Message threads between Statistician and Project Lead to develop final Equation for January 2011 revision of AP-42 section 13.2.1 Paved Roads.

Ron Myers to Benjamin Wells – 10/13/2010 with 5 attached files

Benjamin Wells to Ron Myers – 10/28/2010 with 6 attached files

Ron Myers to Benjamin Wells – 11/26/2010 with one attached file

Benjamin Wells to Ron Myers – 11/30/2010 (11:37am) (includes 11/26/2010 e-mail)

Ron Myers to Benjamin Wells – 11/30/2010 (4:43pm) (includes 11/26/2010 & 11/30/2010 e-mail)

Ron Myers to Benjamin Wells – 12/02/2010

Benjamin Wells to Ron Myers – 12/03/2010 with 4 attached files.

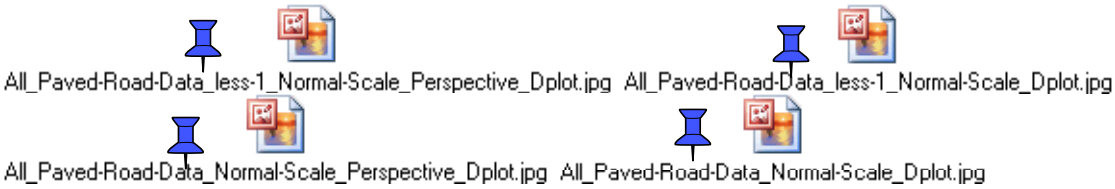
Spreadsheet with Paved Road data

Ron Myers to: Benjamin Wells, Phil Lorang
Cc: Bob Schell, Mark Schmidt

10/13/2010 09:57 AM

Ben:
As requested during the meeting I have attached a spreadsheet with the paved road data. I am only including the parameters that have been used in the past for predicting emissions. There are other meta data associated with the tests but these are not consistently collected (such as number of wheels) or have no influence on the emissions (such as ambient temperature). If you think you will need these, I can assemble them.

Also, for your information I have produced figures that show the data in 3D and 2D. The 2D figures also identify the data name for that point.



Give me a call if you have questions.

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov

Re: Spreadsheet with Paved Road data

Benjamin Wells to: Ron Myers 10/28/2010 03:44 PM

Cc: Bob Schell, Mark Schmidt, Phil Lorang

Hi Ron,

I took a look at the paved road data. Here is a brief summary of what I did/found:

First, I removed the 'Z3' point that looked like an outlier. This point was not considered in any models.

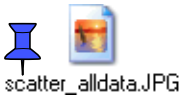
Next, I fit the log-linear model log(PM10) = b0 + b1*log(silt) + b2*log(weight). I should point out that I did not use silt/2, weight/3, or speed/30 as was done in the original publication. Multiplying or dividing the data by a constant is called 'scaling' and does not have any effect on the regression coefficients (the b's), EXCEPT for b0, which is not of interest to us.

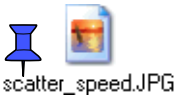
Then I removed all of the points with missing values for speed and fit several variations of the above model with silt, speed, and weight. Silt was always the best predictor, by far. The silt/speed model performed slightly better than the silt/weight model.

However, I would not recommend a model with silt/speed/weight. There is a strong negative correlation between the speed/weight terms, and it causes a phenomenon called multicollinearity. Multicollinearity is best avoided because it means that adding or removing just a few data points can have a huge effect on the parameter estimates, causing the model to behave wildly.

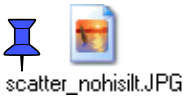
It is possible that both the speed and weight terms are truly meaningful, but we cannot determine this from the available data. To determine if both parameters are meaningful, we would need more data, especially from heavy vehicles moving at high speeds and light vehicles moving at low speeds. Based on the data we have, I can only recommend that we use a log-linear model with silt/weight or a log-linear model with silt/speed.

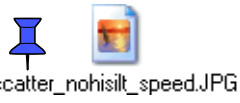
Here is a copy of the model output for each of the different models I looked at, and pairwise scatter plots of the predictors, one with and one without the missing speed values.





Phil suggested that I try removing the high silt values, as they are not realistic of the present-day situation. Attached below is the same analysis, removing all silt values larger than 20 g/m^2. This resulted in 10 additional data points being removed. The effect was that the relationship between PM10 and silt was weakened slightly, but overall the results did not change appreciably.





We can get together and talk after you have had a chance to look at all of the output. I understand that the output by itself is not 100% self-explanatory, so I'll be happy to answer any questions you might have. I will be out tomorrow, but I will be around next week if you'd like to schedule a meeting.

Benjamin Wells, Statistician
US EPA, OAQPS-AQAD-AQAG
Research Triangle Park, NC
919 - 541 - 7507

wells.benjamin@epa.gov

Ron Myers	Ben: As requested during the meeting I have att...	10/13/2010 09:57:48 AM

From: Ron Myers/RTP/USEPA/US
To: Benjamin Wells/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA
Cc: Bob Schell/RTP/USEPA/US@EPA, Mark Schmidt/RTP/USEPA/US@EPA
Date: 10/13/2010 09:57 AM
Subject: Spreadsheet with Paved Road data

Ben:
As requested during the meeting I have attached a spreadsheet with the paved road data. I am only including the parameters that have been used in the past for predicting emissions. There are other meta data associated with the tests but these are not consistently collected (such as number of wheels) or have no influence on the emissions (such as ambient temperature). If you think you will need these, I can assemble them.
[attachment "Paved_Road_Test_Data.xls" deleted by Benjamin Wells/RTP/USEPA/US]

Also, for your information I have produced figures that show the data in 3D and 2D. The 2D figures also identify the data name for that point.

[attachment "All_Paved-Road-Data_less-1_Normal-Scale_Perspective_Dplot.jpg" deleted by Benjamin Wells/RTP/USEPA/US] [attachment "All_Paved-Road-Data_less-1_Normal-Scale_Dplot.jpg" deleted by Benjamin Wells/RTP/USEPA/US] [attachment "All_Paved-Road-Data_Normal-Scale_Perspective_Dplot.jpg" deleted by Benjamin Wells/RTP/USEPA/US] [attachment "All_Paved-Road-Data_Normal-Scale_Dplot.jpg" deleted by Benjamin Wells/RTP/USEPA/US]

Give me a call if you have questions.

