TERM PAPER

ON

WIRELESS COMMUNICATION – BLUETOOTH

SUBMITTED TO

AMITY SCHOOL OF ENGINEERING AND TECHONOLGY

Guided By- Mr Roshan Lal Chokker Submitted By- Reyhan Gupta

Assistant Professor Enrolment Number: A2305214091

Computer Science Department Roll Number:

ASET



AMITY UNIVERSITY UTTAR PRADESH

GAUTAM BUDDHA NAGAR

**DECLARATION**

I, Reyhan Gupta, a student of B.Tech CSE hereby declare that the project titled “Statistical Analysis using Data Mining: A District Level Analysis of Malnutrition” which has been submitted by me to the Department of Computer Science, Amity School of Engineering and Technology, Amity University, Uttar Pradesh, Noida is a partial fulfilment of the award of the degree of Bachelor of Technology in CSE has not previously formed the basis for the award of any degree, diploma or other similar title or recognition.

The Author attests that permission has been obtained for any copyright material appearing in the Project, other than brief excerpts requiring only proper acknowledgements in scholarly writing and all such use id acknowledged.

Signature

Noida

Date Name and signature of student

**CERTIFICATE**

This is to certify that Reyhan Gupta, a student of B. Tech CSE has carried out the work presented in the project of the term paper entitled “Statistical Analysis using Data Mining: A District Level Analysis of Malnutrition” as a part of third year program of Bachelor of Technology in Computer Science Engineering from (Amity School of Engineering and Technology) Amity University, Noida, Uttar Pradesh under my supervision.

Name and signature of the Faculty Guide: Dr Tanupriya Choudhury

Amity School of Engineering and Technology, AUUP

**ACKNOWLEDGEMENTS**

The project came to me as an excellent learning opportunity and gave me such a wonderful exposure. I would like to thank the authorities of Amity University for giving me this opportunity to work in this esteemed institution under the guidance of highly knowledgeable and experienced faculty and industry guide.

I would like to thank our Vice Chancellor Professor Dr. Balvinder Shukla and our Head of The Department Dr. Abhay Bansal for their overall support throughout the project.

I would also like to thank my project guide Dr Tanupriya Choudhury under whose guidance I was able to complete my project. I am wholeheartedly thankful to him for giving me his valuable time and attention and providing me with a systematic way to complete this project on time.

Reyhan Gupta

Computer Science and Engineering

A2305214091

B. Tech CSE (2014 - 2018)

Amity University Uttar Pradesh, Noida (UP)

**Contents**

Abstract......................................................................................................................................6

1 Introduction.............................................................................................................................7

2 Literature Survey.....................................................................................................................8

2.1 Data Mining Methods......................................................................................................9

2.2 Clustering.........................................................................................................................9

2.3 Davies Bouldin Index.....................................................................................................10

3 Methodology.......................................................................................................................11

3.1 Steps Involved................................................................................................................11

3.2 Experimental Setup........................................................................................................13

3.3 Model Structure..............................................................................................................14

4 Result and Analysis.............................................................................................................15  
 4.1 Clusterization.................................................................................................................15

4.2 Correlations....................................................................................................................16

5 Discussion..........................................................................................................................19

Conclusion...............................................................................................................................21

References...............................................................................................................................22

**Abstract**

Children in India suffer from the highest level of undernourishment in the world, having serious effects on health. Bihar in turn has the highest incidence in India. This study was taken up to ascertain the effect of clusterization on a large data set with respect to analyses the association of multitude of variables with malnutrition indices. Raw data was obtained using Secondary data sources, especially Government reports, pertaining to malnutrition indices, demographic, social, nutritional, economic and medical factors causing malnutrition. In stage one the variables from non-clustered data was correlated with malnutrition indices (Stunting, wasting and underweight). Subsequently the data was split using Rapid Miner Studio (version -7.2.003) into three clusters. This segregation was done by the software using k-means and Hierarchical agglomerative clustering. In the second phase each of these clusters were again analyzed using the software and the correlation results were compared. Significant variation was observed in most of the correlations in the clustered and un-clustered data sets. This indicates the importance of clusterization in reaching the truth when a statistical analysis is carried out, as clusterization excludes/segregates the outliers / extremes of values. This has significant implications in policy making for malnutrition control, through identifying the most relevant variables/factors responsible.

**1. Introduction**

As per the World Bank Report, 2015, India fares very poorly in Global Hunger Index. India scores 29 points and its malnutrition condition is stated as ‘serious’, as it is ranked even below Bangladesh, Sri Lanka and Sub-Saharan Africa. Bihar lies at the bottom of Malnutrition table in India, with child undernourishment rates to be as high as 80 percent. [1] Though several polices have been initiated to address malnutrition in the state, it seems either the schemes presently running in the state are not reaching the beneficiaries or they are flawed. It is also possible that the policies for malnutrition control are not based on sound presumptions, which are meant to be dependent and, are as good as, relevant data generation, analysis, interpretation and application. [2]

The latest NFHS 4 data released by the Government of India in 2017, gives the district level details of many determinants of Malnutrition in Bihar. [3] A maze of factors, ranging from poverty, economic status of family, medical condition of child, infections, nutritional status of women/mothers, demographic and social factors (literacy, female literacy), political will, to poor implementation of food-policies qualify as determinants of malnutrition. The relative importance of these determinants, is critical in selecting them to be targeted for effective malnutrition control.

Here lies the importance of efficient data mining which could delineate the most important undernutrition-determinants in a given district (or set of districts with similar social/demographic/medical attributes). Once such a set/cluster of similarly placed districts is segregated, it is easy to analyze it further with respect to the most critical determinants prevalent as a cause of undernutrition in that district/cluster.

Since there are infinite permutations and combinations possible within these clusters, it is not humanly possible to pick the right ones up and process it. The data miner enables us to perform this cluster analysis so as to focus on the most important determinants and to prioritize them for selective policy making and intervention.

The present study is undertaken to appreciate the utilization of Rapid Miner Studio in data analysis with special emphasis on the importance of clusterization in the correct interpretation of results in a large set of data.

**2. Literature Survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TITLE | AUTHOR | JOURNAL | DETAILS | Important Factors |
| 1. Malnutrition among under-five children in India and strategies for control | Swaroop Kumar Sahu, S. Ganesh Kumar | J Nat Sci Biol Med. | 2015 Jan-Jun; 6(1): 18–23:- doi: 10.4103/0976-9668.149072 | Quality of food items, large family and poor sanitary environment |
| 2. Nutritional status of children in India: household socio-economic condition as the contextual determinant | Barun Kanjilal, Papiya Guha Mazumdar | US National Library of Medicine  National Institutes of Health  Search database  Search term | Published online 2010 Aug 11. doi: 10.1186/1475-9276-9-19 | altering the nutritional intake, behavioral changes with nutrition policy implementation |
| 3. Factors affecting prevalence of malnutrition among  Children under three years of age in botswana | Salah E.O. Mahgoub  \*1  , Maria Nnyepi  2  , Theodore Bandeke  3 | AJFAND Online | Vol 6 Number 1 Published 2006 | households with low earnings,  low parental education or households in  rural areas  as maternal income, maternal education, and the creation of employment or economic  engagements |
| 4. Trends and Factors affecting Female Literacy-An inter-district  study of Maharashtra | Aparna Samudra | International Journal of Gender and Women’s Studies | ISSN: 2333-6021 (Print), 2333-603X (Online)  June 2014, Vol. 2, No. 2, pp. 283-296 | Male literacy level(+), age of marriage(+), Workforce participation-Employment(-), Female headed households(-) |
| 5. Factors Influencing Low Level of Women Participation in  Literacy Program in Maiha Local Government Area  Adamawa State | Dr. Aminchi daniel | Journal of Education and Practice(IISTE) | ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)  Vol.6,  No.15, 2015 | Poverty-lack of basic amenities (+), gender stereotyping (+), Socio-cultural belief- unable to contribute in development of society due to lack of education, Lack of awareness |
| 6. Factors affecting female  participation in education in seven  developing countries | Colin Brock & Nadine Cammish | Research Paper No. 09,Department for International Development | Serial No. 9  ISBN: 1861920 65 2  March 1997. | Geographical, Socio-Cultural, Health, Economic, Religion, Legal, Initiatives |

**2.1 Data Mining Methods**

Data Mining is a field which helps us to find unknown patterns in data which may otherwise seem incoherent. The data mining methods that have been used are clustering, regression, association, classification, followed by statistical analysis. [4]

**2.2 Clustering**

Clustering is a technique to group different parameters based on the similarities that exist between them and their interrelationship with each other. It enables us to understand how data is distributed and to understand the similar and dissimilar objects in the data. There are many ways to implement this technique, the important ones are

* Centroid Based Clustering
* Hierarchical Method
* Distributed Clustering
* Density Based Clustering

These help us find clusters in our data following different approaches. The algorithms which are popularly used are the K-means algorithm, Fuzzy C means method, Hierarchical clustering and others alike, which are all used widely for achieving similar goals. [5] For this paper the K-means Clustering algorithm and Hierarchical Agglomerative clustering were implemented which are partition and hierarchical algorithms for clustering, respectively. According to the k means algorithm, if we have a data set having n objects, we can partition it into k different partitions (or clusters) such that k ≤ n. We need to make sure that each data point gets associated with a cluster. [11] The optimal value of k can be determined by the

Davies Bouldin index. According to this index, the most suitable value of k is decided by the smallest DavisBouldin index which tells about the closeness of values within the cluster i.e. lesser the DB index, more optimal the value of k to be selected for further analysis. [9]

The K means algorithm generates clusters based on the following pseudo code:

Procedure k-means

Set ai…ak be distinct randomly selected inputs from xi … xk

repeat

for i=1…n do

yi j = 1 if j = argmin j 2

yi j = 0 otherwise

end for

for j=1…k do

nj =

aj =

end for

until convergence

end procedure

Return a1 … ak

The following expression is used during analysis of k means:

J = 2

J=Objective function, k= number of clusters, n= number of cases, cj = centroid for cluster j, x= case number

The Hierarchical agglomerative clustering algorithm was also applied on the data set to compare and contrast the results obtained with the k-means algorithm. This algorithm uses a bottom up clusterization approach which leads to the creation of sub clusters and repeatedly creates more of these. Clusters are generated using the following steps:

1) Each object is assigned to a separate cluster,

2) Respective distances between various clusters formed is evaluated

3) A distance matrix is then created using the values obtained.

