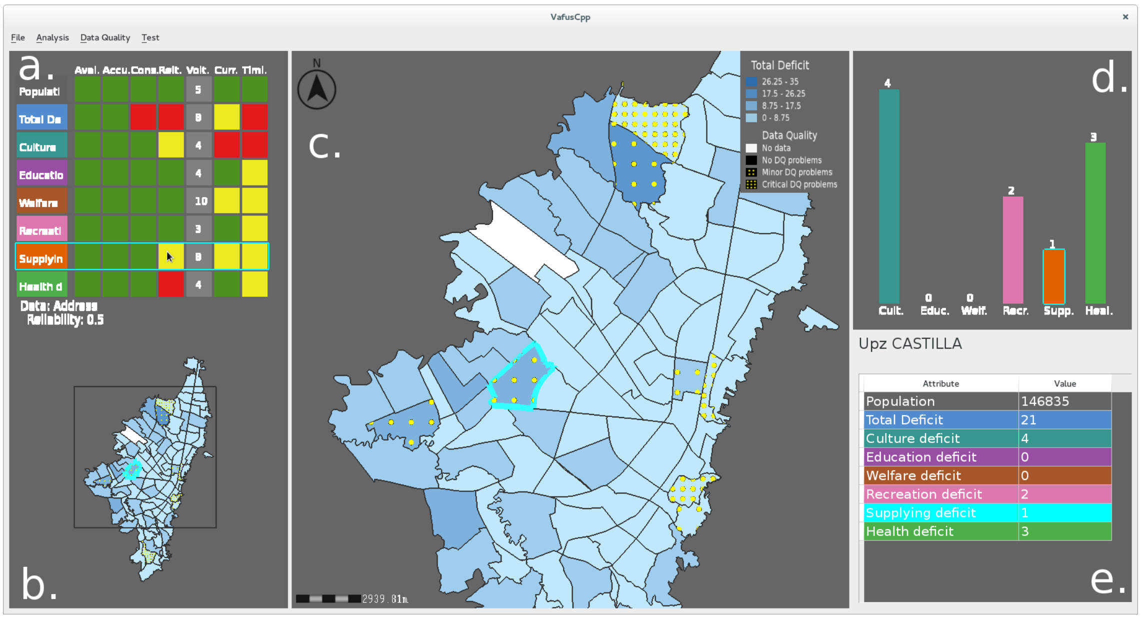
Workshop on Visualisation in Environmental Sciences (EnvirVis) (2015)
A. Middel and K. Rink and G. H. Weber (Editors)

VafusQ: A Visual Analytics Application with Data Quality Features to Support the Urban Planning Process

John A. Triana[1], Dirk Zeckzer[2] Jose T. Hernandez[1] and Hans Hagen[3]

[1]IMAGINE, Universidad de los Andes, Colombia
[2]Institut für Informatik, Universität Leipzig, Germany
[3]Computer Graphics & HCI Group, TU Kaiserslautern, Germany



	Avai.	Accu.	Cons.	Relt.	Volt.	Curr.	Timl.
Populati	🟩	🟩	🟩	🟩	5	🟩	🟩
Total De	🟩	🟩	🟥	🟥	8	🟨	🟥
Culture	🟩	🟩	🟩	🟨	4	🟥	🟥
Educatio	🟩	🟩	🟩	🟩	4	🟩	🟨
Welfare	🟩	🟩	🟩	🟩	10	🟨	🟨
Recreati	🟩	🟩	🟩	🟩	3	🟩	🟨
Supplyin	🟩	🟩	🟩	🟨	8	🟨	🟨
Health d	🟩	🟩	🟩	🟥	4	🟩	🟨

Data: Address
Reliability: 0.5

Upz CASTILLA

Attribute	Value
Population	146835
Total Deficit	21
Culture deficit	4
Education deficit	0
Welfare deficit	0
Recreation deficit	2
Supplying deficit	1
Health deficit	3

Figure 1: *Deficit analysis visualization. a. DQD Deficit Matrix showing seven Data-Quality-Dimension for the required data for the analysis task. b. Map Overview. c. Map View representing the total deficit for Bogotá. d. Deficit Bar Chart and e. Table View showing additional information to support the analysis of the deficit task.*

