Adequate statistical modelling and data selection are essential when analysing abundance and diversity trends

Marion Desquilbet, Pierre-André Cornillon, Laurence Gaume, Jean-Marc Bonmatin

Title page

Journal: Nature Ecology & Evolution, 2021 accepted version

Title. Adequate statistical modelling and data selection are essential when analysing abundance and diversity trends

Matters arising. Arising from Crossley *et al.*, No net insect abundance and diversity declines across US Long Term Ecological Research sites, *Nature Ecology & Evolution* (2020) doi:10.1038/s41559-020-1269-4.

Authors. Marion Desquilbet[1]*†, Pierre-André Cornillon[2]†, Laurence Gaume[3], Jean-Marc Bonmatin[4]

Affiliations

[1] Toulouse School of Economics, INRAE, University of Toulouse Capitole, Toulouse, France.

[2] Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France.

[3] AMAP, University of Montpellier, CNRS, CIRAD, INRAE, IRD, Montpellier, France.

[4] Centre de Biophysique Moléculaire, CNRS, 45071 Orléans, France.

*Correspondence to: marion.desquilbet@inrae.fr.

†These authors contributed equally to this work.

Acknowledgments

Doyle McKey and Emmanuelle Jousselin (University of Montpellier) made suggestions to improve the paper. This work was publicly funded by the French National Research Agency (grants ANR-17-EURE-0010 to M. D. and ANR-16-IDEX-0006 to L. G. under the "Investissements d'avenir" program and grant ANR-15-CE21-0006 to M.D.).

Author contributions

M.D. and P-A.C. performed both the detailed and overall analysis of the article and wrote the original draft. P-A.C. examined and re-programmed the R code. L.G. contributed to the argumentation and extensively edited the manuscript. J-M.B. contributed to the analysis of data selection in the article. All authors contributed to the general comment and reviewed the manuscript.

Title. Adequate statistical modelling and data selection are essential when analysing abundance and diversity trends

Matters arising. Arising from Crossley *et al.*, No net insect abundance and diversity declines across US Long Term Ecological Research sites, *Nature Ecology & Evolution* (2020) doi:10.1038/s41559-020-1269-4.

Abstract

In an analysis of a large number of time series on arthropod abundances in natural and managed areas of the United States, Crossley et al. reported no evidence of an overall decline in insect abundance and diversity[1]. We identified major concerns in the statistical analysis and inconsistencies in the selection of data, which, we argue, invalidate their conclusions. We call for a rigorous methodology in analyses of biodiversity trends because relevant information is crucial for stakeholders and policy makers.

Matters arising

The extent of the decline of insect populations worldwide is much debated[2-5], with major implications for public policy investment in biodiversity protection. Crossley et al. conducted a statistical analysis of 5,375 geographically and taxonomically varied time series on arthropod abundance during 4 to 36 years across the United States[1]. They concluded that there was no significant change in insect populations. However, we argue that issues in the statistical analysis and inconsistencies in data selection invalidate their conclusions.

The modelling proposed by Crossley et al. relied on the following steps: i) collecting data, ii) separating each species of each locale of each LTER (in the R script, a locale could be an arthropod group, a location or a collection method), iii) pre-processing data (Box 1), iv) running a different autoregressive linear model for each species of each locale of each LTER (hereafter, LLS),

v) combining all estimated slopes into a “sample”, vi) analysing this “sample” using violin plots, *T* tests and confidence intervals (Fig. 1). The statistical analysis carried out in this last step relied on the assumption that the observations in the sample were independent and identically distributed (iid). This assumption was violated for two reasons. First, the pre-processing step included a time scaling to change the minimum year of each LLS time series to 0 and its maximum year to 1. As the time length varied from 4 to 36 years depending on LLS, the scaled time *x* varied across LLS time series. Therefore, the estimated slopes did not represent abundance trends per year, but per time units *x* varying over time series and without a clear meaning. Second, according to the linear regression theory, the expectation and variance of the estimated slopes depend on the number of measures of the *x* variable (i.e. the length of the time series) and the distances of *y* measures to the model (i.e. the quality of the model approximation). Among LLS time series, there are different time lengths, and different qualities of approximation. Therefore, the slopes cannot be iid, and estimations and tests used in step vi) are not reliable. To circumvent this problem, it would be more appropriate to use a hierarchical model to analyse the whole dataset.

