NIST Homogeneity Assessor
User's Manual

Luke Benz, Thomas Lafarge, Antonio Possolo

Statistical Engineering Division
Information Technology Laboratory

August 1, 2018

National Institute of
Standards and Technology
U.S. Department of Commerce

Contents

Exhibits

1 Introduction

> "Homogeneity can refer either to variation of a property value between separate units of the material, or to variation within each unit. It is always necessary to assess the between-unit variation. Where the intended use permits the use of part of a unit — for example, a small portion of a solid or liquid material, or a small region of the surface — it is also usually necessary either to assess the within-unit variability of the material (within-unit heterogeneity) or to provide instructions for use that control the impact of within-unit heterogeneity. These instructions can include, for example, remixing of the sample and, for granular materials, a minimum sample size, because the within-unit heterogeneity is directly reflected in the minimum size of subsample that is representative for the whole unit."
>
> — ISO Guide 35 (ISO, 2017, §7)

A reference material produced as a batch, which is distributed and intended to be used in separately packaged portions, or *units*, (which may be ampoules, vials, bottles, etc.), is said to be homogeneous at the physical scale (range of values of mass or volume) of the units, when the between-units variability of measured values is comparable to the typical measurement uncertainty associated with the unit-specific measured values.

In this manual we refer to units as "bottles", to avoid any possible confusion with "measurement units." Bottles often are arranged in boxes, and such grouping may, but need not, reflect the order in which the bottles were filled with portions of the material.

When aliquots may be drawn from each bottle and measured separately under conditions of repeatability, then the uncertainty associated with the reference value must include a contribution corresponding to the within-bottle variability of the values of the property of interest. For this reason, homogeneity studies typically compare the between-bottles variability of measured values, with their within-bottle variability.

Most materials exhibit some level of heterogeneity, and the contribution that this makes to the uncertainty associated with the reference value must be evaluated and incorporated in the evaluation of the uncertainty that will be associated with the reference value.

When the material is found to be significantly heterogeneous, and this heterogeneity is sufficiently pronounced to render it unfit for its intended purpose,

then the batch may be subdivided into several sub-batches whose heterogeneity is acceptable, that will become separate reference materials, each with its own reference value and associated uncertainty.

With the possible exception of solutions that do not involve immiscible liquids and that experience will have shown to remain homogeneous during and after packaging and storing, the homogeneity of reference materials should be evaluated experimentally by conducting a homogeneity study. This need applies both to solid materials (blocks, chips, granules, powders, etc.) and to gas mixtures.

A homogeneity study typically involves three phases: sampling, measurement, and statistical modeling and data analysis of the results:

- In most cases, sampling is either simple random sampling or multistage (nested) random sampling. In all cases, the sampling scheme should be suitably designed to capture all recognized sources of sampling uncertainty. For example, first a sample of boxes is drawn by simple random sampling from the set of boxes containing separately packaged bottles of the reference material, next a sample of bottles is drawn from each of the selected boxes, again by simple random sampling, and finally several aliquots are drawn from each bottle after thoroughly mixing the contents, and a determination of the value of the measurand is made in each one separately. In general, let $x_{1,j}, \ldots, x_{m_j,j}$ denote the m_j replicate determinations made for bottle $j = 1, \ldots, n$. Typically, these determinations are not qualified with evaluations of associated uncertainty.

- If a single measurement result $(y_j, u(y_j))$ is obtained for each bottle, the assumption is that the uncertainty $u(y_j)$ expresses contributions from all recognized, substantively significant sources of uncertainty, not only lack of repeatability. The $\{y_j\}$ may, but need not be averages or some other summary statistic of replicate determinations made of aliquots drawn from each bottle: in some cases a single measurement is made for each bottle, and in such cases only between-bottle variability can be gauged.

- The goal of the statistical modeling and data analysis is to produce estimates of the *variance components* (Searle et al., 2006) attributable to the identified sources of variability that the design of the homogenization study allows estimating.

The *NIST Homogeneity Assessor* (NIHOMA) serves to characterize the homogeneity of a candidate reference material for a particular measurand, based either

(i) on determinations of the value of this measurand made under conditions of repeatability, in aliquots drawn from each of several bottles, or (ii) on individual measurement results for each of several bottles, each comprising a measured value and an evaluation of associated uncertainty.

In the most common case, there is only one level of nesting, say where replicate determinations are made of aliquots sampled from each bottle in a sample drawn from a lot of bottles. The number of replicated determinations may vary among bottles, and the number of bottles sampled from each box may be different for different boxes, etc.

Section 2 summarizes the steps that need to be taken to use the NIHOMA. Section 3 outlines the models and assumptions underlying the data reductions implemented in the NIHOMA, and illustrates and discusses them as they are applied to a set of measurement results for the mass fraction of vanadium in a bituminous coal (Possolo and Pintar, 2017).

After reading Sections 2 and 3, users should be ready to make informed choices to apply the NIHOMA to their own data, and to interpret the results, without further study of this manual. However, for the reader wishing to gain a more thorough appreciation for the technology implemented in the NIHOMA, Section 4 presents additional examples of application using data from the following studies, in all cases providing background information on the study, detailing the data and data reduction techniques that were used, and explaining the meaning of the results:

- *Nitrogen* (§4.1) in a synthetic mixture intended to mimic natural gas, where within-cylinder variability is appreciably larger than between-cylinder variability, hence there is no reason to question homogeneity;
- *Methane* (§4.2) in a synthetic mixture intended to mimic natural gas, which illustrates the case where the variance component attributable to differences between cylinders is significantly greater than zero and the question arises of how to handle the corresponding heterogeneity;
- *Barium* (§4.3) in soil, whose treatment demonstrates how the NIHOMA handles the case where the data comprise individual measurement results for several bottles.

The accompanying graphical representations of the results obtained in these examples, any pre-processing that the data will have had to undergo in preparation

for their use in the NIHOMA, and also alternative analyses that are presented in some cases, using methods not available in the NIHOMA, all were done using the R environment for statistical computing and graphics (R Core Team, 2018).

Section 5 emphasizes that the NIHOMA ought not be misconstrued as a toolbox capable of addressing all cases likely to be encountered in homogeneity studies. For example, the NIHOMA does not allow consideration of covariates that may have been measured alongside the measurand and that could help explain any apparent heterogeneity and account for the corresponding uncertainty components.

Section 6 summarizes technical details of the implementation and deployment of the NIHOMA either as a desktop application on a local computer, or as an application available in the World Wide Web.

2 Quick Start

2.1 Access

The NIHOMA is accessible as an R package, **NIHOMA.App**, that, once installed in a local computer, enables access via a web browser. It is currently in the process of being made accessible on the web at homogeneity.nist.gov,

Clicking on About the NIST Homogeneity Assessor brings up a page containing general information and guidance about the application. After providing the application with inputs on the Enter Data page, as outlined in §2.2, the user should validate the inputs by clicking on the Validate Inputs button at the bottom of the page. Assuming all inputs are valid, the message Model Inputs are Valid! will appear. Otherwise, a red error message will be displayed, detailing why the inputs are invalid. Upon verifying that that inputs to the application are valid, the user simply clicks on the Fit Random Effects Model button to begin the analysis, which will automatically render in a new tab titled, Analysis Results.

Additional buttons at the bottom of the Enter Data page, shown in Exhibit 2, allow the user to load and save data files. Clicking the button labeled Save Data File will parse user inputs and download a data file in `.csv` format. Users may upload previously downloaded data files or may upload data files of their own creation in `.csv`, `.txt`, `.xls`, or `.xlsx` format using the Browse button. Upon uploading a data file, the user is asked to validate the inputs to ensure proper parsing of user data. Users wishing to perform several analyses with different

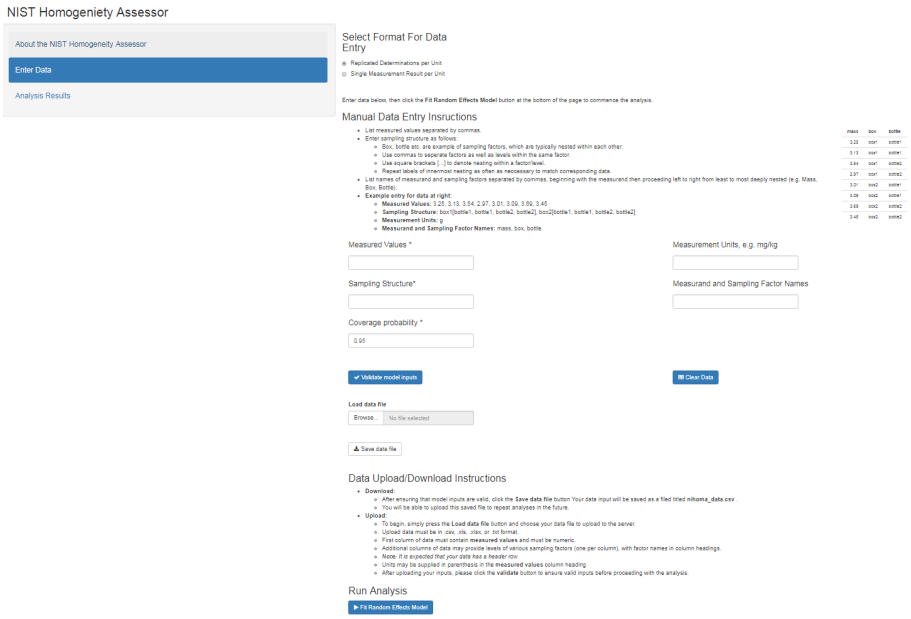

Exhibit 1: User interface for the NIHOMA application

data sets should to use the Clear Data button to delete all previously entered data and reset all parameters.

