

Научная статья
УДК 581.9+58.08
DOI: 10.18101/2542-0623-2022-4-96-104

ОСОБЕННОСТИ РАБОТЫ С БАЗАМИ ДАННЫХ ПО РАСПРОСТРАНЕНИЮ РАСТЕНИЙ И ОПЫТ КОНСОЛИДАЦИИ ДАННЫХ РАЗЛИЧНОГО ФОРМАТА

Д. В. Санданов

© Санданов Денис Викторович

кандидат биологических наук,
Институт общей и экспериментальной биологии СО РАН
Россия, 670047, г. Улан-Удэ, ул. Сахьяновой, 6
Тункинский национальный парк
Россия, 671010, с. Кырен, ул. Ленина, 130
sdenis1178@mail.ru

Аннотация. Работа с базами данных и большими массивами информации имеет определенную специфику. В последние годы нами проводятся исследования по консолидации различных данных по распространению растений для оценки разнообразия и видового богатства, а также их использования для моделирования экологических ареалов видов. Подведены итоги интеграции базы данных по распространению сосудистых растений Азиатской России и рассмотрены особенности работы с различными наборами данных. Приведены сведения о возможности оцифровки карт ареалов растений с использованием ГИС-программ. Отмечаются важность учета разнородности данных и оценка их точности для корректного картографирования.

Ключевые слова: базы данных, карты ареалов, разнообразие растений, сосудистые растения, Западная Сибирь, Алтае-Саянская горная область, Забайкалье.

Благодарности

Исследования выполнены в рамках государственного задания по теме № 121030900138-8.

Для цитирования

Санданов Д. В. Особенности работы с базами данных по распространению растений и опыт консолидации данных различного формата // Природа Внутренней Азии. Nature of Inner Asia. 2022. № 4(19). С. 96–104. DOI: 10.18101/2542-0623-2022-4-96-104

Введение

Разработка баз данных и веб-сервисов позволяет систематизировать имеющую информацию по распространению видов растений, открывает возможности для комплексного анализа их ареалов, позволяет составлять прогнозы по динамике ареалов изучаемых видов [Санданов, 2019]. Большие массивы данных открывают много возможностей, но вместе с этим исследователи сталкиваются с новыми проблемами и трудностями. Помимо работы с большим объемом информации существует проблема консолидации данных, зачастую весьма разнородных. Тут и возникает вопрос о предвзятости и неопределенности использования этих данных. Поэтому при таких исследованиях необходим стандартизированный сбор данных,

адекватные методы и процессы интеграции данных, наработка новых подходов для анализа данных [Wüest et al., 2020].

Одним из ключевых факторов является качество исходной информации, что позволяет при дальнейшем анализе данных избежать появления различных артефактов или закономерностей, не соответствующих характеру ареалов изучаемых видов. В современных биогеографических и макроэкологических исследованиях широко используется информация по распространению видов из различных открытых баз данных, и процесс интеграции таких гетерогенных данных представляет определенную сложность для исследователей [König et al., 2019].

Материалы и методы

Оцифровка карт распространения растений производилась с использованием инструментов «пространственная привязка» и «редактор» в программе ArcGIS 10.6.1. Пространственная привязка осуществляет процесс преобразования растровых данных (в нашем случае карты ареалов видов в формате jpeg) под векторный шейп-файл границ субъектов Российской Федерации. Для того чтобы границы карт ареалов совпадали с изучаемой территорией и для более точной их привязки использовалась система опорных точек. Далее весь массив данных преобразовывался в единую геодезическую систему отсчета WGS-84. Затем с помощью инструмента редактирования создавались объекты, отображающие точки распространения вида. Эти точки сохранялись в виде точечного шейп-файла с информацией по распространению вида (таблица с географическими координатами).

Результаты

Работа с самым крупным массивом разнородных данных проводилась нами при подготовке базы данных по распространению сосудистых растений Азиатской России (табл. 1). Наиболее значимый массив данных получен после оцифровки карт ареалов видов из сводок «Флора Сибири» (1987–2003) и «Сосудистые растения советского Дальнего Востока» (1985–1996), что составило 528 344 точек распространения видов. Предварительный анализ полученных данных выявил большое число точек местонахождений видов на юге Азиатской России и их малую представленность в средних широтах Сибири и на северо-востоке Дальнего Востока [Санданов, 2020].

