# How to Improve Joint Suitability Mapping for Search Space Reduction?

**Haoyu Wang**
Department of Geography and the Environment, University of Texas at Austin, TX, USA

**Jennifer A. Miller**
Department of Geography and the Environment, University of Texas at Austin, TX, USA

## Abstract

Geoforensic analyses are used to identify the location history of objects or people of interest. An effective method for location history identification is to use joint probability or suitability of trace materials. Species distribution models have been used to derive joint suitability distributions using suitable biotic trace evidence such as pollen. One of the key objectives for such analyses is to effectively reduce potential search space and search effort for investigators. This research presents a novel framework for modeling the habitat suitability of pollen identified at the plant species-level to generate joint suitability maps. We provide major limitations and challenges faced by current geolocation analyses based on species distribution models, including opportunities to improve the joint suitability analyses for search space reduction. A conditional probability approach for geolocation identification is also demonstrated for possible future applications in real-world forensic cases.

**2012 ACM Subject Classification** Information systems → Geographic information systems

**Keywords and phrases** forensic geolocation, species distribution modeling, conditional probability, search space reduction

**Digital Object Identifier** 10.4230/LIPIcs.GIScience.2023.77

**Category** Short Paper

**Funding** The research was partly funded by DEVCOM ARL, ARO through a Multidisciplinary University Research Initiative Grant (#W911NF1910231). The research, interpretations, and perspectives reported here are those of the authors and should not be attributed to the Army or the Department of Defense.

## 1 Background

Environmental trace evidence helps link objects or people of forensic interest to time and locations [4]. One such useful candidate for trace evidence usually found on items at scenes is pollen because of their durability on multiple contact carriers such as soil, fabrics, and other materials [3]. The microbial and environmentally ubiquitous characteristics of pollen also make it easy to attach to surfaces. The ability to identify pollen is dramatically improved using DNA-based identification methods. For example, DNA metabarcoding with high-throughput sequencing technologies improved pollen identification in terms of both quantity and accuracy. This improvement can help generate high-resolution plant taxonomic results, leading to potentially more reliable applications using forensic evidence [2]. The practical use of biotic trace materials such as pollen and spores in forensic science has also been discussed in recent research [1, 2, 7]. New methods that estimate suitable habitats of pollen's parent plant taxa using species distribution models for geoforensic location analysis have also been introduced, but have not been widely used [9, 10]. Species distribution models are used in these studies to quantify species-environment correlations which can then be used

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 77; pp. 77:1–77:6
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

to predict the habitat suitability of plants and animals [5]. Joint suitability maps based on species distribution modeling results can then potentially reduce search areas and efforts for investigation purposes.

To test the feasibility of this joint suitability method, a study proposed a geographic attribution framework [10]. The authors collected bees in fieldwork and treated them as objects of interest, and the pollen grains sampled from the bees (pollen profiles of bees) were identified as trace evidence. Species distribution models were estimated for each identified species and combined to reduce the search space for an object that contained these species. Since the location of the bee (object) was known, the authors were able to assess the geolocation accuracy of models by quantifying and mapping the potential search space base on different percentiles. The authors used Google Earth Engine cloud-based geospatial platform that provides petabyte data and algorithms for fast computation to apply geographic attribution at a global scale. The inputs of this framework are georeferenced and filtered occurrences from the parent plant taxa of the recovered pollen species from the Global Biodiversity Information Facility, with more than 2.2 billion taxa information integrated from multiple data sources. The framework combines relative suitability distributions of taxa to a final prediction layer using a scaled-sum method, with percentiles indicating the priority of search areas for investigators, corresponding to different color hues as shown in Figure 1. These processes were also considered a set of methods for the *search space reduction* purpose. The *SSR score* in the top-right corner of Figure 1 shows the metric of joint suitability score, or search score, that indicates the performance of reducing the search space by comparing the joint suitability value between the object's location and all other locations. For more detailed explanations on the model building and accuracy assessment, see [10].

