STATISTICAL MODELING OF THE GROUNDWATER ARSENIC CONTAMINATION LEVEL IN BANGLADESH DUE TO CHEMICAL ELEMENTS

Syed Md. Fakrul AHSAN[1]
MSc, Public Servant, Railcorp, Sidney, Australia

E-mail: chan2001au@yahoo.com

Md. Nazrul ISLAM[2]
MSc, Assistant Professor, Department of Statistics,
Shahjalal University of Science and technology (SUST), Sylhet-3114, Bangladesh

E-mail: nzrul330@yahoo.com

Md. Jamal UDDIN[3,4]
MSc, Lecturer, Department of Statistics,
Shahjalal University of Science and technology (SUST), Sylhet-3114, Bangladesh

E-mail: mjamalu@yahoo.com, jamal-sta@sust.edu

Mohammed Taj UDDIN[5]
MSc, Associate Professor, Department of Statistics,
Shahjalal University of Science and technology (SUST), Sylhet-3114, Bangladesh

E-mail: mtajstat@yahoo.com, taj_stat@sust.edu

Abstract: *This paper intends to offer a modeling of the magnitude of arsenic, by using 54 variables of chemical elements based on secondary data. The samples are collected from 113 wells from different areas of Bangladesh. The study shows that 9 variables are significantly related with the arsenic variable. Analysis also claims that arsenic variable can be modeled with three principal components (a linear combination of independent variables). The study also considered the latitude and longitude variables but there is no spatial autocorrelation between them.*

Key words: *regression model; principal component; spatial autocorrelation; arsenic contamination; chemical elements*

1. Introduction

Arsenic contamination of the groundwater is a major threat to public health in Bangladesh and West Bengal. Groundwater often provides water supply that is more reliable in quantity and more stable in quality than the surface water and thus it has economic and operational advantages due to reduced treatment requirements [8][6]. Until the early 1970s, more than 100 million inhabitants of Bangladesh and the neighbouring West Bengal drank from shallow hand-dug wells, rivers and ponds. These sources of water are generally polluted and account for various water-borne diseases such as diarrhoea, polio, typhoid, amebiasis etc. In order to provide safe drinking water, the government of Bangladesh, international agencies such as UNICEF and various non-government organizations (NGO's) were engaged in setting up tube wells - steel pipes fitted for simple hand pumps in order to tap the plentiful and apparently clean water in the sand and silt of the Ganges flood plain. There are about 3-5 millions tube wells present today, whereas they were only about 50,000 during the British colonial rule [10]. Groundwater provides safe drinking water to over 97% of the rural population in Bangladesh. This extensive coverage is an indication of the country’s successful attempt as to provide safe drinking water to its main population. However, the recent discovery of arsenic in the groundwater has ruined this decade-long success, so the access to safe drinking water has dropped to almost 80% [10].

Arsenic is well-known for its toxicity and carcinogenicity. Clinical effects of arsenic mainly include Keratosis and Melanosis. Other clinical manifestations include body related disturbances, cerebral inffection, cangrene, muscular atrophy, depressive state, auditory problem and various neurological symptoms. The acceptable source of arsenic is the geological route, as it was transported by rivers from the sedimentary rocks in the Himalayas during tens of thousands of years. On the other hand, the anthropogenic sources are other unwanted sources of arsenic. In average, 27% of the shallow tube wells in Bangladesh are producing water with arsenic in excess in comparison to the Bangladesh standard of 0.05mg/l for drinking water.

Non-cancer effects of arsenic can include thickening and discoloration of the skin, stomach pain, nausea, vomiting, diarrhoea, blindness, partial paralysis and numbness in hands and feet. Arsenic has been linked to cancer of the bladder, lungs, skin, kidney, nasal passages, liver and prostate

Bangladesh is a very small country with a total area of 147,570 sq.km and the total population accounts for about 150 millions people. It is the most densely populated country in the world. It is bordered on most sides by India and by Myanmar in the Southeast. Bangladesh has a tropical monsoon climate with a high annual rainfall of 1000-2000 mm or more, falling mainly during the period June- September. Bangladesh has a large area of surface waters taking the form of the major Padma (Ganges), Jamuna (Brahmaputtra), Meghna Rivers and their tributaries. [12]

The number of estimated tube-wells in Bangladesh is around 6-11 millions. The vast majority of these are private tube-wells that penetrate the shallow alluvial wells to depths typically of 10-60 m. In the south and in the Sylhet Basin from northeast Bangladesh, deep tube-wells abstract groundwater from depth of 150m or more. In the south, the tube wells have been installed as to avoid high salinity at shallower levels [1]. Shallow hand- dug wells occur in some areas although they are less common than tube wells.

