A Survey of Requirements of Multivariate Data and its Visualizations for Analysis of Child Malnutrition in India

Jaya Sreevalsan-Nair

Abstract

Multivariate data has been routinely used in the form of feature vectors for identifying/determining, assessing, and analyzing child malnutrition. Using machine learning techniques on such data for eventually deriving social health determinants has been a recent methodology used in concerned research communities. However, data visualization has been grossly underused in this regard, for studying trends and further analysis of the data. Visualization of multivariate data in identifying child malnutrition enables exploration of the data, and summarizes its multidimensional analysis. Even though the multivariate data used for data analytics across various studies, comprise of common variables/parameters predominantly, there are slight variations depending on the purpose of the research study and the control population used. Here, we enumerate all the parameters used in these studies into an exhaustive list of variables, and propose using this list as an*unified* feature vector, which can be used for holistic analysis of the problem of malnutrition. Thus, in this work, we survey existing literature to find and propose the following: (a) a *unified* feature vector, which is a subject-wise data object, derived across several research studies in identifying, assessing, and analyzing child malnutrition, and (b) in the existence of dataset/collection of unified feature vectors across a population, the visualization tasks that can be performed on it, in lines of the five W's of knowledge discovery in journalistic reporting, and (c) the appropriate set of visualization techniques to accomplish those tasks. We further propose the design of the visual analytic framework for such an analysis of dataset.

Keywords: Visual analytics; multivariate feature vector; time series; child malnutrition

I. Introduction

Malnutrition/malnourishment in children under 5 years in developing countries has been a well-researched topic; and it has been identified as an endemic problem in such countries, which includes India. Several reports have stated that India has a larger burden of child malnutrition, in comparison to even relatively less developed sub-Saharan countries, as per an article in the Economist in 2010; where it states that "more than one-third of the world's 150 million malnourished under-fives live in India." Fast forward to 2015, on a positive note, the Global Nutrition Report 2015 [15] has stated that nearly all states in India have shown significant decline in stunting rates in 2006-2014. The progress been attributed to increase in exclusive breastfeeding rates across the nation. However, the progress has been slow and uneven, and the decline in wasting rates is variable across states. In order to understand the current state of affairs as well as plan for the future, we propose the use of visual analytics for a holistic analysis of the burden of child malnutrition in India.

Several research studies have used multivariate data to determine child malnutrition, in the form of feature vectors. These studies have used anthropometric measurements of the child to determine the condition of malnutrition, as the standard classification is based on the z-scores of height-for-age (HAZ), weight-for-age (WAZ), and weight-for-height (WHZ). For values less than -2, HAZ, WAZ, and WHZ indicate stunting, underweight, and wasting conditions, respectively. The extent of malnutrition has been determined using mid-upper arm circumference (MUAC), where low values indicate severely acute malnourishment [31]. In several studies, the causes for child malnutrition have been additionally determined by using a set of variables, which were categorized as household and community variables [3], [23]. In relation to paradigm shifts in global nutrition [23], [12], undernutrition caused by deficiency of micronutrients (iron, vitamin A, iodine) is relevant to the developing countries, and hence tracking the deficiencies in a population is relevant. Bivariate analyses [3], [23] have been used extensively on variants of similar feature vectors. Summarizing, these studies have focused on various aspects of malnutrition.

In a panel in the Global Nutrition Report 2015 [15], John and Menon have done a review of the nutrition surveys done in India and stated that the conductance of the fourteen major surveys and data collection have several flaws, which reduce the usefulness of the data collected. The flaws pertain to six components: *geographic coverage, content, comparability, frequency, ownership, and financing*. To address concerns in the gaps in these components, their recommendations include narrowing down a set of core nutrition outcome indicators and arranging for reasonably frequent data gathering. However, within the set of core variables, several of these studies are based on one or few of these variables, depending on region and time of collection. Now, how do we find the core set of variables, from all a dataset collected in a given scenario ? To answer this question, we propose a holistic study of the problem of malnutrition using a multivariate analysis using exhaustive list of variables. We hypothesize that such a multivariate analysis can reveal more hidden patterns and give more insight to a population-specific study, which takes into consideration the socio-economic and cultural context. Hence, we propose combining all the indicators into a *unified* feature vector for identifying, assessing and analyzing child malnutrition and its remedial programs. We propose using visualization of the multivariate data as the primary tool for its analysis, where interactive visualizations of such data collections of outcome indicators help in narrowing down the set of core variables for specific analysis of the data.

Visual analytics is a class of techniques using data visualizations leading to and/or resulting from data analytics. Its applicability is rapidly gaining popularity in healthcare. Such analytics leads to knowledge discovery from raw data, which can

This document is preprint, as on October 17, 2016. The authors are with Graphics-Visualization-Computing Lab, IIIT Bangalore. email: jnair@iiitb.ac.in

be further enhanced by using derived data models. For instance, hierarchical data models are derived from grouping data based on spatio-temporal characteristics of raw data (e.g. grouping based on geographic region) or from hierarchical clustering. Visual analytics additionally involves appropriate graphical user interface (GUI) design for visualizing either raw data or derived data models. Visual analytics is a relevant tool in healthcare informatics as visualization can aid domain experts in identifying patterns which in turn will help in determining the kind of analysis to be done on the data. Additionally, visualization of the data can be used for information dissemination to the public and policy makers, which helps in sensitizing the concerned stake holders of the state of the problem and progress of viable solutions.

Time-series analysis of such dataset gives more insight to the time-varying trends. While the improvement in mitigation of child malnutrition issue using remedial programmes can be statistically derived, we propose that visualizing these changes over time across various determinants of child malnutrition will give some insight. Such an insight also provides overall improvement in social health indicators in India. We found the motivation for time-series analysis from two different studies. Firstly, Panagriya [25] has combatted the conclusion on the slow progress of India in mitigating child malnutrition, showing that it has been based on flawed methodology of assessing child malnutrition. He has cited evidence based on several detailed comparisons in subsets of the data, including comparing Kerala to Senegal and Mauritania, where the infant and maternal mortality (16 per 1000 live births, and 95 per million live births, respectively) are much lower compared to Senegal (93 and 410, respectively), and Mauritania (117 and 550, respectively), but contrarily, the %age of stunted and underweight children are much higher in Kerala in comparison to Senegal and Mauritania. While Panagriya has argued that the data analysis by the World Health Organization (WHO) is done using its standardized charts [35], the reference data has not been adjusted to reflect the incremental growth that happens over time when the diet of afflicted population is corrected to a nutritional one. Visualizations of these different "slices" of the data can bring out the findings made by Panagriya. Secondly, in the Global Nutrition Report 2015 [15], Blankenship has emphasized on the need for monitoring nutrition remedial programs in countries, showcasing the case-studies in thirteen countries in Africa. However, the statistics of only current data is predominantly published. We hypothesize that a subject-wise analysis of these programs, which can provide better understanding of the efficacy of such programs. Thus, in summary, visual analytics enables viewing subject-wise trends over time, and inferring the progress [3], [23].