## 2 Limitations

Although the geographic attribution framework described here was useful when sufficient quantities of pollen are recovered from bee objects, some assumptions were made when we applied the search space reduction techniques, which bring limitations to the framework that was proposed in previous studies. The most noteworthy and challenging limitations of this framework are summarized below:

1. Current studies that use either probability- or suitability-based approaches (such as the use of species distribution models) to identify the geographic provenance of objects of interest have one common challenge. They can derive one best location or a series of probability-ranked locations. The top percentiles of location history such as the different percentiles/color hues of areas illustrated in Figure 1 are essentially a set of ranked search spaces. Study such as [8] has proposed methods to identify multiple traveled sites by objects of interest through solving geographic optimization problems, where suitability layers generated from species distribution models can be used as inputs. Although capturing any one portion of the total location history would be potentially helpful for investigation, discovering further methods to incorporate multiple location history identification is important. It is also hard for [10] to assess the location identification accuracy with information other than joint suitability of pollen, because the actual travel/foraging pattern or preference of each bee is hard to obtain.
2. Existing studies that generate the *joint* probability or suitability distributions of pollen's parent plants need to be retrospectively assessed for the distribution of each plant taxa. For geolocation analyses that involve combinations of multiple suitability layers, information may not be well analyzed through the combining process. For example,

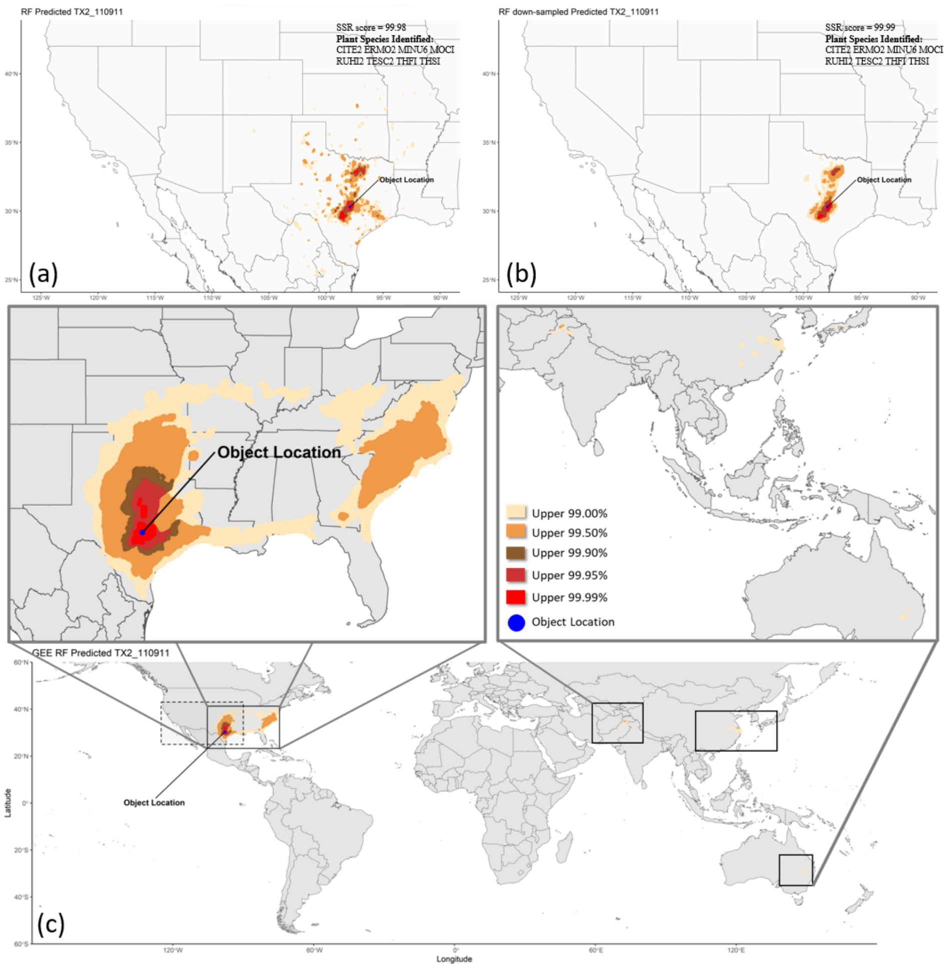

**Figure 1** An example of joint suitability maps of the geographic location history identification of a bee object. The modeling results were made by two widely used species distribution models: (a) Random Forest, and (b) Random Forest down-sampled, at a subcontinental scale. This bee object has nine different pollen genus/species attached. (c): Joint suitability search areas at a global scale. The dashed box shows the subcontinental study area in (a) and (b). Solid boxes indicate potential search areas. Darker hues indicate areas with increasing joint suitability values.

although joint suitability of certain pollen profiles on an object of interest has returned high accuracy of geolocation identification results, additional steps are required to know which one pollen or group of pollen is contributing to the identification, or which pollen is adding noise to the identification.

