УДК 519.24

М.С. Абрамович[1], С.В. Анищенко[2], М.Н. Мицкевич[1], О.И. Быданов[3]

ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ КЛАСТЕРНЫЙ АНАЛИЗ ЗАБОЛЕВАЕМОСТИ

Рассматриваются методы и алгоритмы обнаружения и построения кластеров заболеваемости на изучаемой территории. Предлагается робастная версия пространственной сканирующей статистики для построения кластеров. Алгоритмы пространственно-временного кластерного анализа применяются для обнаружения кластеров заболеваемости карциномой щитовидной железы.

Введение

Во многих областях: медицине, криминологии, археологии, истории, градостроительстве, региональном планировании – исследуемые данные кроме информации об объектах исследования содержат также сведения об их географических координатах. Примером географических данных являются данные о заболеваемости. В простейшем случае такие данные представляют собой совокупность географических координат, соотнесенных с отдельными случаями заболевания. Для исследования географического распространения различных типов заболеваний широко применяются методы пространственно-временного кластерного анализа [1–8]. Эти методы предназначены для определения, является ли наблюдаемое распределение больных равномерным на изучаемой территории или же существуют некоторые кластеры населенных пунктов с повышенным уровнем заболеваемости. Результаты пространственно-временного кластерного анализа могут быть использованы для контроля, диагностики и предотвращения заболеваний, а также для выдвижения гипотез о причинах их возникновения.

Методы пространственно-временного кластерного анализа делятся на глобальные и локальные. Глобальные методы кластерного анализа позволяют определять существование кластеров заболеваемости на изучаемой территории. Недостатком ряда существующих критериев глобальной кластеризации является необходимость задания размеров гипотетических кластеров, которые на практике, как правило, неизвестны [1].

Методы локальной кластеризации позволяют определять местоположение и размер кластеров на изучаемой территории и проверять их статистическую значимость. Наиболее распространенными и эффективными методами построения локальных кластеров являются методы пространственной и пространственно-временной сканирующей статистики [2–5]. Эти методы служат универсальными инструментами кластерного анализа, позволяющими работать с разнообразными типами входных наборов данных, использовать модели заболеваемости, основанные на различных распределениях вероятностей, обнаруживать кластеры любых форм в зависимости от выбранного способа построения множества сканирующих окон.

В настоящей работе предложен критерий глобальной кластеризации, не требующий задания параметра, связанного с размером гипотетических кластеров. При наличии выбросов заболеваемости в кластерах рассмотрена робастная пространственная сканирующая статистика. Разработан также алгоритм построения сканирующих окон для формирования кластеров заболеваемости произвольной формы. Алгоритмы пространственного и пространственно-временного кластерного анализа применены для выявления кластеров заболеваемости карциномой щитовидной железы в популяции населения в возрасте до 18 лет на территории Республики Беларусь в 1989–2005 гг.

1. Критерии глобальной кластеризации

Определим статистическую гипотезу H_0 об отсутствии кластеризации заболеваний на изучаемой территории (т. е. гипотезу о том, что заболеваемость среди исследуемой популяции

населения распределена равномерно) и альтернативную гипотезу H_1 о наличии кластеров с повышенным уровнем заболеваемости.

Предположим, что вся изучаемая территория разделена на m районов. Пусть $\xi_i, i = 1, ..., m,$ – случайная величина, описывающая число случаев заболевания в i-м районе; $\mu_i = E\{\xi_i\}, i = 1, ..., m,$ – ожидаемое число случаев заболевания в i-м районе.

Будем предполагать, что случайная величина ξ_i, описывающая число случаев заболевания в i-м районе, при выполнении гипотезы H_0 об отсутствии кластеризации на изучаемой территории имеет распределение Пуассона со средним $\mu_i, i = 1, ..., m$:

$$H_0 = \{\xi_i \sim Pois(\mu_i), i = 1, ..., m\}.$$

Будем также предполагать, что при выполнении гипотезы H_0 случайные величины $\xi_i, i = 1, ..., m,$ являются независимыми.

