

1 Improving the interpretability of species 2 distribution models by using local 3 approximations

4 **Boyan Angelov¹**

5 ¹**MindMatch GmbH**

6 Corresponding author:

7 Boyan Angelov¹

8 Email address: boyan.angelov@gmail.com

9 **ABSTRACT**

10 Species Distribution Models (SDMs) are used to generate maps of realised and potential ecological
11 niches for a given species. As any other machine learning technique they can be seen as “black boxes”,
12 due to a lack of interpretability. Advances in other areas of applied machine learning can be applied
13 to remedy this problem. In this study we test a new tool relying on Local Interpretable Model-agnostic
14 Explanations (LIME) by comparing its results of other known methods and ecological interpretations from
15 domain experts. The findings confirm that LIME provides consistent and ecologically sound explanations
16 of climate feature importance during the training of SDMs, and that the `sdmexplain` R package can be
17 used with confidence.

18 **INTRODUCTION**

19 In recent years an increased focus has been placed on making machine learning more interpretable
20 (Doshi-Velez and Kim, 2017; García et al., 2009). The main reason for this is that black-box models
21 are often not trusted by domain experts (Ribeiro et al., 2016). Moreover, the visualisation of feature
22 importances can often yield valuable insights. Such an analysis might help a scientist decide which data
23 points are not necessary (so that data collection can be improved), or more importantly, uncover new
24 details about what is contributing to the study phenomenon. In the very least, it can act as a sanity check
25 that the model is learning the right things and successfully avoids bias.

26 Species Distribution Modeling (SDM) is the application of machine learning on estimating the species
27 habitat based on occurrence data and associated environmental features (temperature, humidity etc.) (Elith
28 and Leathwick, 2009). Such models can be used to guide conservation efforts, estimate the effects of
29 climate change and answer other environmental hypotheses (Guisan et al., 2013; Austin and Van Niel,
30 2011). A field of such importance for ecology can benefit greatly from becoming more explainable.
31 Domain experts and people in the field rely on the maps produced by those models, and their trust in their
32 accuracy can be increased if they understand more clearly how models make decisions.

33 Some of the most important breakthroughs in explainable machine learning include the LIME Ribeiro
34 et al. (2016) and IML (Fails and Olsen Jr, 2003) projects¹. For this study we chose the former, due to its
35 more accessible API (Application Programming Interface).

36 Those are the motivations behind the creation of the `sdmexplain` package (Angelov, 2018a). This R
37 package has functions that enable a user to train SDMs and understand which features are most important
38 with various visualisations and an interactive map. In order to prove that the LIME explanations used are
39 indeed consistent with ecological observations, an analysis of two species was performed. Additionally,
40 the LIME results were compared to a standard method for computing feature importance in order to
41 determine in the patterns generated are statistically similar.

¹From this point onwards “interpretable” and “explainable” will be used interchangeably.

42 METHODS

43 Study species

44 In order to ecologically prove the explainability calculations the two species were selected based on their
45 relatively strong dependence on specific climate conditions. The occurrence (observation) data points are
46 shown on the map in Figure 1.

