

Simulation of longitudinal exposure data with variance-covariance structures based on mixed models

PENG SONG^a, JIANPING XUE^{b*}, ZHILIN LI^c

^a Operations Research Program, North Carolina State University, Raleigh, North Carolina, USA

^b Corresponding Author: National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, North Carolina, USA

^c Center for Research in Scientific Computation & Department of Mathematics, North Carolina State University, Raleigh, North Carolina, USA

* Corresponding Author: mailing address U.S. EPA, 109 T.W. Alexander Drive, MD E205-02, Research Triangle Park, NC 27711

Abstract: Longitudinal data are important in exposure and risk assessments, especially for pollutants with long half-lives in the human body and where chronic exposures to current levels in the environment raise concerns for human health effects. It is usually difficult and expensive to obtain large longitudinal data sets for human exposure studies. This paper reports a new simulation method to generate longitudinal data with flexible numbers of subjects and days. Mixed models are used to describe the variance-covariance structures of input longitudinal data. Based on estimated model parameters, simulation data are generated with similar statistical characteristics compared to the input data. Three criteria are used to determine similarity: the overall mean and standard deviation, the variance components percentages, and the average autocorrelation coefficients. Upon the discussion of mixed models, a simulation procedure is produced and numerical results are shown through one human exposure study. Simulations of three sets of exposure data successfully meet above criteria. In particular, simulations can always retain correct weights of inter- and intra- subject variances as in the input data. Autocorrelations are also well followed. Compared with other simulation algorithms, this new method stores more information about the input overall distribution so as to satisfy the above multiple criteria for statistical targets. In addition, it generates values from numerous data sources and simulates continuous observed variables better than current data methods. This new method also provides flexible options in both modeling and simulation procedures according to various user requirements.

Key words: longitudinal data, simulation, mixed models, variance-covariance structure, autocorrelation

1. INTRODUCTION

Longitudinal data on the intensity of time-varying sources of exposures are extremely important for epidemiological studies, environmental exposure modeling and risk assessment since they possess variance and covariance structures that cross-sectional data lack. However, it is very difficult and expensive to track human activities, environmental measurements and other information for the same subject through extended periods of time to obtain observed longitudinal data. Some studies do provide longitudinal data but with small sample size and short duration, such as the Harvard Southern California Ozone Exposure Study ⁽¹⁾, PM2.5 Panel Studies ⁽²⁾ and the Detroit Exposure and Aerosol Research Study ⁽³⁾. This lack of data restricts the applicability of observed longitudinal data for studies of broader spatial or temporal scope.

Computer simulations could help overcome the above limitation. In this paper, we first build statistical models for available observed longitudinal data, and then develop an algorithm to generate large amounts of longitudinal data with flexible numbers of exposure days and close overall distribution with the observed data. Three criteria are used to evaluate this simulation method. First, the overall mean and standard deviation (SD) should be close to those from observed data. Second, the variances due to factors such as inter-, intra- subject, seasonal and other characteristics of the simulated data should be similar to the corresponding variances of the observed data. Third, autocorrelations should be consistent between the simulated and observed data. When these three criteria are met, the generated simulation data can avoid misclassification of variance components and be used for various models, such as SHEDS-Multimedia

(Stochastic Human Exposure and Dose Simulation Model for Multimedia, Multipathway Pollutants)^(4,5,6,7,8).

The challenge of above approach lies in how to model and simulate the complicated variance-covariance structures of observed input data. In longitudinal data, there is variation among subjects. Moreover, there are both variation and correlation within each subject. Traditional methods usually generate all data independently, so that the correlations within subjects are ignored. In statistics, mixed models are developed to make this issue clear. In mixed models, total data variance is divided into that between subjects (inter-subject) and that within subjects (intra-subject). Then, we can model several types of correlations within each subject as necessary, in order to accurately simulate the variance-covariance structures in the observed data. In this paper, we will present the use of mixed methods, and the simulation procedure based on them. Then we evaluate this new method with some observed exposure data and provide simulation results. We also compare this new method with other simulation methods to demonstrate its features.

2. MIXED MODELS AND SIMULATION METHODS

2.1. Basic Mixed Models

Traditional simulation methods for longitudinal data are usually based on the following model under assumption of independence:

$$y_{ij} = \mu + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2) \quad (1)$$

where y_{ij} is the j th observation of the i th subject; μ is the mean of all observations; ε_{ij} is the random term including variation among subjects, variation over time, and variation due to measurement error. In this model, random terms are assumed independent and normally distributed among and within subjects. The normal distribution issue will be discussed later in this section.

As a modified version of (1), mixed model separates inter-subject and intra-subject variances, by splitting ε_{ij} into two terms:

$$y_{ij} = \mu + b_i + e_{ij}, \quad b_i \sim N(0, \sigma_b^2), \quad e_{ij} \sim N(0, \sigma_e^2) \quad (2)$$

where b_i is the random effect of subject i , and e_{ij} is the random term for other variation in its j th observation. Here, b_i 's are assumed to be independent among subjects, and e_{ij} 's are assumed to be independent among and within subjects. In this way, observations from subject i share a common term b_i to retain their correlation, and meanwhile possess their own random terms e_{ij} 's to quantify intra-subject variability. Take an example in which every subject has observations for four consecutive days. Under the assumptions in (2), the variance-covariance matrix of one's 4-day series (4×1 random vector) is:

CS

$$\begin{aligned} \Sigma_{4 \times 4} &= \text{Var} \begin{pmatrix} b_i + e_{i1} \\ b_i + e_{i2} \\ b_i + e_{i3} \\ b_i + e_{i4} \end{pmatrix} = \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 \end{pmatrix} \\ &= (\sigma_b^2 + \sigma_e^2) \begin{pmatrix} 1 & \omega & \omega & \omega \\ \omega & 1 & \omega & \omega \\ \omega & \omega & 1 & \omega \\ \omega & \omega & \omega & 1 \end{pmatrix} \end{aligned}$$

(3)

where

$$\omega = \sigma_b^2 / (\sigma_b^2 + \sigma_e^2) \quad (3a)$$

is the ICC (intra-class correlation coefficient) in statistics ⁽⁹⁾ and exposure studies ⁽¹⁰⁾.

