INTEGRATING POPULATION SURVEYS USING SPATIAL VISUAL ANALYTICS: A CASE STUDY ON NUTRITION AND HEALTH INDICATORS OF CHILDREN UNDER FIVE IN INDIA

A PREPRINT

Harshitha Ravindra **Jaya Sreevalsan-Nair***
Graphics-Visualization-Computing Lab, and E-health Research Center,
International Institute of Information Technology Bangalore, Karnataka 560100, India.
http://www.iiitb.ac.in/gvcl

This is a peer-reviewed article accepted for publication, prior to its camera-ready version, in the Proceedings of the 7th International Conference on Geographical Information Systems Theory, Applications and Management, GISTAM 2021

February 21, 2021

ABSTRACT

Large-scale population surveys are beneficial in gathering information on the performance indicators of public well-being, including health and socio-economic standing. However, conducting national population surveys for low and middle-income countries (LMIC) with high population density becomes challenging. Economizing this activity, multiple surveys with different goals are decentralized and implemented by various agencies. Some of the surveys tend to overlap in outcomes with spatial/temporal or both scopes. Mining data jointly from surveys with significant overlap gives new insights while preserving their autonomy. We propose a three-step workflow for integrating surveys using spatial analytic workflow supported by visualizations. We implement the workflow on a case study using two recent population health surveys in India to study malnutrition in children under five. Our case study focuses on finding hotspots and coldspots for malnutrition, specifically undernutrition, by integrating both surveys' outcomes. Malnutrition in children under five is a pertinent global public health problem prevalent in India. Our work shows that such an integrated analysis is beneficial along with preliminary analyses of existing national surveys to find new insights while maintaining their autonomy.

1 Introduction

Large-scale surveys are implemented to gather information about specific issues on the population. Survey analysis provides time-tested mechanisms for monitoring multi-dimensional indicators of political units, such as countries, geographies, etc. In the public health domain, health surveys are used for public health outcome surveillance [1]. Such surveillance involves quantitative analysis of total population health and indicators [2]. However, despite the central role surveys play in monitoring population trends, implementing surveys is a complex problem owing to the demographic and socio-economic variations in the population, survey design for a multifaceted focus, diversity in handling data, decentralization of survey administration in the field, decisions on publishing data and outcomes, and finally, the economic and time cost of implementing surveys. Hence, we increasingly see that surveys are *owned* by various competent organizations who undertake them for *specific* requirements. This leads us to the case of overlapping surveys, as multiple surveys are implemented, with a focus on different metrics but considerable similarities [3]. Integrating such overlapping surveys is beneficial for gaining new knowledge, *e.g.,* multiple health surveys can be

*jnair@iiitb.ac.in

used to jointly estimate household wealth and expenditures while still maintaining the length of the questionnaires by integrating them [4].

Even though big data is gathered and analyzed in surveys, the scope of reaping integrated benefits from overlapping surveys becomes limited. It requires centralized planning efforts before conducting them. Such centralized activities reduce the degree of the desired autonomy in survey implementation for economic reasons in practice. There are primarily two issues with integrating surveys during its design and administration [3]. Firstly, there is a requirement of concerted effort to determine the scope and extent of overlap between multiple surveys to check the feasibility and benefit of such an integration. Secondly, there is a requirement of efficient government, which fosters such an integration, from planning a survey to publishing its outcomes.

That said, integrating multiple surveys at the data level is more promising, and integrating disparate data sources has been widely practiced. For example, various sources of data, such as geographic information, can be integrated with surveys [5]. One can also link spatial data from surveys and databases for the integration, *e.g.,* health surveys and health facility databases [6]. Since spatial and temporal information are essential to population survey data, they are used for testing the feasibility of direct integration of surveys. They further provide the mappings between the surveys for the implementation of the integration.

It is recommended that the data collection and the reporting systems enable data sharing to improve the adaptation of integrated surveys [2]. As an example, in India, the availability of raw data and reports of the National Family Health Survey (NFHS) in the public domain, has improved the uptake of several researchers working with the data, compared to similar national surveys [7]. The NFHS is favorably implemented at the national scale at a higher frequency, *i.e.,* roughly once in 5 years, aligned with the worldwide data collection efforts. The NFHS data can be strategically used with other national and local surveys to infer health and related socio-economic factors, even though its focus is on maternal-child health indicators. Hence, we choose to integrate the NFHS-4 during 2015-16, the fourth edition of NFHS [8], and the Comprehensive National Nutrition Survey (CNNS) during 2016-18 [9]. These surveys are conducted by the Ministry of Health and Family Welfare (MoHFW), Government of India (GoI), and implemented by the Ministry of Statistics and Programme Implementation (MoSPI), GoI. MoSPI provides access to the demographic survey outcomes. However, studies using the open data have examined these surveys in a silo, based on their specific individual goals. There is also prior work on comparing these surveys, specifically [10], but not integrating them. An integrated analysis of pertinent surveys can effectively reduce the burden of conducting numerous surveys in a populous middle-income country like India. Hence, our goal is to demonstrate a proof-of-concept of a cross-analysis. Our challenge here lies in the difference in the granularity of the open data available in the two chosen surveys, limiting our scope of directly integrating them at the data level. We address this by using spatial statistics and visualizations.

We focus on mining information on various aspects of malnutrition for children under five, in India, through this integrated study. Under-five studies are concluding spatial heterogeneity in various health indicators on malnutrition [11, 12, 13], which can be exploited. The interest in under-five studies is due to the persistence of childhood morbidity and mortality in India, as per NFHS-4 [14]. Wasting has not reduced as much between NFHS-3 and NFHS-4 findings as stunting. In the weighted sample taken in CNNS, the prevalence of anemia is 40.5% amongst children under five, with iron-deficiency anemia being the most prevalent type [15]. The nutritional deficiency affects all age groups, but children under five, particularly those with severe acute malnutrition (SAM), have a higher mortality risk from common childhood illnesses such as diarrhea, pneumonia, and malaria [16]. While the infant mortality rate (IMR) is at 41 per thousand live births, the under-5-mortality rate (U5MR) is at 50. Childhood undernutrition accounts for 45% of U5MR alone and is a crucial public health issue in India. Dietary diversification is an additional solution apart from the focus on infrastructure for food distribution and delivery by the government [14]. There is an emphatic call for more frequent health surveys to be conducted to continuously monitor the progress due to such nutrition programs and infrastructural improvement, motivating our integrated study.

A fine-grained analysis has been done on the occurrence of anemia, stunting, and incomplete immunization in children aged 12-59 months, at district and individual levels, using NFHS-4 data [12]. This study also showed the influence of maternal education on the aforementioned outcomes at the district level. There is also evidence that there is spatial influence on poor sanitation, which is one of the causes of stunting in India, where the extreme temperature is a contextual correlate [17]. We use these analyses of the concerned surveys for identifying *contextual factors* of malnutrition.