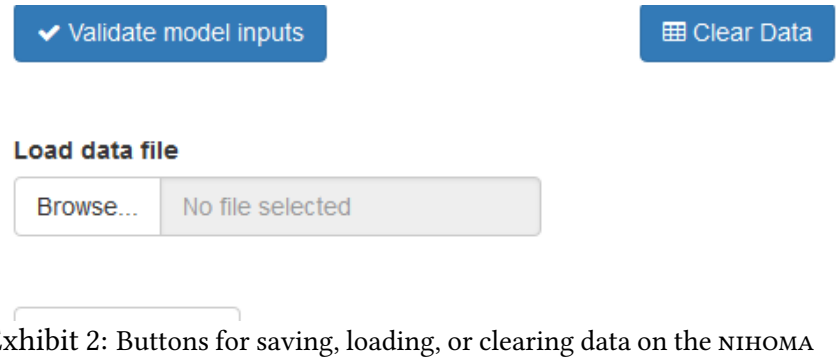

Exhibit 2: Buttons for saving, loading, or clearing data on the NIHOMA Enter Data page.

2.2 General Inputs

Users are first asked to select the format for data entry from among the Replicated Determinations per Unit and Single Measurement Result per Unit options at the top of the page. The default option, Replicated Determinations per Unit, can be used only when the user has several individual determinations available for each bottle. Should the user only possess summary statistics for each bottle, for example a single measured value and associated standard uncertainty, then the user should select the data format option Single Measurement Result per Unit.

Even when individual determinations are available that were made on aliquots drawn from each bottle, the user may opt for the second option because this affords the opportunity to recognize and express (in the input for associated uncertainties) uncertainty components other than just the dispersion of the individual determinations made in aliquots drawn from each bottle.

2.2.1 Replicated Determinations per Unit

General inputs for the Replicated Determinations per Unit format are as follows:

- **Measured Values** (REQUIRED) Magnitudes (numerical values without measurement units) $x_{1,j}, \ldots, x_{m_j,j}$ corresponding to m_j replicates for group j (typically a group of determinations made for bottle $j = 1, \ldots, n$) defined by the sampling factors. Values must be entered separated by commas, and may be written in scientific notation as in 3.52e1 or 352e−1, both meaning 35.2.
- **Sampling Structure** (REQUIRED) A character string providing information on the sampling structure of the measured values. The sampling structure should be specified as follows:
  - Box, bottle etc. are examples of sampling factors, which are typically nested within each other.
  - Square brackets "[...]" are used to denote nesting within a factor/level.
  - Commas separate factors, as well as levels within the same factor.
  - Repeat levels of most deeply nested factor as often as necessary to match the corresponding measured values.

- **Measurement Units** (OPTIONAL) A character string specifying measurement units for the measurand, which will be used in plots produced by the NIHOMA. In addition to plain ASCII character text, units in the form of symbols or Greek letters will be accepted provided that they adhere to the principles governing mathematical annotation in R.

- **Measurand and Sampling Factor Names** (OPTIONAL): A list of names, separated by commas, for the measurand and sampling factors. Names should be ordered from left to right, beginning with the measurand then proceeding left to right from least to most deeply nested factors. If this field is left blank, default names of **measured_value**, **sf_1**, ..., **sf_j** will be used for the measurand and sampling factors 1, ..., j respectively.

- **Coverage Probability** (REQUIRED) A number p between 0 and 1 denoting the probability with which the intervals given for the measurand, and for the variance components, are believed to cover their targets. This user-specified coverage probability, p, also determines the significance level $\alpha = 1 - p$ that is used in hypothesis testing throughout the NIHOMA. Default is 0.95.

2.2.2 Single Measurement Result per Unit

General inputs for the Single Measurement Result per Unit format as as follows:

- **Measured Value for Each Bottle** (REQUIRED) $y_1, \ldots, y_n$ representing estimates of the measurand for each of the n levels of a single sampling factor. Values must be entered separated by commas, and may be written in scientific notation: for example, 35.2, 3.52e1, and 352e−1, all denote the same value.

- **Standard Uncertainty Associated with Each Measured Value** (REQUIRED) $u_1, \ldots, u_n$ representing standard uncertainties associated with the measured values $y_1, \ldots, y_n$ for each of the n levels for a single sampling factor. Values must be entered separated by commas, and may be written in scientific notation: for example, 0.52, 5.2e−1, and 52e−2 all denote the same value.

  NOTE: u_1, ..., u_n are standard uncertainties (not variances), and hence have the same measurement units as the measured values. In the case where the

Mass Fraction	Box	Bottle
3.52	box1	bottle1
3.51	box1	bottle1
3.38	box1	bottle2
3.47	box1	bottle2
3.45	box2	bottle1
3.56	box2	bottle1
3.88	box2	bottle2
3.55	box2	bottle2

Exhibit 3: For a sampling structure with 2 boxes and 2 bottles sampled from each box, as shown above, the input data would be specified as follows:

- **Measured Values:** 3.52, 3.51, 3.38, 3.47, 3.45, 3.56, 3.88, 3.55
- **Sampling Structure:**
  box1[bottle1, bottle1, bottle2, bottle2], box2[bottle1, bottle1, bottle2, bottle2]
- **Units**: g/kg
- **Measurand and Sampling Factor Names:** Mass Fraction, Box, Bottle
- **Coverage Probability:** 0.95

$\{y_j\}$ are averages of determinations obtained under conditions of repeatability, then $u_j = s_j/\sqrt{n_j}$ (Type A evaluation as in Equation (5) of the GUM) is based on $\nu_j = n_j - 1$ degrees of freedom, where s_j denotes the sample standard deviation of the n_j replicates.

- **Sampling Structure** (REQUIRED) A character string providing information on the sampling structure of the measurement values/uncertainties. The sampling structure is expected in the following form.
  - Commas separate levels within the same sampling factor. Only a single sampling factor is accepted in the Single Measurement Result per Unit.
  - Factor levels match corresponding measurement results and uncertainties.
- **Degrees of Freedom** (OPTIONAL) $\nu_1, ..., \nu_k$ representing the degrees of free-

dom on which the standard uncertainties $u_1, \ldots, u_n$ for each of the n sampling factor levels are based. If supplied, degrees of freedom should be entered separated by commas, and must be non-negative numbers.

- **Measurement Units** (OPTIONAL) A character string providing units that qualify the numerical values of the measurement results and standard uncertainties mentioned above. If specified they will be used in labels of axes of plots produced by the App. In addition to standard character text, units in the form of symbols or Greek letters will be accepted provided that they adhere to the principles governing mathematical annotation in R.

- **Column Names** (OPTIONAL): A list of names, separated by commas, for the measurand, uncertainty, degrees of freedom (if supplied) and sampling factor. Names should be ordered from left to right beginning with the name of the measurand, uncertainty, and degrees of freedom, followed by the names of the sampling factor. If this field is left blank, default names of **measured_value**, **uncertainty**, **df**, **sf_1**, will be used for the measurand, uncertainty, degrees of freedom and sampling factor respectively. Note that the user should only supply a column name for degrees of freedom if degrees of freedom $v_1, ..., v_k$ were also supplied. If no degrees of freedom were supplied, the user may enter names for the measurand, uncertainty, followed by the names of the sampling factors, from left to right. In the absence of degrees of freedom, default column names will be **measured_value**, **uncertainty**, **sf_1**.

- **Coverage Probability** (REQUIRED) A number between 0 and 1 used to determine the probability with which the intervals given for the measurand, and for the variance components, cover their targets. The user specified coverage probability, p, also determines the significance level $\alpha = 1 - p$ that is used in hypothesis testing throughout the NIHOMA. Default is 0.95.

2.2.3 Uploading and Downloading Data

UPLOAD: In addition to manual data entry, users may also upload data in `.txt`, `.csv`, `.xls`, or `.xlsx` format. The first column in the uploaded data file must contain values of the measurand (either replicated determinations or or single measurement results depending on the option selected for data entry format),

Mass Fraction	Uncertainty	df	Bottle
3.52	0.11	3	bottle1
3.51	0.18	3	bottle2
3.38	0.31	2	bottle3
3.47	0.20	3	bottle4

Exhibit 4: Suppose we have a sampling structure with averages of determinations made for each of 4 bottles bottles as given above. Data entry would look as follows:

- **Measured Values:** 3.52, 3.51, 3.38, 3.47
- **Standard Uncertainties:** 0.11, 0.18, 0.31, 0.20
- **Degrees of Freedom:** 3, 3, 2, 3
- **Sampling Structure:** bottle1, bottle2, bottle3, bottle4
- **Units**: g/kg
- **Column Names:** Mass Fraction, Uncertainty, df, Bottle
- **Coverage Probability:** 0.95

and must be numeric. If the Single Measurement Result per Unit option is selected for Select Format For Data Entry, then the second column of data must contain bottle uncertainties and must be numeric. A third column containing degrees of freedom may be added, but is not required. If degrees of freedom are included, they must be numeric and non-negative. Any additional columns after the first (Replicated Determinations per Unit) or second/third (for Single Measurement Result per Unit) may contain the names and levels of sampling factors, left to right from least to most nested. As a technical note, it is assumed that any data file has a header row containing column names. Headings **MUST** be supplied on any uploaded file. Users uploading data may supply units, in parenthesis, in the heading for the measured values column. An example might look very similar to Exhibits 3 and 4, with "Mass Fraction" replaced by "Mass Fraction (g/kg)". Users are asked to Validate their inputs after upload before proceeding with any further analysis.

DOWNLOAD: Users may download any data entered into NIHOMA by simply clicking the Save data file button. Data will saved in .csv format as a file entitled **nihoma_data.csv** or **nihoma_data_case2.csv** depending on the structure of the user input.

2.2.4 Advisory

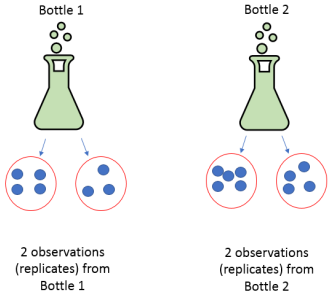

Exhibit 5: NIHOMA interpretation of bottle1, bottle1, bottle2, bottle2 sampling structure.

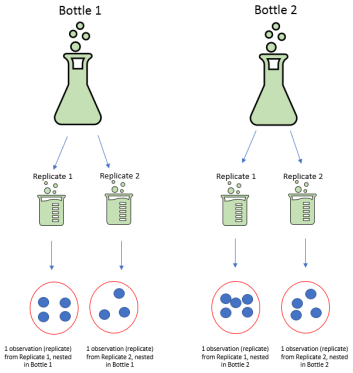

Exhibit 6: NIHOMA interpretation of bottle1[replicate1, replicate2], bottle2[replicate1, replicate2] sampling structure.

Here we offer some clarifications that should help the user avoid any confusion. Suppose we are working with a study design where two replicate determinantions were made for each of two bottles. The following two entries to the sampling structure field ***are not*** parsed equivalently:

- bottle1, bottle1, bottle2, bottle2
- bottle1[replicate1, replicate2], bottle2[replicate1, replicate2]

Exhibits 5 and 6 show how the NIHOMA would interpret each of these two sampling structures. The latter treats replicate as a factor nested within bottle, and will not produce the analysis matching the design of the sampling structure. That is, the NIHOMA would interpret bottle1[replicate1, replicate2], bottle2[replicate1, replicate2] as a sampling structure no different than a sampling structure where

2 bottles were nested within boxes, with a single measurement available for each box-bottle pair. For correct parsing of the desired sampling structure, the user Simply needs to match the factor nesting structure to the individual determinations themselves (entered in the Measured Values field) rather than trying to treat replicate as a sampling factor

A second point of clarification comes on the distinction between the potential download files **nihoma_data.csv** and **nihoma_data_case2.csv**. The naming convention simply reflects which data format the user was working with, to ease in re-upload if the user desires to continue with the analysis in the future. One should be careful to ensure that regardless of the name of their data file, the selected format for data entry does indeed match the format of the data uploaded. Failure to do so would result in the treatment of standard uncertainties and degrees of freedom as levels of nested sampling factors or would interpret levels of a sampling factor as standard uncertainties.

2.3 Further Inputs

No further inputs are required after clicking the Fit Random Effects Model to commence the analysis. If desired, however, the user may supply which sampling factors are to be used in rendering the **Group Subset Plot** as shown in Exhibit 8 and described in more detail in §2.4.1.

2.4 Outputs

The results appear on a refreshed web page under the Analysis Results section. Output differs slightly depending on the format of the data uploaded by the user.

2.4.1 Replicated Determinations per Unit

- **Random Effects Model Summary:** The random effects summary table contains estimates and $(100 \times p)\,\%$ coverage intervals for the consensus value of the measurand as well as within group standard uncertainties and between group standard deviation on the basis of each sampling factor (where p denotes the user entered coverage probability on the Enter Data page).

- **Kruskal-Wallis Procedure Summary:** A summary table for the Kruskal-Wallis procedure contains χ^2 test statistics and p–values for testing whether differences exist between group distributions. Differences are first tested along outermost factor, and then again at each inner level of nesting. The theory underlying the random effects model is reviewed in §3.1.3.

- **Levene's Test Output:** A summary for Levene's Test contains F test statistics and p–values for testing whether differences exist between group variances. Differences are first tested along outermost factor, and then again at each inner level of nesting. The theory underlying Levene's Test is reviewed in §3.1.2.

- **Normal QQ-Plots:** A normal QQ-plot compares theoretical quantiles from a normal distribution with observed quantiles from an input data set. NIHOMA renders normal QQ-plots for each of the random effects and model residuals. If all points fall inside of their respective light-blue bands, there is no reason to reject the assumption of normality for the effect/residuals. If the QQ-plots suggest otherwise, then the linear, Gaussian random effects model may not be an appropriate procedure for assessing homogeneity. The theory underlying the random effects model is reviewed in §3.1.1. Exhibit 7 shows an example of normal QQ-plots for random effects.

- **Group Subset Plots:** Group subset plots display distributions of the measurand broken down by groups created from levels of up to two sampling factors. The default plot places determinations of the measurand on the Y-axis, the first listed sampling factor on the X-axis, and creates separate plotting windows for each level of the second listed sampling factor, if applicable. The user can choose which sampling factor to display on the X-axis with the Primary Sampling Factor drop-down menu. Additionally, the user may choose which sampling factor to create separate plotting windows for by using the Secondary Sampling Factor drop down menu. If the user does not wish to create separate plots on the basis of a secondary sampling factor, they may disable this feature by selecting the Ignore — Single Variable Plot Only option from the Secondary Sampling Factor drop down menu. The type of plot showing group distributions depends on the minimum number of replicates per group, m. If $m \leq 5$, group distributions are rendered as categorical scatter-plots. If $5 < m \leq 30$, group distributions are rendered as box-plots. In the case that $m > 30$, group distributions are

rendered as ridgeline density plots. An example of a group subset plot is shown in shown in Exhibit 8.

All summary tables can be downloaded in `.txt` or `.tex` format, and all plots can be downloaded as `.pdf` files, by clicking the appropriate save output option.

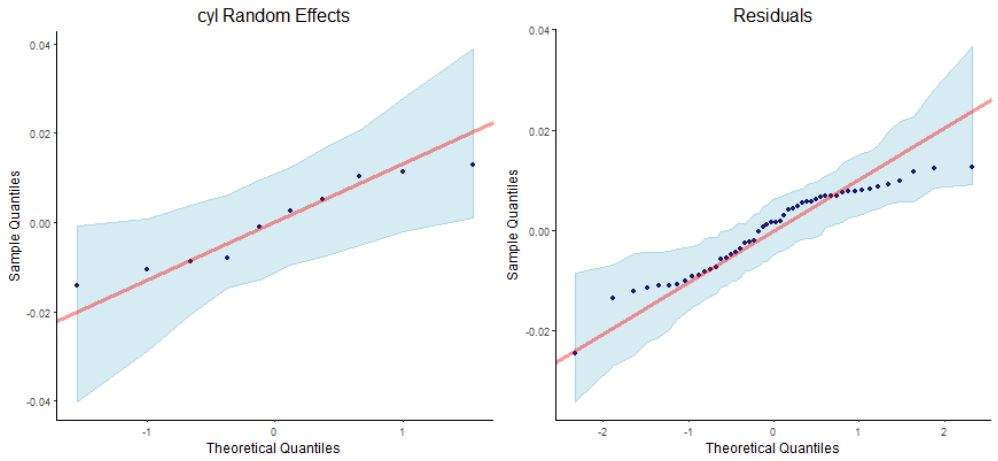

Exhibit 7: Example of normal QQ-plot for random effects output. Neither the random effects due to the **cyl** sampling factor or model residuals lie outside of the light-blue bands, and as such, their is no reason to reject the assumption of normality for either distribution.

Technical Note: The algorithm used to fit the random effects model may not converge. When this happens, a red error message will displayed. A **Convert to Single Measurement Result per Unit** button will pop-up, allowing the user to download their data in the form required for the Single Measurement Result per Unit case, which can then be uploaded to the NIHOMA to try and fit the random effects model using sufficient statistics route. In this case, the following steps are taken to convert data from the Replicated Determinations per Unit format to the Single Measurement Result per Unit format.

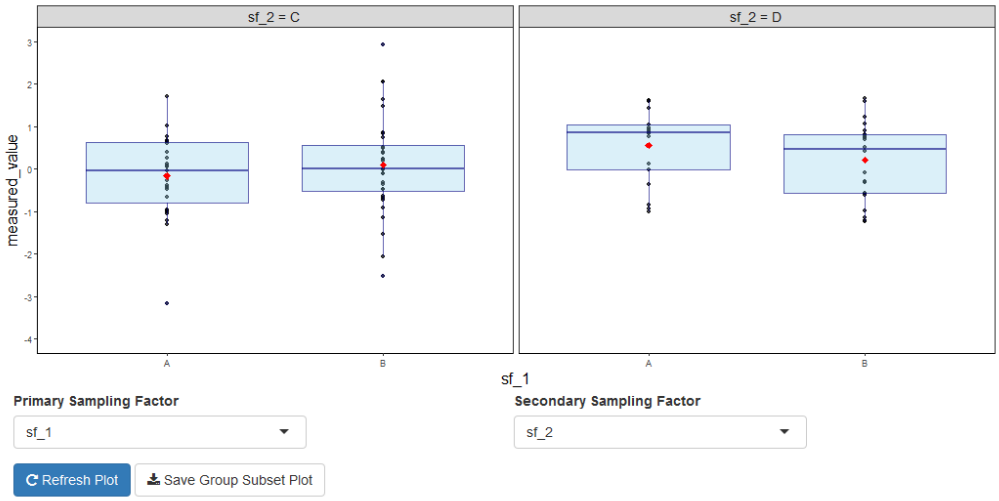

Exhibit 8: Example of a group subset plot. We see determinations of the measurand, **measured_value** is on the Y-Axis while the primary sampling factor, **sf_1** is displayed on the X-Axis. Plots are subset by levels of a secondary sampling factor, **sf_2**. Here $m = 17$, so group distributions are rendered as boxplots. Individual values of the measurand are displayed in black, with group means represented by red diamonds. The user may change the orientation of the plot by using the Primary Sampling Factor and Secondary Sampling Factor drop down menus, then simply clicking the Refresh Plot button to render the new plot.

