

ОБНАРУЖЕНИЕ НЕОЧИЩЕННЫХ СТОЧНЫХ ВОД, СБРАСЫВАЕМЫХ В ВОДОВОДЫ, С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Peter Hammond¹, Michael Suttie², Vaughan T. Lewis³, Ashley P. Smith⁴,
Andrew C. Singer²

NPJ Nature Partner Journals⁵

DETECTION OF UNTREATED SEWAGE DISCHARGES TO WATERCOURSES USING MACHINE LEARNING

npj Clean Water (2021)4:18;
<https://doi.org/10.1038/s41545-021-00108-3>

Мониторинг и регулирование сброса загрязненных сточных вод в водные объекты является обязанностью Агентства по окружающей среде⁶. Выявление и отчетность о случаях загрязнения, вызванного установками по очистке сточных вод, является ответственностью эксплуатирующих организаций. Тем не менее, в 2018 г., по сообщениям общественности, в Англии было свыше 400 случаев загрязнения водных объектов сточными водами. В статье представлены новые методологии отчетности о случаях загрязнения с целью идентификации вероятных разливов неочищенных сточных вод с установок по очистке сточных вод. Ежедневные характеристики очищенных сточных вод с двух установок по очистке были дополнены сообщениями эксплуатирующими организациями случаями сбросов неочищенных сточных вод. С использованием машинного обучения известные случаи разлива служили в качестве набора данных для обучения. Вероятность правильной классификации произвольно выбранных пар моделей очищенных сточных вод с “разливом” и “без разлива” была выше 96%. Из 7160 дней без сообщений эксплуатирующей организацией разливов 926 классифицировались как включающие “разлив”. Анализ также дает основания полагать, что на обеих установках по очистке сточных вод в период с 2009 по 2020 г. были сбросы неочищенных сточных вод. Это доказательство использования машинного обучения для обнаружения сбросов неочищенных сточных вод может помочь водохозяйственным компаниям идентифицировать

¹ Центр экологии и гидрологии Соединенного Королевства, Уоллингфорд, графство Оксфордшир.

² Институт больших данных, Оксфордский университет, Оксфорда, графство Оксфордшир

³ Windrush AEC Ltd, закрытое акционерное общество в области экологического консалтинга, Фордуэллс, Оксфордшир.

⁴ WASP Trust, благотворительное общество, борющееся против загрязнения сточными водами реки Уиндраш (Windrush), притока Темзы, графство Глостершир.

⁵ Серия журналов в открытом доступе в режиме онлайн, публикуемых в сотрудничестве с всемирно известными международными журналами Nature – британским журналом по естественнонаучной тематике, издающимся с 1869 г.

⁶ Агентство по окружающей среде, вневедомственный государственный орган, созданный в 1995г. при поддержке Министерства окружающей среды, продовольствия и сельскохозяйственного развития Соединенного Королевства (DEFRA), ответственный за защиту и улучшение состояния окружающей среды Англии.

неправильно эксплуатируемые очистные сооружения и информировать регулирующие органы о неудовлетворительной работе надзорных органов. Данные в режиме реального времени, открытый доступ, данные аварийной сигнализации, а также аналитические подходы дадут возможность профессионалам и гражданскому обществу провести научный анализ частоты и воздействия сбросов неочищенных сточных вод, в особенности тех, о которых не сообщают эксплуатирующие организации.

Введение

Агентство по окружающей среде Англии (EA) и Управление по регулированию водоснабжения Англии с Уэльсом (OFWAT), соответственно, являются органами, регулирующими экологические и экономические аспекты водохозяйственного сектора [1 – 7]. В 2018 г. было зафиксировано 48 случаев загрязнений категорий 1 и 2 (“крупные и серьезные”) и 1527 случаев категории 3 (“минимальные”), которые воздействовали на качество речной воды, связанных с канализационной сетью и установками по очистке сточных вод (WWTPs) в Англии [1, 8]. Организации, эксплуатирующие WWTP, сами регистрировали от 62 до 84% идентифицированных случаев загрязнений в Англии в 2018 г.; на общественность и третьи стороны приходилась отчетность на оставшиеся 395 случаев [1].

Немаловажная роль общественности, сообщавшей о случаях загрязнений в Англии, выявила, что эксплуатирующие организации могут значительно занижать количество сообщаемых случаев загрязнений; (2) общественность, бессознательно, играет важную роль в осуществлении контроля за деятельностью водохозяйственного комплекса; (3) весьма вероятно, что ежегодная отчетность о случаях загрязнения сточными водами является неполной [9]. Неопределенность, связанная с частотой, длительностью и воздействием случаев загрязнения сточными водами, способствует сохранению пробелов в доказательствах, необходимых для информирования о вмешательстве, капиталовложениях и привлечения к ответственности. В данной ситуации мы применили способы машинного обучения (ML) с целью использования имеющихся потоков данных для акцентирования внимания на предполагаемых случаях загрязнения сточными водами. Авторы также использовали данные об осадках, речном стоке и аварийных ситуациях на WWTP для увязывания с контекстом потенциальных эффектов загрязнений и возможных несоответствий с разрешениями EA на сброс неочищенных сточных вод.

Регламент об экологической информации (EIR) требует в рамках нормативных положений Соединенного Королевства, вводящих в действие Директиву ЕС 2003/4/CE [10], получения ежедневных данных о характеристиках потока очищенных сточных вод и мониторинга длительности событий (EDM), т.е. о времени начала и прекращения сбросов неочищенных сточных вод из ливневых резервуаров. В течение периодов дождей осадков ливневые резервуары используются на многих WWTPs для временного накопления избыточных сточных вод, когда приток воды непомерно нарастает из-за поверхностных сточных вод. Содержимое ливневых резервуаров необходимо, в соответствии с разрешениями на сброс в водотоки, передать на очистку, как только приток воды уменьшится, и производительность WWTP восстановится. В течение интенсивных дождей емкость ливневых резервуаров может не справиться с приемом, что приведет к разрешенным разливам неочищенных, но с частично уловленными сетчатыми

фильтрами загрязнениями, переливами сточных вод из ливневых резервуаров, что приведет к загрязнению принимающих водотоков.

В 2013 г. Министерство окружающей среды, продовольствия и сельскохозяйственного развития Соединенного Королевства (DEFRA) заявило, что к 2020 г. должны быть установлены устройства для EDM на большей части смеси городских и ливневых сточных вод, перетекающих из общесплавной канализации в водный объект, минуя очистные сооружения, на канализационных сетях, переливах ливневых резервуаров WWTP и аварийных переливах канализационных насосных станциях (SPS) в Англии с Уэльсом. Хотя термин “перелив ливневых резервуаров” обычно используется в водохозяйственном секторе, EA и разрешениях на сброс в сочетании грунтового и поверхностного стока, очевидно, что это относится к общесплавной канализации.

С 2016 г. данные о количестве и длительности выявленных разливов EDM ежегодно сообщаются в EA организациями, эксплуатирующими WWTP. В 2020 г. сообщения о EDM были сделаны для 70% таких переливов. Сброс неочищенных сточных вод в течение чрезвычайных дождевых осадков должен проводиться в соответствии с разрешением и до тех пор, пока внедрение устройств EDM не сделает обязательной самоотчетность эксплуатирующей организации перед EA. В течение периодов дождевых осадков, которые не являются чрезвычайными, такие сбросы не соответствуют условиям разрешения и потенциально являются нелегальными в рамках законодательства ЕС и Соединенного Королевства [11, 12].

Требуется также, чтобы организации, эксплуатирующие WWTP, продолжали очищать минимальный поток сточных вод, который является, по крайней мере, привязанным к конкретному для очистных сооружений “уровню ливневого переполнения”, определенному в разрешении EA даже, когда избыточный поток отводится или разливается из ливневого резервуара. Поэтому не будет соответствовать разрешению прием в ливневый резервуар неочищенных сточных вод или сброс перелива в водоток, когда очищенный поток ниже уровня ливневого переполнения. Несмотря на то, что для проверки соответствия требуется, чтобы поток подвергался очистке, обычно при выдаче разрешения не требуется регистрировать это, и многие WWTPs регистрируют только поступление сточных вод. Поток сточных вод, очевидно, тесно связан с потоком, поступающим на очистку, но он не может быть юридически приемлемым заменителем для подтверждения соответствия. В дальнейшем авторы будут цитировать термины “ливнесток” и “разливы” для обозначения возможных ливневых сточных вод, определяемых нашим анализом потоков, но, возможно, не определяемых с помощью устройств EDM, системы аварийной сигнализации отвода из ливневого резервуара и даже имеющих место в течение не чрезвычайных дождевых осадков.

Авторы выбрали две WWTPs, эксплуатируемые одной и той же водохозяйственной компанией, с разным количеством обслуживаемых жителей и наличием данных о принимаемых потоках. Краткая информация о характеристиках WWTP1 и WWTP2 представлена в табл. 1, где намеренно была сохранена анонимность. Ответы на EIRs подтвердили, что на обеих WWTPs регистрируются только очищенные сточные воды, и, следовательно, исключено прямое подтверждение соблюдения минимальной степени очистки для потока, поступающего на очистку во время разливов. Сертификация систем мониторинга (MCERTS) в рамках ежесуточного 15-ти минутного определения характеристик потока сточных вод на протяжении 8000 дней в течение 11-ти лет (2009 – 2020) и данных EDM, определяющих интервалы между разливами, также запрашиваемых EIR на протяжении 900 дней (2018- 2020).

