

УДК 502.36

## АНАЛИЗ СОСТОЯНИЯ ЭКОСИСТЕМ КРУПНЫХ ГОРОДОВ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

© 2018 г. В.В. Рукавицын

*Российский Государственный Геологоразведочный Университет им. Серго Орджоникидзе,  
ул. Миклухо-Маклая д., 23, Москва, 117997 Россия.  
E-mail: vadichruk@list.ru*

Поступила в редакцию 30.03.2017 г.

После исправления 1.07.2017 г.

В работе представлена методика определения уровня стабильности экосистем при помощи инструментов машинного обучения. Дается практическое описание использования технологии автоматизации прогнозов, позволяющих определить риски для проживания населения на определенной территории. Построена модель стабильности экосистем городов Чешской Республики. Полученная точность моделирования превысила 83%. Данная модель может быть применена для градостроения и планирования застройки новых жилых и нежилых кварталов.

**Ключевые слова:** *моделирование, стабильность экосистем, машинное обучение, автоматизация.*

### ВВЕДЕНИЕ

Экологическое районирование территории позволяет оценить степень опасности и возможности проявления различных негативных процессов, наносящих ущерб окружающей природной среде, в том числе здоровью населения. Оно позволяет планировать мероприятия, направленные на снижение негативного техногенного воздействия на окружающую среду и природные системы и является основой для разработки комплекса природоохранных сооружений, необходимых для перспективного планирования территории и, в первую очередь, развития ее инфраструктуры. Кроме этого, результаты экологического районирования могут использоваться при реконструкции существующих промышленных, транспортных, социальных и других объектов, а также при выборе мест расположения проектируемых сооружений.

В данной работе разработана аналитическая модель состояния экосистем на основе методов машинного обучения. Модель позволяет классифицировать территорию, оценивать уровень экологической опасности и распознавать проблемные ситуации и области их локализации.

Машинное обучение — научная дисциплина, которая занимается проектированием и разработкой алгоритмов, позволяющих компьютерам изменить свое поведение на базе некоей информации, например, на базе информационных сенсоров или базы

данных [10]. Одно из основных направлений исследований машинного обучения — 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 в других принимают участие человек и машина. 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.

Машинное обучение исходит из того, что существуют большие объемы данных, полученные при помощи различных методов измерения. Эти данные могут предоставить интересную и полезную информацию при условии, что они обрабатываются квалифицированным персоналом. Эксперты могут выявить такие параметры данных, их значения или их комбинации, которые могут иметь отношение к различным конкретным задачам, и при помощи обобщения эта информация может в дальнейшем использоваться для улучшения и оптимизации конечных результатов различных процессов. Тем не менее, часто возникают препятствия для легкой обработки реальных данных из-за их больших

объемов, недостающих значений, непредставительных параметров классификации, шумов, содержания как цифровых, так и буквенных значений и т.д. Эксперты могут справиться с такими трудностями, если объемы данных не слишком большие, степень размерности не слишком высока, ситуация не очень осложнена большим количеством малозначимых параметров классификации и имеется достаточно времени для решения данной проблемы. В противном случае машина (компьютер) может помочь разобраться в задаче или указать, как это сделать. Ожидаемый результат анализа данных должен быть в форме знания, например, набора правил, деревьев решений или общих классификаторов, которые могут быть использованы в будущем в качестве ценных полезных инструментов [8].

Очевидно, что такие инструменты должны функционировать “разумно”, т.е. так же как их человеческие аналоги. Современная компьютерная наука предоставляет широкий спектр алгоритмов в области искусственного интеллекта. Одна из возможностей – использование набора доступных образцов данных, подготовленных экспертами, чтобы на основе этих обучающих примеров “научить” машину, что и как делать в конкретной ситуации, когда новые неизвестные, но так или иначе подобные ситуации возникнут в будущем. Этот метод называется обучением с учителем. При этом исходные данные обрабатываются и разбиваются на две выборки: обучающую и тестовую [9].

Обучающая выборка – выборка, по которой производится оптимизация параметров модели, которые оптимизируются (обучаются) таким образом, чтобы на объектах выборки определялись нужные классы или близкие к ним.

Если модель зависимости построена по обучающей выборке, то оценка качества этой модели, сделанная по той же выборке, оказывается, как правило, слишком завышенной. Такая ситуация называется переобучением. На практике она встречается очень часто. Хорошую эмпирическую оценку качества построенной модели дает ее проверка на независимых данных, которые не использовались для обучения.

Для этого используется тестовая выборка. Если обучающая и тестовая выборки независимы, то оценка, сделанная по тестовой выборке, является достоверной.

При обучении с учителем значение по каждой точке представляет собой пару “объект, класс”. Требуется найти функциональную зависимость классов от описаний объектов и найти алгоритм, на входе дающий описание объекта, состоящее из формализованного списка параметров, и выдающий на

выходе определение класса объекта с достаточной точностью для каждого объекта. Критерий качества, как правило, определяется как средняя ошибка в определении классов, выданных алгоритмом, по всем объектам выборки. Под учителем понимается либо сама обучающая выборка, либо тот, кто указал на заданных объектах правильные классы [6].

