

## Canton-Massillon PM<sub>2.5</sub> Nonattainment Area Monitor Missing Data Analysis

The current Canton-Massillon nonattainment area is located in northeast Ohio and includes Stark County.

The area has two monitors measuring PM<sub>2.5</sub> concentrations, which are operated by the Air Pollution Control Division of the Canton City Health Department.

### Annual Standard

A listing of the design values based on the three-year average of the annual mean concentrations from 2009 through 2011 is shown in Table 1. The design values calculated for the Canton-Massillon area show that the annual PM<sub>2.5</sub> NAAQS has been attained.

**Table 1 - Monitoring Data for the Canton-Massillon area for 2009 – 2011**

Site	County	Annual Standard			
		Year			Average 2009-2011
		2009	2010	2011	
39-151-0017	Stark	13.1	14.4	12.8	13.4
39-151-0020	Stark	11.9	13.8	11.3	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>

However, based on Section 107(d)(3)(E)(i) of the Clean Air Act (CAA), the PM<sub>2.5</sub> monitoring data 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 µg/m<sup>3</sup>. 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 monitor site 39-151-0017, located at 1330 Dueber Avenue, did not comply with the 75% data capture requirement in 2009. Specifically, the first quarter (January, February, and March) of 2009 has only 67% capture.

In order to comply with U.S.EPA 75% capture requirements, Ohio EPA prepared a statistical analysis using imputation and subsequent Bootstrap analysis. Missing values for site 39-151-0017 were generated and subjected to ordinary analysis as if the imputed values were real measurements (this approach is usually better than excluding subjects with incomplete data). Most methods for the accounting of incomplete data can be complex; the imputation/Bootstrap method prepared by Ohio EPA, however, is a relatively simple method to implement, even though the computations can be slow. This



39-153-0017 as Site C, and 39-153-0023 as Site D. It should be noted here that the second monitor in Canton, 39-151-0020, was not considered as a reference due to a lack of sufficient data in certain quarters of earlier years used in this analysis (Table 3).

## 1. Canton-Massillon Annual Design Value History

**Table 2 – Historic Design Values for Stark and Summit Counties, 2003 to 2010**

Site ID	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	2009-2011
39-151-0017	A	Stark	18.3	18.0	17.3	16.6	16.7	16.0	16.1	14.8	14.3	13.8	13.4
39-151-0020	B	Stark	16.9	16.4	15.8	15.0	15.2	14.2	14.3	12.9	12.9	12.7	12.3
39-153-0017	C	Summit	17.4	17.0	16.6	15.7	15.6	15.0	14.9	14.0	13.7	13.3	12.6
39-153-0023	D	Summit	16.2	16.3	15.6	15.0	14.6	14.1	14.1	13.1	12.7	12.3	11.7

 Less than 75% capture in at least one quarter  
 Violating Design Value

From Table 2, both monitors in Stark County have design values that meet the PM<sub>2.5</sub> annual standard since the 2006-2008 period. However, Site A has not proven clean data in 2009, and therefore it makes the entire nonattainment area ineligible for re-designation based on the 2009-2011 period. As mentioned previously, the lack of clean data in 2009 is due to the low percentage (67%) of data capture in the first quarter of 2009.

The imputation and Bootstrapping procedures will generate the necessary missing data to provide a re-calculated 2009-2011 design value for Site A.

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

Although location is a critical factor in determining the suitability of a monitor to serve as a reference for missing data imputation, a more rigorous statistical analysis was necessary to differentiate between Site C and Site D. To this end, three analyses were performed. Firstly, a linear regression of the reference monitor to Site A was performed. In a linear regression, the relationship between a dependent variable (Site A data), Y, and an independent variable, X (Site C or D data), is assessed. The familiar straight line regression model,  $Y = mX + b$  was used here. Under this model, linear regression finds the straight line that minimizes the sum-of-squares differences between the line and the Y data. The purpose of the regression analysis was to determine the degree of correlation between Site A and Site C and D. The statistic of interest,  $R^2$ , describes the degree of relationship between two variables<sup>1</sup> (variables or site concentrations in Site A and C and D).  $R^2$  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.

