УДК: 519.2+004.93 **DOI: 10.35595/2414-9179-2020-4-26-257-265**

П.А. Украинский[1]

ВЫБОР ОПТИМАЛЬНОГО ПОРЯДКА СОСЕДСТВА ДЛЯ РАЗДЕЛЕНИЯ ПРОСТРАНСТВЕННОГО ТОЧЕЧНОГО ОБРАЗА НА КЛАСТЕРНУЮ И ШУМОВУЮ СОСТАВЛЯЮЩУЮ (НА ПРИМЕРЕ АНАЛИЗА РАЗМЕЩЕНИЯ АНТИЧНЫХ ПОСЕЛЕНИЙ НА КЕРЧЕНСКОМ ПОЛУОСТРОВЕ)

АННОТАЦИЯ

При выделении пространственных кластеров точечных объектов часто возникает проблема наличия шума, который мешает провести чёткие границы. Одним из популярных методов разделения кластерной и шумовой составляющей точечного образа является NNCR (Nearest Neighbor Clutter Removal), предложенный в 1998 г. Bayers и A.E. Raftery. Специфика метода заключается в использовании в расчётах расстояния до ближайшего соседа. При этом результат применения NNCR сильно зависит от выбранного пользователем порядка соседства. В этой работе описывается способ выбора оптимального порядка соседства для NNCR. Этот способ ориентирован на выполнение NNCR с помощью дополнительного пакета spatstat языка программирования R. Предлагается использовать в качестве основного критерия оптимального порядка соседства вероятность наличия в данных кластерной составляющей. При оптимальном порядке соседства её величина достигает максимального значения. В дополнение к этому предлагается анализировать вероятность принадлежности к кластеру для всех точек, причисленных к кластерной составляющей. Для этого строятся графики зависимости медианы и межквартильного размаха вероятности принадлежности от порядка соседства. С ростом порядка соседства медиана вероятности принадлежности для кластерной составляющей увеличивается, стремясь к значению 1,0. Межквартильный размах вероятности принадлежности, наоборот, с ростом порядка соседства уменьшается, стремясь к значению 0,0. Перегиб на этих графиках указывает на оптимальный порядок соседства. На языке программирования R написана пользовательская функция, позволяющая автоматизировать сравнение результатов NNCR, полученных при различных порядках соседства. Она возвращает матрицу, столбцами которой являются медиана вероятности принадлежности, межквартильный размах вероятности принадлежности и вероятность наличия в данных кластерной составляющей. Предложенный метод выбора оптимального порядка соседства опробован для анализа точечного слоя античных поселений Керченского п-ва. Для этих данных оптимальным оказался 3-й порядок соседства.

КЛЮЧЕВЫЕ СЛОВА: анализ точечных образов, античные поселения, пространственная кластеризация, удаление шума

[1] Белгородский государственный национальный исследовательский университет, Федерально-региональный центр аэрокосмического и наземного мониторинга объектов и природных ресурсов, ул. Победы, д. 85, 308015, Белгород, Россия; *e-mail:* **pa.ukrainski@gmail.com**

Pavel A. Ukrainskiy[1]

THE CHOICE OF THE OPTIMAL ORDER OF THE NEIGHBORHOOD FOR SEPARATION A SPATIAL POINT PATTERN INTO A CLUSTER AND NOISE COMPONENT (BY THE EXAMPLE OF ANALYSIS OF LOCATION OF ANTIQUE SETTLEMENTS IN THE KERCH PENINSULA)

