Л.А.Солнцев[1], В.М.Дубянский[2]

ОПЫТ ИСПОЛЬЗОВАНИЯ МЕТОДА МАКСИМАЛЬНОЙ ЭНТРОПИИ (MAXENT) ДЛЯ ЗОНИРОВАНИЯ ТЕРРИТОРИИ ПО РИСКУ ЗАРАЖЕНИЯ ГЛПС НА ПРИМЕРЕ НИЖЕГОРОДСКОЙ ОБЛАСТИ

[1]НИИ эпидемиологии и микробиологии им. акад. И.Н.Блохиной, Нижний Новгород; [2]Ставропольский противочумный институт

Цель. Зонирование территории Нижегородской области по риску заражения ГЛПС с использованием метода Maxent. *Материалы и методы.* Материалами являлись данные Центра гигиены и эпидемиологии по Нижегородской области по каждому случаю заражения ГЛПС за 2010 — 2016 гг.; данные по условиям окружающей среды (Bioclim); данные по вегетационной активности (MODIS). Обработка проводилась в пакетах ArcGIS 10.2.2 и Maxent 3.3.3к. *Результаты.* Получена и валидирована модель для оценки потенциального риска заражения ГЛПС на территории Нижегородской области. *Заключение.* Полученные результаты не противоречат фактически наблюдаемой пространственной локализации случаев заражения ГЛПС (точность предсказания составляет более 75%), выявляют связь между пространственной локализацией случаев заражения ГЛПС и сочетания факторов среды и позволяют формировать прогнозы изменения границ потенциально опасных участков при изменении факторов среды.

Журн. микробиол., 2017, № 5, С. 39—45

Ключевые слова: ГЛПС, Maxent, пространственное моделирование, Bioclim, NDVI, ГИС

L.A.Solntsev[1], V.M.Dubyansky[2]

EXPERIENCE OF USING MAXIMAL ENTROPY METHOD (MAXENT) FOR ZONING OF THE TERRITORY BY HFRS RISK USING NIZHNY NOVGOROD REGION AS AN EXAMPLE

[1]Blokhina Research Institute of Epidemiology and Microbiology, Nizhny Novgorod; [2]Stavropol Institute of Plague Control, Russia

Aim. Zoning of the territory of Nizhny Novgorod region by risk of HFRS infection using Maxent method. *Materials and methods.* Data from Centre of Hygiene and Epidemiology in Nizhny Novgorod region for each case of the HFRS for 2010 — 2016, data on environment (Bioclim), data on vegetation activity (MODIS) were used. ArcGIS 10.2.2 and Maxent 3.3.3к packages were used. *Results.* Model for evaluation of potential risk of HFRS in Nizhny Novgorod was developed and validated. *Conclusion.* The data obtained do not contradict the observed spatial localization of the cases of HFRS infection (prediction accuracy over 75%), detected connection between spatial localization of HFRS cases and combination of environment factors and allow to predict changes in borders of potentially dangerous segments after environmental changes.

Zh. Mikrobiol. (Moscow), 2017, No. 5, P. 39—45

Key words: HFRS, Maxent, spatial modelling, Bioclim, NDVI, GIS

ВВЕДЕНИЕ

При изучении закономерностей пространственного распределения случаев природно-очаговых заболеваний первоочередной задачей является выявление границ очагов. Очаг во многом определяется экологической нишей, которую занимают организмы — носители возбудителя. Знание о простран-

ственных границах природных очагов позволяет более качественно проводить профилактические мероприятия, направленные на снижение риска заражения.

Одной из актуальных природно-очаговых инфекций для территории Российской Федерации и, в частности, Нижегородской области является геморрагическая лихорадка с почечным синдромом (ГЛПС). Надзор за этой инфекцией и ее регистрация регламентируются [2]. Данный документ определяет мышевидных грызунов как единственный источник заражения ГЛПС и указывает связь между локализацией случаев заражения и ареалом переносчика.

