Missing Monitoring Data Imputation Analysis
Monitoring Site 39-057-0005 Greene County, Ohio

Ohio Environmental Protection Agency
Lazarus Government Center
50 West Town Street, Suite 700
Columbus Ohio 43215

Prepared by Ohio EPA

Division of Air Pollution Control
State Implementation Plan Section
March 2011

Dayton-Springfield PM2.5 Nonattainment Area
Incomplete Monitoring Data Substitution Analysis

The current Dayton-Springfield nonattainment area is located in southwest Ohio and includes the following counties: Clark, Greene, and Montgomery.

The area has three monitors measuring $PM_{2.5}$ concentrations, which are operated by the Regional Air Pollution Control Agency (RAPCA)[1], which is part of the Montgomery County Health Department. A listing of the design values based on the three-year average of the annual mean concentrations from 2008 through 2010 is shown in Table 1. The design values calculated for the Dayton-Springfield area shows that the annual $PM_{2.5}$ NAAQS has been attained.

Table 1 - Monitoring Data for the Dayton-Springfield area for 2008 – 2010

Site	County	Annual Standard			
		Year			Average
		2008	2009	2010	2008-2010
39-023-0005	Clark	12.7	12.4	13.1	12.7
39-057-0005	Greene	11.6	11.5	13.2	12.1
39-113-0032	Montgomery	13.2	12.4	14.0	13.2

Less than 75% capture in at least one quarter

Source: U.S. EPA Air Quality System (AQS); http://www.epa.gov/ttn/airs/airsaqs/index.htm

However, based on Section 107(d)(3)(E)(i) of the Clean Air Act (CAA) the PM2.5 monitoring has to show that the three-year average of the annual mean values, based on data from all monitoring sites in the area or its affected downwind environs, are below $15.0/m^3$. Moreover, in accordance with the CAA Amendments, three complete years of monitoring data are required to demonstrate attainment at a monitoring site. In addition, U.S. EPA regulations require at least 75% data capture in each quarter of a consecutive 3-year period in order for a design value to be valid.

Table 1 shows that the monitor site in Greene County (site 39-057-0005) did not comply with the 75% data capture requirement in 2010. Specifically, the third quarter (July, August, and September) of 2010 has only 45% capture. This monitoring site underwent roof repairs (reroofing) from 08/12/2010 through 09/29/2010 for a total of 17 missing 1-in-3 day PM2.5 FRM runs. In 2010 there were scheduled a total of 122 scheduled runs. It is important to mention that this site is also the lowest reading site for $PM_{2.5}$.

In order to comply with U.S.EPA 75% capture requirements, Ohio EPA prepared a statistical analysis using multiple imputations. Imputing missing values for site 39-057-

1 RAPCA is a Local Air Agency that contracts with Ohio EPA and receives grants from U.S. EPA to enforce state and local air pollution control regulations in the six-county region (Preble, Darke, Miami, Montgomery, Clarke, and Greene).

0005 and then doing an ordinary analysis as if the imputed values were real measurements (this approach is usually better than excluding subjects with incomplete data). Most methods for accounting for having incomplete data can be complex; the bootstrapping method however, is an easy method to implement even though the computations can be slow. To use the bootstrap, to correctly estimate variances of regression coefficients, one must repeat the imputation process and the model fitting perhaps 1000 times using a resampling procedure.

Multiple imputations use random draws from the conditional distribution of the target variable given the other variables. When a regression model is used for imputation, the process involves adding a random residual to the "best guess" for missing values, to yield the same conditional variance as the original variable. To properly account for variability due to unknown values, the imputation will be repeated 1000 times.

Imputing missing data for Dayton-Springfield PM2.5 nonattainment area

For simplification purposes we will call Site 39-057-0005 Site A, Site 39-023-0005 Site B, and Site 39-113-0032 Site C.

