



Comparing Discriminant Analysis and Logistic Regression Model as a Statistical Assessment Tools of Arsenic and Heavy Metal Contents in Cockles

Abbas F. M. Alkarkhi (Corresponding author)

School of Industrial Technology, Environmental Technology Division

Universiti Sains Malaysia, 11800 Penang, Malaysia

Tel: 60-4-653-2107 E-mail: alkarkhi@yahoo.com

Azhar Mat Easa

School of Industrial Technology, Food Technology Division

Universiti Sains Malaysia, 11800 Penang, Malaysia

Tel: 60-4-653-2222 E-mail: azhar@usm.my

Abstract

Two statistical techniques; discriminant analysis (DA) and logistic regression model were used to analyze the concentration of arsenic and heavy metal contents in cockles (*Anadara granosa*) from two estuaries in the state of Penang, Malaysia. This study was undertaken to understand the interrelationship between different parameters and also to identify probable source component in order to explain the pollution status. Arsenic (As), chromium (Cr), cadmium (Cd), zinc (Zn), copper (Cu) and lead (Pb) were analyzed using a graphite flame atomic absorption spectrophotometer (GF-AAS) whilst mercury (Hg) was analyzed using a cold vapor atomic absorption spectrophotometer (CV-AAS). Logistic regression model showed that only two explanatory variables Zn ($p < 0.01$) and Cd ($p < 0.05$) exhibited significant effect to discriminate cockles in the two locations and responsible for large variation affording 77.5% correct assignment. On the other hand DA identified the same parameters Zn and Cd which are responsible in discriminating the two locations affording 72.5% correct assignment. Comparison between logistic regression model and DA exhibited that both techniques gave close results in discriminating the two locations.

Keywords: Heavy metals, Arsenic, Estuary, Logistic regression model, Discriminant analysis

1. Introduction

The discharge of effluents and associated toxic compounds into aquatic systems represents an ongoing environmental problem due to their possible impact on communities in the receiving aquatic water and a potential effect on human health (Canivet and Gibert, 2002). Especially in highly polluted and industrial areas, point and non-point sources of anthropogenic chemicals and metals have polluted rivers with highly complex mixtures of chemicals and other anthropogenic perturbations to degree where life in rivers is severely impacted (Smolders et al., 2004).

The application of different statistical methods has increased tremendously in recent years for analyzing environmental data (Hernandez et al., 2005; Vallvey et al., 2006). In the present study logistic regression model was used to identify the most important variables responsible in discriminating between two groups and study the relationship between one or more explanatory variables (here arsenic and heavy metals) and the dependent variable (location). In addition it can also be used for prediction (Erling, 1997; Agresti, 2002). The results of logistic regression model was compared with DA, since both techniques can be used for classification and identifying the contribution of each parameter in discriminating the two locations.

Cockles rearing in Penang that was valued at RM 8.12 million in 2001 is the third largest in Malaysia after the states of Perak and Selangor. Kuala Juru has been one of the main areas of cockle cultivation in Penang. After 1999, cockle production in Penang had undergone a declining trend with the water pollution being the primary reason. This

situation is predicted to further degenerate if Penang coastal waters are being continually polluted from effluents of local industries (Socioeconomic and Environmental Research Institute report, 2002). This study was therefore undertaken to determine whether the concentrations of arsenic and heavy metals in cockles sampled reared in Kuala Juru (Juru River) and Bukit Tambun (Jejawi River) are different based on the concentrations of arsenic (As) and six heavy metals (Cu, Pb, Cd, Cr, Zn, and Hg). In addition it is important to identify the most important explanatory variables which help in distinguishing the cockles according to the above selected parameters.

The objective of this paper is to compare the use of logistic regression model and discriminant analysis (DA) as a statistical tools to identify the contribution of each variable in distinguishing between two groups. This study may also illustrate the usefulness of statistical analysis for evaluation and interpretation of large complex data sets to get better information about arsenic and heavy metal contents in cockles.

2. Materials and methods

2.1 Description of study area

The study site is located on the North West coast of peninsular Malaysia, in the state of Penang and within a coastal mudflat in the Juru and Bukit Tambun district. The sites are located adjacent to industrial areas which were reclaimed from mangrove. The types of industry presently in operation include: electronics; textiles; basic and fabricated metal products; food processing and canning; processing of agricultural products; feed mills; chemical plants; rubber based industry; timber based wood products; paper products and printing works; and transport equipment. Other main activities that are operating in vicinity of the cultured area are a ships' harbor with petroleum unloading and a red earth quarry which extends right up to the coastline. There are three main rivers flowing into the area, Sungai Juru, Sungai Semilang and Sungai Jejawi where some fishing villages are situated.

