Mathematical Modelling of Species Differentiation

Many closely related animal species are hard to distinguish from each other by their exterior. Sometimes "key features" may be found (e.g., some special colouring), but often the biologists have to rely upon sets of measurable characteristics.

You are provided with a set of real data containing the measurements of 564 lizards of 8 species belonging to genus *Darevskia*. The measurements consist of the counts of scales on different parts of the lizards' bodies (pholidosis characteristics) and linear sizes of the lizards' body parts (morphometric characteristics). For each lizard, its number-coded biological species and sex are given.

You are required to develop criteria that enable one to predict biological species and sex of the lizards with the best possible accuracy on the basis of such measurements. These criteria must be relatively simple and obvious, i.e. be realistically calculatable by a biologist in field conditions using nothing more than an engineering calculator. Solution of the computational "black box" type (e.g., artificial neural network classifiers or any other "closed" program that does not explain its "internal logic") are not accepted. It is simple and obvious criteria that may enable the biologists to construct explanatory theories.

Tasks

1. Build a criterion that, with the best possible accuracy, differentiates lizards of species #5 from all other lizards and uses only the femoral pore number on the right side (FPNr).
   Hint: plot and explore the FPNr distribution of the lizards with regard to their species.

2. Build a criterion that, with the best possible accuracy, differentiates lizards of species #5 from all other lizards and uses two variables out of the measured morphometric and pholidosis characteristics.
   Hint: one of the methods to find the best pair of predicting variables (predictors) may be exhaustive search over all possible pairs of variables.

3. Build a criterion that, with the best possible accuracy, predicts lizards' sex regardless of their biological species on the basis of the morphometric and/or pholidosis characteristics.
   Hint from the biologists: it is expected (but not guaranteed!) that sex is correlated with the ratios of some measured linear sizes; but this does preclude from using other predictors in the criterion.

4. Not all lizard species in question are found in the same locations. Therefore, in practice, the tasks of distinguishing species from certain subgroups that live together are most often encountered. Build a set of criteria that, with the best possible accuracy, differentiate all species within the following groups:

   a. species #6 and #7,

   b. species #1 and #2,

   c. species #3, #4, and #5.

5. Build a criterion or a set of criteria that, with the best possible accuracy, predict lizards' species or species and sex in their entire population (this may be useful to the biologists if they don't know the locality where the lizard was captured).

It is quite possible that some pairs or groups of species will be inseparable on the basis of the available data. Provide the best obtained result that may be the most helpful to the biologists.

General Requirement

All your criteria must be accompanied by their performance metrics, i.e. the numbers of correctly and incorrectly classified lizards (preferably with a breakdown by true classes). For example, in Task 1, the "full" metrics is contained in the following table:

	True species #5	**True species #1-4, 6-8**
Classified as species #5	a	b
Classified as species #1-4, 6-8	c	d

Correctly classified are a and d; incorrectly classified (classification errors) are b and c.

Remarks

- **Remark 1.** Neither of the Tasks is strictly mandatory. Although the higher number of successfully solved Tasks improves the work's score.

- **Remark 2.** The Tasks are ordered by their expected difficulty. If you create a method to solve a "hard" Task, it will likely enable you to solve all previous Tasks with little effort.

- **Remark 3.** The problem is a real applied research problem with real data. Thus, the "perfect" solution may not exist at all. In that case, the "best" solution is the "least bad" one.

- **Remark 4.** An example of a "simple criterion": an explicitly specified function f of one or several measured variables $p_1, p_2, \dots$ and a condition like "if $f(p_1, p_2, \dots) > h$, then the lizard belongs to that class, otherwise it belongs to another class". Function f must be "calculatable on an engineering calculator", i.e. it may contain "standard" functions (power functions, trigonometric functions, exponents, logarithms, etc.), but it cannot contain a hidden iterative algorithm or calculations in volume beyond the capabilities of manual implementation.

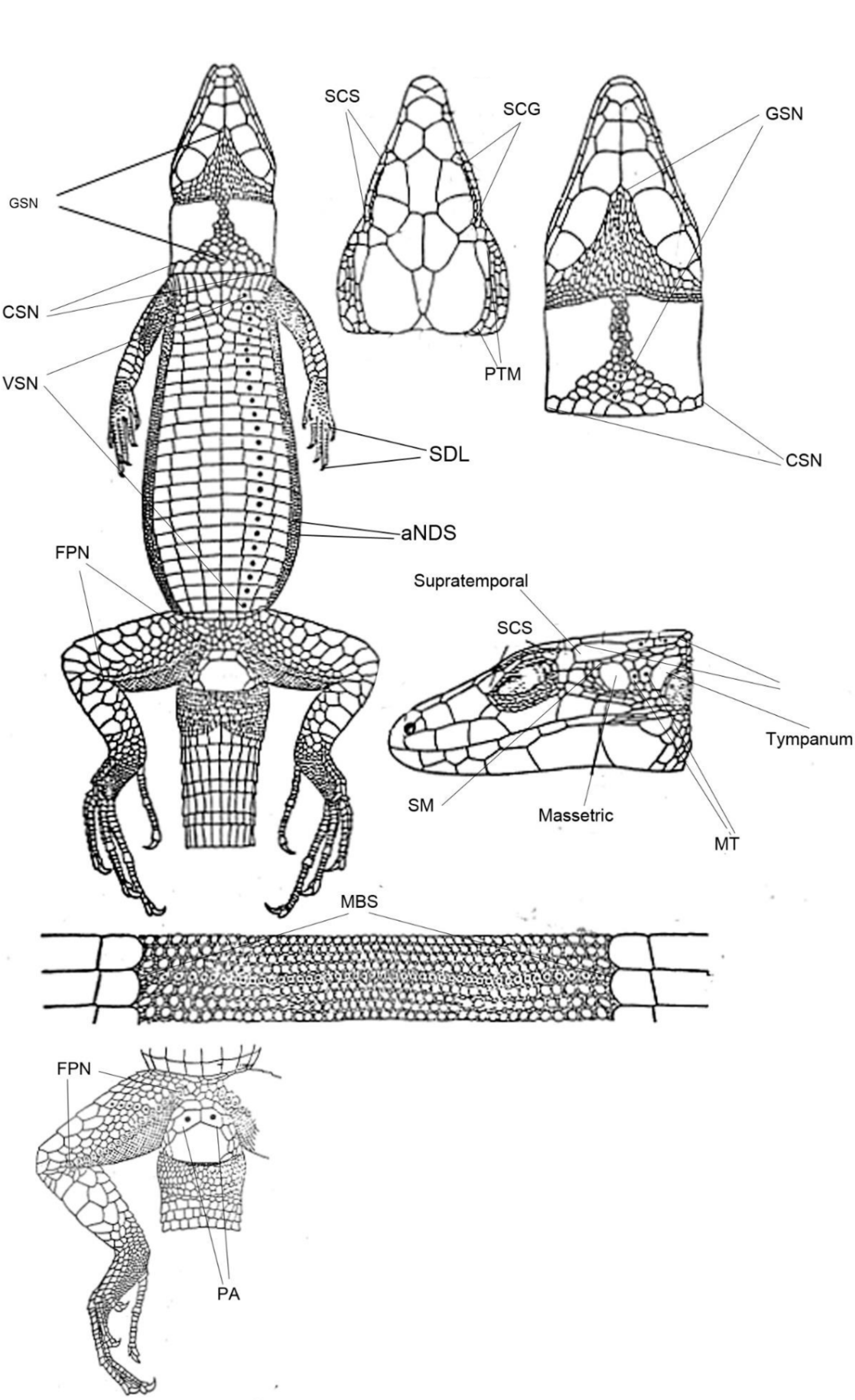
Data Description

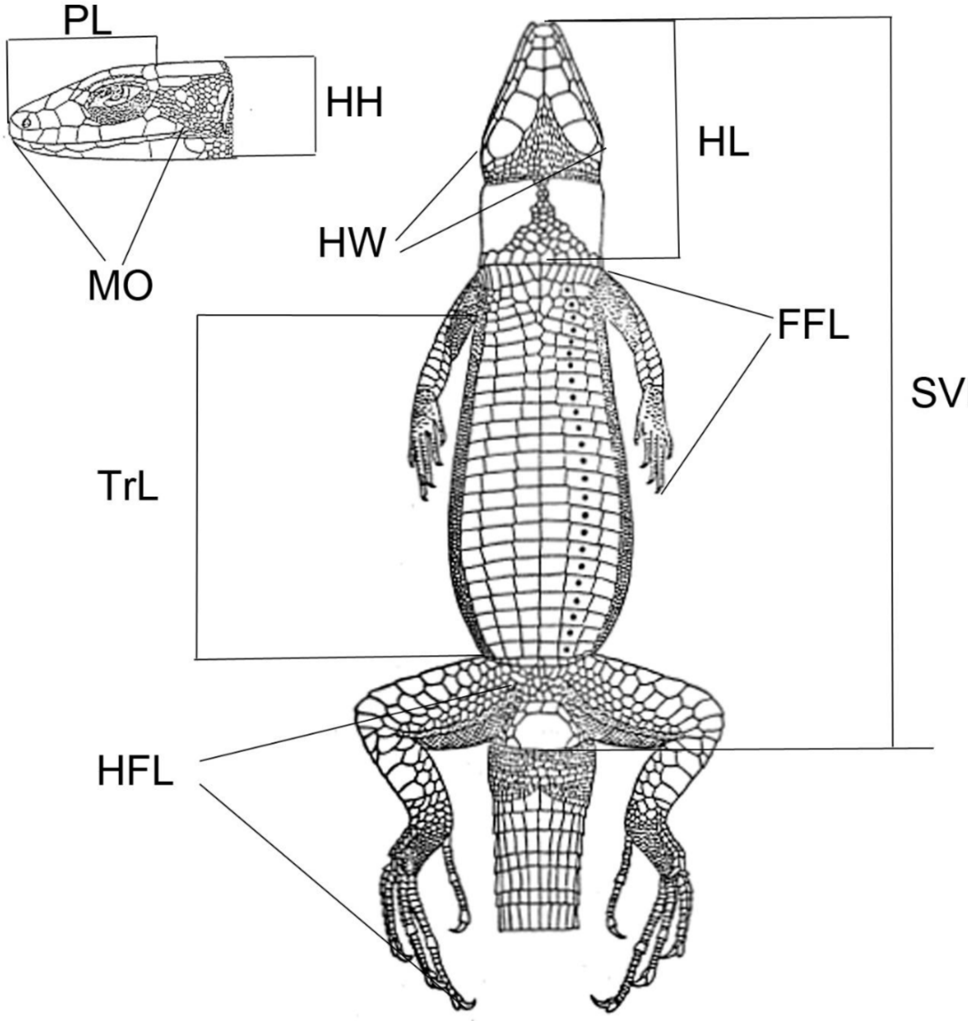
The supplied XLSX- and CSV-files contain the following columns:

- **Species_num** – number-coded species, integer from 1 to 8;

- **Sex_num** – number-coded sex, 1 = male, 2 = female;

- **Sex** – letter-coded sex, M = male, F = female;


SCS
SCG
GSN
GSN
CSN
PTM
VSN
SDL
CSN
aNDS
FPN
Supratemporal
SCS
Tympanum
SM
Massetric
MT
MBS
FPN
PA



Pholidosis characteristics (counts of scales):

1. **MBS** – medium body scales, number of dorsal scales, approximately at half trunk;
2. **VSN** – ventral scale number on the middle line;
3. **CSN** – collar scale number;
4. **GSN** – gular scale number from the angle between the maxillar scales to the collar;
5. **FPN** – femoral pore number (FPNr – FPN on the right side);
6. **SDL** – subdigital lamellae in the 4th toe of the forelimb (SDLr – SDL on the right forelimb);
7. **SCS** – number of superciliary scales (SCSr – SCS on the right);
8. **SCG** – number of superciliary granules (SCGr – SCG on the right);
9. **SM** – number of scales between the masseteric shield and the supratemporal scale (SMr – SM on the right);
10. **MT** – number of scales between masseteric and tympanum shields on the right (MTr – MT on the right);

11. **PA** – preanal scale number;
12. **PTM** – postemporal scale number (PTMr – PTM on the right);
13. **aNDS** – average number of dorsal scales along one abdomen scale near limb (aNDSr – aNDS on the right);

Morphometric characteristics (all lengths are in millimetres):

14. **SVL** – snout-vent length, length of the body from tip of snout to cloaca;
15. **TRL** – trunk length (from the groin to the armpit);
16. **HL** – head length, measured ventrally from the tip of the snout to the posterior margin of the collar;
17. **PL** – pileus length measured dorsally from the tip of the snout to the posterior margin of the parietal + occipital scales;
18. **ESD** – length of the posterior half of the pileus, measured from the anterior margin of the 3rd supraocular scale to the posterior margin of the parietal + occipital scales;
19. **HW** – width of the head before the tympanic hole;
20. **HH** – head height near the occipital plate;
21. **MO** – mouth opening, measured laterally from the tip of the snout to the end of the mouth;
22. **FFL** – total forelimb length, from the base to the tip of the longest toe;
23. **HFL** – total hindlimb length, from the base to the tip of the longest toe.