Analysing global economic costs of invasive alien species with the invacost R package

Boris Leroy[1*], Andrew M. Kramer[2], Anne-Charlotte Vaissière[3], Franck Courchamp[3] & Christophe Diagne[3]

[1] Unité Biologie des Organismes et Ecosystèmes Aquatiques (BOREA, UMR 7208), Muséum national d'Histoire naturelle, Sorbonne Université, Université de Caen Normandie, CNRS, IRD, Université des Antilles, Paris, France.

[2] University of South Florida, Department of Integrative Biology. Tampa Fl, 33620, USA

[3] Université Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique Evolution, 91405, Orsay, France

***Corresponding author:** Boris Leroy, leroy.boris@gmail.com

Abstract

Aim

Large-scale datasets are becoming increasingly available for macroecological research from different disciplines. However, learning their specific extraction and analytical requirements can become prohibitively time-consuming for researchers. We argue that this issue can be tackled with the provision of methodological frameworks published in open-source software. We illustrate this solution with the invacost R package, an open-source software designed to query and analyse the global database on reported economic costs of invasive alien species, *InvaCost*.

Innovations

First, the invacost package provides updates of this dynamic database directly in the analytical environment R. Second, it helps understand the nature of economic cost data for invasive species, their harmonisation process, and the inherent biases associated with such data. Third, it readily provides complementary methods to query and analyse the costs of invasive species at the global scale, all the while accounting for econometric statistical issues.

Main conclusions

This tool will be useful for scientists working on invasive alien species, by (i) facilitating access and use to this multi-disciplinary data resource and (ii) providing a standard procedure which will facilitate reproducibility and comparability of studies, one of the major critics of this topic until now. We discuss how the development of this R package was designed as an enforcement of general recommendations for transparency, reproducibility and comparability of science in the era of big data in ecology.

Key-words: biodiversity threats, biological invasions, drivers of change in biodiversity, economic impacts, economic costs, global change, invasive alien species, multi-scale analyses

Introduction

The global focus of macroecology makes it central to understand responses of biodiversity to global change and their implications for societies. Macroecological approaches tackle complex questions on patterns and processes of biodiversity across time and space with approaches exploiting knowledge and tools from different disciplines (Leitão et al., 2020). These multi-disciplinary approaches are necessary for the development of large-scale multi-driver models and scenarios of change in biodiversity and ecosystem services as required by assessments of the Intergovernmental Platform on Biodiversity & Ecosystem Services (Rounsevell et al., 2018). Typically, macroecological research relies on the unprecedented increase in large-scale datasets emerging from diverse research fields beyond ecology and evolution, such as climate or socio-environmental research (Hallgren et al., 2016; Vilela & Villalobos, 2015). However, analyses and interpretation of such large datasets often present an array of challenges on many fronts, which are generally related to high heterogeneity, varying quality and uncertainty, and intrinsic complexity of recorded information (Farley et al., 2018).

Despite growing frameworks to openly share data, methods, standards and software (e.g., ecoinformatics, Michener & Jones, 2012), many large-scale datasets lack accessibility and their specific nature and jargon can make them overly difficult to grasp for naïve users. This is especially true in macroecology where searching for large-scale datasets and learning their specific extraction and analytical requirements can become prohibitively time-consuming. The most promising solution to tackle this issue relies on applications allowing to access, extract, and analyse such datasets via graphical users interfaces (e.g., Global Biodiversity Information Facility website www.gbif.org) or open-source software (e.g., letsR, Vilela & Villalobos, 2015). In addition, data resources can suffer from a lack of dynamic maintenance for ensuring regular updates, ongoing expert evaluation and contemporary relevance of recorded data. Community-curated platforms can solve these issues by archiving, homogenizing and distributing data (e.g., the plant functional trait database TryDB www.try-db.org).

We illustrate here to bridge complex and dynamic datasets with accessibility in the frame of a large-scale database sufficiently heterogeneous to prevent primary users – i.e., ecologists – from readily using and analysing it. We focus on *InvaCost*, a freely available database on the economic costs of invasive alien species on socio-ecosystems worldwide (Diagne, Leroy, et al., 2020a). The database is global in extent and covers different taxonomic groups, ecosystem types, activity sectors, and temporal and spatial scales. *InvaCost* is an updatable repository that was designed to be a dynamic resource evolving as more information on the economic impacts of biological invasions becomes available. Indeed, the first version of this database has already been expanded by contributions from a large network of ecologists and managers (Diagne, Catford, et al., 2020). Since its recent publication, it has subsequently been complemented by a global effort to compile both English and non-English literature on the topic, which increased its total number of cost estimates from 2,419 to almost 10,000 (Diagne, Leroy, et al., 2020b, Angulo et al. in prep.).

