UCD CSN Technical Information #801B

CSN Data Processing

Chemical Speciation Network
Air Quality Research Center
University of California, Davis

Version 1.0

Prepared By: _____ Date: 2/23/17

Reviewed By: _____ Date: 2/23/17

Approved By: _____ Date: 2/23/17

UCDAVIS
AIR QUALITY RESEARCH CENTER

DOCUMENT HISTORY

Date Modified	Initials	Section/s Modified	Brief Description of Modifications

Table of Contents

1. PURPOSE AND APPLICABILITY

The subject of this technical information (TI) document is processing the sampling and analytical data from the CSN network. The raw sampler operating information is combined with the analytical results to generate concentrations, uncertainties, and detection limits. These calculated variables are then validated in the next step of the process.

2. DEFINITIONS

crocker: A custom software package in the R language that contains the data processing code used to produce, check, and post the final results.

CSN database: A SQL Server database that is the central warehouse of CSN preliminary and final data at UC Davis.

MDL: Minimum detection limit.

3. PROCEDURES

Data processing is performed using the crocker R package, which is developed and maintained by UC Davis specifically for data processing and operational monitoring of the CSN and IMPROVE data. Data processing is performed by the UC Davis data management team on monthly batches of data (a calendar month of sample start dates). Processing occurs in three steps.

1. Calculate concentrations, uncertainties, and MDLs.
2. Perform duplicates check.
3. Post results to CSN database.

The three procedures are outlined below.

3.1 Calculate concentrations, uncertainties, and MDLs

Laboratory results for ions and carbon fractions are stored in the database as mass per filter and elements are stored as mass per cm^2. The *csn_calculate_* methods in the *crocker* package combine per filter analysis results with filter operational data (e.g., flow rates) and corresponding blanks to calculate concentrations, uncertainty estimates, and MDLs. The details and specific equations are provided in Section 4.

To calculate values for all measured and derived parameters, the operator (typically the data validation analyst) will open an R environment (such as RStudio) and run the following command[1]:

[1] Text in [brackets] indicates values that can be changed by the user. Other values should be typed as written.

[month_data] <- crocker::csn_calculate_all([YYYY], [MM], 'production')

This command will calculate concentrations, uncertainties, and MDLs for all measured and derived parameters for the year *([YYYY])* and month *([MM])* and return them (in memory) to the variable *[month_data]*. The last argument in the command specifies that the calculations will use the "production" database (i.e., the CSN operational database).

3.2 Perform duplicates check

Duplicate records in the data produced by *csn_calculate_all* typically indicate a problem and are always investigated. To check for the presence of duplicate records, the operator will execute the following command:

[duplicates] <- crocker::find_duplicates([month_data], c('AqsSiteId', 'POC', 'SampleStartDate', 'Parameter'))

This command will find and count all records with unique combinations of AqsSiteId, POC, SampleStartDate, and Parameter. If the returned variable *(duplicates)* is empty, there are no duplicated records. Otherwise, *duplicates* will contain a list of all records that are duplicated and the operator will investigate the records in consultation with the data manager.

3.3 Post results to CSN database

After all duplicates have been resolved, the results can be uploaded to the CSN database in preparation for level 1 data validation. The results will be uploaded to both the test and production databases. The operator will execute the following commands:

[post] <- crocker::post_results([month_data], 'csn', 'test', delete.file = FALSE, AnalysisQcCode = 1, comment = ['Initial Posting'])

[post] <- crocker::post_results([month_data], 'csn', 'production', delete.file = FALSE, AnalysisQcCode = 1, comment = ['Initial Posting'])

This command appends the processed data to the analysis. Results table in the CSN production database as an analysis set. It also inserts a record into the analysis.Sets table that provides summary information for this set, including the comment and the *AnalysisQcCode*. *AnalysisQcCode = 1* is used for routine data. The command also creates a text file copy (csnAnalysisResults_[Date]_[Time].txt) for archive that is stored in \\CL-SQL\Production_DB_Inserts.

4. DATA PROCESSING EQUATIONS

The following section presents the equations used to calculate aerosol concentrations and the associated uncertainty and mdl. All these calculations are performed by the *crocker* R package.

