*SDR*Querier: A Visual Querying Framework for Cross-National Survey Data Recycling

Yamei Tu, Olga Li, Junpeng Wang, Han-Wei Shen, Przemek Powałko, Irina Tomescu-Dubrow, Kazimierz M. Slomczynski, Spyros Blanas, and J. Craig Jenkins

Abstract—Public opinion surveys constitute a widespread, powerful tool to study peoples' attitudes and behaviors from comparative perspectives. However, even global surveys can have limited geographic and temporal coverage, which can hinder the production of comprehensive knowledge. To expand the scope of comparison, social scientists turn to *ex-post* harmonization of variables from datasets that cover similar topics but in different populations and/or at different times. These harmonized datasets can be analyzed as a single source and accessed through various data portals. However, the Survey Data Recycling (SDR) research project has identified three challenges faced by social scientists when using data portals: the lack of capability to explore data in-depth or query data based on customized needs, the difficulty in efficiently identifying related data for studies, and the incapability to evaluate theoretical models using sliced data. To address these issues, the SDR research project has developed the *SDR*Querier, which is applied to the harmonized SDR database. The *SDR*Querier includes a BERT-based model that allows for customized data queries through research questions or keywords (Query-by-Question), a visual design that helps users determine the availability of harmonized data for a given research question (Query-by-Condition), and the ability to reveal the underlying relational patterns among substantive and methodological variables in the database (Query-by-Relation), aiding in the rigorous evaluation or improvement of regression models. Case studies with multiple social scientists have demonstrated the usefulness and effectiveness of the *SDR*Querier in addressing daily challenges.

Index Terms—Survey data recycling, data harmonization, visual data query, social science, visual analytics.

1 Introduction

Comparative surveys are a powerful tool that researchers in many fields, such as sociology, political science, economics, demography, etc., employ to study how the individual-level conditions (e.g., age, gender) combine with contextual factors (e.g., democracy, economics) to shape social phenomena across cultures and time [1]. While there is a treasure of free, publicly available international survey projects, users often encounter difficulties when doing comparative analyses because single survey projects only cover a limited number of countries and time periods. To expand the scope of comparison, social scientists increasingly harmonize data from existing cross-national datasets that measure the same concepts for different populations and/or at different times into a new integrated database [2], [3]. The Survey Data Recycling (SDR) project is an active research effort that develops ex-post harmonization methods [4], [5] to recode, rescale, or transform variables from 22 international surveys into a single integrated dataset with consistent scales [6]–[8]. The SDR harmonized database is available online through the SDR data portal, allowing scientists to access the data and conduct further analysis. While large-scale harmonized databases have the potential for innovative comparative research, they can also present significant challenges in *understanding* and *exploring* the dataset, as well as *evaluating* their theoretical models built on top of the sliced data. The current online data portal does not provide effective means for *understanding* the complex structure and various types of variables, making it difficult for scholars to choose an appropriate set of variables from the data for their analyses. It is also difficult for researchers to *explore* data availability, considering factors such as source data quality or harmonization features, even with accurate filtering conditions. Lastly, survey data is often used to *evaluate* statistical models proposed by scientists, but it can be challenging to retrieve useful information from the available high-quality data to assess the fit of these models against empirical data.

We believe that visualization is key to addressing the challenges mentioned above from three perspectives. First, since SDR data is harmonized from a set of meta-data (e.g., survey questionnaires, codebooks, and data dictionaries), visualizing the structure of the harmonized data and relating the unstructured text to harmonized variables can significantly improve the effectiveness of data queries. Second, both the meta-data and harmonized data can suffer from issues with data quality, so it is important to visually demonstrate data availability to avoid spending time on less reliable problems and to identify promising research topics with solid data support. Third, visualizations with convenient user interactions can greatly aid in exploring the hidden relationships between meta-data and harmonized data in the dataset. Despite the potential benefits of visualization, we have found that it has not been fully utilized in social science applications. For example, bar charts are often used to show the temporal coverage of surveys, but they do not reveal the surveys' spatial coverage at the same time. Scatterplots

- *Y. Tu, H.-W. Shen, S. Blanas, J.C Jenkins are with the Ohio State University, Columbus. E-mail: {tu.253, shen.94, blanas.2, jenkins.12}@osu.edu.*
- *J. Wang is with Visa Research. E-mail: junpeng.wang.nk@gmail.com.*
- *O. Li, P. Powałko, I. Tomescu-Dubrow, K. M. Slomczynski are with Polish Academy of Sciences Email: {olga.li, ppowalko}@ifispan.edu.pl, {dubrow.4, slomczynski.1}@osu.edu*

are commonly used to qualitatively present the correlation between target variables, but they do not reflect the quality of the underlying data and may present biased results [9].

To address these limitations, we have collaborated with social scientists to develop a new visual analytical system called *SDR*Querier, using the SDR database as a pilot case. The system offers three levels of information queries through visualization and interactions. To facilitate *understanding*, we have proposed a question-driven variable recommendation for efficient data exploration. Users can query relevant variables by inputting their research questions. Based on the related variables, users can then perform accurate queries to check for available data. For *exploring* data availability, *SDR*Querier includes a new design called the Temporal Availability Profiler, which dynamically displays multi-faceted information. Additionally, our system is able to evaluate models and suggest methodological variable improvements by answering the following questions: *What are the relationships between the variables selected for the regression model*? and *What other variables should be included in the model?* We have also conducted extensive case studies with domain experts. In summary, the contributions of our work are as follows:

- We abstract the challenges in analyzing harmonization survey data and propose **a visual analytics system**, *SDR*Querier, to address them. It contains three visual components to assist in different stages: understanding, exploring, and analyzing.
- We propose a new **question-driven variable recommendation** for data understanding, which helps users efficiently identify variables of interest.
- We design the **Temporal Availability Profiler** to visualize available survey projects from different levels and perspectives. Through case studies, we have demonstrated the novelty and usefulness of this design.

2 Related Work

2.1 Survey Data Visualization

To concisely present information, social scientists often use visualizations to create static reports [9]. For example, they may use bar charts or pie charts to display proportions or distributions of different categories for discrete categorical data [10], or use boxplots and error bars to show statistical measurements of quantitative data [10], [11]. These static charts can only convey information pre-selected and filtered by the creators of visualization. To allow human-in-the-loop of the information-seeking process, some visualization tools allow users to flexibly explore survey data with their own questions, e.g., NESSTAR [1], SDA [2]. Jones et al. [10] developed an interactive system for presenting quantitative social and environmental survey data to help explore and understand. Existing works mainly aim to understand the content of traditional survey data through visualization. However, the harmonized data structure is more complex than traditional survey data because it requires recoding, rescaling, and transforming variables when integrating surveys. To the best of our knowledge, *SDR*Querier is the first interactive system that explores large, complex, and high-dimensional harmonized datasets through visualization and visual queries.

1. http://www.nesstar.com/
2. http://sda.berkeley.edu

2.2 Time-Varying Multivariate Data Visualization

Time-varying multivariate data depict how various features evolve over time, and these evolution patterns often provide valuable insights into the data generated by different domains [12], [13]. Because time can be considered either linear or cyclic, visualizations can be categorized into two groups. The Spiral Graph is more efficient in discerning periodic patterns [14]–[16]. For sequential visualization [17], [18], the Theme River [19] is one of the most popular visualizations that maps the frequencies of multiple topics at each time step to the widths of colored currents in the river, depicting the thematic evolution of documents based on a river metaphor. Our novel visual design uses the same metaphor as the Theme River, presenting the data availability as a flow. However, we allow the information to be presented from multiple perspectives, where Theme River and other methods are not applicable [14], [20]. Many visualization tools are designed for capturing multi-level information of time-varying data in the literature [21]. Dasgupta et al. [22] developed coordinated views to illustrate the evolution of chemical species for geologists to observe interactions, including parallel coordinates and matrix views. Wang et al. [23] designed a spiral graph for analyzing the sentiment of time-varying Twitter data. Pena et al. [24] compared three visualizations of geo-temporal multivariate data, considered the most related work to this work. The difference is that geological and temporal information falls into two levels of our analysis in social research. We first illustrate the temporal and high-level spatial availability, and the detailed geological information is presented later to scientists.