The economic costs of invasive alien species are a case of heterogeneous data with different spatio-temporal scales that prevent straightforward global or large-scale study of this topic (Diagne, Catford, et al., 2020; Diagne, Leroy, et al., 2020b). Yet, biological invasions are a major threat to biodiversity (IPBES, 2019) for which there is still insufficient attention from decision makers and the general public, and which requires international coordination in order for actions to be effective and sustainable (Dana et al., 2014). Complementary to the increasing proofs that invasive alien species often jeopardize local biodiversity (Bellard et al., 2016), knowing the economic costs of biological invasions might contribute to make the issue more visible and ensure that the stakeholders address it (Caffrey et al., 2014; Diagne, Catford, et al., 2020; Diagne, Leroy, et al., 2020a; Lodge et al., 2016). Significant costs can justify the need to implement management measures and, for instance, to favour prevention measures. Under tight public monetary conditions, they can help to prioritize or improve the cost-effectiveness of management actions (Hanley & Roberts, 2019). Nevertheless, the prerequisite to all the foregoing is to be able to effectively and accurately respond to a large array of questions, such as the following: how are costs distributed across space, time,

taxonomic groups, and economic sectors? How have these costs evolved over the last decades and will evolve for the decades to come? How do damage and loss costs compare to management expenditures? Indeed, given the intrinsic heterogeneity and complexity of recorded costs, the underlying basic question should be 'what are the costs and how to compare them?'.

In order to ensure the desired results can be obtained and to preclude erroneous estimations, *InvaCost* data requires to be processed beforehand. First, cost data in the database are on varying spatio-temporal scales and were calculated with different assessment methods which make them not directly comparable. Second, there are statistical issues typical of such econometrics data which require specific analytical choices. Therefore, we need to make sure it is straightforward for the users who have not been trained as economists to properly extract information from this database. Furthermore, the absence of a standard procedure to derive cost values for invasive species may lead to the development of idiosyncratic and heterogenous methods, resulting in a lost opportunity for the repeatability and comparativeness of studies, as well as for monitoring the evolution of costs over time, thus reducing value for informing policy and management. For these reasons, we introduce the invacost R package as a robust tool to query and investigate the *InvaCost* database. This package was conceived following a workshop, mainly composed of ecologists and evolutionists working on invasive species, where the inherent difficulties of developing scripts and methods to adequately obtain economic values from *InvaCost* appeared as a limiting factor (Diagne, Catford, et al., 2020).

We developed the invacost R package with three main objectives. The first objective was to conveniently provide the updates of the database directly into R, in order to follow the dynamics of this living database and to relieve users from the burdens of compatibility issues and errors associated with loading such a large dataset in R. The second objective was to help users understand the nature of economic cost data for invasive species and the inherent biases associated with such data. The third objective was to provide two complementary ways

to analyse these data. One the one hand, a standard method to derive cumulative and average cost values over different periods of time, with relevant visualisation methods. On the other hand, a procedure to derive the trend of costs over time with different modelling techniques accounting for the statistical issues of such econometric datasets, such as non-linearities, heteroskedasticity, temporal autocorrelation and outliers. In the following sections, we describe these three objectives with their conceptualising rationale and we detail the methods implemented for these objectives along with relevant literature. Then, we illustrate how to perform the analyses easily with the package with a simple case study on the global economic costs caused by invertebrates (i.e., all non-chordata animals). Ultimately, we discuss why the invacost R package is an enforcement of general recommendations for transparency, reproducibility and comparability of science in the era of big data in ecology.

Requirements and installation of the invacost R package

The package requires a standard installation of R (version≥4.0.0) and is available on the Comprehensive R Archive Network. Upon installation eight dependencies will be automatically installed: dplyr (Wickham et al., 2020), earth (Milborrow et al., 2019), ggplot2 (Wickham, 2016), lmtest (Zeileis & Hothorn, 2002), mgcv (Wood et al., 2016), quantreg (Koenker, 2020), robustbase (Maechler et al., 2020), sandwich (Zeileis, 2004) and scales (Wickham & Seidel, 2020). All the package code is open-source and available from the GitHub repository, which also includes a detailed tutorial: https://github.com/Farewe/invacost.

Objective 1 – Querying, cleaning and filtering the InvaCost database

The *InvaCost* database is a dynamic database where existing information can be corrected and new data are continuously added. The latest version of the *InvaCost* database (accessible at https://doi.org/10.6084/m9.figshare.12668570) is shipped with every release of the R package. The database can be accessed with the command `data(invacost)`.

The database loaded in R contains a total of **64** fields (see here for a full description https://doi.org/10.6084/m9.figshare.12668570.v3). These fields include, for each cost record: bibliographic information; taxonomic information on the focal invasive taxa; geographic information on the impacted area; characteristics and descriptors of cost estimations (e.g., nature of methods, socio-economic sector impacted, type of cost such as damage or management); degree of reliability of cost; as well as several fields related to the cost amounts and their temporal coverage. To ensure reproducibility of published analyses, we provide the function `getInvaCostVersion` to rollback the database to previous major releases (this function currently includes four releases: 1.0, 2.0, 2.1 and 3.0).

The diversity of sources, cases and methods included in *InvaCost* will typically require users to make methodological choices about filters to apply to the database (e.g., reliable vs. unreliable sources, potential vs. observed costs) and about the costs to use (e.g., cost standardized on the basis of exchange rates or purchase power parity, spatial scale of the study). These different necessary choices as well as their required processing are detailed step-by-step in the online tutorial of the invacost R package (https://github.com/Farewe/invacost).

