April 22, 2008

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NC 27711

MEMORANDUM

SUBJECT: Lead NAAQS Review: Development of Pb- PM_{10} to Pb-TSP Scaling Factors

FROM: Mark Schmidt, OAQPS/AQAD/AQAG
Kevin Cavender, OAQPS/AQAD/AAMG

TO: Lead NAAQS Review Docket (EPA-HQ-OAR-2006-0735)

General

This document describes analysis conducted by EPA-OAQPS to identify possible default "scaling factors" that could be used for converting monitored Pb-PM_{10} concentration data to surrogate Pb-TSP levels. The scaling factor estimates generated in this evaluation are simply numeric constants that could be applied to (i.e., multiplied by) Pb-PM_{10} concentration data reported in micrograms per cubic meter ($\mu g/m^3$) to approximate a Pb-TSP level in the same units. Note that two previous memos, also submitted to the current Lead NAAQS Review Docket, also address development of scaling factors (Cavender, 2007, 2008). The analysis described in this memorandum replicates, refines, and builds upon those two earlier preliminary evaluations. The following sections present details about the new and/or refined analysis, and reiterates pertinent information provided in the previous memos.

Data

Data on collocated Pb-TSP and Pb- PM_{10} data were obtained from the Air Quality System (AQS), EPA's repository of ambient air quality data. Due to varying method detection limit issues, the data used in the analysis were limited to those pairs of data where both the Pb-TSP and the Pb-PM_{10} reported measurements were above 0.01 $\mu g/m^3$. Further, only monitoring sites that reported at least 10 pairs of such data between the years 1993 and 2006 were considered. Based on these data requirements, we obtained around 1,200 pairs of collocated Pb-TSP and Pb-PM_{10} data for 22 sites for those years. In addition, we identified data from a study where collocated Pb-TSP and Pb-PM_{10} data were collected near a primary lead smelter in Montana (Brion, 1988). This data set contained data at much higher concentrations than the other data obtained from AQS. In total, 1,222 pairs of collocated Pb-TSP and Pb-PM_{10} daily (24-hour average) concentrations from 23 distinct monitoring sites, located in 19 counties in 7 states, were used in this analysis.

Source-oriented Classification

The older, non-AQS Montana smelter data were included in the study dataset to represent a high-end Pb "source-oriented" perspective. The term "source-oriented" refers to a specific point location (e.g., a monitoring site) and/or defined area that is heavily influenced by emissions of a pollutant(s) of interest (in this case, Pb); the influencing emissions can originate from a single facility or from combinations of sources. For Pb, the general presumption is that source-oriented environments contain relatively more ultra-coarse particles (i.e., those greater than typically captured with PM_{10} monitors) in the ambient air than do non-source environments. Since the EPA is giving consideration to using scaling factors with Pb-PM_{10} concentration data (i.e., to convert them to surrogate Pb-TSP levels) for official NAAQS comparisons, and further, because a proposed Pb monitoring network would probably encompass source- and non source-oriented monitoring locations, it was a goal of this scaling factor analysis to address both categories of sites. Thus, it was important to screen the 22 AQS (non-smelter) monitoring sites to see if any of them also could be regarded as source-oriented. For the purpose of this analysis, we characterized AQS sites as being source-oriented by: 1) comparing the cumulative Pb emissions within a mile radius of the monitoring site (based on emission estimates from the 2002 and/or previous year versions of the National Emissions Inventory, NEI) to a threshold minimum of 0.5 tons per year; 2) reviewing the monitoring sites' AQS classifications; and 3) viewing satellite/aircraft photos for each monitoring location. Based on these evaluations, we classified two additional (AQS) locations as being source-oriented. Thus, 3 of the 23 sites included in these analysis are deemed source-oriented, and the other 20 are assumed to be not source-oriented (i.e., they are monitors that are not dominated by emissions from nearby industrial sources of Pb).

Analysis Methods and Results

While the assembled collocated Pb-TSP and Pb-PM_{10} measurement study data are limited in quantity and in geographic scope, a number of relationships can be developed from them. Basic ratio and regression evaluations were performed on the collocated Pb concentration data, on the overall "all sites" dataset (all 1222 site-day pairs pooled together), on individual sites, and on various subsets and aggregations (e.g., source-oriented and not source-oriented). The chief goals of the analysis were to establish credible ranges for national-level scaling factors and to evaluate the appropriateness of such. The following subsection describes details of the regression analysis and the section thereafter describes the related ratio evaluations.

