

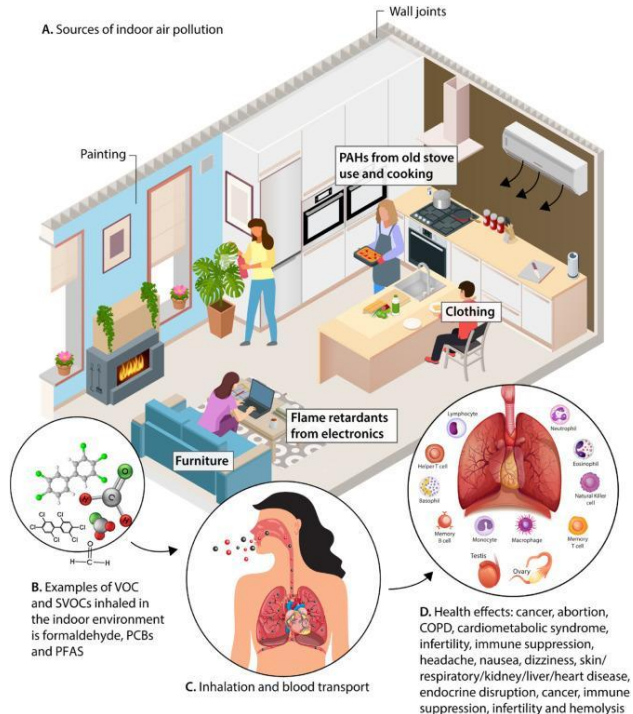


# Chemical Concentrations in Newly Constructed Homes

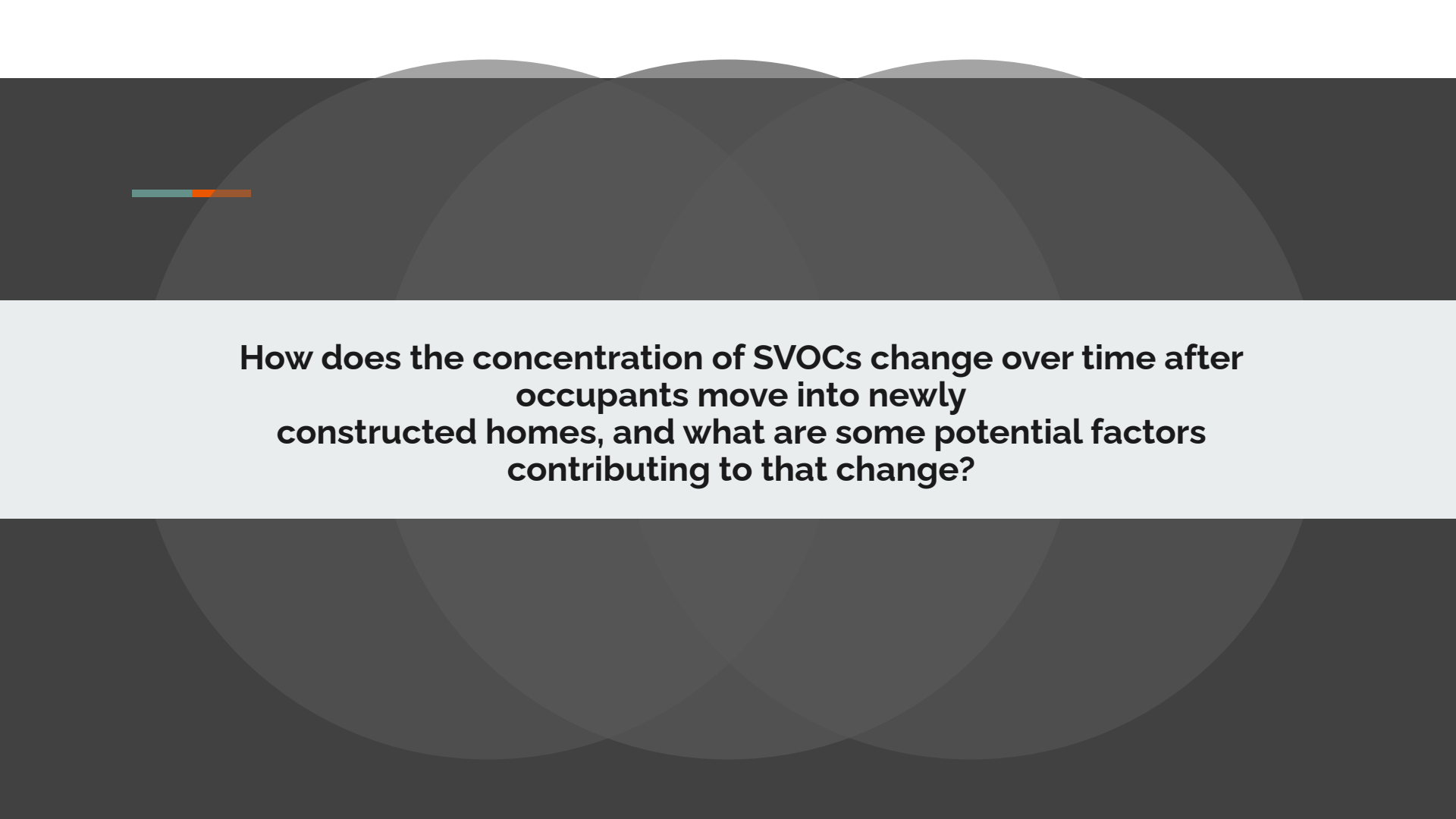
Daihao Wu, Benjamin He, Sehee Kim, You Peng, Tianli Xu



