ISSN:0975-9646



# Crime Trend Analysis and Prediction Using Mahanolobis Distance and Dynamic Time Warping Technique

Anchal Rani[1] , Rajasree S.[2]

[1]*Student, M.Tech, CDAC, Noida, India*
[2]*Director(Associate Professor), M.Tech, CDAC, Noida, India*

***Abstract*-Recently, interest has been increasing in crime analytics in which time series clustering research has been playing an important role, particularly for finding useful similar trends in crime. Time series clustering technique has potential for large volume of data analysis. Multivariate time series data relating to various areas such as finance, health sector, environmental research, and crime is very useful to determine and apply data mining techniques. Analyzing multivariate time series data at different points of time helps reveal crime trends in which law enforcement agencies are interested. Police administration at state and districts level make use of the trend analysis to solve new crime cases and help them prevent future possibilities of similar kind of crime. This paper presents a new approach using dynamic time warping technique and Mahanolobis distance model to use records and statistics from to analyse crime trends and predict future crime.**

***Keywords*: Dynamic Time Warping, Mahanolobis, Euclidean, Minkowski , ANOVA**

## I. INTRODUCTION

Crime Analysis is the systematic study of crime which helps out law enforcement agencies in crime evaluation, crime and disorder reduction, crime prevention. Security agencies and administration are searching for new methods to efficiently find useful patterns from large number of rows of data. New and advanced techniques are required to extract information from the large amount of data. Clustering data can help us find similar trends of data in the existing. It can help us define and find security enhancements in different and similar cities. Here we have studied and implemented a multivariate time series algorithm based on Dynamic Time Warping technique and Mahanolobis distance model. Also we have implemented the same algorithm with different other distances calculation Euclidean and Mahanolobis and a comparative study based on the analysis we calculated the error in the prediction for a test city using the existent cities and applying the algorithm we find out what is the error mean in different methods and which gives the best result. This paper presents a comparative analysis of different distances with the algorithm.

## II. RELATED WORK

Various theories have been proposed for crime. [2]Environmental Criminology gives various theories on criminal activity and victimization and how factors of space influence offenders and victims and increase the probability of occurrences of crime.

- The “opportunity theory” : physical design can be used to prevent crime by reducing the opportunities.
- The “expanded opportunity theory: “certain physical attributes such as specific land uses, street layouts, environmental disrepair and deterioration, and physical features that block visibility and natural surveillance can encourage higher incidence of crime”.
- The “broken windows theory”: physical incivilities and social incivilities result in higher crime and fear of crime.
- The “rational choice”: criminal events are most likely to occur in areas where the activity space of offenders overlaps with the activity space of potential victims/targets.

Various ways have been devised to find out various crimes prone cities [13]:

***Hotspots***

The most common method of "forecasting" crime in police departments is simply to assume that the hot spots of yesterday are the hot spots of tomorrow [13].

***Repeat victimization***

The above research indicates that temporally aggregated hot spots may serve as accurate predictors of crime, but that relying on shorter previous time periods for predictive purposes is less effective. The exception to these research findings relates to "hot dots" rather than hot spots: that is, the repeat victim rather than the high-crime area [13].

***Univariate Methods***

There are a variety of univariate methods available to predict crime. These methods use previous values of one variable to predict its future value. They are attractive because of their straightforwardness: univariate methods require a minimum of data collection since they involve only one variable [13].

***Leading Indicators***

It refers to specific characteristics of areas or neighbouring areas (e.g., shots fired, calls for service, disorderly conduct offenses, etc.) for which their rise or fall in current and previous months can be used to predict future values of the dependent crime variable [13].

***Point Process Model***

The modelling is akin to neural networks and past data are used to predict future events. The output of the model is a

probability surface indicating likely areas of future crime [13].

***Artificial Neural Networks***

One of the earliest efforts to do predictive crime mapping was that of Olligschlaeger (1997), who employed a "feed-forward network with back propagation" to predict areas where future drug markets will emerge. Best known to laypeople as artificial intelligence, the type of neural network model employed by Olligschlaeger is capable of learning extremely complex space-time patterns (Olligschlaeger, 1997)[13].

While the above techniques determine the hotspots on the basis of the frequency of crime, we here are looking forward to analyse crime trend on the basis of which we are going to detect more crime prone city or district using time series clustering. We here find how much similar is a city to the existing known cities.

Here will like to concentrate on analyzing crime trends on the various statistics available on the time series. Analysing on the basis of time series helps us find out various factors determining and leading to crime what kind of factors leads to what kind of crime trends, what kind of crime is more frequent which kind of cities are similar in which kind of crime and this intern will help us to cross examine the existing cases and deploy security in more predicted crime prone area hence we will be able to device a give the common people a more secure environment to live in the cities.

## III. PROPOSED APPROACH

The proposed approach of crime trend analysis and prediction is described in the following lines:

### *A. Methodology*

We have a multivariate time series data from 1971 to 2006 i.e. 36 years. Then a weight matrix is calculated for the above data by multiplying with the weight assigned to the respective type of crime. Then, we compute a distance matrix from the weight matrix using the respective distance formula. DTW matrix is calculated from the distance matrix and a warping path is determined. Then we find the DTW value for the same. Finding DTW for each pair of cities under calculation, hierarchical clustering is applied to it. Following sections define the various techniques and formulae.

### *B. Euclidean Distance*

The Euclidean distance between points p and q is the length of the line segment connecting them ( ). In Cartesian coordinates, [15] if p = ($p1$, $p2$,..., $pn$) and q = ($q1$, $q2$,..., $qn$) are two points in Euclidean $n$-space, then the distance from p to q, or from q to p is given by:

$$\mathrm{d}(p,q) = d(p,q) = \sqrt{\sum_{i=1}^{n}(\mathrm{q_i} - \mathrm{p_i})^2}$$

### *C. Minkowski Distance*

The Minkowski distance is a metric on Euclidean space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance [16].

The Minkowski distance of order $p$ between two points $P=(x_1,x_2,\ldots..x_n)$ and $Q= (y_1,y_2,\ldots.y_n) \in R^n$

Is defined as:

$$\left(\sum_{i=0}^{\mathrm{n}}(\mathrm{x_i} - \mathrm{y_i})^{\mathrm{p}}\right)^{1/\mathrm{p}}$$

### *D. Mahanolobis Distance*

Mahanolobis distance is a descriptive statistic that provides a relative measure of data point distance from a point [18]. It is used to identify and gauge similarity of an unknown sample set to a known one. It takes into account correlations of data sets and is scale-invariant and has a multivariate effect size. Mahanolobis distance is calculated using following formula:

$$D(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2 / {S_i}^2}$$

Where,

$x_i$ = values of time series x

$y_i$= values of time series y

$S_i$=standard deviation

### *E. Dynamic Time Warping*

In time series analysis, dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed[1].

Given two time series, $X = x_1, x_2, \ldots, x_i, \ldots, x_n$ and $Y = y_1, y_2, \ldots, y_j, \ldots, y_m$, DTW aligns the two series so that their difference is minimized[1].

To this end, a distance matrix $D$ of order $n \times m$, where the $(i, j)$ element of the matrix $D$ contains the distance $d(x_i, y_j)$ between two points $x_i$ , and $y_j$. The Euclidean distance is normally used. A warping path, $W = w_1, w_2, \ldots, w_k, \ldots, w_K$, where $max(m, n) \leq K \leq m + n - 1$ that has the minimum distance between the two series is of interest[1].