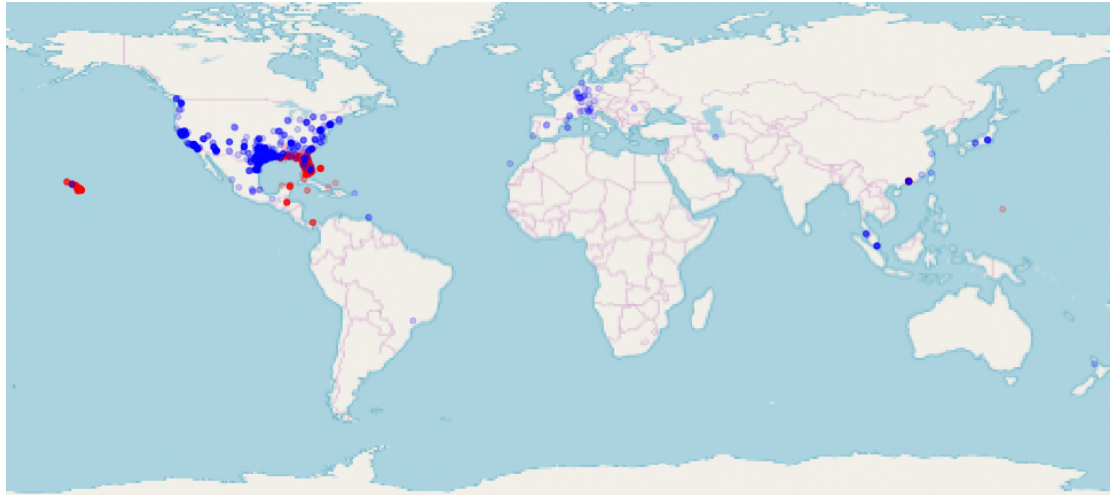


Figure 1. Occurrence map for raw GBIF data (red: *Eleutherodactylus planirostris*, blue: *Trachemys scripta*)

47 *Eleutherodactylus planirostris*

48 *Eleutherodactylus planirostris*, otherwise known as greenhouse frog, is a mostly terrestrial amphibian.
49 It occurs natively in Bahamas, the Cayman Islands, Cuba, and invasively in Jamaica, Guam, and the
50 southeastern United States. An important characteristic of this species is that for its eggs to hatch, 100%
51 humidity is required (Rödder and Lötters, 2010).

52 *Trachemys scripta*

53 Otherwise known as slider turtle, or common slider, *Trachemys scripta* is native to the United States, but
54 invasive in many parts of the world, including Southern Europe, Israel, South Africa and others. Because
55 of the negative effects of its invasive nature, this organism and its response to environmental conditions
56 has been widely studied. The slider turtle is heavily dependent on water availability throughout the year.
57 It's feeding is strongly temperature-dependent. The hibernation of neonates in nests is very sensitive to
58 low temperatures as well (Rödder et al., 2009).

59 Species records

60 Species records have been obtained by using the `sdbenchmark` package (Angelov, 2018b). In the back-
61 ground it obtains data from GBIF (<https://www.gbif.org/>) and performs additional domain-specific
62 preprocessing steps (such as removing occurrences with impossible, incomplete or unlikely coordinates
63 (based on the `scrubr` package, Chamberlain (2016))).

64 Climate data

65 Environmental data has been obtained from WorldClim (<http://www.worldclim.org/bioclim>) by using
66 the `sdbenchmark` package. In order to create variables that are biologically relevant, monthly temperature
67 and rainfall data are processed. Those derived variables are more representative of seasonal and limiting
68 climate characteristics. This preprocessing enables successful downstream species distribution modeling.

69 Species distribution models

70 Random Forests was chosen as the algorithm to create the main species distribution model for the LIME
71 explanations. It is a popular modeling technique that is known to perform well out-of-the box, with a
72 good balance between efficiency, speed and high tolerance for missing data Breiman (2001). It is also one

73 of the MLR “learners” that allow for the calculation of feature importance. The R package MLR (Bischi
74 et al., 2016) is used because of its ease of use and accessible API. In order to compare the results from
75 different methods, several other algorithms which support feature importance computation were also run:
76 RF SRC (Random Forests for Survival), GBM (Gradient Boosting) and rpart (Recursive Partitioning And
77 Regression Trees). Additionally, the Maximum Entropy (MaxEnt, Phillips et al. (2006)) algorithm was
78 also used in the comparison. It is one of the most popular and widely-used SDM techniques, and the
79 contribution of environmental features to the model performance can be extracted from the model.

