A Mathematical model for estimation of Rural House Hold data through different states in India

A.V.N.Krishna
Professor, Computer Science Dept., Indur Institute of Eng. & Tech.,
Siddipet, Medak dist., Andhra Pradesh, India.
Mail: hariavn@yahoo.com .Cell no. 9849520995

Abstract.

In this work a model is going to be developed which helps in measuring household data distributed over a wide area. This model considers the assumption that the house holds data fallows an ordered sequence. The rural house hold data at some states is considered in some ordered sequence from the Census measures. By properly guessing an empirical value, data is derived from the developed model. The derived data from the mathematical model is mapped with the known house hold data from the Census measures. The process is repeated till a sufficient mapping is obtained between the two, which gives a strong empirical value. Known the empirical value, it can be used to generate future rural household data which helps in proper planning of expenditure estimations for rural development in India.

Keywords: *Tridiogonal matrix algorithm, Cubic spline interpolation, Temporal and spatial data, example and Rural Household data.*

1. Introduction

Temporal data mining is an important extension of data mining and it can be defined as the non trivial extraction of implicit, potentially useful and previously unrecorded information with an implicit or explicit temporal content from large quantities of data. It has the capability to infer casual and temporal proximity relationships and this is something non-temporal data mining cannot do. It may be noted that data mining from temporal data is not temporal data mining, if the temporal component is either ignored or treated as a simple numerical attribute. Also note that temporal rules cannot be mined from a database which is free of temporal components by traditional data mining techniques [5],[6],[7].

Partial differential equations to model multiscale phenomena are ubiquitous in industrial applications and their numerical solution is an outstanding challenge within the field of scientific computing. On one hand, a brute force approach to attempt to resolve all the spatial and temporal scales of a typical multi-physics problem requires huge computing resources. On the other hand, simply ignoring those computationally expansive small scales can lead to very inaccurate predictions. The approach is to process the mathematical model at the level of the equations, before discretization, either removing non-essential small scales when possible, or exploiting special features of the small scales such as self-similarity or scale separation to formulate more tractable computational problems[4]. The present work is concerned with a theoretical study related to estimation of empirical formula, which helps in estimation of household data analysis. When more than one independent variable occurs in a differential equation, the equation is said to be partial differential equation. This model generates data over a span of time and space, a partial differential equation is used.

The house holds data for different states and different times will be generally obtained from census data. The census data will be taken for every 10 yrs. Thus by going through census data, house hold data analysis for different states like Utter Pradesh, Bihar, Andhra Pradesh can be obtained for years like 1991 & 2001, [9],[10],[11],[12].In this work a model is outlined which generates data and helps in simulating obtained household data from census with generated data. By suitably mapping this data, an empirical value can be developed which helps in estimating data for any years between 1991 - 2001 and also for future estimates of data.

2. Description

Types of temporal data,

1.Static: Each data item is considered free from any temporal reference and the inferences that can be derived from this data are also free of any temporal aspects [1],[2],[3].

2.Sequence. In this category of data, though there may not be any explicit reference to time, there exists a sort of qualitative temporal relationship between data items. The market basket transaction is a good example of this category. The entry sequence of transactions automatically incorporates a sort of temporality. If a transaction appears in the data base before another transaction, it implies that the former transaction occurred before the latter. While most collections are often limited to the sequence relationships before and after, this category also includes the richer relationships, such as during, meet, overlap etc. Thus there exists a sort of qualitative temporal relationship between data items.

3.Time stamped. Here we can not only say that a transaction occurred before another but also the exact temporal distance between the data elements. Also with the events being uniformly spaced on the time scale.

4.Fully Temporal: In this category, the validity of the data elements is time dependent. The inferences are necessarily temporal in such cases.

3. Some existing techniques used for house hold data analysis in India.

The house hold data analysis is being developed by 'National Sample Survey NSS', under the Union Ministry of Statistics and Programme Implementation, India. One more center that is working in this area is 'The centre for environment and development', an NGO based organization. Some more tools and packages that are used in this area of data analysis are SAS Statistical tools, C-DAC 'Param-10000' and so on. In countries like U.S.A., the house hold data analysis being done using numerical methods like spatial regression analysis.

