IV Всероссийская школа молодых учёных
Системный анализ динамики природных процессов в российской Арктике
г. Видное, 4–7 июня 2024 г.

Применение методов машинного обучения при решении задач в науках о Земле

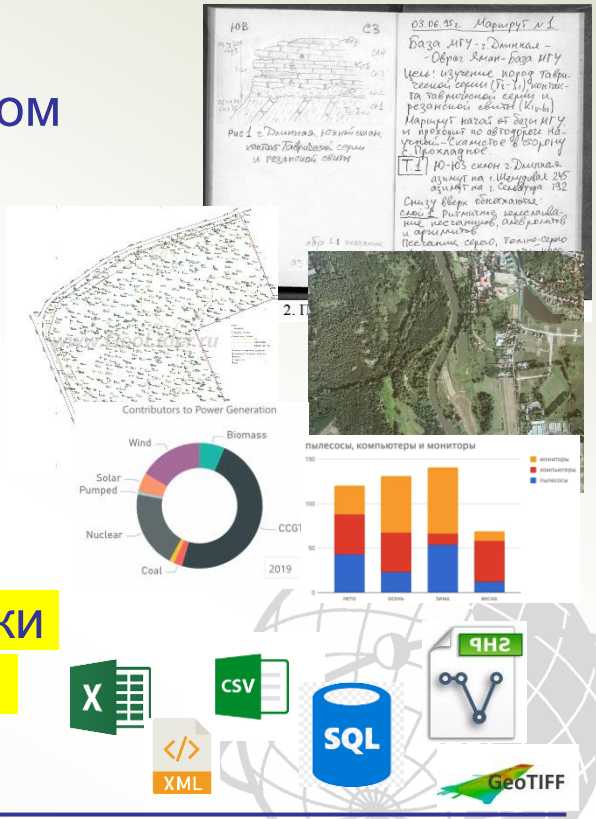
И.А. Лисенков
Геофизический центр Российской академии наук

Категории геопространственных данных

- Геодезия и картография;
- Глобальные навигационные спутниковые системы;
- География;
- Геофизика;
- Геология;
- Полезные ископаемые;
- Гляциология;
- Гидрология;
- Дистанционное зондирование Земли;
- Метеорология и климатология;
- Почвоведение;
- Политическая география;
- Население;
- Промышленность;
- Сельское хозяйство;
- Транспорт;
- Биогеография: география организмов;
- Биогеография: география растительного покрова;
- Экология.

Формат представления геопространственных данных

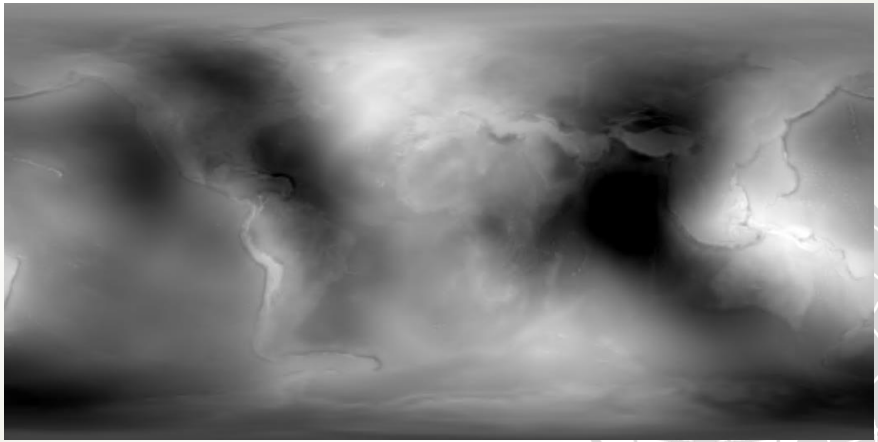
- неструктурированные источники в аналоговом виде (журналы и полевые дневники наблюдений, научные публикации);
- неструктурированные оцифрованные источники (текст в электронном виде, цифровые изображения и видеофайлы);
- структурированные источники в аналоговом виде (таблицы, графики, диаграммы)
- структурированные оцифрованные источники (табличные файлы, базы данных *SQL*, *XML*, *JSON*).



Виды представления геопространственных данных

- Векторные данные (точки, линии, полигоны);
- Растровые данные (массив точек с заданным шагом)

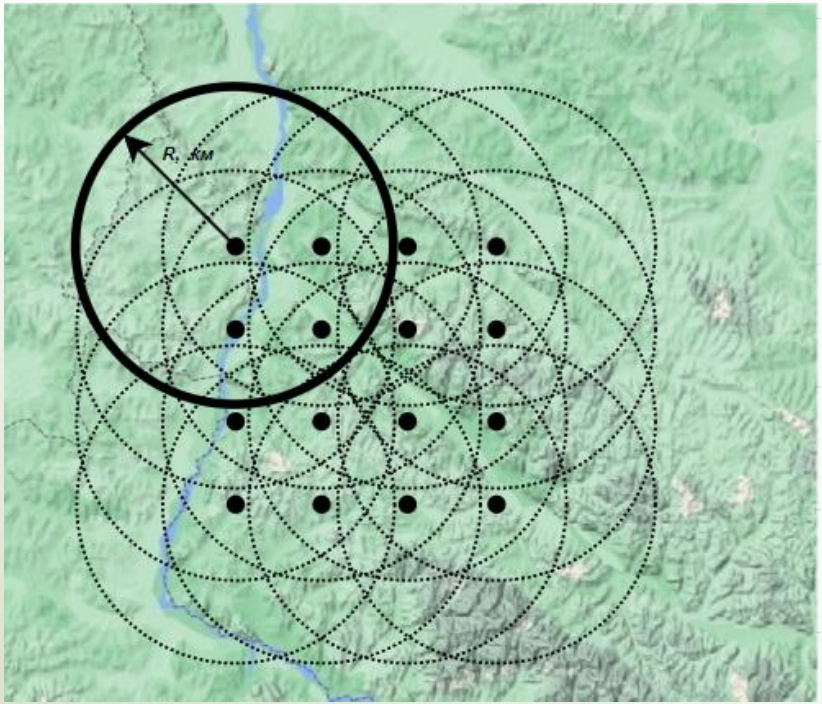
Примеры визуализации геопространственных данных в QGIS

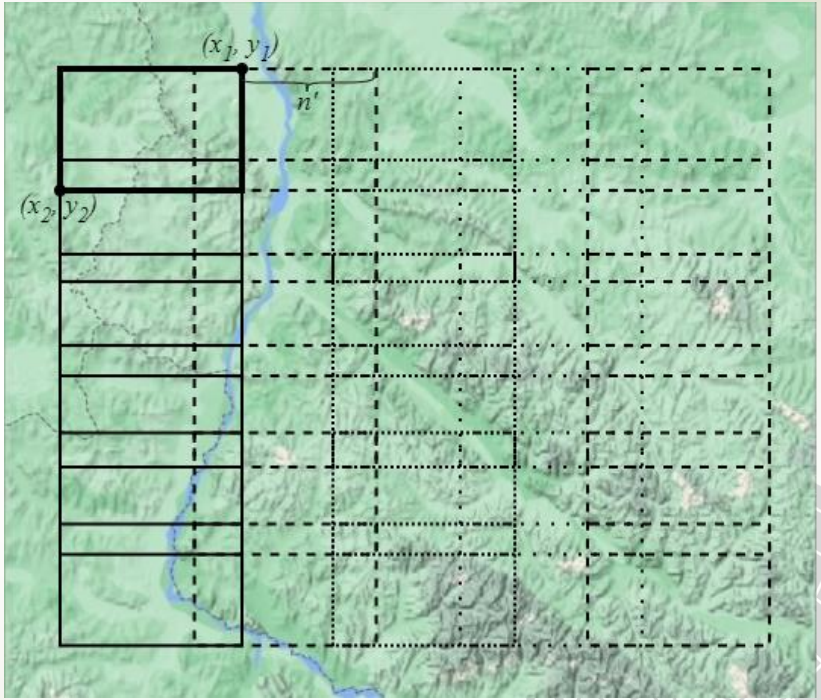


shapefile

geotiff

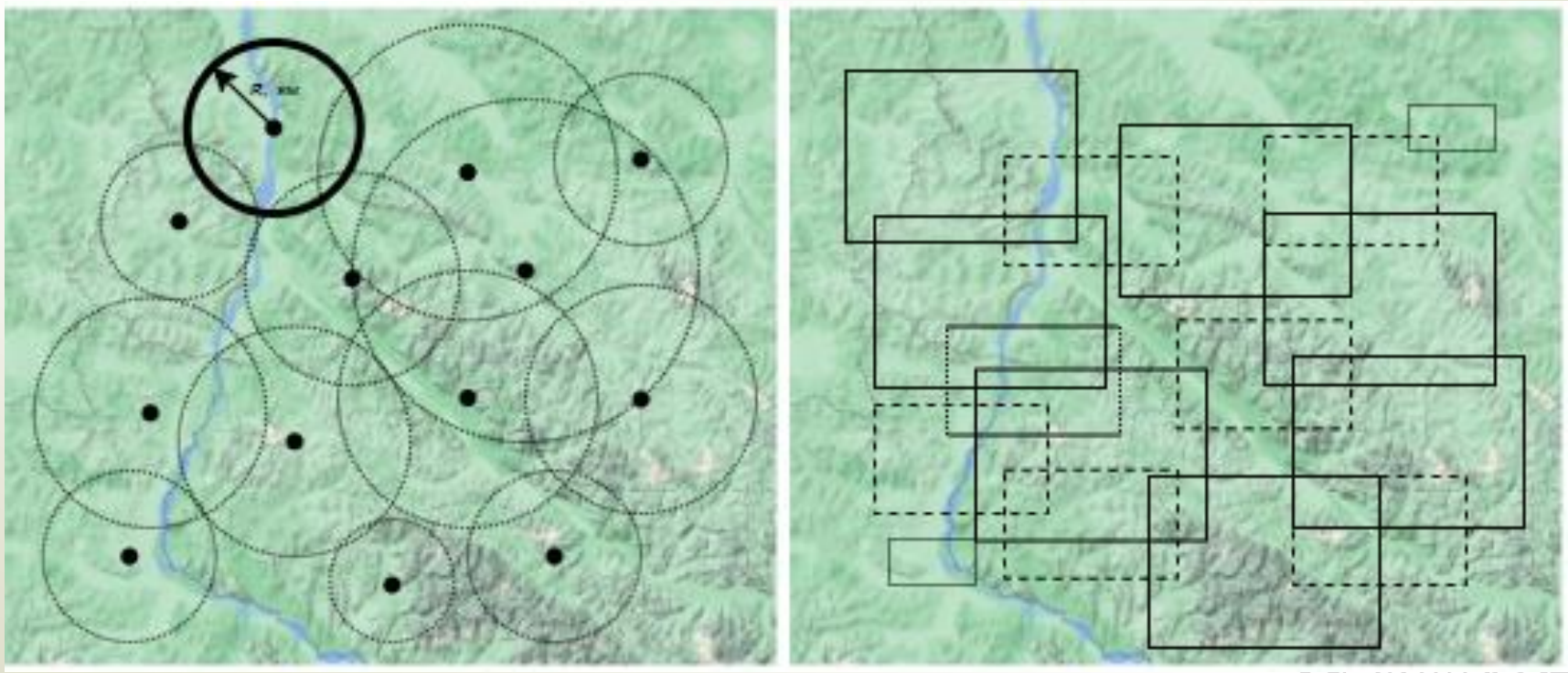
Формирование окрестностей [1]





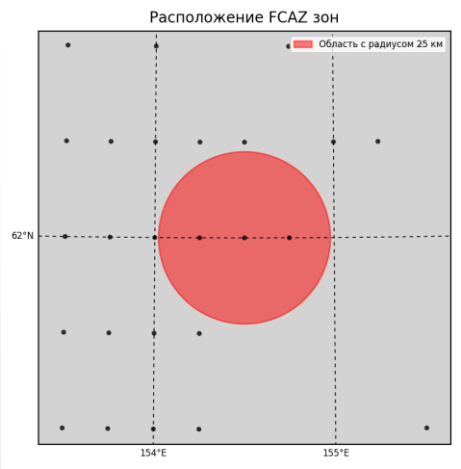
5 Лисенков И.А., Соловьев А.А., Кузнецов В.А., Николова Ю. И. ОБОБЩЕННЫЙ МАССИВ ГЕОЛОГО-ГЕОФИЗИЧЕСКОЙ ИНФОРМАЦИИ ВОСТОЧНОГО СЕКТОРА РОССИЙСКОЙ АРКТИКИ ДЛЯ ПРОВЕДЕНИЯ АНАЛИЗА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ // Геология и геофизика, 2024 (ожидается публикация)

Формирование произвольных окрестностей [1]



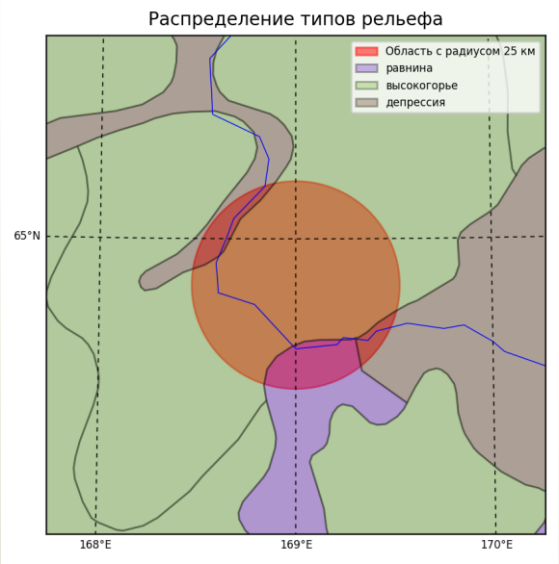
6 Лисенков И.А., Соловьев А.А., Кузнецов В.А., Николова Ю. И. ОБОБЩЕННЫЙ МАССИВ ГЕОЛОГО-ГЕОФИЗИЧЕСКОЙ ИНФОРМАЦИИ ВОСТОЧНОГО СЕКТОРА РОССИЙСКОЙ АРКТИКИ ДЛЯ ПРОВЕДЕНИЯ АНАЛИЗА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ // Геология и геофизика, 2024 (ожидается публикация)

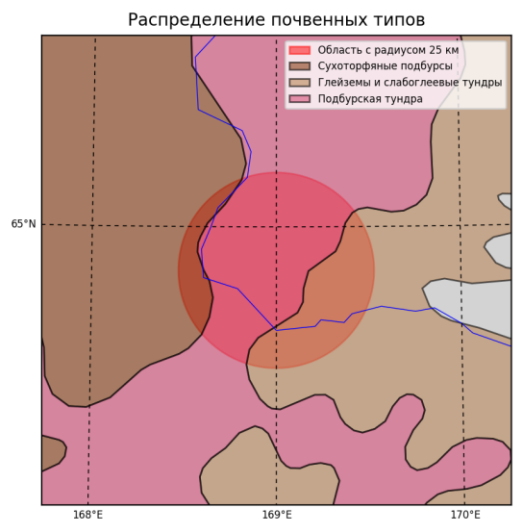
Комбинирование данных в окрестности



n	FCAZ	Разлом (Ранг 1)	Разлом (Ранг 2)	Разлом (Правый сдвиг)	Разлом (вертикальное смещение)
1	3	0	1	0	1
2	1	3	0	2	1

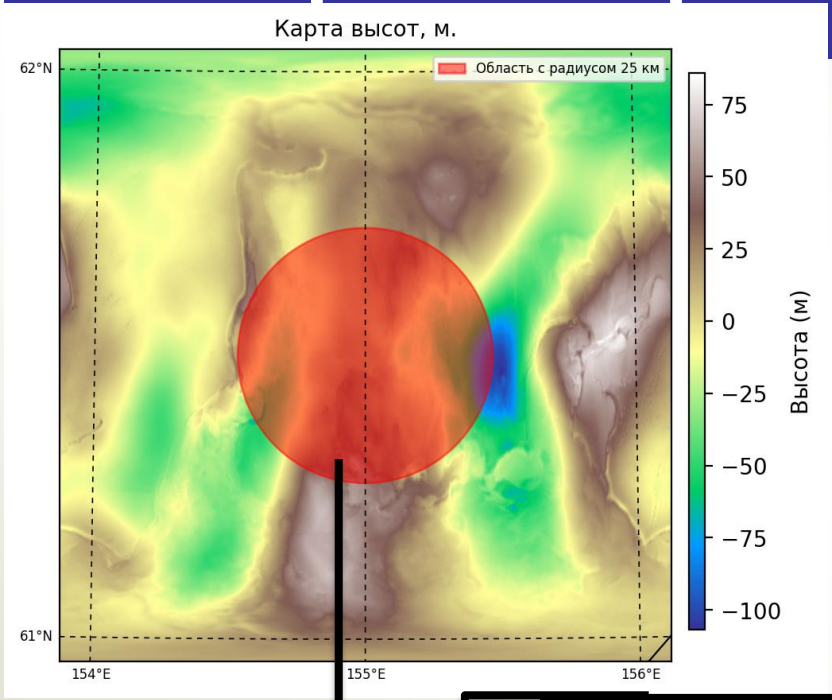
Комбинирование векторных данных (полигоны)

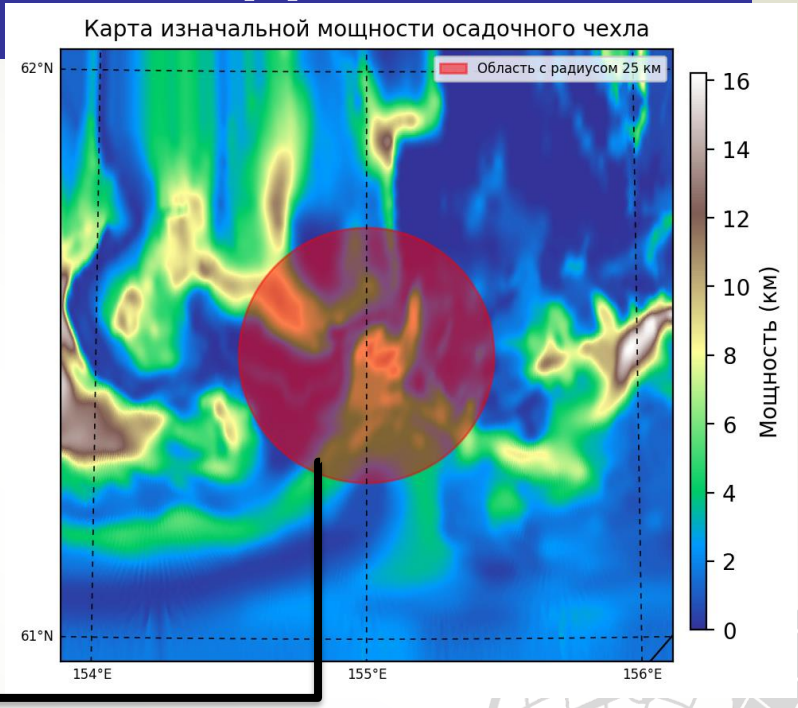




n	Рельеф_равнина	Рельеф_спуск	Рельеф_высокогорье	Почва_тундра	Почва_глейземы
1	2	1	1	1	1
2	0	1	1	2	0

Обработка растровых данных





Из каждой окрестности извлекаются следующие характеристики:

MAX – максимальное значение
MIN – минимальное значение
DIFF – разница MAX и MIN
MEAN – медианное значение

Обобщенный массив данных

- Комплексная совокупность информации, объединяющая все доступные данные об объекте исследования.

FAULTS_SENS1_V	...	grad_height	lat	lon	max_height	max_magn_anom	min_height	min_magn_anom	slope_degree	slope_exposure
NaN	...	1.061218	76.899485	100.1	25.59	20.5611	-39.62	-70.2658	0.416423	-2.342057
NaN	...	0.959253	76.899485	100.3	17.25	22.0803	-39.62	-70.2658	0.290938	-3.642595
NaN	...	0.969691	76.899485	100.5	20.88	29.6652	-37.62	-70.2658	0.346102	-11.039318
NaN	...	0.955165	76.899485	100.7	34.27	15.6623	-22.25	-66.6200	0.579188	0.486362
NaN	...	2.108679	76.899485	100.9	36.02	35.9820	-22.25	-52.5771	1.303514	-27.027916
...	...	...	...	...	...	...	...	...	...	...
NaN	...	0.164680	57.600515	192.9	-61.00	434.1925	-71.38	-256.2720	-0.968239	-147.865660
NaN	...	0.178469	57.600515	193.1	-59.81	505.9238	-70.75	-172.8039	-0.975932	-155.179182
NaN	...	0.182252	57.600515	193.3	-58.19	505.9238	-70.00	-41.4234	-0.897800	-155.456505
NaN	...	0.191191	57.600515	193.5	-56.56	505.9238	-68.88	-50.4846	-0.878275	-151.715476
NaN	...	0.316585	57.600515	194.9	-44.81	254.4485	-64.62	-217.3048	-0.716021	-152.691348

Обобщенный массив Восточного сектора Арктики РФ

- Координаты границ сектора
  60° с.ш., 100° в.д.;
  77° с.ш., 100° в.д.;
  77° с.ш., 165° в.д.;
  57.5° с.ш., 165° в.д.;
  57.5° с.ш., 138° в.д.;
  60° с.ш., 138° в.д
- 7025 круговых окрестностей радиусом 50 км.
- Равномерный шаг - 0.2° по широте и долготе
- В каждой круговой окрестности консолидируются
  данные: Высота рельефа; Геомагнитные характеристики; Каталог землетрясений …
  Всего 107 показателей/514 компонент
  (Июнь 2024)

Базовые задачи анализа данных

- Бинарная и множественная классификация (binary/multi classification)
- Поиск регрессии (Regression)
- Предсказание показателя (time series)
- Кластерный анализ (Clustering)
- Выявление аномалий в данных (Anomaly Detection)
- Извлечение данных (Named Entity Recognition)

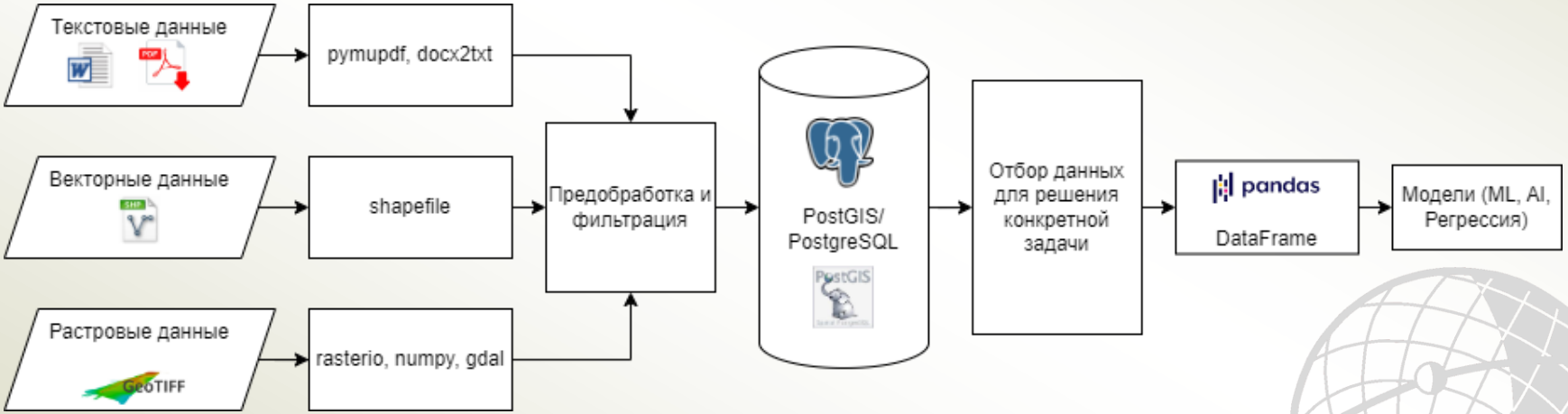
Рекомендуемая литература

1. Гвишиани А.Д., Горшков А. И., Ранцман Е.Я. и др. Прогнозирование мест землетрясений в регионах умеренной сейсмичности. М.: Наука, 1988. 174 с.
2. Гвишиани А.Д., Дзебоев Б.А., Агаян С.М. Интеллектуальная система распознавания FCAZm в определении мест возможного возникновения сильных землетрясений горного пояса Анд и Кавказа // Физика Земли. 2016. № 4. С. 3–23. DOI: 10.7868/S0002333716040013
3. Гвишиани А.Д., Добровольский М. Н., Дзеранов Б.В., Дзебоев Б.А. Большие Данные в геофизике и других науках о Земле // Физика Земли. 2022. № 1. С. 3–34. DOI: 10.31857/S0002333722010033
4. Есин Е.И., Василевский А. Н., Бушенкова Н. А. ПРОСТРАНСТВЕННЫЕ КОРРЕЛЯЦИИ ОСОБЕННОСТЕЙ РЕЛЬЕФА, ГРАВИТАЦИОННОГО ПОЛЯ И АНОМАЛИЙ СКОРОСТЕЙ СЕЙСМИЧЕСКИХ ВОЛН ЦЕНТРАЛЬНОЙ ЗОНЫ КАМЧАТСКОГО РЕГИОНА // Геология и геофизика. №2. 2024. С. 303-318. DOI: 10.15372/GiG2023165
5. Кондорская Н.В., Горбунова И. В., Киреев И. А., Вандышева Н. В. В сб.: Сейсмичность и сейсмическое районирование Северной Евразии. М.: ИФЗ, 1993. В. 1. С. 70–79.
6. Кулаков И. Ю., Гайна К., Добрецов Н. Л., Василевский А. Н., Н. А. Бушенкова. РЕКОНСТРУКЦИИ ПЕРЕМЕЩЕНИЙ ПЛИТ В АРКТИЧЕСКОМ РЕГИОНЕ НА ОСНОВЕ КОМПЛЕКСНОГО АНАЛИЗА ГРАВИТАЦИОННЫХ, МАГНИТНЫХ И СЕЙСМИЧЕСКИХ АНОМАЛИЙ // Геология и геофизика, 2013, т. 54, № 8, с. 1108–1125
7. Соловьев Ал. А., Новикова О. В., Горшков А. И., Пиотровская Е. П. Распознавание расположения потенциальных очагов сильных землетрясений в Кавказском регионе с использованием ГИС-технологий // Доклады Академии наук (2013). 450. 599-601. 10.7868/S0869565213170222.
8. Соловьев Ал. А., Горшков А. И., Соловьев Ан. А. ПРИМЕНЕНИЕ ДАННЫХ ПО ЛИТОСФЕРНЫМ МАГНИТНЫМ АНОМАЛИЯМ В ЗАДАЧЕ РАСПОЗНАВАНИЯ МЕСТ ВОЗМОЖНОГО ВОЗНИКНОВЕНИЯ ЗЕМЛЕТРЯСЕНИЙ // ФИЗИКА ЗЕМЛИ, 2016, № 6, с. 21–27
9. Соловьев А. А., Лисенков И. А. ОБЗОР И ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ СОВРЕМЕННЫХ ПОДХОДОВ КОМПЛЕКСНОГО АНАЛИЗА ГЕОДАННЫХ ДЛЯ ПРОГНОЗА ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ГЕОЛОГО-ГЕОФИЗИЧЕСКИХ ПАРАМЕТРОВ// Геофизические исследования, 2024, №, с.– (принята в печать)
10. Amante, C. and B. W. Eakins, ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24, 19 pp, March 2009
11. Apache Hadoop, accessed 15 March 2024, https://hadoop.apache.org
12. AutoKeras: An AutoML system based on Keras, accessed 07 March 2024, https://autokeras.com/
13. Boehmke, Bradley; Greenwell, Brandon (2019). "Gradient Boosting". Hands-On Machine Learning with R. Chapman & Hall. pp. 221–245. ISBN 978-1-138-49568-5.
14. Chengsheng, Tu & Huacheng, Liu & Bing, Xu. (2017). AdaBoost typical Algorithm and its application research. MATEC Web of Conferences. 139. 00222. 10.1051/matecconf/201713900222

Рекомендуемая литература (2)

16. dBase .DBF File Structure, accessed 15 March 2024, https://www.dbase.com/Knowledgebase/INT/db7_file_fmt.htm
17. ESRI Shapefile Technical Description// July 1998, Available at: https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/whitepapers/pdfs/shapefile.pdf
18. Gvishiani A., Vorobieva I., Shebalin P., Dzeboev B., Dzeranov B., Skorkina A. Integrated Earthquake Catalog of the Eastern Sector of Russian Arctic // Applied Sciences. 2022. Vol. 12. 5010. DOI: 10.3390/app12105010
19. IIASA - The International Institute for Applied Systems Analysis. 2002. Land Resources of Russia, accessed 17 April 2024, https://webarchive.iiasa.ac.at/Research/FOR/russia_cd/download.htm. The International Institute for Applied Systems Analysis (IIASA), Russian Academy of Sciences (RAS). 2002. Land Resources of Russia Available at: https://databasin.org/galleries/44513fc408c149c69b8d72bf0c37ef5f/
20. Keras 3 API documentation, accessed 15 March 2024, < https://keras.io/api/>
21. Hancock, John & Khoshgoftaar, Taghi. (2020). Survey on categorical data for neural networks. Journal of Big Data. 7. 10.1186/s40537-020-00305-w.
22. Lesur, Vincent & Hamoudi, Mohamed & Choi, Yujin & Dyment, J. & Erwan, Thebault. (2016). Building the second version of the World Digital Magnetic Anomaly Map (WDMAM). Earth, Planets and Space. 68. 10.1186/s40623-016-0404-6.
23. NumPy, the fundamental package for scientific computing with Python, accessed 15 March 2024, https://numpy.org
24. OGC GeoTIFF standard, publication date 14 September 2019 https://docs.ogc.org/is/19-008r4/19-008r4.html
25. Pandas – Python Data Analysis Library, accessed 15 March 2024, <https://pandas.pydata.org/>
26. PostGIS 3.3.4dev Manual, accessed 15 March 2024, <http://postgis.net/documentation/manual-3.3/>
27. PostgreSQL: The World's Most Advanced Open-Source Relational Database, accessed 15 March 2024, https://www.postgresql.org/
28. PyShp - The Python Shapefile Library (PyShp) provides read and write support for the Esri Shapefile format, accessed 15 March 2024, <https://github.com/GeospatialPython/pyshp>
29. QGIS A Free and Open-Source Geographic Information System, accessed 15 March 2024, < https://qgis.org/en/site/>
30. Rasterio: access to geospatial raster data, accessed 15 March 2024, https://rasterio.readthedocs.io/en/stable/
31. Roh, Yuji & Heo, Geon & Whang, Steven. (2019). A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Transactions on Knowledge and Data Engineering. PP. 1-1. 10.1109/TKDE.2019.2946162.
32. Scikit-learn machine learning in Python, accessed 15 March 2024, <https://scikit-learn.org/stable/>
33. Zelenin E.A, Bachmanov D.M., Garipova S.T., Trifonov V.G., Kozhurin A.I. The Active Faults of Eurasia Database (AFEAD): the ontology and design behind the continental-scale dataset // Earth System Science Data. 2022. vol. 14. p. 4489-4503

Архитектура ПО



Текстовые неструктурированные данные

Обработка естественного языка (NLP)

- Классификация документов
  - Перевод текст в численный вектор, Bag-of-Words;
  - Подбор оптимальных моделей для классификации, например, LSTM;
- Извлечение значимого текста
  - Разметка текста с помощью тегов B-I-O;
  - Извлечение токенов из текста;