CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report

Author (s): **T. J. Rao**

Title of the Report: **Environmental Sampling Techniques, A Brief Introduction**

Research Report No.: **RR2013-14**

Date: **October 11, 2013**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

ENVIRONMENTAL SAMPLING TECHNIQUES, A BRIEF INTRODUCTION*

by

T. J. Rao**

C.R.Rao Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS), UOH Campus, Prof.C.R.Rao Road, Gachibowli, Hyderabad-500 046 (A.P), INDIA

Abstract

In this paper, we shall briefly outline some of the basic techniques that can be used in environmental sampling with illustrations. We shall also discuss how the sample size is determined.

Key words: Environmental sampling, Simple random sampling, Probability proportional to size sampling, Systematic sampling, Adaptive and network sampling, Spatial sampling, Sample size, Hotspots, Data banks.

1. Historical references to Environmental Sampling.

In the great epic Mahabharata , one comes across the story of king Nala and Bhangasuri taking a walk in the well preserved forests . They loose their way and settle down under a tree for the night. Bhangasuri flaunts his environmental sampling skills by saying that the estimate of total number of fruits on the tree is 2095. Nala counts all night and is amazed at the correctness of the figure. Bhanagasuri with his expertise on throwing of dice , used a random sample of the branches to boost the sample data and get the estimate.

In Kautilya's *Arthashastra*, attributed to 321 B.C., it is mentioned that the Mauryas "*looked at the forests as a resource*". Further, *Arthashastra* unambiguously specifies the responsibilities of officials such as the *Protector of the Elephant Forests*:

"On the border of the forest, he should establish a forest for elephants guarded by foresters. The Superintendent should with the help of guards...protect the elephants whether along on the mountain, along a river, along lakes or in marshy tracts...They should kill anyone slaying an elephant".

Next coming to the Moghul Period, we note that Babur offers his description of fauna of India in a very systematic style. After giving the features of India's physical geography he proceeds to describe first the mammals, then birds, and, finally, aquatic animals.

Jahangir ordered his artists to portray animals and birds as well as prepare accurate paintings of flowers providing all the botanical details necessary for identifying the family of the flower.

In *Ain-i-Akbari*, Abul Fazl (circa 1590 A.D.) describes the Flora and Fauna in detail. The Moghul Emperors and nobles had another interest that indirectly brought them closer to ecology. Laying of gardens whether resting places and parks or flower gardens (*gulistan*), (*bostan*) and orchards was a favourite pastime of the kings, princes, princesses, and nobles.

---

*Paper based on the lectures given during the International Workshop on *Statistical Methods for Environmental Data Analysis* organised by the CR Rao AIMSCS,Hyderabad, 4-7 March 2013.

***Received partial support as Adjunct Professor under a grant from the Department of Science and Technology, Government of India (SR/84/MS:516/07 dated 21/4/08) to CR RAO AIMSCS, Hyderabad, India.*

Turning to a recent British Period, Francis Buchanan's Survey of Eastern India, 1807 describes two surveys he conducted, the first of Mysore in 1800 and the second of Bengal in 1807-14. In 1804, he was in charge of the newly founded 'Institution for promoting the Natural History of India' at Barrackpore near Calcutta. He made a comprehensive survey of Bengal and reported on topography,……., natural productions(particularly fisheries, forests, mines, and quarries), agriculture,……. His research includes an important work on Indian fish species, entitled *An account of the fishes found in the river Ganges and its branches(1822),* which describes over 100 species not formerly recognized.

2. Types of data.

Scientists are generally concerned with different types of data. Some of these refer to geography, geology etc..Most of the data relates to a point of time and some times to a period of time. The first type is more are less constant such as the topography of districts, rivers, locations of mineral reserves and so on. Once the data is collected, for example, by Survey of India or geological expeditions there is no need to collect it for the next few years. To collect data over a period of time , one needs to collect data at the initial point of time and at the relevant end point . Thus one needs to monitor data collection at a point of time through a well defined methodology. However, there are certain types of data which can be obtained only through a well designed experiment such as finding out which combination of N, P and K would give a better yield of a crop. In this paper, we shall look at the methodology for collecting data mostly at a point of time.

