Stoch Environ Res Ris Assess
DOI 10.1007/s00477-006-0098-6

ORIGINAL PAPER

Is correlation dimension a reliable proxy for the number of dominant influencing variables for modeling risk of arsenic contamination in groundwater?

Jason Hill · Faisal Hossain · Bellie Sivakumar

Abstract The correlation dimension (CD) of a time series provides information on the number of dominant variables present in the evolution of the underlying system dynamics. In this study, we explore, using logistic regression (LR), possible physical connections between the CD and the mathematical modeling of risk of arsenic contamination in groundwater. Our database comprises a large-scale arsenic survey conducted in Bangladesh. Following the recommendation by Hossain and Sivakumar (Stoch Environ Res Risk Assess 20(1–2):66–76, 2006a), who reported CD values ranging from 8 to 11 for this database, 11 variables are considered herein as indicators of the aquifer's geochemical regime with potential influence on the arsenic concentration in groundwater. A total of 2,048 possible combinations of influencing variables are considered as candidate LR risk models to delineate the impact of the number of variables on the prediction accuracy of the model. We find that the uncertainty associated with prediction of wells as safe and unsafe by LR risk model declines systematically as the total number of influencing variables increases from 7 to 11. The sensitivity of the mean predictive performance also increases noticeably for this range. The consistent reduction in predictive uncertainty coupled with the increased sensitivity of the mean predictive behavior within the universal sample space exemplify the ability of CD to function as a proxy for the number of dominant influencing variables. Such a rapid proxy, based on nonlinear dynamic concepts, appears to have considerable merit for application in current management strategies on arsenic contamination in developing countries, where both time and resources are very limited.

Keywords Nonlinear deterministic dynamics and chaos · Correlation dimension · Arsenic contamination · Logistic regression · Groundwater · Bangladesh

J. Hill
Department of Civil and Environmental Engineering,
Tri-State University, 1 University Avenue, Angola
IN 46703, USA

F. Hossain (✉)
Department of Civil and Environmental Engineering,
Tennessee Technological University, Box 5015, Cookeville,
TN 38505-0001, USA
e-mail: fhossain@tntech.edu

B. Sivakumar
Griffith School of Engineering, Griffith University, Nathan,
QLD 4111, Australia

1 Introduction

Since the large-scale discovery of arsenic contamination in the alluvial Ganges aquifers of Bangladesh, numerous studies have been conducted to better understand the spatial variability of the contamination scenario (e.g., Biswas et al. 1998; Burgess et al. 2000; McArthur et al. 2001, 2004; Harvey et al. 2002; Mukherjee and Bhattacharya 2002; van Geen et al. 2003; Yu et al. 2003; Ahmed et al. 2004; Hossain et al. 2006a, b). Most of these studies have addressed the 'spatial' pattern of arsenic using geo-statistical tools and the classical notion of linear stochastic dynamics. For example, in the first country-wide study toward spatial (horizontal) characterization of the arsenic calamity, conducted by the British Geological Survey (BGS) in collaboration with the Department of Public Health and Engineering (DPHE) of Bangladesh (hereafter

Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 ☐ LE ☐ TYPESET
MS Code : 98 ✔ CP ✔ DISK

called ‘BGS-DPHE’), an application of kriging (Journel and Huijbregts 1978) was reported to provide the ‘best’ estimate of the whole nation’s arsenic field at the regional scale with limited sampling information. The BGS-DPHE investigation involved the assumption that the arsenic concentration could be treated as a ‘regionalized’ linear stochastic random variable in space.

It must be noted, however, that arsenic in groundwater is not a purely random occurrence and that (hidden) order and determinism may also exist, just as they do in any other natural or man-made phenomenon. Arguing that there existed profound geological and geochemical factors, with possible order, controlling arsenic contamination dynamics (for details, see Hossain and Sivakumar 2006a; McArthur et al. 2004; Zheng et al. 2004), Hossain and Sivakumar (2006b) suggested that it was no longer defensible for the scientific community to continue to use purely geostatistical (linear stochastic) approaches as stand-alone techniques for its spatial interpolation. Our understanding of the role played by these physical factors in arsenic contamination of groundwater continues to be enhanced from recent studies by, for example, Zheng et al. (2004), Akai et al. (2004) and Ahmed et al. (2004). Traditional geostatistical tools are a ‘pattern-filling’ scheme based on the spatial correlation exhibited by two points in space separated by a lag ***h***. This approach simplifies the spatial patterns manifested by the complex interactions between geology and time-sensitive fluid flow dynamics (Christakos and Li 1998). Concerns on the use of purely stochastic approaches and potential for alternative ones have been echoed by a few other studies as well (e.g., Faybishenko 2002; Sivakumar 2004a; Sivakumar et al. 2005).