The SAS procedures PROC MIXED or PROC GLM can provide estimates of the parameters μ , σ_b^2 , σ_e^2 in model (2). ⁽¹¹⁾ To simulate an n -day series for subject i , b_i is generated first for the subject. After that, e_{i1} , ..., e_{in} are generated independently and added to b_i for day 1 through day n . This two-stage simulation procedure maintains both variation and correlation within each subject, as well as the variation among subjects. Finally, we can add the constant μ to every $b_i + e_{ij}$ to meet the overall mean and obtain the simulation longitudinal data set.

2.2. Mixed Models with Autocorrelation

In the basic mixed model (2) and its variance-covariance matrix (3), random terms e_{i1} , ..., e_{in} are assumed to be independent. This means any two sets of days from one subject have equal ICC of $\sigma_b^2 / (\sigma_b^2 + \sigma_e^2)$. However, sometimes this is not true. For example, some high level exposures tend to occur consecutively. If so, data from two closer days are likely to have higher correlations. This property is called autocorrelation in longitudinal data. As a result, random term e_{ij} in model (2) is no longer entirely independent with others. Instead, it partially depends on its preceding term $e_{i,j-1}$. For this,

we further separate e_{ij} into two terms: one is determined by $e_{i,j-1}$, and the other is independent from all:

$$e_{ij} = \rho e_{i,j-1} + s_{ij} \quad (4a)$$

where ρ is the autocorrelation coefficient between two consecutive days, or **lag-one autocorrelation**, $-1 < \rho < 1$. Then model (2) is modified to the following:

$$y_{ij} = \mu + b_i + \rho e_{i,j-1} + s_{ij}, \quad s_{ij} \sim N(0, (1 - \rho^2) \sigma_e^2), \quad (4)$$

where s_{ij} is independent from all. Its re-scaled variance $(1 - \rho^2) \sigma_e^2$ is to keep total variance within one subject equal to σ_e^2 . Under this assumption, the variance-covariance matrix becomes ⁽¹¹⁾:

$$\Sigma_{4 \times 4} = Var \begin{pmatrix} b_i + e_{i1} \\ b_i + \rho e_{i1} + s_{i2} \\ b_i + \rho e_{i2} + s_{i3} \\ b_i + \rho e_{i3} + s_{i4} \end{pmatrix} = \sigma_b^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \quad (5)$$

Among the two matrices in (5), the first one defines **inter-subject** variances equivalently as in (3), and the second one allows observations with a k -day lag to have a correlation coefficient of ρ^k . This is closer to reality in some cases, compared with covariance matrix (3) which requires observations from any two days have equal correlations.

For our simulations, we need to estimate parameter ρ in (4), besides σ_b^2 , σ_e^2 in (2). For input data over short time period, SAS PROC MIXED can estimate all parameters directly by the maximum likelihood method. For input data over long time periods, say, more than 10 days, that algorithm can fail **due to large computation** loads. Instead, we can

use PROC GLM first to get estimates of μ , σ_b^2 , σ_e^2 and save all residuals e_{ij} . PROC

ARIMA can be used on residuals e_{i1}, \dots, e_{in} to calculate the autocorrelation in subject i .

The average of all subjects' autocorrelation coefficients is a reasonable estimate for ρ in the population.

2.3. Mixed Models with Classification Variables

The mixed models discussed above are for longitudinal data with only subject and time information. Oftentimes, some classifications for subjects and observation periods are desired. For example, subjects can be classified by groups according to gender, age, or living district. Similarly, observation days can be classified by treatments applied in various month, season, or different categories. It could be important to study how much variance is attributed to these classification effects. In mixed models, these classifications are modeled as fixed effects, distinct from random effects such as b_i . Suppose a subject k belongs to the group i , and its l th observation is taken under treatment j (for example, the j th season). Then, this observation, labeled as y_{ijkl} , can be modeled as following:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_{k(i)} + \rho e_{ijk,l-1} + s_{ijkl}, \quad (6a)$$

or equivalently,

$$y_{ijkl} = \mu_{ij} + b_{k(i)} + \rho e_{ijk,l-1} + s_{ijkl}, \quad (6)$$

where $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ is the mean of all data from group i and under treatment j ;

$b_{k(i)}$ is the random effect of the subject k from group i ; $\rho e_{ijk,l-1} + s_{ijkl}$ is the random term of

that observation. This model is referred as the Split-Plot model in statistics ⁽¹²⁾.

