

## Моделирование показателей качества воды рек Крымского полуострова с применением методов машинного обучения

**Малыгин Е.В.<sup>1</sup>, Лычагин М.Ю.<sup>1</sup>, Малыгин И.В.<sup>2</sup>**

<sup>1</sup>Московский государственный университет им. М.В. Ломоносова, Москва

<sup>2</sup>Институт физики Земли им. О.Ю. Шмидта Российской Академии Наук, Москва  
[malygin.ev@gmail.com](mailto:malygin.ev@gmail.com)

**Аннотация:** В работе предложена методика оценки влияния природных и техногенных факторов на формирование качества речных вод, описан подход к моделированию показателей качества воды с применением методов машинного обучения на примере рек Крымского полуострова. Показано, что на формирование состава речных вод как в межень, так и в паводковый период наибольшее влияние оказывают природные факторы (близость к морю, горные породы и почвы), из группы антропогенных факторов главное значение имеют виноградники и урбанизированные территории. Применение методов машинного обучения позволило смоделировать внутригодовой ход концентраций некоторых тяжелых металлов (Cu, Co, Ni, Mo, Ba) и рассчитать среднемесячный сток изучаемых элементов.

**Ключевые слова:** моделирование качества воды, машинное обучение, тяжелые металлы, Крым

### Введение

Формирование качества речных вод напрямую зависит от процессов, происходящих на водосборной площади, которые определяются физико-географическими, геологическими, биологическими и техногенными характеристиками водосборов (Никаноров, 2001), при этом вклад каждого из факторов обычно не определяется.

В связи с интенсивным использованием рек и их водосборных бассейнов в различных отраслях хозяйственной деятельности встает актуальная задача оценки антропогенного воздействия на их экологическое состояние. Цель настоящего исследования – оценка влияния природных и техногенных факторов на формирование качества вод рек Крымского полуострова и моделирование потоков загрязняющих веществ.

### Материалы и методы

В работе использованы результаты гидролого-геохимических исследований на реках Салгир, Альма, Кача, Бельбек, проведенных географическим факультетом МГУ имени М.В. Ломоносова в рамках Комплексной Крымской экспедиции Русского географического общества и проекта РФФИ № 16-05-01037 в разные сезоны 2015 – 2016 гг.

Исследования проводились в различные фазы водного режима, в работе использованы данные для трех сезонов: начало зимнего паводкового периода (февраль 2015 г.), летний паводок (июнь 2015 г.), летняя межень (август 2016 г.). Исследования включали измерение расходов и физико-химических параметров воды, отбор проб воды (всего 93). Ионный состав определялся методом ионной хроматографии. Минерализация воды вычислена по сумме главных ионов. Валовое содержание тяжелых металлов (ТМ) определялось масс-спектральным (ICP-MS) и атомно-эмиссионным (ICP-AES) методами с индуктивно-связанной плазмой.

Для решения задачи моделирования качества воды требуется сформировать базу данных по выбранной проблеме. Для этого необходимо определить основные факторы, влияющие на формирование химического состава речных вод на водосборе. Среди них можно выделить геологический, почвенно-растительный, антропогенный. Характеристики

водосборов определены при помощи базы геоданных, составленной на основе картографических и экспедиционных материалов, данных дистанционного зондирования. База данных включает информацию о горных породах, почвах, растительности, численности и плотности населения, урбанизированности территории и сельскохозяйственном использовании земель.

### Методы машинного обучения

Машинным обучением называется систематическое обучение алгоритмов, в результате которого их “знания” или качество работы возрастают по мере накопления опыта (Флах, 2015). Для построения таких методов используются средства математической статистики и теории вероятностей, численные методы, методы оптимизации.

Для оценки влияния природных и техногенных факторов на формирование химического состава рек применен алгоритм машинного обучения, реализованный с применением комбинаторно-логического подхода теории распознавания образов (Малыгин, 2014). Это один из группы методов искусственного интеллекта, у которого имеется полное теоретическое обоснование (Кудрявцев и Андреев, 2006) и практическое применение в смежных дисциплинах (Малыгин, 2015).