Для анализа состояния экосистем выделялись три типа областей с различными уровнями их стабильности (сопротивляемость экосистем и способность восстановления нарушенных связей и функций). Были выделены следующие типы: стабильный, нормальный и нестабильный. Стабильные области приурочены к территориям с высоким уровнем стабильности экосистем в отношении техногенной нагрузки, нестабильные – к территориям с низким уровнем стабильности, где и происходят значительные изменения экосистем, вплоть до их полной деградации. К нормальным областям относим территории со средним уровнем стабильности экосистем и отсутствием значительного техногенного влияния на окружающую природную среду.

*Стабильность территории* характеризуется долгой геологической консолидацией окружающей среды из-за уменьшения активности тепловых потоков и изменения рельефа поверхности земли. Уменьшенный уровень тепловых режимов и устойчивая геодинамическая ситуация определяют условия более длительного развития организма, сокращение геохимической и биохимической скорости реакций, повышение производства фотосинтеза. Все эти факторы стимулируют повышение сопротивляемости окружающей среды техногенному уровню напряжения. В стабильных геодинамических зонах окружающая среда может восстановить свои нарушенные связи и функции. Эта особенность необходима при проектировании новых рекреационных зон, размещении медицинских комплексов и в производстве биопродуктов. Также стабильное состояние экосистем вызывает низкий уровень заболеваемости на данной территории. Это означает, что уровень заболеваемости в регионе – хороший тестер точности модели.

*Нестабильные области* характеризуются быстрым снижением уровня стабильности окружающей среды из-за тепловых, радиоактивных и других воздействий. Повышенный тепловой режим природной системы, геохимическая и биохимическая причины повышения активности реакции, развитие дестабилизации и процессы разрушения окружающей среды. В этих областях существует высокая вероятность появления экологически опасных процессов: засоление почвы, загрязнение

вод, современные тектонические процессы и др. Активация современных тектонических процессов может вызвать появление и развитие оползней, процессов подтопления, разрушение инженерных сооружений и коммуникаций. В этих областях происходит объединенное естественное и техногенное воздействие на ухудшение здоровья населения.

Таким образом, экологически опасные генетические изменения экосистемы и ухудшение здоровья населения вызваны комплексом природных и техногенных факторов. Геодинамические условия формирования природной среды являются очень важным, но не единственным фактором, который влияет на экологическую ситуацию. Однако этот фактор оказывает значительное влияние на все экосистемы и население одновременно. При этом следует отметить, что степень воздействия геодинамических условий не одинакова в разных зонах и зависит от типов и активности полей природных систем. Если есть активные стабильные или активные нестабильные зоны, степень влияния и коэффициент значения данного фактора будет возрастать. В нормальных или неактивных зонах геодинамические поля оказывают незначительное воздействие на экосистему, но в активных зонах их влияние значительно возрастает и не может быть проигнорировано [2].

Тезис о том, что “все в природе взаимосвязано”, требует не только объединения различных областей научных знаний, но и создания некоей базовой, фундаментальной основы, которая позволила бы объединить эти системы и объяснить закономерности. В основе анализа лежит суммация магнитного поля, гравитационного поля и рельефа в единое векторное поле, которое отображает базовое направление развития процессов в экосистеме на заданной территории.

Развитие квантовой механики и необходимость согласования ее с теорией относительности, привело к поиску существования некоего универсального поля, которое взаимодействует со всеми другими полями. Такое поле получило название поле Хигса.

К концу XX в., благодаря работам Л. Пастернака, В. Лимба, Д. Блохина, Г. Шипова и др. в области математики, теоретической и экспериментальной физики, удалось завершить развитие единой теории поля, начало которой было заложено А. Эйнштейном [1]. Содержательной базой новой парадигмы стали принцип всеобщей относительности Шипова, геометрия Римана-Картана-Шипова и *физический вакуум* – реально существующая материальная среда, передающая взаимодействие и рождающая элементарные частицы.

Законы самоорганизации и эволюции энергоинформационного поля и физического вакуума обуславливают единые законы развития живой и неживой природы. Это прекрасно согласуется и объясняет результаты фундаментальных исследований И. Пригожина, основателя термодинамики открытых систем. В его трудах описаны и доказаны единые законы развития систем не только живой и неживой природы, но и социально-экономических, урбанистических и др.

О едином поле Земли – ноосфере писал и В.И. Вернадский. Эта концепция была изложена как философское обобщение учения о биосфере, ее эволюции, о мощном влиянии человека на окружающую среду и преобразовании современной биосферы в ноосферу – “разумную” сферу, которая самоорганизуется (отражение законов развития единого энергетического поля). Сегодня концепция В.И. Вернадского развивается как теория единства (взаимообусловленности) существования и эволюции живой и неживой природы, как составных частей “единого организма Земли (вселенной)”.

Но при этом становится очевидным, что для решения этой задачи сегодня не хватает главного, той базовой основы, которая и предопределяет изменчивость и особенности развития всех материальных систем – описание и картирование динамики изменчивости состояния физического вакуума.

Установлено, что вследствие флуктуаций физического вакуума происходят активизация всей иерархической цепочки энергий (от ядерной до гравитационной) и изменение состояния физических полей, значит в каждом из них должна быть заложена информация о динамике изменчивости физического вакуума. Если это так, то необходимо понять, каким шифром в физических полях закодирована данная информация.