Метаданные⁷ для WWTP1 и WWTP2

Данные эксплуатирующей организации	WWTP1	WWTP2
Дополнительная очистка	Нет	Удаление фосфатов
Эквивалентная численность жителей (population equivalent)	7594	47000
Тип данных о потоке	Очищенные сточные воды	Очищенные сточные воды
	3990 (01/09-03/20)	4083 (01/09-03/20)
Количество дней с сообщенными данными о потоке	С EDM	С EDM
	446	471
	Без EDM	Без EDM
	3544	3612
Расход ливнестока (л/с)	50,52	240
Емкость резервуара для приема ливневых вод (м ³)	Разрешенная	Разрешенная
	364	1728
	Реальная	Реальная
	359	2201
Перелив из дождеприемника	(01/09-03/20)	(01/09-03/20)
Согласованные данные об уровне перелива	04/18 – 03/20	04/18 – 03/20
Количество дней с данными о потоке для анализа формы потока	1511	1527

⁷ Метаданные – информация о другой информации, или данные, относящиеся к дополнительной информации о содержимом или объеме.

Цель работы заключалась в разработке способов для анализа ежегодных характеристик потока и данных EDM, с помощью которых можно обнаружить разливы неочищенных сбросов “переливов из дождеприемника” в водотоки. По мнению авторов, ретроспективное определение таких разливов должно принести пользу как водохозяйственным компаниям и регулирующим органам, так и гражданскому обществу и профессиональным ученым, проявляющим интерес к загрязнению водотоков сбросами сточных вод. Авторы применили подход машинного обучения (ML), в котором использовали зарегистрированные характеристики потока в течение EDM, связанного с разливами от перелива из дождеприемников, для обучения алгоритмов распознавания характеристик с целью определения сходных характеристик потока, когда не было сообщений, или не было известно о сбросах сточных вод. Способы искусственного интеллекта (ИИ), основанные на четко выраженных символических представлениях о нормативных положениях и законодательстве, были успешно разработаны и применяются в течение нескольких десятилетий [13]. Использование количественных методов ИИ, таких как машинное обучение (ML) и распознавание образов для проверки соблюдения нормативных требований в настоящее время получило внимание, так как правительство и промышленность накапливают большие базы данных, доступные для непрофессиональной, научной и нормативной проверки [14, 15]. Например, способы ML недавно применялись для предсказания вероятности того, что организация не пройдет проверку правительственного агентства США на соответствие с экологическими нормами [16]. Авторы использовали общедоступную базу данных обеспечения и соблюдения природоохранного законодательства (<https://eco.epa.gov>) для предсказания соблюдения на основе местоположения, отрасли промышленности и данных о предыдущих инспекциях в течение 5 лет.

В качестве предшественника компонента ML исследования был выполнен анализ формы характеристик потока в течение 3038 суток с 2016 по 2020 г. с целью идентификации представления о компактных очистных сооружениях. Затем авторы использовали обучение с учителем с 20 вариантами стандартных алгоритмов ML на 917 характеристиках потока с 2018 по 2020 г., с данными EDM, с целью разработки классификатора, способного делать различие между теми водными объектами, которые под воздействием разливов неочищенных сточных вод, и теми, которые не были подвергнуты такому воздействию. Оптимальные классификаторы, по одному на каждую WWTP, впоследствии проверенные в рамках полуслеплого анализа на 2121 потоке с 2016 до 2018 г, были использованы для анализа формы характеристик потока, но не для обучения с учителем. После этого классификаторы применяли ретроспективно и в полностью слепом режиме для 5039 ежесуточных характеристик потоков с 2009 по 2015 г., но не для анализа формы и обучения с учителем. И, наконец, общедоступные данные об осадках, речном потоке и уровне воды в реке, а также телеметрические данные аварийной сигнализации, полученные через EIR от эксплуатирующей организации, были введены для увязки с контекстом и для подтверждения потенциальных дней “разливов”, идентифицированных статистическими методами, или с помощью ML, и для информирования при обсуждении соблюдения условий разрешения EA на сброс.

Основной вклад этого исследования состоит в использовании моделей ML для обнаружения несообщенных разливов неочищенных сточных вод с очистных сооружений, на которых имеются ранее обученные модели для определения связи между известными фактами разливов и соответствующими

пциями нарушениями потока очищенных сточных вод. Добавление телеметрических данных аварийной сигнализации и данных об осадках дает возможность компаниям, эксплуатирующим очистные сооружения, и регулирующим органам выявлять неполадки в работе оборудования и несоблюдение условий разрешения, соответственно. Профессионалы и ученые-любители также получают выгоду от улучшенной идентификации возможных проливов, которая может воздействовать на исследование ими потенциально загрязненных водотоков.

Результаты

Анализ формы ежесуточных характеристик потока за 3038 дней (2016 -2020) для WWTP1 и WWTP2

Пример характеристик потока очищенных сточных вод в течение 10-ти дневного периода на WWTP1 приведен на рис. 1. Выявленные интервалы разливов EDM показаны сверху для демонстрации эффекта сглаживания разливов на профиль характеристик потока.

Для WWTP1 (соответственно, WWTP2) данные EDM имелись для 446 (соответственно, 471) календарных дней за 2018 – 2020 гг., в течение которых были зарегистрированы разливы неочищенных сточных вод с интервалами различной протяженности. Для каждого дня интервалы разливов суммировали до общего количества часов сброса. Из дней, использованных для машинного обучения для WWTP1 (соответственно, WWTP2), в 339 (соответственно, 346) не было зарегистрировано случаев разливов и в 97 (соответственно, 117) общая длительность “разливов”, по крайней мере, 3 ч, помечалась как “разливы”, а 349 (соответственно, 354) с общей длительностью “разливов” менее 3 ч, как “нормальный” режим. Суммарный период в 3 ч был выбран в силу того, что это гарантирует обоснованное количество дней “разливов”, на которых основано обучение с учителем, а предварительные попытки предсказать количество часов разливов в день были самыми для суммарных разливов протяженностью менее 3 ч. Средние “нормальные” (голубая линия) и “разлив” (черная линия) ежедневные характеристики как доля ливневого перелива (красная линия) показаны на рис 2. для каждой WWTP. Расход ливнестока – это отметка минимального потока (расхода), который должен быть очищен перед тем, как можно в соответствии с разрешениями ЕА можно сделать сброс неочищенных сточных вод в водотоки.

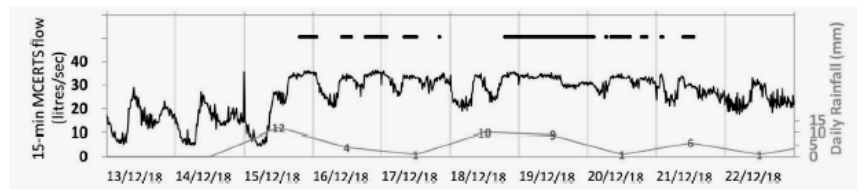


Рис. 1. WWTP1: пример характеристик потока очищенных сточных вод в течение 10 дней, аннотированных с подтвержденными интервалами переливов EDM

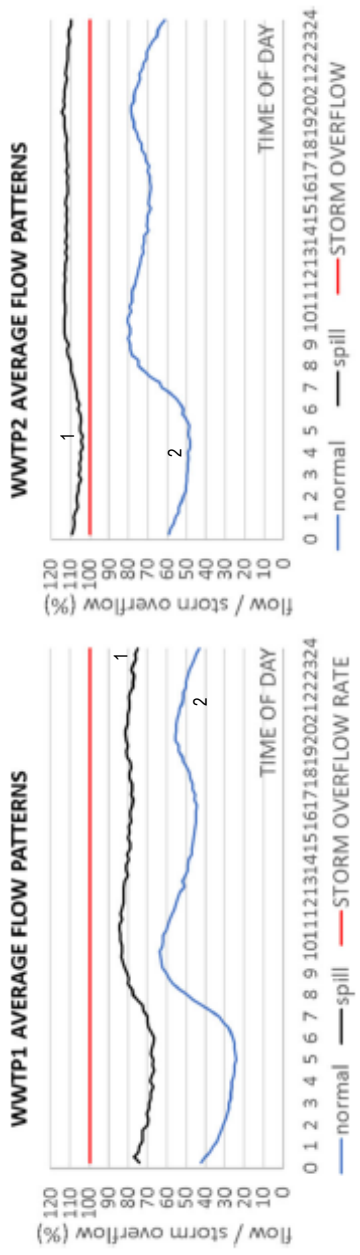


Рис. 2. Среднесуточные характеристики потока.
 WWTP1: кривая 1 для дней с “разливами” (97 дней) и кривая 2 для “нормальных” дней (349 дней),
 WWTP2: кривая 1 для дней с “разливами” (117 дней) и кривая 2 для “нормальных” дней (354 дней)

Пояснения к рисунку 1:

15-min MCERTS flow (litres/sec) – 15-ти минутный расход (л/с) с сертификацией систем мониторинга, Daily rainfall (mm) – суточное количество осадков (мм).

Пояснения к рисунку 2:

flow/storm overflow (%) – поток/ливневосток, Average Flow Patterns – средние характеристики потока, Time of Day- время суток, normal – нормальный (поток), spill-разлив, Storm overflow rate – уровень ливневого переполнения.

