UCD CSN Technical Information #801C

CSN Data Validation

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 document (TI) is validating the analytical data from the CSN network. The analytical results from the previous steps in the data generation process are collected, compared, and summarized in plots. The user will apply qualifier flags and/or null codes as appropriate. This validated dataset will be prepared for delivery 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.

datvalCSN: A custom software package in the R language that contains the data validation code used to collect, compare, and flag the final results.

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

AQS: A database that is the central warehouse of EPA air quality data.

3. PROCEDURES

Data validation is performed using the *crocker* and *datvalCSN* R packages, which are developed and maintained by UC Davis specifically for data processing, monitoring, and validation of the CSN and IMPROVE data. Data validation is performed by the UC Davis data management team on monthly batches of data (a calendar month of sample start dates). Validation occurs in three stages:

1. Collect necessary measurement data
2. Compare results
3. Review summary plots
4. Review every data point via web application

UC Davis is responsible for the first two levels (levels 0 and 1) of the data validation process; however, the levels specified are intended as general guidelines. The CSN data delivered to DART is considered Level 1 data while data delivered to AQS is considered to be Level 2 data. The levels are applied to CSN as follows:

Level 0

Data at this level are, in essence, raw data, obtained directly from the data-acquiring instruments. These data can be reduced or reformatted, but are unedited and unreviewed, without any adjustments for known biases or problems that might have been identified during preventative maintenance checks or audits. These data may monitor instrument operations on a frequent basis (e.g., ambient temperature of sampler every fifteen minutes). Average times represent the minimum intervals recorded, and these data may need to be aggregated to obtain averages for the sampling periods (e.g., 24-hour average). Level 0 data have not been edited for instrument downtime, nor have procedural adjustments for baseline shifts, span changes, or known problems been applied. Examples of data at Level 0 validity in CSN are:

- 24-hour averaged pressure, temperature, and flow data recorded from sampler user interface during sample change procedures;
- balance measurements before automated validity tests;
- XRF raw spectra;
- Sample date and sampling time before consistency checks.

Level 1A

Data at this level have passed several automatic and manual qualitative reviews for accuracy and completeness. The focus of Level 1A validation is to obtain as complete a dataset as possible. CSN Level 1A data validation consists of:

- Replacing invalid data values with NA (the invalid data code) and setting status flags to reflect sampler malfunctions, site or laboratory analyst errors, or power outages;
- Reviewing analyst, Amec, and laboratory comments to verify consistency between records and correct any typographical errors;
- Verifying analyst, Amec, and laboratory comments for questionable records;
- Replacing filter type ID with trip blank ID in the event of failure to sample;
- Identifying, investigating, and flagging data that are beyond reasonable bounds or that are unrepresentative of the variable being measured, including:
  - Sulfur/sulfate and potassium/potassium ion ratios outside a factor of two,
  - Gravimetric or collocated mass/reconstructed mass ratios outside a factor of two,
  - Anion/cation ratios outside the range 0.5 to 2 (temporary limits),
  - Z-scores of the OC/EC ratios outside the range -1 to 1 (temporary limits);
- Examining daily flow rates based on the results of the *flow.check* function that identifies abnormal flow rates and significant variations over 24-hours;
- Setting qualifier flags when deviations from nominal operational settings have occurred (e.g., temperature or pressure outside instrumental tolerances);
- examining the field blank analyses for evidence of swaps with sample filters.

Level 1B

Data at this level have passed additional quantitative and qualitative reviews for accuracy and internal consistency. Objective collocated measurements and internal consistency tests are applied by the data validation manager. Discrepancies that cannot be resolved are reported to the measurement laboratories for investigation. Data that deviate from consistency objectives are individually examined for errors. Obvious outliers (e.g., -85 °C temperature) are noted with a qualifier AQS flag. Changes to the data (e.g., swapping dates on consecutive samples) are recorded and documented by providing comments to the DART reviewers. Level 1B time-series data review is conducted on a site-by-site basis using a combination of *datvalCSN* package functions and Shiny web applications. The historical archive of CSN data is used to place new measurement values in context with previous measurements. CSN level 1B data validation includes:

- Comparing sulfur and sulfate concentrations as well as potassium and potassium ion;

- Comparing organic carbon to elemental carbon for both blank corrected and uncorrected concentrations;
- Comparing anions to cations in a molar mass balance;
- Comparing reconstructed mass and collocated PM2.5 mass;
- Examining individual data points identified by the various checks as potential sample swaps;
- Comparing light absorption to elemental carbon concentrations;
- Comparing the analytical data to expectations based on prior years.

