Компьютерные инструменты в образовании, 2021
№ 1: 45–57
УДК: 004.622
http://cte.eltech.ru
doi:10.32603/2071-2340-2021-1-46-58

информационные системы

ВЫЯВЛЕНИЕ УЧАСТКОВ ПОВЫШЕННОЙ ОПАСНОСТИ НА ДОРОГАХ МАССАЧУСЕТСА В 2013–2018 ГОДАХ

Герштейн А. М.[1], аспирант, ✉ ArkadyGer@gmail.com
Терехов А. Н.[1], доктор физико-математических наук, профессор, ✉ a.terekhov@spbu.ru

[1]Санкт-Петербургский государственный университет,
Университетская набережная, д. 7–9, 199034, Санкт-Петербург, Россия

Аннотация

Для выявления участков повышенной опасности на дорогах штата Массачусетс применяется метод кластеризации DBSCAN. Исследуются серьезные (то есть приведшие к летальному исходу или травмам) дорожно-транспортные происшествия (ДТП) с 2013 по 2018 годы. Алгоритм DBSCAN был также применен к набору равномерно распределенных по дорожной сети точек для определения порога в численности ДТП, после которого кластер можно считать статистически достоверным. Было произведено сравнение двух метрик расстояния: эвклидовой и сетевой. Показано, что обе метрики эквивалентны, если минимальное расстояние между отдельными ДТП в кластере не превышает 10 метров. Последний результат позволяет обосновать гибридный метод кластеризации, применимый для нахождения участков повышенной опасности на дорогах: для нахождения компактных кластеров можно использовать обычные эвклидовы расстояния между ДТП, а дорожную сеть использовать только для генерации равномерно распределенных по сети точек, нужных для выявления достоверных кластеров методом Монте-Карло. Гибридный метод позволяет обработать десятки тысяч ДТП, располагая сравнительно скромными вычислительными ресурсами. Анализ кластеров, выявленных на протяжении нескольких последовательных лет, позволяет сделать вывод об их стабильности и прогностической ценности.

Ключевые слова: *транспорт, ДТП, кластер, DBSCAN, статистическое испытание метод Монте-Карло, Массачусетс.*

Цитирование: Герштейн А. М., Терехов А. Н. Выявление участков повышенной опасности на дорогах Массачусетса в 2013–2018 годах // Компьютерные инструменты в образовании. 2021. № 1. С. 45–57. doi: 10.32603/2071-2340-2021-1-46-58

1. ВВЕДЕНИЕ

Цель данного исследования состоит в том, чтобы найти компактные участки повышенной опасности (УПО) на дорогах, то есть места со статистически достоверной повышенной плотностью серьезных дорожно-транспортных происшествий (ДТП), которые не меняют своего положения в течение нескольких последовательных лет.

Опасность таких участков обусловливается либо дорожной структурой (места пересечения дорог, въезд/выезд на шоссе), либо другими факторами (качеством дорожных покрытий, плохой видимостью и т. д.). Все эти факторы могут быть в известной степени

скорректированы муниципальными или государственными органами, более того, этих опасных участков можно избежать, если исключить их из маршрутов коммерческого и частного транспорта.

Обнаружение УПО на земной поверхности является важным шагом в изучении явлений различной природы. Например, места с повышенной плотностью преступлений выявляют наиболее опасные районы в качестве основной цели для полиции и других организаций, чья деятельность направлена на снижение преступности [1, 2].

Существует несколько методов обнаружения УПО: ядерные оценки плотности KDE [2, 4], I-статистика Морана [5], статистика Getis-Ord Gi* [7], а также различные алгоритмы кластеризации. KDE использует различные ядерные функции для преобразования точек на поверхности (например мест совершения преступлений) в некоторую гладкую функцию в попытке восстановить плотность распределения этих точек. KDE идентифицирует опасные участки на плоскости, но не может оценить их статистическую значимость. Статистики Moran's I и Getis-Ord Gi* позволяют выявлять статистически значимые области (кластеры). В соответствии с их природой нелегко использовать KDE, getis-ord GI* или статистику Морана для обнаружения опасных участков заданного размера. С другой стороны, некоторые алгоритмы кластеризации, такие как DBSCAN [3], могут быть легко настроены для обнаружения кластеров с заданным максимальным расстоянием между точками.

До сих пор мы не касались важной особенности ДТП, а именно того, что ДТП происходят не в евклидовом пространстве, а внутри дорожной сети, которая представляет собой, по существу, одномерный объект с иными представлениями о расстоянии. В случае KDE для дорожной сети следует разработать новую одномерную ядерную функцию [6]. В случае кластеризации нам нужна матрица, которая хранит расстояния между всеми ДТП, вычисленные по дорожной сети. К счастью, эта матрица расстояний между точками, принадлежащими дорожной сети, может быть получена с помощью программного пакета SANET [14].

