Herbicide resistance prediction: a mechanistic model vs a random forest model

Vorhersage der Herbizidresistenz: ein mechanistisches Modell vs. ein Random Forest-Modell

Janin Lepke[1], Johannes Herrmann[2], Roland Beffa[3], Otto Richter[1*]

[1]Technische Universität Braunschweig, Institut für Geoökologie, Langer Kamp 19c, 38106 Braunschweig, Germany

[2]Agris42 GmbH, Arnold-Cahn-Weg 7, 70374 Stuttgart, Germany

[3]Königsteiner Weg 4, 65385 Liederbach, Germany

*o.richter@tu-bs.de

DOI: 10.5073/20220124-063152

Abstract

Herbicide Resistance is a major issue in weed control. Prediction tools can help to detect herbicide resistance development in early stages and enable farmers to take countermeasures. These tools can be simulation models combining population dynamics and genetics or AI methods like random forest. To evaluate and train prediction models the data base used is important. Models using population dynamics and genetics depend on a numerous number of plant physiological traits such as seed dormancy, germination probability or seed production, and knowledge about the genetical inheritance (how many genes are involved). Furthermore, the initial conditions like the distribution of the seedbank or proportion of resistant plants in the population are important. To train an AI based prediction tool a large data set of field history data together with the resistance status of the fields is needed. We show that a random forest model trained with an artificial data set generated by a mechanistic model (HERRMANN, 2000) is able to make predictions of the resistance status for real data sets with acceptable accuracies. Simulated model data are therefore equivalent to field data and can be used to investigate the effect of sampling schemes on the detection of resistance.

Keywords: ALS inhibitor, classification, herbicide resistance prediction, population dynamic, random forest

Zusammenfassung

Herbizidresistenz ist ein aktuelles Problem der Unkrautbekämpfung. Prognosetools können helfen, die Entwicklung von Herbizidresistenzen in frühen Stadien zu erkennen und Landwirte in die Lage zu versetzen, Gegenmaßnahmen zu ergreifen. Diese Prognose-Werkzeuge können Simulationsmodelle sein, die Populationsdynamik und -genetik kombinieren, oder KI-Methoden wie Random Forest-Modelle, die Populationsdynamik und -genetik verwenden. Prognose-Modelle benötigen eine Vielzahl von pflanzenphysiologischen Merkmalen wie Samenruhe, Keimungswahrscheinlichkeit oder Samenproduktion und Kenntnisse über die genetische Vererbung. Weiterhin sind die Ausgangsbedingungen wie die Verteilung der Samenbank oder der Anteil resistenter Pflanzen in der Population relevant. Um ein KI-basiertes Vorhersagetool zu trainieren, wird ein großer Datensatz mit Feldverlaufsdaten zusammen mit dem Resistenzstatus der Felder benötigt. Wir zeigen, dass ein Random-Forest-Modell, das mit einem durch ein Simulationsmodell erzeugten künstlichen Datensatz trainiert wurde, in der Lage ist, Vorhersagen für einen realen Datensatz mit akzeptabler Genauigkeit zu treffen. Simulierte Modelldaten sind daher äquivalent zu Felddaten und können verwendet werden, um den Einfluss von Probenahmeschemata auf den Nachweis von Resistenzen zu untersuchen.

Stichwörter: ALS-Hemmer, Klassifikation, mechanistisches Modell, Populationsdynamik, Random Forest, Vorhersage von Herbizidresistenz

Introduction

The assessment of the resistance status of a field is prone to large uncertainties dependent on both the density of resistant plants and the sampling procedure. Having both a mechanistic population dynamic model (HERRMANN, 2016) and a random forest model trained by field data at our disposal (LEPKE et al., 2020), it was tempting to establish relations between both types of models. The first question was: can a random forest tool trained by model data give reliable predictions of the resistance status observed in the field? The mechanistic model had to be endowed with additional stochastic features such as the emergence of seedlings and the process of establishing the resistance status. Once the practicability of model data as suitable training set is established, data of the mechanistic model can be used to assess the influence of sampling schemes on the accuracy of resistance prediction.