Arsenic presence in the drinking water is a new, unfamiliar problem to the population in Bangladesh, including concerned professionals. There are millions of people who may be affected by drinking arsenic-rich water. The fear for future adverse health effects is a major concern because of the water already being consumed. It has been suggested by WHO that there are between 8-12 million shallow tube-wells in Bangladesh. Piped water supplies are available only to a little more than 10% of the total population living in the large agglomerations of some district towns while up to 90% of the Bangladesh population prefer to drink well water. Until the discovery of arsenic in the groundwater in 1993, well water was regarded as safe for drinking. It is now generally agreed that the arsenic contamination of groundwater in Bangladesh is of geological origin. The arsenic derives from the geological strata underlying Bangladesh. Over the next decade, skin and internal cancers are likely to become the principal human health concern arising from arsenic. According to one estimate, at least 100,000 cases of skin lesions caused by arsenic have occurred and there may be many more. [9]

The number of people drinking arsenic-rich water in Bangladesh has increased dramatically since the 1970s due to well-drilling and population growth. The impact of arsenic extends from immediate health effects to extensive social and economic hardship that affect especially the poor. The costs of health care, the inability of the affected persons to engage in productive activities and the potential social exclusion are important factors to be taken into account.

Several factors, significantly related to arsenic, need to be selected through our proposed model. The level of arsenic in the ground water varies from region to region, in relation with the depth of the tube wells and of other factors, such as the presence of different elements in the ground water, like aluminium, iron, phosphorus, manganese etc. Our interest is to identify the factors that are accountable for the magnitude of arsenic. Although the dimensions of the problems associated with arsenic in Bangladesh are enormous, public awareness of the overall extent of contamination is limited.

The objective is to study the relationship among extent of arsenic, region, depth of tube-wells and other elements in groundwater, and then to conceive a model describing this relationship.

Lot of works have been done on various dimensions of arsenic-related problems at home and abroad, but only few of them are related with fitting the model for the relationship with arsenic magnitude in groundwater and chemical elements. Arsenic, lead, and cadmium contamination in soil samples show strong and statistically high significant correlations for all contaminant pairs. Spearman rank correlations (a nonparametric approach to correlation analysis, applicable to any data distributions) are all significant at $p < 0.0001$; the rank correlations are 0.86 for arsenic and lead, 0.74 for arsenic and cadmium, and 0.74 for lead and cadmium. (The results involving cadmium are, of course, affected by the substitution of one-half of the detection limit for all not detected results). [2]

Arsenic, which is naturally present in soil, can be mobilized and transported, leading to increased concentrations of As in aquifers, that are sources of drinking water [3]. The largest contemporary known mass exposure to it is occurring due to the consumption of tube-well water throughout the Ganges-Brahmaputra Delta in Bangladesh and India. In Bangladesh alone, this exposure is affecting approximately 25–30 million residents. A survey of roughly 6,000 contiguous wells in Araihazar, Bangladesh, reported well-water As concentrations ranging from less than 5 to 860 μg/L. [11]

An approach is described for viewing the interrelationship between different variables and also tracing the sources of pollution of groundwater of north Chennai (India). It was applied the linear regression model (LRM) with correlation analysis in order to to check its validity for prediction of metal speciation and to apply LRM for rapid monitoring of water pollution [6]. An important component of quantitative risk assessment involves characterizing the dose-response relationship between an environmental exposure and adverse health outcome and then computing a benchmark dose, or the exposure level that yields a suitable low risk. This task is often complicated by model choice considerations, because risk estimates depend of the model parameters. A study proposed by the Bayesian methods is meant to address the problem of the model selection and to derive a model-averaged version of the benchmark dose. They illustrate the methods through application to data on arsenic-induced lung cancer in Taiwan. [7]

2. Data and variables

Secondary sample data have been downloaded from the Bangladesh Water development Board web site *http://www.bgs.ac.uk/arsenic/bangladesh/Data/BWDBSurvey Data.csv*. The data were collected from different areas of Bangladesh excepting the hilly regions.

