

Analysis of multivariate extreme intakes of food chemicals

M.J. Paulo ^a, H. van der Voet ^{a,*}, J.C. Wood ^b, G.R. Marion ^b, J.D. van Klaveren ^c

^a *Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, The Netherlands*

^b *Biomathematics and Statistics Scotland, King's Buildings, Edinburgh, EH9 3 JZ, Scotland*

^c *RIKILT Institute of Food Safety, Wageningen University and Research Centre, P.O. Box 230, 6700 AC Wageningen, The Netherlands*

Received 9 December 2004; accepted 16 December 2005

Abstract

A recently published multivariate Extreme Value Theory (EVT) model [Heffernan, J.E., Tawn, J.A., 2004. A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B* 66 (3), 497–546] is applied to the estimation of population risks associated with dietary intake of pesticides. The objective is to quantify the acute risk of pesticide intake above a threshold and relate it to the consumption of specific primary food products. As an example daily intakes of a pesticide from three foods are considered. The method models and extrapolates simultaneous intakes of pesticide, and estimates probability of exceeding unobserved large intakes. Multivariate analysis was helpful in identifying whether the avoidance of certain food combinations would reduce the likelihood of exceeding a threshold. We argue that the presented method can be an important contribution to exposure assessment studies.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Extreme values; Food safety; Exposure assessment; Uncertainty analysis; Pesticides

1. Introduction

In this article we consider the characterization and estimation of the risk of exposure to toxic substances present in food. Many substances, such as pesticides, occur only rarely in human food, but still the occasionally found high levels are a cause of concern, and food authorities need to assess the risks associated with their intake. In animal studies health effects occur due to large exposures to hazardous substances. Whereas some occupational groups (e.g., farmers) may run a direct risk of pesticide exposure, the most likely source of exposure for the general public is food. Food safety authorities are required by the EU General Food Law to base regulations on risk analysis. Any measures taken to reduce risk (risk management) should be based on risk assessments, and these in turn should be

based on the available scientific evidence. Whereas pesticide residues are typically only found at low levels in products on the consumer market and often do not seem to present a large risk, food safety authorities still need to reassess risks when new data arrive. In this paper we present sophisticated tools, and apply it to pesticide data which were available to us.

In risk assessment a distinction is made between acute risks, concerning health effects which arise from a single or short-term intake, and chronic risks, related to long-term intake. For example, high doses of organophosphates cause neurological effects after a single intake in animal studies (Brown and Brix, 1998), whereas many pesticides have carcinogenic effects in animals after a long-term exposure (Dich et al., 1997). The multivariate methods we apply in this paper are relevant in the context of acute risk assessment, based on intakes for a random person at a random day. For chronic risk, residue concentrations can be averaged over time. This means that such risks can be assessed by just multiplying the food consumptions with the average

* Corresponding author. Tel.: +31 317 476938; fax: +31 317 483554.
E-mail address: hilko.vandervoet@wur.nl (H. van der Voet).

residue concentration, and then summing the intakes over all consumed products. Therefore, for a chronic risk analysis, we can use univariate methods.

Risk assessment is often hampered by the fact that available data sets typically have little information about extreme situations. In practice large residue levels are rarely observed, which makes estimation of acute risks hard to achieve. Yet the purpose of risk analysis is to make statements about the probability that such rarely observed situations will occur. Therefore it is often necessary to extrapolate from the data, and the statistical basis on which to make such extrapolations is called Extreme Value Theory (EVT). EVT is a well established framework (Kotz and Nadarajah, 2000; Embrechts et al., 1997) which evolved during the last 50 years into a coherent theory. EVT models have been widely applied in many areas (financial, environmental) where maximal or minimal values (called extremes) are relevant, but to our knowledge they have hardly been applied to food risks—a recent application concerns the exposure to heavy metals from seafood (Tressou et al., 2004). However it should be noted that whilst there is a developed body of theory and application in the univariate case, multivariate EVT is the subject of ongoing research (Coles and Tawn, 1994; Capéraà et al., 1997; de Haan and de Ronde, 1998; Coles et al., 1999; Heffernan and Tawn, 2004; Dupuis, 2005; Martins and Ferreira, 2005).

A potential difficulty therefore arises from the fact that exposure to pesticide may be caused from eating several products containing pesticide simultaneously. Therefore the total intake resulting from all relevant food products should be modelled. One approach is to use a univariate EVT model for the total intake. This, however, obscures the source of the pesticide intake, and does not answer questions about the influence of individual foods or combinations of foods. Consumptions of different food products are typically correlated (Paulo et al., 2005), and therefore also the pesticide intakes from these foods will be correlated. A multivariate model should be able to model correlated intakes. In a recent paper a promising method that models multivariate extreme values was presented (Heffernan and Tawn, 2004). The method models relationships in the tails of the joint distribution (extreme values in EVT jargon) between pairs of variables and unlike other multivariate EVT methods can be used for tail dependence or independence. This paper aims to show how the methodology can be used in dietary exposure assessment, and furthermore we extend the method to the calculation of food risks, conditional on the amounts of eaten products.

2. Data

To illustrate this multivariate extreme intake method we use real pesticide concentration data and food consumption data. However, to avoid the suggestion of a full risk assessment study, we have coded the chosen pesticide and the chosen food products. Pesticide intakes have not been

measured directly, so they are derived from datasets on food consumption and on residue levels. Consumption data are collected in food consumption surveys (Verger et al., 2002), in this case the Dutch National Food Consumption Survey of 1997/1998. In that survey 2564 households were selected as part of a stratified random sample and each individual asked to write down all the food that was consumed over 2 days (DNFCS, 1998). A total of 6250 Dutch individuals completed the questionnaire, which also contained other information about the consumers such as their body weight in kilograms (kg BW). Composite food products such as pizza were transformed into amounts of raw agricultural commodities (e.g., tomato, mushrooms), by use of a food conversion table (van Dooren et al., 1995). In order to work with independent observations, the daily consumption data (in g/kg BW) are restricted to one randomly chosen individual in each household. Although in our residue monitoring dataset 53 foods were found that contained the chosen pesticide in at least one sample, we restrict our analysis to the three main contributors (A, B and C), and, to further illustrate the impact of correlations between intakes of different food products we also show an example with two additional products (D and E) from the same dataset.¹

Residue concentration data comes from monitoring programs run at several stages of the agro-food chain (van Klaveren, 1999). Monitoring data essentially is a collection of measurements of residue concentrations present in samples of raw agricultural products, and it is expressed in milligram per kilogram (mg/kg). A typical feature of monitoring datasets is their sparseness: toxic substances are rarely found in food samples, either because they are not present or because the concentrations are below the limit of reporting. The total number of observations for the pesticide in the three chosen foods A, B and C was 975, 712, and 131, respectively, but only 293, 171 and 36 of these were positive (non-zero) measurements. In this study the non-detects in the sample have been replaced by the value corresponding to the limit of reporting (0.01 mg/kg), which of course corresponds to a worst-case view of the situation. Food processing, such as washing or boiling vegetables or fruit, is likely to reduce or eliminate the residue present in food, but in this study we ignore such factors and work with original concentration measurements (again as a worst-case scenario). In a realistic risk assessment (outside the scope of this paper) a sensitivity analysis can easily be conducted to investigate the effects of processing factors which reduce residue levels, or setting non-detects at zero instead of LOR.

¹ The pesticide is iprodione, and the food products A–E are lettuce, grape, currant, cucumber, and sweet pepper, respectively. However, we want to stress that arbitrary thresholds have been used, and that our analysis omits many practical issues, and does not constitute a full risk assessment. The example only serves to illustrate the proposed methodology.