Measured Values:

$$y_j = \overline{x}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_{i,j}$$

Uncertainties:

$$u_j = \frac{s_j}{\sqrt{m_j}} = \frac{1}{\sqrt{m_j}} \times \sqrt{\frac{1}{m_j - 1} \sum_{i=1}^{m_j} (x_{i,j} - \overline{x}_j)^2}$$

Degrees of Freedom:

$$\nu_j = n_j - 1$$

2.4.2 Single Measurement Result per Unit

- **Random Effects Model Summary:** The random effects summary table contains an estimate and $(100 \times p)\,\%$ coverage interval for the consensus value of the measurand as well as an estimate of σ_b, the between-bottle standard uncertainty (where p denotes the user entered coverage probability on the Enter Data page). The model summary also includes the results of Cochran's Q Test for Heterogeneity.
- **Normal QQ-Plots:** A normal QQ-plot compares theoretical quantiles from a normal distribution with observed quantiles from an input set of data. The NIHOMA renders normal QQ-plots for each of the random effects and model residuals. If all points fall inside of their respective light-blue bands, there is no evidence to reject the assumption of normality for the effect/residuals.
- **Summary Plot:** The single measurement results per unit are plotted on the Y-axis against levels of the sampling factor, displayed on the X-axis. Single measurement results made in each bottle $y_1, ..., y_n$ are marked with red diamonds. Measurement uncertainties are displayed as purple line segments with range $y_j \pm u_j$ for $j = 1, \ldots, n$. Finally, a rectangular area shaded light blue centered around a dark blue line denotes $\hat{\mu} \pm u(\hat{\mu})$, a consensus value and a $(100 \times p)\,\%$ coverage interval for the true value of the measurand.

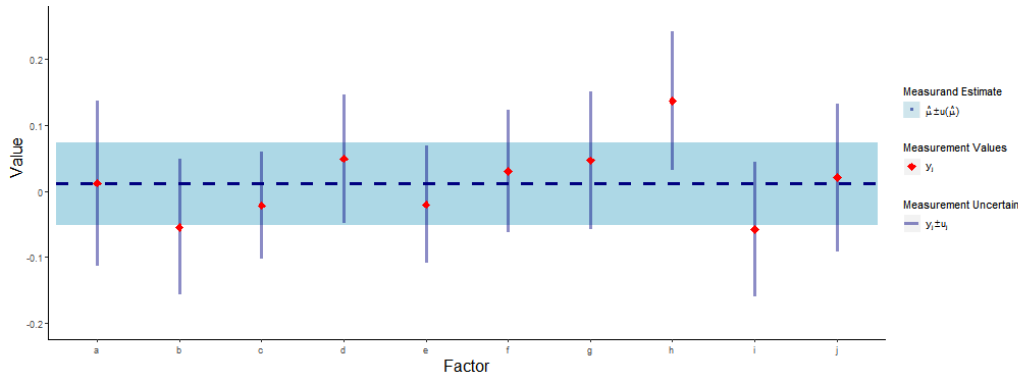

Exhibit 9: Example of a summary plot, part of the NIHOMA output in the Single Measurement Result per Unit case

3 Orientation

3.1 Models & Assumptions

The NIHOMA relies on a linear, Gaussian random effects model (Searle et al., 2006) to estimate and evaluate the significance of variance components that express heterogeneity. Once a variance has been found to be significant, that is attributable to variability at a particular level of sampling, the scientist responsible for the production of the reference material will have to decide whether the apparent heterogeneity is sufficiently small to be acceptable. If it is small enough for the purpose that the material is intended to serve, then it will be incorporated in the uncertainty associated with the reference value. If it is much too large, then the material may be subdivided into two or more subsets that will be certified separately.

A key assumption of the random effects model, as described in §3.1.1, is that factor levels, which in this case may be bottles or groups of bottles boxed together, are representative samples drawn randomly from the population of bottles or boxes containing aliquots of the material. This is ensured by a suitable sampling design.

3.1.1 Linear, Gaussian Random Effects Model

In this section, we review the theory underlying linear, Gaussian random effects models when the material is sampled according to a nested factor design. For the sake of simplicity, we consider the case where there are two sampling factors, one nested within the other, for example bottles in boxes, and where multiple determinations will be made of the contents of each bottle selected for examination.

The factor A denotes the box, and the factor B denotes the bottle, with B nested within A. Suppose that there are I boxes, labeled $a_1, \ldots, a_I$, and J_i bottles $b_1, \ldots, b_{J_i}$ in box a_i, of which n_{ij} will be selected for the homogeneity study, with $1 \leqslant i \leqslant I$ and $1 \leqslant j \leq J_i$. If K determinations are made of the contents of bottle b_j sampled from box a_i, then the random effects model treats the corresponding measured values as outcomes of random variables represented as additive superpositions of four effects:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}.$$

Y_{ijk} The kth determination of the measurand in bottle b_j selected from box a_i.

μ The true value of the measurand in the batch of material that has been subdivided into portions that were bottled and boxed separately.

α_i The effect of box a_i. When different boxes comprise bottles that were filled on different days, some boxes may have average values of the measurand higher than μ, while other may have averages lower than μ. Accordingly, we model $\alpha_1, ..., \alpha_I$ as a sample from a Gaussian distribution with mean 0 and standard deviation σ_{A}.

$\beta_{j(i)}$ The random effect of bottle b_j drawn from box a_i. Since some bottles may have a value of the measurand higher than the corresponding box mean $\mu + \alpha_i$, while others may have it lower than it, we model the bottle effects $\beta_{1(i)}, ..., \beta_{J_i(i)}$ as a sample from a Gaussian distribution with mean 0 and standard deviation σ_{B}.

ϵ_{ijk} Measurement error in the kth determination made for bottle b_j in box a_i. These errors are modeled as a sample from a Gaussian distribution with mean 0 and standard deviation σ.

Note that σ_{B} is assumed to be ***the same*** for all bottles (regardless of the box they were selected from), and that σ also is assumed to be ***the same*** for all determinations (regardless of the bottle they pertain to). We can use Levene Test's, reviewed in §3.1.2, to test the validity of these assumptions. In these circumstances, the individual determinations are like outcomes of independent, identically distributed random variables $\{Y_{ijk}\}$ with common mean μ and variance $\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{B}}^2 + \sigma^2$, and for this reason, σ_{A}^2, σ_{B}^2 and σ^2 are called *variance components*.

This specific example with two nested sampling factors may be generalized to study designs with further levels of nesting, hence with additional effects and corresponding variance components. In the most common case, there is only a single sampling factor B, representing bottles, say. In this most simple study design, the model reduces to $Y_{jk} = \mu + \beta_j + \epsilon_{jk}$.

3.1.2 Levene's Test

Levene's test serves to evaluate the model assumption that the within-group variance, σ^2, is equal for all sampling groups. This assumption is often referred to as homogeneity of variance or *homoscedasticity*.

In the NIHOMA, differences are first tested along outermost factor, and then again at each deeper level of nesting. Suppose we have a sampling design with 2 boxes, with 6 bottles drawn from each box, and 5 replicate determinations made of the contents of each bottle. We first test whether there are significant between-box differences in the dispersions of values of the measurand. Then, we proceed to test whether there are significant between-bottle-box differences in dispersion.

Levene's test statistic is defined as

$$W = \frac{(N - G)}{(G - 1)} \frac{\sum_{i=1}^{G} n_i (z_{i\cdot} - z_{\cdot\cdot})^2}{\sum_{i=1}^{G} \sum_{j=1}^{n_i} (z_{ij} - z_{i\cdot})^2} \ :$$

G — The number of sampling groups;

N — The total number of replicates across all sampling groups;

n_i — The number of replicates in the ith sampling group;

z_{ij} — The absolute value of the difference, $|Y_{ij} - Y_{i\cdot}|$, between the jth determination of the measurand in the ith sampling group and the average determination of the measurand in the ith sampling group;

$z_{i\cdot}$ — The average value of z_{ij} for the ith sampling group;

$z_{\cdot\cdot}$ — The average value of z_{ij} across all N replicates.

When variances are homogeneous, W is approximately distributed as $F_{G-1,N-G}$ which can be used as a reference distribution to obtain approximate p–values for significance tests.

