Sankhyā : The Indian Journal of Statistics
2016, Volume 78-B, Part 2, pp. 341-361
© 2016, Indian Statistical Institute

Abundant Environmental Arsenic Contamination: Some Statistical Perspectives

Pranab K. Sen
University of North Carolina, Chapel Hill, USA

Abstract

There is a severe impact of the current abundant environmental contamination and toxic pollution on ecology as well as our life sustaining resources: drinkable and safely usable water, respirable air, untainted and edible food, and renewable and affordable energy. Population explosion (particularly, in East and South Asia, including India and China), escalating industrialization, immeasurable volume of industrial, human and e-wastes, inadequate sanitary safeguards, our modern life-style, and geo-political undercurrents are threatening a safe propagation of human health and life on earth. Of special interest is the sub-soil and surface water arsenic contamination's deleterious impact on human health and quality of life. Socio-economic, familial and environmental undercurrents are persistent in this respect. A statistical appraisal along with a quantifying contamination severity index constitute the primary objective of this study.

AMS (2000) *subject classification.* Primary 62G35, 62G99; Secondary 62P99.
Keywords and phrases. Absorption toxicology, Affluence index, Arsenic removal plant (ARP), Arsenites, Bio-hazards, Contamination severity index (CSI), Ecology, Epidermal cancer, Gastwirth coefficient, Gini coefficient, Harmonic mean, Ingestion toxicology, Lesion, PBPK modeling

1 Introduction

The water for our domestic use (including drinking) and agricultural needs may not be safe and adequate for all purposes. The air we breathe may be saturated with hydrocarbons, carbon dioxide, respirable airborne particulate matters, and toxic gases. The food we intake may be contaminated with toxic substances and bacterial aggregation due to environmental pollution and improper production, preservation and maintenance standards. Our renewable and somewhat affordable energy sources: coal, wood burning, electricity, natural gas and petroleum products, including diesel may be creating an enormous environmental pollution and toxic contamination. Automobile and industrial exhausts are contaminating our atmosphere at an exponential rate, and industrial and human waste disposition is an alarming global contamination problem. Radiation waste and our rapidly increasing dependence on electronics are leading to an enormous e-waste problem. Sanitary land-fills and improper hygienic standards are contaminating ground (surface as well as sub-soil) water creating bio-hazards. In all respect, there is a major crisis in a major part of the world, more acute in some of the third world countries with a huge population burden. With the explosion of human population, particularly in South and East Asian countries (including China and India), the demand for more food, water, and energy has been sharply increasing. This has led to deforestation in the foot-hill and other areas, resulting in land (or mud-)slides or avalanches, decreasing soil-fertility, and demanding more use of fertilizers etc. Increasing urbanization and changing life-style (to cope with the present e-world) are creating ecological imbalance and huge waste disposition problem. Scarcity of drinkable or agriculturally usable water is a significant health hazard. Climatic disasters, at least partly due to human mishandling of our environment, and flooding are causing serious ground water and drainage problems. In the agriculture sector, the impact of environmental toxicity and contamination, mingled with epidemiological undercurrents is far-reaching to devastating (Sen, 2013a). This is just a brief outline of the abundant environmental contamination that is steadily encroaching all sectors of human life and health.

In the present study, we will, mostly, confine ourselves to the deleterious impact of arsenic (As) contamination on human (quality of) life and survival. Contamination of groundwater and surface water due to arsenites is a major health hazard factor. There is a striking similarity of the health outcomes due to As contamination and endemic *Pemphigus Foliaceus* or Fogo Selvagem which is quite prevalent in certain regions of Brazil, Peru and much of the other sub-tropical areas in Latin America. In this way, there are two principal mode of uptake of poisonous arsenites in human body: (i) *ingestion*, mostly due to chronic exposure to contaminated drinking water, and (ii) *absorption* through epidermal tissues. In this way, both ingestion toxicology and absorption toxicology are pertinent to As contamination. (US) National Research Council (2001) report suggests that arsenic related disease due to chronic exposure through drinking water has a relatively low incidence rate and latency up to decades for most end points significant to a burden of disease assessment (Karim, 2000; Smith et al., 2000). However, in a major part of the Indian sub-continent (including Bangladesh, Myanmar, Nepal, Pakistan and Sri Lanka), the drinking water and other surface waters, though running grossly inadequate for their needs, account for the major battery of diseases and disorders related to arsenic contamination, and the prevalence rate is not low. This is more affected by the lack of public health and medical facilities, more in the rural population. In addition to drinking of arsenic contaminated water, ingestion of arsenic contaminated food is a significant risk factor. In that way, use of arsenic contaminated water for agricultural and food processing purposes is a significant factor. It has been identified that arsenic compounds prevail in human urine, toe-nails as well as hairs. Lesion in major parts of the body leading to cancers is a common outcome of arsenic contamination. Though drinking water may mostly come from deep tubewell or pipeline facilities, other domestic uses of arsenic contaminated water (such as bathing, washing of clothes, cooking and cleaning pots, pans and dishes), particularly in rural areas, is a significant risk factor for both absorption and ingestion toxicity, resulting in skin lesions and cancers, and various cancers relating to blood, kidney and GI tracts. Medicinal preparations may contain low dose arsenic compound and cause ingestion toxicity. Drug addiction may therefore induce some arsenic contamination. Lack of sanitary safeguards, personal hygiene, inadequate public health practice and monitoring, and familial factors may all contribute to the aggravation of arsenite contamination.

Keeping in mind the above perspectives, in Section 2, we identify the usual modes of arsenic contamination and its health impact with special emphasis on the Indian sub-continent. The role of intake, uptake sites, and the formulation of the *bio-concentration factor* (BCF) and *physiologically based pharmacokinetics* (PBPK) modeling are discussed in Section 3. Arsenic mitigation plans and their role in controlling the disorder are discussed in Section 4. At the present, epidemiological investigations are overwhelming, though there is a vast scope for more in depth statistical reasoning. Section 5 deals with the statistical perspectives, and a new *contamination severity index* (CSI), formulated along the lines of affluence index (Sen, 1988) in Section 6 and 7. The concluding section deals with other perspectives of this arsenic contamination and the genuine need for more statistical rationality, interpretation and analysis.

