MAIN MODELING ALGORITHM

Elena Razenkova

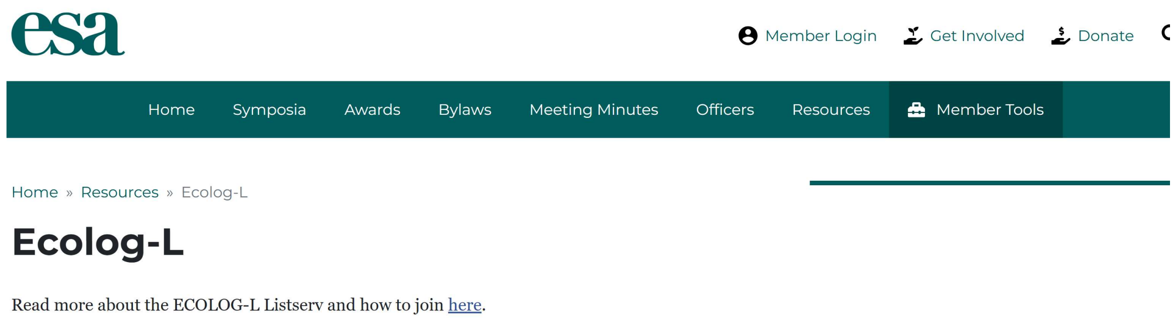
Karaganda, Kazakhstan August 2025

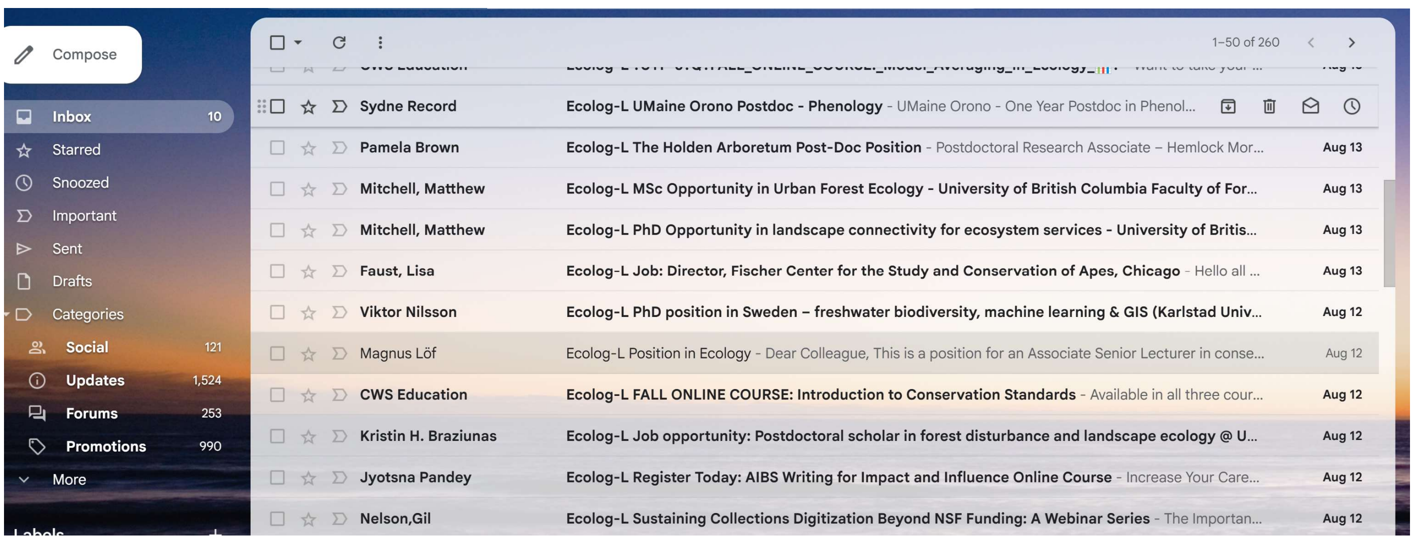
PLAN FOR TODAY

- Some of commonly used methods
- Linear regression
- Generalized linear regression
- Logistic regression
- R code

WHERE TO FIND POSITION

https://esa.org/theory/resources/ecolog-l/





MODELING ALGORITHMS

Statistical methods:

- Generalized Linear Models (GLMs)
- Generalized additive models (GAMs)

Machine-learning techniques:

- Maximum entropy (Maxent)
- Random Forest
- Boosted Regression Trees (BRT)