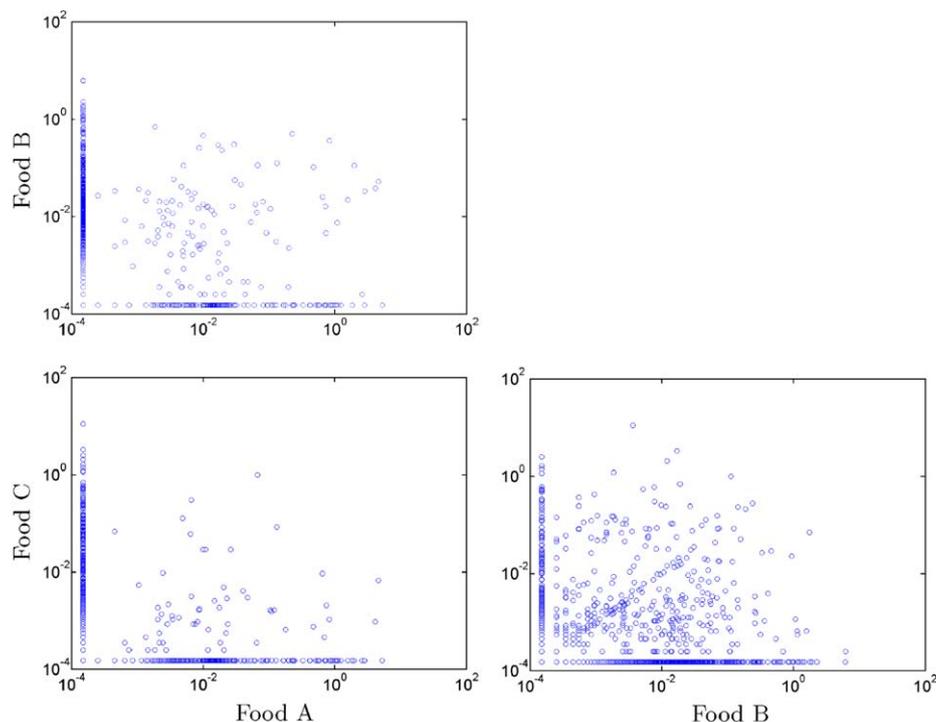


Fig. 1. Bivariate plots for daily pesticide intakes ($\mu\text{g}/\text{kg BW}$) from foods A, B and C. All intakes smaller than 10^{-4} are set to 10^{-4} .

A dataset of 2564 daily pesticide intakes (in $\mu\text{g}/\text{kg BW}$) was constructed by multiplying the (daily) consumption data for each food (in $\text{g}/\text{kg BW}$) by pesticide concentration values (in mg/kg) randomly sampled from the residue monitoring database for that food. Total intake was then obtained by adding the intakes associated with each food. This provided a dataset of total daily intake and daily intakes for each food type used in the multivariate analysis. Bivariate plots of this dataset are shown in Fig. 1. The uncertainty due to the use of randomly selected consumers and residues is addressed in an uncertainty analysis in Section 3.4.

3. Methods

We first explain the traditional Monte Carlo estimate of exposure percentiles using the empirical data (Section 3.1). Then we introduce the univariate EVT approach (Section 3.2) used in this paper, both for the sum of the univariate intakes, and for each food separately as a preparation for the multivariate EVT approach (Section 3.3). The bootstrap method, as explained in Section 3.4, allows comparison of the methods by calculating confidence intervals around the estimates (e.g., of probabilities and quantiles) obtained by each method. Finally, a new method is described to characterize the probability of intakes exceeding a threshold level conditionally as a function of the consumption of individual food products (Section 3.5). All methods described in this section were implemented in Matlab 7.0 (MathWorks, 2004).

3.1. Exposure assessment using empirical data

Empirical quantiles were calculated for the pesticide intakes in the example dataset (constructed as described in Section 2), for each food and also for the total intake. Quantiles are specified using cumulative probabilities, from 0 to 1, e.g., $q_{0.99}$ is the intake (in $\mu\text{g}/\text{kg BW}$) which is not

exceeded in 99% of the cases. For a vector x of n intakes, quantiles are calculated as follows: The intakes ordered by size $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ correspond to $(0.5/n), (1.5/n), \dots, ((n-0.5)/n)$ quantiles. Linear interpolation is used to compute quantiles for probabilities between $(0.5/n)$ and $((n-0.5)/n)$. The minimum intake is assigned to quantiles for probabilities below $(0.5/n)$ and the maximum intake is assigned to quantiles for probabilities larger than $(n-0.5)/n$. The maximum intake in the dataset was also recorded.

Probability of intake smaller or equal to a certain value was given by the proportion of intakes in the data below or equal to that value, e.g., $p_{0.01}$ is the proportion of intakes in the data that do not exceed $0.01 \mu\text{g}/\text{kg BW}$.

3.2. Univariate EVT approach: the generalized Pareto distribution

Extreme value theory (EVT) is essentially about extrapolation. The theoretical basis is that, given a wide class of continuous distributions, extreme values from these distributions can be modelled in a unified way. The limiting distributions for the extreme values from univariate distributions (including for example the lognormal, the normal and the exponential distributions) are the Generalized Extreme Value (GEV) family, and the parent distributions are said to lie in the domain of attraction of a GEV distribution (Pickands, 1975). Rather than modelling specific extreme values such as the maximum value in a dataset, we choose to model the probability of exceeding a threshold, which in food risk analysis could be e.g., an Acute Reference Dose or a lower dose. If our data are assumed to come from that class of distributions in the domain of attraction of a GEV, then for a sufficiently high threshold the conditional distribution $P(X > u + x | X > u)$ can be approximated by a generalized Pareto distribution (GPD) (Pickands, 1975; Coles, 2001): $P(X > u + x | X > u) = \left[1 + \frac{\xi x}{\beta}\right]^{-1/\xi}$. Here u is a suitably chosen high threshold for X , and β and ξ are respectively scale and shape parameters, with $\beta > 0$. Practical application requires choice of the threshold u and use of standard approaches to estimate the parameters β and ξ . Diagnostic techniques exist to judge whether u is chosen high enough (Coles, 2001).

The upper end point of the distribution is ∞ if $\xi \geq 0$ and $u - \beta/\xi$ if $\xi < 0$. In the univariate approach X is taken to be the sum of intakes from the various food products. After estimating parameters β and ξ , quantiles of the estimated GPD are calculated.

3.3. Multivariate EVT approach: the Heffernan–Tawn (HT) model

As noted earlier multivariate extreme value analysis is much less developed than in the univariate case. However, recently Heffernan and Tawn (2004) introduced a conditional approach to model and extrapolate multivariate extremes. Traditional approaches in multivariate EVT (de Haan and de Ronde, 1998; Capéraà et al., 1997) rely on the assumption that all variables become large at the same rate. The HT model, on the other hand, is applicable to any set of variables, which may include dependent and independent subsets. The method uses a semi-parametric model for the univariate distribution of each variable, and then models the distribution of extremes of a variable conditionally on the values taken by another variable in the dataset. Therefore it applies to pairs of variables at a time.

Starting with a random vector (X_1, \dots, X_p) the HT method consists first in fitting the generalized Pareto distribution (GPD) to the extremes of each variable, thereby obtaining p univariate distributions for the tail. A semi-parametric distribution is then defined for each variable, based on its empirical distribution in the centre and on the GPD model for the tail. Then each variable is transformed to a Gumbel scale and models are fitted to one variable conditionally on the value of one other variable, with the restriction that this other variable has a value in the right tail, and the procedure is repeated for each possible pair. This tail is determined by another threshold, which we label u_{HT} because it is specific for the Heffernan–Tawn conditional model. The conditional model, run for $u_{HT} = 0.001, 0.01$ and 0.1 , produced identical results so that the choice of this threshold is not critical. These thresholds correspond to high marginal quantiles as shown in Table 1. The conditional models are then used to extrapolate to new values in the Gumbel scale. The new extrapolated values are then back-transformed to the original scale, and used to estimate tail probabilities and quantiles. The method is described in more detail in Appendix 1.

