AN APPROACH TO THE QUADRATIC NONLINEAR FORMULAE FOR PREDICTING EARTHQUAKE LIQUEFACTION POTENTIAL BY STEPWISE DISCRIMINANT ANALYSIS

Gu Weihua (I)
Wang Yuqing (II)
Presenting Author: Gu Weihua

SUMMARY

This paper discussed the possibility of development of quadratic nonlinear formulae for predicting earthquake liquefaction potential by stepwise discriminant analysis and analyzed two groups of typical data. Furthermore, it also gave a stability limit for nonlinear discriminant functions and offered a suggestion for selecting the optimum discriminant function.

INTRODUCTION

It is customary to assume formulae for predicting earthquake liquefaction potential of sandy deposits to be a linear function, which can be expressed as:

$$Z = \sum_{i=1}^{k} L_i X_i$$

in which Z is liquefaction potential, $L_i (i=1,2,\cdots,k)$ are coefficients, and $X_i (i=1,2,\cdots,k)$ are elementary variables, but this assumption has certain limitations. In practical data, a nonlinear relationship is often noted between certain variables and liquefaction potential, making the study to develop nonlinear discriminant function necessary.

In this paper the quadratic nonlinear form is used, i.e. elementary variables $X_i (i=1,2,\cdots,k)$ are quadratically combined, the combined variables $X_i X_j (i=1,2,\cdots,k;\ j=i,i+1,\cdots,k)$ being new ones are treated equally as elementary variables $X_i (i=1,2,\cdots,k)$ in the analysis. Since number of variables increases considerably following quadratic combination, we cannot have all of them enter the discriminant, but introduce selectively to obtain the optimum discriminant function and ensure its stability. The stepwise discriminant analysis is used here for variable selection and found to be highly efficient according to the results of analysis.

The most important and indispensable problem involved in developing quadratic nonlinear discriminants is how to select the optimum discriminant function and ensure its stability. By analyzing two groups of typical liquefaction data(Ref. 1,2), this paper made a primary approach to the problem and gave a stability limit for

---

(I) Research Engineer, Central Research Institute of Building and Construction, MMI, CHINA
(II) Senior Engineer, Central Research Institute of Building and Construction, MMI, CHINA

nonlinear discriminant functions and offered a suggestion for selecting the optimum discriminant function. In the analysis two nonlinear discriminants are developed appropriate to sandy soil and clayey silt respectively, their discrimination success percentages have improved considerably as compared with the linear formulae.

STEPWISE DISCRIMINANT ANALYSIS (Ref.3)

1. Variable selection

Wilks' criterion $U = |W| / |T|$ can be used to denote separation state between groups, in which W and T are within group and total cross-product matrices respectively.

Given L variables currently in the discriminant, enter a variable X_r, we have

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & w_{rr} \end{bmatrix}$$

in which

$$W_{11} = \begin{bmatrix} w_{11} & w_{12} \cdots & w_{1L} \\ w_{21} & w_{22} \cdots & w_{2L} \\ \cdots & \cdots & \\ w_{L1} & w_{L2} \cdots & w_{LL} \end{bmatrix}$$

$$W_{21} = W_{12}^{T} = \begin{bmatrix} w_{1r} & w_{2r} \cdots & w_{Lr} \end{bmatrix}$$

and

$$|W| = \begin{vmatrix} I & 0 \\ -W_{21}W_{11}^{-1} & 1 \end{vmatrix} \begin{vmatrix} W_{11} & W_{12} \\ W_{21} & w_{rr} \end{vmatrix} = \begin{vmatrix} W_{11} & W_{12} \\ 0 & w_{rr}-W_{21}W_{11}^{-1}W_{12} \end{vmatrix} = |W_{11}| \cdot w_{rr}^{(L)}$$

in which

$$w_{rr}^{(L)} = w_{rr}-W_{21}W_{11}^{-1}W_{12}$$

For the same reason

$$|T| = |T_{11}| \cdot t_{rr}^{(L)}$$

Hence, Wilks' criterion becomes

$$U = \frac{|W|}{|T|} = \frac{|W_{11}|}{|T_{11}|} \cdot \frac{w_{rr}^{(L)}}{t_{rr}^{(L)}}$$