2 Modes of As Contamination

Aquifer is a water bearing stratum of permeable rock, sand or gravel, and it underlies the surface soil in various strata at various depth. Arsenic (As) if present in certain mineral (mostly, sulphide) aquifer, due to lowering of moisture levels, becomes prone to oxidation and adhesion, in a very thin layer, to iron (Fe) hydroxides. This is referred to as *adsorption.* When the groundwater level is raised, albeit seasonally, Fe hydroxide releases As into groundwater. This groundwater on its way to natural groundwater transits (or channels) releases the As to a wider setup. Disposition of this As contaminated groundwater through wells, tube wells or ponds, rivers and streams, creates a major health hazard for human consumption. Uprising of the sea level, as is commonly perceived now a days, makes this As contamination more intense in coastal areas with underground or surface link to this outside water. Sanitary landfill sites and raw sewerage add more misery to this As contamination. Use of deep tube well may lessen the level of As in collected water but it can still be far above the international safe limit 10μg/L. *Arsenic removal plants* (ARP) have been installed in various affected areas to mitigate the As level. Unfortunately, scarcity of water and the cost for mitigation may limit the use of reduced As level water to drinking or related domestic uses. In urban areas, pipe-line served tap water may have the same feature if the collected untreated water is from a source which has the As contamination, as is the case in some of the cities in the Indian sub-continent, and even China is no exception.

As contaminated water is also widely used, more commonly in the suburban and rural areas, and by less privileged people, for washing of clothes, bathing, cleaning and other domestic purposes. In this respect, the land-fill cites and sewerage with direct access to groundwater or sub-soil moisture level, poor hygienic standards, industrialization and lack of environmental hazards awareness can intensify the prevalence of As contamination (Hossain 2006, Chakraborti et al. 2013, and the references cited therein). Improper burial practice, especially in high-moisture low grounds, may generate organic arsenites, though this factor is much less significant compared to others. In a major part part of the third world, irrigation work also involve As contaminated groundwater. This may induce As contamination in plants and crops and vegetation in the affected areas. In that way, cattle and other pets which survive on such As contaminated agricultural products, including straws and water, may contribute to As contamination in foods. Notable cases of this As contamination are the meat products (chicken, ham, veal), fish and other seafood from As contaminated water tracks, cow/goat milk, and nursing mother's breast milk contamination for infants. The endemic Pemphigus Foliaceus (PF) or Fogo Selvagem (FS) disorder, common in certain parts of Brazil and some other countries, have some common features with the As contamination (Aoki et al. 2004, 2011).

Associated with the As contamination of diverse type are the modes of human uptake of *arsinicosis* and their various health hazards. First, the As contaminated drinking water and food items may be labeled under *ingestion.* Ingestion toxicology relates to the internal resistance of the body immunity and the aftermath of As ingestion. The health effect may include tumors in stomach, GI tract, cancer in the urinal area, bladder cancer, blood cancer etc., and those in turn may severely affect the mental state and quality of life of infected people. Other domestic uses of As contaminated water and agricultural work, mostly labeled as *absorption* toxicity, has its aftermaths in epidermal lesion and cancer, loss of hair and deformation of fingers etc. This is very commonly observed in the Indian sub-content, and this feature is also common with the Fogo Selvagem disorder in Brazil. As contamination of some medicinal plants and products may be mostly under the ingestion toxicity category. *Inhalation* toxicity mainly arises in the respiratory intake (inhalation) of dust particles or related products containing arsenites. Of the three modes, absorption and ingestion are the principal ones. Often, absorption and ingestion toxicity work in synergy and thus may have a much serious health impact. We may refer to Sen (2001a, b, c) for some related statistical features. There is a persistent socio-economic factor that makes the health impact of As contamination more severe among the poor people and in areas with significant burden of population. We may refer to Sen (1994, 2012) for some health related problems. Hereditary undercurrents are also perceptible.

3 BCF and PBPK Modeling

As contamination does not impact an affiliated population in a mathematical equation. This is the same problem with other environmental contamination such as air pollution, mesothelioma, and lung cancer, among others. At the very basic stage is the inventory of As contamination, their spatio-temporal disparity, prevalence pattern, and their synergic activities with other modes of environmental contamination. Whereas in the case of air pollution and mesothelioma, the principal mode of human intake is inhalation, followed by absorption and ingestion, in the As contamination case, it is primarily through by ingestion and absorption. Excepting in some cases where As contaminated air particles interact with human inhalation system, its impact is primarily through the consumption of arsenite rich water and food items too. As for the groundwater contamination, the use of tube wells may have a different As impact mode than other sources, such as ponds, lakes or others. Again, even for tube well water, not everybody is fortunate to use it for drinking, as well as other domestic use such as cleaning, bathing, and others. In the Indian sub-continent, especially where water scarcity prevails, tube well water may only be restricted to drinking or at best cooking facilities. Then, the agricultural use of contaminated water may not only affect the agricultural products and food items but also cattle raised on such contaminated water may have perceptible level of As, affecting human consumption through milk products as well as meats. As contamination of surface water has similar effect on fish and other seafood. In this respect, ingestion is most relevant. Those who have to work in As contaminated water either in agricultural fields or in other sectors such as industrial work, absorption mode is very relevant. Based on the above consideration, it is important not only to identify the different sources of As contamination but also related demographic features.

The principal sources of As contamination are (i) groundwater As, (ii) surface water arsenites, and (iii) As transported at the subsoil level by floods, Tsunami and other climatic disasters. These are mostly inorganic As, while organic As may also come from human body and other plants and animals too. In the lower part of Bangladesh and West Bengal, decomposed corpses or buried animals or human being is a significant As contaminator. Among these various sources, there is a lot of variation of the level of contamination and its environmental impact. From human consumption of water point of view, either for drinking and domestic use or agricultural and industrial use, the exposure to As contamination is not uniform across the demographic and socio-economic strata. For example, poor people may have greater chance of having As contamination from domestic as well as non-domestic use of contaminated water. Further, this picture may also be different for male and female population. In essence, not all the environmentally emerging As goes to the human uptake, and only a fraction gets so. Therefore, there is a distribution of As uptake by a designated population, and this must be taken into account in studying the aftermath of As contamination. In environmental health studies, there is a common term, *bio-concentration factor* (BCF) which basically tells about the expected level of contamination that goes to the human system. This BCF is far from being an ultimate yardstick for the As human uptake. Better, we should consider the various demographic strata and for each stratum a BCF or its distribution should be thoroughly studied. Even for the As contaminated water from tube wells, the depth of the wells, their intrinsic As adsorption level, aquifer layer along with their iron and sulphide concentration all constitute important factors for stratification. In practice, specially in the Indian sub-continent, this is very seldom the case, thus raising doubts on the quality and validity of any data collected for administrative or scientific investigation.