Для каждого района известно число наблюдаемых случаев заболеваний $c_i, i = 1, ..., m,$ и численность населения этого района (численность группы риска) $n_i, i = 1, ..., m.$ Введем обозначения: $C = \sum_{i=1}^{m} c_i$ – количество всех случаев заболеваний, $N = \sum_{i=1}^{m} n_i$ – общая численность группы риска, d_{ij} – расстояние между центрами i-го и j-го районов.

Ожидаемое число случаев заболеваний в условиях нулевой гипотезы может определяться как произведение численности группы риска i-го района на общий уровень заболеваемости:

$$e_i = \frac{n_i}{N} C.$$

С учетом пуассоновской модели построим для каждого i-го района следующую статистику [1]:

$$z_i = \frac{c_i - e_i}{\sqrt{e_i}}. \qquad (1)$$

Статистика (1) при выполнении нулевой гипотезы имеет асимптотически стандартное нормальное распределение. Тогда статистику z_i^2 можно рассматривать как случайную величину, имеющую асимптотически χ^2-распределение с одной степенью свободы.

Построим глобальную статистику для определения кластеризации в целом по всей территории:

$$\chi^2 = \sum_{i=1}^{m} \frac{(c_i - e_i)^2}{e_i}. \qquad (2)$$

Эта статистика имеет асимптотически χ^2-распределение с $m-1$ степенями свободы. Отсюда получаем χ^2-критерий с решающим правилом:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P > \alpha; \\ H_1, \text{ если } P \le \alpha, \end{cases}$$

где $P = 1 - F_{m-1}(\chi^2)$, $F_{m-1}(\chi^2)$ – функция χ^2-распределения с $m-1$ степенями свободы; α – уровень значимости.

Предложим еще один подход к построению критерия глобальной кластеризации на основе статистики (1). Для этого определим следующую статистику:

$$M = \max_i z_i^2. \tag{3}$$

Статистику z_i^2 можно рассматривать как случайную величину, имеющую асимптотически χ^2-распределение с одной степенью свободы. В этом случае статистика (3) в условиях нулевой гипотезы имеет распределение

$$F(M) = P\{M \le x\} = P\{z_1^2 \le x,\, z_2^2 \le x, ..., z_m^2 \le x\} = \prod_{i=1}^{m} P\{z_i^2 \le x\} = F_{\chi_1^2}^m(M).$$

Критерий для обнаружения глобальной кластеризации строится следующим образом:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P > \alpha; \\ H_1, \text{ если } P \le \alpha, \end{cases}$$

где $P = 1 - F_{\chi_1^2}^m(M)$, $F_{\chi_1^2}(\cdot)$, – функция распределения χ^2 с одной степенью свободы.

В работе [6] Шварц предложил критерий, основанный на энтропии. Статистика критерия имеет вид

$$SET = \ln(C!) + \ln((N-C)! - \sum_{i=1}^{m} (\ln(c_i!) + \ln((n_i - c_i)!).$$

Нулевая гипотеза об отсутствии кластеризации для критерия Шварца отвергается при малых значениях статистики SET.

Критерии, рассмотренные выше, построены в предположении, что число наблюдаемых случаев заболевания в определенном районе не зависит от числа случаев заболевания в соседних районах. На практике это предположение не всегда выполняется. В работе [7] Уайтмур предложил критерий, который учитывает зависимость случаев заболеваемости в соседних районах. Статистика критерия имеет вид

$$T = \tfrac{1}{2} \sum_{i,j=1}^{m} d_{ij} c_i c_j.$$

Нулевая гипотеза об отсутствии кластеризации для критерия Уайтмура, как и для критерия Шварца, отвергается при малых значениях статистики T.

Так как распределение статистик критериев Шварца и Уайтмура неизвестно, нельзя аналитически вычислить *p*-значения критериев для проверки нулевой гипотезы. В этом случае можно смоделировать наборы данных при выполнении условий нулевой гипотезы и вычислить *p*-значения критериев на основе метода Монте-Карло.

