Министерство Российской Федерации по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий

СОГЛАСОВАНО
Директор
Департамента оперативного управления
МЧС России
генерал-лейтенант
А.В. Елизаров
«18» 06 2020 г.

УТВЕРЖДАЮ
Заместитель Министра
Российской Федерации по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий
генерал-полковник
В.Н. Яцуценко
«22» 06 2020 г.
2-4-71-14-9

СОГЛАСОВАНО
Заместитель начальника
Главного управления «Национальный центр управления в кризисных ситуациях» МЧС России
полковник
С.А. Калугин
«18» 06 2020 г.

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО РЕАЛИЗАЦИИ АЛГОРИТМОВ ПОСТРОЕНИЯ МОДЕЛЕЙ ИЗМЕНЕНИЯ УРОВНЯ ПОДЪЕМА ПАВОДКОВЫХ ВОД, ВЫЗВАННЫХ ДОЖДЕВЫМИ ОСАДКАМИ, ДЛЯ РЕК С СЕВЕРО-КАВКАЗСКИМ ТИПОМ НАВОДНЕНИЙ (НА ПРИМЕРЕ РЕКИ ТУАПСЕ)

Москва 2020

СОДЕРЖАНИЕ

АННОТАЦИЯ

Настоящий документ содержит методические рекомендации по реализации алгоритмов построения моделей изменения уровня подъема паводковых вод, вызванных дождевыми осадками, для рек с Северо-Кавказским типом наводнений (на примере реки Туапсе).

ВВЕДЕНИЕ

Для достижения цели проведения исследовательской работы по моделированию уровня подъема паводковых вод, вызванных дождевыми осадками, для рек с Северо-Кавказским типом наводнений (на примере реки Туапсе) были выполнены следующие задачи:

проведен анализ существующего научно-методического аппарата по прогнозированию подъема уровня воды;

проведен анализ существующих утвержденных методик по оценке уровня подъема паводковых вод, вызванных дождевыми осадками;

проведен системный анализ процесса возникновения подъема уровня воды для рек с Северо-Кавказским типом наводнений (на примере реки Туапсе);

разработаны требования к исходным данным;

разработана структура методики оценки уровня подъема паводковых вод, вызванных дождевыми осадками для рек с Северо-Кавказским типом режимов наводнений (на примере реки Туапсе);

разработана математическая модель прогнозирования уровня подъема паводковых вод, вызванных дождевыми осадками для рек с Северо-Кавказским типом наводнений (на примере реки Туапсе);

разработаны алгоритмы построения моделей прогнозирования уровня подъема паводковых вод, вызванных дождевыми осадками для рек с Северо-Кавказским типом наводнений;

разработана методика оценки уровня подъема паводковых вод, вызванных дождевыми осадками для рек с Северо-Кавказским типом режимов наводнений (на примере реки Туапсе).

Разработка методики оценки уровня подъема паводковых вод, вызванных дождевыми осадками для рек с Северо-Кавказским типом режимов наводнений (на примере реки Туапсе)

1 Общая структура методического аппарата

В настоящих Методических рекомендациях проведен анализ существующих подходов к прогнозированию уровня подъема паводковых вод, вызванных дождевыми осадками. На основе рассмотренных методов разработана комплексная методика определения уровня подъема паводковых вод и оценки последствий паводков. Комплексная методика включает в себя методику прогнозирования непосредственно уровня подъема паводковых вод и методику определения последствий подъема уровня паводковых вод.

Общая структурно-функциональная схема разработанного научно-методического аппарата методики представлена на рисунке 1, а структурно-функциональная схема методики на рисунке 2.

Методика прогнозирования уровня подъема паводковых вод, вызванных дождевыми осадками включает в себя алгоритм обработки исходных данных и алгоритм нахождения параметров модели прогнозирования уровня подъема паводковых вод.

В качестве алгоритмов обработки исходных данных выступают основанные на применении методов корреляционного анализа, комбинаторики, а также информационно-технических методов. В результате работы алгоритма наборы статистических данных по зафиксированным уровням подъема паводковых вод и погодным условиям на метеостанциях проходят процедуру обработки, выделяются наиболее значимые погодные условия, влияющие на уровень подъема паводковых вод, исходные данные в обучающую выборку, пригодную для применения методов машинного обучения.

Алгоритм нахождения параметров модели прогнозирования уровня подъема паводковых вод позволяет получить математическую модель, основанную на методах машинного обучения (построение регрессионных

деревьев решений при помощи алгоритма Cart[1]). Данная модель строится на обучающей выборке и отражает закономерность уровня подъема паводковых вод от зафиксированных значений погодных условий.

Достоверность алгоритма обеспечивается апробацией модели на реальных значениях и определении статистической значимости модели.

Алгоритм оценки последствий паводков определяет зависимость ущерба (в работе понимается площадь затопленной территории и количество объектов, попавших в зону затопления).

В алгоритме используются следующие методы: анализ триангуляционной модели поверхности, дерева связей, бассейновый метод. Достоверность алгоритма обеспечивается применением верифицированных данных о рельефе местности и объектах строительства.