Существует еще одна сложность в обнаружении УПО, принадлежащих дорожной сети: отсутствие соответствующей статистики. Статистики Getis-Ord Gi* и Moran's I работают только в 2d-случае [2], для методов кластеризации подобные статистики весьма редки. Поэтому для статистического обоснования полученных кластеров мы используем методы статистических испытаний [9]. Пакет SANET позволяет относительно просто генерировать огромное количество точек, равномерно распределенных по сети. Таким образом, наш план состоит в том, чтобы выполнить кластерный анализ по выбранным ДТП с использованием расстояний по сети, а затем использовать ту же сеть и то же, что и в реальных данных, количество равномерно распределенных точек для проведения статистических испытаний, по меньшей мере, несколько сот раз, чтобы получить статистику размеров кластеров. Затем мы сравним распределение размеров для реальных и моделируемых кластеров ДТП. В результате будут обнаружены статистически обоснованные кластеры.

2. ВХОДНЫЕ ДАННЫЕ

В данном исследовании мы используем дорожные сети, предоставленные департаментом транспорта Массачусетса [10] в формате Esri shapefile, очень удобном для визуализации с помощью ГИС-приложений, таких как QGIS [11] и OpenJUMP [12]. Также мы использовали данные портала MassDOT [13] о ДТП в штате Массачусетс за 2013–2018 годы в формате электронной таблицы .csv, который легко преобразуется в другие форматы,

например в тот же Esri shapefile.

Как обычно, полученные данные должны быть предварительно обработаны. Изолированные фрагменты дорожной сети (если они есть) должны быть соединены с основной сетью. Для связи фрагментов с основной сетью можно использовать плагин Disconnected Islands и инструменты редактирования QGIS. В файлах, содержащих сведения о ДТП, нужно оставить только те записи, где присутствуют координаты ДТП.

Поскольку сведения о ДТП и файлы, в которых хранится дорожная сеть, получены нами из разных источников, отдельные ДТП не принадлежат в точности элементам сети. Между тем, для некоторых видов анализа важно, чтобы ДТП и дорожная сеть были единым целым. Поэтому точки ДТП должны быть спроецированы на элементы сети, для чего мы использовали плагин QGIS NNJoin, который создает дополнительный слой в QGIS, где сами координаты ДТП не меняются, но появляются дополнительные атрибуты, хранящие начальные и конечные координаты ближайшей к ДТП прямой линии, принадлежащей сети. Эти координаты позволяют спроектировать точку ДТП на соответствующую линию дорожной сети с помощью простого Python-скрипта.

3. МЕТОДЫ ИССЛЕДОВАНИЯ

3.1. Матрица внутрисетевых расстояний (масштаб 20 м)

Подготовив данные, попробуем (с целью подбора алгоритма кластеризации и его параметров) обнаружить зоны повышенной опасности в городе Ньютон, штат Массачусетс, используя данные о серьезных ДТП за 2013 год, всего 369 случаев (см. рис. 1).

В качестве алгоритма кластеризации выберем DBSCAN, поскольку он ориентирован на выявление кластеров с повышенной плотностью точек и очень хорошо реализован в пакете Python sklearn. DBSCAN использует матрицу расстояний между точками или сами координаты (в случае евклидовых расстояний), а также два параметра: eps — максимальное расстояние между точками в кластере и min_samples — минимальное количество точек в кластере. В нашей первой попытке кластеризации будем использовать матрицу сетевых расстояний, вычисленную с помощью пакета SANET, и параметры eps=20 м и min_samples = 3 алгоритма DBSCAN. Выбор параметра eps=20 м представляется разумным, поскольку кластер, в котором минимальное расстояние между точками равно 20 м, достаточно компактен, чтобы располагаться на одной дороге, и, следовательно, его легко можно избежать, если использовать корректный алгоритм маршрутизации.

В результате были получены 63 кластера размерами от 3 до 9 ДТП. Для выделения из этих кластеров статистически значимых сформулируем нулевую гипотезу: ДТП распределены равномерно по дорожной сети. Будем считать, что кластер размером ν статистически значим на уровне α, если вероятность обнаружения хотя бы одного кластера размера, большего или равного ν (в случае, если нулевая гипотеза верна), меньше α [9]. Учитывая, что статистики размеров кластеров в случае алгоритма DBSCAN не существует, мы провели ряд статистических испытаний методом Монте-Карло. С помощью пакета SANET были получены 1024 выборки равномерно распределенных по дорожной сети Ньютона точек (каждая выборка содержит 369 точек, как и в исходных данных), затем эти выборки были использованы для вычисления 1024 матриц расстояний по дорожной сети, и далее к каждой матрице был применен алгоритм DBSCAN с теми же параметрами, как и в случае реальных данных (eps=20 и min_samples=3). Результаты моделирования приведены в таблице 1.

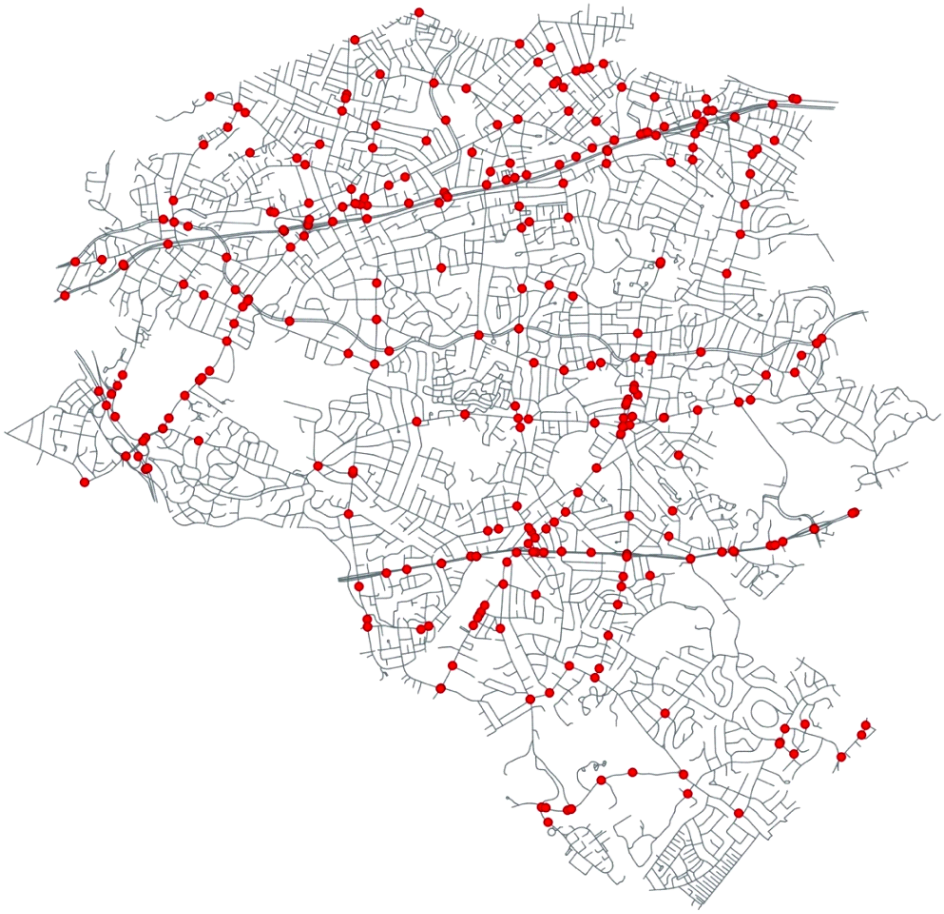

Рис. 1. Серьезные ДТП (Newton MA, 2013)

Таблица 1. Результаты моделирования (1024 испытания, дорожная сеть Ньютона, штат Массачусетс)

Размер	Количество кластеров одинакового или большего размера	P
3	59	0,058
4	0	0

Из таблицы 1 видно, что на уровне 0,05 значимы кластеры с $n \geqslant 4$. Удаление кластеров с $n = 3$ из 63 первоначально обнаруженных дает 5 кластеров.

3.2. Гибридный подход к кластеризации

Кластеризация, использующая матрицу сетевых расстояний, представленная в предыдущем разделе, требует очень больших вычислительных ресурсов и возможна (с учетом необходимости провести, по крайней мере, несколько сотен испытаний) лишь для небольших наборов данных, таких как данные о серьезных ДТП в Ньютоне[1].

[1] Речь, естественно, не идет о суперкомпьютере; все вычисления, результаты которых представлены в данной статье, проведены на рабочей станции с 32 гб оперативной памяти и четырехъядерным процессором.

Поэтому очень заманчиво выглядит замена сетевых расстояний между отдельными ДТП на соответствующие евклидовы. В этом случае алгоритм DBSCAN, реализованный в пакете sklearn (Python), вычисляет матрицу расстояний чрезвычайно быстро, и для масштаба 20 м кажется разумным, что расстояния (евклидовы и по сети) очень похожи. Повторив все вычисления — на этот раз с евклидовыми расстояниями — и получив новый набор статистически значимых кластеров, можно сравнить их с кластерами, полученными в разделе 3.1.

Сравнивая два набора кластеров, получим следующее:

— 5 сетевых кластеров расположены идентично своим евклидовым собратьям;
— 2 евклидовых кластера расположены иначе;
— идентичные кластеры могут сильно отличаться при увеличении масштаба.

Итак, рассматривая сетевую кластеризацию как «истинную», мы можем перечислить некоторые недостатки евклидовой кластеризации:

1. Ложные кластеры — не имеют сетевого аналога.
2. Испорченные кластеры — имеют истинные (те же, что в сетевом случае) и ложные фрагменты (на разных дорожных линиях или на разных дорогах).

Теперь мы можем суммировать два подхода к кластеризации с eps=20 в таблице 2.

Таблица 2. Ньютон 2013. Два метода кластеризации (евклидов и сетевой, eps=20)

Метод	№	% от общего числа серьезных ДТП	Испорченные кластеры число/процент	Ложные кластеры число/процент
Сеть	5	7,3	0/0	0/0
Евклидов	7	10	2/29	3/43

3.3. Масштаб 10 м

Из предыдущего раздела (таблица 2) мы видим существенную разницу между сетевым и евклидовым подходом к кластеризации в масштабе 20 м. Между тем, очевидно, что при некотором достаточно малом масштабе сетевое и евклидово расстояние должны быть идентичны. Конечно, этот масштаб меньше 20 м. Чтобы оценить этот максимальный масштаб, исследуем зависимость разности между евклидовым и сетевым расстояниями от, скажем, евклидовых расстояний и найдем точку расхождения, где евклидовы расстояния становятся неадекватны нашей задаче (рис. 2).

