Saudi Journal of Biological Sciences (2016) 23, S11–S15

King Saud University

Saudi Journal of Biological Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

A new mathematical evaluation of smoking problem based of algebraic statistical method

Maysaa J. Mohammed [a], Isamiddin S. Rakhimov [a], Mahendran Shitan [b], Rabha W. Ibrahim [c,*], Nadia F. Mohammed [a]

[a] Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Malaysia
[b] Institute for Mathematical Research (INSPEM), Malaysia
[c] Faculty of Computer Science and Information Technology, University Malaya, Malaysia

Received 22 April 2015; revised 19 July 2015; accepted 25 August 2015
Available online 1 September 2015

KEYWORDS

Markov basis;
Contingency tables;
Entropy

Abstract Smoking problem is considered as one of the hot topics for many years. In spite of overpowering facts about the dangers, smoking is still a bad habit widely spread and socially accepted. Many people start smoking during their gymnasium period. The discovery of the dangers of smoking gave a warning sign of danger for individuals. There are different statistical methods used to analyze the dangers of smoking. In this study, we apply an algebraic statistical method to analyze and classify real data using Markov basis for the independent model on the contingency table. Results show that the Markov basis based classification is able to distinguish different date elements. Moreover, we check our proposed method via information theory by utilizing the Shannon formula to illustrate which one of these alternative tables is the best in term of independent.

1. Introduction

US Department of Health and Human Services (2014) showed that smoking is responsible for roughly 440,000 deaths each year in the USA. Studies about the effects of smoking were done only on men in 1964 but later on women were included as well when some cigarette companies stated that smoking keeps girls and women thin. It can be noticed that health issues caused by smoking, no longer take into consideration, although most of people are familiar with the smoking consequences. Women who smoke significantly increase the danger of developing heart disease (the leading killer among women) and stroke. The danger increases with the number of cigarettes smoked and the length of time a woman has been smoking, but even people who smoke less than 5 cigarettes a day can have heart and blood vessel diseases. Although most of the women who die of heart disease are past menopause, smoking increases the danger additionally in younger women than in older women. Studies propose that smoking cigarettes increase the danger of heart disease even more among younger women who also take attractive birth control pills. Women who smoke, exclusively after going through menopause, have lower

* Corresponding author.
E-mail address: rahaibrahim@yahoo.com (R.W. Ibrahim).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.sjbs.2015.08.015
1319-562X

bone density (thinner bones). This means they have a higher danger of broken bones, counting hip fracture, than women who do not smoke. They could likewise be at higher danger of receiving rheumatoid arthritis and cataracts (clouding of the lenses of the eyes), as well as age-related macular degeneration, which can cause blindness. Tobacco habit can harm a woman's reproductive health. Women who smoke are more likely to have trouble getting pregnant. Smokers incline to be younger at the start of menopause than non-smokers and may have more unpleasant symptoms while going through menopause.

Nearly most of the people know how harmful it is being addictive smokers, however, few of them really recognize its risks. Some smokers convince themselves that they are social smokers or smoke only outdoors to be distinguished as non-smokers (Harris et al., 2002), and they can control smoking since smoking is not a habit for them. Stopping smoking is not that easy, yet being addictive does not mean a smoker does not have the ability to quit. The process of giving up smoking gets complicated, especially when people get older and older and the condition becomes sometimes irreversible. Some succeed from the early try while others give out more than one try; notwithstanding, some fail to quit smoking (Rollins et al., 2002). Moran et al. (2004) perceived that smokers start having a cigarette in occasional cases which means not daily and by the time they grow up, it gets irregular; then being involved in being motiveless to giving up; finally they reach a higher level (Moran et al., 2004). Due to several negative effects associated with this problem, many of the studies and research have emerged to study this problem in many respects, as in the above studies. This led to the attention of many researchers of the problem and the surrounding circumstances.