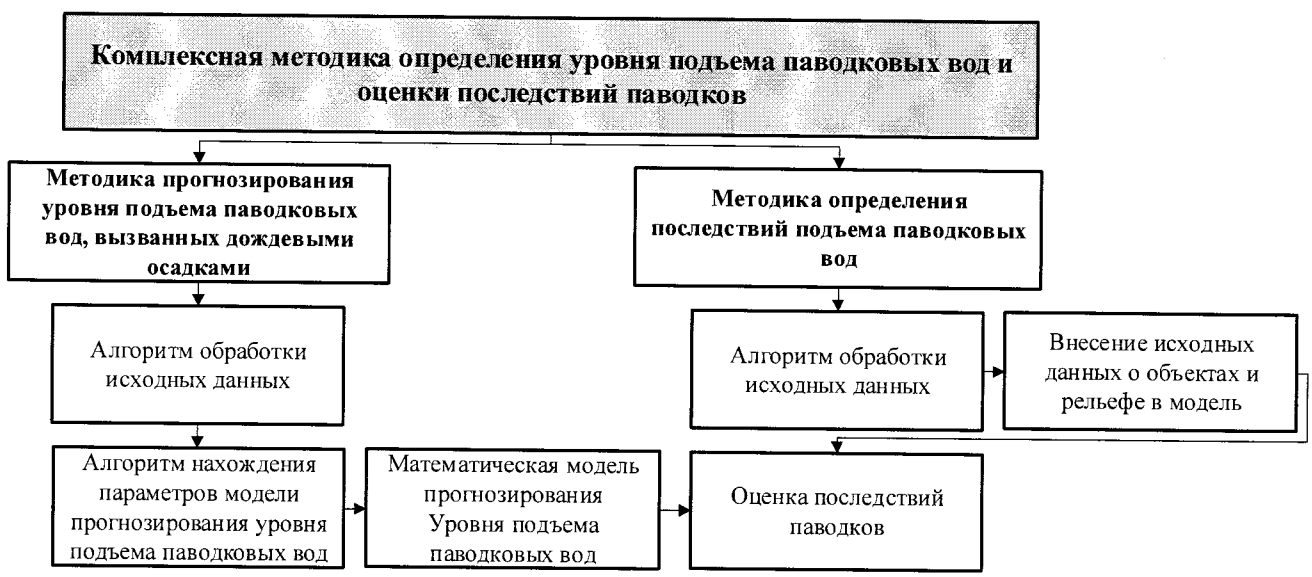

Рисунок 1 – Схема научно-методического аппарата методики определения уровня подъема паводковых вод и ущерба от действия паводка

[1] Справочно: Cart – метод классификации и регрессии

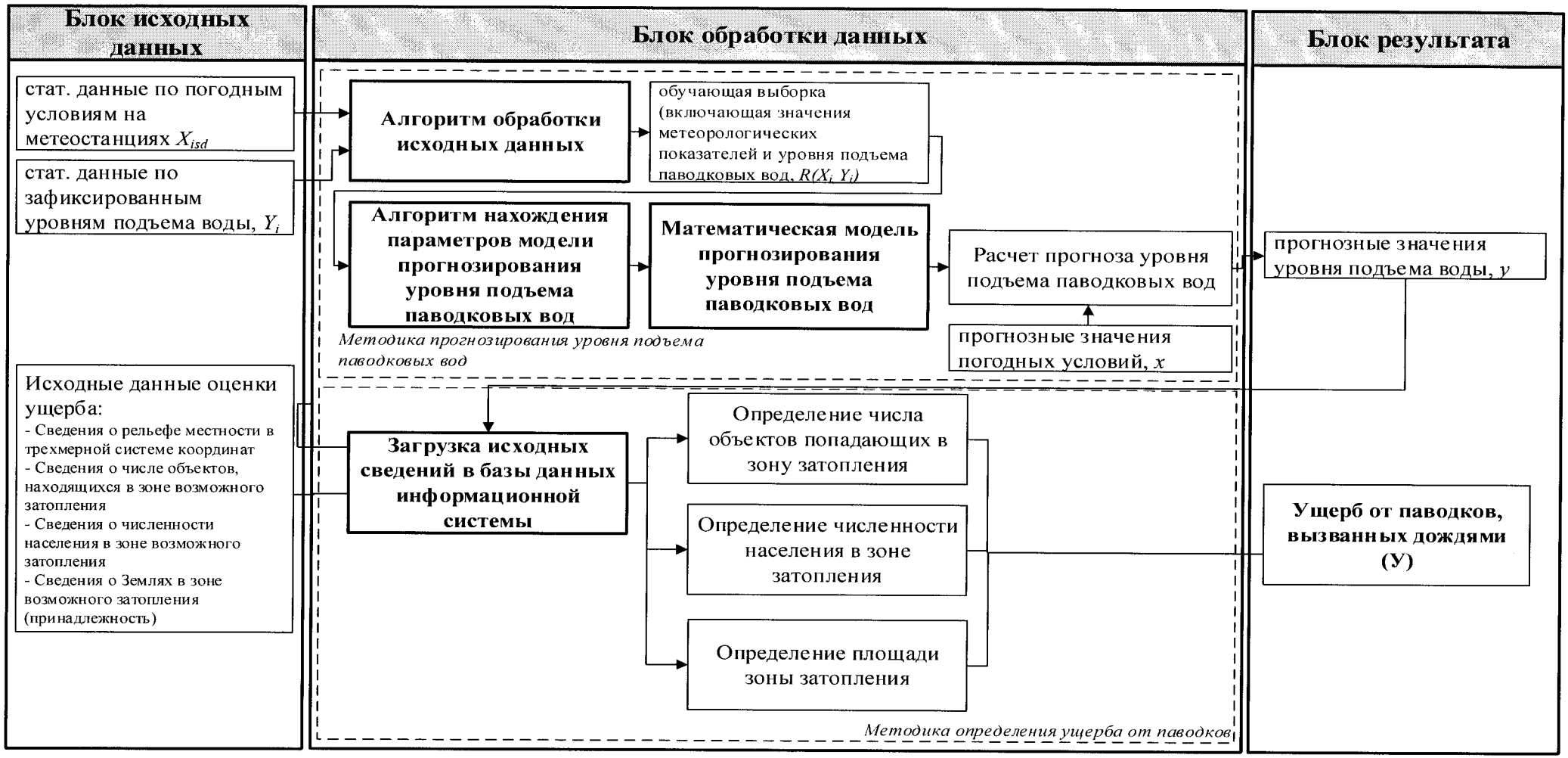

Рисунок 2 – Структурно-функциональная схема методики прогноза подъема уровня паводковых вод, вызванных дождевыми осадками и ущерба, вызванных паводками

Представленная структура методического аппарата позволит сформировать методику прогноза подъема уровня паводковых вод, вызванных дождевыми осадками и ущерба, вызванных паводками.

