

Suggested modification of IMPROVE data patching algorithm

Margaret McCourtney, Minnesota PCA

Donna Kenski, LADCO

Presented to IMPROVE Steering Committee, Oct 16-17, 2018

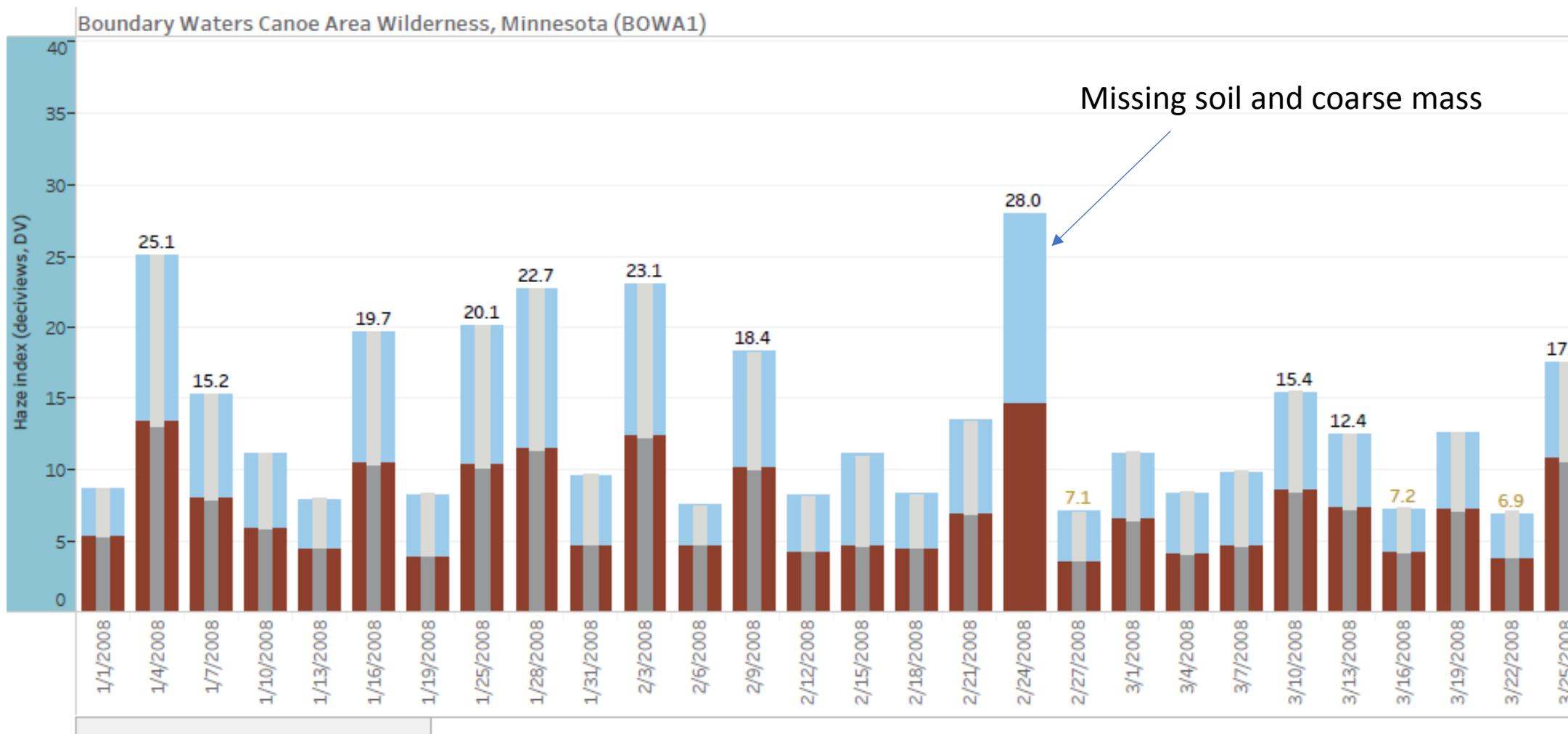
Fort Collins, CO

Issues with Data Patching

- Missing values in the IMPROVE dataset are 'patched' (i.e., filled in) with quarterly 5-year median values
- Process is meant to maximize the number of sample periods available for assessing haze trends without significantly degrading the haze calculation
- Patching with quarterly medians for missing values is currently not done if more than one component of the IMPROVE equation is missing (except OC & LAC count as one component)
- BUT... Guidance for Tracking Progress guidance (2003) permits substitution of more than one component
- In a few cases, substitution of more than one component should be done to retain important days that are otherwise lost
- Multiple substitutions not overly complicated to implement

