Workshop on Visualisation in Environmental Sciences (EnvirVis) (2018)
R. Bujack and K. Rink and D. Zeckzer (Editors)

How To Look at Data: Environmental Practitioners' Lens Through Two Case Studies

Meng Ling [†1], Jeffrey A. Johnson[2], Zhiquan Feng[3,4], and Jian Chen[5]

[1]Environmental consultant, Columbus, OH, USA
[2]NewFields, Houston, TX, USA
[3]School of Information Science and Engineering, University of Jinan, China
[4]Shandong Provincial Key Laboratory of Network-based Intelligent Computing, China
[5]Department of Computer Science and Engineering, The Ohio State University, USA

Abstract

Exploratory data visualization, an idea proposed by Tukey in 1977, is meant to output various types of visualization in order to make the data more understandable. While visualization researchers have generated many novel solutions to critical and complex environmental problems, to everyday environmental consultants, some practical considerations have to be made in the visualization analysis to help stakeholders generate and test hypotheses that would not be possible otherwise. We present two environmental case studies of using visualization to communicate key findings: constructing stratigraphic units (layered) and generating groundwater contaminant plumes (volumetric). These real-world cases show that many times visualization alone may not give us correct answers; often what works is the combination of visualization, domain experts' knowledge, and interpretation of the visualization solutions. The lack of any of them may lead to faulty conclusions. The first case study illustrates together how domain experts, visualization, and contamination conditions assisted in interpreting limited and ambiguous lithologic data. The second case study emphasizes conceptual and technical understanding and discusses some common factors affecting 3D interpolation, which again suggests that we must incorporate domain experts' knowledge as well as analytics into visualization for defensible decision making.

1. Introduction

Three-dimensional (3D) interactive visualization tools have become increasingly accessible to environmental practitioners, from the application areas of air and water quality assessment to recent climate changes and interdisciplinary biological sciences. Fundamental innovative solutions range from multivariate heterogeneous data visualizations [JS14] to remote sensing studies [SDS*17] [CLX*17]. Visualizations have become indispensable in assisting environmental practitioners in interpreting and analyzing complex subsurface problems [CRF05] [LC14].

One key concern for these visualization tools is that other than showing the subject of study in a visually appealing and appropriate form, they should assist in the interpretation and analysis of spatial data. Tukey defined in 1977 [Tuk77] such visually aid analysis as exploratory data analysis (EDA), an approach to analyzing datasets by summarizing data characteristics with visual method. For example, it is always good to examine residues after an insight is achieved from a linear model to see if there are systematic errors due to the analytical process. A deeper visual analysis as such is associated with the concept of *validation*, where visualizations help scientists ensure the correctness of a conceptual or mathematical model with the salient aspects of reality [BO04]. Here, environmental practitioners use visualizations to understand model validity and accuracy.

It is known that a domain expert with no knowledge of visualization can discover a great deal of information in a body of data. In our case, the domain expert is also a visualization expert, allowing discoveries and errors to be revealed more quickly with higher certainty, as the expert knows how and what to look for using the visual tools. This is particularly helpful in a situation involving a large number of potential factors that might influence the response measure individually or in certain combinations. Visualization tools published in the visualization domain tend to support general data analysts to search for data in an organized fashion. How a domain expert searches for goal-driven solutions through visualization analyses is described in this paper.

In the industry, it is common to pass the data and task to visualization experts and let them handle the analysis. This may sometimes result in not being able to explore the data from the domain expert's perspective in the early design stage. Some investigators may believe there is only one best way to look at the data. Such

† Corresponding author email: mling216@gmail.com.

DOI: 10.2312/envirvis.20181136

a belief could be due to limited imagination or over-reliance on the traditional methods developed by others. We show that an iterative process of observing the data using visualization must be introduced to lead to effective and efficient decision making.