2.3 BERT for Information Retrieval

Information Retrieval (IR) is the process of extracting useful information from a collection of resources. Traditional IR methods often rely on lexical attributes, such as word frequency. In contrast, pre-trained language models (PLM) consider the natural language semantics, thus achieving state-of-the-art performance on many tasks. Among those tasks, the most related ones to our automatic recommendation model are search-related, such as document retrieval and question answering [25]–[27]. BERT is a PLM that achieves state-of-the-art performance on many NLP tasks [28], such as question answering, sentiment analysis, and named entity recognition. Researchers have applied BERT to ad-hoc document retrieval by ranking the documents based on inference scores computed for each document and query [29]–[31]. However, different methods have been used to compute these scores. Yang et al. [29] tackled the challenge of long documents by inferring individual sentences first and then computing document scores based on the sentences, while Jiang et al. [32] focus on cross-lingual document retrieval between English queries and foreign-language documents. The above works aim to improve the performance or address challenges in applying BERT to information retrieval. In contrast, our proposed system, *SDR*Querier, aims to use BERT for variable recommendation in information retrieval from two perspectives and identify different scenarios in which this model can be applied.

TABLE 1
The SDR dataset includes both target (T.) and control variables (C.) that can be organized into expert-defined sociology concepts.

Concept		Label	Example T.	#T.	#C.
political attitudes	trust in political instructions	trust in the parliament	T_TRPARL_DISTRIB	2	3
		trust in the legal system	T_TRLEG_DISTRIB	2	4
		trust in political parties	T_TRPARTY_DISTRIB	2	3
		trust in the government	T_TRGOV_DISTRIB	2	4
	interests	interests in politics	T_INTPOL_DISTRIB	2	4
political behavior		participation in demonstration	T_DEMONST	1	6
		signing petitions	T_PETITION	1	3
social-demographics		age	T_AGE	2	2
		gender	T_GENDER	1	0
		living in metropolitan	T_METRO	1	2
		education	T_EDU	2	7

3 Background

Ex-post survey data harmonization is the process of combining information from international survey projects and other sources into a single, integrated dataset. The resulting data are organized in a tabular format, with each column representing a specific question from the original questionnaire, known as a *variable*, and each row represents a respondent's responses to all of the variables. The original variables taken from different surveys for harmonization are called **source variables**, and the resulting indicator in the harmonized dataset, produced from a series of source variables measuring the same concept in different surveys, is called a **target variable**. Ex-post harmonization procedures are necessary because the characteristics of source questions related to a particular concept, such as wording and answer options, often vary between surveys. These procedures help to ensure that the resulting target variable is consistent and comparable across different surveys. When transforming source variables into target variables, the SDR team creates **harmonization controls** to capture primarily inter-survey methodological variability in the formulation of the source questions. These measures are designed to help ensure the validity and reliability of the constructed target variable. The SDR database also provides a set of methodological indicators, i.e., source data **quality controls**. These variables capture biases and errors that stem from differences in the quality of the source survey data.

The SDR dataset v1.0 contains $4,402,489$ rows with 56 target and control variables. They can be aggregated into several expert-defined social concepts, summarized in Table 1. For instance, the target variable `T_TRPARL_DISTRIB` represents trust in parliaments. Different surveys may ask about this concept in various ways, such as with a yes or no answer or by asking respondents to rate their trust on a scale from 0 to 10. To account for this variability, a harmonization control variable called `C_TRPARL_LENGTH` is created to indicate the length of the scale used in the original questionnaire. This variable would have corresponding values of 2 or 11 in the examples provided. In addition, each respondent has 3 quality controls related to the original survey data, e.g., `QR_DUPLICATE_SVY` indicates whether the original survey data contain non-unique records.

4 Requirement Analysis

The SDR portal enables scientists to download harmonized survey data that can be used for comparative empirical research. The challenge is how to help them identify what the related data are and how to use them. We have collaborated with four domain experts for more than one year to identify the requirements, summarized as follows:

- **R1: Identifying related variables to the user's research topic.** Given the large dimensionality of the harmonized dataset, identifying the related columns is important and necessary to acquire meaningful data from the portal. To provide enough guidance for experts, *SDR*Querier should:
  - *R1.1: Give the variable recommendation based on users' needs.* Automatic variable recommendation can help scientists avoid unnecessary exploration and focus on more important variables.
  - *R1.2: Exhibit data provenance of harmonized target variables.* Showing what source variables each target variable links to helps scientists understand the logic and meaning of each target variable. Simultaneously, this information fosters researchers' trust in the harmonized data, as it speaks to the transparency of the harmonization process.
  - *R1.3: Present an overview of the harmonized dataset.* Due to the complexity of harmonized data, it is necessary to present multi-faceted information (e.g., variable types and relationships) in order to fully understand it.
- **R2: Revealing data availability for decision support.** Typically, scholars who conduct quantitative comparative survey research seek data that meet specific conditions. We should reveal data availability to:
  - *R2.1: Facilitate target variables selection.* While several target variables may fit a given research problem, their availability varies a lot. To choose a proper one, scientists need to know their individual and joint availability.
  - *R2.2: Assist with decision making.* Once available data are identified, it is important to assist researchers in deciding whether these data meet formal requirements for the regression analysis. This can be done by providing multi-faceted information, e.g., which data have quality issues?
- **R3: Retrieving underlying patterns for hypothesis testing.** Social scientists use survey data to examine if and to what extent there are empirical support for theoretically-informed relations between various variables, which can be assisted by the hidden patterns in the data. We should:
  - *R3.1 Validate the selected target variables.* Hypotheses propose some associations or causal relationships between variables of interest. Revealing relational patterns from target data can preliminarily evaluate the hypotheses.
  - *R3.2 Describing the potentially related variables to improve the regression model.* Relations between target variables should also consider the potential role of methodological variables. Scientists should take them into account when building theoretical models to test hypotheses.

5 Visual Analytics System: *SDR*Querier

Motivated by these requirements, we design and implement *SDR*Querier with three coordinated visual components to enable multi-granularity queries for social scientists.

5.1 Approach Overview

Figure 1 displays an overview of our framework. We summarize the domain requirements into three challenges in

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3261944

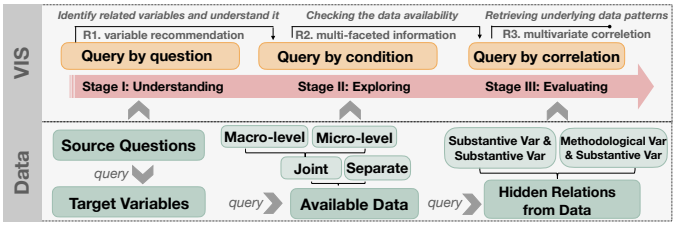


Fig. 1. The overview of our system, which contains three query components with corresponding data.

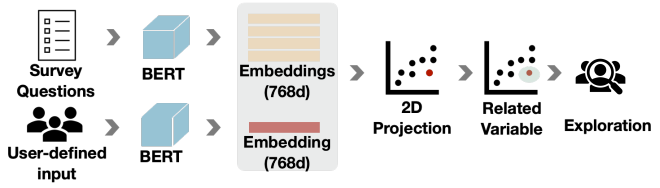


Fig. 2. The pipeline of using the BERT for variable recommendation.

different stages during the analysis pipeline: *understanding, exploring, and evaluating*. To solve the challenges, we propose a framework that contains three corresponding modules. **First**, inspired by conversational Artificial Intelligence, we train a BERT-based model to generate variable recommendations based on the user's input text, either keywords or sentences describing their information of interest (**R1.1**). This process is defined as *Query-by-Question*. Later, the recommendation is combined with visualization and interactions to facilitate harmonized data understanding (**R2.1, R2.2**). **Second**, in the *Query-by-Condition* module, we perform information retrieval based on specific filtering conditions. In order to show the multi-faceted information from the retrieved data, we design a new visualization, Temporal Availability Profiler, to assist scientists in deciding whether data are sufficient to use considering data diversity, coverage, and quality issues (**R2**). **Lastly**, computing the relational patterns from available data samples can verify whether the expected patterns exist or not, which in turn helps scientists to test their hypotheses and choose the variables for their theoretical models (**R3**), defined as *Query-by-Relation*.

5.2 Query-by-Question (QBQ)

Although target variable names in the harmonized data have been carefully chosen, it can be difficult to identify the theoretical concept from the abbreviated names quickly. As explained in section 3, each target variable is summarized from a set of survey questions in the questionnaires. Therefore, the survey questions provide good contexts, accurately reflecting the meaning of target variables. e.g., `T_DEMONST` (a target variable) can be characterized as *authorized demonstrations in democratic countries* or *unauthorized activities in non-democratic countries* given different political backgrounds.