3.4. Uncertainty analysis

Uncertainty is the lack of knowledge the risk assessor has about the state of the system. It is important to assess the uncertainty in the final results of risk assessment in order to avoid false confidence in the results. Food risk assessment is often hampered by lack of adequate data. Therefore one important aspect of uncertainty analysis is to describe how the ultimate results depend on the limited data availability. Bootstrap

Table 1
Empirical distribution and quantiles of daily pesticide intake for foods A, B and C; $p_x = P(X \leq x)$ and quantiles $q_x = P^{-1}(x)$ in $\mu\text{g}/\text{kg BW}$

Summary	Food A	Food B	Food C
$p_{0.00}$	[0.89, 0.90]	[0.61, 0.65]	[0.73, 0.75]
$p_{0.001}$	[0.89, 0.91]	[0.70, 0.73]	[0.84, 0.86]
$p_{0.01}$	[0.94, 0.95]	[0.82, 0.84]	[0.93, 0.95]
$p_{0.10}$	[0.98, 0.98]	[0.95, 0.97]	[0.97, 0.98]
$q_{.75}$	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
$q_{.90}$	[0.00, 0.00]	[0.02, 0.03]	[0.00, 0.00]
$q_{.95}$	[0.01, 0.02]	[0.05, 0.09]	[0.01, 0.04]
$q_{.99}$	[0.27, 0.79]	[0.52, 1.0]	[0.18, 0.52]
Max	[4.6, 23.7]	[3.6, 18.3]	[2.2, 17.8]

Bracketed values correspond to 90% confidence intervals based on bootstrap samples.

methods (Cullen and Frey, 1999; Efron, 1979) can be used for this purpose with any clearly described data-based approach such as the present model. In this study the uncertainty in the estimated quantiles and other results was quantified using a non-parametric bootstrap procedure. A total of 100 samples of size 2564 were obtained. In each bootstrap sample the amounts eaten by 2564 individuals (re-sampled, with replacement from the original set of 2564 individuals) were combined with random concentrations from the set of concentrations in the monitoring data, in a similar way as described in Section 2. The variation present in the bootstrapped samples was used to create 90% confidence intervals for pesticide intake quantiles q_p , and probabilities p_v of exceeding a threshold v . The bootstrap was used to calculate confidence intervals for the estimated probabilities and quantiles obtained with each of the three methods in Sections 3.1–3.3 (see Table 2).

3.5. Probability of exceeding a threshold

So far we have considered the intakes x_i of pesticide from three different food sources and the probability that the total intake will exceed v . However, for practical applications we extend the HT method to model this probability as a function of the consumed amount of a certain food product. So we are interested in the probability of exceeding a specific total intake conditional on eating a fixed quantity of a food: $P\{\sum(X_i) > v | \theta_i\}$, where θ_i is the amount of food i (i.e., A, B or C) consumed in one day.

For a fixed consumed amount θ_i , corresponding intakes X_i are obtained by combining θ_i with empirical concentrations of the pesticide. Other intakes $X_j, j \neq i$ are generated either empirically for $X_i \leq u_{HT}$, or by the conditional model for $X_i > u_{HT}$, where u_{HT} is the threshold used in the

Table 2

Upper posterior distribution for daily total pesticide intake (i.e., summed over foods A, B and C), observed in the data (column 2), estimated with the Generalized Pareto distribution fitted to the sum of daily intakes (GPD, column 3) and estimated with the Heffernan and Tawn method (H–T, column 4); q_x represents the quantile of 100 x % in $\mu\text{g}/\text{kg BW}$

Percentile	Method		
	Data	GPD	H–T ^a
$q_{.900}$	0.06 [0.06, 0.09]	0.06 [0.058, 0.084]	^a
$q_{.950}$	0.22 [0.19, 0.32]	0.18 [0.16, 0.25]	0.19 [0.17, 0.27]
$q_{.975}$	0.59 [0.48, 0.86]	0.46 [0.41, 0.71]	0.45 [0.39, 0.69]
$q_{.990}$	1.2 [1.1, 2.4]	1.6 [1.38, 2.76]	1.6 [1.22, 2.78]
$q_{.999}$	6.1 [4.5, 11.4]	37 [25, 95]	52 [27, 160]
Max	11.0 [6.5, 25.6]		
$P(\sum X_i > 1)$	0.016 [0.012, 0.021]	0.074 [0.069, 0.089]	0.075 [0.074, 0.099]
$P(\sum X_i > 10)$	0.0004 [0, 0.001]	0.003 [0.002, 0.004]	0.003 [0.002, 0.004]
$P(\sum X_i > 60)$	0	0.0007 [0.0005, 0.0013]	0.0009 [0.0006, 0.0015]

Point estimates from the working sample are given, as well as 90% confidence intervals based on bootstrap.

^a For $u_{HT} = 0.01$. Other values $u_{HT} = 0.001$ and 0.1 produced similar values.

conditional model of Heffernan and Tawn (as described in Section 3.3). Each generated vector (of the values of intakes from each food) is also assigned a weight based on empirical and model-derived distributions. The probability of exceeding a threshold is then obtained by averaging the outcomes with their weights. The algorithm is described in detail in Appendix 2.

4. Results

Table 1 shows the univariate empirical probabilities (see Section 3.2) for pesticide intakes from consuming three foods A, B and C. The proportion of zero intakes in the bootstrap samples is around 90% for food A, 63% for food B and 74% for food C. Only 2–3% of the consumers of any of the three foods could have potentially ingested more than $0.1 \mu\text{g}/\text{kg}$ BW of pesticide in a single day. In fact, the maximum pesticide intakes observed in the bootstrap samples were $23.7 \mu\text{g}/\text{kg}$ BW for food A, $18.3 \mu\text{g}/\text{kg}$ BW for food B and $17.8 \mu\text{g}/\text{kg}$ BW for food C. For illustration we consider a low threshold value of $10 \mu\text{g}/\text{kg}$ BW and a high threshold value of $60 \mu\text{g}/\text{kg}$ BW. Note that these thresholds are not related to known levels of acute risk, but are only used as an example how the methodology works when only few data, or no data at all, exist above the threshold value. This only describes the situation for each food separately, but in order to quantify the results of simultaneous occurrence of such high values we need to apply a multivariate model.

Fig. 1 shows 2564 pesticide intakes for pairs of foods. Large intakes can be caused either by consuming a large amount of a food with a moderate pesticide content or

by consuming a moderate amount of a food with a high pesticide content. Simultaneous intakes of pesticide from the consumption of two or more foods containing it are observed, also for large values. In the case of foods B and C some simultaneous intakes lie on a straight line, probably due to the occurrence of the two foods in fixed proportions in a recipe.

The mixed empirical/GPD model fitted to the univariate intake distributions (from step 1 of the Heffernan and Tawn method, see Section 3.3 and Appendix 1) are shown in Fig. 2. We conclude that $0.01 \mu\text{g}/\text{kg}$ BW is a suitable threshold u_{GPD} to separate the empirically modelled and the GPD-modelled parts of these marginal fits. The scale and shape GPD parameters were estimated as $\beta = \{0.018, 0.021, 0.047\}$ respectively for foods A, B and C and similarly $\xi = \{2.000, 1.136, 1.066\}$.

The HT model was fitted to all pairs of bivariate data, and a total of 10,000 extrapolations generated for the marginal tails and the conditional models are shown in Fig. 3 in the Gumbel scale (step 4 in Appendix 1).