4. Some Numerical Methods for Data Mining.

The interpolation techniques like Lagrange, Newton's forward and backward methods fit a single polynomial through all the tabulated points. If the set of points is that of a polynomial, this method works well. One more method of fitting a graph to a set of points is by using piece wise linear segments where the slope of the segments depends on the values of the function at the two closest points.

Even though this is a simple idea, it is not very good as the different line segments have different slopes and the resultant graph does not look smooth. This problem could be solved by drawing a quadratic through (x_i,y_i) &(x_{I+1},y_{I+1}) such that its slope at (x_{I+1},y_{I+1}) matches with that of another quadratic sleeve through (x_{I+1},y_{I+1}) & $(x_{I+2},y_{I+2.})$. A better method is if a cubic curve is drawn through(x_i,y_i) & (x_{I+1},y_{I+1}) and another cubic through (x_{I+1},y_{I+1}) & $(x_{I+2},y_{I+2.})$ such that the slope and the curvature of the two cubes match at the point (x_{I+1},y_{I+1}). Because of the better approximation of cubical curve between the grid points, a cubic spline interpolation technique is used for generation of house hold data analysis over different states. It was assumed that the tabulated values have no errors.

If the deviations are more with the calculated values , the problem of fitting a function f(x) to the tabulated values is done by least square fit, which minimizes the sum of squares of deviations.[14]

5. Numerical Data Analysis

The fallowing are the steps to generate a numerical method for data analysis.

5.1. Discritization Methods.

The numerical solution of data flow and other related process can begin when the laws governing these processes have been expressed in mathematical form, generally in terms of

differential equations. The individual differential equations that we shall encounter express a certain conservation principle. Each equation employs a certain quantity as its dependent variable and implies that there must be a balance among various factors that influence the variable.

The numerical solution of a differential equation consists of a set of numbers from which the distribution of the dependent variable can be constructed. In this sense a numerical method is akin to a laboratory experiment in which a set of experimental readings enable us to establish the distribution of the measured quantity in the domain under investigation.[13]

Let us suppose that we decide to represent the variation of ϕ by a polynomial in x

$$\phi = a_0 + a_1 x + a_2 x^2 + \ldots\ldots\ldots\ldots\ldots\ldots a_n x^n$$

and employ a numerical method to find the finite number of coefficients a1 , a2…….an. This will enable us to evaluate ϕ , at any location x by substituting the value of x and the values of a's in the above equation.

Thus a numerical method treats as its basic unknowns the values of the dependent variable at a finite number of location called the grid points in the calculation domain. This method includes the task of providing a set of algebraic equations for these unknowns and of prescribing an algorithm for solving the equations.

A discretisation equation is an algebraic equation connecting the values of ϕ for a group of grid points. Such an equation is derived from the differential equation governing ϕ and thus expresses the same physical information as the differential information. That is only a few grid points participate in the given differential equation is a consequence of the piecewise nature of the profile chosen. The value of ϕ at a grid point there by influence the distribution of ϕ only in its immediate neighborhood. As the number of grid points becomes large, the solutions of discritization equations are expected to approach the exact solution of the corresponding differential equations.

5.2. Control Volume Formulation.

The basic idea of the control volume formulation is easy to understand and lends itself to direct physical interpretation. The calculated domain is divided into a number of non overlapping control volumes such that there is one control volume surrounding each grid point. The differential equation is integrated over each control volume piecewise profiles expressing the variation a ϕ between grid points are used to evaluate the required integrals.

The most attractive feature of the control volume formulation is that the resulting solution would imply that the integral conservation of quantities such as mass, momentum and energy is exactly satisfied over any group of control volumes and ofcourse over the whole calculation domain. This characteristic exists for any number of grid points, not just in a limiting sense when the number of grid points becomes large. Thus even the course grid solution exhibits exact integral balances.