In this paper we use two applications to demonstrate the practical aspects of *what* to consider when applying visualization analysis to common environmental problems. The first application is about constructing stratigraphic units from limited boring lithology data. The second application discusses generating conservative groundwater contaminant plumes for prediction purposes. How to interpret ambiguous data and balance the factors affecting 3D interpolation are discussed.

2. Related Work

Environmental or geospatial data in general has focused on the multivariate data exploration paradigm of discovery where scientists synthesize diverse data sources to find relationships and distributions [JS14].

Many environmental applications require 3D interpolation of lithologic/structural data (layered) or concentration data (volumetric). Interpolation is typically conducted with 3D kriging, which depends on a geostatistical correlation model (i.e., semivariogram) and is capable of providing some levels of uncertainty analysis [IS89] [DJ92]. In the case when data are limited, pseudo control points are often used to help supplement the data, confine and smooth the spatial distribution, and/or facilitate uncertainty analysis. There are also cases where control points should be avoided.

The EDA paradigm for environment data analysis is based on a desire to let the data speak for themselves without biases. The emphasis is not on creative data display but the use of simple indicators to elicit patterns and produce hypotheses in an inductive manner, while avoiding potential misleading "atypical" observations [Tuk77]. As practitioners may differ in expertise, experience, and objectives, it could be difficult for them to reach a consensus on what data to use and how to interpret, and vastly erroneous estimates may result from unforeseen conceptual and technical errors. In many cases, it is not the perfection of an estimate or visualization, but its practical utility that matters - essentially a validation process. For example, for problems with inherent uncertainty, we found that providing a reasonable range of possible outcomes could be more defensible and acceptable for stakeholders.

Many researchers study semiotics approach (e.g., the study of symbols [Ber83]) to encode data to communicate key ideas or tell stories about data [KW05]. The challenge is to tell a "convincing" story of data. Recent novel and fascinating remote sensing applications that enable capturing multivariate datasets to study environmental changes [SDS*17] and characterization [CDM10] [CLX*17] demand analytical solutions. Complementary to these novel exploratory paradigms, our work here focuses on the use of existing techniques to aid validation.

3. Case Studies

This section presents two applications where visualization was integrated in the analytical activities to assist environmental consultants to revisit modeling results and correct errors. The first case is constructing stratigraphic units. The whereabouts of the constituents of concern had to be evaluated to judge whether the constructed stratigraphic unit supports such a condition. In the second case of generating the methyl tertiary butyl ether (MTBE) plumes for model predictions, it is important to capture the potential upper-limit with a range of estimates that are reasonably but not erroneously large.

3.1. Stratigraphic Units - Interpreting Boring Lithology

Background. At a petroleum-hydrocarbons contaminated site historically used for fuel operations along the west coast of USA, a conceptual site model for the subsurface was developed based on site-specific boring lithologies and regional geology. The boring lithologies for deep borings that penetrate several lithologic units are illustrated in Figure 1. The unconsolidated stratigraphic units from top (youngest) down include Artificial Fill (sand and silt), Young Bay Mud (clay and silty clay), San Antonio Formation (sand), and Old Bay Mud (clay). A clayey unit within the San Antonio Formation ($SA_{Aquitard}$, Figure 1) was observed present under much of the site. This unit functions as an aquitard separating the San Antonio Formation into an upper unit and a lower unit.

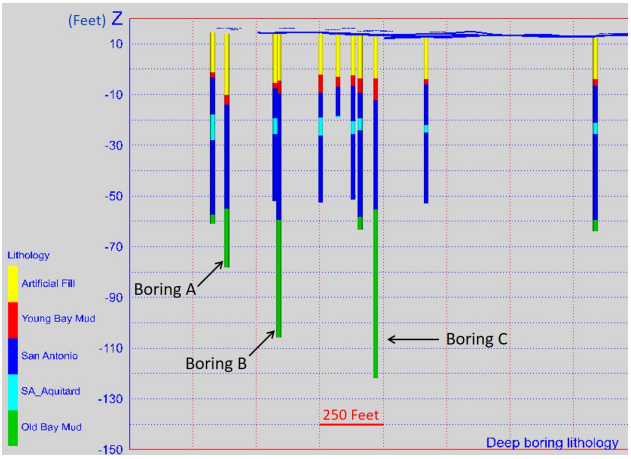

Figure 1: *Deep boring lithologies at the site.*

Key analytical task. The question of concern was whether this clayey unit is continuously present underneath the site. This was important as a continuous aquitard would likely prevent constituents of concern from entering the lower sandy unit.