3. For the geographic attribution framework tested in [10], the sampling locations of bees are assumed to be locations for accuracy assessments. However, a sampling location of a bee should be ideally treated as one of the location history stamps. Although this is not a problem in real-world applications since investigations would usually attempt to identify all meaningful location history of the objects of interest instead of focusing only on sampled/collected location, the misplaced *true* location could have an adverse impact on how we understand the geolocation analysis results.

## 3 Updated Concepts

We provide two concepts based on the existing geographic attribution framework to potentially address some of the limitations mentioned above. For limitation #1, although the travel routes of bees are hard to obtain, this information could be partially available through inference or in some ways calculable in real-world forensic cases. Similar to [10], we set up a study area as a customized spatial domain, where $i, j$ are longitude/latitude grid cells that have $M \cdot N$ total grid cells, where $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$. For each grid location $(i, j)$, we use $\mathcal{L}$ to denote the incident that people or objects of interest have traveled to this specific location. The conditional probability of people and objects that have traveled at a location $(i, j)$ in a spatial domain is then provided as:

$$P(\mathcal{L}|T_1, T_2, \ldots, T_n) = \frac{P(T_1, T_2, \ldots, T_n|\mathcal{L}) \cdot P(\mathcal{L})}{P(T_1, T_2, \ldots, T_n|\mathcal{L})P(\mathcal{L}) + P(T_1, T_2, \ldots, T_n|\mathcal{L}^C)P(\mathcal{L}^C)} \tag{1}$$

where $T_k$ is a set of the distribution of trace evidence such as the pollen or other biotic materials identified on objects of interest or at scenes, where $k = 1, 2, \ldots, n$. Equation 1 is then illustrating how the pollen distribution probability provides an adjustment to probability surface derived by various investigation approaches, for example, criminal geographic targeting or geographic profiling that uses a set of locations from a series of crime [6]. The joint probability of equation 1 could be further computed with for example Bayesian inferences to solve the posterior probability which is the probability of people or objects have traveled to a location given that there is pollen found or corresponding plant taxa growing at this location. The minimal spatial unit for the calculation can be any meaningful size depending on the scales of focus, for example, a 900 m grid cell size used in the global geographic attribution cases.

To address the limitation #2 mentioned above, one would normally want to do repeated sampling of pollen profiles at one location, and need a method to distinguish and quantify the contribution of a single pollen within a pollen profile recovered on an object of interest. To achieve this, for every pollen profile of an object, one can keep one pollen out of the joint suitability combination and calculate the joint suitability score (search score) using the remaining pollen distribution layers, and repeat this process until every pollen found on this object is traversed. This is a methodology similar to leave-one-out cross validation, a procedure widely used in machine learning algorithms. To test the feasibility of this method, we first sampled multiple locations with various pollen profiles and fit species distribution models to obtain joint suitability search scores. Selected preliminary results from the leave-one-out method are shown in Figure 2. Each record at the x-axis of Figure 2 is the pollen that is left out in different pollen profiles. The mean search score difference on the y-axis is the difference in the two search space reduction scores before and after the corresponding pollen is left out. Negative score differences indicate that ignoring this pollen negatively affects geolocation identification, while positive score differences mean the opposite. We can then figure out how several pollen genus/species constantly contribute to or negatively affect the geolocation accuracy. For example, the genus of *Pinus* is always reducing the geolocation accuracy with a mean of around 0.03 for all geolocations we focused on. This corresponds to around five million pixels with a size of 900 × 900 m per pixel at a global scale. A possible reason for the negative contribution of *Pinus* is that pines as plants are growing in a large variety of environments and almost around the globe, contributing noise to most of the joint suitability analyses for geolocation identification. On the other hand, *Amaranthus* and other genus- and species-level pollen taxa with positive search contributions are having

