Assessing Feature Importance in Non-Linear Regression: An Ecology Perspective

Anthony F. Cannistra
tonycan@uw.edu

DEPARTMENT OF BIOLOGY

Motivation:

- For years, linear regression has been used by ecologists to study and quantify observed ecological relationships.
- These tools have been useful in understanding ecological mechanism, but our societal mandate to **predict future ecological state** requires models which offer more predictive power than a simple linear method can (there's reason to believe many relationships are nonlinear).
- To maintain scientific and practical credibility, these models must be at least to some extent *interpretable* to both researchers and the users of their predictions.
- Simple methods to capture the nonlinearities of many ecological relationships with provable properties are kernel regression and SVMs.

Key Question: What methods exist to extract variable importance from a non-linear regression, and how do the results therein compare to coefficients from a standard linear regression model? **Do the non-linear methods still capture theoretically-rigorous biology?**

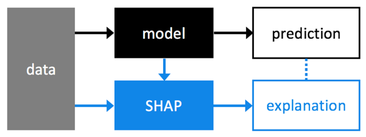
Data:

- **Angert et al. 2011** gathered **functional trait data** for Swiss alpine plants ($N = 133$, **Holzinger et al. 2008**) and computed the *observed elevational range shift* from historical records (i.e: how much has the maximum elevation at which these plants are observed shifted in the past century?). There are $d = 32$ observed traits.
- This is one of two datasets analyzed.

Methods:

- We compare **Ordinary Least Squares' coefficient values** to :
  - Kernel Regression (RBF, Polynomial) coefficients
  - Ridge Regression coefficients
  - Random Forest Gini Scores
  - **Shapley Values[1]** for SVM (RBF, Polynomial)

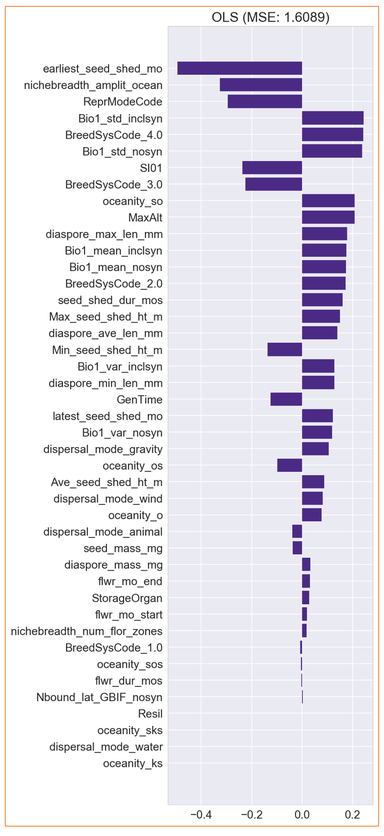
1: The Shapley additive feature importance measure, defined in **Lundberg and Lee 2017**, from cooperative game theory.



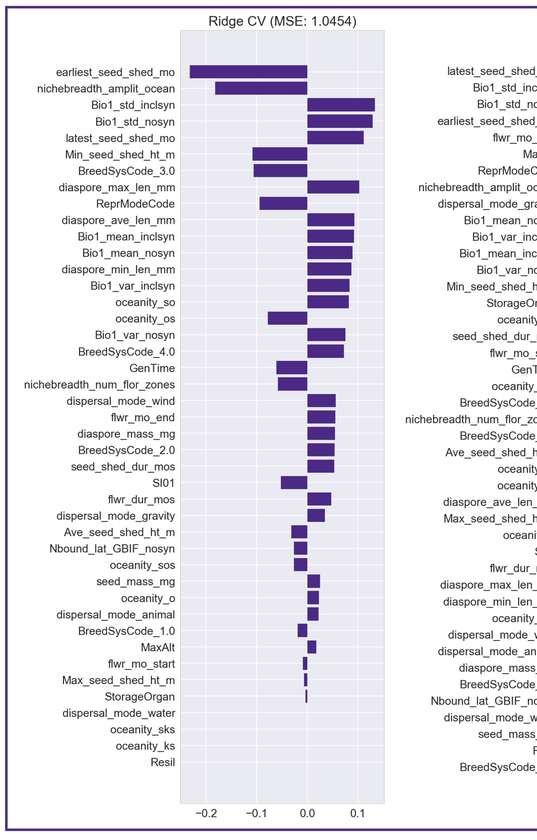
Implementation:

- We remove samples (i.e., plant species) which are missing any feature data, which brings us to $N = 20$.
- We one-hot encode categorical features, $d = 43$.
- We center the numeric features with 0 mean and unit variance for regression.