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov

Subject:	**Paved Road Equation development**
Date:	11/26/2010 04:07 PM
From:	Ron Myers/RTP/USEPA/US
To:	
Cc:	Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA
Bcc:	
Sent by:	Ron Myers/RTP/USEPA/US

Distribution List:

Ben:

Here is a spreadsheet with the final paved roads data. This is in the tab labeled “Test Data & Parameters”. This two left columns in this tab are the test run ID and the measured PM10 emissions both which you may not need but provide linkage to what were in the test reports. The remainder columns are the reported independent variables speed, weight and silt loading with the associated Road Dust emissions (dependent variable). There are other tabs with additional ancillary information that include calculations and Excel output which I describe in my e-mail.

I have some concerns over which of the data to use in developing the final equation. My goal is to provide an equation that provides a good predictive estimate of emissions from public roads with silt loadings typical of most paved roads in the US. Data that we use for the National Emissions Inventory has default silt loadings of between 0.015 and 0.6 g/m2 for most of the year and silt loadings between 0.015 and 2.5 for months with frozen precipitation. While I am not positive, I believe the majority of the emissions inventory (miles driven) is based upon roads with lower silt loadings. For most of the nation the average vehicle weights are between 2 tons and a little more than 4 tons. A secondary goal is to provide a reasonably good predictive estimate from public and private roads with higher silt loadings and also with higher vehicle weights. These are values that are typical of the current estimates used in permit applications. All of the silt loading data in the AP-42 section are quite dated and probably not reflective of current housekeeping practices. I think it is reasonable to assume that the lack of very high silt loading data in the test data and the large gap between 17 g/m2 and 53 g/m2 are indicative that the highest data are anomalous not to mention that all of the tests with high silt loadings were conducted in 1982.

Using Excel, I have explored some of the potential selection of the data to achieve these goals.

First, I have excluded test Z3 (weight of 8 tons, silt loading of 12.4 g/m^2 and PM10 emissions factor of 1819 g/VMT) since it looks like an outlier on a 3D graphic. Is there a means by which we can justify this exclusion with statistical means? Might there be other data points which should be

considered for exclusion? I assume that identification of outliers may be difficult as there seems to be a very significant variation in the data across all weight and silt loading values.

Next, I have recognized that the speed term does not add much value to the final equation. I based that conclusion on a general assessment of the average percentage error of the equation could achieve about the same value without the use of the speed term. Can you provide a more statistically valid rationale for using or not using the speed term as an additional independent variable?

I received two comments that suggested developing separate equations for different classes of data. One class that was identified was the large number of tests conducted at very low speeds (1 and 5 mph). Because of their numbers, they may adversely bias the final equation. Another class that was identified was high weight vehicles. Having multiple equations for different classes was a problem until 1995 when all the data were combined for a single equation. The problem was that people would select the equation that gave them the answer they wanted and then they went to extremes to justify this selection. It also set up disagreements which were difficult to resolve without a long argument. I am not adverse to different classes as long a the overlap between the classes produces about the same emissions. I have attempted to divide the test data to separate data for different scenarios but was unable to find equations that performed acceptably at or near the overlapping conditions (average vehicle weight, silt loadings). Also, by parsing out the “special test data” the correlation coefficient degraded significantly for each of the different categories and it was difficult to define a bright line to separate the categories. Again my decision was based upon the correlation coefficient, the P-values for the individual terms and the Standard Errors for the terms. Again, my decision to keep only one equation is based upon engineering judgment based upon the goals I stated at the beginning of the second paragraph. Is there a means in SAS to make an assessment that would support (or reject my hypothesis) that the data are all related and should continue to be assessed as a single group?

In my data selection, I excluded the use of all data above 20 g/m^2 silt loading. In my judgment (looking at the individual and combined differences from the actual data) the equation developed from this reduced data performed as well on the ten data that were excluded as if I would have used them in the analysis. The equation developed without the high silt data performed better on the lower silt data. Another area is the large number of tests at low speeds. I tried several regressions with only half of the low speed data and averaging the parameters for the equation. That was very resource intensive and I don’t think it generated a better equation for any of the data. Is there some means in SAS where one could evaluate the ability for an equation to perform better for some conditions and still incorporate a reasonable performance for other conditions (i.e. decreased weighting for higher silt loadings, vehicle speeds and/or greater weights)?

If the preceding is not possible in SAS, is there a statistical basis for using only the data with silt loadings below 20 g/m^2 and excluding the 10 data with the very highest silt loadings for the development of the equation. But, incorporating these higher silt loading tests in the assessment of the predictive accuracy of the equation?

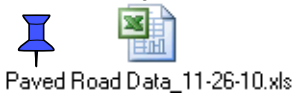
When I used Excel to determine the equation and used only silt loading and average vehicle weight, it produced an equation with a correlation of 0.56, P-values of less than 0.03. See the tab “Output - sL, W” for the Excel equation terms, residuals and graphs. I looked at the residual graphs for silt and they did not seem to curve any and averaged very close to zero. The

residual graph for weight did appear to curve some. I then included the weight term squared as an addition to weight and silt. See the tab “Output - sL, W, W^2”. The correlation improved and the P-values went up to less than 0.035. However, upon reviewing how I squared the term, I was squaring the natural log of the weight and not the weight. When I changed the term, Excel was unable to generate an equation and I could not determine a reason. Also, the form of the equation may be problematic from a calculation standpoint. Your insights on the benefits of this addition are welcome.

Perhaps I am over thinking the issues. I'll call you or come by your office to discuss these issues and what I can expect from you to include in my background report.

_____

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov



Re: Paved Road Equation development

Benjamin Wells to: Ron Myers 11/30/2010 11:37 AM
Cc: Bob Schell, Phil Lorang

From:	Benjamin Wells/RTP/USEPA/US
To:	Ron Myers/RTP/USEPA/US@EPA
Cc:	Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA
History:	This message has been replied to.

