Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Cholera Forecast based on Association Rule Mining

[1]LTN Anh*, [2]HX Dau, [3]NH Phuong
[1]Hanoi Medical University, Hanoi, Vietnam
[2]Post and Telecommunication Institute of Technology, Hanoi, Vietnam
[3]Ministry of Health Hanoi, Vietnam

Abstract— Recently, the epidemic forecast has been interest of many researchers. Traditionally, mathematical epidemiological models have been used for epidemic forecast. However, these models require some functional dependency assumptions that may not be always practical. Other approaches based on data mining have achieved promising results thanks to the development of efficient data mining techniques and computing technologies. This paper proposed a method for cholera forecast based on association rule mining using the non-standard distribution dataset. Experimental results show that the proposed method is suitable for cholera forecast and can be used as an important input in the decision making process of the preventive healthcare.

Keywords— Cholera forecast, data mining, association rule mining

I. INTRODUCTION

In recent years, the epidemic forecast has been interested in research and plays an increasingly important role in the disease prevention activities in the health sector due to the complexity and the quick spreading ability of known and emerging diseases. In the epidemic forecast, the data collection is an important task and the selection of suitable data analysis and processing methods plays a decisive role to the accuracy of the disease forecasting outcome. In many types of diseases, cholera is a dangerous disease because it is capable of spreading rapidly and it greatly affects the health of the community, even to cause significant damage to people. Water has long been identified as a major factor in the spread of cholera and this was also affirmed by John Snow in mid-nineteenth century [12]. The spread of cholera has the epidemiological factor and the interaction network among people in the community. This study uses association rule generation algorithm [9] to generate association rules from datasets of cholera cases in districts of Hanoi from 2001 to 2012 for predicting the occurrences of cholera cases. The rest of this paper is organized as follows: Section II discusses some typical related works. Section III gives a brief introduction to association rules and describes the Apriori and Rule generation algorithms. Section IV presents the application of association rule mining in the cholera forecast for nonstandard distribution cholera dataset of Hanoi city and Section V is our conclusion and future work.

II. RELATE WORK

There have been many cholera forecast models published and they can be divided into 3 major classes. The first class consists of mathematical epidemiological models. Typical examples of these models are variants of SIR-SIS (Susceptible-Infected-Recovered and Susceptible-Infected) forecast models [3][7]. The second class includes data mining based models, in which self-regression models and those based on social media mining [16] have been widely used. The last class contains all other models, in which agent-based models have been considered to be suitable for complex systems. Based on our extensive review, models that belong to the first and second classes have been popularly used in practice of the cholera forecast.

J. Wang and S. Liao [7] used the generalised mathematical epidemiological model that is a combination of a regular SIR model and an environmental component via 4 differential equations as follows:

$$\frac{dS}{dt} = bN - Sf(I,B) - bS,$$
$$\frac{dI}{dt} = Sf(I,B) - (\gamma + b)I,$$
$$\frac{dR}{dt} = \gamma I - bR,$$
$$\frac{dB}{dt} = h(I,B),$$

where, S, I, and R denote the susceptible, the infected, and the recovered populations, respectively, and B denotes the concentration of the vibrios in the contaminated water. The total population $N=S+I+R$ is assumed to be a constant. The parameter b represents the natural human birth/death rate, and γ represents the rate of recovery from cholera. In this model, $f(I,B)$ is the incidence function that determines the rate of new infections and the function $h(I,B)$ describes the rate of change for the pathogen in the environment. By introducing general incidence and pathogen functions, their model can unify cholera studies into a single framework of modelling, simulation and analysis.

In 2011, Mukandavire *et al.* [17] proposed a mathematical epidemiological model to study the 2008–2009 cholera outbreak in Zimbabwe. The model took both human-to-human and environment-to-human transmission pathways into consideration. Their work confirmed the importance of the human-to-human transmission in cholera epidemics, especially in Zimbabwe, a landlocked country in the middle of Africa.

Almost all mathematical epidemiological models require some functional dependency assumptions as follows:

a) f(0, 0)=0, *h(0, 0)*=0,
b) $f(I,B) \geq 0$,
c) $\frac{\partial f}{\partial I}(I,B) \geq 0 \quad \frac{\partial f}{\partial B}(I,B) \geq 0 \quad \frac{\partial f}{\partial I}(I,B) \geq 0 \quad \frac{\partial h}{\partial B}(I,B) \leq 0$
d) *f(I,B)* is concave,
e) *h(I,B)* is concave.