Once arsenites enter human body, be it in the absorption, ingestion or inhalation mode, its propagation through the human body and immune system brings us to the doors of *physiologically based pharmacokinetics* (PBPK). Absorption toxicity is mostly due to epidermal cells, hairs, nails and foot, hands, back, face etc. At the first phase, it reacts with these body cells and tissues, tries to penetrate deeper into body organs, including blood and the circulation system. Ingestion toxicity, mostly through drinking water, foods and some medicines too, takes the mouth to the stomach route, through tongue, epiglottis, esophagus and trachea, thus connecting to the lower abdomen and GI tract. In the inhalation mode, the primary passage is the nasal inlets, and it has the access to the respiratory organs. In either way, the As material enters the human body system, though differently in different modes. In this way, the *toxico-dynamics and toxico-kinetics* (TDTK) play the major role in bodily reaction to the As toxicity, its immunity to the contaminating material and its aftermath with diseases or disorders in various organs. Stomach cancer, cancer in the GI tract (including colon), lesion on the body surface, skin cancer and a battery of mental health disorders. Thus, we have the chain: Uptake to intake of toxins, PBPK/TDTK phase, response in terms of manifested diseases and disorders. The picture is far more complex due to a synergy of various other environmental stressors and socio-economic conditions. It is in this setup, *toxicogenetics* becomes relevant and that might explain the familial impact of As contamination. Like the case of breast or ovarian cancer, here also, genetic effects may be perceptible. It is not uncommon to see multiple members of a family or community suffering from As contamination be it due to family environment or some familial factors carried out by heredity and mingled with environmental impacts. This is particularly noticeable in the agricultural sector where there may still be a pattern of family labors and all suffering from skin lesion to foot disorders. In this complex system, there is a pioneering role of statistical planning and interpretation. This is a rapidly growing field of interdisciplinary research, and in USA (and elsewhere) the environmental health sciences researchers are developing the PBPK/TDTK models based on sub-human primates and other animals, with a view to extrapolating the findings to human being. In the Indian sub-continent, there is a pressing need to undertake serious scientific investigations with these PBPK/TDTK models and integrate the findings with human system. Haphazardly or poorly organized epidemiological studies may not meet the basic requirement for data representation and validity criteria, and thereby, may raise serious question about the scope of derived conclusions from such conducted investigations.

4 Arsenic Mitigation Plants

At the present, in the Indian sub-continent, some arsenic removal plants (ARP) are in operation in specific geographical areas and specifically related to deep tube wells. The very terminology ARP is a bit vague. The international set tolerance limit for safe consumption of As contaminated water is

10μg/L. In the Indian sub-continent, this threshold specification is 50μg/L. Even so, most of the collected records relate to scores far above 50μg/L. Further, the distribution of the pre-ARP contamination level of groundwater arsenite has been observed to be highly positively skewed, and there is a perceptible diversity due to spatio-temporal factors. Accurate measurement of As contamination level may generally need further geological study of the underground soil layer structures and aquifer status. The burden of population using the designated tube well water as well as their volume of consumption may also tell how far the depth of these tube wells need to be calibrated so as to obtain adequate volume of water which can be used for As removal or mitigation. Moreover, this process of pumping water from a deep level has to be sustainable for a period of time. This itself depends on the sub-soil water level, aquifer layers that permit the propagation of ground water to lower level, and manageable underground transportation of collected water from other sources.

The basic objective of ARP is to reduce the As contamination level from a supposedly higher one to below 10μg/L. This reduction of the level depends on the pre-treatment level, volume of water it is expected to be pumped out daily, regeneration of water level at the source of tube well and their temporal As contamination level. Such ARPs have been mostly tried with deep tube well. But a completely different procedure may be necessary for arsenic reduction in other sources of water. For example, for irrigation purpose, often water is pumped out from some major waterways into the agricultural fields, leaving open the possibility of As contamination at the river-bed level as well as the transportation system. Moreover, the pumped water, albeit with reduced As contamination level, may partly go back to the groundwater level through percolation of porous soil and partly to other drainage system. This may generate a recycle pattern of pumping out water from a deeper level, using it, and releasing it back to the system. However, in agricultural use, a greater part is consumed on surface and evaporation is a major factor too. In heavily populated area, the volume of pumped out water may be much larger than the inflow of water to that aquifer level. This is often aided by rains and floods but that might increase the As contamination process in the sub-soil sector. It is not uncommon to see that in many housing complexes in heavily populated area, the underground moisture level recedes to much lower level and makes the tube well less effective. It may need for greater depth of the pipeline. This problem is more serious in areas which are not too close to the water-fronts but has a heavy burden of population. Excessive environmental pollution of some other surface water sources (such as ponds, lakes and rivers) may call for greater dependence on tube well, albeit power (energy) shortages may stand in the way of making full use of tube well water for all domestic uses.

Any mitigation plant relies on some process of neutralizing the level of As contamination to a supposed safe level. This process may have its side-effect on human consumption. Moreover, such a process may be costly and beyond affordability of a greater class of (mostly poor) people. If ARP has to be implemented at a mass level, in a region like the Indian sub-continent, there has to be a government level management plan so that a majority of people can be accommodated. Such a mega management scheme needs a lot of statistical and economic consideration along with the financial support of local as well as state and federal governments. In urban areas, high-rising apartment complexes' dependence on tube well water (in bulk volume), consumption of energy at a much higher level than a few years back, and modern life-style of the affluent class of people have all created an impasse for rural population with much less economic affluence to have the right level of energy consumption for maintenance of daily life and fulfilling the need for As contamination-free (or at least safe level) water for domestic as well as agricultural needs. In the Indian sub-continent, as we look into the Ganges (*Ganga-Padma-Jamuna*) valley, from the Bihar-Uttar Pradesh border to the delta areas in West Bengal and Bangladesh, it becomes quite clear that the rampage of As contamination of groundwater as well as surface water has created a irreparable damage to the environment and human health. ARPs have been used in some areas with some success, but the affordability of the common people remains a glaring question. There must be some other way to reduce this massive As contamination problem and restore a more healthy society.

