Application Statistical Models for Interpretation Toxicological Data

T R Latypova¹, N Yu Stepanova¹

¹Institute of Environmental Sciences, Kazan Federal University, Kremlevskay str., 18, 420008, Kazan, Russia

E–mail: latamri@yandex.ru

Abstract. The results of bioassay on infusoria of 75 samples of bottom sediments from 6 water bodies of the Middle Volga region were analyzed using traditional nonparametric methods and statistical models Generalized linear mixed model (GLMM) and Cumulative link mixed model (CLMM). The ambiguity of the interpretation of the results of biotesting performed by nonparametric methods is due to the fact that the toxicological data often do not correspond to the normal distribution. The use of the GLMM and CLMM models allow analyze data that do not correspond to the normal distribution and made it possible to clarify the level of toxicity of a number of ambiguous samples, which, after processing by mathematical model algorithms, acquire the status of either exactly toxic or exactly non-toxic.

1. Introduction

Many authors noted difficulties in finding statistically significant relationships between the observed toxicity on test organisms and specific pollutants identified during chemical analysis [1-11].

The authors of these studies concluded that it is necessary to take into account the bioavailability of toxic compounds, for example, metals, however, even when measuring and taking into account the bioavailability of pollutants, the problem of finding a relationship between toxicity and the chemical composition of natural objects remains [3].

Therefore, a number of researchers, when interpreting the results of ecotoxicological studies, have moved from expanding the accounting of toxicity factors to the use of additional approaches, including mathematical tools [12-15].

The use of a limited set of standard tools of mathematical statistics is insufficient for identifying the factors of toxicity of natural samples with a complex matrix, for example, bottom sediments, especially taking into account the spatial factors of changes in the geochemical composition.[16-17]

The most of the statistical methods are based on the fact of normal distribution of the data obtained during the study. But in environmental studies, the data obtained often do not correspond to the law of normal distribution. In this regard, the search for statistical models that allow to analyze such ecotoxicological data is an urgent task.[18-19]

The object of the study was to substantiate the use of generalized linear mixed models to assess the results of toxicological experiments on samples of bottom sediments from rivers in the Middle Volga region.
2. Materials and methods
In the work, we used the primary data on the assessment of the toxicity of the extracts of bottom sediments, performed using the ciliates *Paramecium caudatum* (the criteria of toxicity is the intensity of division per one day) [20].

The toxicological data base was 75 values from six water bodies.

To take into account the variability of the response of test organisms and eliminate random effects that arise in conditions with many control experiments, we used a model of the type GLMM (Generalized linear mixed model). To clarify the picture of the reliability of toxicological experiments, a logistic model of the ordinal type was used - a cumulative model of the ratio of probabilities CLMM (Cumulative link mixed model), where for each probable finite number of test organisms a cumulative probability distribution function is constructed.

3. Results and its discussion
To analyze the data of toxicological experiments on the rate of reproduction of ciliates, a GLMM model with a Poisson distribution was used. This distribution simulates a random variable representing the number of events (the fact of division of the ciliates) that occurred in a fixed time (per one day), provided that these events occur with some fixed average intensity and independently of each other.

In order to eliminate the aforementioned heterogeneity within one experiment and between their series, we used normally distributed random effects $RE_j$ with variance $\sigma_{RE}^2$, which equally affect both the control and the sample in one experiment. This model makes it possible to assess the influence of the sampling site and has the form:

\[
N_{i,j,k} \sim \text{Poisson}(\mu_{i,j,k})
\]

\[
\ln(\mu_{i,j,k}) = a_0 + \text{RiverSite}_i + RE_j
\]

\[
RE_j \sim \text{Norm}(0, \sigma_{RE}^2)
\]

(1)

where

- $i$ - the number of the pair (River, Site),
- $j$ - the number of the processing package,
- $k$ - the number of the experiment for the given sample and package.

For control, a special designation "No site" has been introduced. The control was used as a comparison base (the values of all other parameters are compared with it).

To account for the heterogeneity of the experimental conditions between the packages, we used normally distributed random effects $RE_j$ with variance $\sigma_{RE}^2$, which equally affect both the control and the experiment in the same package.

This model is designed to compare each of the sites with the control, in it $\text{RiverSite}_i$ - the change in the average value of the number of cells at the end of the experiment for the $i$-th site in comparison with the control, on a logarithmic scale.

