Journal of Environmental Management 211 (2018) 125–137

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Developing robust arsenic awareness prediction models using machine learning algorithms

Sushant K. Singh [a,*], Robert W. Taylor [a], Mohammad Mahmudur Rahman [b], Biswajeet Pradhan [c,d]

[a] Department of Earth and Environmental Studies, Montclair State University, 1 Normal Ave, Montclair, NJ 07043, United States
[b] Global Centre for Environmental Remediation (GCER), Faculty of Science, The University of Newcastle (UON), Callaghan, New South Wales, Australia
[c] School of Systems, Management and Leadership, Faculty of Engineering and IT, University of Technology Sydney, Building 11, Level 06, 81 Broadway, Ultimo, NSW 2007, Australia
[d] Department of Energy and Mineral Resources Engineering, Choongmu-gwan, Sejong University, 209 Neungdong-ro Gwangjin-gu, Seoul 05006, Republic of Korea

ARTICLE INFO

Article history:
Received 3 June 2017
Received in revised form
14 January 2018
Accepted 15 January 2018

Keywords:
Arsenic
Awareness
Socioeconomic
Demographics
Sociobehavioral
Machine learning algorithms
SVM
RF
GIS
India

ABSTRACT

Arsenic awareness plays a vital role in ensuring the sustainability of arsenic mitigation technologies. Thus far, however, few studies have dealt with the sustainability of such technologies and its associated socioeconomic dimensions. As a result, arsenic awareness prediction has not yet been fully conceptualized. Accordingly, this study evaluated arsenic awareness among arsenic-affected communities in rural India, using a structured questionnaire to record socioeconomic, demographic, and other sociobehavioral factors with an eye to assessing their association with and influence on arsenic awareness. First a logistic regression model was applied and its results compared with those produced by six state-of-the-art machine-learning algorithms (Support Vector Machine [SVM], Kernel-SVM, Decision Tree [DT], *k*-Nearest Neighbor [*k*-NN], Naïve Bayes [NB], and Random Forests [RF]) as measured by their accuracy at predicting arsenic awareness. Most (63%) of the surveyed population was found to be arsenic-aware. Significant arsenic awareness predictors were divided into three types: (1) socioeconomic factors: caste, education level, and occupation; (2) water and sanitation behavior factors: number of family members involved in water collection, distance traveled and time spent for water collection, places for defecation, and materials used for handwashing after defecation; and (3) social capital and trust factors: presence of anganwadi and people's trust in other community members, NGOs, and private agencies. Moreover, individuals' having higher social network positively contributed to arsenic awareness in the communities. Results indicated that both the SVM and the RF algorithms outperformed at overall prediction of arsenic awareness—a nonlinear classification problem. Lower-caste, less educated, and unemployed members of the population were found to be the most vulnerable, requiring immediate arsenic mitigation. To this end, local social institutions and NGOs could play a crucial role in arsenic awareness and outreach programs. Use of SVM or RF or a combination of the two, together with use of a larger sample size, could enhance the accuracy of arsenic awareness prediction.

1. Introduction

Awareness plays a vital role in the adoption of precautionary measures and policies in disease-prone areas (Niksic et al., 2016; Scatà et al., 2016). In real-world conditions, exposed communities' responsiveness to health threats and risks varies with socioeconomic and demographic conditions (Scatà et al., 2016). Mitigation of water-related diseases is a significant challenge for researchers and policymakers, for more than 600 million people still lack potable water globally, creating ideal conditions for the spread of diseases such as diarrhea, dysentery, cholera, polio, and typhoid—and for the loss of millions of lives (Fewtrell, 2004; NGWA, 2016). Arsenic contamination, which has been described

* Corresponding author. Current address: Virtusa Corporation, Irvington, NJ, United States.
E-mail addresses: sushantorama@gmail.com (S.K. Singh), taylorr@mail.montclair.edu (R.W. Taylor), mahmud.rahman@newcastle.edu.au (M.M. Rahman), biswajeet24@gmail.com (B. Pradhan).

https://doi.org/10.1016/j.jenvman.2018.01.044
0301-4797/

as perhaps the worst "global mass poisoning in human history," has left nearly 300,000 arsenicosis patients still uncured—a figure that is only growing (Smith et al., 2000; WHO, 2010). Globally, more than 296 million people in more than 100 countries face mild to severe risks to health because of their consumption of groundwater contaminated by arsenic (Chakraborti et al., 2017b; IWA, 2016; Singh and Stern, 2017).

Arsenic is a known human carcinogen that naturally occurs in the environment and that is released to groundwater under altered redox conditions triggered by high levels of groundwater withdrawal (Chakraborti et al., 2016c; USEPA, 2001). Continued ingestion of more than 10 μg/L of arsenic for a prolonged period can cause various adverse health effects, among them keratosis; melanosis; skin, lung, and bladder cancer; neural and developmental defects; diabetes; reduced IQ; negative obstetric outcomes; and DNA damage (Dutta et al., 2015; Flanagan et al., 2012; Liu et al., 2017; Quansah et al., 2015). A detailed description of arsenic-induced health effects can be found in the literature (Chakraborti, 2011; Chakraborti et al., 2017b).

Bangladesh and India are severely affected by arsenic (Chakraborti et al., 2016c; Singh, 2015). Nearly 70 million individuals' lives are at risk in 7 of India's 29 states because of their consumption of arsenic-contaminated groundwater (Chakraborti et al., 2016c). Elevated levels of arsenic in water, urine, hair, and nails, as well as mild to moderate arsenical skin lesions in children, have been reported in these regions (Chakraborti et al., 2017a). Most of the arsenic-affected population is impoverished and thus has few available resources: Its members are highly vulnerable to the adverse health effects caused by arsenic intake. Poor socioeconomic conditions and social status have generally hindered arsenic-preventive initiatives, leaving arsenic mitigation policies unable to achieve sustainability (Singh, 2015; Singh and Stern, 2017). Indeed, sustainable use of arsenic mitigation technologies depends in large part on behavioral factors in arsenic-exposed communities, where awareness of the problem plays a crucial role (Tobias and Berg, 2011). Although low levels of arsenic awareness are a persistent concern among scientists and policymakers who seek to promote the success of arsenic mitigation programs, few studies address this issue. In addition, the socioeconomic, demographic, and other social factors that underlie the arsenic mitigation regime are still unknown.

A decade ago, Chowdhury et al. (2001) urged that combating the arsenic crisis would require immediate heightening of awareness and education of arsenic-exposed communities (Chowdhury et al., 2001). Later, they observed that perhaps 80% of the afflicted West Bengalese population of India and Bangladesh could have been spared arsenic toxicity had they been afforded better access to nutrition through heightened awareness among and improvements in the education of netizens (Chakraborti et al., 2002). Certainly arsenic mitigation interventions have been technocentric—but awareness campaigns, including education of villagers and participation of the arsenic-exposed communities in arsenic mitigation policies, have been overlooked (Chakraborti et al., 2002). In a study of conditions in Bangladesh, Hadi (2003) reported that arsenic mitigation projects played a significant role in raising communities' level of arsenic awareness (Hadi, 2003). The authors of another study have claimed that public education programs promoting arsenic awareness could lower levels of engagement in risky behaviors by informing affected communities about arsenic contamination and its associated health risks (Hanchett et al., 2002). Higher levels of awareness of and concern about arsenic contamination have also been linked to a greater inclination to choose piped water systems (Ahmad et al., 2005). Still other studies, for their part, have emphasized the importance of awareness in reducing arsenic exposure through adoption of best practices, such as switching to arsenic-free sources that significantly reduce arsenic risk (Ahamed et al., 2006; Rahman et al., 2005; Van Geen et al., 2002).

Although arsenic awareness has emerged as a vital component in the success of arsenic mitigation programs and in the reduction of risk posed by arsenic, few studies have evaluated the association of socioeconomic and demographic factors with arsenic awareness for arsenic-affected communities. In one example, Parvez et al. (2006) reported that awareness was associated with male sex, position as nonlaboring head of household, better housing, and testing of wells for arsenic concentration (Parvez et al., 2006). The authors of another study found that people who had lower levels of literacy were less aware of arsenic-induced health risks (Mondal et al., 2017). Accordingly, it is vital to understand how arsenic awareness varies with socioeconomic and demographic characteristics of arsenic-exposed communities, as well as how other social and behavioral factors, such as water and sanitation behavior, trust in local institutions and agencies, and trust in other members of a community, influence arsenic awareness.

In view of these gaps in the research, this study represents a first step in assessing arsenic awareness and its association with socioeconomic, demographic, and other sociobehavioral aspects of arsenic-exposed communities. This study (1) offers an in-depth analysis of arsenic awareness among studied communities, (2) describes associations among socioeconomic, demographic, and other sociobehavioral factors as they affect communities' arsenic awareness, (3) applies a multinomial logistic regression model to analyze arsenic awareness, and (4) identifies robust arsenic awareness prediction models by comparing six cutting-edge machine learning algorithms. In doing so, it seeks to help researchers (1) understand existing aspects of arsenic awareness among arsenic-affected communities in rural areas, (2) identify significant socioeconomic, demographic, and other sociobehavioral factors that can influence arsenic awareness, (3) produce a statistical model such as could be adapted for use in other arsenic-affected areas to aid creation of awareness campaigns and arsenic mitigation policies, and (4) identify potential machine learning algorithms capable of accurately predicting arsenic awareness.

2. Materials and methods

2.1. Study area

The study area, the state of Bihar in India, is at risk of being the future site of another mass poisoning as a result of its population's level of exposure to elevated levels of arsenic through groundwater and foods (Chakraborti et al., 2003, 2016a, b; Singh, 2011; Singh and Ghosh, 2011). Already the groundwater in nearly 50% of Bihar's districts is arsenic-contaminated, affecting more than 12 million human lives (Singh, 2015). Arsenic contamination in groundwater, soil, and food materials—with associated health impacts—has been reported in many studies of this region (Chakraborti et al., 2016b; Singh, 2011; Singh and Ghosh, 2011, 2012). Several arsenicosis patients have already been diagnosed, and possible cancer and noncancer risks estimated, in certain areas of Bihar (Chakraborti et al., 2003, 2016a, b; Singh et al., 2014; Singh and Ghosh, 2012). Even so, arsenic mitigation programs and socioeconomic studies are only in their preliminary stages in this region.

Thus far, no reports have evaluated arsenic awareness and its association with communities' socioeconomic, demographic, and other sociobehavioral factors (Singh, 2015). The authors of a recent study reported that existing socioeconomic, demographic, and other factors contribute to overall levels of social vulnerability for arsenic-exposed communities who suffer arsenic contamination of groundwater (Singh and Vedwan, 2015). What's more, current