Исходными данными для используемого алгоритма распознавания образов является таблица вида “объект - признак”. В качестве объектов использованы створы экспедиционных измерений, признаки – природные и техногенные характеристики водосборов. Целевая функция для алгоритма распознавания задается в виде классов. На основе данных о минерализации и ионного состава речные воды были классифицированы на 2 типа: фоновые и преобразованные воды.

Алгоритм обучается на исходных данных. В результате обучения вырабатываются правила, по которым новые поступающие на вход системы данные относятся к одному из выделенных классов. Автоматизированный подбор параметров алгоритма проводится с применением процедуры полной кросс-валидации: на каждом этапе работы из обучающей выборки поочередно удаляется один элемент, формируется новая обучающая выборка, удаленный элемент подается для распознавания. Процедура повторяется для всех доступных наблюдений, в конце всех итераций вычисляется оценка достоверности полученного результата как доля верно классифицированных элементов. Найденные параметры представляют собой набор векторов, по которым производится классификация, в рамках используемой системы они называются “экспертами”.

Данный алгоритм позволяет ранжировать признаки по их важности для результатов исходной классификации. Эта функциональность основана на понятии информационного веса признаков (Алешин, 1996). В процессе работы алгоритма вычисляется вектор информационных весов,  $i$ -ая компонента которого называется информационным весом  $i$ -ого признака. В процессе обучения выявляются признаки, которые различают все элементы, относящиеся к разным классам. Информационный вес отражает количественную оценку частоты использования признака в процессе распознавания: чем больше частота, тем признак более информативен в рамках принятия решения о классификации.

Оценку важности признаков недостаточно проводить по одному “эксперту”, для достижения репрезентативного результата необходимо использовать группу “экспертов”. Тогда важность  $i$ -того признака  $FI_i$  будет определяться как

$$FI_i = \frac{|E_{iv}|}{|E|} \cdot 100\%, FI_i = \frac{|E_{iv}|}{|E|} \cdot 100\%,$$

где  $|E_{iv}|$  – количество “экспертов” с ненулевой оценкой информационного веса  $i$ -ого признака,  $|E|$  – общее количество “экспертов”, обеспечивающих необходимую точность классификации.

Для моделирования внутригодового хода концентраций тяжелых металлов использована модель случайного леса (Random Forest). Случайный лес – алгоритм машинного обучения, заключающийся в использовании композиции (ансамбля) решающих деревьев (Breiman, 2001).

Рассмотрим принцип работы одного решающего дерева. Свое название данный алгоритм получил за счет древовидной структуры базовых моделей. Решающее дерево выстраивает иерархию правил “если... то”, приводящую к решению. Структура дерева представляет собой “листья” и “ветви”. На ребрах (“ветвях”) дерева решения записаны правила, от которых зависит целевая функция, в “листьях” записаны значения целевой функции (ответы). Чтобы получить решение на новых данных, необходимо спуститься по дереву от корня до листа.

Основная идея алгоритма Random Forest заключается в использовании ансамбля решающих деревьев, каждое из которых само по себе даёт не очень высокое качество модели, однако хороший результат достигается путем осреднения ответов большого числа независимых решающих деревьев.