Insert Box 1 and Figure 1.

Other problems are as follows. First, most individual time series were too short to provide reliable estimations of the four unknown parameters specific to each LLS (Box 1). Indeed, 44% of LLS time series only had 4 to 9 years of data. While no simple threshold exists, we do not see how to reliably estimate four parameters with less than ten data points, which will only provide a very imprecise estimation. Some limited sensitivity analysis was provided with a stricter data subset involving a minimum of 15 years of data, but this strict dataset only included less than 6% of the time series. It represented a much more limited variety of situations than the total sample and was therefore much less representative. This is another argument in favour of a global modelling approach, which would improve the precision of the trend estimate of any given LLS by using data from other LLS.

Second, the analysis was performed at a very fine taxonomic level, implying that a high proportion of abundance counts was equal to zero (the full dataset contained 49% of zeros and the strict dataset, 30.5%; moreover, 71% of the series in the full dataset, and 84% in the strict dataset, contained at least a zero). As the logarithm of 0 is undefined, all zero abundance values were shifted upwards before being log-transformed by adding an arbitrary value. Such rudimentary log-transformation of count data is to be avoided because results depend on the chosen value and coefficient estimates are inaccurate[6,7]. Zero-inflated models would have dealt appropriately with the problem of high occurrence of zeros in the dataset[8].

Third, the model corrected for scale differences between abundance series without accounting for imperfect detection, which can be of particular concern for rare species. This problem may be illustrated by the case of *Aphis asclepiadis*, NEPAC locale, Midwest STN LTER (external database S1[9]). In the ten years of records, its abundance was 0 for the nine first years, and 1 for the last year. This time series (like the others in the dataset) was not composed of abundance levels, but estimations of abundance. Due to imperfect detection, a shift from an estimated abundance from 0 to 1 provides poor information on the real abundance trend. After scaling log-abundances (Box 1), this uninformative *A. asclepiadis* data series was erroneously modelled as having the highest abundance increase of all the time series (external database S2[9]), while it could just reflect the rarity of the species or its poor detection. The same slope could have been obtained with a time series reflecting a significant abundance change, with for example a hundred insects in all years except the last year with a thousand insects. In total, 16% of time series included only abundance values of 0 and 1, and 27% of time series included only abundance values of 0, 1 and 2. Simple models of occurrence and abundance have already been developed to cope with the problem of imperfect detection.[10]

As all analyses of diversity (richness, evenness and β diversity) in the article relied on these estimations of abundance and on the same modelling, they shared similar methodological problems.

We also point out that we had to re-program the R script provided by the authors using their external databases S1 and S2[9] to make it run.

Regarding data selection, the article is intended to analyse insect abundance trends in US Long Term Ecosystem Research (LTER) sites, but it departs from this description in two ways. First, 39.5% of time series are from the Suction Trap Network (STN). One suction trap of the STN is located in the Kellogg LTER site, and all STN data, encompassing ten US states, were incorporated into the Kellogg LTER dataset up to 2014[11]. But the dataset used in the analysis (https://suctiontrapnetwork.org), spanning up to 2019, is not linked to a LTER. Its inclusion may bias results by minimising the damages of intensive farming, because the STN exclusively provides data on aphids, and primarily aims to document pest aphids[11], which benefit from intensive agriculture[12,13], unlike most insects (e.g. aphid natural enemies[13] or bees[14]).

Second, the reference to insects in the title of the article is confusing as almost 10% of time series were of non-insect arthropods or included insects and other arthropods. In Fig. 2, three of the 22 violin plots concerned or involved crustaceans. Unlike the rest of the dataset, the violin plot from the Coweeta LTER related to aquatic invertebrate communities in terms of functional feeding groups, and not individual species. These inconsistencies add to other criticisms of this article[15] regarding unaccounted-for changes in sampling location and sampling effort at LTER sites and the unaccounted-for impact of experimental conditions on insect populations.

