III-й Международный командный турнир школьников
по математическому моделированию (ТММ-2020)

Задание МаМоНТ-2020

Многие близкородственные виды живых существ довольно трудно отличить друг от друга по общему внешнему виду. Иногда можно найти какие-то «ключевые признаки» — например, особый окрас, — но зачастую биологам приходится полагаться на комплексы измеримых признаков.

Вам предоставлен массив реальных данных, содержащий измерения 564 ящериц, принадлежащих к восьми видам рода *Darevskia*. Измерения содержат подсчёт количества чешуй на разных частях тела ящериц (признаки фолидоза) и измерения некоторых линейных размеров ящериц (морфометрические признаки). Для каждой ящерицы указан условный номер её биологического вида и пол.

Вам необходимо разработать критерии, позволяющие на основании таких измерений наилучшим возможным образом предсказать биологический вид и пол ящериц. Эти критерии должны быть сравнительно простыми и наглядными, т.е. реалистично вычисляемыми биологом в полевых условиях с использованием, в лучшем случае, инженерного калькулятора. Решения, имеющие вид программного «чёрного ящика» (например, классификатор на основе искусственных нейронных сетей или любая другая «закрытая» программа, не объясняющая своей «внутренней логики») не принимаются. Именно простые и наглядные критерии могут позволить биологам строить объясняющие теории.

Задания:

1) Создайте критерий, позволяющий наилучшим возможным образом отличать ящериц вида №5 от всех остальных ящериц и использующий только количество бедренных пор справа (FPNr).
   Подсказка: постройте и изучите распределение ящериц по FPNr в зависимости от их вида.

2) Создайте критерий, позволяющий наилучшим возможным образом отличать ящериц вида №5 от всех остальных ящериц и использующий две переменных из измеряемых морфометрических и фолидозных признаков.
   Подсказка: одним из способов нахождения наилучшей пары предсказывающих переменных (предикторов) может быть перебор всех возможных пар переменных.

3) Создайте критерий, позволяющий наилучшим возможным образом предсказывать пол ящериц вне зависимости от их вида по морфометрическим признакам и/или признакам фолидоза.
   Подсказка от биологов: предполагается (но не гарантируется!), что пол будет взаимосвязан с отношениями некоторых измеряемых длин; но это не исключает участия в критерии и других предикторов.

4) Не все рассматриваемые виды ящериц встречаются в одних и тех же местах обитания. Поэтому на практике чаще всего встречаются задачи различения видов из определённых подгрупп, обитающих совместно. Создайте набор критериев, позволяющих наилучшим возможным образом отличать друг от друга все виды внутри следующих групп:
   a. виды №6 и №7,
   b. виды №1 и №2,
   c. виды №3, №4 и №5.

5) Создайте критерий или набор критериев, позволяющий наилучшим возможным образом предсказывать вид или вид и пол ящериц во всей их совокупности (это может понадобиться биологам, если они не знают место отлова ящерицы).
   Вполне возможно, что некоторые пары или группы видов не будут разделяться на основе имеющихся данных. Приведите наилучший полученный результат, который может в наибольшей степени помочь биологам.

Общее требование. Ко всем построенным вами критериям должны быть указаны показатели качества их работы, т.е. количества правильно и неправильно классифицированных ящериц (желательно с разбивкой по истинным классам). Например, для Задания №1 в «полном» варианте вы должны заполнить числами ячейки *a*, *b*, *c* и *d* в таблице вида

	На самом деле вид №5	На самом деле виды №1-4, 6-8
Классифицирован как вид №5	*a*	*b*
Классифицирован как вид №1-4, 6-8	*c*	*d*

Правильно классифицированные – это *a* и *d*, а *b* и *c* – это ошибочно классифицированные.

Замечание 1. Ни один пункт задания не является строго обязательным. Однако большее количество успешно решённых пунктов улучшает оценку работы.

Замечание 2. Пункты задания идут в порядке возрастания ожидаемой сложности. Если вы разработаете метод решения «сложного» пункта, это, скорее всего, позволит вам практически без усилий решить все предыдущие пункты.