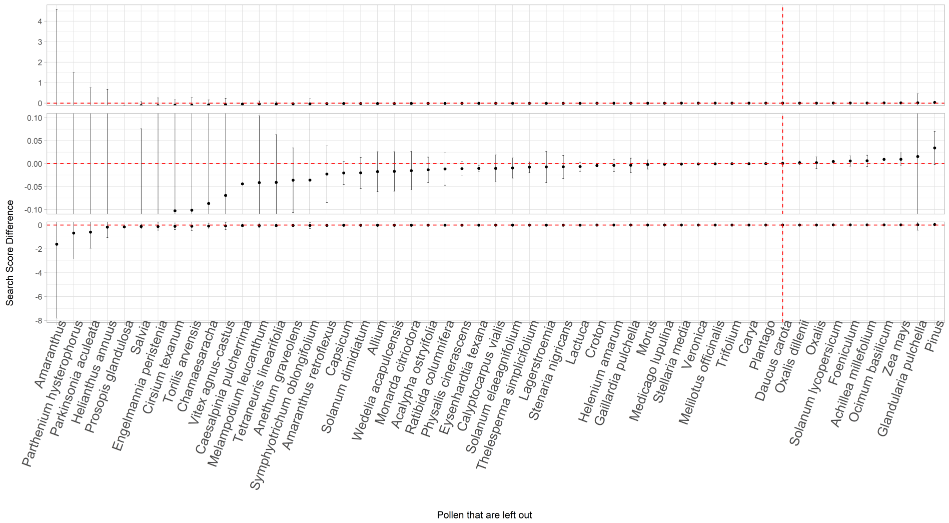

**Figure 2** An example of search score differences after removing pollen from an object's pollen profile using joint suitability analyses. This example was made from species distribution modeling results computed from boosted regression trees (BRT) with multiple geolocations in a global spatial domain. The horizontal red dashed line indicates no search score changed from joint suitability analyses after ignoring this pollen taxa. Pollen taxa at the right side of the vertical red dashed line (including *Daucus carota*) indicate positive search score differences, while those at the left side have negative search score differences.

more fluctuated search score differences. This feature may suggest investigators carefully examine available information from different cases, including objects/people's possible ranges of activities, when such pollen taxa are present on objects or locations of interest.

## References

1 Julia S. Allwood, Noah Fierer, and Robert R. Dunn. The Future of Environmental DNA in Forensic Science. *Applied and Environmental Microbiology*, 86(2):e01504–19, 2020. Publisher: American Society for Microbiology. doi:10.1128/AEM.01504-19.

2 Karen L. Bell, Kevin S. Burgess, Kazufusa C. Okamoto, Roman Aranda, and Berry J. Brosi. Review and future prospects for DNA barcoding methods in forensic palynology. *Forensic Science International. Genetics*, 21:110–116, March 2016. doi:10.1016/j.fsigen.2015.12.010.

3 Marzia Boi. Pollen attachment in common materials. *Aerobiologia*, 31(2):261–270, June 2015. doi:10.1007/s10453-014-9362-2.

4 D. C. Mildenhall. An unusual appearance of a common pollen type indicates the scene of the crime. *Forensic Science International*, 163(3):236–240, November 2006. doi:10.1016/j.forsciint.2005.11.029.

5 Jennifer A. Miller. Species distribution models: Spatial autocorrelation and non-stationarity. *Progress in Physical Geography: Earth and Environment*, 36(5):681–692, October 2012. Publisher: SAGE Publications Ltd. doi:10.1177/0309133312442522.

6 D. Kim Rossmo. *Geographic Profiling*. CRC Press, December 1999. Google-Books-ID: YQlS59Pv35oC.

7 Libby A. Stern, Jodi B. Webb, Debra A. Willard, Christopher E. Bernhardt, David A. Korejwo, Maureen C. Bottrell, Garrett B. McMahon, Nancy J. McMillan, Jared M. Schuetter, and Jack Hietpas. Geographic Attribution of Soils Using Probabilistic Modeling of GIS Data for Forensic Search Efforts. *Geochemistry, Geophysics, Geosystems*, 20(2):913–932,

2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GC007872. doi:10.1029/2018GC007872.

8 Daoqin Tong, Tony H. Grubesic, Wangshu Mu, Jennifer A. Miller, Edward Helderop, Shalene Jha, Berry J. Brosi, and Elisa J. Bienenstock. Identifying the spatial footprint of pollen distributions using the Geoforensic Interdiction (GOFIND) model. *Computers, Environment and Urban Systems*, 87:101615, May 2021. doi:10.1016/j.compenvurbsys.2021.101615.

9 Haoyu Wang, Jennifer A. Miller, Tony H. Grubesic, and Shalene Jha. A Framework for Using Ensemble Species Distribution Models for Geographic Attribution in Forensic Palynology. In *2022 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–7, November 2022. doi:10.1109/HST56032.2022.10025427.

10 Haoyu Wang, Jennifer A. Miller, Tony H. Grubesic, and Shalene Jha. Using habitat suitability models for multiscale forensic geolocation analysis. *Transactions in GIS*, 27(3):777–796, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.13052. doi:10.1111/tgis.13052.