

# On Parameters Repeated Estimation Methods (PREM's Method)<sup>1</sup> And its Applications in Data Mining

Vithanage Pemajayantha<sup>2</sup>

*Modelling and Simulation Research Group  
School of Quantitative Methods and Mathematical Sciences  
University of Western Sydney, P.O. Box 1797, Australia,  
and  
Software Development and Research Centre, IU Japan*

## Abstract

A new class of estimation of parameters is proposed for data mining, analysis and modeling of massive datasets.

With the expansion of Information Technology, the present problem with many scientists is the analysis and modeling with extremely large databases, sometime refers to as data mining or knowledge discovery in databases. It was found that many attempts used to solve this problem were based on classical approaches such as regression, classification and multivariate techniques, and even summary statistics such as mean and standard deviations are still having problem of estimation with extremely large datasets. Because classical statistical approaches were developed historically to cater the limited availability of data, they do not intend to solve the problem with massive dataset.

In this study, certain properties of sub-totaling and repeated estimation of population parameters were used to establish a new statistical method for estimating summary characteristics of populations, and relationships between variables with extremely large datasets. While the method has straightforward applications in data mining and analysis of large databases, it poses the significance of further statistical research.

**Keywords:** *Data Mining, PREM's Estimation, Modeling with Massive Data.*

## INTRODUCTION

Data mining or extracting knowledge from large databases is presently done using classical statistical procedures such as visualization of data in forms of histograms, various types of charts, classification techniques, regression analysis, multivariate analysis together with some network such as neural-network relying on classical methods of sampling or re-sampling from large datasets. Hardly any attempt is made on full use of data found in extremely large database as it is practically difficult or impossible to carry out calculations. It is merely because most of the statistical techniques ended with a sum of an extremely large number of data points thus exceeding the computer memory

---

<sup>1</sup> Paper presented at International Conference in Statistics, Hawaii, June 4-9, 2003

<sup>2</sup> Visiting Professor, IU Japan, Head, Modelling and Simulation Research Group, School of Quantitative Methods and Mathematical Sciences, UWS, Australia

in analyses. Consequently, tera-flops, gega-bytes and dual and multi-processing etc. have reached its supreme at any point of time in data mining. Obviously, many data mining packages use only a part of data, for instance some dedicated statistical packages for data- mining such as Megaputer, Polyanalyst, STATISTICA Data Miner use classical methods with a limited number of data points. Even some articles found in dedicated resources have used only a very limited number of data points with classical statistical methods. It prompts the following question. Why do we need only a pinch of data when the whole set of millions of data points is readily available in the electronic media, sometimes collected electronically and accurately using modern computers that are capable of extracting them in few minutes or milliseconds? Classical argument was that computation of data takes more computer times and so is costly. Clearly, this argument is no longer valid for the difference in calculation of small data to large dataset could be a matter of three to four hours the maximum or few minutes on the average whereas editing or publishing research findings takes years.

### **1.1 Problem with Classical Methods in Data Mining**

The problem is that all classical methods of statistical computing lead to very large sums of numbers or extremely large integers at a certain stage, and handling extremely large numbers is yet a problem beyond human comprehension as well as difficult to handle with the latest computers and analytical packages. However, some mathematical approaches such as quantum computations are being under investigations; for instance, it is expected that quantum computers might partly solve the problem. But, we do not know when it will happen. This means that we cannot handle too many data points regardless of the giga-bites and tera-flops of world largest super computers and the availability of large electronic databases. It points to the fact that world demands refreshment of quantitative techniques, statistics. In this context, it is heartening to observe that statisticians are in good demand today.

### **1.2 Why Data Mining is Important?**

Data mining also known as knowledge discovery in databases has been defined as “non trivial extraction of implicit, previously unknown, and potentially useful information from data”. Today, there are millions of databases containing very valuable pieces of information but real pictures are hidden within these databases. To answer many human problems such as medical problems, economic problems, social and cultural problems, we need to discover the true pictures or relationships hidden in large databases. Medical database itself is an invaluable source of data. For instance blood pressure, ECG, reactivity to some antigens, drug administration and related problems such as effect of treatments as well as their side effects being well recorded and readily available in many advanced countries including Australia. What might cause a disease or an epidemic could be hidden in medical databases. Meteorological data are another source of useful information on environmental problems. These sorts of databases carry pieces of information collected within milliseconds to several years. Regardless of how old they are, some data never get redundant. For instance geological formations of rocks take millions of years and so is natural contamination of lakes and reservoirs. How fast the man can destroy such geological formations or natural environment? What can we do to preserve environment when we really need to interfere with it. The answers to these questions may be hidden in many such databases. That is why we need to extract the hidden knowledge from large databases readily available or yet to build.