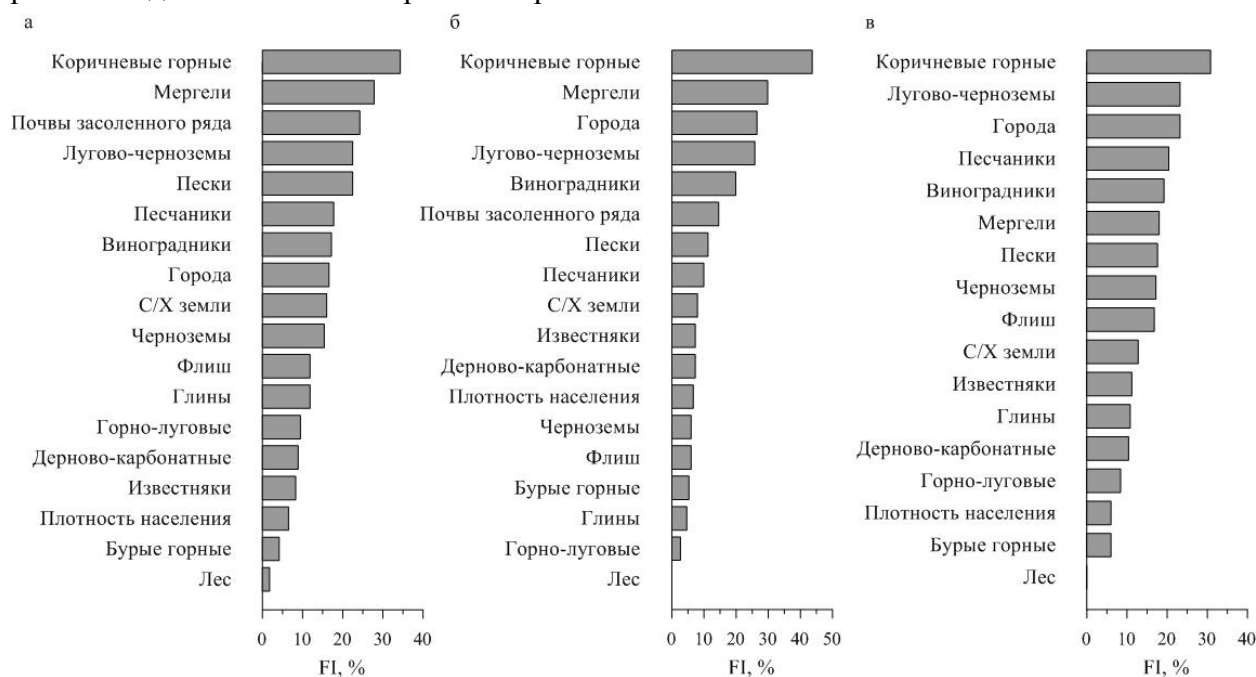
Модель реализована на языке программирования Python 3 с использованием библиотеки машинного обучения scikit-learn. Алгоритм Random Forest используется для решения задачи регрессии. Исходные данные для построения модели: антропогенные характеристики водосборов (плотность населения и доли от водосборной площади урбанизированных территорий, сельскохозяйственных земель, виноградников), среднесуточные расходы воды, среднесуточная температура воздуха, средняя высота водосбора, номер бассейна реки (категориальный признак). Целевая функция – концентрации (мкг/л) тяжелых металлов в речных водах. Оценка качества модели проведена с использованием полной (leave-one-out) кросс-валидации по метрике средней абсолютной ошибки.

### **Результаты исследования**

Для оценки влияния природных и техногенных факторов на формирование химического состава рек проведено обучение алгоритма распознавания и рассчитаны векторы информационных весов. Обучение проведено по критерию “количество верно классифицированных элементов обучающей выборки не менее 75%”. На рис. 1 представлены результаты ранжирования признаков по их влиянию на процесс формирования химического состава речных вод.

