Missing Monitoring Data Imputation Analysis
Monitoring Sites 39-035-0060, 39-035-0065, and 39-035-1002
Cuyahoga 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
August 2011

Cleveland-Akron-Lorain PM2.5 Nonattainment Area Monitor Missing Data Analysis

The current Cleveland-Akron-Lorain nonattainment area is located in northwest Ohio and includes the following counties: Cuyahoga, Lake, Lorain, Medina, Portage and Summit.

The area has twelve[1] monitors measuring $PM_{2.5}$ concentrations, which are overseen by the Ohio EPA and operated by Local Air Agencies and Ohio EPA's Northeast District Office[2]. A listing of the design values based on the 3-year average of the annual mean concentrations from 2008 through 2010 is shown in Table 1. The design values calculated for the Cleveland-Akron-Lorain area shows that the annual $PM_{2.5}$ NAAQS has been attained.

Table 1 - Monitoring Data for the Cleveland-Akron-Lorain area for 2008 – 2010

		Annual Standard			
Site	County	Year			Average
		2008	2009	2010	2008-2010
39-035-0034	Cuyahoga	10.9	10.2	10.9	10.7
39-035-0038		14.1	12.8	14.0	13.6
39-035-0045		13.7	11.8	13.3	12.9
39-035-0060		14.1	12.3	13.7	13.4
39-035-0065		14.6	12.4	13.2	13.4
39-035-1002		12.0	10.9	11.3	11.4
39-085-0007	Lake		10.4	10.4	10.8
39-085-3002		11.5			
39-093-3002	Lorain	11.4	9.9	10.4	10.6
39-103-0003	Medina	11.8	10.8	10.8	11.1
39-133-0002	Portage	12.1	11.1	11.2	11.5
39-153-0017	Summit	13.8	12.6	13.4	13.3
39-153-0023		12.9	11.4	12.5	12.3
	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

1 The redesignation and maintenance plan document refers to thirteen monitors. The thirteenth monitor was permanently shut down in 2009 as a part of Ohio's approved monitoring network. Therefore, this monitor is not included in this analysis.

2 Monitors in Cuyahoga County are operated by the Cleveland Department of Public Health & Welfare; the monitor in Lake County are operated by Lake County General Health District; monitors Medina, Portage, and Summit Counties are operated by the Akron Regional Air Quality Management District; and the monitor in Lorain County is operated by Ohio EPA's Northeast District Office.

However, based on Section 107(d)(3)(E)(i) of the Clean Air Act (CAA) the $PM_{2.5}$ monitoring has to show that the 3-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 three monitor sites in Cuyahoga County, site 39-035-0060, site 39-035-0065, and site 39-035-1002, did not comply with the 75% data capture requirement for the 2008-2010 period. Specifically, site 39-035-0060 has only 48% and 63% capture in the second and third quarter of 2009, respectively. Site 39-035-0065 has only 74% capture in the third quarter of 2010, and site 39-035-1002 has only 71% capture in the first quarter of 2008 and 73% capture in the third quarter of 2009. These monitoring sites experienced instrument malfunctions during the low percentage capture periods with a total of 27 missing 1-in-3 PM2.5 FRM runs in site 39-035-0060 (in 2009), a total of 8 missing 1-in-3 $PM_{2.5}$ FRM runs in site 39-035-0065 (in 2010), and a total of 8 missing 1-in-3 $PM_{2.5}$ FRM runs and a total of 9 missing 1-in-3 $PM_{2.5}$ FRM runs in site 39-035-1002 (in both 2008 and 2009 respectively).

In order to comply with U.S.EPA's 75% capture requirements, Ohio EPA prepared a statistical analysis using multiple imputations. Imputing missing values for sites 39-035-0060, 39-035-0065, and 39-035-1002 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 Cleveland-Akron-Lorain PM2.5 nonattainment area

For simplification purposes we will refer to the sites in the Cleveland-Akron-Lorain nonattainment area by the Site letter denoted in the second column of the table below.