Our novel contribution is in using *visual analytics with spatial context* for integrating surveys, namely NFHS-4 and CNNS in India, for under-five child malnutrition study. Visual analytics is a data analysis workflow where one uses visualization to provide the feedback loop along with other data mining methods [18]. We propose a three-step workflow of (i) using state-wise differences for determining the feasibility of survey integration, (ii) a region-based study to identify variables for integration, and (iii) finding spatial clusters for survey integration outcomes. For (i), we use descriptive statistics, in addition to map-based visualizations of state-wise counts of affected children U5, distribution

counts, and distribution distances between surveys. Once the feasibility of integrating surveys is established, we identify appropriate variables and factors for the integration. Thus, for (ii), we use circular radar plots to investigate the region-wise trends of variables that are not common in both surveys. We also identify the contextual factors, *e.g.,* sanitation facilities, maternal literacy, using literature surveys. We take care that the variable and contextual factors are from different surveys for the sake of integration of the two surveys. Using these selected variables and factors, we achieve (iii) through spatial statistical analysis using global Moran's I and bivariate LISA (Local Indicators of Spatial Association) using local Moran's I. The results of the integrated study are the spatial clusters based on the significance of selected indicators and contextual factors from both surveys.

2 Methodology

Spatio-temporal metadata is familiar to our selected surveys, and appropriate aggregation can alleviate the differences in the spatial granularity of open data and minor overlap of its time-periods in the surveys. Hence, our integration method is driven predominantly by spatial analysis. Our proposed workflow begins with determining integration feasibility and then follows up with determining variables and spatial statistical methods for integration. Given the complexity of the data in such large-scale surveys, visualizations enable a qualitative understanding of the spatial trends. Thus, we propose a *spatial visual analytic* approach for survey integration for identifying high-risk regions in India for under-five child malnutrition in our case study.

Data: NFHS-4, 2015-16 provides information on population, health, and nutrition for women, men, and children under five for all districts in all states and union territories in India. The International Institute for Population Sciences (IIPS), Mumbai, is the nodal agency for conducting different rounds of the survey.

CNNS, 2016-18 is the largest exhaustive nutrition survey including micro-nutrients conducted for the first time in India, led by UNICEF and Population Council, New Delhi. This survey is focused on all children, *i.e.,* population under 18 years of age.

Both surveys overlap in the coverage of nutrition indicators of children under five. The summary analysis reports have been published for both surveys, and the raw anonymized household-level data is available for the NFHS-4. The data and indicators that we use for our case study using NFHS-4 and CNNS surveys are listed in Table 1. While the relevant indicators for undernutrition are present in both surveys owing to their respective scope and goals, certain variables are covered in only one of the two. Our goal is to correlate variables across the two surveys spatially.

Scope of our Study: The focus of our case study is on under-five child malnutrition as recorded in the NFHS-4 and CNNS. We find potential indicators and contextual factors for identifying high-risk regions of under-five child malnutrition using integrated data mining from both surveys. The integration is at the state-level, given the coarsest granularity of data available in both surveys. Since the indicators for undernutrition conditions, except anemia, are available for sub-groups (Table 1) based on gender and urbanization, we use this additional information to study distributions of specific populations. Our proposed spatial analysis, inclusive of visualizations, validates the choice of variables used in the integrated study.

Method: Our three-step workflow consists of feasibility check, choosing variables for integration, and integration using spatial correlation and clustering. Given the difference in the scope and goals, implementation, and time-frame of the surveys, we first check the feasibility of integrating them. The time-frame difference is not highly significant here, given that population surveys in consecutive years will not yield considerable differences. However, since there is a difference in the survey implementation, including population sampling, and differences in publishing data, we undertake the feasibility test.

For the feasibility test, we first visually check if the state-wise sample distributions for both surveys are equivalent and check them against the state-wise population distribution latest official census taken in 2011 in the states. The region-wise grouping of the thirty states and distribution of sample population covered in the surveys are shown in Figure 1, (A) and (B), respectively. We use visualization in addition to quantitative analysis, as visualizing state-wise discrepancies provides a look-up to explain the differences we see in the indicators given in both surveys.

Step-1:- To complete the feasibility test, we identify the common indicators from both surveys. In addition to their absolute count, both the surveys have data of the discrete probability distributions of the severity of each of the malnutrition conditions, namely, stunting, underweight, wasting, and anemia. Except for anemia, we also have data available in the gender- and urbanization-based sampled groups, in addition to the total population. Hence, we find distribution distances for each group using Hellinger distance (HD) to quantify the similarity between the indicators

Table 1: Metadata and overall descriptive statistics (mean μ and standard deviation σ within the corresponding respondents) of selected indicators and contextual factors, and available distribution data on severity with given labels, from NFHS-4 and CNNS, for children under five (U5).

	NFHS-4	CNNS
#Respondents (total)	601,509 households	112,100 children
#Children (U5)	259,628	40,700
Survey Time	2015-16	2016-18
Granularity of available data	Household	State
Indicators for children (U5): **Undernutrition** in $\mu(\sigma)$		
Stunted	32.23 (7.40)	30.27 (6.80)
Wasted	18.30 (5.40)	14.56 (5.43)
Underweight	28.00 (9.90)	26.50 (9.13)
Anemic	42.36 (10.61)	34.06 (10.94)
Micronutrient Deficiency		
Folate	✗	24.98 (21.21)
Low Serum Ferritin	✗	32.27 (16.64)
Vitamin A	✗	17.03 (9.83)
Vitamin B12	✗	10.93 (6.74)
Vitamin D	✗	15.78 (13.49)
Zinc	✗	18.39 (7.81)
Indicators for Children (U5): Distribution Data		
Stunted	—— [“not severe”, “severe”] ——	
Wasted	—— [“not severe”, “severe”] ——	
Underweight	—— [“not severe”, “severe”] ——	
Anemic	- [“mild”, “moderate”, “severe”] -	
Indicators for Children (U5): **Immunization** in $\mu(\sigma)$		
No/Partial Immunization	35.94 (14.00)	✗
BCG	90.93 (08.08)	✗
DPT	79.17 (11.59)	✗
Fully Immunized	64.06 (13.90)	✗
Hepatitis B	64.64 (13.80)	✗
Measles	80.66 (11.51)	✗
Polio	74.05 (11.06)	✗
Contextual Factors *in Mean (SD)*		
Maternal Illiteracy	27.06 (12.03)	✗
Sanitation (Unimproved)	43.77 (20.15)	✗

across the two surveys. The HD between two discrete distributions P and Q, $D_{HD}(P,Q)$, is given as:

$$D_{HD}(P,Q) = \frac{1}{\sqrt{2}}.||\sqrt{P} - \sqrt{Q}||^2 \quad .$$

We choose HD owing to its properties of symmetry and being a bounded metric with the support [0.0, 1.0], where $D_{HD} = 0$ means highly similar distributions, and $D_{HD} = 1$, highly dissimilar. These properties enable comparisons of the HD distances across states, where the HD is computed per state between distributions across surveys.

