Collaborative Design of Visual Analytic Techniques for Survey Data for Community-based Research in Public Health

Jaya Sreevalsan-Nair* Shivam Agarwal
Reddy Rani Vangimalla Sanat Ramesh
Graphics-Visualization-Computing Lab (GVCL) and
E-Health Research Center (EHRC), IIIT Bangalore (IIIT-B), Karnataka, India

Nirmala Murthy
Foundations of Research in Healthcare
Systems (FRHS), Bangalore, Karnataka,
India

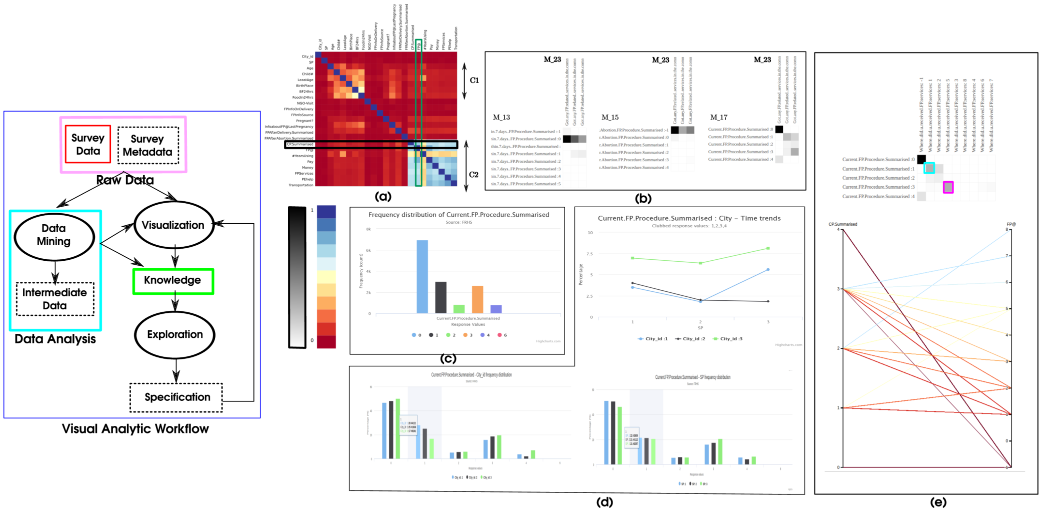

Figure 1: (Left) Our proposed workflow, has been adapted from sense-making loop for visual analytics [5], for exploration of the raw data from surveys. Using both survey data and metadata as the input data (magenta), we propose adding data mining (cyan) to the visual analytic workflow, where we visualize the data mining observations as intermediate data (clusters, matrices). (Right) Visualization techniques included in SurveyVis enable in-depth analysis of the effectiveness of the Urban Health Initiative (UHI) India, for family planning (FP). (a) The association matrix shows clear clusters of variables in the survey, and important variables, such as, M_17 (a metric for the uptake of FP services) is highlighted in black and M_18 (source of FP services) in green. (b) The GPLOM plots of M_23 (on accessibility of FP services) vs M_13 (type of FP method adopted), M_15 (type of FP method adopted after the last failed pregnancy), and M_17, show trends in accessibility of contraceptive methods. (c) The chart shows trends in quality or types of contraceptive methods by using the cumulative trends in M_17. (d) Charts for spatio-temporal analysis of M_17 give the trends in the uptake of contraceptive methods. (e) Parallel coordinate plot is used for multivariate analysis for finding trends in establishing outlets for socially marketed contraceptive measures.

Abstract

Visual analytics is widely adopted in iterative data science/analytic workflows where the human-in-the-loop uses the visualizations to make sense of the data. We are interested in the in-depth analysis of the surveys conducted for studying the effectiveness of public health programs, for which we propose the use of visual analytics. Here, a collaboration between researchers in public health and visualization has led to co-creation of workflows enabling visual analytic techniques for subject-centric raw data from the surveys. We build an appropriate data model which feeds into our proposed visualization techniques. We present a prototype implementation of our tool, SurveyVis, and demonstrate the usage of our tool in analyzing a public health program deployed in India.