Regression Analysis

Figure 1 provides a scatter-plot and linear regression results for the entire 1,222 collocated pair ("all sites") database. The x-axis is the Pb-PM_{10} concentration in $\mu g/m^3$ and the y-axis is the Pb-TSP concentration (also in $\mu g/m^3$). The solid line represents the linear regression "best fit" which is the equation, Pb-TSP = 1.99 * Pb-PM_{10} - 0.02. The regression results show a strong linear relationship between the Pb-TSP and Pb-PM_{10} as indicated by the model R^2 of 0.86; associated F and t values indicate statistical significance in the model and

coefficient. Although the fit is tight, it appears as though there are outliers on the high side (one in particular) that are driving, perhaps exaggerating, the model results; these outliers are almost exclusively associated with the source-oriented monitoring sites. Additionally, there is an abundance of pairs on the extreme low end that might be skewing assumptions made from the relationship results if higher levels were of more interest.

Figure 2a re-plots the same ("all sites") 1,222 data pairs but differentiates between those associated with source-oriented sites (red) and the nonsource-oriented ones (blue). Figures 2b and 2c show data for the two categories separately. Separate linear regressions were run on the two categories and "best fit" result lines for both categories are plotted using the same colors as for the individual data points. The nonsource-oriented regression yields a Pb-PM_{10} coefficient of 1.06, an intercept of 0.01 $\mu g/m^3$, and an R^2 of 0.65. The source-oriented regression shows a Pb-PM_{10} coefficient of 2.00, an intercept of -0.03 $\mu g/m^3$, and an R^2 of 0.95. Thus, based on this dataset, the Pb-TSP: Pb-PM_{10} relationship is stronger (based on R^2) for source-oriented site data than for non-source-oriented site data. The two model intercepts appear comparable and are both "near zero", due to both categories having numerous data pairs with low values, as further discussed below. (A discussion of the phrase, "near zero" is also given some perspective in the following paragraph.) Because both regression model intercepts are "near zero", as is the one for the model generated from the combined "all sites" (1,222 pairs) dataset, the three model slope coefficients could seemingly be considered as central tendency estimates for default scaling factors. Thus, based on this evaluation, central tendency scaling factor estimates would be around 2.0 for both the "all sites" and source-oriented categories, and 1.1 for the non-source-oriented category. More complex evaluations are described below.

The non-source regression has a reduced fit statistic (R^2), as compared to the aggregate and source-only ones, because of the loss of the high value outliers. The source-oriented regression encompasses those high value outliers but also contains numerous low value pairs. Although the high value pairs (red) are predominately associated with source-oriented locations, the low value pairs emanate from both types of sites. Low values at source-oriented sites were not uncommon even during the general timeframe that the associated source facilities are thought to have operated with high emissions. The numerous low values seen at these source-oriented sites may be because the effecting Pb emission sources were not operating or not emitting as highly every day during the entire timeframe being evaluated. For example, a weekday:weekend effect might exist where low values most frequently occur on the weekend if a source facility operated only Monday through Friday. Meteorology might also explain such low values; although a monitoring site might be located in the predominant downwind direction of a source facility, the wind might occasionally blow from a cleaner direction.

The large numbers of low values seen in the datasets, especially for the nonsource-oriented sites, have implications on the interpretation of the regression results. The large quantities of low values can dominate the regression fit. However, in the context of scaling factors, hypothetical future monitoring sites with concentrations similar to these low values are not nearly as important in terms of practical implications as monitoring sites with concentrations similar to the larger, albeit fewer, higher concentration data points. Furthermore, although intercept values of -0.03, 0.01, and –0.01 $\mu g/m^3$ might seem "near zero" in comparison to a current NAAQS level of 1.5 $\mu g/m^3$, these intercept values are indeed substantial when noting

that for the risk assessment, the EPA has considered a NAAQS level as low as 0.02 $\mu g/m^3$, and further, that the average daily TSP value in the collocated database is only about 0.07 $\mu g/m^3$. Also, the envisioned NAAQS application of scaling factors does not include the use of an intercept term. Comparisons of regression slopes for different datasets or subsets can also be misleading when intercepts are substantially different. Because of these numerous factors, we decided to rerun the above noted regression analysis using a forced intercept of zero. Table 1 (columns J through M) shows the results of the initial and revised regressions. Because the R^2's computed from the regression analysis variation with a forced intercept of zero can be deceptive, they are not provided in Table 1. With a forced intercept of zero, the slopes for the overall and source-oriented categories (Table 1, rows 30 and 32) changed very little; in fact, when stated with one decimal place, neither changed (i.e., both are 2.0). The non-source-oriented slope (Table 1, row 31), however, changed (increased) by over 15% as shown in Table 1 with two decimal places, or (as stated with one decimal) increased from 1.1 to 1.2.

In addition to the regression influence associated with the presence of high and low values, there also exists a site-based influence in the Figure 1 and Figure 2 regressions because of varying observation counts for each site. For example, one study site contributes 205 pairs to the overall regressions plots but another site only contributes 10 pairs; thus, assuming similar concentration levels, the former site is essentially weighted by about 20-fold in comparison to the latter site. To give each site more "equal" weighting, regression analysis were also conducted on an individual site-level basis. Summary composite results for these 23 individual analyses provide alternative estimates for default scaling factors (for all sites, and for source / non-source subsets). Comparisons of the individual site-based results can provide indications of variability in scaling factor estimates (i.e., slopes) across individual sites and/or among different classifications of sites, and hence, give insight into the appropriateness of using aggregate or central tendency factors.

Table 1 provides analyses results for the 23 individual study sites (rows 1 through 23) as well as various summarizations of them. As discussed above, regressions were conducted with and without a forced intercept of zero. The R^2's provided are from the unconstrained analysis. In general, the R^2's for most of the 23 individual sites are quite high; the median R^2 is 0.82. (Medians of site-level statistics are shown in rows 27 through 29 of Table 1.) All three of the source-oriented sites have an R^2 of 0.90 or greater. There are several nonsource-oriented sites that have very poor relationships between daily concentrations of Pb-TSP and Pb-PM_{10}. Seven of the twenty nonsource-oriented sites have R^2 values of less than 0.50. However, half of the nonsource-oriented sites have R^2 values of 0.80 or more. Slopes (from the constrained analysis) also varied across sites but except for two sites, they all were in the range of 1.0 to 2.0; data from one (nonsource-oriented) site showed an unrealistic slope of 0.9 and the slope associated with the smelter site was (a realistic) 2.1. The median slope for "all sites" and for nonsource-oriented sites was 1.2 and for source-oriented sites it was 1.4.

Ratio analysis

To provide additional information on Pb-TSP:Pb-PM_{10} concentration relationships, simple ratio analysis techniques were also conducted on the paired Pb-TSP and Pb-PM_{10} data. As with the regression evaluations, the calculations were made on the overall ("all sites") data,

on individual sites, and various summarizations and subsets of each. The results of these evaluations are also provided in Table 1.

Columns N through U of Table 1 show distribution statistics for daily ratios of Pb-TSP to Pb-PM_{10}. For the overall dataset, this median daily ratio was 1.2 for the "all sites" data and the source-oriented data (Table 1, rows 30 and 32) and 1.3 for the nonsource-oriented site data (Table 1, row 31). For individual sites, the median daily Pb-TSP to Pb-PM_{10} ratios ranged from 1.0 to 1.8. The median site-level median daily ratio was 1.2 for "all sites" and non-source-oriented sites (Table 1, rows 27 and 29) and 1.8 for source-oriented sites (Table 1, row 28)

Columns V and W of Table 1 show the average concentration levels for Pb-TSP and Pb-PM_{10}, as computed with the study data; recall that the study data excludes any collocated pair where either the Pb-TSP or Pb-PM_{10} concentration is 0.01 $\mu g/m^3$ or less. Column X contains ratios of the Columns V and W mean values. These ratios can be considered as national scaling factor estimates. For the overall dataset, this ratio is 1.6 for the "all sites" data, 1.3 for the nonsource-oriented site data, and 1.8 for the source-oriented data; see Table 1, rows 30 through 32). For individual sites, the ratios (of means) range from 1.0 to 2.0. The three source-oriented sites had (rounded) ratios of 1.1, 1.6 and 2.0. The 20 nonsource-oriented sites had (rounded) ratios ranging for 1.0 to to 1.9; except for three of the 20 sites, the nonsource-oriented site ratio fall within a narrower range of 1.0 to 1.4.

References:

Cavender, USEPA, November 26, 2007. "Review of Collocated Lead In Total Suspended Particulate Matter Less Than Ten Micrometers". (http://www.epa.gov/ttn/naaqs/standards/pb/data/20071126_collocated_data_memo.pdf)

Cavender, USEPA, March 3, 2008. "Options for Lead NAAQS Indicator: Monitoring Implications". (http://www.epa.gov/ttn/naaqs/standards/pb/data/pb_indicator_implications.pdf)

Brion, Gail, USEPA, July 22, 1988. "Col-located PM-10/Hi-Vol Results for E. Helena".

