Exploratory Data Analysis Report for Group 5

Daihao Wu, Benjamin He, Seehe Kim, Fred Peng, Tom Xu

11/16/2023

Introduction

Semi-volatile organic compounds (SVOCs) are liquids or solids at lower temperatures, that are commonly found in many common products such as pesticides, oil-based products, fire retardants and building materials. SVOCs can volatilize from surfaces and materials and become airborne, which might be exposed to humans through inhalation, skin contact and food contamination. Recent studies have shown that the presence of materials containing SVOCs in homes lead to higher risk of allergies, respiratory symptoms, diagnosed asthma and birth defects.

Building materials, such as paints, adhesives, and wood products may contain SVOCs that can slowly release into the indoor air over time. This release can continue for an extended period after construction or installation. The question of the concentration of SVOCs in newly constructed homes remained unanswered and this poses a health risk to the occupants, so our team's goal is to determine the levels of SVOCs in newly built homes and measure the chemical levels over a period of time after occupants move in. We hope that our research results can contribute to safety measurements and home development regulations.

Therefore, this paper serves as the foundational analysis for investigating the concentrations of Semi-volatile Organic Compounds (SVOCs) in newly constructed homes. As our team delves into the exploratory data analysis of the provided dataset, we aim to unravel the levels of SVOCs and track their variations over time following the occupancy of these homes.

Research Question

How does the concentration of Semi-volatile organic compounds (SVOCs) change before and after occupants move into newly constructed homes, and what are some potential factors contributing to that change?

Data

Our data set mainly consists of 3 Excel spreadsheets as follows: "NHS_2023_Data_October_31_University of Toronto_First phase.xlsx", "20231027_Technician Survey Data.xlsx" and "20231027_0-12 Month Questionnaire Data.xlsx".

SVOCs Concentration Data:

"NHS_2023_Data_October_31_University of Toronto_First phase.xlsx"

The concentration data spreadsheet contains the concentration levels of up to 69 different SVOCs. The data was collected using different methods, including using active samplers, passive samplers, wiping and dusting. Thus, these corresponding SVOCs collected via different methods are separated by sheets. For the passive sampler, wipe sampler and dust sampler, the data collection interval was every three months after move-in up to nine months. For the active sampler, data was collected right before move-in and right after move-in in the newly constructed homes. However, due to COVID-19, many participating houses were unable to complete sampling for the 9-month interval after move-in. As a result, only 3 houses have complete data for the SVOCs in the Phthalates (PAEs) family through passive and wiping data collection methods.

Technical Data:

“20231027_Technician Survey Data.xlsx”

The technician survey data spreadsheet contains the technical details of those newly constructed homes that participated in the research recorded by the survey technicians. It contains several significant attributes such as the number of bedrooms, the number of bathrooms, the size of windows, the size of rooms, type of heating, construction material etc. There is another document (EN-Technician Survey - FINAL_coding_v2.docx) that explains the coding for the response and variables in the technical survey data. There are in total 50 homes recorded in the technical data with some missing data points at different technical variables. This spreadsheet provides data to support analysis of how the building/home itself can affect the SVOC levels.

Questionnaire Data:

“20231027_0-12 Month Questionnaire Data.xlsx”

The questionnaire data spreadsheet contains the survey results that were completed by the survey participants. The data mainly consists of personal details in the surveyed houses, daily habits, personal behaviours, furniture details etc. There are two other documents (EN-0 month Visit Questionnaire_coding.docx/ EN-3-12 month Visit Questionnaire_coding.docx) that explain the coding for the response and variables in the questionnaire survey data. There are in total 44 homes recorded in the data spreadsheet with some missing data points at different survey variables. This spreadsheet provides data to support analysis on how human activity can affect the SVOC levels.

Data Preprocessing:

To prepare for the data exploratory process, we imported the passive/active air datasheet from the concentration dataset as well as the questionnaire dataset. We converted the “Period” column, which represents the time interval the concentration is recorded, to reflect type numeric instead of string. We then merged the selected variables/columns from the technical survey data with the concentration data into a full dataset named “full_data”.

The selected variables from technical survey data are as follows:

“id”: the unique house ID for each participating newly constructed home; equivalent to “House ID” in the concentration dataset

“base_ac_yn”: indication of if the home has air conditioning (0/1)

“heating_type”: the type of heating system used in the home

“cellulose_ins_yn”: indication of if the home uses Cellulose as insulation material (0/1)

“styrofoam_ins_yn”: indication of if the home uses Styrofoam (polystyrene) as insulation material (0/1)

“fiberglass_ins_yn”: indication of if the home uses Fiberglass as insulation material (0/1)

“sprayfoam_ins_yn”: indication of if the home uses Spray polyurethane foam as insulation material (0/1)

“base_ij_osb_yn”: indication of if the home has I-joists or oriented strand board (OSB) panels exposed in the basement (0/1)

“Ukc1_mat”: the material that the upper kitchen cabinetry is made of

“lkc_mat”: the material that the lower kitchen cabinetry is made of

”hrv_on_yn”: indication of if the home has HRV or ERV turned on

Similarly, we then further merge the selected variables/columns from the Questionnaire data with the “full_data”, and the selected variables from the survey data are as follows:

“id”: the unique house ID for each participating newly constructed home; equivalent to “House ID” in the concentration dataset

“low_VOC_home”: indication of if the home is marketed as using low volatile organic compound (0/1/-7)

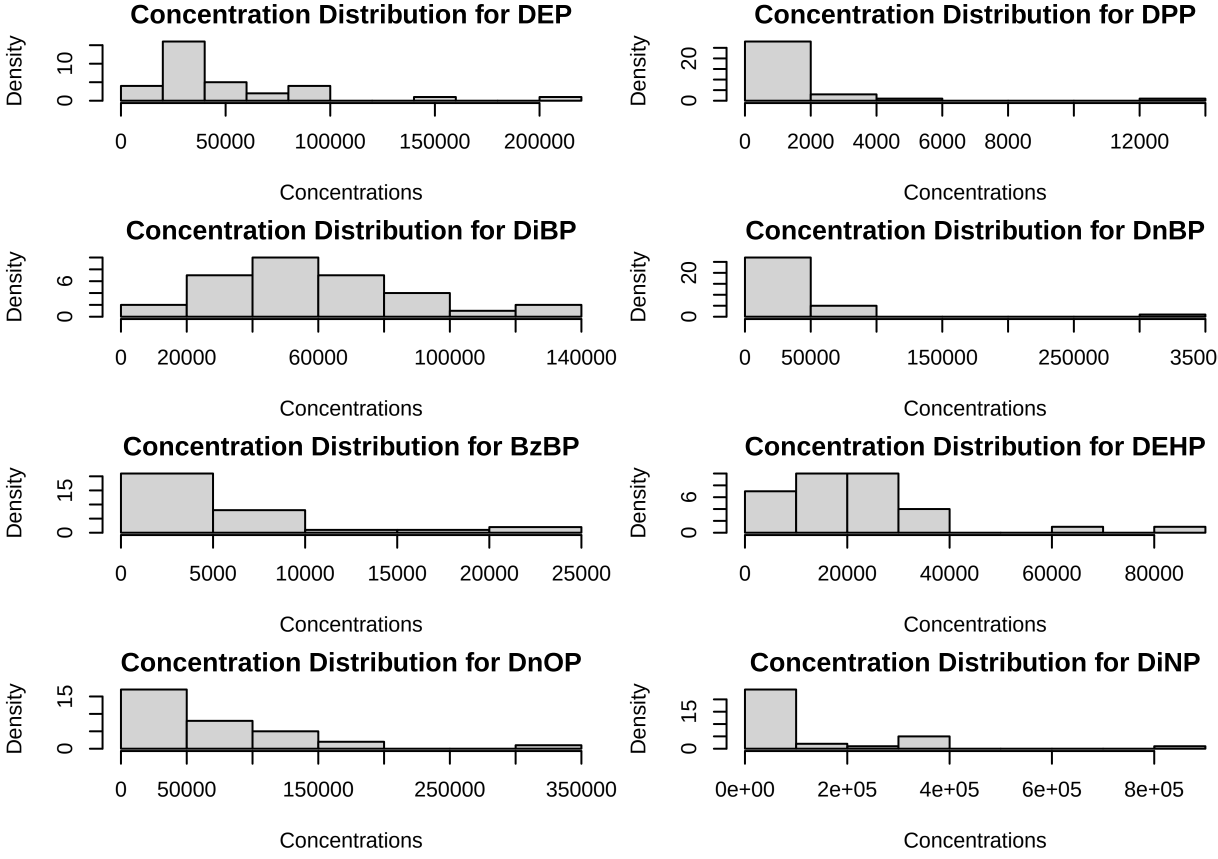
“dogcat_yn”: indication of if the owner keeps dogs or cats as pets inside the home (0/1)

Thus, we have the complete “full_data” with significant variables from technical and questionnaire datasets. We further converted all the concentrations from type string to numeric for analysis and graphical purposes. However, in our “full_data”, most chemicals had missing data or data that was lower than the detection level, we chose to do an analysis on Phthalates (PAEs) in the beginning stage as they had complete data across all the homes at recorded intervals. A new dataset “data_complete_month” was created to only contain the houses with ID “NHAQS-001”, “NHAQS-002” and “NHAQS-003” since they were the only houses in the passive sampler that had complete sampling data. Thus, we could take further exploration of our passive air samples, and see if there are any relationships between the PAE concentrations and human behavior or technical details of the homes.