2. Методы построения локальных кластеров на основе сканирующей статистики

Наиболее распространенным и эффективным методом локального кластерного анализа является метод пространственной сканирующей статистики [2, 3].

Рассмотрим построение круговой пространственной сканирующей статистики, которая определяет круговое окно Z для каждого административного центра района. Для каждого

из этих центров радиус круга может увеличиваться от 0 до значения, при котором в круг попадает заданное максимальное число районов K, включаемых в кластер.

Пусть $Z_{ik}, k = 1, ..., K$ – окно, составленное $(k-1)$ ближайшими к району i соседями. Тогда все окна, которые должны сканироваться круговой пространственной сканирующей статистикой, включаются в множество

$$Z_1 = \{Z_{ik} \mid 1 \le i \le m, 1 \le k \le K\}.$$

Пространственная сканирующая статистика строится с помощью отношения правдоподобия и имеет следующий вид [3]:

$$S = \sup_{Z \in Z_1} \lambda(Z) = \sup_{Z \in Z_1} \left(\frac{c_Z}{\mu_Z}\right)^{c_Z} \left(\frac{C - c_Z}{C - \mu_Z}\right)^{C - c_Z} I\left(\frac{c_Z}{\mu_Z} > \frac{C - c_Z}{C - \mu_Z}\right), \quad (4)$$

где $c_z = \sum_{i \in Z_c} c_i$ и $\mu_z = \sum_{i \in Z_c} \mu_i$ – соответственно наблюдаемое и ожидаемое число случаев заболевания в окне Z; $I(\cdot)$ – индикаторная функция.

Выражение $\frac{c_z}{\mu_z} > \frac{C - c_z}{C - \mu_z}$ в (4) означает, что внутри окна Z количество случаев заболеваний относительно среднего больше по сравнению с областью вне окна.

При решении задачи поиска кластеров с пониженным риском заболевания знак неравенства в этом выражении необходимо поменять на противоположный.

Окно $Z^* \in Z_1$, на котором статистика (4) достигает максимального значения, будет являться наиболее вероятным кластером.

Гибкая сканирующая статистика в отличие от круговой сканирующей статистики определяет для каждого района окно неправильной формы Z путем объединения смежных с ним районов [4, 5]. Для каждого данного района i создается множество окон неправильной формы, состоящих из k объединенных районов, включая район i (т. е. граф с вершинами в центрах этих районов и ребрами, показывающими смежность районов, является связным графом). Чтобы исключить обнаружение кластера маловероятной формы, число объединенных районов ограничивается $(K-1)$ ближайшими к региону i соседними районами. Отметим, что для реализации метода гибкой пространственной сканирующей статистики необходимо иметь матрицу смежности всех районов. Смежность двух районов может определяться как наличие у них хотя бы одной общей граничной точки. В случаях когда построение такой матрицы смежности затруднено, соединенными районами могут считаться такие, расстояние между центрами которых удовлетворяет определенному ограничению.

Пусть $Z_{ik(j)}, j = 1, ..., j_{ik}$, обозначает *j*-е окно, которое является объединением k соседних районов, начиная с района i, где j_{ik} – количество значений *j*, удовлетворяющих условию $Z_{ik(j)} \subseteq Z_{ik}, k = 1, ..., K$. В итоге все окна, которые должны сканироваться, будут включены в множество

$$Z_2 = \{Z_{ik(j)} \mid 1 \le i \le m, 1 \le k \le K, 1 \le j \le j_{ik}\}. \quad (5)$$

Для любого заданного района *i* круговая пространственная сканирующая статистика рассматривает *K* концентрических кругов, в то время как гибкая сканирующая статистика рассматривает также и все множества соединенных районов (включая единственный район *i*), центры которых размещены внутри *K*-го наибольшего концентрического круга. Поэтому мощность множества Z_2 гораздо больше мощности множества Z_1, которая не превышает *mK*.

При альтернативной гипотезе H_1 о наличии кластеризации на изучаемой территории существует по крайней мере одно окно Z, для которого риск заболевания будет более высоким внутри окна, чем за его границами.