The scope of our work is to provide a state-of-the-art report of the data required for determining, assessing, and analyzing the burden of child malnutrition in India and the various remedial programs being implemented, with a proposal of using visual analytics on the data for specific tasks. We survey and analyze relevant literature in arriving at the exhaustive feature vector, the *unified* feature vector, leading to multivariate data, describing the state of child malnutrition in India. We hypothesize that subsets of parameters in such a feature vector are required for determining, assessing and analyzing child malnutrition in India. Depending on the aspect of the problem being studied, the subset will vary, owing to which we support collecting and curating such subject-wise datasets based on the unified feature vector. We identify suitable visualization tasks to be done on such a dataset; and the appropriate visualization and visual analytic techniques which will enable accomplishing those tasks.

II. Related Work

There is a rich body of literature on multivariate visualization. We list out the most relevant ones, where such visualizations have been used for nutrition data. We also mention some of the existing work has also inspired design choices for our proposed visualizations.

A. Visualizations for Nutrition Issues

Pachauri et al. [24] have developed the Ananya data visualization tool which is for inferring knowledge and enabling analytical reasoning from complex data collected from large social programs, such as Ananya. Ananya website[1] shows effective visualization of health statistics related to maternal and child healthcare in Bihar, a state in India. The visualization application showcases (a) statistical information on progress in public health remedial programs, (b) availability of services on the program, and (c) region-wise nutrition and health indicators. The tool uses linked GIS application from Google Maps for (b) and (c); and uses relevant visual metaphors to represent characteristics of remedial programs (e.g. texture of a pregnant woman to indicate maternal health, and similar icons for child nutrition and child weighing), for (a). Our proposed visual analytics solution is similar to Ananya in motivation, albeit is different in the following aspects: (i) ours uses subject-wise data, instead of statistical data, and (ii) ours allows the user to explore data further, such as interactive data analytics, instead of just summary visualizations of the statistical aggregates. Global Nutrition Report 2015 [15] has an accompanied online data visualization platform [2] which shows the country-wise trends in public health nutrition problems, such as under-five stunting, anemia in women of reproductive age (WRA), and overweight issues in adults.

Klimov et al. [18] and Hinz et al. [13] have used multivariate visualization techniques such as parallel coordinates and parallel sets for analyzing effects of medications for diabetes and hemoglobin A1c levels, respectively, for patient data. Borland et al. [2] have shown the use of radial coordinates, which is radial layout of parallel coordinates, to visualizing several aspects of primary care trust in United Kingdom's National Health Service. Chen et al. [4] have used cluster heat map to show the

[1] http://www.ananya.org.in

[2] http://globalnutritionreport.org/the-data/data-and-visualization-platform/

variations in datasets in online communities dedicated to e-cigarettes and hookah versus contextual dimensions on healthy behavior. They have effectively shown popularity of certain topics in different fora. Dabek et al. [7] have used force-directed graph visualization to study trends in longitudinal clinical trajectories of traumatic brain injury (TBI), which is similar to our proposed visual analytics in our case.

In short, several of the visualization techniques used here are static and more with a goal of summarizing. While we propose the use of standard visualization techniques, such as parallel coordinates and scatterplot matrix, we additionally propose their use in a visual analytics workflow so that there is more knowledge discovery than what can be achieved from just a single visualization.

B. Visual Analytics in Healthcare

In general, the trend in research in visual analytics in healthcare has moved from *visualization of raw data* to using *visualization for knowledge discovery*, in the recent past [18]. In this section, we focus on a few works which are relevant in terms of design principles.

a) **Electronic Medical/Health Records:** Zhang et al. [37] have proposed considering the five W's *(who, when, what, where, why)* in visual information displays of a patient's EMR (electronic medical record) within healthcare informatics application. The five W's is a concept inspired from information gathering in journalistic reporting. Here, the *who* referred to the patient, the *where*, the location in the patient's body, and the *(when, what, why)* pertained to the reasoning behind the health condition. The patient is visually represented as a sunburst radial visualization with a body map. A sequential display, similar to parallel sets [19], represent the logical reasoning chain. One of the usecases of the visualization involved collaborative diagnosis between several doctors with different expertise. We have proposed our visual analytics design in lines of the five W's, however the design differences arise from our usecase in public health informatics as opposed to personal medical records in [37].

VisuExplore, a visual exploratory tool for patient medical database, proposed by Rind et al. [30], and evaluated by Pohl et al. [27]. VisuExplore visualizes several medical parameters (multivariate), and time-varying data. It can be used to compare data between multiple patients. It uses standard visual techniques, such as line plots, which is similar to the design proposition in our work. Chetta et al. [5] have integrated several visual interfaces which enable nurses to see the predictive outcome of classifying a patient based on historical EHR (electronic health records) and make decisions based on the predictive outcome and summary of other nurses' interpretation of patient activities.

While we have adopted the principles of visual analytics as found in the aforementioned visual analytics frameworks, our proposal is specific to understanding assessment and monitoring of child malnutrition. We have not seen the use of visual analytics in the context of child malnutrition so far.

b) **Public Health Data:** The targeted audience for patient medical database visualization [37], [30] are specifically, the physicians and patient, whereas public health data visualization is for a larger audience, such as the policy makers, social scientists, domain experts, and the public. Our propositions are in lines of the recommendations made by Shneiderman et al. [32] on the role of visualization in enabling public health tracking/monitoring systems. Sopan et al. [34] have proposed community health map, which has geospatial and multivariate data visualization components, for public health datasets, which we have adopted in our proposal. The community health map is a combination of geographical maps, charts, and tables; whereas our proposal is a combination of parallel coordinates, cluster drift, embedded geographical maps, scatterplot matrix, and heatmap.

III. Child Malnutrition Data

In this section, we describe the descriptors or variables used across studies on malnutrition in children under 5 years of age. The relevance of the variables has been found to vary region-to-region [33], and to be heavily influenced by the purpose of the study [3], [10]. We have found several studies on the topic of child malnutrition, which are done using population study in India, as well as other developing nations, such as Kenya, Senegal, and Mauritania [23], [25]; and based on global standards [35], [9]. We mention many of the findings mentioned, which we propose to include in our proposed unified feature vector (Section V).

A. Research Findings on Descriptors

Here, we present the findings from research literature on different classes of descriptors used for analyzing child malnutrition.

c) **Anthropometric Measurements:** Blössner et al. [1] have focused in studying mortality and morbidity of child malnutrition, which is measured using anthropometry, biochemical indicators and clinical signs of malnutrition. They have stated that while anthropometric measurements are sensitive across variations within malnutrition (stunting, wasting, underweight) and can be used for identifying the condition, the biochemical and clinical indicators can be used to assess the chronicity of condition, once it is established that the child is at least moderately malnourished. Since the short-term response of a child's body to inadequate nutrition is stopped or delayed growth, the measures such as HAZ, WAZ, and WHZ, are good indicators of the condition. To consider for the diversity in the ecosystem in which the child exists, statistical remedies such as z-scores or

percentiles are used [1]. Blössner et al. have additionally suggested the use of weight of infant and his/her gestational age in completed weeks, which is to be recorded at birth, in order to determine the presence and effects of low birth weight (LBW).

