Grant Agreement No: 245163
Project start date: July 2010
Project end date: July 2013

ACROPOLIS

Aggregate and Cumulative Risk Of Pesticides: an On-Line Integrated Strategy

SEVENTH FRAMEWORK PROGRAMME

Deliverable 5.1: Functional design of the cumulative risk model combining useful algorithms of existing software and new functionality defined in other WPs.

Biometris unit of DLO Plant Research International,
part of Wageningen University and Research centre
Contact person: Hilko van der Voet, email: hilko.vandervoet@wur.nl

Project team: Hilko van der Voet, Waldo J. de Boer, Paul W. Goedhart, Gerie W.A.M. van der Heijden, Jaap Kokorian, Hilko van der Voet

Table of contents

ACROPOLIS model: Review and plans

20 June 2011

Hilko van der Voet, on behalf of the project team of DLO/Biometris: Waldo J. de Boer, Paul W. Goedhart, Jaap Kokorian, Gerie W.A.M. van der Heijden, Hilko van der Voet. Thanks to Marc Kennedy (FERA) for help and comments.

1 Introduction

The final ACROPOLIS model is planned to include the following extensions of methodology:

1. Cumulative exposure
2. Aggregate exposure
3. More uncertainty analysis
4. Integrating exposure and effect modelling, Margin of exposure modelling

The ACROPOLIS integrated model will be developed as a web-based platform. It will be an extension (Release 8) of the Monte Carlo Risk Assessment (MCRA) system. MCRA is a ASP.NET website plus C# modules with some R in the background.

This document describes the design for the Acropolis model first phase. This includes modelling cumulative exposure and to a lesser degree integration with effect modelling, but not yet much on aggregate modelling and all aspects of uncertainty analysis.

A separate document will describe the functional design of the software, and constitutes the second part of Deliverable 5.1.

2 Modelling of cumulative exposure

2.1 Review

Recent papers on exposure modelling including cumulative exposure are Boon et al. (2008), Bosgra et al. (2009), Boobis et al. (2008), EFSA (2008a,b), Müller et al. (2009), Petersen (2010), Price and Chaisson (2005), van der Voet et al. (2009), van Klaveren et al. (2009).

From these papers it appears that the most useful model for cumulative exposure assessment in practice currently is the Relative Potency Factor (RPF) approach. RPFs are used to convert exposures of all chemicals in a common assessment group into exposure equivalents of an index compound. The basis of an Acropolis model will be a Monte Carlo (MC) RPF approach: Given a characterization of the (multivariate) residue concentration distribution, this distribution is sampled, then the residue concentrations are multiplied by the RPFs and summed to give a possible realization of cumulative exposure. The resulting MC distribution of cumulative exposure may be characterized by a set of percentiles. Uncertainty can be quantified within a general two-dimensional Monte Carlo framework by repeating the process to yield multiple realizations of these percentiles. Bootstrapping or Bayesian alternatives will be investigated within Acropolis.

It has been stressed by Price and Chaisson (2005) that models should focus on the population of potentially exposed persons, not on the source of the contamination (person-oriented modelling). In practice this works by defining a target population for which a representative sample of consumption records is available. Many programs developed in the USA, such as DEEM, CARES, LifeLine, Calendex are designed especially for the US population or subgroups thereof (see e.g. Price and Chaisson 2005, Petersen 2010). In Europe, the tendency has been to develop programs that can work with user-defined consumption databases (see e.g. Gibney and van der Voet 2003, van Klaveren and Boon 2009).

Statistical model development for cumulative exposure may build on the experiences in the EFSA triazole project (van Klaveren et al. 2009). In that study two possible approaches (Approach 1 and 2) were addressed. The main difference between the approaches was how to deal with samples in which not all triazoles are analyzed. Approach 1 (which is the approach used commonly, e.g. in Boon et al. (2008), Bosgra et al. (2009), Müller et al. (2009)) starts with summing up the concentrations of different triazoles in the same sample according to their corresponding RPF. This accounts for correlations in the use pattern of pesticides. Whereas in practice not all samples are analyzed for all triazoles, Approach 1 considers these triazoles as non-detects (or zero values if we assume that a non-detect is a zero). This might lead to an underestimation of the exposure because in reality those non analyzed triazoles might have been positive values if they had been analyzed. Therefore an alternative approach was devised which simulates all samples of each triazole separately and finally sums the results of the separate simulations according to the corresponding RPF in a later stage of the cumulative exposure calculations (Approach 2). In this approach the calculation was limited to the number of analyzed values for each triazoles and no assumptions were made for non analyzed triazoles.

