Advancing and Accelerating Release Estimations for Chemical Processes: Opportunities for Unit Operations, Data Mining, and Machine Learning

Raymond L. Smith, David E. Meyer, Gerardo J. Ruiz-Mercado, Michael A. Gonzalez, William M. Barrett, John P. Abraham

U.S. Environmental Protection Agency, Office of Research and Development

The views expressed in this poster are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Summary and Context

- Computer-aided process designs, risk assessments, and life cycle assessments can incorporate environmental impacts
- Data needs are large for these assessments, as there are many chemicals and conditions of use where releases can occur:

  Manufacturing, processing, use, and end-of-life
- Approaches such as simulation, data mining, and machine learning offer methods for rapidly estimating releases:

  Simulation offers a unit-operation or bottom-up perspective,

  Data mining uses established EPA databases,

  Machine learning can use classification and regression trees to predict emissions.
- Application of approaches should be fit for purpose, i.e., no single method is appropriate to every set of circumstances
- Future work will explore use of these methods in exposure and risk assessments

Approach

Approaches to design processes and estimate releases include:

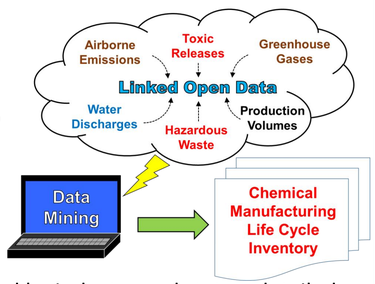
1) Conceptual Chemical Process Design[1]
2) Top-Down Data Mining[2]
3) Bottom-Up Simulation[3]
4) Machine Learning to Predict Releases[4]
5) Evaluation of Release Inventories[5]

Top-Down Data Mining

Data mining: the study of collecting, harmonizing, processing, and analyzing data to gain useful insights.

EPA has ample data for facilities:

CDR, NEI, TRI, DMR, GHGRP, RCRAInfo

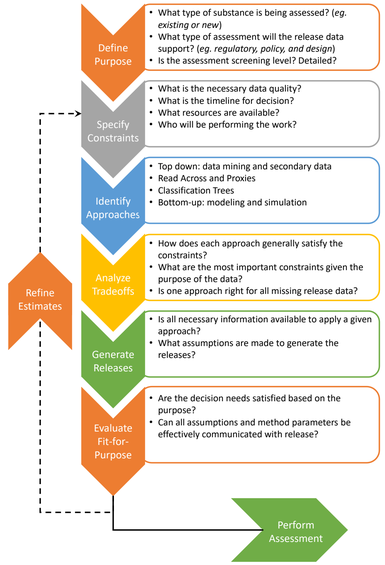


Advantages: primary data reported by industry and states; detailed emission profiles; able to be automated

Challenges: allocating in multi-chemical production facilities; data gaps for inventory inputs; limited to TSCA CDR chemicals (for now)

Release Estimation Framework

Purpose-Driven Framework for Estimating Releases and Example Results



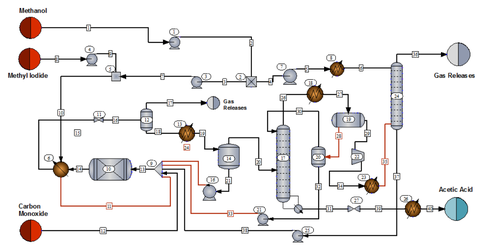
Example Criteria for Selecting an Emission Estimation Approach

	LOW	MEDIUM	HIGH
Data Quality Concern	**DQ Score ≤ 1.7**	**1.7 < DQ Score < 2.3**	**DQ Score ≥ 2.3**
Required Time	**< 5 days**	**5 - 20 days**	**> 20 days**
Required Resources	**< $2,000**	**$2,000 - $10,000**	**> $10,000**
Required Training	***novice*** **scientific/engineering background required (bachelor's degree with no experience)**	***moderate*** **scientific/engineering background required (bachelor's degree with 1-5 years experience)**	***advanced*** **scientific/engineering background required (MS/PhD; bachelor's degree with >5 years experience)**
Required Knowledge	**no activity-specific or data source knowledge required**	**either activity-specific or data source knowledge required**	**both activity-specific and data source knowledge required**

