

ГЕОЭКОЛОГИЯ

УДК 502.36

В.Н. ЭКЗАРЬЯН, В.В.РУКАВИЦЫН

**МЕТОДИКА ОЦЕНКИ УРОВНЯ УСТОЙЧИВОСТИ ЭКОСИСТЕМ
ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ**

Представлена методика определения уровня устойчивости экосистем при помощи инструментов машинного обучения. До настоящего времени проведение подобной оценки было доступно только профессионалам высокого класса, что затрудняло ее широкое применение. Однако использование машинного обучения для обработки исходной информации на основе логики принятия решений позволяет автоматизировать эту процедуру и сделать её более быстрой и доступной для широкого круга специалистов. Дано практическое описание использования технологии автоматизации прогнозов, позволяющих определить риски для проживания населения на определенной территории. Построена модель устойчивости экосистем России. Точность моделирования превысила 80 %. Найден наилучший алгоритм обработки данных и наилучшие его параметры для компьютерного определения уровня устойчивости экосистем, что позволяет автоматизировать сложный процесс экспертной классификации территории.

Ключевые слова: моделирование; устойчивость экосистем; машинное обучение; автоматизация; врожденная заболеваемость.

Машинное обучение — это научная дисциплина, которая занимается проектированием и разработкой алгоритмов, позволяющих компьютерам изменить своё поведение на базе некой информации, например, на базе информационных сенсоров или базы данных [16]. Одним из основных направлений исследований машинного обучения является способность компьютера учиться автоматически решать сложные проблемы и принимать решения, основанные на обработке фактического материала.

Некоторые системы машинного обучения пытаются устранить участие человека в анализе данных, в то время как другие принимают совместное участие человека и машины. Человек, однако, не может быть полностью устранен, так как разработчик системы должен указать, каким образом данные должны быть представлены и какие механизмы будут использоваться для поиска характеристик данных. Машинное обучение может рассматриваться, как попытка автоматизировать часть научного метода.

Авторами рассмотрен прогноз природно-техногенных процессов при помощи машинного обучения.

При исследовании экосистем, которые являются сложными, самоорганизующимися, саморегулирующимися системами необходимо определять свойства и возможности средств наблюдения за динамическими, геофизическими, геодезическими процессами окружающей среды.

Геофизические объекты, процессы и явления характеризуются: сложной уровневой организацией; мощными энергетическими возможностями; широкими диапазонами проявления во времени и в пространстве; чувствительностью к начальным условиям; наличием частично автономных и изменчивых подсистем; массовым случайным взаимодействием элементов; существованием некоего общего начального уровня (информационной составляющей природных геофизических процессов) и обратных связей, обеспечивающих целостность, самоорганизацию и саморегулирование систем.

Геофизические поля являются особой формой материи, которая объединяет массивы горных пород в системы геологических тел. Геофизические поля осуществляют передачу взаимодействия геологических тел между собой, удерживают гидросферу и атмосферу, поддерживают процессы энер-

гообмена, необходимые для существования жизни на Земле. Совокупность всех геофизических полей будем называть геополем. Геополем можно назвать совокупность значений различных физических величин (параметров), количественно определяющих естественное или искусственное наведенное физическое поле (или отдельные его элементы) в пределах определенной территории.

Анализ природных полей, изучение пространственно-временной динамики их развития, это ключ к оценке и прогнозу проявления экзогенных процессов (подтоплений, оползней, провалов), оказывающих большое влияние на процессы урбанизации и безопасность проживания людей. Важность данной проблемы очевидна, так как с ней связаны проявления многих аварийных ситуаций — обрушение жилых домов, повреждения инженерных сооружений и коммуникаций и т. д.

Природные поля имеют различный уровень экологической устойчивости, т. е. способности к ассимиляции техногенной нагрузки и восстановлению нарушенных связей и функций. Такие условия определяются положением геодинамических зон унаследованной стабилизации. Кроме того, геодинамические процессы не только определяют места формирования и время активизаций тектонически-активных зон (а, следовательно, время и место проявления опасных и аварийных техногенных ситуаций), но и сами несут в себе факторы риска хозяйственной деятельности, безопасности проживания и здоровья населения.