In our example on the global costs of invertebrates, we chose to filter out less reliable cost estimates and potential costs, to focus only on reported costs with a better certainty, which yielded a consolidated invertebrate subset of 1113 cost estimates (Fig. 1.1).

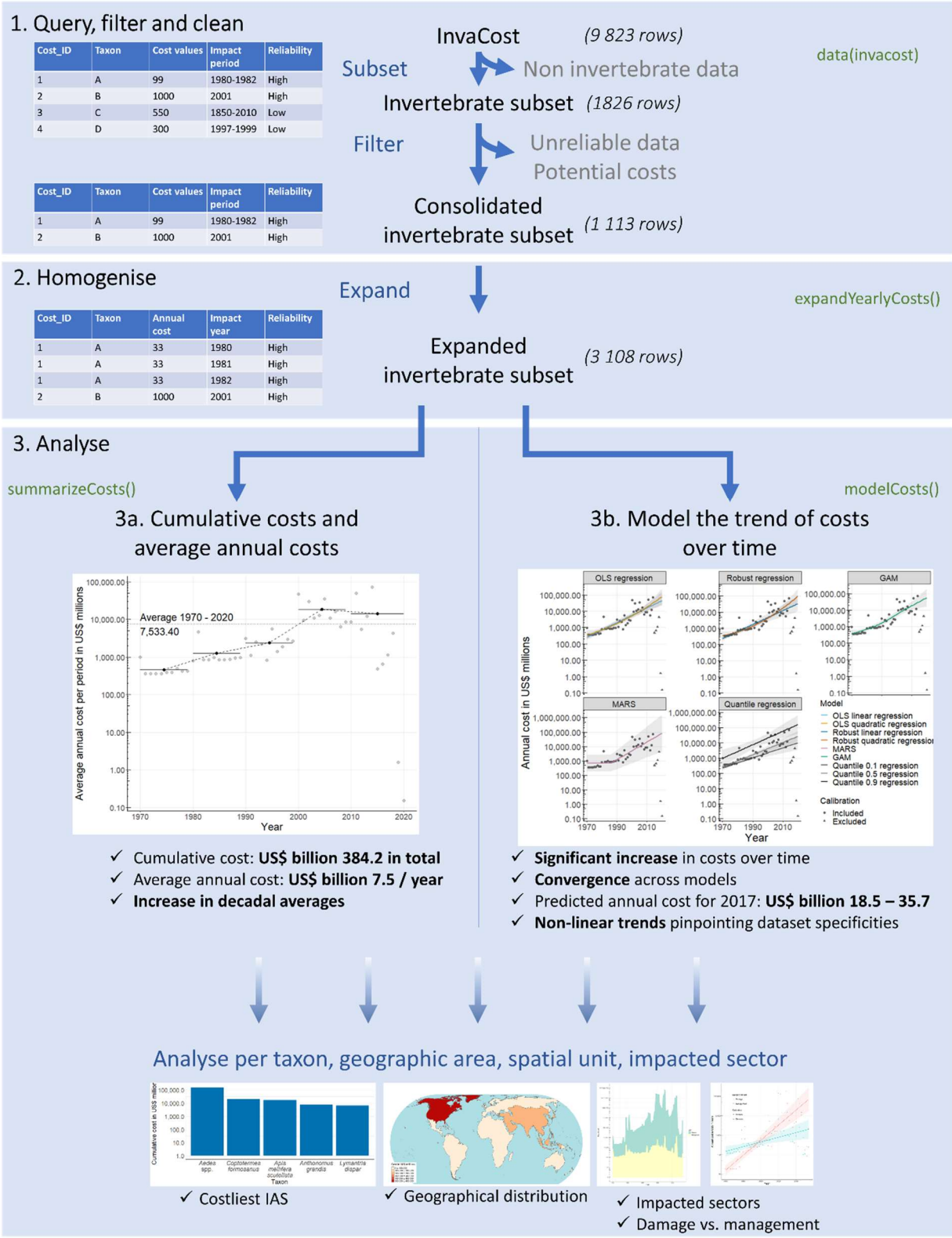

Figure 1. Conceptual framework of analysing economic costs of biological invasions within the invacost R package. We illustrate this framework with the example of a subset of the database (invertebrates, i.e., all non-chordata animals). The framework depicts the three

objectives detailed in the manuscript. Within each objective, functions used in the invacost R package are written in green. In 1 and 2, we illustrate with simplified tables how the structure of the database changes as the database is subsetted, filtered and then expanded. The cost over time graphs in 3a and 3b illustrate the native graphical outputs implemented in the package.

Objective 2 – homogenization of costs: expression in annual costs and expansion to their relevant time periods

Once the relevant filters have been applied to the database, the resulting subset of cost records cannot be used directly to derive cost estimations. Indeed, cost entries in the database have different temporal coverages: entries can be one-time costs, annual costs with repetitions over multiple years, or total costs of impacts which spread over multiple years. Therefore, to be comparable, cost estimates must be homogenized with a two-step process. First, they must be expressed with the same temporal unit, where the most relevant choice is annual costs. This step is already accounted for in the database with fields containing "cost_estimate_per_year" in their names. Second, once they have been homogenized on an annual basis, costs must be applied to their relevant time periods, i.e. repeated for each year over which the monetary impact was reported. This step is performed with the `expandYearlyCosts` function. This function relies on the fields indicating the starting and ending years of the annual costs. For example, reference ID 1619 reports an eradication budget of € 550 000 for *Anoplophora glabripennis* in France between 2003 and 2008. A preliminary step, already included in the *InvaCost* database allows to standardize the costs into a common currency (conversion and inflation into 2017 USD, see Diagne et al. 2020 for details) - it yields an annual cost of 2017 USD 136 437 for that period. The expansion step replicates this standard annual cost over each year of the impact period (2003-2008, Fig. 1.2). The costs are not expanded for the database by default because the database is easier to distribute in the compact form and because expanding the costs requires decisions which should be assessed by the user.