Как видно из рис. 2, сетевое расстояние всегда не меньше евклидова, потому что прямая линия, соединяющая две точки, имеет минимальную длину в евклидовом мире. Также видно, что расхождение между евклидовым и сетевым расстояниями начинается с масштаба 10 м. На графике хорошо заметны большие всплески вблизи 15 м, и это, скорее всего, искажает евклидову кластеризацию с eps = 20. Таким образом, мы можем выбрать 10 м в качестве масштаба, где евклидово расстояние, требующее гораздо меньших вычислений, может заменить расстояние по дорожной сети.

Значение 10 м кажется очень маленьким, но на самом деле это не так. Оно лишь немного меньше ширины 3-полосного шоссе (в Соединенных Штатах каждая полоса имеет ширину примерно 12 футов или 3,6 м. Иными словами, 10 метров — это естественный масштаб для движения транспортных средств в одном направлении по дорожной сети.

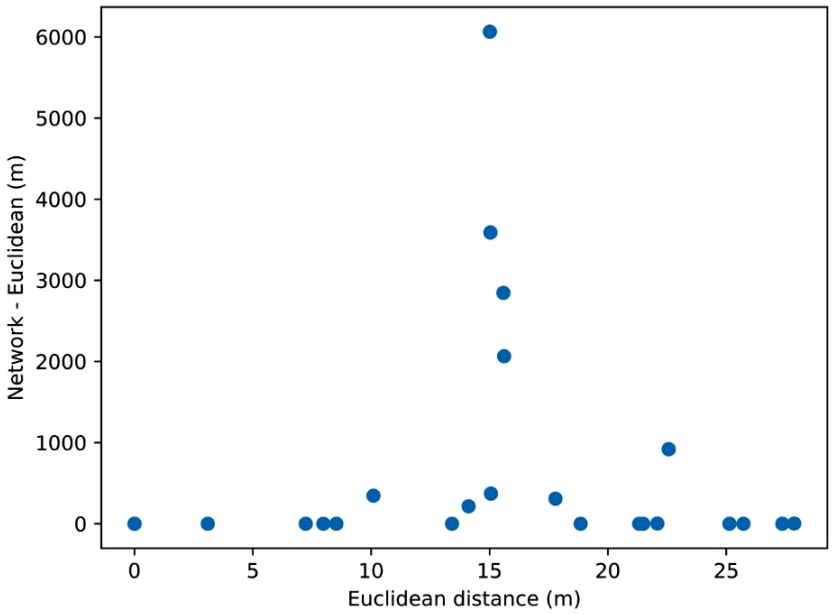

Рис. 2. Ньютон, штат Массачусетс, 2013 год, серьезные ДТП. Зависимость разницы между евклидовыми и сетевым расстояниями, от евклидовых расстояний

3.4. Гибридная кластеризация

Глядя на рис. 2 и принимая во внимание соображения о натуральном для дорожной сети масштабе, приведенные в конце предыдущего раздела, можно обосновать гибридный метод кластеризации: использовать дорожную сеть только для генерации наборов равномерно распределенных точек и выполнять всю кластеризацию (для обнаружения реальных кластеров и для статических испытаний) с eps=10, используя евклидовы расстояния между ДТП.

4. СОДЕРЖАТЕЛЬНЫЙ ПРИМЕР

4.1. Массачусетс 2013, кластеризация серьезных ДТП

Согласно имеющимся данным, в штате Массачусетс в 2013 году зарегистрировано 30696 серьезных ДТП, из них 23964 (78 %) на крупных дорогах.

Применение кластеризации с eps=10 и min_samples=3 к 23964 серьезным ДТП дает 1340 кластеров, из которых нужно выделить статистически значимые. Для этого кластеризация была выполнена 1502 раза на наборах из 23964 равномерно распределенных по дорожной сети точек, сгенерированных программным пакетом SANET. Результаты испытаний методом Монте-Карло приведены в таблице 3.

Таким образом, на уровне 0,05 мы должны рассматривать кластеры размером $\geqslant 5$ как статистически значимые. Удаляя из первоначальных кластеров те, чей размер меньше 5, получим 354 кластера, общее число серьезных ДТП в кластерах составляет 2301, то есть 9,6 % от общего числа серьезных ДТП.

На рис. 3 показаны статистически значимые кластеры серьезных ДТП, случившихся в Массачусетсе в 2013 году.

Таблица 3. Результаты моделирования методом Монте-Карло для Массачусетса, 2013 год (1502 испытания, 23964 точки в каждом)

Размер	Количество кластеров одинакового или большего размера	P
3	1502	1
4	160	0,1
5	6	0,004

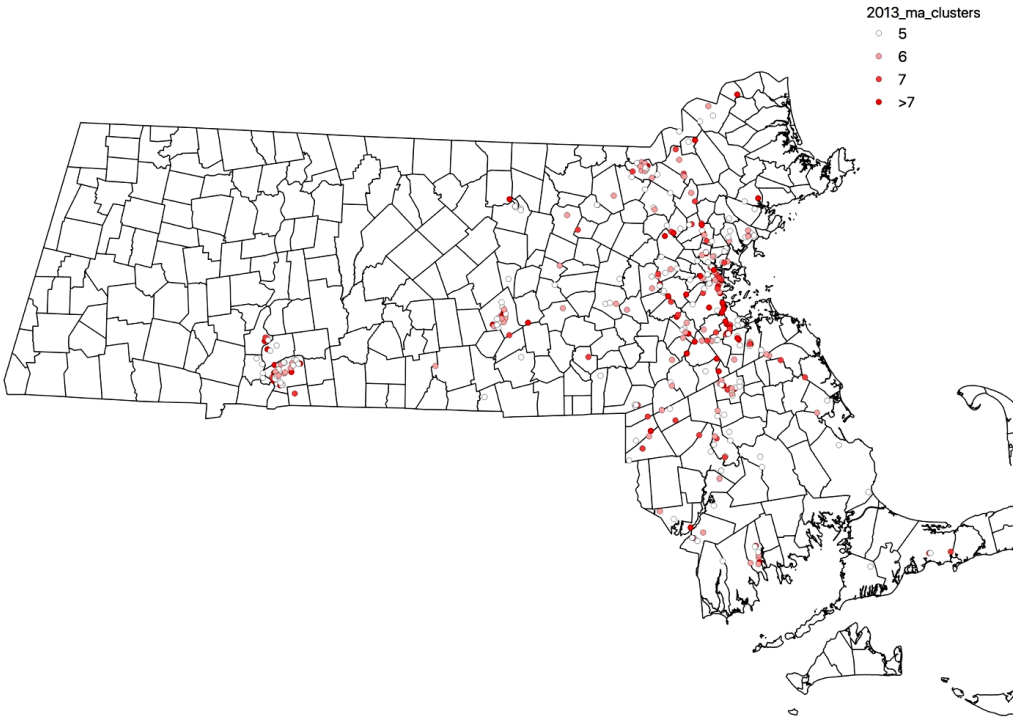

Рис. 3. Массачусетс 2013, значимые кластеры серьезных ДТП (DBSCAN, eps=10)

Как видно из рис. 3, пространственное распределение кластеров по всему штату далеко не однородно. В тринадцати городах найдены 166 (47 %) кластеров (таблица 4). Ньютон с 3 кластерами занимает 38-е место в общем списке и включен в таблицу только потому, что мы подробно изучали его в Разделе 3.

Из рис. 3 также видно, что большинство кластеров имеют размеры от 5 до 7.

4.2. Мета-кластеры в Массачусетсе. Гибридный подход снизу вверх

В предыдущем разделе мы проанализировали кластеры (УПО) в Массачусетсе, используя данные о серьезных дорожно-транспортных происшествиях за 2013 год. Наша следующая цель — использовать данные за годы 2013–2018 для выявления стабильных во времени кластеров, то есть мест, которые остаются опасными как минимум несколько лет. Наш подход заключается в использовании двухэтапной кластеризации, то есть получении значимых кластеров за каждый год, а затем кластеризации этих кластеров еще раз с параметром min_samples = <минимальное число лет>. Например, min_samples = 4 означает, что результирующий кластер содержит данные за 4 разных года.

Таблица 4. Города Массачусетса с максимальным количеством кластеров

Город	Кластеры	Число ДТП
Springfield	38	234
Boston	24	209
Worcester	23	148
Lowell	15	93
Brockton	14	87
Braintree	8	68
New Bedford	9	55
Quincy	7	53
Woburn	5	39
Weymouth	5	36
Holyoke	5	35
Raynham	7	35
Waltham	6	35
Newton	3	19

Как и в разделе 4.1, получим сначала все (в том числе статистически незначимые) кластеры за 6 лет: 2013, 2014, 2015, 2016, 2017, 2018. Затем проведем испытания методом Монте-Карло, чтобы выделить значимые кластеры. Результаты испытаний методом Монте-Карло за все 6 лет собраны в таблице 5.

Таблица 5. Результаты испытаний методом Монте-Карло за шесть лет, DBSCAN eps=10, равномерно распределенные по дорожной сети случайные точки

Год	Всего ДТП	Количество испытаний	Размер кластера	Количество кластеров данного или большего размера	P
2013	23964	1502	3	1502	1,0
			4	160	0,11
			5	6	0,004
2014	24921	1444	3	1442	0,99
			4	171	0,12
			5	6	0,004
2015	25959	1386	3	1384	0,99
			4	208	0,15
			5	11	0,008
2016	26772	1344	3	1344	1,0
			4	214	0,16
			5	8	0,006
			6	1	0,0007
2017	26672	1349	3	1349	1,0
			4	206	0,15
			5	9	0,007
2018	24763	1453	3	1453	1,0
			4	201	0,14
			5	7	0,005

После удаления статистически незначимых кластеров объединим кластеры за все 6 лет, добавив метку года к каждому кластеру, а затем выполним вторую кластеризацию с eps=10 и параметром min_samples=3. Некоторые статистические характеристики

метакластеров показаны в таблице 6. Количество ДТП во всех метакластерах — 7456 и составляет 4,8 % от 153051 — общего количества ДТП, используемых для двухступенчатой кластеризации. На рис. 4 показано распределение кластеров на карте Массачусетса. Очевидно, пространственное распределение метакластеров аналогично 2013 году.