2.2 Analytical procedure

Samplings were carried out during a rainy season in the year 2005. 5 (five) samples of cockles (*Anadara granosa*) were collected from each of the 40 estuarine sites in Juru and Jejawi Rivers (20 sites from each location). The cockles were collected manually at low tide from the inter-tidal flats (slikkes) of the two areas. Cockles were cleaned externally by washing thoroughly through a 1-mm mesh sieve with deionized water before being transferred into a high density polyethylene (HDPE) sampling bag. After a 36 hrs purging period, during which time sediment-bound metals were voided from the gut, the bivalves of approximately similar size (32.6 mm) were rinsed in deionized water. Soft parts (7-10 g of wet tissue) were digested in 10 mL of boiling concentrated nitric acid (Analar grade) to near dryness. Additional digestion was accomplished by the addition of 10 mL of 1:1 nitric acid/deionized water. The resulting residue was diluted to 50 mL in deionized water. The solution mixture was then filtered through a 0.45- μm cellulose nitrate filter prior to arsenic and heavy metals analysis.

Graphite furnace atomic absorption spectrophotometer (GF-AAS; Perkin Elmer HGA-600) was employed for the analysis of arsenic and heavy metals (Cr, Cd, Zn, Cu and Pb) and cold vapour atomic absorption spectrophotometer (CV-AAS) method was employed for Hg analysis after sample digestion in acid solution. Calibration curves from standard mixtures of 0.05, 0.1, 0.2, 0.5 and 1.0 mgL^{-1} of Cu, Zn, Cd, Cr and Hg and 0.5, 1.0, 2.0, 5.0 and 10 mgL^{-1} of As and Pb were prepared in nitric acid solution. The accuracy of the methods was determined by preparing digestion mixture blanks and by spiking the sample with known concentrations of As, Cu, Cd, Cr, Zn, Pb and Hg with mean recoveries of $90.5 \pm 1.2\%$.

2.3 Statistical methods

2.3.1 Logistic regression model

Logistic regression model or logit deals with the binary case, where the response variable consists of just two categorical values. Logistic regression model is mainly used to identify the relationship between one or more explanatory variables X_i and the dependent variable Y . Logistic regression model has been used for prediction and determining the most influential explanatory variables on the dependent variable (Cox and Snell, 1994).

The logistic regression model for the dependence of p_i (response probability) on the values of k explanatory variables x_1, x_2, \dots, x_k is given in below (Collett, 2003):

$$\text{logit}(P_i) = \log \left(\frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

(1)

Or

$$p_i = \exp \left(\frac{\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})} \right)$$

(2)

which is linear and similar to the expression for multiple linear regression.

where, $\left(\frac{p}{1-p}\right)$ is the ratio of the probability of a success to the probability of a failure, and called odds,

β_0, β_i are parameters to be estimated, and p_i is the response probability.

In logistic regression model the predicted values for the response will never be ≤ 0 or ≥ 1 , regardless of the values of the explanatory variables.

3. Result and discussion

The univariate statistics including maximum, minimum, median and Q_1, Q_2 quartiles for Juru and Jejawi Rivers are presented in (Fig.1 a and b) respectively. It can be seen that As and Hg exhibited higher levels in cockles, whereas Cr and Pb exhibited the lowest. The observed variability of arsenic and heavy metal contents in cockle was low and random.

Logistic regression model was applied to study the relationship between the two locations and seven explanatory variables (Arsenic and heavy metal) in order to find the most important variables that discriminate the cockles in the two locations based on selected metals.

Logistic regression was carried out using forward and backward stepwise methods for variable selection to identify the most important explanatory variables that significantly influence the response and help in distinguishing cockles obtained from the two rivers. Both methods identified the same explanatory variables and yielded the following model (Eq.3):

$$p_i = \exp \frac{52.51 - 79.86 \text{ Zn} - 41.23 \text{ Cd}}{1 + \exp (52.51 - 79.86 \text{ Zn} - 41.23 \text{ Cd})}$$

(3)

Only two explanatory variables Zn ($p < 0.01$) and Cd ($p < 0.05$) exhibited significant effect to discriminate cockles in the two locations and responsible for large variation, while other parameters did not show significant effect. This model provided a good fit since the value of Cox and Snell is 0.53 and Negerlkerkr R^2 value is 0.67. The results of classification (Table 1) using logistic regression model showed that more than 77 % of the cockles obtained from different sites were correctly classified to their respective location. It can be said that the differences between cockles obtained from the two rivers belong only to Zn and Cd, while other parameters have almost the same concentration regardless of the location.

These results indicate that cockle sampled from some sites in each location have similar characteristic with cockle obtained from another location.

In summary, the two rivers studied seemed to be polluted and this is in agreement with previous studies at Juru River, Malaysia on metal accumulation on aquaculture (Yahya, 1994, DANCED, 1998). According to report by DANCED (DANCED, 1998) the electroplating, pulp and paper, textiles, food and beverages and auto-workshops industries were closely linked to industrial pollution in the Prai industrial area. This is in line with the report conducted by department of environment (DOE, 1999) that the four predominant industries of Penang are electronics/electrical, textiles, fabricated metal products, plastic and plastic products. Other industries include paper and paper products/printing works, rubber based, chemical/fertilizers and basic metal industries. Illegal mud-dumping activities of industrial waste could be another reason for the river pollution.

In general, some sites in both locations receive pollution from similar sources.

4. Comparison between Logistic regression model and discriminant analysis (DA)

The data was analyzed using discriminant analysis (DA) (Abbas, et al., 2007). The results of DA showed that 72.5% of original cases were correctly classified to their respective group (Juru river 70%, fourteen cases were correctly classified under Juru River and 75% Jejawi River, fifteen cases were correctly classified under Jejawi River). DA also showed that Zn and Cd exhibited strong evidence in discriminating the two locations and account for most of the expected variations in arsenic and heavy metal contents, while other parameters showed less contribution in explaining the variation between the two locations.

The results of logistic regression model showed that 77.5% of the cases were correctly classified to their respective groups (Juru river 75%, fifteen cases were correctly classified under Juru River and Jejawi River 80%, sixteen cases were correctly classified under Jejawi River). Logistic regression model identified only two parameters Zn and Cd responsible in discriminating the two locations. Comparing the results obtained from logistic regression model and discriminant analysis indicate that the two techniques gave almost the same percentage of correct classification and identified the same parameters responsible in discriminating the two locations.

In general, Logistic regression model and DA gave the best results in distinguishing the two locations and their results were almost similar. Both techniques also indicated that some cockles have similar characteristic regardless of the location.

5. Conclusion

Logistic regression has proved to be an efficient tool for source identification of arsenic and heavy metals in cockles. Logistic regression model showed that only two parameters Zn and Cd responsible for distinguishing the cockles affording 77.5 % correct assignment. The results of logistic regression model were close to DA results and either one can be used. This paper also presented some evidence of the bioaccumulation of As and heavy metals in cockles reared in the two rivers.

Acknowledgments

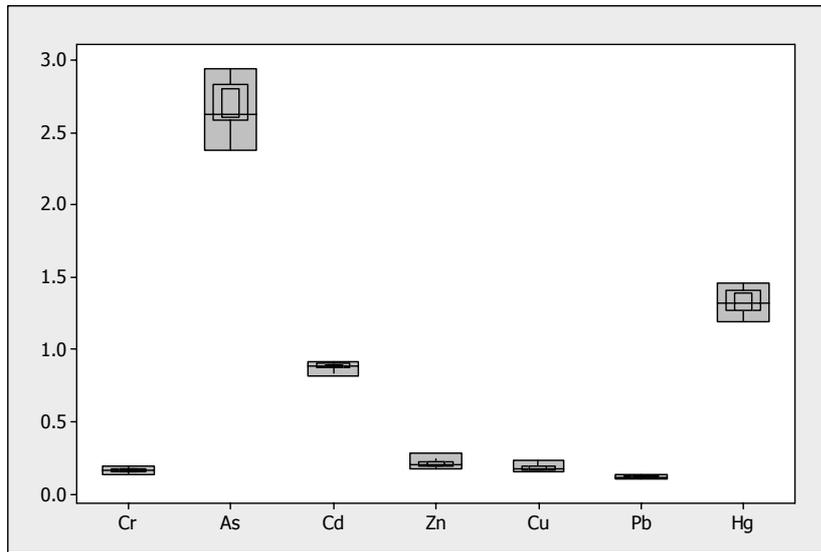
The study was funded through USM short term grant (internal research grant) grant number PTEKIND/ 636054.

References

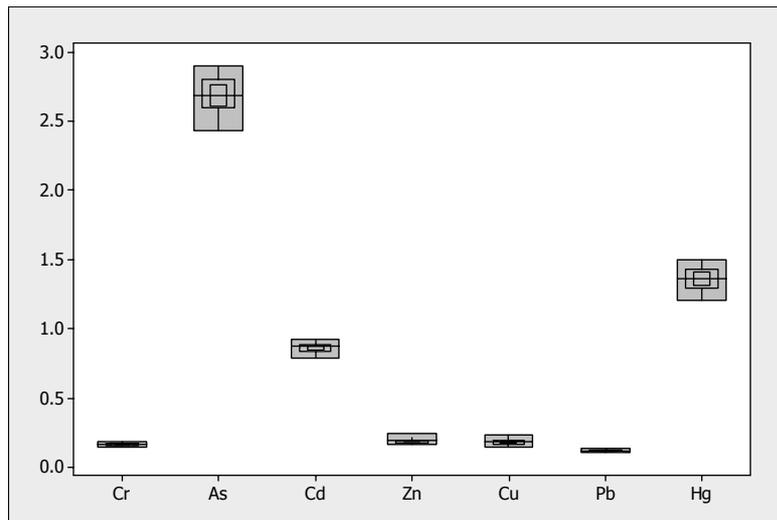
- Abbas F.M. A., Norli, I., Azhar, M. E. (2008). Assessment of arsenic and heavy metal contents in cockles (*Anadara granosa*) using multivariate statistical techniques. *Journal of hazardous material*, 150, 783–789.
- Agresti, A. (2002). *Categorical data analysis*, John Wiley&Sons, Inc., USA.
- Ahmet, D., Fevzi Y., A. L. Tuna, Nedim O. (2006). Heavy metals in Water, sediment and tissues of *Leuciscus cephalus* from a stream in southwestern Turkey. *Chemosphere*, 63, 1451-1458.
- Canivet, V., and Gibert J. (2002). Sensitivity of epigeal and hypogean freshwater macroinvertebrates to complex mixtures, Part I: Laboratory experiments. *Chemosphere*, 46, 999-1009.
- Collett, D. R. (2003). *Modeling Binary data*, London, Chapman&Hall.
- Cox, D. R., Snell, E.J. (1994). *Analysis of binary data*. Chapman&Hall, London.
- DANCED (1998). Penang coastal profile. Integrated Coastal Zone Management Project Report, Penang State Government, Malaysia.
- Department of Environment (DOE) (1999). Annual Report, Ministry of Science and Technology, Malaysia.
- Erling, B. A. (1997). *Introduction to the statistical analysis of categorical data*. Springer- Verlag Berlin, Germany.
- Hernandez, O.M., Fraga J.M.G., Jimenez A.I., Jimenez F., Arias J.J. (2005). Characterization of honey from the Canary Islands: determination of the mineral content by atomic absorption spectrophotometry. *Food Chemistry*, 93, 449-458.
- Smolders, R., De Coen W., Blust R. (2004). An ecologically relevant exposure assessment for a polluted river using an integrated multivariate PLS approach. *Environmental Pollution*, 132, 245-263.
- Socio-economic and Environmental research Institute Report (2002). Economic briefing to the Penang State Government. P 8, 4(10).
- Vallvey, L.F., Guerrero, G.E., Fern, M.D., Cuadros, R. L. (2006). Logit linearization of analytical response curves in optical disposable sensors based on coextraction for monovalent anions. *Analytica Chimica Acta*, 561, 156–163.
- Yahya, M.N., and Zubir, D. (1994). The effect of metals accumulation on aquaculture from Juru River. Research Report.
- Yujing, C., Young-Guan, Z., Rihong, Z., Yizhong, H., Yi Q., Jianzhong, L. (2005). Exposure to metal mixtures and human health impacts in a contaminated area in Nanning, China. *Environmental International*, 31, 784-790.

Table 1. The result of the classification using Logistic regression model

		Predicted location		%
		Juru	Jejawi	
Observed location	Juru	15	5	75
	Jejawi	4	16	80
Overall percentage				77.5



a- Juru River



b- Jejawi River

Figure 1. Box-plot of arsenic and heavy metals along a- Juru River b- Jejawi River