Using geostatistical methods to automatically verify citizen science data on alien species

GI_Forum 2016
Salzburg, 06.07-09.07

Karin Wannemacher[1], Roland Grillmayer[2]

[1]University of Applied Sciences Wiener Neustadt, Austria · karin.wannemacher@fhwn.ac.at
[2]Austrian Federal Environment Agency (Umweltbundesamt) Wien, Austria,
roland.grillmayer@umweltbundesamt.at

Abstract

With the increase of travel and transportation of goods, the distribution and invasion of alien species have increased. While the majority of neobiota do not cause any problems, there are some that are problematic for nature conservation, have negative effects on the economy, or cause health problems.

In Regulation (EU) No 1143/2014 of 22 October 2014 the European Parliament and the Council of the European Union published a set of rules to prevent, minimise, and mitigate the adverse impacts caused by invasive alien species and orders Member States to implement a surveillance system of invasive alien species to prevent the spread of such species into or within the Union.

Citizen Science lends itself to the collation of high-quality information on a wide variety of species over a large scale and long term. However, verification of the collected data is a key challenge for legal monitoring projects, especially regarding costs. To ensure that the data fits the quality standards while also minimising the necessary budget for the project an algorithm for the automated verification of observation data has been developed. It is based on geostatistical methods.

Keywords:

Citizen Science, Alien Species, Geostatistics

1 Introduction

In 1860 a fungal disease, known as crayfish plague, came to Europe from North America. It proved fatal for the noble crayfish (Astacus astacus) population, which then, went into a steep decline. In order to compensate such losses, the signal crayfish (Pacifastacus leniusculus) was introduced to Europe's rivers and crayfish farms in the mid-20th century. The species, which comes from North America, is immune to the fungal disease. This move, however, turned out to be even more bad news for the indigenous species as the signal crayfish, while not affected by the disease, is a carrier of the lethal crayfish plague and also competes for the same habitats as Astacus astacus, where it often has the upper hand against the European species due to high reproductive rates. (ESSL & RABITSCH 2002)

Organisms which have been, intentionally or unintentionally, introduced into a specific area outside their natural range by humans after 1492 are generally known as alien species or neobiota. While the majority of neobiota do not cause any problems, there are some that are problematic for nature conservation, have negative effects on the economy, or cause health problems. (ESSL & RABITSCH 2002)

In Regulation (EU) No 1143/2014 of 22 October 2014 (EU REGULATION 2014) the European Parliament and the Council of the European Union published a set of rules to prevent, minimise, and mitigate the adverse impacts caused by invasive alien species. The paper states three forms of intervention: prevention, early warning and rapid response, and management. EU member states are ordered to implement a surveillance system of invasive alien species, which collects and records data on the occurrence of such species to prevent the spread of invasive alien species into and within the Union.

According to Article 14 (Surveillance system) of the regulation, such a system must cover the whole territory, and determine the presence and distribution of new and already established invasive alien species of the EU's concern (a); the process should be able to quickly detect any new invasive species (b); and take, as far as possible, relevant trans-boundary impacts into account (d); while complying with, but not duplicating, other regulations on species monitoring (c). The surveillance system should also be used to confirm early detection of the introduction or presence of invasive alien species of EU's concern. Any such early detection should be notified to the EU without delay (Article 16 - Early detection notifications).

2 Project Parameters

Species monitoring lends itself to citizen science, as economic and logistical factors prevent scientists from generating the volume of data they need for research on their own. One of the greatest concerns about citizen sciences is the quality of the gathered data. Yet studies have shown that, with proper instructions and statistical methods or expert verification in place, the data collected by citizen scientists can match the quality of data collected by experienced researchers. (JARVIS et al. 2015)

Records collated by volunteers are often the only source for high-quality information on a wide variety of species over a large scale and long term. (ROY et al. 2012)

For this project the incoming data is verified by an algorithm, which derives a number of index values to decide whether the observation is credible. The experts on alien species who are involved in the project do not have to evaluate every observation but only those that fail to meet the criteria set by them as threshold parameters for the algorithm. This will reduce the overall costs and make it possible to draft this as a long-term project.

While the overall aim is to provide a nationwide platform to observe alien species, it is also possible to connect local projects or projects with a limited list of taxa to the kernel that all feed the same database. Moreover, the same framework can also be used to target white or red-listed (endangered) species.

