

Steubenville-Weirton PM_{2.5} Nonattainment Area Monitor Missing Data Analysis

The current Steubenville-Weirton nonattainment area is located in the eastern side of Ohio and includes the following counties: Jefferson in Ohio and Hancock and Brooke in West Virginia.

The area has five monitors measuring PM_{2.5} concentrations, which are operated by the Ohio EPA Division of Air Pollution Control, Southwest District Office (Jefferson County monitors) and by the West Virginia Department of Air Quality. 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 Steubenville-Weirton area shows that the annual PM_{2.5} NAAQS has been attained.

Table 1 - Monitoring Data for the Steubenville-Weirton area for 2008 – 2010

Site	County	Annual Standard			
		Year			Average 2008-2010
		2008	2009	2010	
39-081-0017 [a]	Jefferson, OH	14.3	12.1	12.7	13.0
39-081-1001		14.1	11.2	12.7	12.7
54-009-0005	Brooke, WV	14.7	12.2	14.1	13.7
54-009-0011		13.8	11.9	13.5	13.1
54-029-1004	Hancock, WV	13.3	11.2	12.6	12.4
Less than 75% capture in at least one quarter					
[a] This site has a 73% capture, resulting in one day of missed data. Based on the data from previous years, this site shows a decreasing trend in monitor readings. Ohio believes that had the missing data been collected it would not have resulted in a design value that would exceed the standard.					

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 PM_{2.5} 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/m³. 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 Jefferson County (site 39-081-0017) did not comply with the 75% data capture requirement in 2008. Specifically, the fourth quarter (October, November, and December) of 2008 has only 73% capture. This monitoring site experienced an instrument malfunction during the low percentage capture period with a total of 8 missing 1-in-3 day PM_{2.5} FRM runs.

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 site 39-081-0017 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 Steubenville-Weirton PM2.5 nonattainment area

For simplification purposes we will refer to the sites in the Steubenville-Weirton nonattainment area by the letter denoted in the second column of the table below.

1. Steubenville-Weirton Annual PM2.5 Design Value History

Table 2 – Historic Design Values for the Steubenville-Weirton 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
39-081-0017	A	Jefferson, OH					15.8	15.4	15.5	14.8	14.2	13.0
39-081-1001	B		18.2	17.8	17.8	16.9	17.2	16.3	16.1	14.8	13.6	12.7
54-009-0005	C	Brooke, WV	17.3	16.8	16.8	16.5	16.8	16.4	16.3	15.4	14.4	13.7
54-009-0011	D		16.6	16.1	16.2	15.8	16.4	15.7	16.1	14.9	14.0	13.0
54-029-1004	E	Hancock, WV	17.3	17.5	17.4	17.0	16.6	15.4	15.2	14.3	13.4	12.4

 incomplete data (quarter with <75% capture)
 violating DV

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

Multiple imputations and bootstrapping methods will help to generate values for the missing data to determine 2008 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, A and C, A and D or A and E) has the best data (best fit) from which to impute missing data for site A we determine the correlation between sites A and B, between A and C, between A and D and between A and E.

The correlation value (R^2) is useful because it describes the degree of relationship between two variables¹ (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 (B, C, D or E) 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, C, D or E, have more data when paired with A.

Below are the results for site A vs. the rest of the sites.

Table 3 – Correlation Matrix for site A against all other sites.

	B	C	D	E
A	0.7575	0.7564	0.8449	0.8308

¹ 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 R^2 equals 1.0, all points lie exactly on a straight line with no scatter; knowing X predicts Y perfectly.