1. Cleveland-Akron-Lorain Annual PM2.5 Design Value History

Table 2 – Historic Design Values for the Cleveland-Akron-Lorain_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
39-035-0034	A	Cuyahoga	15.3	15.0	14.2	13.4	14.1	13.5	13.8	12.0	11.6	10.7
39-035-0038	B		20.3	19.2	18.4	17.6	18.1	17.2	16.8	15.1	14.4	13.6
39-035-0045	C		16.7	17.5	16.7	16.0	17.0	16.2	16.2	14.4	13.6	12.9
39-035-0060	**D**		18.6	18.2	17.5	17.0	17.7	16.9	16.8	15.0	14.1	13.4
39-035-0065	**E**		17.6	17.0	16.0	15.5	16.5	15.6	15.8	14.5	14.3	13.4
39-035-1002	**F**		15.1	15.0	14.6	14.1	14.6	13.9	13.9	12.3	12.1	11.4
39-085-0007	G	Lake									11.9	10.8
39-085-3002										12.3		
39-093-3002	H	Lorain				12.5	12.8	12.6	13.0	11.9	11.4	10.6
39-103-0003	I	Medina							13.3	12.1	11.8	11.1
39-133-0002	J	Portage	15.3	15.2	14.2	13.3	13.4	13.2	13.6	12.6	12.3	11.5
39-153-0017	K	Summit	17.4	17.0	16.6	15.7	15.6	15.0	14.9	14.0	13.7	13.3
39-153-0023	L		16.2	16.3	15.6	15.0	14.6	14.1	14.1	13.1	12.7	12.3

incomplete data (quarter with <75% capture)

violating DV

From Table 2, all twelve sites have design values (DV) that meet the PM2.5 annual standard since the 2007-2009 period. However, sites D, E, and F have not achieved clean data in at least on quarter from 2008-2010, and; therefore the entire nonattainment area is not eligible for redesignation absent this analysis. As mentioned before, the lack of clean data is due to the low percentage of data capture in one or more quarters between 2008 and 2010.

Multiple imputations and bootstrapping methods will help to generate values for the missing data to determine 2008, 2009, and 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[3] has

3 Site D vs. Site A or Site D vs. Site B or Site D vs. Site C or Site D vs. Site G or Site D vs. Site H or Site D vs. Site I or Site D vs. Site J or Site D vs. Site K or Site D vs. Site L.
Site E vs. Site A or Site E vs. Site B or Site E vs. Site C or Site E vs. Site G or Site E vs. Site H or Site E vs. Site I or Site E vs. Site J or Site E vs. Site K or Site E vs. Site L.
And Site F vs. Site A or Site F vs. Site B or Site F vs. Site C or Site F vs. Site G or Site F vs. Site H or Site F vs. Site I or Site F vs. Site J or Site F vs. Site K or Site F vs. Site L.
Note that we do not compare Site D vs. Site E, Site D vs. Site F nor Site E vs. Site F since these sites are incomplete and would provide a biased relationship.

the best data (best fit) from which to impute missing data for site D, site E, and site F we determine the correlation between all the sites and site D, the correlation between all the sites and site E and the correlation between all the sites and site F.

The correlation value (R^2) is useful because it describes the degree of relationship between two variables[4] (variables or site concentrations in all sites). 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 has more data completeness to impute for sites D, E, and F.

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 has more data when paired with D, E, and F.

Below are the results for site D vs. the rest of the sites, site E vs. the rest of the sites, and for sites F vs. the rest of the sites.

Table 3 – Correlation Matrix for site D, site E, and site F against all other sites.

	A	B	C	G	H	I	J	K	L
D	0.8841	**0.9154**	0.8965	0.8387	0.8461	0.8165	0.7663	0.8217	0.8079
E	0.8763	0.8484	**0.8864**	0.7833	0.8366	0.8002	0.7628	0.7922	0.7877
F	**0.8804**	0.8136	0.8700	0.8449	0.8528	0.8410	0.7976	0.8143	0.8233

4 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.