Estimates of μ_{ij} can be obtained using SAS PROC MIXED or PROC GLM. For our simulations, we first build variance structures $b_{k(i)} + \rho e_{ijk,l-1} + s_{ijkl}$ as described in the above sections. Then, we assign classification levels (for example, four seasons, two genders) according to their frequency percentages in the input data. Finally, we add up the corresponding μ_{ij} by assigned classification levels to build the fixed effects. If the input data set contains classification structures not as typical as in model (6), we can always convert it to satisfy (6). When there is no group classification in input data, we can simply establish a dummy variable (all zero values) as a virtual group effect ($\alpha_i = 0$). If there are two treatment effects, we can incorporate their information into one virtual effect to fit the model (6). For example, suppose we are not only interested in which season the observations belong to (4 levels treatment effect), but also interested in whether they are observed on a weekday or weekend (2 levels treatment effect). We can combine both of them into one virtual treatment effect with $2 \times 4 = 8$ levels, and then send it to the simulation module as β_j in model (6). After simulation data are generated, we can separate this virtual effect back to the original season effect and weekday/weekend effect according to the combining rules. By repeating this combine-and-separate procedure, we can even handle more complicated classification information in input data. This technique will considerably broaden the applicability of model (6), without any modification on the core simulation program.

2.4. Transforms on Input Data

In the mixed models above, all random terms b_i and e_{ij} are assumed normally distributed. This is the basic assumption of mixed model theory, as well as the requirement of the SAS procedures we used. That means input data y_{ij} or y_{ijkl} also roughly follow normal distributions. If they are not normally distributed, we can apply some mathematical transforms on input variables y_{ij} or y_{ijkl} to re-scale them. For example, some exposure data empirically follow a log-normal distribution. Thus, we can take logarithms of input data, and then build model (6) on the transformed data:

$$\log(y_{ijkl}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_{k(i)} + \rho e_{ijk,l-1} + s_{ijkl}, \quad (7)$$

However, sometimes we find that after the logarithm transform, residuals $b_{k(i)}$ or e_{ijkl} still fail to fit a normal distribution. That may cause some biased simulation results, such as a higher or lower overall standard deviation than the target level of input data. One alternative is to use a broader family of transforms, called the Box-Cox transforms⁽¹³⁾, of which the logarithm transform is only one special case:

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \quad (8)$$

When λ approaches to zero, the Box-Cox transform is almost the logarithm transform, since $\lim_{\lambda \rightarrow 0} \frac{y^\lambda - 1}{\lambda} = \log(y)$. When $\lambda = 1$, the Box-Cox transform just shifts all data down by one unit, without changing the variance-covariance structures of input data. With (8) implemented, model (7) can be generalized to the following:

$$y_{ijkl}^{(\lambda)} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_{k(i)} + \rho e_{ijk,l-1} + s_{ijkl}, \quad (9)$$

The optimal Box-Cox transform can be automatically completed by SAS PROC TRANSREG. This module goes through all possible λ values (for example, from -3 to 3 by 0.1) and evaluates each likelihood for model (9). Then, the λ value with maximum likelihood is selected. In this way, this “smart” Box-Cox transform can determine whether the input data y_{ijkl} already meet the normal assumption ($\lambda=1$ selected), or whether the logarithm should be taken ($\lambda=0$ selected), or another transform ($\lambda \neq 0, 1$ selected) should be applied.

Although the ~~Box-Cox~~ transform can make input data closer to a normal distribution, there are some cases it cannot help with, such as heavy tails or outliers in input data. In these cases, transformed data may not conform to the assumptions of mixed models, resulting in certain bias in the simulation.

2.5. Simulation Procedure

We now formulate a complete procedure to simulate longitudinal data based on mixed model (9) by SAS:

✓ Step 1: Observed input data. Compute target statistics: overall mean and standard deviation, variance components percentages, average lag-one autocorrelation coefficient.

✓ Step 2: If input data have more classifications than model (6), combine them into one group effect and one treatment effect as in (6), and then keep the percentages of all levels of groups and treatments.

✓ Step 3: Find the optimal ~~Box-Cox~~ transform as in (9) by SAS PROC TRANSREG.

Step 4: Estimate all model parameters in (9) by SAS PROC GLM and PROC MIXED.

Step 5: Input required numbers of subjects and days in simulation data, generate that amount of $b_{k(i)}$, $\rho e_{ijk,l-1}$, s_{ijkl} from corresponding normal distributions, organize and sum up as in (2) and (4).

Step 6: Assign group and treatment levels as their percentages in Step 2. Add up proper means $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ for data in each level.

Step 7: Transform obtained simulation data $y_{ijkl}^{(\lambda)}$ back to original scale y_{ijkl} , as inverse of Step 3.

Step 8: Restore group and treatment effects as in input data, as inverse of Step 2.

Step 9: Check simulation results. Compute three aspects of statistics from simulation data and compare with the targets set up in Step 1.

Figure 1 is a flow chart of this procedure from step 1 to step 9:

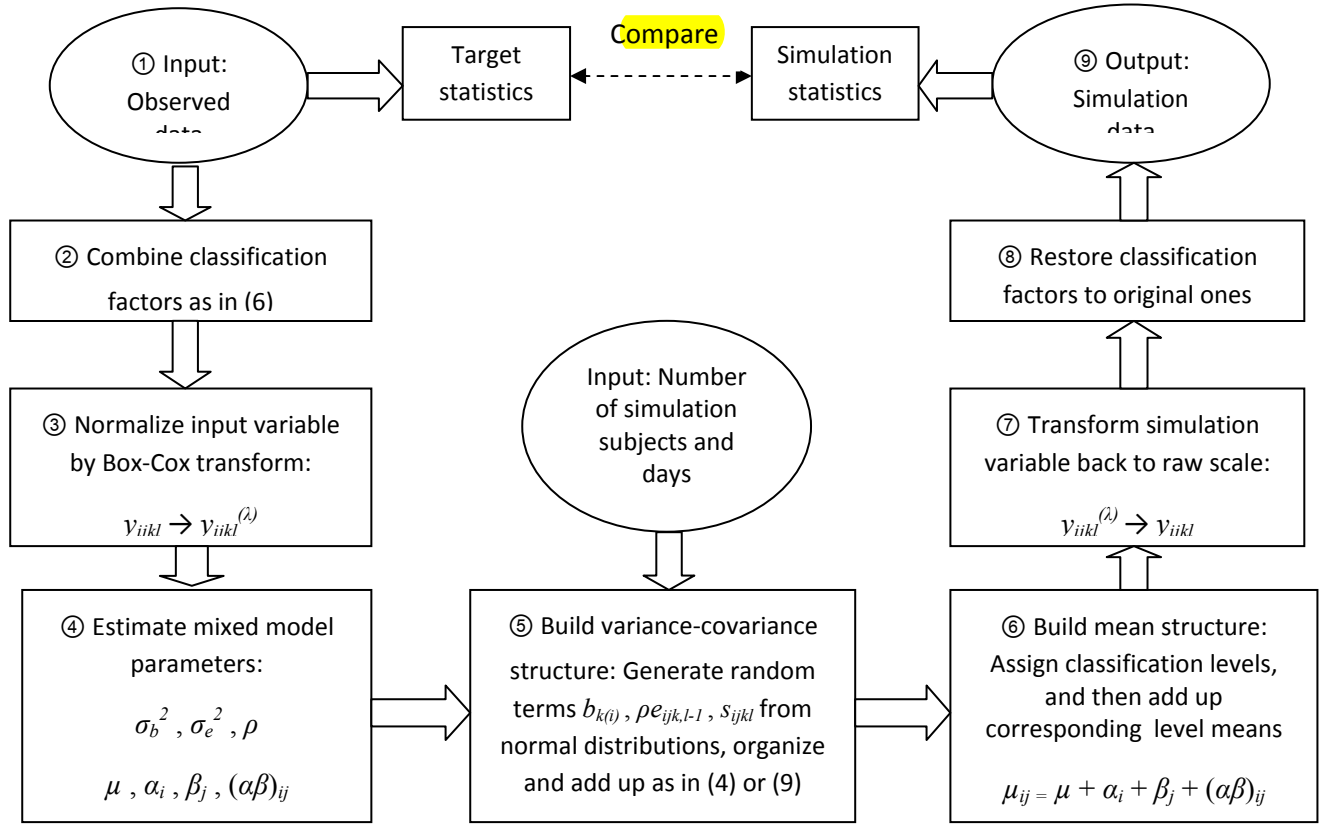


Figure 1. Flow chart of simulation of longitudinal data by the complete mixed model (9). The left side

①~④ include modeling steps, and the right side ⑤~⑨ include simulation steps.

3. RESULTS

The PM_{2.5} (particulate matter less than 2.5 micrometers in diameter) Panel Studies⁽²⁾ took observations on personal, indoor and outdoor PM exposure data and other variables of interest from 37 participants over four seasons from June 2000 to June 2001. The involved subjects came from two socioeconomic cohorts living in Raleigh and Chapel Hill, respectively, both in North Carolina. Each subject was expected to be monitored on seven consecutive days in each season. Due to missing data, there are 23 observations per subject on average.

We conducted simulation experiments on three input variables: **personal PM, indoor PM, and outdoor PM**. We present results of outdoor PM as an example to test four models as described above in a tiered order: the independence model (1), the basic mixed model (2), the mixed model with lag one autocorrelation (AR(1)) as in (4), and the complete model using Box-Cox transformed data as in (9). For a better comparison with (9), we took **logarithms on input data for** the first three models.

$$\log(y_{ijkl}) = \mu_{ij} + \varepsilon_{ijkl} \quad (10a)$$

$$\log(y_{ijkl}) = \mu_{ij} + b_{k(i)} + e_{ijkl} \quad (10b)$$

$$\log(y_{ijkl}) = \mu_{ij} + b_{k(i)} + \rho e_{ijk,l-1} + s_{ijkl} \quad (10c)$$

better ↓

$$y_{ijkl}^{(\lambda)} = \mu_{ij} + b_{k(i)} + \rho e_{ijk,l-1} + s_{ijkl} \quad (10d)$$

In Table I, the first row in bold labeled “Observed” shows target statistics associated with the input data. Below that, the simulation results from the four models are presented in order. The shaded vertical comparisons show how certain target statistics are improved by upgrading the above model to the one below. From the independence model, we see that the inter-subject variance percentage is almost diminished to zero, whereas the intra-subject variance percentage is much higher than its target. When the basic mixed model is used, these two variance components immediately get closer to their target percentages. These two are discussed more at the end of this section. When the mixed model is further modified with autocorrelation, we see that the autocorrelation coefficient is raised to 0.37, comparable to the observed value of 0.39. Finally, when we improve the logarithm transform to the optimal Box-Cox transform, the simulation overall standard deviation is slightly adjusted from 11.0 to 9.5.

Table I. Simulation results of outdoor PM by four models (10a)-(10d) with 1000 subjects over 300 days, compared with target statistics set by actual observed data. Shaded vertical comparisons show how simulation results are improved by each model refinement. In row of (10b), variance components inter- and intra- percentages are corrected when mixed model is used. In row of (10c), observed autocorrelation coefficient is approached when AR(1) is added to model. In row of (10d), overall standard deviation (SD) is adjusted when Box-Cox transform is used to replace logarithm transform.

Method	Overall Scale ($\mu\text{g}/\text{m}^3$)		Variance Components Percentages (%)				AR(1) ρ
	Mean	SD	Inter	Intra	Cohort	Season	
Observed	20.0	9.5	13.5	75.9	0.5	8.9	0.39
(10a) Independent	20.2	11.0	0.3	89.7	0.4	9.6	0.10
(10b) Mixed	20.3	11.1	8.1	82.4	0.5	9.0	0.10
(10c) Mixed + AR(1)	20.2	11.0	8.1	82.2	0.3	9.4	0.37
(10d) Mixed + AR(1) + Box	20.0	9.5	10.2	79.5	0.4	9.9	0.40

Parallel simulations for personal PM and indoor PM gave similar results as in Table I: simulations by the last model “Mixed + AR(1) + Box” work well to approach observed targets in all aspects. When fixed effects such as cohort and season take very few percentages in observed data (less than 1%), they can be considered insignificant and omitted in the simulation.

We also studied the trend of autocorrelation for increasing simulation time periods in Figure 2. The observed outdoor PM data contains much larger autocorrelation ($\rho = 0.40$) than the observed personal PM data ($\rho = 0.08$). From Figure 2, both simulation

autocorrelations start from a low level, increase gradually with time, and eventually become stable. For outdoor PM, the simulation autocorrelation converges to its observed level accurately. For personal PM, the simulation autocorrelation is a little higher than its observed value, which is in fact very weak. If the observed autocorrelation can be inferred insignificant, we can simply force ρ to be zero and apply the model without autocorrelation.

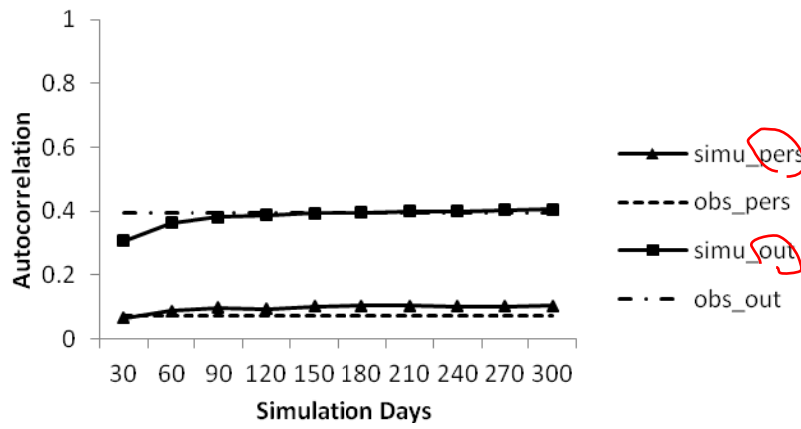


Figure 2. Simulation on autocorrelation of personal PM (labeled simu_pers) and outdoor PM (labeled simu_out) for increasing time lengths with 1000 subjects, compared with their observed autocorrelations.

In Figure 3, we ran the simulation on personal PM for increasing days to explore the trends of inter-subject and intra-subject variance percentages. Simulation results of the independence model (1) are also shown for comparison. The independence model reduces inter-subject variance towards zero as the simulation period is lengthened, whereas the mixed model maintains inter-subject variance close to its observed target. Accordingly, the target intra-subject variance is also well approached by mixed model. Another observation is that simulation inter-subject percentage tends to decrease and then

get stabilized, while intra-subject percentage goes in the opposite direction. This is because inter- and intra- subject variances percentages are also affected by data length, especially in short term data (less than 30 days). We will examine this issue further below.

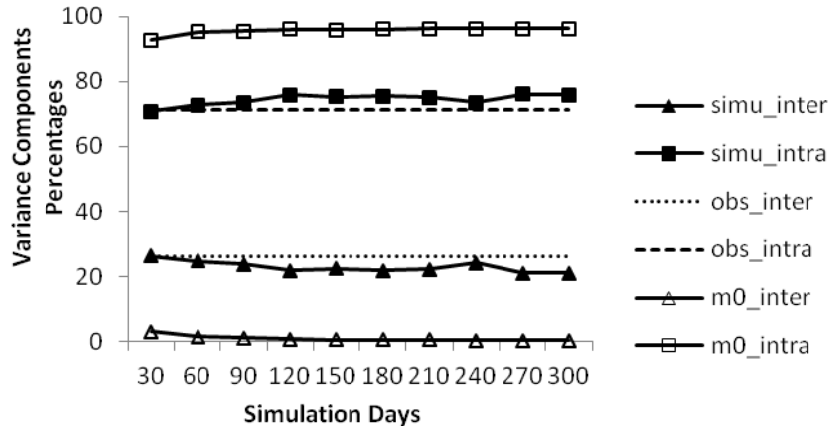


Figure 3. Simulation of personal PM inter- and intra- subject variances percentages with 1000 subjects over increasing time lengths by mixed model (labeled “simu_”) and independence model (labeled “m0_”), compared with their observed target percentages (labeled “obs_”).