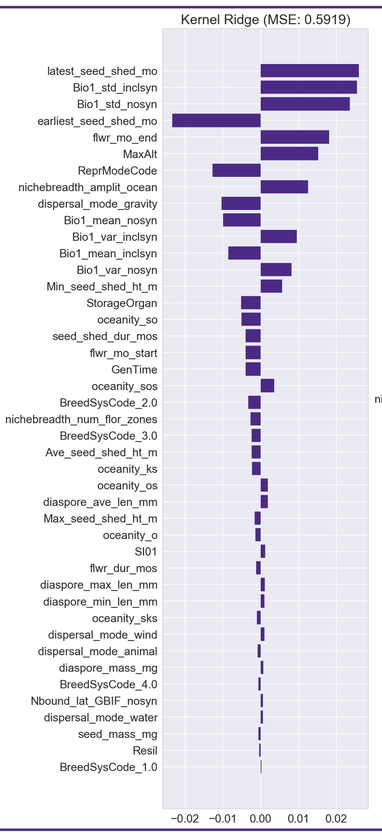
References:
Angert, A. et al., 2011. "Do species' traits predict recent shifts at expanding range edges?" *Ecology Letters*, **14**: 677–689.
Holzinger, B. et al., 2008. "Changes in plant species richness over the last century in the eastern Swiss Alps: elevational gradient, bedrock effects and migration rates." *Plant Ecol.*, **195**, 179–196.
Lundberg, S. and Lee, S., 2017. "A unified approach to interpreting model predictions" *NIPS 2017. arXiv:1705.07874*



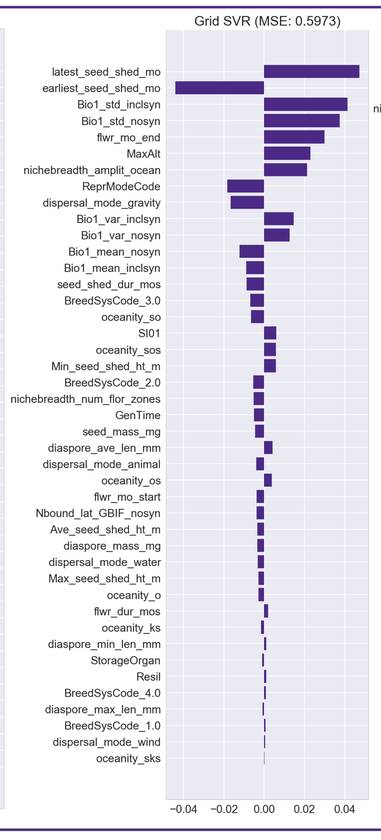
OLS Coefficients



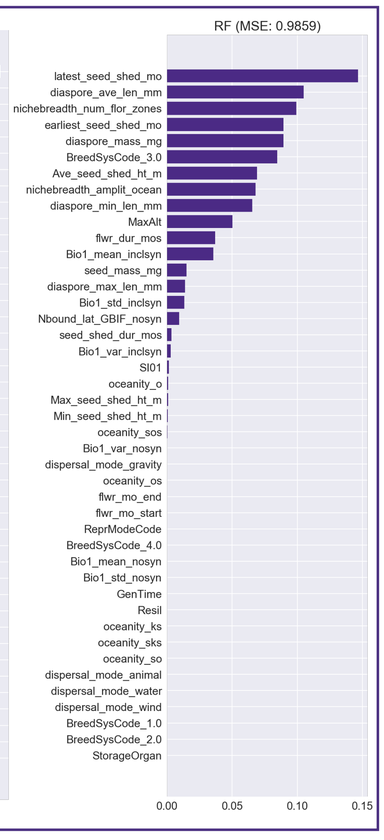
Ridge Regression Coefs.