Index Terms: H.5.2 [Information Interfaces and Presentation (HCI)]: User Interfaces—User-centered design; I.3.8 [Computing Methodologies]: Computer Graphics—Applications; K.4.3 [Computers and Society]: Organizational impacts—Computer-supported collaborative work;

1 Introduction

Community-based research in public health has been an outcome of *critical reflection* of various aspects of the field of public health [4]. Participatory forms of research have been a mainstay in the domain of public health, and our work in itself is an example of collaboration between public health researchers and computing/visualization experts. Our work on techniques and tools for data exploration and analysis resonates with some of the key principles of research in public health. The principles of public health research are collaborative research, knowledge integration for mutual benefit of all stakeholders/partners, co-learning environment for research, and dissemination of key findings to all partners. Systematic collection of data, e.g. through survey instruments, and its analysis contribute to public health *surveillance* [1]. Survey data pertaining to public health programs is important for evaluating the effectiveness of the public health programs, using representative data from sampled regions/sites during specific time periods.

Given the significance of surveys in capturing public health data, we focus on the visual analytics of survey responses. We refer to the collection of responses for the survey questionnaire by each respondent and the survey metadata, as *raw data*. For public health data, data analytic and visualization tools exist separately [7], and currently there is a need to integrate these two sets of tools. Owing

*Corresponding author: e-mail: jnair@iiitb.ac.in ; this work has been supported by a grant from the Foundations of Research in Health Systems (FRHS). Dr. M. Prakash, formerly with FRHS, was instrumental in providing clean datasets for the work. The authors are grateful to the support from members of GVCL, EHRC, IIIT-B, and FRHS.

[1] N. Allee, K. Alpi, K. Cogdill, C. Selden, and M. Youngkin. Public health information and data: a training manual. National Network of Libraries of Medicine and National Library of Medicine (US), 2004. http://www.phpartners.org/pdf/phmanual.pdf

to the gap in such tools, analysis of public health data continues to be limited to descriptive statistics and does not consider the raw data in its entirety. Our work, contrarily, facilitates exploration and analysis of the raw data, by providing affordable, efficient, and popular visualization techniques for community-based research. Our proposed techniques blend with other data analytic techniques in a data science workflow [2], which is shown in Figure 1.

2 Our Approach

We propose that the survey responses be modeled as multivariate data and the survey metadata as geospatio-temporal attributes. Even though the model is straightforward, such a model has not been used so far in survey data analysis, to our knowledge. This model is the backbone of our proposed visualizations. It makes our proposed visual analytic tool generic, so that it can be used with survey data for other public programs, and is extensible to similar requirements in other geographical regions. Slingsby et al. have proposed a visual analytic tool for internal monitoring and external communication of citizen survey outcomes in England [9]. Their tool visualizes statistical analysis of the survey data and is primarily targeted to survey administrators and public health researchers, and secondarily for public consumption. Their tool is not for subject-wise analysis.

A survey questionnaire predominantly consists of multiple choice questions (MCQs), which are of different types based on the number of answers they expect [1], namely, single or multiple answers. For example, questions such as,"Did you have access to any information sources?" expect a single response from its binary or tertiary choices (e.g., yes/no/maybe questions) response type, whereas questions of the type "Which of the information sources did you have access to?" expect multiple responses. In our data model, every question is treated as a variable in the multivariate data. MCQs with multiple responses are nested multivariate data, synonymous to vector data, which we visualize as "details-on-demand" [8].

2.1 Collaborative Design

The need for yet another visualization software in our work stems from four reasons. *Affordability* of existing holistic tools such as Tableau in developing countries, *minimalism* in co-created tools, *familiarizing* public health researchers with appropriate visualization techniques, and *co-ownership* [2] of the tool, which improves its adoption. Collaborative design and data-driven design are two different characteristics of our choice of visualizations of raw data. We integrate two data mining methods, namely, (a) user-defined clustering of the choices of a MCQ to give coarser set of choices; and (b) computation of the association matrix using normalized mutual information (NMI) [10] between variables with categorical and numerical values.

For visual analytics of survey data, we modify the requirements in [11], which were for a similar visual analytic tool created in collaboration with climate researchers. Our requirements are similar to theirs, primarily due to the similar nature of the data (i.e., being multivariate with geospatial context). We modify the requirements to include those based on time, subject-wise analysis, and data processing and mining. Thus, our broad requirements from the visual analytics of survey data are interaction with raw data (R1), summarization of data using a subset of variables (R2), visualization of inter-variable relationships (R3), support for geospatial and temporal contexts (R4), support for subject-wise analysis (R5), support for linked views (R6), and support for data processing and mining (R7).