It is a set of matrix elements that satisfies three constraints:

1) Boundary condition i.e , $w1$ = first element of distance matrix $D$ and $wK$= last element of distance matrix $D$ [1]. The boundary condition enforces that the first elements of $X$ and $Y$ as well as the last elements of $X$ and $Y$ are aligned to each other.
2) Continuity restricts the allowable steps to adjacent cells [1], and
3) Monotonicity forces the points in the warping path to be monotonically spaced in time [1]. It reflects the requirement of faithful timing: if an element in $X$ precedes a second one this should also hold for the corresponding elements in $Y$ , and vice versa.

The step size condition expresses a kind of continuity condition: no element in $X$ and $Y$ can be omitted and there are no replications in the alignment [1].

Mathematically,

$$DTW = min\frac{\sum_{k=1}^{K} w_k}{K}$$

Dynamic programming is generally being used to effectively find this path by evaluating the following recurrence, which defines the cumulative distance as the

sum of the distance of the current element and the minimum of the cumulative distances of the adjacent elements [1].

$$d_c(i,j) = d(x_i, y_i) + \min\{d_c(i-1, j-1), d_c(i-1, j), d_c(i, j-1)\}$$

The major advantage of the DTW is that two sequences need not to be of same length [1].


Fig. 1. Flowchart of the Approach

### *F. Procedure*

The method has been described below in steps:

1) Take multivariate time series data as input for 20 districts and calculate the weight matrix accordingly
2) Calculate distance matrix with Mahanolobis/Euclidean/Minkowski distance calculation.
3) Find warping path by calculating DTW matrix and DTW.
4) Apply hierarchical clustering on the pair of cities matrix having DTW for each pair.
5) Find the similar trends from the dendrogram generated after hierarchical clustering.
6) Applying prediction for a test city.
7) Calculate mean error and accuracy.

We calculated predicted value by checking the nearest city to the test city and finding its value. The value of the nearest city is taken as the predicted value.

The above method is been implemented on the data from national crime records bureau (NCRB) website. Statistics are available on murder, dacoity, riots , arson etc. In this research paper we have worked on only murder statistics.

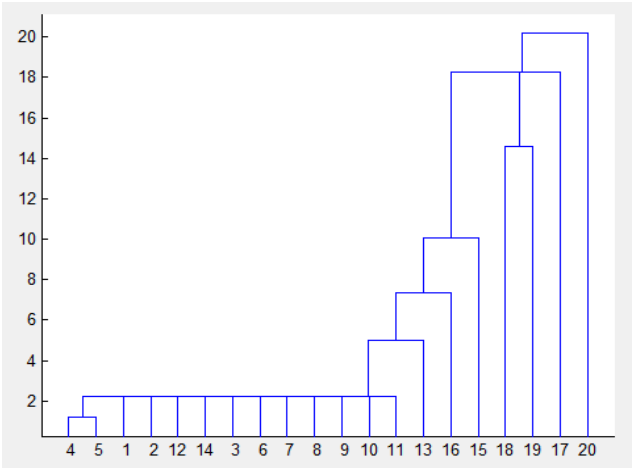


Fig. 2. Dendrogram of Crime Statistics for murder for 1971-2006 using DTW with Mahanolobis distance. X- axis shows the districts and y-axis depicts the relative distance between the districts.

Besides this we applied the above methodology with Euclidean distance and Minkowski distance on the same time series data.

## IV. RESULTS AND DISCUSSION

The algorithm has been applied to 20 districts of UP in three different ways. The Statistics have been taken from National Crime Records Bureau's website. Data is from 1971 to 2006. We have applied clustering on 20 districts initially. The 20 districts 1,2,…..20 taken in fig.2 are :Agra, Aligarh, Allahabad, Bareilly, Bijnor, Bulandshahar, Etawah, Fatehpur, Gorakhpur, Jaunpur, Jhansi, Mathura, Meerut, Moradabad, Muzaffarnagar, Saharanpur, Sultanpur, Unnao and Varanasi, respectively

Three approaches are using three different distances being calculated which are: Mahanolobis distance, Minkowski distance, Euclidean distance.

