www.epa.gov

Informatics-Based Approaches for Collecting and Curating Consumer Product Data for Exposure Assessments

P116

J. T. Wall[1], S. Burns[1], K. Phillips[1], K. Dionisio[1], V. Hull[2], and K. Isaacs[1]

[1]Center for Computational Toxicology and Exposure, US EPA, Durham, NC.; [2]Oak Ridge Associated Universities, Oak Ridge, TN

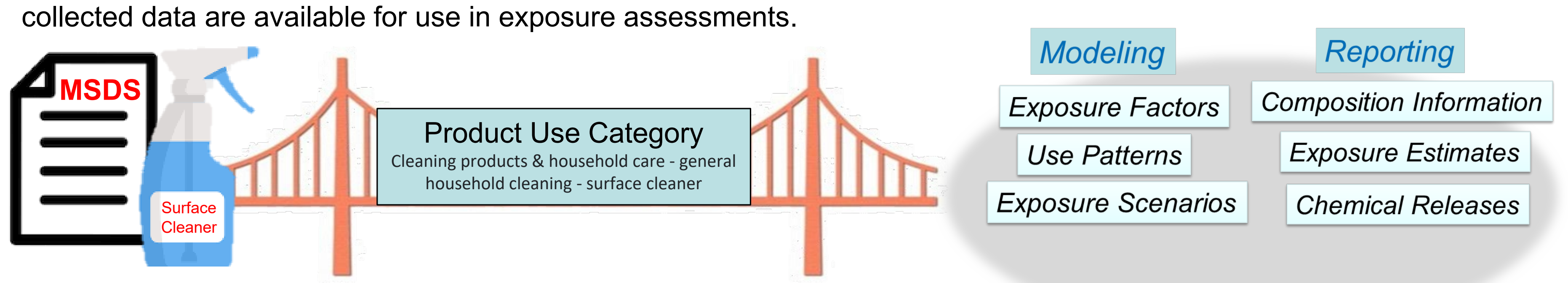
Jonathan Taylor Wall I wall.jonathan@epa.gov I 919-541-0720

Abstract

Quantitative data on product chemical composition is necessary for characterizing consumer exposure to chemicals. EPA's Office of Research and Development (ORD) has built rapid models, including the High-Throughput Stochastic Human Exposure and Dose Simulation model (SHEDS-HT), that use this data to estimate exposures for over 300 hierarchical harmonized product use categories (PUCs). However, this data is often lacking or is in various formats, making it difficult to use in models like SHEDS-HT. To fill this data need, ORD has developed automated approaches for collecting and curating data on thousands of individual products and chemicals from public documents (safety data sheets, manufacturer ingredient disclosures, ingredient lists). However, curation of documents for individual products to PUCs is historically a bottleneck, requiring manual assessment of product names. Here, we use natural language processing machine-learning approaches to hasten this curation step. The model training dataset was comprised of all products within ORD's Chemical and Products Database (CPDat) that had a PUC manually assigned; models were built for PUCs with at least 30 products (63,593 products; 161 PUCs). For modeling, each product-brand name was combined, cleaned, lemmatized to word roots, and converted to a quantitative vector using standard libraries. A Support Vector Machine (SVM) classifier was created for each level of the 3-tier PUC classification, each informed by the higher tier prediction. The probabilistic SVM models were used to generate multiple predictions per tier; the median predicted PUC was selected. Five-fold cross validation was performed (stratified by PUC to ensure proportional representation) resulting in an average 94% classification accuracy. The final models were applied to 460,518 additional products from documents within CPDat, increasing its scope to 524,11 products and 7,134 chemicals associated with PUCs. The expanded data can be used to update consumer exposure predictions using SHEDS-HT, which provided refined aggregate and PUC-specific consumer exposure distributions, particularly for home care and home maintenance PUCs (which were previously data poor in terms of products in CPDat). In summary, implementation of informatics approaches for managing and curating public documents are rapidly expanding the quantity and quality of data available for assessing consumer exposure to chemicals in consumer products.