There were 56 variables and 113 wells. Most of the variables are geological elements in ground water in Bangladesh. The dependent variable was Arsenic which was censored at 0.05 ug/l. The independent variables are latitude, longitude, depth, and geological elements in ground water including Al-Aluminum, As- Arsenic, B-Boron, Ba-Barium, Be-Beryllium, Ca-Calcium, Cd-Cadmium, Ce-Cerium, Cl-Chlorine, Co-Cobalt, Cr-Chromium, Cs-Cesium, Cu-Copper, Dy-Dysprosium, Er-Erbium, Eu-Europium, Fe-Iron, F-Fluorine, Gd-Gadolinium, HCO3- Bicarbonate, Ho- holmium, I-Iodine, K- Potassium, La-Lanthanum, Li-Lithium, Lu-Lutetium, Mg- magnesium, Mn- Manganese, Mo- Molybdenum, Na- Sodium, Nd- Neodymium, $_{NH4N}$, N- Nitrogen, i NO_2N, NO_3N, P- Phosphorus, Pb- Lead, Pr- Praseodymium, Rb- Rubidium, Sb- Antimony, Si- Silicon, Sm- Samarium, Sn- Tin, SO_{4-} Sulphate, Sr- Strontium, Tb- Terbium, Tl- Thallium, Tm- Thulium, U- Uranium, V- Vanadium, Y- Yttrium, Yb- Ytterbium and Zn- Zinc. Some of the independent variables were also censored, but in this analysis we will consider them as the observed value at censored limit. For example, cadmium (Cd) has a censored value of 0.02 ug/l but we will take 0.02 ug/l as an observed value.

3. Methodology

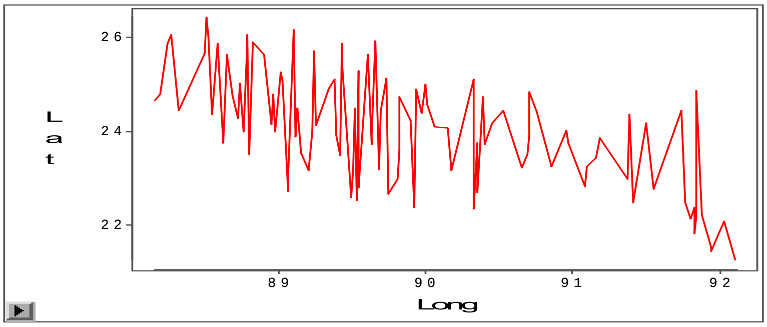
There is a moderate portion of missing values in the data set. To avoid this difficulty, the expectation–maximization (EM) method is used for inputing missing data. The multiple linear regression technique has been applied to modelling the extent of arsenic with various independent variables. There are too many input variables in comparison to the total number of observations in the data set. Regression based on Principal Component Analysis is used to overcome the problem of the large number of input variables. The spatial autocorrelation technique has been performed to find out the possible spatial correlations between the positions of the wells.

There were used automatic model selection procedures, such as the backward elimination method. The software used throughout the analysis was *Microsoft Excel, SAS* and the *S-Plus* module *Spatial*. Excel was used for simple data transformation (i.e. from Mg/L to μ_g/L). The SAS was used for modeling – as there are the Principal Component Analysis and the Multiple Regression. The main procedures from the *SAS* software system were *Lifereg*, *Princomp* and *Reg* procedures. The module *Spatial* in *S-Plus* was used for analyzing spatial correlation, Spatial Autocorrelation and for filling in the missing data by using the Expectation- Maximization Method.

4. Results and Discussions

Possible Spatial Autocorrelation

Spatial data is different when comparing with standard data analysis, especially when spatial information and predictions in the model are being included. Data are often correlated with space. Spatial structure can arise from several different sources, such as measurement error, continuity effects including spatial heterogeneity and space-dependent processes or mechanisms.



The graphic above shows the relationship between latitude and longitude.

The data set contains the latitudes and longitudes of the tested wells. The data have been analyzed for testing autocorrelation, by using the Moran and Greay test statistic. Depending on the outcome of the test, we would decide to either apply spatial modeling or disregard it in case the spatial autocorrelation turns out to be negligible. In the former case, we would fit a Spatial Linear Model describing the dependence of arsenic on the spatial locations [5]. After fitting the model, we would consider the residuals and would fit a standard regression model describing the effects of the remaining dependent variables on the residuals. However, in the alternative case that there was no significant autocorrelation exhibited, we would just include the latitude and longitude variables along all other variables in our regression model without trying to eliminate first their spatial autocorrelation effect.