Minnesota estimated 20 percent most and least impaired visibility days at select IMPROVE monitoring sites

Comparison of Minnesota and CIRA calculated haze index. Each day haze index value is partitioned by episodic and routine, and anthropogenic source estimates. Haze index totals at top of bar flag 20 percent most impaired days (**black font**) and 20 percent least impaired days (**gold font**) by IMPROVE site and year as estimated by Minnesota. Missing CIRA haze index bars indicate a failure to meet criteria for patching quarterly 4-year average median values for any one component species in the CIRA calculations.



Click below on IMPROVE site, Year and/or Month to explore data. Use slider above to scroll through days

Minnesota ■ Episodic & routine ■ Anthropogenic
 CIRA ■ Episodic & routine ■ Anthropogenic

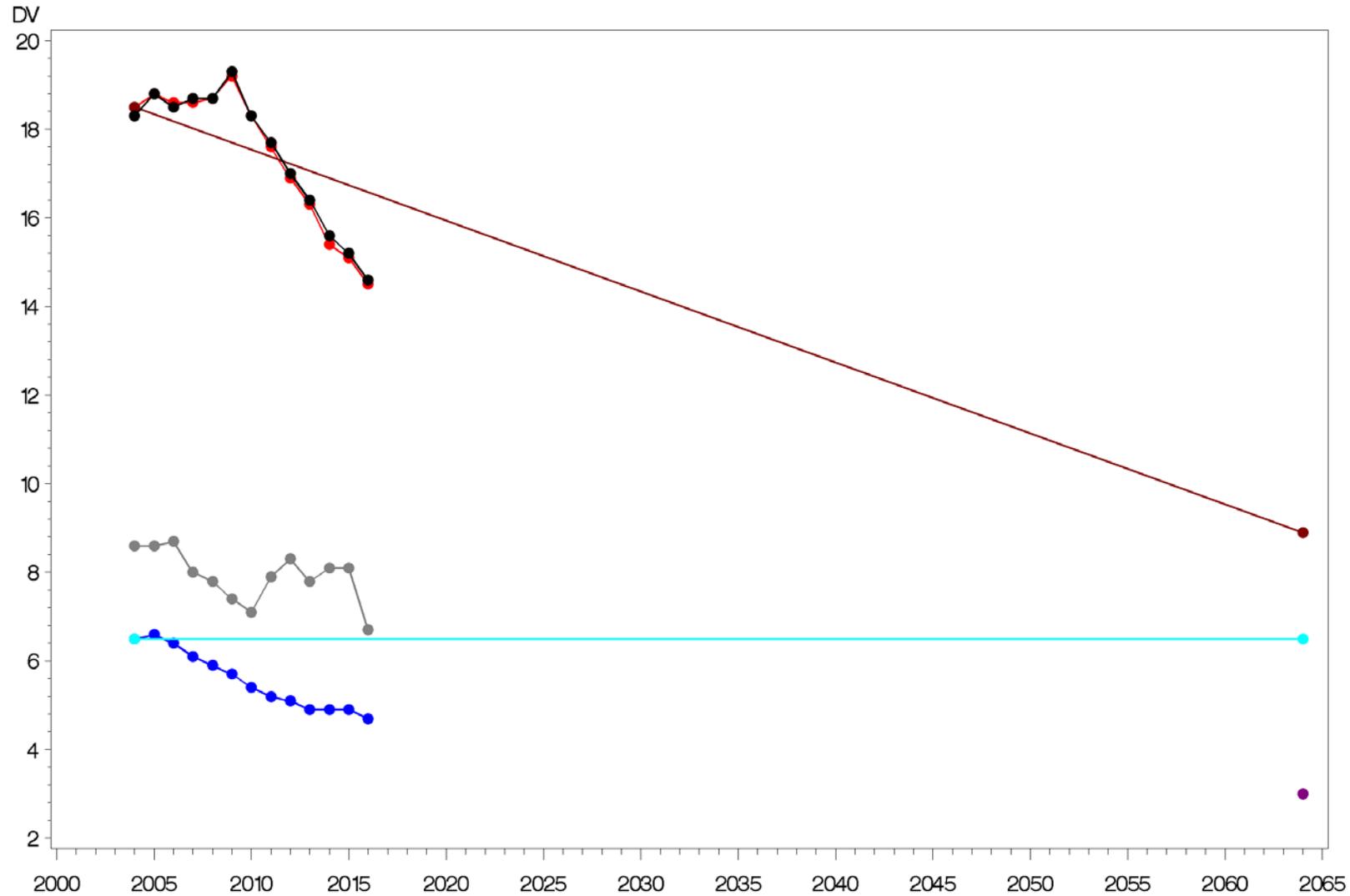
IMPROVE site
 Boundary Waters Canoe Area Wilderness, Minnesota (BOWA1)