Results for the test cities taken are shown in the tables below.

Error in the predicted value and real value has been calculated as:

Variance = |Predicted value – Real value|

Mean Error = Total Variance /No. of test cities

| *CITY* | *REAL VALUE* | *PREDICTED VALUE* | *VARIANCE* |
|---|---|---|---|
| Ballia | 35 | 73 | 38 |
| Barabanki | 79 | 73 | 6 |
| Basti | 30 | 73 | 43 |
| Budaun | 169 | 77 | 92 |
| Deoria | 32 | 73 | 41 |
| Faizabad | 45 | 73 | 28 |
| Hardoi | 87 | 78 | 9 |
| Kheri | 105 | 73 | 32 |
| Mainpuri | 110 | 73 | 37 |
| Rampur | 57 | 73 | 16 |
| Sitapur | 124 | 96 | 28 |
| **TOTAL** | | | 370 |

Table 1: Prediction of crime occurrences values for year 2006 using Mahanolobis distance

| CITY | REAL VALUE | PREDICTED VALUE | VARIANCE |
|---|---|---|---|
| Ballia | 35 | 73 | 38 |
| Barabanki | 79 | 73 | 6 |
| Basti | 30 | 73 | 43 |
| Budaun | 169 | 73 | 96 |
| Deoria | 32 | 73 | 41 |
| Faizabad | 45 | 73 | 28 |
| Hardoi | 87 | 73 | 14 |
| Kheri | 105 | 73 | 32 |
| Mainpuri | 110 | 73 | 37 |
| Rampur | 57 | 73 | 16 |
| Sitapur | 124 | 73 | 51 |
| **TOTAL** | | | 402 |

Table 2: Prediction of crime occurrences values for year 2006 using Minkowski distance

| CITY | REAL VALUE | PREDICTED VALUE | VARIANCE |
|---|---|---|---|
| Ballia | 35 | 96 | 61 |
| Barabanki | 79 | 73 | 6 |
| Basti | 30 | 73 | 43 |
| Budaun | 169 | 123 | 46 |
| Deoria | 32 | 123 | 91 |
| Faizabad | 45 | 77 | 32 |
| Hardoi | 87 | 73 | 14 |
| Kheri | 105 | 96 | 9 |
| Mainpuri | 110 | 96 | 14 |
| Rampur | 57 | 73 | 16 |
| Sitapur | 124 | 96 | 28 |
| **TOTAL** | | | 351 |

Table 3: Prediction of crime occurrences values for year 2006 using Euclidean distance

We calculate the mean error for each distance in the algorithm:

Mean error for Mahanolobis is

Total Variance= 370

Mean Error=370/11=33.63

Mean error for Minkowski is

Total Variance=402

Mean Error = 402/11=36.54

Mean Error Euclidean:

Total Variance=351

Mean Error 351/11=31.90

Calculating the mean error we see that Euclidean distance has least mean error which shows that using Euclidean distance with time series clustering is most effective and efficient.

After applying ANOVA, we see that there is an statistically insignificant difference in these three techniques. ANOVA compares the mean of all three different techniques it says that all three techniques do not differ much and have almost same error at 92.9% of the times.

This is the ANOVA test results:

One-way ANOVA: C1, C2, C3

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Factor | 2 | 88 | 44 | 0.07 | 0.929 |
| Error | 30 | 17849 | 595 | | |
| Total | 32 | 17937 | | | |

Here, C1, C2, C3 are columns of error from the three methods namely Mahanolobis, Euclidean and Minkowski respectively.

With P value at .929 which is greater than .05. Thus, it proves that these three techniques are equally effective and can be used alternatively.

## V. CONCLUSION

In this paper, a new approach based on Mahanolobis distance, Euclidean distance and Minkowski distance model and dynamic time warping technique is used to analyse multivariate time series data on crime from National Crime Records Bureau's website.