Для определения зон стабильного и нестабильного состояния экосистем и использовалось машинное обучение. По уровню стабильности экосистемы территории можно определить интенсивность развития природно-техногенных процессов.

Общую методологию можно разделить на две части: обучение и тестирование. Обучение необходимо для создания базовой модели на обучающей выборке. Тестирование производится для проверки точности модели, её корректировки и оценки возможности решения реальных задач. После опробования модели она будет проверена в решении реальной задачи классификации и должна будет соответствовать критериям, с которыми будут сравниваться значения классов. Таким образом, после прохождения всех уровней будет получена рабочая система автоматизации принятия решения об определении уровня устойчивости экосистемы по параметрам геополей. На основе полученной информации можно будет сделать вывод о риске для населения проживания на изучаемой территории.

Для каждого алгоритма были выбраны наилучшие параметры, подходящие конкретно для имеющейся обучающей выборки. Параметры подбирались также экспериментально и выбирались для каждого алгоритма отдельно до получения лучшего результата классификации.

Таким образом, необходимо выбрать наилучшую модель определения уровня устойчивости экосистем с наиболее подходящим алгоритмом автоматизации логического принятия решения о классификации.

Лучший результат был показан алгоритмом Random Forest [12] с более чем 98 % точности. Это означает, что компьютерная модель может скопировать логику человека-эксперта в такой классификации.

Оценка точности модели

Для решения задачи проверки полученных моделей и окончательного определения лучшего алгоритма классификации использовались данные магнитного поля, гравитационного поля и рельефа для всей территории России. Данные не были классифицированы экспертом, и эта задача была доверена компьютеру. Этот способ позволяет проверить модель в реальной работе. Все оценки тестовых данных должны производиться только с помощью компьютера, а затем результаты могут быть проанализированы отдельно. Дело в том, что метод оценки экосистемы должен быть универсальным и одинаково хорошо работать в разных масштабах и с данными полученными на разных территориях. Именно для тестирования универсальности предполагаемой модели выбраны такие разные наборы данных. Данные включали в себя информацию о координатах точек замера геофизических полей, а также магнитном поле, гравитационном поле и абсолютной высотной отметке в этих точках. Для оценки точности прогнозирования была использована величина уровня врожденной заболеваемости в разных субъектах РФ. В результате сравнивалась предсказанная устойчивость экосистемы с реальной ситуацией и оценивалась точность модели.

Для построения модели территории России были извлечены данные из различных карт: о величине магнитной индукции (Тл), ускорения свободного падения (Гал) и абсолютной высотной отметке (м). Были взяты карты гравитационного поля [2], магнитного поля [4] и карта рельефа России [3]. Эти карты оцифровывались, вся информация представлена в сводной таблице. Разработана единая система точек, поскольку координаты точек замеров каждого из параметров отличались между собой. Затем данные таблицы были обработаны.

Тестовая выборка включала 5900 точек, в каждой из которых были получены данные о геофизических полях, рельефе и координатах. Для построения модели использовано два алгоритма: Bagging [11] и Random forest [12].

На рис. 1 и 2 показаны результаты описанной классификаций.

Для анализа достоверности модели необходимо сравнить результаты моделирования (рис. 1, 2) с данными о врожденной заболеваемости в России.