In our analysis, firstly we have calculated the spatial neighbour and then, with this spatial neighbour, we calculated spatial correlation on arsenic and found that there was no spatial autocorrelation (i.e. spatial autocorrelation is 0.002 and it is statistically insignificant). Therefore, latitude and longitude variables were left in the model, like all the other variables.

The Multiple Regression Model

In our data there are not too many censored variables. This motivated us to also use the standard Multiple Regression Analysis as to the effect of the independent variables on Arsenic and to compare the results. The multiple regression model has the general form:

$$Y = X\beta + \varepsilon$$

where:

-*X* it is a *n x p* matrix of observation on the predictor variables

-*Y* it is a *n x 1* vector of response measurement

-ε it is a *n x 1* random vector; the error term. [4]

The output of the model by using SAS is as follows:

Table 1. Analysis of variance

Source	Degrees of Freedom (df)	Sum of Squares	Average Sum of Squares	F- value	p- value
Model	9	339739	37749	13.05	<0.0001
Error	103	297838	2891.63060		
Corrected total	112	637577			

Table 2: Estimation of parameters and its test result

Variable	Parameter Estimate	Standard Error	Type II SS	t-value	p- value
Intercept	-22.79033	18.18662	4540.87759	1.57	0.2130
B	-0.19611	0.04764	49001	16.95	<0.0001
Dy	-18525	2947.58734	114215	39.50	<0.0001
Er	19816	3536.06507	90814	31.41	<0.0001
Fe	0.00348	0.00123	22979	7.95	<0.0058
K	-0.00362	0.00068856	79710	27.57	<0.0001
Mg	0.00102	0.00026434	43075	14.90	<0.0002
Mn	-0.03408	0.00786	54294	18.78	<0.0001
Mo	39.93068	5.41908	157002	54.30	<0.0001
P	0.06359	0.01133	91059	31.49	<0.0001

The final fitted model based on the multiple regression approach is

$$As = -22.790 - 0.196B - 18525Dy + 19816Er + .0035Fe - .0036K + .00102Mg - 0.0341Mn + 39.931Mo + 0.064P$$

From this model it clearly results that out of 54 variables 9 of them are highly significant for the arsenic contamination level.

Multiple Linear regression with PC variables

When using principal components regression, we transform first the predictor variables into principal components. Then we regress the output on the principal components. The scaling option for PC is *Z = X U*

where $U'U = I$.

To implement principal components regression, we first transform the original data into PC's. Then we obtain

$Y = Zb_z$

where b_z denotes the regression coefficient obtained by using principal components. [4]

The reason to use principal components is that many of the predictors exhibit multicollinearity, which has an effect on the rank of the design matrix and it causes difficulties in the calculation of the inverse of $X'X$ when calculating the smallest squares estimators. Another related reason is that when there are too many potential independent variables (like in our case 56 of them), we would like to reduce the dimensionality of the problem by involving only a small number of principal components as regressors, instead of using the large number of all independent variables.

Since the new predictors based on the principal components are not correlated, the resulting regression coefficients will be not correlated also. PCR will predict the response with the exact same precision as OLS when all of the pc's are used. However we are interested only in including a relatively small number of relevant principal components.

As already pointed out, further benefit of using PCR is the simplification of the model since a small number of pc's are used, because the pc's are so readily interpretable that they became the new variables in the prediction model. In case they are not interpretable at all, one can still relate the responses to the original prediction, as follows:

$Y = Zb_z = XUb_z$

Since $Z = XU$ and therefore

$b = Ub_z$

and

$$b_z = [U'X'XU]^{-1}U'X'Y$$

then

$$b = U[U'X'XU]^{-1}U'X'Y \quad [4]$$

The output of the model by using SAS is as follows:

Table 3. Analysis of variance

Source	Degrees of Freedom (df)	Sum of Squares	Average Sum of Squares	F- value	p- value
Model	3	192835	64278	15.75	<0.0001
Error	109	444742	4080.20423		
Corrected total	112	637577			