80 Feature importances were compared between the MLR models, MaxEnt and LIME. For LIME the
81 individual predictions were gathered and the average contribution values per feature computed. For the
82 MLR models the importances are represented by the mean decrease of the Gini impurity index, shown in
83 Equation 1, where $p(i|j)$ is the proportion of samples of class c for a particular node t of the decision tree.

$$I_{Gini}(j) = \sum_i p(i|j)(1 - p(i|j)) \quad (1)$$

84 For LIME, a single explanation is computed in Equation 2, where $G \in G$ is the model, $\lambda(f, g, p_{i_x})$
85 is the “unfaithfulness” in how the the model approximates the locality p_{i_x} and $\Omega(g)$ is the complexity
86 measure.

$$\varepsilon(x) = \underset{g \in G}{\operatorname{arcmi}} \lambda(f, g, p_{i_x}) + \Omega(g) \quad (2)$$

87 For MaxEnt, the feature importance is represented internally by the software as percent contribution.
88 This is the result of algorithm modifying a single MaxEnt feature coefficient and assigning the gain to the
89 environmental variable the MaxEnt feature depends on (Phillips et al., 2006).

90 RESULTS

91 Model Training

92 The data consists of the aforementioned WorldClim environmental variables and a label (or target
93 variable). The latter contains the observations (encoded as a “positive” label). A common issue in species
94 distribution modeling is the lack of true absences recorded (“negative” label). In those cases, in order to
95 build a classifier, we need to artificially create those by sampling the background data, thus generating
96 pseudo-absences (Barbet-Massin et al., 2012).

97 Before model training the data was randomly split into train and test sets (70% and 30% of the whole
98 dataset respectively). This is a common method to make sure that we are not training and testing a model
99 on the same data, which can result in overfitting (the model learns the data and its noise too well, and
100 is prone to generalize poorly). For both species the models achieved good accuracy (% of correctly
101 classified occurrences) and Area under the curve (AUC ²) on the test set.

102 Feature Importance Analysis

103 The example visualisations of LIME importances (Figure 2) show that there are two types of contributions
104 a feature can make for a prediction: positive and negative, also with different magnitudes. Additionally,
105 we are provided with feature rules (i.e. $bio5 \geq 42$). Those add a new layer of interpretability to the model.
106 A first look at those visualisations show that often a few key features are often contributing the most to the
107 probability of occurrence at a given location. Those are quantified on Figure 3, where the mean feature
108 importances per model are shown.

109 In order to test that the feature importances between the different methods are in agreement, ANOVA
110 was performed on a multiple linear regression of species feature importance versus model type. The
111 alternative hypothesis (that there is a significant difference between model types) was rejected with
112 $p > .05$ for both species.

113 For *Eleutherodactylus planirostris* the most important features were `bio4` and `bio17`, while for
114 *Trachemys scripta* those were `bio19`, `bio17`, and `bio3`. Those results are summarized in Table 1.