Имеющиеся данные дополнены из различных источников и на текущий момент объем базы данных составляет 1 812 474 географических координат для 8 146 видов (табл. 1). Второй по значимости массив геопривязок получен из базы данных растительности Сибири [Kogolyuk, Zverev, 2012]. Общий объем с географическими координатами составил 16113 геоботанических описаний из трех крупных регионов Сибири: Западная Сибирь (10 529 описаний), Алтае-Саянская горная область (1 435), Забайкалье (4 149), что составляет более 50% от всей базы данных растительности Сибири.

Большой объем данных по распространению растений представлен в проекте «Флора России» на платформе iNaturalist. Этот набор данных представляет детальную информацию о местонахождении широко распространенных видов растений, которые зачастую очень слабо представлены в гербарных коллекциях.

Таблица 1

Количественный состав базы данных по распространению
сосудистых растений Азиатской России

Источник	Число видов и подвидов	Число геопривязок
<i>Флористические списки и флоры регионов с картами распространения растений</i>		
Флора Сибири (1987–2003)	2 567	398 946
Сосудистые растения советского Дальнего Востока (1985–1996)	4 084	129 398
Арктическая флора СССР (1964-1987)*	1 031	63 904
Флоры различных регионов Азиатской России	5 872	312 964
Флора Центральной Сибири (1979)	1 284	34 893
Эндемичные высокогорные растения Северной Азии (1974)*	231	13 433
<i>Пользовательские базы данных</i>		
Оцифрованные региональные гербарии	267	6 432
База данных по распространению видов рода <i>Oxytropis</i> Азиатской России	166	6 582
База данных геоботанических описаний Сибири (Korolyuk, Zverev, 2012)	2 381	461 793
<i>Информация по распространению растений в свободном доступе</i>		
Агроатлас (http://www.agroatlas.ru)*	533	49 518
Цифровой гербарий МГУ*	4 838	106 443
iNaturalist «Флора России»*	4 563	228 168
ИТОГО	8 146	1 812 474

Примечание. Звездочками отмечены источники, в которых приводилась информация по распространению растений на территории России и сопредельных стран. В таблице представлены только данные по Азиатской России. Данные по цифровому гербарии МГУ и базе iNaturalist «Флора России» приведены по текущему состоянию на 15.12.2020.

Активное участие пользователей из России в проекте iNaturalist в 2019 и 2020 гг. добавило к базе данных 228 168 наблюдений, и число этих данных продолжает дальше расти. Однако эти данные также оказались пространственно неоднородными, много точек распространения видов отмечается на юге Азиатской России и частично на северо-западе территории.

В дальнейшем имеющийся массив данных был в значительной степени дополнен картами ареалов из Арктической флоры СССР (1964–1987) и данными из цифрового гербария МГУ. Это позволило заполнить недостающие данные по распространению видов на севере изучаемого региона. В финальный вариант были включены все остальные доступные источники, включая материалы различных региональных и локальных флор Азиатской России, что позволило устранить пространственную неоднородность данных.

База данных основана на точечных данных, что позволяет в дальнейшем использовать их для разных целей: моделирование экологических ареалов растений, сеточное картирование флоры, разработка природоохранных мероприятий для редких и исчезающих видов и др. Большая часть имеющихся данных имеет сопоставимый масштаб, так как карты ареалов для многих видов во флористических сводках составлены на основе гербарных коллекций. Геоботанические описания и информация с платформы iNaturalist имеют хорошее разрешение, так как геопривязка в этом случае проводилась с использованием спутниковых навигаторов. Несколько обособленными являются точки распространения видов из Флоры Центральной Сибири, так как в данной сводке для составления карт ареалов растений использовался принцип сеточного картирования [Чепинога и др., 2017]. Вследствие этого оценка распространения сосудистых растений по данным из этой сводки возможна при мелкомасштабном картографировании. Использование этих данных для целей моделирования (особенно для редких видов или узколокальных эндемиков) является нецелесообразным из-за невысокой точности, которая при условиях горного пересеченного рельефа изучаемой территории может привести к ошибочным результатам [Санданов и др., 2020]. Имеющая в базе данных шкала точности позволяет формировать карты и наборы данных с различным разрешением (рис. 1). В дальнейшем данные с высоким разрешением можно использовать для крупномасштабного картографирования.