1. Dayton-Springfield Annual PM2.5 Design Value History

Table 2 – Historic Design Values for the Dayton-Springfield area from 1999 to 2010

Site	Site	County	Annual Design Value									
			1999-2001	2000-2002	2001-2003	2002-2004	2003-2005	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010
39-057-0005	A	Greene					**12.4**	**13.2**	13.6	12.3	12.1	**12.1**
39-023-0005	B	Clark	**15.6**	**15.4**	14.7	14.2	14.8	14.4	14.8	13.5	13.2	12.7
39-113-0032	C	Montgomery			**16.0**	15.5	15.9	15.2	15.5	14.1	13.7	13.2

incomplete data (quarter with <75% capture)

violating DV

From Table 2, all three sites have design values (DV) that meet the PM2.5 annual standard since 2006-2008 period. However, site A has not proven clean data in 2010 and therefore it makes the entire nonattainment area non-eligible for redesignation. As mentioned before, the lack of clean data in 2010 is due to the low percentage of data capture in one or more quarters of 2010.

Multiple imputations and bootstrapping methods will help to generate the necessary missing data to determine 2010 completeness and recalculate the 2008-2010 DV.

2. Correlation, Quarterly Data Capture, and Data Site Pairing

Linear regression analyzes the relationship between two variables, X and Y. For each subject, there is a known X and Y and we want to find the best straight line through the data. The goal of linear regression is to find the line that best predicts Y from X. Linear

regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line. To determine which site combination (A and B or A and C) has the best data (best fit) from which to impute missing data for site A we determine the correlation between sites A and B and the correlation between A and C.

The correlation value (R^2) is useful because it describes the degree of relationship between two variables[2] (variables or site concentrations in sites A, B and C). R^2 is only a descriptive statistics. Roughly speaking, we associate a high value of R^2 with a good fit of the regression line and associate a low value of R^2 with a poor fit.

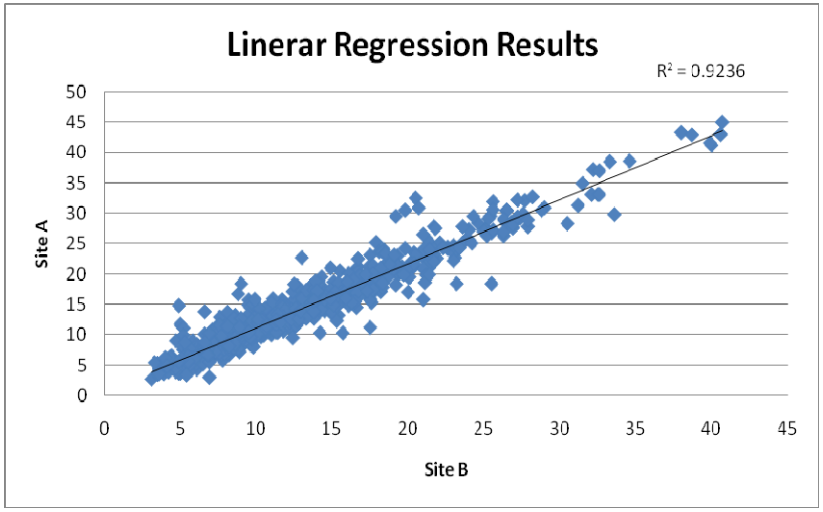
The mean of the quarterly data captured (the mean of the percentage captured) will allow verifying the central tendency of each site, which will help to determine what site (B or C) has a more data completeness to impute for site A.

Finally, although not as statistically significant as the correlation, or as the mean of the percentage captured, pairing the site data seeks to reduce variability in order to make more precise comparisons with fewer observations. Pairing the data will help to determine which site, B or C, have more data when paired with A.

Below are the results for sites A vs. B and for sites A vs. C

2 An R^2 value of 0.0 means that knowing X does not help to predict Y, there is no linear relationship between X and Y. When R2 equals 1.0, all points lie exactly on a straight line with no scatter; knowing X predicts Y perfectly.