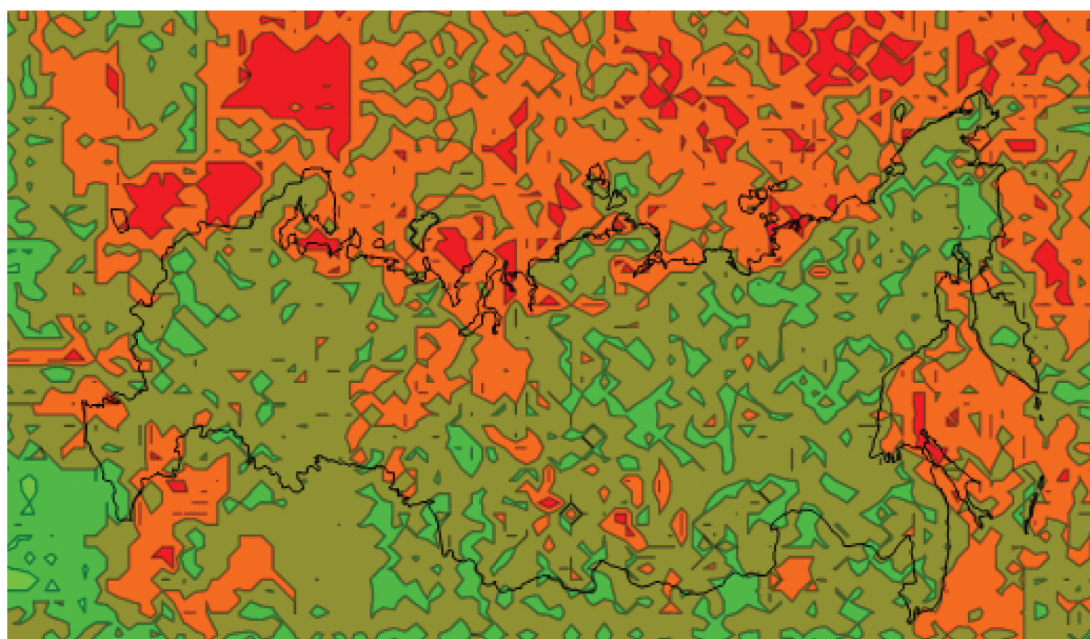


Рис. 1. Результат моделирования состояния экосистем в России, оцененный алгоритмом Bagging: 1–5 – классы состояния экосистемы: 1 – очень стабильный, 2 – стабильный, 3 – нормальный, 4 – нестабильный, 5 – очень нестабильный

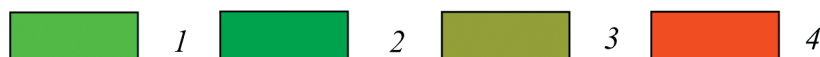
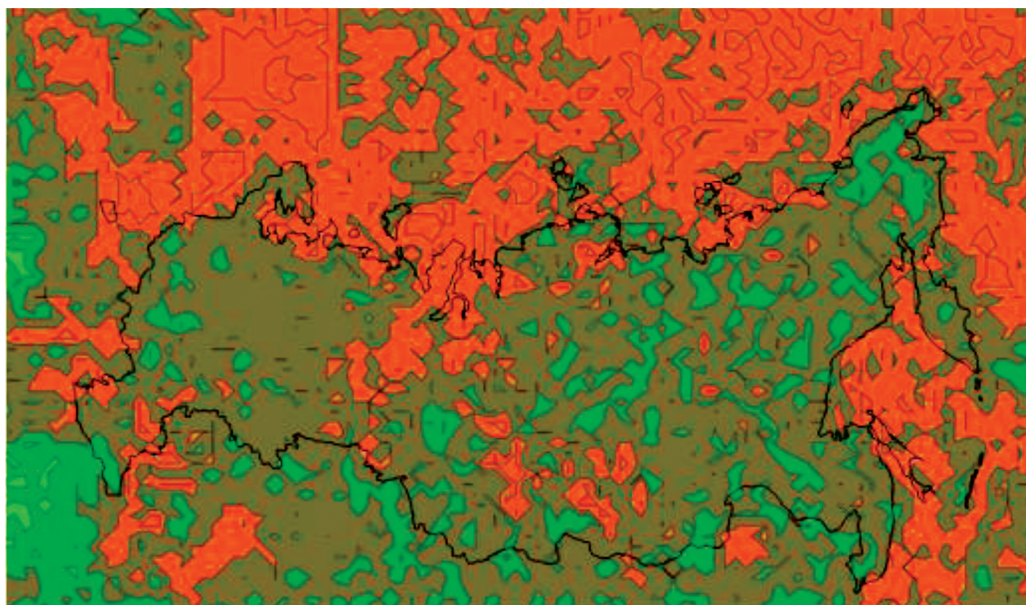


Рис. 2. Результат моделирования состояния экосистем в России, оцененный алгоритмом Random forest: 1–4 – классы состояния экосистемы: 1 – очень стабильный, 2 – стабильный, 3 – нормальный, 4 – нестабильный

К сожалению, врожденную заболеваемость можно оценить только по субъектам РФ. Почти невозможно создать подлинную карту изолиний врожденной заболеваемости, так как вся информация сортируется по субъектам. По этой карте можно заключить, что наиболее достоверная информация сконцентрирована в западной части России, потому что регионы здесь относительно небольшие.