Roberfroid et al. [31] have shown that low MUAC $<$ 115mm has been used as a stand-alone anthropometric measure for determining severely acute malnourished children alongside low WHZ $<$ -3. Even though there is insufficient evidence of using MUAC as it being an effective stand-alone measure, we hypothesize it can be used alongside with other anthropometric parameters for change detection.

d) **Socio-Economic Factors:** Som et al. [33] have studied the effects of the same set of socio-economic conditions in the states of West Bengal and Assam, and arrived at the differences in importance of several variables used for determining malnutrition in children under the age of 3. These differences arise from socio-economic, demographic, and cultural factors, predominantly, which is not captured in the anthropometric data. For instance, increase in maternal illiteracy would increase malnutrition in children in West Bengal, but was found to not have much impact in Assam.

Radhakrishna et al. [28] have shown that, contrary to the general perception of reduction of malnutrition with increase in household income, lower incidence of child malnutrition occurs in middle income states where the government has progressive social policy, as opposed to richer states.

e) **Classification of Descriptors:** Studies on causes for child nutrition predominantly indicate family- and community-based influences. Maternal socio-economic and health conditions have been proven to be a strong influencer of child nutrition. Borooah [3] has shown that proximate literacy, where the father is literate and mother is not, does not alleviate the risk of malnutrition, and can be considered similar to maternal illiteracy. Borooah has performed unit-record analysis in Indian population, and Masibo [23] has performed a similar analysis in Kenya using four surveys. Both the works have given clear classification of parameters/variables that can be used for assessing as well as analyzing causes of child malnutrition. The classification given on unit records of child nutrition data, common to both works, is the following:

- Dependent/anthropometric variables: include z-scores such as HAZ, WAZ, and WHZ. Masibo [23] has additionally used BMI (Body Mass Index)-for-age z-scores (BAZ), based on WHO 2010 recommendations. BAZ serves the same purpose as WHZ, however, in cases of non-linear fitting of data, BAZ will fit better than WHZ.
- Child characteristics: include age, sex, birth order, size at birth. Masibo has additionally included presence of diarrhea, fever, or cough, two weeks preceding the survey, which was found useful for determining child morbidity and mortality. Masibo has analyzed dataset of children under 5 years, however the age-group of the test group used by Borooah has not been specified.

  Borooah [3] has used data on the quality and duration of nutrition in the child's first year with respect to breastfeeding and introduction of solids. Gupta et al. [10] have attested that intervention programs, which ensure adequate and timely introduction of complementary feeding with breastfeeding for infants of age 6-9 months and promoting awareness amongst mothers on exclusive breastfeeding in the first six months of life, can reduce stunting in the nation. They have stated that low rate of exclusive breastfeeding and sufficient complementary feeding have been found prevalent in urban and rural India. The Global Nutrition Report 2015 [15] has reported that the rates of exclusive breastfeeding has increased across India, and hence has significantly brought down stunting rates in the country. Hence the data on the duration of exclusive breastfeeding given to a child in his/her early years and the quality of complementary feeding beyond 6 months of life is significant in analyzing malnutrition.
- Maternal or household characteristics: include maternal age, child-bearing pattern (e.g. average spacing between children, or equivalently number of children under 5 the mother is caring for), birth order of the child. Masibo has documented significance of maternal nutrition using BMI, and socioeconomic status of the household by using maternal education and maternal work status. Borooah has used measures of food stocks in the household, the household income status with respect to above or below poverty line, and the cooking conditions (such as use of chulha, ventilation, etc.). Empowerment of women in the household is a determinant for preventing child malnutrition has been determined using gender of the household head, in both studies, and Borooah has used the prevalence of access of women to information as an additional determinant.
- Background or community characteristics: include urban vs. rural residence of the child, state/province, wealth index, source of drinking water, and toilet facilities. Borooah has specifically used a test population from rural areas, and has additionally used access to hospitals and mother-child welfare centers (anganwaadis) and food security (based on public distribution system) as community characteristics.

Markos et al. [22] have specifically used child's vaccination record, and level of anemia, as individual characteristics, and maternal age and maternal occupation, as household characteristics, when building a predictive model for determining undernutrition in under-five children using data mining, on demographic data in Ethiopia.

f) **Nutritional Abnormalities:** Masibo [23] has documented the prevalence of dual burden of malnutrition in Kenya, where trends indicated children of overweight mothers being stunted. The coexistence of two or three forms of nutritional abnormalities (i.e. overnutrition, undernutrition, micronutrient deficiencies) in a single household has been termed as dual- or triple-burden of malnutrition, which is a recently found phenomenon worldwide, and hence has been considered to be a nutritional paradigm shift [12]. Herforth has stated that the triple burden of malnutrition has been found to be highly prevalent in countries with "high stunting burden", which includes India. There have been several policies undertaken worldwide to combat obesity (a form of overnutrition) and noncommunicable diseases (includes micronutrient deficiencies), and, to a lesser

extent, malnutrition (a form of undernutrition) [15]. However the coexistence of different forms of nutritional discrepancies in households is understudied. Hence, we propose that our unified feature vector should include information which can reveal the dual- or triple-burden scenarios in households.

B. Datasets in Indian Population

Rajan et al. [29] have elaborated about the dubitable quality of data collected in the third National Family Health Survey (NFHS-3) done in 2005-06, which is the last nationwide survey and which indicates worsened nutritional status in children. They have found stark differences in siblings of the same gender in a single household, where they had deliberately kept gender inequality, within a household, out of the picture. They have proposed that quality control is critical for such nationwide surveys.

Nutritional status of women in reproductive age (15-44 years) has been considered as a critical social determinant of malnutrition, as conditions such as LBW and Intra-uterine Grow-th Retardation (IUGR) can be predicted in infants using the data on maternal nutritional status and weight gain during pregnancy [1]. The data for women of all age-groups is necessary to spread awareness on preparing for pregnancies and antenatal care. The NFHS-3 has evaluated this for all women, instead of just focusing on mothers as was done in NFHS-1 and NFHS-2 [29].

