Ecotoxicology and Environmental Safety 145 (2017) 193–199

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv

Species sensitivity distribution for pentachlorophenol to aquatic organisms based on interval ecotoxicological data

Jinsong Zhao[a,*], Run Zhang[a,b]

[a] College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China
[b] Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Keywords:
Species sensitivity distribution
Interval ecotoxicological data
Bayesian statistics
Pentachlorophenol
Aquatic organisms

ABSTRACT

Species sensitivity distribution (SSD) model is often used to extrapolate the chemicals' effects from the ecotoxicological data on individual species to ecosystems, and is widely applied to derive water quality criteria or to assess ecological risk. Because of the influence of various factors, the ecotoxicological data of a specific chemicals to an individual usually exist in a range. The feasibility of interval ecotoxicological data directly applied to build SSD model has not been clearly stated. In the present study, by means of Bayesian statistics, the half maximal effective concentration (EC_{50}) of pentachlorophenol (PCP) to 161 aquatic organisms, which were organized into 7 groups, i.e., single determined value, geometric mean estimation, median estimation, interval data, and combination of single determined data with other groups, were used to develop SSD models and to estimate the minimum sample sizes. The results showed that the interval data could be directly applied to build SSD model, and when combined with single point data could give the narrowest credible interval that indicates a stable and robust SSD model. Meanwhile, the results also implied that at least 6–14 ecotoxicological data were required to build a stable SSD model. It suggests that the utilization of interval data in building SSD model can effectively enhance the availability of ecotoxicological data, reduce the uncertainty brought by sample size or point estimation, and provide a reliable way to widen the application of SSD model.

1. Introduction

Species sensitivity distribution (SSD) model is often used to extrapolate the ecotoxicological effects of a specific chemicals from individuals to ecological community or higher level. It is widely applied to derive ecological or environmental criteria, or to assess the ecological risk of chemicals or other agents. It assumes that the sensitivity or response of different species in an ecosystem with complex structure to certain stress, such as toxic chemicals, obeys a certain probability distribution (Forbes and Calow, 2002; Posthuma et al., 2002; Vighi et al., 2006), e.g., normal distribution (Aldenberg and Jaworska, 2000). Thus, a SSD model describes the sensitivity of different species to a stress by an empirical probability distribution.

The basis for developing a SSD model is sufficient, representative, accurate and reliable ecotoxicological data from a specific ecosystem (Dowse et al., 2013). However, due to realistic conditions, the half maximal effective concentration (EC_{50}) or no observed effect concentration (NOEC), the most common ecotoxicological endpoints used to build a SSD model (Dowse et al., 2013; Duboudin et al., 2004; Hickey et al., 2012; Wheeler et al., 2002), are mainly obtained from certain database, such as ECOTOX from U.S. Environmental Protection Agency, which collects lots of experimental ecotoxicological data of various chemicals from published literatures or reports (U.S. U.S. Environmental Protection Agency, 2017). Although most of them were acquired according to standard protocols, due to many factors affecting biological response, the ecotoxicological data for a certain chemicals to a specific species are quite different in such database. In ECOTOX, for example, the EC_{50} for pentachlorophenol to *Daphnia magna* is in the range of 0.143 μmol L^{-1} to 14.680 μmol L^{-1}, with a difference of more than 100 folds.

The general statistical method used to derive SSD model, such as generalized linear model, cannot deal with the above mentioned interval data (Lind, 2010). Thus, for such data, two options can be chosen. Firstly, the species having interval data are discarded, which will definitely increase the cost to obtain the ecotoxicological data, and reduce the biodiversity involved in SSD model, and thereby affect the application scope of SSD model. Secondly, a point estimation derived from the interval are used, such as (geometric) mean and median (European Commission, 1996; Raimondo et al., 2008; RIVM, 2001). However, the representativeness and effectiveness of a point estimation

* Corresponding author.
E-mail address: jszhao@mail.hzau.edu.cn (J. Zhao).

http://dx.doi.org/10.1016/j.ecoenv.2017.07.029
Received 4 July 2016; Received in revised form 6 July 2017; Accepted 11 July 2017
Available online 19 July 2017

is less than the interval data (Lind, 2010). Therefore, to establish a SSD model based on the interval data has a clear practical significance.

Since the information about empirical judgement on parameters and other interesting aspects can be incorporated into the data analysis process, Bayesian statistics, has been widely applied to SSD model development (Aldenberg and Rorije, 2013; Dowse et al., 2013; Grist et al., 2006; Hayashi and Kashiwagi, 2010a, b; Hickey et al., 2012). However, it is rarely used to build a SSD model based on interval ecotoxicological endpoints. Hayashi and Kashiwagi (2010a) had used Bayesian statistics to establish SSD model based on different taxonomy in which interval data were included. However, they did not compare the effectiveness of interval data with that of point estimation. At present, there is no any detailed comparative study on SSD models based on interval data retrieved from a database. In addition, there is no report on the minimum sample size for building a SSD model with interval data, though a lot of researches had been performed in the aspect of the minimum sample size requirement.

Although it is already banned as pesticides and disinfectants, pentachlorophenol (2,3,4,5,6-Pentachlorophenol, PCP) is still widely used in wood preservation, *Oncomelania* snails killer, and many others (Crosby, 1981). Because of significant toxicity and low biodegradability, the fate and the ecological effects of PCP have been caused widespread concerns (Yadid et al., 2013; Zheng et al., 2012). Currently, only a handful of countries have developed ecological criteria or water quality standards for PCP.

