Disaster Prediction System Using IBM SPSS Data Mining Tool

B.Renuka Devi[1], Dr.K.Nageswara Rao[2], Dr.S.Pallam Setty[3], Dr.M.Nagabhushana Rao[4]

[1]*Assoc. Prof, Dept. of CSE, VLITS, Vadlamudi, Guntur.*

[2] *Professor, Dept. Of CSE, PSCMR College of Engineering, Vijayawada.*

[3] *Professor, Dept. Of CS & SE, Andhra University, Visakhapatnam.*

[4]*Professor, Dept.of CSE, SCET, Narsapuram*

***Abstract*— Data mining, or knowledge discovery in databases, refers to the discovery of interesting, implicit, and previously unknown knowledge from large databases. Spatial data mining presents new challenges due to the large size of spatial data, the complexity of spatial data types, and the special nature of spatial access methods. Spatial data mining is the task of unfolding the knowledge hidden in the spatial databases. By collecting spatial data i.e patients' data, we analyze, predict and interpret the data to the health organizations for conducting Campaigns. The spatial Databases contain both spatial and non-spatial attributes .In this paper we focus effectively in designing disaster prediction system to identify the Dengue disease using Data mining tools i.e SPSS Modeler and Data mining algorithms.**

***Index Terms*— C5.0 Algorithm, Dengue Fever, Spatial Data Mining, SPSS Modeler, SPSS Statistics.**

1 INTRODUCTION

Dengue virus and dengue hemorrhagic fever are amongst the most important challenges in tropical diseases due to their expanding geographical distribution, increasing outbreak frequency, hyperendemicity and evolution of virulence [Dengue Bulletin – Vol 28, 2004]. Artificial Intelligence (AI), with its various subfields, has a long history of knowledge extraction, representation, and inference in medicine. In dermatology, applications of AI methods, the focus has traditionally been on image analysis and understanding, aimed at providing decision support for physicians. The field of computer-assisted dermatology has thus benefited greatly from advances in knowledge representation techniques and machine learning algorithms. Recently, increased connectivity and the ubiquitous availability of internet access have resulted in new opportunities for distributed and collaborative diagnosis. Clinical dermatology is mainly a visually dominated discipline. The recognition of signs and symptoms as well as their interpretation of patterns typical for specific diseases remains the core task for diagnosis. As such, dermatology benefits from intelligent computer applications that structure and analyze visual information, allowing the combination of this information with test results containing physical, biochemical or genomic information. During the last decade computer-assisted applications have proven to be of value for the diagnosis of various forms of skin cancer, especially cutaneous melanoma, Cholera, Dengue Fever, Malaria, Polio, etc... The increasingly large amount of geographical data available to physicians calls for computer-assisted methods to extract information and knowledge from the available data.

Geography is an integrative discipline and geographic data under analysis often span across multiple domains. The complexity of spatial data and geographic problems, together with intrinsic spatial relationships, constitute an enormous challenge to conventional data mining methods and call for both theoretical research and development of new techniques to assist in deriving information from large and heterogeneous spatial datasets (Han and Kamber 2001; Miller and Han 2001; Gahegan and Brodaric 2002**).** Health maps have become available as the use of geographical information systems in health related contexts increased [1][4][11].

A formula implemented as Hazard science to Risk Science, towards understanding the hazards and their consequences (risks), following a probabilistic approach using spatial data mining[1].Due to larger heterogeneity of spatial data, the providers of geographic data specify different models for same spatial objects. Context specific semantics is one of the best approach suggested which deals with provision of feature space derivations. Unknown and unexpected patterns, trends or relationships can hide deep in a huge feature space and make it very hard for analytical methods or visual approaches to find [Miller and Han 2000].

A hypothesis space is formed by all possible configurations of the tools used to detect patterns in a feature space. Characteristically, however, the hypothesis space for a large and high dimensional geographic dataset has an extreme degree of complexity. This is caused by several factors. First, each pattern may involve a different subset of variables from the original data, and the number of such subsets (hereafter subspaces), i.e., possible combinations of attributes, is huge. Second, inside a subspace, potential patterns can be of various forms (e.g., clusters can be various shapes). Third, for a specific pattern form (e.g., cluster of a specific shape), its parameter space is still huge, i.e., there are many ways to configure its parameters. Fourth, patterns can vary over geographic space, i.e., patterns can be different from region to region.

