Using SAS to improve the quantification of environmental chemistry samples.
Charles L. Dow and Anthony K. Aufdenkampe
Stroud Water Research Center, Avondale, PA

ABSTRACT
Quantifying chemistry samples, the process of applying a calibration curve defining the relationship of an instrument response across a set of known standard concentrations to sample values, is an often overlooked aspect of laboratory work. Instrument software packages have calibration routines embedded but these packages offer limited flexibility in applying and evaluating different types of calibration curves. We have developed a semi-automated routine that uses various SAS products/procedures to capture instrument data, develop calibration equations, run diagnostics on the developed calibration equations, and present the output in an easily accessible and interpretable format. We used the following products/procedures (among others) in developing the quantification routine: SAS Macros at various points for performing iterative functions, PROC REG for developing the calibration equations, PROC REPORT/SAS GRAPH in conjunction with ODS for presenting diagnostics and final calibration curves as HTML output. The entire process improved both accuracy and precision of laboratory analyses as well as allowing for faster turn around in producing final analysis results.

INTRODUCTION
The process of quantifying chemical samples inevitably involves multiple software packages beginning with the instrument analysis software used to establish the calibration curve. From the instrument, data are then exported to a software package for post processing of the data to obtain final concentration values, and ultimately to another package performing statistical analyses of the data. Beyond the cumbersome nature of data migration between software packages, and keeping track of the data stream, an overarching concern in this process is data integrity. A secondary concern is flexibility, from how calibration curves are calculated, to rejecting calibration standards based on defined quality control criteria, to post-processing of concentration data. SAS has allowed us to combine all these steps under the umbrella of single, unified, data processing routine.

We are presenting an overview of the SAS routines used to capture instrument data, establish calibration curves, perform quality control checks of the resulting calibration curves, and calculate final sample concentration values for the analysis of molecular tracers in environmental samples. Occasional examples of SAS code are also provided in this overview. In describing the SAS routines used in our quantification process it will be necessary to describe certain components of the laboratory and subsequent instrument analyses. We will not go into great detail regarding the specific laboratory or instrument analyses; rather just enough detail will be given in order to provide proper context for the use of any specific SAS routine.

BACKGROUND
The specific laboratory analyses we cover in this paper are for a group of chemicals referred to as molecular tracers which are processed by liquid-solid extraction techniques and analyzed by capillary gas chromatography – mass spectrometry (GC-MS) (Aufdenkampe et al. 2006). Molecular tracers, such as caffeine or certain fragrances found in detergents, are measured in stream samples as rather specific 'tracers' of contamination sources. The data output from the GC-MS is a chromatogram, which is a plot of instrument detector response against time (chromatography is the process of separating chemicals within a mixture). Individual chemicals are identified by a 'peak' in the chromatogram at a give time, and the area under that peak, which is related to the chemical’s concentration, is calculated by the instrument software’s peak identification and integration routines. To arrive at chemical concentrations, the peak area values have to be converted to concentrations via a relationship between the known concentrations of the compound, analyzed in a reference standards mixture at 3-6 concentration levels, and their respective peak areas; i.e. a calibration curve. Instrument software is generally used to produce these calibration curves, however, little flexibility is allowed within the software to modify the curves if needed, either by eliminating calibration standards that do not pass certain quality controls, or flexibility in defining different types of curves.
**DATA CAPTURE**

The instrument data for a given run are output as text files referred to as quantitation reports (Figure 1). These reports contain sample and instrument-run information (upper highlighted portion of Figure 1) and compound-specific information extracted from the chromatogram (lower highlighted portion of Figure 1). Most notable for our purposes are the peak areas (‘Resp’ in lower highlighted box of Figure 1) for the target ion (the compound of interest, ‘Tgt’ in Figure 1) plus ‘confirmation’ ion information (i.e. ‘Q1’ in Figure 1) needed to positively identify the target compound.

All relevant information in the quantitation reports are captured and placed in tabular format using a series of if/then statements and proc sql statements that ultimately put all of this information for a given compound into a single row of data. Proc sql served as a (relatively) efficient alternative to a series of proc transpose statements by reading in the same input dataset multiple times and extracting different compound-specific data using table aliases:

```sql
proc sql;
    create table gcms3 as
    select distinct g1.projidsamp, mean(g2.rt_exp) as rt_exp, t.qion as qion,...
    from gcms2 as g1, gcms2 as g2, gcms2 as t,....
    where (g1.cmpd=g2.cmpd=t.cmpd) and (g2.ionid='T' or g2.ionid='C2') and ...
    group by g1.cmpd
    order by g1.cmpd;
    quit;
```

where:
- projidsamp = identifier for the specific sample
- cmpd = specific molecular tracer (i.e. compound)
- rt_exp, ionid, qion,... = other molecular tracer-specific information