Notwithstanding the quality, we have found that the raw data for the variables that we propose to include in the unified feature vector (in Section V) exists in NFHS-3, except for the data on dual- and triple- burden in households.

g) **Nature of Data for Nutritional Abnormalities:** The Food and Agriculture Organization (FAO) [9] of the United Nations define prevalence of undernourishment (PoU) indicator using a probability density function (pdf) of per capita calorie consumption and minimum dietary energy requirement (MDER). The pdf and MDER are measured for a representative individual of the population. We propose that calorie consumption of each of the members in a household of a subject must be collated in our unified feature vector for child malnutrition. The rationale is that, even though the subject-wise calorie consumption data, for children and adults, is available in NFHS-3 [16], we have found no explicit studies on dual- or triple-burden of malnutrition in India, owing to it being a recent phenomenon globally [12]. Thus, we propose including prevalence and extent of each of the three forms of nutritional abnormalities (overnutrition, malnutrition, micronutrient deficiency) in the household of a subject and using these measures as household characteristics in our proposed unified feature vector.

IV. Proposal for Visual Analytics

The existing studies of analysis of nutrition data have been predominantly based on statistics, and recent developments lean towards machine learning [22]. We are interested in taking a new perspective on the data analysis, which is to understand the multivariate model of the factors involved in identifying, monitoring, and predicting malnutrition in children using visual analytics. Our hypothesis is that the multivariate visualization of a feature vector for a population, from a *subject-centric* dataset and which *unifies the purposes of identifying, assessing, analyzing and monitoring*, will reveal holistic trends and patterns in the population.

We propose data analysis using visualization of subject-wise or record-wise trends as opposed to trends of representative values obtained from statistical analysis of the data. The latter is sufficient when there is uniform behavior in the data, which is usually not the case, owing to which, the former will give a more accurate analysis of trends in the entire population. Visual analytics is a class of techniques of analyzing data where visualization complements traditional data analytics, such as statistical analyses, machine learning, and data mining. The need for using visual analytics is two-fold: (a) it enables data exploration with human-in-the-loop, and (b) it gives visual summary of trends and causes, in the form of either intermediate or final outcomes. The human-in-the-loop enables a wider range of usability of an analytic tool to serve different target audiences, such as social scientists, nutritionists, policy makers, and the public.

We have identified the following outcomes (O1-O3) which can be achieved using visual analytics of the child malnutrition data:

O1: identification of patterns in the data, which can further be applied to new streaming data (modeling),
O2: Analysis of causes for the patterns (causation),
O3: Tracking of changes over time (monitoring).

h) **Identifying patterns:** In the context of child malnutrition data, we can visualize the membership of a subject to the category of malnutrition (stunting, wasted, underweight) or well-nourished, by looking simultaneously for different anthropometric measurements, where the data points are falling with respect to the prescribed thresholds. This categorization based on visual assessment can be done for both static as well as streaming data. These patterns can help in finding appropriate models for the data, say for children in same region, socio-economic conditions, or age-group. The patterns help in *identifying and assessing* the extent of nutritional problems in India.

i) **Finding causes for the patterns:** We can identify the reasons for the children to fall under one of the categories in the malnutrition by looking at the similarity in trends of children from the same region, socio-economic class, and/or age-group, where these trends are determined by observing the household, child, and community characteristics, e.g. how socioeconomic factors have been proven to make the difference between West Bengal and Assam [33]. For instance, visualization can show how progressiveness in government policies or women empowerment influences nutrition metrics in certain regions. These are trends not available directly from the raw data, but derived from it. We propose using this aspect to identify the extent of dual-

and triple-burden of malnutrition in causing child malnutrition. We hypothesize visual analytics will allow finding appropriate subsets of parameters which influence specific regions, and thus help in *analyzing child malnutrition in India*.

j) **Tracking changes over time:** Visualization of time-series subject-wise data shows how a certain subject who has been categorized as malnourished has responded over time upon receiving remedial measures. Similar studies have been done on a test population in the past [25]. A visualization, such as an animation, enables policymakers and domain experts to monitor success of remedial programs. Such analysis of remedial programs helps in better *assessment and monitoring* of the problem of child malnutrition in India. Time-series data and its visualization show the status of the issue of malnutrition over time, as well as the factors, which would have facilitated static, positive or negative trends of progress.

V. Proposed Data & Its Analysis

The goal of our study is to consider data in the Indian scenario, after incorporating changing global trends, and arrive at the most optimum *unified* feature vector. Our rationale is that our proposed unified feature vector exhaustively covers all the relevant attributes of all aspects of nutrition issues (Section III) and provides relevant data for achieving the outcomes (O1-O3) of our proposed visual analytics (Section IV). We further propose specific visualization tasks and techniques for achieving outcomes O1-O3.

A. Multivariate Nutrition Data

Our proposed unified feature vector is the multivariate data required for determining, assessing, and analyzing the incidence of malnutrition in children under five in India. *As can be seen here, the nutrition data is not pertaining to "nutrition" of child alone, but also encapsulates various conditions that influence the nutritional intake of the child.* Here, we enumerate 46 variables/parameters which we propose to use in our nutrition data analysis[3]:

1) **Anthropometric/dependent characteristics** (5 variables): BAZ, WHZ, WAZ, HAZ, MUAC,
2) Child characteristics (15 variables): gender, age, square of age, size at birth, age when solid food was introduced, age in months until when exclusively breastfed, prevalence and quality of complementary feeding, prevalence of edema, color changes in hair and skin, prevalence of diarrhea, fever, or cough preceding the survey, prevalence of anemia, weight at birth, gestational age in completed weeks at birth, vaccination record being on track,
3) **Maternal/household characteristics** (17 variables): maternal literacy level (illiterate, proximate literate, literate), maternal BMI, maternal work status, access of the mother to antenatal care, optimal maternal nutrition during the pregnancy, access of the mother to postnatal care, average spacing between maternal pregnancies, birth order of the child, gender of the household head, prevalence of empowerment of women in the household with respect to access to information and general awareness, economic status of the household (above or below poverty line), prevalence of access of the household to a fair-price shop, existence and quality of sanitation facilities in the household, quality of housing conditions (kind of stove, use of chulha, ventilation, etc.), prevalence and extent of overnutrition in the household, prevalence and extent of undernutrition in the household, prevalence and extent of micronutrient deficiencies in the household,
4) **Community characteristics** (5 variables): urban or rural residence, access of the household and neighborhood to safe drinking water, access of the household and neighborhood to hospital or substation of the household, prevalence of anganwaadis (early childhood centers in villages) in the community, prevalence of progressive social policy of the state government,
5) **Biochemical and micronutritional characteristics** (4 variables): prevalence of vitamin A deficiency, prevalence of iron deficiency, prevalence of iodine deficiency, serum albumin level.

The proposed unified feature vector is an exhaustive unit record for a single subject under five years of age, just as in [3]. Such records for a test population, across different districts in different states in India, yield the ideal dataset. The dataset can be prepared on a nationwide sample population, similar to the NFHS [29]. The proposed dataset is **time-series of multivariate data**, which is needed for studying progression of the issue and its remedial programmes.

B. Visualization Tasks

As explained in Section IV, visualization complements machine learning and statistical analysis of the multivariate data of child malnutrition. The five W's (who, when, what, where, why), inspired from journalistic reporting, have used in knowledge discovery in healthcare informatics [37]. We use the five W's in the following exploratory tasks:

Task1: (what) visually summarize or analyze the incidence and extent of malnutrition,
Task2: (where) subselect regions using geographic maps to study patterns of nutrition in children there,
Task3: (why) subselect or analyze variables, thus visually identify plausible causes for malnutrition using trends in a population and their clustering patterns, observed from correlation between variables,
Task4: (who) subselect subjects to be explored for (what), (where), (why), and (when) – which enables understanding: (a) what conditions prevail for selected subjects, (b) if the region they come from has an influence in their conditions, (c) the causes for their conditions, and (d) how their conditions have changed over time.
Task5: (when) subselect interval in the timeline for tracking changes in selected subjects or selected clusters.

[3] Since the writing of this paper, we have used this exhaustive list to get data from a research organization, for a subset of variables. The variables, for which we have data for, form a logical category of information on child and maternal healthcare. The data has been collated from relevant data gathering organizations.

C. Proposed Visual Analytic Framework

In this section, we deliberate the design of visual analytic framework for the dataset proposed in Section V-A. The framework includes visual analytic processes to support the visualization tasks and a user interface to combine the processes. We propose the following visual analytic processes:

P1: Deriving data models from the dataset, such as, (a) hierarchical model using natural grouping based on spatial or temporal parameters of the dataset, and (b) data matrix, which is a rectangular matrix, with subjects along the rows and variables along columns.

P2: Performing clustering in the dataset using (a) partition methods such as k-means, Expectation-Maximization (E-M), fuzzy c-means, etc. and visualize the clusters, and (b) outcome of various seriation techniques on the data matrix.

P3: Finding correlation between subsets of parameters in the unified feature vector, as per the visualization task, using popular multivariate visualization techniques.

We propose the use of juxtaposed views of the proposed visualization techniques [17] in the visual analytics tool. Juxtaposed views are composed of multiple visualizations placed side-by-side, with implicit linking, i.e. synchronization updates across visualizations in the view. We propose linked visualizations for *brushing*, i.e., to select subsets of data of interest (subselection), which can be based on spatial or temporal parameters, or for selecting specific clusters of data points formed from the clustering of the data. The rationale behind choosing juxtaposed view is due to its simplicity which complies well with the complexity of the subject-wise nutrition data.

D. Proposed Visualization Techniques

We identify the following visualization techniques for implementing the processes proposed in Section V-C: (a) parallel coordinates plot [11], [14], (b) scatterplot matrix [8], [6], (c) cluster heat map [36], and (d) parallel sets-like representation of clusters [26]. Figure 1 shows samples of identified visualization techniques.

Additionally, to cater to the (where) and (when) in the five W's of knowledge discovery of the dataset, we propose including a georeferential map and a timeline in the linked views, respectively, similar to the modifiable maps with controls, proposed by MacEachren et al. [21] (shown in Figure 2). The map and timeline can be used for both (a) referencing spatio-temporal characteristics of data being visualized; or (b) for subselecting data in regions or time-intervals of interest, respectively.

k) **Parallel coordinates plot (for (what),(why),(who)):** Parallel coordinates is a multidimensional coordinate system representation where linearly independent variables are represented as parallel axes, and a point in a n-dimensional plane is represented as a polyline connecting points on each axis, equivalent to its magnitude. While there are standard anthropometric measures to identify malnutrition in a subject, the values in the different parameters BAZ, HAZ, WAZ, WHZ, and MUAC, can summarize the trends for a population pertaining to a region and/or socioeconomic group. Depending on the values of these parameters, the parallel coordinates plot can inform the user of the extent of malnutrition in a subject, e.g. cases of severe acute malnutrition. Since parallel coordinates plot indicates subject-wise data, one can subselect subjects using brushing.

As shown in Figure 3, the correlation between two variables/parameters can be identified using commonly observed patterns in parallel coordinates. Users can explore these patterns by reordering the axes, e.g. when axes corresponding to quality of complementary feeding, maternal literacy, and HAZ, are reordered together in the parallel coordinates, one can notice the pattern for positive correlation. The correlation has been found in different works: (a) Gupta et al. [10] have found that improved complementary feeding reduces anemia and stunting and (b) Borooah [3] has found positive correlation between HAZ and maternal literacy level.

l) **Scatterplot matrices (for (why), (who)):** Scatterplots show bivariate trends, and a scatterplot matrix is a symmetric matrix layout of the scatterplots, where n rows and n columns correspond to n-dimensions of the data. Thus the diagonal elements of the matrix show auto-correlation scatterplots, and the non-diagonal plots show correlation between any pairwise combination of variables. Similar to the case of positive correlation that can be inferred from reordering of parallel coordinates plot, it can be done for pairwise variables using scatterplot, e.g. positive correlation is expected from HAZ and maternal literacy [3]. Multivariate analysis of prevalence of different nutritional abnormalities can be used to identify dual- or triple-burden of malnutrition in the households. Scatterplot matrices can effectively isolate outliers.

m) **Cluster heat map (for (who), (what)):** Two-mode matrix is used for representing multivariate data. The two modes are the subjects and the variables in the unified feature vector, along rows and columns of the two-mode matrix. Seriation of such a rectangular matrix can showcase patterns based on subjects chosen [20], where basis of selection could be parameters, such as, age, region, and socio-economic class. Performing seriation on two-mode matrix involves performing SVD (singular value decomposition). Cluster heat map of the matrix can showcase clustering patterns in the subjects in the data [36]. Cluster heat map includes a dendrogram which indicates permutation of objects in one of the two modes (row- or column-mode). Dendrogram allows the grouping at each level of the hierarchical clustering tree, thus adding providing more information about the trends in the light of the hierarchical model.

For instance, using hierarchical model which is built with the last level of the tree represents the household level, one can visualize the anthropometric measurements and calorie consumption in the data matrix as cluster heat map. Such a visualization will enable finding occurrence of dual- or triple-burden of malnutrition in the household, which can be considered to be a cause for stunting in children in the household [23].