Linear Regression and Correlation for Site A vs. Site B



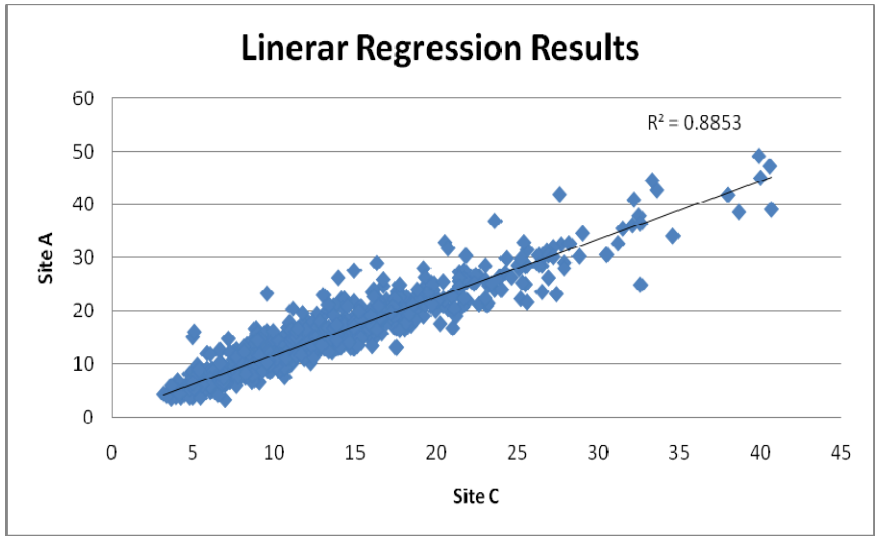
Regression Statistics	
Multiple R	0.961043954
R Square	**0.923605481**
Adjusted R Square	0.92351105
Standard Error	1.780706901
Observations	811

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.521517632	0.136976611	3.80734805	0.00015108	0.252646159	0.790389106
X Variable 1	0.873190988	0.00882923	98.89775017	0	0.855860087	0.890521889

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	31013.99463	31013.99463	9780.764989	0
Residual	809	2565.271907	3.170917067		
Total	810	33579.26654			

Linear Regression and Correlation for Site A vs. Site C



Regression Statistics	
Multiple R	0.940886234
R Square	**0.885266906**
Adjusted R Square	0.885120562
Standard Error	2.193683537
Observations	786

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.81715995	0.171189431	4.773425231	2.16082E-06	0.481116	1.153203844
X Variable 1	0.81295687	0.010452417	77.77692926	0	0.792439	0.833474906

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	29110.49143	29110.49143	6049.250725	0
Residual	784	3772.802007	4.812247458		
Total	785	32883.29344			

Table 3 - Quarterly Data Capture

Quarterly Data Capture 2003-2010	Monitoring Sites		
	A	B	C
2003 Q1	-	100%	87%
2003 Q2	-	93%	96%
2003 Q3	-	100%	100%
2003 Q4	73%	100%	89%
2004 Q1	71%	100%	100%
2004 Q2	93%	100%	90%
2004 Q3	100%	97%	97%
2004 Q4	90%	100%	90%
2005 Q1	97%	100%	100%
2005 Q2	97%	100%	100%
2005 Q3	100%	100%	100%
2005 Q4	100%	90%	97%
2006 Q1	100%	100%	100%
2006 Q2	97%	100%	97%
2006 Q3	94%	100%	100%
2006 Q4	100%	100%	97%
2007 Q1	97%	100%	100%
2007 Q2	93%	100%	100%
2007 Q3	100%	100%	100%
2007 Q4	93%	100%	97%
2008 Q1	97%	97%	93%
2008 Q2	100%	87%	98%
2008 Q3	94%	100%	98%
2008 Q4	100%	100%	95%
2009 Q1	100%	100%	94%
2009 Q2	100%	97%	100%
2009 Q3	100%	100%	90%
2009 Q4	97%	100%	100%
2010 Q1	97%	100%	97%
2010 Q2	100%	100%	100%
2010 Q3	45%	100%	97%
2010 Q4	100%	100%	100%
MEAN	**94%**	**99%**	**97%**

Table 4 - Paired Data Site

		29 - Quarter look (2003 Q4 - 2010 Q4)				
Site	**Site**	**pairs Q1**	**pairs Q2**	**pairs Q3**	**pairs Q4**	**Total**
A	**B**	197	199	191	224	811
A	**C**	186	194	188	218	786

The correlation between A and B is 0.9236 and the correlation between A and C is 0.8853. This shows that the relationship between A and B is stronger than the relationship between A and C. From the mean of the quarterly data captured (Table 3), we can observe that Site B has more data completeness (99%) than Site C (97%); therefore Site B has a more complete data from where to impute for site A. In addition, from the data site pairing (Table 4) it can be observed that Site A and B have more paired data (in each quarter and in total) to establish a better relationship than Site A and Site C.

