Interactive Data Visualization for Census Data

Naina Bharadwaj
Department of Information Technology
St. Francis Institute of Technology

Prachi Mishra
Department of Information Technology
St. Francis Institute of Technology

Prajyoti Dsilva
Assistant Professor
Department of Information Technology
St. Francis Institute of Technology

ABSTRACT
Interactive data visualization is a technique of analyzing data, where a user interacts with the system that results in visual patterns for a given set of data. In this paper, seven basic modules and their corresponding operations have been proposed that an interactive big data visualization tool for Census dataset should possess. The current visualization tools for Census dataset are limited in their results due to lack of interactivity. This paper aims to eliminate the limitation by enhancing the interactive visualization process with more relevant operations for manipulation of resultant visuals according to various attributes. Gaps and discontinuities in data have also been considered for visualization. Reliability factor for the sources of the big data has been introduced. It also explains why Census dataset requires additional features and modules in comparison to the ones in existing visualization tools. The working of every module and operations associated with it has been described using a real life example of Census Data for a country

Keywords
Big data; Census data; interactive visualization; visualization tool

1. INTRODUCTION
In this digital era, data surrounds us in all dimensions; hence it is important to extract sensible patterns and relationship using visual representation. Also, visual representation is more appealing and easily understood by humans as compared to textual formats [1]. Existing visualization tools are unable to analyze patterns for a Census dataset because unlike other datasets which use online monitored behavior and extraction from online surveys, Census dataset is a combination of datasets and acquired from multiple sources that in addition to online monitored or survey data include manually collected data from regions that are not technically advanced but still form a significant part of the population.

Manually collected data needs to be filtered well to extract patterns that are useful, also the reliability of data collected varies according to factors like means of collection and transmission. In addition to this, Census data requires analysis of data that can go unnoticed in the traditional methods of visualization, these include gaps in data and overlapping of various attributes. Contemporary tools of visualization for Census lack flexibility and possess low usability; hence they are incapable of discovering valuable patterns and trends [1]. An interactive visualization tool for this area should be able to overcome these limitations and provide results in the most efficient manner. The ideas reflected in this paper can be helpful for people building data visualization tools for Census dataset to incorporate methodologies that improve the interactivity of existing visualization tool.

Example: Consider an example of a population census that includes all the data about the population of an area. Various parameters are considered in Census Data such as age, gender, educational qualifications, etc. for people, as well as parameters of regions/places such as weather conditions, topology, development status, area, etc. [2]. In reality these numbers go large and hence data visualization is required here for effective analysis and understanding interesting patterns. Also, such data has to be analyzed over various parameters, different scales, boundaries with various kinds of visualization environments, thereby creating a need for interactive data visualization tool independently for analyzing Census data.

2. LITERATURE REVIEW
Census data visualization has always been restricted with respect to interactivity and flexibility. In [2], The Hurricane Sandy Census Viewer offers an online system to visualize the big population data on the region map. [4], the User Manual to use [2] specifies the operations that can be performed on the system and describes what each operation indicates. Some of the operations described in [4] are density, map dot size, transparency, count by, boundaries, data sources, filter, etc. These operations are used in an unorganized manner in this manual. Also, this system is inadequate in providing flexibility and user interaction at a larger level. These act as the key elements in the proposed system. Operations like analyzing the metadata, considering the reliability factors and using color combinations to analyze multiple attributes simultaneously are not a part of [2][4] but are integrated in the proposed work. L. Wang et al in [3] have specified four steps for interactive data visualization, viz. Selecting, Linking, Filtering and Rearranging or Remapping. These steps do not involve any steps mentioned for integrating the various data sets, analyzing the metadata and storing the views generated by the tool. It is important to remove the inconsistencies produced by the heterogeneity of various data sets from multiple sources before visualizing the data. Analyzing the metadata can also produce informative results in Census, which are often ignored. Saving the views is useful for future references. Hence, this system involves modules to address each of these. M. Morgan in [5] explains about some dynamic and interactive visualization methods that can be applied in dashboards. One of the visualizations mentioned in his blog is "Motion chart for Trend Analysis". It is suggested in this paper that "Trend Analysis" must be a part of the step-by-step process of interactive data visualization after the data has been filtered, view has been manipulated and metadata has been well understood, so that the motion chart comprises of the required data only and the reasoning analysis is aided well after metadata analysis. In [7], which is a visualization tool for US Census it is seen that reliability of data sources is not taken into consideration. The gaps and discontinuities in the dataset are not analyzed. Also the tool is less interactive and flexible. Reference [8] describes about the contemporary