Table 4 – Quarterly Data Capture

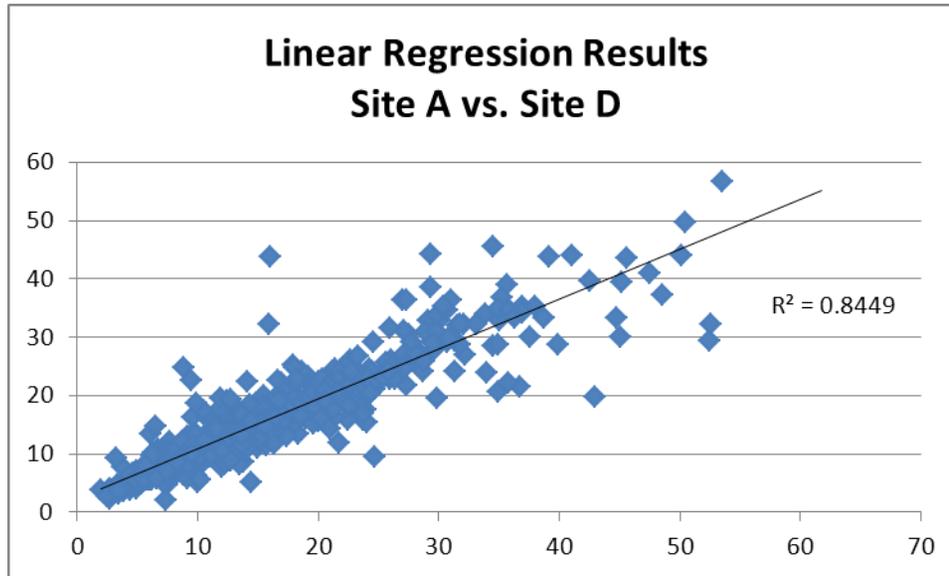
		Monitoring Sites				
		A	B	C	D	E
Quarterly Data Capture 2003-2010	2003 Q1		49%	97%	100%	97%
	2003 Q2		100%	100%	100%	93%
	2003 Q3		99%	100%	100%	97%
	2003 Q4	85%	92%	93%	97%	80%
	2004 Q1	97%	93%	100%	100%	100%
	2004 Q2	87%	100%	100%	97%	93%
	2004 Q3	84%	94%	97%	100%	77%
	2004 Q4	73%	93%	97%	100%	97%
	2005 Q1	93%	93%	87%	100%	100%
	2005 Q2	61%	100%	97%	97%	97%
	2005 Q3	53%	100%	100%	100%	97%
	2005 Q4	100%	100%	100%	100%	100%
	2006 Q1	100%	100%	100%	100%	100%
	2006 Q2	77%	80%	100%	100%	90%
	2006 Q3	100%	93%	100%	100%	94%
	2006 Q4	100%	100%	100%	100%	90%
	2007 Q1	97%	100%	100%	100%	97%
	2007 Q2	87%	100%	100%	100%	93%
	2007 Q3	97%	93%	90%	100%	97%
	2007 Q4	87%	100%	100%	100%	97%
	2008 Q1	100%	94%	94%	97%	94%
	2008 Q2	83%	100%	97%	100%	97%
	2008 Q3	90%	100%	100%	100%	100%
	2008 Q4	73%	100%	93%	100%	100%
	2009 Q1	97%	93%	100%	100%	100%
	2009 Q2	94%	100%	100%	100%	100%
	2009 Q3	83%	93%	100%	90%	100%
	2009 Q4	94%	93%	100%	97%	100%
2010 Q1	93%	100%	100%	100%	100%	
2010 Q2	93%	93%	100%	100%	100%	
2010 Q3	87%	94%	100%	100%	100%	
2010 Q4	87%	93%	100%	100%	100%	
MEAN	88%	95%	98%	99%	96%	

Table 5 – Paired Data Site

		29 - Quarter look (2003 Q4 - 2010 Q4)				
Site	Site	pairs Q1	pairs Q2	pairs Q3	pairs Q4	Total
A	B	102	84	83	101	370
A	C	198	175	180	200	753
A	D	201	172	181	201	755
A	E	202	169	173	196	740

Table 3 shows that the correlation between A and D ($R^2 = 0.8449$) is stronger than the correlation between site A and the other sites. From the mean of the quarterly data captured (Table 4), we can observe that site D has more data completeness (99%) from which to impute for site A. In addition, from the data site pairing (Table 5) it can be observed that site A and D have more paired data to establish a better relationship than site A and the other sites. Based on the information above, we will use multiple imputations and bootstrapping methods to generate the necessary data from Site D to impute on Site A missing values.

Linear Regression and Correlation for Site A vs. Site D



<i>Regression Statistics</i>	
Multiple R	0.919173221
R Square	0.844879411
Adjusted R Square	0.844673408
Standard Error	3.301408004
Observations	755

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2.218223276	0.22529593	9.845820463	1.34608E-21	1.775940467	2.660506086
X Variable 1	0.855135619	0.013352878	64.04129849	6.1227E-307	0.828922326	0.881348912

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	44701.14605	44701.14605	4101.287912	6.1227E-307
Residual	753	8207.168992	10.89929481		
Total	754	52908.31505			

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 $-5.148082e^{-4}$.

Table 5 - Bootstrapping Summary Statistics

Average	-5.1480820 e ⁻⁴
SD	0.1189829967
Max	0.4526497298
Min	-0.3519216842

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 D:

$$\text{Site A concentration} = \text{Intercept} + x \text{ Variable} * \text{Site D concentration} + \text{bootstrapping residual}$$

Where:

Intercept = 2.2182232761572 (from linear regression)
 x Variable = 0.855135618643403 (from linear regression)
 Bootstrapping residual = $-5.1480820307 e^{-4}$
 Site D concentration = values in Site D

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 39-081-0017). Table 6 shows Site 39-081-0017 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 the Steubenville-Weirton area 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	39-081-0017	15.9	16.4	13.8	16.2	14.3	12.1	12.7	15.4	15.5	14.7	14.2	13.0
NEW	39-081-0017	15.8	16.7	14.1	16.7	14.5	12.4	13.2	15.5	15.8	15.1	14.5	13.3

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/m³).

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.