Сама численность переносчиков определяется условиями окружающей среды, формирующей фундаментальную и реализованную экологические ниши [1]. Поэтому, используя данные по факторам окружающей среды и ретроспективные данные локализации случаев заражения человека, можно сделать попытку выделить на изучаемой территории участки, наиболее благоприятные по сочетанию условий для заражения ГЛПС (по сути наиболее благоприятные для обитания грызунов-носителей). Такой опосредованный путь диктуется тем, что сбор зоологического материала, как правило, ведется с низкой степенью детализации. Данные усредняются до административного района области. Границы очагов редко совпадают с административными границами, и один очаг может располагаться на территории нескольких районов. Учет случаев заражения ведется с точностью до населенного пункта (по возможности) и является более детальным с пространственной точки зрения. При этом следует учитывать, что часть случаев имеет систематическую ошибку места локализации, поскольку в материалах регистрации случаев ГЛПС указывается предполагаемое место заражения. Формально задача формулируется так: полагая, что сочетания условий окружающей среды в окрестностях населенного пункта, где был зарегистрирован случай заражения, является оптимальными для обитания переносчика инфекции, нужно на всей исследуемой территории определить участки со схожими условиями среды. Степень такой схожести желательно получать в виде некой безразмерной величины (расчетной вероятности) и иметь возможность провести статистические тесты качества расчетов.

Целью исследования являлось зонирование территории Нижегородской области по потенциальному риску заражения человека ГЛПС. Основной задачей — создание вероятностной модели пространственного распределения риска заражения в зависимости от сочетания условий среды.

Аналогичные задачи уже достаточно давно решаются в экологии с использованием так называемых моделей распределения видов (SDM, species distribution models).

МАТЕРИАЛЫ И МЕТОДЫ

Модели распределения видов — широкий класс моделей, которые позволяют установить связь между фактом встречи вида в какой-либо точке пространства и сочетанием условий окружающей среды в данной точке [7]. В работе [5] приведен большой обзор моделей и примеров их использования. Исходными данными для моделей является информация о точках встречи изучаемого вида. Если данные содержат информацию о присутствии/отсутствии вида, то, как правило, используются статистические модели общего назначения [3, 9]. Для случаев, когда есть данные только о присутствии вида, наиболее применимым методом является Maxent. В работе [6] приведены

ссылки на работы, где данный метод применялся как самостоятельно, так и в сравнении с другими методами.

Метод максимальной энтропии (Maxent), предложенный в работах [12, 13], является SDM, специально разработанной для случаев, когда известны точки присутствия вида, но нет точных данных о его отсутствии на определенной территории [6, 10, 14, 15]. Для моделирования помимо координат точек присутствия вида используется набор прямоугольных решеток, полностью покрывающих исследуемую территорию, каждая из которых описывает тот или иной фактор среды. Значение каждой ячейки решетки представляет собой величину фактора на территории, ограниченной ячейкой. Например, для фактора «годовое количество осадков» значение ячейки будет представлять собой количество осадков (в мм), которое выпадает на территории, ограниченной ячейкой, за год. Чем меньше размер ячейки, тем более детально описывается фактор среды. Все такие решетки должны иметь одинаковый пространственный охват (описывать одну и ту же территорию) и одинаковый размер ячеек. Накладывая на такую решетку (решетки) точки с координатами встречи с видом мы получаем значения набора параметров окружающей среды для точек встречи. Результатом моделирования является пространственная решетка, где для каждой ячейки рассчитана вероятность встречи вида. Интерпретируя терминологию модели в рамках нашего исследования, «точка встречи» будет являться локализацией случая заражения ГЛПС, а «вероятность встречи» — числовой оценкой того, насколько в данной ячейке сочетания условий среды будут благоприятны для риска заражения.

Для оценки модели используется ROC (receiver operating characteristic) — анализ, показывающий соотношение чувствительности модели к специфичности. В случае Maxent в качестве меры чувствительности выступает величина, представляющая собой долю ячеек с известным наличием вида, для которых модель предсказала отсутствие его в данной ячейке [13]. Этот показатель называется «омиссия». Для количественной оценки ROC анализа используется критерий AUC (area under curve).

Можно обозначить следующие диапазоны критерия AUC: 1 — 0.9 — отлично, 0.9 — 0.8 — хорошо, 0.8 — 0.7 — допустимо, 0.5 — 0.7 — плохо, 0 — 0.5 — недопустимо (модель дает менее точный прогноз, чем случайное предсказание)

Следует отметить, что модель Maxent не позволяет решить все проблемы, связанные с пространственным моделированием. Так, критический анализ данной модели и сопоставление ее с моделью логистической регрессии даны в работе [8]. В работе [11] помимо детального анализа самой модели рассмотрены вопросы, связанные с характером исходных данных, выбором параметров для моделирования, интерпретацией получаемых результатов.

