MAPCLUST - СЕРВИС ДЛЯ ЭПИДЕМИОЛОГИЧЕСКОГО И ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА

Смирнов Антон Сергеевич1,Захаров Олег Станиславович1,Пятницкий Алексей Михайлович1, Гукасов Вадим Михайлович2

1 - РНИМУ им. Н.И. Пирогова, Москва, 2 - ФГБНУ НИИ РИНКЦЭ, Москва

# Аннотация

Эта статья посвящена описанию разработанного авторами веб-сервиса для кластеризации пространственных и пространственно-временных данных MapClust. MapClust является практической реализацией исследований, проведенных нами в предыдущих работах. В этой работе мы демонстрируем применение сервиса для решения задач эпидемиологического и экологического мониторинга.

Ключевые слова: MapClust; эпидемиологический мониторинг; экологический мониторинг; алгоритм кластеризации.

# Введение

Поиск кластеров в пространственных данных является нетривиальной задачей, решение которой важно для мониторинга ситуации в некоторой изучаемой области. В текущий момент, в связи с неблагоприятной эпидемиологической и экологической ситуацией в мире остро стоит задача обнаружения областей, на которые в первую очередь необходимо обратить внимание для принятия мер. Один из наиболее популярных видов методов поиска кластеров, например, в анализе эпидемиологии туберкулеза – сканирующие статистики[7]. Самый известный метод сканирующей статистики – метод Куллдорфа[5]. Этот метод широко используется в современных эпидемиологических исследованиях[3; 4; 6]. Однако, метод зависит от формы и размера сканирующую окна, выдает один наиболее вероятный кластер и обладает довольно большой вероятностью ложноположительного результата, то есть в кластеры часто попадают области, которые не имеют никакого отношения к ним[8].

Для преодоления этих ограничений был разработан MapClust, новый веб-сервис для кластеризации пространственных и пространственно-временных данных. MapClust обладает рядом преимуществ по сравнению с существующими методами. Во-первых, он не зависит от формы и размера кластера, что позволяет более точно определять кластеры с различной структурой. Во-вторых, он позволяет пользователю регулировать чувствительность алгоритма, обеспечивая лучший контроль над балансом между ложноположительными и ложноотрицательными результатами. Наконец, MapClust разработан для одновременного поиска нескольких кластеров при сохранении заданной величины ошибок I рода, что позволяет получить более полное представление о данных.

Данная работа организована следующим образом: сначала мы описываем принципы работы алгоритма MapClust, включая расчеты для определения p-значений и уравнения, используемые для обнаружения и оценки кластеров. Далее мы приводим несколько примеров и приложений, демонстрирующих использование MapClust для задач эпидемиологического и экологического мониторинга. Наконец, в заключение мы суммируем ключевой вклад и выводы данного исследования, а также обсуждаем практические последствия применения метода MapClust в различных областях.

# Принцип работы

Данный сервис получает на вход данные о взаиморасположении областей исследования и о значении исследуемой величины в этой области. Критерий подразумевает, что исследуемая величина - частота события в каком-либо регионе. Данные должны быть представлены в специальном общеиспользуемом геоинформационном формате – ESRI Shapefile. Формат Shapefile - это популярный формат геопространственных векторных данных, разработанный компанией ESRI для использования в программном обеспечении географических информационных систем (ГИС). Shapefile используются в различных приложениях, включая анализ пространственных и пространственно-временных данных, картографию и геопространственную визуализацию. В контексте веб-сервиса MapClust формат Shapefile используется для предоставления входных данных о расположении интересующих областей и связанных с ними анализируемых значений. Благодаря использованию широко распространенного геопространственного формата Shapefiles, MapClust обеспечивает совместимость с различным программным обеспечением ГИС и упрощает подготовку и интеграцию данных для пользователей.

Принцип поиска кластеров описан в статьях[9–12]. Для каждого региона рассчитывается значение p для сравнения величин в регионах ([Equation 1](#eq-p-spat)) или для сравнения моментов времени ([Equation 2](#eq-p-time)), где - функция Лапласа. Для каждой области, для которой p меньше или выше некоторого порога, заданного пользователем, рассчитываются статистики для разряжений ([Equation 3](#eq-discharges-stat)) и для кластеров ([Equation 4](#eq-clust-stat)).

Программа проверяет вхождение значения статистик в критическую область S(n), где n – число регионов соседей, для которых p меньше или выше определенного порога. Программа строит эту область методом Монте-Карло. Кластеры или разрежения, выходящие за критическую область, являются значимыми.

