MCRA 8.1

a web-based program for Monte Carlo Risk Assessment

Reference Manual

September 2015
(updated 13-10-2016)

WUR/Biometris, Wageningen University and Research centre
FERA, Food and Environmmental Research Agency
RIVM, National Institute for Public Health and the Environment

Contributors to MCRA
Main programmers of MCRA 8 are:
Waldo de Boer, Johannes Kruisselbrink, Marco van Lenthe

Many people contributed to the MCRA code over the years: Frits van Evert, Jack van Galen, Paul Goedhart, Gerie van der Heijden, Paul Keizer, Marcel Koenders, Jaap Kokorian, Sanne Korzec, Helen Owen, Gerrit Polder, Pim Reijersen, Willem Roelofs, Gert-Jan Swinkels, Jac Thissen, Hilko van der Voet.

MCRA has been tested in practice by many people. Feed-back over many years from the team at RIVM is gratefully acknowledged: Jan Dirk te Biesebeek, Polly Boon, Gerda van Donkersgoed, Jacob van Klaveren

Contributors to the MCRA 8.1 Reference Manual:
Waldo J. de Boer, Paul W. Goedhart, Andy Hart, Marc C. Kennedy, Johannes Kruisselbrink, Helen Owen, Willem Roelofs, Hilko van der Voet

---

WUR/Biometris is the unit for **Mathematical and Statistical Methods of Wageningen University & Research centre**
P.O. Box 16, 6700 AA Wageningen, Netherlands
WUR Campus, Building 107 (Radix), Droevendaalsesteeg 1, 6708 PB Wageningen
Telephone: +31 (0)317 476925
http://wageningenur.nl
http://www.biometris.nl

Fera Food and Environmental Research Agency
Sand Hutton, York, YO41 1LZ, United Kingdom
Telephone: +44 (0)1904 462000
http://fera.defra.gov.uk

RIVM National Institute for Public Health and the Environment
P.O. Box 1, 3729 BA Bilthoven, Netherlands
Antonie van Leeuwenhoeklaan 9, 3721 MA Bilthoven
Telephone: +31 30 2749111
http://rivm.nl

Contents

1 Introduction

This reference manual describes MCRA Release 8.1. The acronym MCRA stands for Monte Carlo Risk Assessment. MCRA is a web-based platform for probabilistic risk assessment of substances in the diet (and optionally also from other routes of exposure). The MCRA platform brings together statistical models, shared data and data uploaded by the user. MCRA 8.1. has been implemented in the flexible environment for high-performance computing at RIVM, allowing the use of a flexible number of simulation worker services to address simultaneously submitted jobs in parallel (Figure 1).
MCRA 8 is available at https://mcra.rivm.nl.

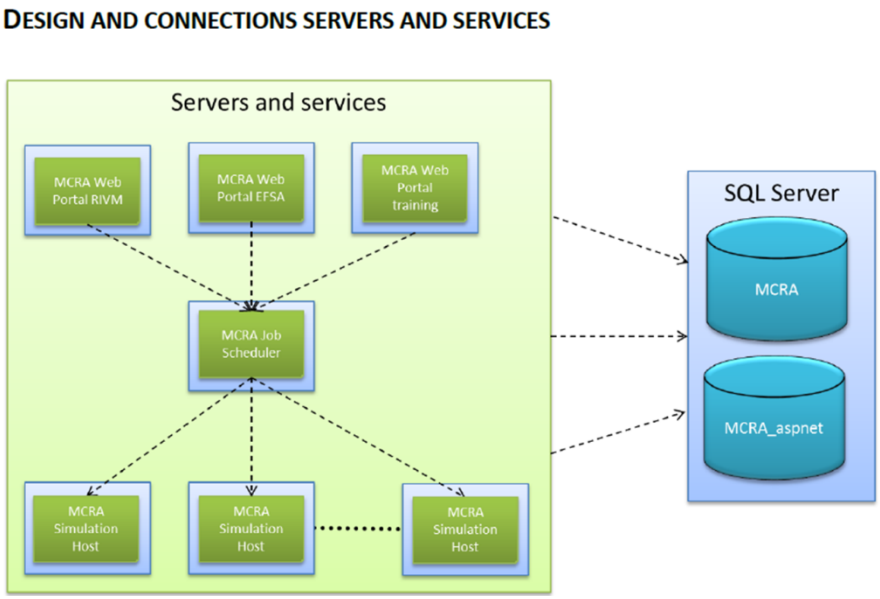


Figure 1. MCRA 8.1 infrastructure

2 Assessment types

In probabilistic risk assessment we consider a population of individuals. Risk assessment with MCRA can address **acute risk** or **chronic risk**. Acute risk is relevant when the short-term effect on individuals is relevant, chronic risk when the long-term effects on the individuals matter. In MCRA short-term is operationalised as one day, so effectively acute risk assessment is concerned with a population of person-days, whereas chronic risk assessment is concerned with a population of persons.

Risk assessors may approach food safety in a population on at least three levels:

1. **Consumption Assessment**: quantify the consumption of a possible risk food.
2. **Exposure Assessment**: quantify the exposure to a substance (or group of substances) in the diet.
3. **Health Impact Assessment**: quantify and integrate the impact on one or more health parameters from exposure to one or more substances in the diet.

The hierarchical nature of these assessments is illustrated in Figure 2, also listing the required (colour) and optional (grey) data used for the assessments.

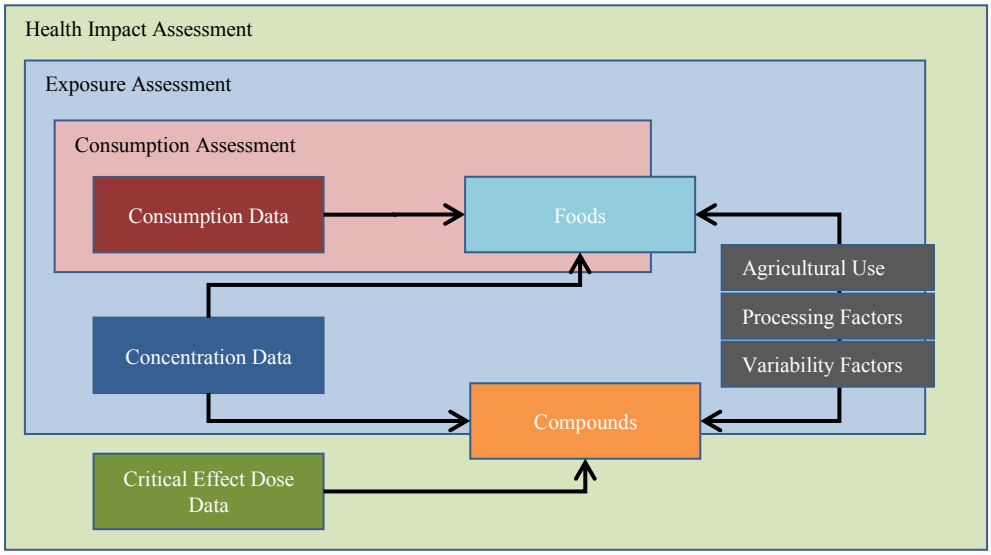


Figure 2. Assessment types

Currently, MCRA 8 implements Exposure Assessments and Health Impact Assessments only. Consumption Assessments can be performed by using a dummy compound with concentration 1. It is the intention to add Consumption Assessments to the MCRA 8 platform.

The basic operation in exposure assessment is integrating consumption and concentration data per food. With multiple foods, consumptions are typically correlated, therefore MCRA works with the multivariate distribution of a consumption vector, as represented by the consumption data of individuals in a consumption survey. In contrast, the distributions of concentration for each food are typically considered to be independent between foods, *e.g.* eating an apple with an accidentally high residue concentration does not predict that another food eaten on the same day will also have a high residue concentration. As a consequence of this assumption, MCRA models concentrations of substances for each food independently.

Exposure assessments are most often concerned with a single chemical substance. MCRA 8 also provides **Cumulative Exposure Assessment** for groups of substances which have been grouped in a Cumulative Assessment Group (CAG) for which a single health effect is considered relevant. This requires the availability of Relative Potency Factors (RPFs) for all the substances in the CAG.

MCRA was originally designed for dietary exposure assessment. In MCRA 8 it is also possible to perform **Aggregate Exposure Assessment**, combining multiple routes of non-dietary exposure (dermal, inhalatory, oral) with dietary exposure to aggregated internal exposure. This requires the specification of aborption factors for the relevants routes of exposure.

In practice often a single food may be of special interest, for example if the effects of spraying a particular pesticide on a specific crop have to be assessed. For such cases MCRA 8 provides **Focal Food Assessments**, which allows concentration data for the focal food to be specified separately, and replace the concentration data for that food in the main concentration data.

Total Diet Studies (TDS) are complement to traditional monitoring and surveillance programs and provide a scientific base for dietary exposures by analysing foods as consumed. For more information about Total Diet Studies, visit the TDS-Exposure website http://www.tds-exposure.eu.

3 Food coding and conversion

MCRA is intended to retain complete transparence of the results of risk assessment in terms of the foods that were actually consumed (foods-as-eaten). In many cases measurements of substances have not been made on the **food-as-eaten**, *e.g.* pizza, but on a raw agricultural commodity (RAC), *e.g.* tomato, onion etc. The food on which the concentration measurements have been made is termed the **food-as-measured**. MCRA implements a recursive search algorithm to link foods-as-eaten to foods-as-measured. This means that there can be intermediate steps, *e.g.* if unpeeled *apple* and *grapes* are the foods-as-measured, the food-as-eaten *apple pie* contains *peeled apple* and *raisins*, *peeled apple* is linked to unpeeled *apple*, and *raisins* are dried *grapes*.

3.1 Food code definition

In MCRA, a food code is a string consisting of symbols. Some special symbols (. $ - *) are reserved for special use (see below), and can not be used freely in own codes.

Codes can be hierarchical. Any code can be followed by $ or . plus a subtype code. This can be repeated any number of times, e.g. ABC$D.

Codes can specify food processing. Any code can be followed by - plus a processing code. Only one level of processing code is allowed (*e.g.* FP0226-2). Subtype codes should precede processing codes (*e.g.* NL005$123$456-2).
The asterisk (*) serves as a wildcard for the preceding code: the processing information is valid for all codes that start with the code preceding the *.

3.2 Food codes in consumption surveys

Any coding system for foods-as-eaten can be used in MCRA. For example, in Europe EFSA develops a Food Classification and Description System for exposure assessment named FoodEx 2 (EFSA 2011a,b), featuring a hierarchical system of a core list of foods, an extended list, and domain-specific hierarchies.

3.3 Food codes in concentration data

Any coding system for foods-as-measured can be used in MCRA.

3.4 Food processing

Concentrations of substances in foods may change when foods are processed. Examples of processing types are peeling (*e.g.* of apples), *cooking* (*e.g.* of spinach), drying (*e.g.* of grapes), juicing (*e.g.* of oranges). In MCRA a **processing factor** can be specified for any food. Processing factors specify the ratio of concentrations in the processed and unprocessed food. The food code of the processed food will be converted to the food code of the unprocessed food. In simulations the concentration in the unprocessed food will then be multiplied by the processing factor.
Special attention is needed if food processing also changes the weight of the food. Traditionally, processing factors combine the effects of chemical alteration and weight change, so the weight change should not be double-counted. The processing correction factor is introduced to correct processing factors that combine both effects, *e.g.* when 100g *raisins* (dried grapes) are translated to 300g *grape* (food-as-measured) and the processing factor for drying combines both effects, the processing correction factor is 3.

3.5 Recipes and food translation

Recipes specify the composition of composite foods, *e.g. pizza*, in terms of relevant ingredients, *e.g.* 100g pizza contains 10g *tomato*, 5g *cheese* and 50g *flour*. Recipes are also used to specify weight changes, *e.g.* to obtain 100g *raisins* (dried grapes) 300g of the food-as-measured *grape* is needed.

A special use of recipes and food translation is found in Total Diet Studies. Here, the composition of a Total Diet Study food is specified, *e.g.* TDS-food *FruitMix* is composed of *apple*, *orange* and *pear* with a default translation proportion of 100%. So in MCRA, the food-as-eaten *apple* is converted to *FruitMix* (100%) and *FruitMix* is considered as the food-as-measured (TDS-food). A conversion from *apple-pie* (food-as-eaten) to *FruitMix* (food-as-measured) is based on a recipe for apple-pie and a TDS composition for FruitMix.
Another use of converting foods (as-eaten or as an intermediate step), is through the specification of so-called read across translations, *e.g.* for *pineapple* no measurements are found but by specifying that *pineapple* is converted to *FruitMix* (with a default proportion of 100%), the TDS sample concentration value of *FruitMix* will be used for *pineapple* (as-eaten or as ingredient).

3.6 Market shares and brand loyalty

Sometimes measurements of substances in food are available at a more detailed food coding level than consumption data. For example, measurements may have been made for specific brands of a food whereas the consumption survey did not record the brand. MCRA allows to specify **market share** data for subtypes of a food (e.g. A$1, A$2, A$3 are three brands of food A), and to calculate acute exposure based on such market shares.
For chronic assessments **brand loyalty** should be specified according to a simple Dirichlet model (Goodhardt *et al.*,1984). Technically, the Dirichlet model for brand choice needs *nbrand* parameters α_i (which should be positive real numbers). The average brand choice probability for each brand is α_i/S, where $S = \sum \alpha_i$. By definition, the market shares m_i should be proportional to the brand choice probabilities, and thus to the parameters α_i. Thus means that S, the sum of the alphas, is the only additional parameter that should be specified, and indeed this is the parameter that determines brand loyalty. S=0 corresponds to absolute brand loyalty, and brand loyalty decreases with increasing S. We define $L = (1+S)^{-1}$ as an interpretable brand loyalty parameter, where now $L = 0$ and $L = 1$ correspond to the situations of no brand loyalty and absolute brand loyalty, respectively.
Given empirical or parametric distributions of consumption and concentration values, the algorithm for chronic exposure assessment now operates as follows:

1. Simulate consumptions for a large number *n* of individuals.
2. Simulate *n* selection probabilities from the Dirichlet distribution
3. For each individual, simulate *d* brand choices from a multinomial distribution using the individual specific selection probabilities from step 2.
4. For all individuals and days simulate values from the appropriate concentration distribution.
5. Multiply consumption with concentration to obtain exposure.

3.7 Supertypes

Sometimes measurements of substances on food are available at a less detailed food coding level than consumption data. MCRA allows to use the concentration data of a supertype for all underlying food codes. However, this is not the default, and an explicit permission should be given to allow this feature.

3.8 Maximum Residue Levels

Maximum residue levels are the upper legal levels of a concentration for substance residues in a food, *e.g.* pesticide, or feed based on good agricultural practices and to ensure the lowest possible consumer exposure.

3.9 MCRA food code conversion algorithm

Food codes are linked using a 7-step procedure. For each foods-as-eaten, the code is matched according to the following steps.

1. Identical code: try to find the code in the **Concentration** values group (step 1). The algorithm checks whether the measurements on a food are all non-detect (< LOR) or not. For pessimistic

scenario's, consider to proceed with foods with only non-detect measurements as well. This step is optional, see advanced settings if you want to use this.
If successful, a food-as-measured has been found, and the current search stops.

2. Processing link: try to find the code in the **ProcessingFactors** table (step 2a). If successful, try to find find the code in the FoodTranslations table (step 3a) to correct for weight reduction or increase.
   Processing link wildcard match: try to find a wildcard match in the processing table (step 2b). Wildcard match codes consist of an initial string (*startcode*, may be empty), an asterisk (*), and possibly a processing part (*-processingtype*). * may be any string *endcode* (not containing a -) such that *code* equals *startcodeendcode* or *startcodeendcode-processingtype*.
   a. If *code* contains a processing part (*-processingtype*), then the wildcard match code should also end with *-processingtype.* Convert to the code specified in the field **foodunprocessed**, where *endcode* is substituted for any * in the new code.
   b. If *code* contains no processing part, then the wildcard match code should also contain no processing part. Convert to the code specified in the field **foodunprocessed**, where *endcode* is substituted for any * in the new code.

   If successful, restart at step 1 with the new code of the unprocessed food.
3. Food translation link: try to find the code in the **FoodTranslations** table (step 3a). This may result in 1 or more food codes for ingredients, and the iterative algorithm will proceed with each of the ingredient food codes in turn. Try to find the code in the **TDSFoodSampleCompositions** table, a default translation proportion of 100% is assumed. The iterative algorithm will proceed with a TDS food sample (step 3b). Try to find the food in the **ReadAcrossFoodTranslations** table, a default translation proportion of 100% is assumed (step 3c). If successful, restart at step 1 with each of the new codes of the ingredient foods, TDS foods or Read Across foods.
4. Subtype link: try to find subtype codes, *e.g.* 'xxx$*' in the **MarketShares** table. In general, marketshares should sum to 100%. Foods with marketshares not summing to 100% are ignored in the analysis unless the checkbox 'Allow marketshares not summing to 100%' is checked. This step is optional, see advanced settings if you want to use this.
   If successful, restart at step 1 with each of the new codes of the subtype foods.
5. Supertype link: try to find supertypes, *e.g.* 'xxx$yyy' is converted to 'xxx'. This step is optional, see advanced settings if you want to use this.
   If successful, restart at step 1 with the new code of the supertype food.
6. Default processing factor 1: remove processing part (-xxx) of the code.
   If successful, restart at step 1 with the new code without processing part.
7. Maximum residue limit: try to find the code in the **MaximumResidueLimits** table.
   If successful, the current search stops. If not successful, then stop anyway and the search is marked as failed food conversion.

4 Food classification: FoodEx2

'The collection and evaluation of data on levels of chemical occurrence or presence of biological agents in food and feed are important tasks of EFSA. By combining the data with information on food consumption allows for detailed intake and exposure estimates crucial to any food and feed safety risk assessment or nutrient adequacy analysis. The EU Member States provide an increasing volume of data to EFSA and other European bodies. To provide a common link to all the diverse food and feed databases, a system for the unique and universal identification and characterisation of food and feed items is essential.
EFSA has developed a preliminary standardised food classification and description system called FoodEx2 (version 2 of the EFSA Food Classification and Description System [FCDC] for exposure assessment). The system consists of descriptions of a large number of individual food items aggregated into food groups and broader food categories in a hierarchical parent-child relationship. Central to the system is a common 'core list' of food items or generic food descriptions that represent the minimum level of detail needed for intake or exposure assessments. More detailed terms may exist in addition to the core list and these are identified as the 'extended list'. A parent-child relationship exists between a core list food item and its related extended list food items. The terms of the core and extended list may be aggregated in different ways according to the needs of the different food safety domains. In the present version four hierarchies are proposed: three domain-specific and a general purpose one. Facets are used to add further detail to the information provided by the food list term. Facets are collections of additional terms describing properties and aspects of foods from various perspectives'. For more information visit: http://www.efsa.europa.eu/en/datex/datexfoodclass.htm.

4.1 FoodEx2 in MCRA

For MCRA, having a different set of food codes is in itself not a problem. That is, for MCRA, it does not matter how foods are coded, as long as they can be linked to consumptions and concentrations within an exposure assessment. What makes FoodEx2 different from other food coding systems is that it provides additional food hierarchies, food facets, and a combined food/facet coding system. Below follows a brief summary of these main features of the FoodEx 2 coding system from the perspective of exposure assessment using MCRA.

4.1.1 Foods and food hierarchies

FoodEx 2 contains different food hierarchy definitions and allows for creation of custom food hierarchy definitions. These hierarchies could, for exposure assessment, allow to assess intake or consumption data based on the groups defined by these hierarchies.

Table 1shows a part of the FoodEx 2 *Exposure Hierarchy* exported from the FOODEX 2.0 Browser version 0.1.3.

Code	Level	Name	ParentCode	Scopenotes
A000J	1	Grains and grain-based products	ROOT	The category covers all …
A000K	2	Cereals and similar	A000J	…
A000L	3	Cereal and cereal-like grains	A000K	…
A000M	4	Amaranth grain	A000L	…
A000N	4	Buckwheat grain	A000L	…
A000P	4	Barley grain	A000L	…
…	…	…	…	…

Table 1: Food hierarchy export from FOODEX 2.0 Browser version 0.1.3

4.1.2 Facets and facet descriptors

FoodEx 2 allows to provide supplementary details on specific aspects of foods by means of so-called facets and facet descriptors. Facets are collections of terms defining specific characteristics of food

from particular points of view and facet descriptors describe specific characteristics foods. For example, *processing technology* is a facet, and *baking* is a facet descriptor belonging to this facet. Currently, 26 facets are defined, containing in total 2172 descriptors (EFSA 2011b).
Facets are also defined in a hierarchical system. For instance, c*ooking in fat* (*A07GR*) and *baking* (*A07GX*) are sub-items of the descriptor c*ooking and similar thermal preparation processes* (*A0BA1*).
Facets are coded as small strings that consist of a facet code and a facet descriptor code separated by a '.'-character. For example, the facet code *F28.A07GX* holds 1) the facet code *F28*, which is the facet code for *process technology*, and 2) *A07GX* , which is the descriptor code for *baking*.
Table 2 shows a part of the FoodEx 2 facet descriptor codes of the *source* facet (F01).

Code	Level	Name	ParentCode	Scopenotes
A04SF	1	Animals	ROOT	…
A056H	2	Mammals (food source animal)	A04SF	…
A056Z	3	Farmed / non-game mammals (food source animal)	A056H	…
A057A	4	African buffalo (food source animal)	A056Z	…
A057B	4	American buffalo (food source animal)	A056Z	…
A057C	4	Buffalo (food source animal)	A056Z	…
A057D	4	Cape buffalo (food source animal)	A056Z	…
A057E	4	Cattle (food source animal)	A056Z	…
…	…	…	…	…

Table 2: Facet descriptor export of the source facet (F01) from FOODEX 2.0 Browser version 0.1.3

4.1.2.1 *Implicit facets*

Implicit facets are facets of a product that are already implied by the food product itself. Consider, for example, *potato boiled* (*A011P*), where *boiling* (*A011P*) is an implicit facet, because boiling is already implied by the product.
According to EFSA (2011a) 'inclusion of implicit facets in the string recorded for each food database record is not encouraged' and it is suggested to identify and record the implicit facet descriptors in a separate table.

4.1.2.2 *Foods as facets*

Foods and facet descriptors share the same unique alphanumerical coding system; in some cases, like *characterising ingredient* or *sweetening agent* food list elements may be used as facet descriptors.

4.1.3 The FoodEx 2 coding system

In the coding system, facets can be added to the primary food codes to provide supplementary detailed information of particular data records. The structure of the FoodEx 2 codes is:

idFood#idFacet.idFacetDescriptor$idFacet.idFacetDescriptor$....

The code starts with the primary FoodEx2 food code. Then, when there are supplementary facets, the food code is followed by a '#'-character and the facets string. The facets string is constructed as a concatenation of the individual facets strings, separated by means of the '$' character.
As an example, consider the string *A011P#F28.A07GL$F28.A07KQ* which is composed of:

- Food: *A011P - Potato boiled*
- Facet 1: *F28.A07GL - Process technology - Boiling*
- Facet 2: *F28.A07KQ - Process technology - Freezing*

4.2 FoodEx2 in MCRA

For MCRA, FoodEx 2 introduces the following points of attention:

- Reading and dealing with FoodEx 2 coded data sets
- Reading and dealing with food facets
- Reading and exploiting food hierarchy data

4.2.1 Reading and dealing with FoodEx 2 codes

All data entities that contain foods data are potentially affected by the introduction of FoodEx 2. In MCRA, the following data tables are adapted to allow for input of full FoodEx 2 food codes:

- Foods
- Consumptions
- Concentrations

For these tables, the food code is allowed to be the complete FoodEx 2 food code and automatically recognized as such. As an example, Table 3 shows how the FoodEx 2 coded consumptions should be provided to the system.
On important note: the maximum field length of the food code is 50. This means that there is a maximum of five facets that can be specified for a food.

Individual	DayOfSurvey	Food	Amount	FoodSurvey
14233701	1	A011R# F28.A07GX	153.43	FS01
18843004	1	A011R# F28.A07GX	125.23	FS01
34025701	1	A011R# F28.A07GX	153.6	FS01
14720005	2	A011P# F28.A07GX	105.00	FS01
49174010	1	A011P# F28.A07GX	140.00	FS01
62794010	1	A011P# F28.A07GX	67.00	FS01
61392002	1	A011P# F28.A07GL$F28.A07KQ	104.72	FS01
61281231	1	A011P# F28.A07GL$F28.A07KQ	104.72	FS01

Table 3: Integrated coding of the facets in the consumed foods field of food consumptions. Implementation

4.2.2 Reading and dealing with facets data

Within MCRA, the following facets related aspects are accounted for:

- Reading facets data
- Dealing with facets
- Facets in concentration data
- Facets in food conversion
- Using facets as processing factors
- Using hierarchy data in the output

4.2.2.1 *Reading facets data*

To incorporate input of facets data in MCRA, two tables *Facets* and *FacetDescriptors* are introduced as optional tables of the Foods data group:

Column name	Key	Required	Type	Size	Description
idFacet	Yes	Yes	String	5	The id of the facet of this definition.
Name	No	Yes	String	200	The name of the facet.

Table 4 Facets table definition of the Food data group.

Column name	Key	Required	Type	Size	Description
idFacetDescriptor	Yes	Yes	String	5	The id of the facet descriptor of this definition.
Name	No	Yes	String	200	The name of the facet descriptor.