When a scientist or a statistician is collecting data with a specified objective in mind we call it primary data, while if the aim is to utilize already existing data (collected by some one else, such as Population Census of the office of the Registrar General of India or Large Scale Sample Surveys such as those conducted by National Sample Survey Office), we call it secondary data. While in both cases, it is very important to *cross examine* data before drawing inferences based on them, as emphasized by R. A Fisher, the Father of Modern Statistics , P.C. Mahalanobis, the Founder of the Indian Statistical Institute and C.R. Rao, the Doyen of Statistics, one has to scrutinize more carefully the secondary data collected by some one else before making use of it. We shall illustrate this by a couple of examples:

***Example* 1:**

A question is set as follows: An environmentalist recorded maximum temperatures from 15-24,May 2012 in deg. Celsius in Kolkata as 22,25,24,29,31,26,30,24,29 and 30 while in Kota as: 30,31,31,30,32,31,30,30,31 and 30. Which data is **more** variable?

***Example* 2:**

Hts. of four boys of class ten in a school are: 5ft.4in., 5ft.6in. 5ft.4in. and 4ft.9in., what is their mean?

In the first example, before jumping to a quick conclusion that the temperatures in the second list of Kota are less variable than in the first list of Kolkata, one should question the correctness of the data for Kota which are never as given here, the time specified being a midsummer in Rajasthan. The data is wrongly recorded or the place is not Kota but somewhere else.

For the second example, a good mathematician will add all the four values and divide by four, while a good statistician will notice that the last value is an outlier and delete and calculate the average of the first three and call it a modified mean. On the other hand, an environmental specialist would contact the family of the fourth child and find out whether there is any environmental effect that caused the stunted growth, such as a pesticide factory nearby or high tension power lines or low nutrition etc. before calculating the mean.

Primarily, Inference is based on Descriptive Statistics computed from the data such as measures of central tendency, namely mean, median , mode or measures of dispersion like standard deviation, range etc. . Inference based on Analytical Statistics answers questions by testing hypotheses as evidenced by the data.

3. Complete Enumeration *Vs*. Sample Surveys.

There are two methods of data collection namely, Complete Enumeration(C.E.) and Sample Surveys(S.S.).

From a Sample Survey, we can obtain the data with a greater speed at a less cost and with a greater scope. However, since different samples give different estimates, we come across what are called sampling errors(s.e.). There are no sampling errors when one deals with Complete Enumeration, but when data is collected on the same population units by different individuals or different survey instruments , one may not obtain the same answer, because of various factors such as non response, non availability, measurement errors, recording errors etc. . Thus even though there are no sampling errors present in complete enumeration , we have 'non sampling errors' that creep in. Since it is possible to have non response, non availability, measurement and recording errors etc. in sample surveys too, we encounter non sampling errors when we conduct a sample survey as well, but to a lesser extent since the size of the sample is much smaller than the population size.

It is interesting to note that way back in late thirties and early forties , Mahalanobis organized the Jute Survey in Bengal . The survey gave a production of 7540 bales which employed around 600 to 700 employees at a cost of 8 lakh rupees while the official plot-plot enumeration gave a production figure of 6304 bales, which employed 33,000 personnel at a total cost of 82 lakh rupees. It turned out that the more reliable customs and trade figure was 7562 bales (1 bale = 400 lbs.) showing that Mahalanobis's sample survey estimate was an underestimate by a mere 0.3 % while government's complete enumeration was an underestimate by 16.6%.

Unlike Socio economic surveys where the population units are households or establishments or individual persons, environment surveys refer to air, water, soil and biota etc. .Scientists are interested in assessing the characteristics of the units, extent of resources, hazard levels , risk factors etc..Thus it makes it difficult to obtain proper sampling frame from which selection of units can be made.

Population to be sampled may consist of discrete and finite entities such as lakes, wet lands, villages, agricultural plots, households etc. or continuous such as streams, forests, volume of water, timber or chemical concentrations etc..

Besides conventional sampling frames such as list frames and maps, we may also have Remote Sensing Data, GIS etc..Data may be collected at a particular point of time as in the form of a 'one-shot' survey or long term survey to continuously gauge the trends and change.