Исходя из предложенной структуры, методика будет состоять из трех основных алгоритмов, которые последовательно позволят решить следующие задачи:

построение модели прогноза подъема уровня паводковых вод;

проведение расчетов по построенной модели;

проведение расчетов по определению прогнозируемого ущерба, вызванного подъема уровня паводковых вод.

2 Требования к исходным данным

В качестве исходных данных при реализации предложенного методического аппарата выступают:

1. Сведения о гидрологических характеристиках рассматриваемого объекта исследования (реки Туапсе):

- сведения о уровне подъема воды (для обеспечения требуемого уровня достоверности разрабатываемой модели данные должны предоставляться за интервальные периоды в 3 часа, с точностью до 10 мм. Размер выборки для построения модели должен составлять не менее 10 лет непрерывных наблюдений по каждому рассматриваемому гидропосту);

- сведения о скорости течения (для обеспечения требуемого уровня достоверности разрабатываемой модели данные должны предоставляться за интервальные периоды в 3 часа, с точностью до 0,1 м/с. Размер выборки для построения модели должен составлять не менее 10 лет непрерывных наблюдений по каждому рассматриваемому гидропосту);

- сведения о суммарном потоке (для обеспечения требуемого уровня достоверности разрабатываемой модели данные должны предоставляться за интервальные периоды в 3 часа, с точностью до 0,1 $м^3/с$. Размер выборки для

построения модели должен составлять не менее 10 лет непрерывных наблюдений по каждому рассматриваемому гидропосту).

2. Сведения о метеорологических характеристиках в районе для которого осуществляется прогноз (район водосборного бассейна):

- сведения о интенсивности осадков (для обеспечения требуемого уровня достоверности разрабатываемой модели данные должны предоставляться за интервальные периоды в 1 час, с точностью до 1 мм/м^2. Размер выборки для построения модели должен составлять не менее 10 лет непрерывных наблюдений по каждому рассматриваемому метеопосту (метеостанции));

- сведения о интенсивности испарения (для обеспечения требуемого уровня достоверности разрабатываемой модели данные должны предоставляться за интервальные периоды в 1 час, с точностью до 0,1 мм/м^2. Размер выборки для построения модели должен составлять не менее 10 лет непрерывных наблюдений по каждому рассматриваемому метеопосту (метеостанции));

- сведения о температуре, (интенсивности солнечного излучения) (для обеспечения требуемого уровня достоверности разрабатываемой модели данные должны предоставляться за интервальные периоды в 1 час (для параметров солнечного излучения в пределах светового дня), с точностью до 0,1^0 С (Вт/м^2). Размер выборки для построения модели должен составлять не менее 10 лет непрерывных наблюдений по каждому рассматриваемому метеопосту (метеостанции));

- сведения о скорости ветра (для обеспечения требуемого уровня достоверности разрабатываемой модели данные должны предоставляться за интервальные периоды в 1 час, с точностью до 0,1 м/с. Размер выборки для построения модели должен составлять не менее 10 лет непрерывных наблюдений по каждому рассматриваемому метеопосту (метеостанции)).

3. Сведения о характеристиках, необходимых для расчета ущерба в районе для которого осуществляется прогноз (объекты населенных пунктов, находящихся в зоне возможного затопления и инфраструктуры):

- сведения о объектах (привязка (по координатам широты и долготы) объектов к применяемой картографической подложке информационной системы. Архитектурные решения, применяемые при создании объекта, такие как этажность, высота, основные несущие конструкции, материал и пр.) (сведения вносятся по каждому объекту, могут подгружаться автоматически из внешних баз данных);

- сведения о численности населения, находящегося в объектах, внесенных в базу данных (сведения вносятся по каждому объекту, могут подгружаться автоматически из внешних баз данных);

- сведения о Землях, находящихся в зоне возможного затопления (принадлежность к землям сельскохозяйственного пользования, лесного фонда и т.п.) (сведения вносятся по каждому объекту, могут подгружаться автоматически из внешних баз данных);

- сведения о рельефе местности в трехмерной системе координат (требования к точности зависят от предъявляемой точности прогноза, рекомендуется точность определения широты и долготы в 0,5 метра, по высоте 0,1 м) (сведения вносятся по каждому объекту, могут подгружаться автоматически из внешних баз данных).

Порядок работы с исходными данными и их обработки будет представлен ниже.

При этом работа с исходными данными при создании моделей сопряжена с рядом трудностей, которые необходимо учитывать при их обработке и формировании. Проведенный анализ позволил сформулировать следующие требования к исходным данным.

1. Данные обладают некоторой структурой, которую необходимо учитывать в математической модели или алгоритме. Поэтому очень тяжело

представить модель, которая будет описывать или обрабатывать неизвестные данные.

2. В данных (особенно во временных рядах) может быть корреляция (автокорреляция), что крайне важно учитывать при построении моделей и отборе признаков (предикторов, независимых переменных). Корреляция в независимых переменных может давать негативный эффект при построении модели. При отсутствии данных невозможно численно измерить корреляцию.