Ежесуточные характеристики потока длительностью 24 ч (с полуночи до полуночи), состоящие из 96 15-ти минутных интервалов со средним расходом (л/с) очищенных сточных вод показаны голубым цветом. Черные горизонтальные линейные пояснения представляют собой зарегистрированные интервалы EDM, означающие сброс из ливневого резервуара (т.е. разрешенный разлив или потенциально не разрешенный разлив неочищенных сточных вод), самый короткий из которых составляет 15 мин, а самый длинный- 24 ч. Общее суточное количество осадков (мм/сутки) показано зеленым цветом. Первые двое суток, без обнаруженных разливов, характеризуются суточными характеристиками низкого расхода между полуночью (предыдущего дня) и первым пиком после середины утра (9 ч), с последующим затишьем до второго меньшего пика вечером. Следующие семь дней (с 15 до 21 декабря 2018 г.) характеризуются интервалами разливов различной длительности (черная линия EDM), демонстрирующими сглаживание расхода, которое является типичным ливневым в течение интенсивных осадков. Последние сутки характеризуются повышенными расходами и частичным возвратом к суточным характеристикам расхода, без выявленных разливов.

Отдельные формы моделей были сгенерированы из характеристик потока с 2016 по 2020 гг. для WWTP1 (1511 дней) и WWTP2 (1527 дней). Первый метод главных компонент⁸ изменения формы, PCA 1, в обеих моделях связан с величиной и временным изменением утреннего пика расхода (дополнительное видео 1.mp4), а также с “сезонными” изменениями, вызванными переходом на летнее время, праздничными данными, периодом отпусков (дополнение к рис. 1). Несмотря на различия в количестве населения, обслуживаемого WWTP1 и WWTP2, на рис. 3а, б показаны сходные распределения для графиков рассеяния PCA1 по сравнению с PCA2 для 2121 потоков за 2016 – 2018 гг. без данных EDM. Аналогичные графики PCA1 по сравнению с PCA2 для 917 потоков за 2016 – 2018 гг. без данных EDM (рис.3с, d) дают основание полагать, что для обоих WWTPs PCA 2 коррелируется с различием формы между “нормальным” потоком (светлые точки) и потоком под воздействием “разлива” (закрашенные треугольники). Это сглаживание, связанное с разливами, иллюстрируется видоизменением общих среднесуточных характеристик потока для WWTP1 между стандартными отклонениями PCA2 от -1 до +1 (дополнительное видео 1.mp4). Интересно, что площадь под графической характеристикой эксплуатируемого приемника потока, связанная только с использованием PCA2 для дифференциации “нормальный”/”разлив”, составляет 0,88 и 0,91 для WWTP1 и WWTP2 (это оцениваемая вероятность правильно классифицируемой пары случайно выбранных характеристик потока из подмножеств, обозначенных “нормальный” и ”разлив”).

⁸ Метод главных компонент – один из основных способов уменьшения размерности данных, с потерей наименьшего количества информации.

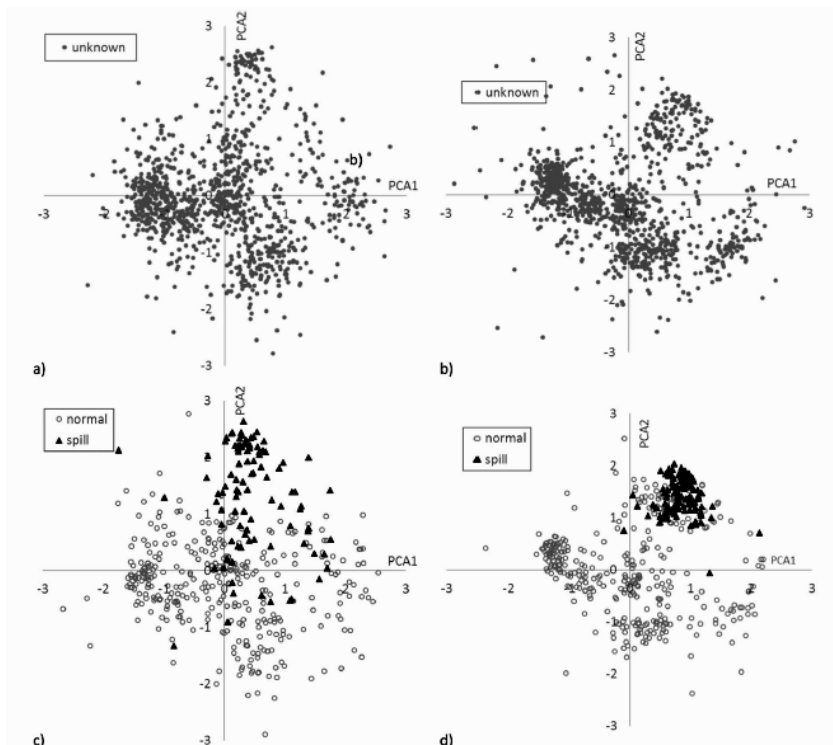


Рис. 3. PCA1 по сравнению с PCA2 для ежесуточных характеристик потока.
 Неизвестное состояние разлива (серые закрашенные точки);
 разлив, подтвержденный EDM до WWTP2 (1056 дней);
 c) WWTP1 с данными EDM за 2018 – 2020 гг. (466 дней);
 d) WWTP2 (471 день)

Пояснения к рисунку 3:

unknown - неизвестное, normal - нормальный, spill – разлив

Воздействие обучения с учителем на разливы сточных вод для 917 характеристик потока очищенных сточных вод

Эффективность 20-ти кратной перекрестной проверки обучения с учителем для табл. 3 и 4 для 20 вариаций метода опорных векторов⁹ (SVM), в то время как остается до 15 состояний PCA для синтеза характеристик потока. Количество остающихся состояний PCA для синтеза формы оказывает воздействие на точность преобразования каждой ежесуточной характеристики

⁹ Метод опорных векторов – набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа.

потока, и, следовательно, точность классификации. Для трех наиболее эффективных алгоритмов дополнительно на рис. 2 показаны изменения в точности классификации ежедневных характеристик потока для различного количества оставшихся состояний РСА, оцененных как средняя площадь под 20 эксплуатируемыми приемниками потока, связанных с кратностью перекрестных проверок. Для оптимальных классификаторов средняя площадь под графикскими характеристиками составляла 0,97 для WWTP1 и 0,97 для WWTP2.

Для подтверждения, перед более широким применением использовали оптимальные классификаторы ML для каждой WWTP для перекалификации характеристик потока, использованных при их получении. На рис. 4 показаны эти характеристики потока в непрерывной временной последовательности с аннотациями для каждого дня, отражающего определенные интервалы разливов EDM (горизонтальные черные сегменты) и подтверждение ML “разливов” (не заполненные золотистые точки). В течение этого периода было 97 (соответственно, 117) дней с подтвержденными EDM суммарными разливами, по крайней мере, в течение 3 ч, на WWTP1 (соответственно, WWTP2). Соответствие между оптимальной классификацией ML и отмеченными днями разливов, полученными по данным EDM, было чрезвычайно высоким (WWTP1: чувствительность = 0,91; специфичность = 0,95; (WWTP2: чувствительность = 0,98; специфичность = 0,98), как и следовало ожидать для таких “тренированных” данных.

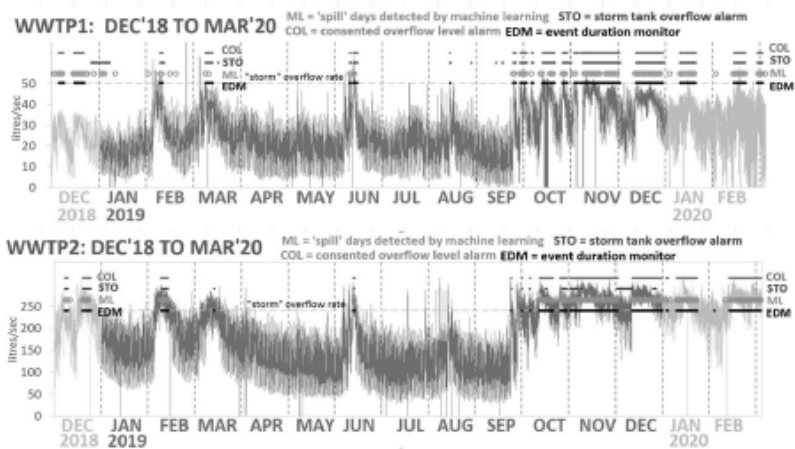


Рис. 4. Ежесуточные характеристики потока очищенных сточных вод и интервалы обнаруженных разливов при мониторинге длительности событий (EDM) на WWTP1 и WWTP2, используемые как тренировочные данные (декабрь 2018 г. – март 2020 г.)

На рис. 4 включены также данные другой аварийной информации, связанной со сбросами неочищенных сточных вод, которые могли подтвердить классификацию ML характеристик потока для прошлых периодов без данных EDM. Для WWTP1 имеется соответствие, близкое к идеальному (ко-

эффицент капша Коэна¹⁰: 0,81 – 1,00) между EDM, STO (перелив из ливневого резервуара) и COL (разрешенный уровень перелива) и классификацией ML с 19 по 20 декабря 2020 г. (рис. 4 и табл. 2). В течение двух месяцев (декабрь 2018 и январь 2019г.) данные EDM и COL имели соответствие, близкое к идеальному (коэффициент капша Коэна: = 0,95), данные STO в основном расходятся (коэффициент капша Коэна: ≤ 0), а классификатор ML указывал на инциденты, определенные тремя способами. Эти результаты дают основание полагать, что STO является хорошим кандидатом, а аварийные данные COL являются превосходным кандидатом, для подтверждения обнаруживаемых с помощью ML возможных разливов на WWTP1, когда не имеется данных EDM.