Для того чтобы ответить на этот вопрос, попробуем смоделировать цепочку последствий таких изменений, происходящих из-за нарушения равновесия системы. Оставим без комментариев причину возмущения физического вакуума. Но очевидно, что следствием такого возмущения будет нарушение динамического равновесия, деформация и разрыв внутренних взаимосвязей по всей иерархической цепочке организации материи. Подобный процесс приводит к формированию структур избыточной концентрации энергии (избыточным напряжениям), последующей релаксации напряжений и деформации пространства. Данный процесс следует понимать, как нарушение (деформацию) связей элементарных частиц, сопровождающееся высвобождением энергии, формированием новых молекулярных структур и химических соединений, что

изменяет структурные и вещественные параметры материального мира от микро до галактического уровня. Подобные деформации реализуются в виде направленных линейных или вихревых волновых динамических процессов концентрации энергии (процессы сжатия пространства) и ее рассеивания (растяжения). Поэтому любые структурно-вещественные преобразования и пространственно-временное развитие материальных объектов и природных систем (в широком понимании этого слова) происходят в пределах структурно ограниченных энергетических (квантово-волновых, деформационных) потоков.

Следовательно, любое физическое поле есть отражение пространственно-временной эволюции квантово-волнового (энергетического) поля, динамики его напряженно-деформированного состояния и уже как следствие структурно-химических и структурно-вещественных особенностей материальных объектов.

Очевидно, что любой физический объект может быть описан целым набором физических характеристик. Например, поверхность Земли может быть охарактеризована цифровыми многоканальными космическими снимками, картами гравитационного, магнитного, электрического полей, поверхности рельефа и пр. Каждая из этих характеристик описывает объект на своем иерархическом уровне и в своем резонансно-частотном диапазоне. Работа максимально полного комплекса характеристик физических полей объекта позволяет получать отражение динамики изменчивости физического вакуума в различных резонансно-частотных диапазонах, что значительно приближает нас к расшифровке реальной картины процесса.

Очевидно, что любой антропогенный комплекс формируется в пределах различных, реально функционирующих геоприродных полей, которые, прежде всего, являются отражением динамики и интенсивности флуктуаций физического вакуума. Именно это определяет специфику условий формирования и характерные особенности природных экосистем (набор факторов с определенной степенью их взаимосвязей).

Каждая из таких систем на фиксируемых и изучаемых уровнях их проявления характеризуется различной интенсивностью эндогенных и экзогенных процессов, спецификой тектонического строения и тектонических напряжений, различными уровнями и особенностями теплового, радиационного, гравитационного, магнитного и электрического полей и т.д. Комплексное сочетание этих и многих иных факторов отображает условия формирования природных систем и определяет

специфику таких характеристик, как типы и химизм грунтов, гидродинамические и гидрохимические характеристики, состав и биохимические особенности растительности и биоты и пр. Кроме того, геоприродные системы определяют и контролируют пространственно-временное развитие техногенных процессов, в том числе направление и режимы миграции химических элементов в различных средах, места их концентраций, накопления и т.д.

Важным фактором является то, что природные системы имеют различный уровень экологической устойчивости, т.е. способности к ассимиляции техногенной нагрузки и восстановлению нарушенных связей и функций. Природные системы не только определяют и контролируют зоны максимального техногенного риска (пути миграции, особенности развития и места формирования устойчивых концентраций химических элементов-загрязнителей), но и сами создают условия повышенного риска для проживания людей. Для того чтобы понять, как и почему природные условия влияют на здоровье людей, и насколько определяющими являются эти факторы, необходимо рассмотреть условия и особенности функционирования природных систем.

В пределах активных геологических структур происходят продолжительные и нарастающие всплески (высокоамплитудные изменения) магнитных и электрических полей, проявления микросейсмичности, активизируется интенсивность эманаций природных газов (радоновой группы), изменяются режимы термобарических, тепловых, радиационных, геохимических и других потоков, интенсивность (резонансно-частотные характеристики) энергетических полей. Подобные изменения природных условий в значительной мере контролируют интенсивность химических и биологических процессов, что предопределяет различную скорость химических и биохимических реакций, время развития и рефлекторного поведения живых организмов, различную продолжительность процессов фотосинтеза и т.д.

Биологические системы, которые развиваются в нестабильных условиях, в пределах активных геодинамических структур, вынуждены приспосабливаться к таким условиям. Это приводит к определенным генетическим мутациям растительности, микроорганизмов и животных. Например, изменяется биологическая активность живых клеток и тканей, чаще наблюдаются процессы биологической мутации клеток, грибковых микроорганизмов и растительности. И биологическая система человека не является исключением.

Следует отметить, что по результатам исследований [2], у людей, проживающих в пределах активных

геологических структур, прежде всего, происходит значительное ослабление иммунной системы. В тех же исследованиях было установлено, что беря за основу данные магнитного и гравитационного полей и рельефа территории, преобразуя их определенным образом и добавляя ряд характеристик, получаемых путем математических преобразований исходных данных, можно получить модель, описывающую стабильность экосистемы, которая в свою очередь взаимосвязана с заболеваемостью населения на изучаемой территории.