Materials and Methods

Data

Field history information over a period of 6 years was obtained from farm management records. The data comprise field histories and the resistance status of *Alopecurus myosuroides* of 98 fields from the Hohenlohe area in Germany and 131 from the Champagne area in France. They are described in detail in the foregoing Proceedings (LEPKE et al., 2020). Predictor variables comprise in particular crop species rotation, number of crops used, seeding date, soil cultivation like ploughing or shallow tillage, and herbicide applications. In total, there are 20 predictors as described in Table 1 in LEPKE et al. (2020). From these data, an input vector for the random forest procedure was generated by allocating scores.

Model

The model simulates the population dynamics and genetics of *A. myosuroides* with target site resistance against three herbicides. The model structure and the parameterization was taken from the PhD thesis of HERRMANN (2016). It is assumed that resistance is conferred by one specific locus for each herbicide. The weed population is composed of 6 cohorts representing different seasonal phases. The general model structure is shown in Fig. 1. Population dynamic parameters depend on the crop, soil cultivation, and herbicide use, for details cf. HERRMANN (2016).

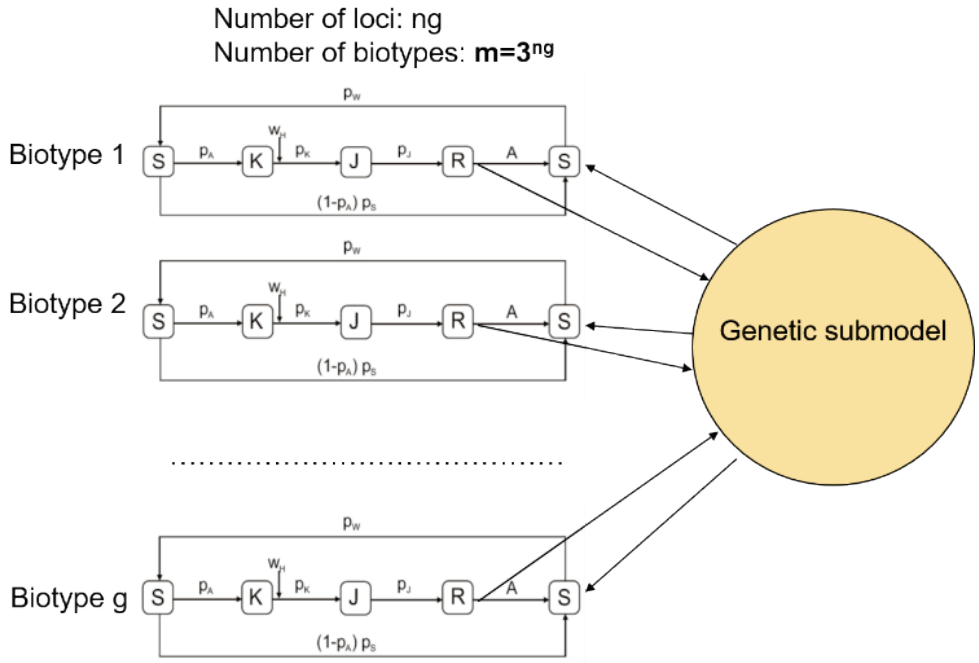

Figure 1 General model structure. Biotypes are coupled via a genetic sub model. This scheme is applied to each cohort. Cohorts are coupled via the seed bank. Notations: S: seedbank, K: seedling, J: young plant, R: mature plant, pi: transition probabilities.

Abbildung 1 Allgemeine Modellstruktur. Biotypen werden über ein genetisches Submodell gekoppelt. Dieses Schema wird auf jede Kohorte angewendet. Kohorten werden über die Samenbank gekoppelt. Notationen: S: Samenbank, K: Keimling, J: junge Pflanze, R: reife Pflanze, p_i: Übergangswahrscheinlichkeiten.

