Natural Gas Ambient Air Monitoring Initiative (NGAAMI) Appendix B: Longitudinal Data Analysis for Volatile Organic Compounds

ADDITIONAL STATISTICAL METHODS USED AND DATA ANALYSES CONDUCTED [3]

Longitudinal Data Analysis for Volatile Organic Compounds

Methods

The dataset containing no tentatively identified compounds and all remaining actual and estimated values (NoTICs) contains Multi-level clustered data. They are unbalanced and incomplete- unbalanced in that the various compounds might not have data from all the same days. They are incomplete not because any data are missing, but as certain observations were removed from the analysis because of their estimated identification or value, some intended measurements could not be available. If this is not accounted for 1) the fundamental assumption of independence is violated 2) sampling variability estimates may be incorrect leading to 3) estimates that were less precise which could result in making an incorrect interpretation. Earlier analysis of variance was used for explanatory purposes only; it violated the assumption of independent observations.

To conduct a longitudinal data analysis the data must meet some assumptions [1]. First, the sampled air represents a random sample of all the air that could have been collected. Second, the resultant concentration levels had a multivariate normal distribution (after log transformation). Third, resultant concentration levels from different compounds are independent, while repeated measurements of the same compound are not assumed to be independent. This also holds true for concentration levels from different days and for different locations. Lastly, missingness was assumed to be completely at random as mentioned above.

This is an observational study and so there is no aspect of randomization. As such, we do not require an interaction term to study the association. Additionally, the interest is in group differences overall, not in time. The source of contamination was continuous in time; there was no one singular event that occurred. The group effect, the effect of each compound, and the differences between the specific compounds are the main interest here.

Data Analysis

Compounds are clustered by location and by day. Days form the lowest level of the clustering, nested by compounds which are nested within location. The analysis appropriately accounted for the correlation between the observations. The main source of correlation in the analysis was 'between-individual heterogeneity' or, here, 'between-compound heterogeneity'. A random effects model was used to describe the effect of the explanatory variable (wind) on each compound's outcome. We used a linear mixed effects model with 3 levels, where:

i = day

j = compound

k = location

Let Y_{ijk} be the resulting concentration level of day i compound j and location k. We assume:

$Y_{ijk} \sim N(\mu_{ijk}, \sigma^2)$

and form the model: $\mu_{ijk} = \beta_o + \beta_1 wind + b_{jk} + c_k + \varepsilon_{ijk}$.

Here, b_{jk} is the random intercept for the compound and c_k is the random intercept for the location. This model does not include any effects of a random slope.

As an example, a given day and compound, would result in the following matrices and vectors:

$$\mu_{ijk} = \begin{bmatrix} \mu_{ij1} \\ \mu_{ij2} \\ \mu_{ij3} \end{bmatrix} \qquad X_i = \begin{bmatrix} 1 & wind_{ij1} \\ 1 & wind_{ij2} \\ 1 & wind_{ij3} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$b_{jk} = \begin{bmatrix} b_{j1} \\ b_{j2} \\ b_{j3} \end{bmatrix} \qquad Z_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad c_k = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

$$\varepsilon_{ijk} = \begin{bmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \\ \varepsilon_{ij3} \end{bmatrix}.$$

This random effects model used wind as the only predictor. With this complete, the covariance model must also be determined. Using a random effects structure for the

model, the number of covariance parameters is the same regardless of the number and timing of the measurement occasions. There are 2 random effects, q=2 and $q*(q+1)/2 + 1 = 4$ covariance parameters. Even so, unstructured, compound structure, and autoregressive covariance structures were explored using restricted maximum likelihood (REML). The autoregressive covariance structured is nested within unstructured and these can be compared by -2 log likelihood.

