BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 9: e182246
doi: 10.3897/biss.9.182246

OPEN ACCESS

Conference Abstract

Predicting the Number of Coleoptera Species for a State Province in Colombia Using Deep Neural Networks and Open Data

Esteban Marentes Herrera‡, Mario J Mora Cardona§, Jennifer C. Girón Duque|

‡ Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Fusagasugá, Colombia
§ Pontificia Universidad Javeriana Cali, Cali, Colombia
| Natural Science Research Laboratory, Lubbock, United States of America

Corresponding author: Esteban Marentes Herrera (emarentes@humboldt.org.co)

Received: 12 Dec 2025 | Published: 17 Dec 2025

Citation: Marentes Herrera E, Mora Cardona M, Girón Duque JC (2025) Predicting the Number of Coleoptera Species for a State Province in Colombia Using Deep Neural Networks and Open Data. Biodiversity Information Science and Standards 9: e182246. https://doi.org/10.3897/biss.9.182246

Abstract

Colombia is a mega-biodiverse country, however the number of known species present in the country remains underestimated due to a lack of information, especially for understudied groups that are difficult to sample and identify, such as beetles (Coleoptera). It is not possible to obtain a precise number of species for this group using conventional methods, due to their high diversity and lack of resources for taxonomic research. For this reason, a modelling exercise was conducted using the Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al. 1999), with occurrence records published through SiB Colombia (Sistema de información sobre biodiversidad de Colombia) and GBIF (Global Biodiversity Information Facility). These records were labeled using checklists of various families and subfamilies compiled by experts of Coleoptera de Colombia, and enhanced by relevant abiotic measurements available in open repositories, which were identified as having an impact on the distribution of beetles. The best model was used to indirectly predict the number of beetle species in the Antioquia department, which has a large amount of available data and a variety of representative habitats.

© Marentes Herrera E et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Occurrence data for the order Coleoptera were downloaded directly from GBIF as of March 2025 (GBIF 2025a), using a global filter for the order Coleoptera. This dataset contains 32,356,574 rows, which include information on taxonomy, locality, coordinates, date, publishing entity, collector and type of record. Occurrence records of plant species associated with beetles were obtained from GBIF by downloading all biological records of the kingdom Plantae for Colombia (GBIF 2025b) and were spatially intersected with the Coleoptera data using a buffer of 1000 m. The abiotic variables selected included wind speed, relative humidity, precipitation, solar radiation, temperature, ecosystems/ habitats, elevation above sea level and their relevant information can be found at Herrera (2025).

Finally these data were labeled with the number of species per family of Coleoptera in Colombia compiled from available checklists published in GBIF via SiB Colombia by the research group Coleoptera de Colombia*[1] (Coleoptera de Colombia 2025). This data was used to annotate the final dataset and served as the dependent variable used in the regression models.

Four commonly used regression techniques were selected (linear regression, multilayer perceptron, deep neural networks and Random Forest) after a literature review of the most common methods used for species prediction Herrera (2025). These were built in Python, fine-tuning them for optimal performance. Data was split into training and testing sets to ensure validity and the best model was selected based on standard accuracy metrics and confirmed using cross-validation.

An escalated deep neural network was selected after getting the best values in the three metrics and was used to make predictions for the number of species per family in Antioquia and Colombia. The script with all the details is available in Suppl. material 1.

For Antioquia, the model predicted a minimum of 2,007, a maximum of 9,381, and an average value of 4,210 species of beetles. Using the same model for Colombia, there was a minimum of 3,226, a maximum of 14,991 and an average value of 6,420 species predicted. These predictions were compared to the last published checklist, which listed 6,170 species recorded from Colombia and was consistent with the number of species described from the country since 1945. These predictions were shared with expert coleopterists, who qualitatively assessed the results.

However, depending on the particular beetle family, experts thought the models may have overestimated or underestimated predicted values, highlighting the general uncertainty in species estimates and how artificial intelligence (AI) models are also unreliable in predicting a clear value, serving only as a starting point or providing a wide range to work with.

Keywords

biodiversity, modelling, regression

Presenting author

Esteban Marentes Herrera

Presented at

Living Data 2025

Acknowledgements

Thanks to the experts Juliana Cardona, Diego Martínez, Larry Jiménez, Juan Botero, John Neita and Claudia Uribe, who kindly assessed the results. To Ming and Camila Plata, who helped me review my thesis and to GBIF for the Graduate Researchers Award.

Funding program

2025 GBIF Graduate Researchers Award

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Chapman P, Kerber R, Clinton J, Khabaza T, Reinartz T, Wirth R (1999) The Cross-Industry Standard Process Model: CRISP-DM. CRISP-DM consortium, 91 pp. [In en]. URL: https://www.keithmccormick.com/wp-content/uploads/CRISP-DM%20No%20Brand.pdf
- Coleoptera de Colombia (2025) Conjuntos de datos Coleoptera de Colombia. GBIF. URL: https://www.gbif.org/dataset/search?type=CHECKLIST&publishing_org=2c39be5c-c11e-46d0-bcb4-552f2072d19f
- GBIF (2025a) Occurrence data Coleoptera. GBIF Occurrence Download[In en]. https://doi.org/10.15468/dl.retzxm
- GBIF (2025b) Occurrence Plantae Colombia. GBIF Occurrence Download https://doi.org/10.15468/dl.j6xumw
- Herrera EM (2025) Modelo de predicción para el número de especies de Coleoptera en el Departamento de Antioquia. Pontificia Universidad Javeriana Cali, Cali, 62 pp. [In es]. URL: https://vitela.javerianacali.edu.co/server/api/core/bitstreams/25858a62-8c79-4e11-a7ed-6a6d650895d5/content

Supplementary material

Suppl. material 1: Scripts and final models doi

Authors: Esteban Marentes Herrera
Data type: scripts
Brief description: The script used to run these predictions, best resulting models, test file with the labels by family and information for the training data is available at the GitHub repository https://github.com/EstebanMH-SiB/modelPredictColeopteraSpecies
Download file (1.30 MB)

Endnotes

*1 https://sites.google.com/view/coleopcol/inicio