In this study, SSD models based on interval data published in ECOTOX were developed by means of Bayesian statistics, and compared with those based on single determined data or point estimations. The purposes were to confirm the feasibility of developing a SSD model and to determine the minimum sample size required for a stable SSD model based on interval data.

2. Materials and methods

2.1. Data sources

The ecotoxicological data, i.e., EC_{50} of PCP to aquatic organisms in this study, were retrieved from ECOTOX (U.S. Environmental Protection Agency, 2017). All EC_{50} were obtained in laboratory with exposure type of flow-through, renewal, or static, and concentration type of active ingredient. Generally, each sample for a specific species in ECOTOX database has 3, i.e., minimum, mean and maximum EC_{50}. If the mean EC_{50} had " < " or " > " operator, it would be treated as missing value. When there were multiple samples for one species, the interval (i.e., the upper and lower limit) and point estimation (i.e., geometric mean and median) of EC_{50} were calculated based on the gather of all the three EC_{50} if available. Finally, the EC_{50} of PCP to 161 aquatic organisms were obtained (cf. Table S1).

In order to verify the objects of this study, the whole dataset was divided into 7 subsets: D1, with only one determined EC_{50} for one species in ECOTOX; D2, with the geometric mean for multiple EC_{50} for one species; D3, same as D2 but with median; D4, same as D2 but with interval data; D5, the combination of D1 and D2; D6, the combination of D1 and D3; D7, the combination of D1 and D4. The sample size of D1 was 65, D2 to D4 was 96 and D5 to D7 was 161.

2.2. SSD model development based on Bayesian statistics

In this study, the EC_{50} of PCP to aquatic organisms were assumed to follow normal distribution (Aldenberg and Jaworska, 2000). The cumulative probability function of normal distribution is:

$$F(X;\mu,\sigma) = \Phi\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{X-\mu}{\sigma\sqrt{2}}\right)\right] \quad (1)$$

where, X is logarithmic EC_{50} of PCP to aquatic organisms; μ and σ are 2 parameters of normal distribution; erf is the error function.

Because of the interval data, the parameters of normal distribution would be estimated by means of Bayesian statistics.

2.2.1. Prior distribution of parameters

Since no effective prior information about each parameter had been reported, non-information prior distribution was used as follows (Gelman et al., 2004):

$$\mu \sim \mathrm{dnorm}(0,\ 0.001) \quad (2)$$

$$\sigma \sim \mathrm{dgamma}(0.001,\ 0.001) \quad (3)$$

where, dnorm and dgamma are probability density functions for normal and gamma distribution.

2.2.2. Likelihood function

For dataset D1, D2, D3, D5 and D6, the likelihood function was defined as:

$$X_i \sim \mathrm{dnorm}(\mu, \tau) \quad (4)$$

For dataset D4, the likelihood function was defined as:

$$Z_i \sim \mathrm{dinterval}(T_i, I_i) \quad (5)$$

$$T_i \sim \mathrm{dnorm}(\mu, \tau) \quad (6)$$

For dataset D7, its likelihood function was the combination of formulae (4)–(6). That meant, the formulae (4) was used to deal with the mean part of D7, and formulae (5) and (6) were used to process the interval part of D7.

In the above likelihood functions, Z_i was an indicator variable that took the value 1 if X_i was an interval data and 0 otherwise; $\tau = 1/\sigma^2$ was the precision of normal distribution; I_i was the interval data of the i^{th} species. dinterval implemented the discretization and set up the observed values (T_i) as stochastic nodes as required for the Gibbs sampler to work.

Because its theoretical expression was unknown, the posterior distribution for each parameter is simulated by Markov Chain Monte Carlo (MCMC) with Gibbs sampler (Kruschke, 2011). In the simulation, the number of Markov chain was set to 3; the iterations for each Markov chain was set to 60,000 times, with initial 10,000 discarded to obtain stable estimates. The convergence for each parameter was identified when Gelman-Rubin convergence index fell in the range of 1.0–1.1 (Gelman et al., 2004).

2.2.3. Credible interval

The credible interval, which represented the degree of centralization of each parameter or sample, was represented by the 95% interval of highest posterior density (HPD). The width of the HPD credible interval was an alternative way to measure the uncertainty of the parameters of SSD model (Kruschke, 2011). The credible interval of μ and σ was the estimation of HPD interval based on 150,000 pairs of μ and σ in the MCMC sample. The credible interval of HC_5 (hazardous concentration at which 5% species in an ecosystem may be affected) was the estimation of HPD interval based on 5% quantile values calculated from 150,000 pairs of μ and σ in the MCMC sample.

2.3. Determination of minimum sample size

To determine the minimum sample size that was required to build a stable SSD model, the parameters and HC_5 of SSD models based on a series of simulated dataset with different size were estimated. The simulated dataset with a specific sample size were generated from each of the 7 ecotoxicological datasets using the procedure of basic bootstrap (Davison and Hinkley, 1997). For each sample size, totally 5000 simulated samples were generated. SSD models were built based on those simulated samples, and then mean and credible interval of parameters and HC_5 were estimated. A change point analysis was used to determine