in which $|W_{11}| / |T_{11}|$ denotes separation state between groups for L variables currently entered, while $w_{rr}^{(L)} / t_{rr}^{(L)}$ denotes separation power between groups for L variables selected with entry of variable X_r. The latter can be written as:

$$U_{r|1,2,\cdots,L} = w_{rr}^{(L)} \Big/ t_{rr}^{(L)}$$

the corresponding F-statistic

$$F = f(N-G-L) = \frac{1-U_{r|1,2,\cdots,L}}{U_{r|1,2,\cdots,L}} \cdot \frac{N-G-L}{G-1}$$

can be used to test the significance of the change in U resulting from the addition of X_r and will be used here to guide variable selection. In the formula N, G, L are number of cases, groups and variables respectively.

If $F \geqslant Fa$, it implies that with entry of the variable X_r good separation between groups will result; if $F < Fa$, poor separation between groups will result. Fa is the specified threshold.

2. Stepping procedure

(1) Entry of variables (Given L variables selected in the function)

Find the variable with the smallest $U_{r|L}$ among all the variables not to be entered, compute $F=f(N-G-L)$. If $F \geqslant Fa$, enter the variable, turn to test separation power between groups for L variables selected with entry of the variable, and then delete variables following step (2). If $F < Fa$, terminate variable selection and turn to develop the discriminant.

(2) Deletion of variables

Find the variable with the smallest $U_{r|(L-1)}$ among L variables selected, compute $F=f(N-G-L+1)$. If $F < Fa$, we consider separation power between groups for the variable to be poor and have it delete. Meanwhile, find the variable with the smallest $U_{r|(L-2)}$, repeat step(2) until no variables are to be delected, then turn again to enter variables following step(1).

(3) Development of discriminant

To save internal storage location we adopt the scheme variable selection - inversion, i.e. transform the matrix W as follows (for step L+1):

$$w_{ij}^{(L+1)} = \begin{cases} w_{rj}^{(L)} \Big/ w_{rr}^{(L)} & (i = r,\ j \neq r) \\ w_{ij}^{(L)} - w_{ir}^{(L)} w_{rj}^{(L)} \Big/ w_{rr}^{(L)} & (i \neq r,\ j \neq r) \\ 1 \Big/ w_{rr}^{(L)} & (i = r,\ j = r) \\ -w_{ir}^{(L)} \Big/ w_{rr}^{(L)} & (i \neq r,\ j = r) \end{cases}$$

Similarly, transform the matrix T . The corresponding inverse matrix W^{-1} has been obtained following the conclusion of variable selection, the corresponding inverse covariance matrix is $S^{-1} = (N-G)W^{-1}$. Hence, coefficients for the discriminant are $L=S^{-1}d$ in the case of two groups in analysis, in which $L^T = [L_1, L_2, \cdots, L_k]$, $d^T = [d_1, d_2, \cdots, d_k]$ and $d_i = \bar{X}_i^{(1)} - \bar{X}_i^{(2)}$, $\bar{X}_i^{(g)}$ denotes the mean value of the ith variable in the gth group, k is number of variables to be used in the final analysis.

ANALYSIS ON LIQUEFACTION DATA OF 93 CASES OF CLAYEY SILT IN TIANJIN AREA (Ref.2)

In the analysis 10 elementary variables were considered, including depth of stratum ds(M), number of SPT blows N, ground water level dw(M), thickness of overlying cohesive soil strata S(M), thickness of stratum under consideration h(M), plasticity index Ip, content of cohesive grain(<0.005mm) Ps(%), mean grain size d_{50}(mm), uniformity coefficient Cu, content of silt grain (0.05——0.005 mm) P_2(%).

Following quadratic combination of elementary variables, 65 variables are obtained. The results of stepwise discriminant analysis on these variables are tabulated in Table 1.

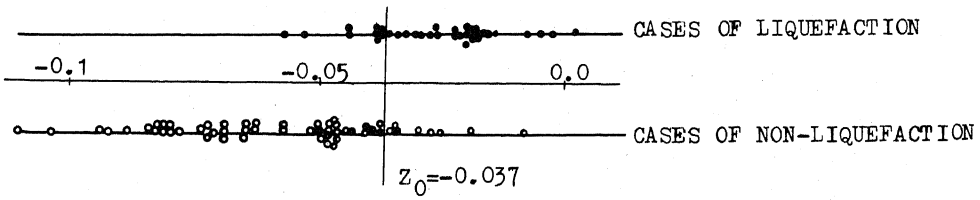
Table 1 Results of stepwise discriminant analysis on liquefaction data of 93 cases of clayey silt

Quadratic combination	Fa		Variables entered	P (%)
	Fin	Fout		
No	5.0	5.0	ds, N, Ps	82.8
No	0	0	ds,N,dw,S,h,Ip,d_{50},Cu,P_2,Ps	81.7
Yes	6.97	6.97	Cu, Ndw, Sh	76.3
Yes	3.96	2.77	Cu, Ndw, Sh	76.3
Yes	2.77	2.77	N,h,Cu,N^2,Sdw,hdw	88.2
Yes	1.5	1.5	N,h,N^2,NS,NCu,Sdw,hdw,Cudw,hd_{50},Ps^2	91.4
Yes	1.0	1.0	N,dw,S,h,N^2,NS,NIp,NCu,dwh,dwCu,S^2, Sh,SIp,h^2,hIp,$P_2{}^2$	94.6
Yes	0.5	0.5	N,dw,S,h,Ps,N^2,NIp,NCu,dw^2,dwS,dwh, dwCu,S^2,Sh,h^2,hIp,hd_{50},hP_2,IpCu,Ps^2, PsCu,PsP_2,d_{50}^2,Cu^2	97.8
Yes	0	0	10 elementary variables and 55 combination variables	100

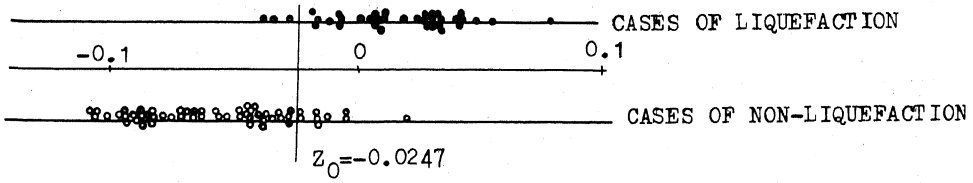
AS seen from Table 1, discrimination success percentages P have not improved for linear functions with variables S,h,dw,Ip,d_{50},Cu,P_2 entered, but have improved considerably for nonlinear functions with combined variables entered. It further demonstrates the limitations of linear function assumption.

It can also be observed that when Fa decreases from 6.97 to 1.0, discrimination success percentages have improved markedly with increasing number of variables. But when Fa is taken below 1.0, discrimination success percentages have not improved markedly despite substantial increase in number of variables. Consequently, there exists a problem how to select the optimum discriminant function, which is defined as one that reaches the highest possible discrimination success percentage while minimizing number of variables, espesially elementary ones.

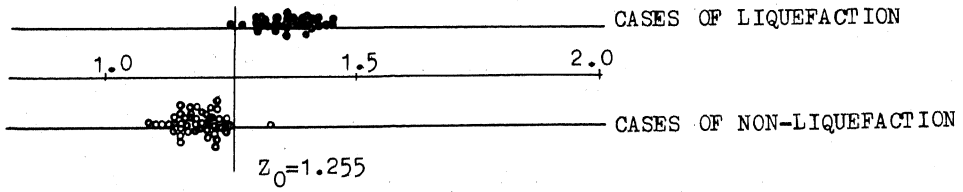
Figure 1 shows distribution of liquefaction potential Z obtained from linear formula corresponding to Fa=5.0 and nonlinear formulae corresponding to Fa=2.77,0.5,0 in Table 1. It can be seen from Figure 1



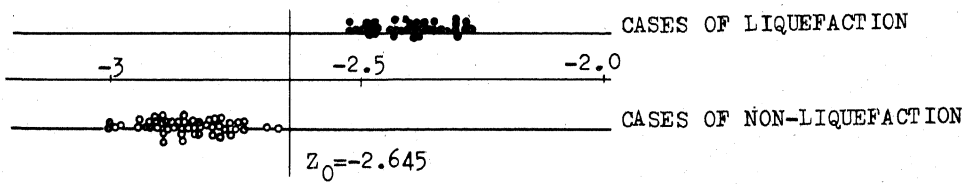
(a) Distribution of Z obtained from linear formula (F = 5.0)