Fig. 4 shows a randomly selected 10^3 of the 10^4 extrapolations of conditional intakes in the original scale of intakes (step 5 in Appendix 1). The intakes in the vertical axis are conditional on the intakes in the horizontal axis being above threshold u_{HT} . The plots also show the original data for comparison, shown as dots, with the only extreme shown as a triangle. Extrapolated total intakes larger than $10 \mu\text{g}/\text{kg}$ BW were classified as extreme values, as they are relative extremes in the dataset, and are shown as stars in the figure. Extrapolated total intakes lower than $10 \mu\text{g}/\text{kg}$

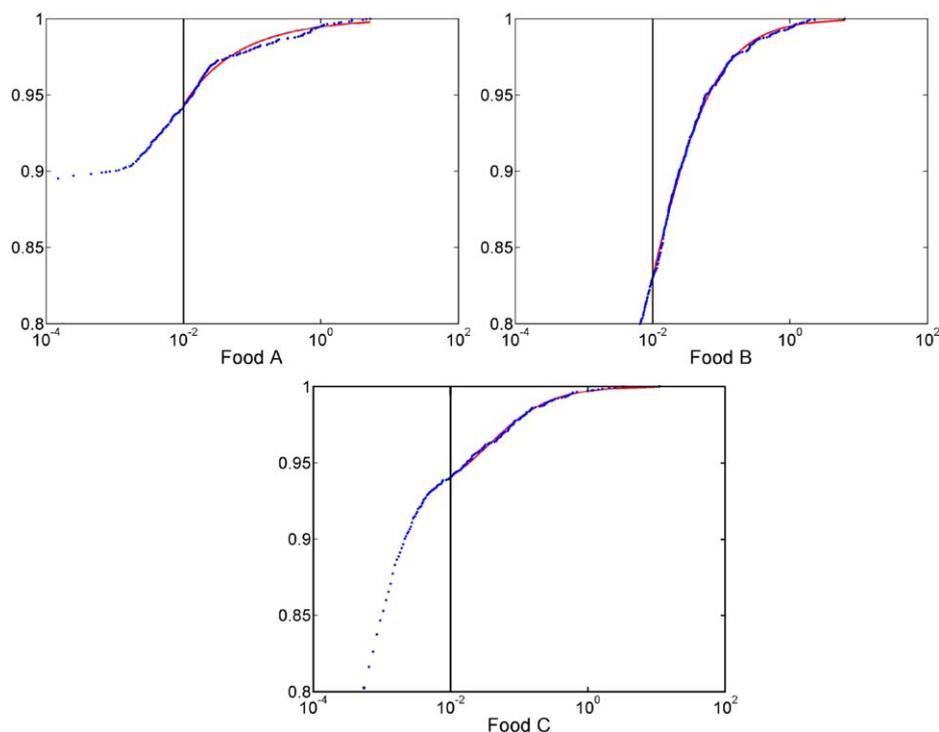


Fig. 2. Empirical distributions (dotted line) and GPD-fitted (solid line) for daily pesticide intakes ($\mu\text{g}/\text{kg}$ BW). The marginal threshold $u_{\text{GPD}} = 0.01$ is also shown (vertical line).

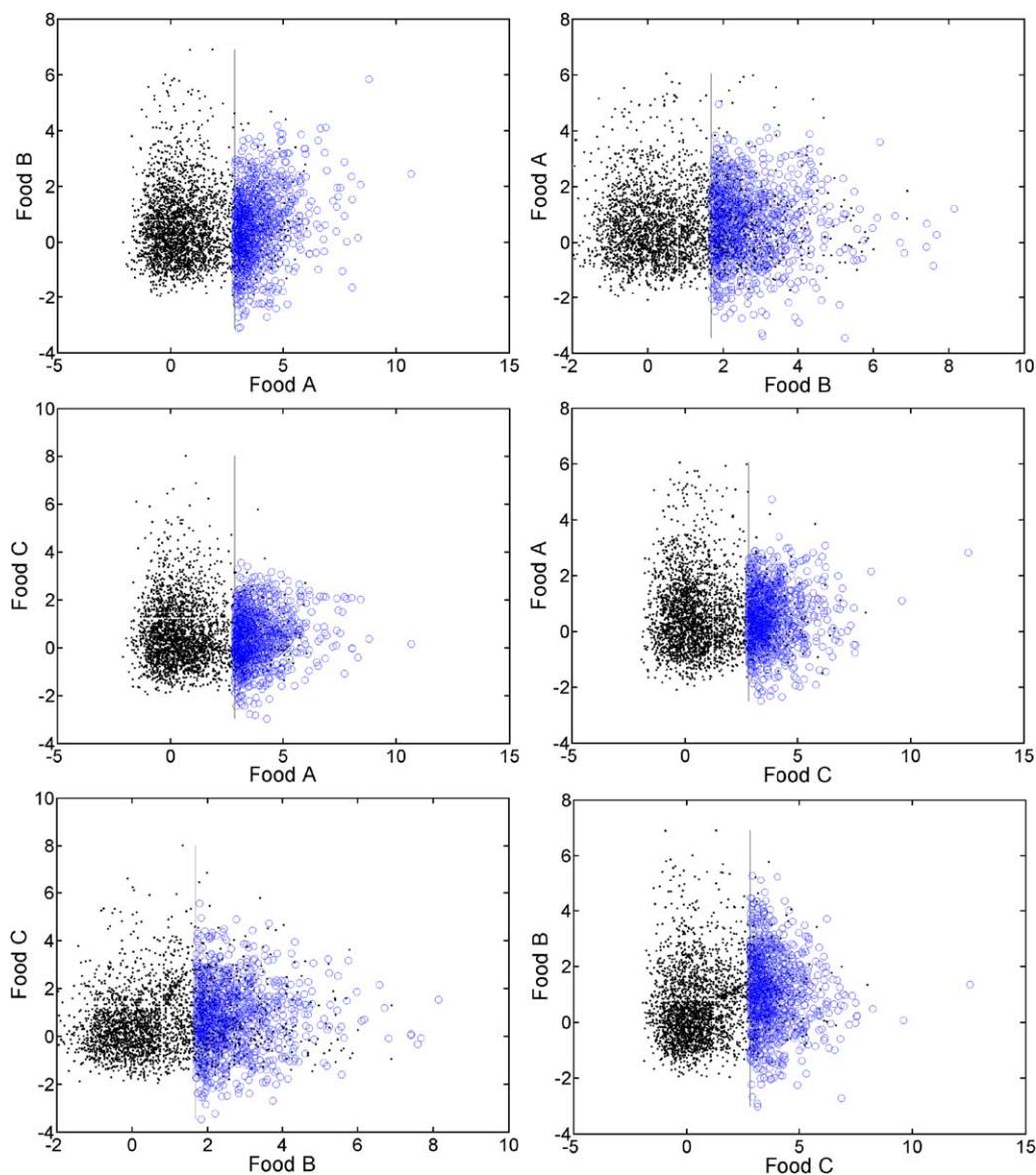


Fig. 3. Daily pesticide intake data (dots) and extrapolations of the 99% tail (circles), in the Gumbel scale. The vertical solid line indicates the conditional threshold u_{HT} as well as the extrapolation threshold v , which are both equal to $0.01 \mu\text{g}/\text{kg BW}$ in this example.

BW are displayed in the figure as circles. The amount of extrapolated extremes clearly depends on the GPD-fit (Appendix 1—step 1) to the variable which is fixed in the conditional model (i.e., variable shown in the horizontal axis in Fig. 4). In this example, many more extremes are extrapolated in models conditional on values of food A, because the GPD model for food A has a longer tail than the models for the other two variables.

Two types of calculations were performed based on these extrapolations—calculation of the probability that total intake exceeds the low and high thresholds introduced in Section 4 and calculation of tail quantiles. The calculation of tail probabilities and quantiles was based on the extrapolations at the tail using Eq. (1) shown in Appendix 1. A comparison of the outcomes of the HT method with the two other methods (empirical data, and univariate

EVT model) is given in Table 2. Both EVT model results agree with the data, especially for lower quantiles. For example, 95% of total intakes are estimated to be lower than $0.2 \mu\text{g}/\text{kg BW}$. However, although the higher threshold of $60 \mu\text{g}/\text{kg BW}$ was never exceeded in the observed data, the EVT models estimate the probability of exceeding this value as 0.1%.