ABSTRACT

When allocating spatial clusters of point objects, the problem of noise in the data often arises. This noise prevents clear boundaries of the clusters. One of the popular methods for separating the cluster and noise components of a point image is NNCR (Nearest Neighbor Clutter Removal), proposed in 1998 by Bayers and A.E. Raftery. The method is based on using the distance to the nearest neighbor in the calculations. The result of applying NNCR is highly dependent on the user selected neighborhood order. This paper describes a method for selecting the optimal neighborhood order for NNCR. This method focuses on the implementation of NNCR using the optional spatstat package of the programming language R. It is proposed to use the probability of the presence of a cluster component in the data as the main criterion for the optimal order of the neighborhood. With an optimal order of neighborhood, its value reaches its maximum value. In addition to this, it is proposed to analyze the probability of belonging to a cluster for all points assigned to the cluster component. For this, graphs of the dependence of the median and interquartile range of the probability of belonging on the order of the neighborhood are built. With an increase in the order of neighborhood, the median of the probability of belonging to the cluster component increases, tending to a value of 1.0. The interquartile range of the probability of belonging, on the contrary, decreases with an increase in the order of neighborhood, tending to a value of 0.0. The inflection in these graphs indicates the optimal order of the neighborhood. A user function is written in the programming language R, which makes it possible to automate the comparison of the NNCR results obtained in various orders of the neighborhood. It returns a matrix whose columns are the median of the probability of belonging, the interquartile range of the probability of belonging, and the probability of the presence of a cluster component in the data. The proposed method for choosing the optimal neighborhood order has been tested to analyze the point layer of ancient settlements of the Kerch Peninsula. For this data, the third order of neighborhood was optimal.

KEYWORDS: point pattern analysis, antique settlement, spatial clustering, clutter removal

ВВЕДЕНИЕ

При выделении пространственных кластеров точечных объектов нередко приходится сталкиваться с невозможностью однозначно провести границы кластеров. Такие ситуации возникают, когда между кластерами (плотными скопления точек) находится не пустое пространство, а шум — сильно разреженные группы точек [*Hennig, Coretto,* 2008]. Удаление шумовой составляющей из точечного образа позволяет чётче выделить существующие пространственные закономерности. В этом отношении разделение шумовой и кластерной составляющей аналогично по своему предназначению удалению выбросов из статистической выборки.

Визуальное разделение шумовой и кластерной составляющей является ненадёжным и субъективным. Количественные методы решения этой задачи разрабатывались в рамках такого раздела пространственной статистики, как анализ точечных образов. M. Ester с

[1] Belgorod State National Research University, Federal and Regional Centre for aerospace and ground monitoring of objects and natural resources, Pobedy str., 85, 308015, Belgorod, Russia; *e-mail:* **pa.ukrainski@gmail.com**

соавторами в 1996 г. разработали метод DBSCAN (Density-based Spatial Clustering of Applications with Noise), основанный на плотности точек [*Ester et al.,* 1996]. D. Allard и C. Fraley предложили в 1997 г. Метод, основанный на использовании полигонов Вороного [*Allard, Fraley,* 1997]. S. Bayers и A.E. Raftery в 1998 г. предложили метод NNCR (Nearest Neighbor Clutter Removal), основанный на расстоянии до ближайшего соседа [*Bayers, Raftery,* 1998].

Все 3 метода различаются не только по критериям, используемым для разделения кластеров и шума, но и по параметрам, которые задаются пользователем. Метод D. Allard и C. Fraley не требует от пользователя настройки параметров процедуры. Для DBSCAN необходимо задать радиус поиска, а для NNCR — порядок соседства. Вопрос выбора оптимального радиуса поискав настоящее время глубоко проработан и имеет множество решений *[Heidenreich et al.,* 2013]. А вот выбор оптимального порядка соседства проблема менее изученная. В этой работе мы предлагаем способ выбора оптимального порядка соседства для разделения точечного образа на кластерную и шумовую составляющую методом NNCR.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЙ

Чтобы изучить, как влияет порядок соседства на результаты NNCR, были использованы данные о расположении античных поселений на территории Керченского п-ва. Данные были собраны из различных источников, сведены в единый каталог и опубликованы Д.В. Бейлиным с соавторами. В этот каталог к поселениям был отнесён максимально широкий круг объектов — всё, что не может быть отнесено к каким-либо иным категориям археологических памятников [*Бейлин и др.*, 2014].