Для создания карты заболеваемости были использованы медицинские статистические материалы России [5–7].

Была собрана информация об относительной врожденной заболеваемости в каждом субъекте РФ с 2009 по 2011 гг., подсчитано число врожденных заболеваний на 100 000 человек.

Далее для оценки точности модели необходимо провести сопоставление карты врожденных заболеваний в РФ (рис. 3) и результатов моделирования состояния экосистем РФ (рис. 1, 2)

Для определения результатов моделирования была оценена корреляция моделей и данных по заболеваниям. Для этого на территории РФ были выбраны 83 точки, так чтобы на один оцененный субъект приходилась одна точка. В каждом субъекте точка выбиралась случайно. В таких точках значения, рассчитанные по модели, сравнивались с данными по врожденным заболеваниям.

Сравнивались предсказанные классы устойчивости экосистем и определенные уровни заболеваемости. При этом получились следующие значения:

Коэффициент корреляции по алгоритму Bagging (R_b) 0,74;

Коэффициент корреляции по алгоритму Random forest (R_{rf}) 0,65.

Коэффициент корреляции был рассчитан по формуле [7]:

$$R_{X,Y} = \frac{M[XY] - M[X]M[Y]}{\sqrt{(M[X^2] - (M[X])^2)(M[Y^2] - (M[Y])^2)}}, \quad (1)$$

где M — математическое ожидание, а X и Y — случайные величины, определённые на одном вероятностном пространстве.

Следовательно, корреляция между врожденными заболеваниями и оценкой территории по уровню устойчивости экосистем моделью машинного обучения очень высокая. При сопоставлении результатов анализа соответствия карты врожденных заболеваний РФ и карт прогнозного состояния экосистем можно сделать вывод, что точность модели примерно соответствует её оцененной точности 74 %. Особенно это хорошо показано в западной части России, где регионы с небольшой площадью, но с большой плотностью населения.

По оценкам погрешности при моделировании экосистем РФ лучшим был алгоритм Bagging, при анализе экосистем городов алгоритм Random forest оказался более точным. Это показывает, что применение алгоритма Bagging с алгоритмом Random forest повышает эффективность модели и делает оценку более точной при работе с масштабами, превышающими масштабы обучающей выборки. При идентичности масштабов модель оказывается более точной и выбором является применение Random forest.