innovations that the U.S. Census Bureau has incorporated in statistical mapping and data visualization to analyze census data. Census Bureau's Topologically Integrated Geographic Encoding and Referencing (TIGER) web combines the geospatial data with statistical data. It provides an interactive map to visualize the data without any extra GIS software requirements. But the smallest unit of mapping in this system is an entire state. It plots only one attribute at a time with a palette scheme to distinguish between a fixed set of classes of values. The proposed system focuses on plotting more than one attribute at a time to help analyze that section of data which is the intersection of the two attributes. And the smallest unit of mapping can be set by user which is visible in the form of pixel size. C. Ling et al have compared two visualization tools viz. Gapminder and Tableau Public at simple and integrated level. The major parameters considered in [9] are accuracy and ease-of-use. Whereas in the proposed system the "reliability factor" of every source is considered. The proposed system takes into consideration the drawbacks of all the existing Census visualization tools and provides a more systematic approach concentration on the integrity of the data visuals.

3. PROPOSED WORK

This proposed system for interactive data visualization for Census dataset includes seven basic modules, as shown in Figure 1.

Also a loop has been incorporated in the modular structure of the process, as shown to make it evident that four of the modules may be required to be performed iteratively to generate the required visual.

The various operations to be performed in each of the modules are explained as follows:

3.1 Data Source Selection

As Census data (i.e. real time or passive user data) is obtained from various sources, this step includes an option of selection of datasets from multiple data sources for making the visualization tool more flexible. This helps in identifying patterns between more varieties of data thus making the observations more concrete for real-world applications.

For example: For using the visualization tool to determine the effect of literacy rate on the economy of the country, user requires datasets from two data sources, i.e. Education dataset and the Economy dataset. These datasets belong to different sources such as manually updated database files, online survey results, online monitored data and metadata files.

3.2 Data Integration

This step involves the integration of:

- Multiple datasets from the same data source
- Multiple datasets from different data sources

This is done on the basis of a single attribute that is common to link all the datasets [4]. This step requires the conversion of datasets from all data sources into a single format for analysis purpose. A single format for conversion also makes it easier to remove redundancy or duplication in the data.

For example: To analyze the relationship between literacy rate and population of the country, the Education and Population datasets of all states will be integrated. The datasets collected from different regions will first be converted to a single file format, for e.g. '.csv' and then integrated on the basis of common attributes and filters.