Monitoring potentially dangerous and invasive species will help us to establish an early warning system, which could make early responses and counter-measures to threats as effective as possible.

2.1 Project Species List

Many invasive species cause damages to managed and natural ecosystems or are responsible for the extinction of native species. The most problematic aliens are those that manage to fill a biological gap within an established ecosystem. Incidentally their habits, characteristics, and lifespans can have an effect on the surrounding fauna and flora. Neobiota can severely affect nutrient cycles if they are competing for the same resources or prey on native species. Another point of concern is the transmission of parasites and diseases (i.e., Hepatitis E) to species that have not had a chance to develop resistances over many generations. (ESSL & RABITSCH 2002)

As the list of invasive species mentioned in the EU Regulation 1143/2014 was yet to be developed at the time of the research project, a reduced list of species that has been agreed upon with specialists from the Austrian Federal Environment Agency. It included species that citizen scientists can easily identify and/or that are pretty common like Ailanthus altissima (Tree of heaven), Ambrosia artemisiifolia (Ragweed), Bunias orientalis (Turkish rocket), Potentilla indica (Indian mock strawberry) and Robinia pseudoacacia (False acacia). Ragweed, which also has a negative impact on human health (allergies), and false acacia are two of 17 alien plant species that are considered to pose a threat to biodiversity in Austria.

The project list also included Aethina tumida (Small hive beetle), a species where every occurrence in Austria ("Bienenseuchengesetz" [rules for the protection of bees against epizootic diseases], §3.1. and 2.) and the European Union is compulsorily notifiable.

Taxa are referenced internally via one of the three reference lists (EU-Nomen, EUNIS or Natura 2000) recommended in the INSPIRE scheme.

2.2 System Architecture

Users can submit their sightings by giving the location, species, date, and at least one picture of the specimen via a mobile-ready website. The upload of a photo of the monitored taxon is mandatory so that the classification can, if necessary, be verified.

Once submitted, the data will be processed by the pre-validation algorithm. The algorithm is designed to consider all the factors that an expert would inspect, apart from the observation picture, and return a value of certainty on whether the observation is credible or not (see section 3 and Figure 1). The combined value will be compared with a species-specific acceptance threshold value. If the credibility index is below the threshold, the observation requires additional verification by an expert user.

Additionally, other users or experts can confirm or oppose to the categorisation of observations. Once an expert confirms an observation, it is no longer open for user verification.

All observations are stored in a spatial database and their derivatives (heat maps, gridded data and features, among others) are accessible as OGC Webservices via Geoserver.

The gridded observation data is modelled taking the concepts of the INSPIRE Data Specification on species distribution into account. In this context, species distribution is defined as a geographical distribution of the occurrences of aggregated animal and plant species by using grids or polygons. (INSPIRE 2013)

The base grid used in this project is a grid for floristic observation data provided by the Federal Environment Agency. If a quadrant contains observations verified by an expert, then the polygon is assigned the value 1. If a quadrant contains observations which have not been verified by an expert, then the polygon is assigned the value 2. Both 1 and 2 indicate the presence of a species.

In accordance with the Inspire Data specification for species observation a distinction has to be made between areas where a thorough search for a particular species yielded no results and areas that have not been searched at all. As this kind of citizen science projects are not designed to confirm the true absence of alien species, quadrants that do not contain any observations are assigned the value 0 which means that the species was not searched for.

3 Pre-Validation Algorithm

Every submitted observation will be sent through the pre-validation algorithm (Figure 1).