Inspired by conversational AI, we train a BERT-based classification model on survey questions to predict the target variables. With such a model, we can infer the target variable from various text inputs, e.g., research questions, descriptions, or a set of keywords for a sociological concept. For example, when a researcher studies if life conditions can influence political participation, they might type in the sociological concept, i.e."political participation", or the descriptions of life condition indicators, i.e."how much are you satisfied with your life?" or "are you living in metropolitan or not?" to retrieve the related target variables. Based on the model, we can recommend a target variable in two ways (**R1.1**): (1) *the hard recommendation*, which outputs the target variable with the highest probability from the classification model; (2) *the soft recommendation*, which converts one-to-one prediction problem to a one-to-many clustering issue by allowing users to flexibly explore the semantic similarity between their inputs and survey questions.

5.2.1 BERT-Based Model for Target Variable Prediction

To automate the QBQ process, we train a BERT-based model to relate the survey questions with target variables. The model (1) takes a survey question as input, (2) converts input into a $768d$ embedding, which represents the entire text sequence and is then used for sequence classification tasks, and (3) classifies the embedding to a target variable. A pre-trained BERT model is employed to perform (1)→(2), and a classification layer is appended to the model to conduct (2)→(3). A set of question and target variable pairs, labeled by our social scientists, are used to train the classification layer with a cross-entropy loss.

5.2.2 BERT-Based Model for Soft Recommendation

The *soft recommendation* qualitatively measures the semantic similarity between user-defined text input and the survey questions. We extract the hidden states, i.e., embeddings from the trained model, which is promised to capture the semantic information. As shown in Figure 2, the embeddings of users' input and survey questions are extracted and jointly projected to 2D for visual exploration. tSNE [33] is employed here to interactively update the projection result, given its superior performance over UMAP for non-linear projections. We perform the embedding updates in an iterative manner following algorithm 1, aiming to reduce the running time and acquire stable results.

Algorithm 1: Embedding Iterative Updating Alg.

Input: question projection coordinates at timestamp t, $\mathbb{P}^t$: (p_1^t,...,p_N^t), new input sentence: s
Output: whole projection coordinates set $\mathbb{P}$ at timestamp t+1, $\mathbb{P}^{t+1}$: (p_1^{t+1},...,p_N^{t+1}, p_s^{t+1})

e_s^{t+1} = BERT(s) //Generate embedding for input s
p_s^{t+1} = random_init(e_s^{t+1}) //Initialize position for s
$\mathbb{P}^t$ = (p_1^t,...,p_N^t, p_s^{t+1}) //Adding coordinate of s into $\mathbb{P}$
$\mathbb{P}^{t+1}$ = tSNE(init=$\mathbb{P}^t$) //Init tSNE with $\mathbb{P}$ from time t

5.2.3 Visual Design for Understanding Harmonized data

There are three views to help understand the harmonized data (**R1**): the *Scatterplot* (Figure 6-a_1), the *Information Table*(Figure 6-a_2), and the *Circular Graph* (Figure 6 a_3-a_4).

The *Scatterplot* displays the embedding projection result, revealing semantic similarity among survey questions and user inputs. Each dot represents a source question and is colored based on its related target variable. As shown in Figure 6-a_1, questions of the same color form clusters, verifying that our BERT model captures their semantic similarity. Users can brush the dots of interest, which will automatically update the *Information Table*.

The *Information Table* connects source and target information, helping scientists identify which variables to query from the SDR portal. The columns of the tabular data are *year*, *survey wave*, *source question*, *target variable*, *label of target variable*. As confirmed by domain experts, individual source question varies across surveys. Hence, it is helpful to present this variation to scientists in order to help them better understand the data pre-processing process and improve the credibility of the harmonized data **(R1.2)**.

The *Circular Graph* is proposed to handle the complexity and dimensionality of the harmonized data, which targets to: (1) indicate diverse types of variables, including *source-, target-, harmonization control- or quality-variables*; (2) illustrate the relationships of different variables (**R1.3**). As shown in Figure 6-a_3, the circular bar chart represents the target variables. The length of the bar implies the overall availability of each target variable, i.e., how frequently the corresponding target variable is measured in international surveys. The color indicates the topic of the target variables, which is consistent with the color schema used in the *Scatterplot*. Once the user triggers the query from *scatterplot*, only the predicted bar will be highlighted in orange, while others fade out. Several target variables can describe the same topic from different perspectives. For example, `T_HAPPY_11` and `T_HAPPY_DISTRIB` both measure respondents' self-reported happiness, but using different specifications. As described in section 3, target variables capturing the same theoretical concept can share one or several harmonization control variables, which record the variance in source variable properties. These controls are visualized as the orange arcs (●). The number of arcs in the same radial position reflects the number of harmonization control variables in the group. When clicking an arc, the right panel will pop up to show its value distribution with labels of the harmonization control variable (Figure 6-a_4). As proved by experts, showing this information is extremely helpful when querying data from the SDR portal. The intermediate circle (●) conveys that all the target variables are related to the *quality control* variables. The inner network represents the demographics of the respondents. It includes *age*, *birth year*, *sex* of respondents; their color is also consistent with the *Scatterplot*. For example, for *age*, surveys can ask about age in many ways as reflected by the numerous red points (●) in the *Scatterplot*.

5.3 Query-by-Condition (QBC)

Core to social science quantitative comparative research is to assess the extent to which empirical data support their hypotheses. While these postulated hypotheses often refer to specific countries and certain year-range, it is a common practice among social scientists to blindly download the full harmonized data without any filtering conditions (from data portals) and then check if downloaded data fit their research needs, e.g., in terms of country and time coverage. However, the process is inefficient and can be greatly improved if data availability can be effectively and user-friendly checked from multiple perspectives before downloading **(R2)**.

5.3.1 Temporal Availability Profiler

We propose a new design, *Temporal Availability Profiler*, to reveal the availability of the harmonized data at multiple levels. The design comprises two sub-components: (Separate Availability and Join Availability), sharing the same x-axis to reveal data samples density evolution across time (i.e., temporal availability).

The *Separate Availability* view (Figure 6-b_1) illustrates the number of valid samples for **each** target variable over time, helping users to decide among multiple alternative variables **(R2.1)**. The *Joint Availability* view (Figure 6-b_2) exhibits the available samples for **all** the selected target variables, indicating the precise pool of valid samples that can be used to evaluate the multi-variate relations **(R2.2)**.

Before constructing the view, we have condition sets $\mathbb{C}$ and selected target variable sets $\mathbb{T}$. Condition sets are used to filter rows in the harmonized dataset. For example, a scientist wants to study political protests in Russia under Putin's regime. The condition sets $\mathbb{C}$ = ("country=Russia", "year$\leq$2020", "year$\geq$2000"). For *Joint Availability* view, the available samples should satisfy all condition sets and contain data at all target columns. While in the *Separate Availability*, each row indicates one corresponding target variable t_j. The samples in each row should be valid for both condition sets and the corresponding columns.

Given one specific year, the connection between the two sub-components can be summarized as follows:

1) case1: Each target variable t_i has data d_i, and there are also jointly available samples, i.e., $\forall t_i \in \mathbb{T}, d_i \neq \varnothing \rightarrow d_1 \cap d_2 \cap ... d_N \neq \varnothing$. This is the ideal case where there exist data of high quality to use.
2) case2: At least one target variable t_i does not have data, resulting in the lack of jointly available data, i.e., $\exists t_i \in \mathbb{T}, d_i = \varnothing \rightarrow d_1 \cap d_2 \cap ... d_N = \varnothing$. In other words, the lack of available data to use comes from specific variables, helping scientists to decide whether to omit the unavailable variable or impute the missing data.
3) case3: Each target variable t_i has data d_i, but there is no overlap among them, i.e., $\forall t_i \in \mathbb{T}, d_i \neq \varnothing \rightarrow d_1 \cap d_2 \cap ... d_N = \varnothing$. This scenario indicates we have low-quality data since they do not contain all the variables of interest.

5.3.2 Visual Design of Temporal Availability Profiler

As visualized in Figure 6-b_1, *Separate Availability* presents the available data for each variable over time. The color summarizes aforementioned connections with *Joint Availability*, including blue (case1) and orange (case2 & case3). For each variable, the width of the flow illustrates how many samples are covered each year. In the *Joint Availability*, each row represents one valid survey project, which may cover a period of years. A user may click the survey project name to show the background information of each survey. Incorporating survey documentation into *SDR*Querier is highly recommended by domain experts (Figure 7C-D). Given one survey, there is multi-faceted information to be presented properly. Through the discussion with domain experts, they prefer simple but efficient visualization to delicate glyph design for multi-faceted information. To do this, we propose interactions with a responsive bar chart to show information from multiple perspectives.