Замечание 3. Предлагаемая задача является реальной прикладной исследовательской задачей с реальными данными. Это означает, что «идеальное» решение может вообще не существовать. В таком случае «наилучшим» является «наименее плохое» решение.

Замечание 4. Пример «простого критерия»: некоторая явно заданная функция *f* от одной или нескольких переменных-измерений p_1, p_2, … и условие вида «если $f(p_1, p_2, \ldots) > h$, то ящерица принадлежит к такому-то классу, иначе – к другому классу». Функция *f* должна быть «вычисляема на инженерном калькуляторе», т.е. в ней могут присутствовать «стандартные» функции (возведение в степень, тригонометрические функции, показательная функция, логарифм и т.п.), но не может быть скрытого итеративного алгоритма или вычислений в объеме, недоступном для реализации вручную.

Описание данных:

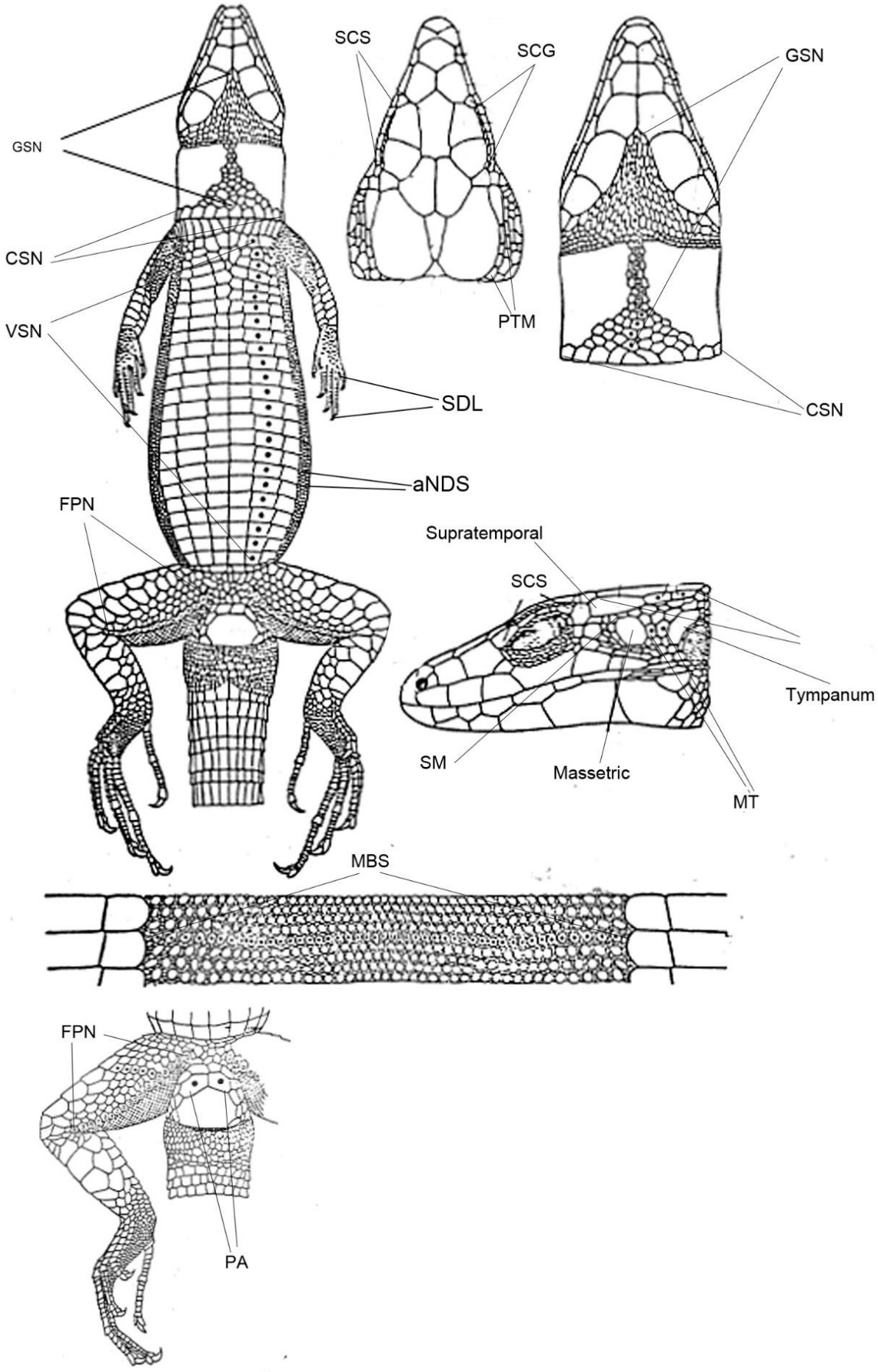
В прилагаемых XLSX- и CSV-файлах содержатся следующие столбцы:

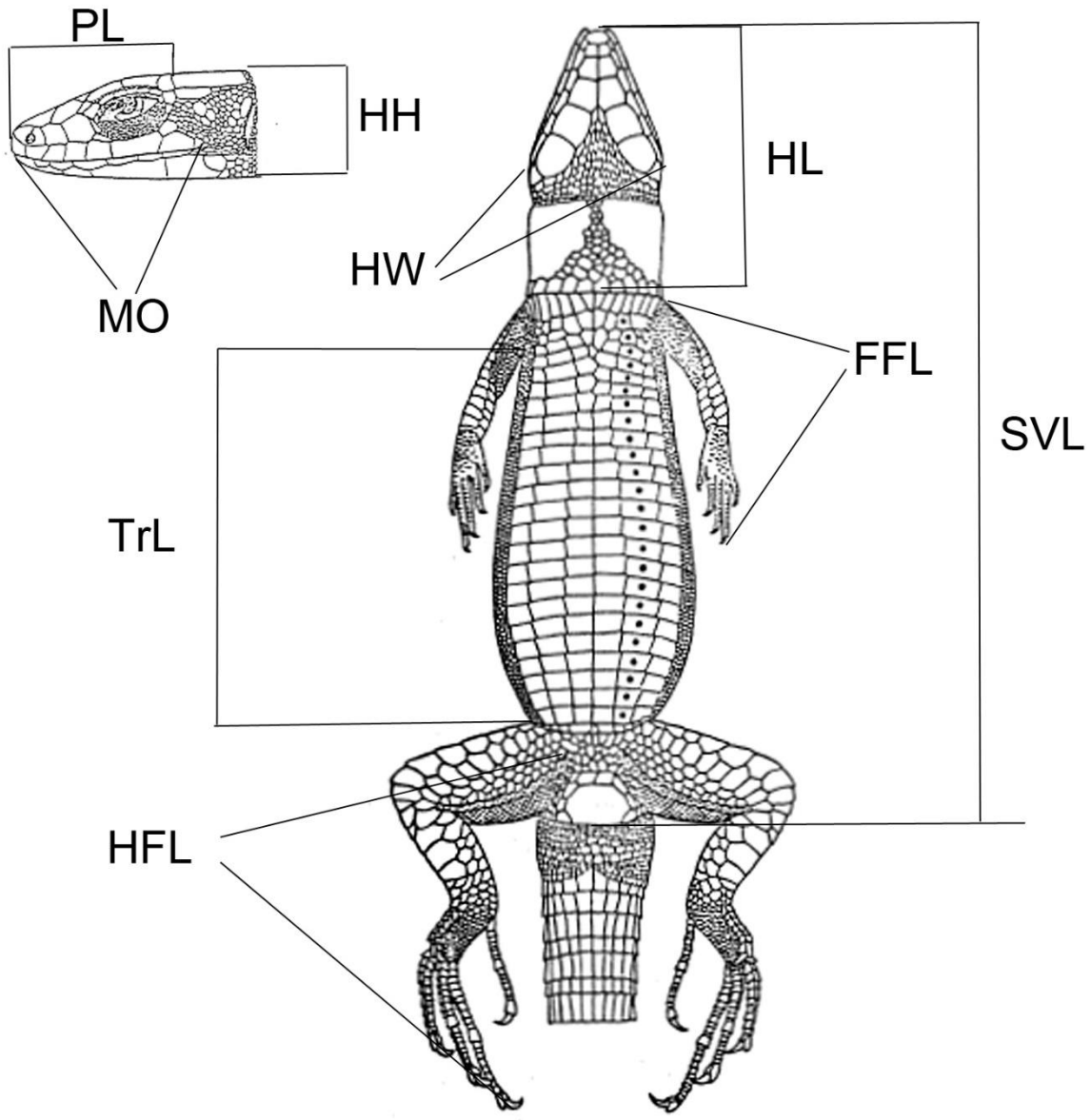
Species_num – номер вида, целое от 1 до 8;

Sex_num – номер пола, 1=муж., 2=жен.;

Sex – буквенное обозначение пола, M=муж., F=жен.;

Остальные столбцы являются измеренными признаками ящериц и описаны ниже:


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



Признаки фолидоза (подсчет чешуй на теле):

1. **MBS** - Количество чешуй вокруг середины тела / medium body scales, number of dorsal scales, approximately at half trunk;
2. **VSN** - Количество брюшных чешуй / ventral scale number on the middle line;
3. **CSN** - Количество чешуй на воротнике / collar scale number;
4. **GSN** - Количество чешуй по средней линии горла до воротника / gular scale number from the angle between the maxillar scales to the collar;
5. **FPN** - Число бедренных пор / femoral pore number) (**FPNr** – FPN справа / FPN on the right side);
6. **SDL** - Количество чешуй под 4-м пальцем передней лапы / subdigital lamellae in the 4th toe of the forelimb (**SDLr** – SDL справа / SDL on the right forelimb);
7. **SCS** - Количество верхнересничных чешуй / number of superciliary scales (**SCSr** – SCS справа / SCS on the right);
8. **SCG** - Количество верхнересничных зернышек / number of superciliary granules (**SCGr** – SCG справа / SCG on the right);

9. **SM** - Количество рядов чешуй между центральновисочным и верхневисочным щитками / number of scales between the masseteric shield and the supratemporal scale (**SMr** – SM справа / SM on the right);
10. **MT** - Количество рядов чешуй между центральновисочным и барабанным щитками / number of scales between masseteric and tympanum shields on the right (**MTr –** MT справа / MT on the right);
11. **PA** - Количество крупных прианальных чешуй / preanal scale number;
12. **PTM** - Количество задневисочных чешуй / postemporal scale number (**PTMr –** PTM справа / PTM on the right);
13. **aNDS** – Среднее количество спинных чешуй вдоль края брюшной чешуи / average number of dorsal scales along one abdomen scale near limb (**aNDSr –** aNDS справа / aNDS on the right);

Морфометрические признаки (все длины даны в миллиметрах):

14. **SVL** - Длина тела от клоаки до кончика морды / snout-vent length, length of the body from tip of snout to cloaca;
15. **TRL** - Длина туловища (от паха до подмышки) / trunk length (from the groin to the armpit);
16. **HL** - Длина головы измеренная от воротника до кончика морды / head length, measured ventrally from the tip of the snout to the posterior margin of the collar;
17. **PL** - Длина головного щита от заднего края теменных щитков до кончика морды / pileus length measured dorsally from the tip of the snout to the posterior margin of the parietal + occipital scales;
18. **ESD** - Длина задней части головного щита от задней границы теменных щитков до передней границы 3-го надглазничного щитка / 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** - Ширина головы перед тимпанальным отверстием / width of the head before the tympanic hole;
20. **HH** - Высота головы за глазами / head height near the occipital plate;
21. **MO** - Длина рта / mouth opening, measured laterally from the tip of the snout to the end of the mouth;
22. **FFL** - Длина передней конечности от основания до 4 пальца / total forelimb length, from the base to the tip of the longest toe;
23. **HFL** - Длина всей задней конечности / total hindlimb length, from the base to the tip of the longest toe.