Another important HD property is that it follows the triangle inequality property, which implies that the HD between the two empirical discrete probability density distributions is not greater than the HD between each of the discrete distribution and the actual parameterized distribution. Thus, the use of HD ensures the comparison of the lower bound of distances here. We compute the HD distances between distributions of [non-severe, severe, absence] for each malnutrition condition in the two surveys. We compute distances for selected group (female, male, urban, rural), wherever applicable, as well as the for the total population, *e.g.,* we find the HD of distribution of [non-severe, severe, absence] of stunting for female children under five, given in percentages, between NFHS and CNNS data (Figure 2, first row, (a)).

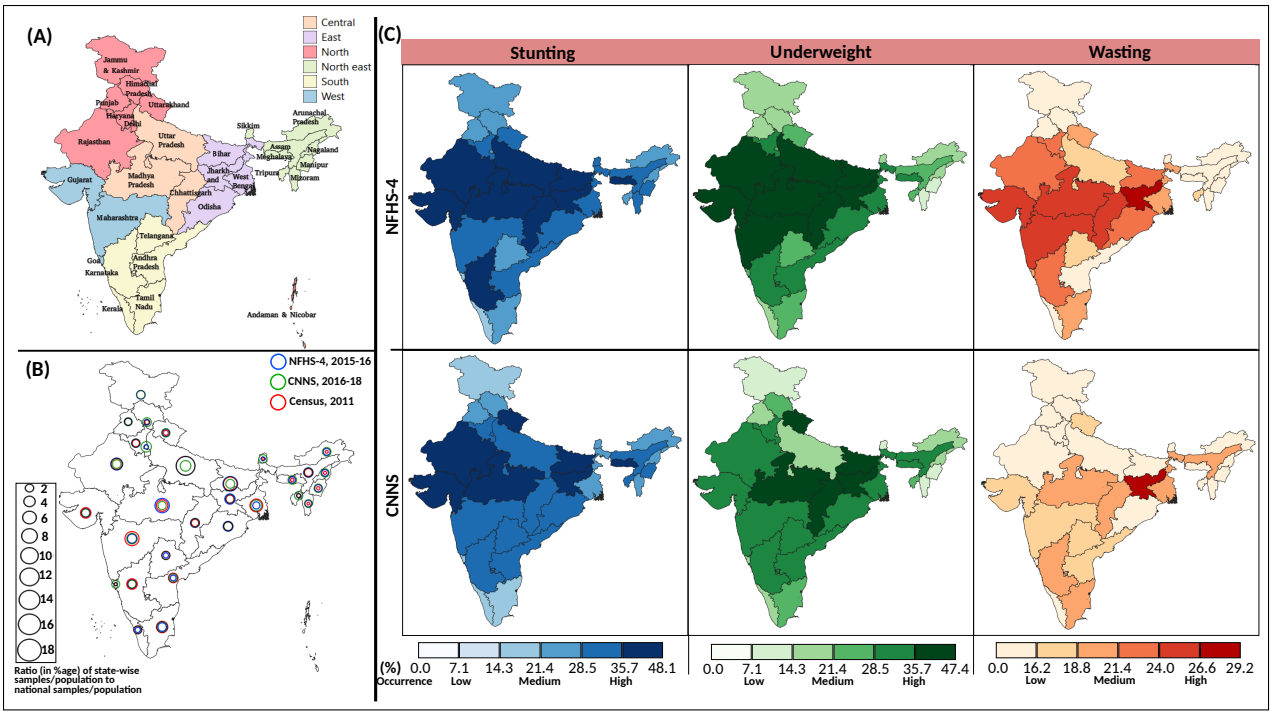


Figure 1: Data from the selected surveys. (A) Region-wise grouping of states in the political map of India. (B) Comparison of sampled population distribution for NFHS-4 and CNNS using ratios of state-wise count with respect to the that of the country, against baseline ratios using the population size from Census 2011, using percentage format. (C) Percentage of children under five who are stunted, wasted and underweight across all states in India, as reported by the surveys.

Unlike other undernutrition conditions, the data for anemia in the CNNS report is both sparse and at the coarser level. Hence, we use pie-chart glyphs in maps to visualize the relative distribution of the severity of anemia in different states and compare the surveys' distributions.

Step-2:- The second step in our workflow is the variable selection for the integrated analysis of surveys. Given the spatial local heterogeneity in undernutrition [11, 13] in India, we study the region-based trends in variables exclusive to each of the surveys. We use the immunization status from NFHS-4 and micronutrient deficiency from CNNS. The immunization status includes the percentage of children under five completing [BCG, DPT, Hepatitis B, Measles, Polio] vaccinations and achieving *fully immunized* status. The micronutrient deficiency includes the percentage of children under five with deficiencies in [folate, low serum ferritin, vitamin A, vitamin B12, vitamin D, Zinc]. Higher percentages for immunization status and lower percentages for micronutrient deficiencies imply better health indicators for children under five in the region. Since the variable analysis is for choosing a variable for integrating surveys, we use visualizations using a circular radar plot for qualitative comparisons. We choose a circular plot to visually represent percentage data. The choice of radar plot is owing to its compactness, where a region-wise radar plot has each spoke or axis representing a state in the region.

Step-3:- The third step in our workflow is the integrated analysis using spatial correlation using global Moran's I and localized cluster maps using bivariate LISA (Local Indicators of Spatial Association) computed using local Moran's I [19]. We perform the spatial correlation analysis of the indicators common to both surveys and the variables identified in Step-2. Moran's I is a weighted correlation coefficient, where the weights are provided based on spatial locations of the entities, given by:

$$I = \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}.(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}.\sum_{i=1}^{N}(x_i-\bar{x})^2}.$$

where N is the number of observations, $\bar{x}$ is the mean of the variable x, x_i and x_j are the values of x at locations i and j, respectively, and w_{ij} is a weight indexing location i with respect to location j. We compute the Moran's I for each common indicator between its values from both the surveys, using states as observations. Moran's I values significantly less than $\tau = \frac{-1}{N-1}$ imply negative spatial autocorrelation, and significantly higher than τ imply positive spatial autocorrelation. Moran's I values transformed to z-scores, and its p-value provides information about spatial

clustering and statistical significance, respectively. (p-value < 0.05) implies the variable is statistically significant in rejecting the null hypothesis that the spatial distribution of features is an outcome of random spatial processes. A positive z-score indicates more spatially clustered patterns, and a negative z-score indicates more spatially dispersed patterns.

We use bivariate LISA to identify the high-risk (hotspots) and the low-risk (coldspot) regions. These values are computed between each of the common undernutrition indicators in both surveys, identified in Step-1. We then identify high-risk and low-risk regions with the indicator selected from Step-2 and corresponding contextual factors determined from the literature survey. We ensure that the indicator and its corresponding contextual factor are not from the same survey. We make inferences from these identified hotspots and coldspots.

