Iranian J Publ Health, Vol. 41, No.6, Jun 2012, pp.58-64

Original Article

Classification of Death Rate due to Women's Cancers in Different Countries

*M Farhadian [1], *H Mahjub [2], A Moghimbeigi [1], J Poorolajal [2], GH Sadri [2]*

1. *Dept. of Biostatistics & Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*
2. *Research Center for Health Sciences, Department of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*

(Received 11 Jan 2012; accepted 22 Apr 2012)

Abstract

Background: The two most frequently diagnosed cancers among women worldwide are breast and cervical cancers. The objective of the present study was to classify the different countries based on the death rates from sex specific cancers.

Methods: In this cross-sectional study, we used dataset regarding death rate from breast, cervical, uterine, and ovarian cancers in 190 countries worldwide reported by World Health Organization. Normal mixture models were fitted with different numbers of components to these data. The model's parameters estimated using the EM algorithm. Then, appropriate number of components was determined and was selected the best-fit model using the BIC criteria. Next, model-based clustering was used to allocate the world countries into different clusters based on the distribution of women's cancers. The MIXMOD program using MATLAB software was used for data analysis.

Results: The best model selected with four components. Then, countries were allocated into four clusters including 43 (23%) in the first cluster, 28 (14%) in the second cluster, 75 (39%) in the third cluster, and 44 (24%) in the fourth cluster. Most countries in South America were to the first cluster. In addition, most countries in Africa, Central, and Southeast Asia were located to the third cluster. Furthermore, the fourth cluster consisted of Pacific continent, North America and European countries.

Conclusion: Considering the benefits of clustering based on normal mixture models, it seems that can be applied this method in wide variety of medical and public heath contexts.

Keywords: Neoplasm, Finite mixture models, Model-based clustering, BIC criteria

Introduction

Distribution of various types of disease in different populations varies based on associated factors such as social, cultural, racial, geographical, and nutritional characteristics. Cancer imposes a major disease burden worldwide, with variation among countries and regions. Around the world, the two most frequently diagnosed cancers among women are breast and cervical cancers. Breast cancer is the main leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths, which is more than double the second most common cervical cancer that made up 10 percent. Other common cancer sites among women included colorectal, respiratory, ovarian and stomach cancers (1, 2). The study of cancer distribution in regional and

***Corresponding Author:** Tel: +98 811 83800909, E-mail address: mahjub@umsha.ac.ir

global levels is difficult because of their relation with vast variety of phenomena. Thus, the first step to study such phenomena is to detect and classify the regions with common characteristics. Different studies have been carried out to investigate the geographical distribution of various cancers. However, studies on simultaneous distribution of cancer are limited. In most studies, having considered a specific type of cancer, different area of countries has been clustered (3, 4). Whereas if the goal of the study showing pattern of various type of cancers in different countries, multivariate methods of statistical techniques must be applied.

Cluster analysis is one of the prevailing methods that widely used for classification of such phenomena (5). Multivariate Statistical methods in order to classification of diseases and the various indicators are used in different studies. For example, Farhadian et al. used multivariate method of factor analysis to examine the relation between social economic indicators and the indicators of child mortality in different provinces of Iran (6). Yazdi et al. used cluster analysis technique to classify different provinces of Iran based on the health indicators of mother (7). Mahjub et al. also used multivariate method of factor analysis to assessment women's health needs in different provinces of Iran (8).

Clustering methods are widely used for classification of objects of similar kind into respective categories. Cluster analysis consists of different algorithms and methods such as fuzzy clustering, K-Means clustering, hierarchical clustering, and model-based clustering (5).

One drawback of various clustering method is ignoring the distribution of variables. In order to solve this problem, using the method "model based clustering" was purposed. The model-based clustering was first introduced by Wolfe and then was revised by McLachlan and Peel in 2000 (9) and by Fraley and Raftery in 2002 (10). Model-based clustering is widely used in multivariate analysis such as density estimation and discriminant analysis (10), magnetic resonance imaging (11, 12), and microarray data analysis (13-15).

Classification of the disease distribution among different countries worldwide based on single factor is straightforward. However, considering the classification of countries based on different indices need to apply multivariate methods such as clustering method. Model-based clustering is a powerful multivariate method based on mixture distribution, which can efficiently classify the countries considering several indices (9). Furthermore, no literature was found for classifying death rates from women's cancers in different countries in the world. Therefore, the main objective of the present study is to classify countries based on the death rates from women's cancers using model-based clustering.