The statistical generalized linear model for comparison the toxicological response of ciliates in the experiment with the control of each site of the water body allowed us to determine the relative amount of toxic samples. So, for the Sumka River, no significant differences were found between the data and the control, the largest number of reliably toxic samples according to the results of this model was observed in the Sheshma River in the amount of 40% (table 1).
Table 1. Statistical generalized linear model of comparison of toxicological response of ciliates with control in samples to each site of a water body.

| Water body               | Simulated value | Standard deviation of the average | Percentage of reliable results, % |
|--------------------------|-----------------|-----------------------------------|----------------------------------|
| r. Kichuy                | -0.580          | 0.183                             | 0.112                            | 0.227                            | 25 |
| Kuibyshev water reservoir| -17.361         | 0.163                             | 0.091                            | 1.753                            | 26 |
| r. Kazanka               | -0.900          | 0.086                             | 0.122                            | 0.328                            | 30 |
| r. Sumka                 | -0.351          | 0.076                             | 0.128                            | 0.262                            | 0  |
| r. Sheshma               | -1.013          | 0.404                             | 0.111                            | 0.365                            | 40 |
| r. Yushut                | -1.123          | 0.047                             | 0.105                            | 0.364                            | 13 |

Analysis of the data comparing the toxicological response of ciliates in bottom sediments samples with control obtained in GLMM showed that there was an underestimation of the variance of the model residues. This is due to the actual distribution of toxicological data differs from Poisson's and has less variance. This leads to an overestimation of the variance when calculating the significance level, and a decrease in the significance of the difference with the control.

Thus, the impact assessment results obtained using GLMM models were conservative (underestimated), and if a difference from the control was established, then there was a toxic effect, and if not, then it was probably absent.

An alternative to GLMM would be to use the ordinal Cumulative Link Mixed Model (CLMM), in which toxicological data are simply considered a “scoring” of environmental quality. CLMM is a cumulative probability ratio model, in which the probability of an increase in \( Y \) (the final number of cells) by one division of the "point" scale is estimated on a logistic scale.

This model has the form

\[
\text{logit}(P(Y_{i,k,m} \leq j)) = \theta_j - \text{RiverSite}_i + RE_k
\]

\[RE_k \sim \text{Norm}(0, \sigma_{RE}^2)\]

(2)

where

\( j \) - the number of ciliates,
\( i \) - the sample number,
\( k \) - the processing package number,
\( m \) - the experiment number for a given river and package.

For control, a special variable "No site" was introduced, which is used as a comparison base (the values of all other parameters are compared with it).

To account for the heterogeneity of the experimental conditions between the packages, we used normally distributed random effects \( RE_k \) with variance \( \sigma_{RE}^2 \), which equally affect both the control and the experiment in the same package.

Since the model is designed to compare the experimental data of samples from each river with the control, it shows the change in the average value of the number of cells in the experiment in comparison with the control for the \( i \)-th pair on the logistic scale.

In this model, the reliability of the experiment is determined by comparing the probabilities of obtaining in the control a certain final number of ciliates at the end of the experiment and the same value in the test sample at a significance level of 0.05.
Table 2. Results of a statistical cumulative probability ratio for comparing toxicological response of ciliates with control in samples to each site of a water body.

| Water body       | Simulated value | Standard deviation of the average | Percentage of reliable results, % |
|------------------|-----------------|-----------------------------------|----------------------------------|
|                  | min.            | max.                | min. | max. | without methods | with methods of | Bayevsky analysis | Bayevsky analysis |
| r. Kichuy        | -1,189          | 0,549               | 0,253| 0,312| 58              | 93               |                   |                   |
| Kuibyshev water  | -9,156          | 0,443               | 0,197| 2,024| 48              | 92               |                   |                   |
| reservoir        | -2,197          | 0,227               | 0,255| 0,464| 40              | 91               |                   |                   |
| r. Kazanka       | -0,926          | 0,266               | 0,252| 0,455| 29              | 50               |                   |                   |
| r. Sumka         | -9,206          | 2,492               | 0,313| 27,460| 67           | 80               |                   |                   |
| r. Sheshma       | -0,639          | 1,378               | 0,262| 0,464| 50              | 100              |                   |                   |
| r. Yushut        |                 |                     |      |      |                 |                  |                   |                   |

Data, presented in the table 2 shows that the use of the CLMM model can significantly increase the accuracy of identifying the toxicity of natural samples.