3.1.3 Kruskal-Wallis

The Kruskal-Wallis procedure is a non-parametric rank-based test for testing whether differences exist between-group distributions corresponding to levels of a single grouping factor. In the cases that Levene's Test yields statistically significant differences between group dispersions, or when the effects of the sampling factors, or the measurement errors, are not consistent with the assumption that they are samples from Gaussian distributions, the Kruskal-Wallis procedure is preferable to the random effects model described in §3.1.1 for identifying the existence of heterogeneity.

In the NIHOMA, differences are first tested for the outermost factor, and then again at each successively deeper level of nesting. Suppose we have a sampling design with 2 boxes, with 6 bottles nested within each box, and 5 replicates taken per bottle. We first test whether there are significant differences between the 2 boxes' distributions of determinations of the measurand. Then, we proceed to test whether there are significant differences between the 12 box-bottle distributions of determinations of the measurand. The test statistics is given by

$$X^2_{kw} = (N-1)\frac{\sum_{i=1}^{G} n_i(\overline{r}_{i\cdot} - \overline{r})^2}{\sum_{i=1}^{G}\sum_{j=1}^{n_i}(r_{ij} - \overline{r})^2}$$

G — The number of sampling groups;

N — The total number of replicates across all sampling groups

n_i — The number of replicate determations made in the ith sampling group;

r_{ij} — The rank, among all replicates, of the jth replicate from the ith sampling group;

$\overline{r}_{i\cdot}$ — The average replicate rank (among all determinations of the measurand) from the ith sampling group, equal to $\sum_{j=1}^{n_i} r_{ij}/n_i$;

$\overline{r}$ — The median rank among all replicates, that is $(N+1)/2$.

While asymptotically the probability distribution of X^2_{kw} is approximated by χ^2_{G-1}, it can be quite different in cases where the number of replicates per group are small. Hence, we used a bootstrapped version of the probability distribution of X^2_{kw} when computing the p–value associated with the test.

It should be noted that the Kruskal-Wallis procedure is not a perfect alternative to the random effects model in the case that the assumptions of the random effects model are not met. The Kruskal-Wallis procedure tests for differences in distribution, and there are many ways that group distributions can differ from one another that do not involve differences between their means.

3.2 Illustration and Overview of Methods

In this section, we walk through an example demonstrating the use of the NIHOMA and how one would interpret the results that it produces. The example

worked in this section is of the simplest, and most common variety, where multiple determinations of the value of the measurand are available for levels of a single sampling factor. Additional examples, including those covering the Single Measurement Result per Unit are given in §4.

3.2.1 Data

In this example, we will use the NIHOMA to assess homogeneity of the mass fraction of vanadium in a bituminous coal (NIST SRM 2684c). The results of two determinations are given for 7 bottles, plus an additional single replicate available for an eighth bottle (Possolo and Pintar, 2017). The data used in this example is presented in Exhibit 10.

3.2.2 Results and Analysis

The NIHOMA's first output is the random effects model summary, presented below in Exhibit 11. The random effects model summary provides estimates and 95 % coverage intervals for the the **Bottle** uncertainty component, σ_B, the **within-bottle** uncertainty, σ, and the consensus value of the **measurand**, μ. Note that in the App, these three terms would appear as `Sigma_Bottle`, `Sigma_Within_Group`, and `Measurand`, respectively, due to limitations on rendering symbols in R model outputs.

We see that the 95 % coverage interval for σ_B contains 0, meaning that there is no evidence to suggest σ_B differs significantly from 0 at the 0.05 significance level. One might be tempted to immediately conclude there is no significant heterogeneity, but such a conclusion should not be reached without some further analysis. Levene's Test yields p–value $< 2.2 \times 10^{-16}$, suggesting that dispersions differ by bottle, and that the assumption of homogeneous variance is violated.

These results need to be taken with a grain of salt, as there are only two replicates per bottle, which is the bare minimum whereon to evaluate the dispersion of the determinantions per bottle. At the very least we should turn to the Kruskal-Wallis procedure to see if there are significant differences between the bottle-specific distributions of determinations. The resulting X^2_{kw} produces p–value of 0.13, suggesting that there are no such differences are significant. Notice that this is the same conclusion one would have drawn when examining the random effects model summary even though the validity of the model assumptions is questionable.

Mass Fraction (mg/kg)	Bottle
15.93	B1
16.09	B1
15.90	B2
16.44	B2
16.16	B3
16.18	B3
17.14	B4
16.55	B4
16.01	B5
15.57	B5
16.77	B6
16.56	B6
15.70	B7
15.98	B7
16.89	B8

Exhibit 10: Determinations of the mass fraction of vanadium bituminous coal from eight bottles. A table in this format could be uploaded to the Enter Data page. In the case of manual data entry, the user would enter the mass fractions, separated by commas, in the **Measured Values** field, and in the **Sampling Structure** field: B1, B1, B2, B2, B3, B3, B4, B4, B5, B5, B6, B6, B7, B7, B8.

term	group	estimate	2.5 %	97.5 %	std.uncertainity
σ_{B}	Bottle	0.3848	0.0000	0.6300	—
σ	Residual	0.2672	0.1263	0.4024	—
μ	Fixed Effects	16.29	15.90	16.62	0.1531

Exhibit 11: NIHOMA random effects model summary for the mass fraction in bituminous coal.

The NIHOMA normal QQ-plots produced for the bottle random effects, $\alpha_{B1}, ..., \alpha_{B8}$, as well as for the residuals $\{\epsilon_{ij}\}$ are show below in Exhibit 12. Given that both sets of points are well within their respective blue bands, there is no reason to reject the normality assumptions of the random effects model. It seems as though the results of Levene's test could be due to such a small number of replicates per bottle, and there is support for homogeneity across the eight bottles.

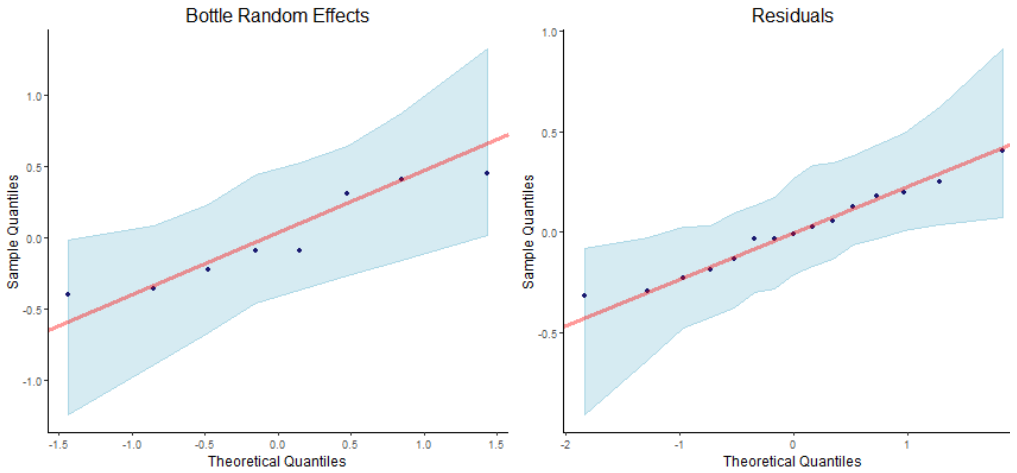

Exhibit 12: Decorated normal QQ-plots for the Bottle random effects and model residuals

The purpose of the NIHOMA is not simply to offer scientists a binary determination on the existence of heterogeneity in their data, but rather to help them evaluate how much any heterogeneity may contribute to the uncertainty of the result, and to understand the underlying distribution of their data to identify and size sources of heterogeneity when it is present.

Even though this example suggests the absence of any significant heterogeneity, how much additional uncertainty one should still add to the uncertainty budget on account of apparent heterogeneity is a task left to the scientist. Given the results suggested by the NIHOMA, a scientist would have ample reason not to add additional uncertainty due to heterogeneity to their uncertainty budget. However, the more conservative scientist may choose to take $\widehat{\sigma}_b$ as an appropriate estimate of the amount of additional uncertainty that should be factored into the uncertainty budget.

This vanadium example is illuminating, as it demonstrates the proper sequence of analysis and considerations that should be taken into account when making assessments of heterogeneity for candidate reference materials. As a final output in this example, the NIHOMA renders the group subset plot displayed in Exhibit 13.

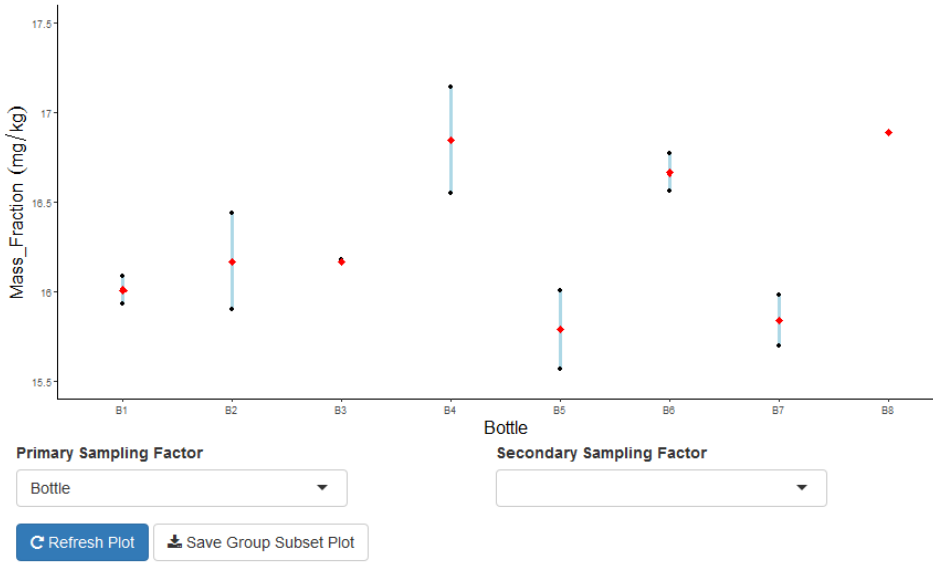

Exhibit 13: Group Subset Plot showing distributions of determinations of the value of the measurand by bottle.