Secondly, the mean of the quarterly data captured (the mean of the percentage captured) allowed the central tendency of each site to be verified, providing a second

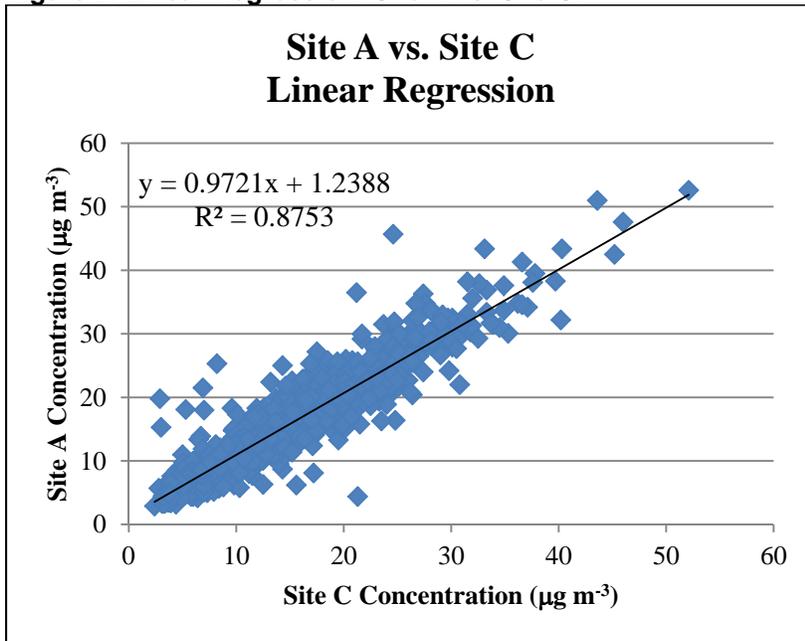
<sup>1</sup> An  $R^2$  value of 0.0 means that knowing X does not help to predict Y, there is no linear relationship between X and Y. When  $R^2$  equals 1.0, all points lie exactly on a straight line with no scatter; knowing X predicts Y perfectly.

means of determining what site (C or D) has a more complete data record from which to impute data for Site A.

Finally, although not as statistically significant as the correlation or mean of the percentage captured, pairing the site data seeks to reduce variability between data sets. Particular focus was placed on 2009-2011, the period for which the redesignation request is based upon and the period in which Site A demonstrated less than 75% capture in the first quarter of 2009.

Below are the results for Site A vs. C and for Site A vs. D.

**Figure 2: Linear Regression: Site A vs. Site C:**



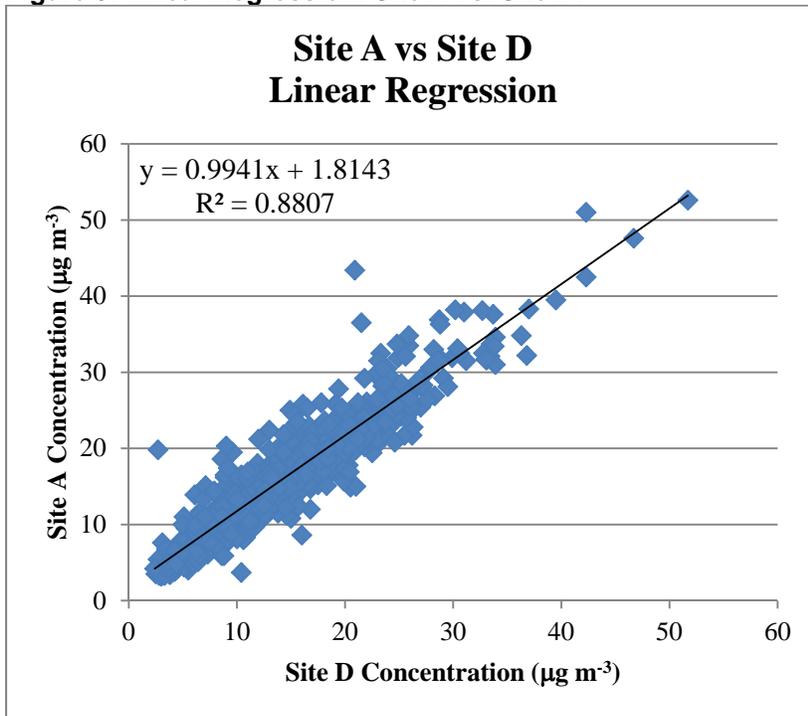
<i>Regression Statistics</i>	
Multiple R	0.935568718
R Square	0.875288826
Adjusted R Square	0.875193481
Standard Error	2.60878312
Observations	1310