Table 5 FacetDescriptors table definition of the Food data group.

Within MCRA, the facets of FoodEx 2 coded foods, consumptions, and concentrations are automatically linked to the provided facets and facet descriptors. Also, the facet descriptor names are added automatically to the foods containing these facets.

4.2.2.2 *Dealing with facets*

The introduction of food facets allows for much more detailed specifications of consumption and concentration data. However, it introduces the problem of deciding on which level of detail the exposure assessment should be performed.
That is, should concentration models be generated on the level of foods-without-facets or on the level of foods-with-facets? E.g., should the concentrations of *clementine peeled* (A01CE#F28.A07LC) and *clementine unprocessed* (A01CE#F28.A0C0S) be modelled separately or should one model be constructed for *clementine* (A01CE)? Treating all clementine's as equal may yield over-simplified conversions, whereas treating all separately may lead to many concentration models based on only few measurements.
In MCRA, no implicit grouping of concentrations of equal foods with different facets is applied. If concentrations are provided for both *clementine peeled* (A01CE#F28.A07LC) and *clementine unprocessed* (A01CE#F28.A0C0S), then these are modelled separately.
Another question is whether the order of the facets is relevant or not. E.g., is A0BYV#F02.A06GF$F03.A06HY the same as A0BYV#F03.A06HY$F02.A06GF?
Regarding this matter, MCRA considers the facet order to be important. I.e., A0BYV#F02.A06GF$F03.A06HY is not the same as A0BYV#F03.A06HY$F02.A06GF.

4.2.2.3 *Facets in food conversion*

For conversion of foods-as-eaten to foods-as-measured, MCRA considers foods with different facet strings as different foods. I.e., there is no implicit conversion of foods-with-facets to foods-without-facets and also the order of the facets is important.
However, as it is realistic to convert food-with-facets to the base food without facets, an additional (explicit) conversion step *remove-all-facets* is added that converts foods with facets to the base foods. I.e., the action is "remove all". There is no conversion step for "stripping off one facet at a time". The reason for this is that there is no good way of deciding which facet to strip off first. This new conversion step is somewhat equivalent to the already existing *default processing* conversion step (step 6), and is therefore implemented as step 6b of the conversion algorithm.
Particular rules followed by this step:

- Conversion of food-with-facets to food-without-facets.

4.2.2.4 *Using facets that reveal processing data*

Facets containing processing information, such as *part-consumed-analysed* (F20) and *processing technology* (F28) could be integrated with processing data. As an example, consider *clementine peeled* (A01CE#F28.A07LC). This could be linked to *clementine* (A01CE), with processing type *removal of external layer* (A07LC). Linking to processing data could be achieved by entering processing data using the facet codes.
As an alternative to the current processing factor tables, a facet-based processing factors table is defined for processing facets. That is, the codes for food processed and unprocessed are implicitly defined for FoodEx 2.

FacetCode	Compound	FoodCode	Proc Nom	Proc Upp	ProcNom UncUpp	ProcUpp UncUpp
A07LC	CompoundX	A01CE	0.5	0.6	0.05	0.06
F28.A07GV	CompoundX	A0BY	0.2	0.1	0.03	0.04

Table 6: Example of a MCRA processing factors table using FoodEx 2 foods and facets codes.

Note that in the example, the facet code could be specified as the full facet code, or just the code of the facet descriptor.
As a more elaborate example consider

French fries from cut potato (A0BYV#F02.A06GF$F03.A06HY$F04.A00ZT$F28.A07GR)

For this food code, the substring of the processing facet is extracted from the list of facets.

A0BYV#F02.A06GF$F03.A06HY**$F28.A07GR**$F04.A00ZT
→ (processing facet link A07GR)
A0BYV#F02.A06GF$F03.A06HY$F04.A00ZT

In MCRA, a table *FacetProcessingFactors* is introduced that allows for specification of processing factors by means of facets. This table has the following structure:

Column name	Key	Required	Type	Size	Description
idProcessingType	Yes	Yes	String	5	The facet code of this processing factor definition. May be specified as full facet code, i.e., facet code plus facet descriptor code, or as the facet descriptor code.
idFood	Yes	Yes	String	200	The food code.
idCompound	Yes	No	String	50	The compound for which this processing factor is defined.
Nominal	No	Yes	Double		Nominal value (best estimate of 50th percentile) of processing factor (defines median processing factor)
Upper	No	Yes	Double		Upper value (estimate of 95th percentile or "worst case" estimate) of processing factor due to variability
NominalUncertaintyUpper	No	Yes	Double		Upper 95th percentile of nominal value (Nominal) due to uncertainty. A standard deviation for uncertainty of the nominal value (Nominal) is derived using the nominal value (Nominal) and upper 95th percentile (NominalUncertaintyUpper).
UpperUncertaintyUpper	No	Yes	Double		Upper 95th percentile of upper value (Upper) due to uncertainty. From the nominal value (Nominal), upper value (Upper) and the specified uncertainties of these values (NominalUncertaintyUpper and UpperUncertaintyUpper, respectively) the degrees of freedom of a chi-square distribution describing the uncertainty of the standard

	deviation for variability is derived.

Table 7 Table FacetDescriptors of the Food data group.

The integration with the food conversion algorithm is as follows: Conversion step 2 (*processing*) is extended with a step 2c (*processing facet*) that attempts to match facets of a food code to processing data provided in the processing facets table. The following important rules are followed:

- Processing factors can be defined for base-food-code/facet-code combinations and translate as food-with-processing-facet to food-without-processing-facet.
- If multiple processing facets are present in the food-as-eaten code, then the last processing facet is used first for conversion.
- Facet processing factors can be specified using the full facet code (i.e., facet-code plus facet-descriptor-code) or just the facet descriptor code. If both are specified for the same food, the full facet code is used.
- Facet processing factors can be defined compound-specific, and non-compound-specific. Processing factors that are defined compound-specific always precede non-compound specific processing factors.
- Processing factors defined by a food-processed/food-unprocessed combination precede processing factors defined through facets.

Weight reduction factors for processing factors defined for facets should be included in the food translation table and should match exactly.

4.2.3 Reading and dealing with food hierarchy data

Within MCRA, the following hierarchy related aspects are accounted for:

- Reading food hierarchy data
- Using hierarchical data for conversion of foods
- Using hierarchy data in the output

4.2.3.1 *Reading food hierarchy data*

A new data group named *Food hierarchy* is added. In this group, a new table *FoodHierarchy* is proposed for input of food hierarchies. This table has the following structure:

Column name	Key	Required	Type	Size	Description
idFood	Yes	Yes	String	50	The id of the food/node of this definition.
idParent	Yes	No	String	50	The parent of the food of this definition.

Table 8: Proposed new table definition: table FoodHierarchy of the new data group Food hierarchy. This table contains food hierarchy node-definition records that reflect a hierarchical structure. For foods that are not in this list as idFood, it is implicitly assumed that these foods are root items.

Note: It is common practice to describe hierarchies using tree structures. Here, the elements of the tree are named *nodes*, the lines connecting the nodes are named *branches*, and nodes without children are *leaf nodes*/*end-nodes*. This terminology is also used throughout the remainder of this document.

4.2.3.2 *Using food hierarchies for food conversion*

The introduction of the hierarchy structure allows for integration with step 4 and step 5 of the food conversion algorithm; the *subtype* and *supertype* linking steps. That is, when no concentration data is found for a certain product, the concentration data of a (according to the hierarchy) related product could be used.
In MCRA, the *supertype* conversion step also contains a *hierarchy-supertype* step based on the food hierarchy.
Supertype link (step 5):

a) **Supertype:** Try to find supertypes base on '$'-coded strings, e.g., 'xxx$yyy' is converted to 'xxx'
b) **Hierarchy-supertype:** try to find the supertype of the current food based on the food hierarchy (i.e., convert the current food to its parent).

Note 1: the *supertype* conversion step is optional and should be specified in the conversion settings panel.
Note 2: the *hierarchy-supertype* step only applies for foods-without-facets. The reason for this is that for the conversion, the base type of a food-with-facets can be considered as a better conversion candidate than the parent food with the same facets.

4.2.3.3 *Using hierarchy data in the output*

Food hierarchy information could be used in presentation of various tables of the output of MCRA. That is, in the tables in which foods data is presented, these records could be grouped based on the hierarchy and/or a tree-like display can be built for the presentation of this data. Tables that are candidate for being extended are, for example, the input data tables foods-as-eaten/foods-as-measured and the exposure by food-as-eaten/food-as-measured output tables.

Summarizing over the food hierarchy is many cases not a straightforward task. Consider, for instance, the statistic *number of consumption days* given the artificial hierarchy of *Citrus Fruits* containing two child-nodes *Mandarin* and *King Mandarin*: the number of consumption of *Citrus Fruits* is not "just" the sum of the consumption day of *Mandarin* and *King Mandarin*.
A difficulty for summarizing based on a hierarchy arises when a node contains both data and child-nodes with data. E.g., concentrations are defined on the level of *Citrus Fruits* and on the level of *Mandarin*. In this case, the hierarchy view should ideally summarize for both *Citrus Fruits* as data record and *Citrus Fruits* as summary node.
An additional complication is the status of facet-coded foods within the hierarchy. In a hierarchical view, foods-with-facets should ideally be added to their base-foods for visualization.

In MCRA, an alternative view (treetable) is added that can display hierarchical data. This alternative view is used to present a hierarchical view based on the foods hierarchy for the consumption input summary tables *food as eaten* and *food as measured*. The data summary methods for these tables are updated such that the data is also summarized per hierarchy-node.

Food name	Food code	Mean consumption (g)	Mean consumption days (g)	Consumption days	Percentage consumption days	Total weights consumption days	Percentage total weights consumption days
⊟ Fruit and fruit products	A01BS	167	200	5	83.3 %	5.0	83.3 %
⊞ Fresh fruit	A04RK	167	200	5	83.3 %	5.0	83.3 %
⊟ Starchy roots or tubers and products thereof, sugar plants	A00ZR	100	600	1	16.7 %	1.0	16.7 %
⊟ Starchy root and tuber products	A011B	66.7	400	1	16.7 %	1.0	16.7 %
⊟ Processed root and tuber products	A04MJ	66.7	400	1	16.7 %	1.0	16.7 %
⊟ Potato boiled	A011P	66.7	400	1	16.7 %	1.0	16.7 %
Potato boiled Tuber (as part-nature)	A011P#F02.A067V	16.7	100	1	16.7 %	1.0	16.7 %
Potato boiled Tuber (as part-nature), Potatoes, Boiling	A011P#F02.A067V$F27.A00ZT$F28.A07GL	16.7	100	1	16.7 %	1.0	16.7 %
Potato boiled Tuber (as part-nature), Potatoes, Boiling	A011P#F02.A067V$F28.A07GL$F27.A00ZT	16.7	100	1	16.7 %	1.0	16.7 %
Potato boiled Tuber (as part-nature), Potatoes, Boiling, Baking	A011P#F02.A067V$F27.A00ZT$F28.A07GL$F28.A07GX	16.7	100	1	16.7 %	1.0	16.7 %
⊟ Starchy roots and tubers	A00ZS	33.3	200	1	16.7 %	1.0	16.7 %
⊟ Tubers	A04MC	33.3	200	1	16.7 %	1.0	16.7 %
⊟ Potatoes	A00ZT	33.3	200	1	16.7 %	1.0	16.7 %
Potatoes Potatoes (food source plant), Tuber (as part-nature)	A00ZT#F01.A05KG$F02.A067V	16.7	100	1	16.7 %	1.0	16.7 %
Potatoes Potatoes (food source plant), Tuber (as part-nature), Baking	A00ZT#F01.A05KG$F02.A067V$F28.A07GX	16.7	100	1	16.7 %	1.0	16.7 %

Figure 3 Hierarchy view for the foods as eaten input summary table.

If a node contains both data and a child record, then this node is split-up in two nodes: a summary node that summarizes the data of the node and all of its child nodes, and a data record with the string "(unspecified)" added as a child of this summary node. See Figure 3 for an example (*Citrus Fruits* versus *Citrus Fruits (unspecified)*).
In MCRA, foods-with-facets are added as child nodes of the foods-without-facets

5 Consumption data and modelling

Twenty-four hour dietary recall data are stored in table **Consumptions**. For an acute exposure assessment, the interest is in a population of person-days, so one day per individual may be sufficient. For chronic exposure assessments, the interest is in a population of person, so preferably two or more days per individual are needed.
Table **FoodSurvey** is used to specify the number of days of the survey (obligatory field *NumberOfSurveyDays*).
Table **Individuals** lists individual id's and characteristics like gender, age, body weight and/or sampling weight. If the number of survey days varies between individuals, field *NumberOfSurveyDays* can be used to overrule *NumberOfSurveyDays* from table FoodSurvey.

In acute exposure assessments, importance sampling is used to sample some individuals more frequently than others based on given sampling weights. The resulting set of individual days is weighted and approximates the true population of individual days. In MCRA, importance sampling is applied for a specified number of Monte Carlo iterations, typically 100.000. Optional is to take the data as such, statistical sampling weights enter all calculations where weighing is involved *e.g.* the estimation of percentiles or summary statistics of foods and compounds. In chronic exposure assessments, sampling weights enter the simulation as statistical weights in regression and variance components.

MCRA does not implement consumptions assessments. However, by introducing a dummy compound with concentration value 1, the estimated distribution is the result of consumption patterns only.

6 Concentration data and modelling

MCRA models concentrations of substances or compounds independently for each food. In cumulative assessments the modelling is in principle also independent for each compound, but there is an option to correlate the concentration distributions of compounds in the same cumulative assessment group.

A basic distinction is between using the empirical concentration data (**empirical model**) or fitting a statistical model to the concentration data (**parametric model**).

6.1 Sample-based or tabulated concentration data, focal food data, total diet studies

The recommended way to enter concentration data in MCRA is to report a list of analysed **samples** for all measured foods specifying the **analytical method**, and to report separately all positive values found in specific samples. For each sample the date and location of sampling can be specified. For each analytical method it should be specified which compounds it can analyse. MCRA follows exactly the EFSA Guidance (pp.56-58) for cumulative assessments. This implies a sample-based approach, *i.e.* the EFSA method considers a collection of samples on which all substances are measured (there may be missing values in which case there is an imputation procedure based on the measured residues in the same sample collection).

Alternatively, concentrations can be specified as **tabulated concentrations**, specifying for each food and compound (and possibly sampling location and period) a list of concentration values together with their absolute frequency. This way of entering data has the disadvantage that the co-occurrence of compounds is not recorded (which could be a problem in cumulative assessments).

In some assessments one food-as-measured is of special interest, and concentration data for this **focal food** have to be combined with concentration data for other foods from a background (*e.g.* monitoring) database. Examples are the approval scenario, MRL setting scenario, authorisation scenario and high-residue event scenario as identified in EFSA (2012). MCRA has an option to replace sample-based food concentration data for a specified focal food with concentration data in a separate data table. Note that the entire collection of samples from the Concentration data group (typically monitoring data) are replaced by the collection of samples from the focal food data (typically field trial data). So in the case of cumulative assessments the assumption is that the focal food data are representative for the population of the respective commodity.
For **Total Diet Studies** (TDS), the food-as-measured is represented by a TDS food sample which is composed of food subsamples. In a TDS study, the selection of foods that in the end constitute the TDS food samples, is based on national consumption surveys. Ideally, the selection covers 90-95% of the foods found in the dietary survey. The selected foods that represent the diet for a specific target population are collected, prepared as consumed and related foods pooled prior to analysis. The exposure is based on whole diets rather than on raw agricultural commodities and results in a more realistic measure of exposure to substances.

6.2 Limit of Reporting and non-detects

A complication in concentration modelling occurs if results are reported as being below a limit. Different names may be used for such a limit, *e.g.* limit of detection or limit of quantification. For the purpose of exposure assessment it is only relevant whether results are reported as a positive value or as a non-detect, therefore we refer to any limit as the **Limit Of Reporting** (LOR), and any result reported as '<LOR' is termed a **non-detect**. The value of LOR should always be known for the particular analytical method used.

Non-detects are a very common phenomenon for some classes of substances like pesticides. Non-detects can be handled by replacing them with a given value (**imputation**), or by incorporating them

in a parametric model. In the imputation approach, non-detects (values reported less than LOR) can be replaced in simulations by any value between 0 and LOR * *constant*.

6.3 Agricultural use and food origin

Sometimes it can be assumed that the majority of non-detects represent true zeroes. For example, this will be the case when the agricultural use of a pesticide on many crops is not allowed, or when no agricultural use exists of this pesticide together with another pesticide for which positive values have been found in the same food sample. In such cases a realistic model is to consider the concentration distribution as **a mixture** of a distribution of positive values (some of which may be non-detects) and a spike of true zeroes.

In MCRA it is possible to specify **agricultural uses** (one pesticide, or a group of pesticides, in combination with a food), whether they are legal or not (legal is assumed if not specified), and if legal the relative frequency of use (100% is assumed if not specified). Agricultural uses can be given for specific locations (*e.g.* countries) and time periods. This can be done at multiple hierarchical levels, or even in general (not specifying location and period in the data is interpreted to indicate general default values).

Data specifying percentages of **food origin** may be used in the simulations to assign a location and/or period of origin to each simulated food sample, *e.g.* 40% of bananas in the United Kingdom in 2009 originated from Ghana, therefore a simulated banana sample will be considered to be from Ghana with 40% probability. Then data on agricultural use for this specified origin can be used to simulate a correct proportion of true zeroes for the compounds considered, leaving the remaining proportion of samples to be handled either by imputing a value, or sampling from a parametric model.

If agricultural use data are available, then these data will specify the expected minimal proportion of true zeroes (p_0) in the concentration dataset for each food-compound combination. Agricultural use information can be specific for *e.g.* the origin location of samples.The algorithm used is to find the origin of all samples in the concentration data set, and average the relevant p_0 values over all samples. For a single origin, $p_0 = 1 - \sum_{j \in use} p_{AU,j}$, where $p_{AU,j}$ is the fraction of agricultural use for the specific combination j of compounds on a crop (agricultural use), and *use* is the set of agricultural uses in the crop where the relevant compound is included. If p_0 is lower than the proportion of non-detects (p_{ND}), then the surplus proportion of simulated sample concentrations (p_0 - p_{ND}) is handled according to the specifications made for non-detects (either imputation or sampling from a parametric distribution). If p_0 is higher than the proportion of non-detects (p_{ND}), then priority is given to the actual concentration data, and the value of p_0 is adjusted to this lower value.

6.4 Available concentration models

6.4.1 Concentration Empirical model

In the empirical (non-parametric) approach, simulated concentrations are sampled at random from the available data. Non-detects are handled by imputation. If agricultural use data have been used a proportion p_0 / p_{ND} of non-detects is set as 0.

6.4.2 Concentration NonDetectSpike-LogNormal model

A binomial model is used to estimate the proportion p of positive values (detects). This is just the proportion observed in the data (unless agricultural use data have been used to set a proportion of true zeroes).
A lognormal model is fitted to the positive data. This provides estimates of μ and σ, which are the mean and standard deviation of the natural logarithm of the concentration.
Simulated concentrations are a non-detect with probability p_{ND} = 1-p or a value sampled from the fitted lognormal distribution with probability p. Non-detects are handled by imputation. If agricultural use data have been used a proportion p_0 / p_{ND} of non-detects is set as 0.

Minimum requirements: at least two positive concentration values. See also paragraph 15.1 and further.

6.4.3 Concentration NonDetectSpike-TruncatedLogNormal model

A binomial model is used to estimate the proportion p of positive values (detects). This is just the proportion observed in the data (unless agricultural use data have been used to set a proportion of true zeroes in which case p is calculated on the remaining proportion).
A truncated lognormal model, with LOR as the truncation limit, is fitted to the positive data, leading to estimates of μ and σ, which are the mean and standard deviation of the natural logarithm of the concentration.
Simulated concentrations are a non-detect with probability $p_{ND} = 1-p$ or a value sampled from the fitted lognormal distribution with probability p. Non-detects are handled by imputation. If agricultural use data have been used a proportion p_0 / p_{ND} of non-detects is set as 0.
Minimum requirements: at least two positive concentration values, all non-detects must have one LOR value. See also paragraph 15.1 and further.

6.4.4 Concentration CensoredLogNormal model

A censored lognormal model, with LOR as the censoring limit, is fitted to the data, both positives and non-detects. This provides estimates of μ and σ, which are the mean and standard deviation of the natural logarithm of the concentration.
If agricultural use data are being used, then a proportion p_0 / p_{ND} of non-detects will be excluded, where p_0 will be lowered to p_{ND} if it would be higher.
Simulated concentrations are sampled from the fitted lognormal distribution. If agricultural use data have been used, simulated concentrations are 0 with probability p_0 or are sampled from the fitted lognormal distribution with probability $1-p_0$.
Minimum requirements: at least one positive concentration value. See also paragraph 15.1 and further.

6.4.5 Concentration ZeroSpike-CensoredLognormal model

A mixture distribution of a spike of true zeroes and a censored lognormal model, with LOR as the censoring limit, is fitted to the data (non-detetcs and positives). This provides estimates of p_0, which is the proportion of true zeroes, and of μ and σ, which are the mean and standard deviation of the natural logarithm of the concentration.
Simulated concentrations are 0 with probability p_0, and are sampled from the fitted lognormal distribution with probability $1-p_0$.
Minimum requirements: at least one positive concentration value, no agricultural use data for the food-compound combination (which directly specify p_0, therefore it should not be estimated from the data). See also paragraph 15.1 and further.

6.4.6 Concentration Bayesian ZeroSpike-CensoredLognormal model

A mixture distribution of a spike of true zeroes and a lognormal model is fitted to the data (non-detects and positives). The model uses a data augmentation algorithm to account for the limited amount of information provided by the <LOR data when inferring the proportion of true zeroes, p_0, and the mean and standard deviation of the natural logarithm of the concentration, μ and σ respectively. The uncertainty algorithm is external to MCRA. Example methods are Paulo *et al.* (2005). Simulated concentrations are 0 with probability p_0, or sampled from the fitted lognormal distribution with probability $1-p_0$.
Minimum requirements: at least one positive concentration value. Agricultural use data for the food-compound combination can be used to specify a prior distribution for p_0.

6.4.7 Concentration NonDetectSpike-MRL model

This model simply takes values specified in an input table as Maximum Residue Limit (MRL) to be used for the proportion of positive values in the concentration dataset, and can be used to force the use of a pessimistic value.

6.4.8 Concentration Summary Statistics model

For this model, no individual measurements on raw agricultural commodities are needed. The final estimates of μ and σ are simply provided or pooled or estimated using *e.g.* a coefficient of variation. Specific use of this model is found in Total Diet Studies. In general, each TDS food sample is prepared only once, yielding one measurement for a TDS food sample. The variability of the underlying distribution is unknown. However, a rough guess can be made using the *e.g.* coefficient of variation of the subsamples (in general raw agricultural commodities) that compose the TDS food sample. The estimated standard deviation is calculated as a pooled estimate using the coefficient of variation and the count of each subsample in the TDS food.

6.5 Choice of concentration models

For each food-compound combination there is a choice of model. One option is to choose the same model for all food-compound combinations. For example,

- The **EFSA (2012) basic optimistic model** is to use empirical sampling, and to impute non-detects by 0.
- The **EFSA (2012) basic pessimistic model** is to fit a NonDetectSpike-LogNormal model, and to impute non-detects by LOR.

Parametric models have minimum data requirements, or the fitting of the model may fail for technical reasons in case of a severe lack-of-fit. MCRA uses a fall-back scheme to a simpler model if a model cannot be fitted:

- If the ZeroSpike-CensoredLognormal model fails, then try the CensoredLognormal model.
- If the CensoredLognormal model fails, then try the NonDetectSpike-Lognormal model.
- If the NonDetectSpike-TruncatedLognormal model fails, then try the NonDetectSpike-Lognormal model.
- If the NonDetectSpike-Lognormal model fails, then use the Empirical model. However, for the EFSA basic pessimistic model the first fall-back option is NonDetectSpike-MRL.

MCRA will try to fit the specified default model to all combinations. If a model cannot be fitted for technical reasons, a simpler model will be fitted according to the scheme above.
In the EFSA basic pessimistic model, the cumulative equivalent concentration values are fitted using the NonDetectSpike-Lognormal model.

A second option is to choose a model for each specific combination of food and compound. For this, MCRA provides a graphical overview of available concentration data for each combination. As a starting point a default model can be chosen and MCRA will try to fit this model to all combinations. If a model cannot be fitted for technical reasons, a simpler model will be fitted according to the scheme above. After this, all available models can be fitted for any desired food-compound combination, and a model can be selected for use in the simulations.

6.6 Maximum rank correlation

In a cumulative exposure assessment, MCRA offers the possibility to do a sensitivity analysis by ranking the sampled concentrations within co-occurrence patterns for substances within the same cumulative assessment group.
For *e.g.* three substances, $2^3 = 8$ patterns of co-occurrence exist. The relevant concentration patterns may be represented by (*1,1,0*), (*1,0,1*), (*0,1,1*), (*1,1,1*), where indicator *1* and *0* denote a vector of positive or zero concentrations, respectively, for the 1^{th}, 2^{nd} en 3^{rd} substance. Within each pattern, concentrations per substance are ordered, such that high concentrations are more likely to occur together on the same food. The option Maximum rank correlation is not available when *screening and the two-step approach for large CAGs* (see paragraph 9.1) is applied.