Furthermore, data collection need not necessarily be on units from well defined localized and identifiable sites (lakes, rivers, forests etc.). It could be from units relating to global challenges such as global warming, industrial waste disposal, arsenic poisoning, CO2 emissions etc. which do not refer to a particular locality.

One is usually interested in Parameters such as Mean,Variance(sampling error), Coefficient of Variation (C.V.), Distribution Function (D.F.), Proportions .The primary objective is to obtain estimates of descriptive statistics together with estimates of sampling errors or to determine trends looking at the data at different points of time. A scientist is also interested in the relationship between variates/attributes measured at a site. In situations where physical data collection is not possible or difficult, Remote Sensing data, where available could be made use of.

For any project, an idea of Cost function is necessary which involves identification costs, sampling costs, laboratory costs. Even though sampling errors are reduced by adopting good design, non sampling errors which creep in at different stages are to be identified and taken care of. Some of the sources of non sampling errors are non response from denied access , safety, health hazards, missing data, measurement and recording errors, defective frames, data handling, lack of supervision and training ,political and economical problems.

4. Basic sampling designs

Consider a finite population of N identifiable (labelled) units $U = \{ U_1, U_2, ...,U_N \}$. The characteristic of interest y, called the study variable takes values Y_i on unit U_i, i= 1,2,...,N. In most of the surveys an auxiliary variable x related to the study variable is readily available taking values X_i on unit U_i. We shall first briefly discuss three basic sampling designs.

Simple Random Sampling.

In Simple Random Sampling(SRS) or Equal Probability Sampling , all the units are given equal chance of selection, namely $P(U_i) = 1/N$, for all i = 1,2,....,N. If unit is allowed to repeat , the

selection is called With Replacement (WR), if it is removed before selecting the next unit, it will be referred to as selection WithOut Replacement(WOR).

Suppose that the sample consists of n units labelled as ($u_1, u_2, ..., u_n$) with the corresponding y- values of the study variable as $(y_1, y_2, ..., y_n)$. For both SRS WR and SRSWOR schemes , the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$. Also the sampling error measured by variance of $\bar{y}$,

$$V(\bar{y}) = \sigma^2 / n , \qquad (4.1)$$

where σ^2 is the population variance given by $\sigma^2 = \sum_1^N (Y_i - \bar{Y})^2 / N$. Since this is an unknown parameter, we use the sample data to get an unbiased estimate of σ^2 which is given by

$s^2 = \sum_1^n (y_i - \bar{y})^2 / (n - 1)$.

Thus an unbiased estimate of the variance is given by

$$\hat{V}(\bar{y}) = s^2 / n \qquad (4.2)$$

And this can be computed from the sample data itself. However for SRSWOR , the variance is given by

$$V(\bar{y}) = \{(N-n)/(N-1)\} \sigma^2 / n \qquad (4.3)$$

with a multiplicative constant $\{(N-n)/(N-1)\}$ which is less than unity , thereby implying that WOR sampling is less variable than WR sampling.

Here σ^2 is estimated by $\{(N-1)/N\} s^2$ unbiasedly and thus an unbiased estimate of the variance in this case is given by

$$\hat{V}(\bar{y}) = \{ (1/n) - (1/N) \} s^2 . \qquad (4.4)$$

Example A.

A simple random sample (SRS) of 10 zones from a population of 105 Tiger Reserve Zones is selected WithOut Replacement (WOR) and the number of tigers found in these sampled zones are 24,12,17,16,03, 31,14,18,16 and 21. Estimate the total number of Tigers in the Population. Also calculate an unbiased estimate of the sampling error (measured by variance).

Instead of estimating the population mean, sometimes the scientist faces with the problem of estimating the Population Proportion *P*.

Example B.

There are 400 (big and small) lakes around Hyderabad. A sample of size 20 is drawn by Simple Random Sampling With Out Replacement with Sl. Nos.

216,318,314,118,020, 306,229,224,044,168, 111,144,266,348,167 288,349,279,045,298.

It is found that those marked RED are highly polluted. Estimate unbiasedly the Proportion P of lakes in Hyderabad which are highly polluted. Also estimate the sampling error of your estimate.