Первичная обработка исходных данных и картографирование конечных результатов работы выполнены в программе ArcGIS 10.5. Разделение данных на кластерную и шумовую составляющую выполнено с помощью языка программирования R 3.4.4[1] в интегрированной среде разработки RStudio 1.1.453. Использованы дополнительные пакеты spatstat [*Baddeley, Turner,* 2005] и rgdal.

Исходные данные представляли собой координаты поселений в географической системе координат WGS-84. Из них был создан шейп-файл с точечной геометрией. Его система координат была преобразована в систему координат проекции UTM 36N WGS-84. Далее шейп-файл был импортирован в R при помощи инструментов пакета rgdal[2].

Для разделения кластерной и шумовой составляющей точечного образа методом NNCR использовалась функция nnclean из дополнительного пакета spatstat. Она принимает на вход объект для анализа и порядок соседства. Возвращает функция следующие значения:

- вероятность наличия в пространственном точечном образе кластерной составляющей;
- вероятность принадлежности каждой точки к кластерной составляющей;
- класс каждой точки (принимает два значения — кластер или шум).

При определении класса точки функция nnclean причисляет к кластерной составляющей все точки с вероятностью принадлежности более 0,5.

Главная цель подбора порядка соседства при анализе методом NNCR — это получение результата с оптимальной степенью генерализации, без чрезмерной детализации или огрубления. Выбрать наилучший порядок соседства можно путём сопоставления результатов NNCR, полученных с разными порядками соседства. Визуальный анализ результатов, нанесённых на карту, для этого не подходит. Необходимо сравнение количественных

[1] R Core Team (2018). The R project for statistical computing. Web resource: https://www.R-project.org/ (accessed 15.01.2020)

[2] Rgdal: bindings for the geospatial data abstraction library. R package version 1.3-6. . URL: https://CRAN.R-project.org/package=rgdal / (Available at 15.01.2020)

показателей. Можно проанализировать вероятность наличия кластерной составляющей в данных. Оптимальным будет порядок соседства, при котором эта величина принимает наибольшие значения.

Также можно проанализировать, какие значения вероятности принадлежности к кластерной составляющей имеет совокупность точек, отнесённая функцией nnclean к классу «кластер». Обратить внимание следует на медиану и межквартильный размах этой величины. Можно также использовать среднее и стандартное отклонение, но эти показатели хуже, так как более чувствительны к наличию выбросов в данных.

Чтобы автоматизировать процесс сравнения результатов NNCR при разных порядках соседства была написана пользовательская функция на языке R. Ниже приведён её код.

```
OptimK <- function(X, k, ...) {
  library(spatstat)
  A <- vector("list", length = k)
  for (i in 1:k) A[[i]] <- nnclean(X, k = i)$marks$prob[nnclean(X, k = i)$marks$class ==
    "feature"]
  M <- matrix(nrow = k, ncol = 3)
  for (i in 1:k) M[i, 1] <- median(A[[i]])
  for (i in 1:k) M[i, 2] <- IQR(A[[i]])
  for (i in 1:k) M[i, 3] <- attributes(nnclean(X, k = i))$theta$p
  colnames(M) <- c("median", "IQR", "Cluster probability")
  return(M)
 }
```

Приведённая выше пользовательская функция возвращает матрицу, столбцами которой являются медиана вероятности принадлежности, межквартильный размах вероятности принадлежности и вероятность наличия в данных кластерной составляющей. На основе этих показателей для исследуемых объектов был выбран оптимальный порядок соседства и проведено разделение кластерной и шумовой составляющей. Результаты разделения были экспортированы из R в формате шейп-файла с помощью инструментов пакета rgdal.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ И ИХ ОБСУЖДЕНИЕ

Чтобы опробовать предложенный подход к определению оптимального порядка соседства, были сопоставлены результаты NNCR при порядках соседства от 1 до 15. На рис. 1 представлены результаты разделения шумовой и кластерной составляющих при крайних значениях этого показателя — 1-ом и 15-ом порядке.

При первом порядке соседства результат получается излишне детализированным. Кластерная составляющая, помимо крупных скоплений точек, включает в себя большое количество групп из 2–4 точек, окружённых шумом. При этом внутри крупных массивов кластерной составляющей обнаруживаются шум в виде отдельных точек или групп из 2–3 точек. Таким образом, кластерная и шумовая составляющие образуют ажурное переплетение.

При 15-ом порядке соседства результат получается излишне генерализированным. Кластерная и шумовая составляющие оказываются чётко разделены в пространстве. Они образуют почти непересекающиеся ареалы. Кроме того, по периметру кластерной составляющей в отдельных местах выделяется окаймляющая полоса шума. И плотность точек в этой полосе выше, чем остальной части шумовой составляющей, что заставляет подозревать ошибочное причисленные этих точек к шуму.

С ростом порядка соседства медиана вероятности принадлежности для кластерной составляющей увеличивается, стремясь к значению 1,0. Межквартильный размах вероятности принадлежности, наоборот, с ростом порядка соседства уменьшается, стремясь к значению 0,0. Чем больше медиана и меньше межквартильный размах, тем надёжнее разделение

кластерной и шумовой составляющих. Но стремиться к увеличению 1-го и уменьшению 2-го до бесконечности бессмысленно. Достаточно дойти до того порядка соседства, после которого изменения становятся незначительными. Если нанести эти показателя на графики, то на них можно обнаружить перегиб, после которого скорость изменений резко уменьшается. Точка перегиба указывает на оптимальный порядок соседства. Для античных поселений Керченского п-ва оптимальным оказался 3-ий порядок соседства (рис. 2).

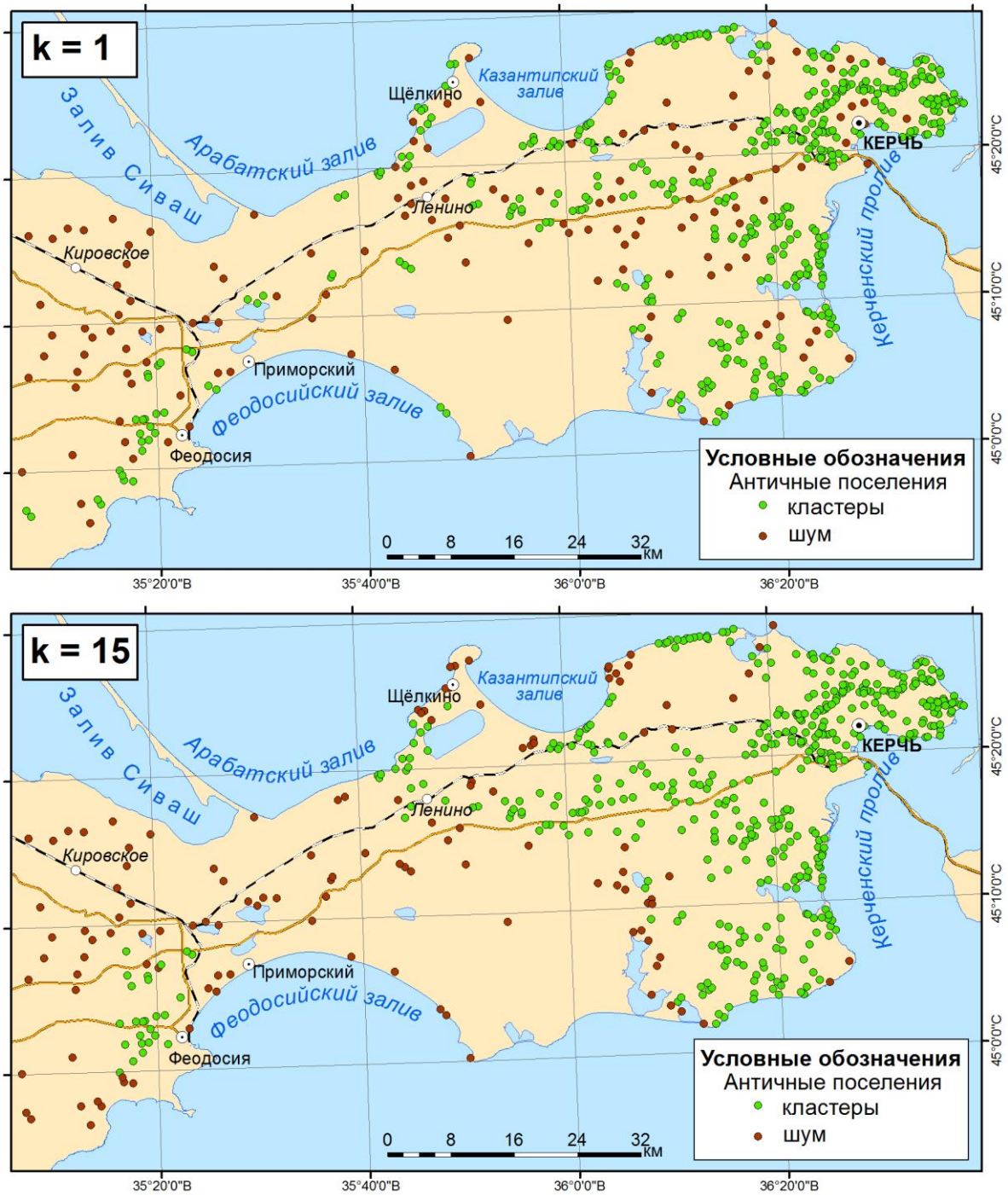

Рис. 1. Результат разделения кластерной и шумовой составляющей методом NNCR при 1-ом (вверху) и 15-ом (внизу) порядке соседства
Fig. 1. The result of separation of the cluster and noise components by the NNCR method in the 1st (top) and 15th (bottom) order of neighborhood

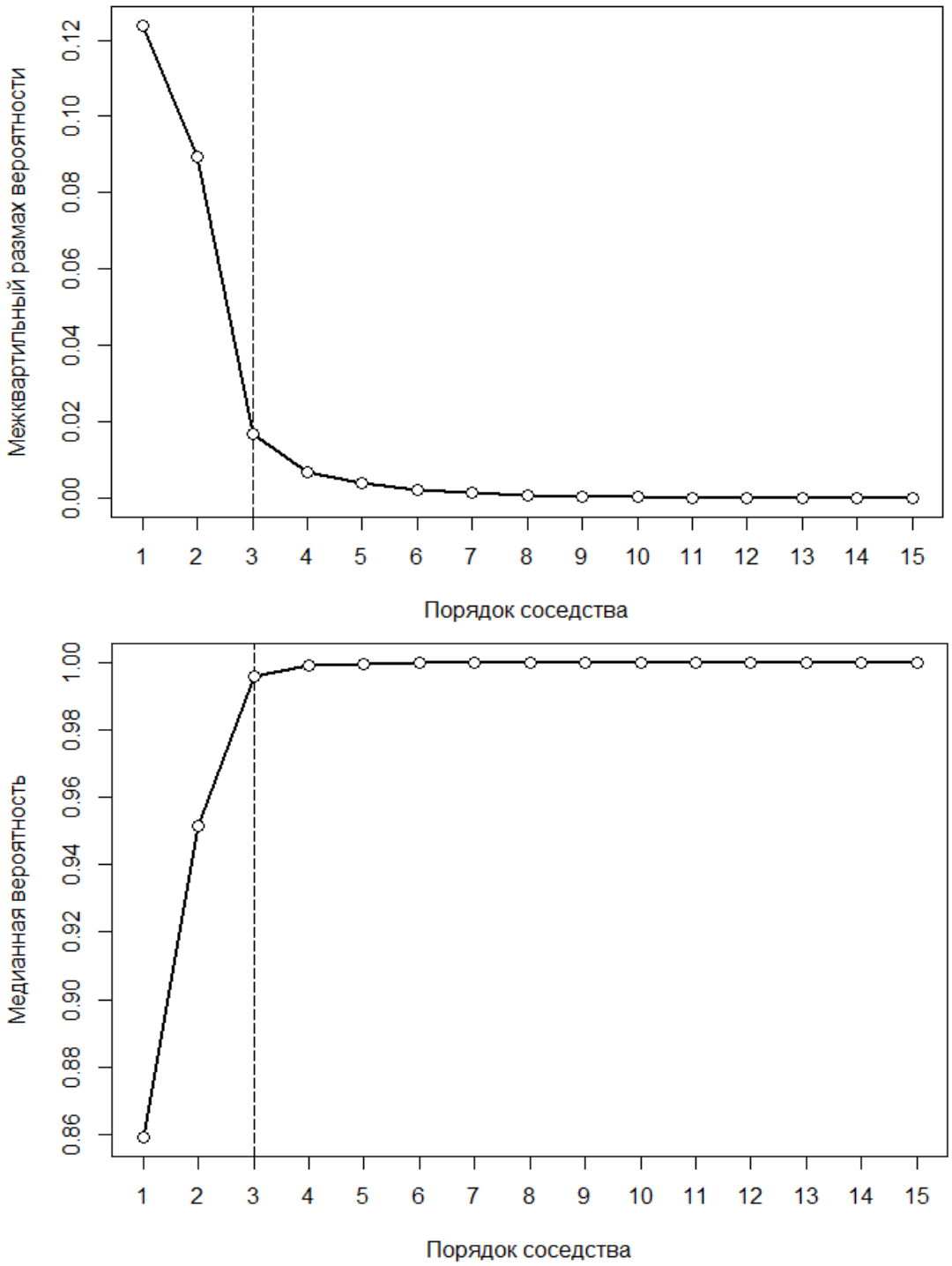

Рис. 2. Графики изменения медианы и межквартильного размаха вероятности принадлежности для кластерной составляющей
Fig. 2. Graphs of changes in the median and interquartile range of probability of belonging for the cluster component

Подобный графический метод известен в кластерном анализе как метод «локтевого сгиба» (elbow method). Он применяется при определении оптимального числа кластеров и часто критикуется за субъективность интерпретации. Перегиб на графиках может быть нечётким. В этом случае интерпретация графиков становится неоднозначной. Так, для античных поселений Керченского п-ва как перегиб графика межквартильного размаха можно трактовать точку и при 3-ем, и при 4-ом порядке соседства.

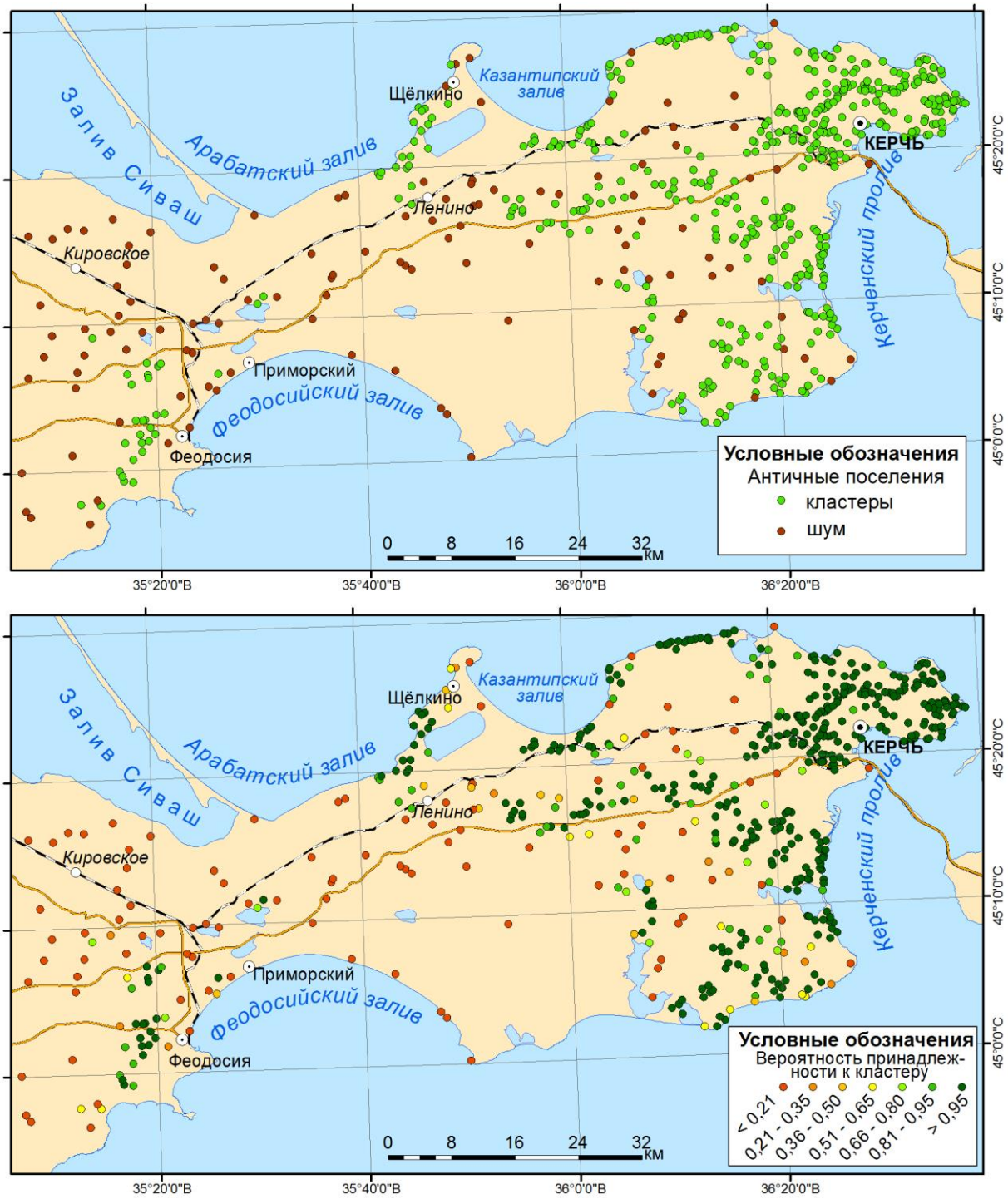

Рис. 3. Результат NNCR, полученный при оптимальном порядке соседства
Fig. 3. NNCR result obtained in optimal neighborhood order

Вероятность наличия в данных кластерной составляющей позволяет более однозначно установить оптимальный порядок соседства. Для анализируемых данных она колеблется от 0,713 до 0,816 и достигает максимума при 3-ем порядке соседства (табл. 1).

Для случаев, когда оптимальные порядки соседства, определённые по разным критериям, не совпадают, можно рекомендовать два варианта действий. Если совпадают 2 из 3-х значений, то они признаются оптимальными по принципу консенсуса. Если не совпадают все 3 значения, то в качестве оптимального порядка соседства принимается промежуточное значение.

Табл. 1. Результаты NNCR при разных порядках соседства
Table 1. The results of the NNCR at different orders of the neighborhood

Порядок соседства	Количество точек		Вероятность наличия кластерной составляющей
	Кластеры	Шум	
1	413	108	0,765
2	398	123	0,713
3	410	111	0,816
4	414	107	0,801
5	419	102	0,777
6	444	77	0,777
7	454	67	0,797
8	451	70	0,782
9	451	70	0,797
10	423	98	0,767
11	403	118	0,770
12	403	118	0,770
13	396	125	0,758
14	390	131	0,783
15	363	158	0,785

Результат NNCR для античных поселений Керченского п-ва, полученный при оптимальном (3-ем) порядке соседства показан на рис. 3. Картина взаимного расположения шумовой и кластерной составляющих является промежуточной между вариантами, показанными на рис. 1. Отдельно необходимо обратить внимание на географию вероятности принадлежности к кластерной составляющей. Точки, имеющие вероятность принадлежности от 0,5 до 0,95 — это в основном те точки, которые при крайних порядках соседства (1-ом и 15-ом) меняли свою принадлежность.

ВЫВОДЫ

При выполнении анализа точечного образа методом NNCR необходимо выбирать порядок соседства, который обеспечит оптимальный уровень генерализации результата, не допуская ни чрезмерной детализации, ни чрезмерного огрубления. В качестве критериев оптимального уровня соседства можно использовать вероятность наличия кластерной составляющей в точечном образе, медианное значение и межквартильный размах вероятности принадлежности для кластерной составляющей. Эти показатели просты в интерпретации, а написанная на языке R пользовательская функция убыстряет и упрощает их расчёт.

БЛАГОДАРНОСТИ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-00-00562.

ACKNOWLEDGEMENTS

The study was funded by the Russian Foundation of Basic Research, grant No 18-00-00562.

СПИСОК ЛИТЕРАТУРЫ

1. *Бейлин Д.В., Ермолин Е.Л., Масленников А.А., Смекалов С.Л.* Античные поселения Европейского Боспора эллинистического времени (каталог памятников). Древности Боспора, 2014. Т.18. С. 35–72.

2. *Allard D., Fraley C.* Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoi tessellation. Journal of the American statistical Association, 1997. V. 92. P. 1485–1493. DOI: 10.2307/2965419.
3. *Baddeley A., Turner R.* Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software, 2005. V. 12. No 6. P. 1–42. DOI: 10.18637/jss.v012.i06.
4. *Byers S., Raftery A.E.* Nearest-neighbor clutter removal for estimating features in spatial point processes. Journal of the American Statistical Association, 1998. V. 93. P. 577–584. DOI: 10.2307/2670109.
5. *Ester M., Kriegel H.-P., Sander J., Xu X.* A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland: AAAI Press, 1996. P. 226–231.
6. *Heidenreich N.B., Schindler A., Sperlich S.* Bandwidth selection for kernel density estimation: A review of fully automatic selectors. AStA Advances in Statistical Analysis, 2013. V. 97. No 4. P. 403–433.
7. *Hennig C., Coretto P.* The noise component in model-based cluster analysis. Data Analysis, Machine Learning and Applications. Berlin–Heidelberg: Springer, 2008. P. 127–138.

REFERENCES

1. *Allard D., Fraley C.* Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoi tessellation. Journal of the American statistical Association, 1997. V. 92. P. 1485–1493. DOI: 10.2307/2965419.
2. *Baddeley A., Turner R.* Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software, 2005. V.12. No 6. P. 1–42. DOI: 10.18637/jss.v012.i06.
3. *Beilin D.V., Ermolin E.L., Maslennikov A.A., Smekalov S.L.* Antique settlements of European Bosporus of Hellenistic time (catalog of monuments). Antiquities of the Bosporus, 2014. V. 18. P. 35–72 (in Russian).
4. *Byers S., Raftery A.E.* Nearest-neighbor clutter removal for estimating features in spatial point processes. Journal of the American Statistical Association, 1998. V. 93. P. 577–584. DOI: 10.2307/2670109.
5. *Ester M., Kriegel H.-P., Sander J., Xu X.* A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland: AAAI Press, 1996. P. 226–231.
6. *Heidenreich N.B., Schindler A., Sperlich S.* Bandwidth selection for kernel density estimation: a review of fully automatic selectors. AStA Advances in Statistical Analysis, 2013. V. 97. No 4. P. 403–433.
7. *Hennig C., Coretto P.* The noise component in model-based cluster analysis. Data Analysis, Machine Learning and Applications. Berlin–Heidelberg: Springer, 2008. P. 127–138.