# Semi-volatile organic compounds (SVOCs)



- Prevalent in everyday products like electronic devices, fire retardants, and building materials.
- Exposed to humans through inhalation, skin contact and food contamination
- Linked to increased risks of **respiratory issues, allergies, diagnosed asthma**, and even worse, **birth defects**.



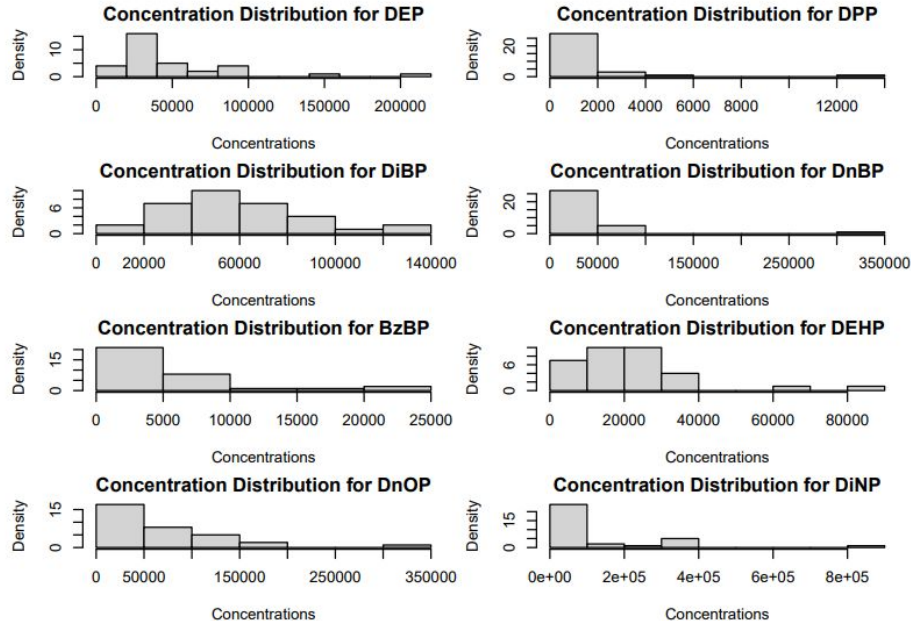
**How does the concentration of SVOCs change over time after occupants move into newly constructed homes, and what are some potential factors contributing to that change?**