4 Examples

This section provides several further examples to demonstrate a wider range of scenarios that may arise in homogeneity studies. §4.1 presents a case similar to the one presented in §3.2, where no statistically significant heterogeneity is present. §4.2 presents a study in which significant heterogeneity is present, and offers suggestions on how to proceed in the presence of heterogeneity. §4.3 demonstrates how to use NIHOMA when data is entered in the Single Measurement Result per Unit format.

4.1 Nitrogen

The data comes from the Dutch National Metrology Institute (VSL), and consists of measured values of the amount fractions of nitrogen in a synthetic reference mixture designed to mimic natural gas (Beelen, 2016; van der Veen , 2017). Five determinantions of the value of the measurand are available from each of 10 separate gas cylinders, depicted in Exhibit 14. The NIHOMA random effects model summary is show in Exhibit 15.

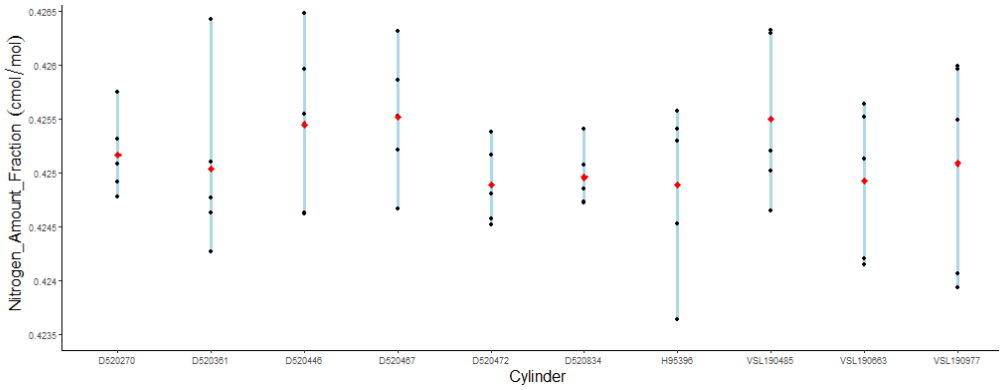

Exhibit 14: Group Subset Plot showing distributions of determinations of the value of the measurand by cylinder. Cylinder means are denoted by red diamonds

term	group	estimate	2.5 %	97.5 %	std.uncertainty
σ_C	Cylinder	7.558×10^{-13}	6.075×10^{-13}	8.994×10^{-13}	—
σ	Residual	6.782×10^{-4}	5.451×10^{-4}	8.070×10^{-4}	—
μ	Fixed Effects	0.4251	0.4250	0.4253	9.591×10^{-5}

Exhibit 15: NIHOMA random effects model summary for the amount fraction of nitrogen in a synthetic gas mixture simulating natural gas.

We see estimates and 95 % coverage intervals for the three parameters of interest, σ_C, σ, and μ in the table below. The parameter most of interest in our assessment of homogeneity is of course, σ_C, the between cylinder standard uncertainty. While the left most end of the 95 % coverage interval for σ_C is not 0, NIHOMA indicates that σ_C doesn't differ from 0 significantly, at least practically. More on

how such a determination of significance is made is presented in §6. This is readily seen however, by simply comparing $\widehat{\sigma}_C$ with $\widehat{\sigma}$, the within-group standard uncertainty. Namely, $\widehat{\sigma}_C$ is roughly 9 orders of magnitude smaller than $\widehat{\sigma}$, so within-cylinder variation dominates between-cylinder variation, and as such, it seems that there is no reason to reject the assumption of homogeneity. Before we conclude any assessment of homogeneity, we must carefully check that the assumptions made in fitting the random effects model are valid.

Levene's Test yields a p–value of 0.6585, hence there is no reason to suggest cylinder variances are heteroscedastic. Furthermore, the normal QQ-plots shown in Exhibit 16 support the notion that the cylinder random effects, $\{\alpha_i\}$, and the measurement errors $\{\epsilon_{ik}\}$, are like samples from Gaussian distributions.

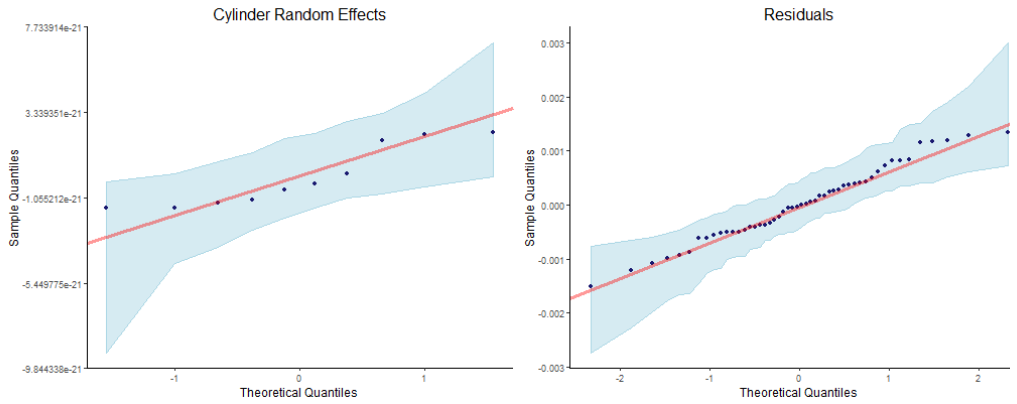

Exhibit 16: Normal QQ-Plots for cylinder random effects and model residuals. Since neither plot has any points falling outside their blue coverage bands, the assumptions of the random effects model, presented in §3.1.1 are not violated.

In this case, the variance component attributable to differences between cylinders does not differ significantly from 0 and its magnitude is negligible. In general, when the between-groups variance component still does not differ significantly from 0, but its apparent magnitude is not negligible, it is up to the scientist responsible for the production of the material to decided whether the corresponding uncertainty contribution should or should not be recognized in the combined uncertainty for the value assigned to the material.

Here, and indeed in the majority of homogeneity assessments, statistical significance is of secondary importance to the practical significance of the size of any

estimated heterogeneity for the purpose that the reference material is intended to serve. Therefore, the output of the NIHOMA serves as guidance and suggestions helping scientists to make informed decisions — not to make decisions for them.

4.2 Methane

The data in this subsection comes from the same VSL study (Beelen, 2016; van der Veen , 2017) used as an example in §4.1, and consists of measured values of the amount fractions of methane in a synthetic reference mixture designed to mimic natural gas. Five determinantions of the value of the measurand are available from each of 10 separate gas cylinders, depicted in Exhibit 17. Exhibit 18 shows the NIHOMA random effects model summary.

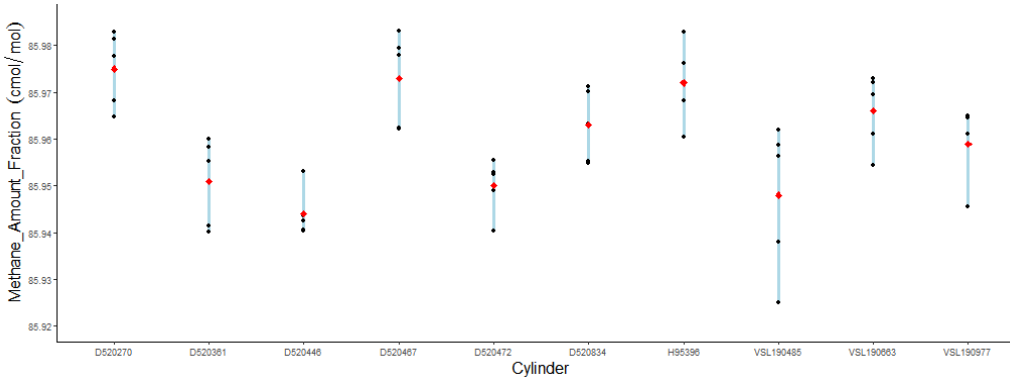

Exhibit 17: Group Subset Plot showing distributions of determinations of the value of the measurand by cylinder. Cylinder means are denoted by red diamonds

We see estimates and 95 % coverage intervals for the three parameters of interest, σ_C, σ, and μ in the table below. The parameter of greatest interest for the assessment of homogeneity is, of course, σ_C, the between-cylinder standard uncertainty. The NIHOMA indicates that σ_C differs significantly from 0. More on how such a determination of significance is made is presented in §6. This is readily seen however, by simply comparing $\widehat{\sigma}_C$ with $\widehat{\sigma}$, the within-group standard uncertainty. Specifically, the facts that the left endpoint of the 95 % coverage interval for σ_C is greater than 0, and that $\widehat{\sigma}_C > \widehat{\sigma}$, suggest that within-cylinder variation

term	group	estimate	2.5 %	97.5 %	std.uncertainty
σ_C	Cylinder	0.01062	0.004695	0.01610	—
σ	Residual	0.009104	0.007122	0.01117	—
μ	Fixed Effects	85.9601	85.9529	85.9673	0.003598

Exhibit 18: NIHOMA random effects model summary for the amount fraction of methane in a synthetic gas mixture simulating natural gas.

is dominated by between-cylinder variation. Therefore, there seems to be sufficient evidence to reject the assumption of homogeneity. Before explaining how one might proceed in such case, we should first validate the assumptions of the random effects model.

