# Models in 'morse' package

September 26, 2019

This document describes the statistical models used in 'morse' to analyze survival and reproduction data, and as such serves as a mathematical specification of the package. For a more practical introduction, please consult the "Tutorial" vignette ; for information on the structure and contents of the library, please consult the reference manual.

Model parameters are estimated using Bayesian inference, where posterior distributions are computed from the likelihood of observed data combined with prior distributions on the parameters. These priors are specified after each model description.

## 1 Survival toxicity tests

In a survival toxicity test, subjects are exposed to a measured concentration of a contaminant over a given period of time and the number of surviving organisms is measured at certain time points during exposure. In most standard toxicity tests, the concentration is held constant throughout the whole experiment, which we will assume for *1.1 Analysis of target time survival toxicity tests*, but not for *1.2 Toxicokinetic-Toxicodynamic modeling* which can handled time variable exposure. In the case of constant exposure, an experiment is generally replicated several times and also repeated for various levels of the contaminant. For time-variable exposure, a profile of exposure is usually unique, and the experiment is repeated with several profiles of exposures.

In so-called *target time* toxicity tests, the mortality is usually analyzed at the end of the experiment. The chosen time point for this analysis is called *target time*. Let us see how this particular case is handled in 'morse'.

### 1.1 Analysis of target time survival toxicity tests

A dataset from a target time survival toxicity test is a collection $D = \{(c_i, n_i^{init}, n_i)\}_i$ of experiments, where $c_i$ is the tested concentration, $n_i^{init}$ the initial number of organisms and $n_i$ the number of organisms at the chosen target time. Triplets such that $c_i = 0$ correspond to control experiments.

#### 1.1.1 Modelling

In the particular case of target time analysis, the model used in 'morse' is defined as follows. Let $t$ be the target time in days. We suppose the *mean survival rate after $t$ days* is given by a function $f$ of the contaminant level $c$. We also suppose that the death of two organisms are two independent events. Hence, given an initial number $n_i^{init}$ of organisms in the toxicity test at concentration $c_i$, we obtain that the number $N_i$ of surviving organisms at time $t$ follows a binomial distribution:

$$N_i \sim \mathcal{B}(n_i^{init}, f(c_i))$$

Note that this model neglects inter-replicate variations, as a given concentration of contaminant implies a fixed value of the survival rate. There may be various possibilities for $f$. In 'morse' we assume a three

parameters log-logistic function:

$$f(c) = \frac{d}{1 + (\frac{c}{e})^b}$$

where $b$, $e$ and $d$ are (positive) parameters. In particular $d$ corresponds to the survival rate in absence of contaminant and $e$ corresponds to the $LC_{50}$. Parameter $b$ is related to the effect intensity of the contaminant.

### 1.1.2 Inference

Posterior distributions for parameters $b$, $d$ and $e$ are estimated using the JAGS software [10] with the following priors:

- we assume the range of tested concentrations in an experiment is chosen to contain the $LC_{50}$ with high probability. More formally, we choose:

  $$\log_{10} e \sim \mathcal{N}\left(\frac{\log_{10}(\min_i c_i) + \log_{10}(\max_i c_i)}{2}, \frac{\log_{10}(\max_i c_i) - \log_{10}(\min_i c_i)}{4}\right)$$

  which implies $e$ has a probability slightly higher than 0.95 to lie between the minimum and the maximum tested concentrations.

- we choose a quasi non-informative prior distribution for the shape parameter $b$:

  $$\log_{10} b \sim \mathcal{U}(-2, 2)$$

The prior on $d$ is chosen as follows: if we observe no mortality in control experiments then we set $d = 1$, otherwise we assume a uniform prior for $d$ between 0 and 1.

## 1.2 Toxicokinetic-Toxicodynamic modeling

For datasets featuring time series measurements, more complete models can be used to estimate the effect of a contaminant on survival. We assume the toxicity test consists in exposing an initial number $n_i^0$ of organisms to a concentration $c_i(t)$ of contaminant (constant or time-variable), and following the number $n_i^k$ of survivors at time $t_k$ (with $t_0 < t_1 < \cdots < t_m$ and $t_0 = 0$), thus providing a collection $D = (c_i, t_k, n_i^k)_{i,k}$ of experiments. In 'morse', we propose two Toxicokinetic-Toxicodynamic (TKTD) models belonging to the General Unified Threshold model for Survival (GUTS) [4, 5]. One is known as the *reduced stochastic death* model [6] or GUTS-SD and the other is the *reduced organism tolerance* model or GUTS-IT, which we describe now.

**GUTS Modelling** The number of survivors at time $t_k$ given the number of survivors at time $t_{k-1}$ is assumed to follow a binomial distribution:

$$N_i^k \sim \mathcal{B}(n_i^{k-1}, f_i(t_{k-1}, t_k))$$

where $f_i$ is the conditional probability of survival at time $t_k$ given survival at time $t_{k-1}$ under concentration $c_i(t)$. Denoting $S_i(t)$ the probability of survival at time $t$, we have:

$$f_i(t_{k-1}, t_k) = \frac{S_i(t_k)}{S_i(t_{k-1})}$$

The formulation of the survival probability $S_i(t)$ in GUTS [4] is given by integrating the *instantaneous mortality rate* $h_i$:

$$S_i(t) = \exp\left(\int_0^t -h_i(u)\mathrm{d}u\right) \tag{1}$$

Table 1: Parameters and symbols used for GUTS-SD and GUTS-IT models. Alternative symbols are used within pubications (see for instance [4, 3, 5]). The unit $[D]$ refers to unit of actual damage, $n.d$ for non dimensional. For GUTS-IT model, we assume a log-logistic distributions, but other distributions are occasionally used [1].

| Parameters | Symbols | Alternative symbols | Units | Models |
|---|---|---|---|---|
| Background hazard rate | $h_b$ | $m_0$ | time$^{-1}$ | SD and IT |
| Dominant toxicokinetic rate constant | $k_d$ | $k_e$ | time$^{-1}$ | SD and IT |
| Threshold for effects | $z_w$ | *NEC*, $z$ | $[D]$ | SD |
| Killing rate constant | $b_w$ | $k_s$, $k_k$ | $[D]^{-1}$.time$^{-1}$ | SD |
| Median of the threshold effect distribution | $m_w$ | $\alpha$ | $[D]$ | IT |
| Shape of the threshold effect distribution | $\beta$ | - | n.d. | IT |

In the model, function $h_i$ is expressed using the internal concentration of contaminant (that is, the concentration inside an organism) $C_i^{\text{INT}}(t)$. More precisely:

$$h_i(t) = b_w \max(C_i^{\text{INT}}(t) - z_w, 0) + h_b$$

where (see Table 1):

- $b_w$ is the *killing rate* and expressed in concentration$^{-1}$.time$^{-1}$ ;
- $z_w$ is the so-called *no effect concentration* and represents a concentration threshold under which the contaminant has no effect on organisms ;
- $h_b$ is the *background mortality* (mortality in absence of contaminant), expressed in time$^{-1}$.

The internal concentration is assumed to be driven by the external concentration, following:

$$\frac{\mathrm{d}C_i^{\text{INT}}}{\mathrm{d}t}(t) = k_d(c_i(t) - C_i^{\text{INT}}(t)) \tag{2}$$

We call parameter $k_d$ of Eq. (2) the "dominant rate constant" (expressed in time$^{-1}$). It represents the speed at which the internal concentration in contaminant converges to the external concentration. The model could be equivalently written using an internal damage instead of an internal concentration as a dose metric [4].

If we denote $f_z(z_w)$ the probability distribution of the no effect concentration threshold, $z_w$, then the survival function is given by:

$$S(t) = \int_0^t S_i(t) f_z(z_w) \mathrm{d}z_w = \int \exp\left(\int_0^t -h_i(u)\mathrm{d}u\right) f_z(z_w)\mathrm{d}z_w \tag{3}$$

Then, the calculation of $S(t)$ depends on the model of survival, GUTS-SD or GUTS-IT [4]:

**GUTS-SD** In GUTS-SD, all organisms are assumed to have the same internal concentration threshold (denoted $z_w$), and, once exceeded, the instantaneous probability to die increases linearly with the internal concentration. In this situation, $f_z(z_w)$ is a Dirac delta distribution, and the survival rate is given by Eq. (1).

**GUTS-IT** In GUTS-IT, the threshold concentration is distributed among all the organisms, and once exceeded for one organism, this organism dies immediately. In other words, the killing rate is infinitely high (e.g. $k_k = +\infty$), and the survival rate is given by:

$$S_i(t) = e^{-h_b t} \int_{\max\limits_{0<\tau<t}(C_i^{\text{INT}}(\tau))}^{+\infty} f_z(z_w)\mathrm{d}z_w = e^{-h_b t}(1 - F_z(\max_{0<\tau<t} C_i^{\text{INT}}(\tau)))$$

where $F_z$ denotes the cumulative distribution function of $f_z$.

Here, the exposure concentration $c_i(t)$ is not supposed constant. In the case of time variable exposure concentration, we use an midpoint ODE integrator (also known as modified Euler, or Runge-Kutta 2) to solve models GUTS-SD and GUTS-IT. When the exposure concentration is constant, then, explicit formulation of integrated equations are used. We present them in the next subsection.

### 1.2.1 For constant concentration exposure

If $c_i(t)$ is constant, and assuming $C_i^{\mathrm{INT}}(0) = 0$, then we can integrate the previous equation (2) to obtain:

$$C_i^{\mathrm{INT}}(t) = c_i(1 - e^{-k_d t}) \tag{4}$$

**GUTS-SD** In the case $c_i < z_w$, the organisms are never affected by the contaminant:

$$S_i(t) = \exp(-h_b t) \tag{5}$$

When $c_i > z_w$, it takes time $t_i^z$ before the internal concentration reaches $z_w$, where:

$$t_i^z = -\frac{1}{k_d} \log\left(1 - \frac{z_w}{c_i}\right).$$

Before that happens, Eq. (5) applies, while for $t > t_i^z$, integrating Eq. (1) results in:

$$S_i(t) = \exp\left(-h_b t - b_w(c_i - z_w)(t - t_i^z) - \frac{b_w c_i}{k_d}\left(e^{-k_d t} - e^{-k_d t_i^z}\right)\right)$$

In brief, given values for the four parameters $h_b$, $b_w$, $k_d$ and $z_w$, we can simulate trajectories by using $S_i(t)$ to compute conditional survival probabilities. In 'morse', those parameters are estimated using Bayesian inference. The choice of priors is defined hereafter.

**GUTS-IT** With constant concentration, Eq. 4 provides that $C_i^{\mathrm{INT}}(t)$ is an increasing function, meaning that:

$$\max_{0<\tau<t}(C_i^{\mathrm{INT}}(\tau)) = c_i(1 - e^{-k_d t})$$

Therefore, assuming a log-logistic distribution for $f_z$ yields:

$$S_i(t) = \exp(-h_b t)\left(1 - \frac{1}{1 + \left(\frac{c_i(1-\exp(-k_d t))}{m_w}\right)^{-\beta}}\right)$$

where $m_w > 0$ is the scale parameter (and also the median) and $\beta > 0$ is the shape parameter of the log-logistic distribution.

### 1.2.2 Inference

Posterior distributions for all parameters $h_b$, $b_w$, $k_d$, $z_w$, $m_w$ and $\beta$ are computed with JAGS [10]. We set prior distributions on those parameters based on the actual experimental design used in a toxicity test. For instance, we assume $z_w$ has a high probability to lie within the range of tested concentrations. For each parameter $\theta$, we derive in a similar manner a minimum ($\theta^{\min}$) and a maximum ($\theta^{\max}$) value and state that the prior on $\theta$ is a log-normal distribution [3]. More precisely:

$$\log_{10}\theta \sim \mathcal{N}\left(\frac{\log_{10}\theta^{\min} + \log_{10}\theta^{\max}}{2}, \frac{\log_{10}\theta^{\max} - \log_{10}\theta^{\min}}{4}\right)$$

With this choice, $\theta^{\min}$ and $\theta^{\max}$ correspond to the 2.5 and 97.5 percentiles of the prior distribution on $\theta$. For each parameter, this gives:

- $z_w^{\min} = \min_{i,c_i \neq 0} c_i$ and $z_w^{\max} = \max_i c_i$, which amounts to say that $z_w$ is most probably contained in the range of experimentally tested concentrations ;
- similarly, $m_w^{\min} = \min_{i,c_i \neq 0} c_i$ and $m_w^{\max} = \max_i c_i$ ;
- for background mortality rate $h_b$, we assume a maximum value corresponding to situations where half the indivuals are lost at the first observation time in the control (time $t_1$), that is:

  $$e^{-h_b^{\max} t_1} = 0.5 \Leftrightarrow h_b^{\max} = -\frac{1}{t_1} \log 0.5$$

  To derive a minimum value for $h_b$, we set the maximal survival probability at the end of the toxicity test in control condition to 0.999, which corresponds to saying that the average lifetime of the considered species is at most a thousand times longer than the duration of the experiment. This gives:

  $$e^{-h_b^{\min} t_m} = 0.999 \Leftrightarrow h_b^{\min} = -\frac{1}{t_m} \log 0.999$$

- $k_d$ is the parameter describing the speed at which the internal concentration of contaminant equilibrates with the external concentration. We suppose its value is such that the internal concentration can at most reach 99.9% of the external concentration before the first time point, implying the maximum value for $k_d$ is:

  $$1 - e^{-k_d^{\max} t_1} = 0.999 \Leftrightarrow k_d^{\max} = -\frac{1}{t_1} \log 0.001$$

  For the minimum value, we assume the internal concentration should at least have risen to 0.1% of the external concentration at the end of the experiment, which gives:

  $$1 - e^{-k_d^{\min} t_m} = 0.001 \Leftrightarrow k_d^{\min} = -\frac{1}{t_m} \log 0.999$$

- $b_w$ is the parameter relating the internal concentration of contaminant to the instantaneous mortality. To fix a maximum value, we state that between the closest two tested concentrations, the survival probability at the first time point should not be divided by more than one thousand, assuming (infinitely) fast equilibration of internal and external concentrations. This last assumption means we take the limit $k_d \to +\infty$ and approximate $S_i(t)$ with $\exp(-(h_b + b_w(c_i - z_w))t)$. Denoting $\Delta^{\min}$ the minimum difference between two tested concentrations, we obtain:

  $$e^{-b_w^{\max} \Delta^{\min} t_1} = 0.001 \Leftrightarrow b_w^{\max} = -\frac{1}{\Delta^{\min} t_1} \log 0.001$$

  Analogously we set a minimum value for $b_w$ saying that the survival probability at the last time point for the maximum concentration should not be higher than 99.9% of what it is for the minimal tested concentration. For this we assume again $k_d \to +\infty$. Denoting $\Delta^{\max}$ the maximum difference between two tested concentrations, this leads to:

  $$e^{-b_w^{\min} \Delta^{\max} t_m} = 0.001 \Leftrightarrow b_w^{\min} = -\frac{1}{\Delta^{\max} t_m} \log 0.999$$

- for the shape parameter $\beta$, we used a quasi non-informative log-uniform distribution:

  $$\log_{10} \beta \sim \mathcal{U}(-2, 2)$$

# 2 Reproduction toxicity tests

In a reproduction toxicity test, we observe the number of offspring produced by a sample of adult organisms exposed to a certain concentration of a contaminant over a given period of time. The offspring (young organisms, clutches or eggs) are regularly counted and removed from the medium at each time point, so that the reproducing population cannot increase. It can decrease however, if some organisms die during the experiment. The same procedure is usually repeated at various concentrations of contaminant, in order to establish a quantitative relationship between the reproduction rate and the concentration of contaminant in the medium.

As already mentionned, it is often the case that part of the organisms die during the observation period. In previous approaches, it was proposed to consider the cumulated number of reproduction outputs without accounting for mortality [7, 8], or to exclude replicates where mortality occurred [9]. However, organisms may have reproduced before dying and thus contributed to the observed response. In addition, organisms dying the first are probably the most sensitive, so the information on reproduction of these prematurely dead organisms is valuable ; ignoring it is likely to bias the results in a non-conservative way. This is particularly critical at high concentrations, when mortality may be very high.