Таблица 6. Статистика метакластеров

Разные годы в кластере	Кластеров	ДТП	ДТП / кластер	Медиана ДТП	Стандартное отклонение ДТП
23	115	2269	19,7	19	3,29
4	74	1979	26,7	25	5,77
5	36	1412	39,2	36	10,68
6	29	1796	61,9	56	21,71
Всего	254	7456	29,3	24	16,23

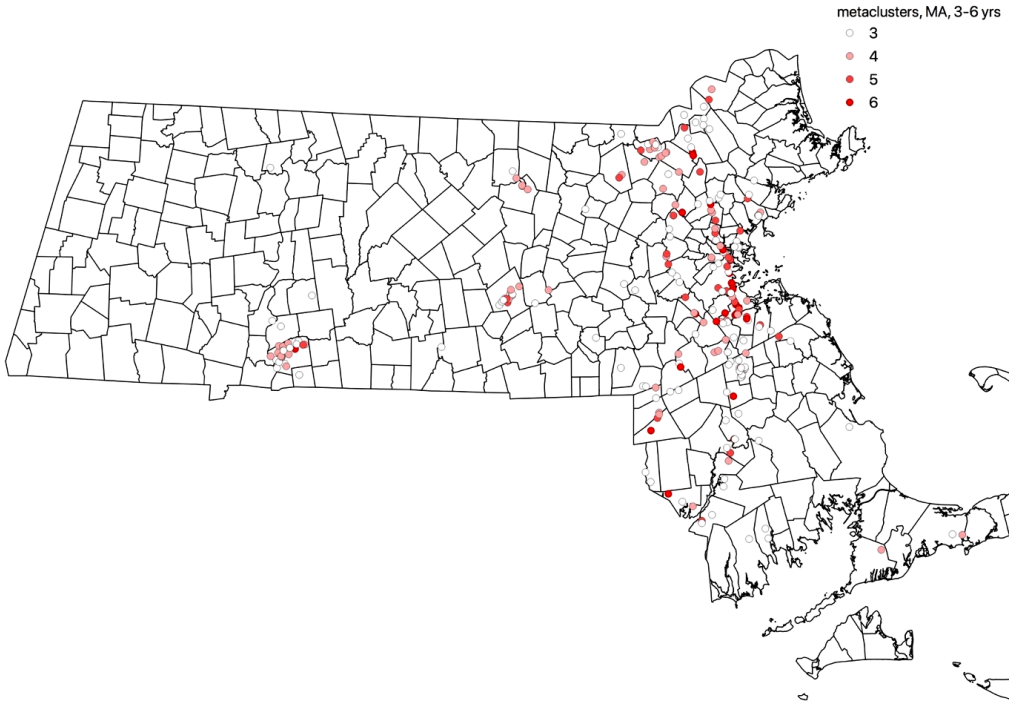

Рис. 4. Мета-кластеры в Массачусетсе, 3–6 разных лет в каждом кластере

Кроме статистических характеристик метакластеров, содержащих различное число лет, когда произошли ДТП, весьма интересны характеристики повторяемости кластеров из года в год. Можно, например, для каждого кластера ДТП, случившихся в одном году, найти (если он существует) пространственно эквивалентный кластер ДТП, случившихся в другом году. Результаты представлены в таблице 7, которая напоминает матрицу расстояний и содержит относительное количество кластеров, наблюдавшихся в течение одного года и оставшихся в последующие годы. Например, 0,36 на пересечении строки 2018 и столбца 2017 означает, что отношение (количество кластеров в 2017 году, найденных в 2018 году / общее количество кластеров в 2017 году) равно 0,36. Общее количество значимых кластеров в соответствующем году показано в колонке «Кластеры».

Таблица 7. Относительная повторяемость кластеров из года в год

	2017	2016	2015	2014	2013	Кластеры
2018	0.36	0,35	0,35	0,33	0,36	378
2017		0,35	0,35	0,34	0,32	369
2016			0,39	0,33	0,3	374
2015				0,3	0,33	334
2014					0,36	349
2013						354

Интересно также выяснить, как повторяемость кластеров зависит от годового интервала. В таблице 8 данные таблицы 7 представлены несколько иначе. Столбец «1 год» содержит «расстояния» в один год между наборами кластеров (то есть повторяемость кластеров за 2013–2014, 2014–2015 и так далее), столбец «2 года» представляет данные о двухлетней повторяемости (то есть 2013–2015, 2014–2016 и так далее). Последняя колонка («3 года») представляет данные для трехлетней повторяемости. Последние две строки таблицы содержат элементарную статистику повторяемости для разных годичных интервалов. Видно, что «расстояние» между наборами кластеров остается относительно постоянным, незначительно уменьшаясь с увеличением годового интервала. По-видимому, дорожные условия для исследуемых ДТП существенно не меняются из года в год, то есть относятся к одному распределению и могут быть объединены при необходимости для дальнейшего анализа.