Responsive Bar Chart The meaning of bar can be embedded as either *macro-level* or *micro-level* by users, defined as the *responsive bar chart*. Scientists can select to see how many respondents are available *(micro-level)* or how many countries are available *(macro-level)* through the drop-down

selector ❯ at the top of this module. To present the country coverage, we further allow users to click the bar for the detailed information on a map, where green means covered country (Figure 6-b_3). Also, we allow two different sorting methods of rows: availability-based and quality-based. For the availability-based method, if a project covers more distinct years, it will have higher availability. For quality-based sorting, we compute a quality indicator q_i for each survey project ($\mathbb{S}$) as follows:

$$quality_i = \frac{\sum_{w_i \in \mathbb{S}} N_{w_i} \sigma(q = 0)}{\sum_{w_i \in \mathbb{S}} N_{w_i} \sigma(\varnothing)} \quad (1)$$

where N is the number of samples, w_i is the waves of a project $\mathbb{S}$ conducted in different years. In each wave, some samples do not have quality issues, which are recorded as quality variables $q = 0$ in the harmonized dataset. The quality indicator is the faction of the samples without quality issues to the total samples given survey.

5.4 Query-by-Relation (QBR)

Social scientists propose theoretically derived hypotheses, which are then tested through statistical models using appropriate data. However, building these models often requires a long time to process the survey data and identify the related variables. Effective and accurate variable selections become crucial.

To accommodate this, we propose a third module, *Query-by-Relation (QBR)*, to query the hidden patterns from data for model verification and improvements. QBR is performed by answering two questions: (1) *what are the relationships between the selected target variables?* Whether the variables are correlated can help scientists to preliminary test their hypotheses. For example, a researcher wants to test whether respondents' resources have a negative effect on an individual's trust in political institutions. Checking the correlation strength among these variables can help scientists to determine whether the selected variables are appropriate for hypotheses testing. (2) *what are the potentially related variables?* In the SDR harmonized dataset, both types of methodological variables can affect the relationship between substantive variables, so they should be included in the regression analysis when constructing regression models. Understanding the correlation between them is important for scientists to include appropriate additional methodological variables in their models. In the end, we designed two subviews in QBR to answer the above two questions.

Correlation Matrix. Driven by **R3.1**, our first subview presents the pairwise correlations for user-selected target variables, allowing scientists to check if these variables are correlated with each other and to further determine what to keep for their regression analysis. We compute several necessary and common statistics for pairwise relations, including the Pearson correlation coefficient, p-values, which can be grouped into categories based on expert-suggested p-value thresholds, and standard errors. To flatten the learning curve of visual encoding, as suggested by our domain experts, we show the computed information with texts and only incorporate two visual channels in the matrix, i.e., position for pairwise relation and responsive color. Users are allowed to select one of the computed information and map it to the color interactively. After several key design iterations with the domain experts, we determined to show one-half of the symmetric matrix to reduce redundant information and avoid unnecessary interpretation of the position.

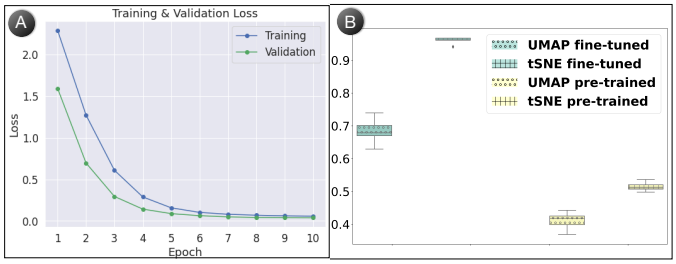


Fig. 3. (A) The training & validation Loss of the BERT-based model. (B) Adjusted Mutual Information(AMI) score for our model and baseline model with two projections, i.e., UMAP and tSNE.

Network Visualization. Variable selection is extremely important when building regression models to test research hypotheses. Due to the dissimilar structure of harmonized data to typical survey data, users may not know what methodological variables should be considered together with substantive variables for robust model analysis. Thus QBR facilitates to query of the complex relations given one pair of target variables (**R3.2**). As shown in Figure 6-c_2, we apply the same color scheme to the nodes as QBQ: harmonization control (●), quality control (●), target (●). We label the significance level for each edge in the network, defined by domain experts. If the correlation coefficient is not defined for an edge, we highlight it with red.

6 Evaluation

This section first evaluates the efficiency and usefulness of the backend algorithms employed in *SDR*Querier. Then, we worked with the domain experts on the case studies to demonstrate how *SDR*Querier can help them in the process of understanding and exploring harmonized data, as well as evaluating social science models via effective visual queries and friendly user interactions.

6.1 Evaluation of the BERT-based Model

We fine-tune a pre-trained BERT model to classify each survey question to the most related target variable, aiming to recommend variables given multiple types of user-defined inputs. To achieve this, both the final prediction and intermediate result, i.e., embeddings are extracted from the tuned model for *hard* and *soft* recommendation, respectively. To perform a comprehensive evaluation, we demonstrate the effectiveness of two recommendations through qualitative and quantitative measures.

6.1.1 Quantitative Evaluation

The performance of *hard recommendation* can be revealed from the successful loss converging pattern and high prediction accuracy. To train the model, we used a dataset of 1591 survey questions that were divided into 15 categories. These categories were either a single target variable from Table 1 or a combination of target variables. The categorization was done by domain experts during the pre-processing phase of harmonization. The dataset is split into 90% training and

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3261944

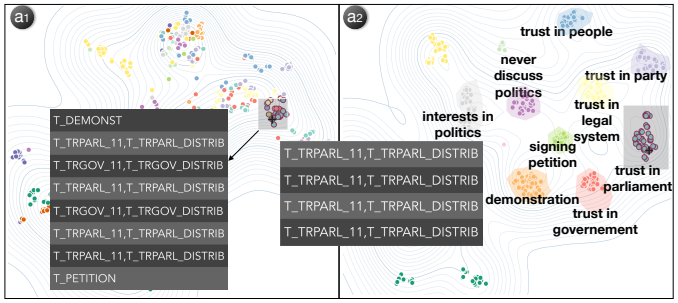


Fig. 4. tSNE space of our (a_2) model compared to (a_1) baseline model.

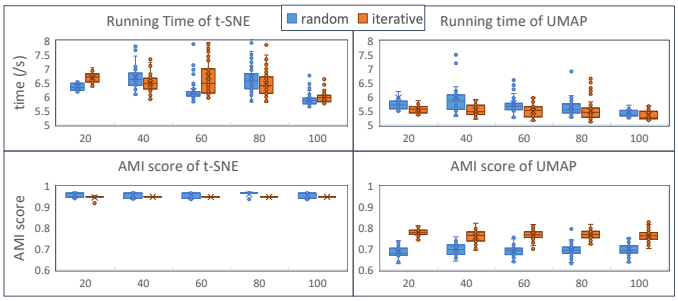


Fig. 5. The running time and Adjusted Mutual Information (AMI) score of tSNE (left) and UMAP (right) with different parameter settings.

10% validation. The model converging process is depicted by the loss shown in Figure 3A. The validation accuracy reaches 99% in the final epoch, which promises good performance on the classification task.

Despite the effectiveness of hard recommendations, some user-defined text may be related to multiple target variables, in which case soft recommendations may be more suitable. Soft recommendations aim to capture semantic similarity in the embedding projection space, allowing users to identify multiple related survey questions. To verify that the clusters formed in the embedding space accurately capture this semantic information, we calculated the similarity between the clustering of survey question embeddings and the groups of ground truth, expecting that questions with the same labels would be grouped into the same cluster in the projection space. We used K-Means clustering (with K=15) to compute the results, and also compared pre-trained BERT embeddings with fine-tuned embeddings as a baseline. In addition to the embedding representation, the performance of the recommendation also depends on the projection method used. Therefore, we also considered two projection methods: tSNE [33] and UMAP [34].

We choose Adjusted Mutual Information (AMI) to quantitatively compare the clusters from embedding and ground truth labels. AMI is widely used to compare different partition/clustering results of the same dataset [35]. It adjusts mutual information (MI) through the following equation:

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{max\{H(U),H(V)\} - E\{MI(U,V)\}} \quad (2)$$

where U, V are the results of two clustering methods, MI computes the mutual information, H is the entropy of the clustering result, and E is the expected value. The AMI value is in the range of [0, 1], with a value of 1 indicating that U and V are identical.

As shown in Figure 3B, color is used to differentiate fine-tuned (●) and pre-trained model (●). Within each group, the projection method is encoded in the textures of the boxplot. It is clear that fine-tuned model improves the results for both tSNE and UMAP because the results of embedding clustering from the fine-tuned model much better match the ground truth. Also, we can observe that tSNE outperforms UMAP in both models. We can conclude that the fine-tuning improves the *soft recommendation* regardless of the projection methods.

