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[1]. Самый известный метод сканирующей статистики – метод Куллдорфа[2]. Этот метод широко используется в современных эпидемиологических исследованиях[3–5]. 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[6].

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

Принцип поиска кластеров описан в статьях[7–10]. При изучении пространственно-временной изменчивости в числах событий используются два представления исходных данных:

1. event-based datа, когда известны координаты событий (точечные данные)
2. population-based data, когда известно лишь общее число событий в каждой области за определенный промежуток времени (частотные данные).

Предлагаемый критерий может быть использован в обоих случаях, однако, данный сервис предполагает работу только с частотными данными.Даны частоты событий для каждой i-ой области, обозначим их . Возможны две постановки задачи:

1. разыскиваем пространственные неоднородности в один и тот же момент времени (hot spots, cold spots, то есть кластеры и разряжения)
2. разыскиваем области, в которых количество событий значимо увеличилось или уменьшилось.

В первом случае области должны быть выбраны так, чтобы численности популяций в них были одинаковы. Во втором случае области произвольны и могут, например, соответствовать административным регионам (предполагается, что размер популяции в каждой области не изменился).

1. Даны частоты для каждой области, назовем их
2. Допустим, размеры популяций в каждой области одинаковыми и средняя частота по всем ячейкам не мала
3. Тогда для каждой области вводятся вероятности (см. ниже), которые в случае справедливости имеют равномерное распределение и две пороговые вероятности , . Пороговые вероятности выделяют области, для которых есть подозрение на то, что имеется неоднородность (кластер или разряжение, увеличение частоты событий с течением времени или уменьшение).
   * Для пространственной неоднородности
   * Для сравнения моментов времени
   * - функция Лапласа.
4. Соприкасающиеся ячейки(группы), попадающие в диапазон - разряжения
5. Соприкасающиеся ячейки(группы), попадающие в диапазон - кластеры
6. Статистика для кластеров
7. Статистика для разряжений
8. При справедливости справедливо и
9. 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.

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. MapClust предполагает, что исследуемые величины уже находятся внутри файлов, связанных с Shapefiles. Помимо непосредственно shapefile с разрешением .shp, программе необходимо подать на вход файлы с расширениями .shx, .prj, .dbf и опционально файл .cpg, в котором содержится информация о кодировке файлов. Подготовка файлов для использования выходит за рамки функционала MapClust, так как зависит от целей исследователя и имеющихся у него данных. Примеры скриптов для подготовки данных представлены в репозитории GitHub проекта. Примерная инструкция на основе готовых примеров, выглядит следующим образом:

1. Пример готовых данных можно скачать в вкладке “Примеры данных”
2. Распаковываем архив в любую пустую папку
3. Необходимо определить распределение размеров кластеров/разряжений и распределение максимального размера кластеров/разряжений методом Монте-Карло. Для этого
   * Загружаем в сервис **все** распакованные файлы из архива.
   * Выбираем необходимые параметры (про пределы см. принцип работы критерия)
   * Запускаем симуляцию
   * Сохраняем результат
4. Далее можно воспользоваться самим критерием:
   * Загружаем в сервис **все** распакованные файлы из архива.
   * Выбираем необходимые параметры: столбцы с данными, режим работы, т.е. какой столбец будет использован для расчетов)
   * При необходимости внести изменения или посмотреть значение в регионе, можно кликнуть на регион и поменять значение величины при необходимости.
   * Запускаем расчет
   * Сохраняем результат

# Примеры работы

Приведем несколько примеров работы. Для первого примера возьмем данные по заболеваемостью вирусом иммунодефицита человека (ВИЧ) в Российской Федерации (РФ) с портала ЕМИСС[11]. Веб-сервис может работать в двух режимах работы: пространственном и пространственно-временном. Пространственный режим заключается в поиске кластеров по величине, измеренной 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 только разряжения или только кластеры. Для примера возьмем , и и запустим критерий для в первом режиме для данных 2011, 2012 и 2013 годов. Данные представляют из себя количество случаев за год на 100000 населения, но население не распределено равномерно по областям. Для того, чтобы нивелировать эффект различных дисперсий в разных регионах, мы немного модифицировали расчет р-значения в этом случае, применив арксинус-преобразование Фишера. Стоит отметить, что требование равномерности распределения популяции стоит соблюдать только в режиме поиска пространственной неоднородности. Преобразование данных описано в приложении. С этими же параметрами мы посмотрели разницу между 2011 и 2012 годом, 2012 и 2013 годом. Карта РФ приведена в соответствии с государственными границами на момент 2011-2013 годов. Результаты представлены на [Figure 1](#fig-first), [Figure 2](#fig-second), [Figure 3](#fig-third), [Figure 4](#fig-fourth) и [Figure 5](#fig-fifth). Программа также позволяет получить список областей, составляющих кластеры или разряжения. Из полученных данных можно сделать выводы о регионах с благополучной и тревожной эпидемиологической обстановкой.