Results

Comparing covariance models suggests an unstructured model is best. Both the autoregressive and compound symmetry models have similar likelihood values and AICs, [(Akaike's Information Criterion) used to compare different models fit to the same data] and these both have fewer parameters typically suggesting their superiority. However, the convergence criteria are met but the final hessian is not positive definitely suggesting these models are overfit. As such, the unstructured model was used. Using this model, we found significant values for the fixed effects of the intercept (p-value <0.0001 and wind (p-value = 0.0102) using an alpha value of 0.05. As such, wind was a confounder for this analysis and must be taken into consideration. With a parameter estimate of 0.2472, if a location experiences 10% more wind in its direction for a given day, the concentration level of a given compound was $e^{2.472}$ ppbv higher, or 11.85 ppbv higher. However, it's the specific compounds and their activity that we are most interested in and not the overall effect.

The following compounds had a significant p-value for at least one of the locations: 1,4-Dichlorobenze (overall location p-value = 0.0323), m-Xylene (overall location p-value 0.0173), Toluene (overall location p-value <0.0001), Ethanol (overall location p-value <0.0001, location 2 p-value 0.0229), Benzene (overall location p-value = 0.0011), Methylene Chloride (overall location p-value <0.0001), Carbon disulfide (overall location p-value = 0.0007), Dichlorodifluoromethane (overall location p-value <0.0001), and Methyl Ethyl Ketone (overall location p-value = 0.0046). Only ethanol had a location specific significant estimate at location 2, all other significant values were for the compound overall, not location-specific. Six of these eight compounds are present most frequently and, as such, were used earlier for the descriptive statistics.

SAS output for the Longitudinal Analysis using the model $\mu_{ijk} = \beta_0 + \beta_1 wind + b_{jk} + c_k + \varepsilon_{ijk}$

Tables A-20: SAS output results from Longitudinal model $\mu_{ijk} = \beta_0 + \beta_1 wind + b_{jk} + c_k + \varepsilon_{ijk}$

Class Level Information		
Class	**Levels**	**Values**
cas_number	30	100-41-4 100-42-5 106-46-7 107-06-2 108-10-1 108-38-3 108-88-3 120-82-1 127-18-4 56-23-5 64-17-5 67-66-3 71-43-2 71-55-6 74-83-9 74-87-3 75-09-2 75-15-0 75-34-3 75-69-4 75-71-8 76-13-1 76-14-2 78-93-3 79-00-5 79-01-6 79-34-5 91-20-3 95-47-6 95-63-6
location	3	1 2 3

Dimensions	
Covariance Parameters	3
Columns in X	2
Columns in Z Per Subject	4
Subjects	30
Max Obs Per Subject	127

Number of Observations	
Number of Observations Read	1081
Number of Observations Used	1081
Number of Observations Not Used	0

Iteration History			
Iteration	**Evaluations**	**-2 Log Like**	**Criterion**
0	1	3097.64750331	
1	3	2051.72972830	7.51085557
2	2	2050.84899578	2.93411251
3	2	2050.42930416	10.73195894
4	2	2045.21567181	0.00316516
5	1	2045.13076362	0.00008454

Iteration History			
Iteration	**Evaluations**	**-2 Log Like**	**Criterion**
6	1	2045.12825337	0.00000006
7	1	2045.12825166	0.00000000

Convergence criteria met.

Estimated G Matrix							
Row	**Effect**	**cas_number**	**location**	**Col1**	**Col2**	**Col3**	**Col4**
1	Intercept	100-41-4		0.3876			
2	Intercept	100-41-4	1		0.01102		
3	Intercept	100-41-4	2			0.01102	
4	Intercept	100-41-4	3				0.01102

Estimated V Matrix for cas_number 100-41-4									
Row	**Col1**	**Col2**	**Col3**	**Col4**	**Col5**	**Col6**	**Col7**	**Col8**	**Col9**
1	0.7562	0.3986	0.3876	0.3986	0.3876	0.3876	0.3986	0.3986	0.3986
2	0.3986	0.7562	0.3876	0.3986	0.3876	0.3876	0.3986	0.3986	0.3986
3	0.3876	0.3876	0.7562	0.3876	0.3876	0.3986	0.3876	0.3876	0.3876
4	0.3986	0.3986	0.3876	0.7562	0.3876	0.3876	0.3986	0.3986	0.3986
5	0.3876	0.3876	0.3876	0.3876	0.7562	0.3876	0.3876	0.3876	0.3876
6	0.3876	0.3876	0.3986	0.3876	0.3876	0.7562	0.3876	0.3876	0.3876
7	0.3986	0.3986	0.3876	0.3986	0.3876	0.3876	0.7562	0.3986	0.3986
8	0.3986	0.3986	0.3876	0.3986	0.3876	0.3876	0.3986	0.7562	0.3986
9	0.3986	0.3986	0.3876	0.3986	0.3876	0.3876	0.3986	0.3986	0.7562