6.1.2 Qualitative Evaluation

To demonstrate the qualitative improvements of the fine-tuned model, *scatterplots* with the two clustering results from different models are shown in Figure 4. The figure discloses several advantages of our fine-tuned model (a_2) compared with the pre-trained BERT (a_1). First, there is a clear boundary between clusters in Figure 4-a_2. While in the pre-trained model, several groups interfere with each other, and it is hard to differentiate them without coloring. Second, related groups are also closer to each other in Figure 4-a_2, which corresponds to high-level sociology concepts. As pointed out by the experts, "trust in people", "trust in party", "trust in government", and "trust in legal system" form the concept of "trust in political institutions". "interest in politics" is close to "never discuss politics", both of them depict the attitude of respondents in politics. "demonstration" and "signing petition" comprise the concept of "political behaviors".

We would also like to give an example to compare the quality of *soft recommendation*. As shown in Figure 4, ✚ indicates the projected embedding of a user's input, i.e., "trust in parliament". From the result of the pre-trained model (Figure 4-a_1), it is clear that the queried topic is isolated from multiple topics in the projection. After brushing the surrounding circles, the table presents some non-related target variables. However, in our fine-tuned BERT model (Figure 4-a_2), the queried topic falls into a small cluster of circles. The cluster (i.e., trust in parliament) presents related target variables (`T_TRPARL_11, T_TRPARL_DISTRIB`) for the queried topic. We can conclude that the training not only teaches the model to better predict target variables but also significantly improves the performance of *soft recommendation*.

6.2 Embedding Iterative Updating Algorithm

This section measures to what extent our *Embedding Iterative Updating Algorithm* can stabilize the projection results and how much it can improve the projection efficiency. The study was conducted by running the algorithm with different iterations (i.e., 20, 40, 60, 80, and 100). To show the effectiveness of our algorithm, a baseline of updating the embedding projections with random initialization was also conducted.

We compute both the running time and AMI score of tSNE and UMAP with different parameter settings, as shown in Figure 5. The x-axis indicates the increasing number of iterations utilized in our iterative updating algorithm. The *random initialization* and *iterative updating* are encoded in the colors. Compared to UMAP, tSNE takes a longer time but generates better clustering results. When it comes to the efficiency between *random initialization* and *iterative updating*, tSNE cannot guarantee *iterative updating* will help the algorithm converge faster (upper left). But it is clear that *iterative updating* decreases the running time for UMAP

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3261944

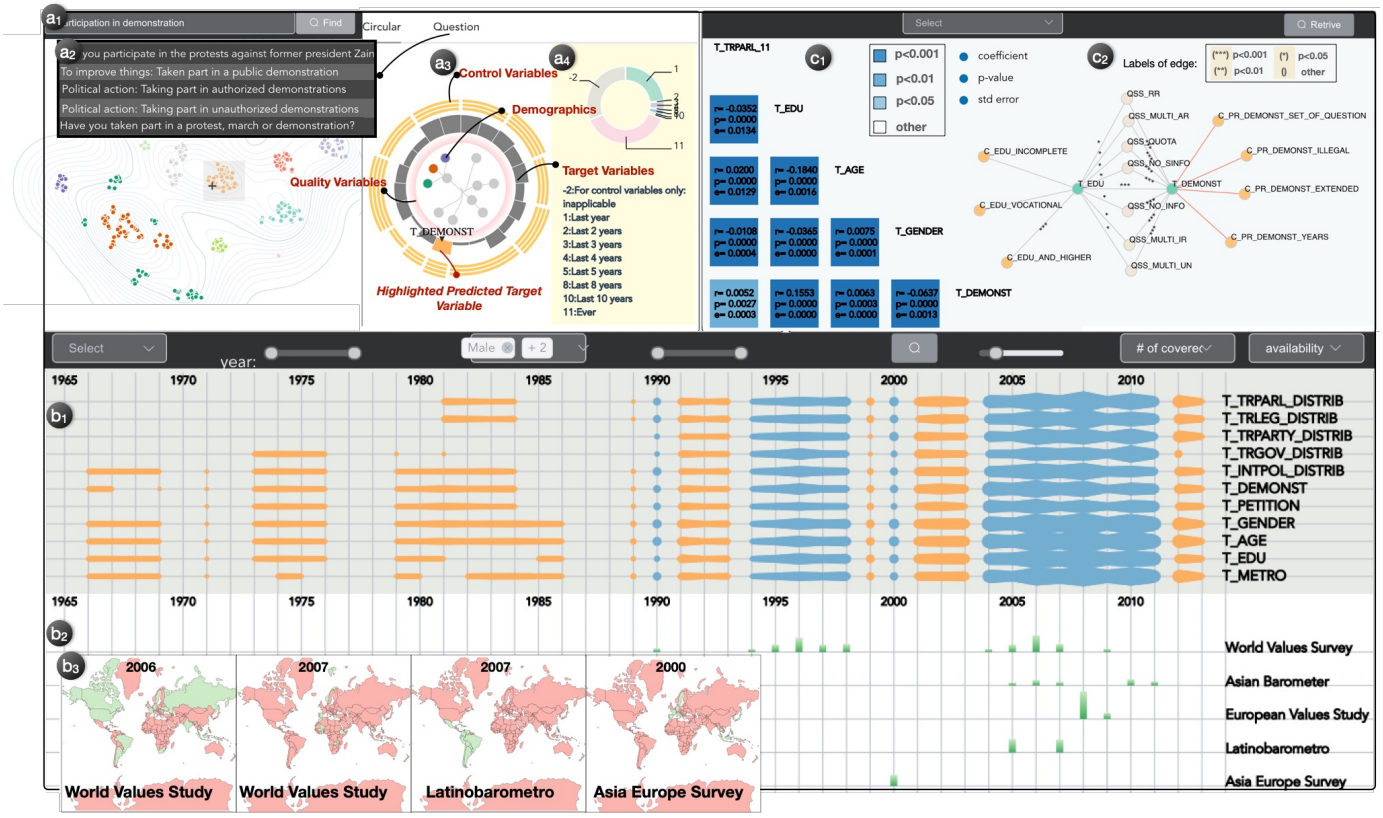


Fig. 6. case study regarding political engagement. (a1-a4) *Query-by-Question* recommends target variables based on user input and allows further visual explorations; (b1-b3) *Query-by-Condition* includes the new design, Temporal Availability Profiler, revealing the multi-faceted information of available data given user's conditions; (c1-c2) *Query-by-Relation* presents the relational patterns from the available data to assist social research.

(upper right). For the AMI score, with tSNE projection, *iterative updating* does not improve accuracy given the fact that *random initialization* already reaches a high level (bottom left). But we can also infer from the figure that *iterative updating* makes the iteration stable by decreasing the variance of running time. For UMAP, *iterative updating* can improve the clustering results regardless of the running iterations (bottom right). In addition, we found that the EIU algorithm significantly improves the projection stability. More information can be found in supplemental materials.

6.3 Data Availability Checking

We illustrate how the Temporal Availability Profiler can be utilized in different scenarios with two cases: one emphasizes how it can help the SDR research group to summarize the possible directions for social science research; the other describes how it can help scientists to decide whether data are sufficient or not for comparative analysis. All the names used in the case studies are pseudonyms for privacy issues.

6.3.1 Case1: Political engagement: attitudes and behaviors

We invite an expert, Arya, who works at the SDR project and deeply understands its harmonization process. She wants to propose future research directions according to data from the SDR regarding political engagement.

Since Arya knows the variables very well, she jumps to the QBC directly to check the data availability. First, Arya selected some target variables to form the high-level theoretical concepts, summarized in Table 1. Without adding filtering conditions, Arya clicked the 🔍 button, the availability of selected variables is displayed in Figure 6 ($b_1 - b_3$) From the color (●) in Figure 6-b_1, the joint available data for all the selected variables are available for years 1990, 1995-1998, 2004-2011. Arya checked each concept separately. In the concept of socio-demographics, the data for "age" and "gender" are pretty complete. However, there are no sufficient data for "living in metropolitan" in the 70s and in the beginning of the 80s. Also, "education" has a deficiency gap during 1982-1984. Given the incomplete socio-demographics of respondents, Arya concluded that those years with data deficiency should be imputed or excluded by researchers. For the concept of political attitudes, "interest in politics" has the greatest data coverage. Based on the observation that "trust in the parliament" and "trust in the legal system" share the same temporal coverage, Arya confirmed that it allows researchers to conduct a study about trust in political institutions from the 80s, even with the gap from 1985-1988. Compared to the "political attitudes", "political behaviors" can be analyzed more comprehensively from the 60s, given the higher temporal coverage.