Finally we analysed the effect of eating different amounts of the three foods to acute exposure to the pesticide. Some consumers combine two or three foods, and therefore the corresponding pesticide intakes are associated. For relatively low intakes we can observe this association directly in the data, for large intakes this association can be predicted by the conditional model. The algorithm used to calculate exposure is given in Appendix 2. The probabilities of an intake greater than 10 or $60 \mu\text{g}/\text{kg BW}$

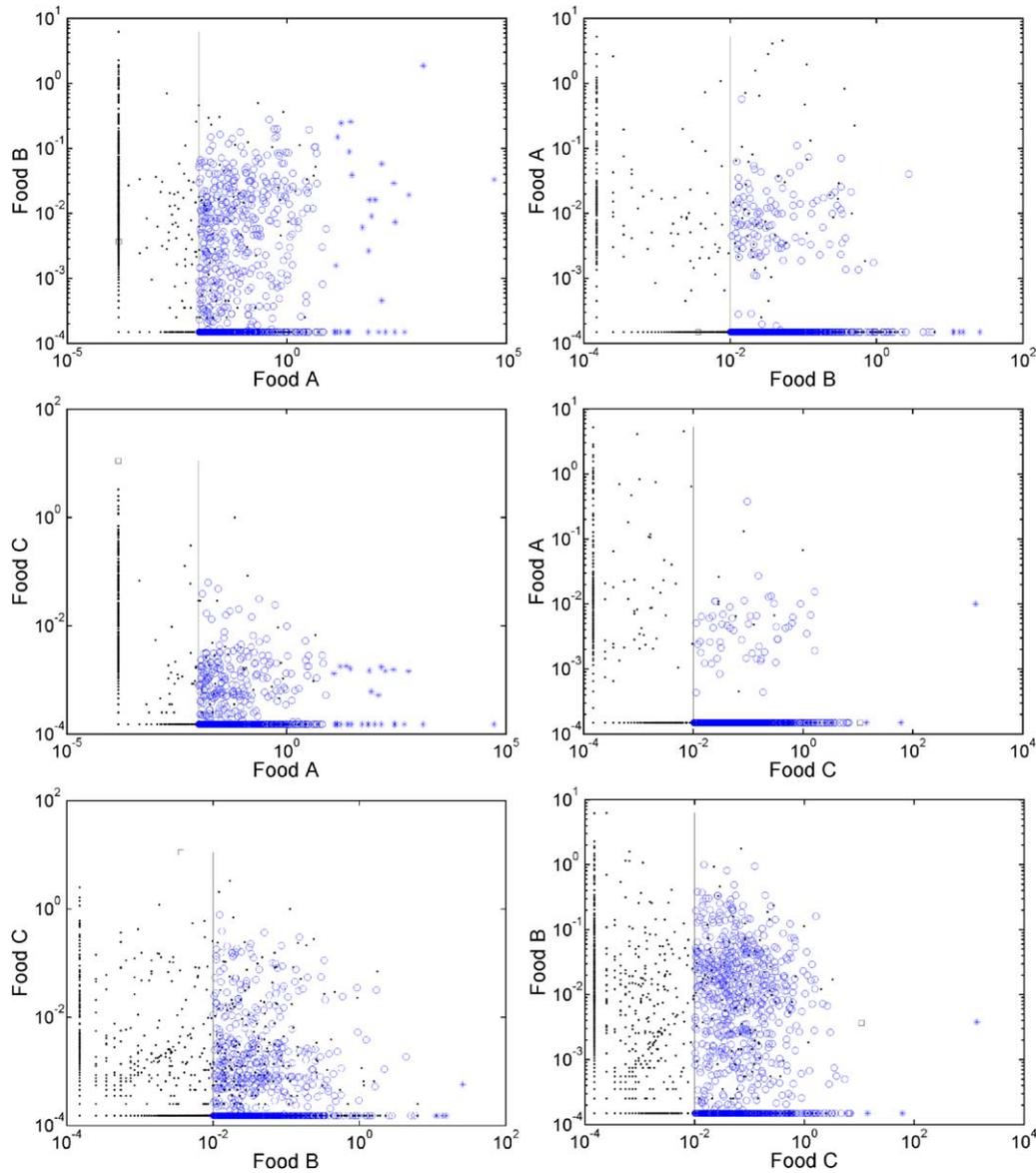


Fig. 4. Daily pesticide intake data (dots and triangle) and extrapolated intakes of the 99% tail (circles and stars), original scale. Intakes exceeding $10 \mu\text{g}/\text{kg}$ BW are marked as a triangle (data) or as a star (extrapolations). The vertical solid line indicates the conditional threshold μ_{HT} as well as the extrapolation threshold v , which are both equal to $0.01 \mu\text{g}/\text{kg}$ BW in this example.

as a function of consumption of one of the foods is shown in Fig. 5. The amounts consumed are given in gram per kilogram body weight (g/kg BW). Among the 2564 consumers in the dataset the maximum amounts consumed were 2.4, 14.8 and 2.8 (g/kg BW) for foods A, B and C, respectively. In Fig. 5 values are shown up to a maximum of 6 g/kg BW (300 g of a commodity for a 50 kg person, 600 g for a 100 kg person). Large pesticide intakes in this case were more likely to occur for a large consumption of food A—consuming 6 g/kg BW of food A would lead to a 25% probability of exceeding an exposure of $10 \mu\text{g}/\text{kg}$ BW, and nearly 2% probability of exceeding $60 \mu\text{g}/\text{kg}$ BW. Food B and food C are safer commodities, as the probability of exceeding an amount of $60 \mu\text{g}/\text{kg}$ BW by consuming large amounts of these foods is negligible.

In the previous example the correlations between the intakes were not very strong, but nonetheless the multivariate analysis reveals a stronger association between food A and B than between A and C. Next, we show the application of the conditional model in another example using data from the Dutch National Food Consumption Survey of 1997/1998 where there is a strong positive correlation between two different food stuffs for small intakes and slightly negative correlation in large intakes. Fig. 6 shows the relation between pesticide intakes from eating two other commodities, (foods D and E) and the GPD fit to the two univariate empirical distributions. The scale and shape GPD parameters for foods D and E were estimated as $\beta = \{0.022, 0.005\}$ and $\xi = \{-0.037, 0.388\}$, respectively. Pesticide intakes from food D up to $0.005 \mu\text{g}/\text{kg}$