4) The clusters with the shortest distance between them are selected and removed from the matrix.

5) These are then merged into a single cluster.

6) The above process is repeated till the matrix is left with a single element and further comparisons are not possible.

**2.3 Davies Bouldin Index**

The K means algorithm can be implemented for varying degree of clusters expressed by k. The optimal value of K can be found by considering the lowest Davies Bouldin Index which can be found using the expression shown hereunder:

DB =

Here, the count of clusters is shown by N, Di is the cluster distance ration such that average distance between each point and its corresponding centroid in that cluster are considered, for all the clusters in the data set.

**3. Methodology**

**Scope of Study**: The present work is an in depth district level exploratory analysis where information has been gathered applying the 'Secondary data collection' approach. The data pertained to all the 38 districts of Bihar. Twenty seven potential determinants of malnutrition of economic, health, medical, social and demographic categories were shortlisted and studied. Correlations were studied for all variables and the significant ones were analyzed in depth.

**Period of Study**: The present study was undertaken in late 2016, and results analyzed and interpreted in 2017.

**Data Sources**: All requisite variables were not available at one place. Relevant social, medical and malnutrition related data was accessed primarily from NFHS4. Demographic, social, medical and economic data were obtained from Indian Census, RSOC and Economic Surveys.

**Data Extraction**: The data with reference to Bihar was analyzed and the pertinent variables were listed in an Excel sheet and further imported into the Rapid Miner Studio.

**Statistical Analysis and Interpretation**: Data was analyzed using Rapid Miner Studio-7.2.003(rev: a8c41d-WIN64). This is an open source tool which provides an environment for data mining, business analytics and other applications alike. An in depth analysis was undertaken using clustering algorithms, correlation, regression, plotting etc., to understand critical determinants of malnutrition.

**3.1 Steps Involved**

1) **Obtaining the Data**: The primary source for social and medical data was NFHS4, of 2017 Government of India (GoI). The demographic data was derived from Census of India, 2011 and the economic data from Economic Survey of GoI. [3]

2) **Data Cleaning**: Data was cleaned and prepared before analysis. The few missing entries were replaced with respective averages. Columns with no significance to the study were not considered. All values were brought to percentage form for uniformity and ease of analysis. Certain data which are traditionally depicted as ‘per 1000’ (e.g. Sex Ratio) was also converted into percent for the sake of uniformity in analysis.

3) **Importing Data**: Data was imported in .xlsx format (Excel file) and saved in the Local Repository of the Rapid Miner Studio, ready for analysis.

4) **K means Clustering Algorithm**: To proceed with the analysis, the data is divided into clusters by using the k-means clustering algorithm.

5) **Hierarchical Agglomerative Algorithm**: The same data was run through the Hierarchical Agglomerative algorithm to compare the results with that of k-means, and selecting the most suitable approach.

6) **Davies Bouldin Index**: The K means algorithm can be implemented for different count of clusters denoted by k. The optimal value of k was found by considering the smallest Davies Bouldin Index value obtained after calculating it for different k values. The optimal K value was then selected.

7) **Correlation Analysis**: Correlation is found out by using the Pearson’s correlation coefficient (embedded in the software) to determine how strongly the values are correlated with dependent variables (malnutrition). [6]

A flowchart has been shown alongside which shows the steps involved in the process.

No

Yes

Visit NFHS4 website, Census India website, RCOS and various Economic Surveys

Compile the data in an Excel File

Import data into Rapid Miner Studio

Perform k-means clustering

Data Inconsistent?

Perform Correlation Analysis

Clean Data

Minimum DB Index?

No

Yes

Fig. 1. Flowchart showing steps involved

**3.2 Experimental Setup**

The Dependent variables for the study were Malnutrition indicators, i.e., Stunting, Wasting, Severe Wasting and Underweight. These were analyzed against the independent variables i.e. Demographic, Education, Mother/Child Health and Health Services related factors. The important parameters have been described in Table 1.