3. Некоторые явления обладают накопительным эффектом, некоторые происходят внезапно (мгновенно). Это необходимо учитывать при агрегации и отборе признаков модели. Соответственно, для изучаемого региона и построения модели необходимо определить, какие явления более характерны: явления с эффектом накопления или мгновенные.

4. Данные могут быть сильно "зашумлены", поэтому очень часто требуется предварительная обработка данных. Это также отдельное исследование. Очень часто бывает так, что нельзя просто взять и подставить данные в модель. Их предварительно приходится как-то обрабатывать, например логарифмировать и т.д. В любом случае "зашумленность" данных нуждается в проверке, причем четкого универсального алгоритма проверки в настоящее время нет.

5. Очень часто строится ансамбль моделей, т.е. одна модель получает одну величину, вторая – другую величину, а их ансамбль способен получить третью (нужную). Очень трудно сказать, как будет устроен этот ансамбль и как на его основании делать выводы.

При условии соблюдения требований к входящей информации (и их адекватной обработке), построенные с применением статистических и физико-статистических методов должны давать удовлетворительную сходимость прогнозов.

Кроме того, требуют решения проблемы, представленные выше и касающиеся вопросов как требований к исходным данным, так и механизмов их обработки.

3 Математическая модель прогнозирования уровня подъема паводковых вод, вызванных дождевыми осадками для рек с Северо-Кавказским типом наводнений (на примере реки Туапсе)

3.1 Математические модели прогнозирования уровня воды с применением методов Data Mining

3.1.1 Математическая модель на основе регрессионного дерева решений

Среди множества существующих на данный момент методов Data Mining[2], деревья решений является одним из наиболее популярных методов решения задач классификации и регрессии. Деревья решений также называют деревьями решающих правил, деревьями принятия решений, деревьями классификации.

Деревья решений применяются во многих сферах, где имеется большая выборка статистических структурированных данных, содержащая данные наблюдений нескольких независимых факторов и соответствующее им значение ключевого (зависимого) – обучающая выборка. В контексте работы независимыми факторами являются данные значений метеорологических показателей за определенный период, а зависимым – соответствующее значение уровня подъема воды в реке.

Задача прогнозирования уровня воды на основе данных обучающей выборки решается в два этапа: построение модели (дерева решений) и ее использование.

Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение (рисунок 3).

[2] Справочно: Data Mining – интеллектуальный анализ данных

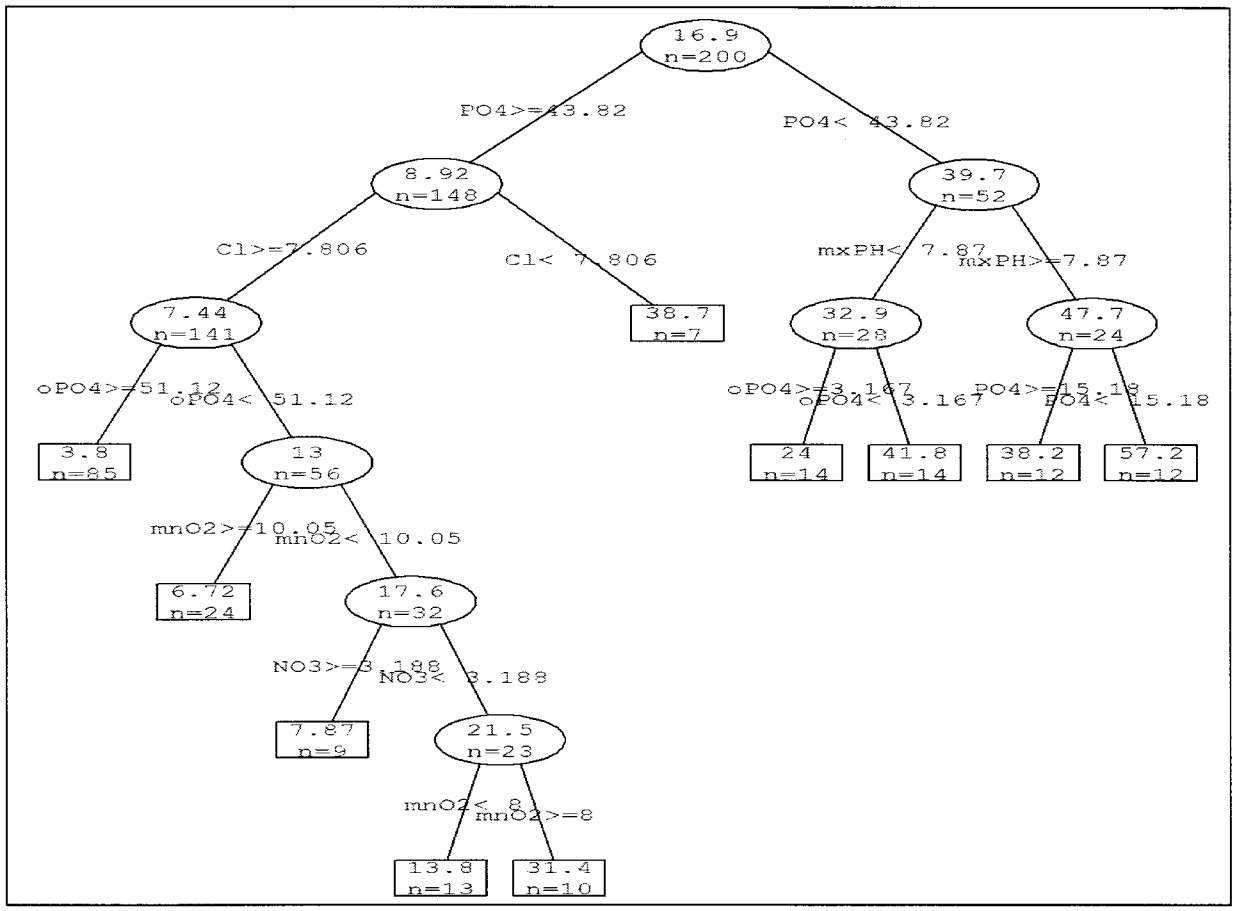

Рисунок 3 – Дерево решений (пример)