Ron,

See comments in bold below.

Benjamin Wells, Statistician
US EPA, OAQPS-AQAD-AQAG
Research Triangle Park, NC
919 - 541 - 7507
wells.benjamin@epa.gov

Ron Myers | Ben: Here is a spreadsheet with the final paved r... | 11/26/2010 04:07:27 PM

From:	Ron Myers/RTP/USEPA/US
To:	Benjamin Wells/RTP/USEPA/US@EPA
Cc:	Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA
Date:	11/26/2010 04:07 PM
Subject:	Paved Road Equation development

Ben:

Here is a spreadsheet with the final paved roads data. This is in the tab labeled “Test Data & Parameters”. This two left columns in this tab are the test run ID and the measured PM10 emissions both which you may not need but provide linkage to what were in the test reports. The remainder columns are the reported independent variables speed, weight and silt loading with the associated Road Dust emissions (dependent variable). There are other tabs with additional ancillary information that include calculations and Excel output which I describe in my e-mail.

I have some concerns over which of the data to use in developing the final equation. My goal is to provide an equation that provides a good predictive estimate of emissions from public roads with silt loadings typical of most paved roads in the US. Data that we use for the National Emissions Inventory has default silt loadings of between 0.015 and 0.6 g/m2 for most of the year and silt loadings between 0.015 and 2.5 for months with frozen precipitation. While I am not positive, I believe the majority of the emissions inventory (miles driven) is based upon roads with lower silt loadings. For most of the nation the average vehicle weights are between 2 tons and a little more than 4 tons. A secondary goal is to provide a reasonably good predictive estimate from public and private roads with higher silt loadings and also with higher vehicle weights. These are values that are typical of the current estimates used in permit applications. All of the silt loading data in the AP-42 section are quite dated and probably not reflective of current housekeeping practices. I think it is reasonable to assume that the lack of very high silt loading data in the test data and the large gap between 17 g/m2 and 53 g/m2 are indicative that the highest data are anomalous not to mention that all of

the tests with high silt loadings were conducted in 1982.

Using Excel, I have explored some of the potential selection of the data to achieve these goals.

First, I have excluded test Z3 (weight of 8 tons, silt loading of 12.4 g/m^2 and PM10 emissions factor of 1819 g/VMT) since it looks like an outlier on a 3D graphic. Is there a means by which we can justify this exclusion with statistical means? Might there be other data points which should be considered for exclusion? I assume that identification of outliers may be difficult as there seems to be a very significant variation in the data across all weight and silt loading values.

Data exclusion is really more of an art than a science. Statistical methods for data exclusion exist but in my experience they do not perform any better than 'common sense' approaches. My recommendation is that you decide on range of values that constitute reasonable silt loadings should be, provide a scientific rationale for your decision, then exclude all silt loading values outside of that range. The 3D scatter plot of weight, silt loading, and PM10 emissions factors should be sufficient to demonstrate that the Z3 test is a clear outlier and should be excluded.

Next, I have recognized that the speed term does not add much value to the final equation. I based that conclusion on a general assessment of the average percentage error of the equation could achieve about the same value without the use of the speed term. Can you provide a more statistically valid rationale for using or not using the speed term as an additional independent variable?

Yes, the speed term should not be excluded because speed is highly correlated with weight. Speed and weight are providing the same information to the model and it is only necessary to use one of them.

I received two comments that suggested developing separate equations for different classes of data. One class that was identified was the large number of tests conducted at very low speeds (1 and 5 mph). Because of their numbers, they may adversely bias the final equation. Another class that was identified was high weight vehicles. Having multiple equations for different classes was a problem until 1995 when all the data were combined for a single equation. The problem was that people would select the equation that gave them the answer they wanted and then they went to extremes to justify this selection. It also set up disagreements which were difficult to resolve without a long argument. I am not adverse to different classes as long a the overlap between the classes produces about the same emissions. I have attempted to divide the test data to separate data for different scenarios but was unable to find equations that performed acceptably at or near the overlapping conditions (average vehicle weight, silt loadings). Also, by parsing out the “special test data” the correlation coefficient degraded significantly for each of the different categories and it was difficult to define a bright line to separate the categories. Again my decision was based upon the correlation coefficient, the P-values for the individual terms and the Standard Errors for the terms. Again, my decision to keep only one equation is based upon engineering judgment based upon the goals I stated at the beginning of the second paragraph. Is there a means in SAS to make an assessment that would support (or reject my hypothesis) that the data are all related and should continue to be assessed as a single group?

Again, choosing which data to use in a model is more of an art than a science. My suggestion is to look at the residual plots, on the original scale. To get

the residuals, take exponents of the predicted values from the model and subtract them from the original data. Then plot the residuals vs. speed and weight. If there is a clear trend in the residuals as a function of speed or weight, then stratifying the model based on speed and weight classes could be worth exploring. If there are no trends, then the simpler model is always the better option.

In my data selection, I excluded the use of all data above 20 g/m^2 silt loading. In my judgment (looking at the individual and combined differences from the actual data) the equation developed from this reduced data performed as well on the ten data that were excluded as if I would have used them in the analysis. The equation developed without the high silt data performed better on the lower silt data. Another area is the large number of tests at low speeds. I tried several regressions with only half of the low speed data and averaging the parameters for the equation. That was very resource intensive and I don’t think it generated a better equation for any of the data. Is there some means in SAS where one could evaluate the ability for an equation to perform better for some conditions and still incorporate a reasonable performance for other conditions (i.e. decreased weighting for higher silt loadings, vehicle speeds and/or greater weights)?

If the preceding is not possible in SAS, is there a statistical basis for using only the data with silt loadings below 20 g/m^2 and excluding the 10 data with the very highest silt loadings for the development of the equation. But, incorporating these higher silt loading tests in the assessment of the predictive accuracy of the equation?

For the silt loading exclusions, it looks like you are taking the 'common sense' approach I mentioned before. If you can provide a scientific rationale for excluding these points (for example, they do not represent typical silt loading values found in the current vehicle fleet on the current road system), then exclude them.

For the low speed data, I wouldn't try anything fancy because more complex modeling techniques often require more explaining and additional justification.

When I used Excel to determine the equation and used only silt loading and average vehicle weight, it produced an equation with a correlation of 0.56, P-values of less than 0.03. See the tab “Output - sL, W” for the Excel equation terms, residuals and graphs. I looked at the residual graphs for silt and they did not seem to curve any and averaged very close to zero. The residual graph for weight did appear to curve some. I then included the weight term squared as an addition to weight and silt. See the tab “Output - sL, W, W^2”. The correlation improved and the P-values went up to less than 0.035. However, upon reviewing how I squared the term, I was squaring the natural log of the weight and not the weight. When I changed the term, Excel was unable to generate an equation and I could not determine a reason. Also, the form of the equation may be problematic from a calculation standpoint. Your insights on the benefits of this addition are welcome.

The reason squaring weight then taking the log won't work is because log(weight^2) = 2*log(weight). So on the log scale, weight and weight squared contain exactly the same information. You could use a log(weight)^2 term, but in my opinion it isn't necessary because it is a rather awkward term and it would difficult to come up with a scientific rationale for including it.

Perhaps I am over thinking the issues. I'll call you or come by your office

to discuss these issues and what I can expect from you to include in my background report.

I just thought of a new idea that would improve the model performance:

In the data you sent, I'm assuming that the letters (BH, M, etc.) in the run ID variable refer to different groups of tests carried out on different roads, with different vehicles, by different researchers, etc. and that the numbers refer to the individual tests within each group. If this is the case then there may be considerable variablility between test groups due to various other factors such as the road, the vehicle, the researchers, etc. The amount of 'between group' variability may be considerably larger than the 'within group' variability, creating additional noise.

If we assume that these factors are constant within each group (identified by the start letters in the run ID column), then we can use the groups themselves as a 'factor' variable. By doing so, we can use the group term to first eliminate the variability due to differences between groups, then look at how much of the remaining variability is explained by silt loading, weight, etc.

Here is an example in SAS (using all 93 data points):

Dependent Variable: logpm

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	652.6490009	38.3911177	19.69	<.0001
Error	75	146.1986525	1.9493154		
Corrected Total	92	798.8476535			

R-Square	Coeff Var	Root MSE	logpm Mean
0.816988	53.83523	1.396179	2.593430

Source	DF	Type I SS	Mean Square	F Value	Pr > F
testclass	15	622.6406235	41.5093749	21.29	<.0001
logsilt	1	29.9352431	29.9352431	15.36	0.0002
logweight	1	0.0731344	0.0731344	0.04	0.8469

Source	DF	Type III SS	Mean Square	F Value	Pr > F
testclass	15	152.7379936	10.1825329	5.22	<.0001
logsilt	1	30.0083760	30.0083760	15.39	0.0002
logweight	1	0.0731344	0.0731344	0.04	0.8469

Notice that the R-square is now 0.817, much higher than before. The 'testclass' variable represents the different groups. log(silt loading) is still a good predictor of log(PM10 emission factor) but now log(weight) is not useful.

Another example (using only the 71 data points with recorded speeds):

Dependent Variable: logpm

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	488.6139226	30.5383702	14.59	<.0001
Error	54	113.0203847	2.0929701		
Corrected Total	70	601.6343074			

R-Square	Coeff Var	Root MSE	logpm Mean
0.812144	79.99762	1.446710	1.808442

Source	DF	Type I SS	Mean Square	F Value	Pr > F
testclass	13	437.6013497	33.6616423	16.08	<.0001
logsilt	1	35.4477555	35.4477555	16.94	0.0001
logweight	1	0.0038577	0.0038577	0.00	0.9659
logspeed	1	15.5609597	15.5609597	7.43	0.0086

Source	DF	Type III SS	Mean Square	F Value	Pr > F
testclass	13	78.95456614	6.07342816	2.90	0.0030
logsilt	1	22.49931842	22.49931842	10.75	0.0018
logweight	1	0.19437674	0.19437674	0.09	0.7617
logspeed	1	15.56095974	15.56095974	7.43	0.0086

Here, log(silt) and log(speed) are important predictors of log(PM10 emission factor) but log(weight) is still not useful. The R-square is 0.812, about the same as in the model above.

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov [attachment "Paved Road Data_11-26-10.xls" deleted by Benjamin Wells/RTP/USEPA/US]

Re: Paved Road Equation development

Ron Myers to: Benjamin Wells 11/30/2010 04:43 PM
Cc: Bob Schell, Phil Lorang

From: Ron Myers/RTP/USEPA/US
To: Benjamin Wells/RTP/USEPA/US
Cc: Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA

Ben:
Thanks for the thoughts on my questions and concerns.

As far as the winnowing of the data. I have somewhat decided that the ten data with silt loadings in excess of 20 g/m2 were not representative of the sources that should be included in the final regression. In Excel, I have come to the conclusion that including this data skews the equation for the lower silt loadings and that excluding them does not adversly impact the ability of the equation to predict these high silt loading data. Also, given the large number of data and their distribution, removing Z3 does not change the equation much either. So, I guess the final data set will exclude Z3 and all data with silt loadings over 20 g/m2.

As far as the speed and weight terms being highly correlated, I think that is a figment of the data collection and not any limitation of the way public roads and industrial roads behave within the physical influences. On the other hand there is a physical and rational explanation for the correlation between vehicle speed and silt loading. The issue in the past has not been whether to add weight or speed but between silt and vehicle speed. In the correlation matrix I have generated, the Emissions Factor and silt loading have the greatest correlation followed by weight and then speed. The comment that we received was that with the addition of speed to the equation using silt and weight the P values increased to an unacceptable level. As a result the MRI statisticians stated that that indicated that speed did not add any predictive accuracy.

As far as the residuals assessment, I did that in several ways. I looked at the performance within different criteria (absolute difference in EF by silt loading and by weight and by percentage difference from actual by silt loading and by weight). My assessment was that the best estimator was to combine the data into a single group, not use Z3 or any data with silt over 20.

Looking at the whole use of the squared terms and the speed term and multiplicative (silt times weight etc) terms, I did not find a form that was better than using just silt and weight.

Yes, the different codes are different test programs at different types of locations. The same test method was used at all locations by the same contractor. There are some letter combinations that are tests of different types of sources. If you think it is good, I can provide a more detailed breakdown of the tests that were condunced of similar types of sources.

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov

Benjamin Wells — Ron, See comments in bold below. — 11/30/2010 11:37:40 AM

From: Benjamin Wells/RTP/USEPA/US

To: Ron Myers/RTP/USEPA/US@EPA
Cc: Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA
Date: 11/30/2010 11:37 AM
Subject: Re: Paved Road Equation development

Ron,

See comments in bold below.

Benjamin Wells, Statistician
US EPA, OAQPS-AQAD-AQAG
Research Triangle Park, NC
919 - 541 - 7507
wells.benjamin@epa.gov

Ron Myers | Ben: Here is a spreadsheet with the final paved r... | 11/26/2010 04:07:27 PM

From: Ron Myers/RTP/USEPA/US
To: Benjamin Wells/RTP/USEPA/US@EPA
Cc: Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA
Date: 11/26/2010 04:07 PM
Subject: Paved Road Equation development

Ben:

Here is a spreadsheet with the final paved roads data. This is in the tab labeled “Test Data & Parameters”. This two left columns in this tab are the test run ID and the measured PM10 emissions both which you may not need but provide linkage to what were in the test reports. The remainder columns are the reported independent variables speed, weight and silt loading with the associated Road Dust emissions (dependent variable). There are other tabs with additional ancillary information that include calculations and Excel output which I describe in my e-mail.

I have some concerns over which of the data to use in developing the final equation. My goal is to provide an equation that provides a good predictive estimate of emissions from public roads with silt loadings typical of most paved roads in the US. Data that we use for the National Emissions Inventory has default silt loadings of between 0.015 and 0.6 g/m2 for most of the year and silt loadings between 0.015 and 2.5 for months with frozen precipitation. While I am not positive, I believe the majority of the emissions inventory (miles driven) is based upon roads with lower silt loadings. For most of the nation the average vehicle weights are between 2 tons and a little more than 4 tons. A secondary goal is to provide a reasonably good predictive estimate from public and private roads with higher silt loadings and also with higher vehicle weights. These are values that are typical of the current estimates used in permit applications. All of the silt loading data in the AP-42 section are quite dated and probably not reflective of current housekeeping practices. I think it is reasonable to assume that the lack of very high silt loading data in the test data and the large gap between 17 g/m2 and 53 g/m2 are indicative that the highest data are anomalous not to mention that all of the tests with high silt loadings were conducted in 1982.

Using Excel, I have explored some of the potential selection of the data to achieve these goals.

First, I have excluded test Z3 (weight of 8 tons, silt loading of 12.4 g/m^2 and PM10 emissions factor of 1819 g/VMT) since it looks like an outlier on a

3D graphic. Is there a means by which we can justify this exclusion with statistical means? Might there be other data points which should be considered for exclusion? I assume that identification of outliers may be difficult as there seems to be a very significant variation in the data across all weight and silt loading values.

Data exclusion is really more of an art than a science. Statistical methods for data exclusion exist but in my experience they do not perform any better than 'common sense' approaches. My recommendation is that you decide on range of values that constitute reasonable silt loadings should be, provide a scientific rationale for your decision, then exclude all silt loading values outside of that range. The 3D scatter plot of weight, silt loading, and PM10 emissions factors should be sufficient to demonstrate that the Z3 test is a clear outlier and should be excluded.

Next, I have recognized that the speed term does not add much value to the final equation. I based that conclusion on a general assessment of the average percentage error of the equation could achieve about the same value without the use of the speed term. Can you provide a more statistically valid rationale for using or not using the speed term as an additional independent variable?

Yes, the speed term should not be excluded because speed is highly correlated with weight. Speed and weight are providing the same information to the model and it is only necessary to use one of them.

I received two comments that suggested developing separate equations for different classes of data. One class that was identified was the large number of tests conducted at very low speeds (1 and 5 mph). Because of their numbers, they may adversely bias the final equation. Another class that was identified was high weight vehicles. Having multiple equations for different classes was a problem until 1995 when all the data were combined for a single equation. The problem was that people would select the equation that gave them the answer they wanted and then they went to extremes to justify this selection. It also set up disagreements which were difficult to resolve without a long argument. I am not adverse to different classes as long a the overlap between the classes produces about the same emissions. I have attempted to divide the test data to separate data for different scenarios but was unable to find equations that performed acceptably at or near the overlapping conditions (average vehicle weight, silt loadings). Also, by parsing out the “special test data” the correlation coefficient degraded significantly for each of the different categories and it was difficult to define a bright line to separate the categories. Again my decision was based upon the correlation coefficient, the P-values for the individual terms and the Standard Errors for the terms. Again, my decision to keep only one equation is based upon engineering judgment based upon the goals I stated at the beginning of the second paragraph. Is there a means in SAS to make an assessment that would support (or reject my hypothesis) that the data are all related and should continue to be assessed as a single group?

Again, choosing which data to use in a model is more of an art than a science. My suggestion is to look at the residual plots, on the original scale. To get the residuals, take exponents of the predicted values from the model and subtract them from the original data. Then plot the residuals vs. speed and weight. If there is a clear trend in the residuals as a function of speed or weight, then stratifying the model based on speed and weight classes could be worth exploring. If there are no trends, then the simpler model is always the better option.

In my data selection, I excluded the use of all data above 20 g/m^2 silt loading. In my judgment (looking at the individual and combined differences from the actual data) the equation developed from this reduced data performed as well on the ten data that were excluded as if I would have used them in the analysis. The equation developed without the high silt data performed better on the lower silt data. Another area is the large number of tests at low speeds. I tried several regressions with only half of the low speed data and averaging the parameters for the equation. That was very resource intensive and I don’t think it generated a better equation for any of the data. Is there some means in SAS where one could evaluate the ability for an equation to perform better for some conditions and still incorporate a reasonable performance for other conditions (i.e. decreased weighting for higher silt loadings, vehicle speeds and/or greater weights)?