|  |
| --- |
| Figure 1: Результаты кластеризации при заданном уровне значимости, равным 0.05, для данных на 2011 год. На рисунках 1-5 синим цветом обозначены статистически незначимые кластеры, черным - значимые. Цветовая шкала соответствует уровню р-значения для этих данных. Статистически значимый кластер состоит из Свердловской и Тюменской областей и Ханты-Мансийского автономного округа. Также следует обратить внимание на соседний кластер из Новосибирской и Кемеровских областей. О них пойдет речь далее. |

|  |
| --- |
| Figure 2: Результаты кластеризации при заданном уровне значимости, равным 0.05, для данных на 2012 год. Статистически значимый кластер, представленный на рисунке 1, “потерял” свою значимость, а значимость кластера из Новосибирской и Кемеровской областей возрасла. Следовательно, эпидемиологическая обстановка там ухудшилась. |

|  |
| --- |
| Figure 3: Кластеризация данных по разнице случаев в 2011 и 2012 годах. Можно заметить, что статистически значимое ухудшение обстановки произошло в Кемеровской области. За счет этого и возрасла сила кластера из двух соседних регионов из рисунка 2. |

|  |
| --- |
| Figure 4: Результаты кластеризации при заданном уровне значимости, равным 0.05, для данных на 2013 год. Заметен всплеск случаев на Урале, Западной Сибири и в западных областях Восточной Сибири, который объеденен в один большой кластер неправильной формы. |

|  |
| --- |
| Figure 5: Кластеризация данных по разнице случаев в 2012 и 2013 годах. аномальный всплеск и огромный кластер связан с резким ростом числа случаев в 2013 году в сравнении с 2012 году в этих регионах. Нельзя сделать вывод, что причиной этому послужила неблагоприятная обстановка в Кемеровской области, но на неё необходимо обратить особое внимание, как и на другие соседние сибирские регионы. |

Следует отметить, что программа ищет кластеры, исходя из соседства регионов и не учитывает миграцию населения, транспортные потоки и другие факторы, которые могут служить причиной развития эпидемий. Программа позволяет выявлять аномальные очаги. Выводы о причинах и механизмах их развития необходимо делать, основываясь на дополнительных исследованиях.

В качестве задачи экологического мониторинга рассмотрим поиск очагов ухудшения качества воздуха в Румынии. Данные взяты с сайта Европейского агентства по окружающей среде[12]. 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 6](#fig-sixth) Наблюдается статистически значимое улучшение качества воздуха в районе Плоешти и Брашова, который исторически был промышленным в Румынии.

|  |
| --- |
| Figure 6: Сравнение PM10 в Румынии за 2018 и 2021 год. |

Рассмотрим значимость параметра . Для этого возьмем данные о количестве лесных пожаров по штатам Индии за 2008-2009 и 2009-2010 года и запустим критерий с параметрами или [13]. – это уровень надежности, определяющий вероятность ошибки I рода, или ложного определения области как значимого кластера, когда это не так. Его изменение может оказать значительное влияние на выявление статистически значимых областей. Одна и та же область: кластер или разряжение - может быть статистически значимым при , но не быть статистически значимым . Подобный пример представлен на [Figure 7](#fig-seventh). Использование более низкого уменьшает вероятность ошибки I рода, но увеличивает риск ошибки II рода, или неспособности определить истинный кластер. Таким образом, тщательный выбор параметра необходим для обеспечения точной идентификации значимых кластеров при минимизации риска ложноположительных и ложноотрицательных результатов. Сервис MapClust позволяет пользователям регулировать чувствительность алгоритма путем изменения параметра , обеспечивая больший контроль над балансом между ошибками I рода и II рода.

|  |
| --- |
| Figure 7: Сравнение количества лесных пожаров в Индии за 2008-2009 и 2009-2010 года. Слева график при уровне значимости равным 0.05, справа - 0.01. Можно заметить, что один кластер “потерял” значимость при уровне значимости в 0.01. Второй кластер, расположенный рядом с ним, её сохранил: точка находится чуть выше границы критической области |

# 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 доступен 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, добавить расчет p-value для структур данных и распространить используемый подход на другие типы данных.

# Приложение

## Расчет арксинус-преобразования для данных с неоднородно распределенной популяцией.

Допустим дано число случаев на 100000 населения для каждого региона, обозначим эту величину буквой . Тогда нам необходимо выразить число случаев в i-ом регионе - . Сосчитаем среднее число случаев по стране . Тогда , где - функция Лапласа.

# 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. 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.

2. 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.

3. 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.

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. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

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

13. Kamal S. [Forest fires in india](https://www.kaggle.com/datasets/sanikamal/forest-fires-in-india) / S. Kamal.