Каждое разбиение выборки называется узлом дерева. Все узлы дерева, кроме конечных, содержат в себе правило типа *(X>C)*, где *X* – наименование фактора (измерения метеорологического показателя), а *C* – значение этого фактора (при этом знак неравенства может быть любой). На каждом шаге построения дерева правило, сформированное в узле, делит входную выборку на две части: часть, для которой правило выполняется и часть, для которой правило не выполняются. Самый первый узел называется корнем. Каждый узел дерева в дальнейшем разбивается на 2 узла нижестоящего уровня, которые называются потомками. А, наоборот, узел, который объединяет двух потомков, называется родителем или родительским узлом. Конечные узлы, которые не имеют потомков называются листьями дерева. Листья содержат правила типа *(Y=D)*, где *Y* – целевой показатель (уровень подъема вода за заданный интервал времени), а *D* – значение этого показателя. Листья

определяют выходное (прогнозное) значение модели. Количество уровней дерева решений называется глубиной дерева.

Дерево решений можно представить в виде набора логических конструкций (правил) вида «если …, то…». Для этого необходимо указать логически каждый путь от корня к листу.

Формализованный вид дерева решений.

Дерево решений разбивает множество входных элементов R на J непересекающихся подмножеств $R_j, j = 1 \dots J$, где J – количество листьев в дереве, при этом $J < m$, то есть количество листьев меньше количества обучающих прецедентов. При этом каждой области R_j соответствует некоторая константная переменная φ_j, которая является ответом дерева для всех векторов, принадлежащих R_j. Таким образом, дерево решений это совокупность множеств векторов R_j и соответствующих им ответов φ_j. Формально, дерево решений может быть записано в следующем виде (1):

$$\Theta = \{(R_j, \varphi_j), j = 1 \dots J\}. \quad (1)$$

Функция ответа для дерева решений принимает вид (2):

$$T(w; \Theta) = \sum_{j=1}^{J} \varphi_j \cdot I(w \in R_j), \quad (1)$$

где Θ – набор параметров дерева решений;

w – вектор значений метеорологических факторов, не обязательно входящий в множество векторов W, использованных для построения дерева,

$I(w \in R_j)$ – индикаторная функция, такая, что $I(w \in R_j) = 1$, если вектор w принадлежит подмножеству R_j, и $I(w \in R_j) = 0$ в обратном случае.

Θ определяется в процессе настройки модели на основе данных R:

$$\Theta = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{m} E\big(Y_i, T(W_i; \Theta)\big), \tag{3}$$

где E – функция ошибки;
Y_i – уровень подъема воды в i-ый момент времени;
W_i – набор наблюдений погодных условий для i-го момента времени;
$T(W_i; \Theta)$ – функция ответа дерева решений (предсказание уровня подъема воды в i-ый момент времени), где $T(W_i; \Theta) = \varphi_j$, j является номером подмножества $R_j \in \Theta$, которому принадлежит вектор W_i.

Прогнозирование уровня подъема воды на участке реки при помощи рассмотренной модели включает несколько основных этапов:

1. Построение модели прогнозирования (обучение модели) на статистических данных.

2. Использование модели прогнозирования на прогнозных данных по погодным условиям.

3. Уточнение модели прогнозирования (перестроение модели).

Для обучения дерева решений используется *алгоритм CART* .

Алгоритм CART позволяет построить дерево решений, последовательно определяя правила, которые будут разделять строки обучающей выборки на две части по критерию максимального отличия целевого показателя – значения уровня подъема воды. Алгоритм состоит из 3-х основных этапов.

На 1 этапе происходит построение полного дерева решений на основе данных обучающей выборки. В качестве критерия разбиения применяется правило, которое определяет какое разбиение будет максимально точно относить левых потомков к одним значениям подъема уровня воды, а правых к другим на основе дисперсионной оценки. Такое правило проверяется для всех возможных разбиений и определяется наилучшее. Получаемое дерево дает наилучшее качество прогноза на обучающих данных.

На 2 этапе происходит отсечение дерева, то есть устранение излишней сложности дерева посредством обрезания некоторых ветвей дерева. Для отсечения формируется последовательность поддеревьев полного дерева. Из последовательности выбирается то поддерево, которое обеспечивает наименьшую ошибку на тестовой выборке. Это позволяет получить лучшее качество прогноза на данных, не входящих в обучающую выборку.

На 3 этапе проверяется достоверность модели прогнозирования и делается вывод о дальнейшем ее использовании.

3.1.2 Математическая модель на основе нейронной сети

Нейронная сеть – это математическая модель, построенная по принципу организации и функционирования биологических нейронных сетей – сетей нервных клеток живого организма.

Понятие нейронных сетей возникло при изучении процессов, протекающих в мозге при мышлении, и при попытке смоделировать эти процессы. Нейронная сеть представляет собой систему соединенных между собой простых процессоров. Они довольно просты, и каждый из них обрабатывает входящие сигналы и посылает их другим процессорам. Будучи соединенными в достаточно большую сеть с управляемым взаимодействием, такие локально простые процессоры вместе способны выполнять довольно сложные задачи.

Нейронные сети имеют возможность обучаться, в чем и заключается одно из главных их преимуществ перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. Это значит, что, в случае успешного обучения, сеть сможет вернуть верный результат на основании данных, которые отсутствовали в обучающем наборе данных.