Van Klaveren et al (2009) concluded that statistical models should be further optimized to handle unbalanced data sets, and to address possible correlations in pesticide uses. It was recognised that neither of the two approaches mentioned in that report was optimal in the presence of data gaps (missing values). In addition to missing values also nondetects are a problem.

Statistical modelling can only be successful if sensible assumptions can be made on the reasons why data are missing. In practice these reasons are currently often unknown, but it is common practice to assume that data are missing completely at random.

2.2 Data sources

A major statistical problem in cumulative exposure assessment is the lack of data for estimating the multivariate residue distribution. Very often the number of positive values is small. This is partly due to the fact that a specific pesticide is commonly not used on 100% of the crop, and partly to the inability to measure very low levels. It is well documented that parametric modeling of residue distributions is challenging when a high proportion of residues are censored even in the univariate case (EFSA 2010, Helsel 2005). The multivariate case is even more challenging, so it will be necessary to utilize additional information within the statistical model.

To address the lack of data it is important to consider which data can be used. Possibly useful data sources include:

1. Pesticide monitoring (PM) data: a set of concentrations of all pesticides per sample. Observations may be present values (PV), nondetects (ND) or missing values[1] (MV). Nondetects are characterized by a Limit Of Reporting (LOR). The following scheme summarises the possible correspondence between observations and true values (ignoring measurement uncertainty):

   Table 1. Correspondence between observations and true values

observation:	PV	ND	MV
true: 0		x	x
true: <LOR		x	x
true: >LOR	x		x

2. Pesticide usage survey (PUS) data: a set of indicator variables of all pesticides per field, plus the area of the field. Values may be 0 (not applied) or 1 (applied). These data may be summarized to specify the fraction of the total crop area for each use pattern (possible combination of applied pesticides).
   PUS data are usually available on a national scale.
3. Food origin (FO) data: data specifying for a certain country (or other group relevant for the assessment population) the percentages of the crop originating from specific countries (or other groups). These data could be

[1] Missing data is not always an issue, in many countries a complete set of pesticides is tested for in all samples. However, the pattern of tested pesticides may change over time.

useful in conjunction with PUS data, as usage surveys are specific to food crops produced in individual countries.

PUS and FO data are currently rare, but it is expected that these can be made available in the near future.

As a typical example of available data, consider triazole on Brussels sprout in the UK. In the monitoring data 2004-2009 only 54 samples have been measured on 8 triazoles (see Table 2). These were 52 samples from the UK, 1 sample from the Netherlands and 1 from Poland. Only 10 positive values were found, 9 for tebuconazole, and 1 for difenoconazole (on another sample). In these UK monitoring data there is hardly any information for the estimation of correlations between positive concentrations (see Appendix 1).

Table 2. Number of positive values found on Brussels sprouts, UK 2004-2009, on 1 NL, 1 PL and 52 UK samples in total. The LOR was 0.05 for bitertanol and 0.01 for the other triazoles.

pesticide / country:	NL	PL	UK	positive values
bitertanol	0	0	0	
cyproconazole	0	0	0	
difenoconazole	0	0	1	0.04
epoxiconazole	0	0	0	
flusilazole	0	0	0	
myclobutanil	0	0	0	
propiconazole	0	0	0	
tebuconazole	0	0	9	0.02 0.01 0.01 0.02 0.02 0.01 0.04 0.01 0.02

An example of the aggregated PUS data is shown in Table 3, which shows that from the more than 20 triazole pesticides only 3 have been used on Brussels sprout. On 55 % of the crop no triazole pesticides have been applied and on 41 % of the crop two or three triazole pesticides have been applied. Obviously, such information cannot be estimated from the available monitoring data.

Table 3. Example aggregated PUS data.