7 Acute exposure assessment

In an acute exposure assessment, the short term exposure to a substance or group of substances is estimated. The interest is in the distribution of individual day exposures and derived statistics like the fraction of days that exceed an intake limit or acute reference dose (ARfD).

The basic model for the exposure to a compound in an acute exposure assessment is:

$$y_{ij} = \frac{\sum_{k=1}^{p} x_{ijk}\ c_{ijk}}{bw_i}$$

where y_{ij} is the intake by individual i on day j (in microgram substance per kg body weight), x_{ijk} is the consumption by individual i on day j of food k (in g), c_{ijk} is the concentration of that substance in food k eaten by individual i on day j (in mg/kg), and bw_i is the body weight of individual i (in kg). Finally, p is the number of foods accounted for in the model. Within parenthesis, the default unit definitions are assumed, but decimal multiples or submultiples of units are easily specified using the relevant tables.

In the exposure assessment, individual days enter the Monte Carlo sample using the inverse of the sampling weights w_i when the number of MC iterations is > 0 (see table **Individuals**, field *SamplingWeight*).

7.1 Unit variability

In the basic model for an acute exposure assessment, it is assumed that the concentration of the substance displays the variation of residues between units in the marketplace. In general, both monitoring data and controlled field trial data are obtained using composite samples and, as a result, some of the unit to unit variation is averaged out. The model for unit variability aims to adjust the composite sample mean such that sampled concentrations represent the originally unit to unit variation of the units in the compositie sample.

MCRA offers three distributions to sample from. The beta distribution simulates values for a unit in the composite sample and requires knowledge of the number of units in a composite sample and of the variability between units. The lognormal distribution simulates values for a new unit in the batch and requires only knowledge of the variability between units. The bernoulli distribution is considered as a limiting case of the beta distribution when knowledge of the variability between units is lacking and only the number of units in the composite sample is known. For the beta and lognormal distribution, estimates of unit variability are realistic (no censoring at the value of the monitoring residue) or conservative (unit values are left-censored at the value of the monitoring residue). For the lognormal distribution, sampled concentrations have no upper limit whereas for the beta distribution, sampled concentration values for a unit are never higher than the monitoring residue * the number of units in the composite sample.
Variability between units is specified using a variability factor v (defined as 97.5th percentile divided by mean) or a coefficient of variation cv (standard deviation divided by mean). Following FAO/WHO recommendations, for small crops (unit weight < 25 g), a default variability factor $v = 1$ is used, for large crops (unit weight ≥ 25 g), a variability factor $v = 5$ is used. For foods which are processed in large batches, *e.g.* juicing, marmalade/jam, sauce/puree, bulking/blending a variability factor $v = 1$ is proposed. See also paragraph 15.2 .

7.1.1 Estimation of intake values using the concept of unit variability

- For each iteration i in the MC-simulation, obtain for each food k a simulated intake x_{ik} , and a simulated composite sample concentration cm_{ik} .
- Calculate the number of unit intakes nux_{ik} in x_{ik} (round upwards) and set weights w_{ikl} equal to unit weight wu_k, except for the last partial intake, which has weight $w_{ikl} = x_{ik} - (nux_{ik} - 1)wu_k$.

- For the beta or bernoulli distribution: draw nux_{ik} simulated values bc_{ikl} from a beta or bernoulli distribution. Calculate concentration values as $c_{ikl} = bc_{ikl} * cm_{ik,\,max} = bc_{ikl} * cm_{ik} * nu_k = svf_{ikl} * cm_{ik}$, where svf_{ikl} is the stochastic variability factor for this simulated unit, i.e. the ratio between simulated concentration c_{ikl} and the simulated composite sample concentration cm_{ik}. Sum to obtain the simulated concentration in the consumed portion:

$$c_{ik} = \sum_{l=1}^{nux_{ik}} w_{ikl} c_{ikl} \Big/ x_{ik}$$

- For the lognormal distribution: draw nux_{ik} simulated logconcentration values lc_{ikl} from a normal distribution with (optional) a biased mean $\mu = \ln(cm_{ik})$ or (default) unbiased mean $\mu = \ln(cm_{ik})$ - ½ σ^2 and standard deviation σ. Calculate concentration values as $c_{ikl} = \exp(lc_{ikl}) = svf_{ikl} * cm_{ik}$, where svf_{ikl} is the stochastic variability factor for this simulated unit, i.e. the ratio between simulated concentration c_{ikl} and the simulated composite sample concentration cm_{ik}. Back transform and sum to obtain the simulated concentration in the consumed portion:

$$c_{ik} = \sum_{l=1}^{nux_{ik}} w_{ikl} c_{ikl} \Big/ x_{ik}$$

For cumulative exposure assessments, a sensitivity analysis may be performed by specifying a full correlation between concentrations from different substances on the same unit. As a result, high (or low) concentrations from different substances occur together on the same unit. In MCRA, for each unit the random sequence is repeatedly used to generate concentration values for all substances.

7.2 Processing

Concentrations in the consumed food may be different from concentrations in the food as measured in monitoring programs (typically raw food) due to processing, such as peeling, washing, cooking etc. In general, we assume the model:

$$cpos_{ijk} = f_k \cdot c_{ijk}$$

where c_{ijk} is the concentration in the raw food, and where f_k is a factor for a specific combination k of food as measured and processing.
For fixed processing factors, $f_k = f_{k,upper}$, where $f_{k,upper}$ is typically some sort of central value from an experimental study. For distribution based processing factors, f_k is sampled from a normal distribution. The first two moments are defined through the specification of $f_{k,nominal}$ and $f_{k,upper}$.
See also paragraph 15.3 .
The processing correction factor is introduced to correct for double counting the effects of chemical alteration and weight change *e.g.* for a dried food with a consumption of 100 gram which is translated to 300 gram raw agricultural commodity, the correction factor is 3.

7.3 Acute exposure as a function of covariates

In MCRA, acute exposure values may be modelled as a function of covariates. The modelling is restricted to main effects models, additive models and/or interaction models including one continuous and/or one discrete covariate at the same time. Continuous covariates may be modelled by a polynomial function. The order of the polynomial allows for nonlinear effects, starting from a linear relation (one degree of freedom), quadratic (two degrees of freedom) up to higher order polynomials. To determine the optimal number of degrees of freedom, likelihood ratio tests are used to compare the fit of two adjacent models. MCRA can automatically select the best fitted model, starting from a full model (backward selection) or starting from a empty model (forward selection). To decide on the effect of a qualitative covariate, fit alternative models and perform a likelihood ratio test using the log-likelihoods as shown in the output (Mood *et al.,* 1974; Snedecor *et al*., 1980)
For an acute exposure assessment with covariates, two models are available, the betabinomial-normal (BBN) model and the logisticnormal-normal (LNN0) model. See also paragraph 15.6 .

8 Chronic exposure assessment

8.1 Introduction

In a chronic exposure assessment, the main interest goes to the fraction of individuals with a usual intake per day higher than an intake limit *e.g.* the acceptable daily intake (ADI). Usual intake is defined as the long-run average of daily exposure to a substance or group of substances by an individual. Usually, for an individual, dietary recall data are available on 2 (or more) consecutive days. We assume an equal number of days for each individual, unless specified differently in table **Individuals**.

For a chronic exposure assessment the available data are used to calculate exposures per person-day (daily intake):

$$y_{ij} = \frac{\sum_{k=1}^{p} x_{ijk}\ c_k}{bw_i}$$

where y_{ij}, x_{ijk} and bw_i are defined as before but now concentrations of the substance found in food k enter the model as the estimated mean value c_k .

Using the person-day intakes MCRA uses one of the following models to calculate the distribution of usual intake at the person level:

1) the observed individual means (OIM) model;
2) the logisticnormal-normal model, in a full version that includes the estimation of correlation between intake frequency and amount (LNN), and in a simpler version without this estimation (LNN0);
3) the betabinomial-normal (BBN) model;
4) the discrete/semi-parametric model known as the Iowa State University Foods (ISUF) model. For this model, an equal number of days per individual is assumed.

In modelling usual intake, two situations can be distinguished. Foods are consumed on a daily basis or foods are episodically consumed. For the logisticnormal-normal model and the betabinomial-normal model, the latter requires fitting of a two-part model, 1) a model for the frequency of consumption, and 2) a model for the intake amount on consumption days. In the final step, both models are integrated in order to obtain the usual intake distribution. For daily consumed foods, fitting of the frequency of consumption is skipped and modelling resorts to fitting the model to daily intake amounts only. Note that the distinction between BNN, LNN and LNN0 disappears and modelling will give equivalent results.

8.2 Model based and model assisted

Following Kipnis *et al.* (2009), some of the models available in MCRA are extended to predict individual usual intakes. This model assisted approach has been added to BBN and LNN0 and may be a useful extension in evaluating the relationship between health outcomes and individual usual intakes of foods. In contrast, the estimation of the usual intake distribution in the general population is called the model based approach. Summarizing, we get Table 9:

model based approach	model assisted approach
	observed individual means (OIM)
betabinomial-normal (BBN)	betabinomial-normal (BBN)
logisticnormal-normal without correlation (LNN0)	logisticnormal-normal without correlation (LNN0)
logisticnormal-normal with correlation (LNN)	
Iowa State University Foods (ISUF)	

Table 9: model based and assisted approach available for chronic exposure models

The model assisted approach builds on the proposal of Kipnis *et al*. (2009), but is modified to ensure that the population mean and variance are better represented. The method is based on shrinkage of the observed individual means (modified BLUP estimates) and shrinkage of the observed intake frequencies. The model-assisted usual intake distribution applies to the population for which the consumption data are representative, and automatically integrates over any covariates present in the model. Model-assisted intakes are not yet available for LNN, and when a covariable is modelled by a spline function of degree higher than 1.
In case of a model with covariates the usual intake is presented in graphs and tables as a function of the covariates (conditional usual intake distributions).

8.2.1 Observed individual means (OIM)

The usual intake distribution for a population is estimated with the empirical distribution of individual means. Each mean is the average of all single-day intakes for an individual. The mean value for an individual still contains a considerable amount of within-individual variation. As a consequence, the distribution of within-individual means has larger variance than the true usual intake distribution and estimates using the OIM-method are biased, leading to a too high estimate of the fraction of the population with a usual intake above some standard.
Despite its known tendency to over-estimate high-tail exposures, the OIM method is the method to be used in EFSA (2012) basic assessments.

8.2.2 Betabinomial-Normal model (BBN)

The Betabinomial-Normal (BBN) model for chronic risk assessment is described in de Boer *et al.* (2009), including its near-dentity to the STEM-II model presented in Slob (2006). See also paragraph 15.5.2.1

8.2.3 Logisticnormal-Normal model (LNN with and without correlation)

An alternative to the betabinomial modelling of intake frequencies in BBN model is modelling these frequencies by a logistic normal distribution. In notation, for probability p:

$$\text{logit}(p) = \log(p/1\text{-}p) = \mu_i + \underline{c}_i$$

where μ_i represents the person specific fixed effect model and $\underline{c}_i$ represent person specific random effects with estimated variance component $\sigma^2_{between}$.

This model is referred to as the LogisticNormal-Normal (LNN) model. The full LNN model includes the estimation of a correlation between intake frequency and intake amount. This is similar to the NCI model described in Tooze et al. (2006). See also paragraph 15.5.2.3 and 15.5.2.2 .
A simple and computationally less demanding version of the LNN method which does not estimate the correlation between frequency and amount is termed LNN0, where the '0' indicates the absence of correlation. The models are fitted by maximum likelihood, employing Gauss-Hermite integration. See also paragraph 15.5.3 .

For chronic models amounts are usually transformed before the statistical model is fit. The power transformation, given by y^p, has been replaced by the equivalent Box-Cox transformation. The Box-Cox transformation is a linear function of the power transformation, given by $(y^p\text{-}1)/p$, and has a better numerical stability. Gauss-Hermite integration is used for back-transformation. See also paragraph 15.4 .

8.2.4 Discrete/semi-parametric model (ISUF)

Nusser *et al.* (1996) described how to assess chronic risks for data sets with positive intakes (a small fraction of zero intakes was allowed, but then replaced by a small positive value). The modeling allowed for heterogeneity of variance, *e.g.* the concept that some people are more variable than others

with respect to their consumption habits. However, a disadvantage of the method was the restricted use to contaminated foods which were consumed on an almost daily basis, *e.g.* dioxin in fish, meat or diary products. The estimation of usual intake from data sets with a substantial amount of zero intakes became feasible by modeling separately zero intake on part or all of the days via the estimation of intake probabilities as detailed in Nusser *et al.* (1997) and Dodd (1996). In MCRA, a discrete/semi-parametric model is implemented allowing for zero intake and heterogeneity of variance following the basic ideas of Nusser *et al.* (1996, 1997) and Dodd (1996). This implementation of the ISUF model for chronic risk assessment is fully described in de Boer *et al*. (2009).

8.3 Model-Then-Add

The traditional approach can be termed the Add-Then-Model approach, because adding over foods precedes the statistical modelling of usual exposure. MCRA 8.0 offers, as an advanced option, an alternative approach termed Model-Then-Add (van der Voet et al. 2014). In this approach the statistical model is applied to subsets of the diet (single foods or food groups), and then the resulting usual exposure distributions are added to obtain an overall usual exposure distribution. The advantage of such an approach is that separate foods or food groups may show a better fit to the normal distribution model as assumed in all common models for usual exposure (including MCRA's BBN and LNN models). That this principle can work in practice was shown in previous work (de Boer et al. 2009, Slob et al. 2010, Goedhart et al. 2012), and a simulation model was developed and implemented in MCRA 7.1 to show how multimodal distributions can arise from adding unimodal distributions of foods that are not always consumed (Slob et al. 2010, de Boer and van der Voet 2011). For specific cases involving separate modelling of dietary supplements and the rest of the diet, proposals have been made (Verkaik-Kloosterman et al. 2011). However, a practical approach to apply the Model-Then-Add approach to general cases of usual exposure estimation was still missing. Therefore a module in MCRA 8.0 was developed to implement such an approach based on a visual inspection of a preliminary estimate of the usual exposure distribution using the Observed Individual Means (OIM) method.

The Model step
At this stage of development the division of foods into a number of food groups is performed in an interactive process, where the MCRA user is presented with a visual display (see example in *Figure 4*) which shows:

1. The OIM distribution represented as a histogram, where each bar shows the frequency of exposures (summed over foods) of individuals in a certain exposure interval; each bar is subdivided according to the contributions of the individual foods contributing to those exposures (left panel of Figure 2).
2. The contributions graph, where each of the bars in the OIM histogram is expanded to 100%. This graph allows a better view of the lower bars in the OIM histogram (right panel of Figure 2).

The visual display identifies the nine foods that contribute most to the total exposure; the remaining foods are grouped in a rest category to avoid identification problems because of too many colours.

The user has now the possibility to select one or more foods and to split these from the main exposure histogram. A separate graph shows the OIM distribution for the split-off food or food group. The graphs for the main group (now called the rest group) are adapted to show the OIM distribution and the contributions for the remaining foods only (see Figure 5 upper two panels). This splitting-off can be repeated several times for other foods or food groups. In this way the user can try to obtain foods or food groups that show unimodal OIM distributions. If the result is not what is intended, a food or food group can be added again to the rest group. Per split-off food or food group the usual

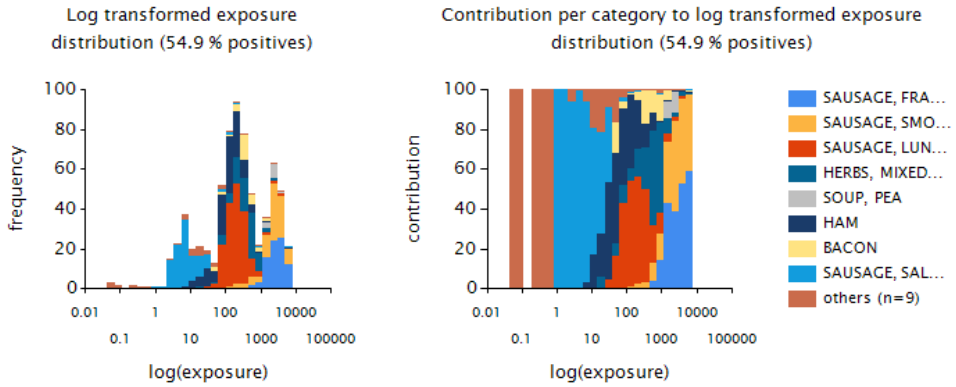


Figure 4. Left panel: OIM usual exposure distribution to smoke flavours via the different foods (excluding the zero exposures) in young children; right panel: Contribution of foods to exposures within each bar of the OIM distribution histogram.

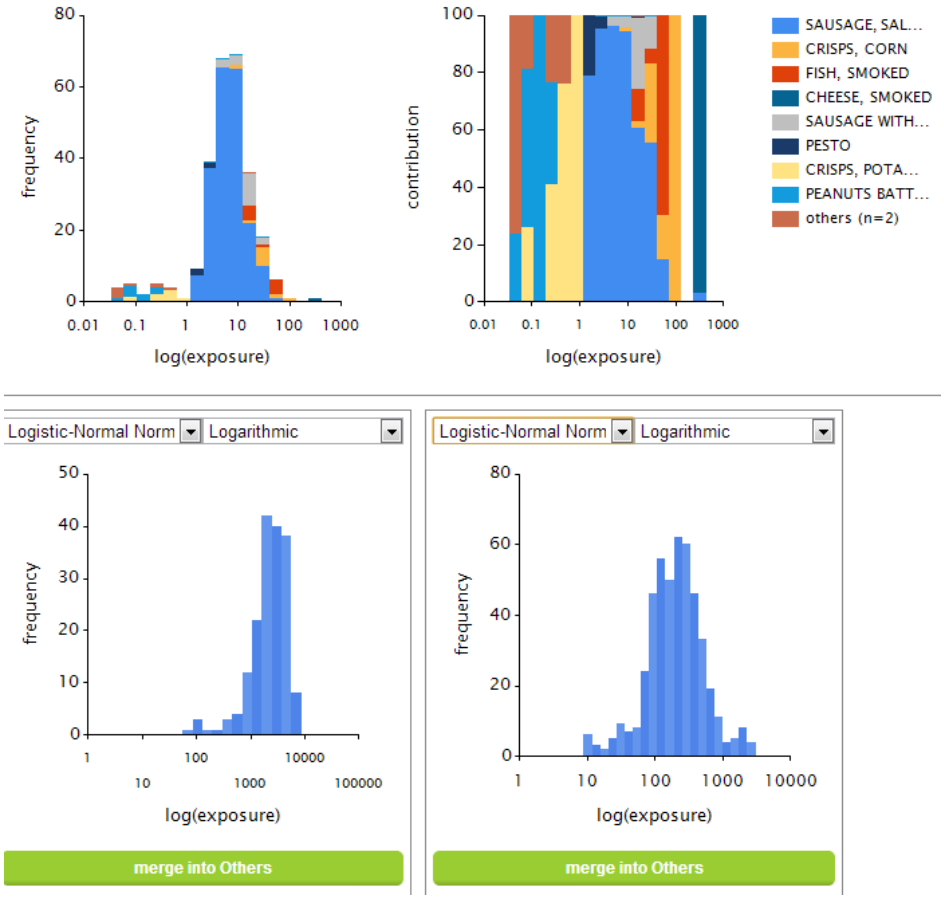


Figure 5. Result of a selection into two split-off groups and a rest group. The graph bottom left represents the exposure via a food group containing 'Sausage, frankfurter' and 'Sausage, smoked cooked'. The graph bottom right represents the exposure via a food group containing 'Sausage, luncheon meat', Herbs, mixed, main brands, not prepared', 'Soup, pea', 'Ham', and 'Bacon'. The top graph represents the exposure via the rest group.

exposure can be modelled using either BBN or LNN, with a logarithmic or power transformation. The rest group will always be modelled as OIM. It is possible that the rest group is empty, when the total exposure via the different split-off foods and /or food groups is modelled with BBN or LNN.

After a split-off selection has been made, the OIM distribution is summarised in terms of the defined grouping (Figure 6), and the usual exposure distribution per split-off food or food group is fitted according to the chosen modelling settings.

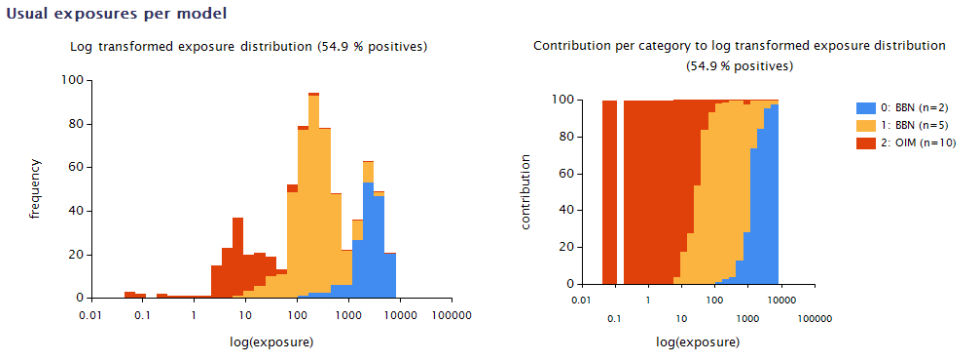


Figure 6. OIM usual exposure distribution showing the contributions from the three food groups as constructed in Figure 5.

The Add step

Consumptions of foods may be correlated. In the traditional Add-Then-Model approach the Add step automatically reflects any correlations that are apparent in the consumptions at the individual-day or individual level. In the Model-Then-Add approach the estimated usual exposure distributions for different foods or food groups have to be combined to assess the total usual exposure. Two approaches are available for this:

1. Model-based approach: adds independent samples from the usual exposure distribution per food or food group, ignoring any correlations in consumption;
2. Model-assisted approach: adds the model-assisted, person-specific usual exposure estimates per food or food group, taking correlations in consumptions into account.

Before the addition is made, in the model-based approach, model-based estimates of the usual exposure amounts distribution per food or food group are back-transformed values from the normal distribution assumed for transformed amounts per food or food group, and the model-based frequency distribution is sampled to decide if a simulated individual has exposure via the food or food group or not. Model-assisted estimates of the usual exposure distribution are back-transformed values from a shrunken version of the transformed OIM distribution, also done per food or food group, where the shrinkage factor is based on the variance components estimated using the linear mixed model for amounts at the transformed scale (van Klaveren et al. 2012). For individuals with no observed exposure (OIM=0) no model-assisted estimate of usual exposure can be made and a model-based replacement is used.

The model-based approach was investigated in Slob et al. (2010) and performed surprisingly well, even if correlations in consumptions of foods were present. The model-assisted approach adds exposures at the individual level, and therefore retains effects of correlations between foods in the usual exposure distribution.

MCRA 8.0 calculates both the model-based and model-assisted usual intake distributions (see Figure 7).

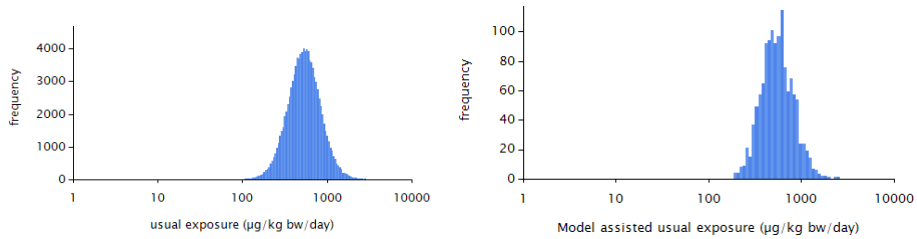


Figure 7. Model-based (left) and model-assisted (right) estimated usual exposure distributions (excluding the zero exposures).

8.4 Chronic exposure as a function of covariates

The intake frequency and transformed intake amounts may be modelled as a function of covariates. MCRA allows one covariable and/or one cofactor.

For frequencies:

cofactor: $\text{logit}(\pi) = \beta_{0l}$

covariable: $\text{logit}(\pi) = \beta_0 + \beta_1 f(x_1; df)$

both: $\text{logit}(\pi) = \beta_{0l} + \beta_1 f(x_1; df)$

interaction: $\text{logit}(\pi) = \beta_{0l} + \beta_{1l} f(x_1; df)$

For amounts:

cofactor: $\text{transf}(y_{ij}) = \beta_{0l} + c_i + u_{ij}$

covariable: $\text{transf}(y_{ij}) = \beta_0 + \beta_1 f(x_1; df) + c_i + u_{ij}$

both: $\text{transf}(y_{ij}) = \beta_{0l} + \beta_1 f(x_1; df) + c_i + u_{ij}$

interaction: $\text{transf}(y_{ij}) = \beta_{0l} + \beta_{1l} f(x_1; df) + c_i + u_{ij}$

where $l=1...L$ and L is the number of levels of the cofactor, y_{ij}, the intake amount, x_1 is the covariable, f is a polynomial function with the degrees of freedom df, c_i and u_{ij} are the individual effect and interaction effect, respectively. These effects are assumed to be normally distributed N(0, $\sigma^2_{between}$) resp. N(0, σ^2_{within}). The degree of the function is determined by backward or forward selection.

In the output, the usual intake is displayed for a specified number of values of the covariable and/or the levels of the cofactor.

8.5 Usual intake estimation when there are no replicated data

When a chronic model LNN, LNN0 or BBN is applied to consumption data with just one day per person, MCRA asks for the input of a variance ratio for the amount model and a dispersion factor for the frequency model. These values can for example be taken from the output of similar exposure assessments of datasets with multiple days per person.

9 Cumulative exposure assessment

In MCRA cumulative assessments can be performed if a Cumulative Assessment Group (CAG) is specified in terms of a list of compounds and corresponding **Relative Potency Factors** (RPFs) relating to a particular health effect. Cumulative exposure is expressed as equivalents of one of the compounds in the CAG, termed the reference or index compound.

The occurrence and concentrations of compounds in the same samples may be correlated, which is of importance for acute exposure assessments (Note that chronic assessments only use mean concentration values). Theoretically, this could be modelled and fitted to datasets. However, in practical applications (regarding pesticide residues) the number of positive values is commonly too low to allow such detailed modelling.

Ideally all samples have been measured for all compounds in the CAG (although part or all of the results may be non-detects). However, some samples may have missing values (MVs) because not all compounds in the CAG were analysed: in this case values can be imputed for the MVs. The imputation value may be zero, or a positive number drawn from a distribution.

Imputing MVs with 0 is correct if it is assumed that measurements have not been made because it is a priori known that the sample will not contain the compound. This approach is used in the **EFSA Guidance basic optimistic model** for acute exposure assessments.

In the **EFSA Guidance basic pessimistic model** for acute exposure assessments MVs in the data are imputed by sampling values at random from the distribution of concentrations as fitted on other samples. A pessimistic model is obtained by using the highest sampled values for imputing the MVs of samples which already are calculated to have a high RPF-weighted exposure based on the present values. MCRA implements the precise algorithm as documented in EFSA (2012).
In MCRA, both EFSA Guidance basic models are denoted as a sample based approach.

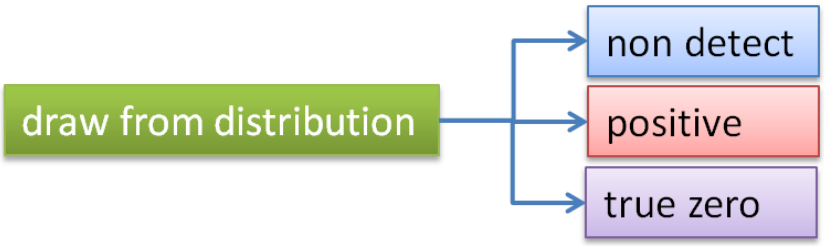


Figure 8: EFSA Guidance basic pessismistic model, sample based

MCRA also implements **custom methods** including those developed in the EU Acropolis project. In the **Acropolis model for cumulative exposure** for each food-compound combination a choice from the available models is made, which may be based on visual inspection. In fitting the chosen distribution MVs are ignored. In the basic uncorrelated Acropolis model MCRA simulates a cumulative exposure by drawing values for all compounds in the CAG from the respective distributions, and then calculating the RPF-weighted sum.

In addition, the distribution might be dependent on the presence of other compounds in this sample and/or on known patterns of co-occurrence of compounds. For example, it might be known that among triazole pesticides used on pineapple triadimefon is never combined with other compounds. In that case any pineapple sample on which triadimefon has been found cannot contain other compounds. And if no triadimefon has been found (either because it was not measured or because it was measured as '<LOR'), then there are two possibilities: either only triadimefon is present, or triadimefon is 0 but other compounds may be present.

To model this dependence additional data are needed about patterns of Agricultural use and their relative frequencies, which specify if and how often combinations of compounds can occur together in a sample. In MCRA this is modelled in a **multivariate ZeroSpike model**, which defines a mixture of 2^p components (p is the number of compounds). In each component 0, …, p specific concentrations are 0, and the remaining concentrations follow a univariate distribution as selected earlier. In practice, only a minority of the 2^p components of the mixture distribution will be specified, and all non-specified combinations will have probability 0%. If no data on Agricultural use are specified, MCRA assumes that all combinations are possible, so effectively the component where none of the concentrations are excluded has probability 100%, and the other 2^p-1 components of the mixture distribution have probability 0%.

When simulating cumulative exposures, MCRA will first draw an agricultural use pattern from the multinomial distribution specifying the 2^p (or less) components. This draw determines which compounds will have a zero concentration. For the remaining compounds draws are made from the respective univariate distributions as selected earlier.

Figure 9: Acropolis model, non sample based

9.1 Screening and the two-step approach for large CAGs

A full Monte Carlo analysis can be unwieldy for large cumulative assessment groups (CAGs) and/or large number of foods or concentration data. An algorithmic approach was developed to handle large CAGs. Two unique features of MCRA are: (1) contributions to the exposure results can be seen both in terms of food-as-eaten (e.g. white bread) and foods-as-measured (e.g. wheat), and (2) a drill-down can be made into the exact foods and compounds contributing for simulated individuals or individual-days in the upper tail. The number of combinations of simulation, compound, food-as-measured and food-as-eaten can be very large. To avoid memory problems with very large datasets, an additional optional modelling step, named Screening, was added to MCRA. Screening should be used if the data dimensions are too large for a direct analysis. Screening identifies risk drivers. A full analysis based on screened risk drivers will still retain all food/compound combinations in the exposure calculation, and will therefore produce exactly the same cumulative exposure distribution, and allow to see contributions of all compounds and all foods-as-measured. Details with respect to foods-as-eaten are however restricted to the risk drivers selected in the screening step.

The two-step approach consists of:

Step 1: Data screening and selection of risk drivers

Run a simple analysis for each potential source/compound combination (SCC). Here source means the combination of food-as-eaten and food-as-measured, for example apple in apple pie. The screening is based on this combination, and not just foods-as-measured, to avoid problems with potentially multi-modal consumption distributions as much as possible (see van der Voet et al. 2014). SCCs are also referred to as risk driver components.

The screening step in MCRA implements a simple model that is applied to each SCC. The model calculates a percentile of interest in a distribution, consisting of a spike of zeroes (non-consumptions), and a mixture of two lognormal distributions for the exposure related to non-detects and positive concentrations, respectively. For more details see Appendix A.

SCCs (risk driver components) can be combined to measured source/compound combinations (MSCCs, risk drivers). For example APPLE/apple juice/captan and APPLE/apple pie/captan combine to APPLE/captan.
MCRA has an interface which identifies the Top-N SCCs (based on a chosen exposure percentile, e.g. p95) with an option to select N based on cumulative importance according to some criterion. Remark: Screening is performed before concentration modelling. Therefore there is no correction for processing factors at the screening stage.

Step2: Full MC analysis

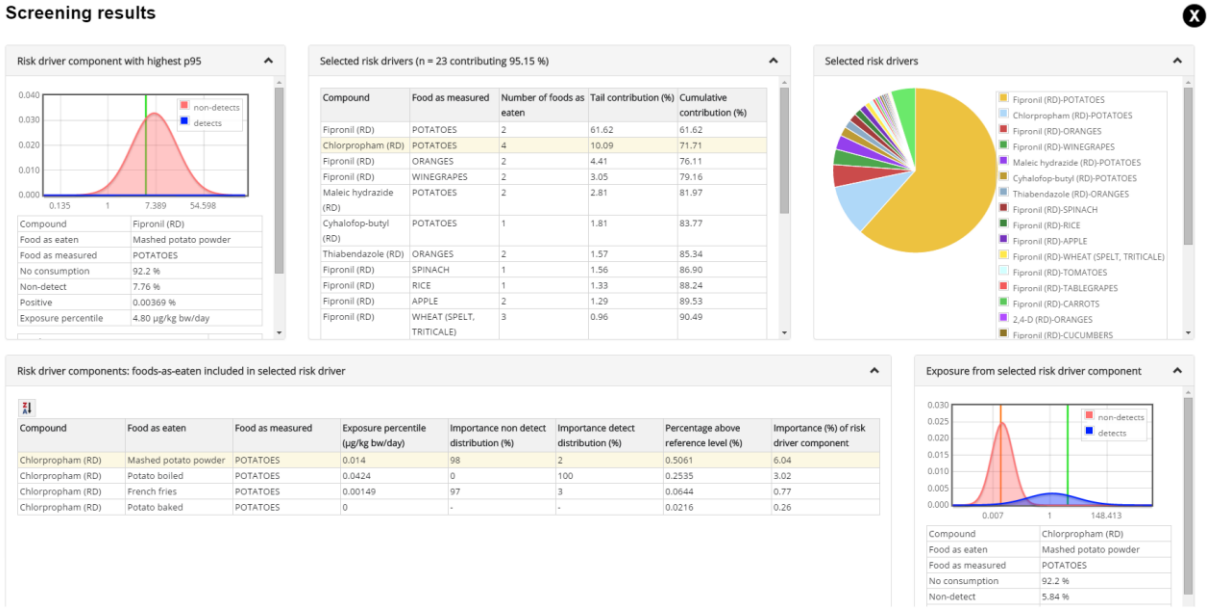
Perform the standard MC to all combinations of compounds and foods, but restrict the stored information regarding foods-as-eaten to the SCCs selected in step 1.

The screening method requires to specify:

- Which percentile to consider for each single Source-Compound Combination (SCC, potential risk driver component) (default p95)
- Which percentage of all exposures (according to the screening model) should be covered by the selected set of SCCs (default 95%)
- How to impute non-detect concentrations (c < LOR) in the screening step. The choice of a factor 0 (default) represents optimistic imputation, the choice of a factor 1 represents a pessimistic imputation. It may be noted that a factor 1 (pessimistic imputation) may select many SCCs (risk driver components) with relatively high LORs and high RPFs, but with only nondetect measurements. Choosing a lower fraction, e.g. 0.25 can be useful if a more realistic method is sought.

Based on limited experience with the EFSA test data, useful settings of these three screening parameters were found to be (95, 95, 0) in preparation for an EFSA optimistic run, and (50, 95, 0.25) in preparation for an EFSA pessimistic run.

The screening can be performed interactively (click green button 'Screen exposures') , and the results will appear in a pop-up window:



This screen shows in the upper left corner the screening distribution for the SCC with the overall highest percentile of interest (indicated by the green line). In the upper middle panel and the pie chart the SCCs are grouped into risk drivers (MSCCs) in decreasing order of contributing to the exposures higher than this reference exposure value. Clicking on a risk driver in the upper middle panel allows to zoom in on the risk driver components in the lower part of the screen.

In a full run the screening approach can also be performed, and the results will be used in the second step to allow a full analysis for large CAGs.

The details of the statistical models for screening are described in Appendix 15.7 .

9.2 Co-exposure

Co-exposure of compounds is defined as the pattern of compounds occurring together on a single individual day. Co-exposure can enter the risk assessment through the use of mixtures of substances on a single food or by combining different food sources on a single day (through consumption). In MCRA, an overview is given of the numbers of compounds occurring together without specifying the specific compound combinations. In addition, some specific output is displayed for mixtures with the highest number of compounds occuring together and a summary of the most frequent mixtures of compounds occuring together in the analyis. This is both done for the total exposure distribution and the upper tail of the exposure distribution (see also chapter 11).

10 Aggregate exposure assessment

Aggregate exposure assessments can be performed in MCRA when pre-calculated non-dietary exposure estimates are supplied. The user has flexibility to input either single deterministic non-dietary exposure or probabilistic exposure estimates (this section describes variability in non-dietary intakes and 14.1.3.1 describes uncertainty analysis of non-dietary intakes). When including multiple non-dietary surveys it is possible to supply some with uncertainty/variability and others without variability/uncertainty according to the requirements and data availability.

Non-dietary exposure estimates may be specified for multiple routes of exposure (dermal, oral and inhalation), for multiple compounds, and for multiple exposure sources. The multiple sources case will be relevant, for example, in modelling individuals taking part in various activities involving pesticide use or incidental exposures as a resident. Each non-dietary source is characterised in a particular user-selected or user-supplied 'non-dietary survey'. Exposures are included for each compound separately in the multi-compound case because the combination with RPFs takes place after the aggregation with dietary sources. By default, exposures from separate non-dietary surveys (sources) are considered to be independent events, but as explained below correlations between compounds and/or activity types per individual can be represented if generated prior to uploading to MCRA. Examples are presented as case studies in Kennedy *et al*. (2013a) and R code to generate these examples is available for general use.

10.1 Matched and unmatched aggregation

To create aggregate exposure estimates, the non-dietary exposures can either be matched to specific individuals in the food survey or they can be randomly assigned. For example, if both dietary and non-dietary information is available for a known population of individuals, the user may switch 'matching on' such that specific dietary and non-dietary estimates are aggregated for each individual in the food survey population. If matching is enabled, any non-dietary exposures that do not correspond to individuals from the food survey will be ignored (see Example 2 below), unless an individual is specified with id = General. In that case, the dietary individual should meet the criteria of the non-dietary survey, specified by the survey properties, to be assigned.

If the non-dietary data relates instead to a population in which individuals have no corresponding records in the food survey, the user may choose to randomly assign the non-dietary exposures to the individuals from the food survey. This is the unmatched case. When multiple nondietary surveys are available, the options with or without correlation are important (not relevant when matching is switched on). When option correlation is chosen, the exposure contributions of non-dietary individuals with identical id's in different surveys are combined and allocated to a randomly selected dietary individual. When option without correlation is chosen, the nondietary exposures of randomly selected individuals from different surveys are combined and allocated to a dietary individual. The user may also define demographic criteria for the assignment (for each source of non-dietary exposure) to indicate that those exposures are relevant only to a defined sub-population. Only those individuals in the food survey who meet the criteria of the non-dietary survey will be assigned non-dietary exposures from that source e.g. only males aged 18 to 65 (see Example 1 below). The simplest assessment consists of a single (deterministic) non-dietary exposure estimate which is assigned to all individuals in the food survey (idIndividual = General). This case, and more complex possibilities are illustrated below using hypothetical examples.

10.2 Internal and external doses

Non-dietary exposures may be represented as internal (absorbed) or external doses, with the interpretation inferred from the type of user input. Absorption factors are used to convert external doses to internal doses. If absorption factors are not specified, MCRA will assume that the supplied exposures are already internal doses (see Example 1 below). Absorption factors can vary between compounds and sources. For example, the absorption could be different when working with concentrates and dilutions. Thus, absorption factors may be specified for each compound and non-

dietary survey in an assessment (see Example 2 below). For aggregation, the dietary exposure estimates will be converted, if necessary, to internal doses using the oral absorption factor for dietary exposure (which may be supplied by the user).

The non-dietary exposures may be short term (acute) or longer term averages (chronic). The user must ensure that they supply appropriate non-dietary data for the type of exposure assessment they wish to conduct. For chronic assessments this means the non-dietary exposure is averaged over an appropriate time interval.

Example 1: Deterministic cumulative (multi-compound) non-dietary exposure input, adult male sub-population. Internal dose. Unmatched case.

idIndividual	idNonDietarySurvey	idCompound	Dermal	Oral	Inhalation
General	1	011003001	10	5	17
General	1	011003002	34	20	18
General	1	011003003	56	43	19

Table 10: NonDietaryExposures

idNonDietarySurvey	Description	Location	Date	NonDietaryIntakeUnit
1	BROWSE, acute, cumulative, operators	York	09/10/2012	µg/day

Table 11: NonDietarySurveys

idNonDietarySurvey	IndividualPropertyName	IndividualPropertyTextValue	IndividualPropertyDoubleValueMin	IndividualPropertyDoubleValueMax
1	Age		18	65
1	Gender	Male		

Table 12: NonDietarySurveyProperties

In this example, there are exposure values for multiple compounds in the NonDietaryExposures table and the user has provided a NonDietarySurveyProperties table which specifies that the non-dietary exposures given in survey number 1 relate to males aged 18 to 65.

When this assessment is performed, only those individuals whose property values fit the criteria in the NonDietarySurveyProperties table will receive the non-dietary exposures in survey 1. The use of idIndividual = General indicates that this is the default exposure. All individuals in the dietary survey meeting the criteria receive all exposures given in the 3 rows, corresponding to 3 compounds. The following should be noted:

- There should only ever be 1 General entry in the dietary exposures table per compound, survey combination.
- The property names and the values of any text properties must precisely match those given in the IndividualProperties and IndividualPropertyValues tables for this to work.
- The minimum and maximum values for numeric properties are both inclusive boundaries.

So in this example, all males aged 18 to 65 will receive the given exposures of all three compounds and the other members of the population will receive no non-dietary exposure. Note that example 1 describes the unmatched case.

Example 2: Variability (but no uncertainty) in cumulative non-dietary exposure input (matched to dietary survey individuals). External dose.

idIndividual	idNonDietarySurvey	idCompound	Dermal	Oral
5432	1	011003001	10	5
5432	1	011003002	33	21
5433	1	011003001	12	7
5433	1	011003002	34	23
5434	1	011003001	18	9
5434	1	011003002	35	25
5435	1	011003001	10	5
5435	1	011003002	33	21

Table 13: NonDietaryExposures

idNonDietarySurvey	Description	Location	Date	NonDietaryIntakeUnit
1	BROWSE, acute, cumulative, operators	York	09/10/2012	µg/day

Table 14: NonDietarySurveys

idNonDietary Survey	idCompound	DermalAbsorption Factor	OralAbsorption Factor	InhalationAbsorption Factor
1	011003001	0.1	1.0	1.0
1	011003002	0.1	1.0	1.0

Table 15: NonDietaryAbsorptionFactors

In this example the non-dietary exposures are external doses (because table NonDietaryAbsorption-Factors has been supplied) and the non-dietary exposures are being specified explicitly for individuals in the dietary population. Switch 'matching' on to allow exposure variability to be specified at the individual level. For the purposes of illustration, the population is extremely small, consisting of only four individuals. The values in the idIndividual column of the NonDietaryExposures match those in the Individuals table (dietary population).

It is not mandatory to specify exposures for every individual in the population. Those not included will simply receive a zero non-dietary exposure, unless there is also a default exposure value ('General' row(s) in the NonDietaryExposures table) and the individual matches the specified demographic criteria for the survey, as specified in the NonDietarySurveyProperties table. (In this example, a default exposure would apply to all individuals not listed in the NonDietaryExposures table because the NonDietarySurveyProperties table has not been used).

There is variability between individuals in this example, but no uncertainty in exposure. Note that these data could also be used with matching switched off. This would be the same as treating the idIndividual values as generic individuals, so that each pair of exposure lines would be assigned at random to individuals meeting the criteria.

Example 3: Variability (no uncertainty) in cumulative non-dietary exposure input (unmatched individuals). External dose.

idIndividual	idNonDietarySurvey	idCompound	Dermal	Oral	Inhalation
ND1	1	011003001	10	5	17
ND1	1	011003002	33	21	45
ND2	1	011003001	12	7	18
ND2	1	011003002	34	23	47
ND3	1	011003001	18	9	19
ND3	1	011003002	35	25	49
ND4	1	011003001	10	5	17
ND4	1	011003002	33	21	45

Table 16: NonDietaryExposures

idNonDietarySurvey	Description	Location	Date	NonDietaryIntakeUnit
1	BROWSE, acute, cumulative, operators	York	09/10/2012	µg/day

Table 17: NonDietarySurveys

idNonDietary Survey	Individual PropertyName	Individual PropertyText Value	IndividualProperty DoubleValueMin	IndividualProperty DoubleValueMax
1	Age		50	65
1	Gender	Female		

Table 18: NonDietarySurveyProperties

idNonDietary Survey	idCompound	DermalAbsorption Factor	OralAbsorption Factor	InhalationAbsorption Factor
1	011003001	0.1	1.0	1.0
1	011003002	0.1	1.0	1.0

Table 19: NonDietaryAbsorptionFactors

This example is similar to example 2, except that the values in the idIndividual column of the NonDietaryExposures do not match those in the Individuals table. In this instance, 'matching' would not be an option, and the non-dietary exposures would be randomly assigned to individuals who meet the criteria in the NonDietarySurveyProperties table. (In fact for the same result rather than changing the values in the idIndividual column the NonDietaryExposures table from the previous example may be used with matching switched off). Exposures in the NonDietaryExposures table will be recycled if the number of exposure rows is less than the number of dietary records with which to aggregate exposures.

Again, there is variability between individuals in this example, but no uncertainty in exposure.

By allowing generic idIndividual values in this way, correlations between different sources (within individual) can be accounted for even in the unmatched case. If the same idIndividual value is used in different surveys, then the corresponding exposure values will be kept together and assigned to an eligible individual as a combined exposure.

So for option matching switched of, it is relevant whether individuals are correlated or not.

In the following example, two nondietary surveys are available, per survey three individuals are specified.

idIndividual	idNonDietarySurvey	idCompound	Dermal	Oral	Inhalation
ND0	1	011003001	10	5	17
ND1	1	011003001	23	21	45
ND2	1	011003001	12	7	18
ND0	2	011003001	34	23	47
ND3	2	011003001	18	9	19
ND4	2	011003001	33	16	35

Table 20: matching switched of, with correlation or without.

When a correlation is applied, the nondietary exposure for individual ND0 from survey 1 and 2 are combined and allocated to a dietary individual. For individual ND1, ND2, ND3 and ND4 just a single nondietary exposure is found and allocated to a dietary individual.
When no correlation is applied, the exposure for individual ND0 from survey 1 is combined with one of the exposures of ND0, ND3 or ND4 from survey 2; exposure of ND1 from survey 1 is combined with one of the exposures of ND0, ND3 or ND4 from survey 2, etc.
When the intention is to sample just one exposure for a dietary individual, specify per survey different codes, e.g. ND1, ND2, ND3 for survey 1, ND4, ND5, ND6 for survey 2 and apply correlation, or specify 6 different individual codes and just one idNonDietarySurvey. Then, options with or without correlation are irrelevant and sampling results are identical no matter which option is chosen.

11 Mixture Selection

The most common model of cumulative risk assessment is to focus on substances that belong to the same common assessment groups (CAG). Substances in such a group belong to the same chemical family and may or may not have a similar mode of action. In assessing the risk, possible interactions between substances are often ignored and, moreover, little information is available about synergistic effects at low doses. More information is needed about the combined effects of such substances, but it is impractical to investigate all possible mixtures, and therefore instruments are needed to select the most relevant compounds for further investigation. This selection should not only be based on the hazard (highest relative potencies) but also on the exposure of the population to these substances. The potential risk of being exposed to chemicals in a mixture depends on the food consumption patterns of individuals in a population. A regular diet can contain hundreds of substances, so the number of combinations of compounds to which an individual in a population is exposed can be numerous. Therefore, it is essential to identify the most relevant mixtures to which a population is exposed.

In MCRA three approaches are available which may help to identify and select mixtures contributing to the exposure of a target population:

1. qualitative approach: **counting of co-exposure**. To which combinations of compounds are individuals exposed?
2. quantitative approach: **maximum cumulative ratio (MCR)**. To what degree are mixtures more important than single compounds?
3. quantitative approach: **sparse non-negative matrix underapproximation (SNMU)**. What mixtures predominantly determine the underlying patterns in the exposure matrix (compound x person (day))?

11.1 Counting co-exposure

In this qualitative approach, the number of combinations of compounds to which an individual is exposed are recorded. There is no cut-off level, the only criterion is the presence of a compound in the simulated daily diet or not. For an acute or short term exposure assessment, a simulated individual day is smallest entity to determine co-exposure. For a chronic or long term exposure assessment, co-exposures are summarized at the individual level, e.g. co-exposure is determined combining all consumption days of an individual. In Table 21, an example for acute exposure is shown.

compound	day 1	day2	day 3	...	day n
A	✓	✓		...	
B	✓		✓	...	✓
C	✓			...	✓
...	...	...	...	...	...

Table 21: counting combinations of compounds in the exposure matrix: for example, om day 1 there is co-exposure to compounds A, B and C.

For chronic exposure the data are summarised at the individual level.

In the current implementation in MCRA the co-exposure counts are summarized in four tables:
1. A table listing the frequencies of compound combinations grouped by the number of compounds.
2. A table listing the exposures with the highest numbers of compounds.
3. A table showing the most frequent unique combinations of compounds.
4. A table showing the most frequent combinations of compounds which may be part of larger groups.

11.2 Maximum Cumulative Ratio (MCR)

Price and Han (2011) propose the Maximum Cumulative Ratio (MCR) which is defined as the ratio of the cumulative exposure received by an individual on an intake day to the largest exposure received from a single compound:

MCR = Cumulative exposure/ Maximum exposure

This MCR statistic is also picked up as a practical device in a recent JRC report (Bopp et al. 2015) to investigate cumulative exposure. If MCR is large, it is important to consider cumulative effects, if MCR is close to 1, the individual exposure will not be much different from a single-compound assessment. The MCR can therefore be interpreted as the degree to which the risk of being exposed is underestimated by not performing a cumulative risk assessment.

The MCR statistic is implemented in MCRA for both the acute risk and the chronic risk cases. In the acute risk case the short-term (single-day) exposures are used, in the chronic case the long-term individual exposures (estimated by aggregating over the available survey days of each individual).

Table 22 shows an artificial example how the MCR is calculated in the acute risk case. First the cumulative exposure per day is calculated by cumulating the exposure of each substance multiplied by the relative potency factors (RPF). Then, for each day, the cumulative exposure (in equivalents of the reference compound) is divided by the maximum exposure of a single compound on that day. The last column shows the MCR values within parenthesis the compound with the highest exposure. The MCR has a value of 1 or close to 1 for mixtures where the exposure is dominated by one compound (*e.g.* day 1, compound B). When all compounds have approximately equal exposure (*e.g.* day 3) the MCR value is equal or close to the number of compounds, here 4. Day 2 represents an intermediate

case. The MCR suggest that for exposure days (or persons) with MCR values close to 1, the need for a cumulative risk assessment is low.

	compound A	compound B	compound C	compound D	total exposure	ratio
day 1	0.01	0.99	0	0	1.00	1.01 (B)
day 2	0.10	0.20	0.30	0.40	1.00	2.50 (D)
day 3	0.25	0.25	0.24	0.26	1.00	3.99 (D)

Table 22: Maximum Cumulative Ratios

In the example of Table 22, all days have equal values for total exposure. For real data, total exposure will vary. It is obviously of interest to know if the MCR is high or low at those days (or individuals) where the total exposure is highest.