Наиболее значимый фактор формирования химического состава речных вод рек Крымского полуострова – горные коричневые почвы. Однако, вероятнее всего, на процесс формирования состава вод влияют не сами свойства данных почв, а их географическое положение. Горные коричневые почвы располагаются в приустьевых районах, где атмосферные осадки обогащаются морскими солями над акваторией моря. Минерализацию речных вод в прибрежных районах могут повышать и грунтовые воды, дренирующие засоленные морские отложения. Лесная растительность (точнее, залесенность территории) – наименее значимый фактор формирования химического состава речных вод для всех сезонов, она не оказывает влияние на преобразование речных вод, так как в основном на территории изучаемых речных бассейнов она приурочена к верховьям рек, где формируются фоновые воды. На формирование состава речных вод в начало зимнего паводкового периода наибольшее влияние оказывают геологический и почвенный факторы, что обусловлено преобладанием грунтовых вод в питании рек. Из группы антропогенных факторов значительное влияние оказывают виноградники, урбанизированные территории и сельскохозяйственные земли. В период летнего паводка 2015 г. среди антропогенных факторов наибольшее влияние на формирование состава вод оказали городские и сельские поселения, что обусловлено смывом пыли и различных

загрязняющих веществ с урбанизированных территорий во время дождевого паводка. В летнюю межень наибольшее влияние оказывают природные факторы, однако на них накладываются антропогенные. Среди почвенного фактора наиболее значимый – лугово-черноземные почвы. Они приурочены к речным долинам, их влияние обусловлено протеканием рек в пределах естественного русла. Существенное влияние оказывают виноградники и урбанизированные территории. В период низкого стояния уровня рек и небольших расходов воды поступление даже небольшого количества загрязненных вод с урбанизированных территорий и виноградников может оказывать сильное воздействие на речные воды за счет малой кратности разбавления.



**Рис. 1. Вклад природных и техногенных факторов в формирование химического состава речных вод Крымского полуострова: а – зимний паводковый период, б – летний паводок, в – летняя межень.**

Применение методов машинного обучения позволило смоделировать внутригодовой ход концентраций Cu, Co, Ni, Mo, Ba. На основе данной модели рассчитан среднемесячный сток растворенных форм этих элементов. В табл.1 приведен пример результатов моделирования стока меди, никеля и кобальта для гидрологического поста р. Альма – Суворово.

Анализ внутригодового распределения показывает, что в 2015 г. около 90% стока данных металлов прошло в многоводный период: 50% в зимне-весенний паводковый период (январь – май) и 40% во время ливневого летнего паводка. Используемый подход также позволяет исследовать влияние признаков на итоговый результат моделирования. Главные факторы, влияющие на содержание меди в речной воде – сток и количество виноградников на водосборе, которые являются основным антропогенным фактором поступления Cu на изучаемой территории. Для борьбы с вредными насекомыми и оидиумом (грибком винограда) плантации обрабатывают медным купоросом и медьсодержащими препаратами. Увеличение склонового стока в многоводный период приводит к увеличению поступления меди с водосборного бассейна.

На содержание никеля наибольшее влияние оказывают урбанизированные территории и температура воздуха. Загрязнение никелем происходит при сжигании твердого и жидкого топлива, с выхлопными газами автотранспорта, а также за счет износа автомобильных шин и деталей автомобилей. Отмечается повышение концентраций

никеля в период летней межени, что может быть связано со стоками населенных пунктов и малой кратностью их разбавления в период низкой водности.

*Табл. 1. Среднемесячный сток растворенных форм тяжелых металлов для р. Альма – Суворово, кг/мес*

| Месяц | 1    | 2    | 3    | 4     | 5     | 6     | 7     | 8    | 9    | 10   | 11   | 12   |
|-------|------|------|------|-------|-------|-------|-------|------|------|------|------|------|
| Co    | 0.62 | 1.37 | 1.26 | 2.07  | 1.86  | 4.17  | 1.42  | 0.14 | 0.14 | 0.12 | 0.31 | 0.31 |
| Ni    | 3.26 | 6.81 | 6.38 | 10.54 | 10.15 | 22.34 | 8.92  | 1.21 | 0.90 | 0.67 | 1.64 | 1.64 |
| Cu    | 3.93 | 9.86 | 8.46 | 14.99 | 16.48 | 39.61 | 10.95 | 0.77 | 0.89 | 0.81 | 2.02 | 2.06 |

Основные факторы, влияющие на содержание кобальта, – сельскохозяйственные земли и орографическое положение водосбора. На повышение урожайности сельскохозяйственных культур влияет внесение в почву различных минеральных удобрений, в том числе и кобальтсодержащих. Кобальт положительно влияет на течение многих физиологических процессов, происходящих в растениях, повышает активность ферментов, увеличивает засухоустойчивость растений. К наиболее слабо освоенным в хозяйственном отношении территориям Крымского полуострова относятся горные районы, таким образом, орографическое положение водосбора может служить индикатором антропогенной нагрузки.