Field treatment type (triazoles)	**Percent of total GB Brussels sprouts crop (2007)**
none	55.48
Difenoconazole	2.98
Tebuconazole	0.56
Difenoconazole, Flusilazole	0.29
Difenoconazole, Tebuconazole	38.17
Flusilazole, Tebuconazole	1.93
Difenoconazole, Flusilazole, Tebuconazole	0.59

2.3 Modelling options

For a univariate model (underlying the multivariate generalisation needed for cumulative assessment) we consider two options:

1. CensoredLogNormal (CLN). This model assumes that in reality all pesticide residue concentrations in all samples are positive values.
2. Spike-CensoredLogNormal (SCLN). This model is a mixture model, and assumes that the pesticide residue concentration is 0 with a certain probability, or else is a positive value from a lognormal distribution.

When the number of observed positive values is small, and their range is not relatively far above the LOR, then it is difficult to discriminate between the CLN and the SCLN model on the basis of monitoring data alone. Both models then tend to fit reasonably, but the implications for percentiles may be very different, in particular outside the data range.
On prior grounds the SCLN model seems more reasonable than the CLN model in the case of pesticides which are not naturally occurring substances. As is clear from e.g. the PUS data for triazoles for Brussels sprouts in the UK, most pesticides are not used on all fields. It is therefore likely that residue concentrations will be 0 for some samples and the SCLN seems more appropriate.
When PUS data are available, these can be used to estimate the probability of a zero concentration in the SCLN model. It is also possible to estimate the zero probability from the combination of PUS and monitoring data.

The univariate CLN and SCLN models are already implemented in the existing software (MCRA 7). However, use of external PUS data (agricultural use data) and FO data is not yet integrated with fitting the SCLN model. Approaches for cumulative exposure will initially be implemented by adapting these existing models assuming no correlation.

Further model development will continue, for example to incorporate additional information sources as outlined above, and also to include correlations if these are found to be important. These system will be designed to allow future extensions in residue modeling through a flexible interface that may call external routines or input pre-generated residue simulations.

In the multivariate case there are two major modelling options, similar to the univariate case:

1. Multivariate CensoredLogNormal (MCLN). This model has some additional complexity, because in any sample there can be a combination of PVs, MVs and NDs. The parameters of the MLN distribution are a vector μ and a covariance matrix $\sum$. They can be estimated by maximum likelihood (ML), expectation-maximization (EM) or stochastic expectation-maximization (SEM). For all three we have GenStat programs for the bivariate case. For SEM there is a C# program for any number of pesticides. Note that EM and SEM both yield maximum likelihood estimates. Alternatively, a Bayesian approach could be used. A description of the SEM algorithm is given in Appendix 2, as well as a comparison between the Bayesian approach and the SEM estimation method.

2. Multivariate Spike-CensoredLogNormal model (MSCLN). This model is a mixture of 2^p components (p is the number of pesticides), where in each component 0, …, p specific concentrations are 0, and the remaining concentrations follow a multivariate lognormal (MCLN) distribution (with dimensionality varying between 0 and p). If PUS data are available these can be used to provide prior use pattern probabilities, which under many assumptions can be used to specify the mixture probabilities of the MSCLN model. An alternative would be to estimate the mixture probabilities from monitoring data, but it is not expected that actual data will have enough information. Note: 2^p components may be a high number, but typically most combinations will have zero probability. For example, 12 triazoles were used for wheat in the UK in 2008, but the number of occurring combinations is only 112 (including 'no triazole used') out of the theoretical $2^{12} = 4096$.

In principle the MSCLN model is preferable over the MCLN model if estimated pesticide usage probabilities can be estimated, e.g. from PUS and FO data. For the MSCLN model there is a further modelling choice:

a) Ignore correlations between positive concentrations;
b) Don't estimate the correlations, but set them to 0 and 1 (or values in between) in a sensitivity analysis;
c) Estimate the correlations from available data.

It is expected that in most cases the data contain insufficient information to estimate correlations reliably. In such cases it cannot be expected that any multivariate model that includes correlation will perform better than a model that ignores the correlation. Therefore the basic Acropolis model (a) will ignore correlations between the positive concentrations. If ignored, draws from the multivariate distributions can be replaced by independent draws from the univariate distributions. The potential effect of correlations on the higher percentiles of the cumulative exposure distribution can be examined by drawing from multivariate distributions with pre-set correlations in a sensitivity analysis (b). In Acropolis we will develop options a and b. A potential additional model (c, still under development) may include modelling of correlations for situations where the monitoring data convey enough information (such data are not yet available). An investigation of UK PUS data and related residue data is being carried out to understand the patterns of usage that may lead to correlated residues in consumed items. The aim is to assess the likely impact of the modeling assumptions being made.

In all models, univariate as well as multivariate, lognormal as well as spike-lognormal, the assumption is made that the distribution for any pesticide is independent of the use pattern of the other pesticides. For example, we assume the same mean and variance for pesticide A used alone and used in combination with pesticide B. In principle models can be formulated to estimate differences between those situations, but it seems very unlikely that enough data is available to estimate such differences from the actual monitoring and PUS data.

3 Modelling of aggregate exposure

For a general discussion on modelling aggregate exposure see e.g. Peterson (2010) or Price and Chaisson (2005).
In the Acropolis project specific models for the European context are developed in work package 3. There is a need to define which (sub)populations are addressed in specific cases (e.g. operator, worker, bystander, resident exposure), e.g. by specifying covariables such as sex and age group. Most likely the non-dietary models will deliver a deterministic estimate for specific subgroups of the general population together with an uncertainty specification. Median and high quantile exposures could be precalculated for various typical scenarios and built into MCRA as a user-selected default option. It is needed to determine how these estimates (which might be dermal exposures or internal exposures) can be aggregated with the dietary (oral) exposure, and the role of bioavailability factors should be elucidated. In future it may be possible to input richer information about variability and uncertainty of exposure (such as the outputs that will be generated as part of the EU project BROWSE www.browseproject.eu).

4 Modelling of uncertainty

The progress on identification, prioritisation and quantification of uncertainties within the ACROPOLIS project has been described in a FERA report (Flari et al. 2011). Appendix III of that document desribes a tool to tabulate expert opinions on uncertainty in a qualitative format. This tool will be added as a module to MCRA, enabling risk assessors to accompany each quantitative analysis with a structured table on unquantified uncertainties.

A quantitative approach (model-based) will be attempted for

1. Handling of data below the LOR
2. Measurement uncertainties in pesticide concentrations
3. Sampling uncertainty due to limited monitoring data
4. Treatment of unit-to-unit variation

Structured methods for eliciting expert judgment will be applied by FERA, and these may lead to other possibilities to quantify uncertainties. Alternatively, they will contribute to the qualitative uncertainty tables.

5 Integrated modelling of exposure and health effects

The integrated modelling of exposure and health effects in a probabilistic manner is known as probabilistic health impact assessment (van der Voet et al. 2009). For the case of a single substance and a single health effect the Integrated Probabilistic Risk Assessment (IPRA) software has been developed (van der Voet and Slob 2007). IPRA will become part of the MCRA platform in the near future.

The dose level at which a critical effect size (CES) of the chosen health effect occurs, is termed the critical effect dose (CED) or benchmark dose (BMD). Useful algorithms for probabilistic analysis of dose effect relations have already been implemented in software such as PROAST (www.rivm.nl/proast) and EPA-BMDS (www.epa.gov/ncea/bmds). In Acropolis an interface will be made to import results from PROAST into the MCRA environment.

For cumulative effects, the simple Approach 1 has been combined with IPRA in several applications (Bosgra et al. 2009, Müller et al 2009). As described before, Approach 1 considers missing observations as non-detects which may lead to an underestimation of the exposure, and therefore also of the health impact. Consequently, an improvement of the approach for cumulative exposure as envisaged in Acropolis will also benefit the health impact assessments.

Cumulative exposure and health impact assessments are dependent on the availability of estimates of relative potency factors (RPFs). In the current state of the art, these uncertainties are not yet handled.
Acropolis will address these uncertainties. One approach is to characterise the set of RPFs by an appropriate uncertainty distribution. In the case of health impact assessment there is also the opportunity to eliminate the RPFs from the calculations if the data are available that were used to estimate the RPFs. The principle for this has been described in van der Voet et al. (2009), and is similar to the principle used in deterministic assessments of the total or combined Margin of Exposure (see e.g. Boobis et al. 2008, Petersen 2010).

Appendix 1. Modelling in relation to data scarcity of pesticide monitoring data

In probabilistic modelling we need a distribution of residue concentrations per food (or food group). There are many combinations of residue and food, for example in the NL triazole data 238 (foods) x 23 (residues) = 5474, and in the UK triazole data 119 (foods) x 10 (residues) = 1190 distributions have to be specified.

One source of uncertainty is that no or few measurements are available for part of the residue-per-food distributions. For example, in the UK data there are no measurements for 223 distributions and only one measurement for 61 distributions. For 851 of the 1190 combinations (72% of all) there are at least 10 measurements as an empirical basis for estimating the distribution.

Even if measurements have been made, it is common that most results are non-detects (reported as 'below limit of reporting'). For most of the residue-per-food distributions there are no reported positives (869 for the UK data). Therefore an estimate of a positive residue level is only available for 1190-869 = 321 distributions. For 24 distributions there is only one reported positive, so an elementary estimate of variability is only available for 321-24 = 297 distributions (25% of all). However, in the majority of these cases the number of reported positives is less than 10. For only 14 food-residue combinations (1.2% of all) we have at least 10 positive values to characterise the variability of the residue-per-food distribution (see Table 1).

Table 1. UK data triazoles. Food-residue combinations with at least 4 positive values.

Food	residue	npos
Table-grapes	myclobutanil	117
Banana	bitertanol	76
Parsnip	tebuconazole	58
Apple	myclobutanil	37
Strawberry	myclobutanil	35
Carrot	tebuconazole	34
Table-grapes	triadimenol	32
Table-grapes	tebuconazole	29
Tomato	triadimenol	24
Celery	difenoconazole	22
Peas (pods and	tebuconazole	20
Pear	difenoconazole	16
Pear	tebuconazole	14
Peppers, sweet	triadimenol	10
Brussels sprout	tebuconazole	9
Lettuce, Head	tebuconazole	9
Strawberry	triadimenol	9
Tomato	tebuconazole	9
Currants, black	triadimenol	8
Kale	difenoconazole	8
Lettuce, Head	difenoconazole	8
Peppers, Chili	triadimenol	8
Parsnip	difenoconazole	7
Peach	propiconazole	7
Peas (pods and	triadimenol	7
Peppers, sweet	myclobutanil	7
Herbs	difenoconazole	6
Peppers, Chili	cyproconazole	6

Pineapple	triadimenol	6
Banana	myclobutanil	5
Celeriac	difenoconazole	5
Nectarine	tebuconazole	5
Passion fruit	difenoconazole	5
Peach	bitertanol	5
Apple	triadimenol	4
Cabbages, Head	tebuconazole	4
Herbs	tebuconazole	4
Herbs	triadimenol	4
Onion, Bulb	triadimenol	4
Peach	tebuconazole	4
Tomato	bitertanol	4

T

The data scarcity problems described above are equally relevant for univariate (single-compound) and multivariate (multi-compound) modelling. In traditional (single-compound) exposure assessments the data scarcity, though well-known and noted in a qualitative uncertainty table) is usually not considered as prohibiting.

In multivariate modelling of residue concentrations we would like to know the correlations between all pairs of pesticides per food. For the UK example these are 119 (foods) x 10 x 9 / 2 = 5355 pairwise correlations. For a quantitative estimate we need at least two positive values for each of two pesticides. Even if the positives were obtained on different samples there is some information in the direction of a negative correlation. For the UK data example the information is summarised in Table 2.

Table 2. UK data triazoles. Food-residue combinations with at least 2 positive values for at least 2 pesticides per food.

food	origin	p	n
Apple	all	difenoconazole	2
Apple	all	myclobutanil	37
Apple	all	triadimenol	4
Apricot	all	bitertanol	3
Apricot	all	myclobutanil	3
Apricot	all	tebuconazole	2
Banana	all	bitertanol	76
Banana	all	epoxiconazole	2
Banana	all	myclobutanil	5
Beans, except b	all	cyproconazole	2
Beans, except b	all	tebuconazole	2
Cabbages, Head	all	difenoconazole	2
Cabbages, Head	all	tebuconazole	4
Carrot	all	difenoconazole	2
Carrot	all	tebuconazole	34
Currants, black	all	myclobutanil	3
Currants, black	all	triadimenol	8
Herbs	all	difenoconazole	6
Herbs	all	tebuconazole	4
Herbs	all	triadimenol	4
Lettuce, Head	all	difenoconazole	8
Lettuce, Head	all	tebuconazole	9
Lettuce, Head	all	triadimenol	2
Mandarins	all	difenoconazole	3
Mandarins	all	myclobutanil	3
Melons, except	all	myclobutanil	3
Melons, except	all	tebuconazole	2
Nectarine	all	bitertanol	3
Nectarine	all	propiconazole	2
Nectarine	all	tebuconazole	5
Parsnip	all	difenoconazole	7
Parsnip	all	tebuconazole	58
Peach	all	bitertanol	5
Peach	all	propiconazole	7
Peach	all	tebuconazole	4
Pear	all	difenoconazole	16

Pear	all	myclobutanil	3
Pear	all	tebuconazole	14
Peas (pods and	all	tebuconazole	20
Peas (pods and	all	triadimenol	7
Peppers, Chili	all	cyproconazole	6
Peppers, Chili	all	difenoconazole	3
Peppers, Chili	all	myclobutanil	3
Peppers, Chili	all	triadimenol	8
Peppers, sweet	all	cyproconazole	2
Peppers, sweet	all	myclobutanil	7
Peppers, sweet	all	triadimenol	10
Strawberry	all	myclobutanil	35
Strawberry	all	triadimenol	9
Table-grapes	all	cyproconazole	2
Table-grapes	all	difenoconazole	2
Table-grapes	all	myclobutanil	117
Table-grapes	all	tebuconazole	29
Table-grapes	all	triadimenol	32
Tomato	all	bitertanol	4
Tomato	all	difenoconazole	2
Tomato	all	tebuconazole	9
Tomato	all	triadimenol	24

Table 3. UK data triazoles. Number of correlations per food estimable with at least 2 (a), 4 (b) or 10 (c) positive values for each pesticide.

a. at least 2 positive values for each pesticide.

food	**ncorr**	
Apple	3	
Apricot	3	
Banana	3	
Beans, except b	1	
Cabbages, Head	1	
Carrot	1	
Currants, black	1	
Herbs	3	
Lettuce, Head	3	
Mandarins	1	
Melons, except	1	
Nectarine	3	
Parsnip	1	
Peach	3	
Pear	3	
Peas (pods and	1	
Peppers, Chili	6	
Peppers, sweet	3	
Strawberry	1	
Table-grapes	10	
Tomato	6	
Total	**58**	**(1.1% of 5355)**

b. at least 4 positive values for each pesticide.

food	**ncorr**	
Apple	1	
Banana	1	
Herbs	3	
Lettuce, Head	1	
Parsnip	1	
Peach	3	
Pear	1	
Peas (pods and	1	
Peppers, Chili	1	
Peppers, sweet	1	
Strawberry	1	
Table-grapes	3	
Tomato	3	
Total	**21**	**(0.4% of 5355)**

c. at least 10 positive values for each pesticide.

food	**ncorr**

Pear	1	
Table-grapes	3	
Total	**4**	**(0.1% of 5355)**

The result of this exercise is shown in Table 3. Only a very small fraction of all correlations (at most 1.1%) between residue concentrations appears to have an empirical basis for estimation. Of course, not all correlations are equally important for the upper tail of the distribution of cumulative exposure and it might be true that most positive values are indeed obtained for the most relevant pairs of pesticides.

In the above examples we have assumed that one set of parameters (means, standard deviations and correlations) describes all positive residue concentrations. So no distinction is made between samples from different origins (e.g. countries). This is a common approach in current exposure assessment. It does not preclude the possibility to use origin-specific datasets on pesticide usage for defining the mixture probabilities in a model where the data are described as a mixture of true zeroes for some pesticides and positive values (possibly censored) for the others. Our assumption also implies that the distribution of a pesticide is independent of the usage of other pesticides on the same crop. This may be less realistic, but the data currently available (Pesticide Usage Survey data and monitoring data) do not seem to allow more complex models.

How to handle the parameters (mean levels, standard deviations, correlations) that are not estimable from the data? There are two main approaches:

a) Traditional: set unestimable parameters to zero. So mean levels are 0 if there are no positives, standard deviations are 0 if there is only one positive, and correlations are 0 if no correlation can be estimated.
b) Bayesian: specify prior distributions for all parameters. If the data contain no information, the posterior distribution will be the same as the prior distribution (although data on other parameters may influence the posterior if parameters are correlated).

Clearly, the traditional approach ignores part of the uncertainties. In the Bayesian approach uncertainty is included but depends on subjective prior specifications. In comparison to the data-rich situation the choice of priors will be very influential for final parameter estimates, although it may be less influential on final outcomes which depend mostly on rich parts of the data.
Another difference between approaches is that the Bayesian approach requires a fully parametric model, whereas the traditional approach can be used either with parametric or non-parametric (e.g. resampling) approaches.

The Bayesian approach seems to provide the most possibilities for addressing all uncertainties in modelling. However, it is well-known to be computationally heavy, and the Acropolis project was not set up to implement Bayesian approaches in a web-based environment.

Therefore, the traditional approach seems to be the feasible solution in this project. It may be noted that for univariate parameters (mean levels and standard deviations) we just apply the approaches currently in common use, so that there is little reason to follow another approach for the bivariate parameters (correlations) alone.

Appendix 2. Fitting models based on the censored lognormal distribution

In Acropolis WP5 methods have been investigated and tested for estimating cumulative residue concentration based on fitting the CLN and MCLN models to PM data. This work is described in draft Biometris and FERA documents (Goedhart and de Boer 2011, Kennedy and Roelofs 2011). In short, a comparison was made between

a) several variations of combining univariate analysis;
b) a Bayesian analysis using WinBugs;
c) a stochastic EM (SEM) analysis.

The impression was that methods b and c behaved similarly on the cases investigated with regard to estimating percentiles of the exposure distribution. The basic step is the same in both cases, i.e. sampling from a conditional normal distribution for MVs and NDs (which will be the truncated distribution for NDs). However, both methods need some information about correlations in the data. If the data do not allow the estimation of correlation for any two pesticides the univariate method a can be used instead.

The stochastic EM (SEM) method to cumulative exposure assessment is basically an implementation of Monte Carlo EM (Wei and Tanner 1990). In the Simulation Expectation-step (SE-step), missing or non-detect data are simulated from the appropriate conditional distributions. In the maximization-step (M-step), the mean vector and covariance matrix of the multivariate normal distribution are estimated. The SE-step and M-step are repeated until convergence. The approach is stochastic because the expectation-step is replaced by a stochastic counterpart. The approach is a single imputation method.

Let X denote a p-multivariate random vector of *p* variables, normally distributed: X ~ N(μ, $\sum$)

Partition X, μ and $\sum$ as follows: $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

where X_1 is a q-vector and X_2 is a (p-q)-vector. The distribution of X_1 conditional on X_2 = a is multivariate normal:

$$(X_1 | X_2 = a) \sim N(\mu_c, \Sigma_c)$$

with mean $\mu_c = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$ and covariance

$$\Sigma_c = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

The SEM algorithm proceeds as follows. Suppose that the data vector for a record can be partitioned into X_2 with values, say x_2, larger than their respective LORs, and X_1 with values that are either missing or equal to the corresponding LOR. Suppose that at some stage we have an estimate for μ and $\sum$.

1. (Stochastic Expectation step)

   a. In the first step the conditional distribution of $X_1 \mid X_2=x_2$ is calculated for each record separately according to the formula given above. Note that in case X_2 is an empty set, this conditional distribution equal $N(\mu, \sum)$.
   b. In the second step, for each record separately, a random sample is drawn from the **truncated** conditional distribution as calculate in step 1. Truncation is from above and truncation limits are equal to the LORs when data are censored and equal to infinity when data are missing.
2. (Maximization step) The dataset is now complete and new estimates for μ and $\sum$ are obtained..
3. Repeat step 1-2 and continue until convergence.

The random draw in step 1b can be handled by means of rejection sampling but this can be very slow when the probability of sampling in the truncation region is low. If this is the case the Gibbs sampling approach as implemented in the R package 'tmvtnorm' is used. At the moment we do not have a clear definition of convergence of the SEM algorithm. We used a burn-in of 500 iterations and took the mean parameters of the next 500 iterations as our estimates for μ and $\sum$.

The Bayesian method also uses the standard conditional normal distribution forms listed above, but conditioning is on values of μ, $\sum$ and the unknown censored residues simulated in a preceding iteration of a Markov Chain Monte Carlo (MCMC) algorithm. The multiple realizations from the MCMC algorithm are used as an approximate sample from the posterior distribution. A difference between SEM and the Bayesian method is that only the latter includes an automatic calculation of uncertainties (in the form of the posterior distribution). Uncertainties for either method can also be obtained by using a non-parametric bootstrap approach, in which the set of chemical samples (concentration vectors) is resampled many times.

References

- Boobis AR, Ossendorp BC, Banasiak U, Hamey PY, Sebestyen I, Moretto A (2008). Cumulative risk assessment of pesticide residues in food. *Toxicology Letters*, 180: 137-150.
- Boon, P.E., van der Voet, H., van Raaij, M.T.M. and van Klaveren, J.D. (2008). Cumulative risk assessment of the exposure to organophosphorus and carbamate insecticides in the Dutch diet. *Food and Chemical Toxicology*, 46: 3090-3098.
- Bosgra, S., van der Voet, H., Boon, P.E. and Slob, W. (2009). An integrated probabilistic framework for cumulative risk assessment of common mechanism chemicals in food: an example with organophosphorus pesticides. Regulatory Toxicology and Pharmacology. 54: 124–133.
- EFSA (2008a). Opinion of the Scientific Panel on Plant Protection products and their Residues to evaluate the suitability of existing methodologies and, if appropriate, the identification of new approaches to assess cumulative and synergistic risks from pesticides to human health with a view to set MRLs for those pesticides in the frame of Regulation (EC) 396/2005. http://www.efsa.europa.eu/en/efsajournal/doc/705.pdf.
- EFSA (2008b). Scientific Opinion on Risk Assessment for a Selected Group of Pesticides from the Triazole Group to Test Possible Methodologies to Assess Cumulative Effects from Exposure through Food from these Pesticides on Human Health. http://www.efsa.europa.eu/en/efsajournal/doc/1167.pdf.
- EFSA (2010) European Food Safety Authority; Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal 2010; 8(3):. [96 pp.]. doi:10.2903/j.efsa.2010.1557. Available online: www.efsa.europa.eu.
- Flarie V, Glass R, Hart A, Kennedy M, Owen H (2011). Summary of progress on identification, prioritisation and quantification of uncertainties within the ACROPOLIS project. Report March 2011, FERA.
- Gibney, M.J. and van der Voet, H. (2003). Introduction to the Monte Carlo project and the approach to the validation of probabilistic models of dietary exposure to selected food chemicals. *Food Additives and Contaminants*, 20 (Suppl. 1): S1-S7.
- Goedhart P, de Boer WJ (2011). DRAFT: WinBUGS and Stochastic EM, some experiences, 15-02-2011 (updated 22-03-2011).
- Helsel, D.R. (2005). Nondetects and Data Analysis: Statistics for Censored Environmental Data. Wiley and Sons, New York.
- Kennedy M, Roelofs V (2011). Bayesian approach to imputation of missing values and non-detects (Draft of first results), 2011-01-12.
- Müller, A.K., Bosgra, S., Boon, P.E., van der Voet, H., Nielsen, E. and Ladefoged, O. (2009). Probabilistic cumulative risk assessment of anti-androgenic pesticides in food. *Food and Chemical Toxicology*, 47: 2951-2962.
- Petersen BJ (2010). Modeling dietary exposure with special sections on modeling aggregate and cumulative exposure. In: Hayes' Handbook of Pesticide Toxicology, Third Edition.
- Price PS, Chaisson CF (2005). A conceptual framework for modeling aggregate and cumulative exposures to chemicals. *Journal of Exposure Analysis and Environmental Epidemiology*, 15: 473-481.
- 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. and Brüschweiler, B.J. (2009). A model for probabilistic health impact assessment of exposure to food chemicals. *Food and Chemical Toxicology*, 47: 2926-2940.
- van Klaveren JD, Boon PE (2009). Probabilistic risk assessment of dietary exposure to single and multiple pesticide residues or contaminants: Summary of the work performed within the SAFE FOODS project. *Food and Chemical Toxicology*, 47: 2879-2882.
- van Klaveren JD, van Donkersgoed G, van der Voet H, Stephenson C, Boon PE (2009). Cumulative exposure assessment of triazole pesticides. Report 2009.008.

RIKILT, Wageningen. Submitted to EFSA. http://www.efsa.europa.eu/en/supporting/doc/40e.pdf.

- Wei GCG, Tanner MA (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 85: 699–704.