As a conclusion, the methodology chosen in this article is very approximate with several identified problems likely to substantially bias the results. The analysis would have required an adequate global model for all data, considering all our criticisms and those of ref. [15]. We call for the application of rigorous standards for analyses on global change, especially because results could have a significant impact on policy decision-making and the fate of biodiversity.

References

1 Crossley, M. S. *et al.* No net insect abundance and diversity declines across US Long Term Ecological Research sites. *Nat Ecol Evol* **4**, 1368-1376, doi:10.1038/s41559-020-1269-4 (2020).

2 Cardoso, P. *et al.* Scientists' warning to humanity on insect extinctions. *Biological Conservation* **242**, doi:10.1016/j.biocon.2020.108426 (2020).

3 van Klink, R. *et al.* Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. *Science* **368**, 417-420, doi:10.1126/science.aax9931 (2020).

4 Desquilbet, M. *et al.* Comment on “Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances”. *Science* **370**, eabd8947, doi:10.1126/science.abd8947 (2020).

5 Jähnig, S. C. *et al.* Revisiting global trends in freshwater insect biodiversity. *WIREs Water*, doi:10.1002/wat2.1506 (2020).

6 O’Hara, R. B. & Kotze, D. J. Do not log-transform count data. *Methods in Ecology and Evolution* **1**, 118-122, doi:10.1111/j.2041-210X.2010.00021.x (2010).

7 St-Pierre, A. P., Shikon, V. & Schneider, D. C. Count data in biology-Data transformation or model reformation? *Ecol Evol* **8**, 3077-3085, doi:10.1002/ece3.3807 (2018).

8 Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1-14, doi:10.2307/1269547 (1992).

9 Crossley, M. et al. No net insect abundance and diversity declines across US Long Term Ecological Research sites, Dryad, Dataset, https://doi.org/10.5061/dryad.cc2fqz645 (2020).

10 Royle, J. A., Nichols, J. D. & Kéry, M. Modelling occurrence and abundance of species when detection is imperfect. *Oikos* **110**, 353-359, doi:10.1111/j.0030-1299.2005.13534.x (2005).

11 Lagos-Kutz, D. and D. Voegtlin. Midwest Suction Trap Network. *Iowa State Research Farm Progress Reports*, 2203, http://lib.dr.iastate.edu/farms_reports/2203 (2015).

12 Simon, J. C. & Peccoud, J. Rapid evolution of aphid pests in agricultural environments. *Curr. Opin. Insect Sci.* **26**, 17-24, doi:10.1016/j.cois.2017.12.009 (2018).

13 Zhao, Z. H., Hui, C., He, D. H. & Li, B. L. Effects of agricultural intensification on ability of natural enemies to control aphids. *Scientific Reports* **5**, 7, doi:10.1038/srep08024 (2015).

14 Woodcock, B. A. *et al.* Impacts of neonicotinoid use on long-term population changes in wild bees in England. *Nat. Commun.* **7**, 8, doi:10.1038/ncomms12459 (2016).

15 Welti, E. A. R. *et al.* Meta-analyses of insect temporal trends must account for the complex sampling histories inherent to many long-term monitoring efforts, *EcoEvoRxiv Preprints,* doi:10.32942/osf.io/v3sr2 (2020).

Box 1. Model used by Crossley et al. (2020)

Each time series i was composed of abundance levels A_{it} for LLS i and for years t_{i1} to t_{iT_i}.

The first step of data pre-processing consisted in log-transforming abundances. For LLS i in year t, the abundance value A_{it} was replaced either by its logarithm, log A_{it}, or, if $A_{it} = 0$, by the logarithm of a constant, log c_i, where c_i was half the minimum non-zero abundance in time series i, to obtain a series of log-transformed abundances a_{it}.

In a second step, log-abundances were scaled: the empirical mean $\bar{a}_i$ of log-transformed abundances of the series was subtracted to each a_{it} and this difference was divided by the empirical standard deviation s_i of log-transformed abundances. This yielded the scaled logarithm of abundance of LLS i in year t, y_{it}, defined as $y_{it} = (a_{it} - \bar{a}_i)/s_i$.