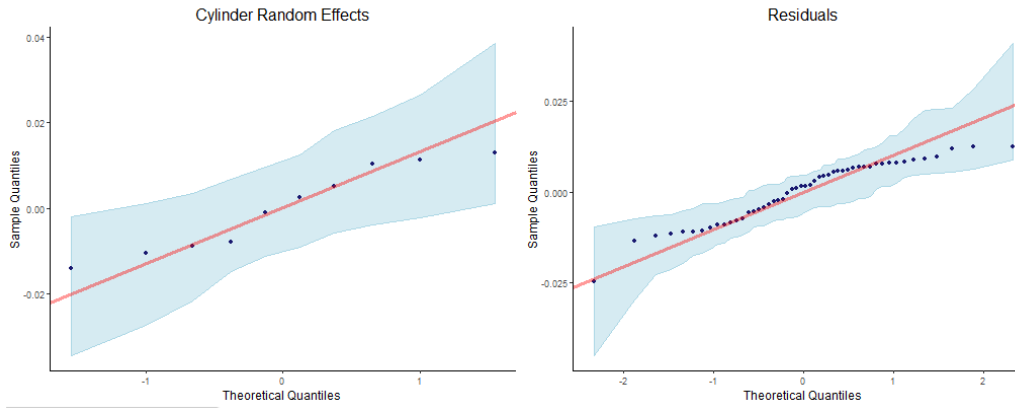

Exhibit 19: Normal QQ-Plots for cylinder random effects and model residuals. Since neither plot has any points falling outside their respective blue bands, the assumptions of the random effects model, presented in §3.1.1 are not violated.

Levene's Test yields a p–value of 0.8, hence there is no reason to suggest that cylinder-specific variances differ from one another. Furthermore, the normal QQ-plots shown in Exhibit 19 support the notion that the cylinder random effects, $\{\alpha_i\}$, and the measurement errors $\{\epsilon_{ik}\}$, are like samples from Gaussian distributions.

Regarding the heterogeneity that appears to be significant, there are two possi-

ble courses of action. If the apparent heterogeneity is excessive, then one option is to split the lot of cylinders into two or more lots that are comparatively less heterogeneous. Another option, which is far more practical, particularly in examples such as this one, where $\widehat{\sigma}_C$ is relatively small, is simply to add $\widehat{\sigma}_C$ to the uncertainty budget and to propagate it to the final result, thus recognizing and expressing this source of uncertainty.

4.3 Barium

The data explored in this subsection are from a VSL homogeneity study (van der Veen, Linsinger, and Pauwels , 2001; Linsinger et al., 2001), although this example is unrelated to work presented in §4.1 and §4.2. The data comprise 3 replicate determinations of the mass fraction of barium in each of 18 different soil samples. For the purposes of demonstrating how to use the NIHOMA when data is uploaded in the Single Measurement Result per Unit format, the original data has been summarized as shown in Exhibit 20.

Entering the above into the NIHOMA produces a summary table with the following information:

- **Estimate of Between-Sample Standard Uncertainty:** $\widehat{\sigma}_S = 5.825$ mg/kg
- **Measurand Estimate and 95 % Coverage Interval:** $\hat{\mu} = 318.73 \pm 2.15$ mg/kg
- **Cochran's Q Test for Heterogeneity:** $Q(df = 17) = 36.75, p\text{–value} = 0.003638$

The results of Cochran's Q Test are strongly suggestive of heterogeneity across the 18 soil samples, a hypothesis that is also supported by the summary plot shown in Exhibit 21. A consequence of the reduction to the case of Single Measurement Result per Unit is that evaluating the assumptions of the random effects model becomes somewhat more difficult. For example, we are not able to apply Levene's Test to assess the assumption of common variance for within-sample replicates, as we don't explicitly have access to the factor level distributions. (We do have such access in this example because we have reduced the original, granular data to the summary statistics in Exhibit 20, but in general the more granular data may not even exist, only a single measurement result for each unit.) Nevertheless, one can still build and examine normal QQ-plots for sample random effects and model residuals, as shown in Exhibit 22.

Mass Fraction	Uncertainty (mg/kg)	df (mg/kg)	Sample
311.33	6.39	2	S0118
330.00	7.21	2	S0201
316.67	3.84	2	S0383
324.67	6.89	2	S0442
329.67	4.18	2	S0557
310.33	7.36	2	S0666
324.00	4.04	2	S0791
317.00	7.00	2	S0918
328.33	4.33	2	S1026
316.67	5.70	2	S1133
310.00	4.00	2	S1249
309.33	9.84	2	S1464
320.00	5.20	2	S1581
315.00	3.51	2	S1607
316.00	9.54	2	S1799
300.00	6.51	2	S1877
324.33	6.06	2	S1996
326.33	7.97	2	S2000

Exhibit 20: Barium data table for upload to the NIHOMA in the Single Measurement Result per Unit format.

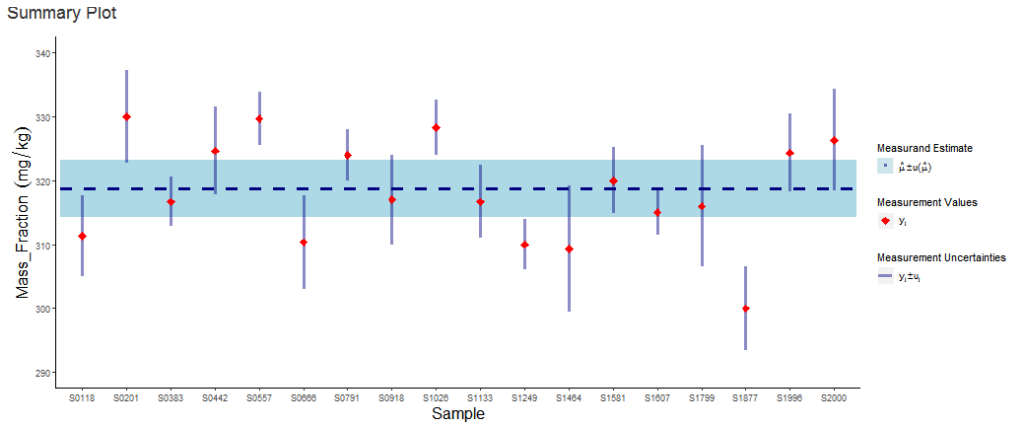

Exhibit 21: Averages of replicates per sample, and associated standard uncertainties, for each of the 18 soil samples, as well as an estimate and 95 % coverage interval for the consensus value of the measurand.

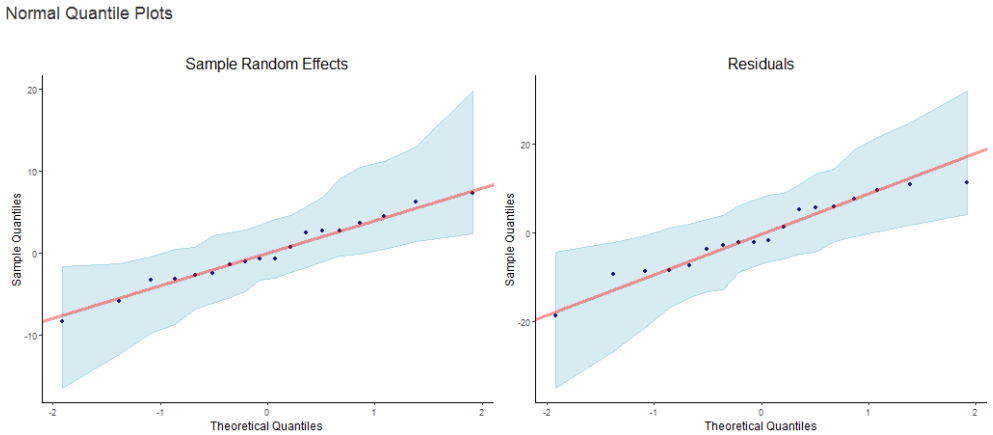

Exhibit 22: Normal QQ-plots for the 18 sample random effecs and model residuals

The normal QQ-plots support the notion that the cylinder random effects, $\{\alpha_i\}$, and the model residuals, $\{\epsilon_i\}$ are like samples from Gaussian distributions. Assessing the statistical significance of σ_S is slightly different in the Single Measurement Result per Unit case, because the NIHOMA does not produce a coverage interval for σ_S. Rather, the determination of statistical significance is done through Cochran's Q Test. If the level of heterogeneity, captured in $\widehat{\sigma}_S$, is acceptable for the purpose the reference material is intended to serve, then the scientist responsible for the production of the reference material can simply add $\widehat{\sigma}_S$ to the uncertainty budget to reflect the heterogeneity in the material, and fold it into the evaluation of uncertainty associated with the reference value.

Again, as has been discussed with previous examples, in cases where there is no statistically significant heterogeneity, scientists have the option to factor in $\widehat{\sigma}_S$ anyway as just described, for a more conservative uncertainty evaluation.