Figure 1. Scatter-plot of cCollocated Pb-PM_{10} versus Pb-TSP

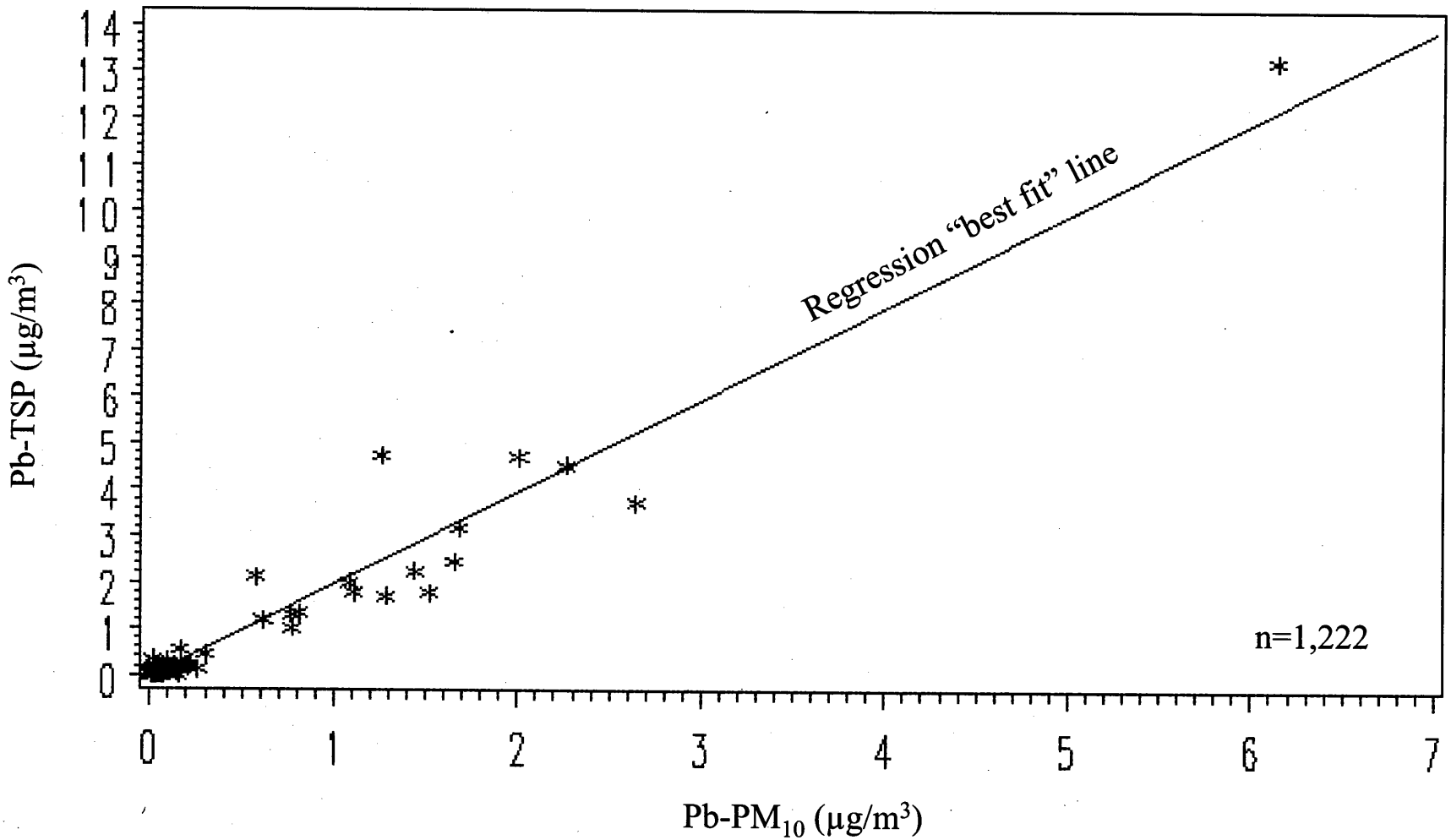

Figure 2a. Scatter-plot of collocated Pb-PM_{10} versus Pb-TSP – source-oriented site data (red) versus nonsource-oriented site data (blue)