Assessment of the resistance status is a two-stage process consisting of i) taking seed plant samples in the field and ii) performing bioassays with the seeds grown in a greenhouse. This stochastic process was implemented into the model. Further stochastic elements are the emergence of seedlings and the efficacy of herbicide spraying. The model input is a vector of field management in accordance to real farm management records comprising herbicides, crops, soil cultivation and others (Table 1 in LEPKE et al., 2020). The simulations yield artificial weed data including the distribution of genes of herbicide resistance for several succeeding years.

For the classification problem the random forest method (BREIMAN, 2001) was used. In all analyses, the data were split into training data (75%) and test data (25%). To test which predictor variables have a significant impact on the response, importance measures such as the Gini index and the mean decrease accuracy were analysed.

Data generated with the simulation model are used to analyse the detection of resistance depending on the sampling scheme comprising the number of plants harvested randomly in the field and the number of plants analysed in the greenhouse.

Results

Example of a model run

Figure 2 shows an example of a model run for the most unfavourable management scheme: no crop rotation in combination with application of ALS inhibitors each year. In addition to the time courses of biotype and seedbank densities and others the model shows the time course of the result of resistance assessment (0 sensitive, 1 resistant) assuming that a test is performed each year.

Figure 2 underlines the effect of the sampling scheme on the classification. Although resistant plants are evolving, they remain undetected when 10 plants from the field and 8 plants in laboratory are taken. Augmenting the number of plants in the lab gives a positive result. Note that this is a random result serving as an illustration.

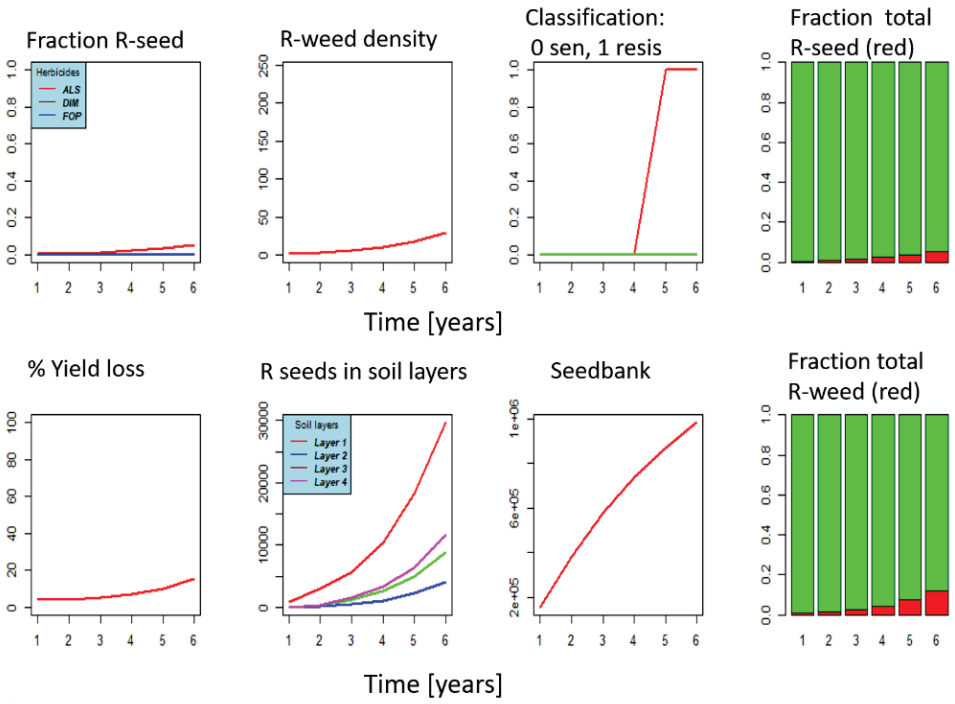

Figure 2 Model run for a simple management: no crop rotation, application of ALS inhibitors each year. The 3rd figure in the upper row demonstrates the effect of sampling on the detection of resistance. Green curve: 10 plants from the field, 8 plants in laboratory (no detection). Red curve: with 10 plants in laboratory resistance is detected.

