Documentation of the template scripts for the data analysis

EMPIR 19NRM03 SI-Hg

National
Metrology
Institute

Federica Gugole

21 September 2023

VSL
Thijsseweg 11
2629 JA Delft
the Netherlands
phone +31-15-2691500
fax +31-15-2612971
vsl@vsl.nl

Summary

As part of the EMPIR 19NRM03 SI-Hg project a calibration protocol for elemental mercury gas generators [1]. Together with the protocol two software templates have been developed using Excel and Python 3: one template for the single point calibration, and one for the multi point calibration. The template for the single point calibration calculates

- the output ratio between the measurements of the candidate and the measurements of the reference generator;
- the corrected candidate concentration;
- the complete uncertainty budget associated to the measurement.

The template for the multi point calibration calculates the above listed items for each setpoint plus

- the coefficients of the calibration curve
- the uncertainty associated to the coefficients of the calibration curve.

Details about the equations implemented can be found in [1].

Contents

Summary II

Contents IV

1 Setup 1

1.1 Software requirements . . . 1

1.1.1 Interaction between Excel and Python . . . 1

1.2 Input from user . . . 2

1.3 Data . . . 3

1.3.1 Manual pre-processing of the data . . . 3

1.4 Code execution . . . 4

2 Data processing 5

2.1 Single point analysis . . . 5

2.2 Multi-point analysis . . . 5

3 Validation 6

1 Setup

1.1 Software requirements

The templates for the processing of measurement data taken for the calibration of mercury generators are Python 3 powered Excel files. Therefore, to be able to execute them you need to enable macros in Excel, and have Python 3 and the following libraries installed.

The scripts have last been tested using **Python 3.11.5** and with the following packages:

- `logging`
- `matplotlib` 3.8.0
- `numpy` 1.26.0
- `os`
- `pandas` 2.1.1
- `seaborn` 0.12.2
- `scipy` 1.11.2
- `statsmodels` 0.14.0
- `sys`
- `xlwings` 0.30.12

As libraries are updated regularly, the scripts might need to be checked and in case adapted to new versions of the libraries. For Windows users, we advise to install Python via Anaconda© `https://www.anaconda.com/download`.

It is necessary to install also the xlwings add-in for Excel, see `https://docs.xlwings.org/en/stable/installation.html`. If the xlwings add-in is correctly installed, you should see it in the Excel ribbon as shown in figure 1.1.

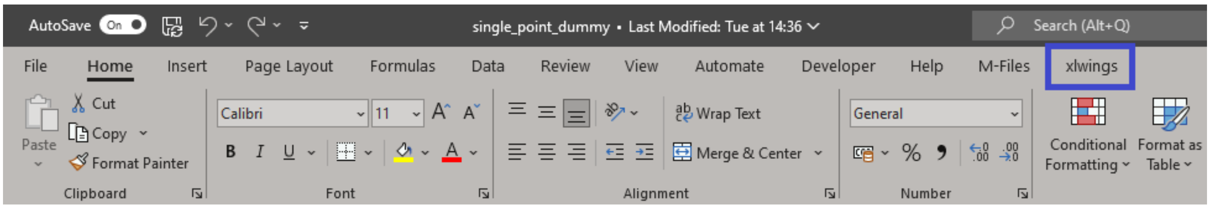
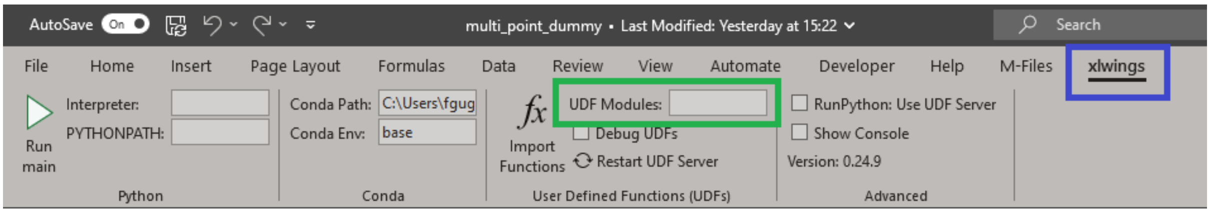


Figure 1.1: Excel ribbon with the xlwings addin correctly installed.

1.1.1 Interaction between Excel and Python

When executing the macro (that calls the Python code) from Excel, by default it looks for a `.py` file with the same name and in the same folder as the Excel file. It is possible to tell Excel to look for a `.py` file with a different name. This has to be specified in *UDF Modules* under the xlwings tab in the Excel ribbon (see figure 1.2) without the `.py` ending. For more information, see xlwings official documentation `https://docs.xlwings.org/en/stable/udfs.html`.



AutoSave On

multi_point_dummy • Last Modified: Yesterday at 15:22

Interpreter:

PYTHONPATH:

Conda Path: C:\Users\fgug

Conda Env: base

UDF Modules:

□ Debug UDFs

□ RunPython: Use UDF Server

□ Show Console

Version: 0.24.9

Figure 1.2: UDF Modules location in the Excel ribbon with the xlwings addin correctly installed.

1.2 Input from user

Both the template for the single point and for the multi point analysis include a sheet named *Input_info* with a list of inputs that the user should provide in the yellow colored cells. The column *Data type* says the type of variable the Python script is expecting when reading the data inserted by the user. Likewise, the column *Unit* says the measuring unit expected by the script associated to the input provided by the operator.

The pieces of information that the user should input are:

- *Reference concentration* (only for the single point analysis): concentration measured by the reference generator;
- *Setpoints reference* (only for the multi point analysis): set of concentrations measured by the reference generator;
- *Setpoints candidate* (only for the multi point analysis): set of concentrations measured by the candidate generator;
- *Extended uncertainty reference generator* $U(c_{ref})$ *(k=2)*: relative expanded uncertainty of the reference generator, including a coverage factor k=2;
- *Reference generator channel A*: string, contained in the `Name` column of the data set, which identifies the measurements of channel A of the reference generator;
- *Reference generator channel B*: string, contained in the `Name` column of the data set, which identifies the measurements of channel B of the reference generator;
- *Candidate generator channel A*: string, contained in the `Name` column of the data set, which identifies the measurements of channel A of the candidate generator;
- *Candidate generator channel B*: string, contained in the `Name` column of the data set, which identifies the measurements of channel B of the candidate generator;
- *Zero channel A*: string, contained in the `Name` column of the data set, which identifies the zero-concentration measurements of channel A;
- *Zero channel B*: string, contained in the `Name` column of the data set, which identifies the zero-concentration measurements of channel B;
- *Apply zero correction*: value stating if the zero correction should be applied (Apply zero correction = 1) or not (Apply zero correction = 0);
- *Reproducibility uncertainty (k=1)*: type B estimation of the relative reproducibility uncertainty provided by the user including a coverage factor k=1;

- *Maximum time between two measurements belonging to the same bracket* (only for the multi point analysis): maximum time interval elapsed between one zero-concentration measurement and the next (this is used to correctly group the zero-concentration measurements, and thus correctly estimate the zero correction);
- *DateTime format*: (Python readable) format used to record the `DateTime` column in the data set (`Y`: year, `m`:month, `d`:day, `H`:hour, `M`:minute, `S`:seconds).

In order to apply the zero correction, two measurements of the zero flux (one before and one after the measurements of non-zero mercury concentration) are necessary. We advice the user to look carefully at the data before applying any data processing analysis.

Note 1. The script does not check if the user provided data matching the described unit, that is responsibility of the user. The only unit that might be changed by the user without breaking the script is the unit of the reference and candidate concentrations (which shall match the measuring unit in which the data are expressed), since this is not used in the script but it is good to know in which measuring unit the results are. If any other change is done to the unit (e.g., the uncertainty is not reported as relative uncertainty and not with the specified coverage factor), the results given by the script are not to be relied on.

1.3 Data

Both in the template for the single point and for the multi point analysis it is assumed that the data are entered directly in the template in a sheet named *Data*. It is further assumed that the data contain the following columns:

- `DateTime`: column with time and date of the measurement.
- `Name`: column indicating whether the measurement was taken by the reference generator or by the candidate, or if the measurement corresponds to a no-flux measurement.
- `Pk Area`: column with the response of the generators.
- `Setpoint_candidate` (only for the multi point analysis): column indicating the candidate setpoint associated to the measurement.
- `Valid`: booleand flag stating if the measurement shall be used in the calculations or not.

Note 2. The data processing routines assume that the valid data to be processed follow the protocol sequence, i.e., reference meter - candidate meter - reference meter - candidate meter - reference meter - etc. It is responsibility of the user to mark as invalid (by setting as `False` the relative entry in the `Valid` column) the measurements that do not follow that scheme.