Данные о ежесуточных потоках и разливах EDM оценивались с интервалами в 15 мин. Дополнительные пояснения - это данные телеметрической аварийной сигнализации, предоставленные эксплуатирующей организацией. Эти данные обладают потенциалом подтверждать предсказания ML о днях с “разливами” для невиденных характеристик потока с 2009 по 2018 гг., для которых имеются данные EDM. Сходные графики, показывающие невиденную классификацию ML для ежесуточных характеристик потока представлены на рис. 5.

Пояснения к рисунку 4:

WWTP1: Dec'18 to Mar'20 - WWTP1: декабрь 2018 – март 2020 г., litres/sec – л/с, ML = ‘spill’ days detected by machine learning – машинное обучение = дни “разлива”, определенные с помощью машинного обучения, STO = storm tank overflow alarm – аварийная сигнализация о переливе из ливневого резервуара, COL = consented overflow level alarm- аварийная сигнализация о достижении разрешенного уровня перелива, EDM – event duration monitor – мониторинг длительности события, “storm” overflow rate – уровень дождеприемника, WWTP2: Dec'18 to Mar'20 – WWTP2: декабрь 2018 – март 2020 г.

Для WWTP2 имеется почти идеальное соответствие между EDM и аварийной сигнализацией о достижении разрешенного уровня перелива COL (коэффициент капша Коэна= 0,87) (рис. 4 и табл. 2). Не были предоставлены данные аварийной сигнализации о переливе из ливневого резервуара STO за 2020 г., а в период с декабря 2018 по декабрь 2019 г. выявлена возможность только случайного соответствия с другими устройствами и классификатором ML (коэффициент капша Коэна $< 0,1$). Эти результаты дают основание полагать, что STO является плохим кандидатом, в то время как COL является COL превосходным кандидатом для подтверждения возможных разливов, определенных с помощью ML, когда не имеются данные EDM.

Определение разливов в 7160 ежесуточных характеристиках потока (2009-2018 гг.), не использованных для обучения алгоритмов ML

Классификация 2121 характеристик потока с января 2016 г. по ноябрь 2018 г. считалась полуслепой, так как эти характеристики использовали при анализе формы, но не в “обучении” ML, в то время как классификация 5039 характеристик потока считалась полностью слепой, так как эти характери-

¹⁰ Коэффициент капша Коэна предложен американским психологом и статистиком Джейкобом Коэном для сравнения наблюдаемой и ожидаемой точностью. Чем больше капша, тем ближе сравниваемые точности

стики не использовали ни для того и другого. В табл. 3 подытожено ежегодное количество потенциальных дней ”разливов”, обнаруженных с помощью алгоритмов ML.

Подмножество из 327 дней ”разливов”, обнаруженных с помощью анализа ML в период с 2009 по 2018 гг. на WWTP1, было подтверждено данными по переливу из ливневого резервуара STO или данными аварийной сигнализации о достижении разрешенного уровня перелива COL. Для того же самого периода подмножество из 128 дней ”разливов”, обнаруженных на WWTP2, было подтверждено данными STO или сигнализации COL. Аварийная сигнализация COL подтвердила все обнаруженные разливы, для которых были доступны данные, в то время как ненадежность, показанная ранее для аварийной сигнализации COL на WWTP2, предполагает использование альтернативного подхода для подтверждения разливов, обнаруженных с помощью анализа ML. Для обеих WWTP имелось примерно 3 дополнительных месяца данных для потока и EDM (87 дней в период с 7 марта до 1 июня 2020 г.) после окончания наличия данных обучения ML. Этот период был пропущен из исходных данных обучения ML, поскольку ежедневные расходы потока на WWTP1 были равны нулю или менее 1% ожидания на протяжении более 50% времени, и, следовательно, очистные сооружения не использовались. Такие аномальные данные в любом случае – следствие нарушения требований разрешения EA, и в течение каждого года только на протяжении 37 дней данные обучения отсутствовали, или вызвали подозрение. Однако была возможность провести слепое тестирование ежесуточных характеристик потока в течение 87 дней на WWTP2 относительно моделей классификации, сделанных для периода с декабря 2018 по март 2020 г., и подтвердить соответствие с данными EDM на протяжении 93 % срока наблюдений.

Статистика капша Коэна оценивает соответствие между двумя показателями и интерпретируется следующим образом:

≤0 нет соответствия;

0,01-0,20 от отсутствия до пренебрежимо малого;

0,21-0,40 удовлетворительное;

0,41-0,60 умеренное;

0,61 -0,80 существенное;

0,81- 1,00 почти полное соответствие

N/A – данные отсутствуют, остальные обозначения см. выше

Когда на WWTP1 происходил разлив неочищенных сточных вод, определенный с помощью аварийной сигнализации COL/EDM либо классификации ML, это обычно происходило при расходе сточных вод, который был значительно ниже уровня ливневого стока (50,52 л/с), обусловленного разрешением EA в качестве минимального расхода для непрерывной очистки (расход сточных вод, поступающих на очистные сооружения, или PFF). Это можно было также увидеть в течение периода мониторинга EDM с 2018 по 2020 гг. Сравнение средних характеристик потока с ”разливом” или ”нормального” (рис. 2) показывает, что средний поток очищенных сточных вод для дней с ”разливом” на WWTP1 никогда не был выше уровня перелива из ливневого резервуара, в то время как на WWTP2 он всегда был выше. Характерно, что на WWTP1 141 из 274 (51,5%) необобщенных (т.е. отдельных) разливов, обнаруженных с помощью EDM на WWTP1, начинались, когда расход очищенных сточных вод был меньше, чем 80% уровня перелива из ливневого резервуара, по сравнению с отсутствием таковых на WWTP2.

Таблица 2

**Соответствие классификации ML, EDM, COL и STO
для обучения с учителем**

	Каппа Коэна	ML-EDM	ML-COL	EDM-COL	STO-EDM	STO-COL	STO-ML
- 0-	12/18-01/19	0.53	0.57	0.95	-0.25	-0.24	0.00
	02/19-03/20	0.87	0.87	1.00	0.93	0.93	0.81
WWWTP2	12/18-03/20	0.78	0.78	0.87	N/A	N/A	N/A
	12/18-12/19	N/A	N/A	N/A	0,03	0,07	0,06

Таблица 3

**Количество потенциальных дней "разливов", определенных
с помощью машинного обучения**

Год	Машинное обучение (ML) для слепого тестирования										Итого	
	09	10	11	12	13	14	15	16	17	18	01/18-09/18	ML/EDM для обучения с учителем 12/18-03/20
WWWTP1	53	70	33	89	70	80	62	74	44	44	65	107/97
WWWTP2	18	21	6	56	41	56	4	31	13	40	40	121/117

^aДни разливов, определенные с помощью ML с длительностью 3 или более часов. До декабря 2018 г. не было данных для EDM, которые, по мнению эксплуатирующей, организации, были бы надежными

Вследствие режима самоизоляции, связанного с пандемией COVID-19 с марта 2020 г., разрешение для обоих WWTP, действительное на период до 2018 г., может быть не предоставлено в ответ на запрос EIR к EA, поскольку оно не в электронном формате и исходные положения не действуют. Однако для обоих WWTP нынешние разрешения, которые включают прошлые изменения, предполагают, что условия для ливнестоков не изменились по сравнению с условиями до 2009 г. Поэтому, по-видимому, разливы на WWTP1 были уже более 12 лет, в то время как на WWTP2 они происходили редко, и то в незначительной степени.

Определение с помощью ML отдельных и непрерывных серий суточных разливов

Для каждого WWTP характеристики ежесуточных потоков, определенные с помощью EDM или анализа с использованием ML, были распределены по степени сглаживания характеристик потока, что определяли по стандартному отклонению 96 15-ти минутных расходов потока. Для определенных с помощью ML “разливов” на WWTP1 без данных EDM 20 самых “сглаженных” ежесуточных характеристик потока очищенных сточных вод сравнивали на рис. 5 со средним потоком в сухую погоду. Каждый поток отражает непрерывный разлив в течение 24 ч для расхода очищенных сточных вод в диапазоне от 60 до 80% от минимально требуемого. Напротив, для 20 самых “сглаженных” ежесуточных характеристик потока очищенных сточных вод WWTP1 без данных EDM расход очищенных сточных вод был больше или равен соответствующему уровню ливнестока, и, вероятно, совпадал с минимальным расходом для условий очистки. Тем не менее, два из этих 20 “основных” 24-х часовых разливов на WWTP2 на рис. 5b 5-го и 12-го мая 2012 г. имели место в день без дождя после предыдущих 24 сухих часов. По всей видимости, это связано с просачиванием грунтовых вод, которое не разрешено EA. Широко признано, что просачивание грунтовых вод в канализационную сеть происходит, особенно в Англии, где многие канализационные сети находятся в работе более 100 лет (www.swig.org.uk/wp-content/uploads/2014/10/David-Walters-2015/pdf; <https://wwtonline.co.uk/news/thames-water-trials-sewer-infiltration-survey-system>; www.theguardian.com/environment/2020/oct/09/oxford-stop-thames-water-firm-dumping-sewage-river; www.southernwater.co.uk/help-advice/sewers/combined-sewer-overflow-csos).

Трудно получить данные об уровне грунтовых вод в конкретном месте за конкретные дни, когда происходит разлив. Кроме того, геологическое строение грунта для канализационных сетей и канализационных насосных станций (SPSs), обслуживающих две WWTP, в этом исследовании, существенно различалось, без скважинных данных для каждой SPS.

