An Analysis of Change in Chemical Concentrations in Newly-Constructed Homes

Group 5

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

Department of Statistical Sciences
University of Toronto

April 2024

Abstract

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. Our objective is to document the trend in SVOC levels post-occupancy and identify key factors influencing these changes. In our study, we found out that there is evidence that the concentrations of some SVOCs increase over time post-occupancy. By understanding the dynamics of SVOCs' presence in homes, this research aims to contribute valuable insights toward improving safety standards and informing regulatory practices in home construction and material usage.

Contents

1 Introduction

Building materials, such as paints, adhesives, and wood products may contain SVOCs that can slowly be released into the indoor air or absorbed onto the surface over time (Xu and Zhang, 2011). This release can persist for an extended period after construction or installation, leading to prolonged exposure for the inhabitants. Consequently, there is a growing concern about the potential health impacts associated with long-term exposure to SVOC emissions. It is imperative, therefore, to monitor and understand the behaviour of these compounds in indoor environments to develop effective strategies for mitigating exposure risks.

In a groundbreaking study conducted by Health Canada, researchers delved into the dynamics of chemical substances within newly-constructed homes, focusing on the period before the COVID-19 pandemic. This investigation aimed to shed light on the exposure levels of semi-volatile organic compounds (SVOCs) within indoor environments, a concern given that individuals spend approximately 80-90% of their time indoors (Mercier et al., 2011).

While the literature provides a comprehensive overview of chemical exposures in indoor environments, as seen in the work of Dodson et al. (2017), which methodically differentiates between chemicals emanating from building materials and those introduced by occupants, it predominantly focuses on renovated low-income housing and the influence of occupant activities. However, the longitudinal behavior of SVOCs in newly constructed homes and their health impacts over time remains insufficiently explored, and this poses a health risk to the occupants. Hence, our goal is to determine the levels of SVOCs in newly built homes and measure the chemical levels over time after occupants move in.

In this paper, we will analyze the levels of chemical concentration in newly constructed homes post-occupancy as collected by Health Canada. However, the dataset has a small sample size with a large number of missing values. Data imputation techniques are commonly used to replace missing observations in a dataset, but it is unclear how these methods perform with small sample sizes. Therefore, the objectives of this paper are to (1) identify a suitable data imputation technique for our dataset and (2) propose a linear mixed-effects framework to analyze possible trends in chemical concentrations.

2 Data

The dataset, collected from detached houses in close proximity to Ottawa, offers a unique insight into the indoor chemical landscape. The data collection process for this research paper was conducted in two distinct phases. In Phase 1, data were gathered from 18 houses prior to the Covid-19 pandemic, while Phase 2 involved the collection of data from 44 houses after the pandemic had ended. Although various sample collection methods were employed, our study primarily focuses on passive air sampling. It's important to highlight that different equipments were used to record passive air samples across the two phases, leading to discrepancies in measurements. Consequently, the data from the two phases cannot be universally applied or directly compared due to these equipment differences.

Our dataset is composed of three sources of information: chemical concentration data, detailing the levels of various semi-volatile organic compounds (SVOCs) at 0, 3, 6, and 9 months post occupancy; technical survey data, providing technical specifications of the homes, including the types of building materials utilized; and questionnaire data, offering insights into the occupants' characteristics, such as smoking habits.

The concentration data spreadsheet records levels for 69 distinct SVOCs. However, the COVID-19 pandemic disrupted the completion of sampling at many participating houses during the 9-month post-move-in period. Moreover, each sampling device possesses a specific detection limit for various chemicals, leading to numerous data points being marked as 'DL' (below detection limit), rendering their exact values uncertain. Consequently, missing and invalid data points are a widespread issue within the SVOC concentration dataset.

2.1 Data Preprocessing

For the analysis and modelling in our research, we have chosen to focus exclusively on chemicals from the Phthalates (PAEs) group collected using a passive sampler device. This decision was informed by the phase 1 concentration data, which indicated that PAEs collected by the passive sampler had more complete and valid data points compared to other chemical groups, many of which were compromised by extensive missing or invalid values.

Furthermore, to facilitate a deeper investigation into the factors influencing PAEs chemical concentration levels, we integrated select variables from the technical survey (e.g. "base ac yn": indication of if the home has air conditioning) and questionnaire datasets (e.g. "dogcat yn": indication of if the owner keeps dogs or cats as pets inside the home) with the chemical concentration data. The specific variables selected for this merged analysis are detailed in Appendix A.1.

2.2 Concentration Data Overview

For our Phase 1 concentration data, only 3 houses have the complete data from month 0 up to month 9 for PAEs. For Phase 2 concentration data, 15 houses have the complete data from month 0 to month 9 for PAEs.