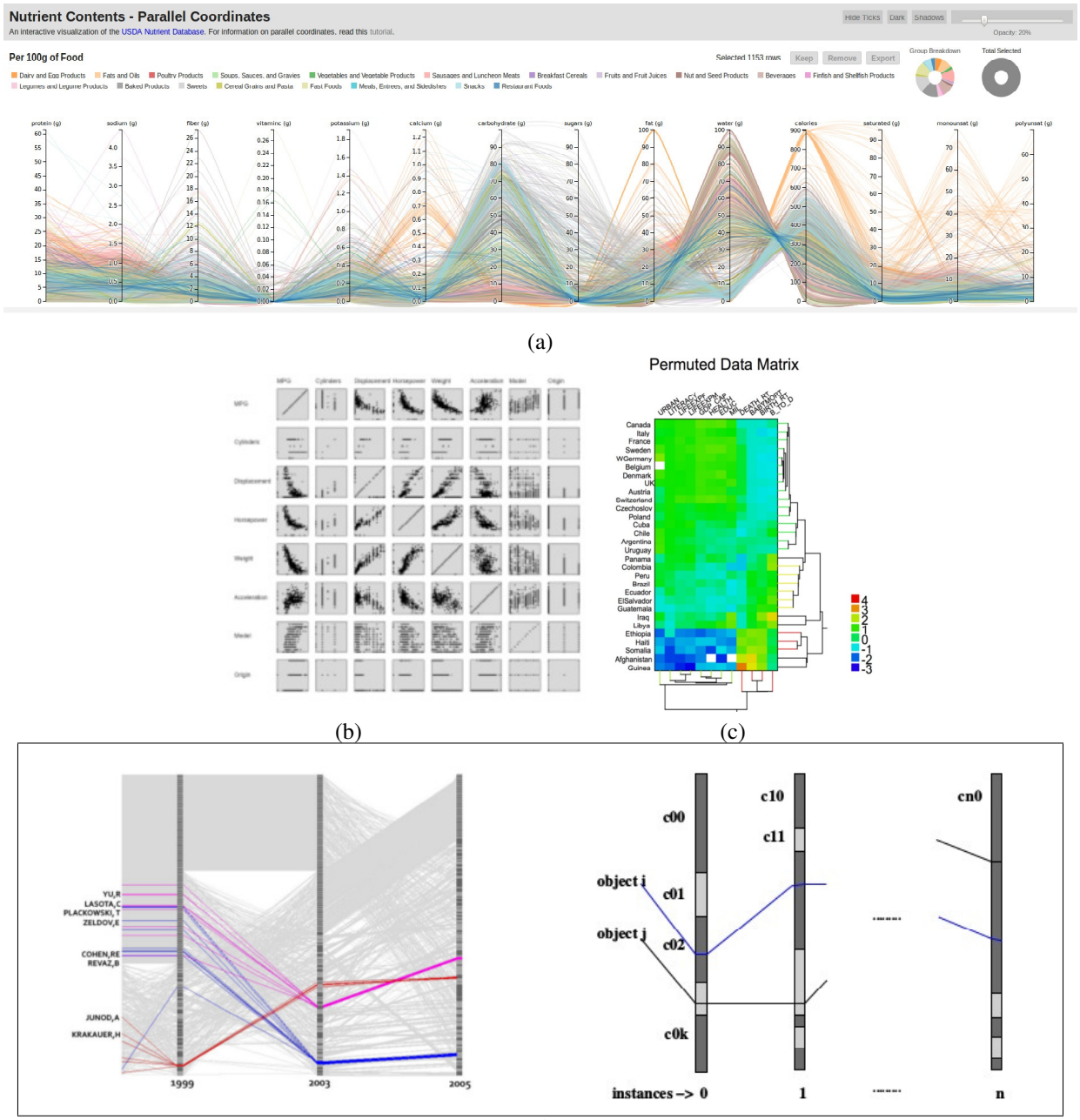

Fig. 1. Proposed techniques to be used for multivariate visualization: (a) Parallel coordinates plot, used for analyzing nutrients in different foods (http://exposedata.com/parallel/); (b) Scatterplot matrices [8] showing 7-dimensional car dataset; (c) Cluster heat map [36] showing social statistics in a United Nations survey of world countries; (d) Time-series visualization using cluster drift shown by parallel sets-like representation [26] – the visualization (shown in the left) uses alternate colors to indicate bands for different clusters (shown in the right) and the polyline shows how a data object has drifted from cluster to cluster in different time instances, represented by the axes.

n) **Parallel sets-like representation of cluster drift (for (when)):** In parallel sets-like representation, the time-series of any variable/parameter can be studied by plotting the variable across multiple instances of time, as parallel axes, and placing values belonging to a cluster together on each axis. The data when loaded on this visualization will showcase how a particular subject has moved away or stayed in the cluster, which is shown in one of the axes. The clusters themselves are determined using appropriate partitioning/clustering algorithms (k-means, E-M, fuzzy c-means, etc.) and distinguished by using blocks of alternating color along each axis. Figure 1 shows a zoomed-in version of the parallel sets-like representation. For instance, plotting variables such as HAZ, where z-scores are computed from the same subject population, plotted across time instances, visually represents the rate of growth of each subject; thus enabling the study of the effectiveness of remedial measures to reduce stunting [25].

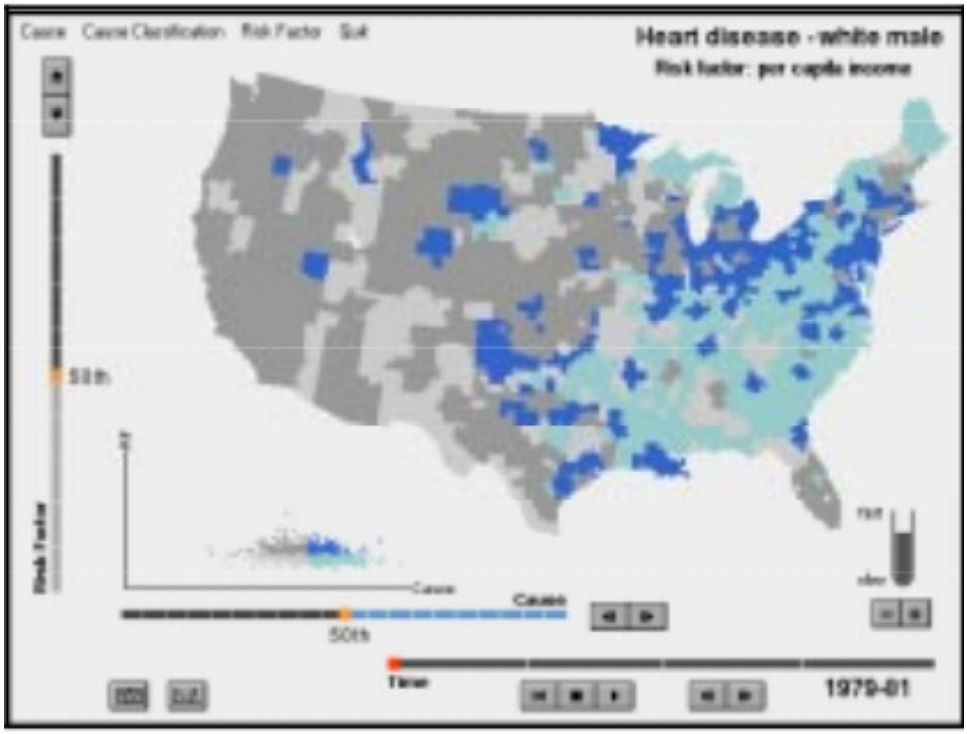

Fig. 2. Modifiable maps with several display controls, including timeline and animation playback, can be used for reference and brushing. This image shows the heart disease statistics in the U.S.A. for white male population over 1978-81 [21].