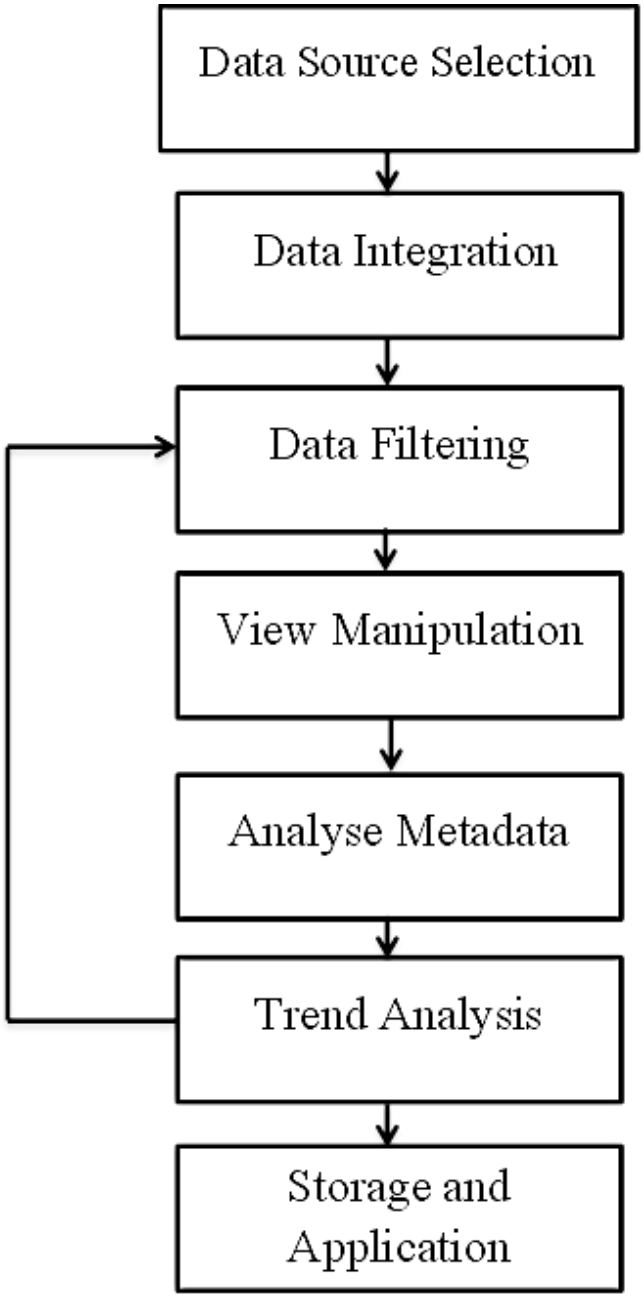

Fig 1: Proposed Work

3.3 Data Filtering

This step focuses on the quality of information to be displayed. The quantity is decreased by focusing only on the crucial information from the datasets using feature extraction and geometric modeling [1]. Filtering can be done based on any specific attribute or multiple attributes. In traditional visualization tools filtering can only be applied for dataset from a single data source. For better understanding of data with respect to different sources, this system allows the user to apply filters to multiple datasets simultaneously, this ensures that no pattern or relationship goes unidentified and helps gain a better insight regarding the analysis. The operations in filtering include:

3.3.1 Sample size and sample selection

The user need not always visualize the entire data all at once. So user can specify the sample size.

User must select the range of records to be displayed, or discrete records can be selected.

For example: If the user wants to analyze the literacy rate for any 15 states, the sample size is then 15. He does the selection of records on a random basis or by selecting any 15 discrete records.

3.3.2 Range

User can set the values for data to be displayed himself for range dependent data i.e. user can set the extreme boundaries for range dependent data.

For example: To categorize states into those needing/ not needing development in education sector on the basis of literacy rate, user can select the range, for e.g. Need Development as 0 to 65%, Not Needing Development as 66% to 100%.

3.4 View Manipulation

In this module, a visual is generated based on the filters selected by the user that plots all the selected data after filters have been applied.

This module provides operations to the user for manipulation of the visualizations based on attributes of their choice. The operations that aid to better analyze the visual are as follows:

3.4.1 Density

When zoomed out, the density is depicted using intense colors to show the clusters or concentration of data values according to the specified filters [4].

This feature is enhanced by classifying density into levels such as high, low or moderate according to the values. This will make the analysis easier.

For example: To view the "Literacy Rate" distribution throughout the country, the user simply zooms out to view the areas as high, low or medium on the basis of colorful clusters.

3.4.2 Environment Settings

The following parameters are used in the system, where the pixels visibly describe the data set in more than just one dimension:

- When it comes to pixel plotting mechanism of data visualization, the user is allowed to choose the color of pixel for different parameters. Such that on merging two parameter colors, a third color pixel comes up to indicate the intersection of data based on the sets of two parameters. The new colors and their corresponding parameter representations will be available in a tabular format to the user.
- Transparency of the pixels is also considered for denoting layering of various attributes i.e. the opacity of the pixel.
- User is allowed to set the ratio of pixel size to sample size. So the user himself understands what number the pixel size indicates.
- The base on which the pixels are plotted can also be changed by the user. In the example given, the region of the map can be changed, or set by the user.

For example: To analyze the relationship between Topography and Development of states, the user chooses the real map of the country as the base and every other attribute to be displayed such as Literacy Rate, Weather, Health Conditions can be depicted with different colors, their extent can be shown with transparency. Pixel size indicates the number of records which can initially be specified by the user.