All the following tests are implemented on the French dataset. This dataset is constituted of the exposure of the French general population to 83 pesticides. The total population (4079 individuals) is composed of 2624 adults (between 18 and 79 years) and 1455 children (between 3 and 17 years). Consumption data for these two populations are available from the INCA2 survey using a seven-day, open-ended food record. Data for pesticide residues in food were recorded for the 83 substances in the annual monitoring programmes implemented between 2009 and 2013 by the French administrations (Ministry of Economy, Ministry of Agriculture, Ministry of Health). Only 39 pesticides had quantified values (> of the limit of reporting). Thus, tests were done on 39 pesticides among the 83.

In **Figure 10**, French steatosis data (39 compounds, 4079 persons) are used to calculate the chronic exposure matrix. For each indiviual the MCR is calculated and plotted against the total exposure. The different colors are used to identify the single compounds with maximum exposure. From the original 39 compounds, 10 different compounds have the largest exposures. For the total exposure and MCR, the p5, p50 and p95 percentiles are indicated with the black line segments. The red line indicates the ratio with value 5. The dashed green lines indicate the p95 percentiles for the MCR value for different ranges of the total exposure. The plot shows that MCR values with Imazalil as risk driving compound (purple) are predominantly found in the lower part of the plot for relatively high values of the total exposure. A second finding is that MCR values decline when total exposure increases. This implies that cumulative exposure for most individuals is driven by multiple compounds. At the right site of the plot, individuals are found with high exposure. Because MCR values tend to be lower here, higher exposures are received from one predominant compound and not because many compounds are above the average level. For those individuals a cumulative risk assessment has less value.

Because Figure 10 can be very dense, in Figure 11, 95% confidence regions representing bivariate lognormal distributions of MCR and total exposure are plotted. This figure facilitates interpretation of the first figure. Note the two lines for Tetraconazole and Tebuconazole due to two observations available. Note that compounds with just one observation cannot be plotted in this display, and compounds with 2 observations are represented by a line.

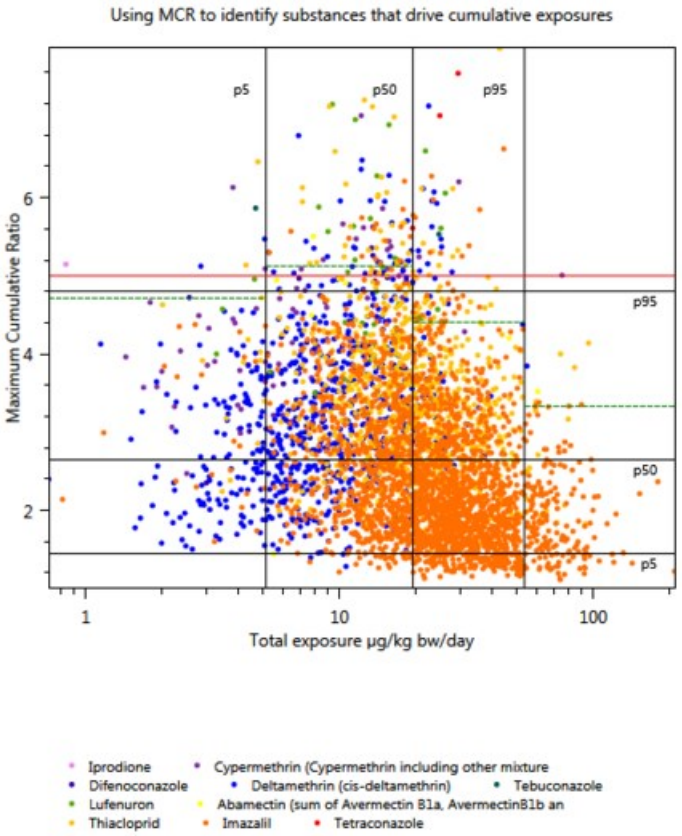


Figure 10: Maximum Cumulative Ratios vs total exposure

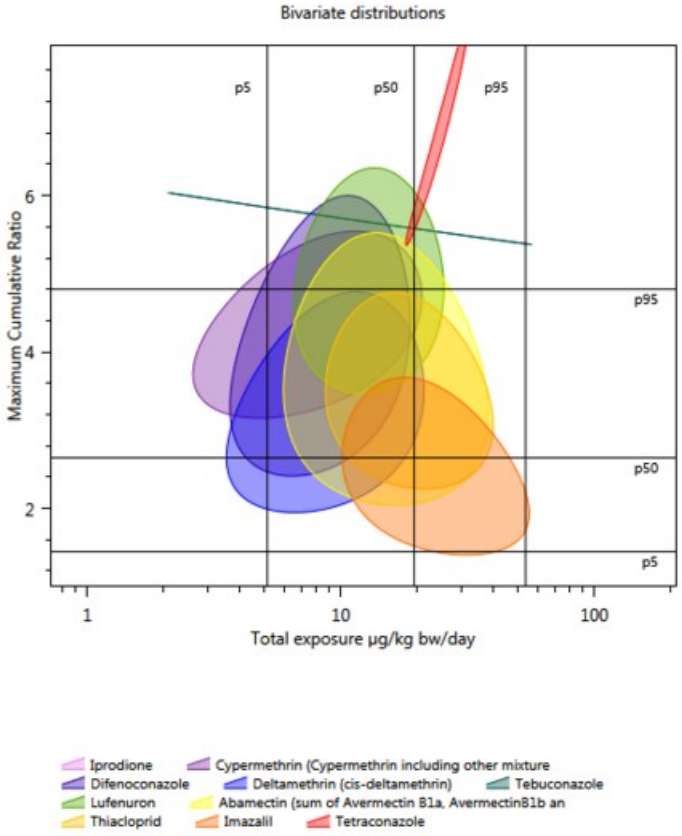


Figure 11: Bivariate distributions MCR vs total exposure

11.3 Matrix factorization

Starting point to identify major mixtures of substances using exposure data was to use Non-negative Matrix Factorization (NMF). Non-negative Matrix Factorization was proposed by Lee & Seung (1999) and Saul & Lee (2002) and deals specifically with non-negative data that have excess zeros such as exposure data. Zetlaoui et al. (2011), Sy et al. (2013), introduced this method in food risk assessment to define diet clusters.

The NMF method was then implemented by Béchaux et al. (2013) in order to identify food consumption patterns and main mixtures of pesticides to which the French population was exposed using TDS exposure to 26 priority pesticides.

At the start of the Euromix project ideas had evolved: to obtain less components per mixture experiments were made using Sparse Nonnegative Matrix Factorization (SNMF) (Hoyer 2004). This method was found not to give stable solutions. Better results were obtained with Sparse Nonnegative Matrix Underapproximation (SNMU) (Gillis and Plemmons 2013). This model also fits better to the problem situation because also the residual matrix after extracting some mixtures should be nonnegative. The SNMU method has been implemented in MCRA.

Indeed, NMF may be used to identify patterns or associations between substances in exposure data. NMF can be described as a method that finds a description of the data in a lower dimension. There are standard techniques available such as principal components analysis or factor analysis that do the same, but their lower rank representation is less suited because they contain negative values which makes interpretation hard and because of the modelling with a Gaussian random vectors which do not correctly deal with the excess of 0 and non-negative data . The NMF solution approximates a non-negative input matrix (*i.c.* exposure data) by two constrained non-negative matrices in a lower dimension such that the product of the two is as close as possible to the original input matrix. In Figure 12, the exposure matrix M with dimensions m (number of compounds) and n (number of intake days or persons) is approximated by matrix U and V with dimensions (m x k) and (k x n) respectively, where k represents the number of mixtures. Matrix U contains weights of the compounds per mixture, matrix V contains the coefficients of presence of mixtures in exposure per intake day or person. M is non-negative (zero or positive) and U and V are constraint to be non-negative. The minimization criterium is: $||M - UV||^2$ such that $U \geq 0$ and $V \geq 0$.

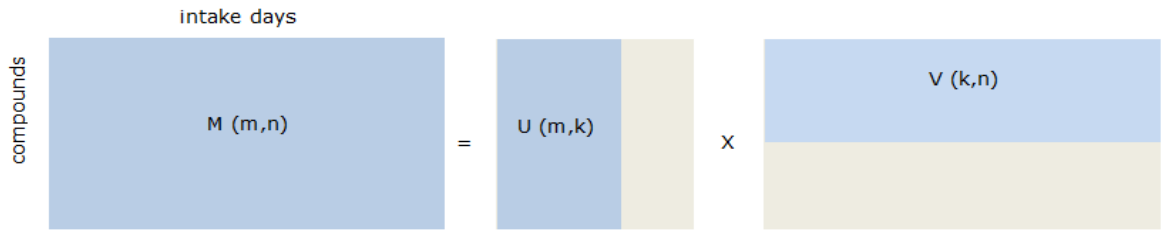


Figure 12: NMF approximation of exposure data

The minimization criterium is: $||M - UV||^2$ such that $U \geq 0$ and $V \geq 0$.

The solution found by NMF contains zeros, but mixtures still contain many components which complicates interpretability. Therefore, the Sparse Nonnegative Matrix Underapproximation (SNMU) (Gillis and Plemmons 2013) which also provide sparse results was investigated. The SNMU has also some nice features well adapted to the objective of the Euromix project: the solution is independent of the initialization and the algorithm is recursive. Consequently, the SNMU method which is based on the same decomposition process as the NMF was the one implemented in MCRA.

SNMU is initialized using an optimal nonnegative rank-one approximation using the power method (https://en.wikipedia.org/wiki/Power_iteration). This initialization is based on a singular value decomposition and results in general in a unique solution that is sparse. The SNMU algorithm is called recursive because after identifying the first optimal rank-one underapproximation u_1v_1, the next rank-one factor is recovered by subtracting u_1v_1 from M and applying the same factorization algorithm to the remainder $M - u_1v_1$. The solution u_1v_1 is called a rank-one underapproximation because an upper bound constraint is added to ensure that the remainder $M - u_1v_1$ is non-negative. For Matlab code see: https://sites.google.com/site/nicolasgillis/code.

11.3.1 Exposure matrix

Basically, exposure is calculated as consumption x concentration. By summing the exposures from the different foods for each compound per person day separately, the exposure matrix for mixture selection is estimated:

$$y_{ijc} = \frac{\sum_{k=1}^{p} x_{ijk}\ c_{ijkc}}{bw_i}$$

where y_{ijc} is the exposure to compound c by individual i on day j (in microgram substance per kg body weight), x_{ijk} is the consumption by individual i on day j of food k (in g), c_{ijkc} is the concentration of compound c in food k eaten by individual i on day j (in mg/kg), and bw_i is the body weight of individual i (in kg). Finally, p is the number of foods accounted for in the model.
More precisely, for an acute or short term risk assessment, the exposure matrix summarises the y_{ijc} with columns denoting the individual-days (ij) and rows denoting the compounds (c). Each cell represents the sum of the exposures for a compound on that particular day. For a chronic or long term risk assessment, the exposure matrix is constructed as the sum of all exposures for a particular compound averaged over the total number of intake days of an individual (compounds x persons). So each row represents a compound and a column an individual. Each cell represents the observed individual mean exposure for a compound for that individual. Note that in the exposure calculation, the concentration is the average of all residue values of a compound.

When relative potency factors (RPF) are available then exposures are multiplied by the RPF and thus exposures to the different substances are on the same and comparable scale (toxicological scale). In this case, the selection of the mixture is risk-based. In some cases, RPFs may not be available. In this case exposure to different substances, even in the same unit, may lead to very different effect. To give all compounds an equal weight a priori and avoid scaling effect, a normalization of the data can be applied as done in Béchaux et al. (2013). In this case, all compounds are assigned equal mean and variance, and the selection of the mixtures will work on patterns of correlation only. Then mixture selection is not risk-based anymore but, what could be called, co-exposure-based.

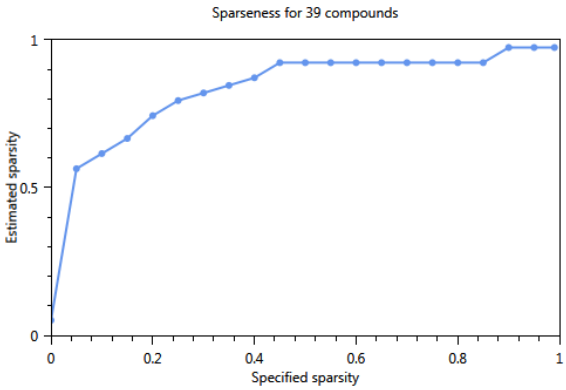
Finally, in the mixture selection module of MCRA, the SNMU expects RPF data for a risk-based selection. If not available, the user should provide alternative RPF data, e.g. values 1 for a purely exposure-based selection, or lower-tier estimates (e.g. a maximum value on RPF thought possible).

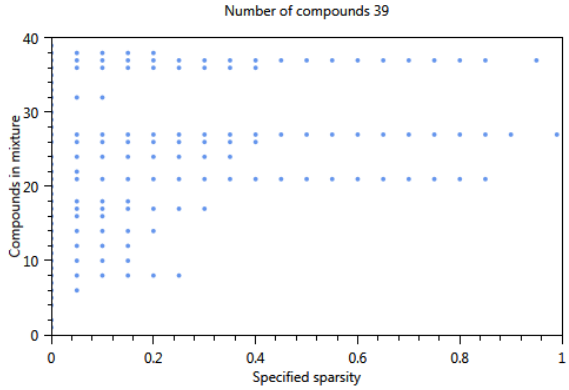
11.3.2 Mechanisms to influence sparsity

Two mechanisms to influence sparsity are available. The SNMU algorithm incorporates a sparsity parameter and by increasing the value, the final mixtures will be more sparse. The other approach is by using a subset of the exposure matrix based on a cut-off value for the MCR. High ratios correspond to high co-exposure, so it is reasonable to focus on these (person) days and not include days where exposure is received from a single compound (ratio close to 1). To illustrate the combined use of MCR and the sparsity parameter, the French steatosis data (39 compounds, 4079 persons) are used and person days with a ratio > 5 (see Figure 10) are selected yielding a subset of 139 records.

In Figure 13, the effect of using a cut-off level is immediately clear. The number of compounds of the first mixture is 17 whereas in the unselected case only 4 compounds were found The three plots show the influence of increasing the sparsity parameter from 0 to 1 on the number of compounds in the mixture. For values close to 0, the mixture contains 17 compounds. For values > 0.4 the number of compounds in the mixture drops to 3.

As mentioned before, one of the nice features of the SNMU algorithm is its recursive character which results in identical mixtures. In Figure 14, two *U* matrices are visualized. In the upper plot 5 mixtures are estimated, in the lower plot the solution for 10 mixtures is shown. Because of the ordering the plots look different, but a closer inspection of the first 5 mixtures of each solution shows that they are the same.





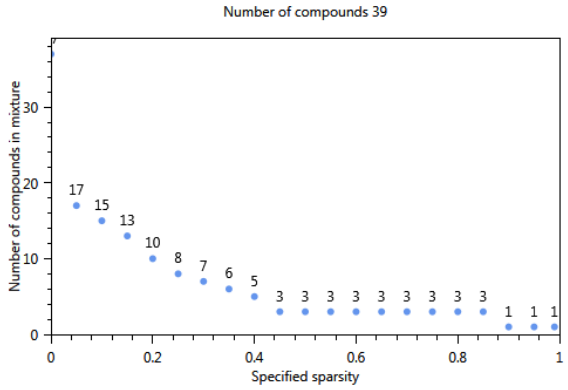


Figure 13: Influence of the specified sparsity parameter on the realized sparsity, n = 139

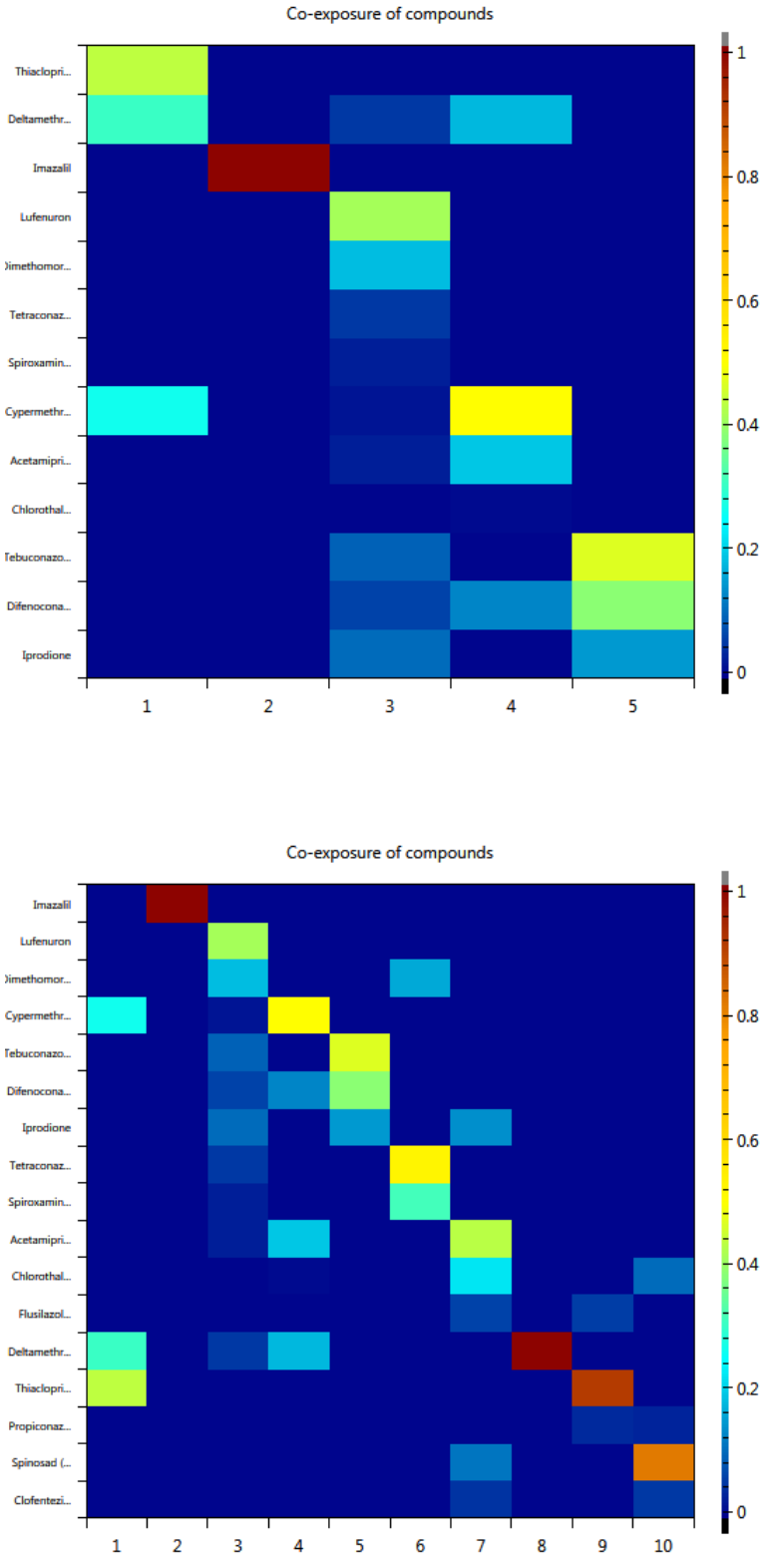


Figure 14: Heatmaps for solutions with 5 and 10 mixtures. The first 5 mixtures of both solutions are identical.

12 Total Diet Study

In Total Diet Studies (TDS), dietary exposure is based on whole diets as consumed. TDS offers a more realistic measure of exposure compared to traditional monitoring and surveillance programs, that is concerned with contamination of raw agricultural commodities. In a Total Diet Study, food selection is based on national consumption data in such a way that 90 to 95% of the usual diet is represented. Selected foods are collected, prepared as consumed and related foods are pooled prior to analysis. The composition of a TDS food sample is used in the conversion algorithm in an analogous manner as recipees describing the composition of a composite food (table FoodTranslations). The main difference is that the translation proportion is always 100% (default). Take TDS food *FruitMix* which is composed of *apple, orange* and *pear* (table TDSFoodSampleCompositions), then food-as-eaten *apple-pie* is converted to *apple, wheat* and *butter* and subsequently, *apple* to food-as-measured *FruitMix*. Not necessarily all foods as consumed are represented in a TDS food sample. Through the use of read across translations, these foods may be directly linked to a TDS food sample, *e.g.* by specifying that *pineapple* is translated to *FruitMix* (table ReadAcrossFoodTranslations), pineapple or foods containing pineapple as ingredient enter the exposure assessment. The default translation proportion is 100%.

The TDS approach for assessing risks are associated with chronic exposures only, in a single compound context or cumulative exposure assessment. In MCRA, Total Diet Studies are implemented in the chronic risk assessment module.

For more information about Total Diet Studies, visit the TDS-Exposure website http://www.tds-exposure.eu.

12.1 Scenario analysis

The outcome of a MCRA risk assessment may be that some foods dominate the right upper tail of the exposure distribution. A scenario analysis answers the question to what extent the risk of foods with a high exposure would have been diminished by an intervention or by taking any precautions. To be able to do so, some information is needed about the variability of the concentration distribution of the raw agricultural commodities that make up the TDS food sample. These distributions may be characterised by a mean and a dispersion factor, the standard deviation or, preferably, a percentile point *e.g.* p_{95}. Monitoring samples may be used for this purpose. In addition, for each subsample food an upper concentration limit is needed. This value is interpreted as the concentration that is considered a high risk. The decision to intervene or not is based on the comparison between this upper limit and p_{95}.

For $p_{95} \leq limit$, most concentration values are below the value that is considered as a potential risk, so there is no urgency to take any precautions. When the opposite is true, *i.c.* $p_{95} > limit$, there may be an argument to intervene for this specific food. In MCRA, limits and p_{95}'s are supplied in table ConcentrationDistributionsValues. In the MCRA interface, a scenario analysis is checked (optionally) and in the scroll down menu only foods are shown with $p_{95} > limit$. Selected foods enter the risk assessment with a reduced concentration value:

$$c_{TDS} \ / \ reductionfactor,$$

where c_{TDS} is the concentration value of the TDS food with $reductionfactor = p_{95}/limit$.

12.2 Read across versus TDS compositions

After the conversion from food-as-eaten to food-as-measured, part of the foods or their ingredients are not linked to a TDS food. These are so-called failed foodconversions. All other foods or ingredients enter the risk assessment and contribute to the exposure distribution. The total exposure can be traced back to foods that enter the risk assessment through read across translations or through TDS compositions. In the *FruitMix* example, the total exposure for *FruitMix* is split into a part due to the consumption of *apple, orange* and *pear*, the remaining part relates to the consumption of *pineapple.*

The first part is summarized as exposure due to all TDS sample compositions, the second as Read Across translations.

12.3 Uncertainty

In MCRA, uncertainty of TDS food sample concentrations is specified through the use of table ConcentrationDistributionsValues. For each subfood, *e.g. apple (*subfood of TDS food *FruitMix)*, a coefficient of variation, CV, is specified that is derived using the available monitoring samples. Note that monitoring samples may be composite samples. For *apple*, composite food samples are measured and each sample contains *e.g.* 12 apples with unit weight 200 g. So monitoring concentrations, $c_{m\_i,}$ are based on composite samples with a total weight w_{m_i} = 2400 g each.

A TDS food sample is composed of w_i g of food i with $i = 1...k$, w_i represents the PooledAmount in table TDSFoodSampleCompositions. Then, the concentration of a TDS food sample may be represented as:

$$c_{TDS} = \sum_{i=1}^{k}(w_i * c_i) / \sum_{i=1}^{k} w_i$$

with variance:

$$var(c_{TDS}) = \sum_{i=1}^{k}(w_i * var(c_i)) / \sum_{i=1}^{k} w_i$$

and $var(c_i)$ is the variance of concentrations c_i of food i with portion sample size w_i.

It is expected that increasing the number of units in a composite sample will have a reverse effect on the variation between concentrations.
Suppose TDS food *FruitMix* is composed of 2 x 200 = 400 g *apple.* The expected variation between portion sizes of 400 g will be larger than between portion sizes of 2400 g:

$$var(c_i) = var(c_{m\_i}) * w_{m\_i} / w_i$$

The variance of the monitoring samples are corrected as follows, calculate:

1. $var(c_{m\_i}) = log(CV^2_{m\_1} + 1)$
2. $var(c_i) = var(c_{m\_i}) * w_{m\_i} / w_i$
3. $CV_i = \sqrt{(exp(var(c+i)) - 1)}$

Specify CV_i in table ConcentrationDistributionsValues.