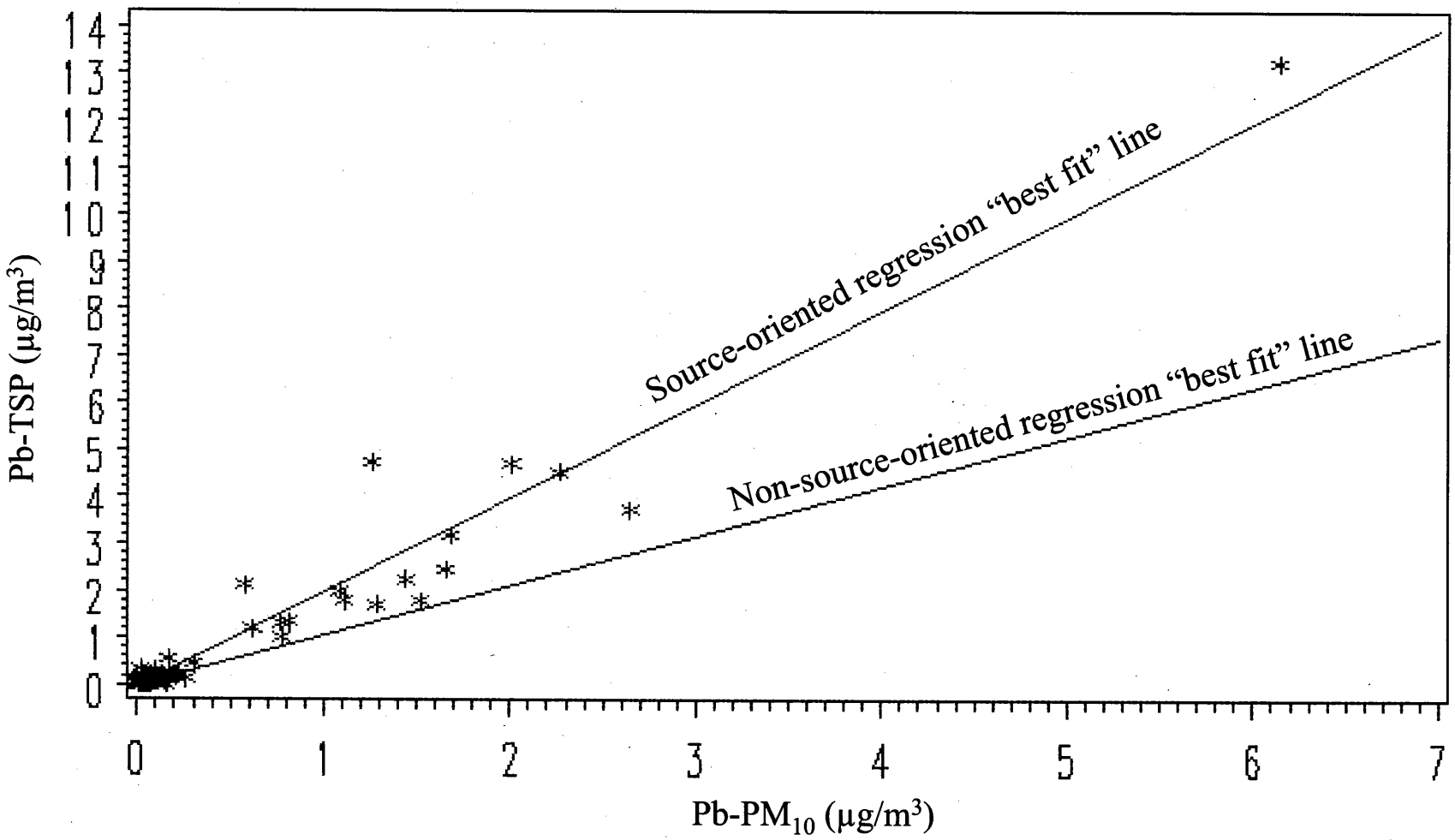

Figure 2b. Scatter-plot of collocated Pb-PM_{10} versus Pb-TSP – nonsource-oriented site data