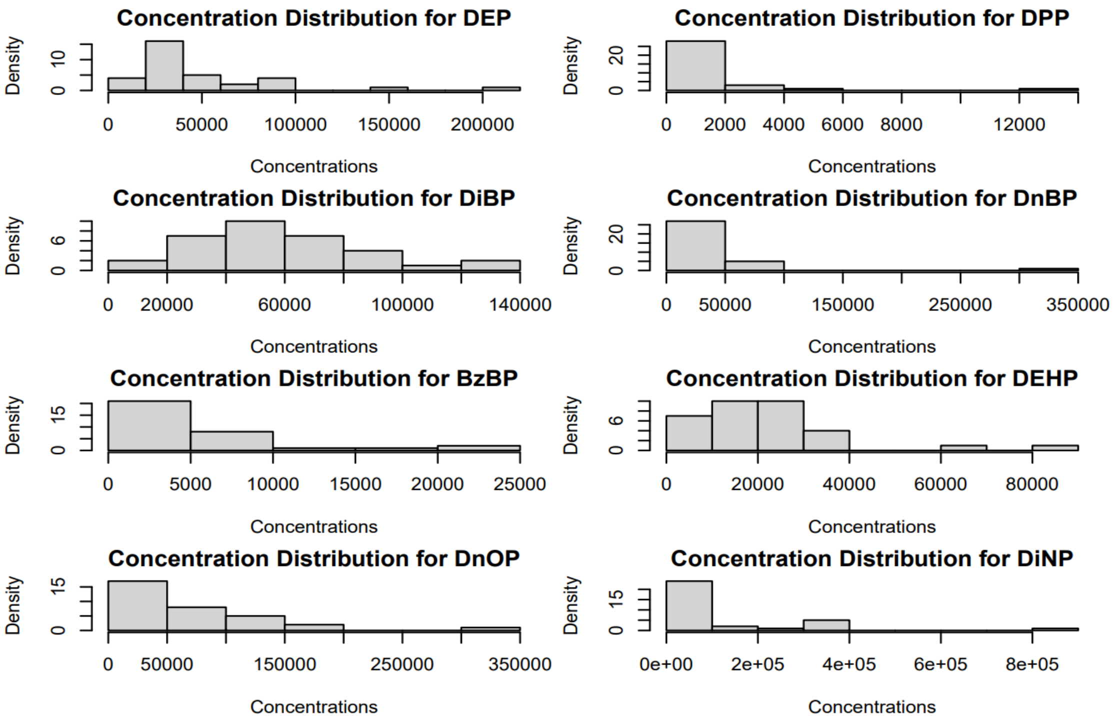
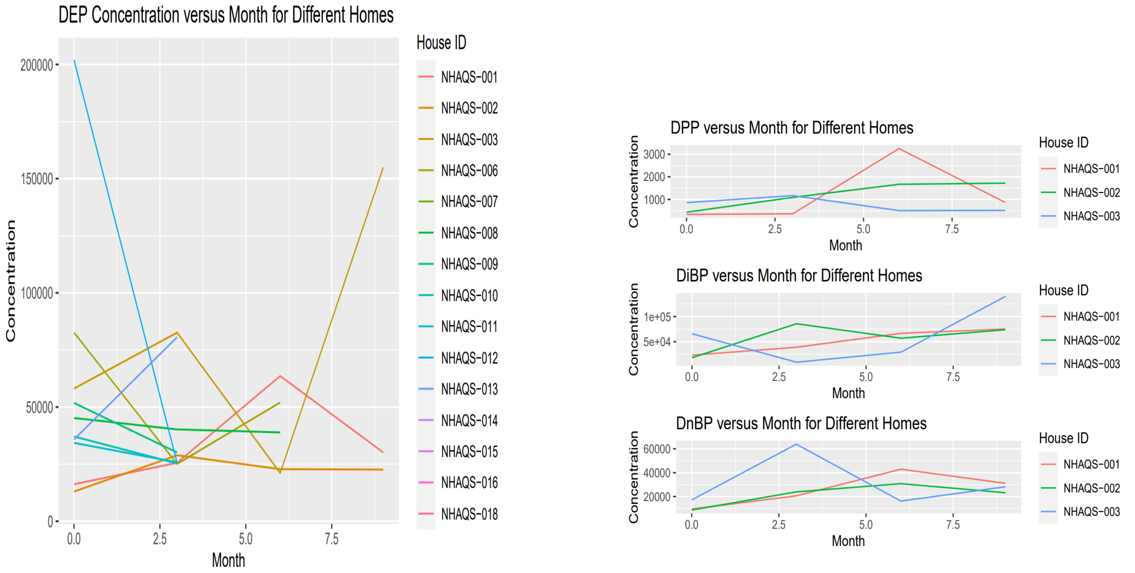


Figure 1: Concentration Distribution of Different PAEs

Figure 1 is the chemical concentration level distribution for different PAEs in phase 1 data. Note that most of the distributions are heavily right-skewed with some outliers and unusual patterns.



(a) DEP Levels for All Homes

(b) Chemical Levels for Homes with Complete Data

Figure 2: Chemical Concentration over Period in all Homes and also Homes with Complete Data

We also visualized the concentration levels of PAEs over time for all the homes and also for the 3 homes with complete data (Figure 2b). No significant trend was observed.

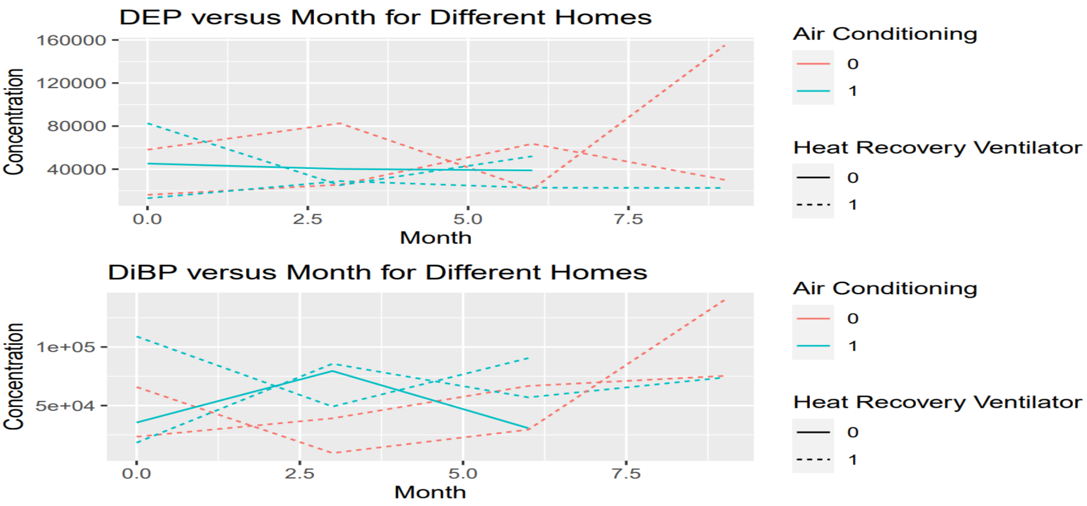


Figure 3: DEP DiBP Levels for Homes with or without AC and Heat Recovery Ventilator

Additionally, from Figure 3, when we also include the technical details such as whether or not the homes are equipped with AC or heat recovery ventilators, no apparent trend could be observed. For more figures exploring relationships between PAE and other technical variables, refer to Appendix A.2

We also fitted a cross-correlation function to examine if the presence of one chemical causes another to change, and all PAE variables are independent of each other and have no effects on each other (Appendix A.3).

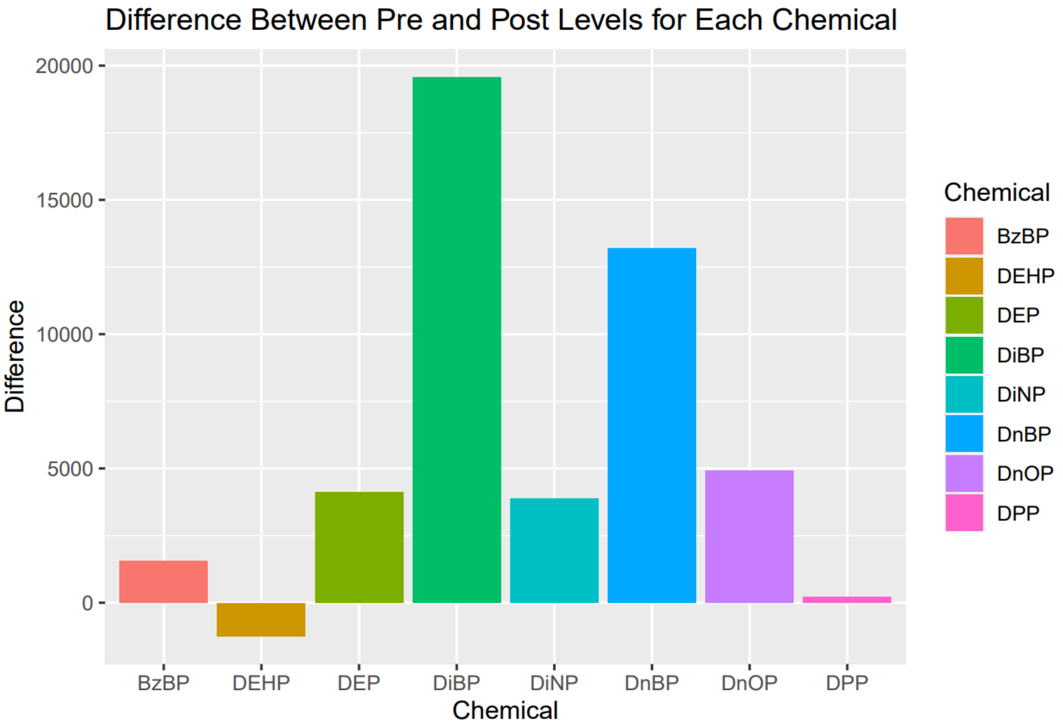


Figure 4: Difference in PAE Concentration Levels Pre and Post Occupancy

However, looking at figure 4, when we examined the PAEs concentration levels before and after occupants move in, 7 out of 8 PAEs showed an increase in concentration levels, suggesting that high PAE chemical level may be associated with human activity and is likely to increase after owners move into a newly constructed home.

Thus, despite the initial analysis of our data and visualizations not revealing any clear trends or significant outcomes, the observed variations in PAE levels after occupants move in offer valuable insights into how chemical concentrations may fluctuate over time. This observation underscores the importance of investigating these changes further.

3 Methods

3.1 Linear Mixed Model & Simple Linear Regression Model

In the introduction, we propose to utilize both **linear mixed models(LMMs)** and **simple linear regression models(SLRs)** to analyze the variation in chemical concentrations over time. These frameworks allow us to model the linear trend of the concentrations, with the advantage of LMM being that it accounts for variability that occurs between houses (Baayen et al., 2008).

Simple Linear Regression:

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

Where:

$$\begin{aligned} Y_i &: \text{Chemical concentration for observation } i \\ X_{ij} &: \text{predictor (Time period) for observation } i \\ \beta_0 &: \text{Intercept} \\ \beta_1 &: \text{Coefficient for } X_i \\ \epsilon_i &: \text{Error term for observation } i \end{aligned}$$

Linear Mixed Model:

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

Where:

$$\begin{aligned} Y_{ij} &: \text{Chemical concentration for observation } i \text{ in house } j \\ X_{ij} &: \text{Fixed effect predictor (Time period) for observation } i \text{ in house } j \\ \beta_0 &: \text{Intercept (fixed effect)} \\ \beta_1 &: \text{Coefficient for } X_{ij} \text{ (fixed effect)} \\ u_j &: \text{Random intercept for house } j \\ \epsilon_{ij} &: \text{Error term for observation } i \text{ in house } j \end{aligned}$$

Note that we have performed a log transformation on our response variable due to the skewness of our data (see Figure 1). Also note that we only have one fixed effect predictor, which is the time period of the observation. We added random effects to the intercept term to account for differences in the average levels of concentrations between houses. Moreover, we assumed that $u_j \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, this assumes independence between and within houses.

3.2 Imputation Methods

Given the presence of substantial missing data within our dataset, directly applying modelling techniques, particularly linear mixed models, could lead to issues of singular fit, compromising the validity and reliability of our analyses. To mitigate this challenge, we applied imputation methods to increase the sample size, which allowed us to proceed successfully with fitting the model. Informed by the insights from Jahangiri et al.'s work on missing imputation approaches in longitudinal data (Jahangiri et al., 2023), we selected five imputation methods for our analysis, incorporating two multilevel imputation techniques, and multilevel imputation involves estimating missing data by considering the hierarchical structure of the dataset, allowing for a more nuanced handling of data dependencies across different levels, such as within and between groups of observations.

1. The interpolation method assumes the data has an underlying linear trend. It uses the slope calculated between the first and last seen non-missing value to estimate the following missing values (refer to Appendix B.1).

2. The Copy-Mean method calculates the mean of all non-missing values within a single household, depending on the presumed stability of the data. The computed mean is then used to fill in the missing values, assuming that these gaps can be reasonably approximated by the overall average observed (refer to Appendix B.2).

3. The Cross Hot Deck imputation method locates a non-missing data ("donor") with similar attributes and the same period as the missing data within the same dataset. Once a donor is located, the donor's value will be used to impute the missing data. In this implementation, if multiple donors are found, we randomly select one donor to avoid bias toward any particular donor.

4. The Fully Conditional Specification with Linear Mixed Models(FCS-LMM) is a multilevel imputation method that employs a chained equations process, iteratively imputing missing values for each variable by modelling it with linear mixed models that account for both fixed effects and random effects. (For details, refer to Appendix B.3)

5. The joint modelling imputation is a multiple imputation that is extended from multivariate multilevel imputation. The method uses a multivariate linear mixed effect model to impute missing data. In here, the model is simplified to

$$Y_{ij} = \mu + Y_{B_j} + Y_{W,ij}$$

where Y_{Bj} is the random effects between groups and $Y_{W,ij}$ residuals within groups, making this as a decomposition model that decomposes the multivariate outcome into between groups and within groups (Grund et al., 2017). This enables multivariate imputation to work naturally with multilevel data. There are many approaches and packages to compute multivariate multilevel imputation such as pan (Grund et al., 2016) and mitm (Grund et al., 1997). In the present study, we used jomo which is derived from REALCOM (Carpenter et al., 2011) in Matlab, which allows us to use a combination of multilevel continuous and categorical data which gives flexibility when choosing an imputation model (Quartagno et al., 2019). The imputation model is chosen based on our analysis model (model of interest) for our multiple imputation results in accurate inference.

When taking the chosen imputation model, it uses Bayesian methods and Markov Chain Monte Carlo (MCMC) to fit the multivariate normal model and impute the missing data (Schafer, 1997). For accurate data imputation, the number of burn in iterations and between iterations are chosen by checking convergence with trace plots.

4 Simulation Study

4.1 Set up

Table 1: **Parameter Setting for Two Scenarios**

Parameters	Scenario #1 (DEP)	Scenario #2 (DiNP)
α_0	10.5	9.8
σ_0	0.2	0.3
β_1	0.0	0.2
σ_2^2	1.4	1.1

The previous studies about imputation methods are conducted on large data sets, while our dataset is small. To make sure our imputation methods can also capture the original pattern of the data when the sample size is small, we performed a simulation study. The study consists of two simulations. The first simulation assumes the underlying complete data doesn't present a significant linear relationship between the fixed effect predictor and the response. This corresponds to the scenario when the chemical concentration does not change over time. On the other hand, the second simulation assumes the underlying complete data has a significant linear relationship between the period and the chemical concentrations. The purpose is to examine whether these methods are able to accurately impute the missing observations such that the fitted model on the imputed data set captures the true underlying trend.

Table 1 shows the parameter setting for two simulations. In the first simulation, we first find a chemical data (DEP) that doesn't show a significant linear relationship between its chemical concentrations and period. Then we construct the pseudo dataset from the following model:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \epsilon_{ij}$$

Where:

$$\begin{aligned} Y_{ij} &: \text{Chemical concentration for observation } i \text{ in house } j \\ X_{ij} \in (0,3,6,9,12) &: \text{Time period for observation } i \text{ in house } j \\ \beta_{0j} &: \text{Intercept generated from } N(\alpha_0, \sigma_0^2) \\ \beta_1 &: \text{Slope estimate for time period obtained from previous Linear Mixed Model} \\ \epsilon_{ij} &: \text{Error term for observation } i \text{ in house } j \text{ generated from } N(0, \sigma_2^2) \end{aligned}$$

The parameters were fixed to the estimates obtained from fitting a linear mixed effect model to the incomplete data set, where α_0 and σ_0 represent the intercept estimate and its corresponding standard deviation. And σ_2^2 represents the residual variance.

For each scenario, we simulated 1000 datasets. For each generated dataset, we remove observations based on the patterns observed on the home concentration data set. Therefore, we will have 1000 datasets that have missing values in the same places as our actual dataset. Then we apply the 5 imputation methods discussed under Method Section on these 1000 datasets and obtain 1000 imputed resulting datasets for each method under consideration. For these, we fit the LMM, and we obtain the estimated coefficient for time, which is β_1, as well as its corresponding confidence interval (95%CI). Lastly, we calculate the coverage probability defined as the proportion of 95%CIs that capture the underlying true value of β_1.

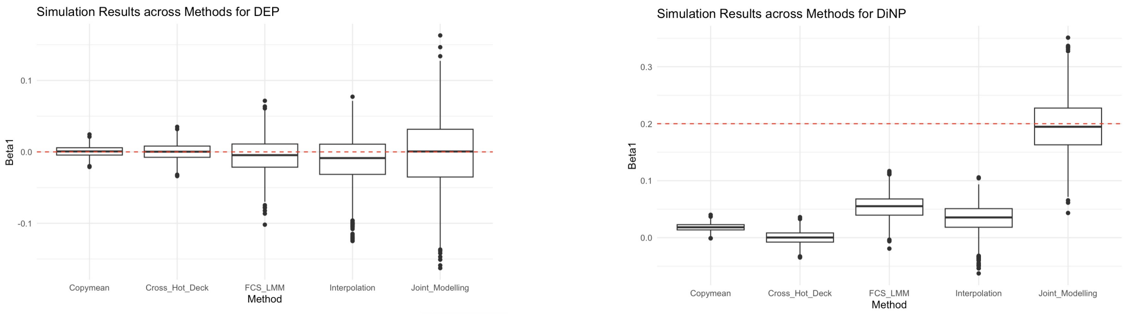
In the second simulation, we follow the same procedure as the first simulation, except that the model parameter we used to generate the completed pseudo dataset is different. This time, the parameter is based on the chemical data DiNP, where time was significant in the model fitted using the incomplete data.

4.2 Simulation Results

Table 2: **Coverage Percentages for Different Imputation Methods**

Method	Scenario #1 (%)	Scenario #2 (%)
Cross Hot Deck	95.6	80.5
Joint Modelling	93.9	93.8
Interpolation	75.0	38.0
Copymean	98.9	42.5
FCS_LMM	73.0	8.9

From table 2, the coverage represents the proportion of CI's that could capture the underlying true value of β_1 in our original model before simulations. All of the methods performed relatively well in the first simulation, where all of them had more than 75% of coverage. However, only the Cross Hot Deck and Joint Modelling methods are consistently effective across different data structures with over 80% of coverage across different set-ups of simulations, making them reliable choices for imputing missing SVOC data.



(a) Simulation Results across Methods for DEP

(b) Simulation Results across Methods for DiNP

Figure 5: Side by Side Boxplots of β_1 from Simulated Models for DEP and DiNP

Moreover, the estimates of the coefficient β_1 were derived from models using the five imputation methods, across two scenarios. Figure 5 displays the distribution of the estimated β_1 coefficients under scenarios 1 and 2. The red dotted line represents the actual β_1 value used to generate observations under scenarios 1 and 2. Notice from the boxplots for For DEP (no significant linear relationship with time i.e. $\beta_1 = 0$), all of the β_1 distributions from different imputation methods lie around the red dotted line, implying that all performed well in capturing the actual β_1 value.

However, looking at the side-by-side boxplots for the second scenario (significant linear relationship with time i.e. $\beta_1 = 0.2$), only the distribution of β_1 from the joint modelling method lies around the red dotted line. Therefore, for data that lacks a strong linear trend (DEP), most methods perform reasonably well. However, for data with a significant linear trend (DiNP), Joint Modelling seems to offer the best performance. Furthermore, Joint Modelling also appears to have a higher variability around the true coefficient β_1 across two different scenarios.

Based on the simulation results presented in both the table and the boxplots, the Joint Modelling imputation method appears to be the most consistent and reliable across different scenarios. It maintains coverage percentages close to the desired 94% level and centers its estimates around the actual β_1 value for both scenarios. This method effectively captures the variability inherent in the dataset while closely approximating the true parameter value, indicating its robustness regardless of the underlying data trend. Therefore, for a dataset characterized by semi-volatile organic compound (SVOC) concentrations in indoor air that may or may not exhibit strong linear trends over time, Joint Modelling stands out as the preferred imputation method due to its overall accuracy and consistency.

