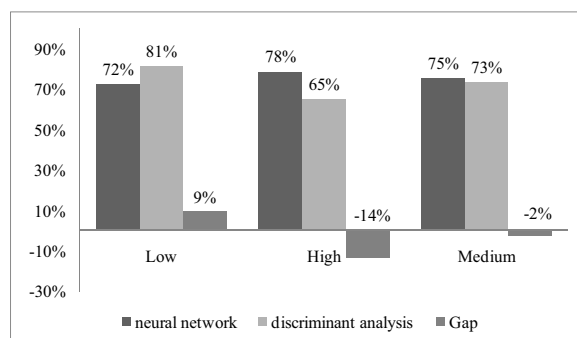

Figure 5. Comparative analysis

The discriminant analysis-based model for prediction correctly classified 81% of the cases in the low-risk group. On the other hand, the neural network-based model correctly classified only 72% of the cases. Thus, there was a difference of 9% in the classifications outlined by the models. For the classification of jobs in the high-risk group, there was a difference of 14% between the models. The model obtained using neural networks had a higher success rate (75%) than shown by the discriminant analysis model (73%). The difference in overall performance between the models resulted in a non-significant advantage for the neural network model.

5. Conclusions

The great contribution of the discriminant analysis model was in relation to reduction of the factors considered in the risk classification. As can be seen in Table 5, only the variables PMOMENT and PLVMAX were included in the model. In both cases, the variables had high discriminatory power, as shown by the Wilk's Lambda values. Moreover, both variables were considered statistically significant at the significance level < 0.05.

Thus, the classification made by the discriminant analysis-based model would be more advantageous for cost and time savings in future data gathering, since only two independent variables were included in the model.

It is emphasized that this model could also be used to handle qualitative data, which are common in ergonomic interventions.

Moreover, this application could support early stages of ergonomic analysis, especially with regard to understanding the demands from specific production situations and assisting in designing workplace developments.

References

[1]   C. Biernacki and G. Govaert, Choosing models in model-based clustering and discriminant analysis, Journal of Statistical Computing Simulation 64 (1999), 49-71.

[2]   C. J. Huberty, Applied discriminant analysis, John Wiley, NY, 1994.

[3]   G. E. Gürcanli and U. Müngen, U, An occupational safety risk analysis method at construction sites using fuzzy sets, International Journal of Industrial Ergonomics, 39 (2009), 371-387.

[4]   G. J. MacLchlan, Discriminant analysis and statistical pattern recognition, John Wiley, NY, 1992.

[5]   I. Aluclu, A. Dalgic, A. and Z. F. Toprak, A fuzzy logic-based model for noise control at industrial workplaces, Applied Ergonomics 39 (2008), 368-378.

[6]   J. D. Jobson, Applied multivariate data analysis, v.I e II. Springer Verlag, NY, 1996.

[7]   J. F. Jr. Hair, R. Anderson, R. Tatham and W. C. Black, Multivariate Data Analysis, Prentice Hall, NJ, 1998.

[8]   J. W. M. Bertrand and J. C. Fransoo, Modelling and Simulation: Operations management research methodologies using quantitative modeling, International Journal of Operations & Production Management 22 02 (2002), 241-264.

[9]   J. Zurada, W. Karwowski and W. S. Marras, A neural network-based system for classification of industrial jobs with respect to risk of low back disorders due to workplace design, Applied Ergonomics 28 1 (1997), 49-58.

[10]   L. Hanson, W. Wienholt and L. Sperling, A control handling comfort model based on fuzzylogics, International Journal of Industrial Ergonomics 31 (2003), 87-100.

[11]   M. Kolich, N. Seal and S. Taboun, Automobile seat comfort prediction: statistical model vs.artificial neural network, Applied Ergonomics 35 (2004) 275-284.

[12]   N. K. Malhotra, Marketing Research: An Applied Orientation, 6 ed., Prentice-Hall, NY, 2010.

[13]   P. A. Lachenbruch, Discriminant Analysis, Hafner Press, NY, 1975.

[14]   W. R. Klecka, Discriminant Analysis for Social Sciences, Sage Publications, CA, Beverly Hills, 1980.

[15]   W. S. Marras, S. A. Lavender, S. Leurgans, L. R. Sudhakar, W. G. Allread, F. Fathallah and S. Ferguson, The role of dynamic three-dimensional trunk motion in occupationally-related low back disorders, Spine 18 (1993), 617-628.