Visual Support for Rastering of Unequally Spaced Time Series

Christian Bors, Markus Bögl, Theresia Gschwandtner, and Silvia Miksch
[christian.bors,markus.boegl,theresia.gschwandtner,silvia.miksch]@tuwien.ac.at
Vienna University of Technology (TU Wien)
Favoritenstrasse 9-11/188
Vienna, Austria 1040

ABSTRACT

Preprocessing is a mandatory first step to make data usable for analysis. While in time series analysis many established methods require data that are sampled in regular time intervals, in practice sensors may sample data at varying interval lengths. Time series rastering is the process of aggregating unequally spaced time series into equal interval lengths. In this paper we discuss critical aspects in the context of time series rastering, and we present a visual design which supports the parametrization of the rastering transformation, communicates the introduced uncertainties and quality issues, and facilitates the comparison of alternative rastering outcomes to achieve optimal results.

CCS CONCEPTS

•**Human-centered computing** →**Visual analytics;** •**Mathematics of computing** →*Time series analysis;*

KEYWORDS

visual analytics, time series analysis, time series rastering

ACM Reference format:
Christian Bors, Markus Bögl, Theresia Gschwandtner, and Silvia Miksch. 2017. Visual Support for Rastering of Unequally Spaced Time Series. In *Proceedings of VINCI '17, Bangkok, Thailand, August 14-16, 2017,* 5 pages.
DOI: 10.1145/3105971.3105984

1 INTRODUCTION

Preprocessing data, like cleansing and wrangling [7, 11], is the task of preparing data and transforming it into a usable form for subsequent analysis. Specifically with time series data, established analysis methods require the data to be structured appropriately, e.g., being equally spaced. By rastering a time series, unevenly distributed time points and their corresponding values are being aggregated and binned into evenly spaced time intervals, while still retaining the original data's structure. Rastering transforms the original data for the sake of consistent value distribution, smoother value curves, and reduced data size. However, to adequately transform a time series for subsequent analysis requires extensive knowledge about the data domain as well as temporal data characteristics.

A Motivating Example. For illustrating the challenges and critical aspects in time series rastering we give a particular example of unequally spaced time series sensor measurements from the environmental domain. Such measurements contain various formats and are used in many application domains. We illustrate our Visual Analytics (VA) approach on a dataset of the Opensense Project in Zurich [15], which measures different environmental variables, like meteorological data, air pollutants such as O_3, NO_2, NO, SO_2, VOC, and fine particles. The interval length, with measurements varying around 20 seconds (s), has the following properties: median of 20s, interquartile range (IQR) of 15s, and median absolute deviation (MAD) of 1.4826s.

Finding an adequate interval length for the rastering transformation is context specific and depends on domain properties. In this illustration, choosing a too short interval length, e.g., less than 20s in this example, would generate many raster intervals with no data, and therefore introduce missing values. On the other hand, too long intervals will mask interesting patterns in the time series. Immediate visual feedback on the new raster aggregation and important quality information are required to find an optimal configuration. This quality information includes introduced missing values, value ambiguity, or reduced temporal granularity.

In this paper (1) we list critical aspects and considerations in the context of time series rastering and suggest ways to handle intricacies in the data. (2) We present the conceptual design of an interactive visual framework for time series rastering, that streamlines the rastering process and communicates all the necessary information to the user. To facilitate the selection of appropriate parametrizations and and allow users to make an informed decision on the rastering outcome, (3) we provide quality measures and uncertainty information in a visual representation. This information is crucial for appropriately assessing the outcome of the time series rastering. In the course of this paper we present a conceptual design that is subject to be implemented and evaluated in more detail in the future.

2 RELATED WORK

Rastering unequally spaced time series tackles the areas of statistical time series analysis, data wrangling and cleansing, and data uncertainty.

Data wrangling is concerned with transforming and reformatting data into a different structural representation, Kandel et al. [11] characterize it as "a process of iterative data exploration and transformation that enables analysis". Data cleansing describes the task of validating data and dealing with erroneous entries, existing approaches allow users to investigate possible error sources and their causes in different domains, e.g., time oriented data [7], heterogeneous, multivariate, tabular data [12, 17]. Kandel et al. [12, 13]

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
VINCI '17, Bangkok, Thailand
© 2017 ACM. 978-1-4503-5292-5/17/08…$15.00
DOI: 10.1145/3105971.3105984

present sophisticated probing and overview features paired with transformations, Heer et al. [9] provide users with predictions for data transformations. Gschwandtner et al. [7] put particular focus on cleansing and profiling time-oriented data, signaling the importance to handle time’s intricacies. In wrangling and cleansing applications, time-oriented data is represented differently, depending on the goals for cleansing/wrangling: heatmaps for observing the time distribution, line charts for detecting anomalies [7]; Bernard et al. [4] employ line charts to assess the impact of sampling on time-series data.

Numerous taxonomies characterize data quality and systematic classifications of error types [2, 3, 14, 18], Gschwandtner et al. [8] derived a specific error characterization for time-oriented data. In data quality assessment quality metrics are utilized to retain consistent and reusable quality measures [3]. Sacha et al. [21] examined the role of uncertainty for generating awareness and building trust in the analysis process. MacEachran [16] debated the challenge of enabling reasoning under uncertainty and coping with uncertainty during decision-making. With our proposed workflow we aim to externalize and communicate to the user the risks and caveats that arise when analyzing data without adequate knowledge about the data’s quality and possible inherent uncertainties.

In statistics the majority of methods focus on equally spaced time series analysis [6], while for unequally spaced time series, there are only few specialized methods, e.g., [10]. Thus, unequally spaced time series are rastered into equally spaced intervals before applying these well-established statistical methods. There are several disadvantages that come with the process of rastering, that require consideration, e.g., data loss, dilution, and loss of time information (cf. [6]). Some of these cannot be avoided, but others just need special consideration. With our visual rastering framework enriched by quality and uncertainties information, we try to compensate these disadvantages. For instance, transforming unequally to equally spaced time series relates to interpolation problems in many domains, like signal processing and geostatistics [5]. With our approach we allow the integration of these interpolation techniques into the aggregation step.

3 RASTERING TIME SERIES DATA

We propose a VA approach for time series rastering that integrates data quality and uncertainty measures to provide essential information for the rastering process, and to produce output metadata that give comprehensive information on the preprocessing itself. These measures increase the awareness of the introduced uncertainties and quality issues for further analysis. For discussing the critical rastering aspects below, we consider relevant characteristics of time and time series data from the work by Aigner et al. [1].

3.1 Critical Rastering Aspects

In many cases, automatic methods for rastering time series data are not effective due to mutually exclusive dependencies, e.g., reducing the amount of empty rasters and minimizing loss of accuracy. During data transformation and aggregation uncertainty information is likely to be introduced, as the data’s structure is altered and sampling operations are applied. By sampling or aggregating values, the original measurement accuracy is lost. Current data processing systems merely store this information indirectly (i.e., provenance aware systems) if at all. By externalizing this uncertainty information, users are made aware of the impact of different rastering operations.

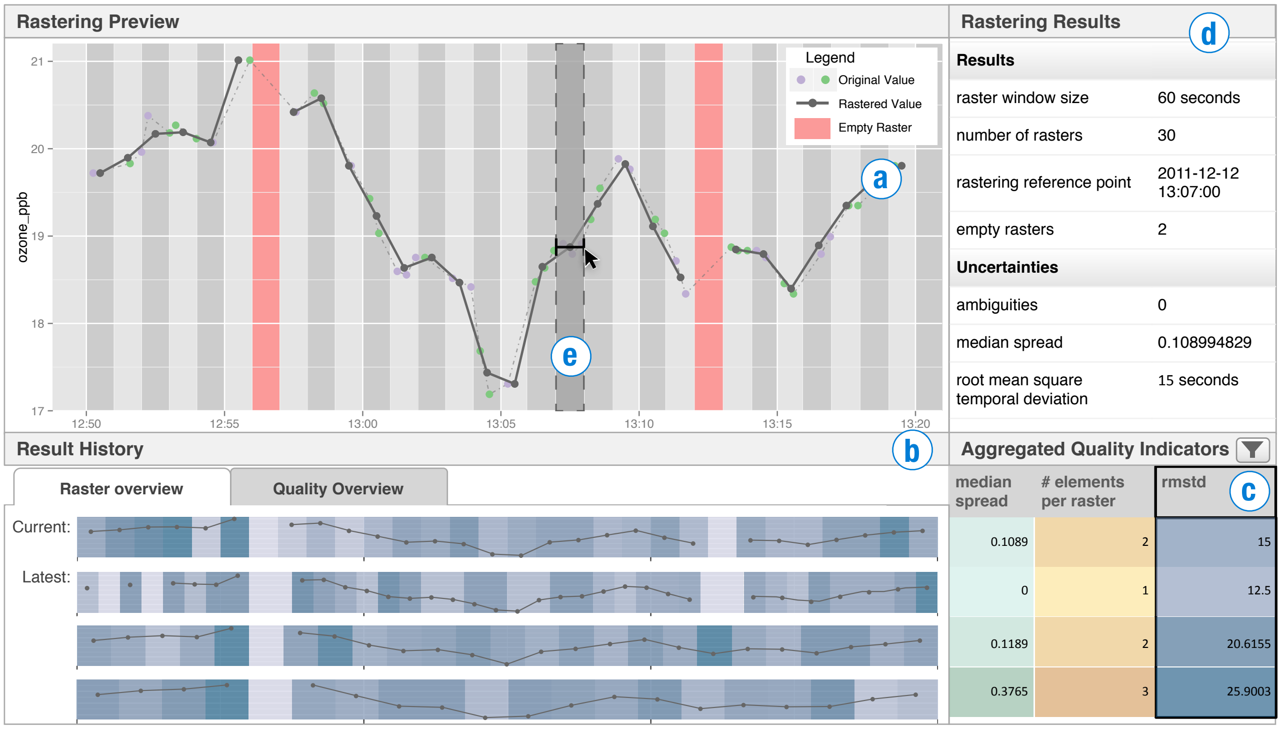
Data quality information can also be helpful to assess effects of rastering operations on datasets. Data quality metrics [19]—proportional measures of data quality dimensions [20]—quantify quality aspects to give expressive assertions to certain data properties. We aim at introducing quality metrics that are specific to time-series data to allow informed rastering. Iterating upon these different aspects we present a list of contingencies that need to be considered when rastering different types of time-oriented data.

Types of Time-oriented Data. Time series data can reflect either state or interval recordings (cf. [1]). State changes occur either at the exact time an entry was recorded, or have changed at any time since the last measurement. This implies an inherent uncertainty within the value domain. Inappropriate aggregation further increases error margins. When rastering a time series, the user needs to be aware of varying uncertainties with respect to different input time series and different rastering parametrizations. Consequently, the time series visualization requires appropriate representation considering these influential factors and results. When considering aggregation on time series data containing intervals, original intervals are potentially split up due to incompatible raster lengths. The time series visualization should represent the time intervals appropriately, and the rastering algorithm should feature options to allow retaining the original intervals’ sizes or proportionately creating new rasters from multiple intervals.

Temporal Granularity. If time series need to be rastered with finer granularity than provided by the original data, data values of one interval need to be divided into smaller intervals – this division must be done based on assumptions, e.g., by computing a time series model based on the input data and super-sampling entries. Analogously, if the time series is rastered into a coarser granularity, details can get obfuscated, e.g., masking outliers by smoothing the time series through aggregation, and classical error margins may get broader. Depending on the goal of the user this is undesirable and should be indicated accordingly.

Ambiguity. It is implied that ambiguities might be introduced into the data during rastering, specifically when dealing with qualitative or discrete data values. Aggregating or sampling values during rastering often requires imputation from time series, or reducing raster granularity. Introduced ambiguity should be marked as such and explicitly communicated in further analysis steps. This information potentially influences analysis, specifically if users are unaware of inherent ambiguities and assume the data as explicitly correct.

Quality and Uncertainty Measures. Statistical measures, like data spread, temporal deviation, or data point density per interval provide important domain independent quality information for finding adequate rasterings. The amount of uncertainty that is introduced by different configurations needs to be considered when trying to identify a suitable rastering of the time series at hand. To construct more expressive measures, quality metrics should be calculated on different granularity levels to give both local information and



Results	
raster window size	60 seconds
number of rasters	30
rastering reference point	2011-12-12 13:07:00
empty rasters	2
Uncertainties	
ambiguities	0
median spread	0.108994829
root mean square temporal deviation	15 seconds

median spread	# elements per raster	rmstd
0.1089	2	15
0	1	12.5
0.1189	2	20.6155
0.3765	3	25.9003

Figure 1: An overview of our interactive time series rastering approach. (a) The interactive rastering preview allows defining the raster window size through *drag&drop* interaction as well as comparing the current raster configuration to the original data. Alternating consecutive raster backgrounds and original value point colors per raster facilitates distinction. Empty raster intervals are highlighted by red segments. (b) In the *result history* view users can compare previous rastering results represented by small multiple line charts. Selecting a quality indicator (in *c*) overlays the small multiples with a heatmap of individual quality measures per raster. This view can be switched to the *quality overview* which gives multivariate quality and uncertainty information on recent raster configurations (cf. Fig. 2). (c) The *aggregated quality and uncertainty indicator* view features a sortable and customizable heatmap view representing the aggregated quality and uncertainty measures for each raster configuration in the history view. Color intensity corresponds to higher values (cf. Fig. 2). (d) Overview information of rastering results, including meta information, general uncertainty measures, and introduced quality issues based on calculated quality metrics. (e) *Drag&Drop* interactive selection of rastering window length.

allow comparisons to overall granularity measures. This allows the user to find local anomalies which could result in reconsidering the current rastering configuration.

Missing Values. Empty intervals can signal quality issues or inappropriate raster window size. The distribution and amount of empty rasters offers valuable clues for finding a suitable rastering. Direct comparison to the original data allows users to assess if empty values are caused by missing values or inappropriate rastering settings.

Temporal and Data Anomalies. With robust outlier detection measures, outliers can be automatically identified and highlighted. However, judging if these outliers – either in the temporal domain or in the data domain – represent anomalies requires additional contextual information. Thus, it takes the user's domain knowledge to reason about the identified outliers. As such, marking outliers as well as rasters which contain outliers and saving this meta information for subsequent analysis is advisable and allows more informed decisions.

3.2 Visual Analytics Approach

In this section we conceptualize a workflow for rastering unevenly spaced time series data and illustrate the application of these principles in our VA approach. In Figure 1 we show a mockup with a design that supports the workflow discussed below. A description of the composed multiple-coordinated views and their use can be found in the respective caption.

For the rastering transformation, an interval length needs to be determined that is appropriate for the original dataset and the usage of the transformed data. The optimal raster window size depends on the data domain, different data characteristics, the further usage of the data, quality information, and introduced uncertainties. In the current state, the user can interactively choose a raster window size (cf. Fig. 1e). For future work an algorithm could suggest several adequate raster window sizes, depending on specific properties and/or data quality metrics specified by the user. Subsequently, the provided quality and uncertainty measures need to be interpreted in the light of the users' domain knowledge in order to draw correct conclusions from the rastering result. To assist the user in supervising the rastering process and determining optimal rastering results, our approach considers the special characteristics of time-oriented data to provide important contextual information. Moreover, time's inherent structure is used for calculating statistical measures (cf. Fig. 1c,d).

The time series *rastering preview* (cf. Fig. 1a) is interactively browsable, showing a superimposition of both the original time series and a preview of the results of the current rastering configuration. This view also serves as input interface for defining the raster length and initial raster anchor point. These parameters are selected via *drag&drop* (cf. Fig. 1e) in the rastering preview. During dragging, the raster values are calculated and interactively updated based on the current configuration (grey dotted line in Fig. 1a). The multiple-coordinated views are dynamically updated during the dragging interaction to show the impact of the chosen configuration on the rastering outcome, like raster length, distribution, and possible empty rasters.

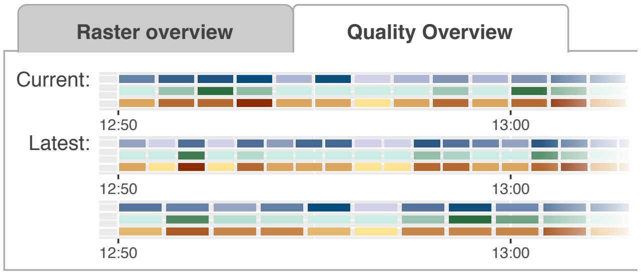


Figure 2: Result quality overview (*truncated*): In this view the user can directly compare quantified quality and uncertainty information of the raster result history for individual rasters encoded in a colored heatmap. The context of the coloring corresponds to the aggregated indicators in Figure 1c and helps to identify conspicuous entries.

We argue that knowledge about data quality and uncertainty facilitates the rastering and assessment process for users. Contextual information on time-oriented data in the form of quality metrics and uncertainty information allow the user to appropriately prioritize certain characteristics and assess rastering results, e.g., minimizing the median spread but consecutively disregarding the actual raster window size. These measures are shown in an aggregated overview to allow comparison with previous configurations (cf. Fig. 1c).

Besides showing the global quality and uncertainty information, the *result history view* (cf. Fig. 1b) shows a juxtaposition of previous rastering results as either small multiple time series line charts (*raster overview*) detailing single quality measures, or as heat bands for multiple quality and uncertainty measures (*quality overview*, cf. Fig. 2). The view is interactively browsable and helps users to visually assess different rastering parametrizations. The *result history* view furthermore allows for comparison between the latest rastering results to determine an optimal configuration where e.g., empty rasters are minimized without losing too much detail information due to value aggregation. The view can be interactively browsed to facilitate the exploration and validation of large time series.

To compare quality and uncertainty information, the *quality overview* (cf. Fig. 2) allows to analyze quality information for individual rasters. E.g., if quality metrics indicate that ambiguities or missing values are less frequent in a particular raster configuration, it could pose significant benefits over small decreases of accuracy. With these analysis options at the user's disposal, the awareness about the influence of rastering transformations on the quality and uncertainty measures is increased. The approach allows to compare alternative rastering configurations with respect to the critical aspects outlined above and to the desired properties of the data for subsequent analysis tasks. Figure 1d gives a comprehensive overview of different properties from the current raster parametrization.

4 CONCLUSION AND FUTURE WORK

This paper conceptualizes a VA approach for rastering unequally spaced time series. We discussed critical aspects of time and time series data and suggested to use uncertainty and data quality information to support an informed rastering process, and thus, improve rastering results. Our solution provides immediate visual feedback to facilitate the complex task of adjusting rastering parameters and assessing the quality and suitability of the rastered output time series for further use.

What we have not yet considered in our visualization design is the proper handling of interval type time series. We plan to address this issue in future work by adapting our visualization design to appropriately represent this type of time series data. We will implement the presented prototype and extend its means to support the rastering of different types of time series. Moreover, we plan to add more sophisticated uncertainty and quality measures and plan to extend its functionality in order to provide a comprehensive tool-set for this kind of task. For instance, implementing automatic suggestions for optimal raster configurations that take the respective uncertainty and quality indicators into account would be a substantial addition. However, this is contemplated as a semi-automatic approach. In order to determine optimal results for the time series at hand, users still need to leverage their contextual domain knowledge and fine-tune the parameters and constraints of the rastering transformation via dynamic querying. Thus, we consider a VA approach with immediate visual feedback to user actions as key to support this complex task.

With the work presented in this paper we (1) outline important considerations and critical aspects in the context of time series rastering and (2) we propose a visual design that communicates these aspects to the user to better support the complex task of identifying optimal results.

ACKNOWLEDGEMENTS

This work was supported by: Austrian Federal Ministry of Science, Research, and Economy via CVAST (#822746), a Laura Bassi Centre of Excellence; and Austrian Science Fund (FWF) via VISSECT (#I2850-N31).

REFERENCES

[1] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. 2011. *Visualization of Time-Oriented Data.* Springer, London, UK. 286 pages. https://doi.org/10.1007/978-0-85729-079-3

[2] Jos Barateiro and Helena Galhardas. 2005. A Survey of Data Quality Tools. *Datenbank-Spektrum* 14, 15–21 (2005), 48. http://dc-pubs.dbs.uni-leipzig.de/files/Barateiro2005ASurveyofDataQuality.pdf

[3] Carlo Batini and Monica Scannapieco. 2006. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications).* Springer Verlag New York, Inc., Secaucus, NJ, USA.

[4] Jürgen Bernard, Tobias Ruppert, Oliver Goroll, Thorsten May, and Jörn Kohlhammer. 2012. Visual-Interactive Preprocessing of Time Series Data. In *Proceedings of SIGRAD: Interactive Visual Analysis of Data (Linköping Electronic Conference Proceedings)*, Andreas Kerren and Stefan Seipel (Eds.), Vol. 81. Linköping University Electronic Press, Linköping, Sweden, 39–48. http://dblp.uni-trier.de/db/conf/sigrad/sigrad2012.html#BernardRGMK12

[5] Jean-Paul Chilès and Pierre Delfiner. 2012. *Geostatistics: modeling spatial uncertainty* (2. ed. ed.). Wiley, Hoboken, NJ.

[6] Andreas Eckner. 2014. A Framework for the Analysis of Unevenly Spaced Time Series Data. (July 2014). Retrieved from http://eckner.com/research.html on (22. Feb. 2017).

[7] Theresia Gschwandtner, Wolfgang Aigner, Silvia Miksch, Johannes Gärtner, Simone Kriglstein, Margit Pohl, and Nik Suchy. 2014. TimeCleanser: A Visual Analytics Approach for Data Cleansing of Time-oriented Data. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business (i-KNOW ’14).* ACM, New York, NY, USA, 18:1–18:8. https://doi.org/10.1145/2637748.2638423

[8] Theresia Gschwandtner, Johannes Gärtner, Wolfgang Aigner, and Silvia Miksch. 2012. A Taxonomy of Dirty Time-Oriented Data. In *Lecture Notes in Computer Science (LNCS 7465): Multidisciplinary Research and Practice for Information Systems (Proceedings of the CD-ARES 2012)*, Gerald Quirchmayr, Josef Basl, Ilsun You, Lida Xu, and Edgar Weippl (Eds.). Springer, Berlin / Heidelberg, Prague, Czech Republic, 58–72. https://doi.org/10.1007/978-3-642-32498-7_5

[9] Jeffrey Heer, Joseph Hellerstein, and Sean Kandel. 2015. Predictive Interaction for Data Transformation. In *Conference on Innovative Data Systems Research (CIDR).* Asilomar, California, USA. http://idl.cs.washington.edu/papers/predictive-interaction

[10] Richard H. Jones. 1985. Time series analysis with unequally spaced data. In *Handbook of statistics*, E. J. Hannan, P. R. Krishnaiah, and M. M. Rao (Eds.). Vol. 5. Elsevier, Amsterdam, Netherlands, 157–178.

[11] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 10, 4 (2011), 271–288. https://doi.org/10.1177/1473871611415994

[12] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’11).* ACM, 3363–3372. https://doi.org/10.1145/1978942.1979444

[13] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI ’12).* ACM, 547–554. https://doi.org/10.1145/2254556.2254659

[14] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. 2003. A taxonomy of dirty data. *Data mining and knowledge discovery* 7, 1 (2003), 81–99. http://link.springer.com/article/10.1023/A:1021564703268

[15] Jason Jingshi Li, Boi Faltings, Olga Saukh, David Hasenfratz, and Jan Beutel. 2012. Sensing the Air We Breathe: The Opensense Zurich Dataset. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI’12).* AAAI Press, 323–325. http://dl.acm.org/citation.cfm?id=2900728.2900775

[16] Alan M. MacEachren. 2015. Visual Analytics and Uncertainty: It’s Not About the Data. In *EuroVis Workshop on Visual Analytics (EuroVA)*, E. Bertini and J. C. Roberts (Eds.). The Eurographics Association. https://doi.org/10.2312/eurova.20151104

[17] John Morcos, Ziawasch Abedjan, Ihab Francis Ilyas, Mourad Ouzzani, Paolo Papotti, and Michael Stonebraker. 2015. DataXFormer: An Interactive Data Transformation Tool. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD ’15).* ACM, New York, NY, USA, 883–888. https://doi.org/10.1145/2723372.2735366

[18] Paulo Oliveira, Ftima Rodrigues, and Pedro Rangel Henriques. 2005. A Formal Definition of Data Quality Problems.. In *IQ.* http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/AFormalDefinitionofDQProblems.pdf

[19] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data Quality Assessment. *Commun. ACM* 45, 4 (April 2002), 211–218. https://doi.org/10.1145/505248.506010

[20] Thomas C. Redman. 2012. Data Quality Management Past, Present, and Future: Towards a Management System for Data. In *Handbook of Data Quality*, Shazia Sadiq (Ed.). Springer Berlin Heidelberg, 15–40. http://link.springer.com/chapter/10.1007/978-3-642-36257-6_2

[21] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. 2016. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan 2016), 240–249. https://doi.org/10.1109/TVCG.2015.2467591