Для повторения логики моделирования без применения сложного экспертного анализа данных, который использовался в вышеуказанных исследованиях, были применены методы машинного обучения. Машинное обучение в данном случае позволяет проводить классификацию территории, определяя стабильные и нестабильные геодинамические зоны, которые в свою очередь определяют заболеваемость населения на изучаемой территории.

#### МЕТОДИКА АНАЛИЗА

Определение состояния экосистем сводится к экологическому прогнозированию, в данном случае произведенному при помощи методов машинного обучения. При этом экологический прогноз рассматривается как предсказание изменений состояния природной или природно-техногенной системы в целом. При анализе стабильности экосистем были рассмотрены территории городов: Киев [7], Прага и Брно. Каждый из этих объектов исследования – сложная природно-техническая система, состоящая из множества уровней, сложно связанных между собой. На развитие экосистемы подобного рода объектов влияют различные факторы [5]. Среди них можно выделить группы:

- геодинамические условия,
- природные процессы и явления,
- техногенные факторы,
- социально-демографические факторы.

Автором было выдвинуто предположение об иерархической связи между этими факторами в процессе формирования экосистемы. Здесь факторы перечислены в иерархическом порядке по глубине влияния на экосистему. Таким образом, первая группа факторов (геодинамические условия) начинает процесс формирования экосистемы, закладывая основные тренды ее развития. На основе этих трендов формируются природные процессы и явления. Затем включаются техногенные факторы, способные притормозить или даже остановить развитие экосистемы, но все равно ее развитие будет

продолжаться в пределах изначально заложенных параметров. Впоследствии на техногенные факторы накладываются социально-демографические факторы, способные ввести некоторые коррективы в их влияние на природные процессы. Каждая из этих групп факторов имеет свою иерархическую структуру и включает множества взаимосвязанных параметров. Учитывать все возможные источники влияния на развитие экосистемы при моделировании практически невозможно вследствие их большого количества, разнообразности воздействия и недостаточности изученности степени влияния каждого из них. Вследствие вышесказанного задача состоит в определении наиболее значимых источников воздействия на экосистемы или индикаторов этого воздействия для прогнозирования развития экосистем. Для решения этой задачи первым делом выбиралась та группа факторов, которая определяет изначально развитие экосистемы. Смоделировав геодинамические условия развития экосистемы, можно по иерархической цепочке определить дальнейший тренд развития экосистемы по другим группам факторов.

Ранее подобная работа выполнялась исключительно отдельными экспертами, что не позволяло применять подобное моделирование широко. Автором была предложена методика автоматизации подобного анализа при помощи алгоритмов машинного обучения.

В методике можно выделить две составные части: *обучение и тестирование*. Обучение необходимо для создания базовой модели на обучающей выборке. Тестирование производится для проверки корректности модели, ее совершенствования и оценки возможности решения реальных задач. После опробования модели она проверялась при решении реальной задачи классификации. Таким образом, после прохождения всех уровней была получена работающая система автоматизации принятия решения для определения уровня стабильности экосистемы [3]. Эта система была использована для определения состояния экосистем городов Прага и Брно, а также оценки риска для проживания населения в разных районах городов и проверки точности этого прогноза.

В качестве обучающей выборки использовались данные магнитного поля, гравитационного поля и рельефа поверхности земли на территории г. Киева. Данные содержали координаты точек, в которых были произведены замеры параметров геофизических полей и сами значения параметров (193830 значений). В качестве первых исходных параметров использовались только вышперечисленные сведения. Все остальные параметры для модели были получены на основе преобразований исходных данных [3].

В каждой точке экспертами был определен уровень состояния экосистемы. Такая классификация ранее была проведена и подтверждена реальными замерами. При классификации в качестве классов системы были использованы 5 уровней стабильности экосистемы территории: *очень нестабильный, нестабильный, нормальный, стабильный, очень стабильный*.

Далее были определены параметры классификации для обработки данных при помощи машинного обучения. Прежде всего, список параметров классификации содержал параметры магнитного поля (значение магнитной индукции в каждой исследуемой точке, Тл), гравитационного поля (значение ускорения свободного падения в каждой исследуемой точке, Гал) и рельефа местности (абсолютная отметка местности в каждой точке, м), так как они лежат в основе модели. Остальные параметры были результатами преобразований и компонентного анализа первых трех параметров.

После подготовки данные обрабатывались программным продуктом WEKA, где был выбран алгоритм моделирования и его параметры, максимально подходящие для решения поставленной задачи. Лучший результат показал алгоритм Random Forest, который показал более 98% точности. Это означает, что компьютерная модель может скопировать логику человека-эксперта в такой классификации с точностью в 98%. Таким образом, была получена работающая модель, что дало возможность испытать ее при реальном анализе состояния экосистемы городов.