Materials and Methods

In this study, which was performed in 2009, we used dataset regarding death rate from different variables of breast, cervical, uterine, and ovarian cancers in 192 countries worldwide reported by World Health Organization (WHO) (16). From the data Kiribati and Sanmario countries were excluded because of missing observations in some variables. Hence, 190 countries remained for analysis.

In order to perform model-based clustering, initially normal mixture models were fitted with different numbers of components to the data in order to determine the optimal number of clusters. Then, the best-fitted model was selected with appropriate number of components using Bayesian Information Criteria (BIC). Next, the model parameters were estimated using maximum likelihood via the EM algorithm. Finally, based on the best-fitted model, the countries were allocated into four distinct clusters based on death rate from women's common cancers. Accordingly, the regions with homogenous characteristics were determined. An important aspect of model-based clustering is that the former is based on a statistical model. In other word, in model-based clustering, a postulated statistical model is considered for the population from which the sample data is obtained. Accordingly, can be displayed the probability density function of any given data X_i with mixture distribution as follows:

$$f(x_i;\theta) = \sum_{k=1}^{g} p_k f_k(x_i;\theta)$$

Where, k is the number of components in mixture model, $f_k(x_i;\theta)$ is mixture density, and P_k is the probability of the observation that comes from a gth mixture component and called mixing proportions. The parameters of the model (P_k and θ) were estimated using the EM algorithm. When the mixture model is fitted, the data can be categorized into g clusters. The posterior probability for membership of each observation can be obtained using following formula:

$$P(Z_k = k|x;\hat{\theta}) = \frac{\hat{p}_k f_k(x|\hat{\theta}_k)}{\sum_{k=1}^{g} \hat{p}_k f_k(x|\hat{\theta}_k)}$$

For each observation, the posterior probability was calculated based on the mentioned formula for each cluster. Then, we assigned the observation to the cluster in which the posterior probability was the highest. Accordingly, cluster k includes observations that are devoted to component k. For multivariate data analysis of a continuous nature, attention has been concentrated on the use of multivariate analysis with normal component distribution because of their computational convenience. In that case, the probability density function of component distribution with the mean μ_k and the covariance matrix Σ_k will be as follows (9):

$$\phi(x_i;\ \mu_k;\Sigma_k)$$

The EM algorithm is the reference tool by which the maximum likelihood in a mixture model can be derived (17). Geometric features of the clusters, including shape, volume, and orientation, can be specified by the covariance matrices, Σ_k. Banfield and Raftery in 1993 (18) and Celeux and Govaert in 1995 (19) suggested a basic design for geometric constraints in multivariate normal mixtures. They parameterized covariance matrices through eigenvalues decomposition in the following form:

$$\Sigma_k = \lambda_k D_k A_k D_k'$$

Where, λ_k is an associated constant of probability, D_k is the orthogonal matrix of eigenvectors, $A_k$is a diagonal matrix. The parameters λ_k, D_k and A_k determine size, direction and the shape of cluster k respectively. By analyzing constraints and applying various elements, 28 different models were obtained (18, 19). For the present data, we fitted various mixture normal models (g=2 to 6) considering 28 different forms of decomposition of related matrix of variance covariance components. Model selection was based on minimum Bayesian Information Criteria (BIC) amount as well as the minimum value of Entropy Index (EI) in different number of mixture components in 28 suggested models (20). We used MIXMOD version 2.1.1 and MATLAB version 7.0 software for data analysis.

Results

A total number of 140 models with different components and decomposition matrix of variance covariance were fitted to the data. Among these models, the highest and lowest values of BIC were 3553.04 and 3118.66 respectively. So a mixture model of four components with different mixing proportion, shapes, and directions having minimum BIC=3118.66 and Entropy Index=24.99 was selected.

The estimated parameters of the selected model are shown in Table 1.

Based on the information in the table, posterior probabilities for each country were computed. Then, each country was allocated to one of the four mixture components based on the highest posterior probability. Accordingly, the world countries were allocated to four distinct clusters, including 43 (23%) countries in cluster 1, 28 (14%) countries in cluster 2, 75 (39%) countries in cluster 3, and 44 (24%) countries in cluster 4 as it can be seen in Table 2.

Table 1: Estimated parameters of the best-fit model

				Variance covariance matrix			
Component	Mixing proportion	Variable	Mean	Breast cancer	Cervical cancer	Uterine cancer	Ovary cancer
Component 1	0.24	Breast	9.05	25.58	-3.51	4.04	4.41
		Cervical	6.32	-3.51	15.05	-2.47	0.53
		Uterine	2.75	4.04	-2.47	3.09	0.51
		Ovary	1.78	4.41	0.53	0.51	1.26
Component 2	0.14	Breast	4.67	2.50	2.59	0.31	1.45
		Cervical	4.00	2.59	4.16	0.08	2.50
		Uterine	1.07	0.31	0.08	0.17	0.05
		Ovary	2.21	1.45	2.50	0.05	2.03
Component 3	0.38	Breast	4.77	4.03	-0.50	0.17	0.31
		Cervical	4.66	-0.50	6.20	-0.02	0.36
		Uterine	0.33	0.17	-0.02	0.02	0.03
		Ovary	1.23	0.31	0.36	0.03	0.27
Component 4	0.23	Breast	18.24	15.40	-0.48	1.03	4.10
		Cervical	2.68	-0.48	2.15	1.02	1.05
		Uterine	3.57	1.03	1.02	1.16	1.13
		Ovary	5.60	4.10	1.05	1.13	2.73

Table 2: Classification of women's cancers in different countries worldwide

Component 1				
Albania	Brazil	Guatemala	Paraguay	Serbia
Antigua	Burkina Faso	Guyana	Rep of Korea	South Africa
Argentina	Cuba	Haiti	Romania	Suriname
Azerbaijan	Dominica	Indonesia	Saint Kitts	Swaziland
Bahamas	Dominican	Jamaica	Saint Lucia	Trinidad
Barbados	Ecuador	Lesotho	Saint Vincent	Uruguay
Belize	El Salvador	Mauritius	Sao Tome	Venezuela
Bosnia	Georgia	Nauru	Seychelles	
Bolivia	Grenada	Nicaragua	Singapore	
Component 2				
Cameroon	Fiji	Niger	Samoa	Tonga
Chile	Honduras	Niue	Solomon	Tuvalu
Colombia	Kyrgyzstan	Palau	Sri Lanka	Uzbekistan
Cook Islands	Maldives	Panama	Syrian Arab	Vanuatu
Costa Rica	Marshall	Philippines	Tajikistan	
Ethiopia	Micronesia	Moldova	Timor-Leste	
Component 3				
Afghanistan	Comoros	Iran	Mongolia	Sierra Leone
Algeria	Congo	Iraq	Morocco	Somalia
Angola	Côte d'Ivoire	Jordan	Mozambique	Sudan
Bahrain	Demo Korea	Kenya	Myanmar	Thailand
Bangladesh	Djibouti	Kuwait	Namibia	Togo
Benin	Egypt	Lao	Nepal	Tunisia
Bhutan	Eritrea	Lebanon	Nigeria	Turkey
Botswana	Equatorial	Liberia	Oman	Turkmenistan
Burundi	Guinea	Libyan	Pakistan	Uganda
Brunei	Gabon	Madagascar	Papua	UnitedEmirates
Cambodia	Gambia	Malawi	Peru	Viet Nam
Cape Verde	Ghana	Malaysia	Qatar	Yemen
CentralAfrican	Guinea	Mali	Rwanda	UnitedTanzania
Chad	Guinea Bissau	Mauritania	Saudi Arabia	Zambia
China	India	Mexico	Senegal	Zimbabwe
Component 4				
Andorra	Cyprus	Iceland	Malta	Slovenia
Armenia	Czech Republic	Ireland	Monaco	Spain
Australia	Denmark	Israel	Netherlands	Sweden
Austria	Estonia	Italy	New Zealand	Switzerland
Belarus	Finland	Japan	Norway	Yugoslav
Belgium	France	Kazakhstan	Poland	Ukraine

Bulgaria	Germany	Latvia	Portugal	UK
Canada	Greece	Lithuania	Russian	USA
Croatia	Hungary	Luxembourg	Slovakia	

According to the results of model based clustering, most countries in South America allocated to the first cluster. In addition, most countries in Africa, Central, and Southeast Asia were located to the third cluster. Furthermore, the fourth cluster consisted of Pacific continent, North America, and European countries (Fig. 1).

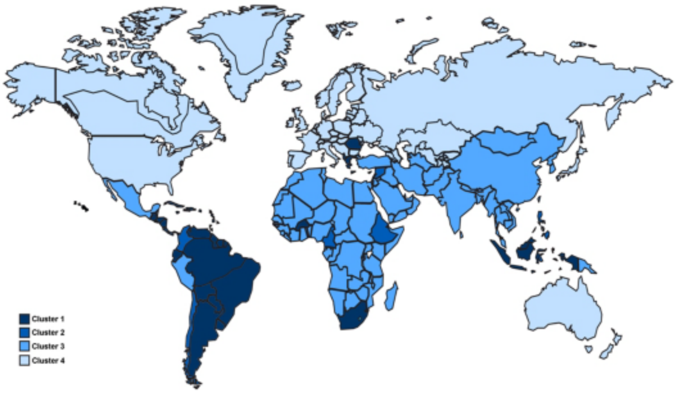

Fig. 1: Model-based clustering of the countries worldwide according to the death rates from women's common cancers (Breast, Uterine, Cervix, and Ovary)

Discussion

Various studies indicated that mode-based clustering methods have better performance than other methods when clusters are overlapping with different shape and size (21). In addition, model-based clustering are increasingly preferred over other procedures because variance-covariance matrices of the model simplify the interpretability of the results (22).

We could not find any study to use model-based clustering to classify regions or countries based on cancer data. However, several studies used other kinds of cluster models for classification of different regions such as hierarchical clustering methods, K-Means and fuzzy clustering. For example, Abadi et al. used cluster analysis to classify universities of medical sciences and faculties of medicines (23). Babaee et al. used fuzzy clustering and hierarchical clustering method to classify the provinces based on population and health indicators (24). Vahedi et al. applied hierarchical and non-hierarchical clustering methods on DNA microarray data to classify patients with breast cancer (25).

In addition, there is an increasing preference to use model-based clustering over other methods worldwide. Mar et al. in 2003 applied model-based clustering method for clustering genes associated with breast cancer (26). Pan et al. in 2002 applied model-based clustering method to analyze gene microarray data. They used log likelihood ratio and BIC criteria to select the number of components of the mixture model method (27). McLachlan in 2002 used EMMIX GENE software for model-based clustering method to classify data of micro array gene (28), whereas, in our work, we used MIXMOD software for data analysis. Furthermore, Chen et al. in 2008 applied model based clustering method for diagnosis of cancer patients (29), while we classified the observations in more than two groups. More recently publication in the field of model based clustering is related to Haibe-Kains study that used model-based clustering to identify molecular species in breast cancer (30) as well as Muna et al. in 2008 applied model-based clustering method for clustering adolescent behavioral problems during adulthood (31).

One limitation of the model-based clustering is the maximum number of parameters needs to be estimated. That means relatively more data points are required in each component (32). Despite its limitation, a main contribution of the present study was introduction of an appropriate and flexible method of clustering that might be used in vast variety of public health contexts. One advantage of model-based clustering is its simplicity and flexibility. Another advantage of this model is that, like other statistical models, it is possible to impose restriction on the parameters to obtain more parsimony (21). The third advantage of the model-based clustering is that there is no necessity to make decision on scaling of the observed variables while in standard non-hierarchical cluster methods like K-means, scaling of the observed variables is always an important issue (22).

In conclusion, we showed that model-based clustering could be easily used to classify geographical regions appropriately based on various sample data. Considering the benefits of clustering based on normal mixture models over other conventional clustering methods, it seems that this method can be applied in wide variety of medical and public heath contexts.

Ethical considerations

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely observed by the authors.

Acknowledgments

This article is a part of MSc thesis supported by Hamadan University of Medical Sciences. We would like to thank Deputy of Education as well as Deputy of Research and Technology of Hamadan University of Medical Sciences for financial support of this study. The Deputy of Research and Technology of Hamadan University of Medical Sciences funded this study. The authors declare that there is no conflict of interests.

References

1. Disease Control Priorities Project. Controlling cancer in developing countries: prevention and treatment strategies merit further study. 2007. http://www.dcp2.org/file/79/DCPP-Cancer.pdf.
2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011). Global Cancer Statistics. *CA Cancer J Clin,* 61: 69-90.
3. Ahmedin J, Martin K, Susan SD, Richard B, HAYES F, Joseph F (2002). A geographic analysis of prostate cancer mortality in the united states, 1970–89. *Int J Cancer,* 101: 168-74.
4. Geoffrey MJ, Dunrie AG (2003). Local clustering in breast, lung and colorectal cancer in Long Island, New York. *Int J Health Geogr,* 2: 1-12.
5. Everitt BS, Landau S, Leese M, Daniel S (2001). *Cluster analysis.* 5th ed. London: John Wiley & Sons, pp 111-43.
6. Farhadian M, Mahjub H, Sadri GH, Aliabadi M (2010). Ranking health status of children in iran's provinces and assessing its relation with socio-economic indicators. *Hakim Research Journal,* 13 (2): 38-44.
7. Yazdi M, Mahjub H (2011). Classification of maternal health status in the rural province of Iran using multivariate methods of factor and cluster analysis. *Iran J Epidemiol,* 7 (1): 7-14 [Persian].
8. Mahjub H, Sadri GH (2001). Evaluation method of women's health needs in Iran provinces. *Hakim Research Journal,* 4: 46-50.
9. McLachlan GJ, Chang SU (2004). Mixture modelling for cluster analysis. *Stat Methods Med Res,* 13 (5): 347-61.
10. Fraley C, Raftery A (2002). Model-based clustering, discriminant analysis and density estimation. *J Am Stat Assoc,* 97: 611-31.
11. Forbes F, Peyrard N, Fraley C, Georgian-Smith D, Goldhaber D, Raftery A (2006). Model-based region-of-interest selection in dynamic breast MRI. *J Comput Assist Tomogr,* 30 (4): 675-87.
12. Fraley C, Raftery AE (2006). Model-based microarray image analysis. *R News,* 6: 60-3.
13. Li Q, Fraley C, Bumgarner RE, Yeung KY, Raftery AE (2005). Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics,* 21: 2875-82.
14. Muira WM, Rosa GJ, Pittendrigh BR, et al (2009). A mixture model approach for the analysis of small exploratory microarray experiments. *Comput Stat Data Anal,* 53: 1566-76.
15. Wong T-T, Liu K-L (2010). A probabilistic mechanism based on clustering analysis and

distance measure for subset gene selection. *Expert Systems with Applications,* 37: 2144-9.

16. World Health Organization. WHO statistical information system 2009; http://www.who.int/whosis.
17. Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Appl Stat,* 39 (1-38).
18. Banfield JD, Raftery AE (1993). Model-based gaussian and non-gaussian clustering. *Biometrics,* 49: 803-21.
19. Celeux G, Govaert G (1995). Gaussian parsimonious clustering models. *Pattern Recognit,* 28: 781-93.
20. Schwarz G (1978). Estimating the dimension of a model. *The Annals of Statistics.*6: 461-4.
21. Fraley C, Raftery A (1988). How many clusters? which clustering method?answers via model-based cluster analysis. *Computer Journal,* 41: 578-88.
22. Bensmail H, Celeux G (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *J Am Stat Assoc,* 91: 1743-7.
23. Abadi AR, Azam K (2000). Classification of universities and faculties of medical sciences in basis of cluster analysis and comparing to peresent categorization. *Hakim Research Journal,* 2 (4): 246-53 [Persian].
24. Babaee G, Feizi A (2005). Fuzzy classification of Iran provinces based on health and demographic indices. *Hakim Research Journal,* 7 (4) :1-7 [Persian].
25. Vahedi M, Alavi Majd H, Mehrabi Y, Naghavi B (2008). Gene expression data clustering and it's application in differential analysis of leukemia. *Journal of Semnan University of Medical Sciences,* 9 (2): 163-9.
26. Mar JC, McLachlan GJ (2003). Model-based clustering in gene expression microarrays: an application to Breast cancer data. *International Journal of Software Engineering and Knowledge Engineering,* 13: 579-92.
27. Pan W, Lin J, Le CT (2002). Model-based cluster analysis of microarray gene expression data. *Genome Biol,* 3 (2): 1-8.
28. McLachlan GJ, Bean RW, Peel D (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics,* 18: 413-22
29. Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X (2008). A clustering approach in developing prognostic systems of cancer patients. *Seventh International Conference on Machine Learning and Applications,* 723-8
30. Haibe-Kains B (2010). Classification models for breast cancer molecular subtyping:what is the best candidate for translation into clinic? *Women's Health,* 6 (5): 623-5.
31. Muna EY, Windlea M, Schainkera LM (2008). A model-based cluster analysis approach to adolescent problem behaviors and young adult outcomes. *Dev Psychopathol,* 20: 291-318.
32. Yeung kY, Fraley C, Murua A, Raftery AE, Ruzzo WL (2001). Model-based clustering and data transformations for gen expression data. *Bioinformatics,* 17 (10): 977-87.