На рисунке 4 представлена простая нейронная сеть. Нейроны в ней расположены по L уровням, на каждом из которых находятся I_k нейронов. Для наглядности каждый нейрон, кроме входных и выходных, представлен в виде двух узлов: один суммирует входящие сигналы, второй преобразует их.

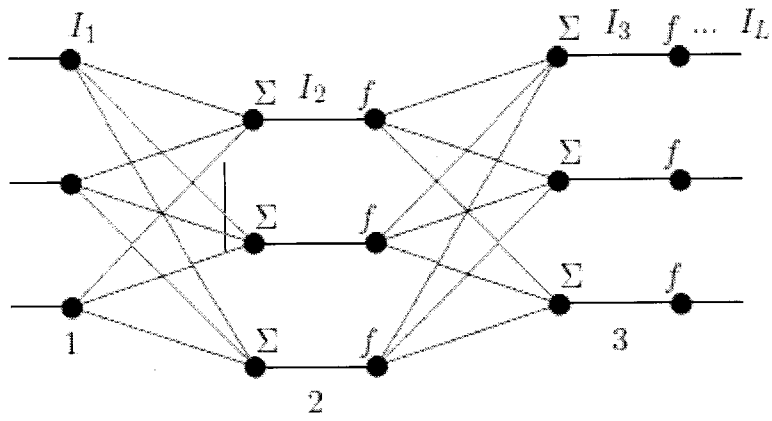

Рисунок 4 – Пример нейронной сети

Обучение нейронной сети – это обобщенный метод оптимизации многомерной функции, то есть нахождения минимума функционала (4):

$$\sum_{i=1}^{N} L\big(y_i, F(x_i)\big), \tag{42}$$

где $L(y_i, F(x_i))$ – это функция ошибки. В данном случае, в качестве функции ошибки используется сумма квадратов ошибки $L\big(y, F(x)\big) = (y - F(x))^2$.

Особенностью нейронной сети является то, что функция $F(x)$ задается не явно, а в виде структуры, состоящей из нейронов и связей между ними (иначе говоря, математических функций и последовательностей их применения).

Обучение нейронной сети происходит в три этапа:

1. Подача на входы сети обучающих данных.
2. Обратное распространения ошибки.
3. Корректировка весов.

Изначально, каждый входной нейрон x_i получает сигнал и транслирует его каждому из скрытых нейронов $z_1, z_2, \dots, z_p$. В свою очередь, каждый скрытый нейрон вычисляет результат его функции активации и рассылает свой сигнал z_j всем выходным нейронам. Каждый выходной нейрон y_k вычисляет результат своей активационной функции, который является выходным сигналом данного нейрона. В процессе обучения, каждый выходной нейрон вычисляет значение на выходе и сравнивает это значение с эталонным значением, которое берется из обучающего набора данных. Исходя из значения на выходе и эталонного значения определяется значение ошибки для данного входного набора данных. На основании ошибки вычисляются значения составляющих корректировки весов связей σ_k, которые используются при распространении ошибки от нейрона на выходе до всех элементов сети предыдущего слоя. После расчета всех σ_k происходит корректировка весов всех связей.

Для вычисления выходного сигнала нейрона используется функция активации, которая должна быть непрерывной, дифференцируемой и монотонно убывающей на всей области определения. Наиболее распространены и используются 4 основные функции активации:

1. Логистическая функция (5)

$$f_1(x) = \frac{1}{1 + e^{-x}}, \tag{5}$$

2. Функция гиперболического тангенса (6)

$$f_2(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{6}$$

3. Линейная функция (7)

$$f_3(x) = x, \tag{7}$$

4. Экспоненциальная функция (8)

$$f_4(x) = e^x. \tag{8}$$

3.2 Математическая модель прогнозирования уровня воды на основе метода регрессионного анализа

Предлагаемая методика краткосрочного прогнозирования стока рек Черноморского побережья Кавказа основана на модели его формирования, разработанной под руководством А.В. Христофорова в отделе речных гидрологических прогнозов ФГБУ «Гидрометцентр России». В этой модели используется опыт построения некоторых блоков модели Гидрометцентра СССР и других, более современных моделей, в частности, модели Сакраменто. При этом имеется ряд отличий, обусловленных следующими обстоятельствами:

1. Для каждого водосбора могут использоваться только один гидрометрический пост и одна метеостанция. В результате возможно только достаточно приблизительное описание процессов формирования речного стока, а результатом моделирования может быть не сама методика прогноза, а физическое обоснование некоторых полуэмпирических зависимостей прогнозируемой величины от ее предикторов.

2. Осреднение по пространственному распределению переменных, входящих в физически обоснованные формулы, и по случайному распределению во времени не учитываемых факторов формирования речного стока приводит к значительной трансформации этих формул. Сохраняются только линейные относительно входящих в них переменных формулы. Все иные формулы могут преобразовываться различными способами в зависимости от используемых распределений вероятностей, которые не имеют достаточного теоретического обоснования и имеют лишь частичное статистическое подтверждение. Следовательно, происходит достаточно произвольная аппроксимация физически обоснованных формул, описывающих процессы формирования речного стока.

3. Среди возможных аппроксимаций рядом преимуществ обладают формулы полиномиального вида. При полиномиальном виде зависимости

прогнозируемая величина выражается в виде линейной комбинации заранее заданных функций от используемых предикторов. Входящие в нее коэффициенты и свободный член являются параметрами формулы получения прогноза. Такой вид формулы получения прогноза существенно упрощает и делает более надежной статистическую оценку входящих в нее параметров, поиск оптимального ее варианта и определение погрешности прогноза.