Level 2

This level of data review is applied after the submission of concentration files to DART, when the results from the analytical laboratories are made available to the reporting agencies. At this level, the data are reconciled with local events and expertise. The first assumption upon finding a measurement that is inconsistent with physical expectations is that the unusual value is due to a measurement error. If, upon tracing the path of the measurement, nothing unusual is found, the value can be assumed to be a valid result of an environmental or statistical cause. Upon completion, Level 2 validated data are sent back to UC Davis for a final review. This review is to ensure consistency between user comments and data transfers. UC Davis will not make any changes to DART-reviewed data unless there has been a documented discussion of the changes with the DART reviewers.

The process of performing the three validation procedures is outlined below.

3.1 Collect Measurement Data

In the previous step, laboratory results for elements, ions and carbon fractions are processed into mass per volume of air and posted into the *analysis.Results* table. Additionally, collocated data should be collected from the AirNowAPI web service via file transfer protocol (FTP). All data can be collected using functions in the *datvalCSN* package.

To collect records for all measured and derived parameters from the laboratories, the analyst (either the Data Manager or a member of their staff) will open an R environment (such as RStudio) and run the following command[1]:

[monthData] <- datvalCSN::csn_get(start.date = ['YYYY-MM-DD'], end.date = NULL, server = 'production')

This command collects records from the *analysis.Results* table starting from date (*YYYY-MM-DD*) and ending either with another date (*YYYY-MM-DD*) or all available records past the start date (input *NULL*). The results will be returned in memory to the variable *monthData*. The last argument in the command specifies that the calculations will use the “production” database (i.e., the CSN operational database).

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

To collect data from the AirNowAPI web service, the analyst can run the following command:

```
airnow <- datvalCSN::get_airnow(dest = '', res = 'Daily', param = ['PM2.5-24hr'], Site = 'CSN', start = ['YYYYMMDD'], end = ['YYYYMMDD'], make.file = FALSE, days = 45)
```

This command will extract the desired data from the posted values on the AirNowTech server. This command is a versatile tool to collect different types of data (*res = 'Daily'* or *'Hourly'*) and different parameters (*param* can be set to collect all available parameters or can be set to various available parameters such as *'OZONE-1HR'*, *'NO'*, and *'PRECIP'*). Specific sites can be chosen in un-hyphenated form (*'AABBBCCCC'* for AA State Code, BBB County Code, and CCCC Site ID). Start and end dates can be chosen or start can be set to *NULL* and the days variable will determine the start date from the end date specified. See help page for additional information by typing *?get_airnow* into the command prompt.

The *csn_validation_mass* function simplifies collection of data from *AirNowTech* with optimal settings for data validation. The input for this function is the resultant variable from the *csn_get* command. For example:

```
airnow <- datvalCSN::csn_validation_mass([monthData])
```

3.2 Perform Checks and Comparisons

The bulk of the validation process is performed by various checks. Each check compares resulting values against pre-defined limits as well as comparable parameters. To perform all automated checks and prepare data for graphical interpretation, the analyst can execute the following command:

```
[allData] <- datvalCSN::csn_validate(['MM'], ['YYYY'])
```

This command will collect the necessary data from the *analysis.Results* table and AirNowTech, perform all automated checks and comparisons for the month and year(*['MM'], ['YYYY'])* to be validated, and can write the output file for delivery to a specified location. Essentially, the *csn_validate* and *plot_csn* functions are the only two commands an analyst may need to use and it is advised for the analyst to review the help pages for these two functions. The individual checks are described below.

Operational parameters such as sample flow rate and operating temperature should have been checked both at the sampling site as well as in the filter handling laboratory. Two functions in the *datvalCSN* package, *Flow.check* and *Amb.check*, compare reported values with instrumental limits to determine if additional flagging is needed. Both functions require an input from the filter.Filters table, which can be retrieved either through a SQL query or by running:

```
devtools::load.all('.')
```

Running the *Flow.check* command will reveal a list of records that have aberrant flow values but have not been invalidated. The Null flag "AH" will be applied to all of these filters, which represents "Sample Flow Rate out of Limits" in the AQS database. At this time, the *Amb.check* function will not invalidate samples based on ambient temperature and pressure, but instead append the appropriate flag notifying end users that the instrumental operating parameters were outside the manufacturer's specifications.