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr > Z**
UN(1,1)	cas_number	0.3876	0.1201	3.23	0.0006
UN(1,1)	Location (cas_numbe)	0.01102	0.01148	0.96	0.1685
Residual		0.3576	0.01600	22.35	<.0001

Fit Statistics	
-2 Log Likelihood	2045.1
AIC (smaller is better)	2055.1
AICC (smaller is better)	2055.2
BIC (smaller is better)	2062.1

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	0.4293	0.1304	29	3.29	0.0026
Wind	0.2822	0.09798	1012	2.88	0.0041

Solution for Random Effects							
Effect	**Compound**	**Sample location**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
Intercept	Ethyl Benzene		-0.05441	0.2315	1012	-0.24	0.8142
Intercept	Ethyl Benzene	1	0.006038	0.1036	1012	0.06	0.9535
Intercept	Ethyl Benzene	2	-0.06643	0.1015	1012	-0.65	0.5128
Intercept	Ethyl Benzene	3	0.05885	0.1026	1012	0.57	0.5663
Intercept	Styrene		-0.01923	0.2238	1012	-0.09	0.9315
Intercept	Styrene	1	-0.00706	0.1025	1012	-0.07	0.9451
Intercept	Styrene	2	-0.1043	0.1010	1012	-1.03	0.3019
Intercept	Styrene	3	0.1108	0.1025	1012	1.08	0.2798
Intercept	1,4-Dichlorobenzene		0.9425	0.4397	1012	2.14	0.0323
Intercept	1,4-Dichlorobenzene	3	0.02680	0.1042	1012	0.26	0.7971
Intercept	1,2-Dichloroethane		0.1040	0.2736	1012	0.38	0.7040
Intercept	1,2-Dichloroethane	1	-0.01951	0.1030	1012	-0.19	0.8497

Solution for Random Effects							
Effect	**Compound**	**Sample location**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
Intercept	1,2-Dichloroethane	2	-0.01554	0.1037	1012	-0.15	0.8809
Intercept	1,2-Dichloroethane	3	0.03801	0.1029	1012	0.37	0.7120
Intercept	Methyl Isobutyl Ketone		0.04676	0.2580	1012	0.18	0.8562
Intercept	Methyl Isobutyl Ketone	1	0.003478	0.1028	1012	0.03	0.9730
Intercept	Methyl Isobutyl Ketone	2	-0.02579	0.1028	1012	-0.25	0.8020
Intercept	Methyl Isobutyl Ketone	3	0.02364	0.1028	1012	0.23	0.8182
Intercept	m,p-Xylene		0.4460	0.1870	1012	2.38	0.0173
Intercept	m,p-Xylene	1	-0.04051	0.09904	1012	-0.41	0.6826
Intercept	m,p-Xylene	2	-0.02869	0.09777	1012	-0.29	0.7693
Intercept	m,p-Xylene	3	0.08188	0.1022	1012	0.80	0.4233
Intercept	Toluene		1.1930	0.1499	1012	7.96	<.0001
Intercept	Toluene	1	-0.02746	0.08557	1012	-0.32	0.7483
Intercept	Toluene	2	0.04389	0.08156	1012	0.54	0.5906
Intercept	Toluene	3	0.01750	0.08478	1012	0.21	0.8365
Intercept	1,2,4-Trichloro-benzene		-0.1505	0.3270	1012	-0.46	0.6454
Intercept	1,2,4-Trichloro-benzene	2	-0.00428	0.1039	1012	-0.04	0.9671
Intercept	Tetrachloro-ethylene		0.1415	0.4396	1012	0.32	0.7477
Intercept	Tetrachloro-ethylene	2	0.004023	0.1042	1012	0.04	0.9692
Intercept	Carbon Tetrachloride		-0.3580	0.2386	1012	-1.50	0.1339