3 Results

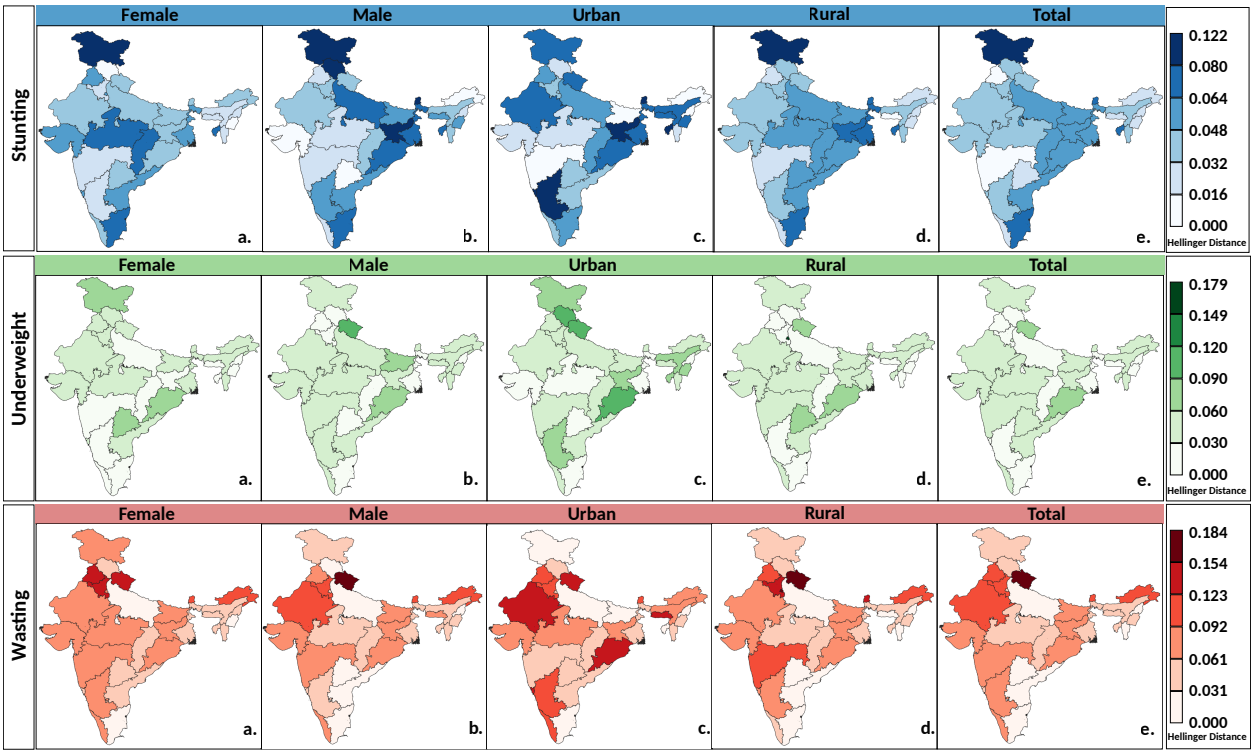


Figure 2: Hellinger distance between discrete probability distribution of different levels of severity [non-severe, severe, absence] of different undernutrition conditions, namely, stunting, underweight and wasting in children under five in the states of India. The distances are computed for different populations of the children, namely, female, male, urban, rural and total.

We have used Python 3.0 implementation with the Scipy package for computing HD. The map-based visualizations have been generated using QGIS version 3.8.3, the circular radar charts using R, and the spatial autocorrelation and cluster maps using GeoDa 1.14.

Step-1: Implementing our proposed workflow in our case study of integrated analysis of NFHS-4 and CNNS for malnutrition in children under five in India, we first evaluate the feasibility of such a study. We observe that the statistical descriptors of stunting, underweight, and wasting are comparable (Table 1), but there are state-level variations across surveys for the percentage of occurrence of these malnutrition conditions (Figure 1,(C)). We observe that NFHS-4 captures more regions for the high-occurrence of each of these conditions than CNNS, especially in the west and central regions. The low-occurrence states are captured more accurately across both surveys. These variations in medium- and high-occurrence states can be attributed to the differences in sampling, survey administration, data processing, reporting, and sampling (Figure 1,(B)) across the states. But still, we need a fine-grained analysis to improve the feasibility of our study. Hence, we use the distribution of different levels of severity of stunting, underweight, and wasting occurring in sub-populations of children under five.

