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[6]. Самый известный метод сканирующей статистики – метод Куллдорфа[4]. Этот метод широко используется в современных эпидемиологических исследованиях[2; 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[7]. Наш критерий не зависит от формы и размеров кластеров, позволяет регулировать свою чувствительность, осуществляет поиск нескольких кластеров при заданной ошибке первого рода.

# Принцип работы

Данный сервис получает на вход данные о взаиморасположении областей исследования и о значении исследуемой величины в этой области. Исследуемая величина зависит от цели исследования: летальность заболевания, количество случаев, площадь лесных пожаров и многое другое. Данные должны быть представлены в специальном общеиспользуемом геоинформационном формате – ESRI Shapefile. Принцип поиска кластеров описан в статьях[8–11]. 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 ([Equation 3](#eq-clust-stat)) для кластеров и ([Equation 4](#eq-discharges-stat)) для разряжений.

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[1]. 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. |

# Заключение

# 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

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

# Информация об авторах

# Список литературы

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

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. 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.

11. 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.