The expansion step requires adequate information with respect to the beginning and ending years of cost impacts. However, information on the beginning and ending years was not directly provided in the literature sources of economic costs for 23% of entries in the database (2302 rows of data). Therefore, for papers for which it was not available, we made educated guesses on the probable starting and ending years, and included these guesses in

the columns "*Probable_starting_year_adjusted*" and "*Probable_ending_year_adjusted*" columns. Because these columns are based on conservative assumptions (Diagne, Leroy et al, 2020b), they will limit over-estimation; hence, we recommend using these columns. Once the homogenization step has been performed on all cost entries in the user's consolidated subset of the database, extractions and analyses can be performed to explore the patterns of costs of invasive alien species. In our example, after expansion the data on invertebrate costs contained 3108 rows (Fig. 1.2).

Objective 3a – Estimating the cumulative and average available costs of invasions

The first method to analyse economic costs of invasive species consists of calculating the cumulative and average costs over time using cost estimates, as they appear in the collected material. These costs can be investigated on an annual basis, over the entire period covered by the database, or over a series of time intervals to account for the evolution of costs over time. All these alternatives are performed simultaneously with the function `summarizeCosts`. First, this function calculates the sum of costs for each year of the time period requested by the user (by default, from 1960 - first year with available exchange rates/inflation rates in the official websites, e.g. www.worldbank.org) to 2017 (year of currency standardisation used when building *InvaCost*). Second, it computes the cumulative total costs and average annual costs over the requested period. Last, it computes cumulative and average annual costs for user-defined time intervals (by default, 10 years) in the requested period.

The function returns an object that is a list composed of tables corresponding to the three aforementioned analyses, as well as the filtered and homogenized *InvaCost* table analyzed – for reproducibility – and the parameters used to run the function. The package further includes several generic functions to summarise and visualise this object. Typing the object name in the R console triggers a summary of the main results and useful information to

understand these results (e.g. time period and intervals). To visualise the results, users can use the generic `plot` function. This visualisation function returns a figure illustrating all the information at once (Figure 1.3a): the total annual costs of invasions over time as points on the graph, the average cost over the entire time period as a dashed line labelled with its value, and the average annual costs of each time interval illustrated with horizontal segments with value labels. We developed this figure following the recommendations of Weissgerber et al. (2015) with respect to the visualisation of small sample sizes. Indeed, time intervals over which average annual costs are calculated are generally small (10 by default), and individual annual cost estimates can be highly variable and not normally distributed – therefore, we voluntarily omitted error bars which can be misleading about the distribution of underlying data. Both the `summarizeCosts` function and its generic `plot` function can be customised with a variety of parameters, such as log-transformation of costs, visualisation of cumulative or average costs, and completely open customisation of graphical parameters (see examples in the online tutorial https://github.com/Farewe/invacost#customising-graphs).

A typical usage of this function is to quickly derive cumulative or average costs for specific subsets of the database, such as for specific geographical areas, type of cost, or on a per-species basis (see the detailed example on how to derive per-species cumulative cost estimates in the online tutorial https://github.com/Farewe/invacost#example-on-many-subsets-all-taxaspecies-in-the-database). In our example on the costs of invasive invertebrates, the function yielded a cumulative cost of 2017 USD billion 384.2 for the 1970-2017 time period, which corresponded to an annual cost of 2017 USD billion 7.5 per year (Figure 1.3a).

Objective 3b – Modelling the trend of economic costs of invasive species over time

The second analytical method implemented in the package consists of modelling the long-term trend in economic impacts of invasive alien species by fitting models of annual costs as a function of time. Such a modelling approach is appealing because it accounts for the

dynamic nature of costs, and thus can provide a robust estimate of the annual cost of invasions for any year within the calibration interval. In addition, under the appropriate statistical settings, it can reliably estimate whether the costs of invasive alien species are significantly changing over time. The package implements such a modelling procedure in the `modelCosts` function, which includes four different modelling techniques with specific parameterisation resulting in a total of nine models fitted. We chose these different statistical methods because they are complementary in their description of the trend of costs over time, and robust to the statistical issues of econometrics data: heteroskedasticity, temporal autocorrelation and outliers. We expect they will fulfil the general needs of most users. Below we briefly describe each modelling technique.