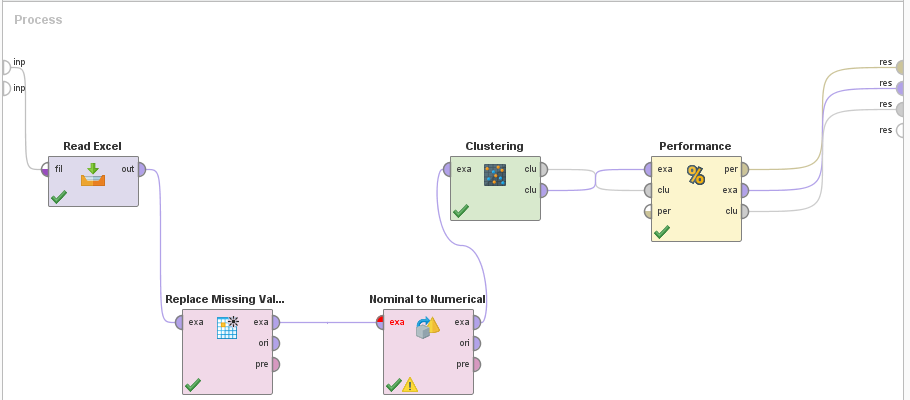
|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Definition** | **Parameter** | **Definition** |
| Electricity% | Percentage of villages with electricity | Exclusive breastfed % | Infants fed only breast milk till 6 months of age |
| Water source % | Percentage of villages with clean water source | Full ANC visits % | Women who visited Dr 4 times for check-up and took TT, Iron tablets during pregnancy |
| Sanitation  % | Percentage of villages with toilets at home | Full immunization  % | Children receiving full course of vaccination |
| Clean fuel % | Percentage of villages with clean fuel | Women BMI <18.5 % | Women with BMI<18.5, (malnourished) |
| Female Literacy % | Women who could read/ write(Census) | Women anemic % | Women with Haemoglobin <10g% |
| Anemia Children % | Children with Haemoglobin <10g% | Women married before 18 % | Women aged 19 to 45 years of age, married before age of 18 years |
| Women 10th class % | Women who had passed Class 10th or more | Adequate diet  % | Infants getting adequate diet |
| Colostrum % | New-borns given the first breast milk | Sex Ratio % | Number of women per 1000 men in the district |
| Under-weight % | Children falling below standards for weight for age criteria | Severely Under-weight  % | Children falling below standards for weight for age criteria (<3SD) |
| Wasted % | Children falling below standards for Weight for height (<2SD) | Severely wasted % | Children falling below standards for Weight for height criteria(<3SD) |
| Stunted  % | Children falling below standards for height for age (<2SD) | Severely Stunted % | Children falling below standards for height for age criteria(<3SD) |

Table 1: - Attributes of Dataset (The determinants of malnutrition have been highlighted in green, and we try to find out how the other independent factors-shown in white- affect them**)**

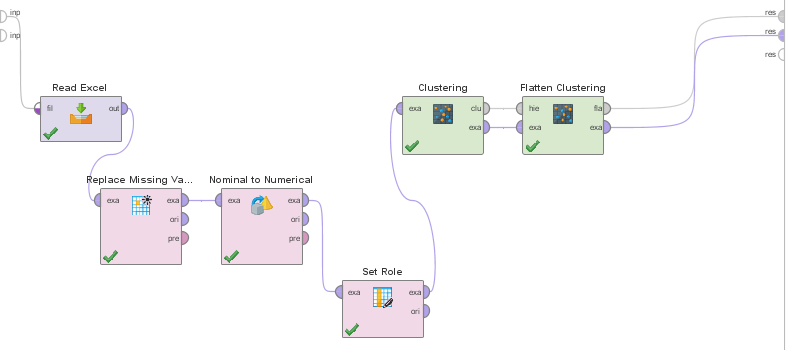
**3.3 Model Structure**

Rapid Miner Studio was used for analyzing the data, running clustering and correlation algorithms and to generate results. The model structures used in Rapid

Miner are shown in figure 2 and 3.



**Fig. 2.** Experimental model used in Rapid Miner Studio for implementing the k-means algorithm



**Fig.3.** Experimental model used in Rapid Miner Studio for implementing Hierarchical Agglomerative algorithm

The excel data sheet was imported and saved in the local repository. It was then passed through a Replace Missing Value attribute where the data fields with no values were fixed by replacing them with average values.

This data was then plugged into a Nominal to Numerical Filter. This was carried out as the data set comprised of alphanumeric and numeric entries and the algorithm was not able to work on such datasets.

After cleaning and preparing the data, the K means clustering algorithm was applied to the dataset using the k-means clustering operator. The results generated were fed into a Cluster Distance Performance Operator to analyze the cluster with respect to centroid distance, Davies Bouldin index, etc. to get a clearer understanding of the clustering carried out by the operators. [7]

The same data set was passed through a hierarchical agglomerative algorithm. The flattening of clusters was used to limit the number of clusters to the most appropriate value and a complete link mode was implemented, subsequently. To determine the optimal value of k, the clustering algorithm was run for various different values of k. Davies Bouldin index was noted afterwards.