In this paper, we impose the algebraic statistical method using Markov basis for the independent model to find alternative data models that simulate the original data models and maintain the same statistical and mathematical qualities of the original data. On the other hand, our method is to find alternative data tables depending on the original data by using Markov basis for independent model. Moreover, we choose the best table in term of independent that liaises the qualities and characteristics is the same of the original data and the value of information on this table is more than the original data by using (information theory) as a measure for the independent. Due to the importance of passive smoking and its impact on all groups of society, we have selected data related to clarify the relationship between sex and smoking status (see Table 1).

Table 1 Contingency table for relationship between sex and smoking status.

Smoking status	Sex		
	Man	Woman	Total
Non smoker	10	6	16
Light smoker	8	8	16
Heavy smoker	4	8	12
Total	22	22	44

2. Material and methods

In 2007, Hannelore (2007) used Breathing test model. The data disquiet the relations among smoking position and breathing test outcomes for employers (under 40 years old) in a certain industrial plant. His study can be extended by presenting a further variable, namely the age. Observed are not the only employers under 40 years, but likewise employers from the age group 40–59. There are no restraints on the rows and column totals, and a simple model is that the count in the (i, j) cell y_{ij} is an understanding of a Poisson variable with expectation μ. The consequential likelihood is

$$L(\mu) = \prod_{i,j} \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \exp(-\mu_{ij}) \tag{1}$$

where μ is the vector of the expectations.

In 2013, Jolanta et al. (2013) imposed the relationship between Peripheral vascular and smoking, and that smoking contributes significantly to the increase of cardiovascular disease and may increase the risk of holding such kinds of diseases by using the prevalence ratio calculation for cross section study.

$$\text{Prevalence} = \frac{\text{All new and preexisting cases during a given time period}}{\text{population during the same time period}} \times 10^n$$

The odds ratio (OR) is calculated as:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc} \tag{2}$$

In 2014, Suzan et al. (2014) found the relationship between the effect of smoking and disease, incontinence and the impact of smoking with the disease using hypothesis testing depending on the Fisher exact test. The study has shown that smoking has to do with the impact of a negative increase in case of this disease. The P-value calculated by formula of Fisher's exact test as:

$$P\frac{\left(\binom{a+b}{a}\binom{c+d}{c}\right)}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{3}$$

In 2015, Anthony et al. (1999) studied that smoking during pregnancy has harmful effects and a large impact on the mother's weight at birth and also has significant side effects in children which may cause a respiratory problem and this leads to the lungs not working well at birth. Thus, in their studies, they have selected a sample of mothers either directly during a hospital admission or within 24 h of birth. The sample was taken from 100 mothers of non-smokers and 100 mothers who smoked 1–10 cigarettes per day and 100 mothers who smoked an average of more than 10 cigarettes per day. The aim of the study is to investigate the effects of smoking on lung function in newborns by knowing whether smoking during pregnancy affects or has to do with the functions of the lung in newborns. Using test hypotheses in statistics depending on chi square test. The formula of chi square test is:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{4}$$

Table 2 Entropy values for some alternative tables $(A_2, \ldots A_{10})$, with original matrix A_1.

Table(CT)	Entropy
A_1	0.7627409102
A_2	0.7551117264
A_3	0.765254924
A_4	0.7527180692
A_5	0.7665592067
A_6	0.7475626142
A_7	0.7715469493
A_8	0.7421311053
A_9	0.7564160091
A_{10}	0.7370474462

where $O_{i,}$ stands for observed frequencies, E_i, stands for expected frequencies, and i, runs from 1, 2, …, n where n is the number of the cells in the contingency table.

2.1. Proposed method

In this section, we illustrate our proposed method, which was based on Markov basis for the independent model.

Step 1: Organize data as in Table 1, (Polit et al., 2013), thus we need a table to represent the relationship between smoking and sex.