Зависящий от погоды поток из WWTP1, каждый разлив длится 4 ч, в течение которого расход очищенных сточных вод составляет от 60 до 80% от уровня перелива ливнестока. Для WWTP2, напротив, расход очищенных сточных вод был равен или выше соответствующего уровня перелива ливнестока. Два 24-х часовых разлива (5-го и 12-го мая 2012 г.) также демонстрируют “разливы при сухой погоде”, так как не было осадков ни в этот день, ни в предыдущий.

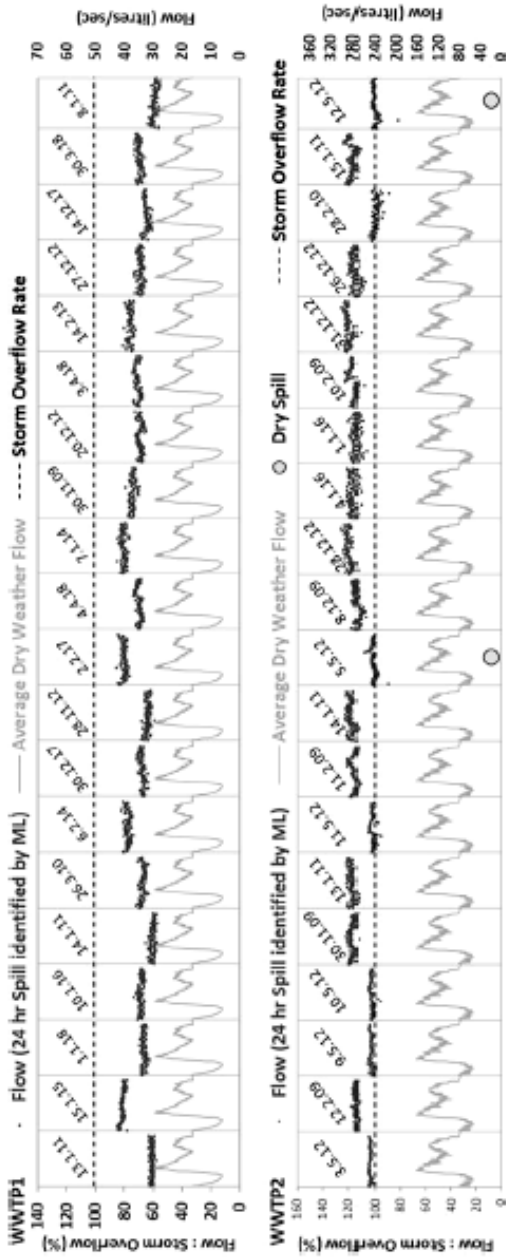


Рис. 5. Ежесуточные характеристики потоков очищенных сточных вод за 20 дней по сравнению со среднесуточными днями сухой погоды

Пояснения к рисунку 5:

Flow: Storm Overflow – поток ливнесток, Flow (24 hr Spill identified by ML) – поток (разлив в течение 24 ч, идентифицированный с помощью ML, Average Dry Weather Flow – средний поток в сухую погоду, Storm Overflow Rate – уровень аварийного перелива, Flow (litres/ sec) – расход (л/с), Dry Spill – разлив в сухую погоду.

Отдельный 24-х часовой разлив не очищенных сточных вод охватывает полный суточный цикл водостока, и, таким образом, включает два пика максимального входного потока, когда разлив разбавленных сточных вод, вероятно, наименьший, а риск ущерба от загрязнения наименьший. Но что еще хуже, потенциал загрязнения, вызванный непрерывной серией из 24-х часов разливов неочищенных сточных вод, в течение которых принимающий водоток не имеет ни отсрочки, ни возможности восстановления.

2009 – 2018 гг. С помощью ML было обнаружено свыше 160 суточных разливов на WWTP1, из которых 105 были подтверждены предупредительной сигнализацией STO или COL. Подобным образом, 200 суточных разливов было обнаружено на WWTP2. Имеется много примеров непрерывных суточных разливов на протяжении более чем 10 дней.

На WWTP2 с помощью классификатора ML был обнаружен примечательный почти непрерывный разлив в течение 60 дней в период с 21 декабря 2013 г. по 22 февраля 2014 г. Представители общественности информировали EA о существенных грибковых образованиях в принимающих водотоках (27 января 2013 г. и 3 февраля 2014 г.) перед тем, как EA посетило очистные сооружения 6 февраля 2014 г. для проведения исследований. В докладе EA о подтверждении соответствия был сделан вывод, что “имели место обширные грибковые образования на протяжении 1,5 км водотоков с соответствующим негативным воздействием на водную экосистему. Наши специалисты в области рыболовства и биоразнообразия выразили глубокую обеспокоенность воздействием, которое было классифицировано как происходящий инцидент категории 2¹¹”.

Не было предъявлено обвинения. На протяжении более чем 20 дней в течение 60-ти дневного разлива уровень атмосферных осадков был ниже 2 мм. Подобные серии непрерывных суточных разливов были обнаружены с помощью анализа ML в 2012 г. (14 дней), 2013 г. (16 дней, 8 дней), 2015-2016 г. (17 дней). В каждом из этих разливов были серии из 2 или более дней без осадков.

2018 – 2020 гг. По запросу EIR было установлено, что в 2019 г. на WWTP1 разливы происходили на протяжении более чем 1000 ч в течение 72 дней (в среднем: разлив в течение 15 ч в сутки), в том числе 21 разлив, обнаруженный с помощью ML с непрерывной серией из 2 – 11 дней; подобным образом, на WWTP2 разливы происходили на протяжении более чем 1390 ч в течение 76 дней (в среднем: разлив в течение 18,3 ч в сутки), в том числе 32 разлива, обнаруженный с помощью ML с непрерывной серией из 2 – 14 дней. Почти непрерывный разлив течение ~30 дней происходил в ноябре 2019 г., в том числе 14 дней, в течение которых максимальный уровень осадков не превышал 2 мм (см. рис.4). В 2014 г. такой разлив привел к обширным грибковым образованиям, о которых представители общественности информировали EA (рис. 6).

Эти длительные разливы и рост грибковых образований включали периоды заурядных осадков. Наш анализ предполагает, что, по крайней мере, в течение 9 лет WWTP2, вероятно, были и продолжают оставаться субъектом просачивания грунтовых вод- движущей силы разливов сточных вод, которое EA считает не разрешенным.

¹¹ На основе правил отчетности Национальной системы регистрации инцидентов (NIRS2) Агентство по окружающей среде Англии подразделяет экологические инциденты (преимущественно связанные с загрязнением) на категории 1 (серьезные) и 2 (значительные).

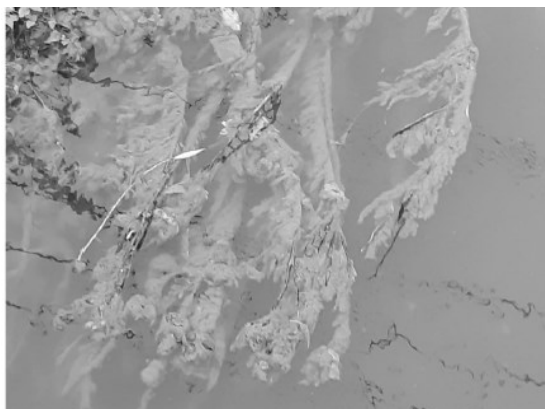


Рис. 6. Фотография грибковых организмов в сточных водах. Грибковые организмы образовались в результате 30-ти дневного разлива неочищенных сточных вод на WWTP2 в ноябре 2019 г.

Обсуждение

В Англии с Уэльсом имеется ~17000 систем общесплавной канализации, и у всех у них возможен сброс неочищенных сточных вод в реки или прибрежные воды. На WWTPs такие сбросы происходят в силу разных причин: чрезмерный приток из-за атмосферных осадков или инфильтрации, неадекватной пропускной способности, неисправного оборудования, неудовлетворительного обслуживания и иногда вследствие преднамеренного уклонения от очистки [2, 17 – 19]. Несмотря на продвинутое экологическое законодательство и требования по оценке воздействия на окружающую среду, остаются пробелы в наших знаниях о частоте, объеме и последствиях загрязнения от разливов неочищенных сточных вод. Самоотчетность водохозяйственного сектора о случаях загрязнения сточными водами как часть проверки соответствия с требованиями EA и ежегодного обзора эффективности работы, требуемого OFWAT (см. введение), например, Thames Water¹² [20]. EA признает факт предоставления неполной отчетности: “Показатели работы компаний в секторе водоснабжения и канализации: Ежегодный отчет” [1]. Перед применением подхода машинного обучения единственное представление о не сообщенных “разливах” можно было получить из сообщений общественности, которые охватывали 38% “разливов” сточных вод в 2018 г., например, в регионе, обслуживаемом Anglian Water¹³ [9]. В данной статье авторы сообщают о 926 предполагаемых “разливах”, выявленных с помощью ML только в двух из 3817 WWTP в

¹² Точное название *Thames Water Utility Limited* – британская компания, оказывающая услуги по водоснабжению, водоотведению и очистке сточных вод в Лондоне и соседних графствах Южной Англии, основанная в 1989 г., со штаб-квартирой в г. Реддинг, церемониальное графство Беркшир.

¹³ Водохозяйственная компания, работающая на Востоке Англии, основанная в 1989 г., со штаб-квартирой в г. Хантингдон, графство Кембриджшир.