#### **Выводы**

Предложенный подход является эффективным инструментом для научных исследований и позволяет количественно оценить влияние природных и техногенных факторов на качество воды. Показано, что на формирование состава речных вод как в межень, так и в паводковый период наибольшее влияние оказывают природные факторы (близость к морю, горные породы и почвы), из группы антропогенных факторов главное значение имеют виноградники и урбанизированные территории. На примере тяжелых металлов предложена методика расчета внутригодового стока загрязняющих веществ по данным единичных наблюдений, основанная на методах машинного обучения. Показано, что основное количество изучаемых тяжелых металлов проходит в многоводный период, на формирование стока данных элементов влияют различные факторы.

Предложенный подход в совокупности с географическим анализом может применяться для комплексных геоэкологических исследований. Количественное моделирование выноса загрязняющих веществ может быть использовано для построения сценарных прогнозов, а полное понимание всех характеристик, влияющих на процесс формирования химического состава воды, позволяет разрабатывать рекомендации для более грамотного управления водными ресурсами.

**Работа выполнена при поддержке гранта РФФИ 16-05-01037.**

#### *Список литературы*

- Алешин С.В. Распознавание динамических образов. Ч. 1. М.: Издательство МГУ, 1996г. 98 с.
- Кудрявцев В.Б., Андреев А.Е. Теория тестового распознавания // Интеллектуальные системы, 2006, Т.10: 95–140.
- Малыгин И.В. Логический подход к созданию экспертных систем прогнозирования опасных природных явлений // Естественные и технические науки, 2015, №2: 102-112.
- Малыгин И.В. Свидетельство о государственной регистрации программы для ЭВМ №2014614960 Экспертная система прогнозирования ледового заторообразования. Заявка №2014612636, дата поступления 24.03.2014. Дата гос. регистрации в Реестре программ для ЭВМ 14.05.2014.
- Никаноров А.М. Гидрохимия: учебник. – СПб.: Гидрометеиздат, 2001. – 444 с.



Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. Пер. с англ. А. А. Слинкина. М.: ДМК Пресс, 2015. – 400 с.

Breiman L. Random Forests. Machine Learning , 2001, vol. 45, no.1, pp. 5–32

## Machine learning approach for river water quality modeling: the Crimean peninsula case study

**Malygin E.V.<sup>1</sup>, Lychagin M.Yu.<sup>1</sup>, Malygin I.V.<sup>2</sup>**

<sup>1</sup>*Lomonosov Moscow State University*

<sup>2</sup>*Schmidt Institute of Physics of the Earth of the Russian Academy of Sciences*

*[malygin.ev@gmail.com](mailto:malygin.ev@gmail.com)*

**Abstract:** This paper studies the impact of natural and technogenic factors on the chemical composition of river waters of the Crimean peninsula using machine learning approach for river water quality modeling. It is shown that natural factors (proximity to the sea, bedrock types and soils) have the greatest influence on the chemical composition of river waters both in the low water period and during the flood. The most important of all the technogenic factors are the shares of the catchment area attributable to agricultural land and to urban areas. Using machine learning techniques allowed to simulate annual variability of several trace metal concentrations (Cu, Co, Ni, Mo, Ba) and to calculate the average monthly ion runoff of the studied elements.

**Keywords:** water quality modeling, machine learning, trace metal, Crimean peninsula.