Visual validation process to resolve mismatch between model and source data. From the boring lithologies shown in Figure 1, the answer is no as it was not identified at three boring locations A, B, and C (i.e., no cyan segments within the blue intervals). During validation, multiple testing methods were adopted by the domain expert through interpolation methods.

The boring lithology data were translated into pinched-out or zero-thickness zones in this clayey unit around the three boring locations. These pinched-out zones were visualized by inserting pseudo lithologic points at the three locations representing the top and bottom of the clayey unit. The vertical locations of these points

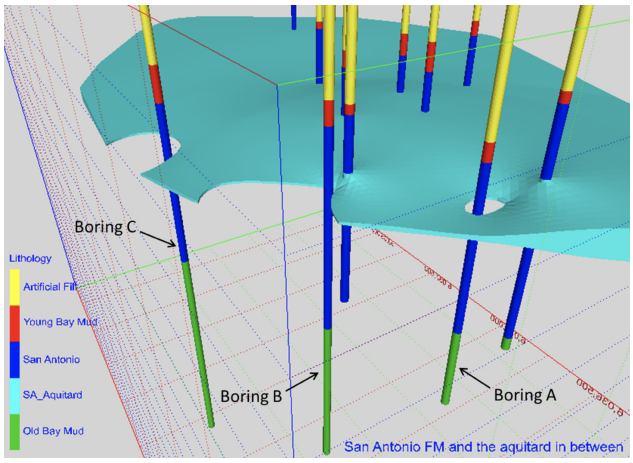

Figure 2: *Pinched-out aquitard at boring locations A, B, and C. Spacings of the horizontal and vertical axes are 250 feet and 10 feet, respectively*

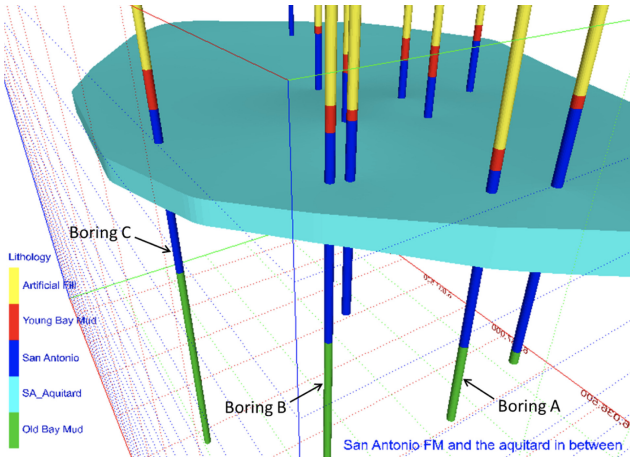

Figure 3: *Continuous aquitard interpreted from other borings. Spacings of the horizontal and vertical axes are 250 feet and 10 feet, respectively*

were to some extent a guess work that is more of art than science, although considerations were given to follow the spatial trends of the surrounding borings. The pseudo unit was made very thin (e.g., 0.2 feet) at boring locations A, B, and C, and a slightly larger cut-off thickness (e.g., 0.5 feet) was used to visualize the pinched-out zones (Figure 2).

Groundwater monitoring data, however, seemed to contradict the pinched-out version of the aquitard. Constituents of concern were detected in some wells screened in the upper unit of the San Antonio Formation but were never detected in wells screened in the lower unit of the formation. Although arguments could be made as to why the constituents of concern had not been detected in the lower unit albeit the pinched-out zones, a further investigation was deemed necessary to try to reconcile this contradiction.