If the preceding is not possible in SAS, is there a statistical basis for using only the data with silt loadings below 20 g/m^2 and excluding the 10 data with the very highest silt loadings for the development of the equation. But, incorporating these higher silt loading tests in the assessment of the predictive accuracy of the equation?

For the silt loading exclusions, it looks like you are taking the 'common sense' approach I mentioned before. If you can provide a scientific rationale for excluding these points (for example, they do not represent typical silt loading values found in the current vehicle fleet on the current road system), then exclude them.

For the low speed data, I wouldn't try anything fancy because more complex modeling techniques often require more explaining and additional justification.

When I used Excel to determine the equation and used only silt loading and average vehicle weight, it produced an equation with a correlation of 0.56, P-values of less than 0.03. See the tab “Output - sL, W” for the Excel equation terms, residuals and graphs. I looked at the residual graphs for silt and they did not seem to curve any and averaged very close to zero. The residual graph for weight did appear to curve some. I then included the weight term squared as an addition to weight and silt. See the tab “Output - sL, W, W^2”. The correlation improved and the P-values went up to less than 0.035. However, upon reviewing how I squared the term, I was squaring the natural log of the weight and not the weight. When I changed the term, Excel was unable to generate an equation and I could not determine a reason. Also, the form of the equation may be problematic from a calculation standpoint. Your insights on the benefits of this addition are welcome.

The reason squaring weight then taking the log won't work is because log(weight^2) = 2*log(weight). So on the log scale, weight and weight squared contain exactly the same information. You could use a log(weight)^2 term, but in my opinion it isn't necessary because it is a rather awkward term and it would difficult to come up with a scientific rationale for including it.

Perhaps I am over thinking the issues. I'll call you or come by your office to discuss these issues and what I can expect from you to include in my background report.

I just thought of a new idea that would improve the model performance:

In the data you sent, I'm assuming that the letters (BH, M, etc.) in the run ID variable refer to different groups of tests carried out on different roads, with different vehicles, by different researchers, etc. and that the numbers

refer to the individual tests within each group. If this is the case then there may be considerable variablility between test groups due to various other factors such as the road, the vehicle, the researchers, etc. The amount of 'between group' variability may be considerably larger than the 'within group' variability, creating additional noise.

If we assume that these factors are constant within each group (identified by the start letters in the run ID column), then we can use the groups themselves as a 'factor' variable. By doing so, we can use the group term to first eliminate the variability due to differences between groups, then look at how much of the remaining variability is explained by silt loading, weight, etc.

Here is an example in SAS (using all 93 data points):

Dependent Variable: logpm

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	652.6490009	38.3911177	19.69	<.0001
Error	75	146.1986525	1.9493154		
Corrected Total	92	798.8476535			

R-Square	Coeff Var	Root MSE	logpm Mean
0.816988	53.83523	1.396179	2.593430

Source	DF	Type I SS	Mean Square	F Value	Pr > F
testclass	15	622.6406235	41.5093749	21.29	<.0001
logsilt	1	29.9352431	29.9352431	15.36	0.0002
logweight	1	0.0731344	0.0731344	0.04	0.8469

Source	DF	Type III SS	Mean Square	F Value	Pr > F
testclass	15	152.7379936	10.1825329	5.22	<.0001
logsilt	1	30.0083760	30.0083760	15.39	0.0002
logweight	1	0.0731344	0.0731344	0.04	0.8469

Notice that the R-square is now 0.817, much higher than before. The 'testclass' variable represents the different groups. log(silt loading) is still a good predictor of log(PM10 emission factor) but now log(weight) is not useful.

Another example (using only the 71 data points with recorded speeds):

Dependent Variable: logpm

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	488.6139226	30.5383702	14.59	<.0001
Error	54	113.0203847	2.0929701		
Corrected Total	70	601.6343074			

R-Square	Coeff Var	Root MSE	logpm Mean
0.812144	79.99762	1.446710	1.808442

Source	DF	Type I SS	Mean Square	F Value	Pr > F
testclass	13	437.6013497	33.6616423	16.08	<.0001
logsilt	1	35.4477555	35.4477555	16.94	0.0001
logweight	1	0.0038577	0.0038577	0.00	0.9659
logspeed	1	15.5609597	15.5609597	7.43	0.0086

Source	DF	Type III SS	Mean Square	F Value	Pr > F
testclass	13	78.95456614	6.07342816	2.90	0.0030
logsilt	1	22.49931842	22.49931842	10.75	0.0018
logweight	1	0.19437674	0.19437674	0.09	0.7617
logspeed	1	15.56095974	15.56095974	7.43	0.0086

Here, log(silt) and log(speed) are important predictors of log(PM10 emission factor) but log(weight) is still not useful. The R-square is 0.812, about the same as in the model above.

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov [attachment "Paved Road Data_11-26-10.xls" deleted by Benjamin Wells/RTP/USEPA/US]

Re: Paved Road Equation development

Ron Myers to: Benjamin Wells 12/02/2010 02:07 AM
Cc: Bob Schell, Phil Lorang

From: Ron Myers/RTP/USEPA/US
To: Benjamin Wells/RTP/USEPA/US@EPA
Cc: Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA

Ben:
I have extracted information from the test reports to segregate the paved road test data into what I believe are homogeneous groups of tests that were performed at different locations at different times and under somewhat differnent conditions. While many of the groups that you established based upon different letter identifiers were of one type of condition, some of the letter groupings contain multiple conditions. For example tests preceded with M are four types of conditions rather than one, tests preceded with F are five different conditions, tests preceded with B are five conditions, and tests preceded with BH are two types of conditions. Below is the grouping that I could identify. Some of my classifications may already be differentiated in that the groups are different speeds at the same location (primarily the tests at corn refiners). There are several groups of only one or two test series. Would this be a problem in performing the randomized block analysis?

Here are the groupings that I have established.

Group 1 - Integrated Iron & Steel Plant with mix of light duty and medium duty vehicles
AUC3
AUC4
AUC5
AUC6
AUC7
AUC8

Group 2 - Integrated Iron & Steel Plant with primarily plant equipment vehicles
AUE1
AUE2
AUE3
AUE4

Group 3 - Commercial Industrial Site in Missouri
M-1
M-2
M-3
M-9

Group 4 - Commercial Residential Sites in Missouri & Illinois
M-4
M-5
M-6
M-7
M-13
M-14
M-15
M-17

M-18
M-19

Group 5 - Expressway in Missouri
M-10
M-11
M-12
M-16

Group 6 - Rural Town in Kansas
M-8

Group 7 - Asphalt Batch Plant with Medium Duty Vehicles
Y1
Y2
Y3
Y4

Group 8 - Concrete Batch Plant with Medium Duty Vehicles
Z1
Z2
Z3

Group 9 - Copper Smelting Plant with Medium Duty Vehicles
AC4
AC5
AC6

Group 10 - Sand and Gravel Plant with Heavy Duty Vehicles
AD1
AD2
AD3

Group 11 - Integrated Iron & Steel Plant - No Control
F34
F35

Group 12 - Integrated Iron & Steel Plant - Vacuum Sweeping
F36
F37
F38
F39

Group 13 - Integrated Iron & Steel Plant - No Control
F61
F62

Group 14 - Integrated Iron & Steel Plant - Water Flushing Control
F74

Group 15 - Integrated Iron & Steel Plant - - No Control
F27
F45
F32

Group 16 - Integrated Iron & Steel Plant - Flushing & Broom Sweep
B50
B51
B52

Group 17 - Integrated Iron & Steel Plant - Water Flushing Control
B54
B55
B56

Group 18 - Integrated Iron & Steel Plant - - No Control
B58

Group 19 - Integrated Iron & Steel Plant - Flushing & Broom Sweep
B53

Group 20 - Integrated Iron & Steel Plant - - No Control
B59
B57
B60

Group 21 - Denver Colorado limited access interstate road during periods following snowfall - sand application
BH1
BH2
BH3 - 0 data

Group 22 - Denver Colorado one lane road (two lanes with a wide median) following snowfall - sand application
BH6

Group 23 - Raleigh, NC - Urban Roadway Principal Arterial
BJ6
BJ7
BJ9 - 0 data
BJ10 - 0 data
BJ11 - 0 data

Group 24 - Reno, NV - Urban Roadway Principal Arterial
BK7
BK8

Group 25 - Minnesota Corn Processor stop-and-go semi-trailer trucks
CE-1
CE-2
CE-3

Group 26 - Minnesota Corn Processor slowly moving (5 mph enforced speed limit) semi-trailer trucks
CE-11
CE-12
CE-15
CE-16
CE-17
CE-19

Group 27 - Columbus, Nebraska Corn Processor stop-and-go semi-trailer trucks
CF-1/South
CF-2/South
CF-3/South
CF-5

Group 28 - Columbus, Nebraska Corn Processor slowly moving semi-trailer trucks
CF-1N

CF-2N
CF-3N
CF-4N

Group 29 - Blair, Nebraska Corn Processor slowly moving semi-trailer trucks
CI-7
CI-8

Group 30 - Marshall, Minnesota Corn Processor stop-and-go semi-trailer trucks
CM-2

Group 31 - Marshall, Minnesota Corn Processor slowly moving semi-trailer trucks
CM-1
CM-4

_____

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov

Re: Paved Road Equation development

Benjamin Wells to: Ron Myers 12/03/2010 01:29 PM

From: Benjamin Wells/RTP/USEPA/US
To: Ron Myers/RTP/USEPA/US@EPA

Ron,

After looking at your groupings carefully and playing around in SAS, I have decided on some final recommendations for the paved roads model. The text files include the SAS output from each model I tried.

1) Remove the 10 points with silt loading values greater than 20.

2) Keep the 'Z3' point. My reasoning for this is that on the log scale, the emissions for the three Z tests are not much different. The smallest emission factor (Z1) was 319, which is 5.76 on the log scale, while the largest emission factor (Z3) was 1819 which is 7.51 on the log scale. These differ by 1.75 on the log scale. By contrast, emission factors of 0.01 and 0.1 convert to -4.61 and -2.30, a difference of 2.31 on the log scale.

3) Splitting the data into the 31 groups you specified below was too many. There were 27 groups left after the high silt points were removed, many with only one or two points representing them. Fitting a baseline parameter for each of these groups used up 27 out of 83 (about 1/3) degrees of freedom, and the group effects were so strong that the silt and weight effects ended up getting completely washed out. No good.


model1_all_groups.txt

4) Next, I tried splitting the data by time and location. Since you included information on test locations along with the groups, I identified 17 different groupings based on where and when the tests were performed. It looks like the only location with two different times was the Integrated Iron & Steel plant. I also removed the location in rural KS with only one data point. After the high silt points were removed, there were 13 locations used in the model. This model performed just about as well as the groups model, but it had the same problem: the silt and weight effects still ended up getting washed out. Still no good.


model2_locations.txt

5) The next model I tried was the silt and weight only model, ignoring speed. This time, I left out the intercept term (the k term in the draft AP-42 section 13.2.1 document). This model performed quite well. Both the speed and weight terms were highly significant, and the R-squared coefficient was about 0.72. It appears that this very simple model is probably the best choice. The equation for this model is:

log(PM10 emissions factor) = 0.9118*log(silt loading) + 1.0213*log(weight) + (residual error)

or equivalently,

(PM10 emissions factor) = (silt loading)^0.9118 * (weight)^1.0213 * (residual error)

This model is also very sound from a scientific standpoint because it is saying that PM10 emissions increase with both silt and weight.

model3_silt_weight.txt

6) Finally, just for exploration I looked at stratifying the model based on speed. I threw out all of the observations without speed observations, and grouped the remaining data into low speed (< 10 mph) and high speed (>= 10 mph). The results were interesting.

At low speeds, only the weight term was important. The only trouble is the low speed model only had 20 data points, which may cause scrutiny if this approach is used.