5 Results

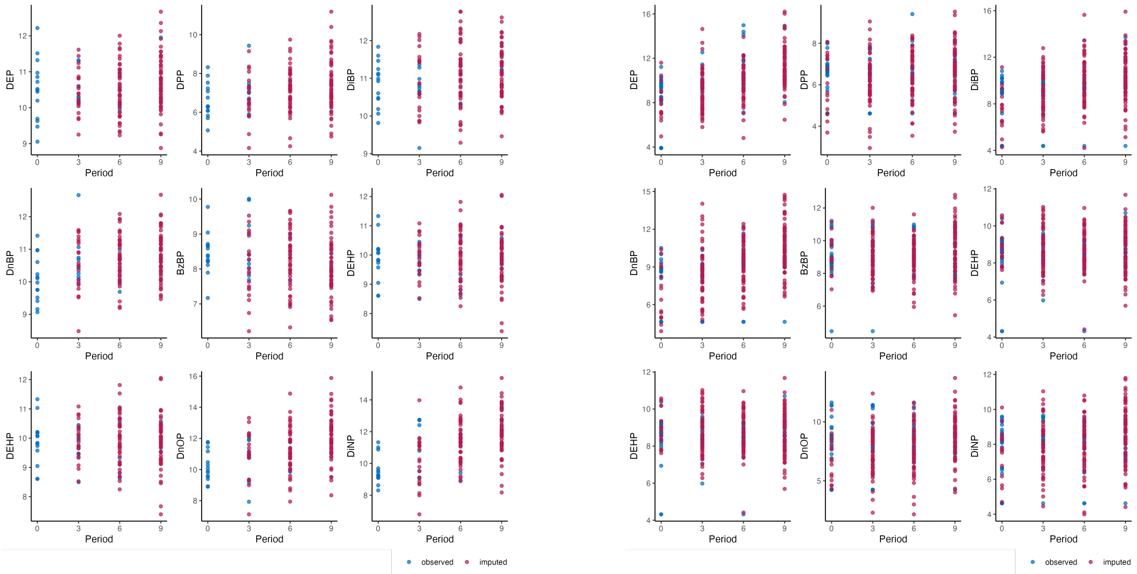
5.1 Model Selection

In finalizing our model selection process, we employed a likelihood ratio test to compare the efficacy of a linear mixed model (LMM) against a simple linear regression (SLR) model both fitted using the incomplete (before imputation) Phase 1 data, and only the houses with fully recorded data were included. This decision-making step was grounded in our analysis of the first group of chemicals, which includes DEP, DPP, DiBP, DnBP, BzBP, DEHP, DnOP, and DiNP. We posited that these chemicals are representative for the behavior of all chemicals under study. We present the analysis results(Table 3) for DEHP, which is noteworthy for not exhibiting issues related to singular fit— a problem that did not affect this particular analysis. Details for the other chemicals can be found in Appendix B. The output from the R anova() function helps us in this comparative analysis. Based on the test statistics(p-value = 0.009456) presented in Table3, we conclude that there is sufficient evidence to prefer the linear mixed model over the simple linear regression model. This conclusion stems from our inability to reject the alternative hypothesis that favors the linear mixed model, indicating its better fit for our data and research context.

Table 3: Model Comparison for DEHP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
slr_DEHP	3	29.286	30.741	-11.6431	23.286			
lmm_DEHP	4	24.552	26.491	-8.2758	16.552	6.7345	1	0.009456**

5.2 Imputation and Model Fitting



(a) Phase 1 (missing ratio = 0.43)

(b) Phase 2 (missing ratio = 0.27)

Figure 6: Imputation results

The data from Phase 1 and Phase 2 were imputed using joint modelling multiple imputation with a simple random intercept model (1), as selected in Section 5.1. Each imputation cycle produced 10 datasets. The number of burn-in iterations was determined individually for each chemical by assessing convergence through trace plots and auto-correlation function (ACF) plots. Figure 6 shows imputed data from phase 1 and 2.
The utility of including building materials and home characteristics (low_VOC_home, paints_yn, base_ac_yn) in the model was evaluated and the following models were fitted and compared:

$$\text{Model\_null}: Y_{ij} = \beta_{0j} + \beta_1 \text{Period}_{ij} + u_j + \epsilon_{ij} \tag{1}$$

$$\text{Model\_ac}: Y_{ij} = \beta_{0j} + \beta_1 \text{Period}_{ij} + \beta_2 \text{base\_ac\_yn}_{ij} + u_j + \epsilon_{ij} \tag{2}$$

$$\text{Model\_low}: Y_{ij} = \beta_{0j} + \beta_1 \text{Period}_{ij} + \beta_2 \text{low\_VOC\_home}_{ij} + u_j + \epsilon_{ij} \tag{3}$$

$$\text{Model\_paints}: Y_{ij} = \beta_{0j} + \beta_1 \text{Period}_{ij} + \beta_2 \text{paints\_yn}_{ij} + u_j + \epsilon_{ij} \tag{4}$$

$$\begin{aligned} \text{Model\_full}: Y_{ij} &= \beta_{0j} + \beta_1 \text{Period}_{ij} + \beta_2 \text{paints\_yn}_{ij} \\ &\quad + \beta_3 \text{low\_VOC\_home}_{ij} + \beta_4 \text{base\_ac\_yn}_{ij} + u_j + \epsilon_{ij} \end{aligned} \tag{5}$$

Using `jomo`'s flexibility, along with those five models, the various interactions between predictors were tested additionally. Model estimates are pooled following Rubin's rule (Rubin, 1987), using the function `testEstimates`, and models are compared using the pooled Wald test implemented in `testModels`. Additionally, the complete case data in Phase 2 were analyzed using the same models and compared through likelihood ratio tests (LRTs). We evaluated the model fitting results across these three datasets to validate our methodological approach. For a comprehensive presentation of model comparison and fitting results, please refer to Appendix D in the research paper.

5.2.1 Phase 1

In phase 1 data, 24 out of 56 (43% missing ratio) chemical concentrations were imputed using joint modelling multiple imputation. All of chemicals favoured null model (1) over models that included technical variables (e.g., including building materials) with significance level 0.05 (17).

Model	F.value	df1	df2	P(>F)	RIV
Model_ac	0.017	1	5648.56	0.90	0.042
Model_low	1.25	2	36.25	0.30	0.82
Model_paints	0.14	1	8043.60	0.71	0.035

Table 4: likelihood test results comparing each of models(2)-(4) with model (1) in phase 1

Table 4 presents the results of the model comparison using pooled Wald test for Phase 1 DiBP concentration against the model detailed in Equation 1. The p-values for each test provide evidence against the null hypothesis of no effect of the corresponding variable; thus, null model 1 with additional covariates did not enhance the model fit and null model explains is as good as the model with other covariates.

5.2.2 Phase 2

In phase 2, 36 out of 132 chemical concentrations(27% missing ratio) were imputed. Similar outcomes were observed in the model comparisons for other chemicals in phase 2, where the null model (1) could not be rejected (18). Table 5 shows the model comparisons with pooled estimates for DiBP. Including building materials did not result in better model fit. Consequently, we will employ model 1 to assess the model fit across various datasets.

Model	F.value	df1	df2	P(>F)	RIV
Model_ac	0.28	1	829.59	0.60	0.12
Model_low	0.63	2	437.27	0.53	0.15
Model_paints	0.079	1	15773.33	0.78	0.024

Table 5: Likelihood test results comparing each of models(2)-(4) with model (1) in phase 2

In the analysis of Phase 2 complete data, we focused on 15 houses with complete observations, where the concentration was measured at every time point. Concentrations falling below the detection limit were substituted with the detection limit value.