The domain expert's knowledge plays a key role in the analysis process. The domain expert provided an in-depth review of available information and indicated that the lithology data from borings A, B, and C did not support interpreting a zero-thickness aquitard within the San Antonio Formation at these locations. These borings were deep geotechnical borings drilled in the late 1970s, and the original logs were generalized and hand-drafted. There was no sample interval information on the original logs, and the limited number of blow counts (the number of hammering on the rig while drilling down a sampler) recorded suggested that the sampling was sparse at depth. In particular, there were few blow counts recorded near the aquitard depth.

Insights acquired. After coupling visual analysis with domain knowledge, it was determined that the sampling in these logs was too sparse to allow picks for the aquitard within the San Antonio Formation. A new interpretation of this aquitard was conducted by using only the lithologic picks from other borings, and the aquitard thickness at boring locations A, B, and C was determined by the interpolation/extrapolation (Figure 3). Although this represents only one possibility, decades of monitoring data have not proven a pinched-out aquitard to be more plausible.

One of the most important things is the way simple visualizations can support science communication. Even the underlying sampling data are complex, easy-to-understand visualization has helped make complex issues more graspable and help the team arrive consensus. The analysis process is iterative. Having a double-expert who understands both visualization and the application domain helps make visualization reach the audience aiming at specific problems. The work however does not necessarily follow perceptual principles. So visual literacy become more and more important because good visualization could have supported more accurate and faster understanding of data.

3.2. Groundwater Contamination - Distribution, Mass, and Prediction

Background. A public well field in the western USA was impacted by MTBE from operations of nearby petroleum products service stations. The well field, consisting of several production wells screened in deep aquifers, was shut off since the detection of MTBE impact in the 1990s. Remedial activities involving extraction of the impacted groundwater at the petroleum operation facilities had drawn the MTBE plume away from the well field towards the remedial systems. Pressing needs to supply more water to the region required resuming operation of the well field. Restart of the operation would draw the MTBE plume back into the well field because its pumping rate would be much larger than that of the remedial systems. The proposed solution was to build a treatment plant to treat the groundwater from the well field before distributing. The quantity and concentrations of the pumped groundwater were key factors affecting the scale and treatment capacity of the proposed treatment plant. In particular, the maximum possible concentrations affected the selection of treatment technologies.

Analysis methods. To predict the MTBE concentrations from the well field and answer what-if type of questions, a groundwater flow and transport modeling study was conducted, and its success

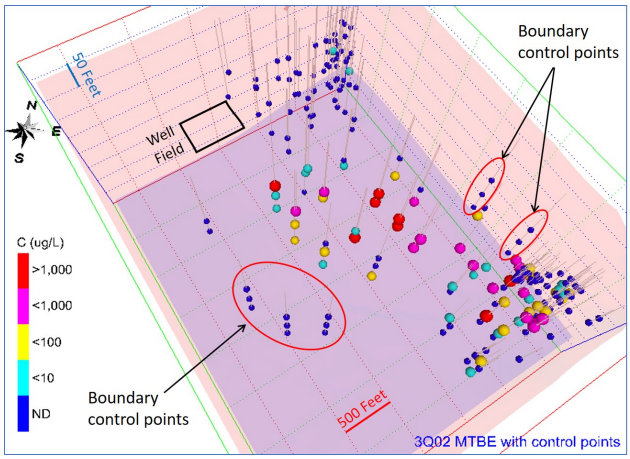

Figure 4: *Monitoring data used for generation of the expected case MTBE plume. MTBE concentrations are visualized as color spheres at the mid-points of well screens. Spacings of the horizontal and vertical axes are 500 feet and 50 feet, respectively*

depended on a reasonable estimate of the distribution and mass of the 'initial condition' MTBE plume. The estimate started with selecting a set of monitoring data that contained more data points at relatively high concentrations. This conservatively selected dataset used to estimate the initial condition are illustrated in Figure 4.