Drilling down to the country coverage, Arya clicked several available surveys in the *Joint Availability* to check it. The result is presented in Figure 6-b3, it is clear that the available data (●) cover Latin America (Latinobarometro), Europe (World Values Study), and Asia (Asia Europe Survey). She concluded that researchers could conduct various analyses on political attitudes and behaviors controlling for socio-demographic characteristics for a period of over 40 years in different regions. She also pointed out that even in the same survey project conducted in different years, the covered region can fluctuate. The beauty of the SDR harmonized data is that different surveys can complement each other

TABLE 2
The study of protest participation in Russia was conducted using SDRQuerier. Several findings and derived insights from Figure 7B.

Survey	Background	Insights
European Social Survey (ESS)	ESS aims to examine stability and change in social structure, conditions, and attitudes in Europe, which is conducted in most European countries since 2002.	Jimmy noticed that the available surveys contain several well-known and widely used surveys. ESS is one of the examples, shown in Figure 7C.
International Social Survey Program (ISSP)	This survey is a continuous program of cross-national collaboration running surveys, covering multiple issues related to social structure.	Some of them are cross-national collaboration surveys while others are conducted in specific regions.
Life in Transition Surveys (LITS)	LITS was carried out by the European Bank for Reconstruction and Development in collaboration with the World Bank in 2006 and 2010 in central-eastern Europe and the Baltic States, south-eastern Europe, etc.	Jimmy learned that some unavailability comes from the surveys, not the selected target variables. The reason is that they were not conducted as continuous programs. For example, LITS was conducted only in 2006 and 2010.
World Values Survey (WVS) European Values Study (EVS)	WVS focused on a wide range of topics, including economic life, religion, basic values relating to politics. EVS are conducted every 9 years in most European countries since 1981, which examines social, political, and economic values and attitudes, as well as living conditions.	Jimmy identified some relationships between surveys from the background information. WVS can be merged with EVS and used together since they share similar survey questions. From the Separate Availability in \autoref{fig:Russian}B, both cover different years, facilitating extensive evolution analysis.

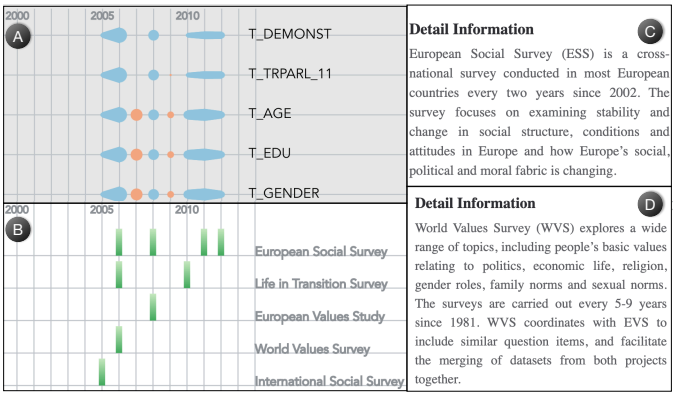


Fig. 7. Case study of protest participation in Russia. (A) available data for each variable. (B) available surveys in responsive bar charts. (C-D) showing survey documentation when users click survey names.

not only temporally but also spatially. For example, when Arya hovered over the bar chart, it showed that World Values Survey (WVS) covered 23 countries in 2006. From the detailed coverage map visualized in Figure 6-b_3, it is obvious that the covered area includes Latin America. However, it only covers 9 countries in 2007 without Latin America, which can be supplemented by another survey project in 2007, i.e., Latinobarometro. Arya summarized that regarding political engagement, the available surveys offer a substantive set of variables for comparative cross-national research.

6.3.2 Case2: Protest participation in Russia

Given the possible research directions of the SDR dataset, Jimmy wants to study protest participation in autocratic states by analyzing contemporary Russia. He needs data that cover Russia in the 2000s∼2010s, the period of the tightening autocratic measures in the country. Based on the literature review, Jimmy proposes that several determinants, such as trust in political institutions, satisfaction with the democratic performance in the country, and economic hardship, can influence protest participation in autocracies differently.

After exploring the system with QBQ, he decided to use `T_DEMONST` as a protest indicator and `T_TRPARL` as an indicator for trust in political institutions. He also identified a set of necessary socio-demographic variables for his study. However, the SDR data lack two variables that measure potential protest determinants, theorized by Jimmy based on the literature review, namely, the subjective perception of democracy and the individual's economic situation. After detecting variables in the SDR data, Jimmy applied two filtering conditions to check the data availability via QBC: (country=="Russian", year∈[2000, 2019]).

While examining the joint availability displayed on Figure 7A, Jimmy identified that the unavailability gaps in 2007 and 2009 come from the lack of `T_TRPARL_11` and `T_DEMONST`. He concluded that sufficient for his study data on Russia are available only from 2005 to 2012 with gaps in "demonstration" and in "trust in parliament". Therefore, he needs to make the decision whether to use biannual data (i.e., data from year 2006, 2008, 2010, 2012), impute the missing data from 2007 to 2009, or search for another dataset because of the missing data. Also, it is clear to see that the size of available samples fluctuates over time, which might be a concern regarding how to use the data properly, as pointed out by Jimmy. The information derived from our visualization helps Jimmy to structure his research, and the prior knowledge of the data availability leads to more reasonable expectations of the final outcome. Drilling down to the *Joint Availability* in Figure 7B, several findings and derived insights are summarized in Table 2.

Based on these observations and conclusions, Jimmy agreed that the combination of these surveys ensures sufficient high-quality samples. However, Jimmy's main concern is the lack of key variables, i.e., two potential determinants of protest participation. He concluded that using only SDR data is not sufficient for his research due to the time and variable coverage limitations. He can probably try to harmonize the data for missing variables by himself from other sources.

6.4 Participation in Demonstrations worldwide

We invite expert Kiara to demonstrate how *SDR*Querier can assist scientists in the social research process following one previous study [36]. It is focused on participation in demonstrations worldwide, which requires samples with a high regional diversity for comparison. Kiara hypothesized that resources and political attitudes have different effects on the levels of participation in demonstrations in democratic and non-democratic countries. Before downloading the dataset to conduct further analysis, she used our *SDR*Querier to check if she could rely on the SDR data.

To begin with, Kiara utilized QBQ to explore the variables she could use. Kiara first typed in the most important

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3261944

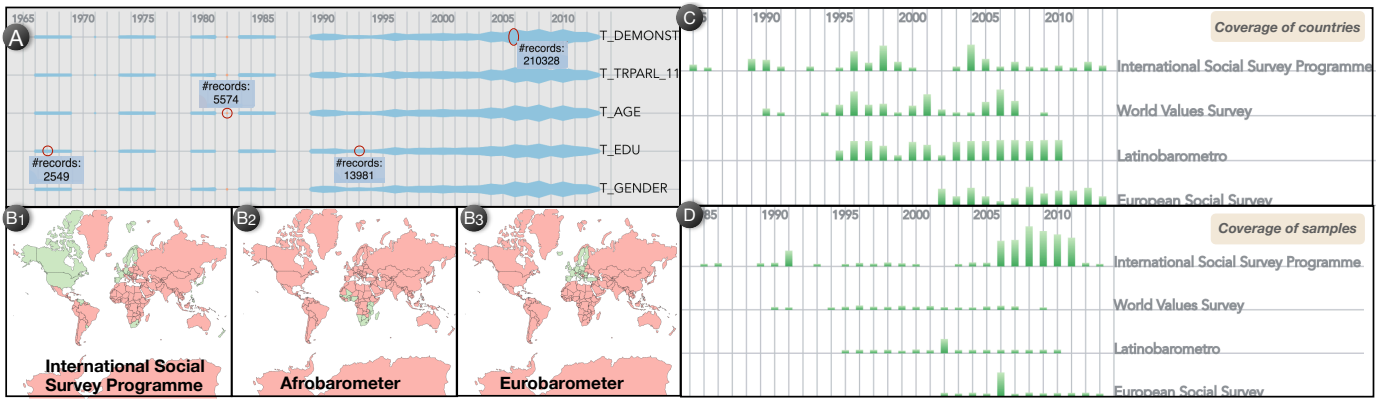


Fig. 8. The study on participation in demonstrations worldwide. (A) *Separate Availability* shows the possible time period of the study. (B1-B3) available data also covers diverse regions and countries. (C-D) top-3 survey projects with the highest availability in (C) macro-level and (D) micro-level.

concept of her research, "participation in demonstration". The prediction from "hard recommendation" is "T_DEMONST" in Figure 6-a_3, which conforms to her domain knowledge. Kiara knew that the meaning of demonstration could vary a lot depending on political backgrounds. Thus, she was wondering what the definition of "T_DEMONST" is in the SDR. From Figure 6-a_1, it is clear that ✚ falls into one cluster, from which Kiara brushed circles (●) to check the detailed information. The table in Figure 6-a_2 showed that the source variables for "T_DEMONST" include not only demonstrations but also protests and marches. It also varied between participation in public, authorized, or unauthorized demonstrations, as well as very specific protests (e.g., against the former president). The source questions also varied in terms of the time length. Generally, respondents were asked in the format of *"Have you performed [action type] in the last [time period]?"*, where [time period] varied across "twelve months", 1, 2, 3, 4, 5, 8, 10 years or ever in different surveys.