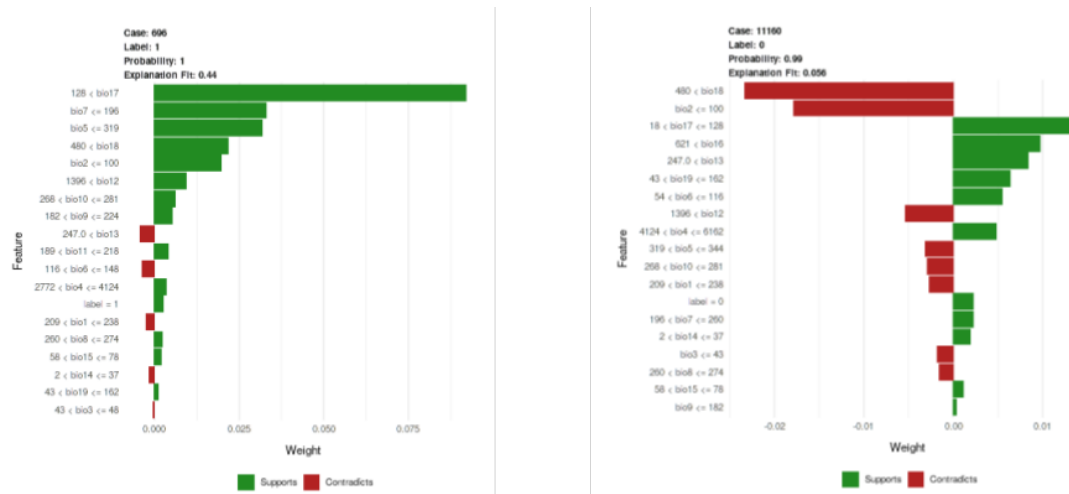


Figure 2. Example LIME plots

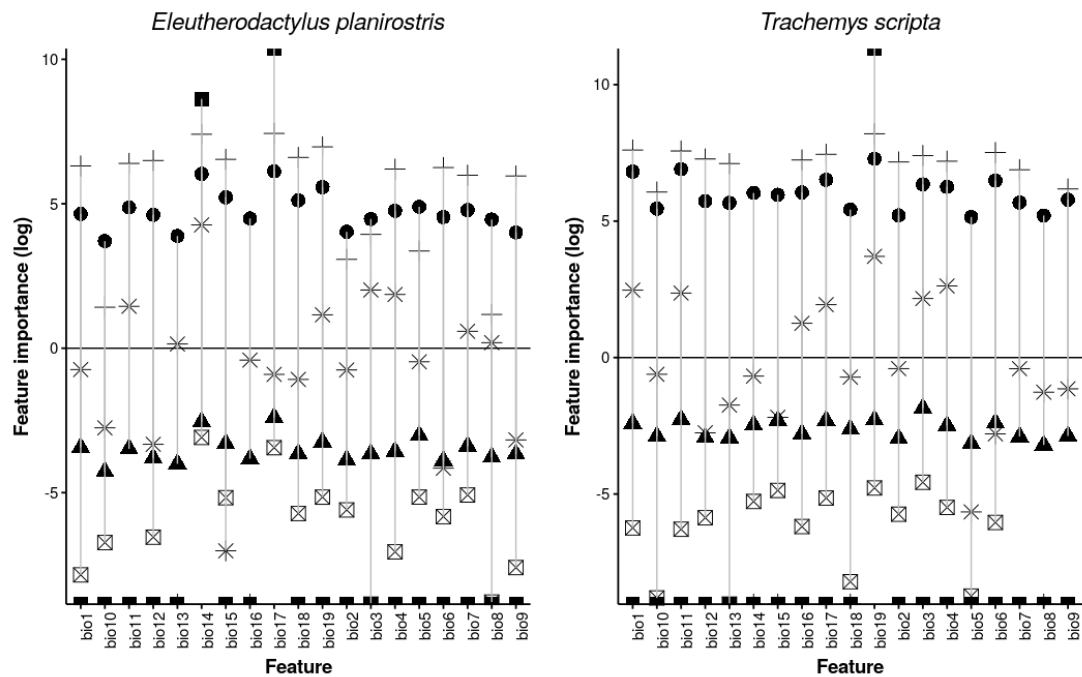


Figure 3. Feature importances calculated by different models (represented by shapes)

DISCUSSION

115

116 We can think of the model interpretability in three stages (Figure 4). In the first stage we are dealing with
 117 the raw data. Depending on how descriptive and intuitive the features are (i.e. the bioclimatic variables
 118 from WorldClim might be derived, but still can make sense to a domain expert), a user can already see
 119 some patterns in the data. Thus we can label this phase as a having a “medium” level of interpretability.
 120 The next step in a normal pipeline consist of the training of a SDM. This is the least understandable part
 121 of the pipeline, and most of the traditional model output is focused on various performance metrics. Some
 122 specific models, such as the ones in this study, allow for feature importance calculation (those are the
 123 various “tree-based” algorithms), but even this might not be enough, since all you get are averages across

²A standard method of estimating the performance of a binary classifier.