Part of the process of capturing instrument data are a series of quality control measures written into the SAS program such as a flag for those compounds having a low signal-to-noise ratio (i.e. the chromatogram peak for that compound does not plot much higher then the noise in the chromatogram baseline). Another flag is for poor confirmation, where the ratio of confirmation (i.e. ‘Q1’ in Figure 1) to quantitation (‘Tgt’) ion peak areas is significantly different then observed in standards.
Additional sample information, such as sample amount, dilution factor, etc., is entered into a text file (or an Excel file) that is converted into a SAS data file using a SAS program referred to as a 'builder' which tracks changes to the resulting SAS data file each time new information is added to the text file and the corresponding program is rerun. All instrument files (as SAS files) are combined and linked to the file containing additional sample information using the instrument file name as the primary key.

CALIBRATION ROUTINE

CALIBRATION STANDARD QUALITY CONTROL

There are two quality control steps involving the calibration standards that are run prior to the calculation of calibration curves. The first of these two quality control steps is an iterative process that looks at the internal standard peak area for each calibration standard. The internal standard is added to each and every sample (including calibration standard mixtures) just prior to injection into the instrument, and is used as a means of normalizing the instrument response between samples to account for drift in detector sensitivity over time. The internal standard should have a similar response from one sample to the next, such that any large deviation from this instrument response is cause for concern. The first quality control check looks at the deviation in the peak areas of the internal standards from across all calibration standard runs, in the order that the calibration standards were injected into the instrument. If the deviation of any one internal standard is >30% of the mean of the group of internal standards within a given iteration, that calibration standard injection is rejected and the check is rerun. If no deviations exist, the next calibration standard injection is brought into the check, with the iterations ending once the last calibration standard injection (again, in terms of the analysis order) has been checked.

The iterative check process is done using the following macro:

Macro variables are:
- insql = incoming dataset containing calibration standard data
- out1sql = 1st internally (within the macro) generated dataset – maintaining naming convention of the incoming dataset.
- out2sql = 2nd internally generated dataset
- out3 = outgoing dataset containing standards that were not rejected; used in the next iteration of the macro, or becomes final calibration dataset to use in remainder of program
- xout4 = outgoing dataset containing the rejected standards

```
%macro istdsel(insql, out1sql, out2sql, out3, xout4);
proc sql;
  create table &out1sql as
  select istdcmpd, runnum, stdconc1, stdconc_istd, pka_istd,
       abs((100*(pka_istd-mean(pka_istd))/mean(pka_istd))) as istddev format 8.2
  from &insql
  group by runnum, istdcmpd;

  create table &out2sql as
  select *, max(istddev) as maxistddev
  from &out1sql
  group by runnum, istdcmpd
  order by runnum, istdcmpd, stdconc1;
quit;

  data &out3 &xout4;
  set &out2sql;
  if istddev = maxistddev and maxistddev > 30 then do;
    xoutcal = 'X1';
    output &xout4;
  end;
  else output &out3;
run;
%mend;
```
The second set of quality control measures applied to the calibration standards involves the response ratio and the amount ratio of the individual standard compounds. The response ratio is the ratio of the peak area for the compound of interest to the peak area of the internal standard; the amount ratio is the corresponding ratio between concentrations (that is, ‘known concentrations’ since these are prepared standards). The response and amount ratio values are used in calculating the calibration curve. There are two criteria used in this second quality control check; (1) the response ratio for a given compound should not be greater than 30% different from the mean of the response ratios of that compound in the three preceding injections of calibration standard mixtures of the same concentration, and (2) the response factor (the response ratio divided by the amount ratio) for a given standard compound should not be greater than 60% different from the mean of all response factors for different concentrations of standard mixtures injected on the same day.

Once these two sets of quality control checks have been applied to the set of calibration standards, the development of a calibration curve begins.

ESTABLISHING THE CALIBRATION CURVE

The power and flexibility of this entire calibration routine is most evident in the ability to define any number of calibration curves, reflecting the nuances in the relationship between the instrument response (i.e. peak area) and concentration of the standards. For the molecular tracer analyses, these nuances are shown in the example calibration curve diagnostic plots for a specific polycyclic aromatic hydrocarbon (PAH) (Figure 2). When viewing the scatter plot of the instrument response ratio to the amount ratio (upper lefthand plot in Figure 2), it appears that the relationship is linear through all the points. Yet when examining the residuals from the linear regression fit using all of the calibration standards (upper right-hand plot, Figure 2) it becomes apparent that the relationship is in fact curvilinear. Not only is the relationship not linear across all calibration standards, the residual plot for all standards suggests that the ‘low-end’ calibration standards (shown in blue in Figure 2) have a different relationship than the ‘high-end’ points (shown in red in Figure 2). Therefore the standards can be split into two groups, each fitted with a unique calibration curve. The specific point at which the calibration curve was split was determined from examining initial calibration curves of all the molecular tracers to be analyzed. Further, these unique relationships were found to be repeatable and consistent over time. We used a quadratic equation to fit all curvilinear relationships.