Компьютер сам определил значения трех исходных параметров, а также значения дополнительных расчетных параметров, соответствующие каждому определяемому уровню стабильности экосистем. Изначально такое деление на уровни стабильности носит условный характер и определяет стабилизацию природных процессов на территории или их дестабилизацию. Для обучающей выборки уровни стабильности в каждой точке были определены экспертно путем анализа, наложения и сопоставления множества различных карт, приводя их к некому "общему знаменателю". Таким образом, уровень стабильности экосистем представляет собой не количественную величину, а качественную, которую при этом можно увязать с другими, уже количественными величинами. Из-за отсутствия единиц измерения у данного параметра он был разделен на 5 условных уровней. Для характеристики каждого уровня и его соотношения с количественными значениями он сравнивался со значением количества врожденных заболеваний на исследуемых территориях. Корреляция между значениями уровня

стабильности и врожденной заболеваемости показывает объективность такой классификации.

Однако показать количественную взаимосвязь между исходными тремя параметрами и установленным уровнем стабильности экосистемы было бы некорректно, поскольку необходимо учитывать весь комплекс критериев, включая математические преобразования исходных данных при помощи вычисления главных компонент и дивергенции. И так как первоначально классификация территории имела качественный характер и основывалась на экспертной оценке, задача компьютера была в попытке воспроизвести логику эксперта и повторить такой качественный анализ.

### ОПРЕДЕЛЕНИЕ ЭКОЛОГИЧЕСКОГО СОСТОЯНИЯ ГОРODOV

Модель была опробована на данных по городам Прага и Брно. Это дало возможность оценить репрезентативность метода при работе с данными сходного масштаба, более близкого к данным обучающей выборки. Перед классификацией данные из указанных регионов были собраны и подготовлены. Исходные данные тестовой выборки, как и в случае с обучающей выборкой, содержали данные магнитного поля, гравитационного поля и о рельефе этих территорий [3], но без проведенной классификации.

Данные по г. Брно содержали 7209 значений, по г. Прага – 17061 значение. Перед расчетом параметров классификации было необходимо создать общую систему координат, так как все части исходных данных имеют разные точки замера на исследуемой территории, т.е. на каждой территории магнитное поле, гравитационное поле и рельеф измерялись в разных точках региона. Поэтому изначально были рассчитаны параметры каждого из полей, и путем интерполяции выбраны общие точки, значения которых использовались для дальнейшего расчета.

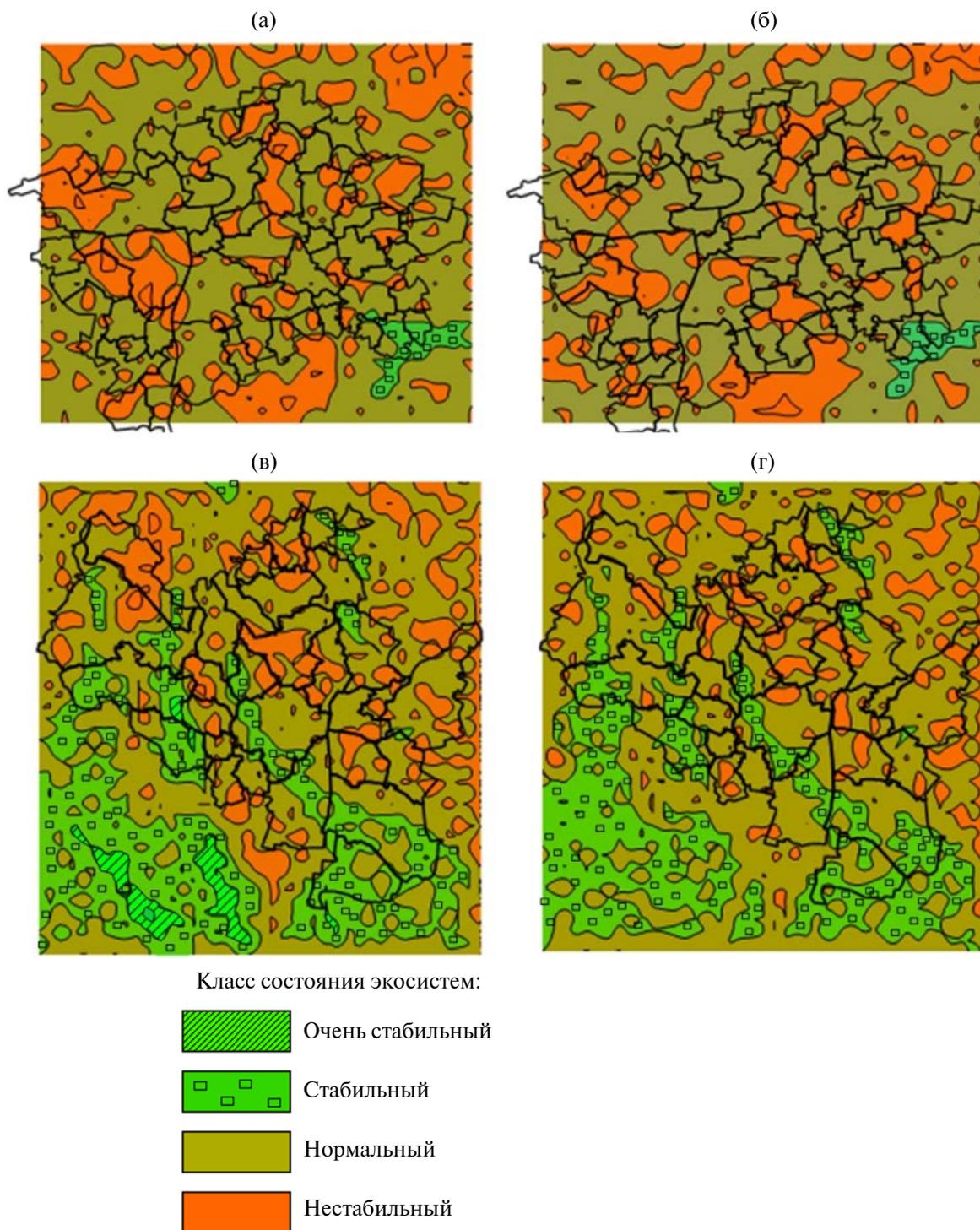
После подготовки данных были рассчитаны описанные ранее параметры классификации и произведено моделирование. Обучающая выборка по данным Киева была использована как основа модели (обучающая выборка), а данные по Праге и Брно являлись тестовыми выборками.