**ANOVA**

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	62478.3728	62478.37	9180.234	0
Residual	1308	8901.920173	6.805749		
Total	1309	71380.29298			

	<i>Coefficient</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	1.23884	0.155837522	7.949532	4.02E-15	0.933116533	1.54455418	0.933116533	1.544554181
X Variable 1	0.97209	0.01014561	95.81354	0	0.95218335	0.99199025	0.95218335	0.991990246

**Figure 3: Linear Regression: Site A vs. Site D:**



<i>Regression Statistics</i>	
Multiple R	0.938477
R Square	0.88074
Adjusted R Square	0.880592
Standard Error	2.67247
Observations	812

**ANOVA**

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	42723.00423	42723	5981.858	0
Residual	810	5785.097836	7.1421		
Total	811	48508.10207			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	1.8142945	0.193758463	9.3637	7.42E-20	1.43396659	2.19462241	1.43396659	2.19462241
X Variable 1	0.99412681	0.012853569	77.342	0	0.968896579	1.01935704	0.968896579	1.019357044

**Table 3: Quarterly Data Capture**

		Monitoring Sites			
		A	B	C	D
Quarterly Data Capture 2003-2011	2003 Q1	93%	87%	86%	90%
	2003 Q2	97%	97%	89%	88%
	2003 Q3	90%	90%	91%	91%
	2003 Q4	87%	97%	90%	97%
	2004 Q1	87%	87%	98%	84%
	2004 Q2	93%	90%	92%	92%
	2004 Q3	97%	90%	96%	99%
	2004 Q4	70%	73%	98%	97%
	2005 Q1	80%	83%	92%	80%
	2005 Q2	84%	90%	94%	100%
	2005 Q3	93%	90%	100%	100%
	2005 Q4	87%	77%	90%	94%
	2006 Q1	93%	97%	100%	100%
	2006 Q2	90%	83%	100%	93%
	2006 Q3	97%	23%	100%	90%
	2006 Q4	90%	100%	97%	97%
	2007 Q1	97%	93%	97%	100%
	2007 Q2	93%	93%	100%	100%
	2007 Q3	74%	58%	100%	97%
	2007 Q4	30%	30%	87%	93%
	2008 Q1	0%	0%	87%	97%
	2008 Q2	40%	53%	100%	90%
	2008 Q3	84%	71%	100%	94%
	2008 Q4	97%	90%	100%	93%
	2009 Q1	67%	87%	77%	93%
	2009 Q2	91%	84%	97%	90%
	2009 Q3	99%	90%	91%	97%
	2009 Q4	93%	90%	100%	94%
	2010 Q1	84%	97%	99%	90%
	2010 Q2	88%	77%	100%	90%
	2010 Q3	96%	97%	100%	97%
	2010 Q4	96%	97%	100%	84%
2011 Q1	97%	97%	91%	100%	
2011 Q2	96%	97%	93%	93%	
2011 Q3	84%	87%	100%	90%	
2011 Q4	92%	97%	100%	100%	
<b>MEAN</b>		<b>84%</b>	<b>82%</b>	<b>95%</b>	<b>94%</b>
<b>MEAN: 2009-2011</b>		<b>90%</b>	<b>91%</b>	<b>96%</b>	<b>93%</b>

**Table 4: Paired Data by Site and Quarter**

SITE	All Quarters, 2003-2011				
	Pairs Q1	Pairs Q2	Pairs Q3	Pairs Q4	Total
<b>A vs C</b>	308	324	355	332	<b>1319</b>
<b>A vs D</b>	178	215	211	205	<b>809</b>

The R<sup>2</sup> value of the linear regression for Site A and Site C is 0.8753, and the R<sup>2</sup> value for Site A and Site D is 0.8807. Thus, the linear relationship between Site A and Site D is stronger than that of Site A and C, although this difference is marginal. By examination of the mean quarterly data capture (Table 3), in particular between 2009 and 2011, Site C demonstrates 96% data capture, and Site D 93% data capture. Lastly, Table 4 shows that significantly more data pairings occurred between Site A and Site C (1319 pairings) than the number of pairings between Sites A and D (809).