Abstract

Fast changing urban systems pose huge challenges for planners and governments. One major challenge is to provide optimized facilities systems fulfilling all the basic citizen needs such as food, education, security, and health. To provide these, the deficit of the complete system needs to be analyzed and quantified. An additional, important problem is the quality of the underlying data influencing the analysis. Often, the data is, e.g., incomplete, not accurate, or not reliable. The goal of this paper is to support the analysis of the deficit for the facilities system of Bogotá by taking into account data quality issues. Our contributions are: the inclusion of data quality in the urban planning process, the design of a novel visualization technique to represent data quality, the implementation of an application to support the analysis of the facilities system, and a case study with experts assessing the usability and usefulness of the application. As a conclusion, the experts find the application useful for the analysis tasks and the inclusion of data quality features important and comprehensible.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—Urban planning

DOI: 10.2312/envirvis.20151091

1. Introduction

Uncertainty visualization has been studied over the last 20 years. Several theoretical analyses have been performed [MRH*05, Pan01, GS11]. Moreover, a considerable number of applications and visualizations have been developed by a variety of researchers [CWRY06, CCP07, XWRH07, OM02, PKRJ10, WYM12, WYM12, HMC*13]. Although many suggestions exist for defining uncertainty (e.g. Pang and Pham et al. [Pan01, PSB09]), none is yet widely used. This situation has influenced research, such that the visualizations and applications proposed are based on different definitions from statistics [HMC*13] and computer science [KPP12, SEG05]. In computer science, uncertainty is described by using metrics of quality called Data Quality Dimensions (DQD) proposed [BS06]. Based on the definitions of Eppler [Epp01] and Batini and Scannapieco [BS06], we define a data quality dimension as a measurement or a perception of the degree of the data's fitness in a particular context.

Instead of using the classical approach to try to clean the data and to discard unreliable data in the preprocessing stage, this work will deal with the possibly unreliable data after the preprocessing step, representing and communicating the unavoidable errors to the users by using interactive visualizations that include data quality features.

2. Urban Planning Background

Around the world there are cities that have high growth rates. Consequently the requirements to keep a minimum quality of life in this kind of cities is a challenge for governments and urban planners. On the other hand, there are also cities that are changing their characteristics continuously within short periods of time. For this reason the urban planning process for these cities is more complex than usual. We call these cities *Fast Growing Urban Systems* and *Fast Changing Urban Systems*. They pose huge challenges for the urban planners. The problems range from data gathering to the decision making process, fast changes, and limited resources. Currently, *Fast Growing Urban Systems* and *Fast Changing Urban Systems* are mainly situated in Latin America and Asia but there are examples of them around the world.

Every city in the world requires a facilities system that comprises all the public and private institutions called facilities.This paper takes as reference the facilities system of Bogotá, which is subdivided into 12 sectors: food supply, social welfare, cemetery, religious, culture, basic education, higher education, fairgrounds, recreation and sports, health, administrative services and security. In 2009, a study of the facilities system of Bogotá was developed at Universidad de los Andes [dlA09]. The study evaluated the deficit for every UPZ (units of planning defined by the local authorities) in the city, as well as the requirements to achieve the standards. The study was developed using Geographical Information Systems (GIS), and other specialized applications. However, the generated visualizations were not interactive and they did not consider the quality issues present in the input data sets.

Taking into account this study [dlA09], three analysis tasks were extracted performing a survey and interviews with experts involved in the study according to the methodology presented by Fernandez et al. [FPZH14]: (1) Identification and proposal of facilities clusters, (2) Analysis of the deficit of facilities, and (3) Placement of new facilities. The visualization and interaction proposed in this paper are targeted to support the second analysis task: analysis of the deficit of facilities. The deficit analysis aims at evaluating and identifying zones in the city that do not adhere to the recommended standards of facilities system in Bogotá. From the 12 sectors, the experts selected 6 priority sectors: food supply, social welfare, culture, education, recreation and sports, and health. Furthermore, the target analysis task is subdivided into seven subtasks: ST_1: How is the total deficit distributed throughout the city? ST_2: How is the deficit by sectors distributed throughout the city? ST_3: How is the deficit by sectors distributed in one UPZ? ST_4: Which UPZ(s) has the highest deficit? ST_5: Which UPZ(s) has the lowest deficit? ST_6: Which UPZ(s) has data quality problems? ST_7: How is the data quality in one UPZ?