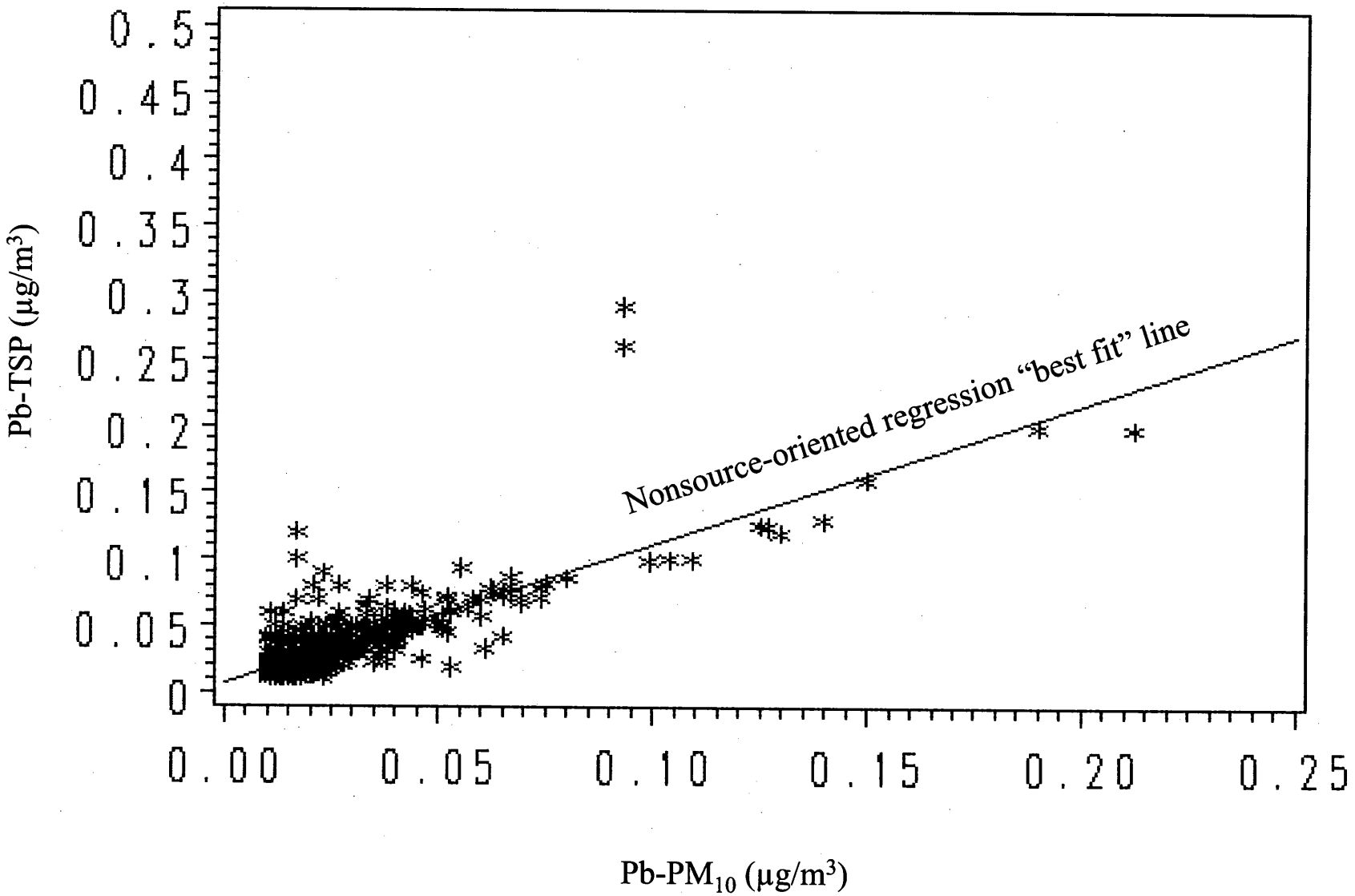

Figure 2c. Scatter-plot of collocated Pb-PM_{10} versus Pb-TSP – source-oriented site data