The assumption (a) makes sure that the vector equation always has a unique solution and assumption (b) ensures a non-negative incidence rate. The first 2 inequalities in assumption (c) indicate that increased infection and pathogen concentration lead to higher incidence rate, while the third inequality indicates that increased infection results in higher growth rate for the pathogen in the environment. The last inequality in assumption (c) indicates a positive net death rate of the vibrios. Condition (d) is a common assumption for nonlinear incidence and condition (e) is an additional assumption for the regulation of the environmental function *h(I,B)*. In practice, not all of these assumptions can hold and that may make mathematical epidemiological models less reliable.

In 2012, R.C. Reiner *et al*. [15] proposed a mathematical epidemiological model for cholera forecast in Dhaka, a mega-city of Bangladesh. The proposed model supports spatial factors, in which the research area was divided into 21 regions. The model needs to determine the probability of shifting cholera level among regions monthly. In this model, cholera status $X_{m,t}$ of region *m* in month *t* depends on the cholera status of that region in previous months and cholera status of neighbouring regions in the neighbouring region set *N(m)* of region *m*. The research results found that a climate-sensitive urban core that acts to propagate cholera risk to the rest of the city of Dhaka.

Y. Yue *et al*. [8] presented a cholera forecast model based on the effect of climate factors in the estuary of Pearl River, South China in the period 2008-2009. The climate data were collected daily at 2 meteorological stations in Guangzhou and Shenzhen. Then the collected daily data were converted to monthly data. The data of positive cholera cases were provided by the China Center for Disease prevention. Parameter values, including water temperature coefficient, coefficient of cholera shifting in the regions, were determined by linear regression.

R. Chunara *et al*. [16] built a model for cholera forecast using data retrieved from Twitter, a well-known social network. The authors believed the proposed approach could provide a chance to collect early information on how an epidemic was happening, and therefore create a chance to implement timely and effective intervention measures. The research has discovered time-series data matching an exponential distribution. In this case, the following formula is used to calculate the number of infected cases of SIR model: $R_e = 1 + rT_c$, where, $T_c = 1 / b$, *b* is the rate of leaving the infected stage in a SIR model and *r* is the growth rate. The confidence of the results is 68% with 1 day delay. However, the authors also pointed out some limitations of cholera forecast using data collected from social or informal media. First, it is the low level of use of social media in the epidemic area. Second, data contributed by individuals from informal media may be more prevalent from certain age or other demographic groups. This may lead to biased data distribution and does not represent the full picture of the ongoing epidemic. Third, informal media reports may contain false positives. Finally, there was a large difference between the number of cholera cases reported by informal media sources and correct number of cases in the epidemic.

Previously, most of cholera forecast researches were based on mathematical epidemiology because data mining techniques and tools have not been well developed. In addition, the collected epidemic data sets may not be suitable to build reliable epidemic forecast models based on data mining. Mathematical epidemiological models require some assumptions that may make them less reliable. Recently, data mining based epidemic forecast has been the interest of many researchers thanks to the development of efficient data mining techniques. Hence, assumptions of mathematical epidemiological models can gradually be removed. Mining of association rules is an important component of data mining [1][2]. The purpose of association rule mining is to find out the association or the correlation between data items.

III. ASSOCIATION RULES AND ASSOCIATION RULE GENERATION

A. Introduction to association rules

An association rule has the form of $X => Y$, where *X* and *Y* are itemsets. The rule $X => Y$ can be understood as items that have an attribute *x* in item set *X* also have attribute *y* in item set *Y*. On the statistical point of view, *X* is considered as the independent variable and *Y* is the dependent variable. *Support*, *Confidence* and *Lift* are 3 measures of association rules [10].

Support of rule $X => Y$ is the frequency of events that contain all items in both *X* and *Y* item sets. *Support* written as $Supp(X \rightarrow Y)$ is computed as:

$$Supp\ (X \rightarrow Y) = P\ (X \cup Y) = \frac{n(X \cup Y)}{N}$$

where *N* is the total number of events and $n(X \cup Y)$ is the number of events that contain all items in both *X* and *Y* itemsets.