13 Health impact assessment

‘A method is proposed for integrated probabilistic risk assessment where exposure assessment and hazard characterization are both included in a probabilistic way. The aim is to specify the probability that a random individual from a defined (sub)population will have an exposure high enough to cause a particular health effect of a predefined magnitude, the critical effect size (CES). The exposure level that results in exactly that CES in a particular person is that person’s individual critical effect dose (ICED). Individuals in a population typically show variation, both in their individual exposure (IEXP) and in their ICED. Both the variation in IEXP and the variation in ICED are quantified in the form of probability distributions. Assuming independence between both distributions, they are combined (by Monte Carlo) into a distribution of the individual margin of exposure (IMoE). The proportion of the IMoE distribution below unity is the probability of critical exposure (PoCE) in the particular (sub)population. Uncertainties involved in the overall risk assessment (i.e., both regarding exposure and effect assessment) are quantified using Monte Carlo and bootstrap methods. This results in an uncertainty distribution for any statistic of interest, such as the probability of critical exposure (PoCE). The method is illustrated based on data for the case of dietary exposure to the organophosphate acephate. We present plots that concisely summarize the probabilistic results, retaining the distinction between variability and uncertainty. We show how the relative contributions from the various sources of uncertainty involved may be quantified.’ (abstract from van der Voet & Slob, 2007).

‘A statistical model is presented extending the integrated probabilistic risk assessment (IPRA) model of van der Voet and Slob (2007) The aim is to characterise the health impact due to one or more chemicals present in food causing one or more health effects. For chemicals with hardly any measurable safety problems we propose health impact characterisation by margins of exposure. In this probabilistic model not one margin of exposure is calculated, but rather a distribution of individual margins of exposure (IMoE) which allows quantifying the health impact for small parts of the population. A simple bar chart is proposed to represent the IMoE distribution and a lower bound (IMoEL) quantifies uncertainties in this distribution. It is described how IMoE distributions can be combined for dose-additive compounds and for different health effects. Health impact assessment critically depends on a subjective valuation of the health impact of a given health effect, and possibilities to implement this health impact valuation step are discussed. Examples show the possibilities of health impact characterisation and of integrating IMoE distributions. The paper also includes new proposals for modelling variable and uncertain factors describing food processing effects and intraspecies variation in sensitivity.’ (abstract from*:* van der Voet *et al*, 2009).

14 Uncertainty analysis

14.1 Quantifying uncertainties

In this section, uncertainty due to limited sampled data is covered, not the uncertainty of model outcomes that may arise by conducting different modelling approaches or applying alternative assumptions in a dietary exposure assessment.

The basic acute exposure distribution is estimated in a Monte Carlo simulation by combining dietary consumption records (person-days) with sampled residue values. The resulting distribution represents a combination of variability in consumption within the population and between residues in a food lot. Percentiles may be used for further quantification *e.g.* the median or 99^{th} percentile. Due to the limited size of the underlying data, these outcomes are uncertain. Confidence (or uncertainty) intervals reflect the uncertainty of these estimates, where MCRA uses bootstrap methodology and/or, depending on the available data, parametric methods to estimate the uncertainty.

14.1.1 Empirical method, resampling

The empirical bootstrap is an approach to estimate the accuracy of an outcome. In its most simple, non-parametric form, the bootstrap algorithm resamples a dataset of *n* observations to obtain a *bootstrap sample* or *resampled set* of again *n* observations (sampling with replacement, that is: each observation has a probability of 1/*n* to be selected at any position in the new resampled set). By repeating this process *B* times, one can obtain *B* resampled sets, which may be considered as alternative data sets that might have been obtained during sampling from the population of interest. Any statistic that can be calculated from the original dataset (*e.g.* the median, the standard deviation, the 99^{th} percentile, etc.) can also be calculated from each of the *B* resampled sets. This generates a *uncertainty distribution* for the statistic under consideration. The uncertainty distribution characterises the uncertainty of the inference due to the sampling uncertainty of the original dataset: it shows which statistics could have been obtained if random sampling from the population would have generated another sample than the one actually observed (Efron, 1979, Efron & Tibshirani, 1993).

14.1.1.1 *Consumption data*

In MCRA 8, in an acute exposure assessment individual consumption **day** data are resampled, thus preserving the multivariate consumption patterns and associated weights and/or other individual characteristics. The method of resampling is changed compared to MCRA 7, where we actually resampled the set of individuals. In MCRA we resample the set of individuals x number of survey days. We think that the new implementation better reflects the notion of acute exposure which is expressed as the normalized intake per day. For chronic exposure assessments the resampling algorithm remained unchanged and the set of individuals (with corresponding days) is resampled.

14.1.1.2 *Concentration data*

Depending on the chosen concentration model, *e.g.* EFSA Guidance basic models or custom Acropolis model, resampling is done on a sample-based basis preserving co-occurrence of residue values on the same sample or, for the non-sample-based approaches, on a univariate collection of concentration values. For the last approach, the uncertainty algorithm is applied to the dataset consisting of both non-detects and positive values; in practice, for a dataset with n_0 non-detects and n_1 positive values, the number of positive values in a resampled set is obtained as a draw from a binomial distribution with parameter $n_1/(n_0+n_1)$ and binomial total n_0+n_1. Then, this number of values is selected randomnly from the set of n_1 positive values.

14.1.2 Parametric methods

Instead of bootstrapping the observed data, inference about parameters is based on parametric methods. For processing, where factors are specified through a nominal and/or upper value this is the

natural choice. For concentration data, where the lognormal model is used to represent less conservative scenario's (EFSA, 2012), the parametric bootstrap may be an alternative, especially when data are limited and the empirical bootstrap fails.

14.1.2.1 *Concentration models*

Let x denote a random variable from the specified distribution. The log transformed variable $y = \ln(x)$ is normally distributed with mean μ_y and variance σ_y^2. The maximum likelihood estimates are $\hat{\mu}_y$ and $\hat{\sigma}_y^2$. In each bootstrap sample, values are drawn from a normal distribution where the maximum likelihood estimates are replaced by ($\hat{\mu}_y^*$, $\hat{\sigma}_y^{*2}$), see also paragraph **Error! Reference source not ound.**.

14.1.2.2 *Processing factors*

Processing effects are modelled either by a fixed processing factor, or by a lognormal or logistic-normal distribution (depending on the distribution type as set in table ProcessingTypes).
In case of a fixed factor, the uncertainty distribution is lognormal or logistic-normal with the same mean μ as the fixed value, and with a standard deviation σ_{unc} which is calculated from the specified central value μ (or nominal) and an estimate of the p95 of the uncertainty distribution (set *NominalUncertaintyUpper* in table ProcessingFactors).

The calculation is: $\sigma_{unc} = \frac{f(NominalUncertaintyUpper) - f(\mu)}{1.645}$

with *f() = logit* for the logistic-normal distribution (distribution type 1) and *f() = ln* for the lognormal distribution (distribution type 2). Values lower than 0.01 or higher than 0.99 (distribution type 1 only) are replaced by default values (0.01 and 0.99); this is useful computationally to avoid problems. In each iteration of the uncertainty analysis a new value is drawn from this distribution to be used as a fixed factor in the Monte Carlo calculation.
In case of distribution based processing factors (describing the variability of processing factors) two uncertainties can be specified.
For $\sigma_{unc,}$ specification and calculation is as before (set *NominalUncertaintyUpper* in table ProcessingFactors).
The uncertainty about the variability standard deviation $\sigma_{var} = \frac{f(Upper) - f(\mu)}{1.645}$ can be specified by the *UpperUncertaintyUpper* value in table ProcessingFactors. This value is specified as the p95 upper limit on *Upper*. The specified value is used to derive in a iterative search the number of degrees of freedom *df* (van der Voet *et al.* 2009). In the uncertainty analysis, a modified chi-square distribution with *df* degrees of freedom is used to generate new values of σ_{var}. A very high value of *df* means litte uncertainty and σ_{var} will be almost equal in all iterations of the uncertainty analysis. A *df* close to 0 means a large uncertainty and very different values of σ_{var} will be obtained in the iterations of the uncertainty analysis. The p95 upper limit on *Upper* is set through parameter *UpperUncertaintyUpper* in table ProcessingFactors.

14.1.2.3 *Portion sizes*

In the context of the European Food Consumption Validation Project (EFCOVAL) the MCRA model for uncertainty has been adapted specifically to the six quantification methods of EPIC-SOFT (Table 23). Using EPIC-SOFT for 24-hour recall consumptions are quantified using portion size and amounts of portions consumed. Although individual consumption data are expressed in grams per day, the primary data may be associated with uncertainty in portion size and amount or number of portions consumed. So, the primary data are unitweights (e.g. the weight of a portion shown on a photo, or the weight of a standard household measure) and amounts of units (e.g. the number of shown portions or the number of cups), the multiplication of both values is the amount consumed in grams. The corresponding portion size uncertainty is primarily connected with unitweights and amounts.

Method	Unitweight (*uw*)	Amount (*a*)
Photographs (P)	Standard portion in grams (Photo 1 of broccoli is 78 g)	Proportion or multiple of standard portion (1 times photo 1 of broccoli)
Household measures (H)	Standard portion in grams (a glass of tea is 150 g)	Proportion or multiple of standard portion (2 glasses of tea)
Standard units (U)	Standard portion in grams (a can of corn is 285 g)	Proportion or multiple of standard portion (1/2 a can of corn)
Standard portion (S)	Standard portion in grams (onion along with fries weighs 10 g)	1
Gram/volume (G)	1	Amount in grams (75 g of potato salad)
Unknown (?)	1	Amount in grams (Salad dressing weighs 15 grams)

Table 23: Overview of EPIC-SOFT quantification methods, with examples in brackets

Three methods (P, H and U) use both unitweights and amounts, one method (S) uses only unitweights, and two methods (G and ?) use only amounts. The difference between unitweight and amount is as follows: unitweights (in grams) are unique for a specific 'food item – quantification method'-combination, but the same for all individuals in the survey, whereas amounts are potentially different for each food item on each eating occasion for each day of an individual. Amounts are in grams (methods G and ?) or in number of units (methods P, H, and U).

For portion size uncertainty analysis of the usual intake assessment of foods and nutrients two sources of uncertainty are modelled:

1. uncertainty in *uw* (for EPIC-SOFT quantification methods P, H, U and S)
2. uncertainty in *a* (for EPIC-SOFT quantification methods G, P, H, U and ?)

For quantification methods P, H and U the uncertainty in *uw* as well as the uncertainty in *a* needs to be specified, for quantification methods G and ? the uncertainty in *a* needs to be specified, and for method S the uncertainty in *uw* needs to be specified. The uncertainty *cv* specifications were obtained using limited expert opinion to provide estimated upper values for *a* and *uw*, and equating these to the p97.5 of the (log)normal uncertainty distribution (the best estimates are interpreted as the mean *m*).

More details of the approach to portion size uncertainty implemented in MCRA are described in Souverein *et al*. (2011).

14.1.3 External uncertainty distributions

14.1.3.1 *Non-dietary data*

When the user supplies non-dietary exposure estimates that have been calculated probabilistically, i.e. there is a distribution for the non-dietary exposure rather than a single nominal value, then this information will be propagated as part of the MCRA exposure assessment. Distributions may be included to represent variability, uncertainty or both, and in these cases the aggregate exposure estimates are reported with variability and/or uncertainty as appropriate. Multiple (uncertain) values from the non-dietary exposure distribution may be supplied per individual and per compound.

Exposures within a dietary survey may be expressed as correlated or independent for the different compounds. For example, if the exposures are a mixture of compounds in a known ratio (e.g. from a specific tank mix of pesticides), or if exposure to one compound strongly implies that exposure to another is likely, these relationships may be included in the non-dietary data supplied by the user. Inference for the matched-case scenario with uncertainty analysis can use exposure sets. These are specific sets of exposures defined *for each individual*, (*e.g.* Table 13, Table 16) and in any uncertainty

iteration an individual will receive exactly one of the exposure sets for that individual. Alternatively, independence may be represented by generating sets drawn from independent distributions when generating these tables. Details, including an example of the input format for implementing uncertainty in non-dietary exposure, are given in Section 11.7.

14.2 Unquantified uncertainties

In any exposure or risk assessment, only a proportion of the uncertainties will be quantified, while others remain unquantified. Even when a source of uncertainty is quantified, there will be further uncertainty (sometimes referred to as 'secondary uncertainty') about how well it is represented. When using a risk or exposure estimate to support decision-making, it is important to consider whether the unquantified uncertainties might be large enough to change the risk management decision. It is therefore important to consider uncertainties at each step in the assessment and document them in a transparent manner (Codex 2011, EFSA 2009).

EFSA's (2006) guidance on uncertainty in dietary exposure assessment suggested a tabular approach for listing the uncertainties and evaluating their individual and combined impact on the estimated exposure. EFSA (2012) emphasises the benefits of providing at least an approximate quantification of the scale on which the evaluations are made. The general form of the basic tabular approach is illustrated in **Table 24**, and an example of a quantitative scale for evaluating the impact of uncertainties is shown in Figure 15. More detailed step-by-step guidance for constructing uncertainty tables of this type (in the context of hazard assessment but equally applicable to exposure assessment) is provided by Edler et al. (2013, section 4.2).

Sources of uncertainty additional to those quantified in the exposure assessment	**Evaluation of uncertainty**
• Uncertainty 1: *very briefly describe the uncertainty and your evaluation of the extent to which it might cause underestimation and/or overestimation of exposure*	*Record here your evaluation as a range of numbers, symbols or words*
• *Insert more rows for additional uncertainties, as needed*	
Overall assessment: *verbal description of your assessment of the overall unquantified uncertainty affecting the exposure estimate and a very brief explanation of how it is derived from the individual uncertainties*	*Record here your evaluation of the overall unquantified uncertainty as a range of numbers, symbols or words*

Table 24. General tabular format for evaluating unquantified uncertainties affecting assessment of a single route of exposure. If symbols or words are used in the right hand column they must be defined in the table legend, in accompanying text, or in a diagram or scale (see Figure 15).

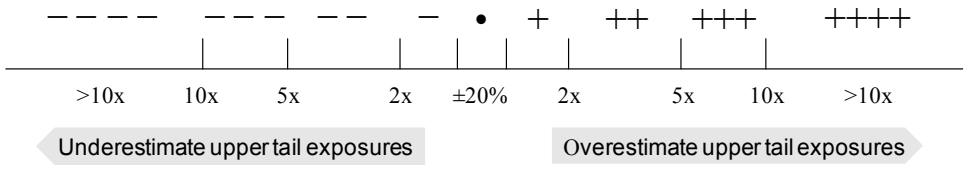


Figure 15. Example of quantitative scale for symbols used to evaluate unquantified uncertainties[1].

[1] Note that some users prefer to define the symbols with opposite meaning, e.g. + for uncertainties that would tend to make the true exposure higher than the estimate (e.g. EFSA, 2013a). There is no general consensus on which approach to use, so it is important to make clear which is used in each assessment.

Appendix 2 of EFSA (2012) included a general assessment of unquantified uncertainties and was designed to be used as a starting point or template for evaluating uncertainties in probabilistic dietary exposure assessments. This can be adapted to specific applications and models such as those implemented in the MCRA software tool. An example of this for dietary exposure is presented in Boon et al. (2015). Similar evaluations may be constructed for non-dietary routes of exposure (Kennedy et al., 2015a).

In an assessment of aggregate exposure it will be necessary to evaluate the overall uncertainty of the estimated aggregate exposure, taking account of the uncertainties associated with each individual route and also how they combine. For this purpose it is useful to develop a separate summary table, along the lines illustrated in Table 25, which summarises the evaluated uncertainty for each route of exposure and provides an evaluation of the overall uncertainty for the aggregate exposure. A complete example, comprising uncertainty tables for individual routes of exposure and for aggregate exposure, may be found in EFSA (2013a, section 4.9.3 and appendix VIII). Other examples are in Boon et al. (2015) and Kennedy et al. (2015a).

Table 25. Suggested format for table summarising the assessment of unquantified uncertainties for each route of exposure together with the assessor's subjective evaluation of their combined impact on the estimate of aggregate exposure (bottom row). For example of scale for symbols see Figure 15.

Route of exposure	**Magnitude and direction of unquantified uncertainties affecting the estimated exposure**
Dietary route. *Copy here the narrative conclusion from the evaluation of unquantified uncertainties for the dietary route (e.g. bottom row of uncertainty table in Boon et al. (2015).*	*Symbols to show overall evaluation of uncertainty for the dietary route, copied from the relevant table (e.g.: +/++)*
Dermal route. *Copy here the narrative conclusion from the evaluation of unquantified uncertainties for the dermal route (e.g. bottom row of uncertainty table in Kennedy et al. (2015b).*	*Symbols to show overall evaluation of uncertainty for the dermal route, copied from the relevant table*
Insert additional rows for further routes of exposure, and/or for uncertainties associated with how the routes are combined to estimate aggregate exposure.	
Overall evaluation of uncertainty affecting the estimate of aggregated exposure. *Add narrative text here, describing the assessor's evaluation of the overall degree of uncertainty affecting the assessment outcome, taking account of the uncertainties for each route as summarised above.*	*Evaluation of overall uncertainty for the estimate of aggregate exposure (e.g., - - - /+)*

Evaluation of unquantified uncertainties is necessarily subjective and approximate, and it is important to make this clear when communicating results, to avoid them being over-interpreted. If a firmer assessment of the unquantified uncertainties is required, consideration could be given to conducting the evaluation using formal methods for expert elicitation (EFSA, 2013b) or to treating some of the unquantified uncertainties deterministically or probabilistically so they become part of the quantified uncertainty. In the latter case it may be efficient to target quantification on the most important unquantified uncertainties, as indicated by the approximate subjective evaluation. This process may be repeated iteratively until the characterisation of uncertainty is sufficiently clear to support the risk management decision in hand. For more discussion of this tiered approach to evaluating uncertainty, see EFSA (2006).

15 Appendices

15.1 Concentration models

Let *x* denote a random variable from a lognormal distribution. Then, the log transformed variable *y* = ln(*x*) is normally distributed with mean μ_y and variance σ_y^2.

The probability density function (p.d.f.) of *y* may be expressed as:

$$f_y(y; p_0, \mu_y, \sigma_y^2) = p_0 I(y;0) + \{(1-p_0)(1-I(y;0)\} * \frac{1}{\sqrt{2\pi}\sigma_y} \exp(-\frac{(y-\mu_y)^2}{2\sigma_y^2})$$

where $p_0 = \Pr(y < \log(x_{lor})$, x_{lor} is the limit of reporting and I(*y*;0) is an indicator function for $y < \log(x_{lor})$. For $p_0 = 0$, the p.d.f. of *y* reduces to the usual lognormal density.

The left truncated density for $y \geq \log(x_{lor})$ may be expressed as:

$$f_y(y; \mu_y, \sigma_y^2) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp(-\frac{(y-\mu_y)^2}{2\sigma_y^2}) / (1-\Phi(z))$$

with $\Phi(.)$ the standard normal c.d.f. and $z = (\log(x_{lor}) - \mu_y) / \sigma_y$.

Model parameters are estimated using maximum likelihood estimation based on the loglikelihood functions specified below. The loglikelihood functions are evaluated in R, using the *optim* algorithm to find estimates for μ_y, σ_y^2 and p_0.

15.1.1 Mixture zero spike and censored lognormal

The loglikelihood may be expressed as:

$$\log L(p_0, \mu_y, \sigma_y^2) = \sum_{i=1}^{n_0} \log(p_0 + (1-p_0)\Phi(z_i)) + n_1 \log(\frac{1-p_0}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n} \frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

where $y_i = \log(x_i)$, $\Phi(.)$ is the standard normal c.d.f., $z = (\log(x_{i,lor}) - \mu_y) / \sigma_y$,

$z_{lor} = (\log(lor) - \mu_y) / \sigma_y$, with n_0 number of censored values ($x_i < x_{i,lor}$), n_1 number of uncensored values ($x_i >= x_{i,l\,or}$) and x_i, $i = 1 \ldots n$.

Multiple values for LOR are allowed.

15.1.2 Censored lognormal

When $p_0 = 0$, the loglikelihood reduces to:

$$\log L(\mu_y, \sigma_y^2) = \sum_{i=1}^{n_0} \log(\Phi(z_i)) + n_1 \log(\frac{1}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n} \frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

Multiple values for LOR are allowed.

15.1.3 Mixture non-detect spike and truncated lognormal

Ignoring the n_0 values below x_{lor}, the loglikelihood may be expressed as:

$$\log L(\mu_y, \sigma_y^2) = -n_1 \log(1-\Phi(z)) + n_1 \log(\frac{1}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n} \frac{(y_i-\mu_y)^2}{2\sigma_y^2}$$

Only one value for LOR is allowed.

15.1.4 Mixture non-detect spike and lognormal

Ignoring the n_0 values below x_{lor}, the loglikelihood may be expressed as:

$$\log L(\mu_y, \sigma_y^2) = n_1 \log(\frac{1}{\sqrt{2\pi}\sigma_y}) - \sum_{i=n_0+1}^{n} \frac{(y_i - \mu_y)^2}{2\sigma_y^2}$$

Multiple values for LOR are allowed.

15.2 Unit variability

A composite sample for food k is composed of nu_k units with nominal unit weight wu_k. The weight of a composite sample is $wm_k = nu_k \times wu_k$ with mean residue value cm_k.

15.2.1 Beta distribution

Under the beta model simulated unit values are drawn from a bounded distribution on the interval (0, c_{max}) with $c_{max} = nu_k * cm_k$. The standard beta distribution is defined on the interval (0, 1) and is usually characterised by two parameters a and b, with $a>0, b>0$ (see *e.g.* Mood *et al.* 1974). Alternatively, it can be parameterised by the mean $\mu = a/(a+b)$ and the variance $\sigma^2 = ab(a+b+1)^{-1}(a+b)^{-2}$, or, as applied in MCRA, by the mean μ and the squared coefficient of variation $cv^2 = ba^{-1}(a+b+1)^{-1}$.
For the simulated unit values in each iteration of the program we require an expected value cm_k. This scales down to a mean value $\mu = cm_k/c_{max} = 1/nu_k$ in the (standard) beta distribution. From this value for μ and an externally specified value for cv_k the parameters a and b of the beta distribution are calculated as:

$$a = b(nu_k - 1)^{-1}$$
$$b = \frac{(nu_k - 1)(nu_k - 1 - cv_k^2)}{nu_k cv_k^2}$$

From the second formula it can be seen that cv_k should not be larger than $\sqrt{nu_k - 1}$ in order to avoid negative values for b.

When the unit variability is specified by a variability factor $v_k = \frac{p97.5_k}{cm_k}$ instead of a coefficient of variation cv_k then MCRA applies a bisection algorithm to find a such that the cumulative probability $P[Beta(a,b)] = 0.975$ for $b = a(nu_k - 1)$.
Sampled values from the beta distribution are rescaled by multiplication with c_{max} to unit concentrations c_{ijk} on the interval (0, c_{max}).

15.2.2 Lognormal distribution

The lognormal distribution is characterised by μ and σ, which are the mean and standard deviation of the log-transformed concentrations. The unit log-concentrations are drawn from a normal distribution with mean $\mu = \ln(cm_{ik}) - \frac{1}{2}\sigma^2$.
The coefficient of variation cv is turned into the standard deviation σ on the log-transformed scale with:

$$\sigma = \sqrt{\ln(cv^2 + 1)}$$

The variability factor is defined as the 97.5th percentile of the concentration in the individual measurements divided by the corresponding mean concentration seen in the composite sample. A variability factor v is converted into the standard deviation σ as follows:

$$v = \frac{p97.5}{mean} = \frac{e^{\mu+1.96\sigma}}{e^{\mu+1/2\sigma^2}} = e^{1.96\sigma-1/2\sigma^2}$$

with μ and σ representing the mean and standard deviation of the log-transformed concentrations. So

$$\ln(v) = 1.96\sigma - 1/2\sigma^2$$

Solving for σ gives: $\sigma^2 - 2*1.96\sigma + 2\log(v) = 0$, with roots for σ according to:

$$\sigma = 1.96 \pm \sqrt{(1.96^2 - 2\log(v))}$$

The smallest positive root is taken as an estimate for σ .

15.2.3 Bernoulli distribution

The bernoulli model is a limiting case of the beta model, which can be used if no information on unit variability is available, but only the number of units in a composite sample is known (see van der Voet *et al*. 2001).
As a worst case approach we may take the coefficient of variation *cv* as large as possible. When *cv* is equal to the maximum possible value $\sqrt{nu_k - 1}$, the (unstandardised) beta distribution simplifies to a bernoulli distribution with probability $(nu_k - 1)/nu_k$ (or $(v_k\text{-}1)/v_k$) for the value 0 and probability $1/nu_k$ (or $1/v_k$) for the value $c_{max} = nu_k * cm_k$..
In MCRA values 0 are actually replaced by cm_k, to keep all values on the conservative side. For example, with $nu_k = 5$, there will be 80% probability at $c_{ijk} = cm_k$ and 20% probability at $c_{ijk} = c_{max.}$.
When the number of units nu_k in the composite sample is missing, the nominal unit weight wu_k is used to calculate the parameter for unit variability.

15.3 Processing

For distribution based processing factors specify $f_{k,nominal}$ and $f_{k,upper}$ (*Nominal* and *Upper* in table ProcessingFactors). Two situations are distinguished depending on the type of transformation.

15.3.1 Nonnegative processing factors

Equate the logarithms of $f_{k,\,nominal}$ and $f_{k,\,upper}$ to the mean and the 95% one-sided upper confidence limit of a normal distribution. This normal distribution is specified by a mean $\ln(f_{k,\,nominal})$ and a standard deviation $\{\ln(f_{k,\,upper}) - \ln(f_{k,\,nominal})\}/1.645$.