On the premise that the current ensemble of proposed ‘theories’ in scientific literature explaining arsenic mobility (e.g., Burgess et al. 2000; McArthur et al. 2001; Harvey et al. 2002; van Geen et al. 2003) can, in principle, be mathematically represented as the cumulative effect of a finite number of dominant processes comprising three or more partial differential equations, Hossain and Sivakumar (2006a) verified the existence of nonlinear deterministic and chaotic dynamic behavior in the spatial pattern of arsenic contamination in shallow wells (depth < 150 m) in Bangladesh. Employing the Grassberger–Procaccia correlation dimension (CD) algorithm (Grassberger and Procaccia 1983), their analysis revealed CD values (i.e., saturation of correlation exponents and a manifestation of ‘determinism’) ranging anywhere from 8 to 11 depending on the region and geology (see, for example, Fig. 1). Their findings suggested that the

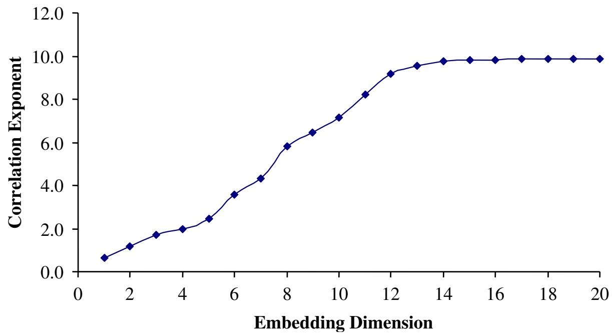


Fig. 1 Relationship between Correlation Exponent and Embedding Dimension for the whole Bangladesh based on BGS-DPHE (2001) arsenic data from shallow wells (after Hossain and Sivakaumar 2006a)

arsenic contamination dynamics in space, from a chaotic dynamic perspective, was a medium- to high-dimensional problem. While it is encouraging to note that the nonlinear CD analysis can reflect the influence of regional geology (and other factors) on arsenic contamination dynamics, the usefulness of the CD and other nonlinear deterministic dynamic techniques to understand the physics of the actual arsenic contamination phenomenon is far from clear, as explained next.

It is well known that the CD of (an attractor of) a time series generally provides information on the number of variables present in the evolution of the underlying system dynamics (e.g., Grassberger and Procaccia 1983; Hao 1984; Fraedrich 1986; Sivakumar 2004b; Hossain and Sivakumar 2006a). However, current environmental literature is largely insufficient in the context of providing links between the CD and the actual physical mechanisms that take place in catchments/aquifers. While some studies have indeed conducted research in this direction, such have essentially been limited to the verification of the reliability of the CD estimate, and especially performed using nonlinear predictions of the respective time series. For example, Sivakumar et al. (2002c) investigated the reliability of the CD estimate of the monthly flow data observed at the Coaracy Nunes/Araguari River watershed in northern Brazil (see also Sivakumar et al. 2001a), using nonlinear local- (chaos theory-based) and global- (artificial neural networks-based) approximation techniques. The study, in fact, focused on the reliability of the CD in the context of short time series, since the data size requirement has been the primary subject of criticism on the reports of low-dimensional chaos in environmental time series (e.g., Ghilardi and Rosso 1990; Schertzer et al. 2002; see also Sivakumar 2000, 2005; Sivakumar et al. 2002a, for details). Similarly, nonlinear predictions of time series have served as the

Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 LE TYPESET
MS Code : 98 CP DISK

basis, implicitly or explicitly, for verification of the CD estimate in other studies as well, albeit in different forms (e.g., Porporato and Ridolfi 1997; Lambrakis et al. 2000; Sivakumar et al. 2001b, 2002b).

With the encouraging results of their preliminary analysis (Hossain and Sivakumar 2006a) regarding the nonlinear deterministic nature of arsenic contamination, Hossain and Sivakumar (2006b) subsequently discussed the potential role the nonlinear deterministic dynamic and related concepts can play in improving our understanding of arsenic contamination patterns in space. They especially highlighted their potential utility in providing improved cost-effectiveness of environmental management in rural and resource-limited settings of developing countries, such as Bangladesh, Vietnam and India. In a related development, Serre et al. (2003) have reported that the spatial interpolation of arsenic contamination, if approached from the conventional paradigm of geostatistical mapping, can be challenging in Bangladesh as most of the variability in arsenic concentration occurs within short distances (2–5 km). Certainly acknowledging the fact that the traditional linear stochastic approaches had generally yielded fairly good and reliable results, Hossain and Sivakumar (2006b) also called for a much-needed change in the current state-of-the-art for spatial interpolation of arsenic contamination, stating that: 'While there is no structural, or even philosophical, flaw in using the conventional geo-statistical approach, there is indeed ample room to argue that the geo-statistical treatment of arsenic contamination in space as a regionalized random (or stochastic) variable may constitute only an incomplete analysis of its spatial variability (even if system-dependent). Incompleteness can potentially arise from the fact that geo-statistics often fails to recognize the random looking but deterministic behavior that may be present due to self-similar (scale-invariant) factors in the continuum of the sub-surface.'