Example C. Consider the following criteria:

a) Usage of plastic bags

b) Printing of train/flight tickets.

c) Printing reprints of published papers to read later

d) Leave the lights on in the office during absence from the room.

If a person satisfies any of the above criteria, he or she is considered as *environment unfriendly*.

Our Population Parameter of interest now is not a Mean or Total of a study variate , but P , the Proportion in the audience that belong to this category. The problem is to estimate P unbiasedly and estimate its sampling error based on a random sample of persons of size n from the audience of size N.

By simple application of the above technique, treating the study variable as dichotomous (y- value 1, if the unit belongs to the category and 0, otherwise) we can show that for SRSWR an unbiased estimate of P is given by

$$\hat{P} = p,$$

the proportion of units in the sample belonging to the category. The sampling error is unbiasedly estimated as

$$\hat{V}(p) = p(1-p)/(n-1). \qquad (4.5)$$

For SRSWOR while $\hat{P} = p$,

$$\hat{V}(p) = \{(1/n) - (1/N)\}\, n\, p(1-p)/(n-1). \qquad (4.6)$$

Example D. From an audience of 120 in a Workshop on Environmental Sampling , a sample of 20 participants is drawn by SRSWOR design. It is found that 14 of them were found to be *environment unfriendly* based on criteria stated in the previous example. Estimate the Proportion P of *environment unfriendly* participants in the population and obtain an unbiased estimate of sampling error.

In this example, $\hat{P} = p = 14/20 = 0.7$ and $\hat{V}(p) = \{(1/20) - (1/120)\}20(0.7)(0.3)/19$

$= 21/2280 = 0.0092$

Probability Proportional to Size (PPS) Sampling.

When the units in the population are of different sizes such as agricultural plots, hospitals, industrial establishments , it is not efficient to give equal chance of selection to all the units. In such a case, a unit with size measure, say X_i is given a chance proportional to X_i , namely $P_i = X_i / X$, where

$X = \Sigma X_i$. Once a sample is selected by PPS WR scheme, the sample mean $\bar{y}$ is no longer an unbiased estimate of the Population mean . In this case, an unbiased estimate of the population total $Y = \Sigma Y_i$ is calculated as

$$\hat{Y} = \sum_1^n (y_i / p_i) / n , \qquad (4.7)$$

based on the sample selected and the population mean is unbiasedly estimated by $\hat{Y} / N$. Further, an unbiased estimate of the sampling error is calculated as

$$\hat{V}(\hat{Y}) = \Sigma_1^n \{(y_i / p_i) - \hat{Y}\}^2 / (n(n-1)). \qquad (4.8)$$

Systematic Sampling.

In forest surveys, plantations are nicely (linearly aligned) and starting at a point every k th unit is selected in a systematic way. Borrowing this technique, when it is required to select sample of size from a population of N units, we define $[N/n]$ = integral part of $N/n = k$, say, as the sampling interval and choose a random start r from $[1,k]$. Select the units U_r, U_{r+k}, U_{r+2k},..... till all the units are exhausted as a sample. This is called a *Linear Systematic Sample* (LSS). The sampling technique is very simple. However here the sample size is not fixed (depending on the random start, it may be n or $n+1$.). Further the sample mean of the associated y- values $\bar{y}$ is not an unbiased estimate of the population mean, unless N is divisible by n in which case the sample size is also fixed and equal to n. To circumvent this, we take the random start r from $[1,N]$ and choose the sample as consisting of units:

$$U_r , U_{r+k}, U_{r+2k}, \ldots\ldots, U_{r+(n-1)k}$$

where U_{N+t} is defined as U_t in a circular way. Thus this is called a *Circular Systematic Sample*(CSS). For this simple scheme, not only sample size is n(fixed) but the sample mean of the associated y – values $\bar{y}$ is an unbiased estimate of the population mean.

Even though it is very easy to apply this procedure in practice , since the sample estimate depends only on *one random start*, it is NOT possible to estimate the sampling error (variance) unbiasedly based on a single sample. Instead we use approximate error estimates. Based on m *independent* samples, the average $\bar{t}$ of m independent estimates t_i gives an unbiased estimate of the mean and the sampling error is estimated unbiasedly by

$$\hat{V}(\bar{t}) = \{\sum_1^m (t_i - \bar{t})^2\} / m(m-1) . \qquad (4.9)$$

When $m = 2$, $\bar{t} = (t_1 + t_2) / 2$ and $\hat{V}(\bar{t}) = (t_1 - t_2)^2 / 4$.

Example E. An early application of systematic sampling can be found in the ICAR Pilot sample survey conducted on the Malabar coast of India for estimating the catch of marine fish (Sukhatme *et al*.,1958).

Example F. A forest ranger has selected a systematic sample of 20 trees from a new plantation of 200 to study their growth. His random start from 1 to 10 is 4 and he has selected units numbered 4,14,24,….., 194. His measurements y on these units are: 22,28,18,26,26,16,21,13,29,24, 22,25,21,25,22,18,17,25,23,22. Estimate the population mean.

Since unbiased estimation of sampling error based on a **single** systematic sample is not possible, he wishes to divide this into two independent half samples of size 10 each. From each independent sub sample estimate the mean and use this information to obtain a pooled unbiased estimate of the population mean and obtain an unbiased estimate of the sampling error.

Spatial Sampling.

As a simple extension of systematic sampling, in 'square grid sampling' in two dimensions , a pair of random numbers are required to fix the unit in the upper square grid. For certain correlograms Mate'rn (1960) has shown that a square grid sample performs better. Some of the earlier studies are due to Haynes(1948), Quenouille (1949) and Das(1950). However, these are based on spatial homogeneity and do not discuss spatial correlation. Rilpey's (1981) work gives the details for spatially auto correlated populations. With correlation between units, there will be a reduction in sample size for estimating population parameters (see for example, Griffith,2005 and Wang *et al.* (2009)). When dealing with populations having spatial heterogeneity, the overall population variance and auto correlation structure need to be taken into account for sampling purposes(Wang *et al*.,2010).

Stratification.

When the population is very heterogeneous, we form sub groups (or sub populations) called *strata* such that the units within each stratum are homogeneous. This increases the efficiency. Sometimes, there is a need for stratification, because estimates are needed for sub populations and also it would be operationally convenient. We may *allocate* the total sample size to each stratum in proportion to the size of the stratum. There are also other optimum allocations. Within each stratum a sampling design is employed and the overall estimate is obtained as a weighted combination of stratum estimates. Our objective is not only to obtain unbiased estimates of parameters using sample data, but also to obtain the variance estimates from the same sample. Thus it is important to select *at least two* units from each stratum in order to estimate the within stratum variances unbiasedly.

It was shown by Cochran(1946) that on an average systematic sampling fares better than stratified sampling which in turn is better than unstratified SRS. When the population is two dimensional, extensions to systematic sampling are available. When it is desired to use Grid

Sampling, it is better to have triangular grids which are found to be more effective. A Radom Tesselated Stratified design is preferred by environmentalists which has a good spread and greater efficiency compared simple systematic and srs designs.

Recap Example G. An environmentalist belonging to Andhra Pradesh who visits temples often, has taken water samples from Godavari river near Kopergaon,, Baasara and Rajahmundry and gets it analysed for pollution levels. Based on the estimates from this sample , he writes a paper on the *Pollution of River Godavari*. Comment on this approach and suggest a proper methodology.

5. Certain other sampling designs.

Adaptive and network sampling designs.

In certain situations , it is very difficult to get a sampling frame. In that case one has to build up a suitable frame. For example let us say that we are interested in factories on a riverside which are polluting the river. We can add units one by one by ascertaining from the factory workers who else in that area is indulged in polluting. Thus we build up a network which forms the basis for an adaptive sampling. Similarly, when dealing with animal abundance, the sampler keeps on adding units to the sample whenever he comes across an abundance. On the other hand , if we are interested in the prevalence of a rare characteristic in a population of households, the adult members are asked to report not only if they possess that characteristic, but whether their siblings also possess this and thus a network of units is obtained so as to utilise a network sampling technique. (See, Thompson and Seber, 1996).