Abbildung 2 *Modelllauf für ein einfaches Management: keine Fruchtfolge, jährliche Anwendung von ALS-Hemmern. Die 3. Abbildung oben rechts zeigt den Einfluss der Stichprobennahmre auf den Nachweis der Resistenz. Grüne Kurve: 10 Pflanzen vom Feld, 8 Pflanzen im Labor (kein Nachweis). Rote Kurve: Bei 10 Pflanzen im Labor wird Resistenz nachgewiesen.*

Model data as training set

Datasets were generated by the model for a variety of management schemes reflecting the information from the field catalogues. These data were used as a training set for the random forest classifier and applied to the common German-French data. It turned out, that classifications based on model training sets gave similar results as training sets from real data concerning accuracy and other measures (LEPKE et al., 2020). Figure 3 shows the performance criteria accuracy, AUC, type I errors and type II errors obtained from 50 repeated Random Forest runs. Mean accuracies range between 75 and 85%. Type II errors are lower than type I errors, i.e. false positive classifications are more frequent. This was also found, if predictions were made on the basis of field data. A possible explanation is that the misclassified sensitive field has features similar to the features of resistant fields, but resistance has not developed yet significantly or was not detected in the plant samples.

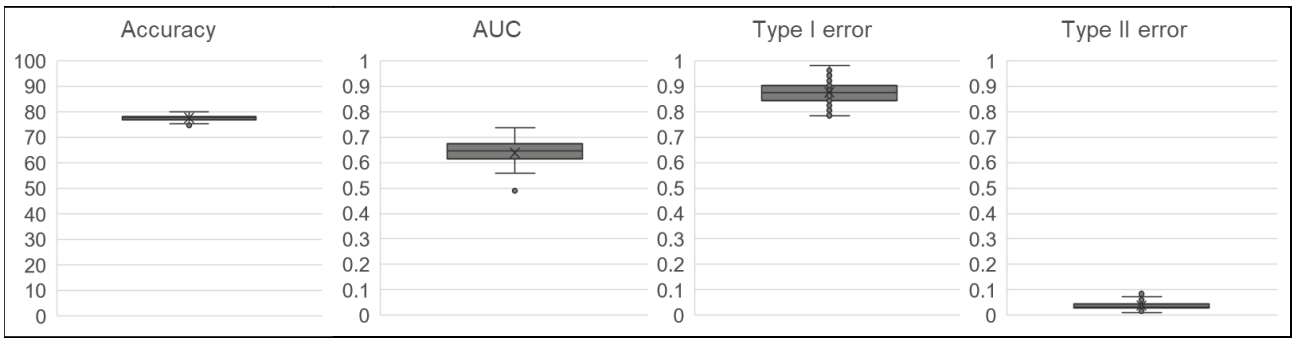

Figure 3 Performance criteria with model data as training set and the combined German-French data as test data obtained from 50 Random Forest runs.

Abbildung 3 *Gütekriterien für die Restenzvorhersage aus 50 Random-Forest-Läufen mit modellbasierten Trainingsdaten und den kombinierten deutsch-französischen Felddaten als Testdaten.*

Effect of sampling on resistance detection

In a first step, model data were generated based on the management information of the field management records. Resistance status of each field is known via the simulation. In a second step, the resistance status was assessed taking into account the two stage stochastic process of taking plant samples in the field and establishing the resistance status of seedling samples in the greenhouse. A field was classified as resistant, if at least two weed plants out of 8 grown from seeds of 10 plant samples from the field were resistant. In a second step, simulations were carried out to obtain sample fields with a large variation of resistance. To each of these fields the sampling procedure with varying number of field samples and greenhouse plants was applied. The results of these simulations (Fig. 4) show the influence of the sampling on the detection of resistance. It is interesting to note that the number of plants in the greenhouse have a larger impact on the rate of correct classifications than the number of plants sampled in the field. The Figure further shows, that for most combinations a saturation effect occurs, i.e. the number of correct classifications does not increase significantly once a resistance degree of 25% is surpassed. For small greenhouse samples (lower curves) a threshold can be seen: below 20-25% there is a high chance of false negative classifications.