The optimal value of k was selected for the point at which the Davies Bouldin index was minimum. As seen in table 2, the optimal k value was 3 and using this value further results were obtained and analyzed.

**Table 2.** Optimal Davies Bouldin Index (shown in green)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K=3 | K=4 | K=5 | K=6 | K=10 |
| DB  Index  =  -0.303 | DB  Index  =  -0.399 | DB  Index  =  -0.457 | DB  Index  =  -0.350 | DB  Index  =  -0.387 |

**4 Result and Analysis**

**4.1 Clusterization**

K means clustering was implemented on the data set with k=3 giving optimal Davies Bouldin Index

The Hierarchical Agglomerative algorithm was also used, with the flattening value set at 3 for most accurate results. It was observed that this approach yielded the same results as the k-means algorithm. Hence we could conclude that the clusterization was carried out in an accurate manner, and the results would be closest to the truth. However, in this paper the k-means algorithm was used for further analysis because it has a few advantages over the other, which include more efficient performance, reduced time for execution, consistent performance results even when data set increases, which are not seen in the Hierarchical agglomerative algorithm. [8]

Since the data set contains multitude of variables, with widely varying figures over a large magnitude, the conventional analysis of this data set as a whole, may give fallacious results. For example, certain extreme values (very low or very high) which are present as outliers do not giv`e the correct interpretation on analysis. It is therefore vital to segregate the dataset into clusters with relatively similar attributes, for more meaningful analysis. [10] Rapid Miner was used to make clusters, applying K Means algorithm and the Hierarchal agglomerative algorithm. The clusters obtained in this data set are as shown in Table 3.

**Table 3**. Clusters and corresponding number of objects

|  |  |
| --- | --- |
| Cluster Number | Objects:- (Districts} |
| Cluster 1 | 1 District:- (Patna) |
| Cluster 2 | 5 Districts:- (Begusarai, Bhagalpur, Munger, Muzzafarpur, Rohtas) |
| Cluster 3 | 30 Districts:- (Araria,Arwal,Aurangabad, Banka, Bhojpur, Buxer,, Darbhanga, Gaya, Gopalganj, Jamui, Jehanabad, Kaimur, Kathiar, Khagaria,Kishanganj, Lahisarai, Madhenpura, Madhubani, Nalanda, Nawada Pash, Champaran, Purb, Champaran, Purnia, Saharsa, Samastipur, Saran, Sheikhpura, Sheohar, Siwan, Sitamarhi) |
|  | Total = 36 districts |

To observe the relevance of results obtained after clusterization, a baseline analysis was carried out on original data prior to clusterization as well. Results have been tabulated in Table 4

**4.2 Correlations**

Correlations were worked out between the Dependent variables (i.e. the malnutrition indicators, namely Stunting, Wasting, Severe Wasting and Underweight with the independent variables namely Demographic, Education, Mother/Child Health and Health Services related factors. The important correlations are depicted in tables below.

The results of statistical analysis, using Rapid Miner Studio are summarized in the subsequent tables. Table 4 depicts results before clusterization, from the original data involving all 36 districts of Bihar

**Table 4**. Correlations before Clusterization (36 Districts): Different Variables and Malnutrition

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Capability Deprivation**  **Indicator** | **Stunting**  **(r value)** | **Wasting**  **(r value)** | **Severe Wasting (r value)** | **Under-weight**  **(r value)** |
| ***Demographic***  ~Sex ratio  ~Women married <18 yrs. | -0.34  0.48 | -0.20  0.06 | -0.10  -0.04 | -0.37  0.32 |
| ***Education***  ~Female literacy  ~Women studied till 10th | -0.31  -0.28 | 0.17  -0.23 | 0.15  0.22 | -0.12  -0.09 |
| ***Mother's Health* ~**Malnourishment (BMI<18.5)  ~ Anaemia | -0.49  0.09 | 0.21  0.24 | 0.90  0.12 | 0.51  0.23 |
| ***Child Health***  ~ Episode of diarrhoea  ~ Episode of Acute Respiratory Infections  ~ Anaemia | -0.14  -0.31  0.07 | -0.53  -0.37  -0.11 | 0.42  -0.35  0.08 | -0.44  -0.41  0.01 |
| ***Availability of Health Services***  ~ AWW  ~ ASHA | 0.15  0.17 | -0.28  -0.16 | -0.20  -0.03 | -0.07  0.08 |
| ***Utilization of Health Services***  ~ANC  ~ 4 ANC Visits  ~Immunization | -0.41  0.26 | 0.11  0.25 | 0.13  0.31 | -0.25  0.21 |
| ***Child Feeding Practices***  ~Breast fed in 1 hr  ~Exclusive breast Feeding  ~Colostrum fed  ~Adequate diet to infant | 0.33  -0.30  0.07  0.03 | 0.16  -0.28  0.14  -0.15 | 0.034  -0.17  -0.01  -0.17 | 0.30  -0.39  0.23  0.01 |

The data is segregated by Rapid Miner Studio into clusters with similar attributes for more efficient analysis with the objective of finding meaningful trends in the data set, and the results are depicted in Tables 5 and 6.