Based on the three statistical categories used to determine the reference monitor, Site C provided more data pairings, as well as a more complete data record, in particular over the 2009 to 2011 period. Although the R<sup>2</sup> value of Site A vs Site C was somewhat smaller than that of Site A and Site D, this difference was considered negligible, and Site C was used as the reference monitor in the data imputation procedure. However, due to the greater R<sup>2</sup> value between Site A and Site D, the data imputation procedure and Bootstrap was also performed using Site D as a reference, for the purposes of comparison and completeness.

### 3. Data Imputation and Bootstrap Analysis

Data imputation was conducted using the mathematical relationship established by the linear regression procedure between Site A and Site C, which takes the form:

$$Y = mX + b$$

where m is the slope, X the value recorded at the reference monitor, and b the intercept. For the imputation of missing values at Site A, m = 0.9721 and b = 1.2388. 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 39-151-0017). Table 5 shows Site 39-151-0017 before and after the imputation of missing data. It should be noted that both before and after inclusion of the imputed data, Site A demonstrated a passing design value for the 2009-2011 period (13.4 and 13.5, respectively).

**Table 5: Annual Design Values Before and After Imputation**

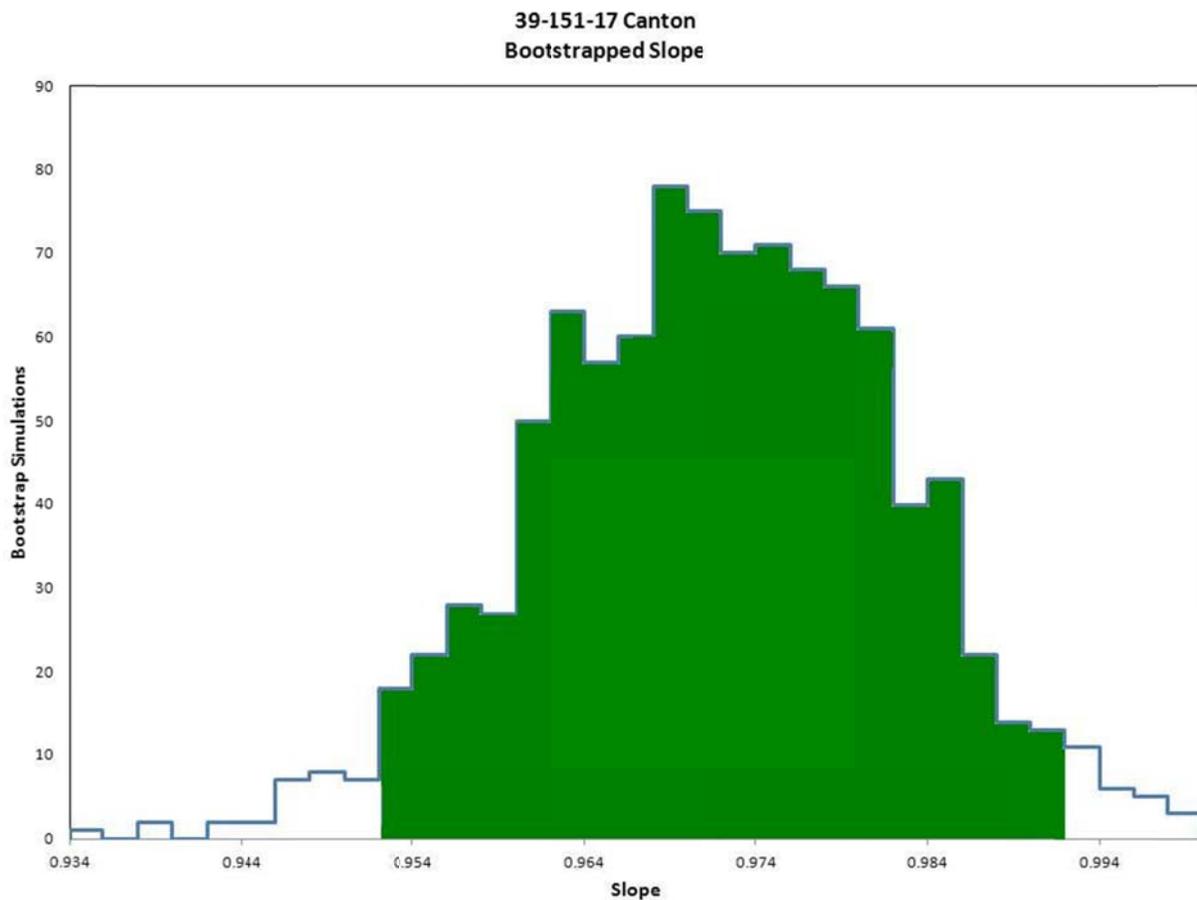
	Site ID	County	Year									Annual Design Value						
			2003	2004	2005	2006	2007	2008	2009	2010	2011	'03-'05	'04-'06	'05-'07	'06-'08	'07-'09	'08-'10	'09-'11
OLD	39-151-0017	Stark	16.8	15.5	17.8	14.6	15.9	13.9	13.1	14.4	12.8	16.7	16.0	16.1	14.8	14.3	13.8	13.4
NEW	39-151-0017	Stark	16.8	15.2	17.8	14.6	15.4	14.2	13.2	14.4	12.8	16.6	15.9	15.9	14.7	14.3	13.9	13.5

 Incomplete data (quarter with <75% data capture)

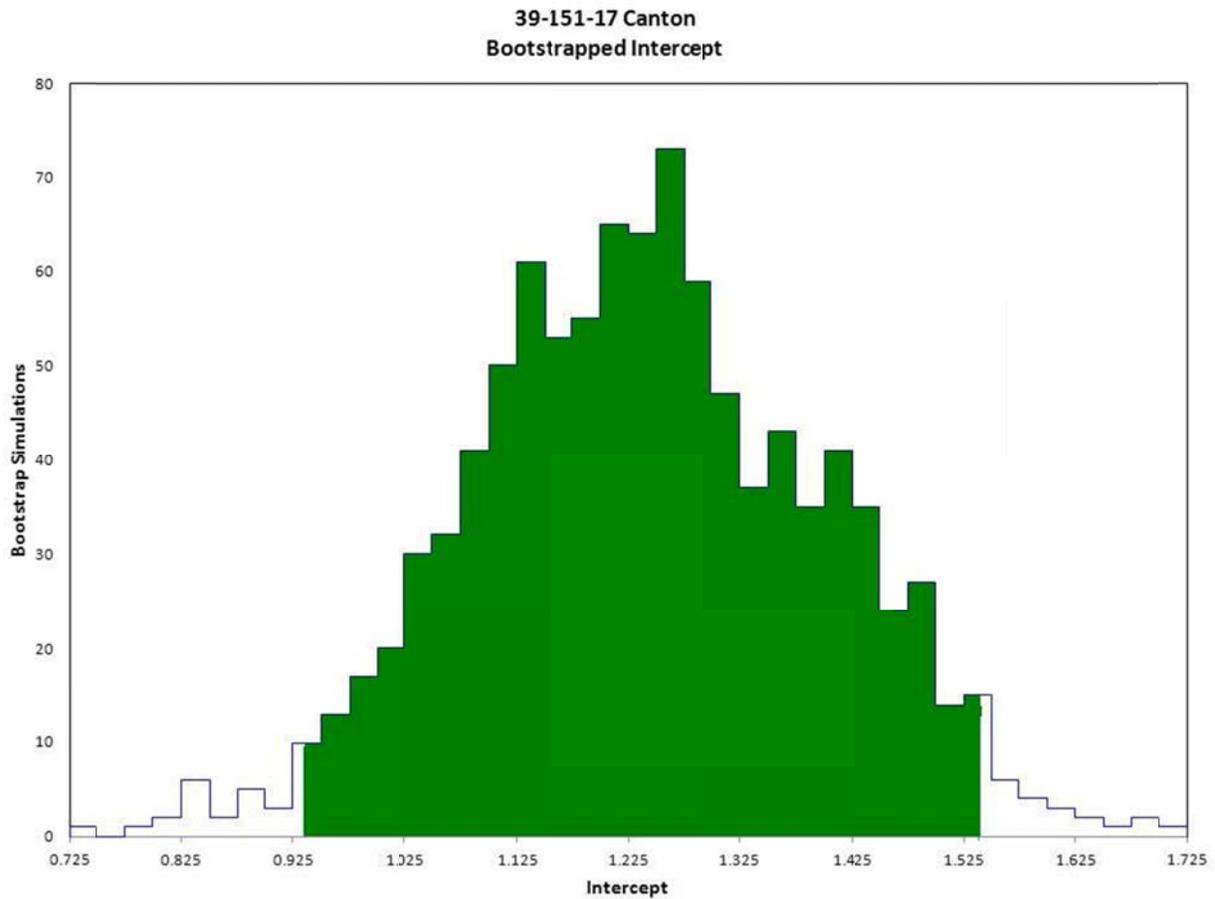
As stated previously, the imputation procedure was also conducted using Site D as the reference monitor for the sake of comparison. For 2004, 2007, 2008, and 2009, the annual averages using imputed data with Site D as the reference monitor were 15.1, 15.3, 14.3, and 13.0, respectively. These values are very similar to those imputed with Site C (15.2, 15.4, 14.2, and 13.2) as the reference, as shown in Table 5.

To provide a measure of confidence in the imputed data as well as the three-year average design values, a Bootstrap analysis was conducted. This analysis provided a mean, standard deviation, and 95% confidence interval for the slope (m) and intercept (b) used to generate the replacement data values at Site A. As detailed above, the Bootstrap analysis randomly resamples the real residuals from the regression analysis, adds those resampled residuals to the imputed data values, and subsequently calculates a new slope and intercept at each iteration. Thus, 1000 pairs of Bootstrapped slope and intercept values were calculated, from which a mean and 95% pseudo-confidence interval can be determined. The distribution of the slope and intercept from the Bootstrap analysis of Site A vs Site C are shown in Figures 4 and 5, respectively.

**Figure 4: Distribution of Slope Values**



**Figure 5: Distribution of Intercept Values**



The resultant mean of the slope,  $m_{boot} = 0.9722$ , and intercept,  $b_{boot} = 1.240$  from the Bootstrap analysis compare favorably to those values actually used for the data imputation performed at Site A,  $m = 0.9721$  and  $b = 1.2388$ . Additional confidence in the accuracy of the regression model used to impute missing data can be gained by using the upper and lower bounds of the 95% confidence interval to calculate a range of design values for all quarters in which Site A demonstrated less than 75% data capture. This analysis is summarized in Table 6.

**Table 6: Imputed Quarter Averages with Upper and Lower 95% Confidence Values**

	Quarter with <75% Data Capture					
	2004 Q4	2007 Q3	2007 Q4	2008 Q1	2008 Q2	2009 Q1
Lower 95%	13.26	18.20	13.43	15.68	12.08	15.27
<b>Imputed Value</b>	<b>13.4</b>	<b>18.4</b>	<b>13.8</b>	<b>16.3</b>	<b>12.4</b>	<b>15.4</b>
Upper 95%	13.56	18.55	14.08	16.91	12.70	15.49

From Table 6, it can be shown that in each quarter for which imputed data were used to replace missing values at Site A, the quarterly average was well within the 95% pseudo-

confidence interval resulting from the 1000 iteration Bootstrap analysis. This suggests that the slope and intercept used to generate the imputed values are representative of the relationship between Site A and Site C, and therefore suitable for generating replacements for missing data values. It should be stated that, even when using the calculated upper boundary for *all* quarters in the 2009-2011 period, the three year design value for this period is 13.9, and therefore still passing the annual PM<sub>2.5</sub> NAAQS of 15.0 µg m<sup>-3</sup>.

In summary, the PM<sub>2.5</sub> three-year average design value from 2009-2011 for monitor 39-151-0017 prior to the inclusion of imputed data was 13.4. Although this value was below the annual standard, the lack of clean data in the first quarter of 2009 made this value invalid. Therefore, an imputation and Bootstrap analysis was performed to replace missing values with valid numbers to achieve the 75% capture criteria. Incorporating these imputed values, a new design value of 13.5 for the 2009-2011 period was calculated. New design values for both Stark County monitors are summarized in Table 7.