Data Structure of PAEs

As discussed earlier, because the chemical elements belonging to Phthalates(PAEs) family have relatively more complete information for the participating homes, so in the early stage of our analysis would only involve PAEs. Therefore, it is essential to look into the data structure of the PAEs, which are in total eight chemicals as follows: DEP, DPP, DiBP, DnBP, BzBP, DEHP, DnOP, DiNP.

Fig 1, Concentration distribution graph for each chemical in PAEs



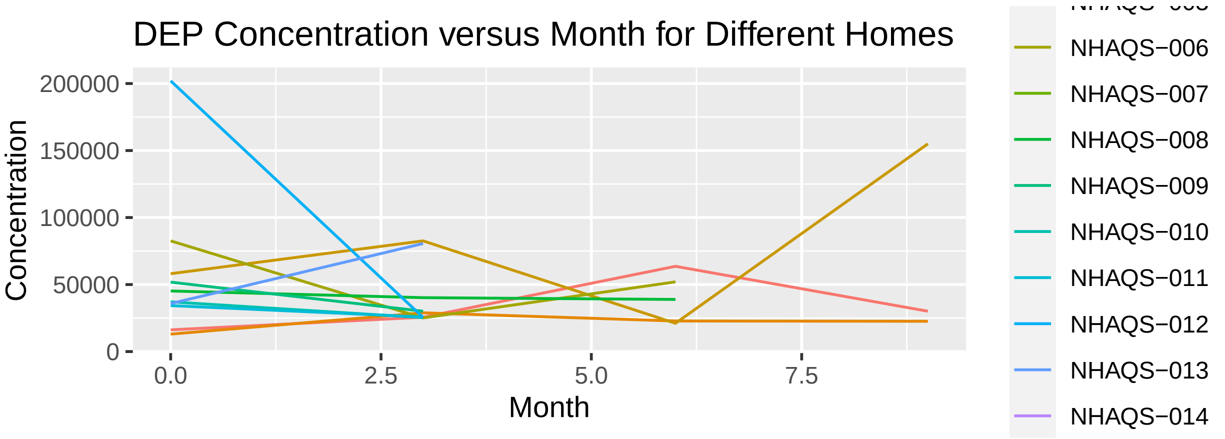
From the above density graph for each chemicals, we could notice that all of distribution are heavily right skewed except for the distribution graph for DiBP. It suggests that there are many outliers and unusual patterns in the concentration of these elements, so we need to be careful about this characteristic during modelling.

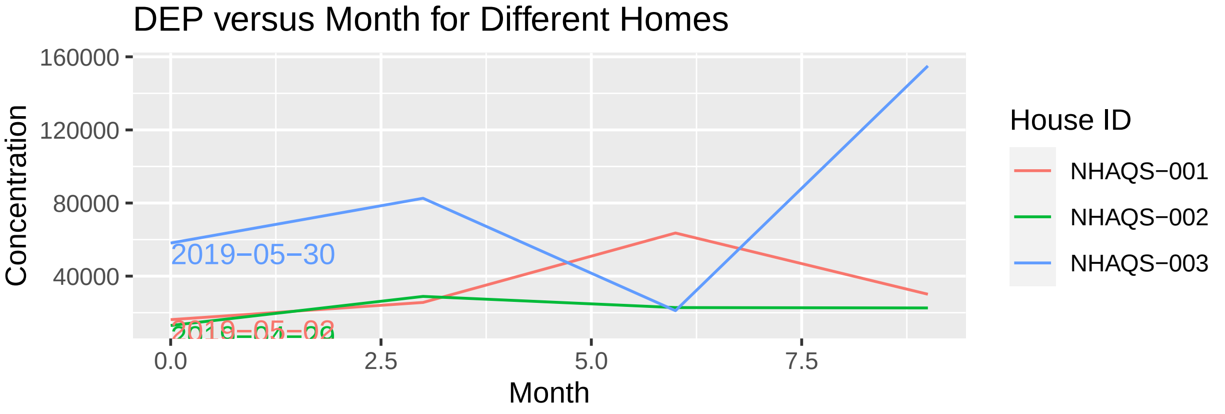
Table 1: Summary Statistics for Phthalates(PAEs)