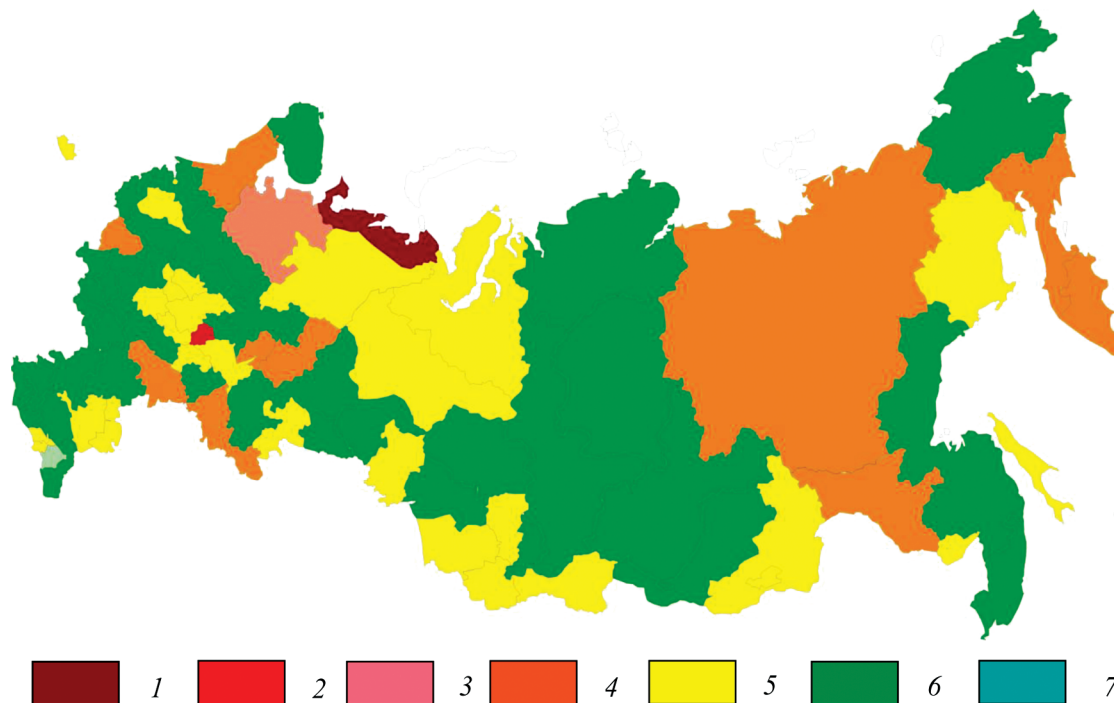


Рис. 3. Число врожденных заболеваний на 100 000 человек в РФ, по данным [5–7]: 1 — более 4000, 2 — от 3000 до 4000, 3 — от 2000 до 3000, 4 — от 1500 до 2000, 5 — от 1000 до 1500, 6 — от 500 до 1000, 7 — менее 500

Заключение

Методика определения уровня устойчивости экосистем путем анализа геофизических и геодезических данных была автоматизирована при помощи средств машинного обучения. Для этого использовались различные методики предварительной обработки данных, выбора необходимых параметров классификации и анализа группы алгоритмов машинного обучения. Проведённые исследования дают возможность построить пространственную модель прогнозирования уровня устойчивости экосистем на территории любого масштаба.

Построенная модель позволила провести классификацию территории и смоделировать экосистемы тестовых территорий. Эта модель в целом совпала с врожденной заболеваемостью на изучаемых территориях. Конечно, множество факторов, которые могут привести к заболеваниям кроме естественного состояния экосистемы, но трудно отрицать корреляцию между ними. Врожденная заболеваемость в этом исследовании выполняла функцию критерия, который определял общее состояние биосферы. Уровень заболеваемости был выбран в качестве такого критерия, потому что такого рода данные легкодоступны и легко обрабатываются. В то же время это хороший показатель состояния экосистемы.

Полученная модель может быть использована для определения качественных рисков для проживания населения и строительства жилых кварталов. Риски, связанные с нестабильным состоянием

экосистемы, вызваны условиями, носящими объективный характер, т. е. не зависят непосредственно от характеристик предполагаемого проекта и от другого антропогенного влияния. Модель позволяет оперативно сделать качественный анализ рисков и определить территории с повышенным риском для организации жилых районов и других видов деятельности. Риски проявления негативных процессов изменяются в зависимости от времени активизации нестабильных зон. Активизация происходит в соответствии с природными циклами различного порядка. От порядка цикла зависит уровень активизации, чем выше порядок цикла, тем сильнее активность нестабильных зон. Поэтому определение уровня устойчивости экосистем во многом снижает риски негативных последствий для населения при активизации нестабильных зон.

Вследствие того, что уровень устойчивости экосистем влияет не только на врожденную заболеваемость, скорее заболеваемость является просто одним из его индикаторов, но и на общее проявление природных процессов построенная модель очень полезна при выборе мест расположения крупных промышленных объектов.

Применение машинного обучения при анализе уровня устойчивости экосистем позволяет выполнять этот анализ быстро и с использованием меньшего количества человеческих ресурсов, что делает его прекрасным оперативным инструментом при проведении рекогносцировки территории.