5 Advisory

This section serves to emphasize that the NIHOMA ought not be misconstrued as a toolbox capable of addressing all cases likely to be encountered in homogeneity studies.

The NIHOMA can handle studies with nested sampling factors, but does not currently support the addition of covariates that may have been measured alongside the measurand and that could help explain any apparent heterogeneity. Furthermore, the NIHOMA only supports "Matryoshka" nesting, one factor within another within yet another etc., but does not support any sampling structures that may be more complex, such as two factors nested within another factor but not nested within each other. Lastly, at present, the NIHOMA can only handle a single sampling factor when there is a single measurement result per unit (rather than multiple replicate determinations).

Data collection in studies producing candidate reference materials is often time and labor intensive, hence very costly. One should realize that the smaller the sample size, the more conservatively one should treat the NIHOMA's output. To best characterize heterogeneity, one may wish to carry out a pilot homogeneity study that will produce preliminary estimates of the relevant variance components, and then use these estimates to plan an optimally designed study, where the number of bottles and the number of replicates per bottle is tailored to the study requirements and to the actual material under examination.

6 Implementation

The implementation of the facilities deployed in the NIHOMA has been done exclusively in the R statistical computing language (R Core Team, 2018), given R's fitness for purpose, wealth of specialized functionality, and universal availability. The following packages are required for running the NIHOMA

Models

- `lme4` (Bates et al., 2018) — Random Effects Model, Replicated Determinations per Unit Case
- `metafor` (Viechtbauer, 2017)— Random Effects Model, Single Measurement Result per Unit
- `coin` (Hothorn et al., 2017) — Kruskal-Wallis Procedure
- `car` (Fox et al., 2018) — Levene's Test
- `boot`(Canty and Ripley, 2017) — Dependency code to bootstap coverage intervals for measurand in Averages and Uncertainties Case, developed for NIST Consensus Builder (Koepke et al., 2017a,b)

Graphics

- `ggplot2` (Wickham et al., 2018a) — Grammar of Graphics framework for all NIHOMA plots
- `ggridges` (Wilke, 2018)— Extension for `ggplot2` for ridgeline density plots
- `gridExtra` (Auguie and Antonov, 2017) — Extension for `ggplot2` for combining several plots into a single plot
- `rlang` (Henry and Wickham, 2018) — Parse text for graphics labels

Data Processing

- `dplyr` (Wickham et al., 2018b) — Data processing throughout NIHOMA
- `readxl` (Wickham and Bryan, 2018) — Read `.xls` and `.xlsx` files into R
- `broom` (Robinson et al., 2018) — Render's tidy summaries of model summary tables

- `xtable` (Dahl, 2016) — Render LaTeXversions of model summary tables

Application Design

- `shiny` (Chang et al., 2018) — Framework for the NIHOMA
- `shinyjs` (Attali, 2018) — Javascript integration with `shiny`

To make the application accessible to users with no knowledge of R we have created an easy-to-use graphical user interface displayed in a web browser employing facilities provided by the R package `Shiny` (Chang et al., 2018). NIST eventually will host the NIHOMA at homogeneity.nist.gov, with several instances of the Shiny app running concurrently for load balancing. NIHOMA is also available as an R package called **NIHOMA.App**, available via GitHub.

For the curious user, we add here a brief explanation of about how we determine whether variance components differ significantly from 0 in the Replicated Determinations per Unit case.

Given a user specified coverage probability $0 < p < 1$, the NIHOMA displays $100 \times p$ % coverage intervals for each sampling factor's variance component.

Determining statistical significance can be established by ascertaining that the lower endpoint of the coverage interval for the between-group standard uncertainty is strictly greater than 0.

However, a statistically significant variance component need not be practically significant. For example, when that lower endpoint is several orders of magnitude smaller than the within-group variance. However, when the magnitude of the measurand is very small, a variance component may be within an exceedingly small distance of 0 but still be both statistically and practically significant.

For these reasons, we determine practical significance as described next, where $\hat{\mu}$ denotes the estimate of the measurand, σ_A the within-group standard uncertainty for some sampling factor A, estimated by $\widehat{\sigma}_A$, and σ the within-group standard uncertainty, estimated by $\widehat{\sigma}$. Furthermore, $L_p(\sigma_A)$ and $L_p(\sigma)$ denote the lower endpoints of $100p$ % coverage intervals for σ_A and σ, respectively.

The NIHOMA decides that σ_A is both statistically and practically different from 0 if at least one of the following criteria is met:

- $L_p(\sigma_A) > 0.1 L_p(\sigma)$;
- $L_p(\sigma_A) > u(\hat{\mu})$.

Acknowledgments

References

D. Attali. shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds, 2018. URL https://cran.r-project.org/web/packages/shinyjs/index.html. R package version 1.0.

B. Auguie and A. Antonov. gridExtra: Miscellaneous Functions for "Grid" Graphics, 2017. URL https://cran.r-project.org/web/packages/gridExtra/index.html. R package version 2.3.

R. Beelen. *Preparation of a homogeneous set of PT materials*, 2006. Technical report, VSL, Delft.

D. Bates, M. Maechler, B. Bolker, S. Walker, R.H.B. Christensen, H. Singmann, et al. lme4: Linear Mixed-Effects Models using 'Eigen' and S4, 2018. URL https://cran.r-project.org/web/packages/lme4/index.html. R package version 1.1-17.

A. Canty and B. Ripley boot: Bootstrap Functions, 2017. URL https://cran.r-project.org/web/packages/boot/index.html. R package version 1.3-20.

W. Chang, J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. shiny: Web Application Framework for R, 2018. URL https://CRAN.R-project.org/package=shiny. R package version 1.1-0.

D. Dahl. xtable: Export Tables to LaTeX or HTML, 2016. URL https://cran.r-project.org/web/packages/xtable/index.html. R package version 1.8-2.

J. Fox, S. Weisberg, B. Price, D. Adler, D. Bates, G. Baud-Bovy et al. car: Companion to Applied Regression, 2018. URL https://cran.r-project.org/web/packages/car/index.html. R package version 3.0-0.

ISO. *Reference materials — Guidance for characterization and assessment of homogeneity and stability*. International Organization for Standardization (ISO), Geneva, Switzerland, Fourth edition, 2017. ISO Guide 35:2017(E).

L. Henry and H. Wickham. rlang: Functions for Base Types and Core R and 'Tidyverse' Features, 2018. URL https://cran.r-project.org/web/packages/rlang/index.html. R package version 0.2.1.

T. Hothorn, K. Hornik, M. van de Wiel, Henric Winell, and A. Zeileis. coin: Conditional Inference Procedures in a Permutation Test Framework, 2017. URL https://cran.r-project.org/web/packages/coin/index.html. R package version 1.2-0.

A. Koepke, T. Lafarge, A. Possolo, and B. Toman. Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia*, 54(3): S34–S62, 2017a. doi: 10.1088/1681-7575/aa6c0e.

A. Koepke, T. Lafarge, B. Toman, and A. Possolo. *NIST Consensus Builder — User's Manual*. National Institute of Standards and Technology, Gaithersburg, MD, 2017b. URL consensus.nist.gov.

T. Linsinger, J. Pauwels, A. van der Veen, H. Schimmel, and A. Lamberty. Homogeneity and stability of reference materials. *Accreditation and Quality Assurance*, 6(1):20—25, 2001. doi: 10.1007/s007690000261.

A. Possolo and A. L. Pintar. Plurality of Type A evaluations of uncertainty. *Metrologia*, 54(5):617–632, 2017. doi: 10.1088/1681-7575/aa7e4a.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

D. Robinson, A. Hayes, M. Gomez, B. Demeshev, D. Menne, B. Nutter, et al. broom: Convert Statistical Analysis Objects into Tidy Tibbles, 2018. URL https://cran.r-project.org/web/packages/broom/index.html. R package version 0.5.0.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. John Wiley & Sons, Hoboken, NJ, 2006. ISBN 0-470-00959-4.

A. van der Veen. Bayesian analysis of homogeneity studies in the production of reference materials. *Accreditation and Quality Assurance*, 22:307-319, 2017. doi: 10.1007/s007690000238.

A. van der Veen, T. Linsinger, J. Pauwels. Uncertainty calculations in the certification of reference materials. 2. Homogeneity study. *Accreditation and Quality Assurance*, 6(1):26—30, 2001. doi: 10.1007/s007690000238.

W. Viechtbauer. metafor: Meta-Analysis Package for R, 2017. URL https://cran.r-project.org/web/packages/metafor/index.html. R package version 2.0-0.

H. Wickham, W. Chang, L. Henry, T. Lin Pederson, K. Takahashi, C. Wilke, et al. ggplot2: Create Elegant Data Visualizations Using the Grammar of Graphics, 2018. URL https://cran.r-project.org/web/packages/ggplot2/index.html. R package version 3.0.0.

H. Wickham, R. Francois, L. Henry, K. Muller. dplyr: A Grammar of Data Manipulation, 2018. URL https://cran.r-project.org/web/packages/dplyr/index.html. R package version 0.7.6.

H. Wickham and J. Bryan. readxl: Read Excel Files, 2018. URL https://cran.r-project.org/web/packages/readxl/index.html. R package version 1.1.0.

C. Wilke. ggridges: Ridgeline Plots in "ggplot2", 2018. URL https://cran.r-project.org/web/packages/ggridges/index.html. R package version 0.5.0.