Kiara intended to learn the structure of harmonized data via *circular graph*. From there, many visual investigations can be done through interactive exploration. With a brief overview, Kiara found that a four-level hierarchical structure corresponds to the type of variables: socio-demographics, quality control variables, target variables, and harmonization control variables, which differs from the common one-survey data structure. She hovered over the highlighted bar to see the predicted variable's name, i.e., "T_DEMONST". It faces four arcs, which indicates that variations in source questions were captured with four harmonization control variables. She was inquisitive about the meaning of control variables, so she clicked one of them, i.e., C_PR_DEMONST_YEARS. The labels and distributions are shown in Figure 6-a_4. She discovered the variance of the time range is captured well, and the most frequently asked year range is "ever". After the in-depth exploration, Kiara confirmed that the information contained in control variables revealed the high quality of the harmonized data and allowed her to conceptualize participation in demonstrations for her study.

After the preliminary examination of the available variables, Kiara decided to choose T_DEMONST as the indicator for participation in demonstrations. When deciding on the measurement for political attitudes, she found a variety of options. During our tutorial session, Kiara learned that the length of each target variable bar indicates its popularity in different surveys and countries. Thus, she selected T_TRPARL (trust in parliament) to measure political attitudes based on the distributions for all the political attitude variables. To measure resources, she picked T_EDU (education). She selected T_GENDER and T_AGE to control for the socio-demographic characteristics of the respondents from the sample. Kiara proceeded further with this set of variables.

From the *Separate Availability* in Figure 8A, Kiara easily identified the possible time period of her study, as SDR provides sufficient data from 1989 to 2013 without any gap. The gap in 1982 can be explained by the deficiency of T_EDU. From the *responsive bar charts*, there are 22 available surveys in total, which convinced Kiara that the increase in data coverage is one of the primary advantages of the harmonized data. We displayed the top-3 surveys with the highest availability in Figure 8C (macro-level) and Figure 8D (micro-level). The pattern indicates some surveys have stable country coverage, e.g., Latinbarometro. In contrast, other surveys fluctuate greatly, e.g., International Social Survey Programme. Kiara found that the available data also cover diverse regions and countries in Figure 8 (B1-B3), which allows her to compare democratic and non-democratic countries from different parts of the world.

Next, Kiara wanted to query the data patterns to verify if the selected variables were correlated with each other, as this was an important condition for including variables in her regression analysis. In the correlation matrix (Figure 6-c_1), the intense color of all cells indicated that the correlations between all the variables were significant. Based on this observation, Kiara concluded that the variables selected in her model were accurate and could be used in her model. Finally, Kiara clicked the cell in the correlation matrix to look deeper into the relationship between T_DEMONST and T_EDU. She then found out that these two target variables are correlated not only with each other but also with their respective harmonization control variables and with quality control variables (Figure 6-c_2). Furthermore, based on the significance level labeled on each edge, she also identified that the T_EDU has weaker relationships with quality control variables compared to T_DEMONST, indicating T_EDU has better quality than T_DEMONST. In conclusion, Kiara verified that she would include those methodological variables (i.e., quality- and harmonization control variables) in her regression analysis as well.

After such detailed exploration, Kiara concluded that the ex-post harmonized survey data were sufficient for her research. The country and time coverage allowed her to study the effect of education and trust in parliament on the probability of individuals participation in demonstrations in democratic and non-democratic countries. She also pointed out that SDR lack macro-level data with democracy indicators, which she needs to add from the other dataset. At the micro-level, SDR data contain all the items she needs. Besides, the interface demonstrates the high quality of the SDR data and the importance of including methodological variables while using a harmonized dataset, such as SDR. She agreed that *SDR*Querier provides accurate guidance for variables identification, supplies efficient decision-making support for relying on the harmonized data or not via visual exploration and contributes a lot to variable selection in regression models.

7 Expert Feedback

We conducted in-depth interviews with the same group of experts to gather their qualitative feedback on the usefulness and usability of *SDR*Querier (*E1*∼*E4*). E1 and E2 are social scientists with more than 45 years of experience studying social movements and contentious politics. E3 has 10+ years of experience in survey data harmonization, and E4 has 5+ years of experience in both survey data transformation and data management. We started the interviews with an introduction to the *SDR*Querier pipeline and the functions of individual visual components. The interviews were conducted in an interactive way to discuss the pros and cons, suggestions, and agreements on *SDR*Querier.

Overall, the experts agreed that the tool *is very helpful in learning the structure and capacity of the harmonized survey data* and *also contributes to data methodology literature by proposing new ways to work with the harmonized dataset.* E1 noted that *given its ambitious goal, SDRQuerier can bring a big contribution to social science as pioneer research.* E2 added that the tool *would increase the popularity of the SDR due to its novel visualizations, making it available to researchers from other fields, such as economics or psychology, who use survey data for analyses.* The experts also noted that the tool is easy to use, even for those without knowledge of computer science and visualization, and that its visual interactions and availability checking features are extremely useful for exploring and identifying desired data within the complex SDR dataset.

All experts expressed interest in the QBQ. E4 mentioned that *the role of the module will become even more important as the harmonized data become more mature and the number of variables increases*. E4 also noted that *the trained model can be used by social scientists during the harmonization process to retrieve relevant questions*. E3 agreed and summarized that *QBQ can be used at three levels: during ex-post analysis, during the harmonization process, and in international surveys containing hundreds of variables.* E3 added that although social scientists may not fully understand the inner workings of the trained model, *it is still effective in identifying related variables and is easy to use*. E2 highly evaluated QBR, stating that in the harmonized data, users should consider including methodological variables in their models and that *it is useful to present the network to scientists due to the complex relationships between variables*. E1 emphasized the importance of visualizing weak relationships between target and quality control variables, as *it indicates the high quality of the samples.*

Additionally, the experts suggested some improvements for *SDR*Querier. E1 expressed a concern that *users might overthink the meaning of each element after learning visual mappings introduced in SDRQuerier*. Take the network in QBR as an example. After noticing the color of the edge indicating the coefficient is valid or not, they may wonder whether the edge length also encodes other information. Both E3 and E2 were first confused about the coloring schema of the Temporal Availability Profiler, though they understood it after detailed explanations. They worried that the *learning curve of the coloring algorithm might be high for social scientists without visualization and database training*. These comments will be considered when deploying *SDR*Querier.

8 Conclusion and Future Work

In this work, we present *SDR*Querier, a visual query system that facilitates scientists to locate target data and evaluate their theoretical models. To achieve this, the system provides visual guidance and queries with three modules: Query-by-Question, Query-by-Condition, and Query-by-Relation. From the solid evaluation and thorough studies, we have identified several applications for QBQ and exemplified how QBC and QBR help scientists to understand, explore, and utilize harmonized survey data in their research. Insightful findings and positive feedback from domain experts demonstrated the novelty, usefulness, and effectiveness of *SDR*Querier.

There are several potential areas for future development of *SDR*Querier based on feedback from domain experts and our own observations. These can be broadly grouped into two categories: (1) Enhancing the functionality of QBQ to better support the harmonization process. Automating some data cleaning processes with minimal user interaction can improve the efficiency of the harmonization process. (2) Conducting a large-scale use case to gather feedback on the visualizations. By collecting input from a wide range of users, we can better understand how to optimize the design of our visualizations to meet the needs of our users and improve their effectiveness.

9 Acknowledgements

This work is supported by the NSF project 1738502, RIDIR: Survey Data Recycling: New Analytic Framework, Integrated Database, and Tools for Cross-national Social, Behavioral and Economic Research. The work is also partically supported by ICICLE project under grant number OAC-2112606.

References

[1] R. Singleton Jr, B. Straits, and M. Straits, *Approaches to social research. 2nd and 5th ed*, 2009.

[2] S. Ruggles, M. L. King, D. Levison, R. McCaa, and M. Sobek, "Ipums-international," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 36, no. 2, pp. 60–65, 2003.

[3] J. R. Frick, S. P. Jenkings, D. R. Lillard, O. Lipps, M. Wooden, *et al.*, "The cross-national equivalent file (cnef) and its member country household panel studies," *Schmollers Jahrbuch: Zeitschrift für Wirtschafts-und Sozialwissenschaften*, vol. 127, no. 4, pp. 627–654, 2007.