Англии [1]. В зависимости от характеристик принимающего водного объекта и погоды это могут быть сбросы с очень большим воздействием. Оставалось определить, смогут ли эти “непонятные” выбросы (т.е. ранее неизвестные) помочь объяснить, почему 80% поверхностных водоемов в Англии имеют статус непригодный, плохой или удовлетворительный по классификации Рамочной директивы по воде [21].

Мы продемонстрировали, что подход машинного обучения, разработанный в этой работе, может ретроспективно помочь в определении “разливов” неочищенных сточных вод. Это может помочь водохозяйственному сектору в идентификации объектов, которые нуждаются в лучшем управлении, помочь регулирующим органам улучшить проверку соответствия нормативам и содействовать общественности в контроле над работой WWTPs. Выгоды от быстрого, автоматизированного подхода машинного обучения к отчетности о “разливах” расширятся до управляющих водосборными бассейнами, природоохранных групп, групп с особыми интересами (например, общества рыболовов), любителей прогулок и клубов (например, по гонкам на каяках и плавания на открытой воде) и консультантов и научных сотрудников, занимающихся моделированием, оценкой качества воды и охраной здоровья дикой природы. Отсутствие подробных данных о частоте и длительности “разливов” может оказать воздействие на общество путем регулирования новостроек и получения необходимых разрешений на планирование. Новое жилищное и коммерческое строительство может оказать значительное воздействие на потоки в канализационных сетях и WWTP. По существу, любое новое капитальное строительство в зоне обслуживания WWTP, которые уже недостаточно эффективны, будет еще больше осложнять ситуацию с частотой и длительностью “разливов” и будет минимизировать прогресс, достигаемый с помощью других инициатив по управлению водосборными бассейнами, например, по восстановлению среды обитания.

Потенциальная неадекватность подхода машинного обучения связана с тем, что он зависит от получения большого количества точных данных. Применение устройств EDM на WWTP началось относительно недавно, и в соответствии с мнением эксплуатирующих WWTPs, рассмотренных в этом исследовании, требуется, по крайней мере, одна переустановка на всех WWTPs, на которых проводится опрос. Этот перезапуск нарушает работу и ограничивает сбор данных с последствиями для ML и подтверждающих анализов. С точки зрения повторного использования EDM, мы провели перекрестное тестирование классификаторов характеристик потока из одних WWTP, влияющих на ежесуточные характеристики потока других WWTP после нормализации для корректировки разности величин. Результаты классификации не были убедительными, и, поэтому, в будущем может случиться так, что подход, принятый в этом исследовании, необходимо будет индивидуально настроить под индивидуальные WWTP или, возможно, класс WWTP с близкими характеристиками, которые должны быть идентифицированы.

Подход ML, представленный здесь, можно будет применять только после продолжительного периода времени рассмотрения в рамках регламента об экологической информации (EIR). В Англии нет стандартного протокола запроса по EIR к водопроводно-канализационной компании для получения данных о потоке (расходе), EDM и показателях аварийной сигнализации. Требуется письменный запрос к одной водопроводно-канализационной компании, в то время как от других компаний требуется запрос об оказании поддержки по электронной почте с последующей электронной передачей или

загрузкой данных. Нормативный период по умолчанию для ответа на запрос EIR в Англии или Уэльсе составляет 20 рабочих дней, но на практике он бывает в 2 раза больше. 2 водопроводно-канализационные компании Великобритании инициировали открытый доступ на облачной основе к данным об умеренном расходе потока и аварийной сигнализации EDM для своих WWTPs и канализационных насосных станций (<https://marketplace.wessexwater.co.uk/dataset>; <https://southernwater.co.uk/our-performance/flow-and-spill-reporting>).

Государственный реестр разрешений EA, относящихся к сбросам в водоемы представляет минимальные итоговые данные в режиме онлайн и признает запросы с анонимной, непомеченной электронной почтой, что усложняет последующий мониторинг исполнения. Разрешение в формате PDF (межплатформенный открытый формат электронных документов) обычно поступает в течение десяти рабочих дней, но после получения может появиться необходимость во внесении общих поправок, относящихся к нескольким сотням участкам, без подробностей, связанных с интересующим участком. Внимательное рассмотрение разрешений в режиме онлайн перед немедленным скачиванием.

Разрешение EA позволяет делать выбросы из ливневого резервуара неочищенных сточных вод в водоемы с условием, чтобы был обеспечена очистка минимального расхода даже тогда, когда ливневый резервуар заполнен, или происходит перелив. Однако в разрешениях не требуется, чтобы поток, направляемый на очистку, измерялся, регистрировался, или подтверждался. В запросе в рамках EIR к эксплуатирующим WWTPs организациям, установлено, что поток, направляемый на очистку, как правило, регистрируется, как “не требующий разрешения”. Поэтому едва ли это условие разрешения можно легко проверить за исключением случая использования потока очищенных сточных вод в качестве заменителя.

Насколько известно авторам, не было аналогичного исследования с применением машинного обучения к потоку очищаемых сточных вод в сочетании с данными об осадках и телеметрическими данными аварийной сигнализации на протяжении длительных временных серий. С использованием подтвержденных EDM данных и данных о характеристиках потока очищенных сточных вод для двух заметно отличающихся WWTPs, авторы этой работы применили стандартные алгоритмы машинного обеспечения для разработки классификаторов, которые исключительно хорошо приспособлены к идентификации разливов, ранее определенных с помощью устройств EDM и сообщенных в рамках обязательств в разрешении эксплуатирующей WWTP организацией. Их применение к 10 дополнительным годам ежесуточных характеристик потока отличается от тренировочных данных, идентифицировавших 926 потенциальных дней “разливов”. Анализ с помощью ML позволил получить представление о “ранних” сбросах из ливневого резервуара, не соответствующих условиям разрешения, в период с 2009 по 2018 гг., в результате которого выявлено отсутствие таковых на WWTP2, но сотни выбросов на WWTP1. Анализ с помощью ML позволил определить много разливов на WWTP2 в течение периодов заурядных дождевых осадков, и это дает основания полагать, что просачивание сточных вод усугубляет ситуацию с разливами неочищенных сточных вод, по крайней мере, в течение 9 лет. В 2012 г. Европейская Комиссия установила, что Великобритания не выполняла свои обязательства в рамках Директивы о городских сточных водах 91/271/ЕЕС, и что сбросы неочищенных сточных вод разрешены только в исключительных обстоятельствах. Кроме того, в результате нашего анализа появились основания полагать, что в период с

2009 по 2018 гг. в нижнем течении после обеих WWTP реки могли принимать более 360 разливов неочищенных сточных вод длительностью в сутки, часто в виде обширных непрерывных разливов в сериях более чем в 10 дней.

Вероятная корреляция между дождевыми осадками и случаями разливов предполагает ее включение в исследование ML. Однако имеется общее согласие в отношении того, что инфильтрация грунтовых вод содействует росту поступающих на WWTP сточных вод, и последующим разливам неочищенных сточных вод. Рост уровня грунтовых вод может быть неожиданно задержан после дождя¹⁴, в связи с чем некоторые разливы не могут быть предсказаны только по местным дождевым осадкам или уровню воды в реке. Кроме того, WWTP2 в данном исследовании рассматриваются как принимающие все сточные воды без появления разливов даже тогда, когда расход сточных вод превышает уровень ливнеприемника. В этом исследовании авторы сконцентрировали внимание на потоке сточных вод на очистку в WWTP и данных EDM, но планируют обратиться к использованию ML для более предсказуемого определения разливов неочищенных сточных вод с помощью параметров качества воды в реке, собранных с использованием многопараметрических зондов, установленных выше и ниже WWTP по течению.

С точки зрения передачи технологий, были получены ценные уроки в результате накопления и анализа данных о многих параметрах и переконфигурации соответствующего визуального отображения для поддержки интерпретации и представления результатов. Наш опыт и методология анализа могут быть использованы канализационно-очистными сооружениями. Мы надеемся, что результаты помогут в улучшении управления WWTP и контроле над выполнением требований, и в конечном итоге, будут содействовать уменьшению сброса неочищенных сточных вод в реки и прибрежные воды.

Методы

Данные, относящиеся к сбросам неочищенных сточных вод и очищенных стоков

Индивидуальные разрешения, контролирующие допустимые сбросы в водоемы, были получены из Государственного реестра Агентства по окружающей среде Англии с Уэльсом (EA) [22]. В общем, такие разрешения определяют:

1. Минимальный расход сточных вод, поступающих на очистку (PFF) перед разрешением на сброс ливнестока.

2. Емкость ливневого резервуара, используемого для хранения неочищенных сточных вод в течение обильных дождей и/или таяния снега.

3. Разрешенный сброс неочищенных сточных вод в ливневые резервуары вследствие “обильных дождей и/или таяния снега”.

4. Требование, чтобы оборудование для мониторинга длительности события (EDM) регистрировало разливы неочищенных сточных вод.

5. Стандарты качества для очищенных сточных вод.

6. Частота отчетности перед EA эксплуатирующей WWTP организации о качестве очищенных сточных вод с данными о EDM и очищенном потоке.

¹⁴ Этот отрезок времени тем больше, чем меньше водопроницаемость пород и больше глубина залегания грунтовых вод.