The comparative analysis on the basis of mean error shows that using Euclidean distance for the calculation of time series data is best of all distances in the specified algorithm. After it we find that using Mahanolobis distance gives better results than Minkowski.

A variance comparison using ANOVA shows us that we have similar results with all the three methods if we work with different samples of data. Hence it can be said that all the three methods does not have much effective difference as the mean error may result in a different output when compared with a different sample of data.

Using the above methodology we can find out similarity factors between the crime scenarios which can help us detect the factors effecting the . This methodology can be applied to various disciplines with time series data available and analyse statistics to reveal useful trends and can help predict trends for future. Also it can help administration and government to make security arrangements.

## REFERENCES


[1] B. Chandra, Manish Gupta, M. P. Gupta, " *A Multivariate Time Series Clustering Approach for Crime Trends Prediction"*, 1-4244-2384-2/08/$20.00_c 2008 IEEE.

[2] Donna R. Tabangin, Jacqueline C. Flores, Nelson F. Emperador, "*Investigating Crime Hotspots Places and their Implication to urban Environmental Design: A Geographic Visualization and Data mining Approach"*, International Journal of Human and Social Sciences 5:4:2010.

[3] Peng Chen, Tao Chen, Hongyong Yuan, "*GIS Based Crime Risk Analysis and Management in Cities"*, 978-1-4244-7618-3/10/$26.00 _c 2010 IEEE.

[4] Shyam Varan Nath, "*Crime Pattern Detection Using Data Mining"* ,Oracle Corporation.

[5] Parang Saraf, Michael W. Mildo , Sarah C. Richards, Tirtha Bhattacharjee, Lonesome Malambo , "*Social Media Analysis and Geospatial Crime Report Clustering for Crime Prediction & Prevention"*, CS/STAT 5525 – DATA ANALYTICS, FALL 2012.

[6] Malathi. A , Dr. S. Santhosh Baboo, "*An Enhanced Algorithm to Predict a Future Crime using Data Mining"*, International Journal of Computer Applications(0975-8887) Volume 21- No.1, May 2011.

[7] Collen McCue, "*Data Mining and Predictive Analytics in Public Safety and Security"*, IT Pro July | August 2006.

[8] Patricia Brantingham, "*Crime and Place: Rapidly Evolving Research Methods in the 21st century"* , A journal of Policy Development and Research, Volume 13, Number 3, 2011.

[9] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padharaic Smyth, "*From Data Mining to Knowledge Discovery in Databases"*, American Association for Artificial Intelligence 0738-4602-1996/$2.00.

[10] Tony H. Grubesic, Alan T. Murray, "*Detecting Hot Spots Using Cluster Analysis and GIS"*, www.tonygrubesic.net/**hot_spot**.pdf

[11] "*GIS Software Requirements for Crime Analysis"*, International Association of Crime Analysts(2012), GIS Requirements for Crime Analysis (White Paper 2012-01).

[12] Mostafa Ahmadi, "*Crime Mapping and Spatial Analysis*, International Institute for Geo-Information Science and Earth Observation Enshede", The Netherlands.

[13] Elzabeth R. Groff, Nancy G. La Vigne, "*Forecasting the Future of Predictive Crime Mapping"*, Crime Prevention Studies, volume 12, pp.29-57.

[14] Chuck Ballard, Dirk Herreman, Don Schau, Rhonda Bell, Eunsaeng Kim, Ann Valencic , "*Data Modelling Techniques for DataWarehousing"*, psproject. googlecode.com /svnhistory/r37/trunk/sg242238.pdf

[15] http://en.wikipedia.org/wiki/Euclidean_distance

[16] http://en.wikipedia.org/wiki/Minkowski_distance

[17] http://en.wikipedia.org/wiki/Dynamic_time_warping

[18] http://en.wikipedia.org/wiki/Mahalanobis_distance

[19] http://www.springer.com/978-3-540-74047-6