In a toxicity test, mortality is usually regularly recorded, *i.e.* at each time point when reproduction outputs are counted. Using these data, we can approximately estimate for each organism the period it has stayed alive (which we assume coincides with the period it may reproduce). As commonly done in epidemiology for incidence rate calculations, we can then calculate, for one replicate, the total sum of the periods of observation of each organism before its death (see next paragraph). This sum can be expressed as a number of organism-days. Hence, reproduction can be evaluated through the number of outputs per organism-day.

In the following, we denote $M_{ijk}$ the observed number of surviving organisms at concentration $c_i$, replicate $j$ and time $t_k$.

## 2.1 Estimation of the effective observation period

We define the effective observation period as the sum for all organisms of the time they spent alive in the experiment. It is counted in organism-days and will be denoted $NID_{ij}$ at concentration $c_i$ and replicate $j$. As mentionned earlier, mortality is observed at particular time points only, so the real life time of an organism is unknown and in practice we use the following simple estimation: if an organism is alive at $t_k$ but dead at $t_{k+1}$, its real life time is approximated as $\frac{t_{k+1}+t_k}{2}$.

With this assumption, the effective observation period at concentration $c_i$ and replicate $j$ is then given by:

$$NID_{ij} = \sum_k M_{ij(k+1)}(t_{k+1} - t_k) + (M_{ijk} - M_{ij(k+1)}) \left( \frac{t_{k+1} + t_k}{2} - t_k \right)$$

## 2.2 Target time analysis

In this paragraph, we describe our so-called "target time analysis", where we model the cumulated number of offspring up to a target time as a function of contaminant concentration and effective observation time in this period (cumulated life times of all organisms in the experiment, as described above). A more detailed presentation can be found in [2].

We keep the convention that index $i$ is used for concentration levels and $j$ for replicates. The data will therefore correspond to a set $\{(NID_{ij}, N_{ij})\}_i$ of pairs, where $NID_{ij}$ denotes the effective observation period and $N_{ij}$ the number of reproduction output. These observations are supposed to be drawn independently from a distribution that is a function of the level of contaminant $c_i$.

### 2.2.1 Modelling

We assume here that the effect of the considered contaminant on the reproduction rate[1] does not depend on the exposure period, but only on the concentration of the contaminant. More precisely, the reproduction rate in an experiment at concentration $c_i$ of contaminant is modelled by a three-parameters log-logistic model, that writes as follows:

$$f(c;\theta) = \frac{d}{1+(\frac{c}{e})^b} \quad \text{with} \quad \theta = (e,b,d)$$

Here $d$ corresponds to the reproduction rate in absence of contaminant (control condition) and $e$ to the value of the $EC_{50}$, that is the concentration dividing the average number of offspring by two with respect to the control condition. Then the number of reproduction outputs $N_{ij}$ at concentration $c_i$ in replicate $j$ can be modelled using a Poisson distribution:

$$N_{ij} \sim Poisson(f(c_i;\theta) \times NID_{ij})$$

This model is later referred to as "Poisson model". If there happens to be a non-negligible variability of the reproduction rate between replicates at some fixed concentrations, we propose a second model, named "gamma-Poisson model", stating that:

$$N_{ij} \sim Poisson(F_{ij} \times NID_{ij})$$

where the reproduction rate $F_{ij}$ at $c_i$ in replicate $j$ is a random variable following a gamma distribution. Introducing a dispersion parameter $\omega$, we assume that:

$$F_{ij} \sim gamma\left(\frac{f(c_i;\theta)}{\omega}, \frac{1}{\omega}\right)$$

Note that a gamma distribution of parameters $\alpha$ and $\beta$ has mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$, that is here $f(c_i;\theta)$ and $\omega f(c_i;\theta)$ respectively. Hence $\omega$ can be considered as an overdispersion parameter (the greater its value, the greater the inter-replicate variability)

### 2.2.2 Inference

Posterior distributions for parameters $b$, $d$ and $e$ are estimated using JAGS [10] with the following priors:

- we assume the range of tested concentrations in an experiment is chosen to contain the $EC_{50}$ with high probability. More formally, we choose:

  $$\log_{10} e \sim \mathcal{N}\left(\frac{\log_{10}(\min_i c_i) + \log_{10}(\max_i c_i)}{2}, \frac{\log_{10}(\max_i c_i) - \log_{10}(\min_i c_i)}{4}\right)$$

  which implies $e$ has a probability slightly higher than 0.95 to lie between the minimum and the maximum tested concentrations.