1.3.1 Manual pre-processing of the data

It is assumed that:

- There is homogeneity, including blanks, in the strings used to identify candidate and reference meters.
- If not all measurements are valid, the operator manually modifies the column *Valid* to indicate which measurements are valid ($Valid = 1$) and which are not ($Valid = 0$).
- Brackets start with valid measurements of the reference meter.

Note 3. If any correction has to be applied to the measurement data (e.g. pressure correction), the corrected data have to be provided in the data set.

1.4 Code execution

The steps to use the templates are the following.

1. Install all necessary software requirements.
2. Open the Excel file either for the single or the multi point calibration. Make sure that the xlwings' add-in is correctly installed and that macros are enabled.
3. Fill (and manually pre-process, if necessary) the measurement data in the *Data* sheet according to the requirements listed in Section 1.3.
4. Fill the required information in the sheet *Input_info*.
5. Click on the button *Perform single point analysis*, or *Perform multi point analysis* depending on which template is being used.

When clicking on the *Perform single point analysis* (*Perform multi point analysis*) button, the script creates sheets *Plot data*, *Channel A*, *Channel B* (*Plot data*, *Channel A*, *Channel B*, *Regression results*, *Calibration results*) to display the results (if they are not already present in the Excel workbook). If these sheets are already present in the Excel file, then the script overwrites the content of the sheets.

2 Data processing

2.1 Single point analysis

The single point analysis is done following *Protocol for the SI-traceable calibration of elemental mercury gas generators used in the field* [1]. For more details on the mathematical formulas, please check the protocol [1].

2.2 Multi-point analysis

In case of the multi-point analysis, the single point procedure is first applied to each point individually. Then the so-calculated concentrations and associated uncertainties (without the reference uncertainty) are given as input to the weighted least square (WLS) algorithm (where the weights are given by the inverse of the uncertainties) to determine the calibration function. Since the reference uncertainty is common to all measurements, it should be considered as a correlation term. Thus the reference uncertainty component should be added after computing the calibration function to avoid an artificial reduction in the candidate uncertainty due to the regression model.

Following the requirements of ISO 6143 [2], WLS is performed only if at least three points are present and the polynomial selection is done according to ISO TS 28038 [3], i.e. the function associated to the lowest value of the corrected Akaike Information Criteria (AICc) is selected. To allow fitting a polynomial of degree 1 using three data points (in which case it is not possible to compute the AICc due to division by zero in the second term of the AICc equation, see equation 10 in [3]), the AIC is displayed instead. Thus we recommend the user to check also other statistics, such as the p-value associated to each polynomial coefficient, to determine whether the selected polynomial is actually the most suitable. A plot of the regressed functions and related residuals is included in the output displayed in the Excel document (sheet *Regression results*). The values of the AICc as well as the sum of the squared residuals are reported for each tested polynomial. Furthermore, the polynomial coefficients, their p-values, and the covariance matrix are displayed for the optimal polynomial (sheet *Regression results*).

The optimal polynomial of channel A and of channel B are then visually compared and the measurement data of both channels are used to determine the calibration polynomial (sheet *Calibration results*). In case of optimal polynomials with different degrees, the polynomial of lower degree is selected to compute the calibration polynomial and a warning message is given to the user. The chi-squared test [4] is used to evaluate the goodness of fit of the calibration polynomial.

3 Validation

Both the single point and the multi point template have been validated. See the validation report for details and results of the validation [5].

Acknowledgements

This project 19NRM03 SI-Hg has received funding from the EMPIR programme co-financed by the Participating States and from the European Union’s Horizon 2020 research and innovation programme.

Bibliography

[1] Iris de Krom, Adriaan van der Veen, Federica Gugole, Warren Corns, Carsten Roellig, Rajamäki, and Reinhold Moesler. Protocol for the SI-traceable calibration of elemental mercury (Hg^0) gas generators used in the field, 2023.

[2] ISO 6143. Gas analysis - Comparison methods for determining and checking the composition of calibration gas mixtures. 2001.

[3] ISO TS 28038. Determination and use of polynomial calibration functions. 2018.

[4] William H. Press, Vetterling William T. Teukolsky, Saul A. and, and Brian P. Flannery. Numerical recipes in C: The art of scientific computing. Second edition. 1992.

[5] Federica Gugole. Validation report of the data analysis template for the calibration of elemental mercury gas generators.