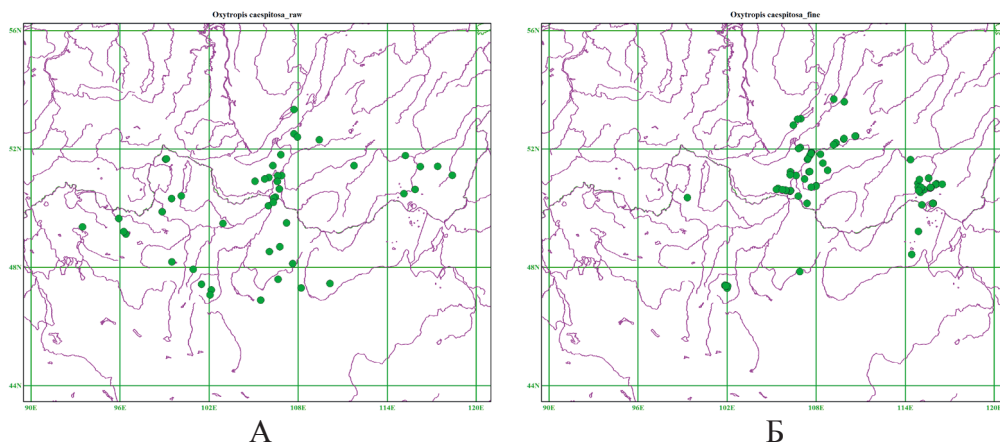


Рис. 1. Точки распространения *Oxytropis caespitosa* (Pall.) Pers. с различной степенью точности геопривязки

Примечание: А — точки с точностью геопривязки от 5 до 15 км, Б — точки с точностью геопривязки по спутниковому навигатору до 100 м.

Точность геопривязок является одним из основных критериев качества полученных данных. При подготовке третьего издания Красной книги Республики Бурятия¹ информация о распространении видов изначально готовилась в формате точечных данных, которая в дальнейшем была включена в ГИС-основу и формировались карты ареалов видов на изучаемой территории. Возможности этой базы

¹ Красная книга Республики Бурятия: Редкие и находящиеся под угрозой исчезновения виды животных, растений и грибов. Улан-Удэ : Изд-во БНЦ СО РАН, 2013. 688 с.

данных описаны ранее [Санданов, 2016]. Наполнение данных в базе продолжается, в частности, расширена и дополнена база данных для Тункинского национального парка [Санданов, 2021]. В конце 2013 г. также была завершена работа с очерками по редким и исчезающим видам растений Забайкальского края, но по техническим причинам Красная книга была издана позже¹. Карты ареалов редких растений в этом издании готовились без геопривязок, и авторы очерков наносили точки распространения видов в графическом редакторе на подготовленную картографическую основу.

Поскольку данные соседствующие регионы по физико-географическим особенностям объединяются как Забайкалье и информация из красных книг является сопоставимой во временном интервале, нами подготовлена единая база данных редких сосудистых растений Забайкалья [Sandanov et al., 2022b]. Для этого в ГИС-пакете ArcGIS 10.6.1. проведена оцифровка карт ареалов изучаемых растений Забайкальского края, полученный массив составил 1705 точек распространения для 165 видов. Итоговый массив общей базы данных по Забайкалью составил 2922 точки для 271 вида.

В процессе разработки этой базы данных возникли вопросы относительно точности геопривязок при оцифровке имеющихся карт. Данных для точной сверки по Красной книге Забайкальского края не имелось, сдвиги в точках распространения видов могли быть допущены на стадии их нанесения. Поэтому сверка точности оцифрованных данных проведена на материале из Красной книги Республики Бурятия (2013) на примере редкого вида астрагала трехгранноплодного *Astragalus trigonocarpus* (Turcz.) — эндемик Баргузинского хребта [Санданов, 2013]. При подготовке карты ареала этого вида были использованы авторские полевые данные с точными географическими координатами со спутникового навигатора. Нами проведена оцифровка карты из печатного издания Красной книги в трех независимых повторностях (три разных исследователя проводили оцифровку на отдельных компьютерах). Полученные результаты сравнивались с оригинальными данными (табл. 2).

Для изучаемого вида отмечено 9 местонахождений в пределах Баргузинского хребта. Можно отметить, что точность оцифровки во всех выборках не превышала 5 км, в некоторых случаях полученные оцифрованные данные были очень близки к оригинальным. Отмечается несколько меньшая точность для агрегированных точек распространения вида по сравнению с обособленными местонахождениями.

Анализ на разных выборках для карт ареалов редких видов Забайкальского края показал разброс значений от 1,5 до 6 км. Этот порог, возможно, является высоким для оценки локальных условий произрастания изучаемых видов, но вполне достаточен для оценки закономерностей пространственного распределения видов на всей территории Забайкальского края.

Включение или исключение даже одной ошибочной точки регистрации вида может оказать большое влияние на дальнейший анализ, особенно если данные используются при моделировании распространения вида. Поэтому на первоначальном этапе необходима сверка данных. На рисунке 2 показан характер изменения модели потенциального распространения остролодочника пузырчатоплодного *Oxytropis ampullata* (Pall.) Pers.

¹ Красная книга Забайкальского края. Растения. Новосибирск : Дом мира, 2017. 384 с.

Таблица 2

Сравнительный анализ точности оцифровки карты ареала
Astragalus trigonocarpus (Turcz.) Bunge

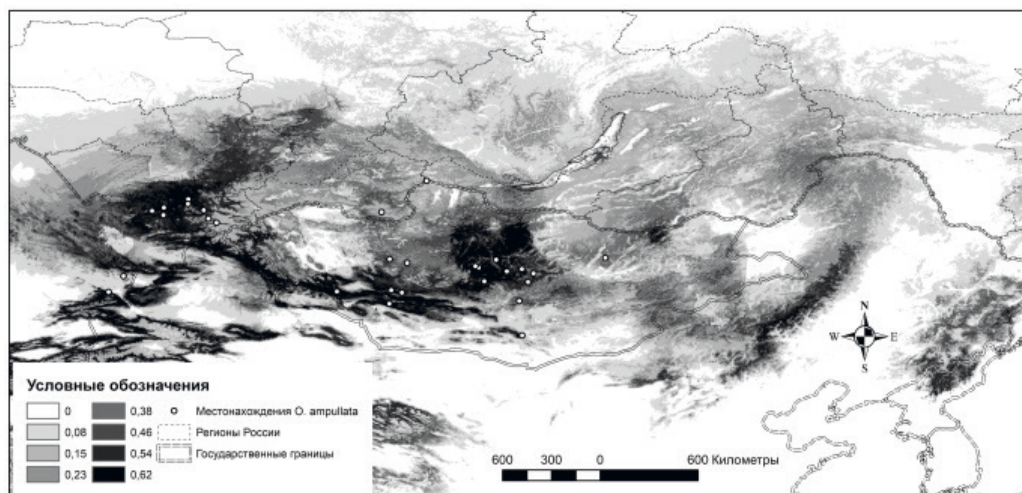
№	Отклонение от оригинальных данных, км		
	Исследователь 1	Исследователь 2	Исследователь 3
1	1.83	1.96	2.43
2	2.21	4.49	4.45
3	3.81	3.1	2.77
4	0.58	2.52	4.66
5	2.15	1.69	2.91
6	1.25	0.76	3.79
7	3.38	1.45	4.07
8	2.65	1.84	4.52
9	1.67	0.6	3.41
Max	3.81	4.49	4.66
Min	0.58	0.6	2.43
M±m	2.17±0.75	2.06±0.88	3.67±0.7

В первоначальный набор данных была включена точка регистрации вида с цифрового гербария МГУ с Хэнтэйского аймака Монголии (образец MW0183647, собран 18.06.1987 и определен Н. Улзийхутагом). Более детальный анализ сканированного изображения выявил несоответствие гербария изучаемому виду вследствие наличия длинных цветоносов и большого числа цветков в соцветии. Габитуально представленный образец более близок к остролодочнику блестящему *Oxytropis nitens* Turcz. и точно не является остролодочником пузырчатоплодным *O. ampullata* (для определения вида необходима непосредственная работа с гербарием). Исключение этого местонахождения вида позволило в значительной степени улучшить модель потенциального ареала изучаемого вида (рис. 2Б).

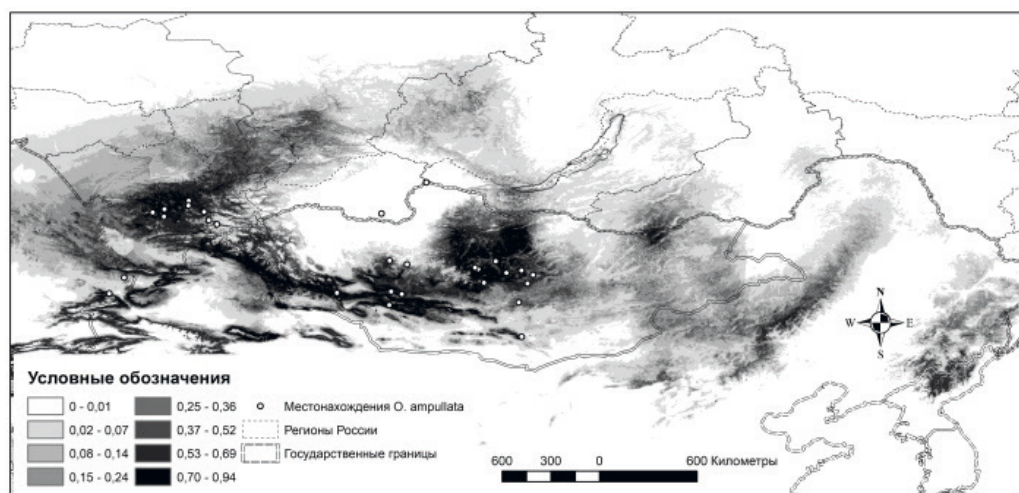
Обсуждение

Формирование баз данных по распространению видов должно основываться на валидизированных данных и иметь четкую структуру. Единый подход обеспечивает возможность сравнительного анализа разнородных данных, а использование шкалы точности позволяет легко отфильтровать нужные данные для дальнейших исследований. Информация из онлайн-баз данных нуждается в дополнительной верификации.

Оцифровка карт ареалов видов с использованием ГИС-программ позволяет получить необходимые данные по их распространению, которые можно использовать в последующем анализе. Полученные после оцифровки данные нуждаются в проверке с различными гербарными коллекциями. Этот процесс стал более доступным с появлением большого числа отсканированных гербарных материалов. За последние годы нами проведена оцифровка карт ареалов растений из различных флористических сводок [Brianskaia et al., 2021, Sandanov et al., 2021, 2022a].



А — предварительная некорректная модель



Б — корректная модель

Рис. 2. Потенциальный ареал *Oxytropis ampullata* (Pall.) Pers.
по результатам моделирования в Maxent на основе переменных ENVIREM

Примечание: цветом обозначена пригодность местообитаний согласно логистической шкале от 0 до 1. Более темным цветом показаны области с более пригодными местообитаниями.

Отметим, что корректное использование данных по распространению видов при моделировании экологических ареалов зависит от их пространственного разрешения и согласно этим показателям проводится подбор необходимых предикторов [Санданов, 2019]. В будущих исследованиях также необходимо уделить внимание качеству исходных данных и воспроизводимости аналитических методов, особенно при компиляции данных различного формата [Wüest et al., 2020]. Использование единого алгоритма подготовки данных и тщательная оценка информации на разных этапах с использованием современных методов является хорошей основой интеграции различных ботанических данных.

Литература

1. Санданов Д. В. *Astragalus trigonocarpus* (Turcz.) Bunge — Астрагал трехгранно-плодный // Красная книга Республики Бурятия. Редкие и находящиеся под угрозой исчезновения виды животных, растений и грибов. Улан-Удэ : Изд-во БНЦ СО РАН, 2013. С. 519–520. Текст : непосредственный.
2. Санданов Д. В. Геоинформационный анализ распространения редких сосудистых растений на территории Бурятии // Ученые записки Забайкальского государственного университета. 2016. Т.11, № 1. С. 38–45. Текст : непосредственный.
3. Санданов Д. В. Современные подходы к моделированию разнообразия и пространственному распределению видов растений: перспективы их применения в России // Вестник Томского государственного университета. Биология. 2019. № 46. С. 82–114. Текст : непосредственный.
4. Санданов Д. В. Разработка базы данных по распространению сосудистых растений Азиатской России // Информационные технологии в исследовании биоразнообразия : материалы III Национальной научной конференции с международным участием, посвященной 100-летию со дня рождения академика РАН П. Л. Горчаковского. Екатеринбург : Изд-во Гум. ун-та, 2020. С. 470–472. Текст : непосредственный.
5. Санданов Д. В., Дугарова А. С., Селюткина И. Ю. Моделирование распространения видов секции *Xerobia* Bunge рода *Oxytropis* DC. на территории Центральной Азии при климатических изменениях в прошлом и будущем // Вестник Томского государственного университета. Биология. 2020. № 52. С. 85–104. Текст : непосредственный.
6. Санданов Д. В. Редкие сосудистые растения Тункинского национального парка // Природа Внутренней Азии. 2021. № 2–3(18). С. 56–63. Текст : непосредственный.
7. Чепинога В. В., Петухин В. А., Стальмакова Д. П. Результаты сеточного картирования сводки «Флора Центральной Сибири» (1979) в цифровом формате: итоги и перспективы использования // Растительный мир Азиатской России. 2017. № 3(27). С. 70–78. Текст : непосредственный.
8. Brianskaia E., Sandanov D., Li Y., Wang Z. Distribution of alpine endemic plants of northern Asia: a dataset. *Biodiversity Data Journal*. 2021; 9:e75348
9. Korolyuk A. Yu., Zverev A. A. Database of Siberian Vegetation (DSV). *Biodiversity & Ecology*. 2012; 4 : 312.
10. König C., Weigelt P., Schrader J., Taylor A., Kattge J., Kreft H. Biodiversity data integration — the significance of data resolution and domain. *PLoS Biology*. 2019; 17(3): e3000183.
11. Sandanov D. V., Brianskaia E. P., Batotsyrenov E. A. Distribution of vascular plants north of Lake Baikal: a new, open access dataset. *Biodiversity Data Journal*. 2021; 9: e77409.
12. Sandanov D. V., Dugarova A. S., Brianskaia E. P., Selyutina I. Yu., Makunina N. I., Dudov S. V., Chepinoga V. V., Wang Z. Diversity and distribution of *Oxytropis* DC. (Fabaceae) species in Asian Russia. *Biodiversity Data Journal*. 2022a. 10: e78666.
13. Sandanov D. V., Brianskaia E. P., Dugarova A. S. Dataset for vascular plants in the Red Data Books of Transbaikalia: species distribution and pathways towards their conservation. *Nature Conservation Research*. 2022b; 7, Suppl. 1: 14–23.
14. Wüest R. O., Zimmermann N. E., Zurell D., Alexander J. M., Fritz S. A., Hof C., Kreft H., Normand S., Cabral J. S., Szekely E., Thuiller W., Wikelski M., Karger D. N. Macroecology in the age of Big Data — Where to go from here? *Journal of Biogeography*. 2020; 47: 1–12.

Статья поступила в редакцию 12.11.2022; одобрена после рецензирования 20.11.2022; принята к публикации 25.11.2022.

FEATURES OF WORK WITH DATABASES ON PLANT SPECIES DISTRIBUTION
AND THE EXPERIENCE IN COMBINING DATA OF VARIOUS FORMATS

D. V. Sandanov

Denis V. Sandanov

Cand. Sci. (Biol.),

Institute for General and Experimental Biology SB RAS

6 Sakhyanovoy St., Ulan-Ude 670047, Russia

Tunkinsky National Park

130 Lenina St., Kyren 671010, Russia

sdenis1178@mail.ru

Abstract. Analysis of databases and other big data has certain specifics. In recent years, we have been studied the consolidation of various plant distribution data to assess diversity and species richness, and use this information for species distribution modeling. The article presents the results of integration of databases on the distribution of vascular plants of Asian Russia, and considers the features of working with various data sets. We have been discussed the possibilities of digitizing plant distribution maps by GIS. It is important to recognize data heterogeneity and precision for better mapping procedures.

Keywords: databases, distribution maps, plant diversity, vascular plants, Western Siberia, Altai-Sayan mountain region, Transbaikalia.

For citation

Sandanov D. V. Features of Work with Databases on Plant Species Distribution and the Experience in Combining Data of Various Format. *Nature of Inner Asia*. 2022; 4(22): 96–104. DOI: 10.18101/2542-0623-2022-4-96-104

The article was submitted 12.11.2022; approved after reviewing 20.11.2022; accepted for publication 25.11.2022.