**Cluster 1**: District Patna, the most urban, economically and socially developed district of Bihar was found to be an outlier with significantly differing parameters as compared to other districts. Hence it was segregated as just one object in that cluster, so no correlations were possible and no tables could be drawn for it and was therefore couldn’t be analyzed further.

**Cluster 2**: This cluster contained 5 districts. Attributes are summarized in Table 5

**Table 5**. Primary: Basic Malnutrition Outcome /Dependent Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Capability Deprivation**  **Indicator** | **Stunting**  **(r value)** | **Wasting**  **(r value)** | **Severe Wasting**  **(r value)** | **Under-weight (r value)** |
| ***Demographic***  ~Sex ratio | -0.54 | -0.55 | -0.11 | -0.86 |
| ***Economic***  ~Per capita income  ~Ruralisation | -0.55  0.53 | -0.30  -0.64 | -0.3  0.02 | -0.13  0.04 |
| ***Mother's Health***  ~Anaemia | -0.19 | 0.74 | 0.24 | 0.30 |
| ***Child Health***  ~Acute Respiratory Infections  ~Anaemia | 0.35  -0.36 | -0.59  0.85 | -0.52  0.83 | 0.05  -0.37 |
| ***Availability of Health Services***  ~ AWW  ~ ASHA | -0.32  0.25 | -0.26  -0.74 | 0.06  -0.13 | -0.75  0.08 |
| ***Utilization of Health Services***  ~ANC | -0.34 | 0.08 | 0.25 | -0.15 |
| ***Child Feeding Practices***  ~Breast fed in 1 hr  ~Colostrum fed  ~Adequate diet to infant | -0.30  0.89  -0.30 | 0.05  -0.43  0.26 | -0.26  -0.30  -0.31 | -0.39  0.82  -0.16 |
| ***Health Facilities***  ~Water  ~Sanitation | 0.18  -0.90 | -0.53  0.31 | 0.23  -0.15 | -0.26  -0.64 |

**Cluster 3**: This cluster contained 30 districts/objects. Attributes are summarized in Table 6.

**Table 6**: Basic Malnutrition Outcome (Cluster 3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Capability Deprivation**  **Indicator** | **Stunting**  **(r value)** | **Wasting**  **(r value)** | **Severe Wasting**  **(r value)** | **Under-weight (r value)** |
| ***Demographic***  Sex ratio  Female Literacy  Women 10th pass  Woman married before 18 | -0.32  -0.22  -0.22  0.54 | -0.32  0.14  0.10  0.12 | -0.21  0.17  0.18  0.01 | 0.28  -0.17  -0.13  0.15 |
| ***Economic***  Per capita income  Below Poverty Line | -0.28  0.09 | -0.01  0.35 | -0.04  0.28 | -0.06  -0.32 |
| ***Mother's Health***  Anaemia  Diarrhoea  Women BMI<18.5 | 0.09  -0.24  0.46 | 0.26  -0.54  0.36 | 0.17  -0.41  0.17 | -0.16  0.30  -0.10 |
| ***Child Health***  Anaemia | -0.01 | -0.04 | 0.21 | 0.12 |
| ***Availability of Health Services***  AWW  ASHA | 0.16  0.02 | -0.28  0.03 | -0.24  0.20 | 0.20  -0.24 |
| ***Utilization of Health Services*** ANC  Four ANC Visits | -0.48  -0.42 | -0.33  -0.09 | -0.23  -0.04 | -0.01  -0.04 |
| ***Child Feeding Practices***  Exclusive Breast Feeding | -0.44 | -0.28 | -0.15 | -0.04 |
| ***Health Facilities***  Water  Clean Fuel | 0.07  -0.32 | -0.45  -0.22 | -0.27  -0.20 | 0.14  0.20 |

**5 Discussion**

If the entire data collected is analyzed as a whole, as one data set containing many variables, with numbers and figures varying over large magnitude, it may end up with erroneous results. For example, certain extreme values (very low or very high) which are present as outliers do not give the correct interpretation on analysis. It is therefore very important to segregate the dataset into clusters with relatively similar attributes, for reaching closer to truth through meaningful analysis.

Inter cluster analysis

The data analysis has been oriented to appreciate this effect. It was therefore decided to compare the correlations obtained in the original information set and the three clusters generated, based on common attributes. This inter-cluster analysis helps the researcher to get a more real effect of different independent variables on the malnutrition indicators.

The effect of inter-cluster analysis is shown in Figure 4. It is observed that for many attributes there is wide variations in the correlations obtained for the original (un-clustered) data set and the two other clusters. For example, the correlation between Sanitation and severe wasting varies starkly for un-clustered data (minimal at r = - 0.19) and for clusters 2 and 3 at r = - 0.90 and r = – 0.05, respectively. This indicates that, had we been interpreting this data merely using the traditional system of analyzing the non-clustered original data, we would have found poor correlation between Sanitation and malnutrition as the Pearson’s Correlation Coefficient is found to only at – 0.19). This indicates that there is negligible association between sanitation and malnutrition in children.