Метод пространственной сканирующей статистики можно модифицировать для анализа пространственно-временных данных, когда известна численность группы риска и число случаев заболевания в последовательные моменты времени. При таком подходе время является третьей координатой, а круговые окна, используемые для вычисления пространственной сканирующей статистики, заменяются цилиндрами. Основания этих цилиндров соответствуют некоторым областям, как и в пространственном случае, а высоты определяют длину потенциального кластера во времени.

Пространственно-временной кластерный анализ позволяет выявить возможные кластеры, существовавшие на исследуемой территории на протяжении определенного временного промежутка. Отметим, что формула для вычисления пространственной сканирующей статистики (4) остается такой же, но меняется множество сканирующих окон. Множество Z_1 в формуле (4) заменяется следующим множеством Z_{ST} [5]:

$$Z_{ST} = \left\{ Z_{ik[a,b]} = \left| 1 \le i \le m, 1 \le k \le K; a, b = T_1, ..., T_p, a \le b \right. \right\},$$

где $Z_{ik[a,b]}$ – цилиндрическое окно, включающее в себя район i и его $(k-1)$ ближайших соседних районов для каждого момента времени T из временного промежутка $[a, b]$. Максимальную длину временного промежутка можно ограничить некоторым значением P, $1 \le T \le P$ (по аналогии с максимальным пространственным размером кластера K).

В пространственном и пространственно-временном кластерном анализе процедура проверки статистической значимости построенных кластеров осуществляется методом статистического моделирования и состоит из следующих шагов:

Шаг 1. Вычисляется значение сканирующей статистики для наиболее вероятного кластера.

Шаг 2. Генерируется n выборок в условиях выполнения нулевой гипотезы.

Шаг 3. Для каждой выборки вычисляется значение сканирующей статистики.

Шаг 4. Из значений сканирующих статистик формируется вариационный ряд.

Шаг 5. Если значение статистики, вычисленное для наиболее вероятного кластера, находится в доле α наибольших значений, нулевая гипотеза отвергается на уровне значимости α.

Использование метода статистического моделирования позволяет также находить статистически значимые второстепенные кластеры, т. е. такие кластеры, значение статистики для которых не является максимальным, но нулевая гипотеза отвергается на уровне значимости α. Самыми важными из вторичных кластеров являются те, которые не содержатся в наиболее вероятном кластере. Способность метода локального кластерного анализа находить несколько непересекающихся статистически значимых кластеров является его важной характеристикой.

3. Построение робастной версии сканирующей статистики

Если вероятностная модель заболеваемости описывает наблюдения с выбросами, то необходимо использовать робастные статистические методы для построения пространственных и пространственно-временных кластеров. Пусть $\bar{c}_Z$ – выборочное среднее число заболеваний в окне Z и $|Z|$ – мощность множества Z. Аналогично пусть $\bar{C}$ – выборочное среднее число заболеваний по всем m районам. Так как $c_Z = \bar{c}_Z |Z|$, выражение (4) может быть записано в следующей форме:

$$S = \sup_{Z \in Z_1} \lambda(Z) = \sup_{Z \in Z_1} \left(\frac{\bar{c}_Z |Z|}{\mu_Z} \right)^{\bar{c}_Z |Z|} \left(\frac{\bar{C}m - \bar{c}_Z |Z|}{\bar{C}m - \mu_Z} \right)^{C - \bar{c}_Z |Z|} I\left(\frac{\bar{c}_Z |Z|}{\mu_Z} > \frac{\bar{C}m - \bar{c}_Z |Z|}{\bar{C}m - \mu_Z} \right). \quad (6)$$

В случае наличия выбросов статистика $\bar{c}_Z$ является смещенной оценкой среднего значения в окне *Z*. Если в выражении (6) вместо среднего значения числа заболеваний $\bar{c}_Z$ в окне *Z* использовать его робастную оценку c_R, получим робастную версию сканирующей статистики S_R:

$$S_R = \sup_{Z \in Z_1} \lambda(Z) = \sup_{Z \in Z_1} \left(\frac{c_R |Z|}{\mu_Z} \right)^{\bar{c}_z |Z|} \left(\frac{\bar{C}m - c_R |Z|}{\bar{C}m - \mu_Z} \right)^{C - \bar{c}_z |Z|} I\left(\frac{c_R |Z|}{\mu_Z} > \frac{\bar{C}m - c_R |Z|}{\bar{C}m - \mu_Z} \right). \quad (7)$$

Если данные содержат хотя бы один выброс, статистика (6) часто определяет кластер, который содержит только этот выброс. Чувствительность пространственной сканирующей статистики к наличию выбросов проанализирована в [8] для случая, когда вместо среднего используются робастные оценки Хампеля, Эндрюса, Хьюбера и винзорированное среднее [9]. Как показано в [8], сканирующая статистика (6) быстро возрастает при увеличении величины выброса.

Если целью исследования является определение кластера, протяженного в пространстве или времени, необходимо определить нижнюю границу числа наблюдений в кластере для уменьшения влияния выбросов.

4. Алгоритм построения множества окон для гибкой сканирующей статистики

При поиске кластеров произвольной формы важную роль играет выбор алгоритма для построения множества областей различных форм при вычислении сканирующей статистики. Необходимо учитывать, что исследуемая территория может быть разбита на большое количество районов, поэтому такой алгоритм должен быть достаточно быстрым и эффективным. Отметим, что при этом должны существовать и ограничения на форму кластеров, чтобы избежать возможности обнаружения кластеров маловероятной формы.

Рассмотрим алгоритм построения множества окон Z_2, задаваемого выражением (5), для метода гибкой пространственной сканирующей статистики. Множество окон представляет собой совокупность соседних районов в пределах заранее определенного максимального числа районов в кластере *K*. Отличие предлагаемого алгоритма от алгоритма из работы [4] состоит в изменении порядка вычислений, которое дает возможность строить только связные множества районов. Это, в свою очередь, приводит к уменьшению количества вычислений. Алгоритм состоит из следующих шагов:

Шаг 1. Строится матрица $A = (a_{ij})$ размерности $m \times m$. При этом полагается $a_{ij} = 1$, если районы i и j являются соседними, и $a_{ij} = 0$ в противном случае.

Множество окон Z_2, задаваемое формулой (5), полагается пустым: $Z_2 = \{\varnothing\}$, $i_0 = 0$.

Шаг 2. Район i_0, $i_0 = 1, 2, ..., m,$ полагается стартовым, $i_0 := i_0 + 1$.

Строится множество W_{i_0}, состоящее из района i_0 и его $(K-1)$ ближайших соседей: $W_{i_0} = \{i_0, i_1, i_2, ..., i_{K-1}\}$.

Шаг 3. Строится множество $Z = \{i_0\}$.

Шаг 4. Множество Z добавляется в множество Z_2, если $Z \notin Z_2$.

Шаг 5. Каждый район j, $j \in W_{i_0}$, $j \notin Z$, который имеет соседний район в множестве Z, добавляется в множество Z, и для каждого такого случая рекурсивно повторяются шаги 4 и 5, если хотя бы один такой район j нашелся.

Шаг 6. Повторяются шаги 2–5 до построения искомого множества окон Z_2.

Вычислительная сложность приведенного алгоритма – $O\{m(K-1)!\}$.

5. Применение алгоритмов локальной и глобальной кластеризации для исследования распространенности карциномы щитовидной железы

Критерии глобальной кластеризации и алгоритмы пространственного и пространственно-временного кластерного анализа были применены для обнаружения кластеризации данных больных карциномой щитовидной железы в популяции населения в возрасте до 18 лет в период с 1989 по 2005 гг. в Республике Беларусь по 119 административным районам. Для каждого года и района были известны численность анализируемой популяции населения и число случаев заболевания.

Для вычисления расстояний в километрах использовались константы 98,699 км (эквивалентно 1° восточной долготы для Республики Беларусь) и 111,272 км (эквивалентно 1° северной широты).

Критерий глобальной кластеризации χ^2; критерий, основанный на статистике (3), и критерий Шварца были применены для определения наличия кластеров заболеваемости карциномой щитовидной железы в исследуемой популяции на территории Республики Беларусь. Уровень значимости α полагался равным 0,05. В табл. 1 приведены по годам число случаев заболеваний, численность группы риска, *p*-значение критериев (значимые значения выделены жирным шрифтом).

Таблица 1

Наличие кластеров заболеваемости карциномой щитовидной железы на территории Республики Беларусь

Год	Число случаев заболеваний	Численность населения	*p*-значение критерия χ^2	*p*-значение критерия, основанного на статистике (3)	*p*-значение критерия Шварца
1989	1	2 905 220	1,000 000	0,073 63	0,551 576
1990	20	2 874 905	**0,003 111**	**0,000 036**	0,372 334
1991	62	2 844 490	**0,000 000**	**0,000 000**	**0,000 250**
1992	68	2 814 088	**0,000 000**	**0,000 000**	**0,008 230**
1993	98	2 783 724	**0,000 000**	**0,000 000**	**0,000 000**
1994	102	2 753 247	**0,000 000**	**0,000 000**	**0,000 000**
1995	103	2 722 849	**0,000 000**	**0,000 000**	**0,000 000**
1996	99	2 692 426	**0,000 000**	**0,000 000**	**0,000 000**
1997	87	2 661 963	**0,000 000**	**0,000 000**	**0,000 010**
1998	81	2 631 345	**0,000 000**	**0,000 000**	**0,009 550**
1999	81	2 607 474	**0,000 000**	**0,000 000**	**0,000 030**
2000	80	2 545 873	**0,000 000**	**0,000 008**	**0,000 040**
2001	72	2 490 440	**0,000 000**	**0,000 001**	**0,000 460**
2002	58	2 422 532	**0,000 000**	**0,000 000**	**0,001 790**
2003	38	2 339 375	0,973 846	**0,006 023**	0,551 376
2004	18	2 257 248	0,698 154	**0,000 001**	0,701 637
2005	6	2 181 276	**0,000 094**	**0,000 000**	0,243 762

Как следует из результатов, приведенных в табл. 1, в период с 1991 по 2002 гг. все три критерия ежегодно показали наличие кластеров больных карциномой щитовидной железы на территории Республики Беларусь.

Для построения кластеров заболеваемости по каждому году отдельно метод пространственной сканирующей статистики применялся с предельным размером кластера $K = 20$. Среди случаев заболевания карциномой щитовидной железы было найдено 13 значимых кластеров при уровне значимости, равном 0,05. Для каждого кластера *p*-значение оценивалось с помощью

999 моделирований по методу Монте-Карло (табл. 2), указаны также центр кластера, количество районов, вошедших в него, число случаев заболеваний в кластере, *p*-значение для проверки значимости кластеров.

Таблица 2

Кластеры заболеваемости карциномой щитовидной железы по каждому году отдельно, построенные с помощью пространственной сканирующей статистики

Год	Центр кластера	Число районов в кластере	Число случаев заболеваний	*p*-значение
1991	Брагин	13	38	**0,000**
1992	Речица	12	24	**0,000**
1993	Столин	5	24	**0,000**
1994	Брагин	15	46	**0,000**
1995	Лоев	12	44	**0,000**
1996	Речица	15	40	**0,000**
1997	Лоев	9	33	**0,000**
1998	Столин	1	9	**0,000**
1999	Брагин	15	41	**0,000**
2000	Хойники	14	35	**0,000**
2001	Лоев	11	27	**0,000**
2002	Столин	4	16	**0,000**
2003	Лоев	13	13	**0,040**

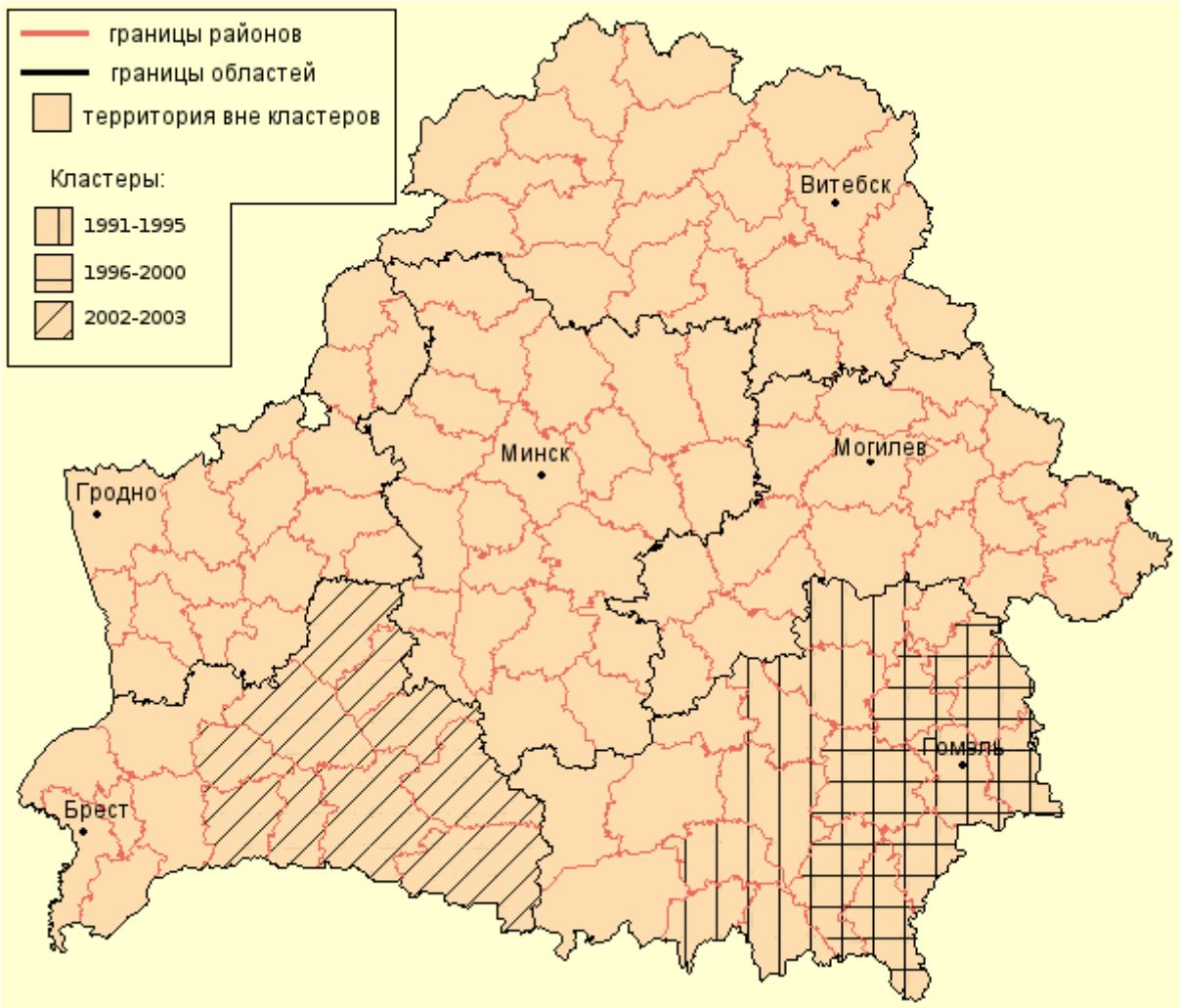
Для исследуемой популяции населения в возрасте до 18 лет был проведен пространственно-временной кластерный анализ для трех периодов: 1990–1995 гг., 1996–2001 гг., 2002–2005 гг. Для первого периода статистически значимый кластер был обнаружен в 1991–1995 гг., для второго периода – в 1996–2000 гг. и для третьего периода – в 2002–2003 гг. (рисунок).

Результаты пространственно-временного кластерного анализа даны в табл. 3. Каждый кластер представлен его центром, числом районов в кластере, *p*-значением для проверки значимости кластера, числом больных в кластере и численностью группы риска.

Таблица 3

Кластеры заболеваемости карциномой щитовидной железы для трех периодов, построенные с использованием пространственно-временной сканирующей статистики

Годы	Центр кластера	Число районов в кластере	*p*-значение	Число заболеваний в кластере	Численность группы риска
Первый период					
1991–1995 гг.	Брагин	15	**0,000 000**	191	1 867 139
Второй период					
1996–2000 гг.	Лоев	9	**0,000 000**	145	1 122 012
Третий период					
2002–2003 гг.	Пинск	10	**0,000 000**	32	450 147



Кластеры заболеваемости карциномой щитовидной железы, построенные с использованием пространственно-временной сканирующей статистики

Заключение

В работе рассмотрены критерии глобальной кластеризации и алгоритмы построения кластеров заболеваемости на изучаемой территории. В случае наличия аномальных наблюдений в данных о заболеваемости предложена робастная версия пространственной сканирующей статистики для построения кластеров. Критерии глобальной кластеризации позволили определить временные периоды, в которые наблюдался повышенный уровень заболеваемости карциномой щитовидной железы у популяции населения в возрасте до 18 лет на территории Республики Беларусь. С применением алгоритмов пространственного и пространственно-временного кластерного анализа построены кластеры заболеваемости карциномой щитовидной железы как отдельно по годам, так и заданным временным периодам.

Список литературы

1. Rogerson, P. A set of associated statistical tests for detection of spatial clustering / P. Rogerson // Ecological and Environmental Statistics. – 2005. – Vol. 12. – P. 275–288.

2. Kulldorff, M. A spatial scan statistic / M. Kulldorff // Common Statistics. Theory Methods. – 1997. – No. 26 (6). – P. 1481–1496.

3. Tango, T. A spatial scan statistic with a restricted likelihood ratio / T. Tango // Japanese Journal of Biometrics. – 2008. – Vol. 29, no. 2. – P. 75–95.

4. Tango, T. A flexibly shaped scan statistic for detecting clusters / T. Tango, K. Takahashi // Intern. J. of Health Geographics. – 2005. – Vol. 4. – P. 115–125.

5. A flexibly shaped scan statistic for disease outbreak detection and monitoring / K. Takahashi [et al.] // Intern. J. of Health Geographics. – 2008. – Vol. 7. – P. 85–98.

6. Swartz, J.B. An entropy-based algorithm for detecting clusters of cases and controls and its comparison with a method using nearest neighbours / J.B. Swartz // Health Place. – 1998. – Vol. 4 – P. 67–77.

7. A test to detect clusters of disease / A.S. Whittemore [et al.] // Biometrika. – 1987. – Vol. 74. – P. 631–635.

8. Abramovich, M.S. Robust spatio-temporal cluster analysis of disease / M.S. Abramovich, M.M. Mitskevich // Proc. of the 10th Intern. Conf. «Computer Data Analysis and Modeling». – Minsk : Publishing center of BSU, 2013. – Vol. 2. – P. 95–98.

9. Хьюбер, П. Робастность в статистике / П. Хьюбер. – М. : Мир, 1984. – 303 с.

Поступила 20.09.2014

[1]*НИИ прикладных проблем математики и информатики Белорусского государственного университета, Минск, пр. Независимости, 4*
e-mail: abramovichms@bsu.by

[2]*ООО «Ай-Джи Дев», Минск, ул. Интернациональная, 36-1*
e-mail: serg.anishchenko@gmail.com

[3]*Научно-практический центр детской онкологии, гематологии и иммунологии, Минский район, д. Боровляны, ул. Фрунзенская, 43*
e-mail: budanov@oncology.by.

M.S. Abramovich, S.V. Anishchanka, M.M. Mitskevich, O.I. Bydanov

SPATIO-TEMPORAL CLUSTER ANALYSIS OF DISEASE

The robust version of the spatial scanning statistics for clustering is proposed. Spatio-temporal cluster analysis algorithms were used for the cluster detection of incidence of thyroid carcinoma. Methods and algorithms of detection and building clusters for disease on studying territories are considered.