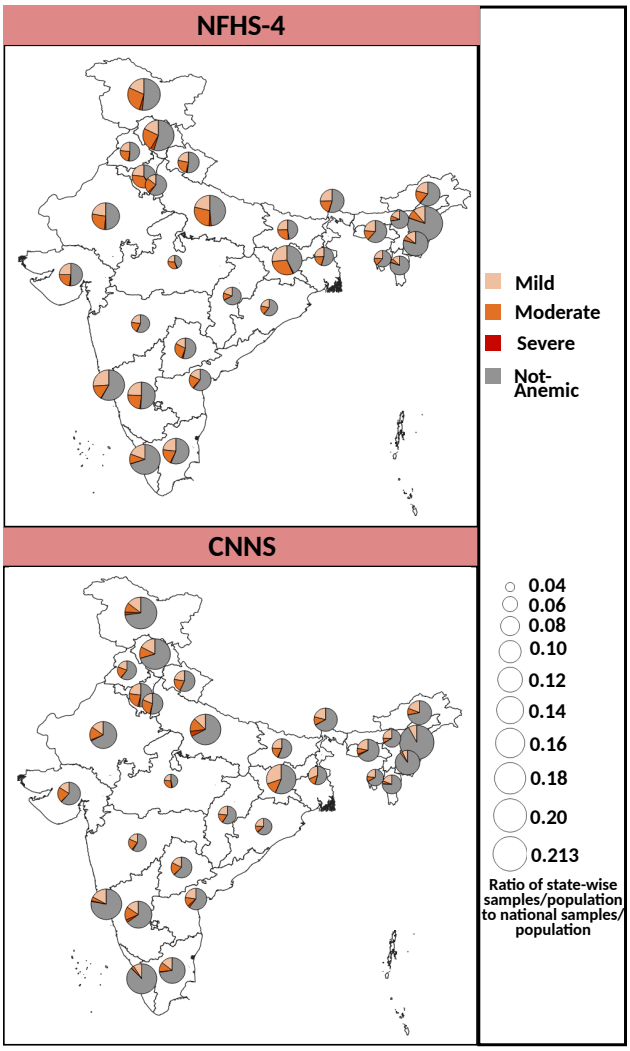


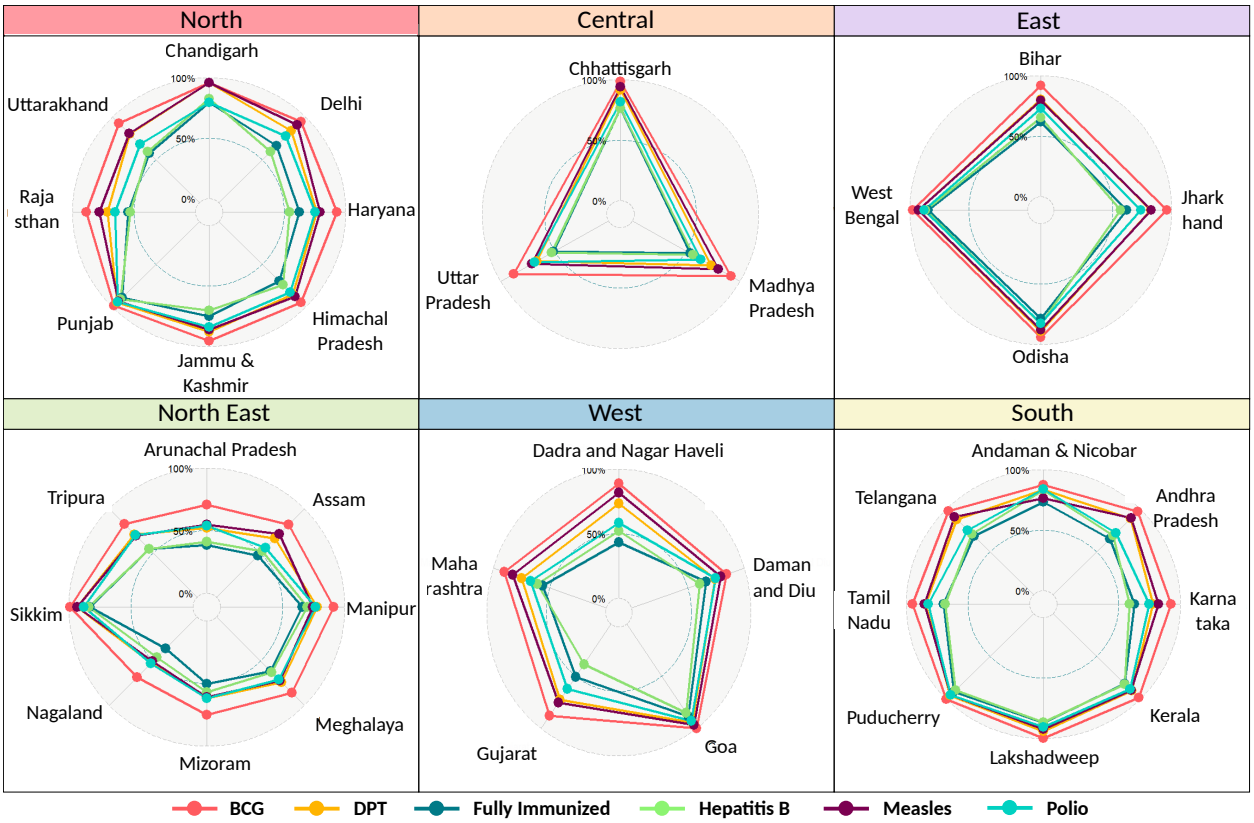
Figure 3: Pie-chart glyphs in map visualization of discrete probability distribution of different levels of severity [mild, moderate, severe, absence] of anemia in children under five in the states of India, where the glyph size is proportional to the fraction of children suffering from anemia relatively in each state of the country.

This additional information is used for computing state-wise Hellinger distances (HD) between the indicators from the surveys, which are visualized using choropleth maps in Figure 2. Here, we observe that the state-wise variations are low, as the HDs are lower than 0.184 overall, much lower than the upper bound, 1.0. We observe that isolated states show relatively higher HDs, namely Jammu & Kashmir for stunting and Uttarakhand for wasting, across all five population groups. This could also be attributed to the lesser number of samples from these regions.

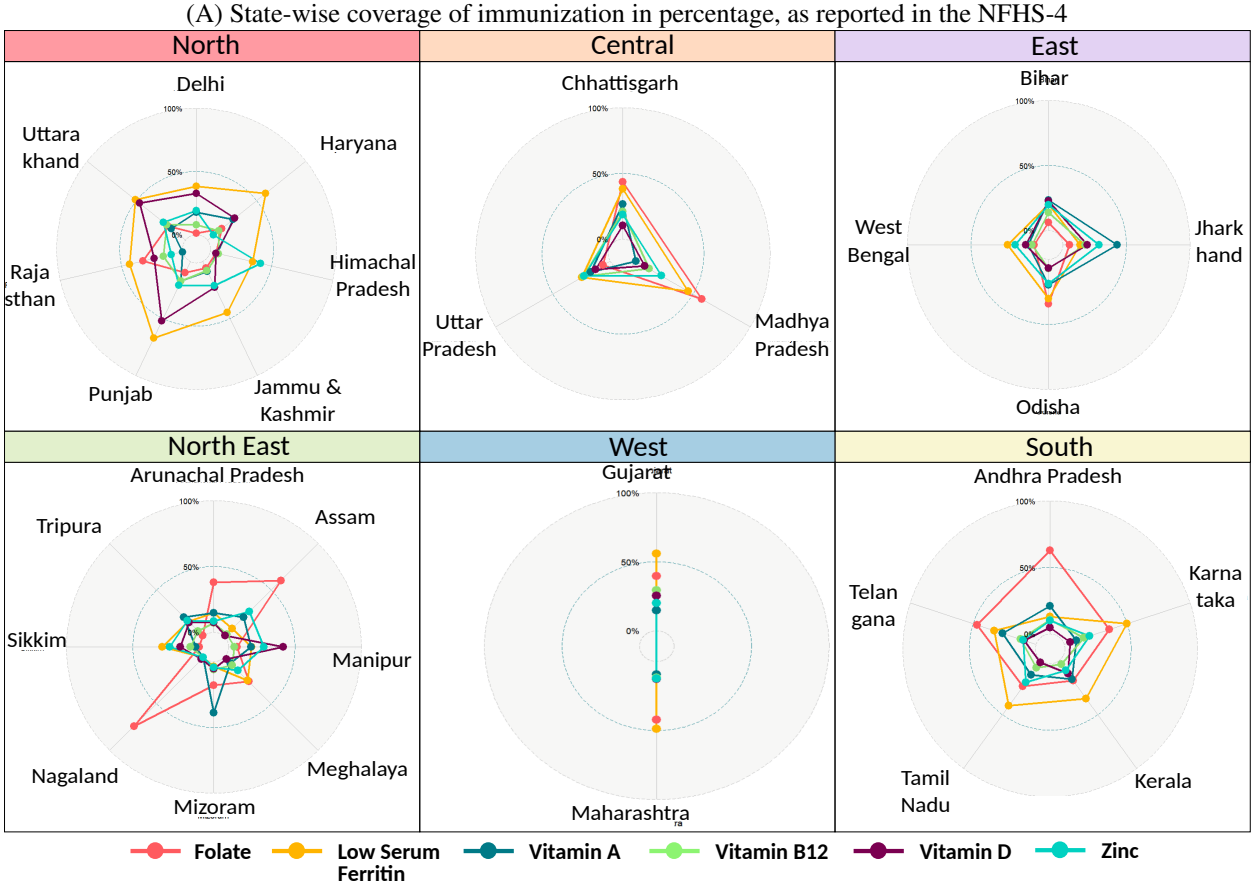
When we consider the data for anemia in Figure 3, we observe from the pie-chart glyph sizes that the occurrence of anemia in each state is similar across the surveys. However, we also observe differences in the distribution of severity of anemia occurring in different states, as seen in the pie-chart glyphs themselves. We do not see salient differences in counts for occurrence of severe-anemia owing to its lesser prevalence. The differences in the prevalence of mild- and moderate-anemia across surveys could be attributed to the lack of information on the population size on which percentages have been computed in the CNNS.