3. Data Quality Dimensions

The inclusion of Data Quality Dimensions (DQD) in the urban planning process helps the experts to make more informed decisions. If there is a data quality issue, the users should be informed about it. The DQD considered for this study are: availability, accuracy, consistency, reliability, volatility, currency, and timeliness. The first four were chosen by their importance in the literature. The last three are time oriented DQD selected because of the importance of time for this kind of applications.

Availability is defined by the presence or absence of data. Missing data is represented by *"NULL"* values. Accuracy is the closeness between a value and its representation [BS06]. This dimension could be estimated by defining a domain D and verifying if the data d is in the domain.The consistency dimension captures the violation of semantic rules defined over (a set of) data items [BS06]. Reliability indicates "whether the data can be counted on to convey the right information; it can be viewed as correctness of data" [BS06]. Furthermore, the reliability is closely related to credibility. Consequently, this dimension will be estimated subjectively by the experts. Volatility is the length of time during which data remains valid [BS06]. For example, a census made every 5 years has a volatility of 5 years. Currency represents how promptly data are updated. The currency could be estimated by using: $Currency = age + (deliveryTime - inputTime)$, where age is how old the data is and $(deliveryTime - inputTime)$ is the time that the data remains in the information system [BS06]. Timeliness expresses how current data are for the task at hand. This dimension is motivated by the fact that it is possible to have current data that are actually useless because they are late for a specific usage.

The codomain of availability, accuracy, and consistency is binary $\{"0","1"\}$. The codomain of reliability and timeliness is a normalized, continuous scale from zero to one $[0,1] \subset \mathbb{R}$. Finally, the codomain of volatility and currency is a temporal scale $\mathbb{R}^+$.

4. Deficit Analysis Visualization

An interactive application was developed for supporting the analysis of the deficit for the facilities system in Bogotá (Section 2). It supports the subtasks ST_1-ST_7. The application consists of 4 views: the Map View, the Map Overview, the DQD-Deficit Matrix, and the Deficit Detail View.

Map View The map view shows a map representing Bogotá and its political division into UPZ. This view supports the subtasks ST_1, ST_2, ST_4, ST_5, and ST_6. The six priority deficit sectors (Section 2) are visually encoded by using color hue, and the deficit value is categorized (four categories) and is mapped to a saturation level. Seven colors from ColorBrewer 2.0 [Bre11] were selected: six for the priority deficit sectors and one for the total deficit. The map is colored according to the deficit sector chosen and the saturation is computed for each UPZ. As a result, the Map View provides seven thematic visualizations, six for the priority sectors and one for the total deficit.

The data quality is integrated into this view to support the subtask ST_6. If the data is complete, its representation is used. If the data is incomplete, a white color is used for filling the area and no other quality value can be computed. Otherwise, a texture of yellow circles is used for representing the other DQ issues. Texture density encodes the index of data quality. This index is modeled using 3 categories according to the type and number of DQ problems: $category_0$: no DQ problem, $category_1$: one DQ problem, $category_2$ more than one DQ problem. For $category_0$ no texture is put, for $category_1$ a low density texture is applied, and for $category_2$ a high density texture is applied (Figure 1c). The texture is small and the color of the texture provides a good contrast to the different deficit colors. Thus, the texture fulfills the requirements that it does not interfere with the color coding of the deficit variables.

Considering the work of Griethe and Schumann [GS05] to DQD visualization, the following design alternatives could be used: (1) free visual variables, (2) integrating additional visual objects, (3) animation, (4) interactive representation, or (5) addressing other human senses. Free visual variables were selected, because this approach is well known and validated in the geographical visualization context. MacEachren et al. [MRO*12] propose the following visual attributes for representing data quality issues: fuzziness, location, value, texture, size, and transparency. Location and size are already used for representing the geographical information in the map. As color hue and saturation are used for representing the data attributes, value (brightness) and transparency are not used, because they would interfere with saturation. The use of fuzziness affects the contours, which are important in the geographical context. Finally, texture interferes least with the other visual variables used, therefore was used here.