### **1.3 Problem of Handling Large Sums In Statistics**

Suppose we need to compute the mean of a large dataset. We need to add all numbers in the dataset and divide that sum by the total number of data points. If all numbers are more than 1, clearly the sum reach to a very large number much bigger than the total number of data points. One simple way to solve this problem is to divide all the numbers by a large number, for example make measurements on inches to yards or centimeters to kilometers. The method works naturally provided that the total number of data points does not exceed the computer memory, say up to  $n$ . It implies that simply  $n+1$  data

points can exceed the memory storage as well as  $2n, 3n, \dots, kn$ . The problem gets far more serious as the databases become increasingly huge. Therefore, the Parameter Repeated Estimation Methods (PREM) of estimating mean and other parameters with manageable amount of data at a time are proposed here as a contribution to solve the above problem.

## 2. PREM ESTIMATION OF MEAN

As the name implies, we repeatedly estimate a parameter in this method using small disjoint sets of data at a time until we complete the whole dataset. The method tunes the estimate until it reaches the parameter using the whole dataset. Amazingly, not only this simple method works very well with mean, but also it provides the actual parameter based on the whole dataset that represent population of interest. Furthermore, PREM estimation for standard deviation, variance, regression parameters are discussed in the subsequent analysis.

Let  $X_i \ i = 1, \dots, N$  be a set of data where  $N \gg n$  such that  $n$  is the largest integer that could be stored in a computer memory (Here we use the notation ' $\gg$ ' to indicate much greater than). The following theorem states the computability of mean of  $X_i$  for a certain subset of  $X_i \ (i = 1, \dots, n) \ n \ll N$ .

### **Theorem 1** (PREM's 1<sup>st</sup> Theorem of Computability)

Mean of  $n$  number of data points  $X_i \ i = 1, \dots, n$  is computable in terms of  $q$  units *iff* there exists  $q$  such that

$$q_0 \leq q \leq n \Leftrightarrow \left( \sum_{i=1}^n X_i / q_0 \right) \leq n \text{ for some } q_0, q_0 \in R$$

where  $n$  being the largest integer that could be stored in a computer memory.

### **Proof:**

Suppose  $n$  is the largest integer that could be stored in a computer memory.

If  $\left(\sum_{i=1}^n X_i\right) < n$ , then  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n} < 1$  and computable

If  $\left(\sum_{i=1}^n X_i\right) > n$ , then the quantity,  $\sum_{i=1}^n X_i$  is uncomputable.

Let  $\sum_{i=1}^n X_i q_o = n$

$\Rightarrow \sum_{i=1}^n X_i q$  is computable  $\Leftrightarrow q \geq q_o$

and for computability  $q \leq n$  completes the proof

Let us consider  $l$  number of largest subsets of  $X_i$  (Say,  $L$  sets) such that for each  $L$  set there exist  $q_o$  such that  $\text{sum}(X_i/q_o) < n$  and  $q_o < n$ . According to PREM's 1<sup>st</sup> theorem, mean values of these subsets are computable if the size of each subset  $L$  is less than  $n$ .

It prompts to the following general Theorem of computation of partial means and recursive estimation of population mean for any dataset.

**Theorem 2** (PREM's 2nd Theorem of Computability)

Let  $\mu = \text{sum}(X_i/N)$  where  $N \gg n$  where  $n$  being the largest integer that could be stored in a computer memory,  $\mu$  is recursively computable:

Suppose partial sums of  $X$  are computable in its original units. For instance,  $\text{SUM}(X_j)/k(t)$  are partial averages of datasets of size  $k(t)$ , and  $s(t)$ 's are the numbers of partial averages at a computational step  $t$ . The following recursive formula computes the mean value of a population.

$$\bar{X}_i(t) = \frac{\sum_{j=1}^{k(t)} X_j}{k(t)} \quad \text{for } i = 1, \dots, s(t)$$

The step  $t$  denotes recursive computation, and  $k(t)$  is a constant for each step  $t$ .

