Readme files used in *Hidden patterns of insect establishment risk revealed from two centuries of alien species discoveries*

This repository includes the code and data used to generate the figures and tables used in the paper *Hidden patterns of insect establishment risk revealed from two centuries of alien species discoveries.*

Figure 2 and Table S.3 are not included in this readme file. Figure 2 is a schematic developed in Latex and Table S.3 includes all data sources in the right-most column.

Software requirements

use three software packages: MATLAB, R, and STATA. The code developed for R and STATA does not require the user to obtain any additional software — the necessary libraries in R are automatically download. The MATLAB program utilizes parallel computing to decrease the time spent on loops, but these loops will run in series if the user does not have the Parallel Computing Toolbox.

In addition, the data are saved as Excel or CSV files.

Initial steps

Several initial steps will be necessary before running the programs.

1. Ensure that all data and program files are stored and accessible.
2. The folders should be organized as they are in the repository. If a folder is missing or the folders are reorganized, the pathway system set in MATLAB will not run correctly.

Programs

Most of the figures and tables are generated using MATLAB. One figure (Fig. 1) is produced in the R and the entries for a table (Table S.2) were produced in STATA.

Generating figure 1

All files needed to generate Fig. 1 are stored within the “Fig_1” folder.

Figure 1 was generated using a combination of data developed in this study (Discoveries-data-v7 and Imports-data-v4) and previously developed shape files. The data developed for this study are stored in the subfolder “Data.” The remaining subfolders contain the shape files. The program file is titled “globe_barplots.R” and calls “Preamble.R.”

The “globe_barplots.R” program will automatically set the working directory to the location where the file is stored and install and call all necessary R libraries. A figure, titled “figure_1” will be saved to the location of “globe_barplots.R,” which will be within the “Fig_1” folder if the original folder system is maintained.

Generating figures 3–5 and S.1–S.7; and Tables S.1 and S.4

The remaining data are generated using the “Main_program.m” file, supporting script files, and the data. The script files are stored in the “Scripts” folder and the data are stored in the “Spreadsheets” folder.

Before running the program, the user should check their machine's memory. The maximum pseudo-profile likelihood estimation approach taken in this program requires a significant amount of memory and run-time. For a machine with 16 GB of RAM, users should consider 25 or fewer values for each α parameter. The number of considered values for each α parameter is defined on line 41 for the aggregate model and line 146 for the regional model. The default is 25 considered values for each α parameter.

Once the user has selected their preferred number of considered α parameters, they can run the program either a cell at a time or all at once. The initial run will generate four .mat files that are stored in the "Output" folder. During subsequent runs these, stored files will be imported rather that generated again. The sizes of these output files depend on the number of considered α parameters and are typically too large to store in this repository (> 25 MB).

Running the entire "Main_program.m" file will generate 10 figures that are stored in the "Figures" folder. Each file is stored as a PNG and named based on their position in the paper.

The program will print the parameter values used in the tables during two programs. The "Fit_effort.m" program will generate the three parameters reported in Table S.1. The "CI_sign.m" program displays the 2.5^{th} percentile, the median, and the 97.5^{th} percentile for each likelihood-ratio-based confidence intervals. A vector, named "star_vec," reports the number of stars associated with likelihood-ratio-based statistical inference.

The "Delay2Discovery.m" script provides the estimated 2.5^{th} percentile, the median, and 97.5^{th} percentile delay between establishment and discovery for each model. This program is run after both the aggregate and regional models.

Generating Table S.2

The STATA file "Quant_val.do" saved in the folder "Quant_value_compare" imports the data from "Value_quant_data.csv," which is included in the same folder. The user will need to change the (following "cd" on line 4) to where they have stored the file. The correlation coefficients reported in Table S.2 will be printed.