# Exploratory Data Analysis

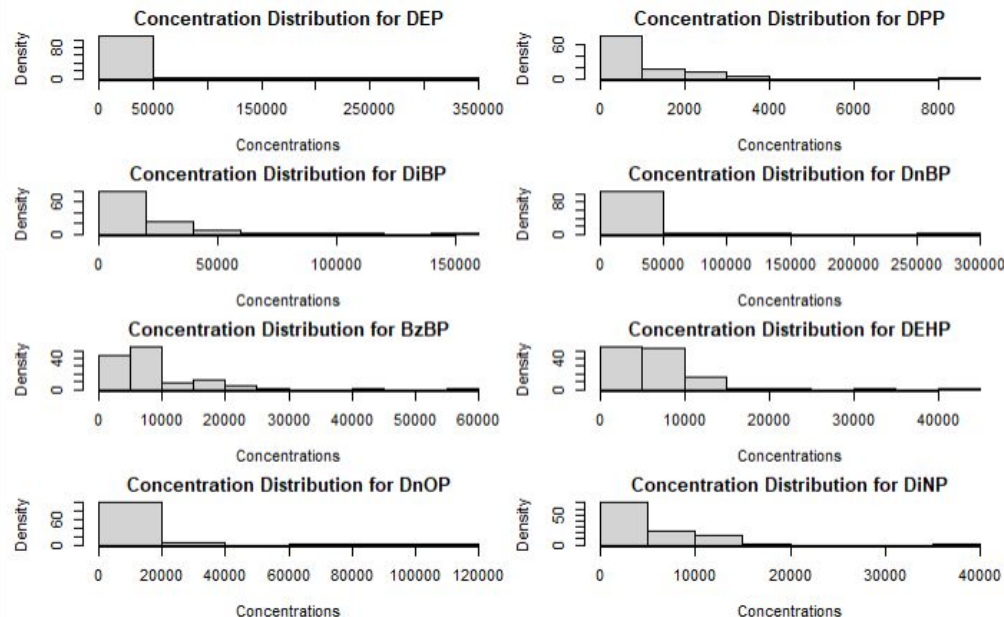
- Concentration Data:
  - Concentration levels collected up to 69 different SVOCs through various methods
  - 2 phases of concentration data available, pre-covid and post-covid
- Technical/Questionnaire Data:
  - The technical details of those newly constructed homes, such as building materials, consists of personal details in the surveyed houses, daily habits, personal behaviours and etc

# Features of Phase 1 Concentration Data



- Due to limited data points for some SVOCs, Only SVOCs belonging to PAE family are examined
- Almost all displayed right-skewed distribution
- Suggests outliers and unusual patterns in the data
- 24 out of 56 PAEs data points are missing

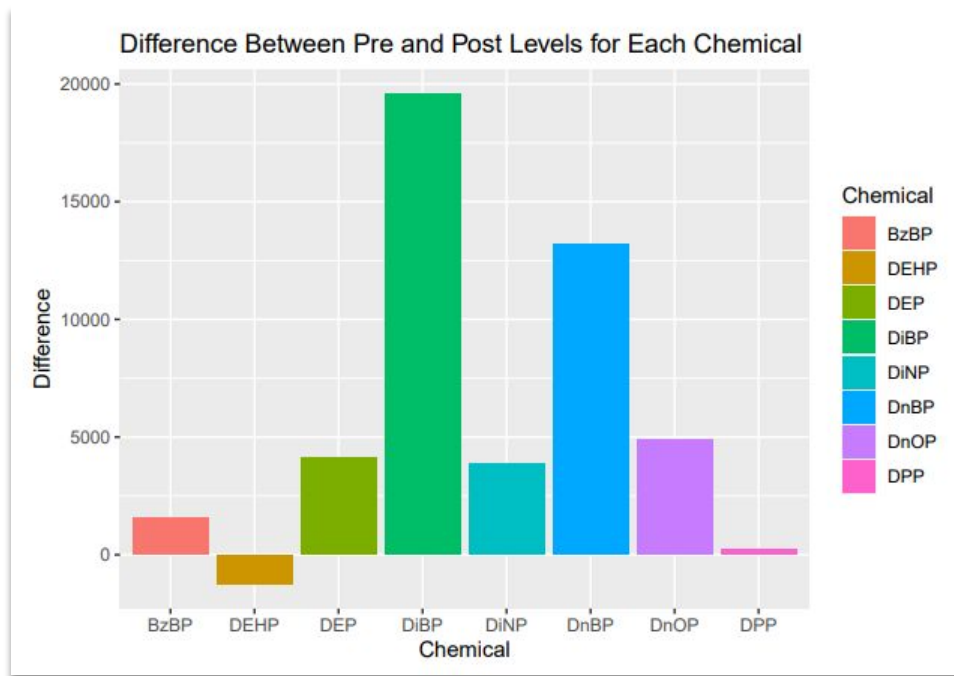
# Features of Phase 2 Concentration Data



- More data points with 36 missing data points out of 132
- Different distribution/shape in 4 of 8 PAEs compared to phase 1 data

