Metadata Supplementary Data 1.

Rds file (for R) with the complete data frame from which all analyses can be reproduced. Code for analysis is available on github: <https://github.com/roelvanklink/final-insect-diversity-trends2>

Column explanation:

|  |  |
| --- | --- |
| Column name | Explanation |
| Plot\_ID | Unique identifier for each site. Always nested in a Datasource\_ID |
| Datasource\_ID | Unique identifier for each dataset. ID’s from BioTIME and GPDD were left as is. |
| Datasource\_name | Name of Datasource\_ID, for ease of recognition. Redundant with Datasource\_ID. |
| Year | Year of sampling event |
| Period | Time of the year of sampling event (month / season) |
| Metric | Metric of assemblage size. Factor:   * + *Abundance*: total number of individuals observed per sampling event   + *Richness*: total number of taxa observed per sampling event   + *rarefiedRichness*: number of species expected to be observed per sampling event had the minimum observed number of individuals in any year been found   + *CoverageRichness.8:* Coverage based richness: Expected number of species if 80% of the community had been sampled   + *Shannon*: raw Shannon -Wiener entropy (not yet converted to its Effective Number of Species)   + *ENSPIE*: inverse Simpson index   + *logNr020*: number of species in SAD interval 0 – 20%   + *logNr2040*: number of species in SAD interval 20 – 40%   + *logNr4060*: number of species in SAD interval 40 – 60%   + *logNr6080*: number of species in SAD interval 60 – 80%   + *logNr80100*: number of species in SAD interval 80 – 100%   + *logNrQ1*: Number of species in first SAD quartile (See Fig. S1)   + *logNrQ2*: Number of species in second SAD quartile (See Fig. S1)   + *logNrQ3*: Number of species in third SAD quartile (See Fig. S1)   + *logNrQ4*: Number of species in fourth SAD quartile (See Fig. S1) |
| Number | Count of insects by any metric observed at this sampling event |
| Realm | Realm of sampling: here, always Terrestrial |
| Continent | Continent at which sampling took place |
| Region | Grouping of Countries or states into geographical units providing sufficient data for analysis |
| Country | Nation state in which sampling took place |
| Country\_State | Country or state in which sampling took place (finer scale than Country) |
| Location | Unique identifier of study area or reserve of the sampling event. Not always nested in Dataset. |
| cYear | variable Year centered around mean for Inla analysis |
| iYear | Index year for random intercept |
| rYear | Copy of iYear for random slope |
| rYear2 | Copy of iYear for random slope |
| Period\_4INLA | Unique number for each Period for random intercept in Inla analysis |
| Plot\_ID\_4INLA | Unique number for each Plot\_ID for random intercept in Inla analysis |
| Datasource\_ID\_4INLA | Unique number for each Datasource\_ID for random intercept in Inla analysis |
| Location\_4INLA | Unique number for each Location for random intercept in Inla analysis |
| Plot\_ID\_4INLAs | Unique number for each Plot\_ID for random slope in Inla analysis |
| Datasource\_ID\_4INLAs | Unique number for each Datasource\_ID for random slope in Inla analysis |
| Location\_4INLAs | Unique number for each Location for random slope in Inla analysis |

Metadata Supplementary Data 2.

|  |  |
| --- | --- |
| Datasource\_ID | Unique identifier for each dataset. ID’s from BioTIME and GPDD were left as is. |
| Plot\_ID | Unique identifier for each site. Always nested in a Datasource\_ID |
| Location | Unique identifier of study area or reserve of the sampling event. Not always nested in Dataset. |
| Realm | Realm of sampling: here, always Terrestrial |
| Year | Year of sampling event |
| Number | Number of individuals (or mean number of individuals over iterations) of this taxon |
| Taxon | Taxon name as provided in original data source |
| Level | Taxonomic level (species, genus, family etc) |
| Rank | Numerical variable for taxonomic level |
| Phylum | Phylum |
| Class | Class |
| Subclass | Subclass |
| Order | Order |
| Suborder | Suborder |
| Family | Family |
| Subfamily | Subfamily |
| Genus | Genus |
| Species | Specific epithet |
| validTaxon | Incompletely harmonized and updated taxonomic names: Genus + species |
| Note | Notes or descriptions from original datasets |
| cYear | variable Year centered around mean for Inla analysis |
| CGYr1.0 | Initial abundance group, based on year 1 only |
| CG1.2 | Initial abundance group, based on years 1 and 2 |
| CG1.5 | Initial abundance group, based on years 1 – 5 |
| CGallYrs | Abundance group, based on all years |
| iYear | Index year for random intercept |
| rYear | Copy of iYear for random slope |
| rYear2 | Copy of iYear for random slope |
| Datasource\_ID\_4INLA | Unique number for each Datasource\_ID for random intercept in Inla analysis |
| Plot\_ID\_4INLA | Unique number for each Plot\_ID for random intercept in Inla analysis |
| Location\_4INLA | Unique number for each Location for random intercept in Inla analysis |
| Plot\_ID\_4INLAs | Unique number for each Plot\_ID for random slope in Inla analysis |
| Datasource\_ID\_4INLAs | Unique number for each Datasource\_ID for random slope in Inla analysis |
| Location\_4INLAs | Unique number for each Location for random slope in Inla analysis |
| TaxonPlot\_4INLA | Unique number for each within a plot for random intercept in Inla analysis |
| TaxonPlot\_4INLAs | Unique number for each within a plot for random slope in Inla analysis |

Metadata Supplementary Data 3.

|  |  |
| --- | --- |
| Datasource\_ID | Unique identifier for each dataset. ID’s from BioTIME and GPDD were left as is. |
| Datasource\_name | Name for easier identification of datasets |
| Location | Unique identifier of study area or reserve of the sampling event. Not always nested in Dataset. |
| Plot\_ID | Unique identifier for each site. Always nested in a Datasource\_ID |
| Plot\_name | Unique plot name |
| TaxonPlot\_4INLAs | Unique identifier for population (a species in a site) |
| CGYr1.0 | Initial abundance group in year 1 |
| Class | Taxonomic class |
| Order | Taxonomic order |
| Taxon | Taxon name as provided in original data source |
| Level | Taxonomic level |
| slope | Mean trend slope of the population, calculated as the model random slope per population + the mean random slopes at the study, location and site levels + correction factor for RtM |
| CImin | Minimum credible interval of trend, calculated the model's lower 90% CI of the population level random slope + the lower 90% CI's of the random slopes at the study, location and site levels + the fixed year effect + correction factor for RtM |
| CImax | Maximum credible interval of trend of trend, calculated the model's upper 90% CI of the population level random slope + the upper 90% CI's of the random slopes at the study, location and site levels + the fixed year effect + correction factor for RtM |
| percchange | Calculated percentage change per year for the trend slope |
| Population trend | Qualification whether there is at least weak evidence (zero outside the 80% CI of the trend slope):  no trend = 80% credible interval overlaps 0  >80% certain positive: 0 < than the lower 80% CI  >80% certain negative: 0 > than the upper 80% CI |