If Weather "hot" is depicted with red color pixels and "poor" health conditions with green pixels, then the region where these pixels overlap with be yellow in color, as red and green mix to give yellow color.

3.4.3 Hiding

Depending on the analysis of the data, the user might want to hide some pixel records from the view; this functionality is provided in the proposed module.

To hide, the user selects the record pixels, he can either hide those selected records from the data or the records except those selected.

For example: A data sample may often contain outliers. Outliers are those values that deviate significantly from other data values as if generated from a different mechanism. To analyze data other than outliers, user selects outliers and hides them to obtain a more accurate visualization.

3.5 Analyze metadata

Metadata can provide some of the most crucial patterns as it describes the analyzed data [3]. This module further analyses the data visualization and produces patterns that were not noticed before. It helps in making more accurate inferences.

3.5.1 Gaps

Gaps in data mean "no data areas" i.e. no pixels marked the said region; they are exploited to understand a pattern in the data.

Analyzing gaps in a Census dataset is crucial as it gives insights about why the discontinuities exist in the data.

There are various reasons for absence of data in a region.

- Missing data
- Issues in data collection for that area
- Value of selected attribute is zero
- Unidentified regions

These reasons need to be identified as they play a significant role in providing inferences about data collection methods as well as information about regions.
An option is provided in the system to explore the gaps, so that the user can highlight and understand the gaps in data and visualization of these gaps gives the user more clarity in analyzing the data as a whole.

First, the gaps in the data are detected in the maps generated from the datasets. Then, these detected regions are analyzed on 2D scatter plots to identify a pattern between them.

The detected patterns will provide the following details about the gaps:

i. Whether the gaps exist in a uniform pattern or for randomly.
ii. Whether the gaps are restricted to a specific geographic region or distributed to various areas.
iii. Whether the same gaps exist for a particular attribute or multiple attributes.
iv. Whether the gaps have been existing since a long time or are recent. (using a motion chart)

For example: Gaps in Census data for an attribute 'Literacy Rate' might indicate one of the following:

- Randomly spaced gaps: Missing values generated due to an error.
- Gaps in a specific region: Data has not been collected in the region. The region has not been identified or considered for the data analysis
- Gaps with attribute value zero: The literacy rate is zero in that region.
- Gaps in a specific region over a long period: The region is uninhabitable.

This feature helps in identifying and working on the drawbacks of the collection system, if any. Exact reasons can be figured from previous results.

3.5.2 Reliability Factor

Reliability factor aids the user in finding out the extent to which the visualization of patterns and trends holds true. The proposed system provides the means for obtaining the reliability of the resultant visualization as follows:

This helps in determining how reliable the visualization patterns are by giving the option of specifying the reliability (in percentage) of each data source to the user and based on that, calculating the total reliability of the data visualized.

Better reliability is provided in this system for visualization patterns in comparison with other systems. When a dataset is being fed to the system, it takes into account, details such as date of collection, efficiency of data collection and transmission, validity period of data and the number of regions covered for collection. The errors due to dataset collection method and extent of completion of the data collection process are taken into consideration in the proposed system.

This system also calculates the effect of the reliability of multiple parent datasets, on the reliability of the resultant visualization.

3.5.3 Count

The count of the records, based on the filters used, that is being displayed can also be shown as metadata. Counts are dynamic and hence updated on every selection and function [4].

These are of three types based on user choice:

- All: Count of all the records of the entire dataset taken into consideration for the entire plotting base.
- Visible: Count of records that are currently visible to the user in the window pane.
- Poly: Count of records in a polygon area of the base, selected by the user.
- Selected: Display the count for those pixels selected by the user
- Queried: Gives a count of the records that are obtained as the result of a query entered by the user.

For example: If the user wants to find the number of illiterates in the country, he can do so by selecting All or Visible, to specify the search to a certain area he uses Poly, to specify an area by selecting he can use Selected. If the user wants to find out the exact number of men who are illiterate in a given city, he applies all the required filters in Queried i.e. "City Name", "Male", "Illiterate" to obtain the count.

3.6 Trend Analysis

This step follows the static data visualization by combining various visualization frames to produce a dynamic motion chart that represents an overall trend in the data [5].