ЛИТЕРАТУРА

- Бондаренко Я.И., Зайонц И.О. Создание многофакторных космофотографических моделей природного риска хозяйственного использования территорий и проживания населения на примере г. Киева/ Бондаренко Я.И., Зайонц И.О. // Проблемы природопользования, устойчивого развития и техногенной безопасности: сб. статей. Днепропетровск, 2001. С. 61–66.
- Демьянов Г.В., Назарова Н.Г., Никольский Н.Г., Никольский Ю.И., Таранова В.А. Карта аномалий гравитационного поля России и прибрежных территорий. Масштаб 1:10 000 000/ Демьянов Г.В., Назарова Н.Г., Никольский Н.Г., Никольский Ю.И., Таранова В.А. СПб: ФГУП «ВСЕГЕИ», 1995.
- Лагутина Н.П., Остовская О.Е. Карта рельефа России. Масштаб 1:10 000 000./ Лагутина Н.П., Остовская О.Е. Омск: Картографическая фабрика Омска, 2005.
- Литвинова Т.П., Шмирянова Н.П. Карта аномалий магнитного поля России и прибрежных территорий. Масштаб 1:10 000 000/ Литвинова Т.П., Шмирянова Н.П. СПб: ФГУП «ВСЕГЕИ», 1995.
- Заболеваемость населения по основным классам болезней, Статистическая информация. [Электронный ресурс] http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/healthcare/#, дата обращения 15.05.2014.
- Заболеваемость населения России в 2010 г. Статистические материалы. Часть 2. М., Минздравсоцразвития. 2011. С. 114–118. [Электронный ресурс]: <http://www.rosminzdrav.ru/documents/6686-statisticheskay-informatsiya>, дата обращения 12.05.2014.
- Общая заболеваемость всего населения России в 2011 г. // Статистические материалы. Часть 2. М., Минздравсоцразвития. 2012. С. 231–235. [Электронный ресурс]: <http://www.old.rosminzdrav.ru/docs/mzsr/stat/46/ZD2.DOC>. Дата обращения 15.05.2013.
- Шестопалов В.М., Зайонц И.О., Бондаренко Я.И. Структурно-геодинамическое и гидрогеологическое районирование с целью выявления структур наиболее перспективных для глубинной изоляции РАО./ Шестопалов В.М., Зайонц И.О., Бондаренко Я.И.// Чернобыль – 97: сб. статей. Киев, Наука. 1998. С. 26–31.
- Шмойлова Р.А. Общая теория статистики: Учебник. 3-е издание/ Шмойлова Р.А. М.: Финансы и Статистика, 2002. С. 472.
- Aha D., Kibler D. Instance-based learning algorithms./ Aha D., Kibler D.// Machine Learning. 1991. No. 6. P. 37–66.
- Breiman Leo. Bagging predictors./ Breiman Leo// Machine Learning. 1996. N 24. P. 123–140.
- Breiman Leo. Random Forests. Breiman Leo // Machine Learning. 2001. N 45. P. 5–32.
- Bondarenko J. The Multifactor Predictive Seis (GIS) Model of Ecological, Genetical and Population Health Risk in Connection with Dangerous Bio-Geodynamical Process in Geopatogenic Hazard Zones/Bondarenko J. Chicago: ECO-INFORMA, 2001. P. 21–34.
- Frank E., Wang Y., Inglis S., G. Holmes, I. Witten H. Using model trees for classification./ Frank E., Wang Y., Inglis S., G. Holmes, I. Witten H.// Machine Learning. 1998. N 32(1). P. 63–76.
- Kampichler Christian, Sas o Dzeroski and Wieland Ralf. Application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and Collemboles community characteristics./Kampichler Christian, Sas o Dzeroski and Wieland Ralf.// Soil Biology and Biochemistry. 2000. Vol.32 Issue 2. P. 197–209.
- Mitchell T. Machine Learning./ Mitchell T.// McGraw-Hill Science/Engineering/Math. P. 5–6.

Российский государственный
геологоразведочный университет
(117997 г. Москва, ул. Миклухо-Маклая д.23;
e-mail: vadichruk@list.ru, ekzar@hotmail.ru)

Рецензент — Т.А. Барабошкина