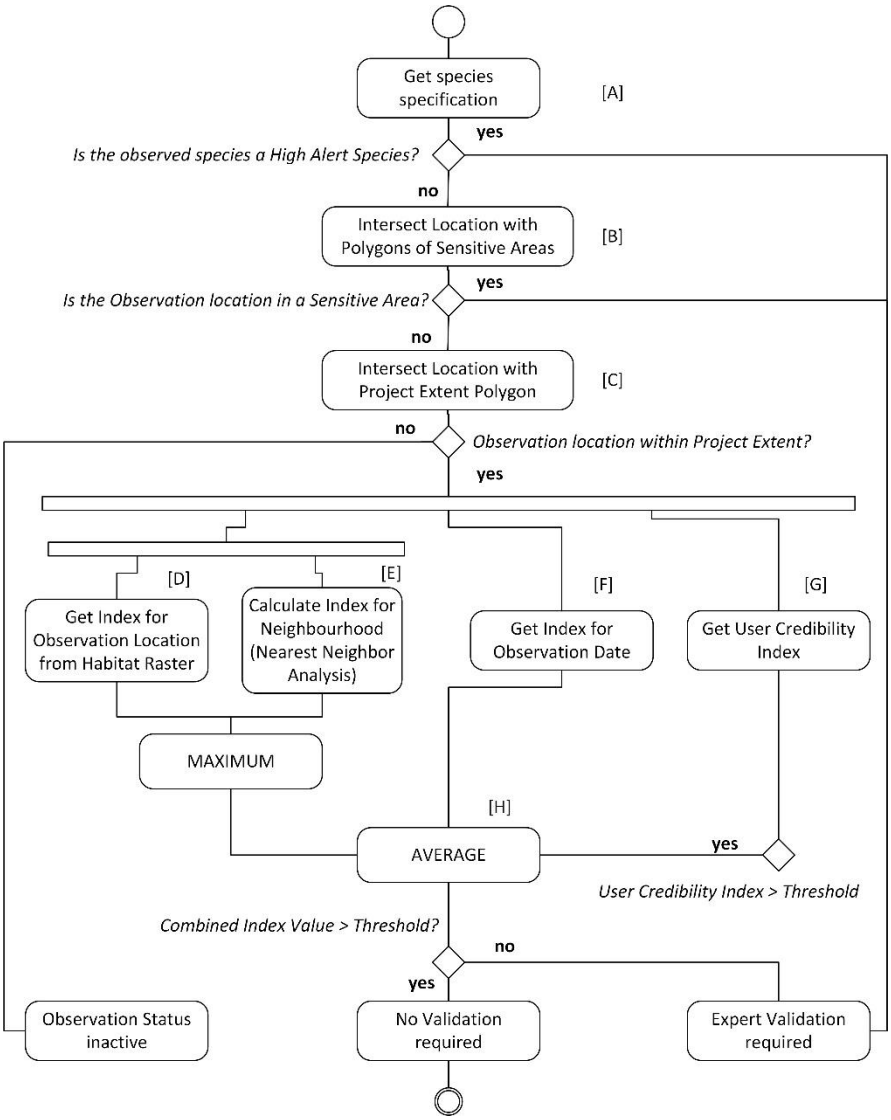

Figure 1: Activity Diagram Pre-Evaluation Algorithm

5.1 High Alert Species

The first module in the algorithm will look for species, where every observation has to be forwarded to a professional. Any such observation will be flagged for expert validation and not go through the rest of the components.

In their respective settings some species can be marked as “high-alert”. Any observation of such a species will be sent for expert validation (Figure 1 [A]). In addition, experts may get a notification via email as an immediate validation and a response might be vital in such a case.

A species may be stamped as “high-alert” because:

- they are of particular interest to scientists
- they are dangerous to the environment
- they are highly invasive
- they are known pests
- they have high impacts on human health or the economy

Furthermore, this part of the process is carried out in accordance with Article 16 (Early detection notifications) of the Regulation (EU) No 1143/2014. The article states that member states have to notify the European Commission of any sighting of an alien species whose presence was not previously known in their territory. (EU REGULATION 2014)

Within the pre-evaluation algorithm, the query for high-alert species is deliberately set before the location of the observation is checked against the extent of the observation area (i.e., Austria). Invasive alien species are, after all, a cross-border challenge and affect not just a single country. Article 22 (Cooperation and coordination) of the Regulation (EU) No 1143/2014 instructs every member state to make every effort to ensure close coordination with other member states, especially if they share the same borders or the same biogeographical region. (EU REGULATION 2014)

If any such observation from outside the project extent is sent via the website, an expert is able to evaluate the data, contact the user who sent it, and forward both, the pictures and data, to colleagues abroad and the authorities concerned.

5.2 Sensitive Areas

There may be areas where any observation of an alien species is of particular interest. Any sighting of a species in a sensitive area will immediately be flagged for expert validation (Figure 1 [B]).

It is possible to define sensitive areas for every species and check the observation points against these areas.

Such areas could be:

- Plantations, farms, etc.
- Protected sites and sanctuaries
- Habitats that house endangered or rare species
- Highly contaminated areas
- Sites with unique/specialised/rare ecosystems

This module is also set before the observation is checked against the extent of the project. However, it will only have an effect if sensitive areas themselves are defined beyond the project scope — for example,

transnational protected sites or national parks. Such an observation can be evaluated by a project expert and if necessary — and in conformance to the Article on “Cooperation and coordination” of the EU regulations regarding alien species (EU REGULATION 2014) — forwarded to the authorities abroad.

5.3 Project Extent

The coordinates of the observation will be intersected with the polygon of the project extent (i.e. Austria) (Figure 1 [C]). If the point lies outside the project polygon, the observation will consequently be flagged as inactive but still be kept in the database.

5.4 Observation Site

This module contains two blocks that are set to produce a one-dimensional parameter for a spatial point by answering the following question: how likely is an observation in the area? The habitat module will match the observation coordinates against the raster of a habitat model. The neighbourhood module will check for similar sightings within a specified distance (which reflects, for example, the maximum home range of a species) of the observations. Only the higher of the two values will be considered for the overall value.

If, for a particular area, the habitat value is repeatedly smaller than the neighbourhood value, it might indicate either an insufficient habitat model or be the first evidence of a species extending its known habitat and pioneering into a new territory. The habitat model itself acts as a safety net, especially during the early stages of the project where the density of observation points and known sightings is low.

5.5 Habitat Module

The habitat maps have to be modelled for each species (for this project the habitats were modelled using geostatistical methods). The index values have to be scaled to decimal numbers ranging between 0.0 and 1.0. The numbers indicate how likely it is for a species to grow or live in a certain area. The value 0 means that the location is considered inadequate for the species, while places with an index of 1 indicates a very suitable location or habitat for the species (Figure 2). Those models have to be prepared by experts who should take known characteristics of a species into consideration.

The precision of the model should be proportional to the quality of the input data as well as the information about the existing potential habitats for the species. The data collected during the monitoring project can later help to enhance or review the existing models.

A raster of the model is transferred into a table in a spatial database. For each observation the value for the given coordinates will be extracted from the raster (Figure 1 [D]).

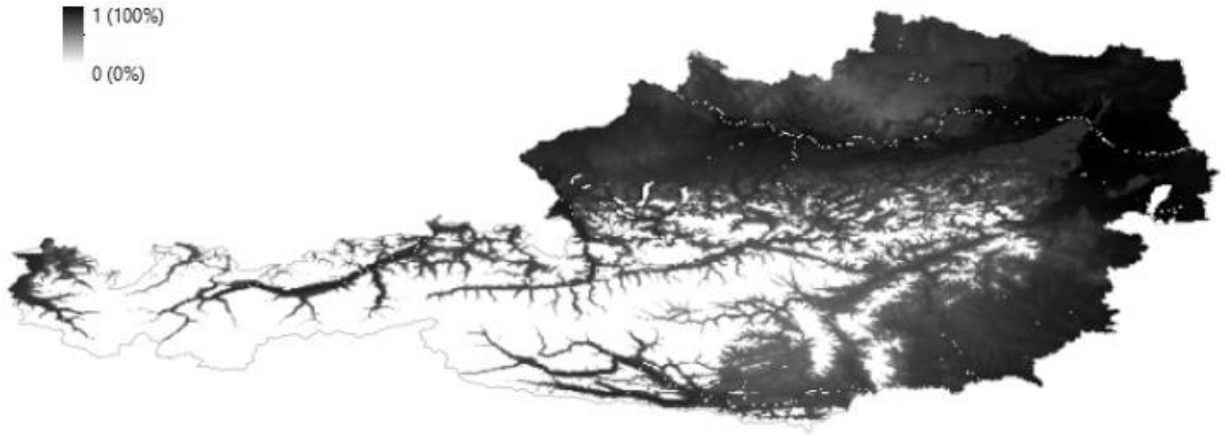

Figure 2: Habitat Model for Robinia Pseudoacacia in Austria

5.6 Neighbourhood Module

Here the algorithm will look for sightings of the same species within a defined radius from the observation point (Figure 1 [E]). It will check for other observations as well as known occurrences that stem from other sources (i.e., open data tree maps). The sighting that is closest to the new observation will be used to calculate the probability value:

$$P(r, d_{min}) = \frac{r - d_{min}}{r}$$

5.7 Observation Date

This module checks whether it is likely to correctly identify this species at the sighting date. (Figure 1 [F])

This block will match the observation date against discrete probability values between 0 and 1 for each month, which will indicate the likelihood of observing and correctly identifying the species at that time. While it may be very easy to correctly identify blooms during summer, it is unlikely that citizen scientists will be able to observe most flowers or hibernating animals during winter.

With discrete values there could be a considerable gap between the values of neighbouring months. However, even finer intervals cannot guarantee a better approximation of the actual number, as the

current development state of the plant or organism will heavily depend on local climate and location characteristics.

For larger surveillance areas it may be necessary to generate different date-related indices for different regions.

5.8 User Credibility

The last parameter reflects a user's credibility.

At the backend a table in the database keeps record of a user's observations by species, awarding 1 point for each categorisation that has been confirmed by an expert. The maximum score is 100 credibility points per species. If a user's score is higher than the species-specific credibility threshold number, the score, divided by 100, will be taken into account for the combined index.

A total of less than the threshold number of accurate observations will not have any impact on the overall probability number, keeping in mind that many potentially knowledgeable users will only sporadically upload data.

5.9 Combined Index

The indices for the observation location, date, and, if applicable, user credibility index are combined as an arithmetic mean (Figure 1 [H]).

$$x_{combined} = \begin{cases} \frac{1}{3}(x_{location} + x_{date} + x_{user}) & x_{user} > x_{threshold}(Species) \\ \frac{1}{2}(x_{location} + x_{date}) & x_{user} \leq x_{threshold}(Species) \end{cases}$$

All the components have been scaled to the same interval where a value of 1 says that a species is very likely to be correctly identified at this time or by this user, and that it is very likely to grow (or live) at the submitted location. An index value of 0 indicates that it is impossible to observe a species or that the location is not a suitable habitat. The threshold element will mark any observation with a combined index below the species-specific threshold value for expert validation.

Additionally, 1% of all observations will be randomly flagged for expert verification.

A feedback system was not included for the prototype installation, but would be recommended so that users get a notification if their observation is confirmed by an expert or why it hast been reclassified or refused.

6 Data evaluation

During a test period of 4 months (May – August 2015) for the project prototype 488 observations have been submitted project by 12 participants who don't have advanced botanical or zoological training or experience in identifying neobiota. All observations are located in the easternmost part of Austria where especially black locust and the tree of heaven show very high habitat indexes in their respective habitat models.

Four users managed to get their user credibility index above the threshold level for at least one species. Therefore, their score improved the combined pre-validation value for any further submissions for that species. For 226 observations the submitting user's credibility index was above the threshold.

On 41 occasions the neighbour index value was higher than the habitat value.

7 Conclusion and Outlook

Unbridled import of animals and plants over the past centuries has shaped today's world, bringing with it access to a wide range of resources and food supplies. But we also have to deal with the impact of new pests, diseases, and invasive species that came in its wake. Monitoring potentially dangerous and invasive species will help us to establish an early warning system, which could make early responses and counter-measures to threats as effective as possible.

The cornerstones of successful citizen science projects are the quantity of volunteers and the quality of the data that they submit.

For this project the latter is met by providing a simple technology in combination with an algorithm that validates the incoming data through a number of derived index values. The experts on alien species who are involved in the project do not have to evaluate every observation but only those that fail to meet the criteria set by them.

This will reduce the overall costs and make it easier to draft this as the long-term project the EU regulation asks for.

References

ESSL, F. & RABITSCH (2002), Neobiota in Österreich, Umweltbundesamt, Wien

EU REGULATION (2014), Regulation (EU) no 1143/2014 of the European parliament and of the council of 22 October 2014 on the prevention and management of the introduction and spread of invasive alien species, Official Journal of the European Communities

INSPIRE (2013), Inspire data specification on species distribution – technical guidelines, INSPIRE Thematic Working Group Protected sites

JARVIS, R. M., BREEN, B. B., KRÄGELOH C. U. & BILLINGTON, D. R., Citizen science and the power of public participation in marine spatial planning, Marine Policy, 2015, Vol. 57, pp. 21 – 26

ROY, H., POCOCK M., PRESTON, C.; ROY, D., SAVAGE, J., TWEDDLE, J. & ROBINSON, L., Understanding Citizen Science Environmental Monitoring, Final Report on behalf of UK-EOF