Table 4. Estimation of parameters and its test result

Variable	Parameter Estimate	Standard Error	Type II SS	t-value	p- value
Intercept	30.66283	6.00899	106244	26.04	<0.0001
Prin3	8.68376	2.87639	37188	9.11	<0.0032
Prin12	23.41578	5.523.4	73340	17.97	<0.0001
Prin13	-25.91756	5.77055	82307	20.17	<0.0001

There are 54 variables in our analysis. By Using Princomp command in SAS we get 20 principal components which cover more than 91% of the variation of the output. Again, with 20 principal components we ran multiple regression model in SAS by Reg command with Backward eliminations method and the first 8 principal components were found significant at 5% level of significance and finally got only three of them as highly significant at 1% level of significance. The fitted model is

As = 30.663+ 8.684prin3 + 23.416prin12 – 25.918prin13

5. Conclusion

In this study two regression models (multiple regression and multiple regression with PC variables) have been applied in order to study the effect of some chemical components on groundwater arsenic contamination level. The two models are found to be highly significant. Only 9 variables in the multiple regression model and 3 principal components in the multiple regression with PC variables are highly significant. So, further research focusing on these variables will be helpful to explore the groundwater arsenic contamination problem related with chemical elements.

References

1. Glass, G. L. **Public health – Seattle & King Country Final Report Vashon/Maury Island Soil Study 1999-2000**, Prepared by Environmental Health Division Public Health - Seattle & King County and Environmental Consultant, Seattle, Washington, JULY 2000, p33, Web Address: http://www.metrokc.gov/health/hazard/finalrpt72500.pdf , (29 August 2007)
2. Harvey, C.F., Swartz, C.H., Badruzzaman, A.B., Keon-Blute, N., Yu W, Ali, M.A., *et al*. **Arsenic mobility and groundwater extraction in Bangladesh,** Science, 298, 2002, pp. 1602–1606
3. Jackson, J. and Edward ***A* User's Guide to Principal Components**, Wiley Interscience, Canada, 2003
4. Kaluzny, S., Vega, S., Cardoso, T. and Shelly, A. **S+ Spatial Stats user's Manual for Windows and Unix**, Springer, New York, 1997
5. Kumaresan, M. and Riyazuddin., P. **Factor analysis and linear regression model (LRM) of metal speciation and physico-chemical characters of groundwater samples**, Environ Monit, Assess, 2007, http://www.springerlink.com/content/v10x8467vmx424x0/ (29 August 2007)
6. Morales, H.K., Ibrahim, G. J., Chen,. J. C., and Ryan, M. L. **Bayesian Model Averaging with applications to Benchmark Dose Estimation for Arsenic in Drinking Water***,* Journal of the American Statistical Association, 2006. Vol. 101, No. 473
7. Robins, N.S., **Hydrology of Scotland**, HMSO, London, 1990
8. Smith, A. H., Lingas, E.O. and Rahman, M. **Contamination of drinking-water by arsenic in Bangladesh: a public health emergency**, Bulletin of the World Health Organization, 78 (9), 2000
9. Van Geen, A., Ahsan, H., Horneman, A. H., Dhar, R. K., Zheng, Y., Hussain, I., et al. **Promotion of well-switching to mitigate the current arsenic crisis in Bangladesh**. Bull WHO, 80, 2002, pp. 732–737
10. * * * British Geological Survey (BGS) **Arsenic Contamination of Groundwater,** 1998, http://www.bgs.ac.uk/arsenic (15 August 2007)
11. * * * UNICEF **Arsenic mitigation in Bangladesh: media brief**, UNICEF, Bangladesh, 1999
12. * * * World health organization (WHO) **Arsenic in Bangladesh**, http://www.who.int/mediacentre/factsheets/fs210/en/index.html. 2001. (1 August 2007)

[1] Mr. Syed Md Fakrul Ahsan, govt. employee at Railcorp in Sydney, Australia. he has completed Bachelor of Statistics in 1992 and Masters in Statistics in 1993, from Jahangirnagar University, Dhaka, Bangladesh. He has also a graduate of the University of New South Wales, Sydney, successfully completing a Masters degree in 2007. His research interests has lead to write two project reports entitled 'Ground Water studies of contamination of Arsenic in Bangladesh' and 'Social Behaviour Pattern of the students of Jahangirnagar University' throughout his postgraduate

studies. Furthermore, he jointly prepared a research article on 'Nutritional Status of Children Under Six, at the Industrial Area in Bangladesh' which was published in the Asian Network for Scientific Information, in Pakistan.