Solution for Random Effects							
Effect	**Compound**	**Sample location**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
Intercept	Carbon Tetrachloride	1	-0.00180	0.1027	1012	-0.02	0.9860
Intercept	Carbon Tetrachloride	2	-0.00625	0.1017	1012	-0.06	0.9510
Intercept	Carbon Tetrachloride	3	-0.00213	0.1036	1012	-0.02	0.9836
Intercept	Ethanol		1.7491	0.1502	1012	11.64	<.0001
Intercept	Ethanol	1	-0.03100	0.08569	1012	-0.36	0.7176
Intercept	Ethanol	2	0.1866	0.08190	1012	2.28	0.0229
Intercept	Ethanol	3	-0.1059	0.08534	1012	-1.24	0.2150
Intercept	Chloroform		-0.00313	0.2932	1012	-0.01	0.9915
Intercept	Chloroform	1	-0.02188	0.1037	1012	-0.21	0.8330
Intercept	Chloroform	2	0.02448	0.1031	1012	0.24	0.8125
Intercept	Chloroform	3	-0.00268	0.1037	1012	-0.03	0.9794
Intercept	Benzene		-0.4957	0.1511	1012	-3.28	0.0011
Intercept	Benzene	1	-0.03987	0.08706	1012	-0.46	0.6471
Intercept	Benzene	2	0.01801	0.08304	1012	0.22	0.8283
Intercept	Benzene	3	0.007758	0.08622	1012	0.09	0.9283
Intercept	1,1,1-Tetrachloro-ethane		-0.3767	0.4402	1012	-0.86	0.3923
Intercept	1,1,1-Tetrachloro-ethane	1	-0.01071	0.1042	1012	-0.10	0.9182
Intercept	Methyl Bromide		-0.4371	0.4402	1012	-0.99	0.3209
Intercept	Methyl Bromide	1	-0.01243	0.1042	1012	-0.12	0.9051
Intercept	Chloromethane		-0.1314	0.1500	1012	-0.88	0.3812
Intercept	Chloromethane	1	-0.02938	0.08585	1012	-0.34	0.7322

Solution for Random Effects							
Effect	**Compound**	**Sample location**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
Intercept	Chloromethane	2	0.02107	0.08165	1012	0.26	0.7964
Intercept	Chloromethane	3	0.004574	0.08483	1012	0.05	0.9570
Intercept	Methylene Chloride		-0.7382	0.1515	1012	-4.87	<.0001
Intercept	Methylene Chloride	1	-0.05362	0.08732	1012	-0.61	0.5393
Intercept	Methylene Chloride	2	0.02607	0.08349	1012	0.31	0.7549
Intercept	Methylene Chloride	3	0.006564	0.08635	1012	0.08	0.9394
Intercept	Carbon Disulfide		-0.6709	0.1978	1012	-3.39	0.0007
Intercept	Carbon Disulfide	1	-0.02380	0.1014	1012	-0.23	0.8144
Intercept	Carbon Disulfide	2	0.03756	0.09998	1012	0.38	0.7072
Intercept	Carbon Disulfide	3	-0.03284	0.09909	1012	-0.33	0.7404
Intercept	1,1-Dichloro-ethane		-0.5055	0.4402	1012	-1.15	0.2510
Intercept	1,1-Dichloro-ethane	1	-0.01438	0.1042	1012	-0.14	0.8903
Intercept	Trichlorofluoro-methane		0.01622	0.1499	1012	0.11	0.9139
Intercept	Trichlorofluoro-methane	1	-0.02301	0.08557	1012	-0.27	0.7881
Intercept	Trichlorofluoro-methane	2	0.01493	0.08156	1012	0.18	0.8548
Intercept	Trichlorofluoro-methane	3	0.008537	0.08478	1012	0.10	0.9198
Intercept	Dichlorofluoro-methane		0.6470	0.1500	1012	4.31	<.0001

Solution for Random Effects							
Effect	**Compound**	**Sample location**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
Intercept	Dichlorofluoro-methane	1	-0.02253	0.08585	1012	-0.26	0.7930
Intercept	Dichlorofluoro-methane	2	0.02476	0.08165	1012	0.30	0.7618
Intercept	Dichlorofluoro-methane	3	0.01617	0.08483	1012	0.19	0.8489
Intercept	Freon-113		-0.1512	0.2959	1012	-0.51	0.6095
Intercept	Freon-113	1	0.004887	0.1037	1012	0.05	0.9624
Intercept	Freon-113	2	-0.00919	0.1032	1012	-0.09	0.9291
Intercept	Freon-114		-0.1503	0.4402	1012	-0.34	0.7329
Intercept	Freon-114	1	-0.00427	0.1042	1012	-0.04	0.9673
Intercept	Methyl Ethyl Ketone		-0.4316	0.1520	1012	-2.84	0.0046
Intercept	Methyl Ethyl Ketone	1	-0.00208	0.08835	1012	-0.02	0.9812
Intercept	Methyl Ethyl Ketone	2	-0.03049	0.08384	1012	-0.36	0.7162
Intercept	Methyl Ethyl Ketone	3	0.02029	0.08677	1012	0.23	0.8152
Intercept	1,1,2-Tricloro-thane		-0.3767	0.4402	1012	-0.86	0.3923
Intercept	1,1,2-Tricloro-thane	1	-0.01071	0.1042	1012	-0.10	0.9182
Intercept	Trichloro-ethylene		0.7394	0.4398	1012	1.68	0.0930
Intercept	Trichloro-ethylene	2	0.02103	0.1042	1012	0.20	0.8402
Intercept	1,1,2,2-Tetrachloro-ethane		-0.2293	0.4402	1012	-0.52	0.6026
Intercept	1,1,2,2-Tetrachloro-ethane	1	-0.00652	0.1042	1012	-0.06	0.9501
Intercept	Naphthalene		-0.3193	0.1894	1012	-1.69	0.0922
Intercept	Naphthalene	1	-0.01055	0.1012	1012	-0.10	0.9170

Solution for Random Effects							
Effect	**Compound**	**Sample location**	**Estimate**	**Std Err Pred**	**DF**	**t Value**	**Pr > |t|**
Intercept	Naphthalene	2	0.009557	0.09795	1012	0.10	0.9223
Intercept	Naphthalene	3	-0.00808	0.09908	1012	-0.08	0.9350
Intercept	o-Xylene		-0.2686	0.2093	1012	-1.28	0.1996
Intercept	o-Xylene	1	-0.02577	0.1015	1012	-0.25	0.7997
Intercept	o-Xylene	2	-0.04203	0.1001	1012	-0.42	0.6747
Intercept	o-Xylene	3	0.06016	0.1024	1012	0.59	0.5569
Intercept	1,2,4-Trimethyl-benzene		-0.1573	0.3226	1012	-0.49	0.6260
Intercept	1,2,4-Trimethyl-benzene	2	-0.01128	0.1038	1012	-0.11	0.9135
Intercept	1,2,4-Trimethyl-benzene	3	0.006809	0.1035	1012	0.07	0.9475

Type 3 Tests of Fixed Effects						
Effect	**Num DF**	**Den DF**	**Chi-Square**	**F Value**	**Pr > ChiSq**	**Pr > F**
Wind	1	1012	8.30	8.30	0.0040	0.0041

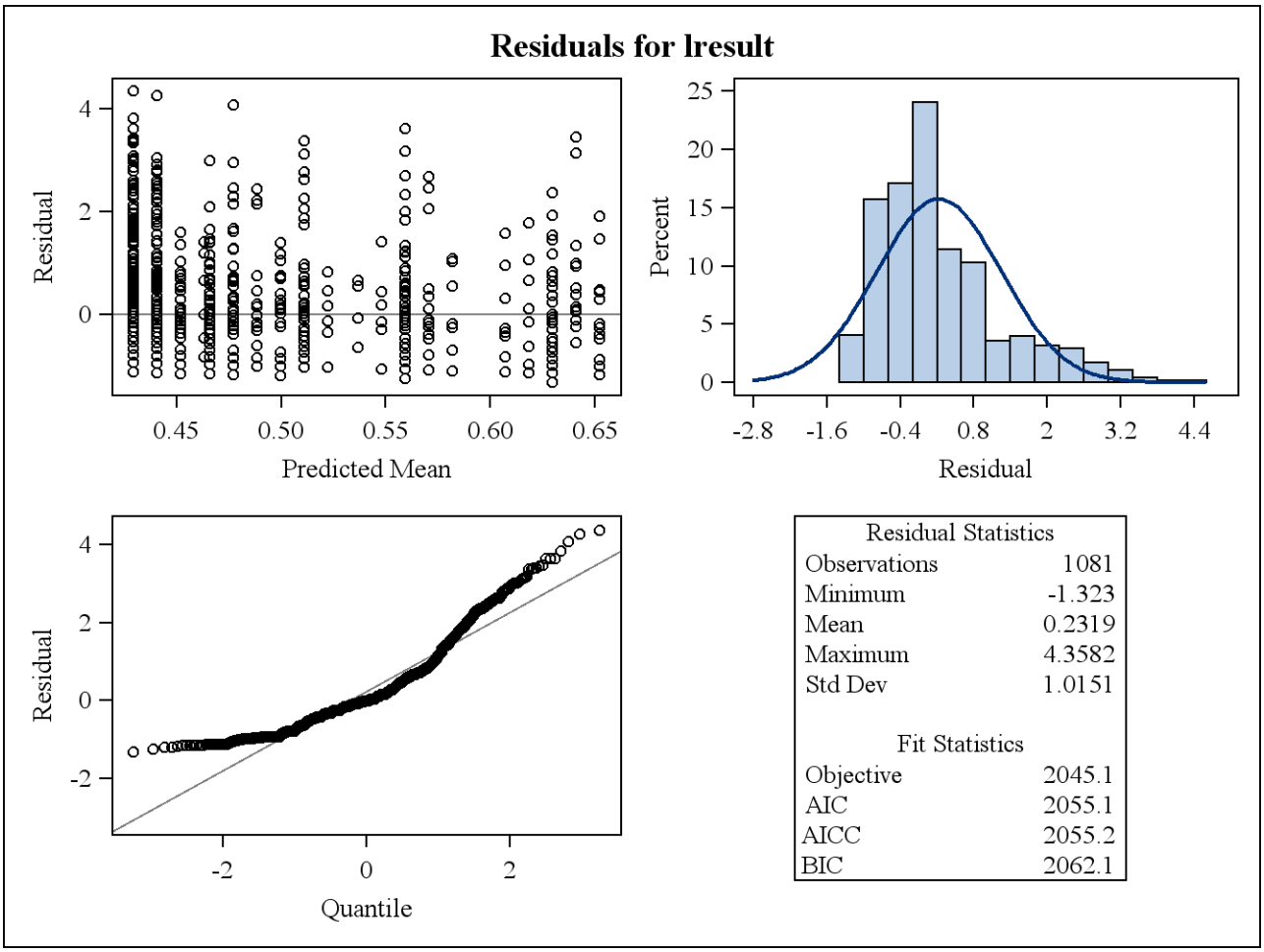

Figure A-2 for modeled residuals for the final dataset "lresult"

References

[1] Song, P, Xue, J, & Li, Z. *Simulation of longitudinal exposure data with variance-covariance structures based on mixed models.*

[2] Fitzmaurice, G. M., Laird, N. M., & Ware J. H. (2011). *Applied Longitudinal Analysis: Second Edition*. Boston, MA. Wiley Series in Probability and Statistics.

[3] Jervis, Allison. Philip-Tab, Loni., Gross-Davis, Carol Ann. *(unpublished thesis) Analyzing the clustered data of the US Environmental Protection Agency's Natural Gas Ambient Air Monitoring Initiative via linear mixed model methods.* For completion of master's degree in Biostatistics. Drexel University School of Public Health