A Spatial Regression Analysis Model for Temporal Data Mining in Estimation of 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.
E-mail: hariavn@yahoo.com Cell no. 9849520995

Abstract

Many applications maintain temporal & spatial features in their databases. These features cannot be treated as any other attributes and need special attention. Temporal data mining has the capability to infer casual and temporal proximity relationships among different components of data.

In this work a model is going to be developed which helps in measuring household data distributed over a wide area. The model considers the assumption that the house holds data follows an ordered sequence. The house hold data at some states is considered from the census data. A grid point identifies each state. Each grid point is identified by a set of coefficients. These coefficients are represented in terms of p1, p2, p3 & p4. Thus known house hold data from the census, a set of simultaneous equations will be developed by multivariate regression model. By solving these simultaneous equations, coefficients of the simultaneous equations will be calculated. These coefficients will be used to generate household data for any years between known data and also for any future house hold data analysis.

Keywords*. Spatial regression analysis , temporal and spatial data, example & 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].

Regression analysis models the relationship between one or more response variables (also called dependent variables, explained variables, predicted variables, or regressands) (usually named *Y*), and the predictors (also called independent variables, explanatory variables, control variables, or regressors,) usually named $X_1,...,X_p$). If there is more than one response variable, it is called as *multivariate regression*.

In general the house holds data for different states and different times will be taken from census data. The census data will be taken for every 10 yrs. 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 going to be discussed which generates data and helps in simulating obtained household data with generated data. By suitably mapping this data, a set of coefficient values can be developed which helps in estimating data for any year between 1991 & 2001, and also for future estimations 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].

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. Different 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 can also be used.

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]

4. Numerical Data Analysis.

4.1 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

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

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.

4.2 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.

4.3 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. 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. The example that is going to be considered in this model is house hold data. It is based on general assumption that the house hold fallows an ordered sequence. And also the data increase depends on the present data. The measured data is mapped with the spatial regression model to generate coefficients p0,p1 ,p2 and p3.

5.1 Description of the problem

The area considered in the problem is divided into set of grid points. In the present problem, a set of 5 states will be considered as grid points. The house hold data is measured at identified points at different intervals of time from census data. Considering this data at different locations and times as input to the model, the coefficients p0,p1,p2 & p3 will be calculated Known the coefficients, house hold data at different locations and different times can be easily estimated.

Multivariate regression model : $Y = p0 + p1*X + p2*Z + p3*X*Z$

The function is a general expression involving one dependent variable (on the left of the equal sign), one or more independent variables, and one or more parameters whose values are to be estimated.

Y=Measured house hold data density at different grid points .

X= Different states being represented as Grid Points

Z=Time Span over different grid points.

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.

Mathematical model of the multivariate regression model can be represented as

$\Sigma y = p0*n + p1\Sigma x + p2\ \Sigma z + p3\ \Sigma xz$

$\Sigma xy = p0\ \Sigma x + p1\ \Sigma x^{2} + p2\ \Sigma xz + p3\ \Sigma x^{2}z$

$\Sigma zy = p0\ \Sigma z + p1\ \Sigma xz + p2\ \Sigma z^{2} + p3\ \Sigma xz$

$\Sigma xyz = p0\ \Sigma xz + p1\ \Sigma x^{2}z + \Sigma z^{2} + p3\ \Sigma x^{2}z^{2}$

Solving the equations will generate coefficients p0, p1, p2& p3.

Case Study : Total House hold data of some states in India

X being represented as states starting from 1-Gujarat, 2-Karnataka, 3-Maharastra, 4-A.P, 5-U.P, Y being represented as house hold data in Multiples of 1 Million,

Z represents years of Census Measured Data, 1-1991, 2-2001.

X	1	2	3	4	5	1	2	3	4	5
Y	07	08	13	15	22	12	13	20	25	34
Z	1	1	1	1	1	2	2	2	2	2

$\Sigma y = p0*n + p1\Sigma x + p2\ \Sigma z + p3\ \Sigma xz$

$\Sigma xy = p0\ \Sigma x + p1\ \Sigma x^{2} + p2\ \Sigma xz + p3\ \Sigma x^{2}z$

$\Sigma zy = p0\ \Sigma z + p1\ \Sigma xz + p2\ \Sigma z^{2} + p3\ \Sigma xz^{2}$

$\Sigma xyz = p0\ \Sigma xz + p1\ \Sigma x^{2}z + \Sigma z^{2} + p3\ \Sigma x^{2}z^{2}$

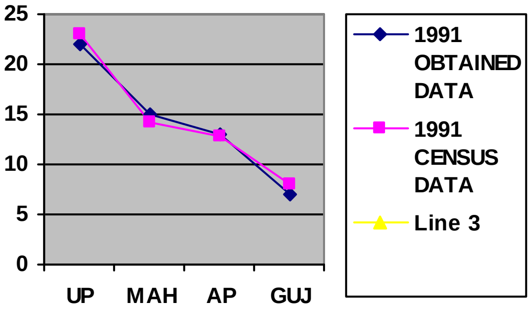
Solving the equations will generate coefficients p0=-7.0792, p1=-4.5056, p2=2.7074& p3=6.6435.

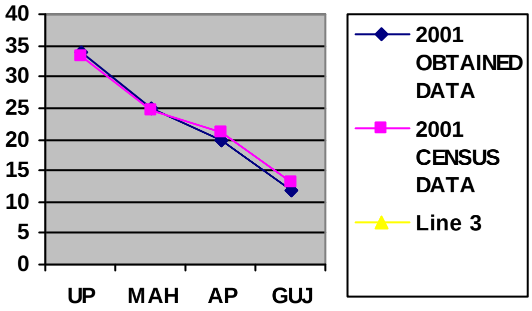
Using the coefficients to generate household data for future estimations like 2011.

Consider Z=2011, X=Different states, Y=Household data.

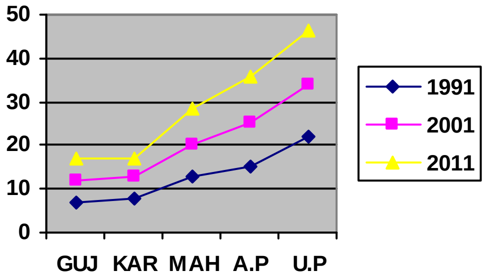
Y=16.83	17.2	28.6	36	46.4

Case Study : Total house hold data in some states in India.





Estimated house hold data through different states in India for 1991,2001 & 2011 from the developed Model.



6. Conclusion & Future work

In the present problem, the total and rural house hold data for different states has been taken from census measures. A model has been developed which generates a set of coefficients. These coefficients are used to estimate total & rural house hold data analysis for 2011, which can be used for expenditure estimations for future developments in house holds in India.

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: 2008-06-17