***Capture recapture techniques*.**

A technique that is adopted to estimate the unknown population size itself is known as capture- recapture technique. Suppose that we wish to estimate the number N of fish in a pond. A simple version of the technique, under certain assumptions is: Capture a first sample of M fish, mark and release them. After a while, take a second sample of size m and let r be the number recaptured. An unbiased estimate of N is given by

$$\hat{N} = \{ (M + 1)(m + 1) / (r + 1) \} - 1.$$

Some of the other practical techniques like *Line Intercept Sampling*, *Composite Sampling* among others are discussed in Thompson (2002) .

6. Determination of sample size.

A question very often posed in many sample surveys is: “What should be the sample size?”. The answer depends on the requirements of the scientist and the available information. We shall briefly discuss some of the simple methods of determining the sample size.

Using Cost Function.

Consider a simple linear cost function

$$C = C_0 + C_1 n$$

where C is the Total Cost(Budget) and C_0 is the overhead(fixed) cost and C_1 is the sampling cost per unit. Given C, C_0 and C_1, we can solve for n as

$$n = (C - C_0) / C_1$$

Using Coefficient of Variation (C.V.).

A more meaningful demand would be : Find the sample size such that the Coefficient of Variation (c.v.) of the estimated mean is a given quantity, say g (this is usually 5% or 10%). For SRSWOR, from the formula we have

C.V. (estimated mean) $= (\sqrt{V(\bar{y})}) / \bar{Y} = \sqrt{[\{(N-n)/(N-1)\}\sigma^2 / n]} / \bar{Y}$.

If this is required to be g, then n, the sample size needed is given by

$n = 1 / \{ (1/N) + (\frac{N-1}{N}) (g^2 \bar{Y}^2 / \sigma^2) \}$ which for large N, becomes $n = \sigma^2 / g^2 \bar{Y}^2$.

This requires the knowledge of the C.V. for the population, namely, $\sigma / \bar{Y}$ which can be obtained from a pilot survey or past data.

Example H. Based on a Simple Random Sample selected WithOut Replacement from a population of N = 491 lakes in a region, it is desired to estimate the average amount of garbage dumped into these lakes. The scientist prefers to have the coefficient of variation (c.v.), g of the estimate not more than 5%. From a previous study , it is known that the region has a c.v. of 0.2771 for this characteristric. Find the sample size required.

Solution. Using the formula, we get $n = 1 / \{ (1/491) + ((490/491) ((0.05)^2) / (0.2771)^2 \} = 29$. Formula for large N, gives the required size as $n = 31$.

Using a probabilty requirement.

Sometimes, It is desired to obtain the sample size when the relative error (r.e.) in the estimated population mean using a design, say SRSWR, is to be controlled at a given level, say r. Further, this r.e. can be allowed to exceed with a chance of α%. Here r.e. $= (\bar{y} - \bar{Y}) / \bar{Y}$ and

$$\text{Prob.} \{ -r \leq ((\bar{y} - \bar{Y}) / \bar{Y}) \leq +r \} = 1 - \alpha.$$

Assuming normality of y- values which gives $\bar{y} \sim N(\bar{Y}, \sigma/\sqrt{n})$ write

$$\text{Prob.}\{-z \leq (\bar{y} - \bar{Y})/((\sigma/\sqrt{n})\} \leq +z\} = 1-\alpha.$$

Comparing the two Prob. Statements we get, $r\bar{Y}/(\sigma/\sqrt{n})\} = z$ giving $\sqrt{n} = (z/r)(\sigma/\bar{Y})$ from which n could be obtained.

Here we should have an idea of population C.V. $=(\sigma/\bar{Y})$ from past data, pilot survey or related auxiliary information as in the *previous* example.

7. Detection of Hot spots.

Square, rectangular or triangular grids are used and samples are collected at the grid nodes. Of these triangular grids are found to be more efficient. It is important to determine the grid spacing and refer to the nomograms which are available. From these nomograms, size of hot spot, probability of 'no hit',etc. can be found.

Also the chance that a hot spot exists when there is 'no hit' can also be calculated by using a simple application of Bayes Theorem. For further details, we refer to Gilbert(1987).

8. Assessing hazards-three illustrations.

We shall now look at a couple of examples to illustrate hazard assessment:

This assessment with respect to glacial lakes in Himalayas is an important problem on which certain attempts are made in the current literature. However, in one such attempt, the models used for the probability of such an occurrence using regression parameters of a logistic fit are meant for Canadian data which may not suit Indian conditions.Thus one has to answer questions such as, whether we have enough data ,whether the assumptions of the model are valid, whether the model itself is relevant etc..

Arsenic contamination especially in the Ganges Bengal Delta Plain (BDP) which became a health hazard for more than 50 million villagers in the state of West Bengal itself , has received much attention by environmental researchers (see, for example, Singh (2006) , Bhattacharya *et al* . (2001), Chakraborti *et al.* among several others). There are several hot spots (As > 0.01 mg/l as per WHO standards and As > 0.05 mg/l as per BIS) in the middle and lower ranges of the river Ganges. Though arsenic contamination due to natural geo chemical processes such as volcanic eruptions or suspended particles in air circulation can not be controlled , it is the industrial pollution of rivers, underground waste disposals that are causes of concern for the environmentalist. Further, another arsenic pollutant , namely, Chromated Copper Arsenate (CCA) in wood processing is still being used in several countries, even though United States EPA , Canadian and European agencies have restricted this.

Referring to the arsenic pollution in the central and lower regions of the river Ganges, we note that not many research studies are available from the statistical point of view. Though several guide lines are

prescribed for collecting samples, it is usually the "grab samples" which are taken for laboratory analyses. Once the major locations of hot spots are known, one could use auto sampler devises for collecting water samples at , say, systematic sampling time intervals. Systematic sampling can be used for selecting the dates of a month as well , starting with a random start date. One of the earlier statistical studies in the eighties ,w.r.t. arsenic pollution in the river Ganges was by Somesh Dasgupta. Purkait *et al.*(2008) used an artificial neural network (ANN) model as multilayer perceptron (MLP) architecture to estimate arsenic concentration in ground water of Malda district.. In continuation , recently Sengupta *et al.* (2010) analysed data based on 700 drinking water samples from tube wells in the hot spots of deltaic alluvial plains of Bengal delta in Malda district. They prepared As zonation map with 6 groups. Statistical tests for isotropy(circular uniformity/randomness) against preferred direction of contamination of arsenic as well as Change Point (CP) test to detect angle at which contamination changes, were used. However, for selecting sample points in a polluted river care has to be taken so that proper statistical analyses can be made. For monitoring pollution along a river , a mathematical formulation of the problem of optimally selecting the sample points was given by Alvarez-Va'zquez *et al.* (2006). They also provided an efficient algorithmic solution.

Next, we shall refer to the hazards of natural floods.In 1926, a catastrophic flood occurred in the Brahmani river in Orissa and several low-lying areas got flooded. It was decided by a group of engineers that the river bed had risen and hence the height of embankments should be raised by several feet. The matter was referred to PC Mahalanobis of the Indian Statistical Institute who did several statistical analyses. Mahalanobis found a significant correlation between the rainfall in the catchment area and height of the river flood in the delta area. His report of 1930 to the government of Orissa led to the construction of the multi-purpose project Hirakud dam, thus avoiding the raising of the embankments. This early work in Operations Research is perhaps very much applicable in the current crisis of frequent flooding of Visakhapatnam (Vizag) Airport in South India. Even though a new elevated runway was built, still the entire surrounding areas get flooded, if there is a heavy downpour. It is not the rainfall in Vizag that is causing the floods, it is a problem arising at the catchment area during the heavy rainfalls there and the operations at the reservoirs. Just as Mahalanobis thought, perhaps there is no need to raise the river embankments all the way up to the point where the river meets the sea in Vizag as was suggested. Further research is needed on this aspect.

9. Available data sets.

In the Indian context, some of the data inventories available are: toxics release inventory (TRI), inventory of hazardous chemicals import in india, national wetland inventory and assessment (NWIA), national forest inventory in India, Inventory of aerosol and sulphur dioxide emissions from India etc.. It may also be noted that the Environment Statistics Section of the United Nations Statistics Division (UNSD) collects environmental data and the sixth round on water and waste statistics was obtained in 2010. UNSD environmental indicators derived from these data are published in ENVSTATS.

Biodiversity indeces such as Shannon's diversity, Simpson's diversity etc. are considered in the literature on the subject , but in the context of India not much information is available due to the complex nature of the problems.

References.

Alvarez-Va'zquez,L.J., Martinez,A., Va'zquez-Mendez,M.E. and Vilar,M.A. (2006). Optimumlocation of sample points for river pollution control, *Maths. and Computers in Simulation,* **71,** 149-160.

Bhattacharya, P., Jacks, G., Jana, J., Sracek, A., Gustafsson, J. P. and Chatterjee, D. (2001). Geochemistry of the Holocene alluvial sediments of Bengal Delta Plain from West Bengal, India: implications on arsenic contamination in groundwater. In *Proceedings of National Seminar on Groundwater Arsenic Contamination in the Bengal Delta Plains of Bangladesh* (eds Jacks, G., Bhattacharya, P. And Khan, A. A.), University of Dhaka, Bangladesh, TRITA-AMI Report, 2001, **3084**, pp. 21–40.

Chakraborti,D., Mukherjee,S.C., Pati, S., Sengupta,M.K., Rahman, M.M.,Chowdhury, U.K., Lodh,D.,Chanda,C.R. ,Chakraborti,A.K. and Basu, G.K. (2003). Arsenic Groundwater Contamination in Middle Ganga Plain, Bihar, India: A Future Danger?, *Environmental Health Perspectives* **111**, 1194-1201.

Cochran, W.G. (1946). Relative accuracy of systematic and Stratified Random Samples for a Certain Class of Populations. *Ann. Math. Statist.,* **17**, 164-177.

Das, A.C. (1950). Two-dimensional systematic sampling and the associated stratified and random sampling, *Sankhya,* **10,** 95-108.

Das Gupta,S., *Unpublished Tech. Report*, Indian Statistical Institute.

Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring* , Chapter 10, Wiley, N.Y.

Griffith, D.A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. *Annals of the Association of American Geographers,* **95,** 740-760.

Haynes, J.D. (1948). *An empirical investigation of sampling methods for an area,* M.S.Thesis, Univ. of North Carolina.

Mate'rn, (1960). *Spatial Variation*, 2nd ed. In: Lecture Notes in Statistics, vol. **36**. Springer, Berlin.

Purkait, B., Kadam, S.S.and Das, S.K. (2008). Application of artificial neural network model to study arsenic contamination in groundwater of Malda District, Eastern India. *Journal of Environmental Informatics,* **12** , 140-149,

Quenouille, M.H. (1949). Problems in plane sampling. *Ann. Math. Statist.* **20**, 355–375.

Ripley, B. (1981). *Spatial Statistics.* Wiley, N.Y.

Sengupta,A.,Purkait,B and Roy,M (2010). *Detection of Arsenic contamination pattern by Change Point Statistics around hotspots in the tubewell water of some parts of the Bengal Delta*, Proc. of the International Conference on Frontiers of Interface between Statistics and Sciences, CRRao AIMSCS,Hyderabad.

Singh, A.K. (2006). Chemistry of arsenic in ground water of Ganges-Brahmaputra river basin . *Current Science,* **91,** 599-606.

Sukhatme,P.V. , Panse,V.G. and Shastri, K.V.R. (1958). Sampling techniques for estimating the catch of sea fish in India, *Biometrics,* **14,** 78-96.

Thompson,S.K. (2002). *Sampling,* Wiley, N.Y.

Thompson,S. And Seber,G.A.F.(1996). *Adaptive Sampling,* Wiley,N.Y.

Wang, J.F., Christakos, G., Hu, M.G. (2009). Modeling spatial means of surfaces with stratified nonhomogeneity. *IEEE Transactions on Geoscience and Remote Sensing,* **47**, 4167–4174.

Wang, J.F., Haining, R.P., Cao, Z.D. (2010). Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. *International Journal of Geographical Information Science* **24,** 523–543.