The smallest percentage of reliably toxic samples was observed in the Sumka River in both calculation variants. In general, the use of the version of the statistical model with using the methods of Bayevsky analysis significantly increased the reliability of detecting the toxic effect of natural samples with a complex matrix, such as bottom sediments.

4. Conclusion
The results of bottom sediment biotesting were analyzed using traditional nonparametric methods and the statistical models. It was shown that errors in the interpretation of toxicity data performed by the traditional method can be associated with data variations in replications of the same experiment, which leads to the identification of the sample as non-toxic. An unreliable experimental result in samples with high toxicity when it is impossible to confirm or refute the presence of toxicity in repeated series demonstrates the insufficiency of using traditional statistical methods. The use of the GLMM and CLMM models allowed to clarify the level of toxicity: after processing by the algorithms of the mathematical model, ambiguous results acquire the status of either exactly toxic or exactly non-toxic.

5. References
[1] Ricking M, Neumann-Hensel H, Schwarz Bauer J and Svenson A 2004 USA. Environ. Chem. Lett. 2 109
[2] Zabetoglou K, Voutsa D and Samara C 2002 Greece. Chemosphere 49 17 http://www.sciencedirect.com/science/article/pii/S0045653502001947
[3] Calace N, Ciardullo S, Petronio B M, Pietrantonio M, Abbondanzi F, Campisi T and Cardellicchio N 2005 Italy. Microchem. J. 79 243 http://www.sciencedirect.com/science/article/pii/S0026265X04001936
[4] Hyotylainen T and Oikari A 1999 Finland. Sci. Total Environ. 243–244 97 http://www.sciencedirect.com/science/article/pii/S0048969799003642
[5] Salizzato M, Pavoni B, Ghirardini A V, Ghetti P F 1998 Italy. Chemosphere 36 2949-2968 http://www.sciencedirect.com/science/article/pii/S0045653598000010
[6] Pedersen F, Bjornestad E, Andersen H V, Kjoholt J and Poll C 1998 Denmark. Water Sci. and Tech. 37 233 http://www.sciencedirect.com/science/article/pii/S0273122398002030
[7] McCarthy L H, Thomas R L, Mayfield C 2004 USA. Lakes Reservoirs: Res. Manage. 9 89
[8] Klamer H J C, Leonards P E G, Lamoree M H, Villerius L A, Åkerman J E, Bakker J F 2005 The Netherlands. Chemosphere 58 1579
[9] Christensen A M, Nakajima F and Baun A 2006 Denmark. Environmental Pollution 144 621-625 http://www.sciencedirect.com/science/article/pii/S0269749106000728
[10] Tomilina I I, Mihajlova L V, Rybina G E and Akaťeva T G 2009 Russia. Toksikologicheskiy vestnik 2 71
[11] Bae M-J and Park Y-S 2014 Republic of Korea. Science of The Total Environment 466–467 635-649
[12] Kudlak B, Tsakovski S, Simeonov V, Sagajdakow A, Wolska L and Namieśnik J 2014 Bulgaria. Chemosphere 95 17-23 http://www.sciencedirect.com/science/article/pii/S0045653513007844
[13] Triebskorn R et al. 2003 USA. Hum. Ecol. Risk Assess. Int. J. 9 171-194
[14] Burkhardt-Holm P and Scheurer K 2007 Switzerland. Aquat. Sci. 69 51-70
[15] Maier D et al. 2015 Germany. Water Research 72 127-144 http://www.sciencedirect.com/science/article/pii/S0043135414006265
[16] Christensen R H B and Brockhoff P B 2013 Denmark. Journal de la Societe Francaise de Statistique & Revue de Statistique Appliquee 154 58-79 https://backend.orbit.dtu.dk/ws/portalfiles/portal/79012845/Christensen+and+Brockhoff+2013.pdf
[17] Hedeker D, Demirtas H and Mermelstein R J 2009 USA. Stat Interface 2 391–401 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847414/
[18] Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens and Jada-Simone S White 2009 USA. Trends in Ecology & Evolution 24 127-135 https://www.sciencedirect.com/science/article/abs/pii/S0169534709000196
[19] Xian Liu 2016 Methods and Applications of Longitudinal Data Analysis (San Diego: Academic Press) https://www.sciencedirect.com/book/9780128013427/methods-and-applications-of-longitudinal-data-analysis#book-description
[20] Stepanova N Yu, Latypova T R and Novikova L V 2016 Russia. Uchenye Zapiski Kazan. Universiteta. 158 416-439