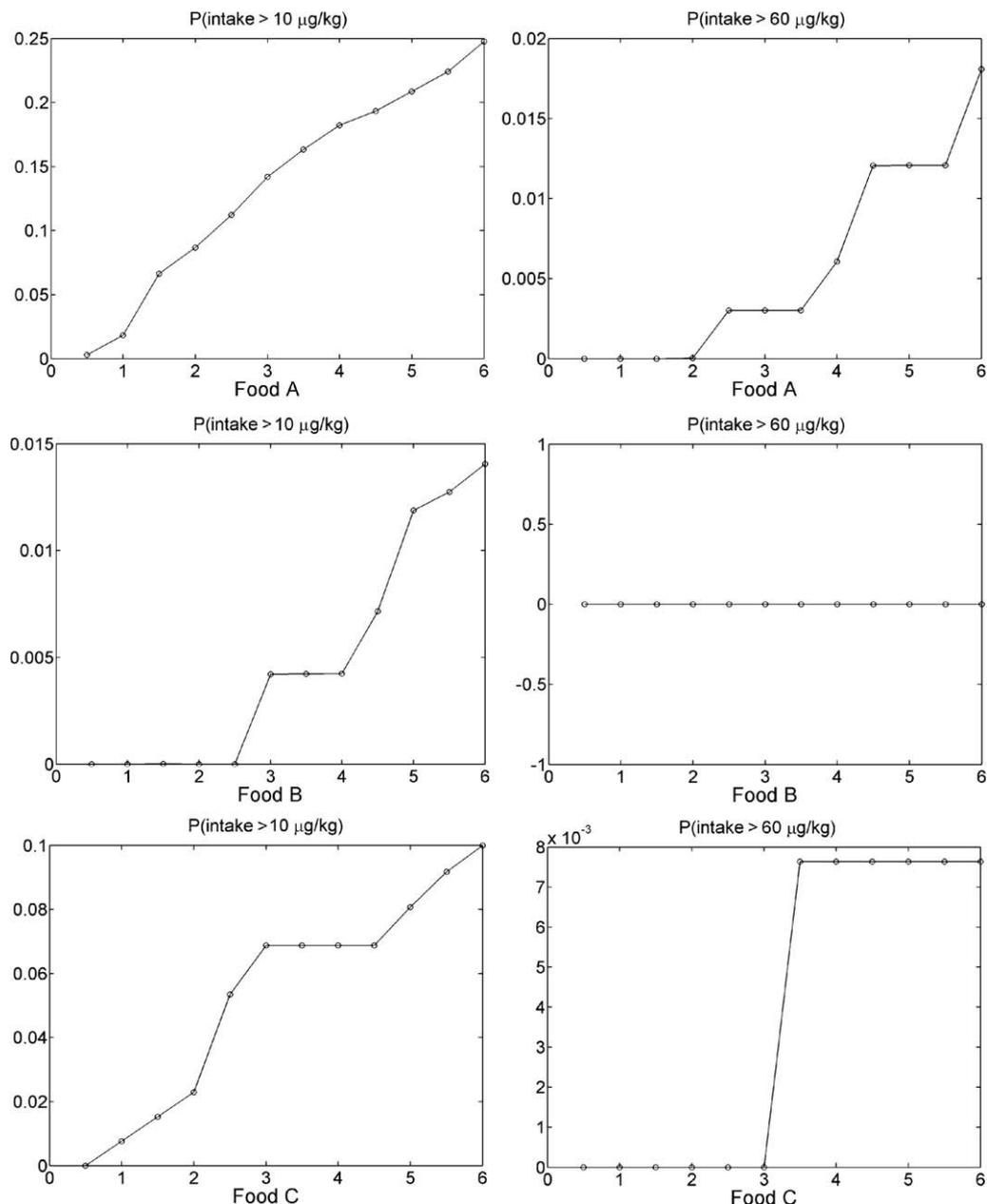


Fig. 5. Probability that daily pesticide intake from all foods exceeds $10 \mu\text{g/kg BW}$ (left) and $60 \mu\text{g/kg BW}$ (right), conditional on the consumption (g/kg BW) of each food.

BW are positively correlated with intakes from food E. The positive correlations are mainly caused by some patterns of a fixed relation (straight line visible in last plot of Fig. 6), and are probably related to standard recipes containing small amounts of the two ingredients. Above the limit of 0.005 the intakes from food D are negatively correlated with intakes from food E (mainly due to many null intakes for E). The extrapolations of intakes from food E conditional on intakes from food D in Fig. 6 display a slight negative correlation for the extrapolated intakes (compare with Figs. 3 and 4). We do not make a complete study of the example, but use it here to show that the conditional tail model correctly captures the relationship in the tails by restricting the information to that region. If univariate

EVT had been used in this example, instead of the multivariate HT model, the two intakes would have been modelled independently, and the risk of exceeding a large intake would possibly be overestimated.

5. Discussion

Extreme value distributions have long been recognised as useful distributions in risk analysis (Konecny and Nachtnebel, 1985; Thas et al., 1997; Piegorsch et al., 1998; Vose, 2000). However, the use of extreme value distributions seems more popular in some fields of applications than in others, and food risk analysis is in this second group. We must ask whether there are good reasons for

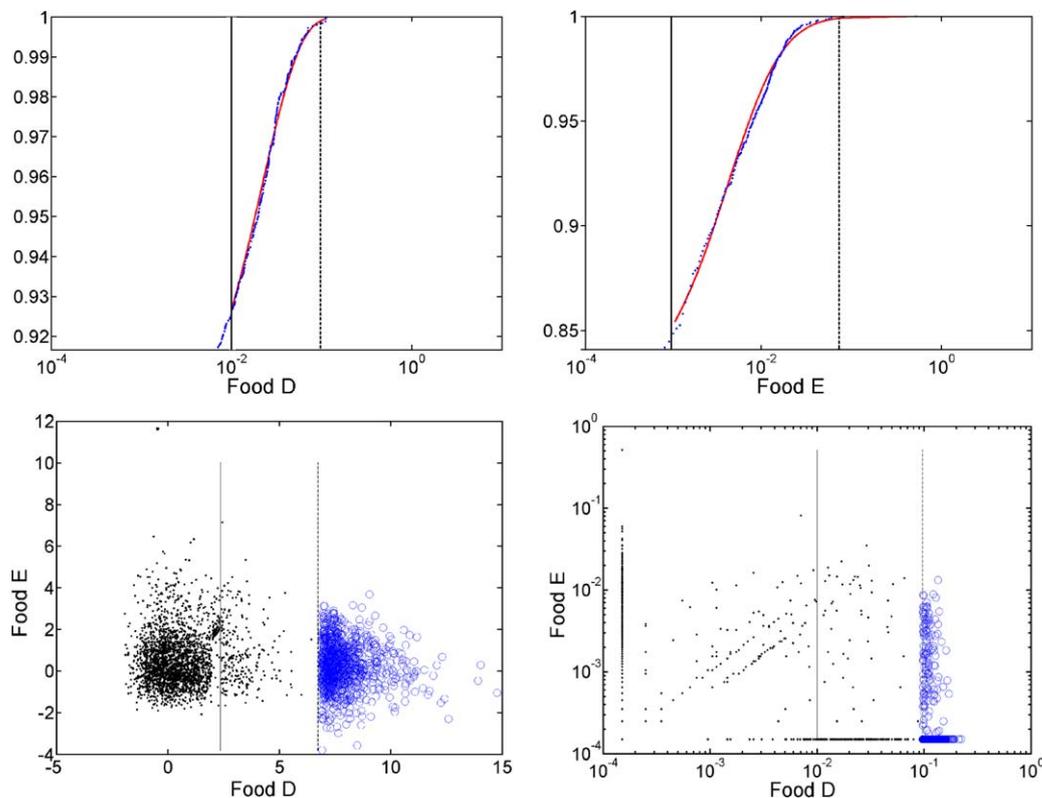


Fig. 6. Application to pesticide intakes from foods D and E: fit of the Generalized Pareto distribution to empirical distributions of intakes from food D (upper left) and food E (upper right), with both marginal and extrapolation thresholds showing (vertical lines, respectively solid and dashed); extrapolations of the 99.9% tail in the Gumbel scale (lower left) and original scale (lower right). The conditional threshold $\mu_{HT} = 0.01 \mu\text{g}/\text{kg BW}$ and the extrapolations threshold are also shown (vertical lines, respectively solid and dashed).

this lack of application. The results of this paper show that fitting extreme value distributions to pesticide intake data is very feasible, and provides more insight in the tail regions that are relevant for assessment of acute risk. This gain is due, of course, to the explicit parametric modelling of values above a well-chosen threshold, where the data below this threshold are just used as they are (in the form of an empirical distribution).

One reason that extreme value methods, which are best developed in the univariate case, have not been used much in dietary risk assessment might be the multivariate character of food consumption and therefore also chemical intake patterns. The recently introduced multivariate Heffernan–Tawn method (Heffernan and Tawn, 2004) was used with two objectives: to extrapolate to extreme cases from observed data and to study tail dependence among pairs of variables. Unlike other EVT methods, it models relationships in the tails whether variables are dependent or independent. Furthermore it can be applied to any number of variables. The conditional model applies to pairs of variables, so different pairs of variables can have different types of relationship. The present method can be very useful to model dietary intake data, where there are groups of correlated variables, and large values are rarely observed.

Although we have used only a small number of variables in an exploratory way to study the applicability of the

model, this method is easily applicable to large numbers of food products. The advantage of the bivariate conditional model (one intake given another) instead of a full joint modelling of the multivariate distribution is the breakdown of dimension into pairs of distinct relations. This is both advantageous in terms of numerical stability of algorithms, and for the interpretability of results, and the restriction to bivariate conditionality seems to present little difficulty in practice.