Table 4 - Quarterly Data Capture

Quarterly Data Capture 2003-2010	Monitoring Sites											
	A	B	C	D	E	F	G	H	I	J	K	L
2003 Q1	100%	94%	100%	90%	87%	93%	-	83%	-	87%	86%	90%
2003 Q2	100%	97%	100%	93%	87%	97%	-	97%	-	90%	89%	88%
2003 Q3	100%	97%	94%	94%	100%	87%	-	90%	-	87%	91%	91%
2003 Q4	90%	72%	100%	93%	93%	97%	-	87%	-	100%	90%	97%
2004 Q1	97%	90%	94%	97%	94%	94%	-	94%	-	90%	98%	84%
2004 Q2	87%	87%	90%	100%	97%	93%	-	100%	-	97%	92%	92%
2004 Q3	97%	86%	100%	100%	100%	100%	-	90%	-	97%	96%	99%
2004 Q4	97%	98%	97%	97%	93%	93%	-	90%	-	90%	98%	97%
2005 Q1	95%	96%	86%	100%	81%	81%	-	93%	-	97%	92%	80%
2005 Q2	100%	94%	93%	100%	93%	100%	-	87%	94%	100%	94%	100%
2005 Q3	100%	97%	93%	83%	100%	93%	-	83%	100%	100%	100%	100%
2005 Q4	100%	94%	100%	97%	100%	100%	-	97%	77%	81%	90%	94%
2006 Q1	87%	87%	100%	93%	93%	100%	85%	97%	90%	97%	100%	100%
2006 Q2	100%	100%	100%	87%	93%	100%	93%	90%	67%	93%	100%	93%
2006 Q3	93%	100%	100%	94%	93%	67%	94%	90%	100%	97%	100%	90%
2006 Q4	94%	100%	94%	100%	100%	88%	90%	94%	94%	100%	97%	97%
2007 Q1	95%	93%	85%	80%	75%	85%	97%	63%	90%	93%	97%	100%
2007 Q2	100%	90%	87%	90%	100%	87%	97%	93%	80%	87%	100%	100%
2007 Q3	90%	77%	100%	100%	100%	94%	97%	94%	100%	97%	100%	97%
2007 Q4	90%	57%	97%	97%	100%	93%	93%	93%	93%	97%	87%	93%
2008 Q1	84%	97%	90%	77%	94%	71%	90%	94%	94%	97%	87%	97%
2008 Q2	100%	100%	97%	100%	100%	90%	100%	97%	100%	100%	100%	90%
2008 Q3	97%	97%	84%	97%	94%	97%	100%	100%	90%	100%	100%	94%
2008 Q4	93%	97%	90%	97%	100%	100%	93%	97%	87%	83%	100%	93%
2009 Q1	93%	93%	97%	90%	93%	87%	90%	100%	97%	93%	77%	93%
2009 Q2	100%	100%	94%	48%	97%	94%	100%	100%	90%	100%	97%	90%
2009 Q3	93%	97%	97%	63%	97%	73%	100%	100%	100%	87%	91%	97%
2009 Q4	100%	100%	100%	97%	100%	97%	87%	94%	94%	97%	100%	94%
2010 Q1	100%	90%	100%	83%	100%	87%	97%	93%	90%	87%	99%	90%
2010 Q2	83%	97%	97%	100%	97%	80%	100%	100%	93%	97%	100%	90%
2010 Q3	94%	87%	100%	97%	74%	100%	100%	100%	100%	97%	100%	97%
2010 Q4	90%	100%	94%	94%	90%	87%	100%	100%	97%	97%	100%	84%
MEAN	**95%**	**93%**	**95%**	**92%**	**94%**	**91%**	**95%**	**93%**	**92%**	**94%**	**95%**	**93%**