В результате обработки тестовой выборки при помощи машинного обучения были получены данные в табличной форме. В таблицы по итогам вычислений добавлены условные координаты точек и значения предполагаемого уровня стабильности экосистемы в этих точках. На основе этих таблиц в программном продукте Surfer строились карты изолиний, описывающие территорию по уровню стабильности

экосистем. В качестве критерия достоверности модели была выбрана врожденная заболеваемость на территории.

Результаты показаны на рис. 1. Как видно на рис. 1, в пределах территории г. Брно больше стабильных районов, чем в Праге. Некоторые районы г. Брно

содержат даже очень стабильные территории. Подтверждение этого моделирования можно найти в медицинской статистике, которая показывает, что в Праге насчитывается около 450 врожденных заболеваний на каждые 10000 человек, а в Брно их менее 260 [4].



**Рис. 1.** Результаты анализа модели состояния экосистемы. а–б – Прага, в–г – Брно; а–в – алгоритм Random Forest; б–г – алгоритм Bagging.

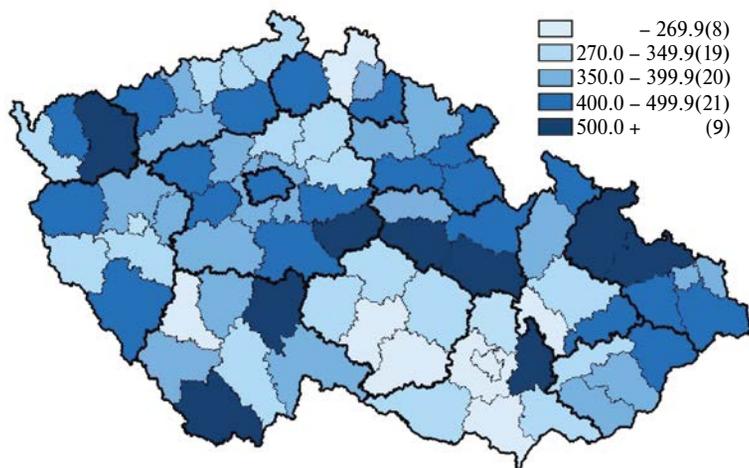


Рис. 2. Распределение врожденных заболеваний в Чешской Республике за 2005–2009 гг.

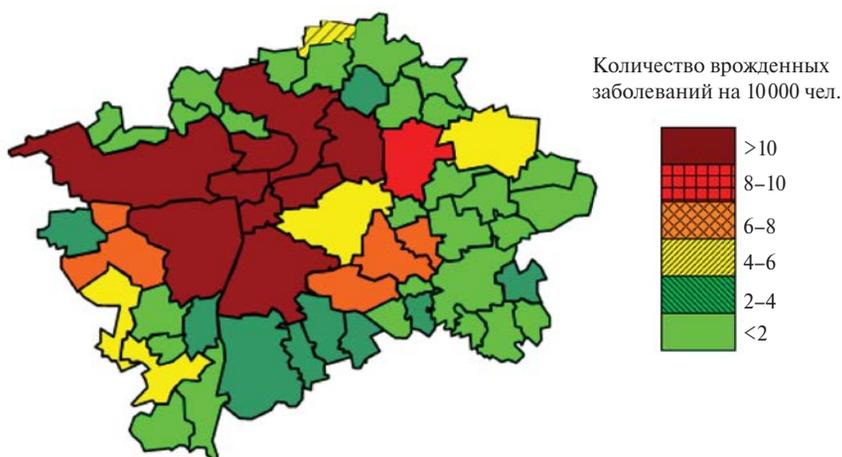


Рис. 3. Количество врожденных заболеваний в Праге в расчете на 10000 человек (2005–2010 гг.).

Были проанализированы данные всех врожденных заболеваний в Чешской Республике с 2005 по 2009 г. Согласно этим материалам, при изменении абсолютного числа заболеваний их относительное количество остается практически неизменным. Это позволяет сделать вывод, что есть некоторые долгосрочные обстоятельства, которые создали подобную ситуацию в каждом районе страны. Причем на территории г. Брно всегда было меньше врожденных заболеваний, чем в Праге (рис. 2).