Species	Features	Description
<i>Eleutherodactylus planirostris</i>	bio4, bio17	Temperature Seasonality ($stdev \times 100$), Precipitation of Driest Quarter
<i>Trachemys scripta</i>	bio3, bio17, bio19	Isothermality ($\frac{bio2}{bio7} \times 100$), Precipitation of Driest Quarter, Precipitation of Coldest Quarter

Table 1. Most important features for both species

Interpretable Machine Learning

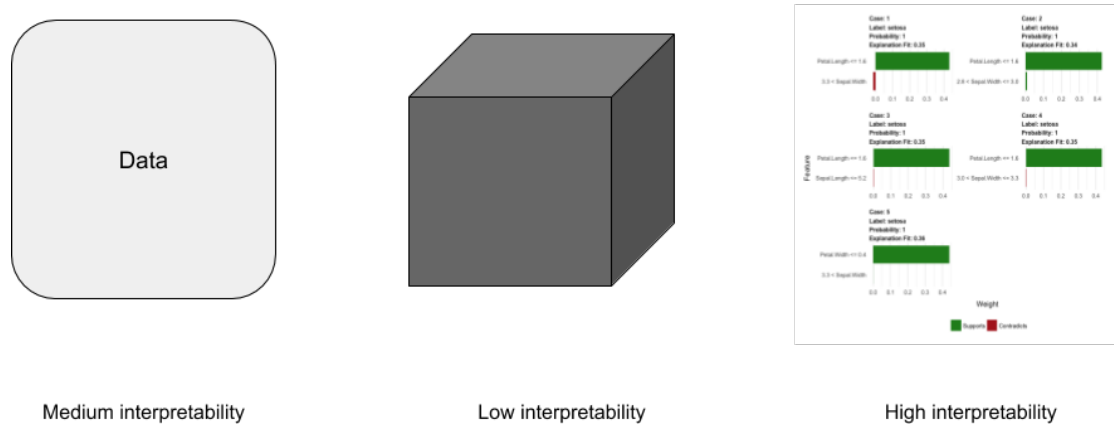


Figure 4. Levels of interpretability in machine learning

124 the whole dataset. Thus we can label this second phase as being the “black-box” and having a “low”
 125 interpretability. The final phase is the one that can shed a light on the black box model is the application
 126 of LIME (or similar methods) on the model, as used in the current study. Here we can achieve a very fine
 127 understanding of feature importance per observation. This can further allow for better understanding of
 128 the model mechanism, a simple sanity check that the model is learning the right things, and maybe further
 129 ecological hypothesis generation.

130 In statistical terms, the machine learning algorithms used are operating by using different mechanisms
 131 (despite all of them being tree based). Moreover the exact measures of “importance” also differ between
 132 the MLR methods, LIME and MaxEnt. Nevertheless, there is no statistically significant difference in
 133 results obtained across them, increasing our confidence that LIME is statistically similar.

134 In the case of *Eleutherodactylus planirostris*, humidity was determined to be the most important factor
 135 in limiting the species distribution. Since the hatching of the eggs requires $\sim 100\%$ humidity this is to be
 136 expected. For *Trachemys scripta* temperature and water availability can have dramatic consequences on
 137 the behavior and reproduction of the species, and this is also confirmed by the study.

138 Conclusions

139 This study provides scientific support for the usage of the `sdmexplain` package by confirming that
 140 LIME explanations are statistically sound and ecologically meaningful. The most important functionality
 141 of `sdmexplain` is the generation of interactive and explainable SDM maps, as shown on Figure 5. Such
 142 maps can be very useful for understanding SDMs, and providing guidance for experts in the field. Model
 143 explainability is a topic of hot research and there is a variety of new methods appearing in the field. Those
 144 can and should be evaluated in the context of species distribution modeling, especially on species of
 145 critical interest and importance.