Occasionally, the quality control checks covered in the previous section actually lead to the elimination of one or more calibration standards which in turn requires modifications to the defined calibration curves. Therefore, even though we established that the calibration curves are curvilinear, eliminating one or more standards may require reverting back to linear curves. In total, fourteen separate curves are defined for each analysis run to cover all possible calibration curve scenarios from having all standards available, to the elimination of one or more standards within either the low-end or high-end groups, to defining ‘below calibration’ curves to cover situations where
the instrument response for a given sample is below the lowest calibration standard. One of the fourteen defined curves is used just for diagnostic purposes; e.g. to provide the linear fit and corresponding residuals using all available calibration standards, as shown in Figure 2.

We used **`proc reg`** to define all of our calibration curves using the following template:

```
proc reg data=in_data outest=output_regstats edf noprint;
  model responser = amountr [amountr2] [/noint];
  by cmpd runnum type;
  output out=residual_data r=resid;
run;
```

where:
- output_regstats = dataset containing, among other items, the actual equations (i.e. regression coefficients) and the resulting R² values (a `quality of fit` measure for each equation, specifically of the variance in the dependent variables; i.e. responser; explained by the resulting equation).
- residual_data = residuals for each equation needed for diagnostic purposes
- responser = calibration standard response ratios
- amountr (amountr2) = the calibration standard amount ratios (plus, if applicable, the squared values of the amount ratios for the quadratic equations)
- / noint = no intercept option; to force regression through the origin; only used in the ‘below calibration’ scenario.

There are two datasets produced from this routine defining the calibration curves. The first dataset contains the calibration equations (i.e. regression coefficients) for each calibration scenario which are ultimately combined with the instrument responses of the samples to produce final sample concentration values. The second dataset contains the evaluation criteria specific to each calibration curve defined in the routine, such as regression residuals, needed to evaluate the adequacy and appropriateness of each calibration curve. The data in this second dataset are what lead to the diagnostic plots shown in Figure 2 as will be described in greater detail in the following section. We use **`proc gplot`** to produce these plots shown in Figure 2.

**DIAGNOSTICS OF THE ESTABLISHED CALIBRATION CURVE**

An important component to the entire process of establishing calibration equations is examining the appropriateness of each equation. Were any standards rejected and why? Does the defined calibration equation have the expected shape? Can the defined calibration equation accurately estimate a separate, known standard (i.e. a check standard)? These are important evaluation questions that have to be asked for each and every calibration equation produced. Having an automated routine in place to address these evaluation criteria and to produce easily interpretable output are also key to making this entire quantification process work properly.

![Example diagnostic tabular output for calibration curve evaluation.](image)

**Figure 3.** Example diagnostic tabular output for calibration curve evaluation. Information provided for calibration standards includes actual concentration, predicted concentration, response factor (response ratio/amount ratio), and relative residuals (100*residual/predicted value). Information is also provided for one or more check standards as an independent check on the accuracy of the calibration equation; this includes known and predicted values; plus percent recovery values.
Along with the plots produced for the calibration data of each compound as shown in Figure 2, are separate tables of diagnostic statistics (Figure 3). The first part of this table presents information for each calibration standard including its response factor, predicted concentration, which calibration equation was used to generate the predicted concentration (and which calibration equation will be used to predict for sample concentrations in that range), and relative residuals (residual divided by the measured response ratio, expressed as a percentage). The second part of this table presents results of applying the appropriate calibration equation to check standards. These check standards are separately analyzed standards having known concentrations that are not included in defining the calibration equation, but rather are used as a check on the resulting calibration equation. As shown in Figure 3, if any of the check standard values are outside of some pre-defined percent-recovery criteria (in this case, estimated value being > ± 20% different from the known value), the values are highlighted to alert the analyst to potential problems with the defined calibration equation for the given molecular tracer.

To generate the table of evaluation data we use `proc report` in conjunction with a `proc format` statement for highlighting check standards that do not meet the aforementioned control criterion:

```sas
proc format;
  value predcrit low-<-20='medium yellow'
   -20-20 =''
    20<high ='medium yellow';
quit;
proc report data=intabledata [other_options]
  style(report)= [cellspacing=1 cellpadding=3in borderwidth=5 bordercolor=blue
                font_size=8pt]
  style(header)= [foreground=black font_size=9pt font_face=arial]
  style(column)= [foreground=green font_face=arial font_size=8pt]
  style(lines)= [foreground=white background=black font_style=italic font_size=2pt];
  column stdtype ..[more_columns... ] chkstdpctd;
define stdtype / group "Standard,Type";
define [more_columns];
define chkstdpctd / display "Check Std.,% Recovery" style=[background=predcrit.]
  center;
run;
```

where:

- `stdtype` = calibration or check standard type (i.e. actual concentration)
- `chkstdpctd` = check standard percent recovery value (the `style=` option is controlled by the proc format that precedes the proc report statements).

The final step in producing the calibration-curve output is to put all the necessary parts, for each compound analyzed, into an organized and easily accessible (by non-SAS users) format. We chose HTML (HTML 4.0 in SAS v.9) using SAS Output Delivery System (ODS) as shown in the following statements:

```sas
ods html frame="CALIBPLOTS_frame.html"
    contents="CALIBPLOTS_contents.html"
    body="CALIBPLOTS_body.html"
    path=gsasfile [the pathway is defined in a filename statement];
ods trace on;
[series of macro statements to produce the plots and tabular output for every compound];
ods trace off;
quit;
ods html close;
```

ODS offers a very simple way of producing non-SAS specific formatted output and it is an important component to making the entire quantification routine user-friendly.
GENERATING THE FINAL DATASET

Having all of the calibration curves generated through this routine creates the challenge of what curve should be used to generate final concentration values for a given sample. Part of this decision process has already taken place in the generation of the calibration curves (e.g. selecting linear versus quadratic fits based on the number of available calibration standards). However, having two sets of curves for ‘low-end’ and ‘high-end’ values requires one final set of decision tools in order to generate final concentration values. As alluded to earlier, the split in calibration standards was based on initial investigations of the calibration results showing two distinct instrument response relationships for values less than or greater than a single standard (the 4 ng standard). The instrument response (e.g. peak area) for the 4 ng standard of each calibration curve therefore establishes the final criterion for selecting which curve to apply to a particular sample value. So, if the peak area is less than the peak area of the 4 ng standard then the low-end calibration curve is used, and likewise, if the sample peak area is greater than the 4 ng peak area the high-end curve is applied to calculate a final concentration. This decision process for applying calibration curves is part of a last SAS program in this quantification routine that produces the final dataset of sample concentration values.

MAKING THE WHOLE ROUTINE USE FRIENDLY (OR ALLOWING NON-SAS USERS IN)

A critical function of this entire quantification process is ease-of-use, specifically for non-SAS users. Towards this end, we’ve established a series of batch files for running the various components of the quantification process. We use batch files as a means of generating permanent log and lst (list) files to document our SAS-driven data management and analysis. Specific to the quantification process, the use of batch files allows the laboratory technician to run the entire process without having to know SAS code or even to open SAS interactively. Separate batch files are established for reading in and creating SAS data files from the instrument files and then for running the calibration routine, generating the diagnostic information, and creating the final SAS dataset of sample concentrations.

CONCLUSIONS

Our original intent in using SAS as part of the quantification routine for molecular tracers was to allow more flexibility in defining calibration curves. What we ended up with though was much more. This routine integrates all of the data management and analysis efforts in taking ‘raw’ instrument data and producing final sample concentration values within a single software platform. This routine has eliminated many sources of error previously part of generating these data by automating data capture, manipulation, calibration curve development, and proper application of the resulting calibration curves. The time it takes to generate final sample concentration values has also dramatically decreased through the use of this quantification routine. However, challenges remain. The laboratory analyst running the routine does not have any mechanism for checking for or troubleshooting data management errors. Such troubleshooting still has to be carried out by someone familiar with SAS. The SAS programs themselves are not terribly efficient, and are not easily adaptable to other instruments and quantification routines (although we have modified the routine for other instruments).

By far, the biggest improvement offered through this quantification process is the more accurate and precise analysis of samples having very small concentrations of molecular tracers. The simple ability to split the calibration curve into two parts, a low-end versus a high-end, has greatly increased our ability to measure lower compound concentrations beyond what was previously possible.

REFERENCES


ACKNOWLEDGMENTS

SAS is a Registered Trademark of the SAS Institute, Inc. of Cary, North Carolina.
CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Charles L. Dow  
Director of Information Services  
Stroud Water Research Center  
970 Spencer Road  
Avondale, PA 19311  
Work Phone: 610-268-2153 (x259)  
Fax: 610-268-0490  
Email: cdow@stroudcenter.org  
Web: http://www.stroudcenter.org/about/Dow.htm

Anthony K. Aufdenkampe  
Principle Investigator,  
Organic and Isotope Geochemistry  
Stroud Water Research Center  
970 Spencer Road  
Avondale, PA 19311  
Work Phone: 610-268-2153 (x263)  
Fax: 610-268-0490  
Email: aufdenkampe@stroudcenter.org  
Web: http://www.stroudcenter.org/about/Aufdenkampe.htm