LINEAR REGRESSION

Effective load (x):	1.72	1.72	1.77	1.78	1.82	1.85	1.88
Terminal velocity(y):	0.85	0.86	0.72	0.79	0.82	0.80	0.99
Effective load (x):	1.93	1.96	1.96	2.00	2.00	2.03	2.06
Terminal velocity (y):	0.94	0.82	0.89	0.95	1.00	0.98	0.99

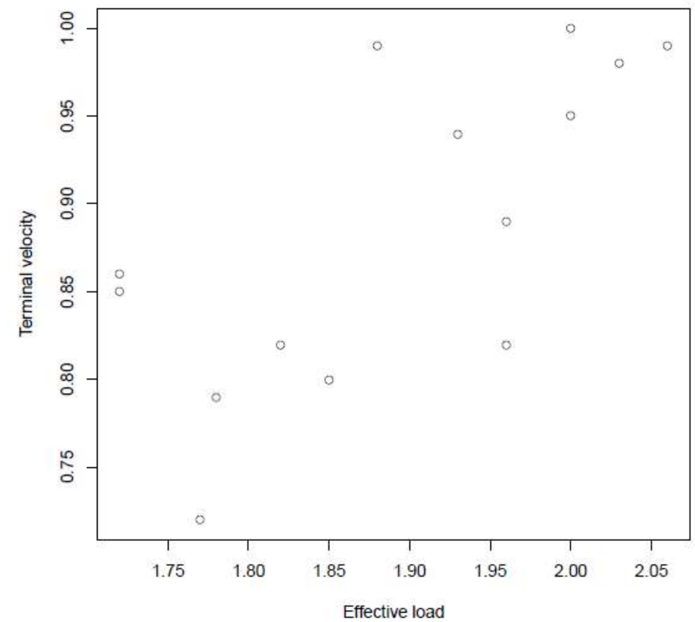


Figure 12.1 *Scatter plot of terminal velocity versus effective load of 14 maple samaras.*

LINEAR REGRESSION

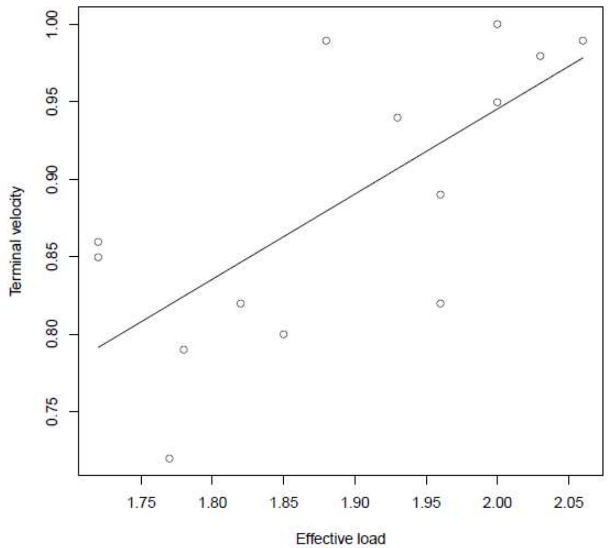
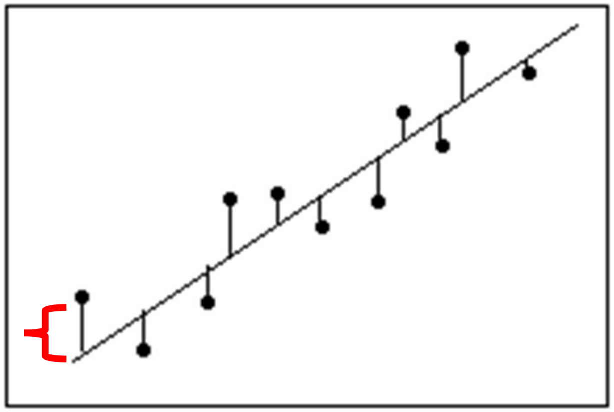


Figure 12.2 *Fitted line superimposed over the scatter plot of terminal velocity versus effective load of 14 maple samaras.*

$$y = \beta_0 + \beta_1 x$$

To estimate the best fitting line:

To determine the estimates of the intercept and the slope that lead to the least squared line



Minimize the distance between each point and line

LINEAR REGRESSION

Response variable
(dependent variable)

Explanatory variable
(independent variable)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