In essence, Hossain and Sivakumar (2006b) argued for the need to couple/integrate the linear and nonlinear concepts/tools, whenever and wherever deemed necessary or appropriate [see also Sivakumar (2004b) for an example of possible integration of different concepts/methods for environmental modeling]. This, however, is easier said than done, since there is still some convincing needed, going by the criticisms, on the utility of the relatively new nonlinear deterministic dynamic concepts for arsenic contamination and other environmental problems in the first place. Roughly speaking, the nonlinear analyses and results need to be verified using the conventional linear techniques, so as to first bring reconciliation between linear and nonlinear concepts and then to bridge the gap between them. With particular reference to the study by Hossain and Sivakumar (2006a), this should obviously start with the verification of the CD values obtained for the arsenic concentration data using any of the available linear tools.

In this spirit, we herein explore possible physical connections between the CD and the mathematical modeling of risk of arsenic contamination in groundwater by applying (the linear) logistic regression (LR) risk assessment technique. Using 11 potentially influencing variables that largely define the geochemical regime of aquifers and, hence, the variability of arsenic concentration, we attempt to provide a possible insightful evidence that the CD can be a proxy for the number of dominant influencing variables required in an LR risk model to optimally predict risk of arsenic contamination at non-sampled wells. To the best of our knowledge, such an insight, although preliminary, constitutes an important finding, with potential implications on the reduction of uncertainty of risk maps produced from conventional (linear stochastic) paradigms. Even though we pursue this task primarily from a data-based perspective, a larger goal of our mission is to encourage greater interactions with the research community traditionally engaged in a more mechanistic understanding of arsenic contamination. We believe that such interactions can play a vital role in the integration of non-linear deterministic dynamic concepts in future groundwater management protocols (discussed in detail later in the paper). In the sections that follow, we provide a systematic overview of our exploratory research to understand the value of CD in modeling risk of arsenic contamination.

2 Study region, data, and CD analysis

We choose to study arsenic contamination over the entire region of Bangladesh, as had been first surveyed by the BGS-DPHE (2001) study comprising 3,534 wells. This is conducted in the manner similar to Hossain and Sivakumar (2006a) for estimating the CD values. The dataset is available (at the time of writing this manuscript) at http://www.bgs.ac.uk/arsenic/bangladesh/datadownload.htm. Wells deeper than 150 m (and consistently below the safe limits) are excluded from the analysis, thus resulting in a set of 3,085 shallow wells. While it is possible that such an exclusion of data based on depth may incur an added bias to our analyses on the application of CD, we believe, to the best of our knowledge, that the impact would be insignificant to alter the overall conclusions of our study, particularly when our goal is to demonstrate a proof-of-concept application of CD in deterministic

Journal : 477 Article No. : 98 MS Code : 98
Dispatch : 20-11-2006 Pages : 9
☐ LE ☑ CP ☐ TYPESET ☑ DISK

modeling. For details on the study region and data, the reader is referred to the works of Hossain et al. (2006b) and Hossain and Sivakumar (2006a).

The CD method employed by Hossain and Sivakumar (2006a) used the correlation integral or function (Grassberger and Procaccia 1983) for distinguishing between chaotic and stochastic behaviors (more specifically, between low- and high-dimensional systems). Although, traditional applications of the phase-space reconstruction and the Grassberger–Procaccia algorithms have been limited to data series in the continuum of time (e.g., Takens 1981; Theiler 1987; Rodriguez-Iturbe et al. 1989; Porporato and Ridolfi 1997; Sivakumar et al. 2001b, 2002c, 2005), Hossain and Sivakumar (2006a) argued that there was no compelling logic that disqualified its application to a data series in space. Their CD analysis revealed positive evidence regarding medium-to-high dimensional chaotic dynamics in arsenic contamination in space, with a country-wide dimension value ranging between 8 and 11. This subsequently led Hossain and Sivakumar (2006a, b) to comment subjectively that the minimum number of variables and hence the number of dominant processes required to model the spatial variability of arsenic contamination should also range from 8 to 11.

It is appropriate to mention, at this point, that questions may be raised regarding the suitability of this data set for CD analysis. Such questions may be related to, among others, the data size (insufficient length) and data quality (presence of noise), as these could potentially influence the CD estimation (e.g., Nerenberg and Essex 1990; Schreiber and Kantz 1996). These issues, and also others, have been and continue to be extensively discussed and debated in the literature, including in the environmental sciences [e.g., Ghilardi and Rosso 1990; Tsonis et al. 1994; Sivakumar et al. 1999, 2001b, 2002a, c; Sivakumar 2000, 2005; Schertzer et al. 2002; see also Sivakumar (2004a) for a review]. Due to space limitations, and also to avoid unnecessary deviation from the main focus of our study, we choose not to discuss such issues, and consequently direct the reader to the above studies and the numerous references therein. We, however, would like to briefly highlight a few points herein, in regards to the reliability of the CD estimates for this data set reported by Hossain and Sivakumar (2006a).

1. We are convinced that the data size, with 3,085 points, is more than sufficient to obtain reliable CD estimates of arsenic contamination in space. In this regard, we are particularly comforted by past studies that have reported reliable CD estimates for much smaller data sizes, albeit in the continuum of time (e.g., Sivakumar 2000, 2005; Sivakumar et al. 2002a, c).
2. While we do admit that the arsenic concentration data are likely contaminated with noise (e.g., measurement errors), we do not believe that it significantly influences our CD estimates [see, for example, Sivakumar et al. (1999)]. Even if it were to influence, the result would be only an overestimation of CD, not underestimation. Therefore, the interpretations and conclusions by Hossain and Sivakumar (2006a) regarding medium-to-high dimensional chaotic pattern would not only stand the test but also be more solidified.
3. Another factor possibly leading to underestimation of CD is the presence of a large number of zeros (or any one particular value) in the data set (e.g., Tsonis et al. 1994). Since there are no zeros (or repetition of a particular value) in the arsenic data set, this problem is also completely eliminated.

3 Logistic regression

The method of LR has been extensively used in epidemiological studies, and more recently, has become a common technique in environmental research on modeling risk of groundwater contamination (Twarakavi and Kaluarachchi 2006). Common regression techniques, such as the classical linear regression, relate the response variables to the influencing variables. LR relates the probability of a response variable to be greater than a threshold value (i.e., a risk) to a set of influencing variables (Afifi and Clark 1984; Helsel and Hirsch 1992). In an LR risk model, regression is linear between the natural logarithm of the odds ratio for the probability of response to be less than the threshold value and influencing variables. Equation 1 mathematically summarizes the LR model used in this study:

$$\ln[p/(1-p)] = \mathrm{logit}(p) = \alpha + \boldsymbol{\beta}\mathbf{x} \qquad (1)$$

where p is the probability of response to be greater than the safety threshold, α is a constant, $\boldsymbol{\beta}$ is a vector of slope coefficients, and $\mathbf{x}$ is a vector of influencing variables. For more details on the use of LR for modeling risk of arsenic contamination, the reader is referred to Twarakavi and Kaluarachchi (2006).

4 The potential influencing variables

Table 1 shows the influencing variables considered herein for defining the geochemical regime of aquifers.

Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 ☐ LE ☐ TYPESET
MS Code : 98 ✔ CP ✔ DISK

These variables were sampled by BGS-DPHE (2001) in Bangladesh. The minimum and maximum values of these variables (Table 1) indicate the range of variability across Bangladesh. The variables chosen are: (1) depth of wells (m), (2) P (Phosphorus) (mg/L), (3) Fe (Iron) (mg/L), (4) Ba (Barium) (mg/L), (5) Mg (Magnesium) (mg/L), (6) Ca (Calcium) (mg/L), (7) SO_4 (Sulfate) (mg/L), (8) Mean annual precipitation (mm/day), (9) Si (Silicon) (mg/L), (10) Na (Sodium) (mg/L), and (11) Mn (Manganese) (mg/L). Although our choice of variables is primarily dictated by literature reports on the causes of arsenic mobility (e.g., Welch et al. 2000; Harvey et al. 2002; van Geen et al. 2003; McArthur et al. 2004; Zheng et al. 2004) and the availability of reliable data, we must also point out to the reader that the selection herein is governed purely from a data-based and qualitative paradigm. As indicated earlier, the larger goal of our study is to encourage greater interactions between the research communities on mechanistic modeling of arsenic contamination and non-linear dynamic analysis. We admit that such a data-based selection without a deeper physical regard for the pertinent mechanics and geochemistry of contamination (as appropriate for Bangladesh) may have potential limitations. However, we also believe that such potential limitations alone should not hamper our ability to investigate the usefulness of the CD value, and particularly so when our intention is to primarily conduct a preliminary exploration. We believe that if there is a weakness in our choice of potential influencing variables, as may be revealed in our results, it only lends greater credibility to our mission in inviting the research community on arsenic contamination to interact more closely with the non-linear deterministic dynamic research community.

As a preliminary step, we first conduct the Spearman's Rank Correlation Coefficient test for these selected variables to identify their non-linear dependence with arsenic concentration. Because all possible combinations of influencing variables are considered during LR modeling of contamination risk (discussed next), results from the Spearman's test are not used in the ranking of the variables according to the order of influence. The precipitation data are obtained from the Bangladesh Meteorological Department (BMD) and Bangladesh Water Development Board (BWDB). The data are derived from a network of 100 recording rainfall gauges that registered less than 5% missing data for the year 2000. The choice of precipitation as an influencing variable is governed by reports that groundwater pumping for irrigation and recharge could be one of the causes of arsenic mobility in the shallow geologic stratum (see Harvey et al. 2002). Because recharge data are not readily available for our study, we choose mean rainfall as a proxy indicator of recharge of aquifers. For consistency, we select precipitation data pertaining to the year 2000 when the BGS-DPHE (2001) survey was completed. The mean annual rainfall value for each well is computed by the method of Thiessen Polygons using the ArcGIS™ software (Ormsby et al. 2004).

Table 1 The selected influencing variables for Logistic Regression Modeling

Variable	Mean	Minimum	Maximum
Well depth (m)	60.550	0.600	362.000
Ba (ppb)	87.340	2.000	1360.000
Ca (mg/L)	51.590	0.100	366.000
Fe (mg/L)	3.353	0.005	61.000
Mg (mg/L)	20.750	0.040	305.000
Mn (mg/L)	0.555	0.001	9.980
Na (mg/L)	88.936	0.700	2700.000
P (mg/L)	0.765	0.100	18.900
Si (mg/L)	20.519	0.030	45.200
SO4 (mg/L)	5.917	0.200	753.000
Annual precipitation (cm)	86.001	25.350	596.140
As (ppb)[1]	55.205	0.500	1660.000

[1] Arsenic (As) is the dependent variable in the LR risk model

5 Method of assessment

The dataset is divided randomly into two equal halves, with one half being employed for LR risk model calibration and the other half for validation. This random selection procedure is repeated 25 times within a Monte Carlo (MC) framework to assess the mean performance of the LR model. Using one-half of each randomly selected dataset, calibration of the LR model coefficients, α and $\boldsymbol{\beta}$, is performed using ordinary least squares technique for a safety threshold of 50 ppb (Bangladesh limit). In the calibration phase, the 'p' values in Eq. 1 are assigned 0–1 binary values depending on the measured concentration of arsenic ($p = 1$ for exceeding the safety threshold; $p = 0$ for being below the threshold). During the validation phase, the LR model is assessed in terms of its ability to successfully predict contamination in 0–1 binary terms according to the safety threshold at non-sampled wells (i.e., over the other half of the dataset not used in calibration of the LR risk models). For this, we employ the notion of contamination risk associated with a pre-assigned probability (i.e., in this case, $p = 0.9$). For example, if the well is predicted by the LR risk model as unsafe with $p = 0.85$ for a given safety threshold, then that well would be flagged uncontaminated according to the high risk criterion of $p = 0.9$. The

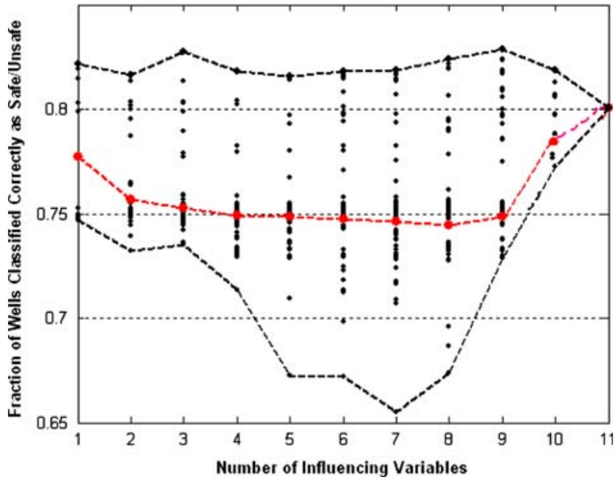
Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 ☐ LE ☐ TYPESET
MS Code : 98 ☑ CP ☑ DISK

predictive power of the LR risk model for a given number of influencing variables is quantified by the probability of successful detection of a well's status as contaminated or uncontaminated at untested well locations. It should be noted that the pre-assignment of a probability value to denote risk category as high(low) is purely subjective and will linearly scale up(down) the predictive behavior of LR model without altering the response pattern to the number of influencing variables. Hence, such a subjective assignment is considered acceptable within the overall scheme of our study as the objective is to delineate the impact of the number of potentially influencing variables and not on the LR risk model performance per se.

The specific question we explore, using LR, in our study is: 'Is CD a reliable proxy for the number of dominant variables required to predict risk of arsenic contamination in groundwater?' We consider all possible combinations of influencing variables from the total set of 11 as candidate LR models. This results in 2,048 LR risk models being evaluated. Each evaluation is repeated 25 times within the MC framework and the mean and range of LR model prediction assessed. For a given number of influencing variables, the mean signified the most probable LR model performance while the range is an indicator of predictive uncertainty to expect. It is important to note that the predictive uncertainty (or range) has important implications for model complexity and parameter optimization. The wider the uncertainty, the more challenging naturally would be the optimization to converge to the best LR model configuration. We discuss this in more detail in the next section.

6 Results and discussion

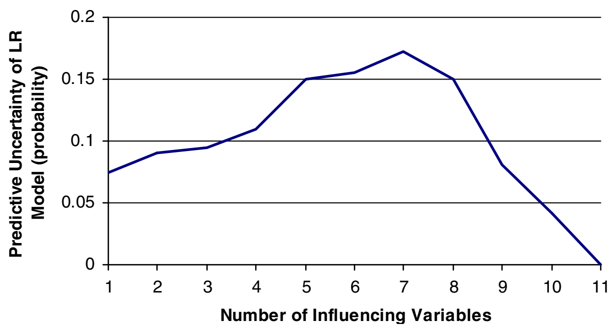
Figure 2 shows the variation of probability of successful detection of wells, or the fraction of validation set wells correctly detected (as contaminated/uncontaminated at the 50 ppb limit) as a function of the total number of influencing variables (Table 1) in the LR model. Basically, the terms 'contaminated/uncontaminated' or 'unsafe/safe' refer to the wells with arsenic concentration exceeding/less than 50 ppb. The mean predictive ability (shown in red circles, Fig. 2) of the LR risk model, while remaining insensitive to number of influencing variables in the ranges of 1–7 variables, is found to noticeably increase in sensitivity when the number of variables is greater than 7. A systematic reduction in the predictive uncertainty is also observed as the number of variables is increased from 7 to 11 (see Fig. 3). The probability of successful detection is