4. Любая, в том числе полиномиальная аппроксимация прогностической зависимости, приводит к систематическим ошибкам, которые будут возрастать по мере увеличения диапазона изменения входящих в нее переменных и характеристик неучтенных факторов формирования речного стока. Уменьшить этот диапазон и, следовательно, повысить точность прогноза можно путем разбиения всего года на несколько расчетных периодов и оценки параметров прогностической зависимости для них в отдельности. При разработке предлагаемой методики в качестве такого периода с относительно постоянными условиями формирования речного стока рассматривались интервалы от недели до квартала.

5. С учетом сроков метеорологических и гидрологических наблюдений процесс формирования речного стока может описываться для расчетного интервала времени продолжительностью 12 часов. Для исследуемых речных бассейнов время формирования стока сопоставимо с интервалами между сроками гидрологических и метеорологических наблюдений и заблаговременностью прогноза. Статистический анализ данных наблюдений для всех шести водосборов показал целесообразность использования в качестве расчетного интервал в одни сутки, что вполне соответствует использованию данных гидрологических ежегодников.

6. Использование для всех рассматриваемых речных створов формулы общего вида для получения прогноза расхода воды с заблаговременностью одни сутки существенно облегчает разработку и оперативное использование автоматизированной системы прогнозирования стока рек Черноморского побережья Кавказа. Специфика условий формирования стока и водного

режима каждой реки, а также конкретного расчетного периода года могут учитываться при использовании достаточно большого числа параметров, входящих в прогностическую зависимость. Как и в большинстве других моделей формирования стока горных рек предполагается линейное уменьшение температуры воздуха и линейное или квадратичное увеличение осадков с высотой местности. Данное предположение означает, что благодаря малым размерам рассматриваемых водосборов для указанных метеорологических характеристик имеет место достаточно высокая корреляция между их колебаниями в разных точках водосбора. Это подтверждается тем, что для расположенных в бассейне р. Мзымта на разных высотах метеостанций Красная Поляна и Лаура коэффициент корреляции между среднесуточными температурами воздуха равен 0,99, а между значениями суточного слоя осадков 0,95.

Для метеостанций Красная Поляна и Ачишхо аналогичные коэффициенты равны 0,92 и 0,86.

Для расположенных в бассейне р. Туапсе метеостанций г. Туапсе и п. Горный аналогичные коэффициенты равны 0,98 и 0,77. На этом основании значения приземной температуры воздуха и количества атмосферных осадков на разных высотах описывались в виде линейной функции от значения той же метеорологической характеристики на метеостанции.

Для расчетного периода времени предполагалось постоянство вертикальных градиентов для каждого метеорологического фактора. Поступление дождевых вод на поверхность водосбора Поступление дождевых вод на поверхность речного бассейна происходит ниже высоты изотермы 2°С. С учетом линейного убывания температуры с высотой для метеостанции определялась минимальная температура $T_{min,P}$, при которой изотерма 2°С проходит через самую нижнюю точку водосбора и, следовательно, еще не происходит поступления дождевых вод на его поверхность. Высота изотермы 2°С является линейной функцией от температуры на метеостанции.

В прогнозах талого стока принято считать, что выпадающие на снег жидкие осадки приравниваются к поступлению воды непосредственно на поверхность бассейна за исключением той их части, которая может быть удержана снежным покровом. Удержание воды в снеге отдельно учитывалось при расчете поступления талых вод.

С ростом температуры изотерма 2°С поднимается, и увеличиваются площадь поступления дождевых вод. Одновременно увеличивается средняя высота части водосбора, на которой формируется дождевой сток и, как результат, растет среднее для этой части водосбора значение слоя осадков. Для описания этого двойного эффекта при аппроксимации зависимости слоя $h_P(t+1)$ от приземной температуры воздуха $T(t+1)$ используется квадратичная парабола.

С учетом данных обстоятельств, слой поступления дождевых вод $h_P(t+1)$ за ожидаемые сутки $t+1$ предлагается определять в зависимости от приземной температуры воздуха $T(t+1)$ и слоя осадков $P(t+1)$ за сутки $t+1$ на метеостанции с помощью формулы:

$$h_P(t+1) = P(t+1)\left[T(t+1) - T_{min,P}\right]\left[a_P T(t+1) + b_P\right], \quad (9)$$

где a_P и b_P – постоянные для каждого расчетного периода параметры модели.

Если на метеостанции температура меньше некоторого значения $T_{min,P}$, то на всем водосборе приземная температура ниже 2°С и, следовательно, поступление дождевых вод отсутствует. Следовательно, в формуле (9) значения $T(t+1) < T_{min,P}$ должны заменяться числом $T_{min,P}$. Влияние температуры $T(t+1)$ на слой поступления дождевых вод происходит до достижения ею некоторого максимума T_{max} , при котором практически на всей территории водосбора снежный покров отсутствует и может формироваться дождевой сток

Потери склонового стока складываются из потерь на поверхностное задержание, испарение и впитывание воды в почву. Последнее расходуется на восполнение запасов воды в почве, инфильтрацию воды в нижележащие слои грунта и аккумуляцию воды в прирусловой делювиальной осыпи. Эти потери косвенно учитываются ходом осадков и температуры воздуха. Изменение свойств почвенного, растительного и снежного покрова, процессы замерзания и оттаивания почвы учитываются путем оценки параметров модели для каждого расчетного периода времени. В виду отсутствия данных наблюдений, которые могли бы характеризовать пространственную изменчивость потерь склонового стока, использовались средние для всего водосбора характеристики, приближенно определяемые данными метеостанции и замыкающего створа.