Таблица 8. Повторяемость кластеров для различных временных интервалов

1 год: 13–14, 14–15, 15–16, 16–17, 17–18 **2 года:** 13–15, 15–17, 14–16, 16–18 **3 года:** 13–16, 14–17, 15–18	**1 год**	**2 года**	**3 года**
	0,36	0,33	0,3
	0,3	0,33	0,34
	0,39	0,35	0,35
	0,35	0,35	
	0,36		
Среднее	0,35	0,34	0,33
Стандартное отклонение	0,03	0,01	0,02

5. ВЫВОДЫ

В данной работе показано, что гибридный метод кластеризации DBSCAN с использованием евклидова расстояния, применяемый к данным о серьезных ДТП в штате Массачусетс, в сочетании с моделированием методом Монте-Карло на дорожной сети штата способен идентифицировать статистически значимые УПО серьезных ДТП. Эти участки (кластеры) компактны, причем примерно треть УПО повторяется в следующем году. Выявленные участки повышенной опасности серьезных ДТП желательно избегать водителям, а городским и государственным органам необходимо использовать информацию о таких участках для планирования мероприятий, имеющих целью снижение травматизма и смертности на дорогах.

Дальнейшие исследования могут быть полезны для выявления пространственных и иных особенностей УПО (например перекресток, въезд/выезд на шоссе, соединение дорог, ширина дороги, ограничение скорости, плохая видимость, время дня, погодные условия и т. д.).

Список литературы

1. *Chainey S, Tompson L, Uhlig S.* The utility of hotspot mapping for predicting spatial patterns of crime // Secur J. 2008. Vol. 21, № 1. P. 4–28. doi: 10.1057/palgrave.sj.8350066
2. *Chainey S., Ratcliffe J.* GIS and Crime Mapping. UK, Chichester: John Wiley and Sons, 2005. doi: 10.1002/9781118685181
3. *Ester M., Kriegel H-P., Sander J., Xu X.* A density-based algorithm for discovering clusters in large spatial databases with noise // Proc. KDD'96: 2nd Int. Conf. on Knowledge Discovery and Data Mining, 1996. P. 226–231.
4. *Gramacki A.* Nonparametric Kernel Density Estimation and Its Computational Aspects. Cham, Switzerland: Springer International Publishing, 2018. doi: 10.1007/978-3-319-71688-6
5. *Moran P. A. P.* Notes on Continuous Stochastic Phenomena // Biometrika. 1950. Vol. 37, № 1/2. P. 17–23. doi: 10.2307/2332142
6. *Okabe A., Sugihara K.* Spatial Analysis along Networks: Statistical and Computational Methods. NJ, USA, Hoboken: John Wiley & Sons: 2012. doi: 10.1002/9781119967101
7. *Songchitruksa P., Zeng X.* Getis–Ord Spatial Statistics to Identify Hot Spots by Using Incident Management Data // Transportation Research Record: Journal of the Transportation Research Board. 2010. № 2165. P. 42–51. doi: 10.3141/2165-05
8. *Yingjie L., Liwei Zh., Junping Y., Pengtao W., Ningke H., Wei Ch., Bojie F.* Mapping the hotspots and coldspots of ecosystem services in conservation priority setting // Journal of Geographical Sciences. 2017. Vol. 27, № 6. P. 681–696. doi: 10.1007/s11442-017-1400-x
9. *Xie Y., Shekhar S.* Significant DBSCAN towards Statistically Robust Clustering // Proc. SSTD'19: 16th International Symposium on Spatial and Temporal Databases, 2019. P. 31–40. doi: 10.1145/3340964.3340968
10. Massgis data-massachusetts department transportation massdot roads // docs.digital.mass.gov. URL: https://docs.digital.mass.gov/dataset/massgis-data-massachusetts-department-transportation-massdot-roads (дата обращения: 10.03.2021).
11. QGIS // www.qgis.org, URL: https://www.qgis.org/en/site/ дата обращения: 10.03.2021).
12. Open jump // www.openjump.org. URL: http://www.openjump.org/ дата обращения: 10.03.2021).
13. MassDOT Crash Open Data Portal // Mass.gov. URL: https://massdot-impact-crashes-vhb.opendata.arcgis.com/searchдата обращения: 10.03.2021).
14. SANET //sanet.csis.u-tokyo.ac.jp. URL: http://sanet.csis.u-tokyo.ac.jp/ дата обращения: 10.03.2021).

Поступила в редакцию 15.02.2021, окончательный вариант — 18.03.2021.