The above recursive formula returns  $\mu$  at  $s(t)=1$  with a certain probability of

$$\mu = \bar{X}(t)$$

**Remark 1**

Let us denote the above repeated estimate of mean as  $PREM(\bar{X})$

When  $k(t)$  divides  $s(t)$  without a remainder in all steps,

$$\mu \equiv PREM(\bar{X})$$

or

$$\mu = PREM(\bar{X}) \text{ with probability } 1$$

**proof:**

*Let  $N$  be the total population,*

*For a subset of population  $N_t$*

*if the following relationship exists at  $t^{th}$  partial average*

$$\mu_{\tau} = PREM(\bar{X}) \Leftrightarrow s(t) = 1$$

*Thus,  $PREM(\bar{X})$  for  $N_{t+1} = N_t k(t+1)$  observations must return*

*$k(t+1)$  number of  $\mu_{\tau}$ 's*

*Thus, if we select  $k(t+1)$  of  $\mu_{\tau}$  at the step  $t+1$ ,*

$$\mu_{\tau+1} = PREM(\bar{X}) \Leftrightarrow s(t+1) = 1$$

*selecting  $k(t+1)$  such that total population  $N = N_t k(t+1)$  completes the proof.*

Based on PREM's theorems of computability, the following two results on computational procedures emerged.

**Result 1**

Suppose a subset of data containing  $g_j$  number of observations is selected at a time where  $g_j < n$  for all  $j$ .

Let  $\text{average}(X_i)$  be  $\sum (X_i/q_o)/g_j$  where  $g_j$  is the number of observations in each dataset. Note that  $g_j < n$  guarantees the computability of each group mean in terms of  $q_o$  scale.

Select  $g_o$  such that  $g_o \leq g_j$  for all  $j$ . Intuitively, the mean values of  $g_o$  subsets of data are computable.

Therefore, repeated estimation of mean values of subsets replacing  $X_i$  with corresponding set of partial averages leads to the population mean for the whole dataset, and setting  $q_o = 1$  provides the actual mean for the whole dataset based on their original scale under certain conditions as illustrated below in Result 2.

## Result 2

Let  $q_o = 1$ , and let us define that

$\text{Mean}_{g_o}(X) = \text{Mean values of each } g_o \text{ subsets of data from } X_i.$

There exists a certain  $g_o$  such that  $\sum(X_i) (i=1, \dots, g_o)$  is computable for all subsets.

At this  $g_o$ , let us define the following recursive formula:

$X_t = \text{mean}_{g_o}(X_t)$  Note that there are  $s(t)$  number of partial averages in this process.

Let us repeat the calculations recursively until  $s(t) = 1$ . At this stage  $X_t = \mu$  with some probability.

According to PREM's 2<sup>nd</sup> theorem, when  $s(t) = 1$ ,  $X_t = \mu$  with probability 1 provided that there is no data loss and that  $g_o$  is a constant within each computational step.

Clearly, the size of the subset  $g_o$  is the same within each computational step but it can vary from one step to another as illustrated below.

Computation steps 1, ..., t, the following recursive formula is applied.

$$X_j(t) = \frac{\text{Sum}_{k_t}(X_i(t))}{k_t}$$

for example  $t = 1, g_o = k_1$

$$X_j(1) = \frac{\text{Sum}_{k_1}(X_i(1))}{k_1}$$

$t = 2, g_o = k_2$

$$X_j(2) = \frac{\text{Sum}_{k_2}(X_i(2))}{k_2}$$

....

In each computation step  $t$ , partial averages  $X(j)$  are computed for  $i = 1, \dots, s(t)$ , where  $s(t)$  is an integer value of  $(N/k_t)$ . In the final computational step, notice that  $g_o = K_t$  and  $s(t) = 1$ , and that the procedure returns the mean.

The integer value of  $(N/k_t)$  may not necessarily be computable, and it simply denotes the total number of possible  $k$  subsets in  $N$  un-computable number of data points.