At high speeds, the silt and speed terms were almost equally important, with the silt term just edging out the speed term. The trouble here was that there was correlation between the silt and speed terms, so that they cannot both be used in the model at the same time. Additionally, the speed term was negative, which may not be justifiable from a scientific standpoint.

The R-squared coefficients for both models were considerably lower than before, 0.55 for the low speed model, and 0.57 for the high speed model.

The equations for this model are:

log(PM10 emissons factor) = 0.6674*log(weight) + error, IF speed < 10 mph

log(PM10 emissions factor) = 1.754 + 1.0584*log(silt loading) + error, IF speed >= 10 mph

or equivalently,

(PM10 emissions factor) = (weight)^0.6674 * error, IF speed < 10 mph

(PM10 emissions factor) = e^1.754 * (silt loading)^1.0584 * error, IF speed >= 10 mph

This model is more difficult to defend than the silt & weight only model. My recommendation is to stick with the simpler approach.

model4_silt_weight_by_speed.txt

I think this should be everything you need to finish the report. Let me know if you need any further clarifications.

Benjamin Wells, Statistician
US EPA, OAQPS-AQAD-AQAG
Research Triangle Park, NC
919 - 541 - 7507
wells.benjamin@epa.gov

Ron Myers | Ben: I have extracted information from the test r... | 12/02/2010 02:07:18 AM

From: Ron Myers/RTP/USEPA/US
To: Benjamin Wells/RTP/USEPA/US@EPA
Cc: Bob Schell/RTP/USEPA/US@EPA, Phil Lorang/RTP/USEPA/US@EPA
Date: 12/02/2010 02:07 AM
Subject: Re: Paved Road Equation development

Ben:
I have extracted information from the test reports to segregate the paved road test data into what I believe are homogeneous groups of tests that were performed at different locations at different times and under somewhat differnent conditions. While many of the groups that you established based upon different letter identifiers were of one type of condition, some of the letter groupings contain multiple conditions. For example tests preceded with M are four types of conditions rather than one, tests preceded with F are five different conditions, tests preceded with B are five conditions, and tests preceded with BH are two types of conditions. Below is the grouping that I could identify. Some of my classifications may already be differentiated in that the groups are different speeds at the same location (primarily the tests at corn refiners). There are several groups of only one or two test series. Would this be a problem in performing the randomized block analysis?

Here are the groupings that I have established.

Group 1 - Integrated Iron & Steel Plant with mix of light duty and medium duty vehicles
AUC3
AUC4
AUC5
AUC6
AUC7
AUC8

Group 2 - Integrated Iron & Steel Plant with primarily plant equipment vehicles
AUE1
AUE2
AUE3
AUE4

Group 3 - Commercial Industrial Site in Missouri
M-1
M-2
M-3
M-9

Group 4 - Commercial Residential Sites in Missouri & Illinois
M-4
M-5
M-6
M-7
M-13
M-14
M-15
M-17
M-18
M-19

Group 5 - Expressway in Missouri
M-10
M-11
M-12
M-16

Group 6 - Rural Town in Kansas
M-8

Group 7 - Asphalt Batch Plant with Medium Duty Vehicles
Y1
Y2
Y3
Y4

Group 8 - Concrete Batch Plant with Medium Duty Vehicles
Z1
Z2
Z3

Group 9 - Copper Smelting Plant with Medium Duty Vehicles
AC4
AC5
AC6

Group 10 - Sand and Gravel Plant with Heavy Duty Vehicles
AD1
AD2
AD3

Group 11 - Integrated Iron & Steel Plant - No Control
F34
F35

Group 12 - Integrated Iron & Steel Plant - Vacuum Sweeping
F36
F37
F38
F39

Group 13 - Integrated Iron & Steel Plant - No Control
F61
F62

Group 14 - Integrated Iron & Steel Plant - Water Flushing Control
F74

Group 15 - Integrated Iron & Steel Plant - - No Control
F27
F45
F32

Group 16 - Integrated Iron & Steel Plant - Flushing & Broom Sweep
B50
B51
B52

Group 17 - Integrated Iron & Steel Plant - Water Flushing Control
B54
B55
B56

Group 18 - Integrated Iron & Steel Plant - - No Control
B58

Group 19 - Integrated Iron & Steel Plant - Flushing & Broom Sweep
B53

Group 20 - Integrated Iron & Steel Plant - - No Control
B59
B57
B60

Group 21 - Denver Colorado limited access interstate road during periods following snowfall - sand application
BH1
BH2
BH3 - 0 data

Group 22 - Denver Colorado one lane road (two lanes with a wide median) following snowfall - sand application
BH6

Group 23 - Raleigh, NC - Urban Roadway Principal Arterial
BJ6
BJ7
BJ9 - 0 data
BJ10 - 0 data
BJ11 - 0 data

Group 24 - Reno, NV - Urban Roadway Principal Arterial
BK7
BK8

Group 25 - Minnesota Corn Processor stop-and-go semi-trailer trucks
CE-1
CE-2
CE-3

Group 26 - Minnesota Corn Processor slowly moving (5 mph enforced speed limit) semi-trailer trucks
CE-11
CE-12
CE-15
CE-16
CE-17
CE-19

Group 27 - Columbus, Nebraska Corn Processor stop-and-go semi-trailer trucks
CF-1/South
CF-2/South
CF-3/South
CF-5

Group 28 - Columbus, Nebraska Corn Processor slowly moving semi-trailer trucks
CF-1N
CF-2N
CF-3N
CF-4N

Group 29 - Blair, Nebraska Corn Processor slowly moving semi-trailer trucks
CI-7
CI-8

Group 30 - Marshall, Minnesota Corn Processor stop-and-go semi-trailer trucks
CM-2

Group 31 - Marshall, Minnesota Corn Processor slowly moving semi-trailer

trucks
CM-1
CM-4

_____

Ron Myers
U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Sector Policy and Programs Division
Monitoring Policy Group, D243-05
RTP NC 27711
Tel. 919.541.5407
Fax 919.541.1039
E-mail myers.ron@epa.gov