Детально были рассмотрены данные по Праге. Данные о врожденных заболеваниях за 2005–2010 гг. были усреднены и сведены в единую таблицу, на основе чего была построена карта врожденных заболеваний в Праге (рис. 3). Ситуация, отображенная на рис. 3, во многом соотносится с рис. 1 (а, б). Можно отметить только один большой стабильный регион на юго-востоке и несколько относительно больших нестабильных регионов на западе, севере

и в центре. На карте видно, что наибольшее количество врожденных заболеваний отмечается в нестабильных регионах. В целом во всей юго-восточной части наблюдается относительно хорошая ситуация с наименьшим количеством заболеваний. Это означает, что модель соответствует реальной ситуации, и ее можно использовать для прогнозирования.

Также точность модели была оценена при помощи расчета коэффициента корреляции. Для каждого района г. Прага было выбрано наиболее распространенное в нем значение уровня стабильности экосистем, после чего рассчитана корреляция между определенным уровнем стабильности экосистем и количеством врожденных заболеваний.

Оценить в процентах уровень точности модели довольно сложно, поскольку данные о заболеваемости предоставляются в среднем по району, в то время как данные, получаемые в результате моделирования, не привязаны к административным

границам. Однако расчет корреляции был произведен. Сравнивались предсказанные уровни стабильности экосистем и определенные уровни заболеваемости. При этом коэффициент корреляции был равен 0.83.

Если учитывать информацию, полученную из сопоставления карт моделирования состояния экосистем и врожденной заболеваемости в Праге и Брно, с уверенностью можно сказать, что модель дает достоверные количественные результаты и может определить уровень стабильности экосистемы. Из анализа следует, что алгоритм Random Forest оказался более приемлемым, а построенная на его основе модель более точная.

### ЗАКЛЮЧЕНИЕ

Проведенное исследование дает возможность построить пространственную модель прогнозирования уровня стабильности экосистем на территории любого масштаба.

Разработана методика применения машинного обучения для определения уровня стабильности экосистем путем анализа геофизической и геодезической информации.

Основная идея работы заключалась в преобразовании исходных данных для определения базовых параметров, на основе которых возможно проводить анализ данных при помощи машинного обучения, и в поиске наилучшего алгоритма классификации данных и лучших его параметров для получения наиболее точного результата оценки.

В процессе исследований:

- определены основные параметры классификации данных для решения поставленной задачи (определения уровня стабильности экосистемы);
- созданы обучающая и тестовая выборки из исходных данных;
- выполнен экспериментальный поиск наилучших алгоритмов и их параметров на обучающей выборке;
- осуществлено применение полученной модели для анализа тестовой выборки;
- выполнена интерпретация результатов и дано обоснование применения выбранного алгоритма для решения указанной задачи;
- осуществлена апробация модели на идентичных по формату данных, взятых с иной географией и масштабом, для определения уровня ее универсальности;
- проведено сравнение результатов компьютерного моделирования со статистикой заболеваний на

исследуемой территории для обоснования точности результатов обработки тестовой выборки.

Построена модель, позволяющая определять уровень стабильности экосистем на различных территориях на основе одной обучающей выборки.

Метод объединяет способы предварительной обработки данных для решения этой проблемы и использования конкретных алгоритмов моделирования процесса. Однако точность модели при ее тестировании была ниже, чем при обучении. Это произошло из-за несовершенства обучающей выборки. Полученная точность моделирования, составляющая примерно 83% при тестировании на различных тестовых объектах, показывает, что эта модель работает и все параметры классификации и алгоритм выбраны правильно. При условии создания большой сбалансированной выборки, которая содержит информацию о многих различных областях, точность модели приблизится к 98%, как это было при обучении.

Найден наилучший алгоритм обработки данных и наилучшие его параметры для компьютерного определения уровня стабильности экосистем, что позволяет автоматизировать сложный процесс экспертной классификации территории. Этим алгоритмом является Random Forest.

Создан механизм, дублирующий логику эксперта при решении проблемы классификации территории по уровню стабильности экосистем, который позволяет решать подобные задачи более оперативно и получать при этом более достоверную информацию.

Показана универсальность модели, построенной на основе машинного обучения. Это позволяет при расширении исходной выборки решать задачи классификации территории по уровню стабильности экосистем в любых масштабах в любой точке Земли без потери точности и с использованием изначально найденной логики классификации.

Использование данного метода может найти широкое применение для оценки экологической безопасности и экологического риска при освоении территорий, планировании новых городов и выборе места расположения жилых зданий.

### СПИСОК ЛИТЕРАТУРЫ

1. Акимов А.Е., Тарасенко В.Я., Шипов Г.И. Торсионные поля как космофизический фактор // Биофизика. Т. 40. Вып. 4. 1995. С. 12–19.
2. Бондаренко Я.И., Зайонц И.О. Создание многофакторных космофотографических моделей природного риска хозяйственного использования территорий и проживания населения на примере г. Киева. Сб.