The concept of motion charts is used in this system for providing flexibility in analyzing dynamic trends. Motion charts can be used for any selected attributes as per user definition. Any attribute selected by the user can be used as the basis for the motion. Motion charts help in analyzing trends over a period, or the effect of a specific attribute on a pattern.

For example: To analyze how women's education has changed over the years, the user can view the trend of literacy rate ratio of male to female, by selecting the motion basis as 'time' attribute and view the motion chart.

In case the user wishes to edit the data selected for the trend right from filtering the data, then steps from data filtering to trend analysis are performed again.

3.7 Storage and Application

After the data has been visualized successfully and satisfactorily, it should be stored such that it can be referred to, as and when required. Census data builds on the previous datasets, thus saving and adding to these datasets should convenient. This can be done in two ways:

3.7.1 Saving

Once the user has generated a visual after applying all the given interactive options, the user is given an option to save the visual along with the filtered data associated with it. (Filtered data set is the data currently visible in the visual.)

To save these two entities(visual and filtered data), a different type of file format is needed which is compatible with the visualizing tool such that on opening the file, the associated data is loaded and visual is generated directly with the applied options.

The file can also support saving an operations' document which specifies:

- the main data set used to generate the visual,
- some primary details about the main data set and
- operations performed on the data set to generate the visual

This document can help the user regenerate the visual from main data set.

NOTE: On opening the saved visual, the user has only the filtered data set loaded and not the main data set.

For example: If a user is working on the trend for "Literacy Rate" in the country, he saves this data and when new information about any city is encountered, he can easily append this information to the already existing visualization by simply opening the previously saved file and integrating the new data.

3.7.2 Embedding

The user might want to display the visualization result or trend on his/her webpage. For this, the user is given an option or a code snippet that when copied to the webpage's source code, displays the static or dynamic visualization on the

webpage. Adding this snippet to the source code links the webpage to the visualization tool.

For example: If the user wants to display a visualization chart of the literacy rate on the "Education" section in census website, he can easily do this by making the required changes in the source code of the web page.

4. RESULTS

The dataset considered for analyzing the results is Climate dataset. It contains the temperature and precipitation of countries in Asia.

Country	MaxTemp	Precipitation
India	51 °C (124 °F)	1083
China	50.3 °C (122.5 °F)	645
Sri Lanka	54.0 °C (129.2 °F)	1712
Syria	42.0 °C (107.6 °F)	252
Thailand	40.1 °C (104.2 °F)	1622
Nepal	47.0 °C (116.6 °F)	1500
Vietnam	42.7 °C (108.9 °F)	1821
UAE	52.0 °C (125.6 °F)	78
Indonesia	39.5 °C (103.1 °F)	2702

Fig 2: The Dataset

4.1 Analyzing Gaps

To analyze gaps, the system first detects the gaps in the map generated from the dataset and analyses them in a 2D scatter plot with respect to Longitude and Latitude.

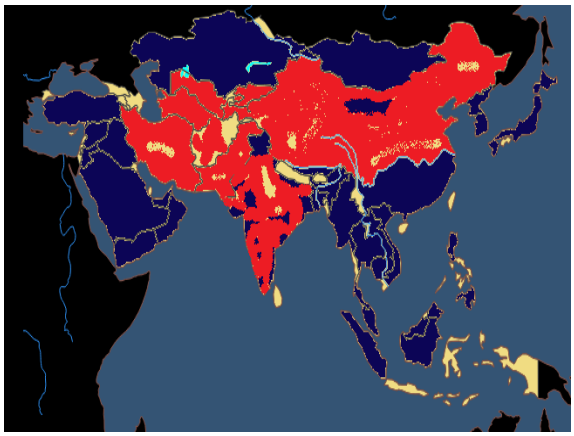

Fig 3: Scatter Plot

From Fig. 3, it can easily be derived whether the gaps are uniform or random, located to a specific region or not. For example, in figure, a map of Asia classifying regions in terms of climate zones on the basis of precipitation. Red and blue based on High precipitation and low precipitation respectively. It can be observed that gaps exist in this map where the regions are uncolored. These gaps need to be analyzed to provide better insight and hence they are converted to 2D scatter plots with Latitude on X axis and Longitude on Y axis to infer patterns (Refer Fig 4).