Year 2008
 Month All

Revised data patching for haze tracking metric newly included 20% most impaired days					IMPROVE 20% most impaired
IMPROVE site	Date	Total DV	Anthropogenic DV	Year average Total DV	Year average Total DV
BOWA1	2000-01-15	24.9	12.2	18.6	18.2
BOWA1	2002-09-08	27.6	11.7	18.8	18.4
BOWA1	2004-03-25	24.0	8.3	17.0	16.5
BOWA1	2008-02-24	28.0	14.7	19.1	18.8
BOWA1	2009-03-17	17.1	8.7	19.5	19.5
ISLE1	2001-03-29	24.4	13.8	22.2	22.0
ISLE1	2001-04-01	20.9	11.9		
ISLE1	2002-09-08	28.4	12.4	19.9	19.2
ISLE1	2003-03-16	31.1	15.0	21.2	19.0
ISLE1	2003-07-26	22.6	7.5		
ISLE1	2003-08-19	23.7	8.4		
ISLE1	2003-09-09	25.4	11.1		
ISLE1	2003-09-12	35.1	17.3		
ISLE1	2004-03-25	26.4	10.8	19.0	18.5
ISLE1	2007-03-13	18.7	9.4	20.5	20.6
ISLE1	2014-10-23	14.9	6.8	18.1	18.2
SENE1	2001-04-04	19.8	12.8	25.2	24.9
SENE1	2001-04-07	20.5	12.5		
SENE1	2001-12-12	30.1	14.7		
SENE1	2002-09-08	36.8	18.5	24.5	24.0
SENE1	2006-02-22	19.3	11.4	23.5	23.6
SENE1	2011-02-17	24.1	10.8	19.7	19.7
SENE1	2012-01-31	22.5	10.2	19.3	18.9
VOYA2	2000-01-15	24.0	11.9	18.0	17.7
VOYA2	2004-03-25	26.2	9.6	17.2	16.8
VOYA2	2006-03-30	24.4	13.1	18.2	17.9
VOYA2	2007-01-21	20.7	10.9	18.1	17.9

27 days added to our dataset from multiple patching; 4 days are newly included because of RHR3 impairment metric

BOWA Glide Paths with/without Additional Patching (RHR3)