Visualization solution. The MTBE plume was estimated through a visual analysis that couples 3-D kriging interpolation and visualization techniques. The 3-D kriging interpolation technique was particularly useful as it not only provided a quantitative estimate of the plume's distribution and mass, but also enabled an assessment of the estimate' uncertainty. Through the assessment of uncertainty, an acceptable level of confidence can be established for the ensuing decision-making.

Main observations and insights introduced by visual analysis. During the visualization analysis, it was noticed that the kriging-estimated distribution and mass were affected by a number of factors and could be vastly different [LJL07]. The analytical methods used are most similar to the insight-based approach of Saraiya, North and Duca [SND05] in that observations are made when the domain experts used the tool.

The factors observed through visual analysis include nearly all analytical stages from grid placements to modeling process.

- Bounding domain and grid spacing. A bounding domain of certain shape and size and the placement of control points are often used to artificially "delineate" the plume, as in practice a plume is rarely 100% delineated in all directions. Grid spacing was found to determine the resolution of the interpolation and affects whether the interpolation is overdone or underdone.
- Vertical representation of well concentration. Concentrations from a traditional monitoring well are usually considered vertically mixed throughout the screen interval. Should the well concentration be assigned to the midpoint of the well screen or represented as a number of points along the well screen? This has little impact on short-screened wells but may have a large impact on long-screened wells.
- Nondetects (ND) cut-off level. Nondetects need to be quantified before being used in kriging interpolation, and are important in delineating the plume extent as different quantitations lead to different kriging results.
- Data transformation. Logarithmic transformation is widely used in the kriging analysis of organic chemical concentrations in groundwater and soil. Linear or other transformation may result in vastly erroneous estimates.

Three generated MTBE plumes that cover a plausible range of the MTBE distribution and mass for input into the groundwater model are presented in Figure 5. Plume A was the expected case based on the quarterly monitoring data shown in Figure 4. Plume B was the expected worst case generated by replacing some of the data points shown in Figure 4 with historical highest concentrations. Plume C was the upper 95% worst case representing the upper 95% confidence limits of Plume B. Figure 6 illustrates the difference in estimated concentrations through the same cross section for Plume A and Plume C. The estimated MTBE mass ranged from several hundreds of kilograms to over one thousand kilograms. These resulted in a range of predictions for the maximum MTBE concentrations and breakthrough curves in the to-be-reactivated well field.

The municipality that owns and operates the well field also made their own predictions on the maximum MTBE concentrations. Its predictions were much higher than those presented above, and thus no easy agreements could be reached between the municipality and the group of potential responsible parties. The case was later settled for hundreds of millions of dollars, and the municipality built the treatment facility and reactivated the well field in late 2010. Production rate for the reactivated well field was initially lower than what was simulated in the model study, but was raised to the simulated rate in less than three years.

A decade later, a retrospective review was conducted to compare past model predictions with actual observations at the well field. The combined MTBE concentrations from the well field were compared to model predicted concentrations based on the above-generated MTBE plumes. The results are presented in Figure 7 and it is clear that the actual concentrations are significantly less than the predicted concentrations for the expected case plume and the expected worst case plume (Plume A and Plume B in Figure 5, respectively). On the contrary, the municipality's predictions are unreasonably high. Note that none of the groundwater model simulations considered biodegradation of the MTBE and thus they overestimated the MTBE concentrations to be conservative. Also, the smaller production rate at the beginning of the well field reactivation would generate an elongated breakthrough curve at a lower concentration. This check against actual data proved that the above-described visualization analysis generated conservative MTBE plumes and reasonably covered the plausible range of the MTBE impact.