Overall, we now conclude that the distribution of the indicators has strong similarities across NFHS-4 and CNNS, thus, making our study feasible.

Step-2: In order to identify indicators and contextual factors across surveys for integrated analysis, we choose immunization record and micronutrient deficiency for indicators of children under five, and maternal illiteracy and poor sanitation facilities for contextual factors [17, 12]. The circular radar plots of region-wise values of the indicators



(A) State-wise coverage of immunization in percentage, as reported in the NFHS-4



(B) State-wise occurrence of micronutrient deficiency in percentage, as reported in the CNNS

Figure 4: Circular radar plots showing the (A) coverage of immunization and (B) occurrence in micronutrient deficiency, given in percentage in different regions in India.

(Figure 4 demonstrate that there is predominantly uniform coverage of immunization in states in each region, whereas micronutrient deficiency shows spatial local heterogeneity even within regions. We observe *spatial local heterogeneity* from the non-uniform patterns in each region, *e.g.,* there is pronounced deficiency in folate in Assam and Nagaland in the north-east, in Andhra Pradesh in the south, and Madhya Pradesh in the central regions. A significant deficiency in low serum ferritin, which is a primary cause of iron-deficient anemia, is observed in Haryana and Punjab in the north and Karnataka in the southern regions.

Overall, we observe high inter-region but low intra-region heterogeneity in immunization for children under five, as per NFHS-4, and high inter- and intra-region heterogeneity in micronutrient deficiency for children under five, as per CNNS. Hence, we use micronutrient deficiencies for indicators in our integrated study. Maternal illiteracy is a contextual factor for malnutrition, in general, and poor sanitation is commonly cited for stunting. The data for both contextual factors are available in the NFHS-4 (Table 1).

Step-3: Our integrated analysis of surveys is based on spatial statistics. The global Moran's I statistics for spatial autocorrelation between common indicators in both surveys for stunting, underweight, wasting, and anemia are given along with the bivariate LISA cluster maps in Figure 5. For N=30 (states), we get $\tau = -0.034$. Thus, we see here that there is low spatial heterogeneity, which is statistically significant, for stunting, underweight, and wasting. The low spatial heterogeneity validates the similarity of indicators for the indicators corresponding to these conditions across the surveys, seen in the Hellinger distance maps (Figure 1,(C)). The spatial auto-correlation results for anemia show more spatial outliers (Figure 5,d.) than the other undernutrition conditions. This result validates the higher differences observed in anemia indicators between the surveys (Figure 3), compared to the other conditions (Figure 2). We observe high-high clusters in central and western regions (Figure 5, a.-c.), which may be attributed to the disparity in sampling (Figure 1,(B)).

The bivariate LISA cluster maps for spatial correlation between an indicator and contextual factor are given in Figure 6. The hotspots are the high-risk regions when both an indicator and contextual factor have high values, *i.e.,* high-high. Bivariate LISA between unimproved sanitation and micronutrient deficiency (Figure 6,(A)) show high-risk clusters in the western region for folate, Rajasthan for low serum ferritin, large parts of northern-central regions for vitamin B12, and Jammu & Kashmir for Zinc deficiencies. We find that 3, 1, 1, 3, 1 out of 30 states have a higher prevalence of folate, low serum ferritin, vitamin A, vitamin B12, and Zinc deficiencies coexisting with unimproved sanitation, respectively. Unimproved sanitation is an important factor of stunting [20], thus, indicating that the hotspots are potential regions for the co-occurrence of both stunting and micronutrient deficiency.

Bivariate LISA between maternal illiteracy and micronutrient deficiencies (Figure 6,(B)) indicate 3, 1, 3, and 1 out of 30 states having a higher prevalence of maternal illiteracy coexisting with folate, low serum ferritin, vitamin B12, and Zinc deficiencies, respectively. We observe the clustering patterns in high prevalence of low serum ferritin, vitamin B12, and Zinc deficiencies coexisting with maternal illiteracy, which is similar to the same with unimproved sanitation. Parental education is an important factor in the occurrence of anemia and stunting [12]. Thus, we can conclude that the hotspots have a high risk of co-occurrence of micronutrient deficiencies and stunting or anemia.

We observe relatively fewer spatial outliers in our integrated analysis (Figure 6,(A)-(B)), reinforcing the feasibility of this integrated analysis. The pie-chart glyph map (Figure 3), the cluster map of correlation of anemic prevalence between surveys (Figure 5,d.), and the cluster map of low serum ferritin deficiency against both contextual factors (Figures 6,(A)-(B)) demonstrate that north-eastern region is a coldspot for prevalence of anemia, *i.e.,* a low-risk region. We also observe that Rajasthan is an outlier for bivariate LISA analysis of anemia across both surveys but is a hotspot in bivariate LISA analysis of variable, low serum ferritin, and contextual factor, for both lack of sanitation facility and maternal illiteracy (Figures 6,(A)-(B), Low Serum Ferritin). This is due to the low occurrence of anemia in Rajasthan recorded in CNNS, in comparison to that NFHS-4 (Figure 3), even though low serum ferritin has been observed in CNNS for Rajasthan (Figure 4, North).

Overall, we conclude that our integrated survey analysis has brought forward findings that could not have been made from either survey in isolation.