For short term longitudinal data, Glen *et al.* ⁽¹⁰⁾ proposed a relationship between true (long run) inter-subject variance percentage D , and its observed value D^* from a sample of length N : $D^* = D + (1 - D) / N$. From this equation, a short term sample tends to give an overestimated inter-subject variability. It is recommended to correct the observed inter-subject variance percentage by $D = (ND^* - 1) / (N - 1)$ for a better target, instead of being very sensitive to data length. The intra-subject variance percentage will be corrected accordingly (shifted up a little bit), since all other fixed effects such as season and cohort always keep their percentages in total variance when data is lengthened. In

Table II, we add the corrected targets in parentheses after observed targets for all three variables and compare simulation results with them. When simulations are run based on entire observed data, the inter- and intra- results successfully approach the corrected target values. Moreover, it is interesting to test how the simulations perform if fewer sample data are available. We use half of the observed data (last two seasons, 13 days) to run our simulations, and find that simulation results on inter- and intra- subject variances percentages have larger relative errors to the corrected targets. Results for the other targets are similar to Table I. Simulations based on even shorter samples (less than 10 days) are not recommended because large biases in the inter- and intra- variances percentages would appear. These results also agree with a previous study ⁽¹⁴⁾ which reported that at least ten days of observations are needed to capture a reliable ICC.

Table II. Simulation inter- and intra- subject variances percentages compared with observed targets and corrected targets (in parentheses) for three input variables. Simulation relative errors to corrected targets are also provided below. The top results are from a 300-day simulation using the entire input observations (23 days); the bottom results are from a 150-day simulation using half of input observations (winter and spring, 13 days). Each simulation includes 1000 subjects.

Data	PM_Personal		PM_Indoor		PM_Outdoor	
	Inter %	Intra %	Inter %	Intra %	Inter %	Intra %
Entire Obs. (23d)	26.1 (23.0)	71.6 (74.7)	29.9 (27.0)	66.5 (69.3)	13.5 (10.3)	75.9 (79.1)
Simulation (300d)	21.6	75.8	26.9	69.6	10.2	79.5
Relative Error (%)	6.3	1.5	0.4	0.4	1.0	0.5
Half Obs. (13d)	26.7(21.4)	67.2(72.5)	30.1 (25.2)	62.6 (67.5)	15.0 (9.6)	67.6 (72.9)

Simulation (150d)	16.6	77.8	26.5	64.8	9.3	73.5
Relative Error (%)	25.3	7.1	5.0	4.1	3.2	0.8

4. DISCUSSION

This paper reports a new simulation method for longitudinal data. A series of mixed models are applied to describe variance-covariance structures of input longitudinal data. As outlined in Figure 1, input longitudinal data are analyzed first to estimate model parameters, and then these parameters are used to generate output longitudinal data with a distribution closely following the input data. The output distribution is checked from aspects of overall mean, standard deviation, variance components percentages, and autocorrelation. Three data sets from the PM2.5 Panel Studies are used to test this new method. Most simulation experiments yield accurate and robust results in approaching input data targets.

Compared with other simulation methods, this new method has the following features.

The first feature is the crucial role of mixed models. Most of our efforts were focused on model refinement to describe the input data accurately and comprehensively. If one model can fit input longitudinal data very well, simulations using that model should produce good results. From this point, this simulation method first serves as a data modeler and then as a data generator. Through the modeling process, this method stores a lot of sample information, so that it can simultaneously satisfy several targets to closely replicate the input distribution. In contrast, other simulation methods are mostly designed from only one or two statistical aspects. There are advantages and disadvantages of other

methods and ours. Our method has the relatively stricter requirements on the input data imposed by mixed models.

The second feature is the simulation data source. Current simulation methods usually sample data from available data pools either by random draw methods or more sophisticated drawing algorithms such as Glen's method ⁽¹⁰⁾. That means simulation data can only take limited values existing in pools. When the simulation size is much larger than the pool size, sampling methods can result in forced repetitions of limited available values, and make the simulation data appear discrete. However, observed exposure data are usually continuously distributed. Our new method can emulate this property, because it starts by generating random numbers from the standard normal distribution, which is actually an infinite data pool. In long simulations, new values will always be produced in the output data set, and make simulation data closer to a continuous distribution.

The third feature is the flexibility in simulation practice. This method has been coded in SAS Macro language. The user only needs to input a longitudinal data set, and specify how many subjects and days to generate. Then, both modeling and simulation steps in Figure 1 run automatically until a simulation data set is output and a comparison table like Table I is displayed. The user has options to specify which variable to simulate, what classification factors to involve, whether the autocorrelation is considered or not, and what kind of transforms to apply on input variable. The user can also simulate a subset of input data with particular properties if necessary.

There are certain requirements to apply the technique to assemble the longitudinal data such as continuous measurements, normal distributions required by mixed model, and

estimated variance-covariance structures. It is important to take co-occurrence into consideration when the technique is used in modeling cumulative exposures for multiple chemicals.

There are possible extensions for our simulation method in future studies. First, some generalized linear mixed model tools have been recently developed for response variables that are not normally distributed. Using these, we could potentially fit the observed data more accurately and obtain better simulation results. Second, continuous variables could be added into current models as covariates, such as air exchange rates that affect indoor pollution levels, in addition to the classification variables already included. Third, related input variables could be simultaneously modeled as a group to maintain inherent correlations among them. These extensions would generalize our method significantly and are well within current-day practices.

5. CONCLUSIONS

The new technique presented in this paper uses variance-covariance structure and autocorrelation coefficients from limited longitudinal data to simulate unlimited longitudinal data. Inter- and intra- personal variances and autocorrelation are close to the observed longitudinal data.