The Map View interactions are navigation and the selection of a UPZ. Map translation is invoked by left mouse click, while zoom in and zoom out are invoked by mouse wheel. In both cases, the map overview is changed showing the visible part of the map using a black rectangle (Figure 1b). A UPZ is selected by left mouse double click and then highlighted using a cyan border. Further, the DQD-Deficit Matrix (Figure 1a) and the Deficit Detail View (Figure 1d) are updated using the data associated to the selected UPZ.

Map Overview The Map Overview visualization (Figure 1b) offers a global visualization of the complete map. The zoomed part of this map shown in the Map View visualization is marked by a black rectangle. It supports partially the subtasks ST_1 and ST_2. In this view two interaction operations are possible: selection and navigation.

DQD-Deficit Matrix The DQD-Deficit Matrix shows all data quality dimensions independently from the geographical data supporting the subtask ST_7 (Figure 1a). The DQD-Deficit Matrix is represented using a heat map. The DQ of each deficit and each DQD is mapped to the colors red, yellow, and green, with a meaning of bad, acceptable, and good data quality, respectively. The DQD are mapped to the X-axis, while the deficit sector and the seven deficit attributes are mapped to the Y-axis. The order of the DQDs was selected according to their importance. Further, the last three DQD are time dependent and they are sorted according to the logical order of estimating them, such as timeliness is calculated as function of volatility and currency. In the current application of this tool, no categorization for volatility is given. As a consequence, volatility is represented by a gray square with the volatility value shown inside. Additionally, the labels of the six priority deficit sectors are surrounded by colored rectangle to ease the relation between the views during the analysis.

Heat maps were chosen, because they are widely scalable [WF09]. They can visualize matrices describing the relation between two different attributes (Figure 1a). Moreover, additional rows and columns (DQD and deficits) can be added until pixel size is reached. Furthermore, the standard color coding used is intuitive and easy to understand by the users. The thresholds needed for this encoding are provided by the users–experts from the application domain.

Moving the mouse over a square in the DQD-Deficit Matrix: (1) highlights the square using a cyan border (Figure 1a), (2) shows deficit and DQ information of the respective row and column of the square (Figure 1a), and (3) highlights the associated bar (Figure 1d) and the associated table row (Figure 1e) of the Deficit Detail View using cyan borders. Selecting a deficit by left mouse click on its label changes the thematic map displayed in the Map View to the one of the selected deficit.

Deficit Detail View The deficit detail view contains a bar chart (Figures 1d) and a table (Figure 1e) and supports the subtask ST_3. In the deficit bar chart, the six priority sectors are represented by six bars and they are mapped to color hue. Additionally, each bar has the name of the sector (below the bar) and the value of the attribute (above the bar) attached. The table visualization uses the same color encoding in order to ease the analysis process. Moreover, this table shows additional attributes such as the population and the total deficit for each UPZ.

Mouse over a bar in this view highlights the deficit bar, the associated row in the DQD-Deficit Matrix, and the associated table row using cyan borders. Selecting a deficit by left mouse click on its bar changes the thematic map displayed in the map view to the one of the selected deficit.

5. Case Study

A case study was performed to assess the usefulness, usability, and effectiveness of the proposed visualizations and interactions. Therefore, it was evaluated, if the inclusion of data quality is beneficial without making the analysis task more difficult. The case selected represents the subtasks introduced in Section 2. Subtasks ST_1 and ST_2 were omitted as they are too general. The participants selected should have experience in working on projects related to the one described in Section 3, know about the current situation in Bogotá, and be familiar with GIS and similar applications.

The protocol of the study consists of three stages. First, an introduction to the test, a consent form, and a tutorial were given to each participant. The second stage comprised the assessment of three activities. Activity A_1 (related to ST_4 and ST_5) asked to identify the most and the least critical UPZ according to the total deficit and the six priority sectors. In Activity A_2 (related to ST_3) the participants had to select a random UPZ and identify the most and the least critical sector. Finally, in Activity A_3 (related to ST_6 and ST_7), the participants had to identify 4 UPZs according to the 4 DQ categories (Section 4) and quantify the number of errors of every DQ category (red squares in the DQD-Matrix). Finally, the third stage was a survey consisting of eight questions to get additional feedback. Six questions examined the usefulness and the usability of the application using a 5-Point Likert-Scale with an additional answer "I do not know". One question asked, which applications are used by the participants for this type of analysis, and one question asked about how often these are used.

Five expert participants, three women and two men, participated in the study. All participants have experience in the urban planning process and they have knowledge about the facilities system and the current situation in Bogotá. Four of them have more than two years of experience, and they obtained the experience by working on projects and doing research. Additionally, all participants are familiar with GIS.

The results of Activity A_1 show that it is easy to find the UPZ with the highest deficit due to the selected colorscale; however, it is hard to identify the UPZ with the lowest deficit because of several low values and similar colors in the colorscale. For Activity A_2, the bar chart facilitates the comparison of the deficits by sectors probably also due to having numbers over the bars. Finally, from the results for Activity A_3 it is possible to conclude that the four categories of DQ are distinguishable in the MapView. From the survey it was found that the layout is helpful and that using texture to map the DQ is effective and that it is easy to perceive the DQ in the MapView. In the DQD-Matrix the three categories are distinguishable, however a legend would be beneficial. Additionally, the participants consider that DQ features are important, complement the analysis, and they are ready to use them. Besides, the participants consider that the DQ analysis does not make the analysis more difficult and the evaluated application makes the analysis tasks easier to perform.

6. Conclusions

We presented an application that supports the specific task to analyze the facilities system for Bogotá. Linked views of specific visualizations eased the analysis process. Data quality dimensions were introduced warning the users of data problems. The DQD-Deficit Matrix, a novel visualization technique to represent data quality dimensions was proposed. It supports the analysis of data quality issues and is widely scalable to more than 100 attributes and 10 DQD. Moreover it is applicable to other domains. Additionally, a set of criteria was established in order to include data quality features into geographical maps. The application makes the analysis tasks easier. The DQ features are important and they do not make the analysis more difficult. The users are ready to include DQ features in their analysis.

7. Acknowledgements

The authors would like to thank all members of the Laboratorio Bogotá project of the Architectural School of Universidad de los Andes, and especially Isabel Arteaga for her remarkable contribution. This work was partially funded by the Colciencias PhD funding program Convocatoria nacional para Estudios de Doctorado en Colombia año 2012 and by the DAAD-Colciencias program PROCOL, the department of systems and computing engineering at Universidad de los Andes, and the Bundesministeriums für Bildung und Forschung BMBF (PROCOL).

References

[Bre11] BREWER C. A.: ColorBrewer 2.0. http://www.colorbrewer.org, accessed November 2013, 2011. URL: http://www.colorbrewer.org. 3

[BS06] BATINI C., SCANNAPIECO M.: Data Quality Concepts, Methodologies and Techniques. In *Data Quality Concepts, Methodologies and Techniques*. Springer Berlin Heidelberg, Berlin, 2006. 2

[CCP07] COLLINS C., CARPENDALE S., PENN G.: Visualization of Uncertainty in Lattices to Support Decision-making. In *Proceedings of the 9th Joint Eurographics / IEEE VGTC Conference on Visualization* (Norrköping, Sweden, Sweden, 2007), Eurographics Association, pp. 51–58. 2

[CWRY06] CUI Q., WARD M. O., RUNDENSTEINER E. A., YANG J.: Measuring data abstraction quality in multiresolution visualizations. *IEEE transactions on visualization and computer graphics 12*, 5 (2006), 709–16. URL: http://www.ncbi.nlm.nih.gov/pubmed/17080791, doi:10.1109/TVCG.2006.161. 2

[dlA09] DE LOS ANDES U.: *Sistema Distrital de Equipamientos - SDE: Componente Urbano*. Tech. rep., 2009. 2