(b) Distribution of Z obtained from nonlinear formula (F = 2.77)



(c) Distribution of Z obtained from nonlinear formula (F = 0.5)



(d) Distribution of Z obtained from nonlinear formula (F = 0)

Figure 1 Distribution of liquefaction potential Z of 93 cases of clayey silt
(Z_0——Critical liquefaction potential)

that for nonlinear formulae, distribution region of liquefaction potential Z extends with increasing number of variables, i.e.sensitivity and hence separation power(or accuracy) of functions are increased.

ANALYSIS ON LIQUEFACTION DATA OF 35 CASES OF SANDY SOIL AFTER SEED AND IDRISS(Ref.1)

6 elementary variables were considered,including Richter Magnitude M,epicentral distance R(km), ground water level dw(M), depth of stratum ds(M),number of SPT blows N, duration of earthquake t(Sec). The results of analysis are tabulated in Table 2.

Table 2 Results of analysis on data of sandy soil

Fa	Number of variables entered L	Discrimination success percentage P (%)	
		With respect to data of Seed and Idriss	With respect to data of Whitman R.V.
2.0	2	77.1	66.7
1.5	9	91.4	66.7
1.0	10	100	88.9
0	27	100	44.4

The results show that the discriminant has reached complete separation for data to be considered in the analysis when Fa is taken as 1.0. In the case of Fa=0, whereas discrimination success percentage of formula with respect to data to be considered in the analysis remains 100 %, there is no separation power at all with respect to data not to be considered in the analysis, discrimination success percentage being only 44.4 %. The decrease by a wide margin in discrimination success percentage with respect to data not to be considered in the analysis is said to be " unstable " for discriminant functions.

AN APPROACH TO SELECTION OF THE OPTIMUM DISCRIMINANT FUNCTION AND ENSURING OF ITS STABILITY

Theoretically, unstability of discriminant function is due primarily to two sources. The first one is inadequate representativeness of data to be considered in the analysis, such as their inadequate number and excessive errors in experiments. The another one is that an excess of variables are selected, causing excessive sensitivity of discriminant functions. In addition, with increasing number of variables, especially combined ones, destruction of independance between variables may result, decreasing computing accuracy during inversion of covariance matrix and hence causing unstability of discriminant functions obtained.

To clarify the effect of these two sources we may plot P against L/N (P — discrimination success percentage, L and N — number of variables and cases respectively) as shown in Figure 2.

From Figure 2 it is apparent that with the increase in L/N, especially the segment with $L/N<0.3$, discrimination success percentages of formulae with respect to data to be considered in the analysis are increased, while maintaining high discrimination success percentages with respect to data not to be considered in the analysis. There is, however, a possibility of instability, e.g. in the case of the segment CD when $L/N>0.3$.

From the viewpoint of statistical theory, the increase in L/N is unfavorable. Consequently, it is reasonable here to take $L/N<0.3$ as a limit to insure the stability of discriminant functions.

By definition, the optimum discriminant function demands L/N be the least and P the highest possible. In Figure 2, if a clear point of inflection (say point E) can be found on the curve of discrimination success percentage with respect to data to be considered in the analysis, it is natural for us to take this point as a basis for selection of the optimum discriminant function. But many other factors should also be considered in selecting the optimum discriminant function, such as number of elementary variables as well as simplicity and cost of testing method. Hence it is preferable to locate roughly a selecting range on the branch of the curve with rapidly increasing P and then make the final choice with other factors comprehensively considered.