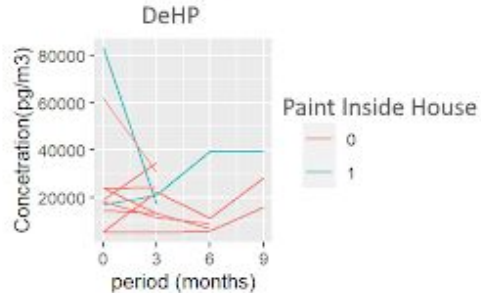
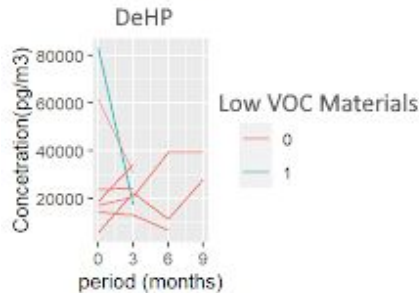
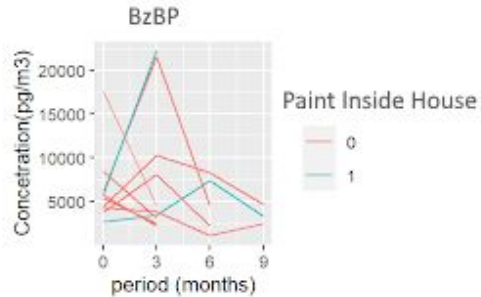
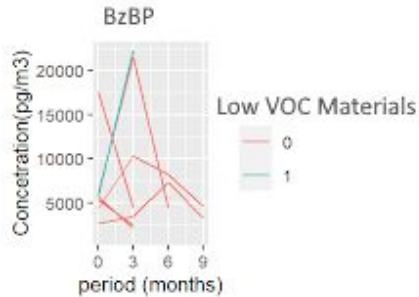


# How do people moving into houses affect PAE chemical levels?



- 7 of 8 PAEs show an increase after move-in
- DiBP and DnBP levels are significantly higher

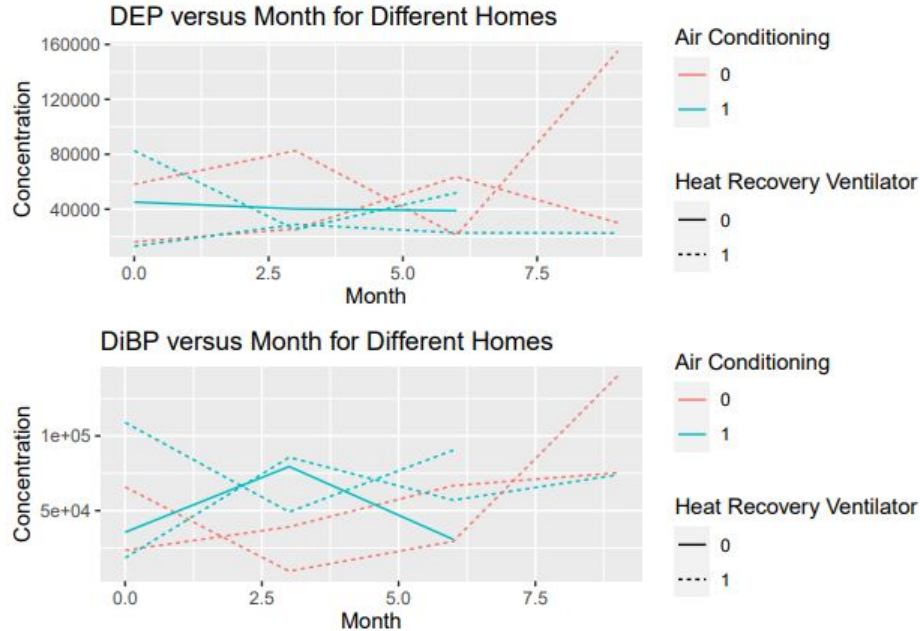
# How does using certain materials affect chemical levels?