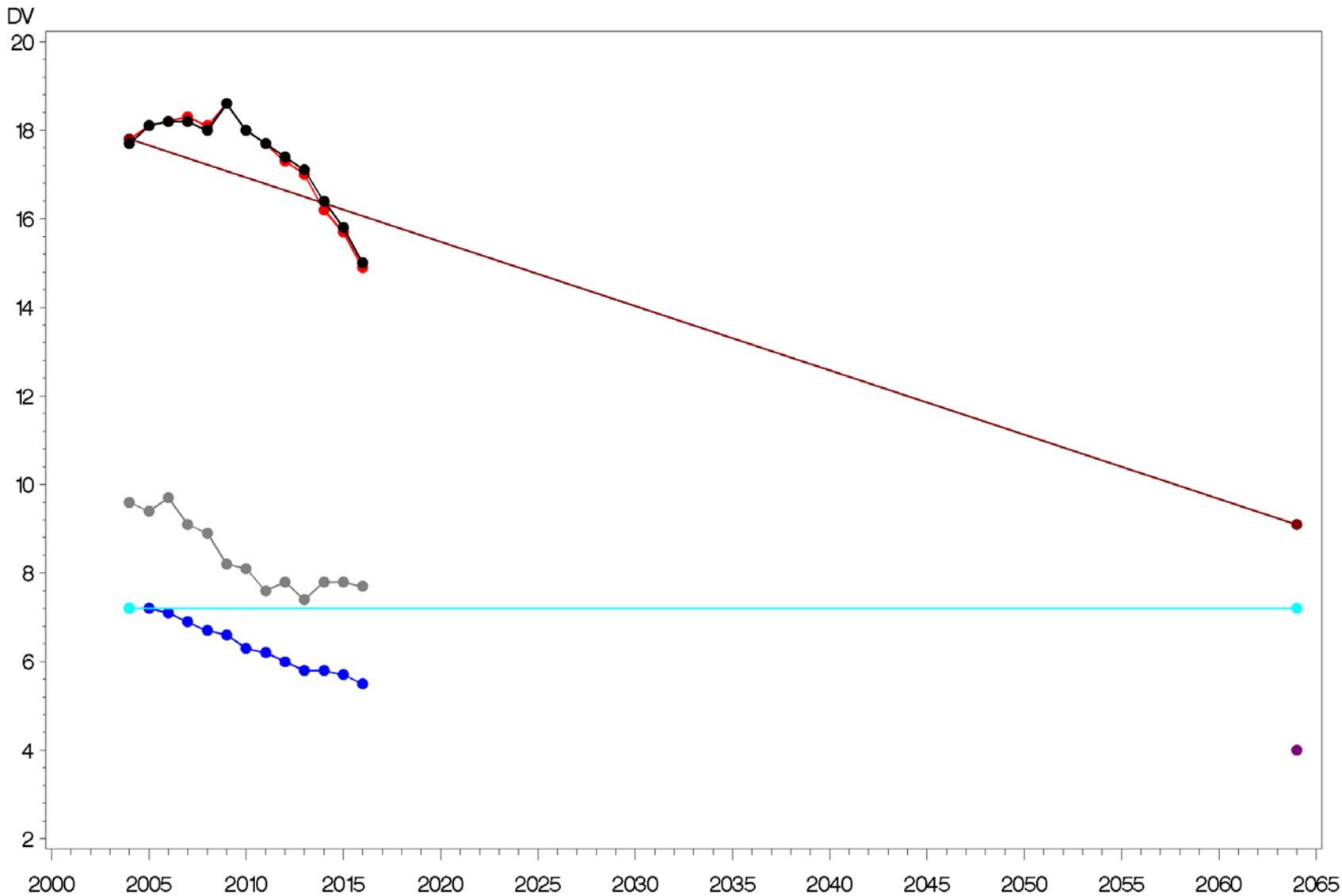


PLOT

- 5 yr avg, best 20%
- Uniform Rate of Progress
- CIRA RHR3, worst 20%
- 5 yr avg, worst 20%
- Natural Conditions
- CIRA RHR3, best 20%
- No Degradation, best 20%

Gray line is 20% least impaired, not 20% clearest!