[4] P. Granda, C. Wolf, and R. Hadorn, "Harmonizing survey data," *Survey methods in multinational, multiregional, and multicultural contexts*, pp. 315–332, 2010.

[5] I. Tomescu-Dubrow, K. M. Slomczynski, and J. C. Jenkins, "Democratic values and protest behavior in cross-national perspective. harmonization of data from international survey projects.," 2016.

[6] M. Kołczyńska and K. M. Slomczynski, "Item metadata as controls for ex post harmonization of international survey projects," *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*, pp. 1011–1033, 2018.

[7] O. Oleksiyenko, I. Wysmulek, and A. Vangeli, "Identification of processing errors in cross-national surveys," *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*, pp. 985–1010, 2018.

[8] I. Wysmulek, K. M. Slomczynski, and I. Tomescu-Dubrow, "Survey data quality in analyzing harmonized indicators of protest behavior: A survey data recycling approach," *American Behavioral Scientist*, 2021.

[9] K. Healy and J. Moody, "Data visualization in sociology," *Annual review of sociology*, vol. 40, pp. 105–128, 2014.

[10] A. S. Jones, J. S. Horsburgh, D. Jackson-Smith, M. Ramírez, C. G. Flint, and J. Caraballo, "A web-based, interactive visualization tool for social environmental survey data," *Environmental modelling & software*, vol. 84, pp. 412–426, 2016.

[11] J. Ryssevik and S. Musgrave, "The social science dream machine: Resource discovery, analysis, and delivery on the web," *Social Science Computer Review*, vol. 19, no. 2, pp. 163–174, 2001.

[12] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski, "Visualizing time-oriented data—a systematic view," *Computers & Graphics*, vol. 31, no. 3, pp. 401–409, 2007.

[13] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman, "Lifeflow: Visualizing an overview of event sequences," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 1747–1756.

[14] K. P. Hewagamage, M. Hirakawa, and T. Ichikawa, "Interactive visualization of spatiotemporal patterns using spirals on a geographical map," in *Proceedings 1999 IEEE Symposium on Visual Languages*, IEEE, 1999, pp. 296–303.

[15] J. V. Carlis and J. A. Konstan, "Interactive visualization of serial periodic data," in *Proceedings of the 11th annual ACM symposium on User interface software and technology*, 1998, pp. 29–38.

[16] M. Weber, M. Alexa, and W. Müller, "Visualizing time-series on spirals.," in *Infovis*, vol. 1, 2001, pp. 7–14.

[17] S. Liu, M. X. Zhou, S. Pan, *et al.*, "Tiara: Interactive, topic-based visual text summarization and analysis," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, pp. 1–28, 2012.

[18] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim, "Eventriver: Visually exploring text collections with temporal references," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 1, pp. 93–105, 2010.

[19] S. Havre, B. Hetzler, and L. Nowell, "Themeriver: Visualizing theme changes over time," in *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, IEEE, 2000, pp. 115–123.

[20] R. L. Harris, *Information graphics: A comprehensive illustrated reference*. Oxford University Press, USA, 1999.

[21] J. A. Fails, A. Karlson, L. Shahamat, and B. Shneiderman, "A visual interface for multivariate temporal data: Finding patterns of events across multiple histories," in *2006 IEEE Symposium On Visual Analytics Science And Technology*, IEEE, 2006, pp. 167–174.

[22] A. Dasgupta, R. Kosara, and L. Gosink, "Vimtex: A visualization interface for multivariate, time-varying, geological data exploration," in *Computer Graphics Forum*, vol. 34, 2015, pp. 341–350.

[23] F. Y. Wang, A. Sallaberry, K. Klein, M. Takatsuka, and M. Roche, "Senticompass: Interactive visualization for exploring and comparing the sentiments of time-varying twitter data," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, IEEE, 2015, pp. 129–133.

[24] V. Peña-Araya, E. Pietriga, and A. Bezerianos, "A comparison of visualizations for identifying correlation over space and time," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 375–385, 2019.

[25] V. Dibia, "Neuralqa: A usable library for question answering (contextual query expansion+ bert) on large datasets," *arXiv preprint arXiv:2007.15211*, 2020.

[26] R. Nogueira and K. Cho, "Passage re-ranking with bert," *arXiv preprint arXiv:1901.04085*, 2019.

[27] W. Yang, Y. Xie, A. Lin, *et al.*, "End-to-end open-domain question answering with bertserini," *arXiv preprint arXiv:1902.01718*, 2019.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] W. Yang, H. Zhang, and J. Lin, "Simple applications of bert for ad hoc document retrieval," *arXiv preprint arXiv:1903.10972*, 2019.

[30] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "Cedr: Contextualized embeddings for document ranking," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1101–1104.

[31] Z. A. Yilmaz, S. Wang, W. Yang, H. Zhang, and J. Lin, "Applying bert to document retrieval with birch," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 2019, pp. 19–24.

[32] Z. Jiang, A. El-Jaroudi, W. Hartmann, D. Karakos, and L. Zhao, "Cross-lingual information retrieval with bert," *arXiv preprint arXiv:2004.13005*, 2020.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[34] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[35] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, 2010.

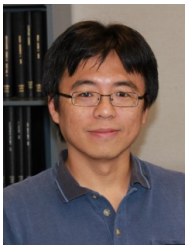
[36] M. Kołczyńska, "Micro-and macro-level determinants of participation in demonstrations: An analysis of cross-national survey data harmonized ex-post," *methods, data, analyses*, vol. 14, no. 1, p. 36, 2020.

Yamei Tu received a BS degree in Software Engineering from East China Normal University. She is a Ph.D. student in the GRAphics & VIsualization sTudY (GRAVITY) Research Group under Computer Science and Engineering at the Ohio State University. Her research interests are Visualization, Text Analysis, and Machine Learning.

Olga Li is a Ph.D. student at the Graduate School for Social Research and is a member of the research unit on Comparative Analyses of Social Inequality at the Institute of Philosophy and Sociology, Polish Academy of Sciences. Her research interests include survey methodology and data analysis, ex-post survey data harmonization and political participation.

Junpeng Wang joined Visa Research as a Staff Research Scientist in June 2019. He received his Ph.D. in Computer Science from The Ohio State University in 2019, M.S. in Computer Science and Application from Virginia Polytechnic Institute and State University in 2015, and B.Eng. in Software Engineering from Nankai University in 2011. His research interests are broadly in explainable artificial intelligence (Explainable AI), visual analytics, and deep learning.

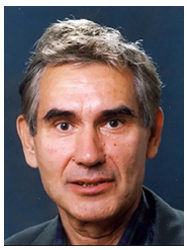
This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3261944



Han-Wei Shen received the BS degree from the Department of Computer Science and Information Engineering, National Taiwan University, the MS degree in computer science from the State University of New York at Stony Brook, and the PhD degree in computer science from the University of Utah. He is a full professor with the Ohio State University. His primary research interests include scientific visualization and computer graphics.

Przemek Powałko is a survey data management and data quality specialist involved in the Survey Data Recycling project at the Institute of Philosophy and Sociology, Polish Academy of Sciences.

Irina Tomescu-Dubrow is a Professor of Sociology at the Institute of Philosophy and Sociology, Polish Academy of Sciences, and Director of the Graduate School for Social Research (IFiS PAN). She got a BA degree from the University of Bucharest in 1997, an MA degree, and the PhD degree in sociology from the Ohio State University.



Kazimierz M. Slomczynski He is currently the Director of CONSIRT (Cross-National Studies: Interdisciplinary Research and Training Program) at Ohio State University. He received his PhD degree from the University of Warsaw in 1997. His research interests include stratification (including race, class, and gender), occupations and work, political sociology, population and health, and social psychology and the methodology of cross-national research.

Spyros Blanas is an associate professor in the Department of Computer Science and Engineering at The Ohio State University. His research focuses on high-performance database systems, and he has received recognition for his work including a Google Faculty Research Award and an IEEE TCDE Rising Star Award. Prior to joining Ohio State University, he received his PhD at the University of Wisconsin–Madison, where he was a member of the Database Systems group and the Microsoft Jim Gray Systems Lab.

J. Craig Jenkins is an Academy Professor Emeritus of Sociology at The Ohio State University. Prior to this, he served as the director of the Mershon Center for International Security Studies from 2011 to 2015. Currently, he works as a senior research scientist. He received his Bachelor's degree from the University of Texas-Austin in 1970 and both his Master's and PhD in sociology from the State University of New York-Stony Brook.