Использованы данные Центра гигиены и эпидемиологии по Нижегородской области по каждому случаю заражения ГЛПС, указанному с точностью до населенного пункта на территории Нижегородской области, за 2010 — 2016 гг. Для получения пространственных координат населенных пунктов, в которых были зарегистрированы больные, проведена процедура геокодирования с использованием сервиса Nominatim (http://nominatim.openstreetmap.org). В качестве топоосновы использовались материалы OpenStreetMap (https://www.openstreetmap.org). Каждый случай заражения представляли в виде точки на карте с координатами, соответствующим координатам населенного пункта, где произошло заражение.

Данные по условиям окружающей среды были получены из банка данных

Biolclim (http://www.worldclim.org/bioclim). Для учета состояния растительности (и типа землепользования) были использованы спутниковые данные MODIS (MOD13A3, https://modis-land.gsfc.nasa.gov/vi.html), представляющие собой многолетние помесячно усредненные данные по значению вегетационного индекса NDVI (Normalized Difference Vegetation Index — Нормализованный относительный индекс растительности [4]).

В качестве обучающей выборки использованы данные по случаям заражения людей за 2010 — 2015 гг. В качестве проверочной выборки послужили данные за 2016 г.

Для предварительной подготовки данных по условиям окружающей среды (обрезки по границе территории, приведение всех растров к единому размеру пикселя в 1 км^2, конвертирование в формат asc) использовалась программа ArcGIS 10.2.2. Моделирование проводилось в программе MaxEnt 3.3.3к (https://www.cs.princeton.edu/~schapire/maxent).

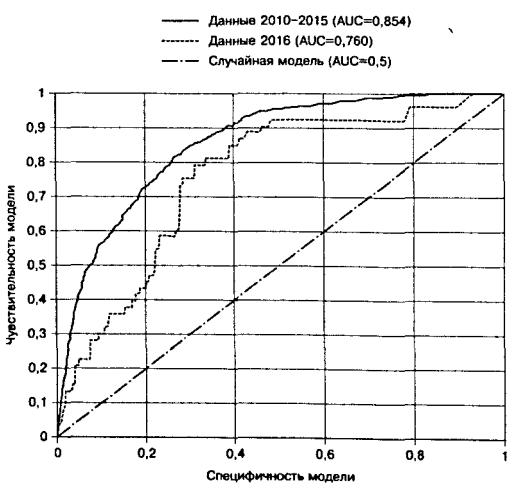
РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Согласно полученным результатам, модель можно оценить как хорошую (AUC=0.854) (рис.). Проверка полученной модели на данных 2016 года показала ее достаточно высокую прогностическую способность (AUC=0.76). Моделирование проводили в 2 этапа. Все результаты доступны для скачивания по следующим ссылкам:

1. Этап 1. Моделирование с использованием всех факторов окружающей среды и с использованием методов для рандомизации для оценки устойчивости (https://mail.nniiem.ru/owncloud/s/f48cEtd8vaf5uAQ) .

2. Этап 2. Моделирование с использованием обучающей и тестовой выборки (https://mail.nniiem.ru/owncloud/s/bf9cOlJGxiovhM0).

Модель выдает значение в безразмерной величине на шкале от 0 до 1. Для удобства дальнейшей работы все значения полученной решетки были разбиты на классы. Выбор границ классов был основан на анализе характера гистограммы распределения полученных данных. Распределение характеризовалось большим количеством ячеек в диапазоне значений от 0 до 0.1, примерно равным числом значений ячеек в диапазоне 0.2 — 0.7 и небольшим числом значений в диапазоне 0.7 — 1. Таким образом, использование метода равных частот на основе квартилей привело бы к необоснованному отнесению всех ячеек с значениями от 0.5 и выше к 4 классу (класс наибольшего риска) (рис.).



Результаты ROC анализа полученной модели для тренировочных и тестовых данных.

Исходя из этого, был применен метод равных интервалов и определены следующие границы классов для перехода от количественных к качественным показателями: 0 — 0.2 — зона очень низкого риска, 0.2 — 0.4 — зона низкого риска, 0.4 — 0.6 — зона среднего риска, 0.6 – 0.8 — зона высокого риска, 0.8 — 1 — зона очень высокого риска.

Проведенные с использованием модели расчеты показывают, что тер-

ритория Нижегородской области не однородна по риску заражения ГЛПС. Можно выделить участки очень высокого риска: восточная часть Борского района вдоль р. Волга, Кстовский район и его граница с Нижним Новгородом, северная часть Богородского района, центральные части Семеновского, Уренского и Тоншаевского районов, территории Краснобаковского, Варнавинского и Ветлужского района вдоль реки Ветлуга.