The following checks exclusively use the results from the *csn_get* command:

- *ions.check(results)*
- *carbon.check(results)*
- *xrfic.check(results)*

For each command, a summary of the results is produced with outliers indicated. Sample records are not invalidated during this process, but may be automatically flagged if comparison results lie outside the predefined limits. If flagged, a comment is added to the record to indicate which check produced the outlier flag.

A particularly important check is the *swap.check* function. This check calculates two indices, one assuming two samples were not swapped and the other assuming the samples were swapped. These indices are combined with the sulfur/sulfate ratio to identify potentially swapped samples for Teflon and Nylon filter samples. If *xrfic* is the variable result of the *xrfic.check* execution, then the analyst can enter the following command:

```
[swapped] <- datvalCSN::swap.check(xrfic)
```

Similar to the *xrfic.check* output, the swapped variable will contain concentrations for sulfur and sulfate, outliers by the sulfur/sulfate ratio, and potentially swapped samples based on the calculated indices.

The other functions performed by the *csn_validate* command include: *get_blank_swaps*, *mass.check*, *csn_complete*, *qualifiers*, and *csn_X*. The *get_blank_swaps* function compares field blanks with sample filters to determine if a filter swap may have occurred. The *mass.check* function uses the collocated mass concentrations retrieved from AirNowTech to compare reconstructed mass with collocated mass. When available, it also compares gravimetric mass with reconstructed mass. The *csn_complete* function checks that all reportable parameters for every sample event have records. If not, the *csn_complete* function fills in with a "Miscellaneous Void" flag for clarification during DART validation. The *qualifiers* and *csn_X* functions make lists of all records with qualifier and null flags, respectively.

3.3 Review Summary Plots

To generate a summary plot, the analyst can use any number of R's plotting tools. There are quite a few custom plotting tools in the d*atvalCSN* package. The analyst should familiarize themselves with each function by reviewing the help pages. To view the results of the *xrfic.check* at the CSN site 01-073-0023, the analyst can type:

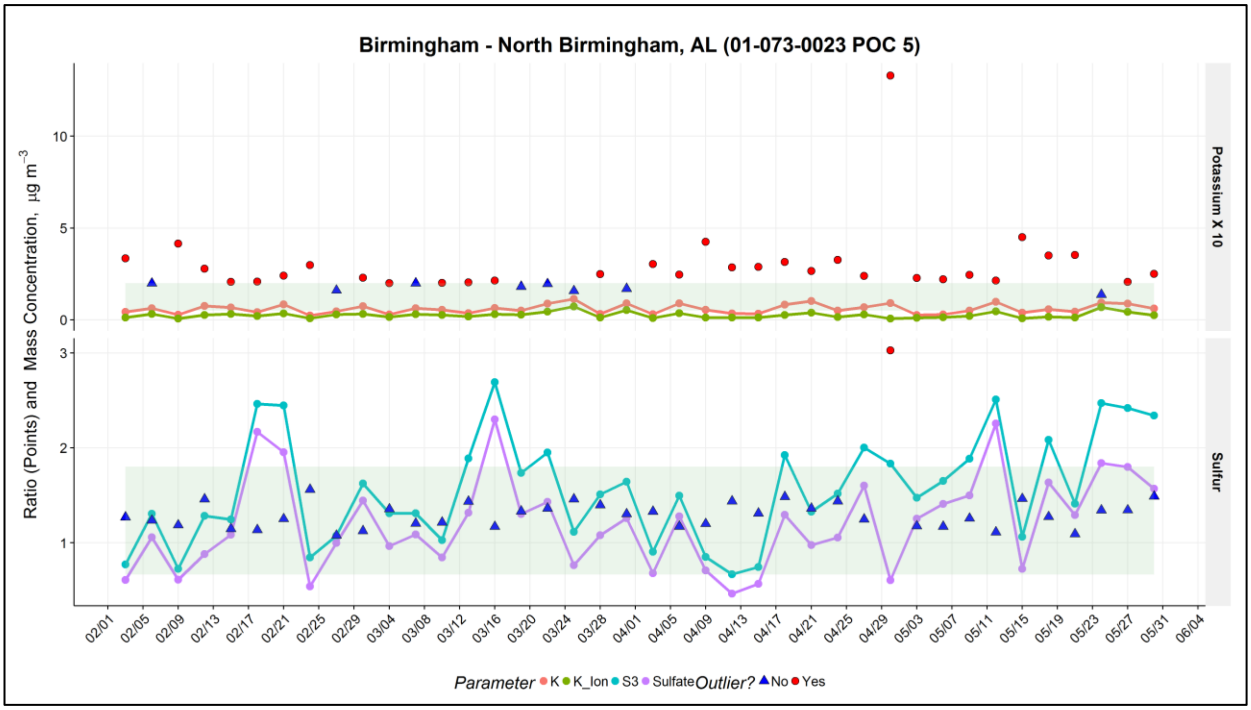
[results] <- datvalCSN::csn_get()

[xrfic] <- datvalCSN::xrfic.check([results])

datvalCSN::plot_xrfic([xrfic], site = '01-073-0023', POC = '5')

The resulting graph would be similar to Figure 1.

Figure 1. Sample output of *plot_xrfic* for comparing XRF analysis to IC analysis. The turquoise and pink lines are the sulfur and elemental potassium measurements by ED-XRF, the purple and green lines are sulfate and potassium ion measurements by IC, and the points illustrate whether the ratio of the two measurements can be considered an outlier or not.



For convenience, the analyst can enter the command:

create_monthly_plots(['Jan'], [janData])

to produce plots for each site and each check. The default directory for saved plots is *U:/CSN/QA/Temporary Plots/*, with a subfolder based on the month of interest (e.g., *Jan Plots*). The*['Jan']* parameter is a character string that will be used to name the plot folder while the *[janData]* parameter is the “validation” object created by the *csn_validate* function. The analyst should cycle through each check and each site to identify potentially swapped samples as well as aberrant data that has not yet been flagged.

3.4 Review all data via web application

In order to facilitate efficient review of all data for a given month, custom web applications (webapps) were developed. Three important webapp tools for the CSN analyst include:

- CSN Data Management Site – csn.crocker.ucdavis.edu

- CSN Data Explorer – analysis.crocker.ucdavis.edu:3838/csnSites/
- CSN Status Explorer – analysis.crocker.ucdavis.edu:3838/csnStatus/

Eventually, the checks previously discussed will be incorporated into the webapps for an even more dynamic and collaborative review process.

The CSN data management site is an online access point to interact with the CSN database. With this tool, the user can look up filter, batch, and site information. Qualifier and null flags can be assigned here with an explanatory comment. Each time a flag is changed, a timestamped record is made of the change. This is also the portal for importing data from Amec and DRI. Files from the respective folders in the networked U drive are uploaded following the instructions in TI 351A.

The CSN data explorer page has multiple tabs for viewing the resultant data in various forms. The analyst should familiarize themselves with the intuitive controls, especially within the "Explorer", 'Validation", and "Early Review" tabs. The Parameter Review tool under the "Validation" tab must be reviewed by the analyst for each monthly dataset. With this tool, every reported data point can be inspected quickly and efficiently. It is also important to note the "Early Review" tab may be used prior to receiving a complete, monthly dataset. This tool utilizes reported sulfur and sulfate mass loadings before the post processing steps described in TI 351B.

The CSN status explorer is useful for monitoring the progress of analyses and timelines for delivery. There are myriad data views and the user should explore the various figures and tables. An important portion to review for validation is the "Analysis Completeness" tab. Here, the reviewer should inspect filter records not analyzed by their respective analyses. All unanalyzed filters should either be flagged as invalid, or be queued for analysis.

4. DATA VALIDATION EQUATIONS

The following section presents the equations used to calculate swap check indices. These calculations are performed by the *datvalCSN* R package.