Confidence of rule $X => Y$ is the conditional probability of *Y*, given *X*. *Confidence* written as $Conf(X \rightarrow Y)$ is computed as:

$$Conf\ (X \rightarrow Y) = P\ (Y \mid X) = \frac{n(X \cup Y)}{n(X)}$$

where *n(X)* is the number of events that contain *X*.

The statistical certainty measure of rule *X => Y*, written as *Lift(X=>Y)* is computed as:

$$Lift(X \rightarrow Y) = \frac{supp\ (X \cup Y)}{supp(X) \times supp\ (Y)}$$

where *supp(X)* is the *support* of item set *X* and it is defined as the ratio of number of events containing items in *X* out of the total number of events. Similarly, *supp(Y)* is the *support* of item set *Y* and it is defined as the ratio of number of events containing items in *Y* out of the total number of events. The higher the value of *Lift(X=>Y)* is the higher statistical meaning of the rule.

In order to mine association rules, it is necessary to select 2 factors, including Minimum support (*minsup*) and Minimum confidence (*minconf*). They are predefined thresholds for the generation of association rules. An itemset that has the occurrence frequency higher than the minimum support is called frequent or large itemset [10]

B. Association rule generation

The purpose of mining association rules is to find out the correlation between data items. For convenience, we use the same notations as that of [9]:

$I = \{i_1, i_2, \ldots, i_m\}$: A set of items;

T: A transaction, $T \subseteq I$;

D: A set of transactions, each of which is *T*;

k-itemset: An itemset having *k* item;

L_k: Set of large *k*-itemsets (those with minimum support). Each member of this set has 2 fields: (1) itemset and (2) support count;

C_k: Set of candidate *k*-itemsets (potentially large itemsets). Each member of this set has 2 fields: (1) itemset and (2) support count;

H_m: A set of *m*-item consequents.

Given the set of transactions *D*, the problem of mining association rules is to generate all rules that have minimum support and minimum confidence, as noted in Section III.A. The process of mining association rules consists of 2 stages: (1) generate all frequent or large itemsets using *Apriori algorithm* and (2) generate association rules from large itemsets using *Rule generation algorithm*. The *Apriori algorithm* **Error! Reference source not found.** is described as follows:

L_1 = {large 1-itemsets}
for (k = 2; $L_{k-1} \neq \varnothing$; k++) do begin
  C_k = apriori-gen(L_{k-1}); // New candidates
  forall transactions $t \in D$ do begin
    C_t = subset(C_k, t); // Candidates contained in t
    forall candidates $c \in C_t$ do
      c.count++;
  end
  L_k = {$c \in C_k$ | c.count ≥ minsup}
end
Answer = $\cup_k L_k$;

The apriori-gen(L_{k-1}) candidate generation function returns a superset of the set of all large *k*-itemsets. The function includes 2 processing steps: the *join* step and the *prune* step. The details of these 2 steps were given in**Error! Reference source not found.**.

In order to generate association rules from large itemsets L_k, $k \geq 2$, the *Rule generation algorithm* was used as follows:

forall large *k*-itemsets L_k, $k \geq 2$ do begin
  H_1 = {consequents of rules from L_k with one item in the consequent } // H_1 is the set of 1-item consequent
  call ap-genrules(L_k, H_1);
end
procedure ap-genrules(L_k, H_m)
  if ($k > m$+1) then begin
    H_{m+1} = apriori-gen(H_m);
    forall $h_{m+1} \in H_{m+1}$ do begin
      conf = *supp*(L_k) / *supp*(L_k - h_{m+1});
      if (*conf* ≥ *minconf*) then
        ouput the rule (L_k - h_{m+1}) => h_{m+1}
          with the confidence = *conf* and
          support = *supp*(L_k)
      else
        delete h_{m+1} from H_{m+1};
    end
    call ap-genrules(L_k, H_{m+1});
  end

IV. ASSOCIATION RULES AND ASSOCIATION RULE GENERATION

We proposed the forecast model of the possibility of cholera occurrence in Hanoi city based on association rule mining from cholera dataset collected in Hanoi's districts from 2001 to 2012. The Apriori algorithm and the Rule generation algorithm [10] described in Section III.B were used to generate association rules

A. Experimental data sets

The experimental data sets used in this research include:

- A data set of cholera cases in Hanoi in the period 2001-2012 provided by the Hanoi Center for Preventive Medicine;
- A data set of hydrological layers of Hanoi city.