However, if the number of observations in the last subset is less than  $g_o$ , the procedure leaves the maximum number of data loss  $(g_o - l)$  in each recursive computation. We simply define this data loss as ‘Information Loss’ in the subsequent discussion. In practice it is possible to select varying sizes of  $g_o$  for each computation step so that information loss is minimum. Alternatively, one can proceed the computation with randomly selected data points for the last subset in each repetitive computation of partial averages in such a way that last subset being complete in each computational step.

### 2.1 Illustrative Example 1

A practical example with a small dataset is given here to illustrate the above methods and the some possible data losses are also included in this illustration (Table 1).

Table 1: An Illustrative Example on PREM’s Estimation of Mean

---

(1.a) Two Points	(1.b) 3,2 Points Averaging
------------------	----------------------------

Observations	$X_i$	Averaging				1	64	$X(1)$	$X(2)$	$X(3)$
		$X(1)$	$X(2)$	$X(3)$	$X(4)$					
1	64	46				2	29			
2	29					3	95	63		
3	95	78	62			4	60			
4	60					5	58			
5	58	66				6	74	64		
6	74					7	29			
7	29	21	44	53		8	13			
8	13					9	42	28	52	
9	42	27				10	13			
10	13					11	6			
11	6	8	17			12	9	9		
12	9					13	58			
13	58	58				14	59			
14	59					15	75	64		
15	75	54	56	37	45	16	33			
16	33					17	99			
17	99	85				18	70	67	47	49
18	70					19	88			
19	88	55	70 <-loser			20	22	<-loser		
20	22									
<i>Average(X1:X20)</i>		50	50	50	45	45	50	49	49	49
<i>T</i>		0	1	2	3	4	0	1	2	3
<i>S(t)</i>		20	10	5	2	1	20	6	2	1

In the above example, small numbers of data points are lost in the process of averaging of equal size subset of data due to the effect of incomplete numbers in the last dataset. This effect causing loss of data points in the Parameter Repeated Estimation is termed as ‘Corner Effect’ and it is further described in the subsequent section.

### 3. RECOVERY OF CORNER EFFECTS

In the above illustrative example, we observed that some data points were lost due to incomplete data in the last subset of data in each recursive computational step due to corner effects. However, in the case of small datasets, the corner effect can be rectified, for example in step 1 in the above example as follows:

Notice that the process of 3 point averaging it left out 88 and 22 thus leading to an estimated value of mean, 49. Therefore, the actual mean  
 $= [49*6 + (88+22)] / (3*6+2) = 50$

The above simple computation is impossible to perform in the case of massive dataset: For instance, suppose that the true mean at  $k$  averaging process of  $s$  subsets leaves  $k-1$  last data points. The following lemma applies for a large dataset.

**Lemma 1**

For  $n \ll N$ , the true mean

$$\mu = \frac{\bar{X}ks + \left( \sum_{i=1}^{k-1} X_i \right) (k-1)}{ks + (k-1)}$$

is un-computable.

*Proof:*

Suppose  $N$  is un-computable and  $n, n \ll N$ , is computable. Let us consider a  $k$  averaging process where  $k < n$ .

If  $s.k < n$ , then  $N-sk > n$ . But the fact that  $N-sk = k-1$  implies that  $k-1 > n$  and  $k > n+1$ , which is contradictory.

Thus, the fact that  $s.k > n$  implies that the above true mean is un-computable.

This means that for an extremely large dataset the above rectification is not valid.

Consequently, in massive data analysis, the corner effect exists and it cannot be simply corrected.

**3.1 Can Corner Effects be Abolished?**

According to Theorem 2, there is no corner effect if the size of subset of data points is so selected that all data points are accommodated in each computational step i.e., there is no data loss in computation. For instance in the above example of 2 point averaging, if we

select the subset of 5 data points in step 3, the procedure returns the actual mean. Therefore, it could be argued that  $k$  must be chosen such that  $N/k = j$ , where  $j$  is an integer. But it is impossible to find such a  $k$  in practice with massive dataset because  $N/k$  is un-computable as  $N$  being un-computable.

Fortunately, the following argument shows that corner effects are negligible for extremely large datasets because it is possible to minimize these effects with proper selection of  $k$  in each computation.

### 3.2 Optimum Corner Effect for Extremely Large Dataset

Let  $X_p$  be data remaining in each computational step of the PREM mean estimated from  $k(t)$  averaging of  $t$  number of recursive computations. If  $X_p$ 's are known, the total number of data loss in this procedure is the sum of  $X_p$ . Yet, unknown  $X_p$  is estimable under certain conditions, and it is interesting to explore extreme values and optimality under these conditions.

Suppose  $k(t) = k$ , and  $k$  is constant in each computation step, leading to the population mean in  $t$  steps. In general, the total number of data loss is theoretically equal to  $N - k^t$ . In the worse case, it is possible to lose  $k-1$  data points in each step. In this case, the maximum information loss is  $(k-1) + k(k-1) + k.k(k-1) + \dots = k^t - 1$ , which could be intolerable in practice.

However, if one selects  $k$  to be moderately large, the procedure reaches to a certain stage at which number of terms could be countable and computable. Consequently, at this stage one may select  $k$  as a factor of the number of terms available in that computational step so that it guarantees that there will be no information loss, say 'Correction Practice'. Clearly beyond that step, same 'correction practice' could be done as the number of data points becomes countable and computable. If the 'correction practice' started at step  $t$ , then the maximum information loss is  $k^t - 1$ . One can guarantee a minimum loss by selecting proper size of  $k$  so that  $k^s \gg k^t$ .

Also, the following Remark 2 applies when either deleting the incomplete dataset or completing the last incomplete dataset using randomly selected data points from the remaining data.

**Remarks 2**

In each computational step, if the last dataset is incomplete, one has to compromise between either the deletion of remaining incomplete dataset or the inclusion of randomly selected data points from previous datasets to complete the last partial average. The following Table shows some extreme cases of information loss due to the deletion of observations or inclusion of randomly selected data points in the last data set in a hypothetical example of  $k(t) = k$ .

**Table 2: Possible Information Loss (L) Due to Deletion or Random Inclusion of Data in the Last Subset.**

	Deletion	Inclusion
Worse Case	$K^t-1 \geq L \gg 1$	$1 \leq L \ll K^t-1$
The best Case	$1 \leq L \ll K^t-1$	$K^t-1 \geq L \gg 1$

Similarly, it is possible to establish upper bounds of information loss and optimal information loss under various corrective actions in general, and it is left open in this study.

While the upper and lower bounds of information loss under each criterion of deletion of remaining data in the last dataset or inclusion of randomly selected data from other used data points to complete the last dataset are of theoretical interest, the following example further illustrates value of corrective action for optimum data loss in practice.

### **3.3 Illustrative Example 2**

Suppose the computer can calculate the average of only 10,000 data point at a time.

For instance if the  $k = 10,000$  and  $s=5$ , the procedure can handle,

10,000,0000,0000,0000,0000,0000,0000,0000,0000,0000,0000,0000,0000,0000,0000,00

0 data point. The resulting computational steps together with corner effects are

summarized in the Exhibit 1 below.

**Exhibit 1: Data Loss and Correction Practice in a Simulated  $10^{20}$  Dataset**

Number of means in each computation steps						Remarks	
Total number	Step 1	Step 2	Step 3	Step 4	Step 5	Steps	
1E+20	1E+16	1E+12	1E+08	10000	1	s(t)	
<b>Maximum Data loss</b>	<b>9999</b>	<b>99990000</b>	1E+12			← Number leaving	
			Fix the problem by proper choice of k, random choice of partial averages to complete the last subset of data or delete the remaining subset of data leading to a loss of 1E+12 data points				Remark: Fix it at Step 3
If we fix the problem at 3 <sup>rd</sup> stage total data loss is only 99999999 That is information loss with respect to the total data points is very close to zero $9.9 \times 10^{-13}$ Total number of data point used for computation of mean is $1E+20-99990000=1E+20$							

In the computational stage 3, if the last group could not be computed due to lack of just one mean value, we may randomly select a point from the remaining computed means to proceed computation using  $1E+20$  data points. The effects of random allocation or deletion of data points are discussed in subsequent Remark 2.