Introduction

- Quantitative data on consumer product chemical composition is necessary for characterizing consumer exposure to chemicals.
- The EPA has developed a curation platform, Factotum, that is an interface to help curate and manage the data for thousands of documents containing product chemical data, including Material Safety Data Sheets, product ingredient lists, and manufacturer disclosure documents. The data curated in Factotum comprise EPA Office of Research and Development's Chemical and Products Database (CPDat)
- Products are curated to Product Use Categories (PUCs) that provide a direct link to key data and algorithms necessary for estimating exposures. PUCs are also used for reporting and summarizing data, e.g., in the CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard).
- Curation of product documents to PUCs is historically a bottleneck, requiring manual assessment of product names and other metadata (e.g., manufacturer, product photos, product descriptions).
- Here, we use natural language processing machine-learning approaches to hasten this curation step, increasing the rate at which collected data are available for use in exposure assessments.



Product Use Category

Cleaning products & household care - general household cleaning - surface cleaner

Modeling

Exposure Factors

Use Patterns

Exposure Scenarios

Reporting

Composition Information

Exposure Estimates

Chemical Releases

Methods

Composition Data Document Collection

- The Factotum application is composed of a variety of curation and document management tools used to extract and clean chemical data from publicly-available documents (**see abstract 2651, poster P117**).
- Thousands of publicly available data documents of multiple source types are downloaded manually and through automated scripting processes (e.g., web scraping). The available product composition documents are summarized in Table 1.
- Document data is manually and automatically curated for relevant data, including extraction of chemical information, (e.g., chemical weight fractions and functional use) and product metadata.
- Historically, products are curated to PUCs via manual processes. Here, we present a new informatics approach to categorizing products.

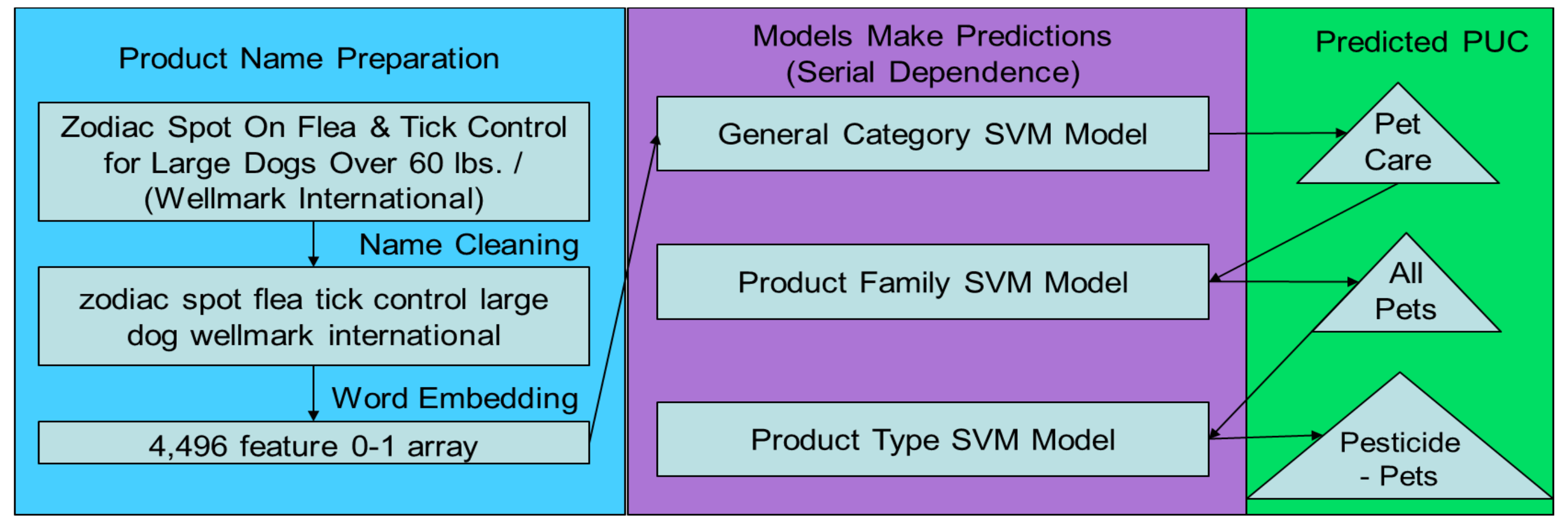
Document Source Type	Document Count	Description
MSDS	240,104	Material Safety Data Sheet. Standard format used by product manufacturers.
Ingredient List	97,430	A list of ingredients in a product
Ingredient Disclosure	5,335	Discloses the functional use of ingredients in products in accordance with California's Cleaning Product Right to Know Act of 2017
SDS	13,057	Safety Data Sheet. Standard format used by product manufacturers.