A feature of the chosen extreme value methodology is that marginal distributions can have either finite upper bounds (e.g., food D in Fig. 6 has an upper bound $0.01 + 0.022/0.037 = 0.60 \mu\text{g}/\text{kg BW}$) or infinite upper bounds (all other foods, Figs. 2 and 6, see Section 3.2). This may run contrary to the intuition that it should be established beforehand whether an upper bound applies. However, it is a difficult matter to quantify biologically or agriculturally plausible upper bounds. We just know that very high consumptions and very high residue concentrations are extremely unlikely. Yet, they can easily be higher than the highest value in the available database, especially when the quantity of data is rather limited. In general maximum values in the data are not suitable upper bounds, and it seems inappropriate to exclude the possibility of more extreme values. Nevertheless the fitting of a Generalised Pareto Distribution has the flexibility to estimate an upper

bound when extremely high values are indeed impossible (see Section 3.2).

In the paper we concentrate on acute risk, using just an example compound. The use of EVT for chronic risk assessment would be different: on the one hand it would be easier because we need not consider a distribution of residue concentrations since fluctuations in intake average out over the long term. Therefore one can multiply the consumption distributions with the average residue level, and then sum over foods, which results in a univariate extreme value problem. On the other hand, a long-term risk assessment is more difficult because the consumption data necessary for chronic risk analysis usually have a hierarchical structure which needs to be modelled (day nested within person). Variance components modelling can solve this under the assumption of normality (Slob, 1993; Nusser et al., 1996), but this problem needs still to be addressed in the context of extreme value theory, and therefore the issue of chronic risk is outside the scope of the current methodology.

Practical risk assessment of pesticides is a field with a long history (FAO, 2004) and many new developments (Ferrier et al., 2002; Hamilton et al., 2004). Whereas practical adoption of probabilistic methods seems to be slow due to the inertia of regulatory communities, it is recognised at the international level that further improvement could be obtained by elaboration of procedures for probabilistic modelling (FAO, 2004). This requires a tiered system such as used in many regulatory frameworks in which risk is first assessed by simple deterministic methods in order to determine where more advanced (probabilistic) modelling is necessary. Multivariate extreme value methods clearly belong to this advanced category, and are not expected to be used routinely. However, they open the possibility of studying important questions concerning risk associated with simultaneous intakes from multiple foods in an objective and data-driven manner.

6. Conclusions

Multivariate EVT models are a suitable method to evaluate risks which are beyond the range of available data. Of course the use of EVT models requires an act of faith in the correctness of a model which seems inevitable in such cases, but EVT provides a sound theoretical basis on which to do this. The approach of Heffernan and Tawn (2004) was found to be a flexible model that allowed visualization of all bivariate relationships in extreme conditions. For example, in our analysis a high pesticide intake from food A is accompanied by a high intake from food B, but not from food C. The probability of a daily total intake larger than 60 μ /kg BW was estimated as 0.1% by applying extreme value models, a result which could not be obtained from direct investigation of the data. A method has been presented to calculate the probability of a total intake (from all foods) exceeding the threshold as a function of the amount consumed of any one food.

Appendix 1

Here we give a description of the Heffernan and Tawn (2004) method for a random vector of pesticide intakes from eating several foods (X_1, \dots, X_p):

1. The Generalized Pareto distribution (GPD) is fitted to the marginal tail of each variable X_i (pesticide intakes from eating food i) for $X_i > u_{\text{GPD}}$, where u_{GPD} is a large threshold, for example the 0.90 quantile. A semi-parametric distribution function \widehat{F}_{X_i} is defined for each marginal, given by the empirical distribution for $X_i < u_{\text{GPD}}$, and by the fitted GPD for $X_i \geq u_{\text{GPD}}$. Fig. 2 shows both the empirical distribution and pareto fit for pesticide intakes from three foods.
2. Obtain new variables $Y_i = -\ln\{-\ln\widehat{F}_{X_i}\}$. The transformed Y_i have a Gumbel marginal distribution.
3. Fit a conditional model to the tails of the Y_i variables i.e., for $Y_i \geq u_{\text{HT}}$, where u_{HT} is another threshold, not necessarily equivalent (when back transformed) to u_{GPD} . Next all $Y_j, j \neq i$ are modelled as semi-parametric functions of Y_i : $Y_{ji} = a_{ji}y_i + y_i^{b_{ji}}z_{ji}$, where a_{ji} and b_{ji} are unknown parameters, and z_{ji} is a standardized residual, independent of Y_i , with mean μ_{ji} and standard deviation σ_{ji} . If $a_{ji} = 0$ and $b_{ji} < 0$ then an alternative model is fitted, $Y_{ji} = c_{ji} - d_{ji} \ln y_i + y_i^{b_{ji}}z_{ji}$, where c_{ji} and d_{ji} are unknown parameters. See Heffernan and Tawn (2004) for the background of the models. First, parameters a_{ji} , b_{ji} , μ_{ji} and σ_{ji} are estimated simultaneously by maximum likelihood assuming that $Y_{ji} \sim N(\mu_{ji}(y_i), \sigma_{ji}(y_i))$. If $a_{ji} = 0$ and $b_{ji} < 0$ then the second model is used and parameters b_{ji} , c_{ji} , d_{ji} , μ_{ji} and σ_{ji} estimated by the same method. As a result of this conditional model, Y_{ji} has mean given either by $\mu_{ji}(y_i) = a_{ji}y_i + y_i^{b_{ji}}\mu_{ji}$, or by $\mu_{ji}(y_i) = c_{ji} - d_{ji} \log(y_i) + y_i^{b_{ji}}\mu_{ji}$ if $a_{ji} = 0$ and $b_{ji} < 0$. Standard deviation of Y_{ji} is $\sigma_{ji}(y_i) = y_i^{b_{ji}}\sigma_{ji}$. For p variables Y_1, \dots, Y_p a total of $p(p-1)$ conditional models are obtained.
4. The conditional model from step 3 can be used to extrapolate to extreme regions. Extrapolation is performed in two steps, first simulate $y_i > v_i$, where the extrapolation threshold v_i is not smaller than u_{GPD} or u_{HT} . Simulate $z_{ji} \sim N(\mu_{ji}, \sigma_{ji})$, independently of y_i . Then, apply the semi-parametric regression model to obtain y_{ji} for all $j \neq i$. For each y_i , a whole vector (y_1, \dots, y_p) conditional on y_i is thus obtained. Repeat to obtain a distribution of extrapolations. A total of p sets of y vectors are extrapolated, conditional respectively on y_1, y_2 , etc. Fig. 3 shows extrapolations from the conditional model for pesticide intakes.
5. Back transform the extrapolations to the original scale: $x_i = \widehat{F}_{X_i}^{-1}\{\exp[-\exp(-y_i)]\}$.

We define extremes here as the set of vectors such that $\{\sum x_i > v\}$, i.e., if total intake of pesticide exceeds v . We shall call this set C . C can be partitioned into p subsets $C = \cup C_i$ where $C_i = \{\sum x_i > v\}, F_{X_i} > F_{X_j}$ for all j . Then

$$P(X \in C) = \sum_{i=1}^p P(X \in C_i) = \sum_{i=1}^p P(X \in C_i | X_i > v_i) P(X_i > v_i) \quad (1)$$

This divides each item of the sum into two components, a marginal model for the tail, i.e., $P(X_i > v_i)$, and a conditional model i.e., $P(X \in C_i | X_i > v_i)$. Here v_i is the extrapolation threshold, so the calculation of tail probabilities are based on expression (1), as the sum of a food of the two components. Probability $P(X_i > v_i)$ is calculated from the univariate GPD fit to X_i . The other component, $P(X \in C_i | X_i > v_i)$ can be calculated from the extrapolations where x is conditional to x_i , as the proportion of extreme points in the simulations (amount of extremes in simulations divided by total number of simulations).

In the pesticide example we calculate $P(X \in C_i | X_i > v_{x_i})$ as the proportion of simulations satisfying both $\{\sum x_i > v\}$ and $\{F_{x_i} > F_{x_j}, \forall j \neq i\}$.

Appendix 2

Here we describe a novel algorithm to estimate the risk of exceeding a large pesticide intake given the consumed amount of a food.

- Let variable θ_i be the amount, in g/kg BW, of food i consumed in one day, for each $i \in \{A, B, C\}$. We take a number of $k = 12$ discrete values of θ_i , to form a regular grid. For each discrete value $\theta_i \in \{0.5, 1, \dots, 6\}$ in the grid we do the following:
 - Generate n_1 intake values by combining θ_i with the empirical concentrations of pesticide. Each observed concentration c_i is given a weight corresponding to its empirical probability, i.e., $w_1 = 1/N$ for positive concentrations, and $w_1 = N_0/N$ for non-detects (N is the total number of concentration samples in the dataset, and N_0 is the number of non-detects). Here all non-detects are replaced by the limit of reporting (LOR). The resulting values are observations from $X_i | \theta_i$, the intake of pesticide through eating amount θ_i of food i . Each intake x_i is given the weight corresponding to the empirical concentrations distribution. The variation in the intakes corresponds thus to variation in pesticide concentrations measured in food i . N.B.: $n_1 = N - N_0 + 1$.
 - If $x_i > u_{HT}$ (where u_{HT} is a very low value in our example, for example 0.02) then apply the previously fitted H-T conditional model a number $n_2 = 100$ times to obtain $\{X_j | X_i\}$ for each $j \neq i$. We obtain a distribution of conditional X_j intakes, $j \neq i$, for each $\{x_i > u_{HT}\}$. If $x_i \leq u_{HT}$ then sample $n_2 = 20$ times the x_j from the nearest neighbours of x_i (at a maximum distance of say 0.02), with replacement, in the dataset of intakes (empirical joint distribution). The sample size n_2 was set equal to 20 because was found large enough to

capture the diversity at the small neighbourhood. Select neighbours randomly, with probability inversely proportional to the distance of x_i to its neighbour. Update corresponding weight: $w_2 = w_1/n_2$.

- Find the distribution of total intakes conditional to a consumed amount θ_i , by summing each x_i to all n_2 corresponding x_j 's. Each total intake has weight w_2 .
 - Find $P\left\{\sum_{j=1}^3 X_j > v | \theta_i\right\} = \sum_{x_i} w_2 I(\sum X_j > v | \theta_i)$ for each fixed $\theta_i, v \in \{10, 60\}$.
- Repeat the previous step for each value of the variable θ_i in turn.
 - Repeat the previous steps for each food.

As outcome of this procedure we obtain a series of total intakes for each θ_i (consumed amount of food i), and respective weights. That is, we obtain a “distribution” of total intakes conditional on θ_i , though the weights do not necessarily add up to one. The probability of exceeding a large intake is obtained by rescaling the weights to sum to one and adding the weights of exceeding intakes.

References

- Brown, M.A., Brix, K.A., 1998. Review of health consequences from high-, intermediate- and low-level exposure to organophosphorus nerve agents. *Journal of Applied Toxicology* 18, 393–408.
- Capéraà, P., Fougères, A.-L., Genest, C., 1997. Estimation of bivariate extreme value copulas. *Biometrika* 84, 567–577.
- Coles, S.G., 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London.
- Coles, S.G., Heffernan, J.E., Tawn, J.A., 1999. Dependence measures for extreme value analysis. *Extremes* 2 (4), 339–365.
- Coles, S.G., Tawn, J.A., 1994. Statistical methods for multivariate extremes: an application to structural design. *Applied Statistics* 43, 1–48.
- Cullen, A.C., Frey, H.C., 1999. *Probabilistic Techniques in Exposure Assessment, A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Plenum Press, New York.
- de Haan, L., de Ronde, J., 1998. Sea and wind: multivariate extremes at work. *Extremes* 1, 7–45.
- Dich, J., Zahm, S.H., Hanberg, A., 1997. Pesticides and cancer. *Cancer Causes and Control* 8, 420–443.
- DNFCS. Zo eet Nederland, Resultaten van de Voedselconsumptie-peiling 1997–1998. Voedingscentrum, Den Haag, 1998.
- Dupuis, D.J., 2005. Ozone concentrations: a robust analysis of multivariate extremes. *Technometrics* 47, 191–201.
- Efron, B., 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1–26.
- Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. *Modelling Extremal Events*. Springer, London.
- FAO, 2004. *Pesticide residues in food—2003*. FAO Plant Production and Protection Paper, vol. 176.
- Ferrier, H., Nieuwenhuijsen, M., Boobis, A., Elliott, P., 2002. Current knowledge and recent developments in consumer exposure assessment of pesticides: a UK perspective. *Food Additives and Contaminants* 19, 837–852.
- Hamilton, D., Ambrus, A., Dieterle, R., Felsot, A., Harris, C., Petersen, B., Racke, K., Wong, S.S., Gonzalez, R., Tanaka, K., Earl, M., Roberts, G., Bhula, R., 2004. Pesticide residues in food—acute dietary exposure. *Pest Management Science* 60, 311–339.

- Heffernan, J.E., Tawn, J.A., 2004. A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B* 66 (3), 497–546.
- Konecny, F., Nachtnebel, H.P., 1985. Extreme value processes and the evaluation of risk in flood analysis. *Applied Mathematical Modelling* 9 (1), 11–15.
- Kotz, S., Nadarajah, S., 2000. *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.
- Martins, A.P., Ferreira, H., 2005. Measuring the extremal dependence. *Statistics and Probability Letters* 73, 99–103.
- The MathWorks, Matlab, the language of technical computing. Version 7 (release 14) edition, June 2004. Available from: <www.mathworks.com>.
- Nusser, S.M., Carriquiry, A.L., Dodd, K.W., Fuller, W.A., 1996. A semiparametric approach to estimating usual intake distributions. *Journal of the American Statistical Association* 91, 1440–1449.
- Paulo, M.J., 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 (8), 759–766.
- Pickands, J., 1975. Statistical inference using extreme order statistics. *The Annals of Statistics* 3 (1), 119–131.
- Piegorsch, W.W., Smith, E.P., Edwards, D., 1998. Statistical advances in environmental science. *Statistical Science* 13 (2), 186–208.
- Slob, W., 1993. Modelling long-term exposure of the whole population to chemicals in food. *Risk Analysis* 13 (5), 525–530.
- Thas, O., Vanrolleghem, P., Kops, B., 1997. Extreme value statistics: potential benefits in water quality management. *Water Science and Technology* 36 (5), 133–140.
- Tressou, J., Crépet, A., Bertail, P., Feinberg, M.H., Leblanc, J.Ch., 2004. Probability exposure assessment to food chemicals based on extreme value theory. application to heavy metals from fish and sea products. *Food and Chemical Toxicology* 42, 1349–1358.
- van Dooren, M.M.H., Boeijen, I., van Klaveren, J.D., van Donkersgoed, G., 1995. Conversie van consumeerbare voedingsmiddelen naar primaire agrarische producten (conversion of foods to primary agricultural products). Report 95.17, RIKILT, Wageningen.
- van Klaveren, J.D., 1999. Quality programme for agricultural products. Results residue monitoring in the netherlands, Technical report, RIKILT Institute of Food Safety, Wageningen University and Research Centre, Wageningen. Available from: <<http://www2.rikilt.dlo.nl/kap/index.html>>.
- Verger, Ph., Ireland, J., Møller, A., Abravicius, J.A., De Henauw, S., Naska, A., 2002. Improvement of comparability of dietary intake assessment using currently available individual food consumption surveys. *European Journal of Clinical Nutrition* 56 (Suppl 2), S18–S24.
- Vose, D., 2000. *Risk analysis: A Quantitative Guide*, second ed. Wiley, Chichester.