15.3.2 Processing factors between 0 and 1:

Equate the logits of $f_{k,\,nominal}$ and $f_{k,\,upper}$ to the mean and the 95% one-sided upper confidence limit of a normal distribution. This normal distribution is specified by a mean $\text{logit}(f_{k,\,nominal})$ and a standard deviation $\{\text{logit}(f_{k,\,upper}) - \text{logit}(f_{k,\,nominal})\}/1.645$.

15.4 Box-Cox power transformation

The Box-Cox power transformation is a data transformation to achieve a better normality and to stabilize the variance. In MCRA, the transformation parameter *p* in $(y^p - 1)/p$ is determined by maximizing the log-likelihood function

$$l(p) = -\frac{n}{2}\log[\frac{1}{n}\sum_{i=1}^{n}(y_i^{(p)} - \overline{y^{(p)}})^2] + (p-1)\sum_{i=1}^{n}\log y_i$$

where *i* indexes the *n* observations and

$\overline{y^{(p)}} = \frac{1}{n}\sum_{i=1}^{n} y_i^{(p)}$ is the average of the $y_i^{(p)}$ (Box & Cox, 1964).

15.5 Chronic exposure assessment

15.5.1 Daily consumed foods

Foods are consumed on a daily basis.

15.5.1.1 *Model*

For individual *i* on day *j* let Y_{ij} denote the 24 hour recall of a food (i=1…n; j=1…n_i). In most cases within-individual random variation is dependent on the individual mean and has a skewed distribution. It is therefore customary to define a one-way random effects model for Y_{ij} on some transformed scale

$$Y_{ij}^{*} = g(Y_{ij}) = \mu_i + b_i + w_{ij} \quad \text{with} \quad b_i \sim \mathrm{N}(0, \sigma_b^2) \quad \text{and} \quad w_{ij} \sim \mathrm{N}(0, \sigma_w^2)$$

Note that b_i represents variation between individuals and w_{ij} represents variation within individuals between days.

The mean μ_i may depend on a set of covariate $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})$:

$$\mu_i = \beta_0 + \boldsymbol{\beta}_1^t \boldsymbol{Z}_i$$

where β_0 and $\boldsymbol{\beta}_1$ are regression coefficients.

The usual intake T_i for an individual *i* is defined as the mean consumption over many many days. This assumes that the untransformed intakes Y_{ij} are unbiased for true usual intake rather than the transformed intakes Y_{ij}^{*}. In mathematical terms T_i is the expectation of the intake for this individual where the expectation is taken over the random day effect:

$$T_i = \mathrm{E}_w\left[g^{-1}(\mu_i + b_i + w_{ij}) \mid b_i\right] \stackrel{\text{def}}{=} F(b_i)$$

15.5.1.2 *Model based usual intake*

For the model based usual intake first note that the conditional distribution

$$(\mu_i + b_i + w_{ij} \mid b_i) \sim \mathrm{N}(\mu_i + b_i, \sigma_w^2)$$

It follows that the usual intake T_i is given by

$$T_i = \mathrm{E}_w\left[g^{-1}(\mu_i + b_i + w_{ij} \mid b_i)\right] = \int_{-\infty}^{\infty} g^{-1}(\mu_i + b_i + w_{ij}) \frac{1}{\sqrt{2\pi\sigma_w^2}} \, exp\left(-\frac{w^2}{2\sigma_w^2}\right) dw$$

Logarithmic transformation

For the logarithmic transform the usual intake T_i can be written in closed form using the formula for the mean of the lognormal distribution:

$$T_i = \exp(\mu_i + b_i + \sigma_w^2/2)$$

In this case T_i follows a log-normal distribution with mean $\mu_i + \sigma_w^2/2$ and variance σ_b^2. This fully specifies the usual intake distribution, e.g. the mean and variance of the usual intake are given by

$$\mu_{iT} = \mathrm{E}[T_i] = \exp(\mu_i + \sigma_w^2/2 + \sigma_b^2/2)$$
$$\sigma_{iT}^2 = \mathrm{Var}[T_i] = \left[\exp(\sigma_b^2) - 1\right] \exp(2\mu_i + \sigma_w^2 + \sigma_b^2)$$

Power transformation

For the power transformation the integral can be approximated by means of N-point Gauss-Hermite integration, see Appendix A. This results in the following usual intake

$$T_i \approx \frac{1}{\sqrt{\pi}} \sum_{j=1}^{N} w_j (\mu_i + b_i + \sqrt{2}\sigma_w x_j)^p$$

with *p* the inverse of the power transformation. A similar approximation can be used for the Box-Cox transformation. There can be a small problem with Gauss-Hermite integration. The summation term

$\left(\mu_i + b_i + \sqrt{2}\sigma_w x_j\right)^p$ can not be calculated when the factor between round brackets is negative and the power p is not an integer. This can happen when $(\mu_i + b_i)$ is small relative to the between day standard error σ_w. In that case the corresponding term is set to zero. This is not a flaw in the numerical method but in the statistical model since the model allows negative intakes on the transformed scale which cannot be transformed back to the natural scale.
The mean and variance of T_i can be approximated again by using Gauss-Hermite integration:

$$\mu_{iT} = \mathrm{E}[T_i] = \frac{1}{\sqrt{\pi}}\sum_{k=1}^{N} w_k \frac{1}{\sqrt{\pi}}\sum_{j=1}^{N} w_j\left(\mu_i + \sqrt{2}\sigma_w x_j + \sqrt{2}\sigma_b x_k\right)$$

$$\sigma^2_{iT} = \mathrm{Var}[T_i] = \frac{1}{\sqrt{\pi}}\sum_{k=1}^{N} w_k \left[\frac{1}{\sqrt{\pi}}\sum_{j=1}^{N} w_j\left(\mu_i + \sqrt{2}\sigma_w x_j + \sqrt{2}\sigma_b x_k\right)\right]^2 - \mu_T^2$$

An alternative method for obtaining model based usual intakes for the power transformation employs a Taylor series expansion for the power, see e.g. Kipnis (2009). This is however less accurate than Gauss-Hermite integration. For the power transformation simulation is required to derive the usual intake distribution: simulate a random effect b_i for many individuals and then approximate T_i for these individuals. The T_i values then form a sample form the usual intake distribution.

15.5.1.3 *Model assisted usual intake*

The model assisted approach employs a prediction for the usual intakes of every individual in the study. This requires a prediction of the individual random effect b_i for every individual.

Model assisted usual intake on the transformed scale

In the one-way random effects model the Best Linear Unbiased Prediction for $(\mu_i + b_i)$ is given by

$$\mathrm{BLUP}_i = \mu_i + (\bar{Y}_i^* - \mu_i)\left(\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2/n_i}\right)$$

in which $\bar{Y}_i^*$ is the mean of the transformed intakes for individual *i*. BLUPs have optimal properties for some purposes, but not for the purpose of representing the variation σ_b^2 between individuals. This can be seen by noting that

$$\mathrm{Var}(\bar{Y}_i^*) = \sigma_b^2 + \sigma_w^2/n_i \quad \text{and thus} \quad \mathrm{Var}(\mathrm{BLUP}_i) = \left(\frac{\sigma_b^4}{\sigma_b^2 + \sigma_w^2/n_i}\right)$$

which is smaller than the between individual variance σ_b^2. As an alternative a modified BLUP can be defined by means of

$$\text{modified BLUP}_i = \mu_i + (\bar{Y}_i^* - \mu_i)\sqrt{\left(\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2/n_i}\right)}$$

which has the correct variance σ_b^2 and also the correct mean μ_i. However these optimal properties disappear when modified BLUPs are directly backtransformed to the original scale.

Logarithmic transformation

For the logarithmic transformation the usual intake T_i follows a log-normal distribution with mean $\mu_i + \sigma_w^2/2$ and variance σ_b^2. If we can construct a BLUP like stochastic variable with the same mean and variance, then this variable be an unbiased predictor with the correct variance. It is easy to see that the following variable has the same distribution as T_i

$$\text{model assisted BLUP}_i = \mu_i + \frac{\sigma_w^2}{2} + (\bar{Y}_i^* - \mu_i)\sqrt{\left(\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2/n_i}\right)}$$

So the model assisted individual intake Exp(model assisted BLUP_i) has the same distribution as the usual intake and is thus the best predictor for usual intake.

Kipnis *et al*. (2009) employs the conditional distribution of b_i given the observations $Y_{i1}, \ldots, Y_{in_i}$ to obtain a prediction. First note that $(b_i \mid Y_{i1}, \ldots, Y_{in_i}) = (b_i \mid Y_{i1}^*, \ldots, Y_{in_i}^*) = (b_i \mid \bar{Y}_i^*)$. Since all distributions in the one-way random effects model are normal it follows that:

$$(b_i, \bar{Y}_i^*) \sim \text{BivariateNormal}(0, \mu_i, \sigma_b^2, \sigma_b^2 + \sigma_w^2/n_i, \sigma_b^2)$$

where the last parameter represents the covariance between b_i and $\bar{Y}_i^*$. It follows that the conditional distribution

$$(b_i \mid \bar{Y}_i^*) \sim N(\mu_c, \sigma_c^2) \quad \text{with} \quad \mu_c = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2/n_i}(\bar{Y}_i^* - \mu_i) \quad \text{and} \quad \sigma_c^2 = \frac{\sigma_b^2 \ \sigma_w^2/n_i}{\sigma_b^2 + \sigma_w^2/n_i}$$

A prediction for the usual intake $T_i = F(b_i)$ is then obtained by the expectation

$$\mathrm{E}[F(b_i) \mid \bar{Y}_i^*] = \int F(b)\, \phi(b;\, \mu_c,\, \sigma_c^2)\, db$$

For the logarithmic transform $F(b_i) = \exp(\mu_i + b_i + \sigma_w^2/2)$ and the expectation reduces to

$$\mathrm{E}[F(b_i) \mid \bar{Y}_i^*] = \exp(\mu_i + \mu_c + \sigma_c^2/2 + \sigma_w^2/2)$$

which is a function of $\bar{Y}_i^*$ through μ_c.To obtain the mean and variance of the prediction note that

$$\mu_i + \mu_c + \sigma_c^2/2 + \sigma_w^2/2 \sim N\left(\mu_i + \frac{\sigma_b^2 \ \sigma_w^2/n_i}{2(\sigma_b^2 + \sigma_w^2/n_i)} + \frac{\sigma_w^2}{2}, \qquad \frac{\sigma_b^4}{\sigma_b^2 + \sigma_w^2/n_i}\right)$$

It follows that the expectation of the prediction equals

$$\begin{aligned}\mathrm{E}[E[F(b_i) \mid \bar{Y}_i^*]] &= \exp\left(\mu_i + \frac{\sigma_b^2 \ \sigma_w^2/n_i}{2(\sigma_b^2 + \sigma_w^2/n_i)} + \frac{\sigma_w^2}{2} + \frac{\sigma_b^4}{2(\sigma_b^2 + \sigma_w^2/n_i)}\right) \\ &= \exp\left(\mu_i + \frac{\sigma_b^2}{2} + \frac{\sigma_w^2}{2}\right)\end{aligned}$$

which equals the mean of the usual intake. However the variance of the prediction equals

$$\mathrm{Var}[E[F(b_i) \mid \bar{Y}_i^*]] = \left[\exp\left(\frac{\sigma_b^4}{\sigma_b^2 + \sigma_w^2/n_i}\right) - 1\right] \exp\left(2\mu_i + \sigma_b^2 + \sigma_w^2\right)$$

Which is less than the variance of the usual intake. The approach of Kipnis et al (2009) will therefor result in too much shrinkage of the model assisted usual intake.

Power transformation

For the power transformation a model assisted BLUP with optimal properties, as derived above, cannot be constructed. The approach of Kipnis *et al.* (2009) can however be used to obtain a prediction in the following way. First approximate $T_i = F(b_i)$ by Gauss-Hermite integration:

$$F(b_i) = T_i \approx \frac{1}{\sqrt{\pi}} \sum_{j=1}^{N} w_j \left(\mu_i + b_i + \sqrt{2}\sigma_w x_j\right)^p$$

Secondly again use Gauss-Hermite to approximate the expectation of the conditional distribution giving the prediction P_i.

$$P_i = \mathrm{E}[F(b_i) \mid \bar{Y}_i^*] = \int F(b)\phi(b;\, \mu_c, \sigma_c^2)\, db \approx \frac{1}{\pi} \sum_{k=1}^{N} w_k \sum_{j=1}^{N} w_j \left(\mu_i + \mu_c + \sqrt{2}\sigma_w x_j + \sqrt{2}\sigma_c x_k\right)^p$$

which is a function of $\bar{Y}_i^*$ through μ_c. It is likely that the thus obtained predictions P_i have a variance that is too small. If we would know the mean μ_{iP} and variance σ_{iP}^2 of the predictions, the predictions could be linearly rescaled to have the correct mean μ_{iT} and variance σ_{iT}^2. The mean and variance of the prediction can be calculated using Gauss-Hermite integration:

$$\mu_{iP} = \frac{1}{\sqrt{\pi}} \sum_{l=1}^{N} w_l \frac{1}{\pi} \sum_{k=1}^{N} w_k \sum_{j=1}^{N} w_j \left(\mu_i + \sqrt{2}\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2/n_i} x_l + \sqrt{2}\sigma_w x_j + \sqrt{2}\sigma_c x_k\right)^p$$

$$\sigma_{iP}^2 = \frac{1}{\sqrt{\pi}} \sum_{l=1}^{N} w_l \left[\frac{1}{\pi} \sum_{k=1}^{N} w_k \sum_{j=1}^{N} w_j \left(\mu_i + \sqrt{2}\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2/n_i} x_l + \sqrt{2}\sigma_w x_j + \sqrt{2}\sigma_c x_k\right)^p\right]^2 - \mu_{iP}^2$$

The proposed prediction then equals

$$P_i^* = \mu_{iT} + \frac{\sigma_{iT}}{\sigma_{iP}}(P_i - \mu_{iP})$$

15.5.2 Episodically consumed foods

For episodically consumed foods we need to take the probability of consumption into account. Define p_i as the probability that individual *i* consumes the food on any given day. The usual intake for this individual is then given by the product of p_i and T_i which is now defined as the usual amount on consumption days. Since individuals will vary in their probability p_i, besides modelling the amounts as for daily consumed foods, it is also necessary to model the frequency of consumption. A three stage analysis of 24-hour recall data is the necessary:

1. A model for the frequency of consumption
2. A model for the intakes on consumption days
3. Integration of both models in order to obtain a usual intake distribution.

Step 2 uses the analysis outlined in the previous section for the positive intakes only. For step 1 two popular models which describe between-individual variation for the probability of consumption are the beta-binomial model and the logistic-normal model.

15.5.2.1 *Beta-Binomial model for frequencies (BBN)*

Let n_i be the total number of recall days for individual *i* and X_i the number of days with a positive intake. The distribution of X_i, with p_i the probability of consumption for individual *i*, is given by

$X_i \sim \text{Binomial}(n_i, p_i)$

In this model the probability p_i varies among individuals according to the Beta distribution:

$$f(p) = B^{-1}(\alpha, \beta)\, p^{\alpha-1}\,(1-p)^{\beta-1} \quad \text{with} \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Combining the binomial and the Beta distribution results in the betabinomial distribution:

$$P(X_i = x) = \binom{n_i}{x} \frac{B(\alpha + x, n_i + \beta - x)}{B(\alpha, \beta)}$$

The mean and variance of the betabinomial distribution are given by

$$\text{E}[X_i] = n_i \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{Var}[X_i] = n_i \frac{\alpha\beta(\alpha+\beta+n_i)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Using the reparameterization $\pi = \alpha/(\alpha+\beta)$ and $\varphi = 1/(\alpha+\beta+1)$, it follows that

$$\text{E}[X_i] = n_i\pi \quad \text{and} \quad \text{Var}[X_i] = n_i\pi(1-\pi)[1 + (n_i - 1)\varphi]$$

This reparameterization enables to model the probability π_i of consumption for individual *i* directly as a logistic regression:

$$\text{logit}(\pi_i) = \gamma_0 + \boldsymbol{\gamma}_1^t\, \boldsymbol{Z}_i$$

Note that the dispersion parameter φ is assumed to be equal for all individuals. The betabinomial logistic regression model can be fitted by means of maximum likelihood.

Model based frequencies for usual intake

For the model based usual intake distribution the estimated parameters π_i and φ are backtransformed using $\alpha_i = \pi_i\,\varphi/(1-\varphi)$ and $\beta_i = (1-\pi_i)\,\varphi/(1-\varphi)$. These can then be used to draw from the Beta distribution.

Model assisted frequencies for usual intake

For the model assisted usual intake distribution a prediction of the consumption probability is required for every individual. Simple predictions are (a) the observed frequencies for every individual or (b) the fitted probability for evey individual. When there are no covariables the fitted probability is the same for every individual. Alternatively (c) one can use the approach outlined in Kipnis et al (2009) employing the conditional expectation of the probability given the observed frequency:

$$\text{E}(p_i \mid X_i = x) = \int_p p\, f(p \mid X_i = x)\, dp = \int_p p\, \frac{f(X_i = x \mid p)\, f(p)}{\int f(X_i = x \mid p)\, f(p)\, dp}\, dp =$$

$$= \frac{1}{P(x_i = x)} \int_p p \binom{n_i}{x} p^x (1-p)^{n_i - x} B^{-1}(\alpha_i, \beta_i) p^{\alpha_i - 1} (1-p)^{\beta_i - 1} dp$$

$$= \frac{B^{-1}(\alpha_i, \beta_i)}{P(x_i = x)} \binom{n_i}{x} \int_p p^{\alpha_i + x} (1-p)^{n_i + \beta_i - x - 1} dp = \frac{B(\alpha_i + x + 1,\ n_i + \beta_i - x)}{B(\alpha_i + x,\ n_i + \beta_i - x)}$$

$$= \frac{\alpha_i + x}{\alpha_i + \beta_i + n_i}$$

For individual with zero intakes on all recall days a prediction for the random individual amount effect b_i is not available. There seem to be two option for predicting the usual intake for such individuals:

- Set the individual intake to zero
- Simulate a model based prediction for the amount and combine this with the conditional expected probability given above to obtain an individual usual intake.

15.5.2.2 *Logistic-Normal model for frequencies (LNN0)*

In this model the distribution of X_i is again binomial:

$X_i \sim \text{Binomial}(n_i, p_i)$

The probability p_i is now given by a logistic regression with a random effect in the linear predictor which represents the between-individual variation in the probability p_i:

$\text{logit}(p_i) = \lambda_i + v_i$ with $v_i \sim \text{N}(0, \sigma_v^2)$ and the regression equation $\lambda_i = \gamma_0 + \boldsymbol{\gamma}_1^t \boldsymbol{Z}_i$

The marginal probability π_i is obtained by integrating over the random effect v_i, i.e. using Gauss-Hermite integration

$$\pi_i = \int H(\lambda_i + v) f(v) dv \approx \frac{1}{\sqrt{\pi}} \sum_{j=1}^{N} w_j H\left(\lambda_i + \sqrt{2}\sigma_v x_j\right)$$

in which $H()$ is the inverse of the logit transformation. Note that this is different from $\text{logit}^{-1}(\lambda_i)$ which is the median probability. The model can be fitted by maximum likelihood using Gauss-Hermite integration. An (approximate) maximum likelihood procedure is implemented in routine glmer of the lme4 package in R.

For a new vector of covariates $\boldsymbol{Z}_i^*$ the linear predictor λ_i^* can be calculated along with its standard error $\text{Se}(\lambda_i^*)$. The marginal predicted probability π_i^* can be calculated by means of Gauss-Hermite integration and the standard error of the predicted probability can be calculated by means of the usual Taylor series expansion:

$$\text{Se}(\pi_i^*) \approx \frac{\text{Se}(\lambda_i^*)}{\sqrt{\pi}} \sum_{j=1}^{N} w_j \frac{d}{d\lambda_i^*} H\left(\lambda_i^* + \sqrt{2}\sigma_v x_j\right)$$

$$= \frac{\text{Se}(\lambda_i^*)}{\sqrt{\pi}} \sum_{j=1}^{N} w_j H\left(\lambda_i^* + \sqrt{2}\sigma_v x_j\right)\left[1 - H\left(\lambda_i^* + \sqrt{2}\sigma_v x_j\right)\right]$$

Model based frequencies for usual intake

For the model based usual intake distribution the estimated parameters λ_i and σ_v^2 can be used to generate individual probabilities.

Model assisted frequencies for usual intake

For the model assisted usual intake distribution simple predictors are (a) the observed frequencies and (b) the marginal probability π_i. The conditional expectation (c) is given by

$$\text{E}(p_i | X_i = x_i) = \int_v H(\lambda_i + v)\, f(v \mid X_i = x_i)\, dv = \int_v H(\lambda_i + v)\, \frac{f(X_i = x_i \mid v) f(v)}{\int f(X_i = x_i \mid v) f(v)\, dv}\, dv$$

$$=$$

$$= \frac{\int_v H(\lambda_i + v)\,[H(\lambda_i + v)]^{x_i}\,[1 - H(\lambda_i + v)]^{n_i - x_i}\, f(v)\, dv}{\int_v [H(\lambda_i + v)]^{x_i}\,[1 - H(\lambda_i + v)]^{n_i - x_i}\, f(v)\, dv}$$

and both nominator and denominator can be approximated by means of the Gauss-Hermite integration. For individual with zero intakes on all recall days see above for the two options.

15.5.2.3 *Logistic-Normal model for frequencies correlated with amounts (LNN)*

This model is extends the LNN0 model with a correlation between the individual random effect b_i for amounts and the individual random effect v_i for frequencies. This model is also known as the NCI model and is introduced by Tooze et al (2006) with further mathematical details in Kipnis et al (2009). The model can be written as

$\text{logit}(P(Y_{ij} > 0) = \lambda_i + v_i$

$g(Y_{ij}) = \mu_i + b_i + w_{ij}$

$(v_i, b_i) \sim \text{BivariateNormal}\left(0, 0, \sigma_v^2, \sigma_b^2, \rho\right)$ and $w_{ij} \sim \text{N}(0, \sigma_w^2)$

The model can be fitted by maximum likelihood employing two-dimensional Gauss-Hermite integration as detailed in Appendix A.

Model based usual intake

Model based usual intake requires generation of the pair (v_i, b_i) for many hypothetical individual. The usual intake U_i for such a hypothetical individual is then given by

$$U_i = H(\lambda_i + v_i)\, T_i = H(\lambda_i + v_i)\, E_w\left[g^{-1}\left(\mu_i + b_i + w_{ij}\right) \mid b_i\right] = H(\lambda_i + v_i)\, F(b_i)$$

The second term can be calculated using the method outlined for daily intakes.

Model assisted usual intake

This requires simulatenous prediction of the random effect for frequency and for amount as outlined in Kipnis et al (2009). We have for individual *i* in the study $(U_i \mid Y_{i1}, \ldots, Y_{in_i}) = (U_i \mid Y_{i1}^*, \ldots, Y_{in_i}^*) = (U_i \mid x_i, \bar{Y}_i^*)$ where x_i is the number of positive intakes and $\bar{Y}_i^*$ is the mean of the transformed **positive** intakes. It follows that the required conditional expectation P_i equals

$$P_i = \text{E}[U_i \mid x_i, \bar{Y}_i^*] = E_{v_i, b_i}[H(\lambda_i + v_i)\, F(b_i) \mid x_i, \bar{Y}_i^*] =$$

$$= \frac{\iint H(\lambda_i + v_i)\, F(b_i)\, f(x_i, \bar{Y}_i^* \mid v_i, b_i)\, \phi(v_i, b_i)\, dv_i\, db_i}{\iint f\left(x_i, \bar{Y}_i^* \mid v_i, b_i\right) \phi(v_i, b_i)\, dv_i\, db_i}$$

Where

$$f(x_i, \bar{Y}_i^* \mid v_i, b_i) = [H(\lambda_i + v_i)]^{x_i}\,[1 - H(\lambda_i + v_i)]^{n_i - x_i}\, \phi(\bar{Y}_i^* - \mu_i - b_i; 0, \sigma_w^2 / x_i)$$

Both nominator and denominator can be approximated by two-dimensional Gauss-Hermite integration. Note that for the log-transform $F(b_i) = T_i = \exp(\mu_i + b_i + \sigma_w^2/2)$ can be calculated exactly; for the power transformation an approximation must be used. It can be expected that the predicted usual intake will not have the correct variance. This can possibly be remedied by equating the mean and variance of U_i and P_i. These are however rather involved to calculate.

For individual with zero intakes on all recall days the model assisted usual intake can be set to zero, or can be simulated as follows

1. Calculate the Model assistefrequency P_0 for usual intake (see LNN0)
2. Transform P_0 back to the logistic scale, i.e. $L_0 = \text{logit}(P_0)$. Get the conditional distribution of $(b \mid v = L_0 - \lambda_i) \sim \text{N}\left(\frac{\sigma_b}{\sigma_v}\rho(L_0 - \lambda_i), (1 - \rho^2)\sigma_b^2\right)$
3. Simulate a draw b_0 from this conditional distribution and obtain the usual intake as $P_0 \exp(\mu_i + b_0 + \sigma_w^2)$

Note that the backtransformation from P_0 to L_0 is according to the median of the distribution rather than the mean.