[2] Mr. Md. Nazrul Islam is an Assistant Professor, Department of Statistics, Shahjalal University of Science & Technology, Sylhet, Bangladesh. He born in 1972 and graduated and completed Master program from Jahangirnagar University, Dhaka, Bangladesh. He joined as a lecturer in Shahjalal University in 1997 and now he is in Department of Statistics, Gauhati University, India as a Ph.D scholar. He has published several research articles in national and international journals. His research interests are: Statistical modeling, nonparametric statistics, Environmental Statistics, Social Statistics and Demography. Currently he is working with Population Aging in Bangladesh.

[3] Corresponding Author

[4] Mr. Md. Jamal Uddin, Lecturer (faculty) in Department of Statistics, Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh. He has completed his Bachelor (B.Sc, with CGPA-3.57 out of 4, credits-120, 2003) and Master Degree (M.Sc., Thesis group, with CGPA-3.79 out of 4, credits- 40, 2004) in Statistics from SUST. It is also mentionable here that he has completed his M.Sc. thesis entitled "Differentials and Determinants of Infant and Child Mortality in Bangladesh". His research interest are Biostatistics/Biomedicine/Public Health, Statistical modeling. Currently he is working in medical science data and some related articles will be published very soon.

[5] Mr. Mohammed Taj Uddin has completed M.Sc. Concordia University, Canada, Statistics (with thesis), 2004, M.Sc. (Statistics), Shahjalal University of Science and Technology (SUST), Bangladesh, First class (in position first) 1995 and B. Sc. (Statistics and Mathematics, Economics and physics) Shahjalal University of Science and Technology, Bangladesh, First class (in position first) 1994. His Research interest: Biostatistics, Medical statistics, statistical inference and econometrics.

[6] Codification of references:

[1]	* * * British Geological Survey (BGS) **Arsenic Contamination of Groundwater,** 1998, http://www.bgs.ac.uk/arsenic (15 August 2007)
[2]	Glass, G. L. **Public health – Seattle & King Country Final Report Vashon/Maury Island Soil Study 1999-2000**, Prepared by Environmental Health Division Public Health - Seattle & King County and Environmental Consultant, Seattle, Washington, JULY 2000, p33, Web Address: http://www.metrokc.gov/health/hazard/finalrpt72500.pdf , (29 August 2007)
[3]	Harvey, C.F., Swartz, C.H., Badruzzaman, A.B., Keon-Blute, N., Yu W, Ali, M.A., *et al.* **Arsenic mobility and groundwater extraction in Bangladesh,** Science, 298, 2002, pp. 1602–1606
[4]	Jackson, J. and Edward ***A* User's Guide to Principal Components**, Wiley Interscience, Canada, 2003
[5]	Kaluzny, S., Vega, S., Cardoso, T. and Shelly, A. **S+ Spatial Stats user's Manual for Windows and Unix**, Springer, New York, 1997
[6]	Kumaresan, M. and Riyazuddin., P. **Factor analysis and linear regression model (LRM) of metal speciation and physico-chemical characters of groundwater samples**, Environ Monit, Assess, 2007, http://www.springerlink.com/content/v10x8467vmx424x0/ (29 August 2007)
[7]	Morales, H.K., Ibrahim, G. J., Chen,. J. C., and Ryan, M. L. **Bayesian Model Averaging with applications to Benchmark Dose Estimation for Arsenic in Drinking Water***,* Journal of the American Statistical Association, 2006. Vol. 101, No. 473
[8]	Robins, N.S., **Hydrology of Scotland**, HMSO, London, 1990
[9]	Smith, A. H., Lingas, E.O. and Rahman, M. **Contamination of drinking-water by arsenic in Bangladesh: a public health emergency**, Bulletin of the World Health Organization, 78 (9), 2000
[10]	* * * UNICEF **Arsenic mitigation in Bangladesh: media brief**, UNICEF, Bangladesh, 1999
[11]	Van Geen, A., Ahsan, H., Horneman, A. H., Dhar, R. K., Zheng, Y., Hussain, I., et al. **Promotion of well-switching to mitigate the current arsenic crisis in Bangladesh**. Bull WHO, 80, 2002, pp. 732–737
[12]	* * * World health organization (WHO) **Arsenic in Bangladesh**, http://www.who.int/mediacentre/factsheets/fs210/en/index.html. 2001. (1 August 2007)