But when the data is segregated into three clusters, the results are starkly different. The 1st cluster of one district (Patna),which had extreme values has been left-out of analysis as it will vitiate the results and give fallacious interpretations. The other two clusters have been defined by the software using certain common attributes. Thus the remaining 35 districts (excluding Patna), have been segregated into two clusters of five and thirty districts. Here the results are totally different. The second cluster of five districts gives a very good Pearson’s correlation coefficient of r = - 0.90. This indicates a very strong association between sanitation and malnutrition.

The third cluster however shows a result with correlation at r = - 0.05. This indicates a negligible correlation for the same- between sanitation and malnutrition.

These varying correlation values have major implications on the prevention and control of malnutrition. In case the correlations are found to be ‘strong’ it indicates that measures are required to be taken against that independent variable (sanitation), and if there is poor/no correlation then the relative importance of that determinant is negligible and it doesn’t really need a modification, in order to control malnutrition. To illustrate the present example, there was a minimal correlation seen between sanitation and malnutrition when the entire (un-clustered) data was analyzed (r= - 0.19). This indicates that since sanitation has minimal effect on malnutrition, no measures are required to be taken to improve malnutrition in the State of Bihar.

However when Cluster 2 was analyzed independently, it indicated a Correlation coefficient of –0.90, which depicts a very strong association, thus highlighting the undisputed importance of sanitation in improvement of malnutrition in this cluster. Hence it is imperative that all measures must be taken to improve sanitation in this cluster of districts, in order to improve malnutrition.

On the other hand in cluster 3, where the r value was found to be minimal, probably that much emphasis on sanitation is not needed.

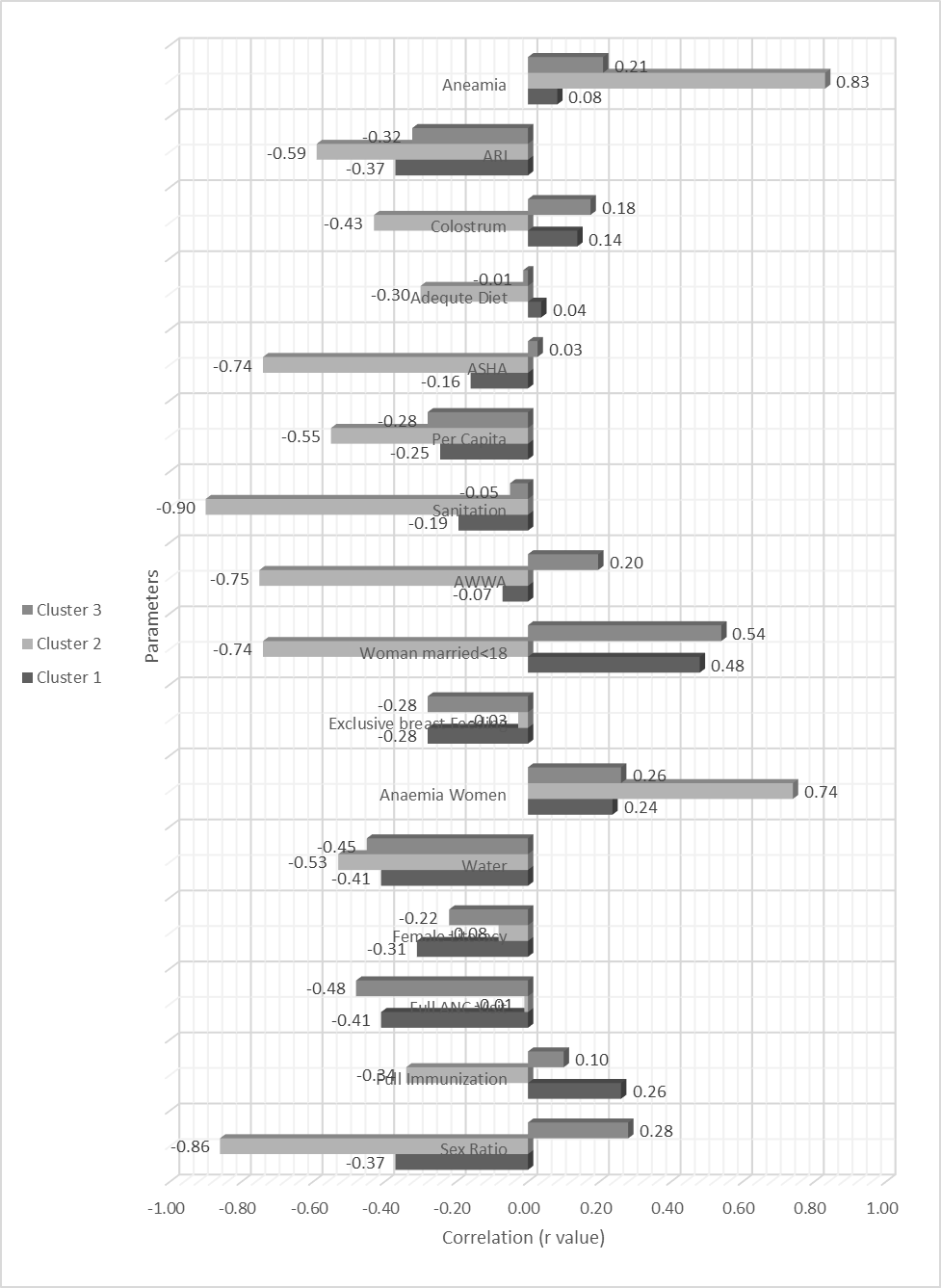
It is inferred from the above discussion that clusterization not only orients the data analysis in the correct direction, but also helps in the most prudent data interpretation, as it is seen in the instant case that the correct data interpretation would lead to correct policy making, for malnutrition control.

**Table 7.** Parameters to target for curbing malnutrition in Bihar.

|  |  |
| --- | --- |
| Cluster with Districts | Parameters to be targeted |
| Cluster 1  (Patna) | Patna, the most urban, economically and socially developed district of Bihar was found to be an outlier with no specific parameters to be targeted. |
| Cluster 2  (Begusarai, Bhagalpur, Munger, Muzzafarpur, Rohtas) | Improving Sanitation and Water source, Adequate diet to infant, Acute Respiratory problems, Maintaining a balance of sex ratio, Preventing anemia in women and children, Increasing Anganwadi workers’ presence. |
| Cluster 3  (Araria, Arwal, Aurangabad, Banka, Bhojpur, Buxer,, Darbhanga, Gaya, Gopalganj, Jamui, Jehanabad, Kaimur, Kathiar, Khagaria, Kishanganj Lahisarai, Madhenpura, Madhubani, Nalanda, Nawada Pash, Champaran, Purb, Champaran, Purnia, Saharsa, Samastipur, Saran, Shikhpura, Sheohar, Siwan, Sitamarhi) | Improving Clean fuel availability, water facilities, breastfeeding to be continued for at least the first six months of infant’s birth for better health, curbing diarrhea, providing Adequate diet for women and awareness should be spread for women to get married only after 18. |

Table 7 mentions the clusters, the districts in them and the parameters to target for curbing malnutrition in Bihar. This was the result obtained after the clusters were formed, dividing Bihar into 3 clusters based on similarity in attributes.

Different parameters have thus been identified from the analysis above, and special emphasis can be given to these specifically instead of targeting all and giving them equal importance, which has not been very successful up till now.



**Fig. 4.** Inter-cluster Variability in Correlation Coefficients

**Conclusion**

The purpose of data collection and its statistical analysis is to reach (closest to) the truth. While there might be many software available for basic data analysis, each one has its own attributes and pros and cons. Analysis of the big data set as a whole has the limitation of having extremes in the data, which produces fallacious results on analysis, thus deviating from the truth. The technique of clustering of a big data set into smaller groups based on common attributes and their subsequent analysis helps overcome this limitation. In the present study clustering enabled us to understand the real correlations between various independent variables and malnutrition. This would help the policy makers’ shortlist the most appropriate factors contributing to malnutrition and take corrective action. Thus this technique of data mining is unique and most appropriate in the given setting.

**Conflict of Interest**: Nil

**References**

1. Gupta A, Gupta SK, Baridylne n. Integrated Child Development Scheme (ICDS): A journey of 37 years. Indian J Community Health. Vol 25 (1) (2013).

2. Report of the Comptroller Auditor General on Performance audit of ICDS. Report No. 22,. New Delhi (2013)

3. National Family Health Survey 4, Government of India, http://rchiips.org/nfhs/nfhs4.shtml

4. Anwesha Mal, Prof (Dr)Bebo White “Analysis and Clustering of PingER Network Data” IEEE Xplore Digital Library, DOI: 10.1109/CONFLUENCE.2016.7508127 (2016)

5. Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." ICML. Vol. 1. (2001)

6. Chandan JS. Statistics for business and economics. Vikas Publishing House Pvt Ltd. New Delhi. (2003)

7. McCallum, Andrew, Kamal Nigam, and Lyle H. Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching."Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2000)

8. Manpreet kaur, Usvir Kaur, “Comparison between K-mean and Hierarchical Algorithm Using Query Redirection”, Volume 3, Issue 7, July (2013)

9. Jian Hua Yeh, Fei Jie Joung, Jia Chi Lin, “CDV Index: A Validity Index for Better Clustering Quality Measurement”, Journal of Computer and Communications, 2014, 2, 163-171 (2014)

10. Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." Expert systems with applications 36.2 2592-2602. (2009)

11. Jyoti Yadav, Monika Sharma, “A Review of K-means Algorithm”, International Journal of Engineering Trends and Technology, Volume4 (2013)