Данные о неочищенных сточных водах и очищенных сточных водах

В рамках регламента об экологической информации (EIRs) от компаний, управляющих WWTP1 и WWTP2, мы получили “Мониторинг выбросов и сбросов в воздух, на почву и водные объекты по схеме сертификации” (MCERTS) с данными о расходе очищенных сточных вод за 2009 – 2020 гг. Данные были предоставлены в файлах as.csv (текстовый формат, предназначенный для представления табличных данных) с временными метками о среднем расходе в течение 15-ти минутных интервалов в л или куб. м в с. Очень редко отсутствовала вся или значительная часть ежесуточных расходов. Такие дни были исключены из анализа, но когда данные были полными, характеристика ежесуточного потока с 15-ти минутными интервалами содержали 96 значений. Общее количество значений потока, доступных для анализа, охватывающих оба WWTPs, содержало около 800 тыс., соответствующих более чем 800 ежесуточных характеристик потока. Выборка из таких данных о характеристиках потока с 15-ти минутными интервалами (MCERTS) представлена в дополнительной табл. 1.

Протокол анализа данных

1. Построение формы модели из ежесуточных характеристик поток за 2016 – 2020 гг. с данными EDM и без них.
2. Построение классификаторов с использованием обучения с учителем на характеристиках потока с данными EDM. Выбор оптимального классификатора и полуслепая проверка на характеристиках потока за 2009- 2015 гг.
3. Ретроспективное тестирование оптимального классификатора в полностью слепом режиме на характеристиках потока за 2009- 2015 гг.

Осадки и уровни воды в реке

Обильные дождевые осадки и таяние снега может оказывать пагубное воздействие на очистку сточных вод за счет стока поверхностных вод, приводящего к перегрузке на входе в сточных вод выше определенного уровня переполнения с отводом в ливнеприемник, при переполнении которого разрешен слив в подключенный водоем. Нормативы Агентства по окружающей среде Великобритании устанавливает, что *“в ливнеприемнике должны осаждаться твердые частицы, и он должен иметь минимальную емкость 68 л/душу обслуживаемого населения в ливнеприемнике или эквивалент хранения в течение 2 ч с максимальным расходом в ливнеприемниках”* [23]. В течение некоторых разливов ливнестоков реки при сильном ливне могут еще больше разбавлять объединенные сбросы неочищенных сточных вод и стоки поверхностных сточных вод. В ожидании проверки классификации с помощью ML ежесуточных потоков как включенные разливы сточных вод мы получили данные о среднесуточных осадках, речных течениях и/или уровнях воды в реке, при их наличии, с помощью общедоступных источников (www.accuweather.com; <https://nrfa.ceh.ac.uk/>; <https://riverlevels.uk/>). Больше подробностей представлено в дополнительной информации.

Телеметрическая связь между WWTPs и Центром контроля и управления операциями по обращению с отходами

Для идентификации сбросов неочищенных сточных вод в водоемы были подготовлены регламенты об экологической информации для канализационно-очистных компаний в виде регистрации данных EDM и обмена телеметрической информацией об аварийной сигнализации между каждой WWTP и центром контроля и управления операциями по обращению с отходами компании (WOOC). Такая информация представляется в файлах as.csv, с регистрацией времени/дат неочищенных сбросов; идентификацией WWTP; указанием времени/дат аварийных сообщений; серьезности аварийных сигналов (отражающей внутренние стандарты компании для соответствующих уровней минимального реагирования и времени вмешательства); источника сообщения (например, соответствующее оборудование, устройство, объект); состояния или изменения состояния соответствующего устройства.

Особенно важными являются аварийные сигналы для ливнеприемников и устройств EDM, установленных на дождеприемнике. Аварийная сигнализация разрешенного уровня перелива (COL) отражает уровень неочищенных сточных вод, отводимых в дождеприемник, для которого EDM определяет интервалы слива в принимающий водоем. Как и следовало ожидать, COL и EDM тесно коррелируют, и так как устройство COL установлено перед EDM на обеих WWTP – это хороший символ-заменитель для использования, когда нет EDM для подтверждения потенциальных дней “разлива”, определенных с помощью классификации ML. Примеры для иллюстрации таких данных представлены в дополнительной табл. 2.

Анализ формы потока

Методы анализа формы и соответствующее программное обеспечение, использованное в этом исследовании, были разработаны за последние 20 лет и активно применялись для определения различий формы трехмерных поверхностей, в особенности тех, которые представляют анатомические структуры. Поэтому вместо разработки программного обеспечения каждая “кривая” ежесуточного потока из значений 96 15-ти минутных интервалов преобразовывалась в тонкую “ленту” потока, содержащую 190 граней треугольных призм, помеченных (автоматически) 192 маркировочными знаками (дополнительный рис. 3), для того, чтобы было соответствие между ежесуточными характеристиками потока, для которых вычислена форма потока.

Плотная модель поверхности (DSM) из множества маркированных поверхностей, подробно описанная в ряде работ [24, 25], состоит из изменения видов формы, происходящего в результате применения метода главных компонент (PCA) (см. сноску 8) для определения различий позиций точек поверхности из усредненных поверхностей в наборе данных. Перед применением PCA, используя полностью плоскую “базу” ленты потока, достигли полного соответствия точек поверхности на всех лентах потока без ручного вмешательства. В течение создания модели формы режимы PCA рассчитывались с точки зрения уменьшения разницы охвата (определяемого как отношение характеристического значения, соответствующего режиму PCA к сумме всех характеристических значений диагонализированной ковариаци-

онной матрицы¹⁵). Достаточные виды в режиме плотной модели поверхности сохранились для охвата 99% изменения формы. Данные расхода очищенных сточных вод, запрошенные для каждой WWTP, для ежесуточных характеристик потока с 2016 по 2020 гг., использовали для построения отдельных плотных моделей поверхности формы потока.

Обучение с учителем для идентификации ежесуточных характеристик потока, связанных с разливами сточных вод

Что касается анализа формы, описанного выше, то программное обеспечение, использованное в этой статье, связано с введением способа обучения с учителем для создания классификаторов для различения формы, и оно было использовано в широком наборе применений в нейролицевой хирургии: изменение формы лица при генетическом синдроме [26]; преждевременное сращивание черепно-лицевых костей [27]; тканевая инженерия¹⁶ челюстно-лицевой формы [28]; лицевая асимметрия, связанная с эпилепсией [29]; развитие раковых заболеваний в раннем детском возрасте [30]; и корреляция изменения формы лица и черепа вследствие нарушения развития плода в результате алкогольного синдрома при беременности [31];

Организация, эксплуатирующая WWTP, сообщила, что в течение начального периода установки EDM результаты были ненадежными, и оба устройства были повторно введены в эксплуатацию в ноябре-декабре 2018 г. По этой причине результаты EDM для обеих WWTP, полученные раньше повторного ввода в эксплуатацию, были исключены из анализа ML.

При обучении с учителем использовали классический способ ML – метод опорных векторов (SVM) (см. сноску 8). Нейронная сеть на основе SVM, или подход классификатора с широкой разрешающей способностью сконцентрирован на отдельных случаях с перекрытием подгрупп для классификации, которая помогает определить отдельную поверхность с самой большой разницей между подгруппами. Классификация на основе SVM использует радиальную базисную функцию¹⁷ с ядром и с 5 эвристическими процедурами для определения ширины разницы (алгоритм Хинтона¹⁸; медианное разделение; среднее расстояние; Jaakkola (алгоритм искусственного интеллекта Яаккола); Jaakkola-mean [последовательный вариационный алгоритм]), каждая с 4 параметрическими вариантами. Таким образом, 20 вариантов алгоритмов “машинного обучения” было использовано для построения классификаторов характеристик потока для каждого WWTP.

Для оценки точности каждого варианта и предотвращения чрезмерного обучения авторы разбитых на слоевые, случайно выбранные подмножества

¹⁵ Ковариационная матрица в теории вероятностей – это матрица, составленная из попарных ковариаций одного или двух случайных векторов.

¹⁶ Тканевая инженерия – подход к созданию имплантируемых тканей и органов, использующий фундаментальные структурно-функциональные взаимодействия в нормальных и патологически измененных тканях при создании биологических заместителей для восстановления или улучшения функционирования тканей.

¹⁷ Радиальная базисная функция создает прогностическую модель для одной или нескольких зависимых (целевых) переменных на основании значений переменных прогностических параметров.

¹⁸ Итеративный градиентный алгоритм, который используется с целью минимизации ошибки работы многослойного перцептрона (математической модели восприятия информации мозгом).

из характеристик потока с “разливом” и “нормальным” режимом, и после классификации оценили общую точность классификации как среднюю площадь под соответствующими графическими характеристиками принимающей эксплуатирующей организации из 20 групп. Для каждого варианта SVM конечная классификация считалась “разливом”, если нижнее значение доверительного интервала (CI) 95% было положительным. Были идентифицированы наиболее эффективные сочетания ядер SVM (медианное, Jaakkola и Jaakkola-mean для обоих очистных сооружений) и с 2 режимами PCA (2 для WWTP1 и 10 для WWTP2) (см. дополнительно табл. 3 и 4 с полными результатами). Для проверки тестирования был принят консервативный подход и определен оптимальный классификатор, который обозначал характеристики потока как “разлив” только в том случае, если три самых эффективных алгоритма классифицировали “разлив”.