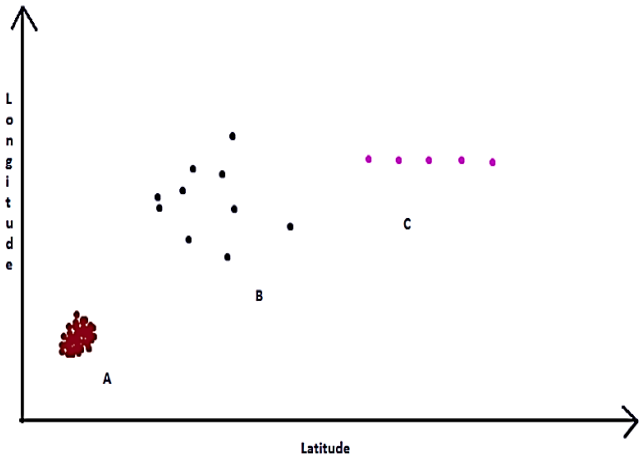

Fig 4: 2D Scatter Plot

From the above scatter plot it can be seen that there are 3 different patterns of Gaps,

In A, gaps are specific to a geographic region, thus this region has not been identified for data collection.

In B, gaps are distributed in a random manner, thus these are missing values due to errors in transmission of data.

In C, gaps are distributed in a uniform manner, thus the attribute follows a specific pattern. Figure 3 shows the information collected when a dataset is fed into the system.

The predefined definitions of variables used in the system for the calculation of the reliability factor are given in the following tables.

4.2 Calculating reliability of visuals:

Figure 3 shows the information collected when a dataset is fed into the system.

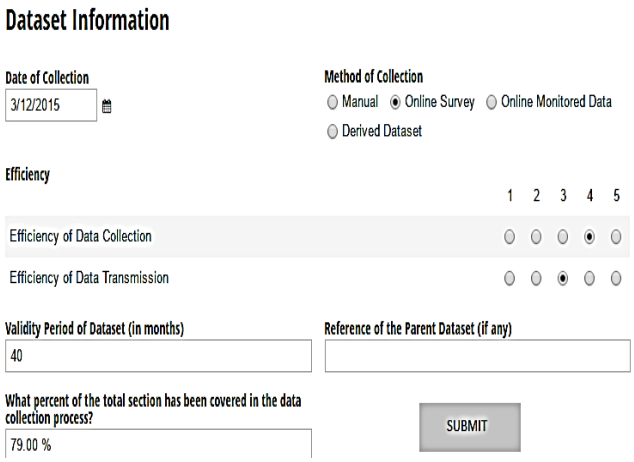

Fig 5: Form to be filled by the user to enter details of data source

The predefined definitions of variables used in the system for the calculation of the reliability factor are given in the following tables.

Table 1 shows the methods of data collection and their corresponding error percentages.

Table 1: Method of data collection and corresponding errors

Method of Data Collection	Approximated Error Percentage
Manual	5%, due to human errors
Online Survey	0.5%, due to server breakdowns
Online Monitored Data	1%, due to real-time network errors

Table 2 shows the value of efficiency selected in the form and the corresponding percentage. The percentage is calculated based on the value selected by the user, using a direct proportion formula.

For the value selected in the form, 'x':

$$\boldsymbol{Percentage\ Efficiency} = \frac{\boldsymbol{x}}{\mathbf{5}} * \mathbf{100} \quad (1)$$

Table 2: Value of efficiency and respective percentages

Value of Efficiency	Percentage
1	20
2	40
3	60
4	80
5	100

4.2.1 Calculation of Reliability Factor :

4.2.1.1 For a visualization based on an independent dataset:

If the dataset is not in its validity period,

$$Reliability\ factor\ of\ the\ visualization\ (Rf) = \mathbf{0} \quad (2)$$

If the dataset is in its validity period,

$$Reliability\ factor\ of\ the\ visualization\ (Rf) = \boldsymbol{Percentage\ of\ section\ covered} * \{(\boldsymbol{Avg.of\ efficiency\ of\ data\ collection\ and\ transmission}) - (\boldsymbol{error\ due\ to\ method\ of\ collection}\} \quad (3)$$

4.2.1.2 For a visualization based on a derived dataset:

If the visualization has been derived from multiple datasets A and B such that A provides x% of the dataset and B provides remaining of the dataset.