Kernel Regression SHAP



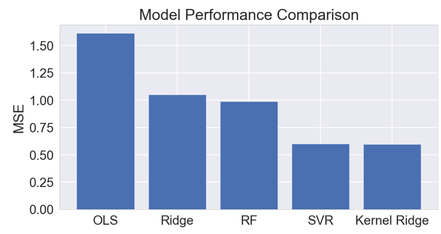
SVM SHAP



Random Forest Gini Scores

	OLS	Ridge	Kernel Ridge	RF	SVR	mean
earliest_seed_shed_mo	0.0	0.0	3.0	3.0	1.0	1.4
Bio1_std_inclsyn	3.0	2.0	1.0	14.0	2.0	4.4
nichebreadth_amplit_ocean	1.0	1.0	7.0	7.0	6.0	4.4
latest_seed_shed_mo	21.0	4.0	0.0	0.0	0.0	5.0
Bio1_std_nosyn	5.0	3.0	2.0	31.0	3.0	8.8
ReprModeCode	2.0	8.0	6.0	28.0	7.0	10.2
BreedSysCode_3.0	7.0	6.0	22.0	5.0	14.0	10.8
Bio1_mean_inclsyn	11.0	10.0	11.0	11.0	12.0	11.0
MaxAlt	9.0	35.0	5.0	9.0	5.0	12.6
Bio1_var_inclsyn	18.0	13.0	10.0	17.0	9.0	13.4
Bio1_mean_nosyn	12.0	11.0	9.0	30.0	11.0	14.6
Min_seed_shed_ht_m	17.0	5.0	13.0	21.0	18.0	14.8
diaspore_ave_len_mm	16.0	9.0	26.0	1.0	23.0	15.0
seed_shed_dur_mos	14.0	24.0	16.0	16.0	13.0	16.6
Bio1_var_nosyn	22.0	16.0	12.0	23.0	10.0	16.6
flwr_mo_end	31.0	21.0	4.0	26.0	4.0	17.2
oceanity_so	8.0	14.0	15.0	36.0	15.0	17.6
dispersal_mode_gravity	23.0	27.0	8.0	24.0	8.0	18.0
SI01	6.0	25.0	29.0	18.0	16.0	18.8
nichebreadth_num_flor_zones	34.0	19.0	21.0	2.0	20.0	19.2
diaspore_max_len_mm	10.0	7.0	31.0	13.0	39.0	20.0
diaspore_min_len_mm	19.0	12.0	32.0	8.0	35.0	21.2
GenTime	20.0	18.0	18.0	32.0	21.0	21.8
Ave_seed_shed_ht_m	25.0	28.0	23.0	6.0	28.0	22.0
oceanity_os	24.0	15.0	25.0	25.0	25.0	22.8
BreedSysCode_2.0	13.0	23.0	20.0	41.0	19.0	23.2
diaspore_mass_mg	30.0	22.0	36.0	4.0	29.0	24.2
oceanity_sos	36.0	30.0	19.0	22.0	17.0	24.8
BreedSysCode_4.0	4.0	17.0	37.0	29.0	38.0	25.0
Max_seed_shed_ht_m	15.0	37.0	27.0	20.0	31.0	26.0
seed_mass_mg	29.0	31.0	40.0	12.0	22.0	26.8
flwr_dur_mos	37.0	26.0	30.0	10.0	33.0	27.2
oceanity_o	27.0	32.0	28.0	19.0	32.0	27.6
flwr_mo_start	33.0	36.0	17.0	27.0	26.0	27.8
Nbound_lat_GBIF_nosyn	38.0	29.0	38.0	15.0	27.0	29.4
dispersal_mode_animal	28.0	33.0	35.0	37.0	24.0	31.4
dispersal_mode_wind	26.0	20.0	34.0	39.0	41.0	32.0
StorageOrgan	32.0	38.0	14.0	42.0	36.0	32.4
oceanity_ks	42.0	41.0	24.0	34.0	34.0	35.0
dispersal_mode_water	41.0	39.0	39.0	38.0	30.0	37.4
oceanity_sks	40.0	40.0	33.0	35.0	42.0	38.0
BreedSysCode_1.0	35.0	34.0	42.0	40.0	40.0	38.2
Resil	39.0	42.0	41.0	33.0	37.0	38.4

Feature Importance Rank per Method, and Averaged Across Methods.



Results | Discussion | Conclusions:

- Across these methods, **dispersal-related traits** like `{latest_seed_shed_mo, earliest_seed_shed_mo}` consistently rank highest in importance across these methods, as measured by the absolute value of the coefficients. For all methods except for the Ridge and OLS, `latest_seed_shed_mo` across was the variable with the highest importance.
- These results are consistent with Angert et al.'s 2011 analysis, wherein seed dispersal period duration was the most important predictor variable.
- This result indicates that the predictive performance gains (see MSE figure above) from the SVM/Kernel Ridge regression maintain their theoretical biological underpinnings while better representing nonlinearities in the relationship between species range shift and trait values.

UNIVERSITY of WASHINGTON

This work is supported in part by an NSF IGERT grant DGE-1258485.

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS