Электронный научный журнал "Математическое моделирование, компьютерный и натурный эксперимент в естественных науках" http://mathmod.esrae.ru/
URL статьи: **mathmod.esrae.ru/44-176**
Ссылка для цитирования этой статьи:
Акчурин А.В., Кондратова Ю.Н. Определение областей повышенной дорожно-транспортной опасности на карте города с помощью кластеризации данных // Математическое моделирование, компьютерный и натурный эксперимент в естественных науках. 2023. №4

УДК 004.89 DOI:10.24412/2541-9269-2023-4-02-11

ОПРЕДЕЛЕНИЕ ОБЛАСТЕЙ ПОВЫШЕННОЙ ДОРОЖНО-ТРАНСПОРТНОЙ ОПАСНОСТИ НА КАРТЕ ГОРОДА С ПОМОЩЬЮ КЛАСТЕРИЗАЦИИ ДАННЫХ

Акчурин А.В.[1], Кондратова Ю.Н.[2]
[1]Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского,
Россия, Саратов, ak.arty@mail.ru
[2]Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского,
Россия, Саратов, kondratovaun@mail.ru

IDENTIFYING AREAS OF INCREASED TRAFFIC DANGER ON THE CITY MAP USING DATA CLUSTERING

Akchurin A.V.[1], Kondratova Yu.N.[2]
[1]Saratov State University of Saratov, Russia,
Saratov, ak.arty@mail.ru
[2]Saratov State University of Saratov, Russia,
Saratov, kondratovaun@mail.ru

Аннотация. Каждый год в России происходят десятки тысяч дорожно-транспортных происшествий, часть из которых заканчиваются смертельными исходами. Аварии происходят по разным причинам: от невнимательности водителей до неисправности транспортных средств и плохого обустройства дорог.

В статье представлен процесс разработки программного обеспечения для выявления таких областей дорожной опасности. Присутствует описание процесса сбора данных о дорожно-транспортных происшествиях, их обработки, поиска региональных кластеров на карте города для выделения областей концентрации и их дальнейшей кластеризации с учетом опасности каждой области, а также отображение областей на карте города. В результате был разработан метод выделения области концентрации дорожно-транспортных происшествий, а также определены уровни опасности областей.

Ключевые слова: дорожно-транспортные происшествия, области дорожной опасности,

DBSCAN кластеризация, кластеризация k-means, коэффициент силуэта, отображение на карте.

Abstract. Tens of thousands of traffic accidents occur in Russia every year, some of which end in fatal outcomes. Accidents occur for various reasons: from driver inattention to vehicle malfunction and poor road construction.

The article presents the process of software development to identify such areas of road hazard. There is a description of the process of collecting data on road accidents, processing them, searching for regional clusters on the city map to identify areas of concentration and further clustering them, taking into account the danger of each area, as well as displaying areas on the city map. As a result, a method was developed to identify the area of concentration of road accidents, and the hazard levels of the areas were determined.

Keywords: traffic accidents, road hazard areas, DBSCAN clustering, k-means clustering, silhouette index, display on the map.

1. Введение

Проблема дорожной опасности остро стоит в наше время, ведь каждый год происходит около сотни тысяч аварий, многие из которых уносят жизни людей. Исследования в направлении анализа ДТП производятся, однако направление слабо развито. А все существующие решения имеют определенные недостатки: плохо продуманная карта опасности областей города, которая только предоставляет информацию о проблемном месте, не проводя более широкую аналитику опасности. Помимо этого, близко находящиеся происшествия никак не объединены и отображаются независимо. А также нигде не учитывается характер опасности, везде анализируется лишь количество аварий на единицу расстояния. В данной статье освещается процесс сбора и анализа данных с целью их отображения на карте города в приложении, которое будет создано для информирования о местах наибольших концентраций дорожно-транспортных происшествий, произошедших в городе Саратов. Данный пример может быть обобщен и на карты других городов.

2. Получение и форматирование данных о ДТП

Для реализации собственного решения необходимо получить данные о произошедших авариях, на основании которых и будет проводиться вся дальнейшая обработка. Для этой цели был использован официальный сайт ГИБДД (stat.gibdd.ru). На этом сайте собрана и сгруппирована информация обо всех зафиксированных происшествиях с учетом временных промежутков, информации о пострадавших и множества других параметров описания ДТП. Ее правильный сбор является первой задачей.

Так как на электронном ресурсе, содержащем информацию, находится много интерактивных элементов, то обычная выгрузка страницы и извлечение требуемых данных из её html-кода не представляется возможной. В этом случае требуется написать программу, имитирующую действия пользователя на странице, чтобы все необходимые скрипты на сайте были выполнены до

попыток сбора информации. Для решения этой задачи был использован язык программирования «Python» и библиотека «Selenium» для него. «Selenium» является инструментом автоматизации действий веб-браузера, в том числе позволяет собирать данные с интернет-страниц и имеет очень обширные настройки взаимодействия [1-3]. С его помощью был составлен алгоритм, описывающий следование к местонахождению нужных данных и их выгрузки.

Полученные файлы содержат много информации, которая не будет использована в решении. Поэтому следующим шагом является ее правильное извлечение и преобразование. Для этих целей написан парсер данных, который в полученной структуре находит требуемые атрибуты и извлекает из них данные. Сначала был составлен список атрибутов, затем для каждого из них указано его местоположение в дереве кода страницы и после этого запущено извлечение элементов всего этого списка. На выходе получили данные, содержащие все необходимые параметры о ДТП, которые преобразовали в файл формата json.

3. Нахождение оптимальных параметров кластеризации

Следующим этапом является нахождение на карте города регионов концентрации происшествий. Это очень важно для выделения областей, связанных единой проблемой. Если в небольшой области находится высокая концентрация аварий, с большой долей вероятности они вызваны какой-то одной определенной проблемой.

Во многих случаях такими областями концентрации являются улицы и перекрестки. И так как области представляют собой различные геометрические формы, то наиболее подходящим методом кластеризации является DBSCAN (Density-based spatial clustering of applications with noise) за счет своей устойчивости к шуму и возможности выделять области разных форм. Данный алгоритм используется для выделения связанных компонент. Его основная концепция состоит в том, чтобы найти области повышенной плотности, которые отделены друг от друга областями с низкой плотностью [4]. Схема кластеризации представлена на рис.1.

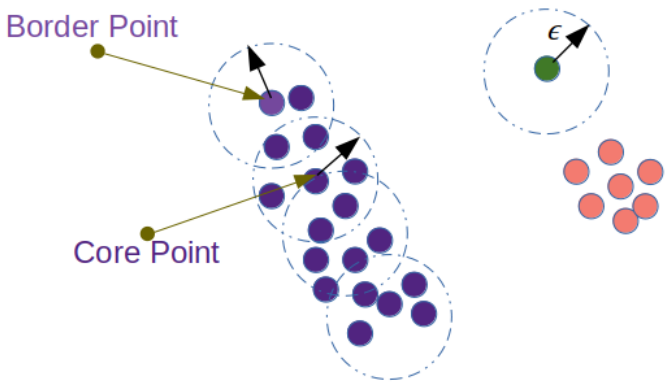

Рис.1. Схема кластеризации методом DBSCAN

Основные параметры, используемые методом: максимальный радиус соседства – предел отдаления, при котором точки считаются соседствующими и минимальное количество соседей в радиусе – число элементов в окрестности, при котором точка рассматривается как основная. В качестве минимального количества соседей было взято значение 2. Так можно находить даже небольшие скопления точек, что является важным, учитывая большой разброс данных по карте. Для нахождения оптимального значения максимального радиуса соседства был построен график K-расстояний, который показывает дистанцию между точкой и ближайшей к ней точкой данных для всех элементов набора. Этот график изображен на рис.2.

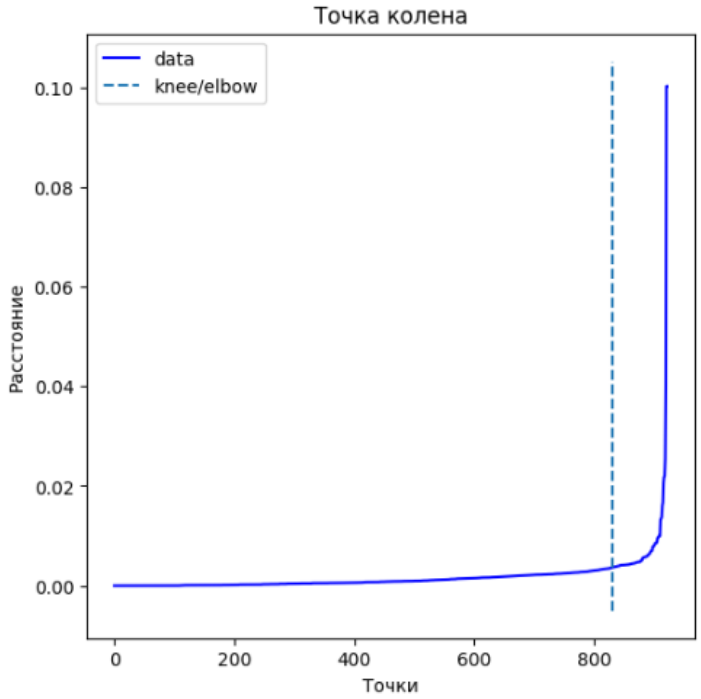

Рис.2. График K-расстояний

Нахождение на графике точки колена – порогового значения, при котором наблюдается резкое изменение, позволяет найти ориентировочное оптимальное значение максимального радиуса соседства. В нашем случае таким значением является 0.00367. Чтобы найти оптимальный радиус более точно, была рассмотрена окрестность найденной точки: диапазон от 0.0015 до 0.006 с небольшим шагом 0.0001. И для каждого значения из диапазона проведена кластеризация методом DBSCAN.

Для понимания, какое из значений разбиения эффективнее, использовалась метрика коэффициента силуэта. Коэффициент силуэта показывает, насколько каждый объект выборки похож на другие объекты в этом же кластере и отличается от объектов других кластеров. С ростом значения коэффициента усиливается эффективность кластеризации [5]. Основным аргументом использования этой метрики является отсутствие требования разметки данных,

что позволяет применить ее в нашем случае. Максимальное значение коэффициента силуэта было найдено при значении радиуса равном 0.0025.

4. Нахождение кластеров областей

С учетом найденных параметров была проведена кластеризация и получено распределение мест дорожно-транспортных происшествий по географическим кластерам. В результате распределения найдено 158 кластеров и 168 точек шума, что, учитывая разрозненность точек в некоторых частях, является хорошим результатом. Отображение результатов кластеризации на графике представлено для наглядности с помощью библиотеки «Matplotlib». Выделение кластеров выполнилось успешно, были выделены скопления близлежащих точек – перекрестков, а также длинные вытянутые участки – улицы. При достаточно высокой разрозненности данных, точки шума, выделенные фиолетовым цветом на графике, составляют всего 15 процентов от всех происшествий. Выделение кластеров областей изображено на рис.3.

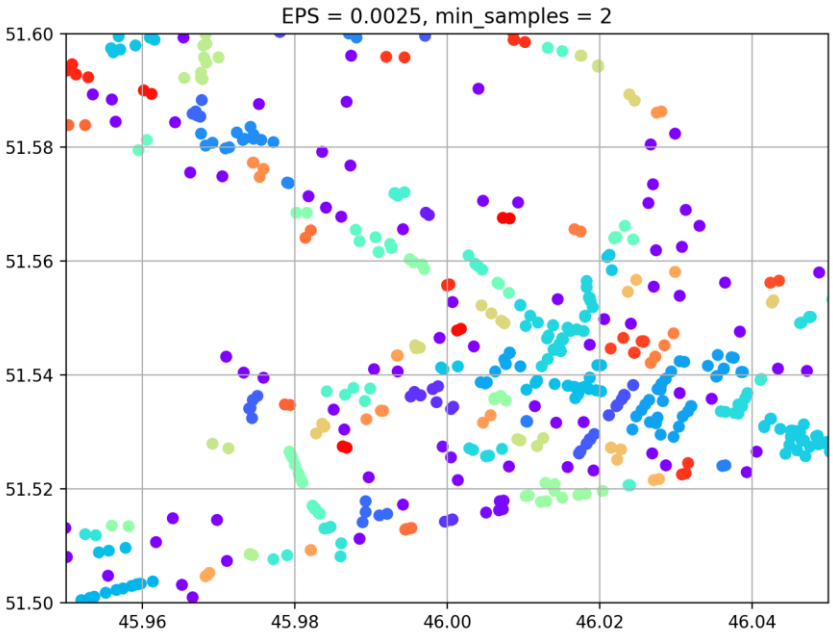

Рис.3. Выделенные кластеры областей

5. Объединение в региональные кластеры

Для дальнейших действий была просуммирована информация обо всех происшествиях, входящих в кластер, и представлена в виде одного элемента. В состав интересующей информации о каждой области опасности входят следующие данные: координаты происшествия, количество транспортных средств, участвующих в ДТП, число пострадавших и число погибших. Полученная информация помогает сделать вывод об опасности всех областей. Для каждой области найдено среднее значение каждого из рассматриваемых параметров. В будущем, при необходимости, набор этих атрибутов может быть расширен.

Полученные данные могут находиться в различном масштабе и вносить разный вклад, поэтому для правильного получения кластеров необходимо предварительно их правильно обработать. С этой целью была проведена стандартизация для приведения данных в стандартный формат, в результате которого данные стали иметь нулевое среднее значение и единичную дисперсию [6-7]. В полученном наборе строк каждая представляет конкретную область на карте и содержит усредненную информацию о происшествиях. Итог нормализации записан в csv-файл для удобства работы.

6. Нахождение лучших разбиений

После получения набора регионов, представляющих собой улицы и перекрестки, следующей задачей является нахождение уровня опасности каждого. Для понимания, какие являются более опасными, а какие менее, нужно провести кластеризацию с учетом найденных ранее характеристик каждой области.

Для этих целей воспользовались методом кластеризации К-средних. Метод был выбран за счет своей простоты и хорошей масштабируемости. Приведенный способ старается так подобрать центры, чтобы критерий суммы квадратов внутри кластера был минимально возможным [8-9].

В результате получаются области. Причем никаких шумовых точек не образуется в процессе использования кластеризации методом К-средних – все области будут классифицированы и им присвоится определенный уровень опасности [10].

Для того, чтобы провести эту кластеризацию, необходимо определить оптимальное число кластеров. С этой целью рассмотрен определенный диапазон кластеров, от 2 до 5, и для каждого значения из диапазона проведена кластеризация. Кластерами выступают уровни опасности, поэтому значения являются небольшими. Чтобы определить какое из них привело к более эффективному результату, снова вычислялся коэффициент силуэта.

На основании результирующих данных был построен график зависимости значения метрики от числа заданных кластеров. График представлен на рис.4.

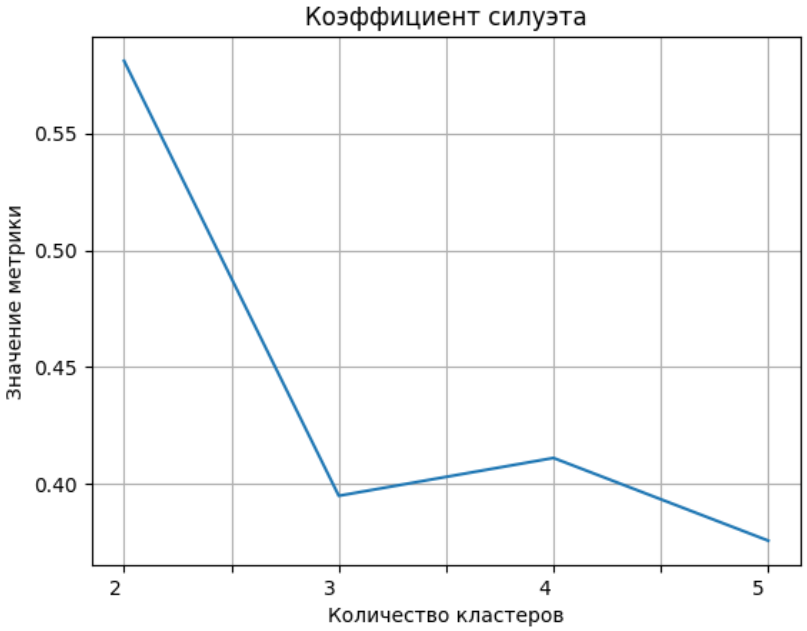

Рис.4. График зависимости значения метрики от количества заданных кластеров

Рассмотрим лучшие значения количества кластеров, которыми являются 2, 3 и 4 кластеров. Для каждого из них проведена кластеризация и построен график для визуальной оценки результатов работы. На рис.5 представлено визуальное выделение кластеров.

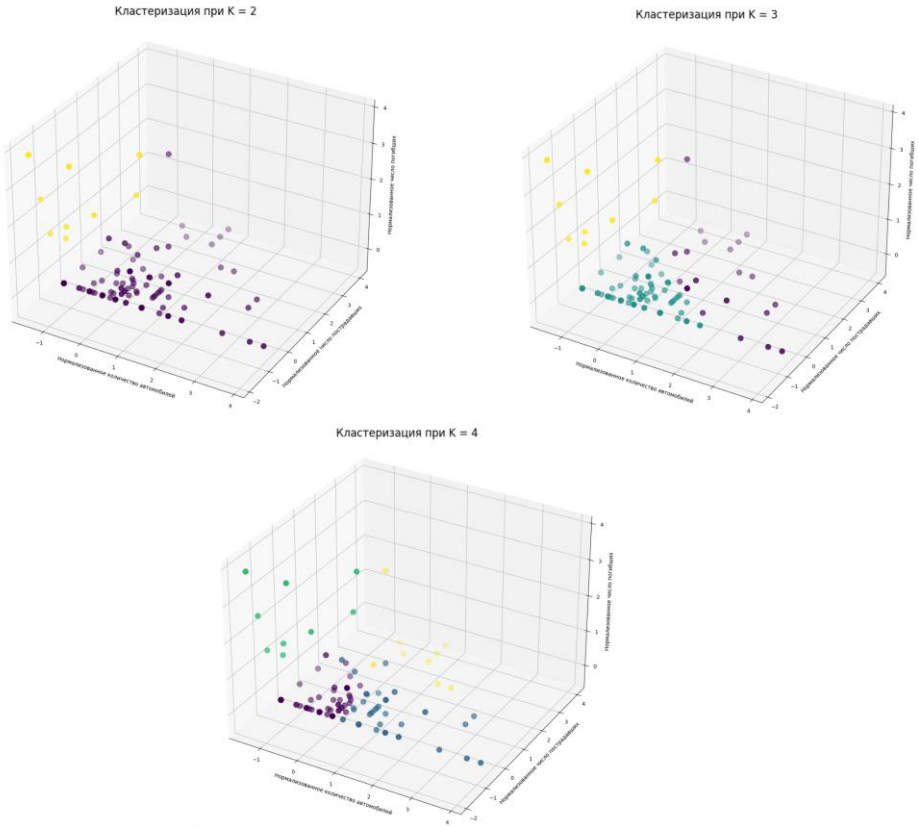

Рис.5. Выделение кластеров опасностей

Несмотря на то, что самым оптимальным значением количества кластеров с точки зрения оценки метрикой является минимальное значение, два кластера, такое распределение не является информативным. По графику видно что деление содержит очень широкий спектр характеристик: как областей, где в происшествиях участвовали малое количество машин и почти отсутствуют пострадавшие, так и те, в которые имеют большие значения данных параметров. Все это будет мешать корректно отделить очень опасные участки от менее. Разбиение же на большее количество кластеров будет путать пользователей, так как разница опасностей областей будет минимальна, а должна быть существенной и отличимой.

График с разбиение на три кластера выглядит наиболее подходящим под поставленные требования, проанализируем его более детально. Самым опасным является фиолетовый кластер, так как содержит наибольшее число участвующих в аварии транспортных средств, а также количество погибших. Малое количество участвующих автомобилей, но большое количество погибших в желтом кластере говорит о том, что такие области представляют большую опасность для пешеходов, но не представляют для машин. Бирюзовый кластер является наименее представляющим опасность, он содержит как пострадавших, так и погибших, но относительно других областей их значение значительно меньше.

Так как по этим данным требуется строить области на карте, то для правильного выделения контуров областей, все точки каждого кластера были представлены в виде альфа-формы – набора точек, представляющих геометрическую форму объекта.

Проделав все необходимые операции, сохранили полную информацию о кластерах в файл, для вывода на карту города. Файл содержит всю информацию о кластере: границы, а также его уровень опасности.

7. Отображение кластеров на карте

Для отображения областей на карте использовался инструмент для разработчиков – API «Яндекс Карты». Это решение имеет расширенный функционал и подробную документацию по сравнению с другими. Использование интерактивной карты позволяет упростить взаимодействие с информацией.

После получения данных на клиентской стороне, отобразили области на карте, представляющие кластеры происшествий. Для этого воспользовались встроенным классом построения полигонов, задав список всех граничных точек региона и некоторые настройки представления, в числе которых цвет и прозрачность области.

Цвет является индикатором уровня опасности: самые опасные области выделены ярко-красным цветом, а наименее – бледно-красным. Участки, представляющие наибольшую опасность для пешеходов, выделены бордовым цветом. Для того, чтобы увидеть сформированный результат необходимо

2023, №4
ISSN 2541-9269

открыть получивший html-файл в любом браузере. Части интерактивной карты с выделенными областями представлены на рис.6 и рис.7.

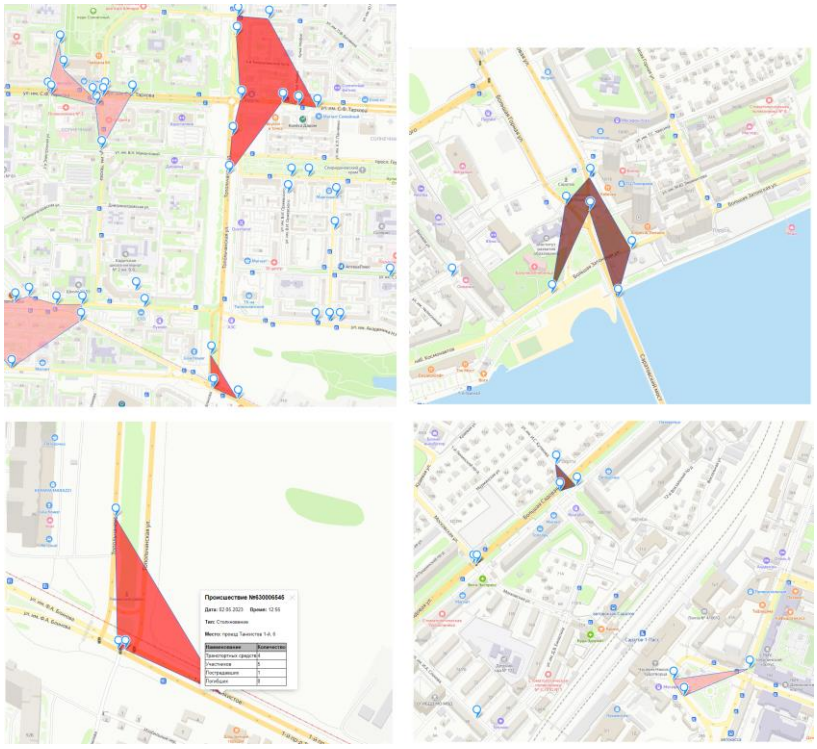

Рис.6. Фрагменты интерактивной карты с выделенными областями

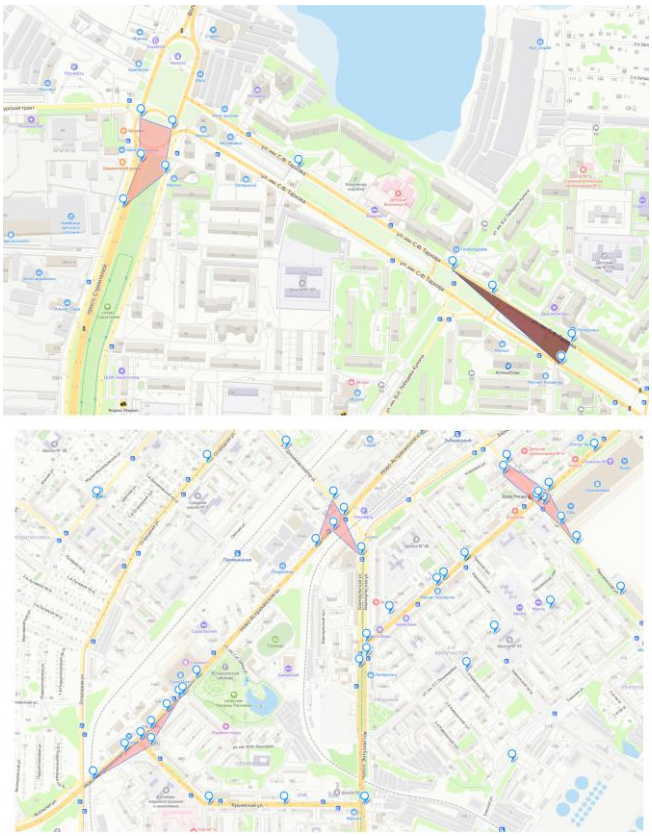

Рис.7. Фрагменты интерактивной карты с выделенными областями

8. Выводы

В работе был описан механизм сбора данных об авариях, а также нахождения регионов концентрации ДТП и определения уровня их опасности с использованием различных алгоритмов кластеризации. Созданное решение позволяет выделить и с помощью цветовой индикации разделить по группам области опасности, на которое в дальнейшем специализированные службы смогут обратить свое внимание.

Литература

1. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011.
2. Урвачева В.А. Обзор методов информационного поиска // Вестник Таганрогского института имени А. П. Чехова. 2016. №1. С. 457-467.
3. Митчелл Р. Скрапинг веб-сайтов с помощью Python. М.: ДМК Пресс, 2016.
4. Chauhan N.S. DBSCAN Clustering Algorithm in Machine Learning [Электронный ресурс]. URL: https://www.theaidream.com/post/dbscan-clustering-algorithm-in-machine-learning (Дата обращения 21.10.2023). Загл. с экр. Яз. англ.
5. Сивоголовко Е.В. Методы оценки качества чёткой кластеризации // Компьютерные инструменты в образовании. 2011. № 4. С. 14–31.
6. Бантикова О.И. Методы кластерного анализа. Оренбург: ГОУ ОГУ, 2011.
7. Тюрин А.Г., Зуев И.О. Кластерный анализ, методы и алгоритмы кластеризации // Вестник МГТУ МИРЭА. 2014. № 2 (3). С. 86–97.
8. Wu X., Kumar V. The Top Ten Algorithms in Data Mining. CRC Press, 2009.
9. Копец Д. Классические задачи Computer Science на языке Python. СПб.: Питер, 2020.
10. Sculley D. Web-scale k-means clustering // Proceedings of the 19th International Conference on World Wide Web. New York, NY, USA: Association for Computing Machinery, 2010. Pp. 1177–1178.