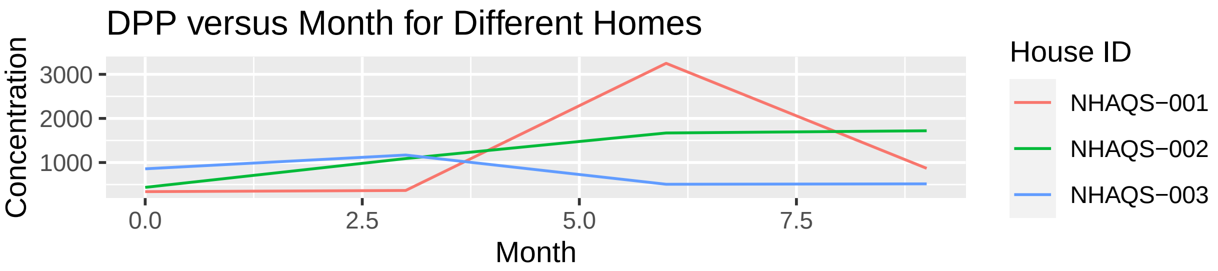
Statistics	Min	Max	Mean	Median	SE
DEP	8550	202000	48329	34300	7136
DPP	159	12500	1513	868	377
DiBP	9420	140000	60713	57000	5475
DnBP	8680	316000	40827	28200	9143
BzBP	1060	22200	5875	4270	903
DEHP	4900	83300	22154	18700	2844
DnOP	2780	302000	67854	48300	11682
DiNP	4050	891000	115036	25100	31831

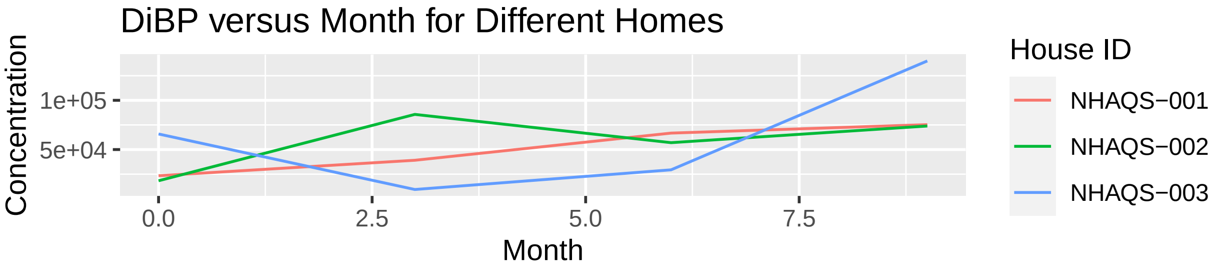
The above summary table for each chemical reveals the central tendency, spread, and variability, providing insights into the distribution of data in the dataset. It also helps identify potential outliers and informs further exploratory data analysis (EDA) and modeling decisions. One key note from this summary table is that the range of concentrations varies widely for each chemical, indicating substantial variability in our chemical concentrations.

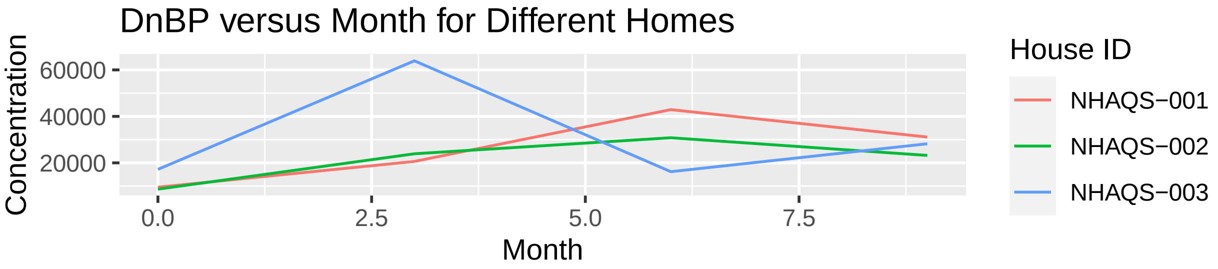
BEN's WORK AFTER THIS________________________






DPP versus Month for Different Homes
Concentration
3000
2000
1000
0.0
2.5
5.0
7.5
Month
House ID
NHAQS−001
NHAQS−002
NHAQS−003


DiBP versus Month for Different Homes
Concentration
1e+05
5e+04
0.0
2.5
5.0
7.5
Month
House ID
NHAQS−001
NHAQS−002
NHAQS−003


DnBP versus Month for Different Homes
Concentration
60000
40000
20000
0.0
2.5
5.0
7.5
Month
House ID
NHAQS−001
NHAQS−002
NHAQS−003