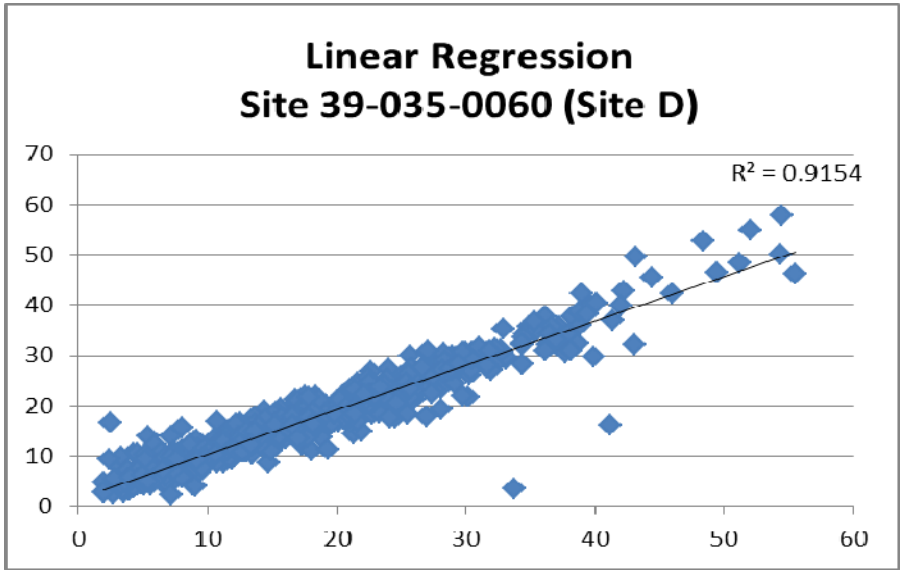
Table 5 - Paired Data Site

		30 - Quarter look (2003 Q1 - 2010 Q4)				
Site	**Site**	**pairs Q1**	**pairs Q2**	**pairs Q3**	**pairs Q4**	**Total**
D	**A**	170	179	185	190	724
D	**B**	193	209	206	207	815
D	**C**	167	178	187	196	728
D	**G**	100	121	136	130	487
D	**H**	180	205	206	214	805
D	**I**	110	140	161	161	572
D	**J**	190	208	213	217	828
D	**K**	187	205	219	220	831
D	**L**	192	203	213	221	829
E	**A**	178	193	192	191	754
E	**B**	179	195	185	177	736
E	**C**	178	189	193	196	756
E	**G**	98	126	124	117	465
E	**H**	164	192	186	185	727
E	**I**	109	132	138	134	513
E	**J**	173	193	191	190	747
E	**K**	172	191	197	192	752
E	**L**	174	188	194	189	745
F	**A**	177	182	184	188	731
F	**B**	170	186	176	174	706
F	**C**	171	178	184	195	728
F	**G**	90	113	119	113	435
F	**H**	157	182	181	182	702
F	**I**	100	121	132	132	485
F	**J**	166	181	184	185	716
F	**K**	167	180	188	188	723
F	**L**	168	177	184	187	716

Table 3 shows that the correlation between D and B is 0.9154, the correlation between E and C is 0.8864, and the correlation between F and A is 0.8804. These correlations show the best data (stronger relationship among sites) from which to impute missing data for all three sites (D, E, and F). From the mean of the quarterly data captured (Table 4), we can observe that sites A, B and C (the sites that show higher correlation with the sites with missing values – D, E, and F) have adequate data completeness[5] from where to impute missing values. In addition, from the data site pairing (Table 5) it can be observed that sites D vs. B, E vs. C, and F vs. C have enough paired data (in each quarter and in total) to establish a good relationship.