Chemical	Model	Df	LogLik	Df	Chisq	Pr(>Chisq)
DEP	Model_null	4	-106.58			
	Model_ac	5	-104.17	1	4.82	0.028
	Model_low	6	-104.29	2	4.57	0.10
	Model_paints	5	-105.38	1	2.40	0.12
DiBP	Model_null	4	-128.86			
	Model_ac	5	-124.26	1	9.20	0.0024
	Model_low	6	-125.61	2	6.50	0.039
	Model_paints	5	-127.52	1	2.68	0.10
	Model_lowac	7	-120.12	3	17.48	0.00056

Table 6: Likelihood test results comparing models (2)-(4) with model (1) in phase 2 complete cases

After conducting likelihood test results comparing each of models (2)-(4) with model 1, DEP and DiBP were identified as the only chemical for which the building materials had a significant effect (p-value < 0.05)

As shown above (Table 6), likelihood ratio test of DEP model 2 with the null model (1), the null model was rejected with p-value of 0.03. However, the model fitting result of model 2 showed that base_ac_yn had no significance (p-value > 0.05). To validate the result, the model 1 and 2 were once again compared with ANOVA (Table 7). The p-values and AIC provide evidence against the null hypothesis of no effect of the corresponding variable.

Table 7: Model_Null vs Model_ac with ANOVA

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
Model_null	4	216.29	224.66	-104.14	208.29			
Model_ac	5	215.48	225.95	-102.74	205.48	2.81	1	0.094

Table 8 displays model fitting result of DiBP. From the table, we can see that houses with air conditioning decreases chemical concentration by 1.95 than those did not have air conditioning. The houses who knew that they did not have low VOC material for their homes had 3.33 higher chemical concentration than those who did not know about their building material.

predictors	Estimate	Std.Error	pvalue
(Intercept)	4.51	1.16	2.8×10^{-4}
Period	0.40	0.071	7.5×10^{-7}
low_VOC_home0	3.33	0.98	0.0013
low_VOC_home1	-0.69	0.63	0.28
base_ac_yn	-1.95	0.76	0.013

Table 8: DiBP model fitting with low_VOC_home and base_ac_yn

5.2.3 Comparing model fitting

The chemical concentrations in both imputed data in phase 1 and phase 2 were better explained with the simple model (1) where phase 2 complete data had DiBP that inclusion of building materials (low_VOC_home, base_ac_yn) improved it's model fit. This difference may come from small sample size and skewed data density. For example, the phase 2 complete data only had 1 house that did not have air conditioning.

Since most chemical concentrations did not have a significant building material effect, we compared the estimates of the time effect obtained when fitting model 1 using the imputed phase 1 and phase 2 data sets, and the complete case phase 2 data set.

Chemical	Phase 1 imputed			Phase 2 imputed			Phase 2 complete case		
	Estimates	Std.Error	p-value	Estimates	Std.Error	p-value	Estimates	Std.Error	p-value
DEP	0.039	0.037	0.30	0.33	0.050	1.73×10^{-10}	0.33	0.044	3.56×10^{-9}
DPP	0.061	0.044	0.18	0.055	0.029	0.063	0.060	0.030	0.050
DiBP	0.049	0.041	0.25	0.28	0.051	1.68×10^{7}	0.40	0.072	1.84×10^{6}
DnBP	0.048	0.061	0.45	0.31	0.052	4.68×10^{8}	0.39	0.069	1.17×10^{6}
BzBP	0.031	0.039	0.43	0.056	0.037	0.14	0.11	0.047	0.022
DEHP	0.014	0.036	0.71	0.026	0.033	0.43	0.014	0.039	0.73
DnOP	0.16	0.090	0.10	0.098	0.057	0.092	0.14	0.066	0.037
DiNP	0.32	0.078	3.92×10^{-4}	0.10	0.045	0.027	0.12	0.048	0.013

Table 9: Period Estimates

From Table 9, the time effects in the complete case data were significant ($p\text{-value} < 0.05$) for all compounds except DEHP. Notably, the time effects for Phase 2 imputed data and Phase 2 complete case data are comparable; for instance, the slope (Period effect) for DEP concentration is identical at 0.33 with similar standard error. This indicates that in Phase 2, the DEP concentration increases by 0.33 for each one-unit increase in the Period with 95% confidence interval (0.237,0.413). The discrepancies between the slopes in Phase 1 and Phase 2 data may be attributed to differing measurement techniques employed in the collection of Phase 2 data. We found a positive trend post-occupancy for DiNP across all three data sets i.e., the concentration of DiNP increases over time post-occupancy. .

6 Discussion

In this article, the primary focus was to examine any trends in chemical concentrations in newly constructed homes and comparing the efficacy of various imputation methods for handling missing data. Our research findings from the phase 2 complete dataset indicate a significant upward trend in concentration levels for most PAE chemicals over time, while upward significance was also found in some chemicals in the phase 1 and 2 imputed dataset. Furthermore, our imputed dataset and complete dataset consistently suggest that factors such as the utilization of low-VOC building materials or the presence of air conditioning units do not exert a significant impact on chemical concentration levels in homes. Nevertheless, it is critical to note that the skewed nature and small sample size of the home property data pose a potential challenge, which may affect the inferences.

This study is limited by the large number of missing observations in the dataset and the assumption of independence within and between homes. Examining house independency can account for a more complex model for future studies. Moreover, experimenting with additional home attributes, such as the presence of smoking, can assess their significance and determine any effect of home properties on chemical concentration levels. Given the adverse health effects of SVOCs, our study highlights the importance of gathering more data to further understand how these chemicals impact households over time.

References

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412. Special Issue: Emerging Data Analysis.

Carpenter, S., Goldstein, H., and Kenward, M. G. (2011). Realcom-impute software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, 45(5):141–165.

Grund, S., Lüdtke, O., and Robitzsch, A. (1997). Mitml: Tools for multiple imputation in multilevel modeling. *Psychological Methods*.

Grund, S., Lüdtke, O., and Robitzsch, A. (2016). Multiple imputation of multilevel missing data: An introduction to the r package pan. 6(4).

Grund, S., Lüdtke, O., and Robitzsch, A. (2017). The current reference style in this template. *Psychological Methods*, 22(1):141–165.

Jahangiri, M., Kazemnejad, A., and Goldfeld, K. S. (2023). A wide range of missing imputation approaches in longitudinal data: A simulation study and real data analysis. *BMC Medical Research Methodology*, 23(1):205–224.

Mercier, F., Glorennec, P., Thomas, O., and Le Bot, B. (2011). Organic contamination of settled house dust, a review for exposure assessment purposes. *Environmental Science & Technology*, 45(16):6716–6727.

Quartagno, M., Grund, S., and Carpenter, J. (2019). jomo: A flexible package for two-level joint modelling multiple imputation. *The R Journal*, 11(2):205–228.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. *Psychological Methods.*

Xu, Y. and Zhang, J. (2011). Understanding svocs. *ASHRAE Journal*, 53(12):121–125.