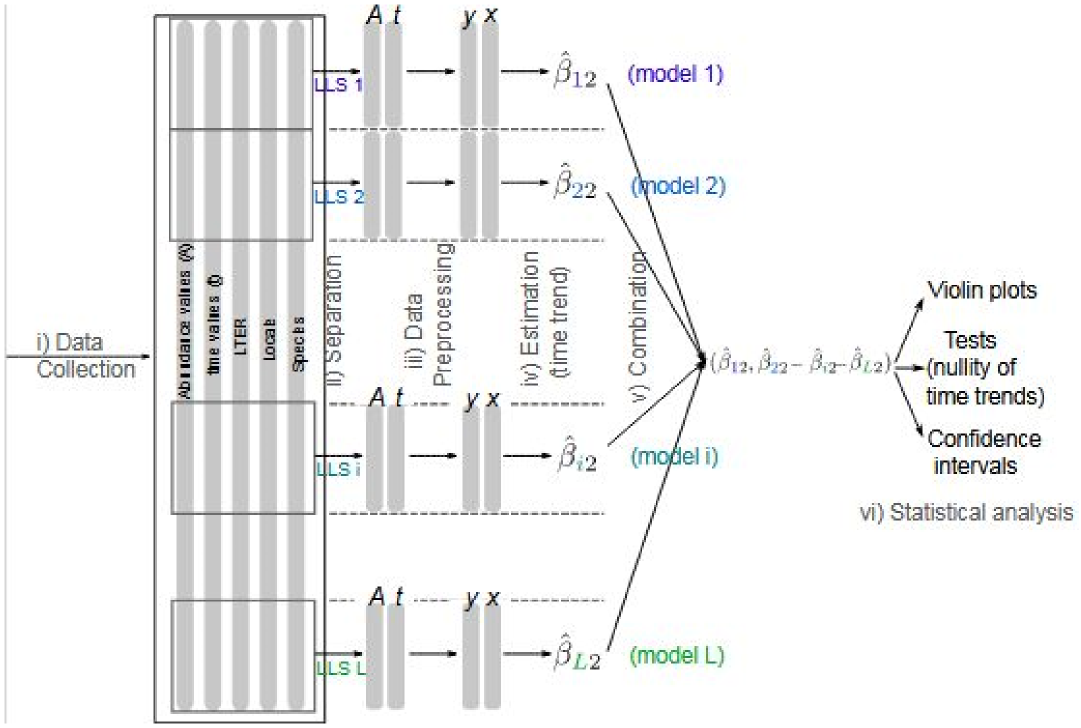
In a third step, the authors transformed all time units using a common scale varying between 0 (the first year of the LLS abundance time series) and 1 (its last year). The scaled year x_{it} was obtained by transforming the first year of time series t_{i1} to 0, and its last year t_{iT_i} to 1, and scaling all years accordingly as $x_{it} = (t_{it} - t_{i1})/(t_{iT_i} - t_{i1})$.

The proposed modelling was a linear model with a Gaussian auto-regressive error of order 1:

$y_{it} = \beta_{i1} + \beta_{i2} x_{it} + \varepsilon_{it}$, where $\varepsilon_{it} = \rho_i\, \varepsilon_{i,t-1} + \eta_{it}$, with $\eta_{it} \sim \mathrm{N}(0,\sigma_i^2)$.

For each individual LLS time series, this model implied the estimation of four parameters, β_{i1}, β_{i2}, ρ_i and σ_i, the slope β_{i2} representing the abundance trend and therefore being the parameter of interest.

Figure 1. Modelling steps in Crossley et al. (2020) and arising problems. Time trends were estimated separately for each species of each locale of each LTER (LLS). The time scaling was performed on LLS of different time lengths and the quality of approximation varied across LLS. Therefore, the abundance time trends were not independent and identically distributed as assumed when calculating the average abundance trends, confidence intervals and significance tests associated with the violin plots of Fig. 2 in Crossley et al. (2020). A global hierarchical modelling would have circumvented this problem.



Competing interests.

The authors declare no competing interests.