5 Site A = 95%, Site B = 93%, and Site C = 95%

Linear Regression and Correlation for Site D vs. Site B



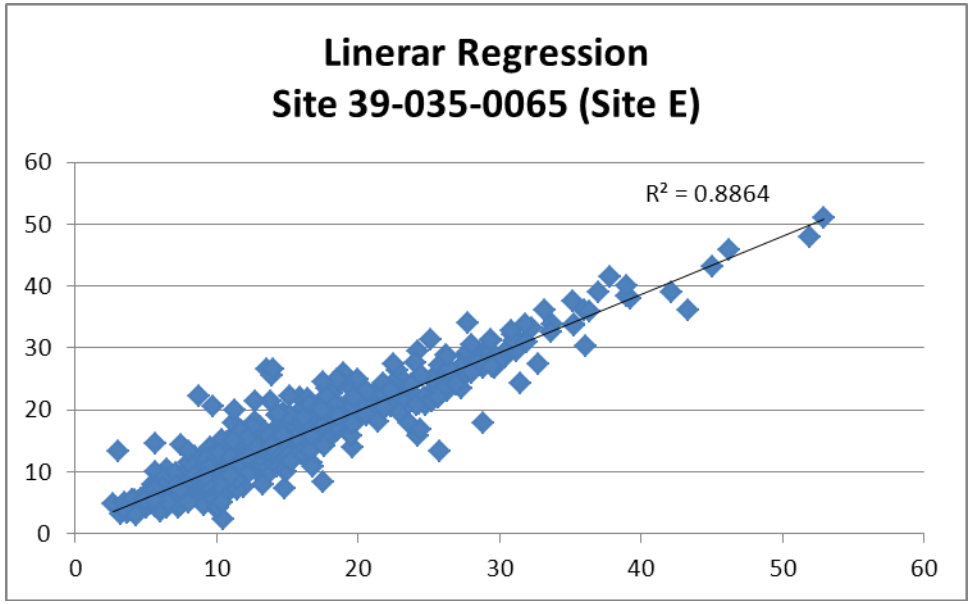
Regression Statistics	
Multiple R	0.9568
R Square	**0.9154**
Adjusted R Square	0.9153
Standard Error	2.5039
Observations	815

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.565566405	0.174087186	8.993001954	1.66443E-18	1.223853073	1.907279737
X Variable 1	0.880442384	0.00938778	93.78600165	0	0.862015241	0.898869529

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	55145.65365	55145.65365	8795.814106	0
Residual	813	5097.130962	6.269533779		
Total	814	60242.78461			

Linear Regression and Correlation for Site E vs. Site C



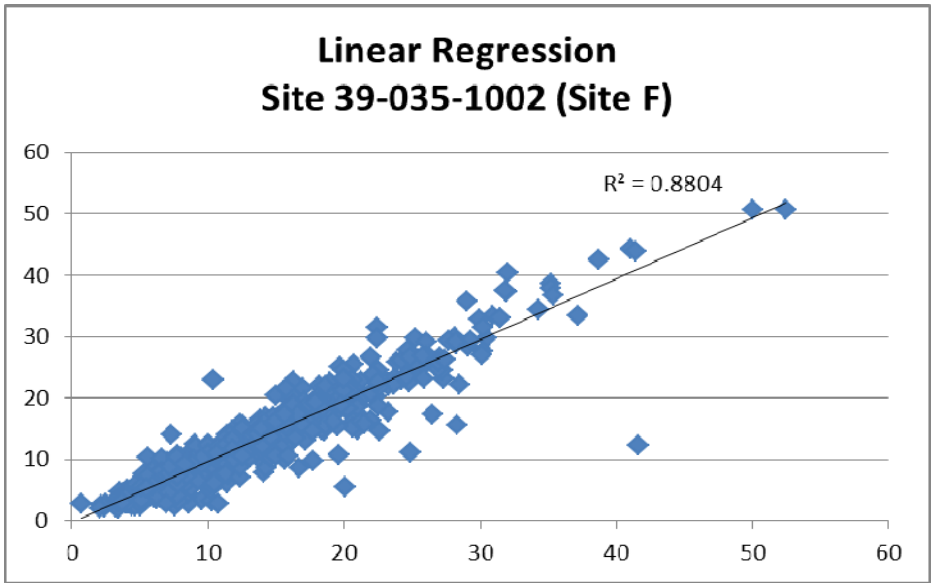
Regression Statistics	
Multiple R	0.9415
R Square	**0.8864**
Adjusted R Square	0.8863
Standard Error	2.5592
Observations	756

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.932587361	0.202223699	4.611662056	4.69183E-06	0.535598942	1.32957578
X Variable 1	0.944152034	0.012306434	76.72019907	0	0.919993087	0.968310981

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	38549.39211	38549.39211	5885.988945	0
Residual	754	4938.208671	6.549348371		
Total	755	43487.60078			

Linear Regression and Correlation for Site F vs. Site A



Regression Statistics	
Multiple R	0.9383
R Square	**0.8804**
Adjusted R Square	0.8802
Standard Error	2.6246
Observations	731

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.31095621	0.198704325	-1.56491920	0.118035816	-0.701057206	0.079144775
X Variable 1	0.991764397	0.013539584	73.24925362	0	0.965183169	1.018345625

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	36959.02579	36959.02579	5365.453155	0
Residual	729	5021.594453	6.888332583		
Total	730	41980.62025			

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

3. Bootstrapping and Imputation

3.1. Site D and Site B

The bootstrapping method randomly applies real residuals from the linear regression to the imputed current-period Site D 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 6, where the bootstrapping residual is 0.00238.