Step 2: We test the data for the independent model using hypothesis tests with the chi-square test (Minhaz, 2007) which are as:

H_0: suppose the relationship between smoking status and sex is independent.

H_1: The relationship between smoking status and sex is not independent. Degree of freedom $D_f = (r-1)(c-1)$, where r is the sum of rows and c is the sum of columns. Hence, $D_f = 2$. Then, we use the formula of chi-square test as above in (4). Therefore, we use this method if the expected frequency is equal or more than 5 where the expected frequency $E_{r,c} = (n_r \times n_c)/n$. Moreover, if the expected frequency is less than 5, we use hypothesis test depending on loglinear model with odds ratio test (Carol, 2001). Hence, the general formula of the odds ratio is: $\frac{P_{ij}P_{kl}}{P_{il}P_{kj}} = 1$, where $1 \leqslant i < k \leqslant r$ and $1 \leqslant j < l \leqslant c$. According to this method, then, the hypothesis test as:

Table 3 Contingency table A_7 for the best relationship between sex and smoking status.

Smoking status	Sex		
	Man	Woman	Total
Non smoker	9	7	16
Light smoker	8	8	16
Heavy smoker	5	7	12
Total	22	22	44

H_0: suppose the relationship between smoking and sex is independent if the odds ratio $(OR) = 1$, then, $log(OR) = 0$.

H_1: If the $(OR) \neq 1$, then, $log(OR) \neq 0$, hence the relationship between sex and smoking status is not independent.

Step 3: According to Table 1, we construct Markov basis elements by using

$$Z_{ij} = \begin{cases} +1 & (i,j) = (i_1,j_1),(i_2,j_2) \\ -1 & (i,j) = (i_1,j_2),(i_2,j_1) \\ 0 & \text{otherwise} \end{cases}$$

where the integer matrix $Z = Z(i_1, i_2, j_1, j_2) = \{Z_{ij}\}$

Step 4: $B = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ with basic moves (see Satoshi et al., 2012; Mathias et al., 2009) such as:

$$e_1 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 0 & 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 & 0 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad e_3 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 0 & 0 \end{bmatrix},$$

$$e_4 = \begin{bmatrix} 0 & 0 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}, \quad e_5 = \begin{bmatrix} 1 & -1 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}, \quad e_6 = \begin{bmatrix} -1 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}$$

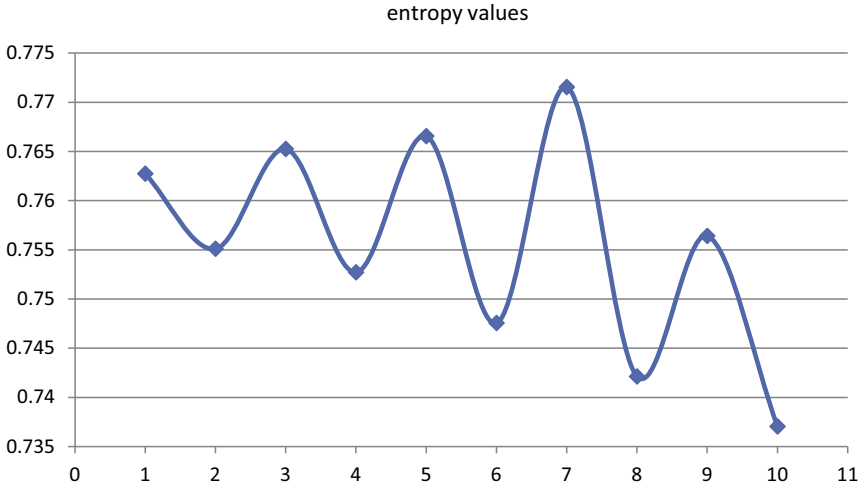

Figure 1 The entropy values for alternative tables $(A_2, \ldots A_{10})$, with original table A_1.

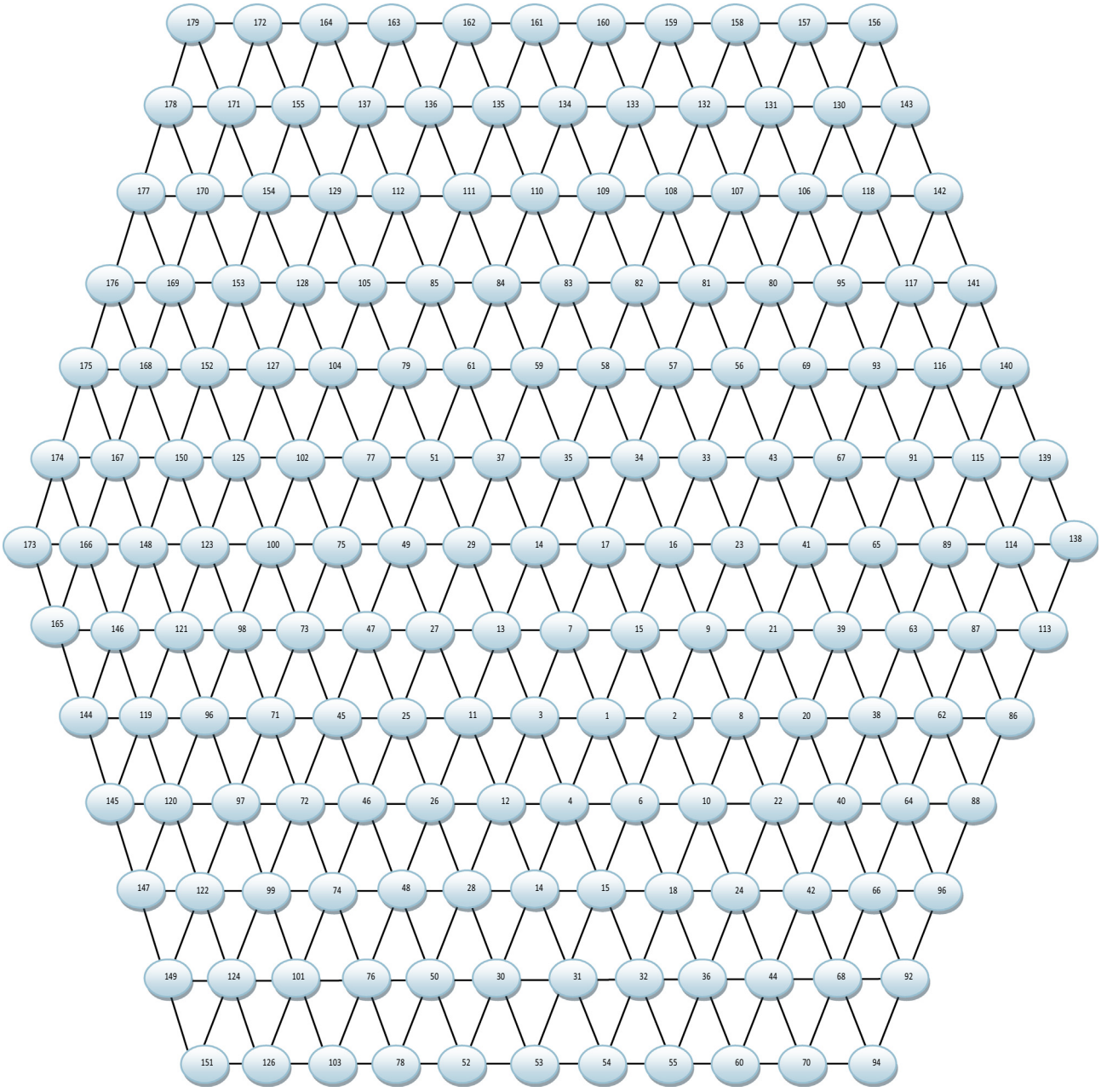

Figure 2 All the elements of fiber with original data (CT construction).