5 Some Statistical Tasks

As contamination (scientific as well as public health) investigation needs an interdisciplinary approach. Geological/geographical science may provide the much needed information on the aquifer structure and groundwater level information. Wetland management agencies can provide information for surface and sub-soil water As contamination. It can also provide the information on some of the major sources of inorganic arsenites which lie in the roots of As contamination. Environmental science can provide us with the needed information on environmental stressors which impact the abundant growth of As contamination. Demographic/population sciences can tell us more about the pattern of population growth and habitation which must bear the health and survival consequences due to this massive arsenic contamination. Environmental health science and epidemiology can focus on adverse health effects while social sciences can cast light on the impact of As contamination on mental health, disability, morbidity as well as mortality. Economic and political science must be a guiding light in scrutinizing the scope for any master plan to reduce As contamination to a safe level and to have sustained efforts to combat this disorder. Social sciences are the basic provider of the impact of As contamination on society in its various strata. Clinical and public health sciences provide the basic and affordable tools for diagnostics and prevention (or at least minimizing) of adverse health effects of As contamination. Computer science and information technology are indispensable in handling massive data sets that crop up in such a massive study. In all respect, statistical planning, interpretation, analysis, and inference are of prime importance in blending an interdisciplinary approach to combat this dreadful contamination problem.

First and foremost, it is imperative to look into the prevalence of As contamination in the whole region (or a country/state) relative to its demographic constitution and socio-economic undercurrents.This itself is a massive statistical task. Collection of relevant data in a valid, representative and interpretable manner is the very first requirement. In this respect, the existing data, however scanty or disorganized they might be, should be carefully scrutinized, and used as a basis for more massive and planned data collection. Training of personnel in that respect is also essential. Information technology is indispensable in this context, but that is needed to be well tuned with statistical reasoning – not just *art for art's sake*. Assuming that the topographical information on aquifer layers and sub-soil moisture level can be obtained from appropriate agencies (of course, incurring cost and ensuring data quality), it should be a reasonable statistical problem to study the spatial pattern of As contamination. Secondly, there is a lot of variation of the As contamination level over time. This may be due to over adsorption of arsenite with time or more transport of As contaminated groundwater from other sources. This aspect is more likely to be observed with As contamination of surface water where industrial waste containing oxidized iron and some other metal scraps may have significant impact. Recent efforts to sort out bottles and cans from trash and human waste dumping may be helpful in reducing the level of As contamination from such cites. This brings us to the possibility of having spatio-temporal statistical models for both groundwater and surface water contamination. In both the contexts, *growth (curve) models* (GCM) in a spatio-temporal response surface setup should be developed. For some related problems, though in a different context, we may refer to Sen (2013b).

The very nature of the As contamination, its spatio-temporal dispersion, and its intrinsic complexities, data sets relating to such investigations could be so overdispersive and their collection could be so costly that it may be necessary for statistical modeling and analysis to have suitable summary measures of the As contamination severity which would relate more meaningfully to the health impact of this contamination. This, on one hand, brings to the study table the battery of adverse health effects impacted by As contamination, and on the other hand, to trace the statistical relationship of the severity of contamination and its health impacts. Although epidemiology is a very useful discipline to study this complex phenomenon, it definitely needs (bio-)statistical conjugation. This is one area where there is ample room for development of statistical models and analysis schemes incorporating possibly more innovative methodology, and stressing on their validity and usefulness in practice. Standard statistical methodology may not be of sufficient help in such ventures.

For studying the diverse health impact of As contamination, even confined to the groundwater tube well problem, much help is needed from clinical sciences. Not only the lesion on body surface and epidermic or skin cancer should be related to As contamination in a more quantitative and statistical way than just reporting the prevalence and mortality in a mere descriptive way. As for the ingestion toxicity arising from As contamination, the BCF to PBPK/TDTK modeling may not suffice. The simple models developed for animal studies in USA and elsewhere may not suffice for human exposure. In the Indian sub-continent, there is not only a pressing need to have objectively conducted animal studies to fathom out the statistical relationship of dose (As contamination) level and health (tumor or other carcinogenic) outcomes, but also to stress how such results can explain carcinogenic outcomes in relation to As contamination. One of the principal route of As contamination is the cow milk when the cattle are exposed to As contaminated water and straw and other foods. The transmission of As contamination in agricultural fields through irrigation or other water pathways, and As accumulation in crops, vegetables and fruits should be thoroughly studied in a statistical way with due consideration of epidemiological undercurrents. Even for nursing mothers, breast milk may contain As acquired from ingestion of contaminated food and drink, as well as some medicines which may contain arsenites, and this may affect the infant quite severely.

As has been stated earlier, As contamination severity is a serious concern not only in the Indian sub-continent but also in many other places all over the world. This itself needs a thorough statistical appraisal. In the rest of this study, we focus on a much needed statistical formulation of *contamination severity index* (CSI) with special attention to the As contamination. This index relates to a summary measure of As contamination in the context of ARP, and can be used effectively for spatio-temporal variation of As contamination. Besides the As contamination level and the amount of reduction achieved, there are other extraneous factors that we need to take into account, and this CSI is adaptable to such quantitative analysis. Some of the details are provided in the next section.