Герштейн Аркадий Михайлович, аспирант математико-механического факультета СПбГУ, ✉ ArkadyGer@gmail.com

Терехов Андрей Николаевич, доктор физико-математических наук, профессор, заведующий кафедрой системного программирования математико-механического факультета СПбГУ, ✉ a.terekhov@spbu.ru

Computer tools in education, 2021
№ 1: 45–57
http://cte.eltech.ru
doi:10.32603/2071-2340-2021-1-46-58

Hotspots of Traffic Accidents that Cause Injuries or Death in Massachusetts from 2013 to 2018

Gershtein A. M.[1], Postgraduate, ✉ ArkadyGer@gmail.com
Terekhov A. N.[1], PhD, Professor, ✉ a.terekhov@spbu.ru

[1]State University, 7–9, Universitetskaya emb., 198504, Saint Petersburg, Starii Petergof, Russia.

Abstract

DBSCAN clustering method is applied to identify severe Traffic Accident (TA) hotspots on roads. The research examines severe TA, defined as those that led to human damage (injury or death), in the city of Newton, MA and in the entire state of Massachusetts, USA from 2013 to 2018. DBSCAN algorithm was also applied to network-constrained uniformly distributed over road network data to locate threshold in number of points per cluster so that all more populated clusters identified in real data can be treated as statistically significant. For DBSCAN algorithm two types of distance metrics, Euclidean and over Network, were compared. It is found that both distances are equivalent on scale of 10 meters, which justifies hybrid approach to clustering: using Network distance only to generate uniformly distributed points needed for Monte-Carlo simulations. All clustering can be performed using Euclidean distances which is much faster and more memory efficient. Subsequent years analysis demonstrates the extend that hotspots identified are stable and occur consecutively for several years and hence may possess predictive value.

Keywords: *vehicle traffic accident hotspot, cluster, DBSCAN, simulation Monte-Carlo, Massachusetts.*

Citation: A. M. Gershteyn and A. N. Terekhov “Hotspots of Traffic Accidents that cause injuries or death in Massachusetts from 2013 to 2018” Computer tools in education, no. 1, pp. 45–57, 2021 (in Russian); doi: 10.32603/2071-2340-2021-1-46-58

References

1. S. Chainey, L. Tompson, and S. Uhlig, “The utility of hotspot mapping for predicting spatial patterns of crime,” *Secur J*, vol. 21, no. 1, pp. 4–28, 2008; doi: 10.1057/palgrave.sj.8350066
2. S. Chainey and J. Ratcliffe, *GIS and Crime Mapping*, Chichester, UK: John Wiley and Sons, 2005; doi: 10.1002/9781118685181
3. M. Ester, H-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proc. KDD’96: 2nd Int. Conf. on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.
4. A. Gramacki, *Nonparametric Kernel Density Estimation and Its Computational Aspects*, Cham, Switzerland: Springer International Publishing, 2018; doi: 10.1007/978-3-319-71688-6
5. P. A. P. Moran, “Notes on Continuous Stochastic Phenomena,” *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950; doi: 10.2307/2332142
6. A. Okabe and K. Sugihara, *Spatial Analysis along Networks: Statistical and Computational Methods*, Hoboken, NJ, USA: John Wiley & Sons, 2012; doi:10.1002/9781119967101

7. P. Songchitruksa and X. Zeng, “Getis–Ord Spatial Statistics to Identify Hot Spots by Using Incident Management Data,” *Transportation Research Record*, no. 2165, pp. 42–51, 2010; doi: 10.3141/2165-05
8. L. Yingjie, et al., “Mapping the hotspots and coldspots of ecosystem services in conservation priority setting,” *Journal of Geographical Sciences*, vol. 27, no. 6, pp. 681–696, 2017; doi: 10.1007/s11442-017-1400-x
9. Y. Xie and S. Shekhar, “Significant DBSCAN towards Statistically Robust Clustering,” in *Proc. SSTD’19: 16th International Symposium on Spatial and Temporal Databases*, 2019, pp. 31–40; doi: 10.1145/3340964.3340968
10. “Massgis data-massachusetts department transportation massdot roads,” in *docs.digital.mass.gov*, [Online]. Available: https://docs.digital.mass.gov/dataset/massgis-data-massachusetts-department-transportation-massdot-roads
11. “QGIS,” in *www.qgis.org*, [Online]. Available: https://www.qgis.org/en/site/
12. “Open jump,” in *www.openjump.org*, [Online]. Available: http://www.openjump.org/
13. “MassDOT Crash Open Data Portal,” in *Mass.gov*, [Online]. Available: https://massdot-impact-crashes-vhb.opendata.arcgis.com/search
14. “SANET,” in *sanet.csis.u-tokyo.ac.jp*, [Online]. Available: http://sanet.csis.u-tokyo.ac.jp/

Received 15-02-2021, the final version — 18-03-2021.

Arkadiy Gershtein, Postgraduate of the Faculty of Mathematics and Mechanics, SPbSU, ✉ ArkadyGer@gmail.com

Andrey Terekhov, PhD, Professor, Head of the Department of System Programming of the Faculty of Mathematics and Mechanics, SPbSU, ✉ a.terekhov@spbu.ru