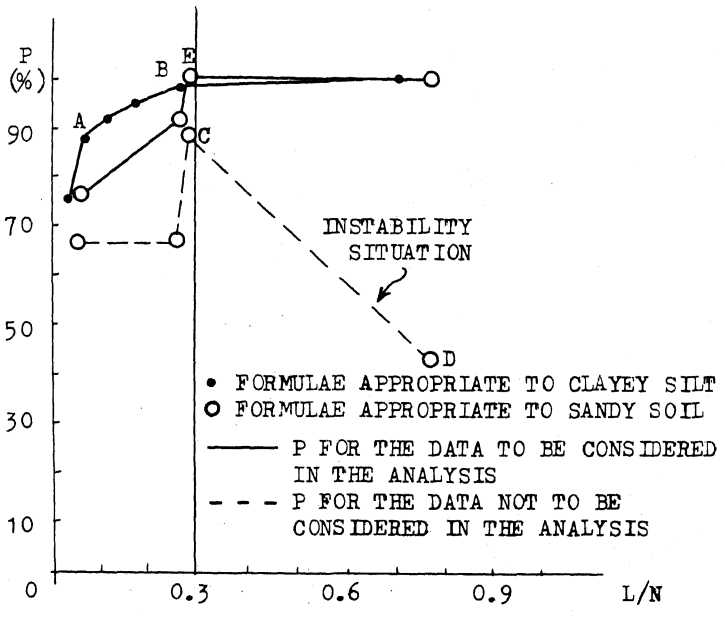


Fig 2. Discrimination success percentage vs number of variables

As for discriminant functions based on liquefaction data of 93 cases of clayey silt, it should be selected from the segment AB in Figure 2. Considering there is no large difference in simplicity of testing methods between elementary variables of formulae on the segment AB, we may select the formula at point B(corresponding to Fa=0.5) with higher discrimination success percentage(P=97.8 %) as the optimum discriminant function, which is written as:

$$Z = -0.00996N+0.12712dw+0.26514S+0.26473h+0.02570Ps+0.00055N^2-$$
$$-0.00201NIp+0.00058NCu+0.04650dw^2-0.01526dwS-0.05387dwh-$$
$$-0.00838dwCu-0.01357S^2-0.02176Sh-0.00252h^2+0.00487hIp-$$
$$-1.30732hd_{50}-0.00071hPs-0.00226IpCu-0.00439Ps^2+0.00310PsCu+$$
$$+0.00059PsP_2+41.83163d_{50}{}^2-0.00039Cu^2$$

$Z_0 = 1.25479$ (liquefaction occurs when $Z>Z_0$)

As for discriminant functions based on liquefaction data of 35 cases of sandy soil after Seed and Idriss, it is proper to select the formula at point E(corresponding to Fa=1.0), its discrimination success

percentage with respect to data to be considered in the analysis being 100 %, while with respect to data of Whitman R.V. being 88.9 %. The formula is written as:

$$Z = -0.02468R+0.97561dw-0.02890N+0.01895M^2+0.00248MR-0.07790Mdw+$$
$$+0.00002R^2-0.05873dwds-0.00597Ndw+0.00870ds^2$$

$Z_0 = 0.8378$ (liquefaction occurs when $Z>Z_0$)

CONCLUSIONS

Discrimination success percentage with respect to data to be considered in the analysis may be increased through the use of quadratic nonlinear formulae for predicting earthquake liquefaction potential. But with increasing number of variables the problem of selecting the optimum discriminant function and insuring its stability becomes more pronounced than that of developing linear discriminant functions.

Stability of discriminant functions is a rather complicated problem. In the paper the suggestion that $L/N<0.3$ be adopted as a limit to ensure stability of discriminant functions is only an approach on a trial basis.

While satisfying stability of discriminant functions as a prerequisite, we should select the optimum discriminant function, in the course of which number of elementary variables, simplicity of testing methods and others should be comprehensively considered in addition to such factors as number of variables and discrimination success percentage.

With the increasing popularization of computers as well as continuous accumulation and perfection of data there are good prospects for nonlinear formulae.

REFERENCES

1. Kiichi Tanimoto and Tsutomu Noda; Prediction of Liquefaction Occurrence of Sandy Deposits during Earthquake by a Statistical Method, PROC OF JSCE, No.256,DEC 1976.
2. Wang Yuqing, Luan Fang, Han Qingyu, Li Guoxin; Formulae for Predicting Liquefaction Potential of Clayey Silt as Derived from A Statistical Method, Proceedings of the Seventh World Conference on Earthquake Engineering, Vol.3, Geotechnical Aspects, September 8—13,1980.
3. Robert I. Jennrich; Statistical Methods For Digital Computers, pp 76 — 95.