6 CSI for As Contamination

In a particular area or location, generally the distribution of the As level in groundwater (denoted by X) has a distribution $F(x), x \geq 0$. For uncontaminated water, $F(x)$ has greater concentration near the origin 0. The more there is As contamination, the more $F(x)$ tilts to the right. Thus, for As contaminated case, this distribution is positively skewed, and with the severity of contamination, it moves away from the origin. This ordering of the distribution of As contamination level underlies a statistical formulation of the severity of As contamination. The international standard for safe consumption of water is 10μg/L. However, in the Indian sub-continent, this safe limit has been raised to 50μg/L. Even though, in a majority of cases, the contamination level is beyond 100μg/L. The situation is quite comparable to the poverty index (Sen 1973, 1976) and affluence index (Sen 1986, 1988), both being based on the income distribution of the poor and affluent people. In the case of the affluent index, because of the heavy tail of the income distribution, and difficulties in obtaining reasonably accurate information for the excessively rich people, a more robust index was proposed. In the present context, an analogous index is proposed for CSI. To set the picture in the proper perspective, let us denote the lower threshold As limit by L and let $F(L)$ denote the proportion of As levels below the threshold limit L. The nature of F below L and its concentration around L (from the left) is not of much use in the present context. It is the dispersion of the distribution above L that characterizes the severity of contamination. As such, we work with the left-censored As level distribution for our formulation. Let

$$\gamma_L = 1 - F(L); \quad \alpha_L = F(L) = 1 - \gamma_L. \tag{6.1}$$

Also, let

$$\mu_L = {\alpha_L}^{-1} \int_0^L x dF(x), \; \beta_L = 1 - \mu_L / L. \tag{6.2}$$

Note that in the context of poverty, β_L is known as the *income gap ratio* (Sen, 1976). Then A. K. Sen's simple poverty index is given by

$$\pi_A = \alpha_L \beta_L. \tag{6.3}$$

This simple index does not take into account the concentration or disparity of wealth below the threshold line L. For that reason, let $F^*(x) = {\alpha_L}^{-1}F(x),\ x \leq L$ be the adjusted distribution below the threshold L. Let $G(L)$ be the *Gini coefficient* for the distribution $F^*(x), x \leq L$. Then A. K. Sen's refined index is given by

$$\pi_A^* = \alpha_L\{\beta_L + (1 - \beta_L G(L)\} = \alpha_L\{G(L) + (1 - G(L))\beta_L\}. \tag{6.4}$$

A robust version of the above index due to Sen (1986) is

$$\pi_S^* = \alpha_L\{(\beta_L)^{1-G(L)}\}. \tag{6.5}$$

These indexes apply to the distribution $F^*(x)$. However, in the present context, As level below L is of no concern and we are more concerned with the complementary part γ_L. For that as in Sen (1988, 1989), we define the harmonic mean for this distribution as

$$\xi_L = \{\int_L^\infty x^{-1} dF(x)/\gamma_L\}^{-1}. \tag{6.6}$$

The mean for the distribution $F^o(x)(= \gamma_L^{-1}\{F(x) - F(L)\}, x \geq L)$ is given by $\mu_L^o = \int_L^\infty x dF^o(x)$, and it is well known that $\xi_L \leq \mu_L^o$. Next, we define the *harmonic gap ratio* (HGR) as

$$\eta_L = 1 - L/\xi_L = 1 - L\gamma_L^{-1} \int_L^\infty x^{-1} dF(x). \tag{6.7}$$

Side by side, we define the *Harmonic Gini coefficient* (HGC) as

$$G^H(L) = E\{|Y_1^{-1} - Y_2^{-1}|\}/E\{Y_1^{-1} + Y_2^{-1}\}, \tag{6.8}$$

where Y_1 and Y_2 are two independent random variables with the distribution $F^o(.)$ as defined above. Though $G^H(L)$ is a suitable measure of concentration, it does not satisfy an invariance property which is enjoyed by the Gastwirth (1975) coefficient, defined below.

$$G^{oH}(L) = E\{|(Y_1^{-1} - Y_2^{-1})/(Y_1^{-1} + Y_2^{-1})|\}. \tag{6.9}$$

It is easy to see that $G^{oH}(L) = E\{|(Y_1 - Y_2)/(Y_1 + Y_2)|\}$, so that it is invariant under the transformation $y \to y^{-1}$, a desirable property for defining CSI in a more robust way. We would prefer to use $G^{oH}(L)$ in the formulation of the CSI. Specifically, we consider the following versions:

$$\kappa_{AL} = \gamma_L \eta_L; \tag{6.10}$$

$$\begin{aligned}\kappa_{SL} &= \gamma_L\{\eta_L + (1-\eta_L)G^{oH}(L)\} \\ &= \gamma_l\{G^{oH}(L) + \eta_L(1 - G^{oH}(L))\};\end{aligned} \tag{6.11}$$

$$\kappa_L^* = \gamma_L \eta_L^{(1-G^{oH}(L))}. \tag{6.12}$$

The first one does not take into account the disparity of the distribution $F^o(.)$, the second one is along the lines of A. K. Sen's index, and the last one is formulated along the lines of Sen (1988). If we look at η_L, it is close to 0 when ξ_L is close to L, and it monotonically goes to 1 as ξ_L moves away from the threshold value L. On the other hand, $G^{oH}(L)$ lies in the interval $[0, 1]$, and it is close to 0 if the disparity of the distribution $F^o(.)$ is small, and it converges to 1 as this dispersion increases. In that way, $1 - G^{oH}(L)$ lies between 0 and 1, it is close to 1 for small disparity of the distribution $F^o(.)$ and it converges to 0 as this dispersion increases. For that reason, in κ_L^*, the adjustment factor $\eta_L^{(1-G^{oH}(L))}$ will be small if either the harmonic gap ratio is small or $G^{oH}(L)$ is small. This explains its rationality as a CSI. The harmonic mean is less affected by outliers or gross error contamination than the arithmetic mean, and also, the harmonic version of the Gastwirth coefficient having the desired invariance property, is unaffected by the choice of the harmonic mean. For this reason, between the two indexes κ_{SL} and κ_L^*, we would recommend the latter one.