# Примеры работы

Приведем несколько примеров работы. Для первого примера возьмем данные по заболеваемостью вирусом иммунодефицита человека (ВИЧ) в Российской Федерации (РФ) с портала ЕМИСС[2]. Веб-сервис может работать в двух режимах работы: пространственном и пространственно-временном. Пространственный режим заключается в поиске кластеров по величине, измерянной в регионах определенной местности. Пространственно-временной режим - поиск кластеров по разнице некоторой величины во времени. Для начала работы необходимо построить критическую область. Для этого необходимо загрузить все файлы из архива, выбрать количество итераций и задать пороговые значения и . Если или , программа соответственно будет искать разряжения или кластеры. Для примера возьмем , и и запустим критерий для в первом режиме для данных 2005 года и в втором режиме для данных 2005 и 2006 года. Результаты представлены на [Figure 1](#fig-first). Программа также позволяет получить список областей, составляющих кластеры или разряжения. Из полученных данных можно сделать выводы о регионах с благополучной и тревожной эпидемиологической обстановкой. Например, на [Figure 2](#fig-second) продемонстрированы регионы с неблагоприятной эпидемиологической обстановкой (кластеры). Нетрудно заметить, что обнаружено несколько кластеров неправильной формы и разного размера.

|  |
| --- |
| Figure 1: Критическая область для обнаруженных кластеров и разряжений |

|  |
| --- |
| Figure 2: Карта расположения статистически значимых кластеров. Границы кластеров обозначены черными линиями |

Также в практическом плане может быть гораздо полезнее смотреть динамику распространения инфекции. Наша программа позволяет сравнивать показатели. Сравним с теми же параметрами 2005 и 2006 год. Результат представлен на [Figure 3](#fig-third). Статистически значимое улучшение обстановки за год наблюдается только в Санкт-Петербурге.

|  |
| --- |
| Figure 3: Сравнение количества новых случаев заражения ВИЧ-инфекцией в РФ в 2005 и в 2006 годах. |

В качестве задачи экологического мониторинга рассмотрим поиск очагов ухудшения качества воздуха в Румынии. Данные взяты с сайта Европейского агентства по окружающей среде[1]. В качестве показателя, отображающего качество воздуха, возьмем концентрацию частиц диаметром меньше или равного 10 микрометрам (PM10). Рассмотрим, как изменилось качество воздуха с 2018 по 2021 год. Запустим критерий с параметрами , и . Результат представлен на [Figure 4](#fig-fourth). Наблюдается статистически значимое улучшение качества воздуха в районе Плоешти и Брашова, который исторически был промышленным в Румынии.

|  |
| --- |
| Figure 4: Сравнение PM10 в Румынии за 2018 и 2021 год. |

# Заключение

В этой работе был представлен веб-сервис MapClust, новый инструмент для эффективной кластеризации и анализа пространственных и пространственно-временных данных, специально разработанный для задач эпидемиологического, экологического и других видов мониторинга. Ключевым вкладом и результат работы являются разработка алгоритма, который преодолевает ограничения существующих методов, таких как статистика сканирования Куллдорфа, не зависящих от формы и размера кластера, предлагающих регулируемую чувствительность и возможность поиска нескольких кластеров при контроле величины ошибок первого типа.

На приведенных примерах мы продемонстрировали эффективность MapClust при выявлении регионов с различной эпидемиологической ситуацией, касающейся распространенности ВИЧ в Российской Федерации и мониторинга качества воздуха в Румынии. Эти практические приложения демонстрируют потенциальные преимущества использования MapClust в здравоохранении и управлении окружающей средой. Позволяя более точно определять области, требующие вмешательства, MapClust может помочь лицам, принимающим решения, более эффективно распределять ресурсы, что в конечном итоге приведет к улучшению состояния здоровья населения и охране окружающей среды.

Разработанный сервис доступен по адресу https://epidemmonitor.ru/MapClust/ . Кроме этого, код сервиса для локального использования по адресу https://github.com/SmirnygaTotoshka/MapClust . Подробная инструкция и тестовые данные представлены на сайте. Ограничения для онлайн-использования - 100 Мб. В дальнейших планах у авторов разработать пакет для языка программирования R и распространить используемый подход на другие типы данных.

# MAPCLUST IS SERVICE FOR EPIDEMIOLOGICAL AND ECOLOGICAL MONITORING

**Smirnov Anton Sergeevich1, Zakharov Oleg Stanislavovich1, Pyatnitskiy Alexey Mikhailovich1, Gukasov Vadim Mikhailovich2,**

1 - PRNRMU, Moscow,

2 - FRCEC, Moscow

# Abstract

This article is the MapClust description. The MapClust is web-service for clusterisation of spatial and spatio-temporal data and it is authors previous research practical realisation. We demostrate web-service work on epidemiological and ecological monitoring tasks.