Table 6 - Bootstrapping Summary Statistics

Average	0.0023787420
SD	0.0849651846
Max	0.2365176245
Min	-0.3681799770

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

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

Where:

Intercept = 1.56556640501966 (from linear regression)
x Variable = 0.880442384961355 (from linear regression)
Bootstrapping residual = 0.0023787420
Site B concentration = values in Site B

After applying the above equation to all missing data in Site D, we recalculated the design values based on the 3-year average of the annual mean concentrations for all existing years in Site D (Site 39-035-0060). Table 7 shows Site 39-035-0060 before and after the imputation of missing data. Also the "new" site (with imputed values) shows a "new" passing DV for 2008-2010.

Table 7 – Comparison Between Original and Imputed Values for Site 39-035-0060 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	39-035-0060	16.4	19.4	15.0	15.9	14.1	12.3	13.7	16.9	16.8	15.0	14.1	13.4
NEW	39-035-0060	15.6	18.5	14.6	15.4	13.8	12.5	13.6	16.2	16.2	14.6	13.9	13.3

incomplete data (quarter with <75% capture)

3.2. Site E and Site C

The bootstrapping method randomly applies real residuals from the linear regression to the imputed current-period Site E 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 8, where the bootstrapping residual is 0.00185.

Table 8 - Bootstrapping Summary Statistics

Average	0.0018513212
SD	0.0900055812
Max	0.2707392673
Min	-0.2713173739

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

Site E concentration = Intercept + x Variable * Site C concentration + bootstrapping residual

Where:

Intercept = 0.932587360923463 (from linear regression)
x Variable = 0.94415203423236 (from linear regression)
Bootstrapping residual = 0.00185132118436311Site C concentration = values in Site C

After applying the above equation to all missing data in Site E, we recalculated the design values based on the 3-year average of the annual mean concentrations for all existing years in Site E (Site 39-035-0065). Table 9 shows Site 39-035-0065 before and after the imputation of missing data. Also the "new" site (with imputed values) shows a "new" passing DV for 2008-2010.

Table 9 – Comparison Between Original and Imputed Values for Site 39-035-0065 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	39-035-0065	15.2	18.6	13.1	15.8	14.6	12.4	13.2	15.6	15.8	14.5	14.3	13.4
NEW	39-035-0065	15.2	18.4	13.2	15.7	14.8	12.4	13.4	15.6	15.8	14.6	14.3	13.5

incomplete data (quarter with <75% capture)

3.3. Site F and Site A

The bootstrapping method randomly applies real residuals from the linear regression to the imputed current-period Site F 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 10, where the bootstrapping residual is 0.00648.

Table 10 - Bootstrapping Summary Statistics

Average	0.0064810934
SD	0.0984385889
Max	0.3018532840
Min	-0.3655841335

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

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

Where:

Intercept = -0.310956215392546 (from linear regression)
x Variable = 0.991764397262166 (from linear regression)
Bootstrapping residual = 0.00648109339124171
Site A concentration = values in Site A

After applying the above equation to all missing data in Site F, we recalculated the design values based on the 3-year average of the annual mean concentrations for all existing years in Site F (Site 39-035-1002). Table 11 shows Site 39-035-1002 before and after the imputation of missing data. Also the "new" site (with imputed values) shows a "new" passing DV for 2008-2010.

Table 11 –Comparison Between Original and Imputed Values for Site 39-035-1002 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	39-035-1002	13.2	16.8	11.6	13.4	12.0	10.9	11.3	13.9	13.9	12.3	12.1	11.4
NEW	39-035-1002	13.1	16.7	11.0	13.3	11.1	10.6	11.1	13.6	13.7	11.8	11.6	10.9

incomplete data (quarter with <75% capture)

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.