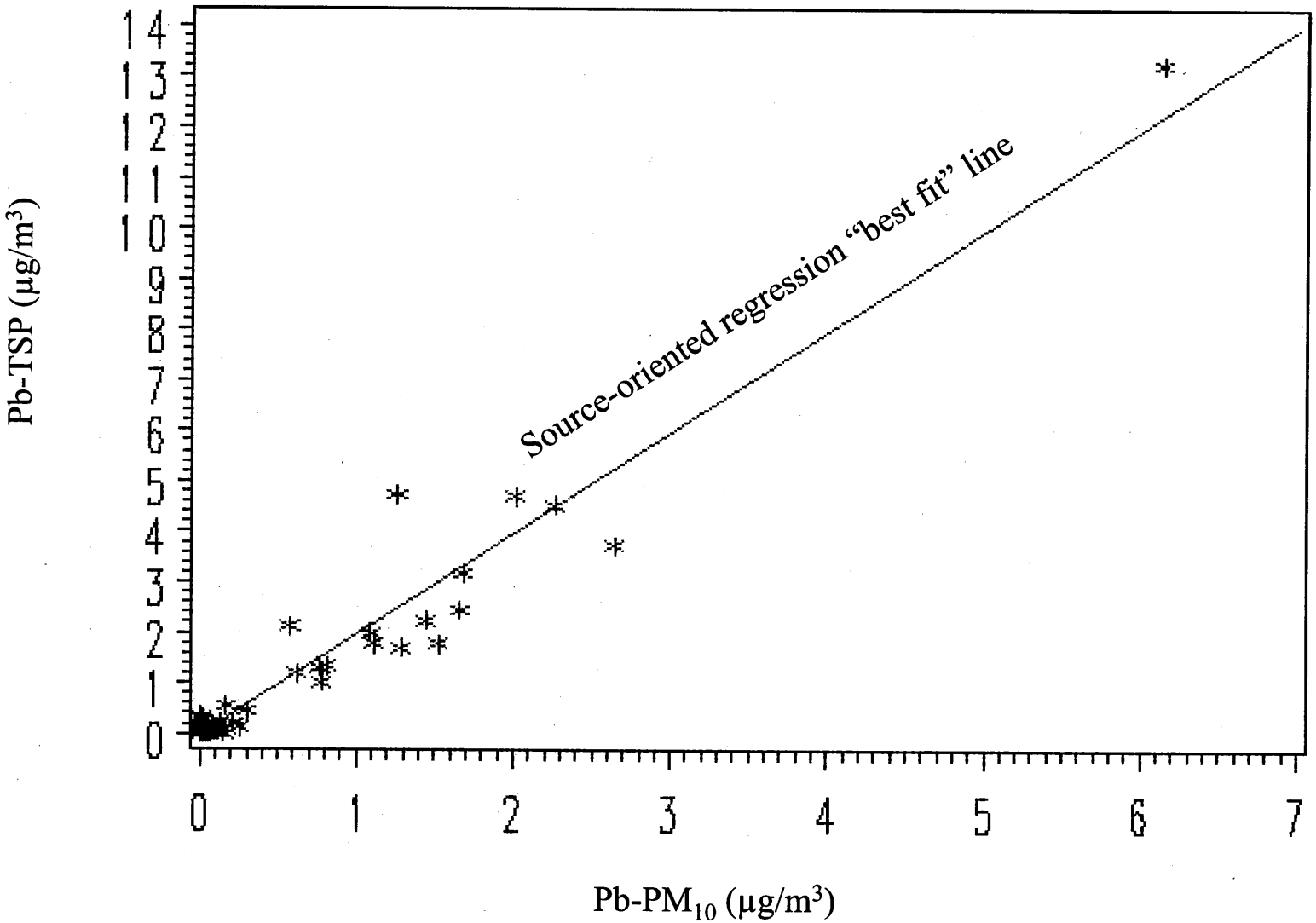

Table 1. Scaling factor analysis statistics

	A	B	C	F	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
							Linear regression results (Pb-TSP = Pb-PM10) w/ out a forced intercept				Daily ratio (Pb-TSP / Pb-PM10) statistics										
	Site	Land Use	Source-oriented?	Timeframe of data collection (years)	Number of sample pairs	r (correlation coefficient) of daily values	Intercept	Slope (Pb-PM_{10} coefficient)	R^2	Slope of linear regression of Pb-TSP = Pb-PM10 w/ forced intercept of zero	min	p5	p25	mean	median	p75	p95	max	Pb-TSP average concentrationg /m3)	Pb-PM10 average concentration (ug/m3)	Ratio of mean concentrations (Pb-TSP / Pb-PM10)
	Sites with at least 10 pairs of same-day Pb-TSP and Pb-PM10 24-hr samples where reported values exceed 0.01 ug/m3 for both Pb-TSP and Pb-PM10.																				
1	060190008	RESIDENTIAL	No	1995 - 2001	32	0.94	0.00	1.00	0.89	1.08	0.75	0.87	1.00	1.12	1.08	1.16	1.63	1.67	0.02	0.02	1.10
2	060250005	RESIDENTIAL	No	1996 - 2001	205	0.95	0.00	0.98	0.91	1.09	0.65	0.80	1.06	1.20	1.18	1.30	1.59	3.73	0.03	0.03	1.15
3	060290014	MOBILE	No	1995 - 2000	32	0.98	0.00	0.95	0.95	1.03	0.69	0.73	1.00	1.13	1.11	1.27	1.45	1.55	0.02	0.02	1.08
4	060374002	RESIDENTIAL	No	1995 - 2000	129	0.76	0.01	1.46	0.57	1.70	0.61	1.05	1.35	1.99	1.67	2.31	4.17	7.06	0.04	0.02	1.88
5	60658001	RESIDENTIAL	No	1995 - 1996	54	0.44	0.02	0.77	0.19	1.67	0.86	1.00	1.42	1.87	1.60	2.35	3.08	4.12	0.03	0.02	1.78
6	060771002	RESIDENTIAL	No	1995 - 2000	53	0.82	0.00	1.05	0.68	1.23	0.54	0.80	1.08	1.33	1.31	1.50	2.08	2.58	0.02	0.02	1.29
7	060850004	RESIDENTIAL	No	1994 - 1999	23	0.72	0.01	0.93	0.52	1.29	0.64	0.65	1.18	1.46	1.36	1.69	2.64	2.65	0.02	0.02	1.39
8	060990002	RESIDENTIAL	No	1995 - 1998	16	0.27	0.01	0.37	0.07	1.14	0.84	0.84	0.97	1.20	1.14	1.33	2.08	2.08	0.02	0.01	1.17
9	201330002	AGRICULTURAL	No	1993 - 1998	13	0.90	0.00	1.28	0.82	1.20	0.69	0.69	0.85	1.18	1.23	1.40	1.82	1.82	0.02	0.02	1.18
10	201730007	COMMERCIAL	No	1993 - 1997	18	0.08	0.01	0.25	0.01	1.32	0.79	0.79	1.06	1.38	1.18	1.45	3.00	3.00	0.02	0.01	1.35
11	201730008	COMMERCIAL	No	1993 - 1997	15	0.03	0.01	0.08	0.00	1.15	0.79	0.79	0.93	1.17	1.14	1.33	2.00	2.00	0.02	0.01	1.16
12	201730009	RESIDENTIAL	No	1993 - 1997	18	0.50	0.01	0.80	0.25	1.28	0.65	0.65	1.08	1.36	1.25	1.64	2.64	2.64	0.02	0.01	1.32
13	201731012	INDUSTRIAL	No	1993 - 1997	21	0.99	0.00	1.00	0.98	1.03	0.67	0.74	1.00	1.11	1.08	1.27	1.46	1.60	0.02	0.02	1.07