Let us now examine the effect of shifting the threshold level L from the international standard level 10μg/L to the Indian standard level 50μg/L; we denote these two levels as L_1 and L_2. The corresponding entities are denoted by $\gamma_{L_j}, \xi_{L_j}, \eta_{L_j}, G^{oH}(L_j)$ and $\kappa^*_{L_j}, j = 1, 2, \ldots$. Then by definition, $\gamma_{L_1} \geq \gamma_{L_2}$. Also, $L_1 < L_2$ and in this specific case, $L_2/L_1 = 5$. Note that

$$(\partial \eta_L / \partial L) = \eta_L f(L)/\gamma_L - 1/\xi_L. \tag{6.13}$$

Note that $f(L)/\gamma_L$ is the failure rate at $x = L$. Also, by definition, ξ_L is greater than L but less than the mean residual life μ_L^o. Hence, whenever,

$$(\mu_L^o - L)f(L)/\gamma_L < 1, \tag{6.14}$$

Equation (6.13) is negative, so that η_L is nonincreasing in Ll. This is, in particular, true for normal as well as gamma family of distributions. In a similar manner, it can be shown that κ^*_L will be non-increasing in L. This may imply that if instead of the international threshold level L_1, we use the Indian threshold level L_2, the resulting CSI may be smaller than it should have been. It is therefore suggested that the CSI should be computed for both the level L_1, L_2 and that would give a better idea of the severity in terms of the health hazards. In this respect, some ordering properties can be studied along the lines of Chatterjee and Sen (2000).

Note that $G^{oH}(L)$ in (6.9) involves a bounded kernel of degree 2, and hence, as in Hoeffding (1948), we define a related measure

$$G^{oH}_L(F, y) = E\{|(y - Y)|/(y + Y)\},\ y \in (L, \infty). \tag{6.15}$$

Note that the distribution of Y over (L, ∞) is denoted by $F^o(.)$, so that

$$\theta_L(F, y) = 1 - G^{oH}_L(F, y) = 2\int_L^\infty \frac{\min(x, y)}{x + y} dF^o(x),\ y \in (L, \infty). \tag{6.16}$$

As such, we consider a robust version of the CSI wherein the contribution of individual observation is weighed by their respective $\theta_L(F, y)$; this is defined by

$$\kappa^{**}_L(F) = \int_L^\infty (1 - L/y)^{\theta_L(F,y)} dF(y). \tag{6.17}$$

Next, note that

$$\begin{aligned} \frac{1}{2}\theta_L(F, y) &= -\int_L^\infty \frac{\min(x, y)}{x + y} d\bar{F}^o(x) \\ &= \frac{L}{L + y} + \int_L^\infty \frac{\max(x, y)}{(x + y)^2} \bar{F}^o(x) dx, \end{aligned} \tag{6.18}$$

where $\frac{\max(x,y)}{(x+y)^2} = x(x + y)^{-2} I(x > y) + y(x + y)^{-2} I(x < y)$ is non-increasing in $y \ \in (L, \infty)$. Also, $\frac{L}{L+y}$ is non-increasing in y. Hence, $\theta_L(F, y)$ is non-increasing in y with

$$\lim_{y\to\infty} \theta_L(F, y) = 0. \tag{6.19}$$

Since for $y \geq L$, $y^{-1}L$ is non-increasing in y, we obtain that

$$(1 - L/y)^{\theta_L(F,y)} \text{ is non-decreasing in } y \in (L, \infty). \tag{6.20}$$

Also, it converges to 1 as $y \to \infty$. This clearly explains the impact of heavy tails of F^o on κ^{**}_L. Further, note that by (6.16),

$$\frac{\partial \theta_L(F, y)}{\partial L} = -2f^o(L)\{1 + \frac{1}{2}\theta_L(F, y)\}\ (\leq 0), \tag{6.21}$$

for all $L \in R^+$. Moreover, $1 - L/y$ is non-increasing in L, so that by (6.20), we conclude that

$$\kappa_L^{**}(F) \;\text{ is }\; \text{non-increasing} \;\text{ in } L \in R^+. \tag{6.22}$$

Although this non-increasing property is quite intuitive, the decrement (over L) is not linear and this has a greater impact on the calibration of the CSI with the threshold value L.

The evolution of $\kappa_L^{**}(L)$ also leads to a measure of CSI for individual units as well as *small area* contamination, quite useful in spatio-temporal variation study of CSI. Let us write

$$(1 - L/y)^+ = \max\{0, (1 - L/y)\}, \quad y \in R^+, \tag{6.23}$$

and using the definition in (6.18), we let

$$G_{L+}^{oH}(F, y) = G_L^{oH}(F, y)I(y \geq L) + 0I(y \leq L) = G_L^{oH}(F, y)I(y > L). \tag{6.24}$$

Then we may rewrite $\kappa_L^{**}F) = \int_0^\infty \{(1 - L/y)^+\}^{1 - G_{L+}^{oH}(F,y)} dF(y)$. As such, for an individual unit with As level y, we may define the CSI score as

$$\kappa_L^{**}(y) = \{(1 - L/y)^+\}^{1 - G_L^{oH}(F,y)}, \quad y \in R^+. \tag{6.25}$$

Note that $\kappa_L^{**}(y) = 0$, $\forall y \leq L$. Basically, for As level less than L, we attach the CSI score 0, while for $y \geq L$, we have the same definition as the integrand at y of (6.20). If we have a region (district) with a considerable variation in As level and possibly in other geological/environmental factors, we can divide the regions in to counties (or sub-divisions), and for each such sub-divisions, we may then obtain the CSI scores. In this context, for the computation of the index $G_{L+}^{oH}(y)$, we will use the sub-divisional distribution of the As levels. This allows us to have a more detailed and refined CSI picture, quite useful for spatial variation studies. The main advantage of this *small area approach* is the utilization of the smooth (and bounded) CSI measures for which more conventional statistical analysis can be made, using suitable transformations if needed.