Table 1: Product Composition Documents Available in Factotum

Methods (Continued)

Automated PUC Curation Model

- A Natural Language Processing model for PUC classification was trained on all products within CPDat that had a PUC manually assigned; the model considered PUCs with at least 30 products (99,705 products; 182 PUCs).
- Each product-brand name was combined, cleaned, lemmatized to word roots, and converted to a quantitative vector using standard Python libraries (Figure 1).
- Support Vector Machine (SVM) classifier models were created for each level of the 3-tier PUC classification, each informed by the higher tier prediction. The SVM generated multiple predictions per tier; the median predicted PUC was selected.
- A five-fold cross validation was performed (stratified by PUC to ensure proportional representation).
- The final model was used to classify to 491,443 additional product composition documents to PUCs.



Product Name Preparation

Zodiac Spot On Flea & Tick Control for Large Dogs Over 60 lbs. / (Wellmark International)

Name Cleaning

zodiac spot flea tick control large dog wellmark international

Word Embedding

4,496 feature 0-1 array

Models Make Predictions (Serial Dependence)

General Category SVM Model

Product Family SVM Model

Product Type SVM Model

Predicted PUC

Pet Care

All Pets

Pesticide - Pets

Figure 1: Model Prediction Workflow Example

Results

Model Performance and QA

- The five-fold cross validation showed no significant difference in model accuracy between 80-20 subset splits of the training dataset.
- After training on 90% of the full training dataset, the remaining 10% external test dataset produced predicted model accuracies of 99.04% (general category), 94.49% (product family), and 90.71% (product type).
- The assignment of PUCs to the 491,443 products using the resulting SVM model took ~48 hours.
- The resulting increases in chemicals represented within PUCs were assessed, and individual product assignments were quality checked
- Following the SVM models' predictions, the unique chemical space within 148 PUCs was increased.
  - The PUCs with the largest increase in chemical count were *caulk and sealant* (2,018), *face cream and moisturizer* (1,844), *bleach* (1,741), and *paint* (1,730).
    - The caulk and sealant and paint category assignments were found to be very accurate, with the increase in chemical space resulting many new products being categorized.
    - The large increase in *face cream and moisturizer* was an artifact of many vitamin supplements being misclassified.
    - The increase in the *bleach* category resulted from industrial bleach products being misclassified.
- New models can be built to further stratify these categories and separate industrial uses from consumer use products using additional model criteria, such as manufacturer. In addition, new categories can be created, and correctly categorized products can be used to increase the model training set. The models can then be improved in an iterative manner (Figure 4).

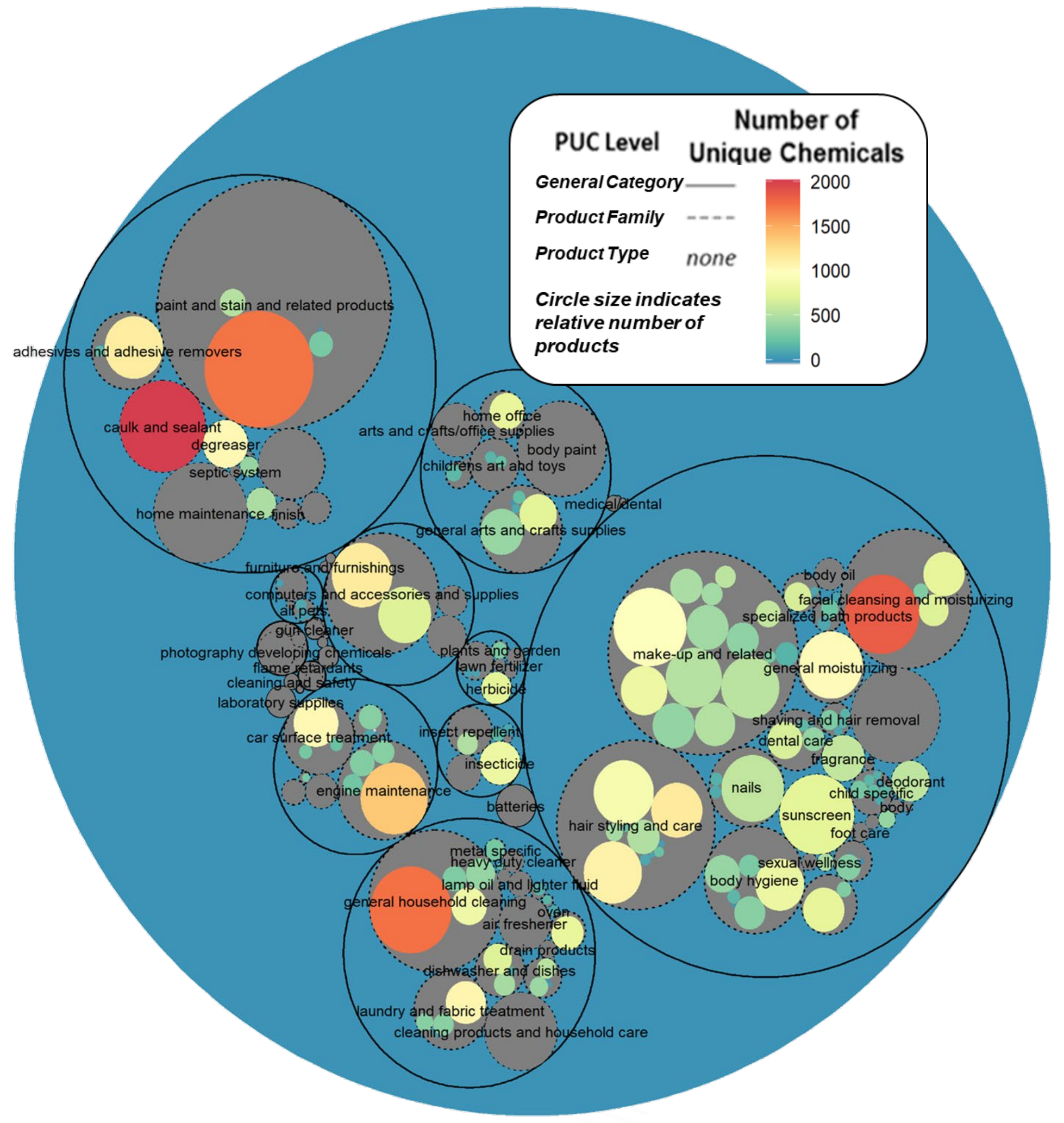


Figure 3: Increase in Number of Chemicals in PUC

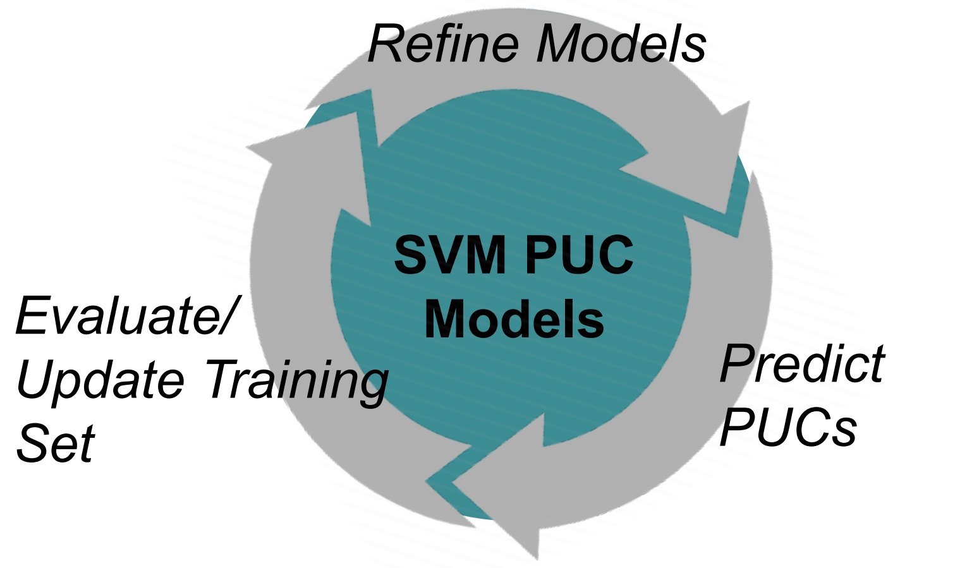


Figure 4: Iterative Model Development

Results (Continued)

Impact on Scope of Curated Data

- The model greatly increased the magnitude and scope of products in Factotum assigned to PUC, thus allowing them to be used in assessments.
- The relative change in magnitude of products assigned to a PUC general category increased by as few as 2.79% (Personal Care) and as large as 2,029.47% (Home Maintenance) (Figure 5).
- The product categorizations performed by the models increased the number of chemicals identified in individual PUCs and provided a notable increase in the number of datapoints (products) for individual chemicals. This allowed for the development of refined exposure estimates for chemicals in products (e.g., as illustrated for surface cleaners in Figure 6.)

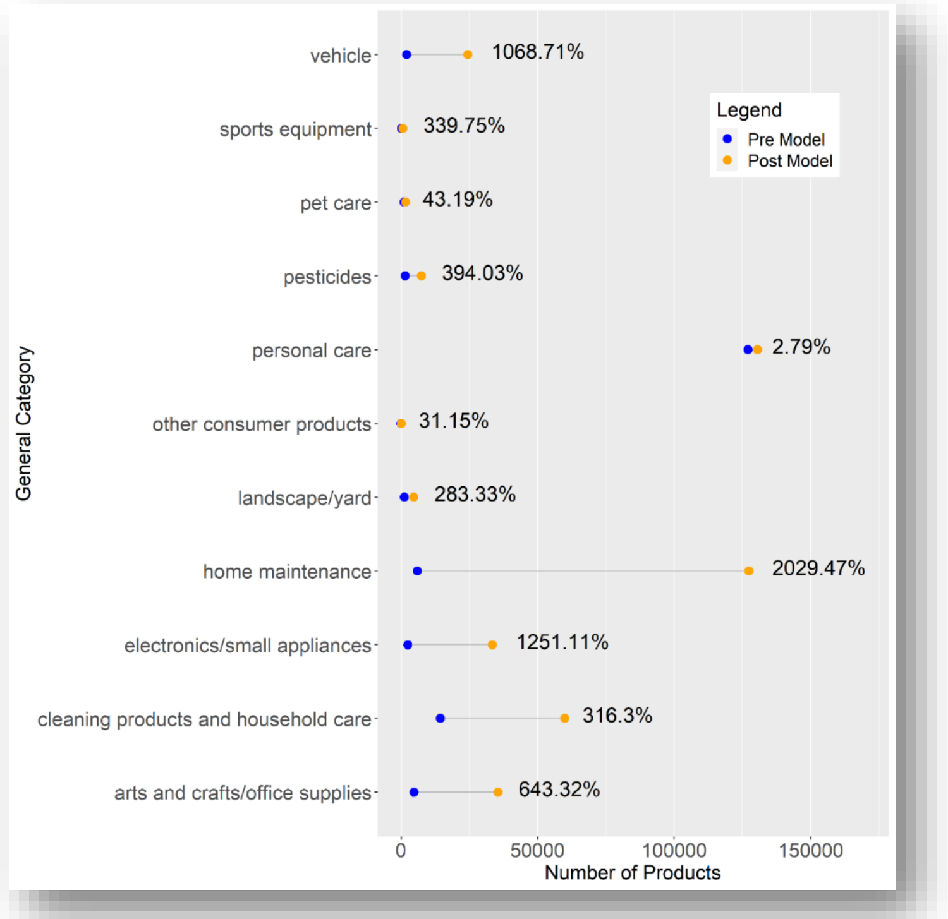
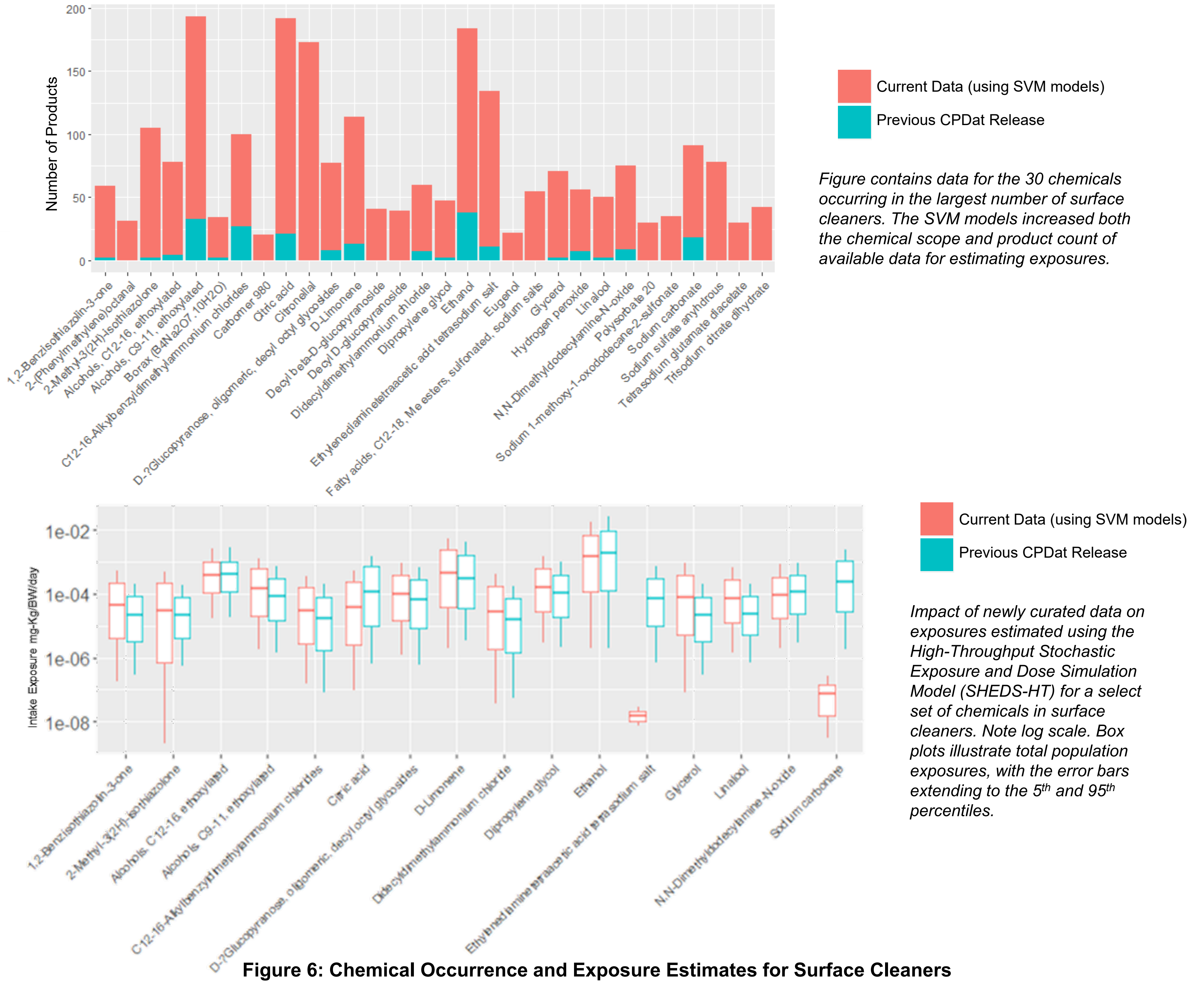


Figure 5: Increase in Number of Curated Products with the Application of the PUC Model



🟥 Current Data (using SVM models)

🟦 Previous CPDat Release

Figure contains data for the 30 chemicals occurring in the largest number of surface cleaners. The SVM models increased both the chemical scope and product count of available data for estimating exposures.

🟥 Current Data (using SVM models)

🟦 Previous CPDat Release

Impact of newly curated data on exposures estimated using the High-Throughput Stochastic Exposure and Dose Simulation Model (SHEDS-HT) for a select set of chemicals in surface cleaners. Note log scale. Box plots illustrate total population exposures, with the error bars extending to the 5th and 95th percentiles.

Figure 6: Chemical Occurrence and Exposure Estimates for Surface Cleaners

Conclusion

Implementation of informatics approaches for managing and curating public documents are rapidly expanding the quantity and quality of data available for assessing consumer exposure to chemicals in consumer products.

This poster does not necessarily reflect EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

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