4 Conclusions

Our study illustrates the integration of national surveys, namely, NFHS-4 and CNNS, using spatial-visual analytics to find high- and low-risk regions of co-occurrence of malnutrition conditions in children under five in India. The analysis is done for undernutrition conditions at the state-level, and resolving the difference in the granularity of the data openly available for both surveys. Our results of hotspots and coldspots using the indicators for micronutrient deficiencies from CNNS and contextual factors from NFHS-4 show the usefulness of our work. We have also shown that the indicators which are commonly available for both the surveys also reveal hotspots and coldspots, where CNNS

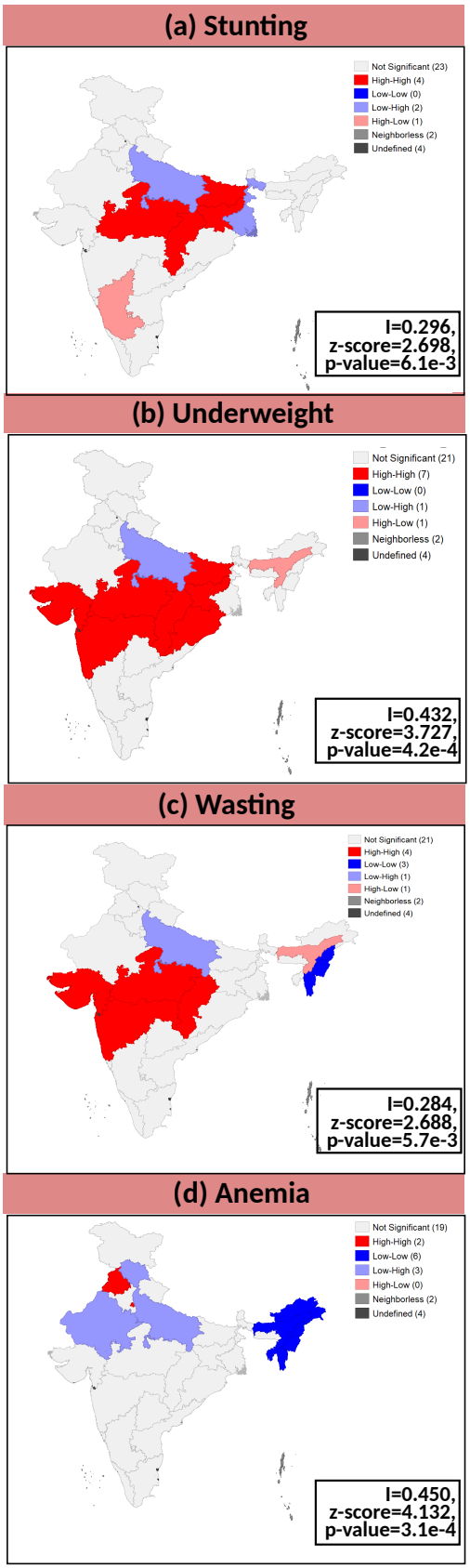
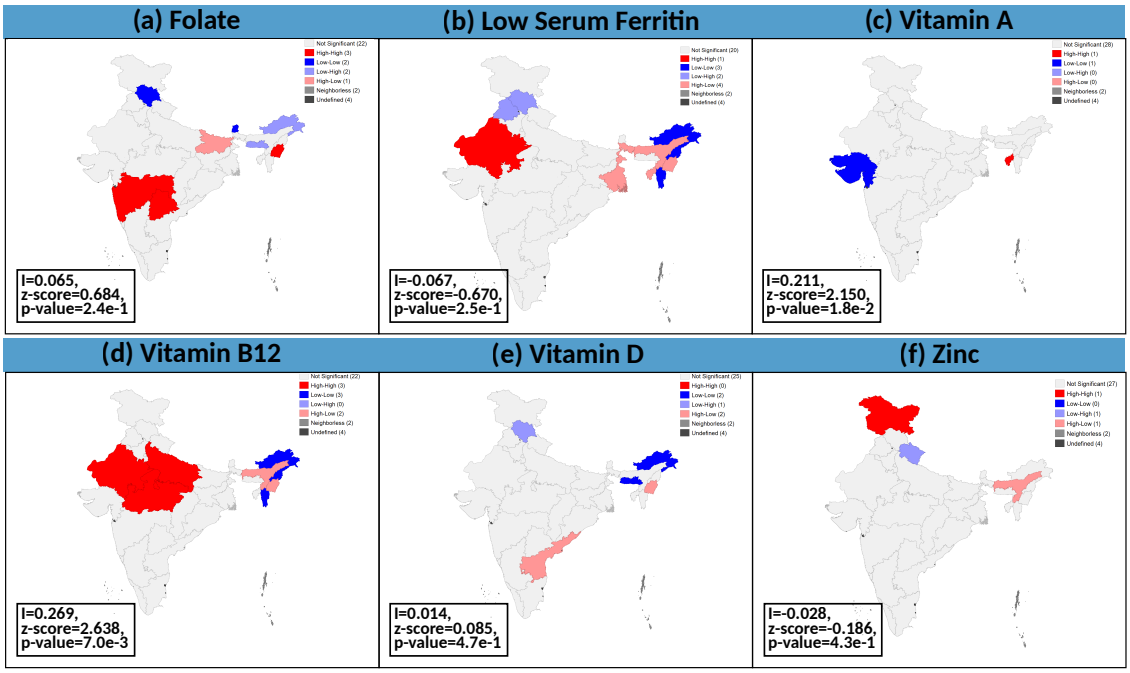
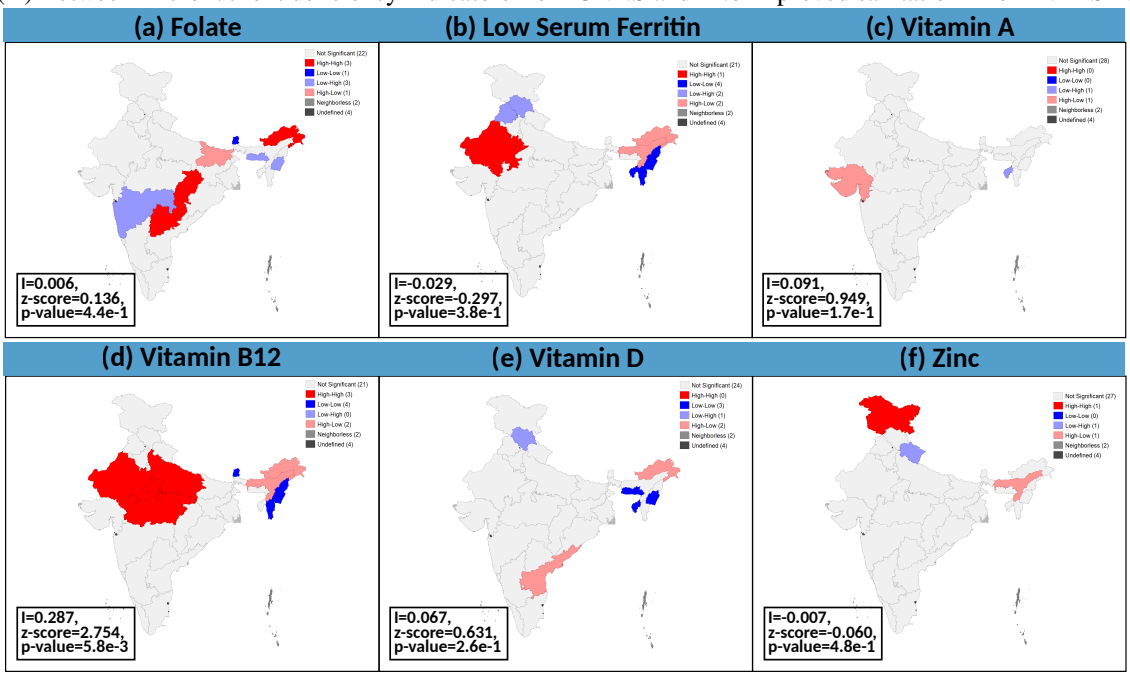


Figure 5: Global Moran's I statistics and bivariate LISA cluster maps of India showing the local clustering (hotspots & coldspots) at the state-level from CNNS and NFHS-4 survey for undernutrition indicators in children under five, for (a) Stunting, (b) Wasting, (c) Underweight, and (d) Anemia.