К зонам высокого риска можно отнести восточную часть Кстовского района вдоль р. Волга, Лысковский район, восточную часть Ковернинского района и западную часть Сокольского района. Сама территория Нижнего Новгорода была исключена из анализа по причине того, что данный населенный пункт, являясь крупной городской агломерацией, обладает резко отличающимися от остальной области условиями среды.

Особое внимание следует обратить на зоны, локализованные на границе Вачского и Навашинского районов, севере Володарского района, западной часть Тонкинского района. Данные участки относятся к территориям высокого потенциального риска, но при этом до сих пор случаев заражения ГЛПС там зарегистрировано не было. Причиной могло послужить недостаточное пространственное разрешение части данных. Примерно 1/3 имеющихся данных по случаям заражения ГЛПС имеет пространственную привязку до уровня административного района. Такие данные не могут быть использованы в процессе моделирования Следовательно, мы можем предположить, что на данных территориях имели место случаи заражения ГЛПС, но данная информация была утрачена в процесс регистрации. Мы полагаем, что на данные зоны следует обратить особое внимание в процессе санитарно-эпидемиологического надзора.

Стоит отметить, что характер исходных данных (а именно их привязанность к населенным пунктам) достаточно сильно влияет на результат моделирования. Фактически модель определяет те населенные пункты, сочетания условий среды в окрестностях которых наиболее благоприятны для риска заражения. Под «благоприятностью» в данном случае следует понимать сочетание условий в данной точке пространства, схожее с таковым в точке пространства, где было фактически зарегистрирован случай заражения ГЛПС.

Процедура моделирования проводилась в 2 этапа. На первом этапе использовались все переменные из набора BioClim (19 переменных) и 9 переменных из набора NDVI. Поскольку метод расчета NDVI в зимние месяцы не позволяет качественно раз-

Вклад переменных по результатам второго этапа моделирования

Переменная	Описание	Вклад, %
ndvi_04	Уровень вегетационной активности апреля	25,4
bio_2	Размах среднесуточной температуры	15,1
ndvi_10	Уровень вегетационной активности октября	14,6
ndvi_08	Уровень вегетационной активности августа	14,6
ndvi_07	Уровень вегетационной активности июля	9,2
bio_9	Средняя температура самого засушливого квартала	5,2
ndvi_09	Уровень вегетационной активности сентября	4,2
ndvi_05	Уровень вегетационной активности мая	2,9
bio_18	Количество осадков самого теплого квартала	2,2
bio_4	Сезонность колебаний температуры	1,9
ndvi_06	Уровень вегетационной активности июня	1,3
bio_19	Количество осадков самого холодного квартала	1,3
bio_6	Минимальная температура самого холодного месяца	0,9
bio_15	Сезонность уровня осадков	0,6
bio_8	Средняя температура самого влажного квартала	0,5
bio_5	Максимальная температура самого теплого месяца	0,1

личить типы растительности, а по большей части, просто дифференцирует водную поверхность, открытую поверхность и поверхность, занятую лесом, это временной период был исключен из моделирования.

В процессе моделирования алгоритм Maxent также определяет условную «важность» каждой переменной для итоговой модели. Суммарный вклад всех переменных принимается за 100%. После первого этапа были выбраны переменные, вклад которых составлял более 1%. Для второго этапа моделирования были использованы только они. В табл. приведены вклады каждой переменной по результатам второго этапа.

Мы ранжировали процентный вклад каждого фактора и условно считали наиболее важными из них те, значения которых превышали медиану (2,55%). Согласно полученным результатам, факторами, значения которых определяют итоговую вероятность, являются вегетационная активность в апреле, июле—августе и сентябре—октябре, размах среднесуточной температуры, средняя температура самого засушливого квартала. Таким образом, определяющими являются факторы, которые непосредственно влияют на процесс формирования кормовой базы мышевидных грызунов — переносчиков возбудителя ГЛПС.

Полученная в результате моделирования карта зон риска заражения ГЛПС на территории Нижегородской области не противоречит данным, получаемым при районировании территории Нижегородской области по числу случаев ГЛПС (суммарно по районам). Выделенные в результате моделирования участки высокого и очень высокого риска располагаются в районах, где фиксируется самое большое число случаев заражения. При этом карта зон позволяет более детально обозначить границы потенциально опасных территорий, не привязываясь к административному делению, что невозможно сделать при использовании данных исключительно на уровне административных районов.