14	201770007	RESIDENTIAL	No	1993 - 1997	19	0.67	0.00	1.00	0.45	1.20	0.75	0.75	0.92	1.25	1.27	1.50	2.00	2.00	0.02	0.01	1.22
15	202090015	INDUSTRIAL	No	1993 - 1997	118	0.65	0.01	0.85	0.42	1.23	0.36	0.68	1.07	1.43	1.30	1.67	2.56	3.42	0.03	0.02	1.33
16	260770905	COMMERCIAL	No	1993 - 1996	78	0.99	0.00	0.96	0.97	1.07	0.65	0.93	1.07	1.21	1.18	1.33	1.50	1.83	0.02	0.02	1.15
17	261390009	COMMERCIAL	No	2000 - 2001	26	0.94	0.01	0.95	0.88	1.15	0.94	0.96	1.08	1.36	1.31	1.53	2.05	2.15	0.03	0.02	1.30
18	270530053	COMMERCIAL	No	1997 - 2001	13	0.93	0.00	1.97	0.86	1.81	1.13	1.13	1.43	1.76	1.56	2.11	3.05	3.05	0.03	0.02	1.77
19	295100085	RESIDENTIAL	No	2004 - 2004	26	0.98	-0.01	1.89	0.95	1.35	0.88	0.98	1.00	1.16	1.00	1.47	1.55	1.69	0.03	0.02	1.24
20	490110001	COMMERCIAL	No	2003 - 2003	19	1.00	0.00	0.93	1.00	0.94	0.91	0.91	0.93	0.97	1.00	1.00	1.04	1.04	0.04	0.04	0.95
21	202090020	INDUSTRIAL	Yes	1993 - 1997	107	0.99	0.01	1.39	0.98	1.40	0.08	0.65	1.24	2.08	1.63	2.05	3.22	25.50	0.10	0.06	1.55
22	261630033	INDUSTRIAL	Yes	2003 - 2006	167	0.90	0.00	0.97	0.81	1.04	0.37	0.84	1.00	1.14	1.07	1.23	1.67	3.04	0.03	0.03	1.10
23	Unknown	Smelter	Yes	1988	20	0.97	-0.18	2.14	0.94	2.07	0.54	0.56	1.50	2.10	1.79	2.25	4.88	6.00	2.46	1.23	1.99
24	Average of all site-level statistics			1988-2006	23 sites	0.76	0.00	1.04	0.66	1.28	0.69	0.82	1.10	1.39	1.28	1.57	2.31	3.75	0.13	0.07	1.80
25	Average of non-source site-level statistics		No	1993-2006	20 sites	0.73	0.01	0.97	0.62	1.25	0.74	0.84	1.07	1.33	1.25	1.53	2.17	2.58	0.02	0.02	1.28
26	Average of source-oriented site-level statistics		Yes	1988-2006	3 sites	0.95	-0.06	1.50	0.91	1.50	0.33	0.68	1.25	1.78	1.49	1.84	3.26	11.51	0.86	0.44	1.95
27	Median of all site-level statistics					0.90	0.00	0.97	0.82	1.20	0.69	0.80	1.06	1.25	1.23	1.47	2.05	2.58	0.02	0.02	1.24
28	Median of non-source site-level statistics		No			0.61	0.00	0.83	0.38	1.09	0.65	0.74	0.99	1.17	1.13	1.32	1.58	1.79	0.02	0.02	1.15
29	Median of source-oriented site-level statistics		Yes			0.97	0.00	1.39	0.94	1.40	0.37	0.65	1.24	2.08	1.63	2.05	3.22	6.00	0.10	0.06	1.55
30	Results for all data points (pooled)			1988-2006	1222	0.97	-0.02	1.99	0.95	1.98	0.08	0.81	1.05	1.44	1.24	1.55	2.64	25.50	0.07	0.05	1.63
31	Results for non-source data points (pooled)		No	1993-2006	928	0.81	0.01	1.05	0.65	1.22	0.36	0.81	1.07	1.40	1.25	1.53	2.58	7.06	0.03	0.02	1.33
32	Results for soure data points (pooled)		Yes	1988-2006	294	0.97	-0.03	2.00	0.95	1.99	0.08	0.81	1.03	1.55	1.20	1.65	2.80	25.50	0.22	0.12	1.79