**Table 7: Historic Design Values and Imputed Design Values, 1999-2011**

Site	County	Annual Design Values										
		1999-2001	2000-2002	2001-2003	2002-2004	2003-2005	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010	2009-2011
39-151-0017	Stark	18.3	18.0	17.3	16.6	16.7	16.0	16.1	14.8	14.3	13.8	13.4
39-151-0020		16.9	16.4	15.8	15.0	15.2	14.2	14.3	12.9	12.9	12.7	12.3
		incomplete data (quarter with <75% capture)										
		violating DV										

Site	County	New Annual Design Values Using Imputed Data										
		1999-2001	2000-2002	2001-2003	2002-2004	2003-2005	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010	2009-2011
39-151-0017	Stark	18.3	18.0	17.3	16.5	16.6	15.9	15.9	14.7	14.3	13.9	13.5
39-151-0020		16.9	16.4	15.8	14.9	15.2	14.5	14.6	13.5	13.2	13.0	12.3
		Imputed data substituted to compensate for <75% capture										
		violating DV										

## **24-hour Standard**

The 24-hour PM<sub>2.5</sub> standard is calculated as the three-year average of annual 98<sup>th</sup> percentile 24-hour average values, recorded at each monitor. As with the annual standard, U.S. EPA regulations require at least 75% data capture in each quarter of a consecutive 3-year period in order for the 24-hour standard to be valid, and, as with the annual standard, monitor 39-151-0017(Site A) did not meet this criteria due to a quarter of less than 75% capture in 2009. Using the same method described above to generate imputed values for the missing data at Site A, a new 24-hour design value was calculated. Table 8 shows the historic record of 24-hour design values for both PM<sub>2.5</sub> monitors located in Stark County.

**Table 8: Stark County 24-hour Design Values**

39-151-17	Year									24-hour Design Value						
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2003-2005	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010	2009-2011
Creditable Samples	111	106	111	111	89	67	320	111	336	39	39	38	35	34	34	30
98th Percentile	34.2	36.3	47.6	32.2	33.4	37.9	30	33	28.1	39	39	38	35	34	34	30

39-151-20	Year									24-hour Design Value						
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2003-2005	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010	2009-2011
Creditable Samples	112	104	104	92	83	65	107	112	114	36	33	33	30	30	30	28
98th Percentile	34.5	32.8	39.3	26.1	32.8	29.8	27.5	32.2	23.1	36	33	33	30	30	30	28

 <75% data capture in a least one quarter

For the 2009-2011 period, monitor 39-151-0020 demonstrates 75% or better data capture, and a valid, passing design value of 28. Monitor 39-151-0017, however, demonstrates an in-valid, passing design value of 30 for this same period of record. Using the same data set with imputed values for missing data at Site A derived from reference monitor 39-153-0017 (Site C) from which the annual design values were calculated, new annual 98<sup>th</sup> percentile values and three-year averages were also calculated from 2003 to 2011. These data are summarized in Table 9.

**Table 9: Monitor 39-151-0017 24-hour Design Values with Imputed Data**

39-151-17 OLD	Year									24-hour Design Value						
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2003-2005	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010	2009-2011
Creditable Samples	111	106	111	111	89	67	320	111	336	39	39	38	35	34	34	30
98th Percentile	34.2	36.3	47.6	32.2	33.4	37.9	30	33	28.1	39	39	38	35	34	34	30

39-151-17 NEW	Year									24-hour Design Value						
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2003-2005	2004-2006	2005-2007	2006-2008	2007-2009	2008-2010	2009-2011
Creditable Samples	111	122	111	111	113	119	334	111	336	39	39	38	35	34	34	30
98th Percentile	34.2	36.3	47.6	32.2	33.4	38.1	30.3	33	28.1	39	39	38	35	34	34	30

 <75% data capture in a least one quarter  
 Year includes one or more quarters with imputed values

Inclusion of imputed data significantly increased the number of creditable samples for each year in which monitor 39-151-0017 did not have sufficient data to meet the 75% data capture criteria, but this did not have a significant impact on the annual 98<sup>th</sup> percentile values or the three-year averages. With imputed data, monitor 39-151-0017 demonstrates a passing value of 30 for the 2009-2011 period, with sufficient data to meet the 75% capture criteria.