The new method will be important for exposure models such as EPA's SHEDS-Multimedia, since it can be used to simulate a series of important input variables by keeping their variance-covariance structure with more accurate prediction of the variance

and high percentiles of exposure output by the models. This could be valuable for linking environmental pollutants with chronic adverse health effects.

DISCLAIMER

This article has been subject to review and approved for publication by the Office of Research and Development, the United States Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ACKNOWLEDGMENTS

This work was funded by the US Environmental Protection Agency under contract EP09D000645. The third author is supported partially by ARO grants 550694-MA, AFSOR grant FA9550-12-1-0188, NSF grant DMS-0911434, and the NIH grant 096195-01. We gratefully acknowledge the careful manuscript reviews provided by Valerie Zartarian, Andrew Geller, Thomas McCurdy and Kristin Isaacs from US EPA. We also benefited from the following individuals for their useful guides and kind help on this study: Huixia Wang, Jason Osborne, Weining Shen and Dehan Kong, all from Department of Statistics in North Carolina State University.

REFERENCES

1. Geyh AS, Xue J, Özkaynak H, Spengler JD. The Harvard Southern California Chronic Ozone Exposure study: assessing ozone exposure of grade-school-age children in two southern California communities. *Environmental Health Perspectives*, 2000; 108: 265-270.
2. Wallace L, Williams R, Rea A, Croghan C. Continuous weeklong measurements of personal exposures and indoor concentrations of fine particles for 37 health-impaired North Carolina residents for up to four seasons. *Atmospheric Environment*, 2006; 40: 399-414.
3. Williams R, Rea A, Vette A, Croghan C, Whitaker D, et al. The design and field implementation of the Detroit Exposure and Aerosol Research Study. *Journal of Exposure Science and Environmental Epidemiology*, 2009; 19: 643-659
4. US EPA SHEDS-Multimedia. Available at http://www.epa.gov/heads/products/sheds_multimedia/sheds_mm.html
5. Zartarian V., Xue J., Glen G., Smith L., Tolve N., Tornero-Velez R. Accepted. Application and Evaluation of EPA's SHEDS-Multimedia Model to an Aggregate Permethrin Exposure Case Study. *Journal of Exposure Science and Environmental Epidemiology*.
6. Zartarian VG, Xue J, Özkaynak H, Dang W, Glen G, Smith L, Stallings C. A Probabilistic Arsenic Exposure Assessment for Children Who Contact CCA-Treated Playsets and Decks, Part 1: Model Methodology, Variability Results, and Model Evaluation. *Risk Analysis*, 2006; 26: 515-531.
7. Xue J, Zartarian VG, Özkaynak H, Dang W, Glen G, Smith L, Stallings C. A Probabilistic Arsenic Exposure Assessment for Children Who Contact CCA-Treated Playsets and Decks, Part 2: Sensitivity and Uncertainty Analyses. *Risk Analysis*, 2006; 26(2): 533-541.
8. Xue J, Zartarian VG, Wang SW, Liu SV, Georgopoulos P. Probabilistic modeling of dietary arsenic exposure and dose and evaluation with 2003-2004 NHANES Data. *Environmental Health Perspectives*, 2010; 118(3): 345-350.
9. Koch GG. Intraclass correlation coefficient. In Samuel Kotz and Norman L. Johnson. *Encyclopedia of Statistical Sciences*. 4. New York: John Wiley & Sons, 1982.

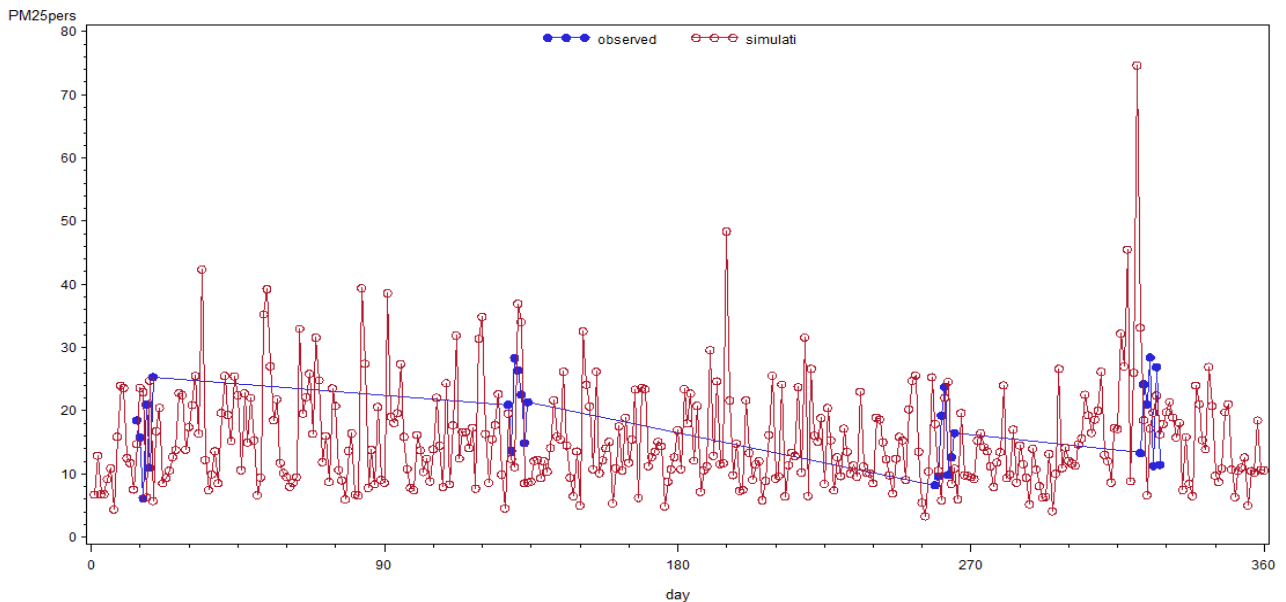
10. Glen G, Smith L, Isaacs K, Mccurdy T, Langstaff J. A new method of longitudinal diary assembly for human exposure modeling. *Journal of Exposure Science and Environmental Epidemiology*, 2008; 18: 299-311.
11. Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O. *SAS for Mixed Models*, 2nd ed. Cary: SAS Institute, 2006.
12. Rao PV. *Statistical Research Methods in the Life Sciences*. Pacific Grove, CA: Brooks/Cole Pub.CO., 1998.
13. Box G, Cox D. An analysis of transformations. *Journal of the Royal Statistical Society*, 1964; Series B 26 (2): 211–252.
14. Xue J, McCurdy T, Spengler J, and Özkaynak H. Understanding variability in the time spent in selected locations for 7–12 year old children. *Journal of Exposure Science and Environmental Epidemiology*, 2004; 14: 222–233.