Finally, based on the above results, it was determined that Site B has the best data, statistically significant (greater correlation, larger data completeness, and larger paired data) from which to impute missing data for Site A. Therefore, we will use multiple imputations and bootstrapping methods to generate the necessary data from Site B to impute on Site A missing values.

3. Bootstrapping and Imputation

The bootstrapping method randomly applies real residuals from the linear regression to the imputed current-period Site A values. A 1000 bootstrap runs adds a random residual to the “best guess” missing values yielding the same conditional variance as the original variable. A summary of the bootstrapping statistics is presented in Table 5, where the bootstrapping residual is $-929.0569514e^{-6}$.

Table 5 - Bootstrapping Summary Statistics

Average	$-929.0569514e^{-6}$
SD	0.0356
Max	0.103
Min	-0.137

Finally to impute the missing data using the bootstrapping residual we use the following equation, based on the linear regression from Site A vs. Site B:

Site A concentration = Intercept + x Variable * Site B concentration + bootstrapping residual

Where:

Intercept = 0.521517632 (from linear regression)
x Variable = 0.873190988 (from linear regression)
Bootstrapping residual = $-929.0569514e^{-6}$
Site B concentration = values in Site B

After applying the above equation to all missing data in Site A, we recalculated the design values based on the three-year average of the annual mean concentrations for

all existing years in Site A (Site 39-057-0005). Table 6 shows Site 39-057-0005 before and after the imputation of missing data. Also the “new” site (with imputed values) shows a “new” passing DV for 2008-2010.

Table 6 – Comparison Between Original and Imputed Values for Monitor 39-057-0005 from 2004 to 2010

	Site	Year							Annual Design Value				
		2004	2005	2006	2007	2008	2009	2010	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010
OLD	39-057-0005	12.1	15.5	11.9	13.3	11.6	11.5	13.2	13.2	13.6	12.3	12.1	12.1
NEW	39-057-0005	11.7	15.2	11.9	13.5	11.6	11.5	12.2	12.9	13.5	12.3	12.2	11.8

incomplete data (quarter with <75% capture)

As we said before, the bootstrapping method randomly applies real residuals from the linear regression to the imputed current-period Site A values. A 1,000 bootstrap runs provided a 95% confidence interval for the imputed current-period design value. Table 7 shows that, since the R^2 value between Site A and Site B was strong, predicting its relationship almost perfectly (R^2 = 0.9236), there is small variation between DV interval (as well as residual intervals) and it was needed to provide at least 3 significant decimals to show the different intervals. Table 7 shows that the DV = 11.763 has the most occurrences, confirming that the “new” DV 2008-2010 of 11.8, in Table 6, is the best guess of imputed data.

Table 7 - Bootstrapping Residuals and DV Calculations

Residual	Design Value	Frequency
-0.14	11.754	3
-0.115	11.755	3
-0.09	11.757	20
-0.065	11.758	87
-0.04	11.760	179
-0.015	11.761	252
0.01	11.763	262
0.035	11.764	143
0.06	11.766	42
0.085	11.767	9

Ohio EPA believes that the above analysis has generated the most statistically significant results yielding to the “best” missing data statistically possible. This analysis should also satisfy US EPA requirements in terms of total percentage data capture and design values under the $PM_{2.5}$ annual standard (15.0µg/m^3).

References:

Brownstone, David and Robert Valleta. 2000. Draft for JEP Econometrics Symposium. “The bootstrap and multiple imputations”.

Harrel, Frank E. Jr. 2001. Springer Series in Statistics. “Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis”.

Mooney, Christopher Z., Robert D. Duval, Robert Duval. 1993. Sage Publications Inc. “Bootstrapping: a nonparametric approach to statistical inference” Issues 94-95.