- «Проблемы природопользования, устойчивого развития и техногенной безопасности». Днепропетровск: Ин-т природопользования и экологии НАН Украины. 2001. С. 61–66.
3. Рукавицын В.В., Эжзарьян В.Н. Методика оценки уровня устойчивости экосистем при помощи машинного обучения // Известия ВУЗов. Сер. «Геология и разведка». 2015. № 1. С. 38–42.
  4. Статистика врожденных заболеваний в Чешской Республике. <http://www.uzis.cz/en/catalogue/congenital-anomalies-births>. Дата обращения 10.02.2016.
  5. Шестопалов В.М., Зайонц И.О., Бондаренко Я.И. Структурно-геодинамическое и гидрогеологическое районирование с целью выявления структур наиболее перспективных для глубокой изоляции РАО. В сб. Научно-практической конференции «Наука. Чернобыль-97» (г. Киев, 11–12 февраля 1998 г). С. 26–31.
  6. Aha D., Kibler D. Instance-based learning algorithms // Machine Learning. 1991. N6. P. 37–66.
  7. Bondarenko J. The Multifactor Predictive Seis (GIS) Model of Ecological, Genetical and Population Health Risk in Connection with Dangerous Bio-Geodynamical Process in Geopathogenic Hazard Zones. Chicago: ECO-INFORMA, 2001. P. 21–34.
  8. Frank E., Wang Y., Inglis S., G. Holmes, I.H. Witten. Using model trees for classification // Machine Learning. 1998. No. 32(1). P. 63–76.
  9. Kampichler C., Dzeroski S., Wieland R. Application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and Collembolan community characteristics // Soil Biology and Biochemistry. 2000. V. 32. Iss. 2. P. 197–209.
  10. Mitchell T. Machine Learning // McGraw-Hill Science/Engineering/Math. 1997. P. 5–6.
  2. Bondarenko, Ya.I., Zayonts, I.O. *Sozdanie mnogofaktornykh kosmofotograficheskikh modelei prirodnoho riska khozyaistvennogo ispol'zovaniya territorii i prozhivaniya naseleniya na primere g. Kieva* [Creation of multifactor models of the area's economic use and the population living risk by the example of Kiev]. Problems of the environmental management, stable development and technological security, Dnepropetrovsk, Institute prirodopol'zovaniya i ekologii NAN Ukrainy, 2001, pp. 61–66. (in Russian)
  3. Rukavitsyn, V.V., Ekzariyan, V.N. *Metodika otsenki urovnya ustoichivosti ekosistem pri pomoshchi mashinnogo obucheniya* [Procedure of ecosystem stability level assessment by machine learning tools]. *Izvestiya vuzov. Ser. "Geologiya i razvedka"*, no. 1, Moscow, 2015, pp. 38–42 (in Russian)
  4. Congenital anomalies in births in Czech Republic. Available at: <http://www.uzis.cz/en/catalogue/congenital-anomalies-births> (Accessed 10.02.2016).
  5. Shestopalov, V.M., Zayonts, I.O., Bondarenko, Ya.I., *Strukturno-geodinamicheskoe i gidrogeologicheskoe raionirovanie s seliyu vyavleniya struktur naibolee perspektivnykh dlya glubinnoi izolyatsii RAO* [Structural geodynamical and hydrogeological zoning for revealing geological structural perspectives for deep isolation of nuclear wastes]. Proc. scientific and practical conference "Nauka. Chernobyl-97". Kiev, 1998, pp. 26–31.
  6. Aha D., Kibler D. Instance-based learning algorithms. *Machine Learning*, 1991, no. 6, pp. 37–66.
  7. Bondarenko J. The multifactor predictive Seis (GIS) model of ecological, genetical and population health risk in connection with dangerous bio-geodynamical process in geopathogenic hazard zones. Chicago, ECO-INFORMA, 2001, pp. 21–34.
  8. Frank, E., Wang, Y., Inglis, S., Holmes, G., Witten, I.H. Using model trees for classification. *Machine Learning*, 1998, no. 32(1), pp. 63–76.
  9. Kampichler, C., Dzeroski S., Wieland R. Application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and Collembolan community characteristics. *Soil Biology and Biochemistry*, 2000, vol. 32, issue 2, pp. 197–209.
  10. Mitchell, T. Machine Learning. McGraw-Hill Science/Engineering/Math., 1997, pp. 5–6.

## REFERENCES

1. Akimov, A.E., Tarasenko, V.Y., Shipov, G.I. *Torsionnye polia kak kosmofizicheskii faktor* [Torsion fields like a cosmophysical factor]. *Biofizika*, vol. 40, issue 4, Moscow, Nauka, 1995, pp. 12–19. (in Russian)

## THE ANALYSIS OF URBAN ECOSYSTEM STATE BY MACHINE LEARNING TOOLS

V.V. Rukavitsyn

*Ordzhonikidze Russian State Geological Prospecting University, ul. Miklukho-Maklaya 23, Moscow, 117997 Russia. E-mail: vadichruk@list.ru*

This work deals with the automation of ecosystem stability assessment. Until now, carrying out such assessment has been available only to high-class experts. The article presents practical description of using a technology of automation of forecasts making it possible to determine risks for the population residence in a particular area. The article also presents a constructed model of ecosystems stability in cities of Czech Republic. Model accuracy was more than 83%. This model can be applied at the cities construction and design of new residential and industrial quarters.

**Key words:** *modelling, ecosystem stability, machine learning, automation.*