The ratio of sulfur to sulfate in the ambient atmosphere is generally well known. Since the majority of sulfur-bearing aerosols by mass are in the form of sulfate, the stoichiometric ratio of 3 x sulfur / sulfate is typically close to one. The calculated indices of the *swap.check* function exploit this tendency. The first index assumes that two subsequent samples are not swapped. The ratio of sulfur by XRF to sulfate by IC is subtracted by one for each sample, and the two results are multiplied. Since both ratios are expected to be close to 1, a small number is expected. Inversely, the second index assumes two samples have been swapped. This irregularity would result in a significantly larger number than index 1. Mathematically,

$$Index1 = \left(\frac{S3_1}{SO4_1} - 1\right) * \left(\frac{S3_2}{SO4_2} - 1\right) \tag{Eqn. 1}$$

$$Index2 = \left(\frac{S3_1}{SO4_2} - 1\right) * \left(\frac{S3_2}{SO4_1} - 1\right) \tag{Eqn. 2}$$

Where,

$S3_x$ = sulfur concentration (µg/m^3) multiplied by 3

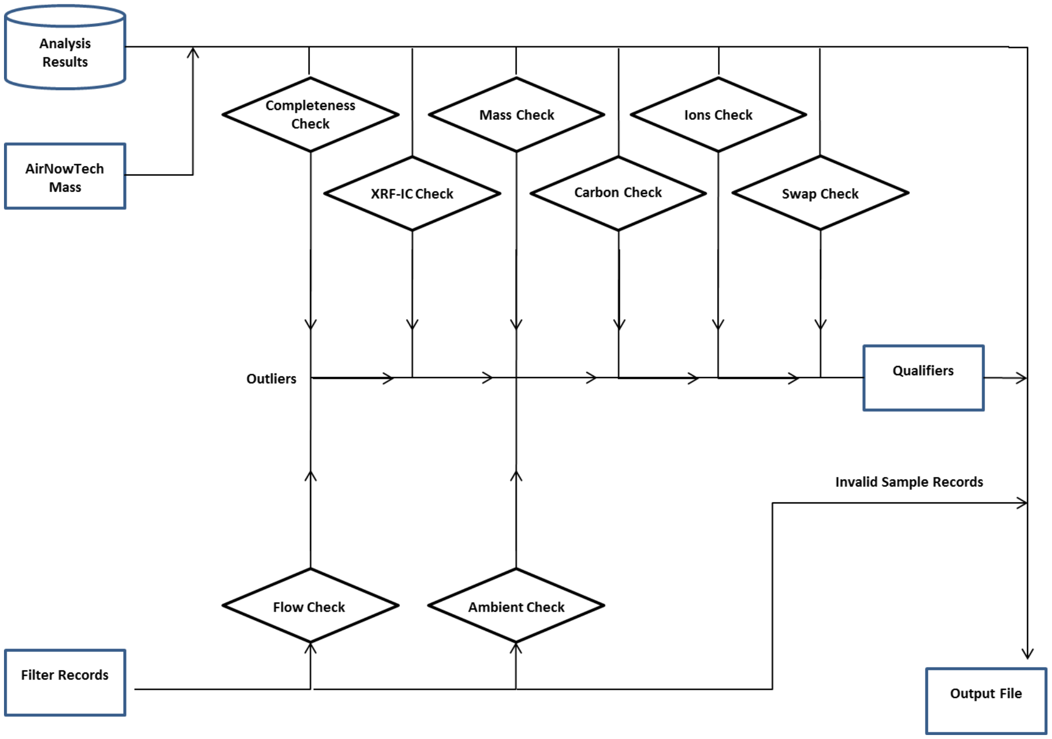
$SO4_x$ = sulfate concentration (µg/m^3)

Subscripts denote the sampling event of interest (1 for target event, 2 for subsequent event). Empirically, potential swaps are indicated by an Index1 < -0.03 and an absolute value of Index2 < 0.05.

5. DATA PROCESSING CODE

This section describes the data flow through the data validation code used to execute all CSN validation checks. **Error! Reference source not found.** outlines the flow of data from the filter and analysis results database tables to final results. The wrapper function *csn_validate* is the only function executed directly by the analyst (see **Error! Reference source not found.**); *csn_validate* in turn calls several functions sequentially to generate data frames with outliers identified. Source code for the functions shown in **Error! Reference source not found.Error! Reference source not found.** is stored in the Crocker source repository.

Figure 2. Flow diagram of the validation code in *datvalCSN::csn_validate*. Rectangles represent data files, diamonds represent R functions, cylinders represent databases, and lines represent inputs and outputs.


Analysis Results
AirNowTech Mass
Completeness Check
Mass Check
Ions Check
XRF-IC Check
Carbon Check
Swap Check
Outliers
Qualifiers
Invalid Sample Records
Flow Check
Ambient Check
Filter Records
Output File