4. Conclusion and Discussion

The case studies presented above demonstrate the utility and value of applying visualization analysis to environmental problems in-

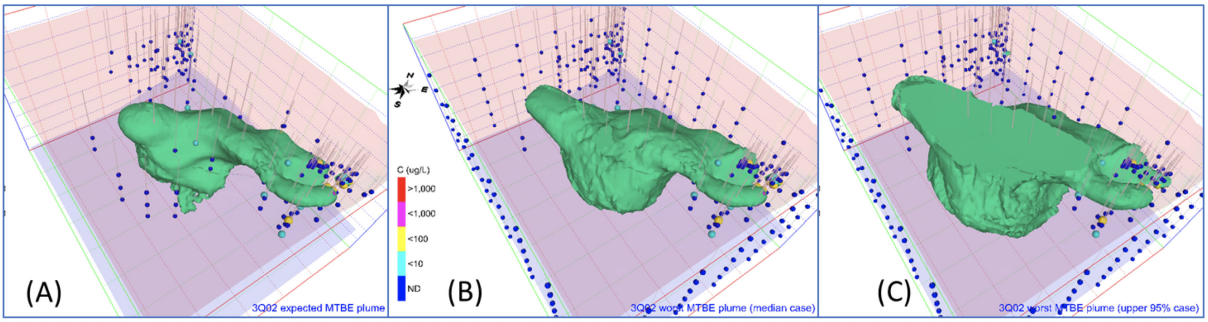

Figure 5: *MTBE plumes at 5 microgram per liter generated for (A) expected case, (B) expected worst case - using historical highest concentrations at certain points, and (C) upper 95% worst case. Spacings of the horizontal and vertical axes are 500 feet and 50 feet, respectively*

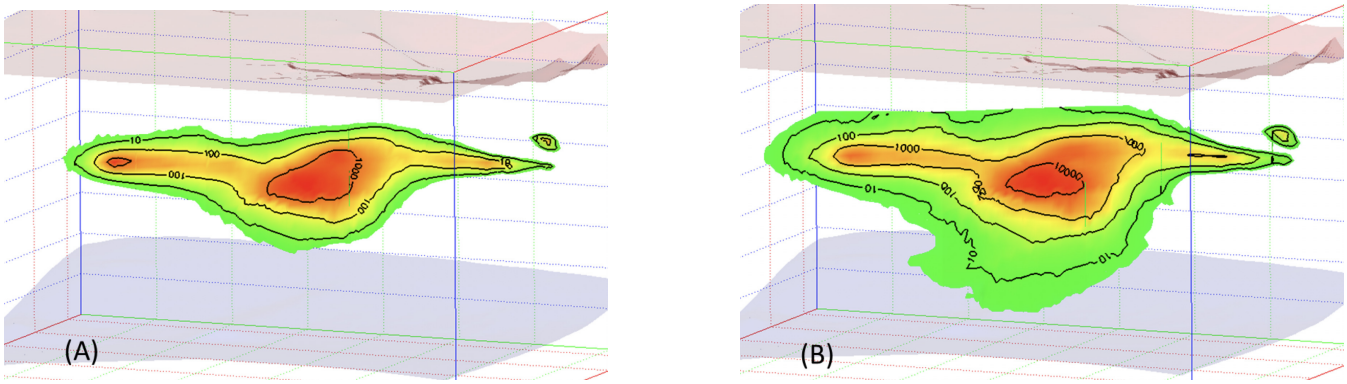

Figure 6: *Figure 6. Cross sections illustrating the difference in plume concentrations for (A) expected worst case and (B) upper 95% worst case. The contour unit is microgram per liter and an order of magnitude difference is obvious. Spacings of the horizontal and vertical axes are 500 feet and 50 feet, respectively*