- we choose a quasi non-informative prior distribution for the shape parameter $b$:

  $$\log_{10} b \sim \mathcal{U}(-2,2)$$

- as $d$ corresponds to the reproduction rate without contaminant, we set a normal prior $\mathcal{N}(\mu_d, \sigma_d)$ using the control:

  $$\mu_d = \frac{1}{r_0}\sum_j \frac{N_{0j}}{NID_{0j}}$$

  $$\sigma_d = \sqrt{\frac{\sum_j \left(\frac{N_{0j}}{NID_{0j}} - \mu_d\right)^2}{r_0(r_0-1)}}$$

[1] that is, the number of reproduction outputs during the experiment per organism-day

where $r_0$ is the number of replicates in the control condition. Note that since they are used to estimate the prior distribution, the data from the control condition are not used in the fitting phase.

- we choose a quasi non-informative prior distribution for the $\omega$ parameter of the gamma-Poisson model:

$$log_{10}(\omega) \sim \mathcal{U}(-4, 4)$$

For a given dataset, the procedure implemented in 'morse' will fit both models (Poisson and gamma-Poisson) and use an information criterion known as Deviance Information Criterion (DIC) to choose the most appropriate. In situations where overdispersion (that is inter-replicate variability) is negligible, using the Poisson model will provide more reliable estimates. That is why a Poisson model is preferred unless the gamma-Poisson model has a sufficiently lower DIC (in practice we require a difference of 10).

# References


[1] Carlo Albert, Sören Vogel, and Roman Ashauer. Computationally efficient implementation of a novel algorithm for the general unified threshold model of survival (GUTS). *PLoS Computational Biology*, 12(6):e1004978, 2016.

[2] Marie Laure Delignette-Muller, Christelle Lopes, Philippe Veber, and Sandrine Charles. Statistical handling of reproduction data for exposure-response modeling. *Environmental Science & Technology*, 48(13):7544–7551, 2014.

[3] Marie Laure Delignette-Muller, Philippe Ruiz, and Philippe Veber. Robust fit of toxicokinetic–toxicodynamic models using prior knowledge contained in the design of survival toxicity tests. *Environmental Science & Technology*, 51(7):4038–4045, 2017.

[4] Tjalling Jager, Carlo Albert, Thomas G Preuss, and Roman Ashauer. General unified threshold model of survival - a toxicokinetic-toxicodynamic framework for ecotoxicology. *Environmental Science & Technology*, 45(7):2529–2540, 2011.

[5] Tjalling Jager and Roman Ashauer. *Modelling survival under chemical stress. A comprehensive guide to the GUTS framework. Version 1.0.* 2018.

[6] Anna-Maija Nyman, Kristin Schirmer, and Roman Ashauer. Toxicokinetic-toxicodynamic modelling of survival of *Gammarus pulex* in multiple pulse exposures to propiconazole: model assumptions, calibration data requirements and predictive power. *Ecotoxicology*, 21(7):1828–1840, 2012.

[7] OECD. Guidelines for testing of chemicals n.220. *Enchytraeid* reproduction test. Technical report, Organisation for Economic Cooperation and Development, 2004.

[8] OECD. Guidelines for testing of chemicals n.226. Predatory mite (*Hypoapsis (Geolaelaps) aculeifer*) reproduction test in soil. Technical report, Organisation for Economic Cooperation and Development, 2008.

[9] OECD. Guidelines for testing of chemicals n.211. *Daphnia magna* reproduction test. Technical report, Organisation for Economic Cooperation and Development, 2012.

[10] Martyn Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2016. R package version 4-6.