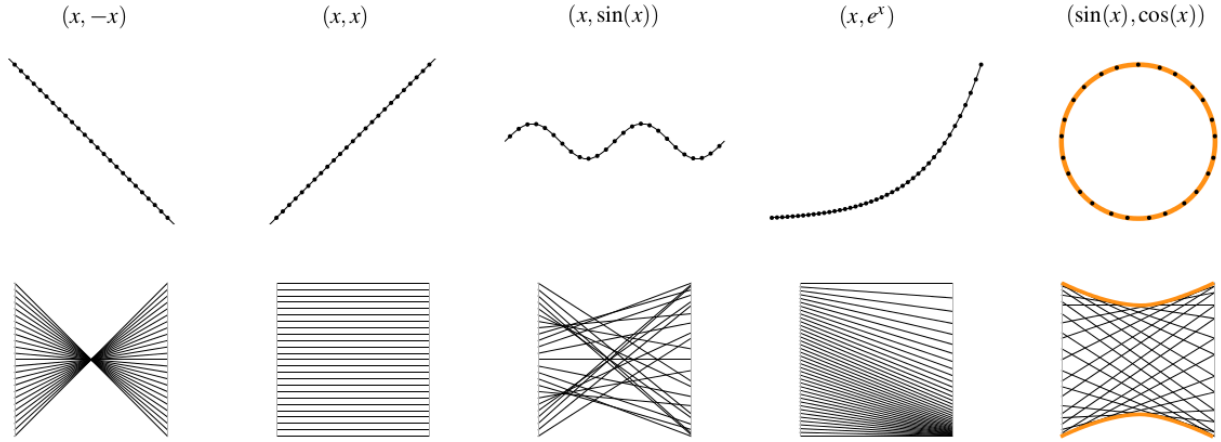

Fig. 3. Patterns in parallel coordinates (bottom), and its equivalent pattern in Cartesian coordinate system [11].

E. Discussions

In this section, we discuss the various aspects of our proposed dataset and proposed visualization techniques in the context of the visual analytic framework.

o) **Regarding proposed dataset:** While the proposed unified feature vector is exhaustive, not all parameters/variables are required for specific analyses. The need for specific subsets of parameters for specific analyses has been documented in several works [3], [23], [33]. Our motivation for using the entire set of variables is for studying various aspects of the problem of malnutrition. Using the classification of variables given in Section V-A in the perspective of the five W's in knowledge discovery given in Section V-B, we can see that normally (a) anthropometric and biomedical and micronutritional characteristics can be used for performing (what) and (when) tasks, (b) child characteristics for (who) and (why) tasks, and (c) maternal/household characteristics and community characteristics, for (why) and (where) tasks. Further, a combination of child and maternal characteristics gives a rich dataset for (who) and (why) tasks.

p) **Regarding proposed visualization techniques:** While some of the visualization techniques accomplish the same task in seemingly similar ways, the techniques also accomplish other unique tasks and hence harmoniously complement each other. An example of similarity in outcomes would be parallel coordinates plot and scatterplot matrix showcasing correlation between

different variables, upon reordering and in a pairwise fashion, respectively. However, in their own merit, parallel coordinates plot can additionally summarize trends and scatterplot matrices can identify outliers more effectively. Hence, having both visualization techniques in a juxtaposed views help in better analysis of the data.

Another such example is understanding the extent of malnutrition using parallel coordinates plot and cluster heat map; where both show subject-wise trends. In the former, the values of anthropometric measures for subjects shown on the parallel axes inform the user which subjects fall in specific ranges of values indicating conditions, such as stunting and wasting. In the latter, the color encoding of the elements of the data matrix indicate the same. However, in their own merit, parallel coordinates plot can show correlation between variables, and cluster heat map can show patterns within the hierarchical model by virtue of seriation of the matrix and graphically represented using a dendrogram. Similarly, both techniques show clustering tendencies in subjects based on parameters, however the latter can give information specific to subtrees in the hierarchical model. Thus, while there is overlap in the functionalities of the selected visualization techniques, individually they offer different perspectives of the same data.

VI. Conclusions

In this work, we have surveyed and identified multivariate data which can describe the situation of child malnutrition in India, in its entirety. We have identified a unified feature vector, using 46 variables, for a child subject under five years of age, for determining malnutrition, and further analyzing its cause and assessing the improvement of the subject in gaining nutrition.

The motivation for this feature vector is to propose a dataset of such data from children under five from different population samples all over India and visualize the dataset in its entirety, where domain experts can identify, assess and analyze the problem of malnutrition in India. The proposed dataset is a collection of the proposed feature vectors for a population. Thus, it is of time-varying multivariate type, as is used in visualization parlance, which shows the complexity of the data. In order to unravel the complexity in the data, we have proposed a set of visualization tasks in line with the five W's of journalistic reporting. The tasks can be performed using a set of appropriate visualizations and a combination of user interactions. We have proposed a visual analytic framework to implement the tasks, which includes (a) processes for data modeling and clustering, and (b) juxtaposed view of visualizations, including facilities for brushing and linking.

We have selected visualization techniques such as parallel coordinates plot and scatterplot matrices to perform multivariate visual analysis of the data. We have proposed derived data model using hierarchical modeling using spatio-temporal aspects of the data, and cluster heat map is a good way to visualize hierarchy in subject-wise representation. While these techniques work for a static dataset, the time-series can be analyzed by performing clustering on each time instance, and visualizing the cluster drift to monitor the improvement or deterioration in malnutrition status of the subject over time.

The unified feature vector, which is exhaustive, can be useful for analytics on the data, as depending on the visualization task, a subset of parameters/variables in the vector can be used. The idea of this work is to additionally motivate the need for creating good quality nationwide data collections [29], which will enable more rigorous study in regional as well as national trends in assessing and mitigating child malnutrition in India. We are aware there exists clean and reliable data in local regions, owing to efforts of local volunteer groups and local government bodies. Here, we are articulating a vision for data collection at a national level to gather reliable data. We have shown requirements of all parameters of the feature vector for different visual analytics tasks, which we hypothesize, provide the motivation to collect such data for studying child malnutrition in India, in its entirety.

VII. Acknowledgements

The author is grateful to the e-health committee and the e-Health Research Center at IIITB for inspiring work in this area, and to the reviewers of the paper for their valuable comments.

References

[1] Monika Blössner, Mercedes De Onis, Annette Prüss-Üstün, Diarmid Campbell-Lendrum, Carlos Corvalán, and Alistair Woodward. Malnutrition: Quantifying the health impact at national and local levels. *Environmental Burden of Disease Series*, (12), 2005.