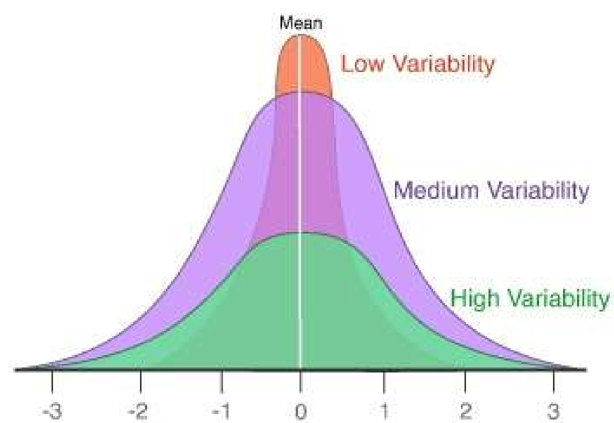
The errors

The y-intercept

The slope of the line

Assumptions:

- The data come from a straight-line model.
- The errors are independent.
- The errors have constant variance.
- The errors follow a normal distribution.



TERMINOLOGY

- The term "general" linear model (GLM) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors.

- The term "generalized" linear model (GLMs) refers to a larger class of model estimated

GENERALIZED LINEAR REGRESSION

Response variable (dependent variable)

Link function

Explanatory variable (independent variable)

$$Y_i = g(\beta_0 + \beta_1 * x_i + \varepsilon_i)$$

The errors

The y-intercept

The slope of the line

Assumptions:

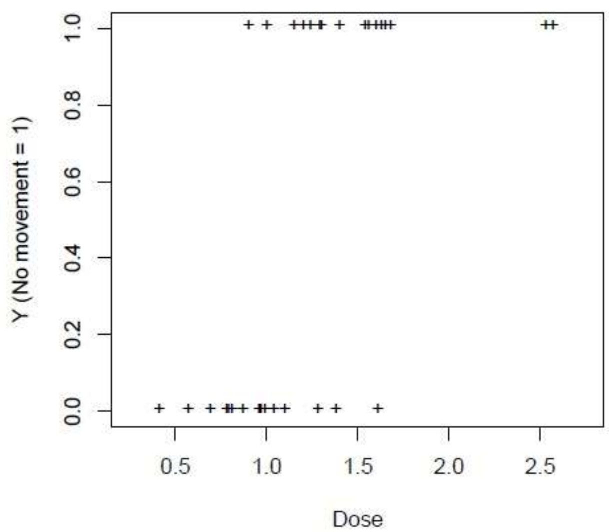
- The errors follow a non-normal distribution.
- Link function connecting the predictors to the response variables
- The errors are independent.

GLMs typically use maximum likelihood to estimate model parameters

LOGISTIC REGRESSION

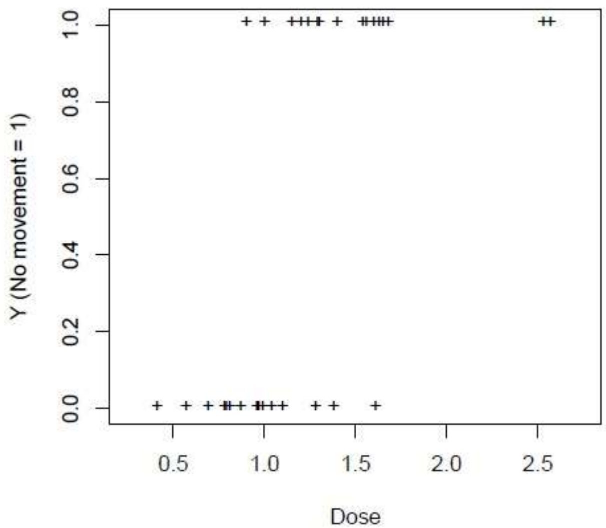
X	0.5	0.6	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1.0	1.0
Gender	1	1	0	0	1	1	1	1	1	0	1
Y	0	0	0	0	0	0	0	0	1	0	0
X	1.0	1.0	1.0	1.2	1.2	1.2	1.2	1.2	1.2	1.4	1.4
Gender	1	0	0	1	0	1	1	0	1	1	1
Y	0	0	1	0	0	1	1	1	1	0	0
X	1.4	1.4	1.4	1.4	1.6	1.6	1.6	1.6	2.5	2.5	
Gender	0	0	1	0	0	1	0	0	1	0	
Y	1	1	1	1	1	1	1	1	1	1	

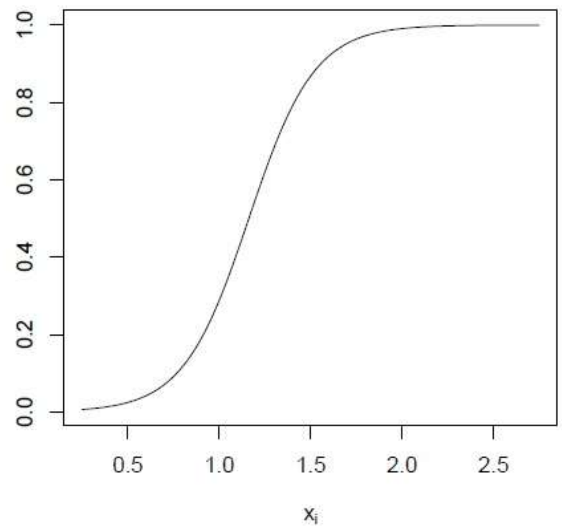
$$Y = \begin{cases} 1 \\ 0 \end{cases}$$



- The probability of a success is constant, and equal to P
- The probability of a failure is 1- P

LOGISTIC REGRESSION





$$\mathrm{P_i} = \beta_0 + \beta_1 * x_i$$

$$Y = \begin{cases} 1 \\ 0 \end{cases}$$

Logit of P or "logit-transformation" of P

- Odds – ration of the probability of success (Y=1) to the probability of failure (Y=0)

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 * x_i$$

MODELING ALGORITHMS

Methods	Software Name	Species data type	Pros/Cons	Reference
Generalized Linear Models (GLMs)	R	P/A	Pros: easy to use, easy to incorporate multiple predictors	Guisan et al.(2002)
Generalized additive models (GAMs)	R	P/A	Pros: powerful for complex ecological responses, high predictive accuracy Cons: no interaction terms, use data defined Smoothing functions to fit nonlinear species-environment relationships	Guisan et al.(2002)

MODELING ALGORITHMS

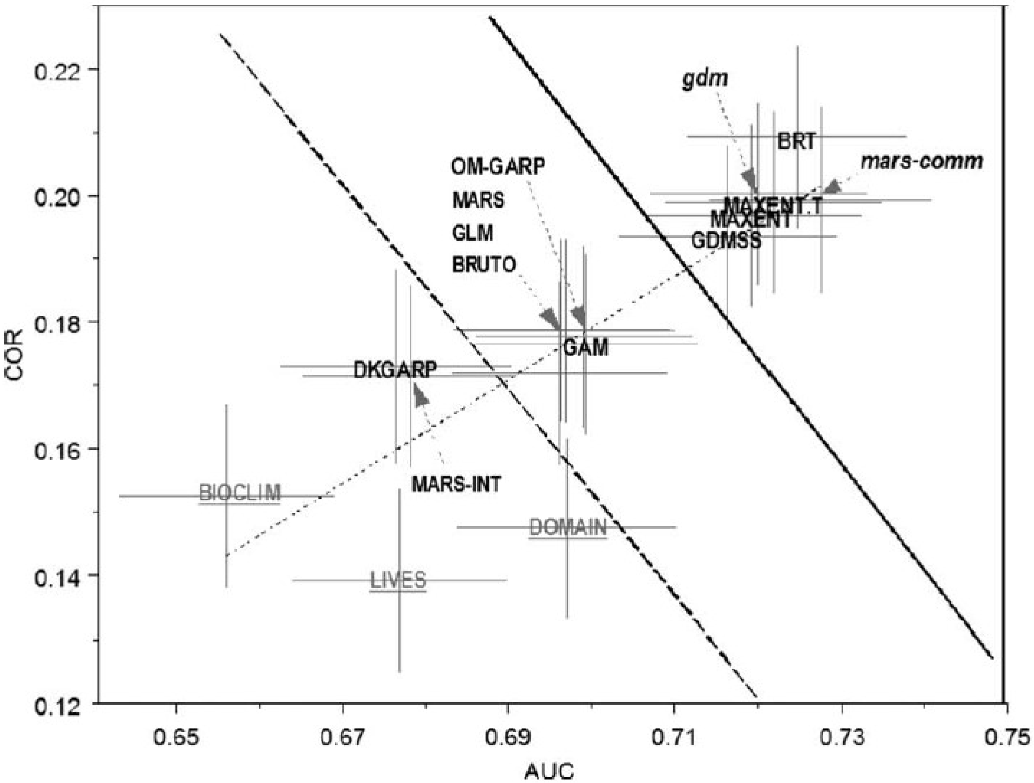
Methods	Software Name	Species data type	Pros/Cons	Reference
Multivariate Adaptive Regression Splines (MARS)	R	P/A	Pros: computationally fast, can include a lot of predictors, includes interactions Cons: (from Franklin 2009) only continuous predictors, requires normally distributed errors	Friedman(1991)
Boosted Regression Trees (BRT)	R	P/A or P/B	Pros: reliable predictive power, good at explaining variance	Elith et al. (2008)
Maximum Entropy	MAXENT	P/B	Pros: easy to use, widely used, robust with small sample size Cons: Sometimes used incorrectly/people think that the output is better that it actually is	Phillips et al.(2006)

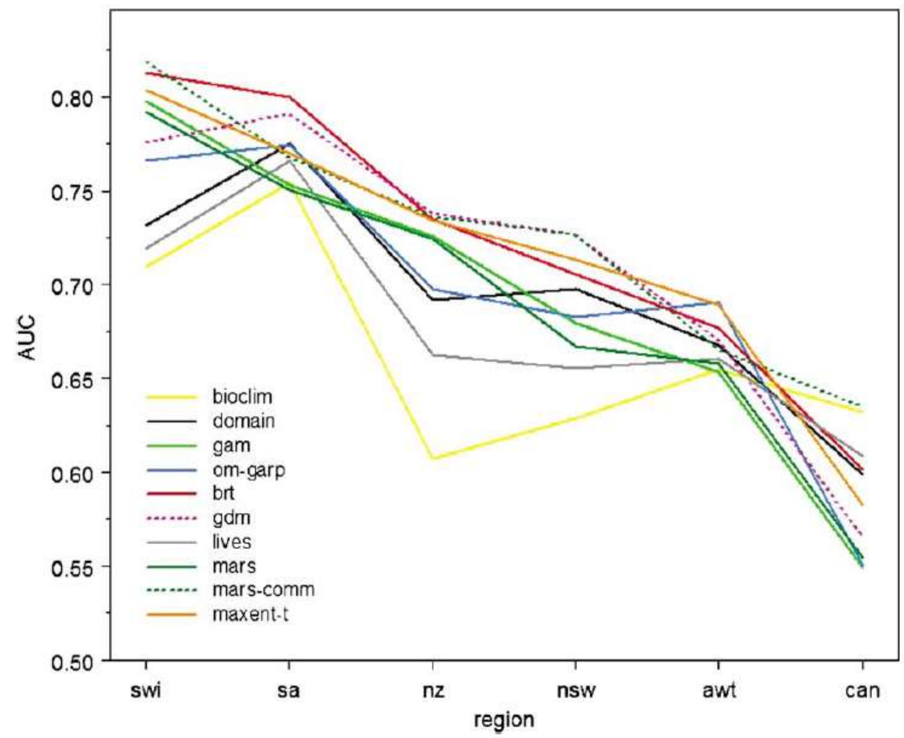
Table 4. Modelling methods implemented.

Method	Class of model, and explanation	Data[1]	Software
BIOCLIM	envelope model	p	DIVA-GIS
BRT	boosted decision trees	pa	R, gbm package
BRUTO	regression, a fast implementation of a gam	pa	R and Splus, mda package
DK-GARP	rule sets from genetic algorithms; desktop version	pa	DesktopGarp
DOMAIN	multivariate distance	p	DIVA-GIS
GAM	regression: generalised additive model	pa	S-Plus, GRASP add-on
GDM	generalised dissimilarity modelling; uses community data	pacomm	Specialized program not general released; uses Arcview and Splus
GDM-SS	generalised dissimilarity modelling; implementation for single species	pa	as for GDM
GLM	regression; generalised linear model	pa	S-Plus, GRASP add-on
LIVES	multivariate distance	p	Specialized program not general released
MARS	regression; multivariate adaptive regression splines	pa	R, mda package plus new code to handle binomial responses
MARS-COMM	as for MARS, but implemented with community data	pacomm	as for MARS
MARS-INT	as or MARS; interactions allowed	pa	as for MARS
MAXENT	maximum entropy	pa	Maxent
MAXENT-T	maximum entropy with threshold features	pa	Maxent
OM-GARP	rule sets derived with genetic algorithms; open modeller version	pa	new version of GARP not yet available

Elith et al.2006

MODELING ALGORITHMS





Elith et al.2006

THANK YOU

Elena Razenkova

erazenkova@gmail.com