- **Ordinary Least Square regression** (OLS). These simple regression models are well understood by users which make them good candidates to illustrate the linear or log-linear trend of costs over time. While the estimation of coefficients is robust to the issues mentioned above, error estimations are not. Therefore, we implemented the covariance matrix with Heteroskedasticity and Autocorrelation Consistent estimators as described in (Andrews, 1991), using the `vcovHAC` function in R package `sandwich`. On the basis of these robust covariance matrices, the function derives 95% confidence intervals and estimates whether the regression coefficients significantly differ from zero with partial t test as described in Zeileis (2004), using the function `coeftest` from package `lmtest`. The function fits two OLS models, with linear vs. quadratic terms.
- **Robust regressions**. Because the econometrics data in *InvaCost* are likely to include outliers, which may significantly bias the estimates of linear regression, particularly when the time period of costs is unclear, we also implemented MM-type regression (hereafter called “robust regressions”). This type of regression model is based on iteratively reweighted least squares which makes them less sensitive to the effect of outliers than OLS regressions (Koller & Stahel, 2011; Yohai et al., 1991). This method estimates standard errors robust to heteroskedasticity and autocorrelation as

described in Croux et al., (2003). We implemented the `lmrob` function from the `robustbase` R package. Like OLS regressions, the function fits two robust models, with linear vs. quadratic terms.

- **Multivariate adaptive regression splines** (MARS). We expect nonlinearities in the cost trends over time to be common, so we implemented the non-parametric MARS model which automatically models nonlinearities, using Generalized Cross-Validation to avoid overfitting (Friedman, 1991; Hastie et al., 2009). We implemented MARS with the `earth` function of the `earth` R package, with the default parameters in order to follow Friedman's parameters, as described in (Milborrow, 2020a) – these default parameters minimize the risk of overfitting. The function provides prediction intervals which account for heteroskedasticity by fitting a linear model on the residuals, fitted with Iteratively Reweighting Least Squares (Milborrow, 2020b). Note, however, that the temporal range of *InvaCost* limits the number of data points such that we can only approximately model the variance, as explained in Milborrow (2020b). Therefore, there is greater uncertainty in the prediction intervals than in the predictions themselves.
- **Generalized additive models** (GAM). GAM models are automatic flexible statistical methods used to identify and characterize nonlinear regression effects (Hastie et al., 2009; Hastie & Tibshirani, 1990). Like MARS, the GAM model was chosen to show non-linear patterns in cost trends over time. To account for heteroskedasticity, we used a location-scale method which consists in fitting two GAMs, one for the average trend and one for the standard deviation. We implemented the `gam` methods from the `mgcv` R package, including the smoothing function `s` therein (Wood et al., 2016). We used a simple Gaussian location scale family (function `gaulss`) because, like MARS, the limited number of data points allows only for an approximate variance model.
- **Quantile regressions**. Contrary to the previous models which estimate the average trend in costs over time, quantile regressions estimate specific quantiles of the

distribution of costs over time. To describe the evolution of the distribution of costs over time, we implemented three quantile regression models, to estimate the conditional median, 0.1 and 0.9 quantiles. We implemented the `qt` function from the `quantreg` R package with default parameters (Koenker, 2020).

The fitting of these different models provides a full description of the linear (with or without outlier correction) and non-linear patterns in the average trend of costs over time, as well as linear trends in the distribution of costs over time. The output of the `modelCosts` function includes all the fitted models with their parameters, a table with predicted values per model over the temporal range chosen by the user, as well as diagnostic tools, such as the summary statistics specific to each model or the root mean square error between observations and predictions. The object also includes the formatted input data and parameters for reproducibility. Several parameters can be modified, including the temporal range of data to use; transformations to apply to cost values beforehand (e.g. by default, costs are log10-transformed); weights or a threshold to reduce the impact of years with incomplete data. For example, there is a lag between the occurrence of a cost and its reporting and publication in the literature. This time lag impacts the most recent years, which consequently constitute obvious outliers with annual costs significantly lower than the rest of the data, a pattern pervasive to all subsets of *InvaCost*. Users can account for this incompleteness of data either by investigating results of models robust to outliers (e.g., robust regressions), by defining an optional threshold to exclude the most recent years from calibration, or by applying optional weights to reduce the influence of incomplete years on model calibration (as illustrated in examples of the online tutorial).

Like in 3a, the output object of `modelCosts` is a list that can be managed in a user-friendly way. Typing the name of the object in the console provides a synthesis of the results, indicating important parameters (e.g., temporal frame of the input data and calibration data; transformations applied to cost values) and predictions for the final year of the input data. A summary of statistical details of fitted models can be accessed with `summary`. The calibrated

models can be visualised with `plot`, which illustrates the annual cost values of the input data along with model predictions and associated error bands (Figure 1.3b). Error bands are specific to each model and include confidence intervals for linear and robust regressions and GAMs, prediction intervals for MARS models and confidence intervals around quantile predictions for quantile regressions. Two types of model visualisation can be chosen (single plot vs. facets for each model type), and the code was designed such that users can entirely customise plots using all the possibilities of ggplot2 (Wickham, 2016). Last, when users are satisfied with their models and want to export results to prepare a manuscript, we provide the function `prettySummary` to export the main statistics for each model into a conveniently formatted table.

In our example on invertebrates, we excluded from model calibration all cost values from 2015 onwards because they constituted obvious outliers with a sudden drop of two orders of magnitude (Fig. 1.3b). We confirmed these outliers by investigating robust regressions calibrated on all data, which had set the weights of years above 2015 to zero. We found significant increases in costs over time, with convergent predictions among modelling methods suggesting annual costs of 2017 USD billion 18.5 to 35.7. Models MARS, GAM, and quadratic regressions identified non-linear patterns in costs, pinpointing an acceleration in cost increase from the mid-1980s onwards. Quantile regressions for 0.1 and 0.9 quantiles had distinct slopes suggesting an increase in between-year cost amplitude over time.

Discussion

The invacost R package provides easy access to the economic impacts of invasive alien species in R with a suite of easy-to-use and complementary robust methods to extract, analyse and report cost data. We expect this tool to be useful for scientists working directly or indirectly on invasive alien species, in at least two ways. On the one hand, it removes the burden of implementing adequate scripts to investigate economic impacts and reduces the associated risks of errors. On the other hand, the consideration of the rationale and functions of our package allows standardisation of the analytical processes underlying cost analysis associated

with IAS worldwide. This standardisation will facilitate reproducibility and comparability of studies, which was one of the intended objectives when the *InvaCost* database has been built. To further facilitate the use of this package, we wrote a complete and up-to-date online tutorial with multiple examples and explanations (https://github.com/Farewe/invacost). On a broader perspective, we consider that the development of this R package is an enforcement of general recommendations for transparency, reproducibility and comparability of science, particularly in the burgeoning era of 'big-data' in global change research. In the following paragraphs, we demonstrate why and how the experience learnt from building this package can serve as a sound basis for developing tools aiming at investigating existing large-scale databases.

First, data used for scientific studies should be archived and made available with enough information and clarity to ensure accurate interpretation by readers (Whitlock, 2011). However, information and clarity can become antagonistic in the case of complex and heterogeneous global datasets, because the amount of information required for an exhaustive description can result in an overflow of descriptive details. For example, there are eight different fields containing economic costs in the database, which makes a steep learning curve before deciphering which cost values should be used. The package and tutorial account for these issues by facilitating the attention of readers towards the basic information required to understand and extract data to fulfil their objectives. A facilitated understanding of the core aspects of the general database is the first step towards an in-depth exploration of the underlying datasets.

Second, cost data - like other heterogeneous measures explored at global scales, such as biodiversity (e.g., McGill et al., 2014) - are characterized by an extreme heterogeneity in the published literature, be it in terms of spatial scales (e.g., from local agricultural yield loss to countrywide syntheses of costs over multiple socio-economic sectors); temporal scales (from singular costs to multi-decadal impacts); taxonomic coverage (from single species impacts to multi-group syntheses); nature and quality of estimations (from accurate field-based

399 estimations to uncertain extrapolations over time and space); and measure complexity (e.g.,
400 types of costs, variety of currencies at different years). This extreme heterogeneity has led to
401 major issues of comparisons between studies as well as opacity over the completeness and
402 precision of cost estimates and extrapolations, which in turn led to repeated criticisms in the
403 past (Bradshaw et al., 2016; Cuthbert et al., 2020; Diagne, Leroy, et al., 2020a; Hoffmann &
404 Broadhurst, 2016). As a consequence, synthesising and comparing costs appeared as an
405 impossible achievement until recently. Yet, this was finally achieved through careful
406 documentation of all characteristics of cost estimates in the published and grey literatures
407 through a major community effort (Diagne, Catford, et al., 2020; Diagne, Leroy, et al.,
408 2020a). Once all characteristics have been categorised into standard descriptors (e.g., spatial
409 scale, temporal range, types of costs, etc.), duplicate or incomparable costs can be flagged and
410 filtered, and comparable costs can be analysed and synthesised. This harmonisation effort
411 would result in a lost opportunity if the methods used to describe and analyse cost data
412 turned out to be extremely heterogeneous as well. Therefore, the development of a consistent
413 analytical framework within the package naturally aligns with the initial harmonisation
414 objectives of *InvaCost*, and further facilitates comparability and reproducibility.
415 Furthermore, the methods presented here and the open-source nature of the code make it a
416 first step towards future improvements and refinements.

417 Third, implementing adequate analytical and modelling methods for data with inherent
418 biases requires significant time investment and statistical background - two requirements
419 that may deter interested users in academic or applied environments. We argue that this
420 issue is tackled by providing a peer-reviewed software where methods have been thoroughly
421 constructed to provide robust analyses which fit the needs of most users. A good illustration
422 for this solution is the popularity of software papers offering access to complex methods in
423 ecology, which accounts for nearly half of top-cited papers (Barbosa & Schneck, 2015). Aside
424 from developing robust analytical and modelling methods, we also seized this opportunity to
425 implement recommendations from data scientists in constructing the visualisation methods
426 (construction of graphs, Weissgerber et al., 2015; colours optimised for colour-blindness,

Wong, 2011), and facilitating the export and reporting of the results, (e.g., automated reporting to improve reproducibility, Makowski et al., 2020). These complementary recommendations should be systematically considered together when developing analysis tools targeting large-scale databases.

Fourth, large-scale datasets compiling current knowledge are generally incomplete and suffer obsolescence unless they are updated. Hence, their usefulness and perennity in the scientific literature is conditional to dynamic updates over time, oftentimes via community efforts (e.g., the Global Biodiversity Information Facility, FishBase, TryDB). The *InvaCost* database is also meant to be a living database with frequent updates - for example, the database has already received two major community-led contributions since its onset (Diagne, Leroy, et al., 2020b). In this context, the availability of latest updates of the database, directly in the analytical environment R, is time-saving and instrumental in facilitating research, which is well illustrated by the R Open Science initiative (ropensci.org) and the popularity of its reviewed packages in multiple scientific domains. However, database updates should not complicate the reproducibility of science. Therefore, the package provides two ways of reproducing analyses. First, the function `getInvaCostVersion` permits rollbacks over major updates of the database. Second, the specific consolidated subset of the database used for analyses is always stored in outputs. Therefore, we encourage users to provide their output objects from the package as supplementary material when submitting their work for publication. More generally, we highlight the importance of close and sustainable collaborations between database curators and the developers implementing the associated analytical tools.

In conclusion, we hope that the invacost R package will be strongly beneficial to researchers and stakeholders, by providing the most up-to-date version of the database on economic impacts of invasive alien species directly in R - with a series of standard and robust methods to extract, analyse and compare cost data. In the context of accelerating alterations in biodiversity and its main drivers of change, we advocate for the necessity of increasing

availability of relevant global-scale knowledge resources with easy access to analytical methods - the invacost R package may serve as an example for developing tools dedicated to investigating other massive data resources.

Acknowledgements

We thank Anna Turbelin and Phillip J Haubrock for their assistance in proofreading and beta-testing the invacost R package. We thank Nicolas Dubos for discussions on statistics.

References

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*. https://doi.org/10.2307/2938229

Barbosa, F. G., & Schneck, F. (2015). Characteristics of the top-cited papers in species distribution predictive models. *Ecological Modelling*, *313*, 77–83. https://doi.org/10.1016/j.ecolmodel.2015.06.014

Bellard, C., Genovesi, P., & Jeschke, J. M. (2016). Global patterns in threats to vertebrates by biological invasions. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1823), 20152454. https://doi.org/10.1098/rspb.2015.2454

Bradshaw, C. J. A., Leroy, B., Bellard, C., Albert, C., Roiz, D., Barbet-Massin, M., Fournier, A., Salles, J.-M., Simard, F., & Courchamp, F. (2016). Massive yet grossly underestimated global costs of invasive insects. *Nature Communications*, *September 16*, 14. https://doi.org/10.1038/ncomms12986

Caffrey, J. M., Baars, J., Barbour, J. H., Boets, P., Boon, P., Davenport, K., Dick, J. T. A., Early, J., Edsman, L., Gallagher, C., Gross, J., Heinimaa, P., Horrill, C., Hudin, S., Hulme, P. E., Hynes, S., & Macisaac, H. J. (2014). *Tackling Invasive Alien Species in Europe : the top 20 issues*. *5*(1), 1–20.

Croux, C., Dhaene, G., & Hoorelbeke, D. (2003). Robust Standard Errors for Robust Estimators. *Katholieke Universiteit Leuven*, *December*, Working Paper.

Cuthbert, R. N., Bacher, S., Blackburn, T. M., Briski, E., Diagne, C., Dick, J. T. A., Essl, F., Genovesi, P., Haubrock, P. J., Latombe, G., Lenzner, B., Meinard, Y., Pauchard, A., Pyšek, P., Ricciardi, A., Richardson, D. M., Russell, J. C., Simberloff, D., & Courchamp, F. (2020). Invasion costs, impacts, and human agency: Response to Sagoff 2020. *Conservation Biology*, *34*(6), 1579–1582. https://doi.org/10.1111/cobi.13592

Dana, E. D., Jeschke, J. M., & García-De-Lomas, J. (2014). Decision tools for managing

biological invasions: Existing biases and future needs. *Oryx*, *48*(1), 56–63. https://doi.org/10.1017/S0030605312001263

Diagne, C., Catford, J. A., Essl, F., Nuñez, M. A., & Courchamp, F. (2020). What are the economic costs of biological invasions? A complex topic requiring international and interdisciplinary expertise. *NeoBiota*, *63*(November 2019), 25–37. https://doi.org/10.3897/neobiota.63.55260

Diagne, C., Leroy, B., Gozlan, R. E., Vaissière, A. C., Assailly, C., Nuninger, L., Roiz, D., Jourdain, F., Jarić, I., & Courchamp, F. (2020a). *InvaCost*, a public database of the economic costs of biological invasions worldwide. *Scientific Data*, *7*(1), 1–12. https://doi.org/10.1038/s41597-020-00586-z

Diagne, C., Leroy, B., Gozlan, R. E., Vaissière, A. C., Assailly, C., Nuninger, L., Roiz, D., Jourdain, F., Jarić, I., & Courchamp, F. (2020b). *InvaCost: References and description of economic cost estimates associated with biological invasions worldwide*. Figshare. https://doi.org/https://doi.org/10.6084/m9.figshare.12668570.v3

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, *19*(1), 1–141. https://doi.org/10.1214/aos/1176347963

Hallgren, W., Beaumont, L., Bowness, A., Chambers, L., Graham, E., Holewa, H., Laffan, S., Mackey, B., Nix, H., Price, J., Vanderwal, J., Warren, R., & Weis, G. (2016). The Biodiversity and Climate Change Virtual Laboratory: Where ecology meets big data. *Environmental Modelling and Software*, *76*, 182–186. https://doi.org/10.1016/j.envsoft.2015.10.025

Hanley, N., & Roberts, M. (2019). The economic benefits of invasive species management. *People and Nature*, *1*(2), 124–137. https://doi.org/10.1002/pan3.31

Hastie, T., & Tibshirani, R. (1990). Exploring the Nature of Covariate Effects in the Proportional Hazards Model. *Biometrics*, *46*(4), 1005–1016. https://doi.org/10.2307/2532444

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer.