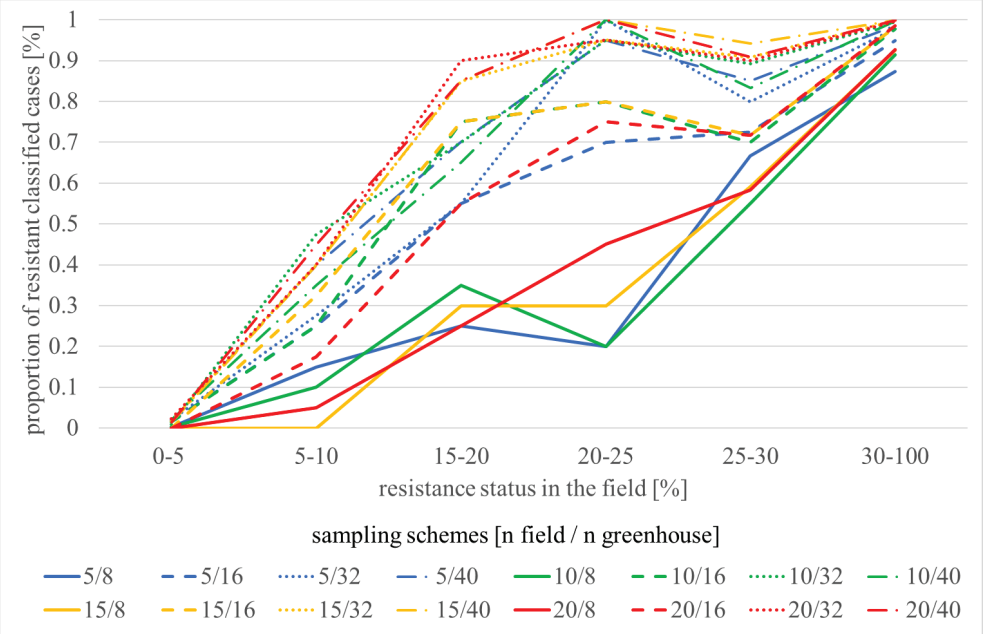

Figure 4 Influence of sampling in the field and in the greenhouse on the classification of fields at different degrees of resistance (abscissa). The pairs of numbers indicate the number of samples taken in the field (1. number) and the number of seedlings tested in the greenhouse (2. number). The curves are based on a large number of simulation runs.

Abbildung 4 *Einfluss der Probenahme im Feld und im Gewächshaus auf die Einstufung von Feldern bei unterschiedlichen Resistenzgraden (Abszisse). Die Zahlenpaare geben die Anzahl der im Feld entnommenen Proben (1. Zahl) und die Anzahl der im Gewächshaus getesteten Setzlinge (2. Zahl) an. Die Kurven basieren auf einer Vielzahl von Simulationsläufen.*

Conclusions

Two major conclusions can be drawn from our work

- Mechanistic population dynamics and genetics models based on human experience, knowledge, and intuition are well comparable with AI classification models solely based on data. Simulated model data can therefore be used as training sets.
- Simulations show that the detection of resistance depends in a non-linear way on the degree of resistance of a field. Below 10%, there is a high chance of false negative classifications. The rate of correct classifications is most sensitive to the number of seedlings tested in the greenhouse.

References

BREIMAN, L., 2001: Random Forests. Machine Learning **45**(1), 5–32.

HERRMANN, J., M. HESS, H. STREK, O. RICHTER, R. BEFFA, 2016: Linkage of the current ALS-resistance status with field history information of multiple fields infested with blackgrass (*Alopecurus myosuroides* Huds.) in southern Germany. Julius-Kühn-Archiv **452**, 42-49.

HERRMANN, J., 2016: Analysis of the spatial and temporal dynamics of herbicide resistance to ACCase- and ALS-Inhibitors in *Alopecurus myosuroides* Huds. and their causes, Dissertation, Technische Universität Braunschweig.

LEPKE, J., R. BEFFA, O. RICHTER, J. HERRMANN, 2020: Transferability of a random forest model for resistance prediction between different regions in Europe. Julius-Kühn-Archiv **464,**490-497.