The mass of material on the filter is equal to the difference between the mass measured on the sample and the mass on the unused filter. The concentration is the mass on the filter divided by the sample volume.

4.1 Ions

Ions are measured by ion chromatography using the nylon filter on the SASS. The ions measurement – chloride (Cl^-), nitrate (NO_3^-), sulfate (SO_4^{2-}), ammonium (NH_4^+), potassium (K^+), and sodium (Na^+) – are delivered as micrograms per filter in the ions data files from DRI. The ion concentration, mdl, and uncertainty in micrograms per cubic meter are calculated using Equations **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.** respectively.

$$C = \frac{\mathrm{A - B}}{\mathrm{V}} \quad (1)$$

Where,

C = ambient concentration ($\mu g/m^3$)

A = mass measured on sample (μg/filter)

B = artifact mass (μg/filter) = monthly median of field blanks (FB).

V = sample air volume (m^3)

The MDL corresponds to three times the standard deviation of monthly field blanks, as shown in Equation **Error! Reference source not found.**.

$$mdl = \frac{3 * S_{FB} * a}{V} \quad (2)$$

Where ,

S_{FB} = standard deviation of at least 50 field blanks for the matching month, including nearby months if necessary (μg/filter)

a = filter sample area (cm^2)

Uncertainty is reported with each concentration. The general model for the uncertainty is a quadratic sum of two components of uncertainty as shown in Equation **Error! R eference source not found.**.

$$\sigma(c) = \sqrt{(S_{FB}\mathrm{a})^2 + (fC)^2} \quad 3$$

Where,

f = proportional uncertainty based on collocated measurements.

S_{FB} = Additive analytical uncertainty, as calculated in Equation **Error! Reference s ource not found.**

$$S_{FB} = \frac{mdl}{1.6449} \quad 4$$

The fractional uncertainty, f, is obtained by taking the root mean square of the preceding five years' collocated precisions (cp).

The fraction uncertainty for each ion species given in Table 1. Note that no collocated precision estimates were available for Chloride, and a value of 0.25 is used. These values are stored in the database.

Table 1. Fractional uncertainty (f) for ions estimated from 2009-2014 collocated data.

Species	f
Chloride	0.25
Nitrate	0.076
Sulfate	0.049
Ammonium	0.071
Sodium	0.247
Potassium	0.126

4.2 Carbon

Carbon is measured by thermal optical reflectance (TOR) using a quartz filter. Carbon measurements are stored in the carbon file as micrograms per filter. For the eight carbon species, the primary source of fractional uncertainty is the separation into temperature ranges. This may be associated with temperature regulation, but it may also be from the inherent variability of the species involved. The concentration, mdl, and uncertainty in micrograms per cubic meter for the carbon fractions and sums – OC1, OC2, OC3, OC4, OPTR, OPTT, EC1, EC2, EC3, OCTR, ECTR, and TCTC – are calculated both with and without an artifact correction using Equations 4-7. For the versions without artifact correction (e.g., OC1_raw, ECTR_raw), B is 0. For the

values with artifact correction (e.g., OC1, ECTR), B is the monthly median mass loading of field blanks across the network.

$$C = \frac{\text{A - B}}{\text{V}} \quad (4)$$

Where,

C = ambient concentration (µg/m^3)

A = mass measured on sample (µg/filter)

B = artifact mass (µg/filter) = monthly median of field blanks

V = sample air volume (m^3)

The MDL corresponds to the maximum of either three times the standard deviation of monthly field blanks or a fixed floor value, as shown in Equation 5. The floor value is the analytical MDL developed by the carbon analysis lab. The values are listed in Table 2.

$$mdl = max\begin{cases}\frac{3 * S_{FB} * a}{V} \\ MDLfloor\end{cases} \quad (5)$$

Where,

S_{FB} = standard deviation of at field blanks for the matching month (µg/filter)

a = filter sample area (cm^2)

$MDLfloor$ = a minimum MDL value (µg/m^3)

Uncertainty is reported with each concentration. The general model for the uncertainty is a quadratic sum of two components of uncertainty as shown in Equation 6.

$$\sigma(c) = \sqrt{(S_{FB}\text{a})^{\,2} + (fC)^2} \quad (6)$$

Where,

f = proportional uncertainty based on collocated measurements.