5.3. Steady One Dimensional data flow.

Steady state one-dimensional equation is given by $\partial/\partial x(k.\partial T/\partial x) + s = 0.0$ where k & s are constants. To derive the discretisation equation we shall employ the grid point cluster. We focus attention on grid point P, which has grid points E, W as neighbors. For one dimensional problem under consideration we shall assume a unit thickness in y and z directions. Thus the volume of control volume is $\Delta x*1*1$.[8]

Thus if we integrate the above equation over the control volume, we get

$$(K.\ \partial T/\partial X)e - (K.\partial T/\partial X)w + \int S.\partial X = 0.0$$

If we evaluate the derivatives $\delta T/\delta X$ in the above equation from piece wise linear profile , the resulting equation will be $Ke(Te - Tp)/(\delta X)e - Kw(Tp - Tw)/(\delta X)w + S\ \Delta X=0.0$ where S is average value of s over control volume.

This leads to discretisation equation

$$ApTp = aeTe + awTw + b$$

Where ae= Ke/δXe

aw = Kw/δXw

ap= ae+aw-spΔX

b=seΔX .

5.4. Grid Spacing

For the grid points it is not necessary that the distances (δX)e and (δX)w be equal. Indeed, the use of non uniform grid spacing is often desirable, for it enables us to deploy more efficiently. Infact we shall obtain an accurate solution only when the grid is sufficiently fine. But there is no need to employ a fine grid in regions where the dependent variable T changes slowly with X. On the other hand, a fine grid is required where the T_X variation is steep. The number of grid points needed for the given accuracy and the way they should be distributed in the calculation domain are the matters that depend on the nature of problem to be solved.

5.5. Solution Of Linear Algebraic Equations

The solution of the discretisation equations for the one-dimensional situation can be obtained by the standard Gaussian elimination method. Because of the particularly simple form of equations, the elimination process leads to a delightfully convenient algorithm.

For convenience in presenting the algorithm, it is necessary to use somewhat different nomenclature. Suppose the grid points are numbered 1,2,3…ni where 1 and ni denoting boundary points.

The discretisation equation can be written as

Ai Ti + BiTi+1 +CiTi-1 = Di

For I = 1,2,3…….ni. Thus the data value T is related to neighboring data values Ti+1 and Ti-1. For the given problem

C1=0 and Bn=0;

These conditions imply that T1 is known in terms of T2. The equation for I=2, is a relation between T1, T2 & T3. But since T1 can be expressed in terms of T2 , this relation reduces to a relation between T2 and T3. This process of substitution can be continued until Tn-1 can be formally expressed as Tn. But since Tn is known we can obtain Tn-1.This enables us to begin back substitution process in which Tn-2,Tn-3…….T3,T2 can be obtained.

For this tridiogonal system , it is easy to modify the Gaussian elimination procedures to take advantage of zeros in the matrix of coefficients.

Referring to the tridiogonal matrix of coefficients above, the system is put into a upper triangular form by computing new Ai.

Ai = Ai – (Ci-1 /Ai)* Bi I = 2,3……ni.

Di= Di – (Ci-1 /Ai) * Di

Then computing the unknowns from back substitution

Tn = Dn / An.

Then Tn = Dk – Ak * Tk+1 / Ak, k= ni-1, ni-2…3,2,1.

6. Mathematical modeling of the problem

The approach to time series analysis was the establishment of a mathematical model describing the observed system. Depending on the appropriation of the problem a linear or non-linear model will be developed. This model can be useful to analyze census data, land use data and satellite meteorological data[8].

6.1. Linear data flow Problem.

In the given work we consider a set of states, which are considered in some sequential order ie in increasing order of house hold data analysis. For example in the given problem the states like Gujarat, Karnataka, Maharastra and Utter Pradesh are considered with Gujarat as the state with least house holds and Utter Pradesh as a state with maximum house holds. All the remaining states are

arranged between these states in sequential order[13]. Considering M states, with each state being represented by a grid point, the population in each state is initialized to 300, i.e.

When t=0, T (I) =Y (I) =300. where I=1,2,……..M.

For each control volume, data leaving the control volume – data entering the control volume data gain per unit time in the control volume. Dividing the problem area into M number of points. The house hold data of the first and Mth grid points are considered to be known and constant.

For the grid points 2, M-1, the coefficients can be represented by considering the conservation equation,

$$k/\partial x\,(T_{I+1}^{\,n+1} - T_I^{\,n+1}) + k/\partial x\,(T_I^{\,n+1} - T_{I-1}^{\,n+1}) = (\rho cp\,\delta x)\,/\,\delta t\,(T_I^{\,n+1} - T_I^{\,n}),$$

where $T_I^{\,n}$ represents data value for the considered grid point for the preceding delt, $T_{I+1}^{\,n+1}$ & $T_{I-1}^{\,n+1}$ represents data values for the preceding and succeeding grid points for the current delt. Dividing all terms by $\rho cp\delta x/\delta t$, the coefficients are obtained for each state (grid point) in terms of $\alpha = k/\rho cp$ and A(I) refers to data value of the corresponding grid point, C(I) and B(I) refers to data values of preceding and succeeding grid points for the current delt, D(I) refers to data value of the considered grid point in the preceding delt.

For the grid point 2,

$T_2^{\,n+1}(1+2\alpha\delta t/(\delta x^2) + T_3^{\,n+1}(-\alpha\delta t/(\delta x^2) = T_2^{\,n} + T_1^{n+1}(\alpha\delta t/\delta x^2)$

$A(2) = 1+2\,\alpha\delta t/\delta x^2$; $B(2) = -\alpha\delta t/\delta x^2$; $D(2) = -\alpha\delta t/\delta x^2 * T_1^{n+1} + T_2^{\,n}$

For the grid points 3 to M-2

$A(I) = 1+2\,\alpha\delta t/(\delta x^2$; $B(I) = -(\alpha\delta t)/(\delta x^2)$; $C(I) = -(\alpha\delta t)/(\delta x^2)$; $D(I) = Ti^n$

For the M-1 th point

$A(I) = 1+2\,\alpha\delta t/\delta x^2$; $C(I) = -(\alpha\delta t)/(\delta x^{2)}$; $D(I) = T_I\text{-}1^n + (-\alpha\delta t)/(\delta x^{2)(}\,T_M^{n)}$

6.2. Procedure for generating data values from coefficients by tridiogonal method.

Using the coefficients of grid points, and by using the tridiogonal matrix algorithm, the data distribution is calculated. The grid points are numbered 1,2,3,…….M. with points 1 and M denoting extreme states.

The discretisation equation can be written as

$$Ai\,Ti + BiTi+1 + CiTi-1 = Di$$

For I = 1,2,3…M. Thus the data Ti is related to neighboring data values Ti+1 and Ti-1. For the given problem C1=0 and BM=0 as T1 & TM represent boundary states.

These conditions imply that T1 is known in terms of T2. The equation for I=2, is a relation between T1, T2 & T3. But since T1 can be expressed in terms of T2 , this relation reduces to a relation between T2 and T3. This process of substitution can be continued until TM-1 can be formally expressed as TM. But since TM is known we can obtain TM-1.This enables us to begin back substitution process in which TM-2, TM-3…T3, T2 can be obtained. This process is continued until further iterations cease to produce any significant change in the values of T's. Finally the data distribution is obtained for all grid points for different times by considering a suitable empirical constant empirical constant α .

7. Results.

Distributed House holds data at different grid points is considered from census measures. The distance between successive states is 900 Km. The time interval considered for the given problem i.e. Del T is 1 year.

Thus δX= 900 Km

δt = 1 year.

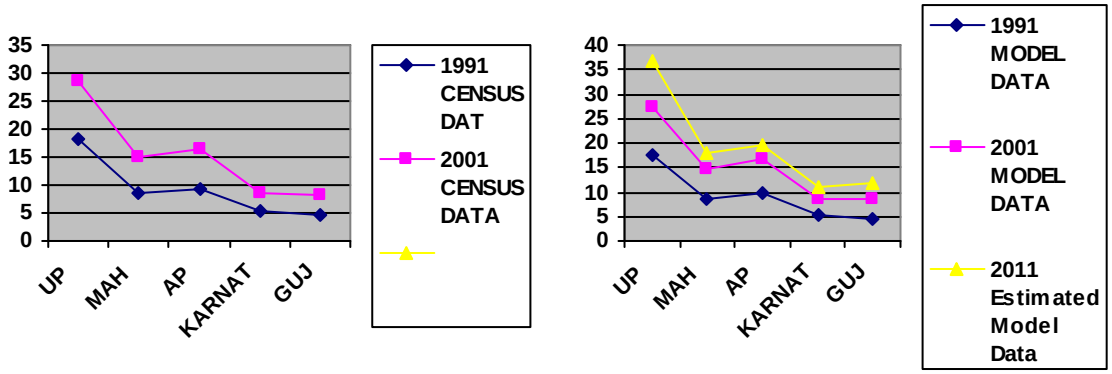
Total time (n) = 10 years.

By suitably assuming empirical constant, house holds data could be generated at different points and at different times from the generated model. The sampled data from census is matched with generated data from the model, which provides for better estimation of empirical constantα.

This procedure is repeated till the empirical value reaches a constant value. Known the empirical constant, the traffic analysis could be made at different grid points at different times.

Empirical value obtained = $4.5 * 10^{2}$.

	Rural house hold data from census measures in 1991 $*10^{6}$	Rural house hold data from census measures in 2001 $*10^{6}$	Derived data for delt=1 Yr (1991) $*10^{6}$	Derived data for time =10yrs (2001) $*10^{6}$.	Expected rural house hold data analysis in 2011. $*10^{6}$
Maharastra	8.6	14.92	8.54	14.85	17.83
Utter Pradesh	18.2	28.6	17.6	27.6	37.62
Gujarat	4.4	8.72	4.6	8.52	11.82
Andhra Pradesh	9.4	16.4	9.96	16.69	19.72
Karnataka	5.48	8.18	5.5	8.42	11.05



8. Conclusion & Future work

In the present problem, the rural house hold data for different states has been taken from census measures. A model has been developed which generates an empirical constant. The empirical constant is used to estimate rural house hold data analysis for 2011 which can be used for expenditure estimations for future developments in rural India.

Future Work

Correlating behavior profiles across multiple locations. By using multiple data from different states at different times, more accurate results can be obtained and the effect of errors can be much lowered. In the present case, only past and present time steps had been used. To get more accurate estimates of house hold data analysis future time steps may also be used. Thus future work on data analysis can be done by using multiple data and present, past and future time steps.

References

[1]. [Patterson, 2001] Dan.W.Patterson *: Introduction to Artificial Intelligence & Expert Systems*, pp 345-385, Prentice-Hall of India Private limited –2001.

[2]. [Pujari, 2002] A.K.Pujari: *Data Mining Techniques*, pp 251-281, Prentice-Hall of India Private limited-2002.

[3]. [Chen and Petronunias, 98] Chen X., and Petronunias I.*Frame work for temporal data mining*. In proceedings of DEXA'98, LNCS-1460, Springer-Verlag, 1998.

[4]. [Abramowitz and Stegun, 64] Abramowitz, M. and Stegun, I.A., *Hand book of mathematical functions with formulas graphs and mathematical tables*, National bureau of standard, vol.55,1964.

[5]. [Al Naemi, 94] Al-Naemi S. *A theoretical frame work for temporal knowledge discover* In proceedings of the international work shop on spatio temporal data bases, pp 23-33, 1994.

[6]. [Davis and Brockwell, 96] Brockwell P.J. and Davis .R *Introduction to time series and forecasting*, Springer- Verlag, 1996.

[7]. [Das et al., 96]Das G., Gunupolos D., Mannila. *Finding similar time series*, Manuscript 1996

[8]. [Hal,] Hal Caswell. *Matrix population models: Construction, analysis & interpretation*, Sinauer Associates, Inc publishers, Sunderland, Mascachusetts.

[9]. www.m.nic.in/health/DATIA.pdf

[10]. www.cyberjournalist.org.in/census

[11]. www.wbcensus.gov.in/mas/Tellmemore.htm

[12]. www.wbcensus.gov.in/publications.htm

[13].[Suhas 1991**]** Suhas V. Patenkar *Numerical Heat Transfer and Fluid Flow* 11-75(1991).

[14]. [Raja 1990] Raja Ramanna *Numerical methods* 78-85(1990).

Article received: 2007-06-13