Considering that, Reliability factor of A is Rf_A and Reliability factor of B is Rf_B

$$Reliability\ factor\ of\ the\ visualization\ (Rf) = \boldsymbol{xRfA} + \boldsymbol{(1-x)RfB} \quad (4)$$

Let A and B be independent datasets, and C be a derived datasets from A.

Dataset information for A as seen in the figure,

As the dataset is in its validity period of 40 days,

The Reliability factor R_{fA} is given by:

$$RfA = 75/100 * \{((60+80)/2) - (0.5/100)\} = 52.496\% \quad (5)$$

A dataset C derived 60% from dataset A and 40% from a dataset B with reliability 62% has the reliability factor

$$RfC = 0.6(52.496) + 0.4(62) = 56.2976\% \quad (6)$$

5. CONCLUSION

Visualization of data has become a necessity for efficient extraction of relevant patterns that can go unnoticed; hence this process needs to have complete accuracy. The interactive data visualization modules have been proposed keeping in mind the flexibility and reliability that Census data visualization requires. These modules can be used to build a tool that can analyze and detect hidden patterns from Census data of various countries. These modules not only provide flexibility and ease of use in the visualization process, but also look after various important factors such as reliability of data, abnormal gaps in data and trends obtained from data. The reliability calculations provided in the proposed system improve the accuracy of the results, thus making the system more efficient for real work applications.

6. FUTURE SCOPE

The proposed system takes into consideration the current requirements of visualization for datasets that are dependent on manual data collection techniques and hence a smart visualization tool providing all the mentioned operations is the need of every organization dealing with such datasets. This system can be implemented for analyzing any dataset that uses real life interaction and collection of data directly from the clients as well as more specific applications where the reliability of the resultant visuals is a priority. Thus, this proposed system can be used to create a visualization tool to overcome the drawbacks of traditional tools or act as an extension for improving the existing Census data visualization tools.

7. REFERENCES

[1] E. Olshannikova, A. Ometov, Y. Koucheryavy and T. Olsson, "Visualizing Big Data with augmented and virtual reality: challenges and research agenda", Journal of Big Data, vol. 2, no. 1, 2015.

[2] "Hurricane Sandy NOAA Forecast :: Census Viewer :: CensusViewer :: Powered by Moonshadow Mobile", Hurricanesandy.censusviewer.com, 2016.[Online]. Available: https://hurricanesandy.censusviewer.com/client. [Accessed: 01- May- 2016].

[3] L. Wang, G. Wang and C. Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress", Digital Technologies, vol. 1, no. 1, pp. 33-38, 2015

[4] ."User Manual", Censusviewer.com, 2016. [Online]. Available: http://censusviewer.com/user-manual/. [Accessed: 01- May- 2016].

[5] M. Morgan, "Dynamic Data Visualizations | Business Intelligence Blog from arcplan", Arcplan.com, 2016. [Online]. Available: http://www.arcplan.com/en/blog/2013/08/dynamic-data-visualizations/.

[6] "UCI Machine Learning Repository: Census Income Data Set", Archive.ics.uci.edu, 2016. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Census+Income. [Accessed: 16- Jun- 2016].

[7] C. US Census Bureau, "Visualization Tools - Business Dynamics Statistics - Center for Economic Studies", Census.gov, 2016. [Online]. Available: http://www.census.gov/ces/dataproducts/bds/visualizations.html. [Accessed: 05- Jun- 2016].

[8] "Data visualization at the US census bureau – an American tradition: Cartography and Geographic Information Science: Vol. 42, No sup1", Cartography and Geographic Information Science, 2016.

[9] C. Ling, J. Bock, L. Goodwin, G. Jackson and M. Floyd, "Comparison of Two Visualization Tools in Supporting Comprehension of Data Trends", Springer International Publishing, pp. 158-167, 2016.