The Journal of Open Source Software

ESAT: Environmental Source Apportionment Toolkit Python package

Deron Smith [1], **Michael Cyterski** [1], **John M Johnston** [1], **Kurt Wolfe**[1], **and Rajbir Parmar** [1]

1 United States Environmental Protection Agency, Office of Research and Development, Center for Environmental Measurement and Modeling

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Source apportionment is an important tool in environmental science where sample or sensor data are often the product of many, often unknown, contributing sources. Source apportionment is used to understand the relative contributions of air sources like vehicle emissions, industrial activities, biomass burning, dust to particulate matter pollution and to identify relative contributions of point sources (e.g., wastewater treatment discharges) and non-point sources (e.g., agricultural runoff) in water bodies such as lakes, rivers, and estuaries. Using non-negative matrix factorization (NMF), source apportionment models estimate potential source profiles and contributions providing a cost-efficient method for further strategic data collection or modeling.

Environmental Source Apportionment Toolkit (ESAT) is an open-source Python package that provides a flexible and transparent workflow for source apportionment modeling using NMF algorithms, developed to replace the EPA's Positive Matrix Factorization version 5 (PMF5) application(EPA, 2014) (Pentti Paatero, 1999). ESAT recreates the source apportionment workflow of PMF5 including pre-post processing analytical tools, batch modeling, model uncertainty estimations and customized constraints. Additionally, ESAT offers a simulator for generating datasets from synthetic profiles and contributions, allowing for direct model output evaluation. The synthetic profiles can either be randomly generated, use a pre-defined set of profiles, or be a combination of the two. The random synthetic contributions can follow specified curves and value ranges. Running ESAT using the synthetic datasets we are able to see how accurately ESAT is able to find a solution that recreates the original synthetic profiles and contributions.

Statement of Need

ESAT has been developed as a replacement to PMF5, and has been designed for increased flexibility, documentation and transparency. The EPA's PMF5, released in 2014, provides a widely-used source apportionment modeling and analysis workflow but is no longer supported and relies on the proprietary Multilinear Engine v2 (ME2) that lacks documentation.

The Python API and CLI of ESAT provides an object-oriented interface that can completely recreate the PMF5 workflow. The matrix factorization algorithms in ESAT have been written in Rust for runtime optimization of the core math functionality. ESAT provides a highly flexible API and CLI that can create source apportionment workflows like those found in PMF5, but can also be used to create new workflows that allow for novel research applications. ESAT was developed for environmental research, though it's not limited to that domain, as matrix factorization is used in many different fields; ESAT places no restriction on the types of input datasets.

Algorithms

Source apportionment algorithms use a loss function to quantify the difference between the input data matrix (V) and the product of a factor contribution matrix (W) and a factor profile matrix (H), weighted by an uncertainty matrix (U) (Pentti Paatero & Tapper, 1994). The goal is to find factor matrices that best reproduce the measured matrix, while constraining all, or most of, the factor elements to be non-negative. The solution, product of the output W and H matrices, can be used to calculate the residuals and overall loss of the model. ESAT has two NMF algorithms for updating the profile and contribution matrices of the solution: least-squares NMF (LS-NMF) (Wang et al., 2006) and weighted-semi NMF (WS-NMF) (Ding et al., 2008) (Melo & Wainer, 2012).

The loss function used in ESAT, and PMF5, is a variation of squared-error loss, where data uncertainty is taken into consideration (both in the loss function and in the matrix update equations):

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{V_{ij} - \sum_{k=1}^{K} W_{ik}H_{kj}}{U_{ij}}\right]^2$$

here V is the input data matrix of features (columns=M) by samples (rows=N), U is the uncertainty matrix of the input data matrix, W is the factor contribution matrix of samples by factors=k, H is the factor profile of factors by features.

The ESAT versions of NMF algorithms convert the uncertainty U into weights defined as $Uw = \frac{1}{U^2}$. The update equations for LS-NMF then become:

$$H_{t+1} = H_t \circ \frac{W_t(V \circ Uw)}{W_t((W_tH_t) \circ Uw)}$$

$$W_{t+1} = W_t \circ \frac{(V \circ Uw)H_{t+1}}{((W_tH_{t+1}) \circ Uw)H_{t+1}}$$

while the update equations for WS-NMF:

$$W_{t+1,i} = (H^T Uw_i^d H)^{-1}(H^T Uw_i^d V_i)$$

$$H_{t+1,i} = H_{t,i}\sqrt{\frac{((V^TUw)W_{t+1})_i^+ + [H_t(W_{t+1}^T UwW)^-]_i}{((V^TUw)W_{t+1})_i^- + [H_t(W_{t+1}^T UwW)^+]_i}}$$

where $W^- = \frac{(|W|-W)}{2.0}$ and $W^+ = \frac{(|W|+W)}{2.0}$.

Error Estimation

An important part of the source apportionment workflow is quantifying potential model error. ESAT offers the error estimation methods that were developed and made available in PMF5 (Brown et al., 2015) (P. Paatero et al., 2014).

The displacement method (DISP) determines the amount that a source profile feature, a single value in the H matrix, must increase and decrease to cause targeted changes to the solution loss value. One or more features can be selected in the DISP uncertainty analysis. The bootstrap method (BS) uses block bootstrap resampling with replacement to create datasets of the original dimensions of the input but where the order of the samples has been modified, in blocks of a specified size. The BS method then goes on to calculate a new model from the bootstrap dataset, and original initialization, to evaluate how the profiles and concentrations changes as

a result of the reordering of samples. The bootstrap-displacement method (BS-DISP) is the combination of the two techniques, where for each bootstrap model DISP is run for one or more features.

These error estimation methods address different uncertainty aspects: DISP targets rotational uncertainty, BS addresses random errors and sample variability, and BS-DISP offers a combined uncertainty estimate, collectively providing a comprehensive understanding of the uncertainty in a source apportionment solution.

Acknowledgements

We thank Tom Purucker and Jeffery Minucci for manuscript and code review and edits. This paper has been reviewed in accordance with EPA policy and approved for publication. ESAT development has been funded by U.S. EPA. Mention of any trade names, products, or services does not convey, and should not be interpreted as conveying, official EPA approval, endorsement, or recommendation. The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of the US EPA.

References

Brown, S. G., Eberly, S., Paatero, P., & Norris, G. A. (2015). Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. *Science of the Total Environment*, *518*, 626–635. https://doi.org/10.1016/j.scitotenv.2015.01.022

Ding, C. H., Li, T., & Jordan, M. I. (2008). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55. https://doi.org/10.1109/TPAMI.2008.277

EPA, U. S. (2014). *Positive matrix factorization model for environmental data analyses*. https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses

Melo, E. V. de, & Wainer, J. (2012). *Semi-NMF and weighted semi-NMF algorithms comparison*.

Paatero, Pentti. (1999). The multilinear engine—a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics*, *8*(4), 854–888. https://doi.org/10.1080/10618600.1999.10474853

Paatero, P., Eberly, S., Brown, S. G., & Norris, G. A. (2014). Methods for estimating uncertainty in factor analytic solutions. *Atmospheric Measurement Techniques*, *7*(3), 781–797. https://doi.org/10.5194/amt-7-781-2014

Paatero, Pentti, & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111–126. https://doi.org/10.1002/env.3170050203

Wang, G., Kossenkov, A. V., & Ochs, M. F. (2006). LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, *7*, 1–10. https://doi.org/10.1186/1471-2105-7-175