In order to build a data table of cholera cases of each district of Hanoi city, we used the *R* language [2][5] to construct a secondary cholera dataset in the form of a transaction list. This list consists of multiple lines and each line is a transaction on a date. Each transaction consists of data fields, including date, month and list of districts, in which there were at list one cholera case on the date.

Fig. 1 presents the number of cholera cases in Hanoi from January 1, 2001 to December 31, 2012 distributed on years and Fig. 2 presents the distribution of the total number of cholera cases based on months.

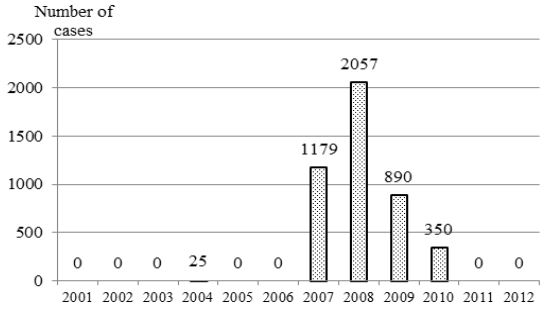

Fig. 1. Number Of Cholera Cases In Hanoi From January 1, 2001 To December 31, 2012

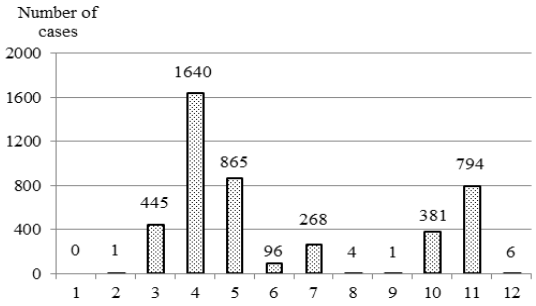

Fig. 2. Distribution Of The Total Number Of Cholera Cases Based On Moths

B. Some experiment results

Mining the secondary cholera dataset using *Rule generation algorithm* with selected parameters, including minimum support of 30%, minimum confidence of 70%, we received a set of association rules. Each rule has 5 elements, including Left Hand Side (LHS), Right Hand Side (RHS), *Support*, *Confidence* and *Lift*. *Support*, *Confidence* and *Lift* are 3 major measures of an association rules discussed in Section III.A. Due to the large number of generated association rules, we only collected 50 rules that have *Lift* measure over 1.2, as shown in Table I.

Taking the rule R1 {Dong Da, Hai Ba Trung, Hoang Mai} => {Thanh Xuan} as an example to explain the meaning of rules given in Table I. Dong Da, Hai Ba Trung, Hoang Mai and Thanh Xuan are districts of Hanoi city. The meaning of rule R1 is "*If there are cholera cases in Dong Da, Hai Ba Trung and Hoang Mai districts, then there also are cholera cases in Thanh Xuan district with support of 30.27%, confidence of 86.15% and statistical certainty of 2.0971*".

TABLE I. SOME SAMPLE RULES OF 50 GENERATED RULES

Rule No.	LHS	RHS	Support	Confidence	Lift
R1	{Dong da, Hai Ba Trung, Hoang Mai}	{Thanh Xuan}	0.3027	0.8615	2.0971
R2	{Dong da, Hoang Mai}	{Cau Giay}	0.3081	0.7307	2.0483
R3	{Hai Ba Trung, Hoang Mai}	{Thanh Xuan}	0.3081	0.8260	2.0108
R4	{Dong da, Hoang Mai}	{Thanh Xuan}	0.3351	0.7948	1.9348
……					

R9	{Tu Liem}	{Thanh Xuan}	0.3027	0.7272	1.7703
R10	{Thanh Xuan}	{Tu Liem}	0.3027	0.7368	1.7703
……..					
R48	{Dong da}	{Hoang Mai}	0.4216	0.7572	1.3735
R49	{Ha Dong}	{Hoang Mai}	0.3027	0.7466	1.3542
R50	{Hai Ba Trung}	{Hoang Mai}	0.3729	0.7113	1.2901

TABLE II. DISTRICTS WITH OLLUTED RIVERS FLOWING THROUGH AND CONTIGUOUS DISTRICTS

Districts	**Polluted rivers flowing through**	**Contiguous districts**
Ba Dinh	To Lich	Hoan Kiem, Cau Giay, Dong Da
Cau Giay	To Lich	Từ Liêm, Ba Dinh, Cau Giay, Dong Da, Thanh Xuan
Dong Da	To Lich	Hoan Kiem, Cau Giay, Ba Dinh, Hai Ba Trung, Thanh Xuan
Ha Dong	Nhue	Tu Liem, Thanh Xuan
Hai Ba Trung	Kim Nguu	Hoan Kiem, Hoang Mai, Thanh Xuan, Dong Da
Hoang Mai	Kim Nguu, To Lich	Hai Ba Trung, Thanh Xuan
Hoan Kiem		Ba Dinh, Hai Ba Trung, Dong Da
Thanh Xuan	To Lich, Kim Nguu	Cau Giay, Ha Dong, Hoang Mai, Dong Da, Tu Liem
Tu Liem	Nhue	Cau Giay, Ha Dong, Thanh Xuan

Reviewing Hanoi's hydrological factors that affect the spread of cholera, there are 3 highly polluted rivers flowing across Hanoi city, including To Lich, Kim Nguu and Nhue rivers. These rivers are flowing through some districts as shown in Table II. Table II also gives list of contiguous districts of each district that has polluted rivers flowing through.

C. Discussion

Based on the observation of the cholera dataset, the number of cholera cases in Hanoi focused mainly in the period 2007-2010 as shown in Fig. 1. The months with the highest number of cholera cases were March, April, May, July, October and November as shown in Fig. 2. The number of cholera cases in other months was very small. In other periods outside 2007-2010, the number of cholera cases was 0 per each year, except there were 25 cholera cases in 2004. So the number of cholera cases in the period 2001-2012 was very small compared to Hanoi's population. Therefore, the variable of the cholera cases in Hanoi did not follow the rule of the standard distribution. Although Y. Yua *et al.* [8] proved that regression and classification approaches for cholera forecast models can produce good results, the application of these approaches are not suitable for Hanoi's cholera non-standard distribution dataset. In addition, there are not significant differences in the natural climate conditions between urban districts and suburbs of Hanoi. Therefore, the application of the cholera forecast model for regions with significant differences in the natural climate conditions, proposed by R.C. Reiner *et al.* [15] is also not suitable. From the results of this research and the Hanoi's hydrological data, two major points can be given:

- Cholera cases tend to co-appear in Hanoi's urban districts and suburbs, where Hanoi's highly polluted rivers, including To Lich, Kim Nguu and Nhue rivers flowing through with high confidence of over 70%;
- Cholera cases in districts with polluted rivers flowing through and cholera cases in contiguous districts without polluted rivers tend to co-appear with high confidence of over 70%.

The proposed approach's results are also in accordance with the research results on cholera outbreaks in Calcutta in1994 [19], in Varanasi in 2006 [20], India and a cholera outbreak in Haiti in 2010 [18]. In Vietnam, cholera bacteria were found in water samples of polluted rivers in cholera outbreaks in Thuy Nguyen district, Hai Phong city in 1976 [12], in Hue city in 1980 [13].

Our association rule mining based cholera forecast model shows that the co-occurrence trend of cholera cases in Hanoi's districts with polluted rivers flowing through is in accordance with the conclusion of previous researches in the world and in Vietnam. This confirms that our proposed approach is suitable for the cholera forecast model on the non-standard distribution dataset and there are not significant differences in the natural climate conditions among geographical regions.

V. CONCLUSIONS AND FUTURE WORK

This paper proposed to use association rule mining in the cholera forecast model on the non-standard distribution dataset and there are not significant differences in the natural climate conditions among geographical regions. The results of mined association rules with high confidence and lift can be used as an important input in the decision making process of the preventive healthcare of Hanoi city.

In order to improve the reliability of the research results, other factors that affect the possibilities of the cholera spread need to be taken into the consideration. These factors include flowing speed of rivers in the regions, wind direction and floods. Furthermore, the research can be extended to cholera datasets of other regions in Vietnam.

ACKNOWLEDGMENT

We are very grateful to Hanoi Department of Science and Technology, Hanoi Centre for Preventive Medicine was supported and providing data to this research.

REFERENCES

[1] J. Han and M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, Third Edition, Morgan Kaufmann, 2011.
[2] L. Torgo, *Data Mining with R: Learning with Case Studies*, Chapman & Hall/CRC, 2011.
[3] F. Brauer, P. V. de Driessche and J. Wu, *Mathematical Epidemiology*, Springer, 2008.
[4] Hanoi's People Committee, *Hanoi's overall environmental report*, Hanoi's People Committee, 2011.
[5] The R Project for Statistical Computing, http://www.r-project.org, accessed in May 2015.
[6] K. Rajamani, A. Cox, B. Iyer and A. Chadla, *Effcient Mining for Association Rules with Relational Database Systems*, International Symposium Proceedings, IDEAS '99, 1999, pages: 148 – 155.
[7] J. Wang and S. Liao, *A generalized cholera model and epidemic- endemic analysis*, Hournal of Biological Dynamics, p.568-589, 2012.
[8] Y. Yue, J. Gong, D. Way, B. Kan, B. Li and C. Ke, *Influence of Climate factors on Vibro cholera dynamics in the Pearl River estuary*, South China, World J. Microliol Biotechnol, 2014.
[9] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A.I. Verkamo, *Fast discovery of association rules*, Book chapter in Advances in knowledge discovery and data mining, pages 307-328, American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.
[10] A. Savasere, E. Omiecinski, and S. Navathe, *An efficient algorithm for mining association in large databases*, In VLDB, 1995.
[11] Kelly-Hope, *Temporal Trends and Climatic Factors Associated with Bacterial Enteric Diseases in Vietnam, 1991–2001*, Environmental Health Perspectives, p. 7–12, 2008.
[12] Nguyen Van Hieu, *Epidemiological characteristics of cholera epidemics in period 1976-1980,1981 caused by V.E.Tor in Hai Phong, Vietnam*, Medical PhD thesis, Hanoi Medicine University, 1984.
[13] Nguyen Dinh Son, *Nguyen Thai Hoa and Duong Quang Minh, Some epidemiological characteristics of cholera epidemics in Thua Thien Hue province*, Journal of preventive medicine, No. 29740, p.194-197, 2005.
[14] E.J. Nelson, J.B. Harri and J.G. Morris, *Cholera transmission: the host, pathogen and bacteriophage dynamic*, Nat Rev Micorobiol 7(10), p. 693-702, 2009.
[15] R.C. Rainer, A. King, M. Emch, M. Yunus, S.G. Faruque and M. Paucula, *Highly localized sensitivity to climate forcing drives endemic cholera in a megacity*, Proc.Nalt. Acd. Sci. U.S.S, 109,2033-2036, 2012.
[16] R. Chunara, J.R. Andrews and J.S Brownstein, *Social and news media enable estimation of epidemiology patterns early in 2010 Haitian cholera outbreak*, The American Journal of Tropical Medicine and Hygiene 86,1, p. 39-45, Jan, 2012.
[17] Z. Mukandavire, S. Liao, J. Wang, H. Gaff, D.L. Smith, and J.G. Morris, *Estimating the reproductive numbers for the 2008–2009 cholera outbreaks in Zimbabwe*, Proc. Natl Acad. Sci. 108 (2011), pp. 8767–8772.
[18] R. Piarroux, R. Barrais, B. Faucher, R. Haus, M. Piarroux, J. Gaudart, R. Magloire and D. Raoult, *Understanding the cholera epidemic, Haiti*. Emerging Infectious Diseases, 2011, 17.7: 1161-1168.
[19] S.K. Bhattacharya, M.K. Bhattacharya, D. Dutta, S. Garg, A.K. Mukhopadhyay, M. Deb, A. Moitra and G.B. Nair, *Vibrio cholerae O139 in Calcutta*, Arch Dis Child, Aug 1994, 71(2):161-2.
[20] S. Hamner, A. Tripathi, R.K. Mishra, N. Bouskill, S.C. Broadaway, B.H. Pyle and T.E. Ford, *The role of water use patterns and sewage pollution in incidence of water-borne/enteric diseases along the Ganges river in Varanasi, India*, International Journal of Environmental Health Research 16.2, 2006, p. 113-132.