Case Study: Cumene Manufacturing

Approach	Emission Factor (kg/kg)
Top-Down Data Mining	2.0×10^{-5}
Bottom-Up Simulation	1.3×10^{-4}
Machine Learning – Regression Tree	9.3×10^{-5}
Machine Learning – Random Forest	2.0×10^{-4}

Bottom-Up Simulation



Simulation: couples engineering material and energy balances with EPA emission modeling

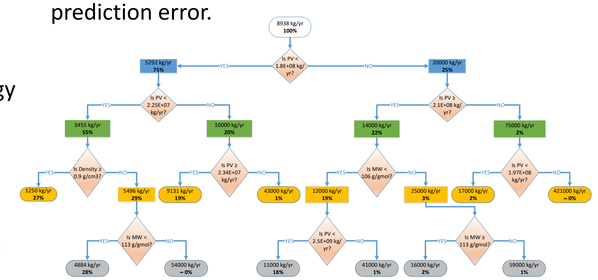
Advantages: improved compared to existing databases; includes storage and fugitive emissions; look-up tables for screening-level inventories

Challenges: knowledge of engineering design; need for chemical synthesis details

Machine Learning Predictions

Regression Trees use predictor variable partitioning and the training and testing of data for emissions. Predictions depend on production volume, molecular weight, vapor pressure, water solubility, and density.

Random Forests create an ensemble of trees with randomly selected predictor sets to lower the average prediction error.



Conclusions and Future Work

Release Estimation Results

Regression approaches offered estimations of emissions for cumene manufacturing that are the same order of magnitude as Top-Down Data Mining and Bottom-Up Simulation methods. For the training and testing data set, regression offers **quick results with relatively low resource needs**, once the required prediction model was developed.

Future work will develop release estimations for exposure and risk assessments. EPA's TSCA Chemical Substance Inventory lists over **32,000 active chemicals**, and CAS registry numbers have been developed for over 150 million organic and inorganic substances. These very large chemical listings point to the need for quick and accurate release estimations.

References and Abbreviations

[1] R.L. Smith (2016). "Conceptual Chemical Process Design for Sustainability," in *Sustainability in the Design, Synthesis and Analysis of Chemical Engineering Processes*, G. Ruiz-Mercado and H. Cabezas, eds., Elsevier: Cambridge, MA

[2] S.A. Cashman et al. (2016). "Mining Available Data from the United States Environmental Protection Agency to Support Rapid Life Cycle Inventory Modeling of Chemical Manufacturing," *Environ. Sci. Technol.*, DOI: 10.1021/acs.est.6b02160

[3] R.L. Smith et al. (2017). "Coupling Computer-Aided Process Simulation and Estimations of Emissions and Land Use for Rapid Life Cycle Inventory Modeling," *ACS Sustainable Chemistry & Engineering,* DOI: 10.1021/acssuschemeng.6b02724

[4] D.E. Meyer et al. (2019). "Purpose-Driven Reconciliation of Approaches to Estimate Chemical Releases," *ACS Sustainable Chemistry & Engineering,* DOI: 10.1021/acssuschemeng.8b04923

[5] R.L. Smith et al. (2019). "Applying Environmental Release Inventories and Indicators to the Evaluation of Chemical Manufacturing Processes in Early Stage Development," *ACS Sustainable Chemistry & Engineering*, DOI: 10.1021/acssuschemeng.9b01961

CDR – Chemical Data Reporting
DMR – Discharge Monitoring Report
GHGRP – GHG Reporting Program
NEI – National Emissions Inventory
RCRAInfo – Resource Conservation and Recovery Act Information
TRI – Toxics Release Inventory
TSCA – Toxic Substances Control Act