Hoffmann, B. D., & Broadhurst, L. M. (2016). The economic cost of managing invasive species in Australia. *NeoBiota*, *31*, 1–18. https://doi.org/10.3897/neobiota.31.6960

IPBES. (2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. (E. S. Brondizio, J. Settele, S. Díaz, & H. T. Ngo (eds.)). IPBES secretariat.

Koenker, R. (2020). *quantreg: Quantile Regression*. http://cran.r-project.org/package=quantreg

Koller, M., & Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis*, *55*(8), 2504–2515.

Leitão, P. J., Andrew, C. J., Engelhardt, E. K., Graham, C. H., Martinez-Almoyna, C., Mimet, A., Pinkert, S., Schröder, B., Voskamp, A., Hof, C., & Fritz, S. A. (2020). Macroecology as a hub between research disciplines: Opportunities, challenges and possible ways forward. *Journal of Biogeography*, *47*(1), 13–15. https://doi.org/10.1111/jbi.13751

Lodge, D. M., Simonin, P. W., Burgiel, S. W., Keller, R. P., Bossenbroek, J. M., Jerde, C. L., Kramer, A. M., Rutherford, E. S., Barnes, M. A., Wittmann, M. E., Chadderton, W. L., Apriesnig, J. L., Beletsky, D., Cooke, R. M., Drake, J. M., Egan, S. P., Finnoff, D. C., Gantz, C. A., Grey, E. K., … Zhang, H. (2016). Risk Analysis and Bioeconomics of Invasive Species to Inform Policy and Management. *Annual Review of Environment and Resources*, *41*, 453–488. https://doi.org/10.1146/annurev-environ-110615-085532

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-barrera, M.,

Verbeke, T., Koller, M., Conceicao, E. L. T., & di Palma, M. A. (2020). robustbase: Basic Robust Statistics. In *R package* (0.93-6).

Makowski, D., Lüdecke, D., & Ben-Schachar, M. S. (2020). *Automated reporting as a practical tool to improve reproducibility and methodological best practices adoption*. https://github.com/easystats/report

McGill, B. J., Dornelas, M., Gotelli, N. J., & Magurran, A. E. (2014). Fifteen forms of biodiversity trend in the Anthropocene. *Trends in Ecology & Evolution*, *30*(2), 104–113. https://doi.org/10.1016/j.tree.2014.11.006

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, *27*(2), 85–93. https://doi.org/10.1016/j.tree.2011.11.016

Milborrow, S. (2020a). *Notes on the earth package*. 1–68. http://www.milbo.org/doc/earth-notes.pdf

Milborrow, S. (2020b). *Variance models in earth*. 1–29.

Milborrow, S., Derived from mda:mars by Hastie T and Tibshirani R, & Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. (2019). *earth: Multivariate Adaptive Regression Splines*. https://cran.r-project.org/package=earth

Rounsevell, M., Fischer, M., Torre-Marin Rando, A., & Mader, A. (2018). *The regional assessment report on biodiversity and ecosystem services for Europe and Central Asia*. https://doi.org/10.5281/zenodo.3237429

Vilela, B., & Villalobos, F. (2015). LetsR: A new R package for data handling and analysis in macroecology. *Methods in Ecology and Evolution*, *6*(10), 1229–1234. https://doi.org/10.1111/2041-210X.12401

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology*, *13*(4), e1002128.

https://doi.org/10.1371/journal.pbio.1002128

Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology and Evolution*, *26*(2), 61–65. https://doi.org/10.1016/j.tree.2010.11.006

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. (R package version 1.0.2). https://cran.r-project.org/package=dplyr

Wickham, H., & Seidel, D. (2020). *scales: Scale Functions for Visualization* (R pacakge version 1.1.1).

Wong, B. (2011). Color blindness. *Nature Methods*, *8*(6), 441. https://doi.org/10.1038/nmeth.1618

Wood, S., Pya, N., & B, S. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*, 1548–1575.

Yohai, V., Stahel, W. A., & R, Z. (1991). A procedure for robust estimation and inference in linear regression. In Stahel & Weisberg (Eds.), *Directions in Robust Statistics and Diagnostics, Part II* (pp. 365–374). Springer. https://doi.org/doi: 10.1007/978-1-4612-4444-8_20

Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software*, *11*(10), 1–17. https://doi.org/https://doi.org/10.18637/jss.v011.i10

Zeileis, A., & Hothorn, T. (2002). Diagnostic testing in Regression Relationships. *R News*, *2/3*, 7–10. https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf