J. Appl. Environ. Biol. Sci., 6(7)80-84, 2016
© 2016, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Classification Accuracy and Factors Affecting Diphtheria Incident Using Multivariate Adaptive Regression Spline (Mars)

Zuryaty[1*], Kuntoro[1], Windu Purnomo[1], Bambang Widjanarko Otok[2]

[1] Faculty of Public Health, Airlangga University in Surabaya (INDONESIA)
[2] Department of Statistic, Sepuluh Nopember Institute of Technology (ITS), Surabaya (INDONESIA)

Received: April 3, 2016
Accepted: May 30, 2016

ABSTRACT

Diphtheria is one of the infectious diseases that are induced by bacterial infection. Bangkalan District has the second largest Diphtheria case after the city of Surabaya, namely as many as 76 cases with 4 cases died. The number of cases Diphtheria in Bangkalan tend to increase each year. Therefore observational research was carried out in 2015. Observations were performed on data holders program 2015 diphtheria cases in each sub-district in Bangkalan. The purpose of this research is to examine the factors that affect the incidentDiphtheria. The results of the study showed that with the approach of MARS through the criteria GCV and the accuracy of classification, Diphtheria influenced by the mobility behavior, shelter density and contact. There is no sick Diphtheria opportunities with high mobility of 0.622, while no sick Diphtheria opportunities with eligible shelter density and mobility medium or high and bad behavior of 0.647, and opportunities of no sick Diphtheria if no contacts of 0.530. Classification of accuracy of the Diphtheria incident is 78.7 percent.
KEY WORDS: Diphtheria Incident, Bangkalan, MARS, GCV, Classification Accuracy

1. INTRODUCTION

Diphtheria is one of the infectious diseases that are induced by bacterial infections *Corynebacterium* namely Diphtheria germs that infect the respiratory tract, especially the tonsils nasofaring (part between the nose and pharynx or throat) and laryngeal [1]. The germs can cause damage to the network in some of the organs in the body causing miocarditis and cardiac that cause death [2].

In Indonesia 2012 Diphtheria spread in 19 provinces. The number of cases candidates in East Java Province compared to other provinces with the number of cases as much as 955 cases [2]. According to the ministry of health (2013) if found one casesDiphtheria in hospital and the community health clinic, then the region is revealed as an extraordinary event (KLB) [3]. In 2011, 2012, and 2013 the entire city district Bangkalan has deadly diseases Diphtheria. Incident Diphtheria begins at the extraordinary (KLB) in the sub-district of Tanah Merah Bangkalan 2005, which cause the spread of the wider to district or other cities [3]. In 2013 Bangkalan district has the second largest Diphtheria case after the city of Surabaya, namely as many as 76 cases with 4 cases died. The number of cases Diphtheria in Bangkalan tend to increase each year [2].

Therefore, relief efforts is expected to stop KLB Diphtheria and not a problem in the future will come. Prevention efforts are done with mass immunization activities and also with the retrospective search for risk factors that cause outbreaks of disease.

Direct contact with the sick person Diphtheria can become carrier. According to[2], region with social and cultural conditions that specifically related to the extraordinary events KLB) influenced by factor population mobility. According to [1], poverty become the eyes of the chain difficult disconnected because of the ability to buy them to meet the needs of the food affect the consumption patterns, access to health services is not adequate education and behavior that less healthy that resulted in a person vulnerable to infectious diseases. To see how big the opportunity to the accuracy in Diphtheriaincident classification can be done with the approach of classification of MARS with the two categories are not sick and sick Diphtheria [4][5][6].

2. RESEARCH METHODOLOGY

The type of building blocks used in this study is *Non Reactive*, because this study is used to refer to the data collected by the way not directly involved to get the information from the subject of the research [7]. The information in the form of a document/ results of annual routine report Bangkalan District health office. This research done by cohort studies on 2015 [2]. The observation done on the case program holder data Diphtheria 2015 on each district in Bangkalan Regency. The variables used in this research consists of: incidentDiphtheria dependent variables and independent variables: behavior, mobility, contact, the source of infection and the density of shelter [3][8].

***Corresponding Author:** Zuryaty, Faculty of Public Health, Airlangga University in Surabaya (INDONESIA).
emails: zur_ya_ty@yahoo.co.id

The first step, done analysis of the descriptive statistics of the variables veriabel perdiktor. Second, to know what factors that affect theDiphtheria incident, made the following procedure [4][5][9][10]

1. The formation of the model MARS for *data sets the* beginning: determine BF; determine MI; determine MO in between the knot.
2. Get the best of MARS model for a single dataset based on the value of the smallest GCV.
3. Get the variables significant effect from the model MARS best for single dataset.

MARS introduced in 1991 by Friedman [9]. MARSmodel useful to resolve the issue of the data dimension high and produce accurate response variable prediction, and produce a continuous in the *knot model* based on the value of the smallest GCV [9][11][12].

The general Model *ARS equation* can be written as follows [11]:

$$\boldsymbol{f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} [s_{km} \cdot (x_{v(k,m)} - t_{km})]_+ + \varepsilon_i}$$

Where,

$\boldsymbol{\alpha_0}$ = function of the parent base (*constant basis function*),
$\boldsymbol{\alpha_m}$ = drag coefficient from the base function to-*m,*
M = maximum base function (*non-constant basis function*),
K_m = degrees of interaction,
$\boldsymbol{s_{km}}$ = worth 1 if data is located to the right of the *knot point*, or worth -1 if data is located at the left of the *knot point*,
$\boldsymbol{x_{v(k,m)}}$ = variables predictors, and
$\boldsymbol{t_{km}}$ = the value of the *knot* from the variables predictors $x_{v(k,m)}$.

MARS can find the layout and number of knot that is required in a step forward / customer whether they stepwise [4]. Forward stepwise done to get the function with the number of maximum base function. The criteria for the selection of the base on the forward stepwise function is to minimize Average Sum of Square Residual (ASSR). Meanwhile, to fulfill the concept of the customer whether they stepwise parsimoni done, aims to select the function of the base is produced from the forward stepwise with minimize GCV value [10].

Determine the model of MARS is optimal among the other models is by selecting the model that has the lowest GCV value. GCV criteria introduced by Wahba in 1979 [11]. The minimum GCV function is defined as:

$$GCV^*(M) = \frac{ASR}{[1 - \frac{C(M)^*}{n}]^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_K(x_i)]^2}{[1 - \frac{C(M)^*}{n}]^2}$$

Where:

M =number of basis functions (*nonconstant* function base) determined on stage *forward*
x_i = variables predictors
y_i = response variable
$C(M)^* = C(M) + dM$
$C(M)$ = Trace $\mathbf{[B(B^TB)^{-1}B^T]}+1$
d = the value of the base function optimization reached ($2 \leq d \leq 4$)

3. RESULTS AND DISCUSSION

The Diphtheria incident relationship with independent variables are presented in the following table 1.

Table 1 shows that the IncidentDiphtheria there is a relationship with the behavior, mobility and contact with the level of the significance of 5 percent, while with the source of infection and there was no shelter density relation. Then the modeling of MARS, but all independent variables included in the modeling. This is due to the approach of MARS is nonparametrik approach.

The procedure for the establishment of the best model obtained with how to compare GCV value and the value of classification accuracy [5][11]. The value of the smallest GCV and high classification accuracy value is the best model. The compared done on various number of basis functions (BF = 10, 15 and 20), and various maximum interaction (MI = 1, 2 and 3). In detail the best model selection is presented in Table 2.

Table 1. The Independency Test of Diphtheria Incident with independent variables

Frequency Percentage of Total Pearson Chi-Square		Diphtheria Incident	
		Not Sick	Sick
Behavior	1: Good	30 27.8	4 3.7
	2: Bad	42 38.9	32 29.6
	Pearson Chi-Square = 10.388 df=1 Asymptotic Significance (2-sided) = 0.001Chi-Square Tabel = $\chi^2_{(0.05;1)} = 3.841$		
Mobility	1: Low	34 31.5	7 6.5
	2: Moderate	30 27.8	11 10.2
	3. High	8 7.4	18 16.7
	Pearson Chi-Square = 20.735 df=2 Asymptotic Significance (2-sided) = 0.000Chi-Square Tabel = $\chi^2_{(0.05;2)} = 5.991$		
Contact	1: No contacts	53 49.1	19 17.6
	2: Contacts	19 17.6	8 15.7
	Pearson Chi-Square = 4.688 df=1 Asymptotic Significance (2-sided) = 0.030 Chi-Square Tabel = $\chi^2_{(0.05;1)} = 3.841$		
Source Infection	1: No	36 33.3	19 17.6
	2: Yes	36 33.3	17 15.7
	Pearson Chi-Square = 0.074 df=1 Asymptotic Significance (2-sided) = 0.785 Chi-Square Tabel = $\chi^2_{(0.05;1)} = 3.841$		
Density Shelter	1: Qualify	62 57.4	32 29.6
	2: Not Qualify	10 9.3	4 3.7
	Pearson Chi-Square = 0.164 df=1 Asymptotic Significance (2-sided) = 0.685Chi-Square Tabel = $\chi^2_{(0.05;1)} = 3.841$		

Table 2. The Best Model Diphtheria Incident Various BF, MI, MO Based GCV and Classification Accuracy

No	BF	MI	MO	GCV	Classification Accuracy	Prediction Model
1	10	3	0	0.178	0.787	Y = 0.273 + 0.224 * BF1 + 0.335 * BF7 - 0.154 * BF9 BF1 = (MOBILITY = 3); BF4 = (BEHAVIOR = 2); BF6 = (MOBILITY=2OR MOBILITY=3*BF4; BF7 = (DENSITY SHELTER = 1) * BF6; BF9 = (CONTACT = 1)
2	15	3	3	0.186	0.787	Y = 0.317 + 0.259 * BF1 + 0.195 * BF7 - 0.344 * BF9 + 0.329 * BF11 – 0.367 * BF13 + 0.208 * BF15 BF1 = (MOBILITY = 3); BF4 = (BEHAVIOR = 2); BF6 = (MOBILITY=2 OR MOBILITY=3)* BF4; BF7 = (DENSITY SHELTER = 1) * BF6; BF9 = (CONTACT = 1); BF11 = (SOURCE INFECTION = 1); BF13 = (BEHAVIOR = 1) * BF11; BF15 = (MOBILITY = 2) * BF13
3	20	3	3	0.186	0.787	Y = 0.292 + 0.198 * BF1 + 0.177 * BF7 - 0.314 * BF9 + 0.358 * BF11 – 0.386 * BF13 + 0.192 * BF15 + 0.168 * BF17 BF1 = (MOBILITY = 3); BF4 = (BEHAVIOR = 2); BF6 = (MOBILITY=2 OR MOBILITY=3)* BF4; BF7 = (DENSITY SHELTER = 1) * BF6; BF9 = (CONTACT = 1); BF10 = (CONTACT = 2); BF11 = (SOURCE INFECTION = 1); BF13 = (BEHAVIOR = 1) * BF11; BF15 = (MOBILITY = 2) * BF13; BF17 = (MOBILITY = 3) * BF10

Table 2 shows the best model with the smallest GCV criteria and the accuracy of the classification of the biggest all models, involving the number of basis functions BF = 10, MI = 3, and MO = 0 with the value of 0.178 GCV and classification accuracy = 0.787. The MARS Model best obtained revealed in the equation as follows.

Y = 0.273 + 0.224 * BF1 + 0.335 * BF7 - 0.154 * BF9

with

BF1 = (MOBILITY = 3);
BF4 = (BEHAVIOR = 2);
BF6 = (MOBILITY=2OR MOBILITY=3)*BF4;
BF7 = (DENSITY SHELTER = 1) * BF6;
BF9 = (CONTACT = 1)

The interpretation of the model of Mars is written on the common (5.11) is as follows.

- BF1 = (MOBILITY = 3)

This means that the coefficient BF3 will mean if respondents have high mobility so every increase of one basis functions (BF3) can raise the value of 0.224 Diphtheria Incident and opportunities respondents not sick Diphtheria is

$$\text{P(not sickDiphtheria)} = \frac{\exp(g(x))}{1+\exp(g(x))} = \frac{\exp(0.273+0.224)}{1+\exp(0.273+0.224)} = 0.622$$

This means that the respondents with high mobility has not sick Diphtheria opportunities of 0.622.

- BF7 = (DENSITY SHELTER = 1) * BF6, BF6 = (MOBILITY=2OR MOBILITY=3*BF4, BF4 = (BEHAVIOR = 2)

This means that the coefficient BF4 will mean if respondents with eligible shelter density and mobility medium or high and bad behavior so every increase of one basis functions (BF7) can raise the value of 0.335 Diphtheria Incident and opportunities respondents not sick Diphtheria is

$$\text{P(not sick Diphtheria)} = \frac{\exp(g(x))}{1+\exp(g(x))} = \frac{\exp(0.273+0.335)}{1+\exp(0.273+0.335)} = 0.647$$

This means that the respondents with eligible shelter density and mobility medium or high and bad behavior have not sick Diphtheria opportunities of 0.647.

- BF9 = (CONTACT = 1)

This means that the coefficient BF9 will mean if respondents no contact so every increase of one basis functions (BF9) can decrease the value of 0,154 Diphtheria events and opportunities respondents not sick Diphtheria is

$$\text{P(not sick Diphtheria)} = \frac{\exp(g(x))}{1+\exp(g(x))} = \frac{\exp(0.273-0.154)}{1+\exp(0.273-0.154)} = 0.530$$

This means that the respondents no contact have the opportunity not sick Diphtheria of 0.530.

Table of classification is another interesting way to reveal the feasibility of a model that is how big the percentage of observations correctly classified. The results of this table in the form of cross-classification of response variables with dikotomus scale as shown in Table 3.

Table 3. The classification accuracy of DiphtheriaIncident Using MARS

DiphtheriaIncident		**PredictionClass**		**Percentage Classification Accuracy**
		0: not sick	**1: sick**	
Actual class	0: not sick	65	7	90.3
	1: sick	16	20	55.6
Percentage of Total Classification Accuracy				78.7

Table 3 it is known that the percentage of the entire observation terklasifikasikan properly is 78.7 percent so great misclassification (APER) is 21.3%. This misclassificationvalue is not too large so that it can be concluded that the model of MARS good enough to classify the observation with the percentage not sick Diphtheria (sensitivity) of 90.3 percent and the percentage of sick Diphtheria (specificity) by 55.6 percent.

The importance level of each variable was assessed by the rise of the value of GCV. Thus the importance level of each variable has the role to minimize the value of GCV in the model. Now the importance level of each of the variables of the model of the finances on the Table 4.

Table 4. The importance level of the variables in the MARS

The variables	Importance	-GCV
Mobility	100.000	0.202
Behavior	64.713	0.180
Density Shelter	48.062	0.173
Contact	30.142	0.168
Source Infection	0.000	0.165

Table 4 can be known that the mobility variable is the most important variable on the model of the MARS incidentDiphtheria with importance level reached 100 percent Furthermore behavior variable reached 64.713 percent, density shelter 48.062 percent, and contact 30.142 percent. Minus GCV values shows that when a variables are removed from the model and GCV value will increase by +GCV on the variable. So the greater the importance level of the variable will be followed by the value minus GCV great and will minimize GCV values in the model are formed.

4. CONCLUSION

There is a relationship between Diphtheriaincident and the behavior, mobility and contact. Factors that affect the Diphtheria incident is mobility, behavior, shelter density and contact. There is no sick Diphtheria opportunities with high mobility of 0.622, while no sick Diphtheria opportunities with eligible shelter density and mobility medium or high and bad behavior of 0.647, and opportunities of no sick Diphtheria if no contacts of 0.530. This model provides the value of the accuracy of the classification of 78.7 percent, the sensitivity of 90.3 percent and specificity by 55.6 percent.

REFERENCES

[1] Kementrian Kesehatan Republik Indonesia. (2013). *Petunjuk Teknis Pelaksanaan Imunisasi Dalam Rangka Penanggulangan Kejadian Luar Biasa (KLB) Difteri.* Jakarta

[2] Dinas Kesehatan Kabupaten Bangkalan. (2014). *Profil Kesehatan Kabupaten Bangkalan Tahun 2014*. Bangkalan

[3] Bres, P. (1995). *Tindakan Darurat Kesehatan Mayarakat Pada Kejadian LuarBiasa.* Universitas Gadjah Mada. Yogyakarta

[4] Otok, B.W., M. Sjahid Akbar, Suryo Guritno dan Subanar. (2007), Ordinal Regression Model using Bootstrap Approach**,** *Jurnal ILMU DASAR, Vol. 8 No. 1, 2007 : 54-67*, UNEJ – Jember.

[5] Otok, B.W., Guritno, S., Subanar, Haryatmi, S.(2006). Bootstrap dalam MARS untuk Klasifikasi Perbankan. *Inferensi Jurnal Statistik*, Volume 2, N0. 1, Januari 2006. FMIPA ITS Surabaya.

[6] Agresti. (2002). *Categorical Data Analysis,* John Wiley & Sons, Hoboken New Jersey

[7] Kuntoro. (2011). *Metode Statistik*. Surabaya: Pustaka Melati

[8] Nailul. (2012). *Analisis Penyakit Difteri Menggunakan Data Spasial*. Tesis. Surabaya: Unair

[9] Friedman, J.H. (1991). Multivariate Adaptive Regression Splines*. The Annals of Statistics*, Vol.19 No.1

[10] Friedman, J.H. and Silverman, B.W. (1989). Flexible Parsimony Smoothing and Additive Modelling. *Technometrics*, 31.

[11] Otok, B.W. (2008). Bootstrap Pada Pemodelan Multivariat Adaptive Regression Spline. *Disertasi UGM*. Yogyakarta. (tidak dipublikasikan)

[12] Jusuf, H. and Bambang W. O. (2014). Optimal input of data series in predicted the number patient of HIV-AIDS in East Java Province Using Multivariate Adaptive Regression Splines. *Journal of Health Sport and Tourism 2014; 5(1), 35-42.* ***DOI:*** *10.7813/jhst.2014/5-1/6*