Keywords: MapClust; epidemiological monitoring; ecological monitoring; clustering algorithm.

# Информация об авторах

* Смирнов Антон Сергеевич – студент 6 курса направления «Медицинская кибернетика» МБФ РНИМУ, Москва.
* Захаров Олег Станиславович – студент 2 курса направления «Медицинская кибернетика» МБФ РНИМУ, Москва.
* Пятницкий Алексей Михайлович – доцент кафедры высшей математики МБФ РНИМУ, кандидат физико-математических наук, Москва.
* Гукасов Вадим Михайлович – доктор биологических наук, главный научный сотрудник Государственного центра экспертизы в сфере науки и инноваций, ФГБНУ НИИ Республиканский исследовательский научно-консультационный центр экспертизы (ФГБНУ НИИ РИНКЦЭ) Министерства науки и высшего образования России, Москва

# Контактное лицо

Смирнов Антон Сергеевич, +79109522086, e-mail: anton.smirnov.9910@gmail.com

# Список литературы

1. [Annual AQ statistics (AirBase & e-reporting merged)](https://discomap.eea.europa.eu/App/AirQualityStatistics/index.html).

2. Александрова Г.А. [Заболеваемость с впервые в жизни установленным диагнозом ВИЧ-инфекции на 100 тыс. Человек населения](https://fedstat.ru/indicator/41718) / Г.А. Александрова.

3. Ito S. [Spatio-Temporal Epidemiology of the Spread of African Swine Fever in Wild Boar and the Role of Environmental Factors in South Korea](https://doi.org/10.3390/v14122779) / S. Ito [et al.] // Viruses. – 2022. – Vol. 14. – № 12. – P. 2779.

4. Kiani B. [Spatio-temporal epidemiology of the tuberculosis incidence rate in Iran 2008 to 2018](https://doi.org/10.1186/s12889-021-11157-1) / B. Kiani [et al.] // BMC Public Health. – 2021. – Vol. 21. – № 1. – P. 1093.

5. Kulldorff M. [A spatial scan statistic](https://doi.org/10.1080/03610929708831995) / M. Kulldorff // Communications in Statistics - Theory and Methods. – 1997. – Vol. 26. – № 6. – P. 1481-1496.

6. Odhiambo J.N. [Spatial and spatio-temporal epidemiological approaches to inform COVID-19 surveillance and control: a systematic review of statistical and modelling methods in Africa](https://doi.org/10.1136/bmjopen-2022-067134) / J.N. Odhiambo [et al.] // BMJ Open. – 2023. – Vol. 13. – Spatial and spatio-temporal epidemiological approaches to inform COVID-19 surveillance and control. – № 1. – P. e067134.

7. Shaweno D. [Methods used in the spatial analysis of tuberculosis epidemiology: a systematic review](https://doi.org/10.1186/s12916-018-1178-4) / D. Shaweno [et al.] // BMC Medicine. – 2018. – Vol. 16. – Methods used in the spatial analysis of tuberculosis epidemiology. – № 1. – P. 193.

8. Tango T. [Spatial scan statistics can be dangerous](https://doi.org/10.1177/0962280220930562) / T. Tango // Statistical Methods in Medical Research. – 2021. – Vol. 30. – № 1. – P. 75-86.

9. Pyatnitskiy A.M. [Searching for clusters in population data](https://doi.org/10.34219/2306-3645-2021-11-3-63-71) / A.M. Pyatnitskiy, V.M. Gukasov, A.S. Smirnov // Medicina i vysokie tehnologii. – 2021. – Vol. 3.

10. Pyatnitskiy A.M. [Searching for event clusters in frequency tables with applications in epidemiology and ecological monitoring](https://doi.org/10.34219/2306-3645-2021-11-2-7-17) / A.M. Pyatnitskiy, V.M. Gukasov, A.S. Smirnov // Medicina i vysokie tehnologii. – 2021. – Vol. 2. – P. 7-17.

11. Пятницкий А.М. Кластеризация данных методом «расширения точек» / А.М. Пятницкий, В.М. Гукасов, А.С. Смирнов // МЕДИЦИНА И ВЫСОКИЕ ТЕХНОЛОГИИ. – 2020. – № 2. – P. 9-14.

12. Пятницкий А.М. Поиск кластеров событий в эпидемиологическом и экологическом мониторинге / А.М. Пятницкий, В.М. Гукасов, А.С. Смирнов // МЕДИЦИНА И ВЫСОКИЕ ТЕХНОЛОГИИ. – 2020. – № 4. – P. 29-37.