7 Further Statistical Discussions

Let us consider some sample counterparts of CSI and other measures introduced in the previous section. In a specific site, let us denote the sample outcome of As contamination level by $Y_j, j = 1, ..., n$. We denote the corresponding order statistics by

$$Y_{(0)} = 0 < Y_{(1)} < \cdots < Y_{(n)}. \tag{7.1}$$

Let $M_k = \max\{i : 1 \leq i \leq n : Y_{(i)} \leq L_k\}, k = 1, 2$. Further, let $F_n(x) = n^{-1}\sum_{i=1}^{n} I(Y_i \leq x), x \geq 0$ be the empirical distribution function. Then, the sample estimates of the γ_{L_k} are the following

$$\hat{\gamma}_{nk} = 1 - F_n(L_k) = (n - M_k)/n, k = 1, 2. \tag{7.2}$$

Similarly,

$$\hat{\xi}_{nL_k} = \hat{\gamma}_{nk}\{\sum_{i>M_k} n^{-1}Y_{(i)}^{-1}\}^{-1} = (n - M_k)\{\sum_{i>M_k} Y_{(i)}^{-1}\}^{-1}, \tag{7.3}$$

for $k = 1, 2$. Further, we note that

$$\hat{G}_{nk}^{oH} = \frac{2}{(n - M_k)(n - M_k - 1)} \sum_{M_k<i<j\leq n} \{Y_{(j)} - Y_{(i)}\}/(Y_{(j)} + Y_{(i)}), \tag{7.4}$$

for $k = 1, 2$. As such, we obtain that

$$\hat{\eta}_{nk} = 1 - (L_k/(n - M_k))\{\sum_{i>M_k} Y_{(i)}^{-1}\}, k = 1, 2, \tag{7.5}$$

$$1 - \hat{G}_{nk}^{oH} = \frac{2}{(n - M_k)(n - M_k - 1)}\{\sum_{M_k<i<j\leq n} 2Y(i)/\{Y_{(i)} + Y_{(j)}\}\}. \tag{7.6}$$

This leads us to the following estimators:

$$\hat{\kappa}_{nk}^{*} = \hat{\gamma}_{nk}\hat{\eta}_{nk}^{1-\hat{G}_{nk}^{oH}}, k = 1, 2. \tag{7.7}$$

We consider a more robust version of $\hat{G}_{nk}^{oH}$. Let

$$\hat{G}_{ki}^{*oH} = (n - M_k - 1)^{-1}\{\sum_{j>M_k} |Y_{(j)} - Y_{(i)}|/(Y_{(j)} + Y_{(i)})\}, \tag{7.8}$$

for $i = M_k + 1, \cdots, n$, so that

$$1 - \hat{G}_{ki}^{*oH} = \frac{2}{n - M_k - 1}\{\sum_{M_k<j<i} \frac{Y_{(j)}}{Y_{(i)} + Y_{(j)}} + \sum_{i<j\leq n} \frac{Y_{(i)}}{Y_{(i)} + Y_{(j)}}\}. \tag{7.9}$$

As such, we consider the following progressively adjusted CSI estimator:

$$\hat{\kappa}_{nk}^{**} = \frac{1}{n} \sum_{M_k<i\leq n} (1 - L_k/Y_{(i)})^{2(n-M_k-1)^{-1}\sum_{j>M_k} \min\{Y_{(i)},Y_{(j)}\}/(Y_{(i)}+Y_{(j)})}, \tag{7.10}$$

for $k = 1, 2$. Note that these functions are all well defined functionals of the empirical distribution function $F_n(.)$ (some of these are U-statistics Hoeffding, 1948) for which standard large sample theory may be incorporated with advantage for statistical inference and model building. As a matter of fact, jackknifing tools can be adopted here to provide (strongly) consistent estimator of the standard error of the estimated CSI.

As in the previous section, we may consider the sample counterpart of the individual SCI scores. We define $(1-L/y)^+$ as in (6.23) andconsider the empirical distribution as in after (7.1). Then the CSI score for a particular unit with As level Y_i is given by

$$\hat{\kappa}_L^{**}(Y) = \{(1-L/Y)\}^{1-\hat{G}_L^{*oH}(Y_i)}\} \tag{7.11}$$

where of course, for $Y_i \leq L$, SCI score is 0. In computing the harmonic Gastwirth coefficient, of course, we may consider the empirical distribution of the As levels in neighborhood network of units. This is in the spirit of nearest neighborhood methodology and can be adapted for large data sets, such as arising in the Bangladesh as well as adjacent India. For small area studies, we can work with the within area empirical distribution and define the CSI score as the average of the $\hat{\kappa}_L^{**}(Y_i)$ as defined above.

As an alternative approach, we may define

$$u_p = \sum_{i \leq [np]} Y_{(i)} / \sum_{i=1}^{n} Y_{(i)}, \ p \in (0,1), \tag{7.12}$$

and consider a Lorenz curve approach for the graph $\{(p, u_p) : 0 < p < 1\}$. However, we need to tune this *area under the curve* (AUC) approach with the L_k in an interpretable manner.

Other statistical issues include the treatise of As level measurements which may be subjected to gross errors, outliers, measurement errors, incompleteness due to missing observations, censoring of various types and administrative errors due to improperly trained personnel for carrying out the vital data collection. These should be taken into account in a systematic way and be addressed as far as possible.

8 Concluding Remarks

It is indeed a challenge, especially in the developing countries, to collect reliable data sets pertaining to the detailed statistical perspective as listed in the preceding section. In most of the cases, there may be data sets pertaining to some administrative objectives and using that in a different context may not necessarily facilitate use for other purposes. Some specific and important purposes for which the CSI measures can be used in practice include the following:

1. Pre-mitigation plan arsenic contamination levels for specific sites, their spatial variation.
2. Variation of the sub-soil moisture level over seasons and their impact on CSI.
3. Regression analysis, albeit in nonlinear forms, of CSI with important socio-economic and environmental factors.
4. Impact of ARP on CSI - this may be a longitudinal cum spatial setup.
5. Other health effects of ARP judged from CSI and side-effects of ARP plants.
6. Spatio-temporal analysis of As contamination effects from a wider health perspective.

As far as the As contamination in the Indian sub-continent, an important statistical question relates to the CSI mapping of the entire region, its socio-economic impacts, adverse health effects, resources to combat this catastrophic hazard, and restore safe arsenic level water for drinking purposes, domestic and agricultural use. On all counts, the situation is very dreadful and it certainly needs an interdisciplinary effort to bring it under reasonable control.