B & W IN PRINT

Fig. 2 Variation of fraction of wells correctly classified by LR model as safe/unsafe (i.e., probability of successful detection) with the number of influencing variables. The *larger black circles* with *dashed line* in the middle indicate mean values. The *upper* and *lower dashed lines* in *black* indicate the range of 25 Monte Carlo realizations for a given number of variables

shown for the mean of the 25 MC simulations on the *y*-axis of Fig. 2. Finally, we observe the best performance of the LR model when the number of influencing variables is 11. (Note that the lines all converge here to a point when the number of variables is 11 because the total number of possible LR model combinations is one. This observation should not be construed as an indication of no uncertainty for an LR model with 11 variables, but rather as an indication of the last point of complex modeling within a set of 11 variables where only one possible model can be constructed). As evident from Figs. 2, 3, an a priori inclusion of CD value in assigning the minimum LR model complexity appears to guarantee global optimization of the model configuration with a considerably higher degree of



B & W IN PRINT

Fig. 3 Predictive uncertainty in terms of probability of successful detection (i.e., the range between upper and lower limits in Fig. 2) as a function of the number of influencing variables. (Note: the value when the number of influencing variables is 11 should be ignored.)

Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 ☐ LE ☐ TYPESET
MS Code : 98 ✔ CP ✔ DISK

success. This empirical observation indicates consistency with the CD concept, according to which the inclusion of any additional variable deemed influential on the dynamics should yield either an improvement or simply no change (unless otherwise significantly influenced by noise) [see also, for example, Sivakumar et al. (2001b, 2002c)]. Overall, this preliminary finding seems to offer credence to the hypothesis that an acceptable number of variables to model the risk of arsenic contamination should range from 7 or 8 to 11 [The LR results also seem to strengthen our earlier point that the CD estimates reported by Hossain and Sivakumar (2006a) may only be an overestimation due to the presence of noise, if any, and not an underestimation].

Currently, there are a number of maps available that characterize the probability of arsenic contamination in non-sampled regions based on kriging [see BGS-DPHE (2001) and McArthur et al. (2001), for example]. Preliminary findings of our study imply that an injection of the chaotic dynamic approach of LR modeling with variables equaling the CD could expedite refinement of the map toward reduction of uncertainty in risk of contamination at non-sampled locations than what would have otherwise been possible by the kriging method alone. Although CD does not offer any physical insight on the variables that need to be chosen or the nature of their integration in risk assessment models, prior knowledge as a proxy for an acceptable number of variables required can be a valuable information that can potentially save considerable time during a rapid assessment of arsenic contamination for remediation management.

7 Conclusion

While applications of nonlinear dynamic concepts, such as the CD method, are gaining momentum in environmental sciences, their usefulness to understand the actual physical mechanisms occurring in our catchments and aquifers remains unclear. With the encouraging results reported recently by Hossain and Sivakumar (2006a) regarding the possible nonlinear deterministic nature of arsenic contamination phenomenon in Bangladesh (with CD values ranging from 8 to 11), we herein have explored the possible physical connection between the CD and the mathematical modeling of risk of arsenic contamination in groundwater. We considered the LR model, with an aim to link the nonlinear CD technique with a linear analysis technique. Using 11 potential influencing variables that largely dictate the variability of arsenic concentration, we observed that the CD may function as an acceptable proxy for the number of variables required in the LR model to accurately predict arsenic contamination at non-sampled wells. Given this preliminary finding, we believe it is time we considered more comprehensive investigations to assess the true merit of non-linear deterministic paradigms in conjunction with the more conventional linear stochastic methods, such as kriging, for reducing uncertainty of risk mapping for groundwater contamination in resource poor countries.

This study is not without its share of limitations. The two primary limitations that should be highlighted herein, so that findings from this study are not quoted out of context, are: (1) selection of potential influencing variables from a purely data-based paradigm; and (2) maximum number of influencing variables being only 11 and barely exceeding the range of CD values. An earlier section (on 'The potential influencing variables') in this paper has already discussed in detail the first limitation with a qualified disclaimer. On the second limitation, we unconditionally recognize that the value of CD could have been more convincingly demonstrated had more than 11 potential influencing variables been analyzed. However, inclusion of a higher number of variables is easier said than done, since there is paucity of quality-controlled data in a rural setting like Bangladesh. For example, an influencing variable such as soil cover is expected to influence recharge and to ultimately affect the water table fluctuations, which may consequently be responsible for the mechanism that mobilizes arsenic (Twarakavi and Kaluarachchi 2006). However, such data are hard to obtain for the case of Bangladesh on a large scale. We believe that inclusion of a larger set of geochemical data is an important area of future study where we, as members of the non-linear deterministic community, should depend on effective feedback from the community engaged in mechanistic understanding of arsenic contamination in order to secure a more complete and appropriate dataset for CD integration. It must be noted, therefore, that more detailed studies are needed to verify the true limitations and strengths of the CD approach to designing LR models for rapid assessment of risk of arsenic contamination. Investigations in this direction are already underway, details of which will be reported elsewhere.

References

Afifi AA, Clark V (1984) Logistic regression in computer-aided multivariate analysis. Lifetime Learning Publications, Belmont

Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 □ LE □ TYPESET
MS Code : 98 ✔ CP ✔ DISK

Ahmed KM, Bhattacharya P, Hasan MA, Akhter SH, Alam SMM, Bhuyian MA, Imam MB, Khan AA, Sracek O (2004) Arsenic enrichment in groundwater of the alluvial aquifers in Bangladesh: an overview. Appl Geochem 19:181–200

Akai J, Izumi K, Fukuhara H, Masuda H, Nakano S, Yoshimura T, Ohfuji H, Anawar MH, Akai K (2004) Mineralogical and geomicrobiological investigations on groundwater arsenic enrichment in Bangladesh. Appl Geochem 19:215–230

Biswas BK, Dhar RK, Samantha G, Mandal BK, Chakraborti D, Faruk I, Islam KS, Chowdury M, Islam A, Roy S (1998) Detailed study report of Samta, one of the arsenic-affected villages of Jessore District, Bangladesh. Curr Sci 74:134–145

Burgess WG, Burren M, Perrin J, Ahmed KM (2000) Constraints on sustainable development of arsenic-bearing aquifers in southern Bangladesh. Part 1: A conceptual model of arsenic in the aquifer. In: Hiscock, Rivett, Davison (eds) Sustainable groundwater development, vol 193. Geological Society of London Special Publication, pp 145–163

Christakos G, Li X (1998) Bayesian maximum entropy analysis and mapping: a farewell to kriging estimators? Math Geol 30(4):435–462

Faybishenko B (2002) Chaotic dynamics in flow through unsaturated fractured media. Adv Water Resour 25(7):793–816

Fraedrich K (1986) Estimating the dimensions of weather and climate attractors. J Atmos Sci 43:419–432

Ghilardi P, Rosso R (1990) Comment on 'Chaos in rainfall.' Water Resour Res 26(8):1837–1839

Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208

Hao B-L (1984) Chaos. World Scientific, Singapore

Harvey CF, Swartz CH, Badruzzaman ABM, Keon-Blute N, Yu W, Ali MA, Jay J, Beckie R, Niedan V, Brabander D, Oates PM, Ashfaque KN, Islam S, Hemond HF, Ahmed MF (2002) Arsenic mobility and groundwater extraction in Bangladesh. Science 298:1602–1606

Helsel DR, Hirsch RM (1992) Statistical methods in water resources. Elsevier, New York

Hossain F, Sivakumar B (2006a) Spatial pattern of arsenic contamination in shallow tubewells of Bangladesh: regional geology and nonlinear dynamics. Stoch Environ Res Risk Assess 20(1–2):66–76. DOI 10.1007/s00477–0055-0012-7

Hossain F, Sivakumar B (2006b) Spatial interpolation based on complementary paradigms: a call for a change in attitude, mathematical geology (in review; available online at http://www.iweb.tntech.edu/fhossain/papers/Math_GeolAs.pdf)

Hossain F, Bagtzoglou AC, Nahar N, Hossain MD (2006a) Statistical characterization of arsenic contamination in shallow tube wells of western Bangladesh. Hydrol Process 20(7):1497–1510. DOI 10.1002/hyp.5946

Hossain F, Hill J, Bagtzoglou AC (2006b) Geostatistically based management of arsenic contaminated ground water in shallow wells of Bangladesh. Water Resour Manage (in press). DOI 10.1007/s11269–006–9079–2

Journel AG, Huijbregts CJ (1978) Mining Geo-statistics. Academic, San Diego

Lambrakis N, Andreou AS, Polydoropoulos P, Georgopoulos E, Bountis T (2000) Nonlinear analysis and forecasting of a brackish karstic spring. Water Resour Res 36(4):875–884

McArthur JM, Ravenscroft P, Safiullah S, Thirlwall MF (2001) Arsenic in groundwater: testing pollution mechanisms for sedimentary aquifers in Bangladesh. Water Resour Res 37(1):109–117

McArthur JM, Banerjee DM, Hudson-Edwards KA, Mishra R, Purohit R, Ravenscroft P, Cronine A, Howarth RJ, Chatterjee A, Talukder T, Lowry D, Houghton S, Chadha DK (2004) Natural organic matter in sedimentary basins and its relation to arsenic in anoxic ground water: the example of West Bengal and its worldwide implications. Appl Geochem 19:1255–1293

Mukherjee AB, Bhattacharya P (2002) Arsenic in groundwater in the Bengal Delta plain: slow poisoning in Bangladesh. Environ Rev 9:189–220

Nerenberg MAH, Essex C (1990) Correlation dimension and systematic geometric effects. Phys Rev A 42(12):7065–7074

Ormsby T, Napolean E, Burke R, Feaster L, Groessl C (2004) Getting to know ArcGIS desktop, 2nd edn. ESRI Press, Redlands (ISBN:1–58948–083-X)

Porporato A, Ridolfi L (1997) Nonlinear analysis of river flow time sequences. Water Resour Res 33(6):1353–1367

Rodriguez-Iturbe I, De Power FB, Sharifi MB, Georgakakos KP (1989) Chaos in rainfall. Water Resour Res 25(7):1667–1675

Schertzer D, Tchiguirinskaia I, Lovejoy S, Hubert P, Bendjoudi H (2002) Which chaos in the rainfall-runoff process? A discussion on 'Evidence of chaos in the rainfall-runoff process' by Sivakumar et al. Hydrol Sci J 47(1):139–147

Schreiber T, Kantz H (1996) Observing and predicting chaotic signals: is 2% noise too much? In: Krastov Yu A, Kadtke JB (eds) Predictability of complex dynamical systems. Springer, Berlin Heidelberg New York, pp 43–65

Serre ML, Kolovos A, Christakos G, Modis K (2003) An application of the holistochastic human exposure methodology to naturally occurring arsenic in Bangladesh drinking water. Risk Anal 23(3):515–528

Sivakumar B (2000) Chaos theory in hydrology: important issues and interpretations. J Hydrol 227(1–4):1–20

Sivakumar B (2004a) Chaos theory in geophysics: past, present and future. Chaos, Solitons. Fractals 19(2):441–462

Sivakumar B (2004b) Dominant processes concept in hydrology: moving forward. Hydrol Process 18:2349–2353

Sivakumar B (2005) Correlation dimension estimation of hydrologic series and data size requirement: myth and reality. Hydrol Sci J 50(4):591–604

Sivakumar B, Phoon KK, Liong SY, Liaw CY (1999) A systematic approach to noise reduction in chaotic hydrological time series. J Hydrol 219(3–4):103–135

Sivakumar B, Berndtsson R, Persson M (2001a) Monthly runoff prediction using phase-space reconstruction. Hydrol Sci J 46(3):377–387

Sivakumar B, Sorooshian S, Gupta HV, Gao X (2001b) A chaotic approach to rainfall disaggregation. Water Resour Res 37(1):61–72

Sivakumar B, Berndtsson R, Olsson J, Jinno K (2002a) Reply to 'which chaos in the rainfall-runoff process?' by Schertzer et al. Hydrol Sci J 47(1):149–158

Sivakumar B, Jayawardena AW, Fernando TM GH (2002b) River flow forecasting: use of phase-space reconstruction and artificial neural networks approaches. J Hydrol 265(1–4):225–245

Sivakumar B, Persson M, Berndtsson R, Uvo CB (2002c) Is correlation dimension a reliable indicator of low-dimensional chaos in short hydrological time series? Water Resour Res 38(2). DOI 10.1029/2001WR000333

Sivakumar B, Harter T, Zhang H (2005) Solute transport in a heterogeneous aquifer: a search for nonlinear deterministic dynamics. Nonlin Process Geophys 12:211–218

Takens F (1981) Detecting strange attractors in turbulence. In: Rand DA, Young LS (eds) Dynamical systems and turbulence, lecture notes in mathematics, vol 898. Springer, Berlin Heidelberg New York, pp 366–381

Theiler J (1987) Efficient algorithm for estimating the correlation dimension from a set of discrete points. Phys Rev A 36(9):4456–4462

Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 ☐ LE ☐ TYPESET
MS Code : 98 ✔ CP ✔ DISK

Tsonis AA, Triantafyllou GN, Elsner JB, Holdzkom JJ II, Kirwan AD Jr (1994) An investigation of the ability of nonlinear methods to infer dynamics from observables. Bull Am Meteorol Soc 75:1623–1633

Twarakavi NKC, Kaluarachchi JJ (2006) Arsenic in ground waters of conterminous United States: assessment, health risk, and costs for MCL compliance. J Am Water Resour Assoc 42(2):275–294

van Geen A, Zheng Y, Vesteeg R, Stute M, Horneman A, Dhar R, Steckler M, Gelman A, Ahsan H, Graziano JH, Hussain I, Ahmed KM (2003) Spatial variability of arsenic in 6000 tube wells in a 25 km^2 area of Bangladesh. Water Resour Res 39(5):1140. DOI 10.1029/2002/WR001617

Welch AH, Westjohn DB, Helsel DR, Wanty RB (2000) Arsenic in ground water of the United States: occurrence and geochemistry. Ground Water 38(4):589–604

Yu WH, Harvey CM, Harvey CF (2003) Arsenic groundwater in Bangladesh: a geo-statistical and epidemiological framework for evaluating health effects and potential remedies. Water Resour Res 39(6):1146. DOI 10.1029/2002WR001327

Zheng Y, Stute M, van Geen A, Gavrieli I, Dhar R, Simpson HJ, Schlosser P, Ahmed KM (2004) Redox control of arsenic mobilization in Bangladesh groundwater. Appl Geochem 19:201–214

Journal : 477 Dispatch : 20-11-2006 Pages : 9
Article No. : 98 ☐ LE ☐ TYPESET
MS Code : 98 ☑ CP ☑ DISK