[Epp01] EPPLER M. J.: A Generic Framework for Information Quality in Knowledge Intensive Industries. In *Sixth Conference on Information Quality (IQ 2001)* (2001), Katz-Haas E. M. P., Raissa, (Eds.), MIT, pp. 329–346. 2

[FPZH14] FERNÁNDEZ-PRIETO D., ZECKZER D., HERNANDEZ J. T.: UCIV 4 Planning: A User-Centered Approach for the Design of Interactive Visualizations to Support Urban and Regional Planning. *IADIS International Journal on Computer Science and Information Systems 8* (2014), 27–39. 2

[GS05] GRIETHE H., SCHUMANN H.: Visualizing uncertainty for improved decision making. *Proceedings of the 4th Conference on Business Informatics Research* (2005). 3

[GS11] GRIETHE H., SCHUMANN H.: The Visualization of Uncertain Data : Methods and Problems. 2

[HMC*13] HOLLT T., MAGDY A., CHEN G., GOPALAKRISHNAN G., HOTEIT I., HANSEN C. D., HADWIGER M.: Visual analysis of uncertainties in ocean forecasts for planning and operation of off-shore structures. *2013 IEEE Pacific Visualization Symposium (PacificVis)* (Feb. 2013), 185–192. 2

[KPP12] KANDEL S., PARIKH R., PAEPCKE A.: Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island (Naples), Italy, 2012). URL: http://dl.acm.org/citation.cfm?id=2254659. 2

[MRH*05] MACEACHREN A. M., ROBINSON A., HOPPER S., GARDNER S., MURRAY R., GAHEGAN M., HETZLER E.: Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science 32*, 3 (Jan. 2005), 139–160. doi:10.1559/1523040054738936. 2

[MRO*12] MACEACHREN A. M., ROTH R. E., O'BRIEN J., LI B., SWINGLEY D., GAHEGAN M.: Visual Semiotics & Uncertainty Visualization: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (Dec. 2012), 2496–2505. doi:10.1109/TVCG.2012.279. 3

[OM02] OLSTON C., MACKINLAY J.: Visualizing data with bounded uncertainty. *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002. 2002* (2002), 37–40. doi:10.1109/INFVIS.2002.1173145. 2

[Pan01] PANG A.: Visualizing uncertainty in geo-spatial data. *Computer Science and Telecommunications Board* (2001), 1–14. 2

[PKRJ10] POTTER K., KNISS J., RIESENFELD R., JOHNSON C.: Visualizing Summary Statistics and Uncertainty. *Computer Graphics Forum 29*, 3 (Aug. 2010), 823–832. doi:10.1111/j.1467-8659.2009.01677.x. 2

[PSB09] PHAM B., STREIT A., BROWN R.: Visualization of Information Uncertainty: Progress and Challenges. In *Trends in Interactive Visualization*, Liere R., Adriaansen T., Zudilova-Seinstra E., (Eds.), Advanced Information and Knowledge Processing. Springer London, London, 2009, pp. 19–48. URL: http://link.springer.com/10.1007/978-1-84800-269-2, doi:10.1007/978-1-84800-269-2. 2

[RWX*07] RUNDENSTEINER E., WARD M., XIE Z., CUI Q., WAD C., YANG D., HUANG S.: XmdvtoolQ: quality-aware interactive data exploration. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data* (Bejing, China, 2007), ACM, pp. 1109–1112.

[SEG05] SULO R., EICK S., GROSSMAN R.: DaVis: a tool for visualizing data quality. *Posters Compendium of InfoVis* (2005). 2

[WF09] WILKINSON L., FRIENDLY M.: The History of the Cluster Heat Map. *The American Statistician 63*, 2 (2009), 179–184. 3

[WYM12] WU Y., YUAN G., MA K.: Visualizing flow of uncertainty through analytical processes. *Visualization and Computer Graphics, IEEE Transactions on 18*, 12 (2012), 2526 – 2535. 2

[XWRH07] XIE Z., WARD M. O., RUNDENSTEINER E. A., HUANG S.: Integrating Data and Quality Space Interactions in Exploratory Visualizations. *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)* (July 2007), 47–60. doi:10.1109/CMV.2007.11. 2