Подтверждение классификации ML для прошлых характеристик потока до EDM с телеметрическими устройствами аварийной сигнализации

Обучение с учителем использовали для разработки классификатора по характеристикам потока, обозначенных как “разлив” и “нормальный”, с использованием данных по интервалам разлива, предоставленных организацией, эксплуатирующей WWTP, и обнаружением устройствами EDM с декабря 2018 по март 2020 г. При отсутствии обнаруженных с помощью EDM интервалов между разливами с января 2009 по начало 2018 г., по-видимому, нет “золотого стандарта”, с помощью которого проводится сравнение, и, следовательно, подтверждение классификации характеристик потока на основе ML. Однако оба WWTP были оснащены другими устройствами, которые можно использовать для подтверждения “прогнозов” ML. Например, незадолго до установки EDM (WWTP1: ноябрь 2017 г., WWTP2: февраль 2018 г.) оба WWTP были оснащены аналогом аварийной сигнализации разрешенного уровня перелива (COL), регистрировавшим уровень сточных вод в ливнеборниках, контролируемых EDM, и посылающим проверенные сообщения в диспетчерский центр. При запросе от системы EIR данных с систем аварийной сигнализации о “заполнении ливневого резервуара”, имеющихся на многих WWTPs, они поступают в этот центр. При отсутствии такой системы эксплуатирующая организация предоставляла данные с помощью системы аварийной сигнализации о переливе из ливневого резервуара (STO), используемой на WWTP1, на протяжении всего периода исследования, используя менее надежную технологию поплавковой сигнализации. Такая же технология используется и на WWTP2.

Определение отдельных и непрерывных серий из 24-х часовых разливов

Разливы сточных вод значительно различаются по длительности, но обычно характеризуются сглаживанием параметров потока вследствие отвода избыточного потока в ливневый резервуар или непосредственно в реку. Это снижает вариацию относительно среднего индивидуального потока, представляющего либо поток, направляемый на очистку, либо очищенные сточные воды. В результате характеристики потока, связанного с соответствующим ливневым стоком в течение 24 ч (или более), часто имеют низкое стандартное отклонение, и в среднем поток близкий, если не выше уровня

аварийного перелива, как определено в разрешении на сброс. 24-х часовые разливы со средним расходом намного меньше, чем уровень аварийного перелива, потенциально не соответствуют условиям разрешения на минимальный PFF (расход сточных вод, поступающих на очистку). Длительные непрерывные серии из 24-х часовых разливов препятствуют восстановлению от воздействия сточных вод, и наиболее вероятно, что в результате произойдет зарастание грибковыми организмами в сточных водах, что вредно как для рыбы, так и для макро беспозвоночных [32]. Кратковременные разливы, намного меньше, чем сутки, могут совпадать с периодами низкого расхода в типичных суточных характеристиках, и потенциально приводят к менее значительному загрязнению или, по крайней мере, имеется возможность восстановления. Однако даже короткие разливы могут оказаться чрезвычайно загрязняющими при первом смыве, если в ливнеприемнике содержатся осажденные твердые вещества предыдущих разливов, поскольку он не был своевременно опорожнен.

Библиография

1. Environment Agency. Water and sewerage companies in England: environmental performance report (Environment Agency, 2019). <https://www.gov.uk/government/publications/water-and-sewerage-companies-in-englandenvironmental-performance-report-2019>. Accessed 01 Feb 2021.

2. Ofwat. Notice of Ofwat's proposal to impose a penalty on Southern Water Services Limited (Ofwat, 2019). <https://www.ofwat.gov.uk/publication/notice-of-ofwatsproposal-to-impose-a-penalty-on-southern-water-services-limited/>. Accessed 01 Feb 2021.

3. Environment Agency, Defra & Gove, M. Thames Water fined £2m for “foreseeable and avoidable” pollution (Environment Agency 2019). <https://www.gov.uk/government/news/thames-water-fined-2m-for-foreseeable-and-avoidable-pollution>. Accessed 01 Feb 2021.

4. BBC News. Maidenhead sewage spills: Thames Water fined more than £600k (BBC News 2019). <https://www.bbc.co.uk/news/uk-england-berkshire-48948413>. Accessed 01 Feb 2021.

5. Environment Agency. Thames Water ordered to pay record £20 million for river pollution (Press release Environment Agency 2017). <https://www.gov.uk/government/news/thames-water-ordered-to-pay-record-20-million-for-riverpollution>. Accessed 01 Feb 2021.

6. Carrington, D. UK water companies receive record sewage leak fines (The Guardian 2016). <https://www.theguardian.com/environment/2016/may/12/ukwater-companies-receive-record-sewage-leak-fines>. Accessed 01 Feb 2021.

7. Meikle, J. & Carter, H. Untreated sewage and cyanide kill thousands of fish in river Trent (The Guardian 2009). Available at <https://www.theguardian.com/uk/2009/oct/07/pollution-river-trent-investigation>. Accessed 01 Feb 2021.

8. Environment Agency. Incidents and their classification: the Common Incident Classification Scheme (CICS) (Environment Agency 2017). <https://www.ofwat.gov.uk/wpcontent/uploads/2017/12/20171129-Incidents-and-their-classification-the-Common-Incident-Classification-Scheme-CICS-23.09.16.pdf>. Accessed 01 Feb 2021.

9. Environment Agency. Water and sewerage companies “Environmental Performance Assessment” (Environment Agency 2018). <https://www.gov.uk/government/publications/water-and-sewerage-companies-in-england-environmental-performance-report>. Accessed 01 Feb 2021.

10. European Parliament. Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC. (2003). <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:041:0026:0032:EN:PDF>. Accessed 01 Feb 2021.
11. UK Government. The Urban Waste Water Treatment (England and Wales) Regulations 1994. (SI 1994/2841). <https://www.legislation.gov.uk/uksi/1994/2841/contents/made>. Accessed 01 Feb 2021.
12. European Commission. Council Directive of 21 May 1991 concerning urban wastewater treatment (91/271/EEC) (1991) <https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31991L0271:EN:HTML>. Accessed 01 Feb 2021.
13. Sergot, M. J. et al. The British Nationality Act as a logic program. *Commun. ACM* 29, 370–386 (1986).
14. Mind Bridge: Ai Auditor. 100 Crosby Street, Suite 605, New York, NY, 10012, USA <https://www.mindbridge.ai/contact/>. Accessed 27 Feb 2021.
15. Browne, R. UK regulator looking to use AI, machine-learning to enforce financial compliance (CNBC 2017).
16. Hino, M., Benami, E. & Brooks, N. Machine learning for environmental monitoring. *Nat. Sustain.* 1, 583–588 (2018).
17. Gaines, M. SWW fined for sewage discharge near Truro (WWT 2017). <https://wwtonline.co.uk/news/south-west-water-fined-for-untreated-sewage-spill-neartruro>. Accessed 01 Feb 2021.
18. Miles, R. Thames Water admits polluting the Windrush. (Oxford Mail 2006). <https://www.oxfordmail.co.uk/news/1060793.thames-water-admits-polluting-windrush/> Accessed 01 Feb 2021.
19. WiredGov. Thames Water fined £380,000 for pollution of a rural stream (WiredGov 2016). <https://www.wiredgov.net/wg/news.nsf/articles/Thames+Water+fined+380000+for+pollution+of+a+rural+stream+09032016072000?open>. Accessed 01 Feb 2021.
20. Thames Water. Building a better future. Annual Report and Annual Performance Report 2018/19 (Thames Water, 2020). <https://www.thameswater.co.uk/medialibrary/home/about-us/investors/our-results/previous-reports/2018-19/full-report-annual-report-and-annual-performance-report-2018-19.pdf>. Accessed 01 Feb 2021.
21. Defra. Biodiversity 2020: A strategy for England’s wildlife and ecosystem service indicators – revised (Defra, 2011). <https://www.gov.uk/government/publications/biodiversity-2020-a-strategy-for-england-s-wildlife-and-ecosystem-services>. Accessed 01 Feb 2021.
22. Environment Agency. Environmental Permitting Regulations – Discharges to water and groundwater (Environment Agency, 2016; updated 2020). <https://www.gov.uk/guidance/discharges-to-surface-water-and-groundwater-environmentalpermits>. Accessed 01 Feb 2021.
23. Environment Agency. Guidance: Water companies: environmental permits for storm overflows and emergency overflows (Environment Agency, 2018). <https://www.gov.uk/government/publications/water-companies-environmental-permits-for-storm-overflows-and-emergency-overflows>. Accessed 01 Feb 2021.
24. Hutton, T. J., Buxton, B. F., Hammond, P. & Potts, H. W. W. Estimating average growth trajectories in shape-space using kernel smoothing. *IEEE Trans. Med. Imaging* 22, 747–753 (2003).
25. Hammond, P. The use of 3D face shape modelling in dysmorphology. *Arch. Dis. Child.* 92, 1120–1126 (2007).

26. Tassabehji, M. et al. GTF2IRD1 in craniofacial development of humans and mice *Science* 310, 1184–1187 (2005).
27. Twigg, S. R. F. et al. Reduced dosage of ERF causes complex cranio-synostosis in humans and mice and links ERK1/2 signaling to regulation of osteogenesis. *Nat. Genet.* 45, 308–313 (2013).
28. Ibrahim, A. et al. Combined soft and skeletal tissue modelling of normal and dysmorphic midface postnatal development. *J. Craniomaxillofac. Surg.* 44, 1777–1785 (2016).
29. Chinthapalli, K. et al. Atypical face shape and genomic structural variants in epilepsy. *Brain* 135, 3101–3114 (2012).
30. Hopman, S. M. J., Merks, J. H. M., Suttie, M., Hennekam, R. C. M. & Hammond, P. 3D morphometry aids facial analysis of individuals with a childhood cancer. *Am. J. Med. Genet. A* 170, 2905–2915 (2016).
31. Suttie, M. et al. Combined face-brain morphology and associated neurocognitive correlates in fetal alcohol spectrum disorders. *Alcohol. Clin. Exp. Res.* 42, 1769–1782 (2018).
32. Curtis, E. J. C. Sewage fungus: its nature and effects. *Water Res.* 3, 289–311 (1969).