What Lies Beneath: Exercising Caution when Repurposing Data in Environmental Analytics

By Melanie Edwards, PSTAT.

We live in a world where data can be easily accessed through public organizations, publications, subscription-based services, and artificial intelligence tools. But we need to ask ourselves, where did the data originate? Is it appropriate and relevant for the current investigation? What is known about the underlying details of these data?

As the analysis of environmental data increases in volume, variety, speed, and complexity, it is critically important for stakeholders to understand the fundamental, underlying information to ensure compatibility.

Environmental scientists can repurpose and combine existing data to prove, or disprove, a conceptual site model without necessarily considering the origins of the data. Beyond the important aspect of verifiable sources to assure data accuracy and integrity is the skill to know when a data source may introduce bias into the analyses used to characterize a site.

Analysts need to trace the data to the source and ideally understand the sampling design, protocol and sample preparation, and laboratory methods, as well as understand how advanced analytics are affected by these characteristics of the data. Without a full understanding of the data, results of the data analyses may not reflect the true uncertainty and conclusions about the site could be dubious, or simply incorrect, and this can be brought to light by opposing experts.

Case Example: Background Screening Level

Modern environmental assessments involve large quantities of data collected over a long period of time, often including multiple studies designed to investigate a variety of different concerns. Combining incomparable data will often have detrimental results, for example, when calculating a background screening level for metals at a Superfund site.

Investigations at the site have targeted geological questions as well as ecological and human health concerns, yet each study included background soil samples for comparison. Logically, a statistically robust background threshold value (BTV) would be based on the combined background samples obtained from all studies. Careful investigation, though, would uncover that the analytical methods for geologic investigations are not equivalent to those used in risk assessment analyses. Combining the studies will significantly bias the BTV because it is like comparing apples and oranges. Risk assessment analytical methods measure the surficial metals on the soil particles that animals are exposed to whereas geologic investigations measure the metals present in the complete soil particles (i.e., dissolve the soil particles entirely).

Case Example: Contaminant Source Identification

Another example relates to the source identification of polycyclic aromatic hydrocarbons (PAHs). PAHs occur in crude oil, coal, and gasoline and are generated during incomplete combustion of organic materials such as wood, garbage, and tobacco. There are hundreds of air quality publications that compare the PAH composition, or “fingerprint,” at a site to the generic fingerprint of specific sources. Many researchers use the Chemical Mass Balance model developed by the U.S. Environmental Protection Agency to help identify the contributions of various pollution sources present in the atmosphere at a site of concern. These analyses frequently rely on detailed chemistry results that characterize the composition of generic sources, such as burning coal or traffic exhaust. However, a look into the origin of these generic fingerprints shows that they were derived for an entirely different purpose by combining incomplete sample results with very high uncertainty (O’Reilly et al. 2023). Therefore, investigations relying on similarity to these sources may be inconclusive when the uncertainty is acknowledged.

In any environmental assessment, it is critical to understand the details of the input data in order to select appropriate data sources. Repurposing and combining existing data to prove, or disprove, a conceptual site model without necessarily considering the fundamentals, can be problematic and result in inaccurate conclusions. Understanding the reliability and comparability of the underlying data used in any model[1] assures confidence in the decisions derived from the analysis results.

What Can Be Done?

A complex environmental investigation requires sophisticated analyses of relevant data that characterize key aspects of the environmental conditions at and near the site. Scientists must go beyond the number crunching of available data sets and instead design and implement analyses that yield insightful solutions. Integral’s planning process starts with information management, progresses through targeted analytics, and culminates in effective communication through compelling visuals. Modern information management skills identify all the environmental aspects relevant to the site, obtain the necessary data, and investigate the origins to identify the most comparable and relevant data sets for analysis. Understanding modern analytics and knowing when these methods will provide insight, as well as when a simpler approach is sufficient, will result in efficiencies and cost savings. Communication of results in plain English with visually compelling graphics to convey the final story for all audiences to understand is critical for all environmental scientists.

For site characterizations, real-time monitoring, impact assessments, remedial designs, and allocation, our approach from start to finish provides insights with confidence. We integrate data with strategic thinking, scientific expertise, and targeted analysis to bring actionable

[1] Including the models that underly modern artificial intelligence assistants.

results that support informed decisions. Our work is collaborative, data-driven, and supports high-value decisions.

Reference:

O'Reilly, K., D. Athanasiou, and M. Edwards. 2023. Evaluation of generic PAH profiles commonly used in receptor models: Implications for source control policy. *Environmental Forensics*. DOI: 10.1080/15275922.2023.2172094