Step 5: In order to obtain alternative data, we add the moves in the fourth step to the original data in Table 1. Repeat this process with each alternate data and then stop, when contingency tables are equal. Fig. 2 shows all the construction of these tables.

Step 6: In order to investigate the original data in terms of the proposed method, we apply information theory by utilizing the Shannon formula (Hnizdo and Gilson, 2010)

$$H(x) = -\sum_{i=1}^{N} p(x_i) log_a(p(x_i)).$$

where the base (a) is a ten and the unit is hartley (Hnizdo and Gilson, 2010; Baez et al., 2011), $p_i(i = 1, 2, \ldots\ldots, N)$. Satisfying $\sum_{i=1}^{N} p_i = 1$.

In addition, we consider that $p_i log_a p_i = 0$ as $p_i = 0$.

3. Results

The main objective of this study is to analyze the original data and assist researchers in finding alternative data bearing specifications and characteristics that are very close to the original data. Therefore, we started the selection of real data on smoking which was a choice of a sample consisting of 44 people, divided by gender and ability to smoke. Then, we test this sample, is it independent or not using the test hypotheses in chi-square test or loglinear model test which depends on the expected value, so if the expected value is more than or equal to 5 then we choose either one of these methods. If not, we

choose the second one for loglinear model of statistical independence. Thus, here the expected value is more than 5. Therefore, we can test by any one of these methods. We see the result of $x_c^2 = 2.3333333334$ and degree of freedom is 2, therefore $x_t^2 = 5.991$, $\alpha = 0.05$ (Minhaz, 2007). Because of $x_c^2 < x_t^2$, the relationship between sex and smoking is independent. After that, we installed fiber that has calculated the Markov basis elements. We add these items to the original data and the result we got 178 of alternative schedules of the original schedule start of $\{A_2 \ldots\ldots\ldots.., A_{179}\}$. Consequently, we applied the Entropy theory of Shannon, to find which one of these alternative contingency tables (CT) is closest or better than the original data table in terms of independence (Tom, 2011). This study is the first one applying Markov basis for the independent model with Shannon information theory. Hence, we hope that it will help all the researchers to find the best alternative data table in terms of independent of their studying to work with it instead of the basic data.

4. Discussion

In this study, we applied the Entropy theory by Shannon for all elements of the fiber on the Markov basis (MB). We find that, for a random sample of MB, say $\{A_{2, \ldots,} A_{10}\}$. We consider, the maximal entropy is in A_7 (see Table 3).

Therefore, the seventh table is the alternative better table with the characteristics and qualities of information and assembly from the original table. Finally, we have found that there is more than one alternative, as shown in the Fig. 1. The information gathering where more than the original table or informational specifications are considered better than the original table.

5. Conclusion

We conclude that our method, which is based on algebraic statistical using Markov basis for the independence model is effective for analyzing and classifying the original data by using the elements of Markov basis with basic moves and fix fiber. Therefore, we chose the real data represented by the relationship between smoking and sex see Table 1. Consequently, we tested it is independent by using chi-square test. Then, we added the elements of Markov basis to get the elements of the fiber, that means we get the alternative tables A_2, …, A_{10} . Moreover, we applied the Entropy theory of Shannon to find which one of these alternative data is closest or better than the original data table in terms of Independence see Table 2. Hence, this method is the first one apply a Markov basis for the independent model to the Entropy theory that helps researchers get the best result which is better than the original data table in terms of independent. We found that A_7 is the best alternative table. Note that there are various approaches for evaluation based on optimization, differential equations and other statistical methods (see Qureshi et al., 2015; Kayani et al., 2014; Noor et al., 2014; Tabassum et al., 2014).

References

Anthony, D., Michael, J., Dorothea, M., Grenville, F., Chakraphan, S., 1999. Effects of smoking in pregnancy on neonatal lung function. Arch. Dis. Child Fetal Neonatal Ed. 80, F8–F14, Downloaded from http://fn.bmj.com/ on March 15, 2015 - Published by group.bmj.com.

Baez, J.C., Fritz, T., Leinster, T., 2011. A characterization of entropy in terms of information loss. Entropy 13, 1945–1957.

Carolyn, J., 2001. Log-linear Models for Contingency Tables. University of Illinois at Urbana-Champaign.

Hannelore, L., 2007. Contingency Tables, Logit Models and Logistic Regression, Loglinear Models. University of Potsdam, February 20.

Harris, K., Wilson, T., Ahluwalia, J., 2002. A qualitative analysis of college students' smoking: Perceptions and interest in change. Poster Presented at: Annual Meeting of the Society for Research on Nicotine and Tobacco.

Hnizdo, V., Gilson, M.K., 2010. Thermodynamic and differential entropy under a change of variables. Entropy 12, 578–590.

Jolanta, D., Kęstutis, Ž., Aušra, B., 2013. Apsvarstė ir rekomendavo išleisti, Vilniaus universiteto Medicinos fakulteto taryba, 2012 m. rugsėjo 25 d., protokolas Nr. 2 (580).

Kayani, S. et al, 2014. Ethnobotanical uses of medicinal plants for respiratory disorders among the inhabitants of Gallies-Abbottabad, Northern Pakistan. J. Ethnopharmacol. http://dx.doi.org/10.1016/j.jep.2014.08.005.

Mathias, D., Bernd, S., Seth, S., 2009. Lectures on Algebraic Statistics. Springer Science & Business Media, Mathematics – pages 171.

Minhaz, F., 2007. CHI-Squared Test of Independence. University of Calgary, Alberta, Canada.

Moran, S., Wechsler, H., Rigotti, N.A., 2004. Social smoking among US college students. Pediatrics 114 (4), 1028–1034.

Noor, M.J. et al, 2014. Estimation of anticipated performance index and air pollution tolerance index and of vegetation around the marble industrial areas of Potwar region: bioindicators of plant pollution response. Environ. Geochem. Health. http://dx.doi.org/10.1007/s10653-014-9657-9.

Polit, Denise F., Cheryl, T., Lippincott, W., Wilkins, 2013. Essentials of Nursing Research: Apprising Evidence for Nursing Practice. Lippincott, Williams & Williams, Copyright 2014 Wolters Kluwer Health.

Qureshi, T. et al, 2015. Decontamination of ofloxacin: optimization of removal process onto sawdust using response surface methodology. Desalin. Water Treat. http://dx.doi.org/10.1080/19443994.2015.1006825.

Rollins, S., Malmstadt Schumacher, J., Ling, P., 2002. Exploring the phenomenon of social smoking – Why do so many young adults socially smoke. In: Abstract MEDI-161 Presented at the 2002 National Conference on Tobacco or Health, San Francisco.

Satoshi, A., Hisayuki, H., Hisayuki, H., 2012. Markov Bases in Algebraic Statistics. Springer, New York, Heidelberg, Dordrecht London. http://dx.doi.org/10.1007/978-1-4614-3719-2, ISSN 0172-7397. ISBN 978-1-4614-3718-5. ISBN 978-1-4614-3719-2 (eBook).

Suzan, A., Mohammad-Reza, S., Theophilus, O., Brendhan, G., Ananias, D., 2014. Stratification of clinical survey data using contingency tables. Int. J. Data Min. Knowl. Manage. Process (IJDKP) 4 (4).

Tabassum, N. et al, 2014. Chemodynamics of methyl parathion and ethyl parathion: adsorption models for sustainable agriculture. Biomed. Res. Int. http://dx.doi.org/10.1155/2014/831989. Article ID831989.

Tom, C., 2011. An introduction to information theory and entropy. http://astarte.csustan.edu/~tom/SFI-CSSS. Complex Systems Summer School.