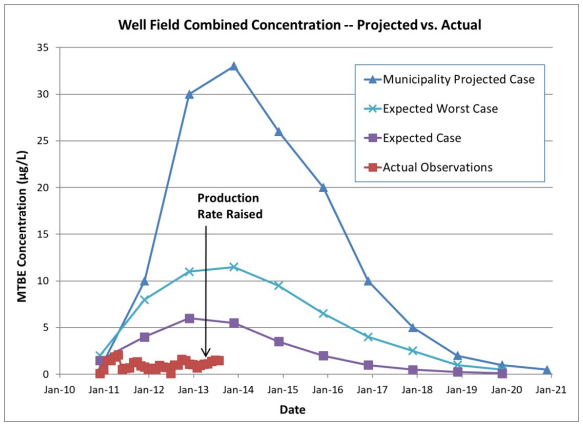

Figure 7: *Model Projected versus Actual Concentrations.*

volving layered and volumetric interpretations. Visualization analysis has supported the process of visualizing data, assessing initial findings, re-interpreting data, adjusting the analysis, and generating outcomes that meet practical needs. To conduct a good visualization analysis for a specific task, one often has to look beyond the task and modeling process to obtain a higher level understanding of the problem and its real-world implications.

The above analyses indicate the needs for visualization experts and environmental practitioners to work together to develop procedures and build tools for validating against reality. In addition, We have observed that environmental consultants tend to use coloring schemes that are simplistic (e.g., categorical colors varying in luminance and the use of rainbow colors) and are unfamiliar with the rules for accurate perception in visualization science. Learning scientific coloring theories and applying them to their visualization analyses can help increase their exploration capacity.

5. Acknowledgement

The work is supported in part by NSF IIS-1302755, NSF CNS-1531491, and NIST-70NANB13H181. Feng is supported by the National Natural Science Foundation of China No.61472163. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and National Institute of Standards and Technology (NIST). Certain commercial

products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

References

[Ber83] BERTIN J.: Semiology of graphics: diagrams, networks, maps. 2

[BO04] BABUSKA I., ODEN J. T.: Verification and validation in computational engineering and science: basic concepts. *Computer methods in applied mechanics and engineering 193*, 36-38 (2004), 4057–4066. 1

[CDM10] CAI S., DU Q., MOORHEAD R. J.: Feature-driven multilayer visualization for remotely sensed hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing 48*, 9 (2010), 3471–3481. 2

[CLX*17] CHANG C.-I., LEE L.-C., XUE B., SONG M., CHEN J.: Channel capacity approach to hyperspectral band subset selection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10*, 10 (2017), 4630–4644. 1, 2

[CRF05] CROSS B., ROGOFF E., FRICKE L.: Data visualization tools for litigation-practical uses and ethical considerations. In *Proceedings of the NGWA Ground Water and Environmental Law Conference* (2005), pp. 1–14. 1

[DJ92] DEUTSCH C., JOURNEL A.: *GSLIB: Geostatistical software library and user's guide*. Oxford University Press, New York, 1992. 2

[IS89] ISAAKS E., SRIVASTAVA R.: *An introduction to applied geostatistics*. Oxford University Press, New York, 1989. 2

[JS14] JÄNICKE S., SCHEUERMANN G.: Utilizing GeoTemCo for visualizing environmental data. *Workshop on Visualization in Environmental Sciences* (2014). 1, 2

[KW05] KAPLER T., WRIGHT W.: Geotime information visualization. *Information visualization 4*, 2 (2005), 136–146. 2

[LC14] LING M., CHEN J.: Environmental visualization: applications to site characterization, remedial programs, and litigation support. *Environmental earth sciences 72*, 10 (2014), 3839–3846. doi:10.1007/s12665-014-3220-y. 1

[LJL07] LING M., JOHNSON J., LIN X.: Contamination distribution and mass estimate via kriging: Pitfalls and lessons. *Proceedings of the Petroleum Hydrocarbons and Organic Chemicals in Ground Water: Prevention, Assessment, and Remediation Conference* (2007). 4

[SDS*17] SINGH S., DASH P., SILWAL S., FENG G., ADELI A., MOORHEAD R. J.: Influence of land use and land cover on the spatial variability of dissolved organic matter in multiple aquatic environments. *Environmental Science and Pollution Research 24*, 16 (2017), 14124–14141. 1, 2

[SND05] SARAIYA P., NORTH C., DUCA K.: An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (2005), 443–456. 4

[Tuk77] TUKEY J. W.: *Exploratory data analysis*, vol. 2. Reading, Mass., 1977. 1, 2