Особый интерес представляют участки, определенные как зоны высокого или очень высокого риска, на территории которых нет населенных пунктов, где были зарегистрированы случаи заражения ГЛПС. На них следует обратить внимание при планировании и проведении мероприятий, связанных с профилактикой ГЛПС в природном очаге. Также выделены участки, которые определенны как зоны высокого или очень высокого риска, локализованные на административной границе области. Можно предположить, что в данном случае мы наблюдаем распространение зоны повышенного риска за границы Нижегородской области. В данном случае для повышения эффективности профилактических мероприятий необходимо проводить изучение и зонирование территории Нижегородской области в сочетании с аналогичными исследованиями на территориях ее ближайших соседей.

Таким образом, мы получили модель, позволяющую разделить исследуемую территорию по степени риска заражения человека ГЛПС с высокой (AUC=0.854) степенью достоверности и значимой прогностической ценностью (AUC=0.76). Использование модели позволяет получить новые, более детальные с пространственной точки зрения данные о границах потенциально опасных в аспекте ГЛПС участков области. В особенности, это касается тех территорий, где ранее не отмечались случаи заражения ГЛПС.

В перспективе модель может быть использована для прогнозирования изменения границ зон риска заражения ГЛПС при изменении условий среды

(увеличение среднемесячной температуры, смена режима увлажнения, смена характера растительности). Появляется возможность разрабатывать превентивные мероприятия по контролю за природно-очаговыми инфекциями, опираясь на общепринятые климатические модели. Подобный подход может быть полезен при моделировании распространения других инфекций, переносчиком возбудителей которых являются клещи и комары.

ЛИТЕРАТУРА

1. Джиллер П. Структура сообществ и экологическая ниша. М., Мир, 1988.
2. Санитарно-эпидемиологические правила СП 3.1.7.2614-10 «Профилактика геморрагической лихорадки с почечным синдромом», утвержд. Постановлением Главного санитарного врача от 26 апреля 2010 г., № 38.
3. Corsi F., de Leeuw J., Skidmore A. Modeling species distribution with GIS. *In*: Boitani L., Fuller T. (Eds.). Research techniques in animal ecology. New York, Columbia University Press, 2000, p. 389-434.
4. Crippen R. E. Calculating the Vegetation Index Faster. Remote Sensing of Environment. 1990, 34: 71-73.
5. Elith J., Leathwick J.R. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution and Systematics. 2009, 40: 677-697.
6. Elith J., Phillips S. J., Hastie T. et al. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions. 2011, 17: 43-57.
7. Franklin J. Mapping species distributions: spatial inference and prediction. Cambridge University Press, 2009.
8. Gastón A., García-Viñas J.I. Modelling species distributions with penalized logistic regressions: A comparison with maximum entropy models. Ecol. Model. 2011, 222 (13): 2037-2041.
9. Guisan A., Zimmerman N.E. Predictive habitat distribution models in ecology. Ecol. Model. 2000, 135: 147-186.
10. Liu H.-N., Gao- L.-D., Chowell G. et al. Time-specific ecologic niche models forecast the risk of haemorrhagic fever with renal syndrome in Dongting Lake District, China, 2005–2010. PLOS ONE. 2014, 9 (9): e106839.
11. Merow C., Smith M.J., Silander J.A. A practical guide to Maxent for modeling species' distributions: what it does, and why inputs and settings matter. Ecography. 2013, 36 (10): 1058-1069.
12. Phillips S.J., Anderson R.P., Schapire R.E. Maximum entropy modeling of species geographic distributions. Ecol. Mod. 2006, 190: 231-259.
13. Phillips S.J., Dudík M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography. 2008, 31: 161-175.
14. Wei L., Qian Q., Wang Z.Q. et al. Using geographic information system-based ecologic niche models to forecast the risk of hantavirus infection in Shandong Province, China. Am. J. Trop. Med. Hyg. Mar. 2011, 84 (3): 497-503.
15. Zeimes C.B., Olsson G.E., Ahlm C. Modelling zoonotic diseases in humans: comparison of methods for hantavirus in Sweden. Int. J. Health Geogr. 2012, 11: 39.

Поступила 05.03.17

Контактная информация: Солнцев Леонид Аркадьевич, к.б.н.,
603950, Нижний Новгород, ул. Малая Ямская, 71, р.т. (831)469-79-01