Appendix

A Data Overview

A.1 Selected Variables from Technical/Questionnaire Dataset

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)

A.2 Concentration of Chemicals over Time for Homes with Different Technical Specs

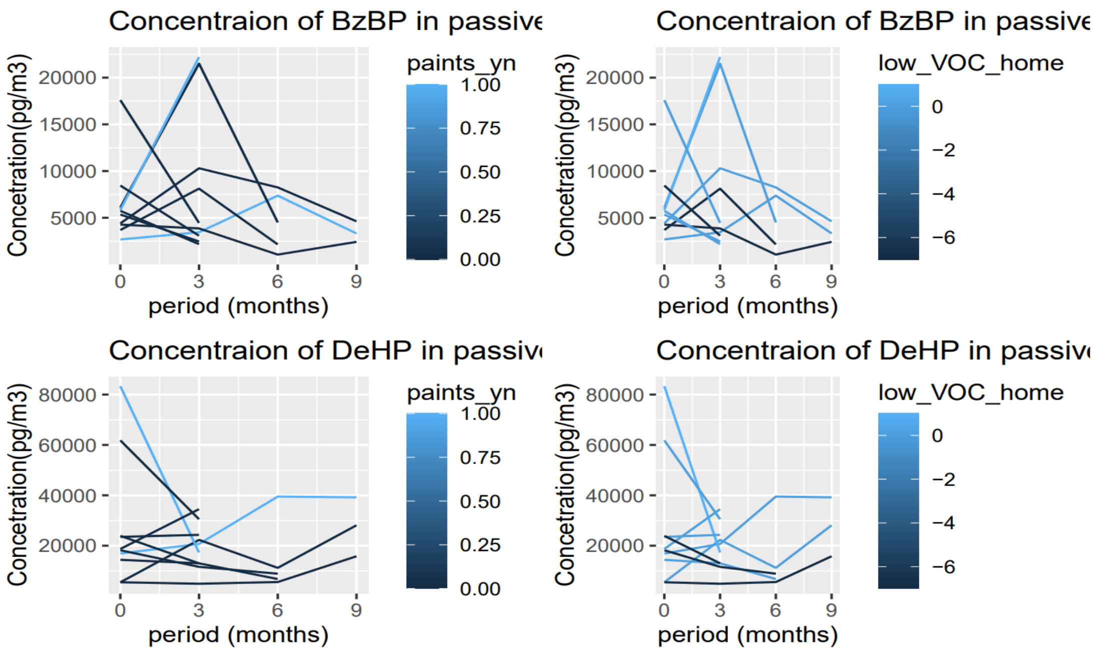


Figure 7: Concentration Levels over Time with Different Paint and Building Materials

A.3 Cross Correlation Plot

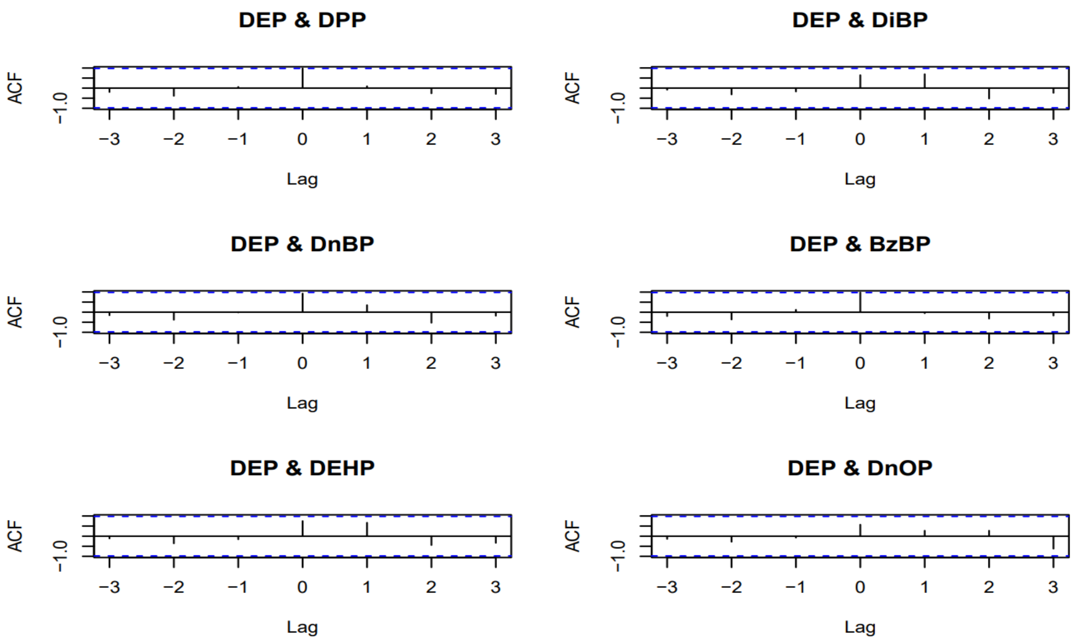


Figure 8: Cross Correlation Plot for Selected PAEs

B Definitions and Technical Details

B.1 Interpolation Method

$$\hat{y} = y_i + (x - x_i)\frac{y_j - y_i}{x_j - x_i}$$

Where:

$$\begin{aligned} \hat{y} &: \text{represents the estimated value to impute.} \\ y_i &: \text{the first known value} \\ y_j &: \text{the last known value before the missing values to be imputed.} \\ x_i &: \text{the corresponding known positions of } y_i \\ x_j &: \text{the corresponding known positions of } y_j \\ x &: \text{the position of the missing value to be imputed.} \end{aligned}$$

B.2 CopyMean Method

$$\hat{y}_i = \frac{1}{n}\sum_{j \neq i}^{n} y_j$$

Where:

$$\begin{aligned} \hat{y}_i &: \text{represents the estimated value to impute.} \\ y_j &: \text{the non-missing value} \end{aligned}$$

B.3 FCS-LMM Imputation Method

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + ... + \beta_p X_{pij} + u_{0i} + u_{1i} X_{1ij} + ... + \epsilon_{ij}$$

where:

$$\begin{aligned} Y_{ij} &\text{ is the response variable for the } i\text{-th subject at the } j\text{-th time point,} \\ X_{kij} &\text{ are the predictors (which could include time, treatment groups, or other covariates),} \\ \beta_k &\text{ are the fixed effects coefficients,} \\ u_{0i} \text{ and } u_{ki} &\text{ represent the random intercepts and slopes for the } i\text{-th subject} \\ \epsilon_{ij} &\text{ is the residual error term.} \end{aligned}$$

The basic idea behind FCS is to impute missing data by specifying a separate model for each variable with missing values, iterating over these variables until the process converges, and the resulting dataset is combined. The LMM part refers to the use of linear mixed models in this process, which are especially useful for accounting for the correlation of different hierarchical data structures in our project such as house_id.