2 APPLYING SPATIAL DATA MINING

Spatial data mining becomes more interesting and important as more spatial data have been accumulated in spatial databases [9].

2.1 Spatial Statistics

Using spatial statistics measures, dedicated techniques such as cross k-functions with Monte Carlo simulations, lattice method have been developed to test the collocation of two spatial features. At the outset the studies include, the spatial data mining problem of how to extract a special type of proximity relationship – namely that of distinguishing two clusters of points , based on the types of their neighboring features is another study[2][6][8]. Classes of features are organized into concept hierarchies [3].A reasonable and rather popular approach to spatial data mining is the use of clustering techniques to analyze the spatial distribution of data. While such techniques are effective and efficient in identifying spatial clusters, they do not support further analysis and discovery of the properties of the clusters.

2.2 Mining Collocation Patterns

Mining collocation patterns give the standard of observing the generic characteristics of a given spatial zone with more relevant Boolean features with their s %(support) and c (confidence)[6]. The work of mining Collocation patterns into spatial statistics approaches and combinatorial approaches [7]. The spatial Collocation pattern mining presented in the erstwhile works has bias on popular events. It may miss some highly confident but **"**infrequent" Collocation rules by using only "support"-based pruning.

In a spatial database S, let F = {f_1,… , fk} be a set of Boolean spatial features. Let I = {i_1,…, in} be a set of n instances in the spatial database S, where each instance is a vector consisting of [instance-id, location, spatial features]. ~ Neighborhood relation R over pair wise locations in S exists ~ is assumed. The object of this collocation rule mining is to find rules. **A** and **B** are subsets of spatial features. **A** determines the set of spatial features that form the antecedent part of the rule and **B** defines the action and its consequential parts the support and the confidence. The rule indicates the coincidence of the spatial collocation rule absorbs the action of the rule in the "nearby" regions of the spatial objects that comply with the collocation rule. A collocation pattern C is a set of spatial features. A neighbor-set L is said to be a row instance of collocation pattern C if every feature in C appears in an instance of L, and there exists no proper subset of L does so. We denote all row instances of a collocation pattern C as row set(C). In other words, row set(C) is the set of neighbor-sets where spatial features in C collocate. The conditional probability is the probability that a neighbor-set in row set (A) is a part of a neighbor-set in row set (B). Intuitively, the conditional probability p indicates that, whenever we observe the occurrences of the spatial features in A, the probability to find the occurrence of B in a nearby region is p.

2.3 Finding/Estimating Symptoms to Build Collocations

Since 1998, we have developed and made use of the PC-based geographical information system (GIS) to manage the huge databases on cases and Aedes mosquitoes island-wide. Examples of information stored on the GIS are: patients' particulars, locations of Aedes breeding, larval densities, species of vectors, habitat types, premises types, and ovitrap locations [3]. The GIS enables us to visualize at a glance "hotspots" where cases or breeding are concentrated so that early control operations can be implemented. We can also perform spatial and temporal analyses of the data for future planning, such as the review of dengue sensitive areas; and for day-to-day operation planning such as the boundary of control operations in outbreak areas, the progression of an outbreak, etc.

The majority of houses have a cement water container located in the bathroom to store water for bathing and a smaller container in the water closet (WC). Water containers made from clay or plastic barrels/jars are also kept in the kitchen for cooking or drinking purposes. Additional water containers may act as potential breeding sites both inside and outside houses. The people are utilizing breed in pools of water.

Dengue (pronounced den' gee) the most prevalent Arthropod-borne viral (Arbor virus) belonging to the family Flaiviridae. The major dengue vector in urban areas is Aedes aegypti but Aedes albopticus is also present. It breeds in pools of water [13]. Only female can transmit the virus. Female mosquitoes can transmit the virus to the next generation of mosquitoes. Symptoms include severe and continuous pain in the abdomen, bleeding from the nose, mouth , skin bruising, frequent vomiting with or without blood, black stools like coal tar, excessive thirst, pale, cold skin. There is no specific treatment for dengue, but closely medical attention and clinical management saves the lives of many patients. At present, the only method of controlling dengue is to combat the vector mosquito through chemical control and environmental management. Remove tires, bottles, cans and other items that catch and retain water, so that potential breeding sites for vector mosquitoes can be eliminated

The disease proceeds in possibly three stages:
(a) Invasion (b) Collapse (c) Reaction

3 THE LAW OF TOTAL PROBABILITY

Although there are many solutions to prevent diseases, finding the right area to apply the prevention measure with right inputs becomes the criterion. The Bayes' theorem evaluates the reverse of conditionality of events; where the symptoms and the causative-agents are analyzed and found with a reciprocal equivalence. The Table-1 describes the most probable symptoms that cause the epidemics. The fact

that the person had a positive reaction to the test may be considered as our data to build the collocation pattern [17].

The conditional probability of the collocation is the probability that a neighbor-set explaining the features of existence of causative agent, infection sources, is a part of the global neighbor-set in the spatial domain for this epidemic application Given a spatial domain in a database view **S**, to measure the implication strength of a spatial feature in a collocation pattern, a participation ratio Pr(C,f) has to be defined. A feature f has a participation ratio Pr(C,f) in pattern C means whenever the feature f is observed, with probability Pr(C,f), all other features in C are also observed in a neighbor-set. In spatial application domain, as there are no natural transactions, for a continuous space, a participation index is proposed to measure the implication strength of a pattern from spatial features in the pattern. For a collocation pattern C, the participation index PI(C) = min fЄC Pr(C, f)}. In other words, wherever a feature in C is observed, with a probability of at least PI(C), all other features in C can be observed in a neighbor-set. A high participation index value indicates that the spatial features in a collocation pattern are likely show up together [14].

4 PROBLEM

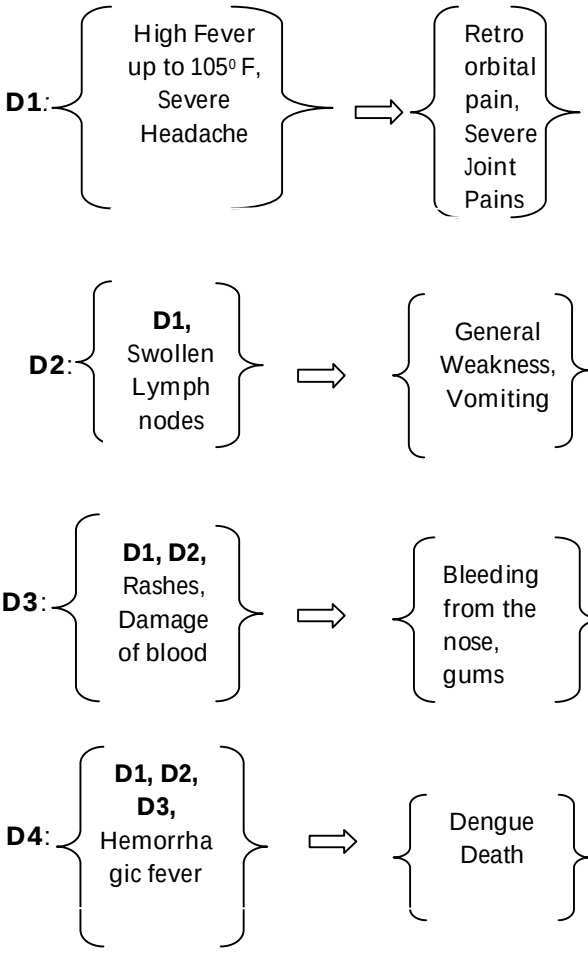
4.1 Detection of the Epidemic

Dengue fever (DF) is a self-limiting disease; however, dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS) are fatal. Since there is no therapy and vaccine against dengue, timely diagnosis is therefore necessary for patient management. Laboratory diagnosis is carried out by virus isolation, demonstration of viral antigen, presence of viral nucleic acid, and antibodies. Further, recombinant dengue envelope protein can be used to detect specific antibodies, both IgG and IgM against all four serotypes of virus using an E. coli vector. Death is generally due to the dehydration caused by the illness. The possibility that dengue outbreaks result from anomalous patterns of precipitation, we analyzed the relationships linking rainfall, the abundance of vector mosquitoes, the degree of source-reduction coverage, and the occurrence of dengue during the couple of year period of observation[16]. The co-location rules of spatial data mining are proved to be appropriate to design nuggets for disaster identification. The state-of-the-art and emerging scientific applications require fast access of large quantities of spatial data. Here both resources and data are often distributed in a wide area networks with components administrated locally and independently. The collocation rules are very useful in detecting the affected areas by finding the symptoms of a disease and influence of symptoms in a disease by using sample identifiers, the collocation can be explained as follows: Assuming firstly, the '**b**' as the consequence of feature '**a**' is developed, forms a first level of collocation, which is identified by **a** → **b**, secondly, if the consequence '**c**' from the feature '**b**' is developed, it forms a collocation, which is identified by **b** → **c**. As '**b**' already has an antecedent '**a**', the consolidated version of collocation, {**a, b**} → **c** can be formed. If '**c**' becomes another feature that can lead to the consequence of'd', then the notation wholly represents the cause of'd' as {**a, b, c**} → **d.** Also implies to {**a** ∪ **b** ∪ **c**} → **d** representation**.**

Similarly, considering the collocation pattern for the problem can be considered

C: {cause of epidemic} {causative agent, infection sources}; in the **nearby** region with high probability.

The collocation pattern is considered with practically proved parameters for dengue as follows



Assuming **X** as defined representation of collocated sequence of patterns i.e., $\mathbf{D_1}$, $\mathbf{D_2}$, $\mathbf{D_3}$, $\mathbf{D_4}$ are the resultant collocation patterns of the disease Dengue. The participation ratio describes the intensities of the symptoms that play

important role to form the collocation rule and builds the reference future.

The general syntax for assessing the reference feature is Pr(C, f).If the probabilities of some features w.r.t $\mathbf{D_1}$ are understood as having the maximum and minimum. If the lead feature of the collocation contains least probability then collocation is considered as feebly important. If the lead feature of the collocation contains higher probability then collocation is considered as highly important [14].The probabilities mentioned in the problem are <excreted along with innumerable Vibrios>, <loss of fluid, electrolyte imbalance>. If one of them or some of them exhibit high probability, then there is a high significance of occurring the disease severely, for low exhibition of probability, the existing of the disease will be indicative. However, the features and the probabilities considered will prove the collocation to be appropriate for the causation of severity of dengue spectrum (simple dengue to dengue death).

5 ALGORITHM

The following algorithm is to find the spatial knowledge i.e. dengue disaster from health demographic data.

- Data collection from the patients.
- Attributes are selected.
- Collocation rule is applied.
- Spatial predicate is applied.
- Source (Area) of disaster identified.

5.1 Introduction to Data Mining Tool SPSS

The "Statistical Package for the Social Sciences" (SPSS) is a package of programs for manipulating, analyzing, and presenting data[18]; the package is widely used in the social and behavioral sciences. There are several forms of SPSS. The core program is called *SPSS Base* and there are a number of add-on modules that extend the range of data entry, statistical, or reporting capabilities. In our experience, the most important of these for statistical analysis are the **SPSS Advanced Models and SPSS Regression Models** add-on modules. SPSS Inc. also distributes stand-alone programs that work with SPSS. SPSS Modeler is a set of data mining tools that enable you to quickly develop predictive models using business expertise and deploy them into business operations to improve decision making. Designed around the industry-standard CRISP-DM model[19], SPSS Modeler supports the entire data mining process, from data to better business results. SPSS Modeler[20] offers a variety of modeling methods taken from machine learning, artificial intelligence, and statistics. The methods available on the Modeling palette allow you to derive new information from your data and to develop predictive models. Each method has certain strengths and is best suited for particular types of problems. While the data mining tools in SPSS Modeler can help solve a wide variety of business and organizational problems. SPSS is a statistical analysis and data management software package. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and conduct complex statistical analyses. SPSS include modules like SPSS Statistics and SPSS Modeler. **SPSS Statistics** is a software package used for statistical analysis. Manipulate highly complicated data and analysis by Simple instructions. There are 4 windows in it. They are: Data editor, Output viewer, Syntax editor, Script window. **SPSS Modeler** offers a strategic approach to find useful relationships in large data sets. Algorithms used in SPSS include

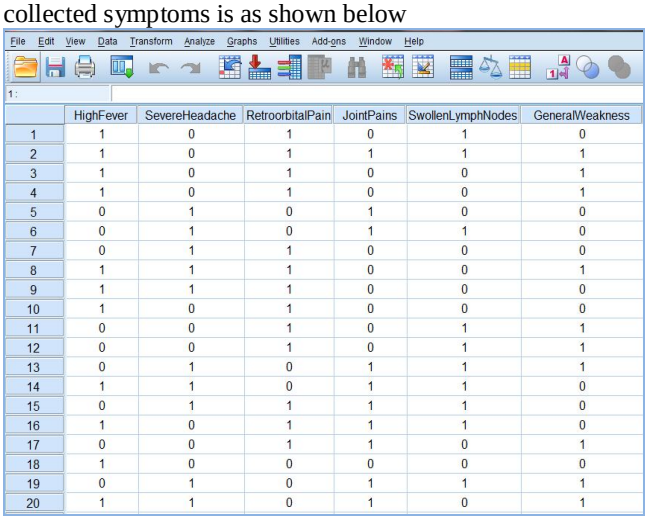
C5.0 algorithm: C5.0 algorithm is used to build either a decision tree or a rule set. A C5.0 model works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.

QUEST algorithm: QUEST—or Quick, Unbiased, Efficient Statistical Tree—is a binary classification method for building decision trees. A major motivation in its development was to reduce the processing time required for large C&R Tree analyses with either many variables or many cases. A second goal of QUEST was to reduce the tendency found in classification tree methods to favor inputs that allow more splits.

Neural net: A neural network can approximate a wide range of predictive models with minimal demands on model structure and assumption.

5.2 Prototype

Data is collected from the patients and read in to SPSS Statistics. In this 2 views are observed. One is the Data View and another one is the Variable View. Data View of the collected symptoms is as shown below



	HighFever	SevereHeadache	RetroorbitalPain	JointPains	SwollenLymphNodes	GeneralWeakness
1	1	0	1	0	1	0
2	1	0	1	1	1	1
3	1	0	1	0	0	1
4	1	0	1	0	0	1
5	0	1	0	1	0	0
6	0	1	0	1	1	0
7	0	1	1	0	0	0
8	1	1	1	0	0	1
9	1	1	1	0	0	0
10	1	0	1	0	0	0
11	0	0	1	0	1	1
12	0	0	1	0	1	1
13	0	1	0	1	1	1
14	1	1	0	1	1	0
15	0	1	1	1	1	0
16	1	0	1	1	1	0
17	0	0	1	1	0	1
18	1	0	0	0	0	0
19	0	1	0	1	1	1
20	1	1	0	1	0	1

5.3 Results

The collected Patients data is obtained in an SPSS statistics file. In the first Stream only historical information has chosen and role of Dengue Death as the target. Applied C5.0, Neural Net and QUEST algorithms for analyzing the efficiency. By analysis, C5.0 is the most efficient algorithm. Collected new patients data and applied C5.0 Algorithm.

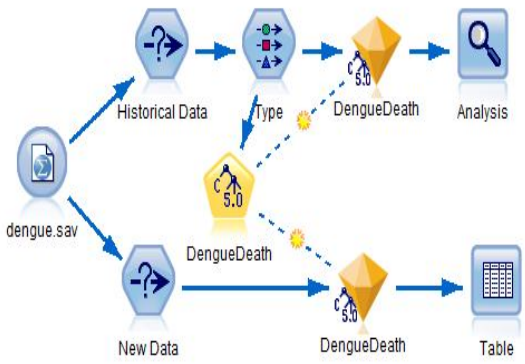


Fig 2 Model Designed for Dengue Prediction Based on C5.0 Algorithm

Based on the historical data the new patients have chances of getting affected by Dengue i.e **80%.**

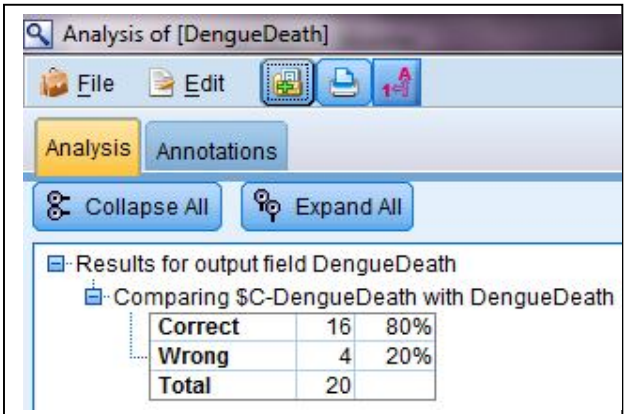


Fig 3 Percentage of People Affected by Dengue

The collocation pattern formed by this sample region acts as a cautious measure or the forecast for the bio-medical researchers, analysts and other health-care-takers of the spatial zone which will be useful for them to take suitable remedial campaigns. Socio-statistical methods related to health-science can be implemented to regulate the input variables that play a parametric role of collocation rule formation, in order to prevent the epidemic in the spatial zone, if not permanently, at least suitable preventive measures can be undertaken for the affect of such candidate epidemic in the interested spatial zone.

6 Conclusion

Epidemics, infectious diseases are generally caused by a change in the ecology of the host population. It spreads rapidly and extensively by infection and affecting many individuals in an area or a population at the same time. A spatial zone probabilistic study is made on the health demographic data. A Collocation rule is defined as a syntactic representation of the parameters in the form of antecedent and consequent. Using the Collocation rule, the affected area of Dengue is found and results are obtained. Using data mining tool SPSS Statistics, the computation is done. Model is designed using C5.0, Neural Net & QUEST algorithms. Among them, C5.0 Algorithm is proved to be more efficient in predicting the epidemic disaster.

7 References

1. Alan T.Murray, Ingrid McGuffog, John S.Western and Patrick Mullins, Exploratory Spatial Data Analysis for Examining Urban Crime, 2001. www.gegraphy.hunter.cuny.edu

2. Bavani Arunasalam, Sanjay Chawla, Pei Sun and Robert Munro, Mining Complex Relationships in the SDSS SkyServer Spatial Database, School of Information Technologies, University of Sydney, proceedings of 28th annual international computer software and application conference (CONPSAC-04), 2004, IEEE.

3. Chawla, Shekkar, Spatial Databases: A Tour, 2003, Prentice all. (ISBN – 013 – 017 480 – 7)

4. Hardy Pundt, Evaluating the relevance of spatial data in time critical situations, University of Applied Sciences and Research, Faculty of Automatisation and Computer Science, wernigerode (DE), pp 779– 788, 2005, www.springer.com/3-540-249988-5.

5. Huang, Shekkar, Xiong, Discovery Collocation Patterns from Spatial Data Sets: A General Approach, IEEE-KDE, volume 16, No: 12, dec 2004.

6. Knorr and Ng, Extraction of Spatial Proximity Patterns by Concept Generalization, proceedings second international conference of KDD, pp. 347 – 350, aug 1996.

7. Koperski and Han, Discovery of Spatial Association Rules in GI Databases, proceedings 4th international symposium large spatial databases, pp. 44-66, aug 1995.

8. Munro, Chawla, Complex Spatial Relationships, proceedings of IEEE international conference on data mining, 2003.

9. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Spatial Data Mining: A Database Approach, Institute for Computer Science, University of Munich, proceedings of 5th international symposium of large special databases, 2007.

10. Masaharu Yoshioka, Yasuhiro Shamoto, Tetsuo Tomiyama, an Application of the Knowledge Intensive Engineering Framework to Architectural Design, a research project of GNOSIS, IMS Program, 1998.

11. Naresh Raheja, Ruby Ojha, Sunil R Mallik, Role of internet-based GIS in effective natural disaster management,R.M.S.I. www.development.net/technology/gis/

12. Nagabhushana Rao, Ramesh babu, Sangameswar, Spatial Knowledge Algorithm For epidemics Using Data Mining Techniques. International conference-ICSCI-2008, Hyderabad, Jan 2008.

13. Nagabhushana Rao M, S.V.V.D.Venugopal, Disaster Management System for Dengue , IJCST/33/4/A-1021 http://ijcst.com/?page_id=2618

14. Shaw C. Feng, Y. Zhang: Conceptual Process Planning - A Definition and Functional Decomposition, Manufacturing Engineering Laboratory, National Institute of Standards and Technology 1997.

15. Yan Huang, Hui Xiong, Shashi Shekhar, Jain Pei, "Mining Confident Collocation Rules without A support Threshold". Symposium on applied computing, proceeding of the 2003 ACM Symposium on Applied Computing, Florida, PP: 497-501, 2003, ISBN: 1-58113-624-2.

16. New Initiatives in Dengue Control in Singapore byTan Boon Teng. Dengue Bulletin – Vol 25, 2001

17. Vector Densities That Potentiate Dengue Outbreaks in A Brazilian City, Ricardo J. S. Pontes, Jonathan Freeman, Jose´ Wellington Oliveira-Lima, J. Christina Hodgson,AndAndrew Spielman, Am. J. Trop. Med. Hyg., 62(3), 2000, pp. 378–383

18. Argyrous, G. Statistics for Research: With a Guide to SPSS, SAGE, London.

19. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step data mining guide*

20. http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03085usen/YTW03085USEN.PDF