□ We cannot observe any significant trend that indicates a difference between chemical levels when using different building materials

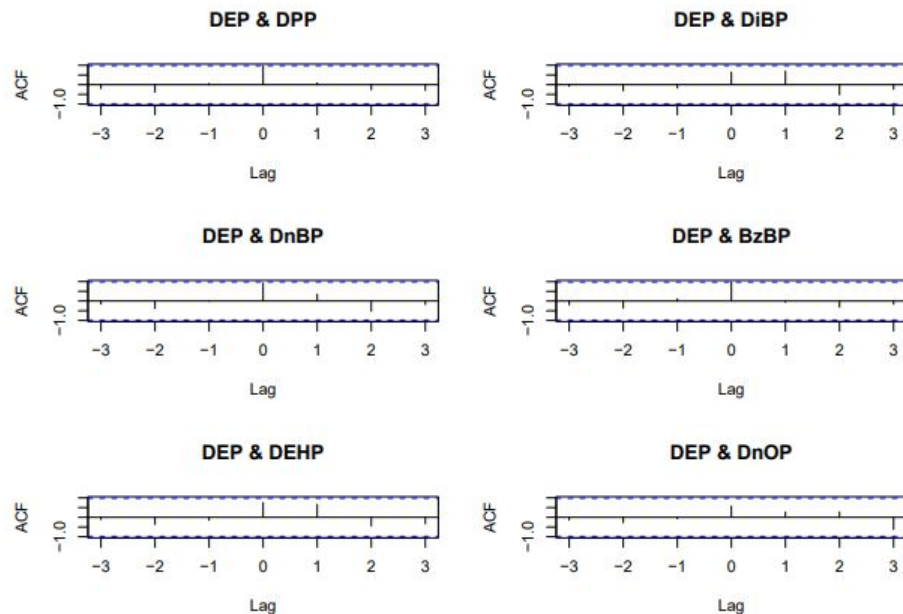


# How does having certain devices such as AC and HRV (Heat Recovery Ventilators) affect chemical levels?



- No observable difference in PAE levels between homes that have AC/HRV installed and homes that don't have AC/HRV installed

# *Does the presence of one chemical affect another?*



- Using cross-correlation analysis, two chemicals with an ACF value greater than 1/-1 indicate a significant correlation
- No ACF values exceed 1/-1, indicating no significant correlation

# Limitations



- ❑ High percentage of missing data in a small sample size
  - ❑ May lead to a statistical model that makes inaccurate conclusions
  - ❑ Poor generalizability, unable to apply the same statistical model to other cases
- ❑ Different trends between the two datasets
  - ❑ Possibly due to different sampling device
  - ❑ Unable to use one dataset to validate the model generated from the other dataset

# Objectives

1. Address the problem of data missingness and sparsity for accurate inferences
  - Investigate different methods of data imputation
  - Simulation study
2. Analyze Linear Trend
  - Compare between Linear Regression Model and Linear Fixed effect Model

# Imputation Methods

---

●

- Uses the value of other subject that has similar properties to fill in missing value

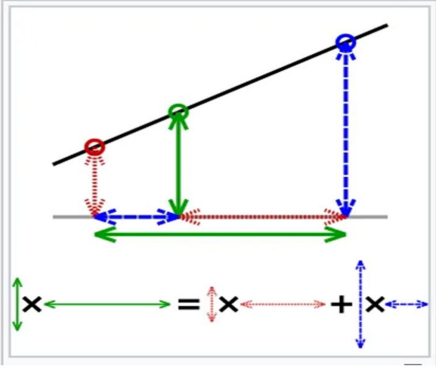
HouseID	Concentration	AC?	Smoking?
1	200	Y	N
2	100	N	Y
3	N/A	Y	N

↓

HouseID	Concentration	AC?	Smoking?
1	200	Y	N
2	100	N	Y
3	200	Y	N

●

- Based on the first and last non-missing value. Connect and follow the trend to impute the missing value



●

- Using the mean of the non-missing data points





# Imputation Methods - Multiple Imputation

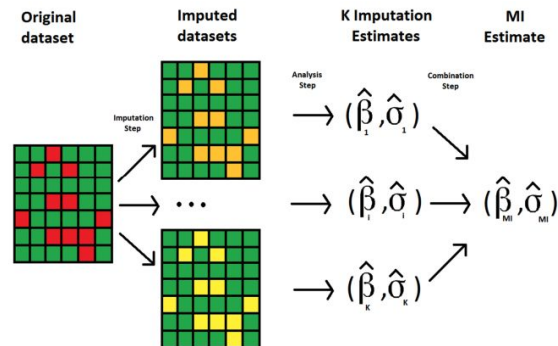
- Joint Modelling

- Create a statistical model that represents the pattern in the data
- Predict the missing values



- FCS-LMM

- Imputed one variable at a time, using the observed values of other variables in the dataset
- Incorporated LMM (linear mixed model) to estimate missing values based on available data



# Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Where:

- Dependent variable is chemical concentration
- Independent variable is Time Period

# Linear Mixed Model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \epsilon_{ij}$$

Where:

- Response is chemical concentration
- Fixed effect is Time Period
- With an additional random intercept for different houses

## Model Comparison Summary

Data: complete\_house

Model	npar	AIC	BIC	logLik	deviance	ChiSq	Pr(> <i>ChiSq</i> )
mod0_DEHP	3	29.286	30.741	-11.6431	23.286		
mod1_DEHP	4	24.552	26.491	-8.2758	16.552	6.7345	0.009456 **

# Likelihood Ratio Test



# Simulations



# Simulation Study

1

## Set the Simulation Parameters

To accurately reflect the linear trend observed in the phase 1 real data  
Eg. slope, intercept

2

## Construct 1000 simulation datasets

Generate a large number of datasets that mimic the characteristics of the original data

Eg. missing value , linear trend

Apply different imputation methods to handle missing data.

3

## Fit the Linear Model Again Using the Simulation Data

4

## Evaluate the effectiveness of each imputation method

Evaluate in terms of its ability to recover the original linear trend.

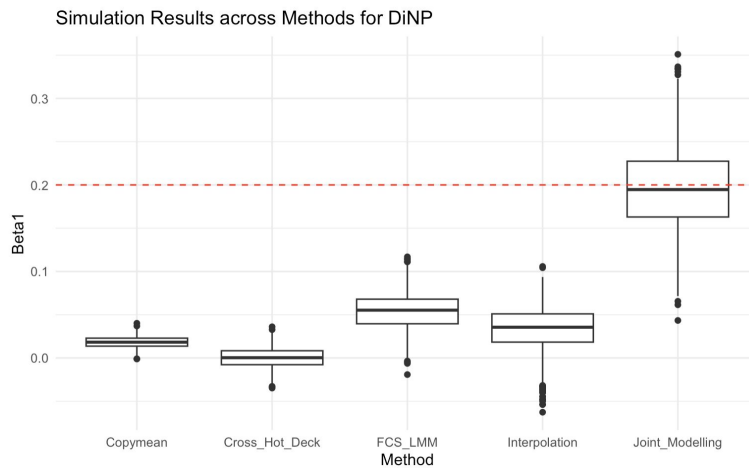
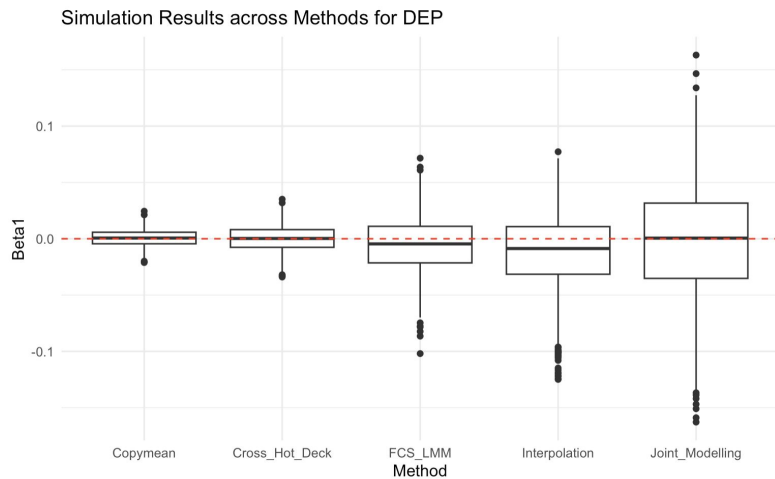




## Simulation results

Method	Coverage #1(%)	Coverage #2(%)
Cross Hot Deck	95.6	80.5
Joint modelling	93.9	93.8
Interpolation	75	38
Copymean	98.9	42.5
FCS_LMM	73	8.9

# Simulation Result Continued



Joint Modelling is preferred as shown from our simulation results

# Going Forward

- Through simulation study, JM was chosen as our imputation method
  - We'll be working on three data sets :
    - Phase 1 imputed data ( n = 56)
    - Phase 2 imputed data (n = 132)
    - Phase 2 complete case data (n = 60)
  - Any measurement below the detection limit was replaced by detection limit value
-

# Model fitting

- Log-transformation on chemical concentration to control skewness of data
- Baseline model includes only Period effect
- Extended our baseline model with building materials and home characteristics
- We have included three covariates that were consistently recorded across all houses, ensuring complete data for each property.
  - **low\_VOC\_home** : homes marketed with low volatile material ( don't know : -7 / no : 0 / yes : 1 )
  - **Paints\_yn** : paints used inside home during monitoring period ( no : 0 / yes : 1 )
  - **Base\_ac\_yn** : homes with air conditioning ( no : 0 / yes : 1 )



## Effect of additional covariates in the model

- In Phase 1 and Phase 2 imputed data sets, adding other covariates did not improve the model fit
  - **These covariates did not have a significant effect on the level of PAEs**
- In Phase 2 complete data, the model including **low\_VOC\_home** and **base\_ac\_yn** was significantly better in **DiBP concentration**
  - In other chemical concentrations, the model with Period was better



## Looking at complete phase 2 data and model (DiBP)

Modelling DiBP			
predictors	Estimate	Std.Error	pvalue
(Intercept)	4.5	1.2	0.00028
Period	0.40	0.071	0.00000075
low_VOC_home0	3.3	0.98	0.0013
low_VOC_home1	-0.69	0.63	0.28
base_ac_yn	-2.0	0.76	0.013

- The low\_VOC\_home and base\_ac\_yn had significant effect in the model
- Since we used log-transformation, exponentiating the estimates give us increase caused by each covariates:
  - Period :  $\exp(0.4) = 1.49$
  - low\_VOC\_home0:  $\exp(3.3) = 27.1$
  - Base\_ac\_yn:  $\exp(2.0) = 7.4$



# Comparing the linear trends of the PAEs across the different data sets.

Model results of phase1 imputed data

Period estimates			
Chemicals	Estimate	Std.Error	pvalue
DEP	0.039	0.037	0.30
DPP	0.061	0.044	0.18
DiBP	0.049	0.041	0.25
DnBP	0.048	0.061	0.45
BzBP	-0.031	0.039	0.43
DEHP	0.014	0.036	0.71
DnOP	0.16	0.090	0.10
<b>DiNP</b>	0.32	0.078	$3.92 \times 10^{-4}$

Model results of phase 2 imputed data

Period estimates			
Chemicals	Estimate	Std.Error	pvalue
<b>DEP</b>	0.33	0.050	$2.47 \times 10^{-8}$
DPP	0.055	0.029	0.063
<b>DiBP</b>	0.28	0.051	$1.68 \times 10^{-7}$
<b>DnBP</b>	0.31	0.052	$4.68 \times 10^{-8}$
BzBP	0.056	0.037	0.14
DEHP	0.026	0.033	0.43
DnOP	0.098	0.057	0.092
<b>DiNP</b>	0.10	0.045	0.027

Model results of phase 2 complete data

Period estimates			
Chemicals	Estimate	Std.Error	pvalue
<b>DEP</b>	0.33	0.044	$3.56 \times 10^{-9}$
<b>DPP</b>	0.060	0.030	0.050
<b>DiBP</b>	0.40	0.072	$1.84 \times 10^{-6}$
<b>DnBP</b>	0.39	0.069	$1.17 \times 10^{-6}$
<b>BzBP</b>	0.11	0.047	0.022
DEHP	0.014	0.039	0.73
<b>DnOP</b>	0.14	0.066	0.037
<b>DiNP</b>	0.12	0.048	0.013

# Findings



- Increasing trends for Phase 1 and Phase 2 are generally the same, but the levels of increase are different due to different measurement scales.
  - However, both indicate that DiNP concentration increases over time.
  - Results are more consistent between Phase 2 imputed and complete data, where 4 chemicals increase over time.
  - There is no evidence that the covariates under investigation affect levels of PAE concentrations.
-

# Future Studies

- Continue on analysing for the other untested chemicals in our dataset
- Consider other factors in technical/survey dataset to determine any significant attributes



# Code

Github link:

<https://github.com/DaihaoWu/Change-in-Chemical-Concentration-in-Newly-Constructed-Homes/tree/main/STA490-Group-5-main>



# Acknowledgement

We wish to acknowledge Dr.Diamond and Sara Vaezafshar for the clarifications and preparation of the dataset, and your help and support throughout the term!

Professor Vianey Leos Barajas for the guidance of the work throughout the year.

Yovna Junglee for the suggestions regarding methodologies and many help during the research process.

We thank you all!

---



# Thank You!