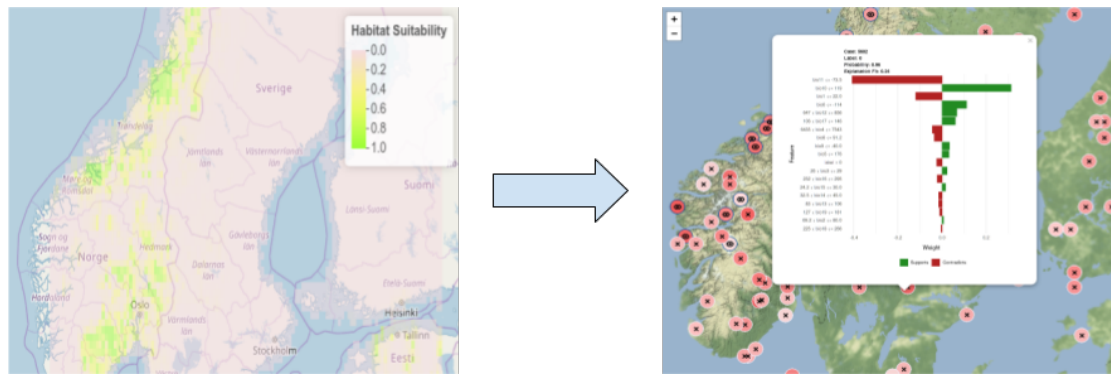


Figure 5. Interactive map with explainable occurrences

146 ACKNOWLEDGEMENTS

147 This research would not be possible without the original LIME package (Ribeiro et al., 2016), and its R
148 port (Pedersen and Benesty, 2018).

149 REFERENCES

- 150 Angelov, B. (2018a). `boyanangelov/sdmexplain: sdmexplain: An R Package for Making Species`
151 `Distribution Models More Explainable`.
- 152 Angelov, B. (2018b). `sdmbench: R package for benchmarking species distribution models`. *Journal of*
153 *Open Source Software*.
- 154 Austin, M. P. and Van Niel, K. P. (2011). Improving species distribution models for climate change
155 studies: Variable selection and scale. *Journal of Biogeography*, 38(1):1–8.
- 156 Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo-absences for
157 species distribution models: how, where and how many? *Methods in ecology and evolution*, 3(2):327–
158 338.
- 159 Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M.
160 (2016). `mlr: Machine learning in r`. *The Journal of Machine Learning Research*, 17(1):5938–5942.
- 161 Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- 162 Chamberlain, S. (2016). `scrubr: Clean Biological Occurrence Records`. R package version 0.1.1.
- 163 Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*
164 *preprint arXiv:1702.08608*.
- 165 Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction
166 across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697.
- 167 Fails, J. A. and Olsen Jr, D. R. (2003). Interactive machine learning. In *Proceedings of the 8th international*
168 *conference on Intelligent user interfaces*, pages 39–45. ACM.
- 169 García, S., Fernández, A., Luengo, J., and Herrera, F. (2009). A study of statistical techniques and perfor-
170 mance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*,
171 13(10):959.
- 172 Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., Regan,
173 T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini,
174 R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier,
175 S., Kearney, M. R., Possingham, H. P., and Buckley, Y. M. (2013). Predicting species distributions for
176 conservation decisions. *Ecology Letters*, 16(12):1424–1435.
- 177 Pedersen, T. L. and Benesty, M. (2018). `lime: Local Interpretable Model-Agnostic Explanations`. R
178 package version 0.4.0.
- 179 Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species
180 geographic distributions. *Ecological modelling*, 190(3-4):231–259.
- 181 Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions
182 of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge*
183 *discovery and data mining*, pages 1135–1144. ACM.

- 184 Rödder, D. and Lötters, S. (2010). Explanative power of variables used in species distribution modelling:
185 An issue of general model transferability or niche shift in the invasive Greenhouse frog (*Eleutherodactylus planirostris*). *Naturwissenschaften*, 97(9):781–796.
186
187 Rödder, D., Schmidlein, S., Veith, M., and Lötters, S. (2009). Alien invasive slider turtle in unpredicted
188 habitat: A matter of niche shift or of predictors studied? *PLoS ONE*, 4(11).