[2] David Borland, Vivian L. West, and W. Ed Hammond. Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates. In *Proceedings of Workshop on Visual Analytics in Healthcare*, pages 53–58, 2014.

[3] Vani Borooah. Maternal literacy and child malnutrition in india. 2009.

[4] Annie T. Chen, Shu-Hong Zhu, and Mike Conway. Text Mining and Visualization to Explore E-Cigarette and Hookah-Related Social Media. In *Proceedings of Workshop on Visual Analytics in Healthcare*, pages 68–69, 2014.

[5] Alessandro Chetta, Jane M Carrington, and Angus Graeme Forbes. Augmenting ehr interfaces for enhanced nurse communication and decision making. In *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare*, page 4. ACM, 2015.

[6] William S Cleveland. Visualizing data. 1993.

[7] Filip Dabek, Jian Chen, Alexander Garbarino, and Jesus J Caban. Visualization of longitudinal clinical trajectories using a graph-based approach. In *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare*, page 5. ACM, 2015.

[8] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1539–1148, 2008.

[9] FAO, IFAD, and WFP. The State of Food Insecurity in the World 2015. *Meeting the 2015 international hunger targets: taking stock of uneven progress*, 2015.

[10] Arun Gupta, JP Dadhich, and MMA Faridi. Breastfeeding and complementary feeding as a public health intervention for child survival in india. *The Indian Journal of Pediatrics*, 77(4):413–418, 2010.

[11] Julian Heinrich and Daniel Weiskopf. State of the art of parallel coordinates. *STAR Proceedings of Eurographics*, 2013:95–116, 2013.

[12] Anna Herforth. Access to adequate nutritious food: new indicators to track progress and inform action. *The Fight Against Hunger and Malnutrition: The Role of Food, Agriculture, and Targeted Policies*, page 139, 2015.

[13] Eugenia McPeek Hinz, David Borland, Hina Shah, Vivian L. West, and W. Ed Hammond. Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels. In *Proceedings of Workshop on Visual Analytics in Healthcare*, pages 17–22, 2014.

[14] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates. In *Human-Machine Interactive Systems*, pages 199–233. Springer, 1991.

[15] International Food Policy Research Institute. Global Nutrition Report 2015: Accounts and Accountability to Advance Nutrition and Sustainable Development. 2015.

[16] International Institute for Population Sciences (IIPS) and Macro International. National Family Health Survey (NFSH-3), 2005-06: India. 1, 2007.

[17] Waqas Javed and Niklas Elmqvist. Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, pages 1–8. IEEE, 2012.

[18] Denis Klimov, Alexander Shkvensky, Robert Moskovitch, and Yuval Shahar. Interactive Analysis of Multiple Longitudinal Records of Diabetes Patients. In *Proceedings of Workshop on Visual Analytics in Healthcare*, pages 11–16, 2014.

[19] Robert Kosara, Fabian Bendix, and Helwig Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(4):558–568, 2006.

[20] Innar Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(2):70–91, 2010.

[21] Alan M MacEachren, Francis P Boscoe, Daniel Haug, and Linda W Pickle. Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pages 87–94. IEEE, 1998.

[22] Z Markos, F Doyore, M Yifiru, and J Haidar. Predicting under nutrition status of under-five children using data mining techniques: The case of 2011 ethiopian demographic and health survey. *J Health Med Informat*, 5(152):2, 2014.

[23] Peninah K Masibo. Trends and Determinants of Malnutrition among Children Age 0-59 Months in Kenya (KDHS 1993, 1998, 2003, and 2008-09). *DHS WORKING PAPERS*, (89):1–48, 2013.

[24] Ash Pachauri, Himanshu Khetarpal, Swati Trehan, and Tanvi Jain. Ananya visual analytics system: Applications for strengthening healthcare delivery in bihar, india. In *Healthcare Informatics (ICHI), 2014 IEEE International Conference on*, pages 206–209. IEEE, 2014.

[25] Arvind Panagariya. Does india really suffer from worse child malnutrition than sub-saharan africa? *Economic & Political Weekly*, 48(18):98–111, 2013.

[26] Saima Parveen and Jaya Sreevalsan-Nair. Visualization of small world networks using similarity matrices. In *Big Data Analytics*, pages 151–170. Springer, 2013.

[27] Margit Pohl, Sylvia Wiltner, Alexander Rind, Wolfgang Aigner, Silvia Miksch, Thomas Turic, and Felix Drexler. Patient development at a glance: An evaluation of a medical data visualization. In *Human-Computer Interaction–INTERACT 2011*, pages 292–299. Springer, 2011.

[28] Rokkam Radhakrishna, K Hanumantha Rao, C Ravi, and B Sambi Reddy. Chronic poverty and malnutrition in 1990s. *Economic and Political Weekly*, pages 3121–3130, 2004.

[29] S Irudaya Rajan and KS James. Third national family health survey in india: issues, problems and prospects. *Economic and Political Weekly*, pages 33–38, 2008.

[30] Alexander Rind, Silvia Miksch, Wolfgang Aigner, Thomas Turic, and Margit Pohl. Visuexplore: gaining new medical insights from visual exploration. In *Proc. Int. Workshop on Interactive Systems in Healthcare (WISH@ CHI2010)*, pages 149–152, 2010.

[31] Dominique Roberfroid, Naïma Hammami, Carl Lachat, Zita Weise Prinzo, Victoria Sibson, Benjamin Guesdon, Sylvie Goosens, and Patrick Kolsteren. Utilization of mid-upper arm circumference versus weight-for-height in nutritional rehabilitation programmes: a systematic review of evidence. *Geneva: World Health Organization*, 2013.

[32] Ben Shneiderman, Catherine Plaisant, and Bradford W Hesse. Improving health and healthcare with interactive visualization methods. Technical report, HCIL Technical Report, 2013.

[33] Suparna Som, Manoranjan Pal, Bishwanath Bhattacharya, Susmita Bharati, and Premananda Bharati. Socioeconomic differentials in nutritional status of children in the states of west bengal and assam, india. *Journal of biosocial science*, 38(05):625–642, 2006.

[34] Awalin Sopan, Angela Song-Ie Noh, Sohit Karol, Paul Rosenfeld, Ginnah Lee, and Ben Shneiderman. Community health map: A geospatial and multivariate data visualization tool for public health datasets. *Government Information Quarterly*, 29(2):223–234, 2012.

[35] WHO Multicentre Growth Reference Study Group and others. WHO Child Growth Standards based on length/height, weight and age. *Acta paediatrica (Oslo, Norway: 1992). Supplement*, 450:76–85, 2006.

[36] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2), 2009.

[37] Zhiyuan Zhang, Bing Wang, Foisal Ahmed, IV Ramakrishnan, Rong Zhao, Asa Viccellio, and Klaus Mueller. The five ws for information visualization with application to healthcare informatics. *Visualization and Computer Graphics, IEEE Transactions on*, 19(11):1895–1910, 2013.