CSI is an index that lies between 0 and 1, with the severity pointing out to the upper end-point. By definition, the CSI is less than γ_k, and even lesser if the η_k is small. However, with heavy tails of the As contamination level distribution, η_K is likely to be close to its upper bound and thus pushing the CSI to a higher level. This is a binding reason for the accurate measurement of the As contamination level, even if they are much higher relative to the threshold levels $L_k, k = 1, 2$. As a matter of fact, in addition to the international threshold level $L_1 = 10\mu$g/L, and Indian level $L_2 = 50\mu$g/L, it would be quite appropriate to consider a staggering set of levels:

$$L_1(= 10) < L_2(= 50) < L_3 < \cdots < L_K, \tag{8.1}$$

where L_3 may be taken as 100μg/L, ..., L_K may be taken as 400μg/L. Ideally, one can define a continuous threshold level sequence $\{L_t : L_1 \leq t \leq L_K\}$, and consider a sequence of CSI $\{\kappa^*_{L_t} : L_1 \leq t \leq L_K\}$ and relate that to the health hazard impacts, some of which can be studied by epidemiological investigations, in conjunction with statistical modeling and analysis. Simply κ_L^* by itself does not convey the full impact. It is the health hazards associated with the CSI quantification that is important to draw the picture in a better and more meaningful way. Spatial heterogeneity of As contamination may throw useful information. Of course, all these refinements can be included in a progressive way once the basic As contamination problem is well sketched in terms of well defined statistical rationality and interpretations. This is by no means confined to the Indian sub-continent, and is equally shared by many other countries including Japan, Taiwan and certain parts of China. The problems might be somewhat different due to geological and climatic differences, but the resolution may be similar. We must bear in mind the affordability of any such investigation consistent with the socio-economic constraints.

References

AOKI, V., MILLIKAN, R. C., RIVITTI, E. A., HANS-FILHO, G., EATON, D. P., WARREN, S. J., LI, N., HILARIO-VARGAS, J., HOFFMANN, R. G., DIAZ, L. A. and AND THE COOPERATIVE GROUP FOR FOGO SELVAGEM RESEARCH (2004). Environmental risk factors in endemic Pemphigus Foliaceus (Fogo Selvagem). *J. Invest. Dermatol. (Symp. Proc.)* **9**, 34–40.

AOKI, V., SOUSA, J. X. JR., DIAZ, L. A. and AND THE COOPERATIVE GROUP ON FOGO SALVAGEM RESEARCH (2011). Pathogenesis of endemic pemphigus foliaceus. *Dermatol. Clin.* **29**, 1–5.

The Bangladesh arsenic mitigation water supply project: addressing a massive public health crisis. (1999) (The World Bank Group).

CHAKRABORTI, D., RAHMAN, M. M., MITRA, S., CHATTERJEE, A., DAS, D., DAS, B., NAYAK, B., PAL, A., CHOWDHURY, U. K., ROY CHOWDHURY, T., AHMED, S., BISWAS, B. K., SENGUPTA, M., LODH, D., DAS, A., CHAKRABORTY, S., CHAKRABORTY, R., DUTTA, R. N., SAHA, K. C., MUKHERJEE, S. C., PATI, S. and KAR, P. B. (2013). Groundwater arsenic contamination in India: A review of magnitude, health, social, socio-economic effects and approaches for arsenic mitigation. *J. Ind. Soc. Agricult. Stat.* **67**, 235–266.

CHATTERJEE, S. K. and SEN, P. K. (2000). On stochastic ordering of a class of poverty indexes. *Calcutta Stat. Assoc. Bull.* **50**, 137–155.

GASTWIRTH, J. L. (1975). A new index of income inequality. *Proc. Int. Stat. Inst.* **1**, 368–372.

HOSSAIN, M. F. (2006). Arsenic contamination in Bangladesh - an overview. *Agri. Ecosyst. Environ.* **113**, 1–16.

HOEFFDING, W. (1948). On a class of statistics with asymptotic normal distribution. *Ann. Math. Stat.* **19**, 293–325.

KARIM, M. (2000). Arsenic in groundwater and health problems in Bangladesh. *Water Res.* **34**, 304–310.

NATIONAL RESEARCH COUNCIL (2001). *Arsenic in drinking water: 2001 update.* National Academies Press, Washington DC.

SEN, A. K. (1973). *On Economic Inequality.* Oxford University Press, London.

sen, a. k. (1976). The measurement of poverty: An axiomatic approach. *Econometrica* **44**, 219–232.

sen, p. k. (1986). The Gini coefficient and poverty indexes: Some reconciliations. *J. Amer. Stat. Assoc.* **81**, 1050–1057.

sen, p. k. (1988). The harmonic Gini coefficient and affluence indexes. *Math. Soc. Sci.* **8**, 65–76.

sen, p. k. (1989). Affluence and poverty indexes. *Encyclop. Stat. Sci., Supplm.* **1**, 1–5.

sen, p. k. (1994). Bridging the biostatistics-epidemiology gap: The Bangladesh task. *J. Stat. Res.* **28**, 21–39.

sen, p. k. (2001a). Toxicology: Statistical perspectives. *Current Sci.* **80**, 1167–1175.

sen, p. k. (2001b). Absorption and ingestion toxicology. *Encyclop. Environ.* **1**, 1–4.

sen, p. k. (2001c). Inhalation toxicology. *Encyclop. Environ.* **2**, 1054–1064.

sen, p. k. (2012) Development and management of national health plans: Health economics and statistical perspectives. *Some Recent Advances in Mathematics and Statistics* (ed. Y. P. Chaubey). World Sci. Publ. Singapore, pp. 195–218.

sen, p. k. (2013a). Agricultural epidemiology and environmental toxicity: Some statistical perspectives. *J. Indian. Soc. Agri. Stat.* **67**, 151–181.

sen, p. k. (2013b) Some statistical perspectives in growth models. *Advances in Growth Curve Models.* (ed. Ratan Dasgupta). Springer, New York, pp. 35–49.

smith, a. h., lingas, e. o. and rahman, m. (2000). Contamination of drinking water by arsenic in Bangladesh: a public health emergency. *Bull. World Health Org.* **78**, 1093–1103.

Pranab K. Sen
Departments of Biostatistics,
and Statistics & Operations Research,
University of North Carolina,
Chapel Hill, NC 27599-7420, USA
E-mail: pksen@bios.unc.edu

Paper received: 21 June 2014.