(A) Between micronutrient deficiency indicators from CNNS and "No improved sanitation" from NFHS-4.



(B) Between micronutrient deficiency indicators from CNNS and women illiteracy from NFHS-4.

Figure 6: Global Moran's I statistics and bivariate LISA cluster maps of India showing the local clustering (hotspots & coldspots) at the state level between indicators from CNNS, and contextual factors from NFHS-4 surveys.

in 2016-18 reinforces the findings of NFHS-4 in 2015-16. Our systematic integration of the surveys uses a three-step workflow involving a feasibility check, variable identification, and the integration using spatial statistics. Further, our spatial clustering results also show the high-risk and low-risk regions identified across the surveys for indicators common in both. Our work has future scope of generalization across any two large-scale population surveys, using a formal abstraction.

In summary, we show a proof-of-concept of integrating existing large-scale population surveys, benefiting the stakeholders. The integrated findings may have been otherwise siloed within the surveys but are significant when observed together. The goal of our work is to demonstrate evidence of such significant integrated results in order to improve the adaptation of survey integration. The responsibility of data collection is split strategically between national and local population health surveys for economic reasons. Planning joint outcomes across different surveys and mining data jointly from multiple surveys can give deeper insights together while preserving the autonomy of each survey in its entirety.

ACKNOWLEDGEMENTS

This work has been supported by the IBM Shared University Grant and the Mathematical Research Impact Centric Support (MATRICS) grant by the Science and Engineering Board (SERB). This paper has also benefited from the inputs from members of GVCL and EHRC, and anonymous reviewers. This study has been possible solely because of the open data available in the public domain – the unit-level NFHS-4 data and fact sheets for NFHS-4 and CNNS.

References

[1] Peter Nsubuga, Mark E White, Stephen B Thacker, Mark A Anderson, Stephen B Blount, Claire V Broome, Tom M Chiller, Victoria Espitia, Rubina Imtiaz, Dan Sosin, et al. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. *Disease control priorities in developing countries*, 2:997–1018, 2006.

[2] Dawn Marie Jacobson and Steven Teutsch. An Environmental Scan of Integrated Approaches for Defining and Measuring Total Population Health. In *National Quality Forum, Washington, DC2012*, 2012.

[3] Marc L Berk, Claudia L Schur, and Jacob Feldman. Twenty-five years of health surveys: Does more data mean better data? *Health Affairs*, 26(6):1599–1611, 2007.

[4] Saul S Morris, Calogero Carletto, John Hoddinott, and Luc JM Christiaensen. Validity of rapid estimates of household wealth and income for health surveys in rural Africa. *Journal of Epidemiology & Community Health*, 54(5):381–387, 2000.

[5] Matthew William Cooper. *People and Pixels: Integrating Remotely-Sensed and Household Survey Data for Food Security and Nutrition*. PhD thesis, University of Maryland, College Park, 2020.

[6] Winfred Dotse-Gborgbortsi, Andrew J Tatem, Victor Alegana, C Edson Utazi, Corrine Warren Ruktanonchai, and Jim Wright. Spatial inequalities in skilled attendance at birth in Ghana: a multilevel analysis integrating health facility databases with household survey data. *Tropical Medicine & International Health*, 25(9):1044–1054, 2020.

[7] Rakhi Dandona, Anamika Pandey, and Lalit Dandona. A review of national health surveys in India. *Bulletin of the World Health Organization*, 94(4):286, 2016.

[8] IIPS and MoHFW. *National Family Health Survey state factsheets, 2015-16*. IIPS, Mumbai, 2016.

[9] MoHFW, UNICEF and Population Council. *Comprehensive National Nutrition Survey (2016-2018) National Report*, 2019.

[10] Komal Rathi, Preeti Kamboj, Priyanka Gupta Bansal, and GS Toteja. A review of selected nutrition & health surveys in India. *The Indian journal of medical research*, 148(5):596, 2018.

[11] Junaid Khan and Sanjay K Mohanty. Spatial heterogeneity and correlates of child malnutrition in districts of India. *BMC public health*, 18(1):1027, 2018.

[12] Parul Puri, Junaid Khan, Apurba Shil, and Mohammad Ali. A cross-sectional study on selected child health outcomes in India: Quantifying the spatial variations and identification of the parental risk factors. *Scientific reports*, 10(1):1–15, 2020.

[13] Himani Sharma, SK Singh, and Shobhit Srivastava. Socio-economic inequality and spatial heterogeneity in anaemia among children in India: Evidence from NFHS-4 (2015–16). *Clinical Epidemiology and Global Health*, 2020.

[14] Nonita Dhirar, Sankalp Dudeja, Jyoti Khandekar, and Damodar Bachani. Childhood Morbidity and Mortality in India – Analysis of National Family Health Survey 4 (NFHS-4) Findings. *Indian pediatrics*, 55(4):335–338, 2018.

[15] Avina Sarna, Akash Porwal, Sowmya Ramesh, Praween K Agrawal, Rajib Acharya, Robert Johnston, Nizamuddin Khan, et al. Characterisation of the types of anaemia prevalent among children and adolescents aged 1–19 years in India: a population-based study. *The Lancet Child & Adolescent Health*, 4(7):515–525, 2020.

[16] UNICEF. *The State of the World's Children 2019. Children, Food and Nutrition: Growing well in a changing world*. UNICEF, New York, 2019.

[17] Rupam Bharti, Preeti Dhillon, and Pralip Kumar Narzary. A spatial analysis of childhood stunting and its contextual correlates in India. *Clinical Epidemiology and Global Health*, 7(3):488–495, 2019.

[18] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[19] Luc Anselin. Local indicators of spatial associationlisa. *Geographical analysis*, 27(2):93–115, 1995.

[20] Laxmi Kant Dwivedi, Kajori Banerjee, Nidhi Jain, Mukesh Ranjan, and Priyanka Dixit. Child health and unhealthy sanitary practices in India: evidence from recent round of national family health Survey-IV. *SSM-population health*, 7:100313, 2019.