Our visualization tasks are organized as multivariate analysis, contextual analysis based on survey metadata, subject-centric analysis and summaries/overview. These visualizations demonstrate the five W's[3] of journalistic reporting for organizing information [9, 12]. "*Who*" encodes the variables which are characteristics of the respondents (i.e. which cross-section of the population?), "*what*" encodes the variables pertaining to effectiveness of the public health program, and "*why*" is captured using the relationships between different variables. "*Where*" and "*when*" are the geospatio-temporal variables of the survey, which are the metadata. The "what" and "why" analysis fulfill R1-R3 and R7; both "when" and "where" fulfill R4; and "who" fulfills R5. Additionally, the visualizations facilitate both summarization as well as data exploration, for which we use the Visual Information Seeking Mantra [8], where, the "overview" visualizations fulfill R2 and R3; and "zoom and filter" and "details-on-demand" fulfill R1.

Thus, we propose the following visualizations and user interactions: (a) parallel coordinates [6] and generalized plot matrices (GPLOMs) [3] to explore subject-wise multivariate data; (b) a map-based (spatial) visualization indicating location-wise frequency of subjects/patients; (c) a double-ended slider widget to select a time interval; and (d) three different options for overviews. Overviews include subject-centric heatmap (or cluster map) for subject-wise analysis, an association matrix of variables for multivariate analysis, and MCQ-based frequency distribution plots demonstrating spatial, temporal, or spatio-temporal trends. The frequency distribution plots also appeal to the familiarity of visualizations used in the existing data science workflow of the public health researchers. Our results in Figure 1 are from our case study of a public health program in India, namely, the Urban Health Initiative India[4]. We have specifically studied the influence of the program in family planning (FP) practices in 11 cities.

3 Conclusions

We have proposed and designed visual analytic techniques for exploring and analyzing survey data for public health programs. Our motivation has been to build a tool as a collaborative effort between the visualization experts and public health researchers. Our prototype implementation, SurveyVis, has been used to demonstrate a case-study. SurveyVis[5] is intended to be a publicly available tool, which is currently being used by public health researchers at FRHS.

References

[1] T. Chiasson and D. Gregory. Data + Design: A simple introduction to preparing and visualizing information. `https://infoactive.co/data-design`, 2014.

[2] P. Guo. Data science workflow: Overview and challenges. *Communications of the ACM*, 2013.

[3] J.-F. Im, M. J. McGuffin, and R. Leung. Gplom: the generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Trans. on Visualization and Comp. Graphics*, 19(12):2606–2614, 2013.

[4] B. A. Israel, A. J. Schulz, E. A. Parker, and A. B. Becker. Review of community-based research: assessing partnership approaches to improve public health. *Ann. review of pub. health*, 19(1):173–202, 1998.

[5] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pp. 154–175. Springer, 2008.

[6] J. Li, J.-B. Martens, and J. J. Van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Info. Visualization*, 9(1):13–30, 2010.

[7] K. Sedig and O. Ola. The challenge of big data in public health: an opportunity for visual analytics. *Online journal of public health informatics*, 5(3), 2014.

[8] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pp. 336–343. IEEE, 1996.

[9] A. Slingsby, J. Dykes, J. Wood, and R. Radburn. Designing an exploratory visual interface to the results of citizen surveys. *International Journal of Geog. Info. Sci. (IJGIS)*, 28(10):2090–2125, 2014.

[10] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[11] Z. Zhang, X. Tong, K. T. McDonnell, A. Zelenyuk, D. Imre, and K. Mueller. An interactive visual analytics framework for multifield data in a geo-spatial context. *Tsinghua Science and Technology*, 18(2):111–124, 2013.

[12] Z. Zhang, B. Wang, F. Ahmed, I. Ramakrishnan, R. Zhao, A. Viccellio, and K. Mueller. The five ws for information visualization with application to healthcare informatics. *IEEE transactions on visualization and computer graphics*, 19(11):1895–1910, 2013.

[2] This is similar to that of *value co-creation* in *customer engaged behavior* through *co-development*, used in marketing literature (E. Jaakkola and M. Alexander, Journal of Service Research, 17(3):247261, 2014)

[3] Five W's refer to variables that answer the "what", "who", "where", "when", and "why" type of questions in the data

[4] `http://www.uhi-india.org.in/`

[5] Demo version and image gallery are available at http://datavis.mybluemix.net