Appendix A

We provide a figure about daily personal PM2.5 profiles of one given individual from observed data (28 days) and simulated data (365 days).

Figure A1.

Example of daily personal pm2.5 profiles with one observed and one simulated



Appendix B

We provide some SAS codes for core simulation steps 4-6 as in Figure 1. We used the version of SAS 9.2 TS Level 1M0.

Suppose we are going to simulate one variable from the input longitudinal data set. We run the following macro to estimate the model parameters as in step 4.

```
/******  
Function: Estimate key model parameters for variance and means;  
Input : sample: objective_sample  
        y: interested variable;  
Output: sigma_b, sigma_e, rho: parameters defined in mixed model  
        means: data set to keep classification means, i.e., means of  
              each group*trt classification;  
*****/  
%macro model_parameters (sample, y);  
%global sigma_b sigma_e rho;  
title 'Estimate Mean and ANOVA parameters';  
proc mixed data = &sample;  
  class group trt subject;  
  model &y = group|trt / s;  
  random subject(group);  
%mend;
```

```

lsmeans group*trt;
ods output covparms = sigma lsmeans = means;
run;
data sigma_2 (keep = sigma_b sigma_e);
array a(2) sigma_b sigma_e;
do _N_ = 1 to 2;
    set sigma;
    a(_N_) = estimate;
end;
run;
data _n_;
    set sigma_2;
    call symput ('sigma_b', sigma_b);
    call symput ('sigma_e', sigma_e);
run;
title 'Estimate AR(1) on Residues';
proc glm data = &sample;
    class group trt subject;
    model &y = group trt subject(group);
    output out=residual r = residual p = predicted;
run;
* Note: in model, rho is defined by correlation of random errors, so
below rho is calculated upon residues, instead of raw data;
%ar_1(sample = residual, y = residual, sub = subject);
proc print data = ar_mean; run;
data _n_;
    set ar_mean;
    call symput ('rho', autocorrelation);
run;
%put &sigma_b &sigma_e &rho;
%mend;

```

Then we can run the following macro for simulation steps 5 and 6. We need to input `n_sub` and `n_day` to specify the size of simulation data set. We also input the model parameters obtained from above. The output is the simulation data set.

```

/*****
Function: Main step of R_A method, generate simulation data set. See
more details step by step below.
Input:  n_sub, n_day: how many subjects and days to be simulated
        sigma_b, sigma_e, rho: key model parameters to build variance-
                                covariance structure
        n_group, n_trt: numbers of group levels and treatment levels
        p_group, p_trt: percentage of each group level and each
                                treatment level
        means: classification means of each group*treatment level;
Output: simulation: simulation data set;
*****/
%macro r_a (n_sub, n_day, sigma_b, sigma_e, rho, n_group, n_trt,
p_group, p_trt, means);

* First, build basic model with proper variance-covariance structure;
data simulation;
    do i = 1 to &n_sub;
        b = rannor(0)*sqrt(&sigma_b);
        do j = 1 to &n_day;

```

```

        subject = i;
        day = j;
        b = b;
        if j = 1 then
            do; e = rannor(0)*sqrt(&sigma_e); s = 0; output; end;
        else do;
            s = rannor(0)*sqrt((1-&rho*&rho)*&sigma_e);
            e = &rho*e+s;
            output;
        end;
    end;
end;
run;

* Second, modify into complete model by assigning classification levels
and level means;
* Assign group number;
%assign (sample = simulation,
        var_1 = subject,
        n_1 = &n_sub,
        var_2 = group,
        n_2 = &n_group,
        proportion = &p_group);
* Assign treatment number;
%assign (sample = simulation,
        var_1 = day,
        n_1 = &n_day,
        var_2 = trt,
        n_2 = &n_trt,
        proportion = &p_trt);
* Distribute classification means according to assigned levels;
%do i = 1 %to &n_group;
    %do j = 1 %to &n_trt;
        data select;
            set &means;
            if group = &i and trt = &j;
                keep estimate;
        run;
        data _Null_;
            set select;
            call symput ('mean',estimate);
        run;
        %put &mean;
        data simulation;
            set simulation;
            if group = &i and trt = &j then mu = &mean;
        run;
    %end;
%end;

* Last, add up above terms following model;
data simulation;
    set simulation;
    y = mu + b + e;
run;
%mend;

```