C Model Comparison Result

Table 10: Model Comparison for DPP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
slr_DPP	3	28.47	29.924	-11.235	22.47			
lmm_DPP	4	30.47	32.409	-11.235	22.47	0	1	1.0

Table 11: Model Comparison for DPP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
slr_DPP	3	28.47	29.924	-11.235	22.47			
lmm_DPP	4	30.47	32.409	-11.235	22.47	0	1	1

Table 12: Model Comparison for DiBP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
mod0_DiBP	3	27.885	29.340	-10.943	21.885			
mod1_DiBP	4	29.885	31.825	-10.943	21.885	0	1	1

Table 13: Model Comparison for DnBP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
mod0_DnBP	3	21.865	23.320	-7.9327	15.865			
mod1_DnBP	4	23.865	25.805	-7.9327	15.865	0	1	1

Table 14: Model Comparison for BzBP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
mod0_BzBP	3	26.914	28.369	-10.457	20.914			
mod1_BzBP	4	28.071	30.011	-10.036	20.071	0.8423	1	0.3587

Table 15: Model Comparison for DnOP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
mod0_DnOP	3	41.927	43.381	-17.963	35.927			
mod1_DnOP	4	43.927	45.866	-17.963	35.927	0	1	1

Table 16: Model Comparison for DiNP Levels

Model	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>Chi)
mod0_DiNP	3	45.562	47.016	-19.781	39.562			
mod1_DiNP	4	47.562	49.501	-19.781	39.562	0	1	1

D Imputation results

D.1 Phase 1

Chemical	Model	F.value	df1	df2	P(F)	RIV
DEP	Mode_ac	0.18	1	514.87	0.68	0.15
	Model_low	0.30	2	192.36	0.74	0.24
	Model_paints	0.24	1	1851.74	0.62	0.075
DPP	Model_ac	0.18	1	749.86	0.67	0.12
	Model_low	0.811	2	22.79	0.46	1.31
	Model_paints	1.06	1	799.28	0.30	0.12
DiBP	Model_ac	0.017	1	5648.56	0.90	0.042
	Model_low	1.25	2	36.25	0.30	0.82
	Model_paints	0.14	1	8043.60	0.71	0.035
DnBP	Model_ac	0.074	1	610.23	0.79	0.14
	Model_low	0.12	2	116.05	0.89	0.34
	Model_paints	0.050	1	854.84	0.82	0.114
BzBP	Model_ac	0.21	1	323.76	0.65	0.20
	Model_low	0.18	2	52.99	0.83	0.59
	Model_paints	1.15	1	1279.22	0.28	0.092
DEHP	Model_ac	0.15	1	482.05	0.70	0.16
	Model_low	0.90	2	97.18	0.41	0.38
	Model_paints	1.05	1	158.74	0.31	0.31
DnOP	Model_ac	0.68	1	334.72	0.41	0.20
	Model_low	0.13	2	34.86	0.88	0.85
	Model_paints	0.27	1	296.83	0.60	0.21
DiNP	Model_ac	0.24	1	373.17	0.62	0.18
	Model_low	1.11	2	376.81	0.33	0.16
	Model_paints	0.21	1	761.06	0.65	0.12

Table 17: Phase 1 Pooled Wald Test

D.2 Phase 2

D.2.1 Imputed data

Chemical	Model	F.value	df1	df2	P(F)	RIV
DEP	Model_ac	0.82	1	163.42	0.37	0.31
	Model_low	0.076	2	1550.55	0.93	0.074
	Model_paints	0.12	1	582.11	0.725	0.14
DPP	Model_ac	0.086	1	1166.96	0.77	0.096
	Model_low	0.38	2	850.35	0.68	0.10
	Model_paints	0.071	1	1617.00	0.79	0.081
DiBP	Model_ac	0.28	1	829.59	0.60	0.12
	Model_low	0.63	2	437.27	0.53	0.15
	Model_paints	0.079	1	15773.33	0.78	0.024
DnBP	Model_ac	0.38	1	427.69	0.54	0.17
	Model_low	0.45	2	473.79	0.64	0.14
	Model_paints	0.38	1	261.11	0.54	0.23
BzBP	Model_ac	0.21	1	323.76	0.65	0.20
	Model_low	0.26	2	224.95	0.77	0.22
	Model_paints	0.17	1	626.63	0.68	0.14
DEHP	Model_ac	0.52	1	237.68	0.47	0.24
	Model_low	1.02	2	66.13	0.37	0.50
	Model_paints	1.32	1	160.12	0.25	0.31
DnOP	Model_ac	0.68	1	334.72	0.41	0.20
	Model_low	0.208	2	442.49	0.812	0.147
	Model_paints	0.51	1	344.94	0.48	0.19
DiNP	Model_ac	0.37	1	363.88	0.54	0.19
	Model_low	0.36	2	34.96	0.70	0.84
	Model_paints	0.37	1	363.88	0.542	0.187

Table 18: Phase 2 Pooled Wald tests

D.2.2 Complete data Likelihood ratio tests

Chemical	Model	Df	LogLik	Df	Chisq	Pr(>Chisq)
DEP	Model_null	4	-106.58			
	Model_ac	5	-104.17	1	4.82	0.028
	Model_low	6	-104.29	2	4.57	0.10
	Model_paints	5	-105.38	1	2.40	0.12
DPP	Model_null	4	-73.360			
	Model_ac	5	-73.174	1	0.37	0.54
	Model_low	6	-73.95	2	1.18	0.55
	Model_paints	5	-73.16	1	0.41	0.52
DiBP	Model_null	4	-128.86			
	Model_ac	5	-124.26	1	9.20	0.0024
	Model_low	6	-125.61	2	6.50	0.039
	Model_paints	5	-127.52	1	2.68	0.10
	Model_lowac	7	-120.12	3	17.48	0.00056
DnBP	Model_null	4	-122.99			
	Model_ac	5	-122.01	1	1.96	0.16
	Model_low	6	-121.00	2	3.9762	0.137
	Model_paints	5	-121.99	1	1.99	0.16
BzBP	Model_null	4	-99.38			
	Model_ac	5	-98.79	1	1.18	0.28
	Model_low	6	-98.60	2	1.55	0.46
	Model_paints	5	-98.50	1	1.75	0.19
DEHP	Model_null	4	-88.98			
	Model_ac	5	-88.01	1	1.95	0.16
	Model_low	6	-87.45	2	3.07	0.23
	Model_paints	5	-88.61	1	0.75	0.39
DnOP	Model_null	4	-121.67			
	Model_ac	5	-119.93	1	3.49	0.062
	Model_low	6	-119.44	2	4.46	0.11
	Model_paints	5	-120.62	1	2.10	0.15
DiNP	Model_null	4	-104.74			
	Model_ac	5	-103.59	1	2.31	0.13
	Model_low	6	-103.61	2	2.26	0.32
	Model_paints	5	-103.89	1	1.69	0.19

Table 19: Likelihood ratio test of models in Phase 2 complete case data