15.5.3 Gauss-Hermite integration

15.5.3.1 *One-dimensional Gauss-Hermite integration*

Gauss-Hermite integration approximates a specific integral as follows

$$\int_{-\infty}^{\infty} f(x)\exp(-x^2)\ dx \approx \sum_{j=1}^{N} w_j\, f(x_j)$$

in which w_j and x_j are weights and abscissas for N-point Gauss-Hermite integration, see Abramowitz and Stegun (1972). N-point integration is exact for all polynomials $f(x)$ of degree 2N-1, see Dahlquist and Björck (1974). This can for instance be used to approximate the mean of a function $F(Y)$ of a normally distributed random variable Y with mean μ and variance σ^2:

$$\int_{-\infty}^{\infty} F(y)\ \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy\ = \int_{-\infty}^{\infty} F\left(\mu+\sqrt{2}\sigma\, x\right)\frac{1}{\sqrt{\pi}}\exp(-x^2)\ dx$$

$$= \frac{1}{\sqrt{\pi}}\sum_{j=1}^{N} w_j\, F\left(\mu+\sqrt{2}\sigma\, x_j\right)$$

15.5.3.2 *Two-dimensional Gauss-Hermite integration*

One-dimensional Gauss-Hermite integration can readily be extended to two dimensions. The following principal result in two dimensions is more or less given in Jäckel (2005) for the standard bivariate normal distribution $\phi(x, y;\ \rho)$ with correlation parameter ρ:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x,y)\ \phi(x,y;\rho)\ dx\ dy \approx \frac{1}{\pi}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j\, F\left(\sqrt{2}\left[ax_i+bx_j\right],\sqrt{2}\left[bx_i+ax_j\right]\right)$$

in which

$a = \frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}$ and $b = \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$ as given in Jäckel (2005).

Jäckel (2005) discusses other Gauss-Hermite approximations to the two-dimensional integral, but found that the approximation given above generally gives the most accurate results. For the general bivariate normal distribution with means (μ_x, μ_y) and variances (σ_x^2, σ_y^2) the integral can be approximated by means of

$$\frac{1}{\pi}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j\, F\left(\mu_x+\sigma_x\sqrt{2}\left[ax_i+bx_j\right],\ \mu_y+\sigma_y\sqrt{2}\left[bx_i+ax_j\right]\right)$$

The product $w_i w_j$ can be very small, especially when many quadrature points are used, thus wasting possibly precious calculation time. This can be remedied by pruning, i.e. by dropping combinations of (i, j) with very small values of the product $w_i w_j$.

15.5.3.3 *Maximum likelihood for the LNN model with two-dimensional Gauss-Hermite integration*

Denote non-consumption on day *j* for individual *i* as $Y_{ij} = 0$. The conditional likelihood, i.e. given random effects b_i and v_i, of a non-consumption on day *j* equals, with $H()$ the inverse of the logit function

$$P\left(Y_{ij} = 0 \mid b_i, v_i\right)\ = 1 - H(\lambda\ +\ v_i).$$

The conditional likelihood of a positive intake $Y_{ij} > 0$ equals, with ϕ the density of the normal distribution

$$f\left(Y_{ij} = y_{ij} \mid y_{ij} > 0, b_i, v_i\right) = H(\lambda + v_i)\ \phi\left(y_{ij} - \mu - b_i; 0, \sigma_w^2\right)$$

The conditional likelihood contribution for individual *i* is the product of the individual contributions for each day. The marginal likelihood contribution for individual *i* is obtained by integrating over the possible values of b_i and v_i. Since the pair (b_i, v_i) follows a bivariate normal distribution, the likelihood contribution for individual *i* can be approximated by means of two-dimensional Gauss-Hermite integration. Individually based covariabels, such as sexe or age, imply that μ_i and λ_i must be used instead of μ and λ. The likelihood must be optimized by means of some general optimization routine.

15.6 Modelling acute exposures as function of covariates

An acute risk assessment may be followed by an analysis where the acute intake distribution is related to a covariable and/or cofactor.

15.6.1 Intake frequency model

Let n_i and $npos_i$ be the total number of simulated intakes per individual, and the number of simulated positive intakes, respectively. Then $npos_i$ is modelled as a function of *e.g.* age (and/or other individual characteristics), using a betabinomial distribution with binomial totals n_i and overdispersion parameter ϕ (independent of age). The fitted binomial probabilities are $\hat{\pi}_x = f(x_i)$, where x_i is the age of individual *i*, and the estimated overdispersion parameter is $\hat{\phi}$.

15.6.2 Intake amount model

For the positive intakes, consider power of logarithmically transformed values y_{ijk}. Average over replicates to obtain individual day averages $y_{ij.}$. These values are modelled in a ML analysis with random terms individual and individual.day as a function of age (and/or other individual characteristics), with the number of values per individual day (n_{ij}) as weights w_{ij} to correct for differences in the precision at the individual day stratum. The fitted values from the model are $\hat{\mu}_x = f(x_i)$, where x_i is the age of individual *i.*

15.6.3 Estimating the acute risk variability of positive intake amounts

Correct the full set of simulated positive intakes by $y'_{ijk} = y_{ijk} - \hat{\mu}_{x(i)}$. Estimate the variance $\sigma^2_{y'}$ of y'_{ijk}. We denote the estimated variance as $\hat{\sigma}^2_{y'}$. Now for each selected age *x* the transformed positive intake distribution is modelled as normal with mean $\hat{\mu}_x = f(x)$ and variance $\hat{\sigma}^2_{y'}$.

15.6.4 Estimating the acute intake distribution

Acute intake distributions dependent on a covariate are obtained by numerical integration. For each combination of levels of the covariable and/or cofactor, intake frequency values and transformed intake amounts are simulated and multiplied. This results in a number of distributions each one representing the acute intake distribution corresponding to a specific combination of levels of the covariates.

15.7 Screening models for large Cumulative Assessment Groups

15.7.1 Statistical model for the screening step (acute exposure)

The screening step implements a simple model that is applied to each SCC. Assume independent NonDetectSpike-LogNormal (NDS-LN) models for both the consumptions of food-as-measured in source S and the concentrations of compound C in source S. A non-detect consumption is assumed to be a zero consumption. A non-detect concentration will be imputed by a user-specified fraction f of the Limit of Reporting. Then the model for consumption has 3 parameters and the model for concentration has four parameters, as specified in Table 26. Note that the parameters of the

consumption distribution are estimated from the consumption data using sampling weights if these have been provided in the consumption data set.

Table 26. Parameters for screening models (per source/compound)

parameter	consumptions	concentrations
probability of a positive	π_x	π_c
mean positives (ln scale)	μ_x	μ_c
standard deviation positives (ln scale)	σ_x	σ_c
value to use for NonDetects (ln scale)		$f * L_c$

Exposure is consumption times concentration, so on logarithmic scale they can be added

$$e = x + c$$

The assessment will focus on a chosen percentile of exposure, e.g. p95. The relevant fraction will be denoted by p, for example $p = 0.95$ for the 95th percentile.
The two NDS-LN models combine to three possibilities, depending on whether there is consumption and if so, whether the concentration is non-detect or positive. In the screening model the two possibilities that lead to potential exposure are modelled with a mixture of two lognormal distribution. For the non-detect case the positive exposure distribution equals the positive consumption distribution modified by the multiplication of a user-chosen factor times an estimate of the average worst-case limit value for concentration (LOR):

$$\pi_1 = \pi_x(1 - \pi_c);\ \ \mu_1 = \mu_x + fL_c;\ \ \sigma_1 = \sigma_x$$

where L_c is the logarithm of the LOR, or, if there are multiple analytical methods with different LOR, a weighted average of these different LORs.
For the detect case the positive exposure distribution is easily combined from the positive consumption distribution and the positive concentration distribution:

$$\pi_2 = \pi_x\pi_c;\ \ \mu_2 = \mu_x + \mu_c;\ \ \sigma_2 = \sqrt{\sigma_x^2 + \sigma_c^2}$$

p can be corrected for the non-consumptions to the appropriate fraction needed in the mixture of the two positive distributions:

$$p' = \frac{p - (1 - \pi_x)}{\pi_x}$$

If $p' \leq 0$ then all positive exposures are beyond the requested fraction, and the estimated exposure is just 0.
If $p' > 0$ then the relevant log exposure e_p satifies

$$(1 - \pi_c) \cdot \Phi\left(\frac{e_p - \mu_1}{\sigma_1}\right) + \pi_c \cdot \Phi\left(\frac{e_p - \mu_2}{\sigma_2}\right) = p'$$

where $\Phi(\cdot)$ represents the cumulative standard normal distribution function. The value of e_p can easily be found in a bisection search within the interval $\left[\mu_{min} - 4\sigma_{max},\ \ \mu_{max} + \max\left(0, z_{p'}\sigma_{max}\right)\right]$. The final exposure percentile estimate then is $\exp\left(e_p\right)$.
Denote by $e_{p,max}$ the highest estimate (for the SCC denoted by $SCC_{highest}$). Then evaluate for each SCC the probability to exceed $e_{p,max}$.

$$P_i = Pr(e > e_{p,max}) = \pi_x \cdot \left[(1-\pi_c) \cdot \Phi\left(\frac{e_{p,max}-\mu_1}{\sigma_1}\right) + \pi_c \cdot \Phi\left(\frac{e_{p,max}-\mu_2}{\sigma_2}\right)\right]$$

P_i is a tentative measure for the 'probability of a high exposure'. For $SCC_{highest}$ $P_i = 1-p$, for all other SCCs it will be lower. The sum of all these probabilities is not a meaningful probability in itself. However, this sum is used to scale the individual P_i values to measures of relative importance for the SCCs

$$Imp_i = P_i / \Sigma P_i$$

Rank all SCCs according to Imp_i and calculate cumulative importance.
The relative importance of the two mixture components at e_p can be estimated as

$$w_{1,2} = \frac{\pi_{1,2} \cdot \phi\left(\frac{e_p-\mu_{1,2}}{\sigma_{1,2}}\right)/\sigma_{1,2}}{\pi_1 \cdot \phi\left(\frac{e_p-\mu_1}{\sigma_1}\right)/\sigma_1 + \pi_2 \cdot \phi\left(\frac{e_p-\mu_2}{\sigma_2}\right)/\sigma_2}$$

where $\phi(\cdot)$ represent the standard normal probability density function.
The user interface should allow to select the top-N SCCs from the list, based on a chosen percentage (e.g. 95%) of cumulative importance included.
The full analysis will calculate exactly the same exposure distribution as a full analysis without screening. However, less information is retained in the output. This concerns tables with information on foods-as-eaten, which is only shown for the selected risk driver components (SCCs).
Risk drivers are groupings of SCCs (risk driver components) at the level of measured-source-compound combinations (MSCCs). Note that output for an MSSC (e.g. APPLE/captan) only covers the selected SCCs (e.g. APPLE from apple juice/captan and APPLE from apple pie/captan), but not unselected SCCs (e.g. APPLE from fruit yoghurt/captan).

15.7.2 Statistical model for the screening step (chronic exposure)

In chronic exposure assessments, the mean concentration of chemicals is calculated first, and combined with the consumption distribution. For this reason a chronic calculation uses less memory, and therefore larger datasets can be handled.
The model described under Acute can be simplified for a chronic screening. The concentration distribution is only used to estimate a mean exposure, incorporating any effect from the imputation of non-detects. The exposure distribution is therefore only a scaled version of the consumption distribution.

$$\pi_2 = \pi_x \pi_c; \;\; \mu_2 = \mu_x + \mu_c; \;\; \sigma_2 = \sigma_x$$

The parameters of the consumption distribution ($\pi_x, \; \mu_x, \sigma_x$) are calculated from the observed individual means (OIMs), i.e. the mean daily consumptions over the survey days of each person in the data (allowing for sampling weights). The percentiles are calculated as $e_p = \mu_2 + z_p, \cdot \sigma_2$ where z is a percentile of the standard normal distribution. The exceedances of the maximum percentile are calculated as

$$P_i = Pr(e > e_{p,max}) = \pi_x \cdot \Phi\left(\frac{e_{p,max}-\mu_2}{\sigma_2}\right)$$

15.8 Parametric uncertainty

According to Cochran's theorem, sample variance $\hat{\sigma}_y^2$ follows a scaled chi-square distribution. In the parametric bootstrap for the lognormal distribution, the sample variance $\hat{\sigma}_y^2$ is replaced by a random draw from a chi-square distribution with n_1 *-1* degrees of freedom; the sample mean $\hat{\mu}_y$ is replaced by

a random draw from a normal distribution with parameters $\hat{\mu}_y$ and $\hat{\sigma}_y^{*2}/n_1$, giving a new set of parameters $\hat{\mu}_y^*$, $\hat{\sigma}_y^{*2}$. This is repeated *B* times.

For the truncated lognormal and censored lognormal, large sample maximum likelihood theory is used to derive new parameters $\hat{\mu}_y^*$ and $\hat{\sigma}_y^{*2}$. This is repeated *B* times.

The binomial fraction of non-detects for the mixture lognormal and mixture truncated distribution is sampled using the *beta* distribution with uniform priors *a = b = 1 (*with the *beta* distribution as the empirical Bayes estimator for the binomial distribution). This is repeated *B* times.

15.9 Uncertainty in aggregate exposure assessment (advanced use case)

Example: Probabilistic (variability and uncertainty) cumulative non-dietary exposure input (matched to dietary survey individuals). Internal dose.

Table 27: NonDietaryExposures

idIndividual	idNonDietary Survey	idCompound	Dermal	Oral	Inhalation
5432	1	011003001	10	5	17
5432	1	011003002	34	20	18
5433	1	011003001	11	6	15
5433	1	011003002	35	22	16

Table 28: NonDietaryExposuresUncertain

idIndividual	idNonDietarySurvey	id	idCompound	Dermal	Oral	Inhalation
5432	1	1	011003001	12	7	18
5432	1	1	011003002	40	22	19
5432	1	2	011003001	18	9	19
5432	1	2	011003002	45	24	20
5433	1	1	011003001	13	8	19
5433	1	1	011003002	42	21	18
5433	1	2	…	…	...	…

Table 29: NonDietarySurveys

idNonDietarySurvey	Description	Location	Date	NonDietaryIntakeUnit
1	BROWSE, acute, cumulative, operators	York	09/10/2012	µg/day

The NonDietaryExposuresUncertain table is similar to the table given in the previous example, except that the user has specified unique id numbers for each alternative uncertainty realisation.

In this example, the user has exposure to two compounds, 011003001 and 011003002. For every uncertainty iteration, all individuals will therefore exposures from both. The exposures to the two compounds will be distributed according to the particular distribution used during the simulation to generate these values (*e.g.* independence or correlation should be included in the simulation as required).

Individuals are assigned a different two-compound exposure value pair per uncertainty iteration.

16 References

Béchaux C, Zetlaoui M, Tressou J, Leblanc JC, Héraud F, Crépet A. 2013. Identification of pesticides mixtures and connection between combined exposure and diet. Food Chemical Toxicology 59: 191-198.

Blom, G. (1958). Statistical estimates and transformed beta-variables. Wiley, New York

de Boer, W.J., van der Voet, H. (2011). MCRA 7. A web-based program for Monte Carlo Risk Assessment. Reference Manual 2011-12-19, documenting MCRA release 7.1 Reportnr: Biometris, Wageningen UR and National Institute for Public Health and the Environment (RIVM), Bilthoven, Wageningen. Available online: https//mcra.rivm.nl .

de Boer, W.J., van der Voet, H., Bokkers B.G.H., Bakker, M.I., Boon, P.E. (2009). Comparison of two models for the estimation of usual intake adressing zero consumption and non-normality. Food Additives and Contaminants. Part A, 26:11,1433 – 1449.

Boon PE, van Donkersgoed G, Christodoulou D, Crépet A, D'Addezio L, Desvignes V, Ericsson BG, Galimberti F, Loannou- Kakouri E, Hamborg Jensen B, Rehurkova I, Rety J, Ruprich J, Sand S, Stephenson C, Strömberg A, Turrini A, van der Voet H, Ziegler P, Hamey P, van Klaveren JD (2015). Cumulative dietary exposure to a selected group of pesticides of the triazole group in different European countries according to the EFSA guidance on Probabilistic modelling. Food and Chemical Toxicology, 79: 13-31. http://dx.doi.org/10.1016/j.fct.2014.08.004.

Bopp S, Berggren E, Kienzler A, van der Linden S, Worth A (2015). Scientific methodologies for the assessment of combined effects of chemicals – a survey and literature review. JRC Technical Report. https://ec.europa.eu/jrc/en/publication/scientific-methodologies-assessment-combined-effects-chemicals-survey-and-literature-review.

Box, G.E.P. and Cox, D.R., (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B, 26, 211-243.

Codex (2011). Working Principles for Risk Analysis. In: Codex Alimentarius Commission, Procedural Manual, 20th edition. FAO/WHO, Rome.

Dodd, K.W. (1996). A technical guide to C-SIDE. Technical Report 96-TR 32, Department of Statistics and Center for Agricultural and Rural Development, Iowa State University, Ames, Iowa. Available online: http://www.card.iastate.edu/publications/DBS/PDFFiles/96tr32.pdf.

Edler L, Hart A, Greaves P, Carthew P, Coulet M, Boobis A, Williams GM, Smith B. (2013). Selection of appropriate tumour data sets for Benchmark Dose Modelling (BMD) and derivation of a Margin of Exposure (MoE) for substances that are genotoxic and carcinogenic: Considerations of biological relevance of tumour type, data quality and uncertainty assessment. Food Chem. Toxicol., http://dx.doi.org/10.1016/j.fct.2013.10.030

Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7: 1-26.

Efron, B. and Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman & Hall, New York.

EFSA (2006). Guidance of the Scientific Committee on a request from EFSA related to uncertainties in dietary exposure assessment. EFSA Journal, 438, 1-54.

EFSA (2009). Guidance of the Scientific Committee on Transparency in the Scientific Aspects of Risk Assessments carried out by EFSA. Part 2: General Principles. EFSA Journal 1051, 1-22.

EFSA (2011a). Report on the development of a food classification and description system for exposure assessment and guidance on its implementation and use. EFSA Journal 2011;9(12):2489. [84 pp.] doi:10.2903/j.efsa.2011.2489. Available online: http://www.efsa.europa.eu/en/efsajournal/doc/2489.pdf.

EFSA (2011b). The food classification and description system FoodEx 2 (draft-revision 1). Supporting Publications 2011:215. [438 pp.]. Available online: http://www.efsa.europa.eu/en/supporting/doc/215e.pdf.

EFSA (2012). Guidance on the Use of Probabilistic Methodology for Modelling Dietary Exposure to Pesticide Residues. EFSA Journal 2012;10(10):2839. [95+47 pp.] doi:10.2903/j.efsa.2012.2839. Available online: http://www.efsa.europa.eu/en/efsajournal/pub/2839.htm.

EFSA (2013a). Draft Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs – Part: exposure assessment. http://www.efsa.europa.eu/en/consultationsclosed/call/130725.htm

EFSA (2013b). Draft Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. www.efsa.europa.eu/en/consultationsclosed/call/130813.htm
Gillis N, Glineur F. 2010. Using underapproximations for sparse nonnegative matrix factorization. Pattern Recognition 43: 1676-1687.
Gillis N, Plemmons RJ. 2013. Sparse nonnegative matrix underapproximation and its application to hyperspectral image analysis. Linear Algebra and its Applications 438: 3991-4007.
Goedhart, P.W., van der Voet, H., Knüppel, S., Dekkers, A.L.M., Dodd, K.W., Boeing, H., van Klaveren, J.D. (2012). A comparison by simulation of different methods to estimate the usual intake distribution for episodically consumed foods. Report: Supporting Publications 2012:EN-299. [65 pp.]. Available online: www.efsa.europa.eu/publications.
Goodhardt, G.J., Ehrenberg, A.S.C. & Chatfield, C. (1984). The Dirichlet: a comprehensive model of buying behaviour. Journal of the Royal Statistical Society A, 147: 621-655.
Hoyer PO. 2004. Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5: 1457-1469.
Kennedy MC, Glass CR, Bokkers BGH, Hart ADM, Hamey P, Kruisselbrink JW, de Boer WJ, van der Voet H, Garthwaite D, van Klaveren JD (2015a). A European model and case studies for aggregate exposure assessment of pesticides. Food and Chemical Toxicology, 79: 32-44. http://dx.doi.org/10.1016/j.fct.2014.09.009.
Kennedy MC, van der Voet H, Roelofs VJ, Roelofs W, Glass CR, de Boer WJ, Kruisselbrink JW, Hart ADM (2015b). New approaches to uncertainty analysis for use in aggregate and cumulative risk assessment of pesticides. Food and Chemical Toxicology. 79: 54-64. http://dx.doi.org/10.1016/j.fct.2015.02.008.
Kipnis, V., Midthune, D., Buckman, D.W., Dodd, K.W., Guenther PM, Krebs-Smith SM, Subar AF, Tooze JA, Carroll RJ, Freedman LS (2009). Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. Biometrics, 65: 1003-1010.
Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401: 788-791.
Mood, A.M., Graybill, F.A. and Boes, D.A. (1974). Introduction to the theory of statistics, New York, McGraw-Hill.
Nusser ,S.M., Carriquiry, A.L., Dodd, K.W. and Fuller, W.A. (1996). A semi-parametric transformation approach to estimating usual daily intake distributions. Journal of the American Statistical Association, 91: 1440-1449.
Nusser, S.M., Fuller, W.A., and Guenther, P.M. (1997). Estimating usual dietary intake distributions: adjusting for measurement error and nonnormality in 24-hour food intake data. In: Lyberg L., Biemer P., Collins M., DeLeeuw E., Dippo C., Schwartz N., and Trewin D. (editors), Survey Measurement and Process Quality, Wiley, New York. p. 689-709.
Paulo, M., van der Voet, H., Jansen, M.J.W., ter Braak, C.J.F., van Klaveren, J.D., (2005). Risk Assessment of dietary exposure to pesticides using a Bayesian method. Pest Management Science, 61, 759–766.
Price PS, Han X. 2011. Maximum cumulative ratio (MCR) as a tool for assessing the value of performing a cumulative risk assessment. International Journal of Environmental Research and Public Health 8.6: 2212-2225.
Saul LK, Lee DD. 2002. Multiplicative updates for classification by mixtures models. Advances in Neural Information Processing Systems 14.
Slob, W. (2006). Probabilistic dietary exposure assessment taking into account variability in both amount and frequency of consumption. Food Chem Toxicol. 44: 933-951 (Corrigendum in Food Chem Toxicol. 44:1952).
Slob, W., de Boer, W.J. and van der Voet, H. (2010). Can current dietary exposure models handle aggregated intake from different foods? A simulation study for the case of two foods. Food and Chemical Toxicology, 48: 178-186.
Snedecor, G.W. and Cochran, W.G. (1980). Statistical Methods (7th edition). Iowa State University Click, Ames, Iowa.

Souverein, O.W., de Boer, W.J., Geelen, A., van der Voet, H., de Vries, J., Feinberg, M. and van 't Veer, P. (2011). Uncertainty in intake due to portion size estimation in 24-hour recalls varies between food groups. Journal of Nutrition, 141: 1396-1401.

Tooze, J.A., Midthune, D., Dodd, K.W., Freedman, L.S., Krebs-Smith, S.M., Subar, A.F., Guenther, P.M., Carroll, R.J., Kipnis, V. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. J Am Diet Ass. 106: 1575-1587.

Tukey, J.W. (1962). The future of data analysis, Ann. Math. Statist. 33: 1-67 and 812.

van der Voet, H., de Boer, W.J. & Boon, P.E. (2001). Modeling exposure to pesticides. Note HVT-2001-03, Centre for Biometry Wageningen, Wageningen.

van der Voet, H. and Slob, W. (2007). Integration of probabilistic exposure assessment and probabilistic hazard characterization. Risk Analysis, 27: 351-371.

van der Voet, H., van der Heijden, G.W.A.M., Bos, P.M.J., Bosgra, S., Boon, P.E., Muri, S.D., Brüschweiler , B.J. (2009) A model for probabilistic health impact assessment of exposure to food chemicals. Food and Chemical Toxicology 47: 2926-2940.

van der Voet H, Kruisselbrink J, de Boer WJ, Boon PE (2014). Model-Then-Add: Usual intake modelling of multimodal intake distributions. RIVM Letter Report 090133001/2014. http://rivm.openrepository.com/rivm/handle/10029/314361.

van der Voet H, de Boer WJ, Kruisselbrink JW, Goedhart PW, van der Heijden GWAM, Kennedy MC, Boon PE, van Klaveren JD (2015). The MCRA model for probabilistic single-compound and cumulative risk assessment of pesticides. Food and Chemical Toxicology, 79: 5-12. http://dx.doi.org/10.1016/j.fct.2014.10.014.

van Klaveren, J.D., Goedhart, P.W., Wapperom, D., van der Voet, H., 2012. A European tool for usual intake distribution estimation in relation to data collection by EFSA. Report: Supporting Publications 2012:EN-300. Available online: www.efsa.europa.eu/publications.

Verkaik-Kloosterman, J., Dodd, K. W., Dekkers, A. L., van't Veer, P. and Ocké, M.C. (2011). A three-part, mixed-effects model to estimate the habitual total vitamin D intake distribution from food and dietary supplements in Dutch young children. J. Nutr. 141:2055–2063.

Zetlaoui M, Feinberg M, Verger P, Clémencon S. 2011. Extraction of food consumption systems by nonnegative matrix factorization (NMF) for the assessment of food choices. Biometrics 67: 1647-1658.