The above example illustrates the value of the PREM’s estimation for mean because the procedure could use with such a huge dataset, in place of only randomly selected 10,000 points. The subsequent section illustrates practical application of PREM’s Estimation in a variety of situations such as estimation of population variance and regression parameters.

## 4. PREM ESTIMATION OF VARIANCE

If several millions of data points are readily available, can we use all of them to estimate the population variance? It is easy to show that PREM's Estimation provides an answer to this problem as follows.

$$\text{Population Variance} = \text{PREM}(\overline{X^2}) - (\text{PREM}(\overline{X}))^2$$

$$\text{where } \overline{X^2} = \frac{\sum_{i=1}^N X_i^2}{N}$$

*Proof:*

$$\text{Let } Z_i = X_i^2,$$

$$\bar{Z} = \frac{\sum_{i=1}^N Z_i}{N} = \frac{\sum_{i=1}^N X_i^2}{N}$$

$$\text{and } \bar{Z} = \text{PREM}(\bar{Z}) = \text{PREM}(\overline{X^2})$$

$$\begin{aligned} \sigma^2 &= \frac{\sum_{i=1}^N X_i^2}{N} - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N^2} \\ &= \frac{\sum_{i=1}^N X_i^2}{N} - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N^2} \end{aligned}$$

$$= \frac{\sum_{i=1}^N X_i^2}{N} - (\overline{X})^2$$

Completes the proof.

## 5. PREM ESTIMATION OF SIMPLE LINEAR REGRESSION COEFFICIENTS

Let us consider a simple linear regression:

$$Y = \beta_o + \beta_1 X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

The model parameters of  $\beta_o$  and  $\beta_i$  in the above model are available as follows

$$\beta_1 = \frac{PREM(\bar{X}\bar{Y}) - PREM(\bar{X}).PREM(\bar{Y})}{PREM(\bar{X}^2) - (PREM(\bar{X}))^2}$$

and

$$\beta_o = PREM(\bar{Y}) - \beta_1 PREM(\bar{X})$$

Similarly, it is a simple exercise to show that correlation coefficient and other related estimates are readily available in terms of PREM's estimation.

## 6. SOME OPEN PROBLEMS AND CONCLUDING REMARK

While there are enormous possibilities for practical applications of PREM's Methods, we leave a number of open problems for future studies. For example, it is possible to find optimum  $k$  and  $s$  for PREM's estimation to guarantee the minimum amount of data loss in computation, which we would leave as an open problem in this study. Furthermore, we simply consider that any impact due to resample points on true mean could be accounted for unbiased estimation of mean because the points are randomly selected. However, any further discussions of this problem are left open in our study. Some asymptotic properties of PREM's method with a possibility of random data filling (and/or deleting) at different stages of computation are also left open. If the data are very badly corrupted and erroneous systematically, it is not unusual that the method may not lead to

a valid estimation of parameters in particular in the regression estimates. The problem of modelling large datasets with corrupted data is a part of our on going research program and would be addressed in a separate study (Pemajyantha, V, and Rajasekera, J., 2003).

Despite those open problems, this method provides a unique answer to handle extremely large datasets breaking a new ground for young statisticians and all.

Since the method presented here is not approached in the past, I could not find any related work for comparisons. Regarding present approaches to data mining, the following websites, statistical packages and publications are suggestive references.

### **References**

Magaputer package and tutorials at  
<http://www.megaputer.com/products/pa/tutorials>

Data Mining and Knowledge Discovery, An International Journal, Kluwer Academic Publishers at  
<http://www.digimine.com/usama/datamine/jdmkd-cfp2.htm>

A Bibliography of Temporal, Spacial and Specio-Temporal Data Mining Research at <http://www.cis.unisa.edu.au/~cisjfr/STDMPapers>

Pemajyantha, V. (2003), Modelling with Massive Datasets, Bulletin of Modelling and Simulation Research, Vol 2, Issue 2, 2003, ISSN 1446 2524

Pemajyantha, V. and Rajasekera, J. (2003), Filtering Corrupted Data at Influencing Points with a Statistical Catalyst Agent, Hawaii International Statistical Conference, June 5-9 Honolulu, Hawaii, USA