S_{FB} = Additive analytical uncertainty, as calculated in Equation 7.

$$S_{FB} = \frac{mdl}{1.6449} \quad (7)$$

Fractional uncertainties, f, were obtained by taking the root mean square of the preceding five years' collocated precisions (cp). These are listed in Table 2.

Table 2. MDL floor and fractional uncertainty for the carbon species estimated from 2009-2014 collocated data.

Species	MDL floor	f
OC1	0.17	0.329
OC2	1.05	0.136
OC3	0.42	0.178
OC4	0.26	0.136
OP (OPTR)	0.23	0.251
OPTT	0.23	0.173
OCTR	1.43	0.116
OCTT	1.48	0.073
EC1	0.22	0.129
EC2	0.22	0.368
EC3	0.06	0.25
ECTR	0.37	0.155
ECTT	0.37	0.128
TCTC	1.48	0.25

4.3 Elements

Elements are measured using X-ray fluorescence (XRF) (PANalytical Epsilon 5) on PTFE filters from the SASS sampler. Because the XRF instrument reports areal densities, concentrations are calculated using Equation 8.

$$C = \frac{(\mathrm{AD} - \bar{\mathrm{L}}) * a}{\mathrm{V}} \qquad 8$$

Where,

AD = areal density (µg/cm^2)

$\bar{L}$ = median areal density of lab blanks for the same lot (µg/cm^2)

a = sample deposit area (cm^2)

V = sample air volume (m^3)

Fractional uncertainties, f, for elements are obtained by taking the root mean square of the preceding five years' collocated precisions (cp). These are listed in Table 3.

Table 3. Fractional uncertainty for elemental species estimated from 2009-2014 collocated data.

Parameter	f
Na	0.164

Mg	0.245
Al	0.252
Si	0.152
P	0.179
S	0.062
Cl	0.342
K	0.106
Ca	0.168
Ti	0.174
V	0.128
Cr	0.389
Mn	0.154
Fe	0.17
Ni	0.178
Cu	0.269
Zn	0.123
Se	0.25
Br	0.15
Pb	0.185

4.4 Reconstructed Mass

Reconstructed mass is compared to mass at collocated continuous monitors during data validation. Reconstructed mass (RCMN) concentration is calculated using Equation 9

$$RCMN = NHSO + NHNO + Soil + 1.8\,Cl + ECTR + OMC \quad (9)$$

Where,

NHSO = Ammonium sulfate concentration (see 4.4.1)

NHNO = Ammonium nitrate concentration (see 4.4.2)

Soil = Soil concentration (see 4.4.3)

Cl = Chlorine concentration as measured by XRF (Section 4.3)

ECTR = Total elemental carbon concentration by TOR (Section 4.2)

OMC = Concentration of organic mass by carbon (see 4.4.4)

For all of the terms in Equation **Error! Reference source not found.**, zero substitution is applied to negative values. The derived components of RCMN and described in the sections below. The mdl for RCMN is 0. Uncertainty is calculated as combination of the individual uncertainties via Equation 10.

$$\sigma_{RCMN} = \sqrt{\sigma_{NHSO}^2 + \sigma_{NHNO}^2 + \sigma_{Soil}^2 + (1.8\,\sigma_{Cl})^2 + \sigma_{ECTR}^2 + \sigma_{OMC}^2} \quad 10$$

4.4.1 Ammonium Sulfate (NHSO)

Sulfur is predominantly present as sulfate in the atmosphere, generally as ammonium sulfate $(NH_4)_2SO_4$, though also as ammonium bisulfate $(NH_4)HSO_4$, sulfuric acid H_2SO_4, gypsum $CaSO_4{\cdot}2H_2O$, and, in marine areas, sodium sulfate $NaSO_4$. In many cases, the particle will include associated water, but we omit this from the calculation. We assume that all sulfur is present as ammonium sulfate, and calculate NHSO concentration, MDL, and uncertainty as:

$$C_{NHSO} = C_S * 4.125 \quad 11$$

$$mdl_{NHSO} = mdl_s * 4.125 \quad 12$$

$$\sigma_{NSHO} = \sigma_S * 4.125 \quad 13$$

4.4.2 Ammonium Nitrate (NHNO)

This is the total dry concentration associated with nitrate, assuming 100% neutralization by ammonium. The concentration, mdl, and uncertainty are derived from the nitrate ion measurement and calculated as:

$$C_{NHNO} = C_{nitrate} * 1.29 \quad 14$$

$$mdl_{NHnO} = mdl_{nitrate} * 1.29 \quad 15$$

$$\sigma_{NSNO} = \sigma_{nitrate} * 1.29 \quad 16$$

4.4.3 Soil

The soil component consists of the sum of the predominantly soil elements measured by XRF, multiplied by a coefficient to account for oxygen for the normal oxide forms (Al_2O_3, SiO_2, CaO, K_2O, FeO, Fe_2O_3, TiO_2), and augmented by a factor to account for other compounds not included in the calculation, such as MgO, Na_2O, water, and CO_2. The following assumptions are made:

- Fe is split equally between FeO (oxide factor of 1.29) and Fe_2O_3 (oxide factor of 1.43), giving an overall Fe oxide factor of 1.36.

- Fine K has a non-soil component from smoke. Based on the K/Fe ratio for average sediment (*Handbook of Chemistry and Physics*) of 0.6, we use 0.6*Fe as a surrogate for soil K and then add the oxide factor for K $\left(\mathrm{K_2O}, \frac{39.1*2+16.0\ \mathrm{g/mol}}{39.1*2\ \mathrm{g/mol}} = 1.2\right)$ to get a total Fe factor of 0.72*Fe (0.6*1.2) for the potassium oxide in soil. This increases the factor for Fe from 1.36 to 2.08.
- The oxide forms of the soil elements account for 86% of average sediment; in order to obtain the total mass associated with soil, the final factors are divided by 0.86 (*Handbook of Chemistry and Physics*). The final equations for fine soil concentration, mdl, and uncertainty are:

$$C_{soil} = 2.2\ Al + 2.49\ Si + 1.63\ Ca + 2.42\ Fe + 1.94\ Ti \quad (17)$$

$$mdl_{soil} = max(2.2\ mdl_{Al}, 2.49\ mdl_{Si}, 1.63\ mdl_{Ca}, 2.42\ mdl_{Fe}, 1.94\ mdl_{Ti}) \quad (18)$$

$$\sigma_{soil} = \sqrt{(2.2\ \sigma_{Al})^2 + (2..49\ \sigma_{Si})^2 + (1.63\ \sigma_{Ca})^2 + (2.42\ \sigma_{Fe})^2 + (1.94\ \sigma_{Ti})^2} \quad (19)$$

For Equations 17-19, zero substitution is applied to any negative terms.

4.4.4 Organic Mass by Carbon (OMC)

To determine the total amount of organic mass associated with organic carbon, we assume a ratio of organic mass to organic carbon of 1.4.

$$C_{OMC} = 1.4 * C_{OCTR} \quad (20)$$

$$mdl_{OMC} = 1.4 * mdl_{OCTR} \quad (21)$$

$$\sigma_{OMC} = 1.4 * \sigma_{OCTR} \quad (22)$$

5. DATA PROCESSING CODE

This section describes the data flow through the data processing code used to calculate concentration, mdl, and uncertainty for all CSN parameters. Figure 1 outlines the flow of data from the filter and analysis specific database tables to final results. The wrapper function *csn_calculate_all* is the only function executed directly by the analyst (see 3.1); *csn_calculate_all* in turn calls several functions sequentially to calculate first measured and then

derived concentrations. Source code for the functions shown in Figure 1 is stored in the UC Davis source repository.

Figure 1. Flow diagram of the processing code in *crocker::csn_calculate_all*. The raw sampling and analysis data are inputs to the process. Rectangles represent R functions and arrows indicate inputs and outputs.