VOYA2 Glide Paths with/without Additional Patching (RHR3)



- | | | | |
|------|---|--|--|
| PLOT | ●—● 5 yr avg, best 20% | ●—● 5 yr avg, worst 20% | ●—● No Degradation, best 20% |
| | ●—● Uniform Rate of Progress | ●—● Natural Conditions | ●—● CIRA RHR3, best 20% |
| | ●—● CIRA RHR3, worst 20% | | |

Gray line is 20% least impaired, not 20% clearest!

Conclusions and Recommendation

- LADCO states have been using multiple substitutions since the first RH implementation period
- Differences are subtle; no extreme changes in trends; biggest changes are in early years when data record is more variable
- Preserving data is important for haze trends, but also for the many other uses of RH data – model validation, source attribution, satellite validation, epidemiology and health studies, ecological studies, etc.
- Even though raw (unpatched) data is available, most people default to the RHR dataset because it has undergone more scrutiny and contains constructed variables like DV and extinction.
- We recommend IMPROVE adopt multiple substitutions (aka extended patching) for RHR calculations going forward.

From Guidance for Tracking Progress Under the Regional Haze Rule, Step 5 - Evaluate Feasibility of Substituting Average Values

“First, for a given IMPROVE site, the total light extinction values (bext, see Section 3) for all days with no missing data are calculated. This calculation is done as described in Section 3 of this guidance document, using the appropriate $f(RH)$ factors and the appropriate calculations of the individual composite components. This calculation produces a list of bext values for that portion of the original data set that had no missing values.

The second step is to recalculate the bext values for the same sampling days, but with the appropriate average of the quarterly median values for a single species (from Step 3 above), substituted in place of all of the individual values of that species. For example, the average of the quarterly median values of sulfate at a site would be substituted for the corresponding individual sulfate values, for all the days from that site with no missing data. The bext values are then calculated for the resulting data set. The product of this step is a second set of bext values, corresponding one-to-one with those calculated from the original values.

Next, these two sets of bext values are compared for each sample day on a calendar year basis. If the difference between the 2nd bext value relative to the 1st bext value is less than 10% for at least 90% of the sample days, then it is acceptable to replace any missing values for that species with the appropriate average of the quarterly medians for that species. If this criterion is not met, then any missing values for that species must remain missing for that calendar year.

The preceding process should be repeated as necessary for each species with any missing data, one at a time, e.g., producing a set of bext values with the average of the quarterly median sulfate concentration substituted for all individual sulfate values, then another set of bext values with an average of the quarterly median nitrate concentration substituted for all individual nitrate values, etc. Each such set of bext values is compared to the original set, to make a judgment about substitution of averages for just one particle species. Note that this process is to be carried out for each composite species for each year, i.e., producing up to 30 tests for each 5-year period, depending on the extent of missing data. It is expected that at any given IMPROVE site, it may be reasonable and appropriate to replace missing data with quarterly medians for some species, but not for others.

Instances in which data on more than one aerosol component are missing in the same sample are likely to be rare. As a result, the process for dual substitution is not presented at length here. However, substitution of two variables in the same sample could be done, subject to adequate justification and testing, such as in the substitution test described previously. The same acceptance criterion of less than 10% difference in bext values in 90% of the data should apply. For example, currently, light absorbing carbon and organic carbon data are likely to either be present or missing in the same samples because of the common analysis method for these species. As a result, this substitution test could also be carried out for those two species simultaneously. That is, the quarterly median values for both species could be substituted for their individual values at a site, the bext values could be calculated, and the comparison made to assess whether simultaneous replacement of missing LAC and OC data with averages is appropriate.”