В течение ожидаемых суток t+1 средний расход воды склонового стока q (t+1) складывается из расхода $q_{пов}$ (t+1) поверхностного стока, расхода $q_{поч}$ (t+1) внутрипочвенного стока и расхода $q_{гр}$ (t+1) грунтового стока. Последний не играет заметной роли в формировании максимального стока и косвенно учитывается предшествующей водностью реки. Все расчеты ведутся для конкретных суток расчетного периода, поэтому последующие водобалансовые уравнения необходимо относить не к среднесуточному расходу склонового стока и его составляющим, а к суточному слою соответствующего вида стока в конкретной точке водосбора. Это достигается умножением расхода q (t+1) на переходный коэффициент k , то есть k q (t+1) - слой склонового стока за сутки t.

Поверхностный склоновый сток формируется, если слой поступления воды на поверхность водосбора h (t+1) превышает слой воды i(t+1), просочившейся в почву за сутки t. В этом случае слой поверхностного стока k $q_{пов}$ (t+1) равен разности h (t+1) - i(t+1).

Если обозначить через W(t+1) запас почвенной влаги к концу суток t+1, то уравнение водного баланса почвенного слоя приобретает вид:

$$W(t+1) = W(t) + i(t+1) - kq_{\text{пов}}(t+1) - kq_{\text{гр}}(t+1). \quad (10)$$

Последовательно повторяя эту процедуру для суток t, …, t - m+1, получаем выражение для запаса почвенной влаги W(t) к началу суток t+1:

$$W(t) = W_0 + I_W(t+1), \quad (11)$$

где W_0 - запас почвенной влаги к началу первых учитываемых суток t-m.

Величина $I_w(t+1)$ определяет средний индекс увлажнения водосбора к началу суток t+1 и является важнейшей характеристикой потерь склонового стока:

$$I_W(t+1) = \sum_{j=1}^{m} [h(t-j+1) - kq(t-j+1)]. \quad (12)$$

Смысл предлагаемого индекса увлажнения вполне определяется разностью между поступлением воды на поверхность водосбора и суммарным склоновым стоком за предыдущие сутки, число m которых зависит от скорости процессов формирования стока.

Суммарный слой внутрипочвенного и грунтового склонового стока предлагается выражать через средний запас почвенной влаги в течение суток t в виде:

$$kq_{\text{поч}}(t) + kq_{\text{гр}}(t) = \beta \left[\frac{W(t-1) + W(t)}{2}\right], \quad (13)$$

где коэффициент β принимается постоянным для расчетного периода времени.

С учетом формулы (13) запас почвенной влаги W(t+1) к концу суток t+1 может быть выражен в виде:

$$W(t+1) = \left(\frac{2-\beta}{2+\beta}\right) W(t) + \left(\frac{2}{2+\beta}\right) i(t+1). \qquad (14)$$

Своего максимального значения i_{max} (t+1) слой поступления влаги в почву достигает в случае, когда к концу суток t запас почвенной влаги W(t) достигает максимальной влагоемкости почвы W_{max}. Из уравнений (10), (11) и (14) следует формула:

$$i_{max}(t+1) = \left(1+\frac{\beta}{2}\right) W_{max} - \left(1-\frac{\beta}{2}\right) [W_0 + I_W(t+1)]. \qquad (15)$$

При моделировании склонового стока необходимо выделить несколько вариантов его формирования. На покрытых луговой растительностью склонах быстро и практически без потерь формируется подвешенный поверхностный сток. На участках с толщами моренных отложений и на прилегающих к русловой сети участках также быстро и практически без потерь формируется подповерхностный сток.

В зависимости от известных на дату составления прогноза t и ожидаемых прогнозируемых на сутки t+1 значений суточного слоя осадков и среднесуточной температуры приземного слоя воздуха определяются величины h(t+1) и h(t). Параметры формул выражаются через параметры модели формирования стока и также принимаются постоянными в течение расчетного периода времени.

Выражение для ожидаемого среднесуточного расхода воды в замыкающем створе определяется следующей формулой:

$$Q(t+1) = d + d_1 Q(t) + d_2 Q(t-1) + d_3 h(t+1) + d_4 h(t). \qquad (16)$$

Ожидаемые значения суточных максимумов расхода воды в замыкающем створе можно получать, опираясь на достаточно тесную корреляцию между максимальными и среднесуточными расходами воды.

Построенная регрессионная модель позволяет осуществлять прогноз уровней подъема воды в зависимости от значений некоторого сочетания факторов. Предлагается осуществлять оценку по текущим, прогнозируемым и ретроспективным значениям факторов.

Дальнейшие работы по построению регрессионной модели сводятся к определению значений коэффициентов для конкретной реки с конкретными гидрологическими и метеорологическими условиями.

3.3 Математическая модель прогнозирования последствий подъема уровня воды на основе триангуляцинного метода

Предлагаемый метод основан на анализе триангуляционной модели поверхности, которую можно нестрого определить как триангуляцию, всем узлам которой поставлена в соответствие их высота (Z-координата). В качестве структуры данных для представления поверхности лучше всего использовать структуру «Узлы, простые рёбра и прямоугольники». В данной структуре каждый прямоугольник содержит ссылки на четыре образующих его узла, на проходящие через него структурные рёбра и на четыре соседних прямоугольника. Использование подобной структуры данных позволяет существенно увеличить скорость работы алгоритмов анализа триангуляционной модели поверхности, на которых основан предлагаемый алгоритм.

Алгоритм расчёта зон затопления:

1. Задается триангуляционная модель рассматриваемой поверхности;

2. Определяется объём выпавших осадков V, мм/м^2 (либо высота подъема уровня воды, полученная в предыдущих пунктах h, м).

Выходные данные: Список полигонов, соответствующих искомым зонам затопления с заданным объёмом воды.

Структура алгоритма выглядит следующим образом:

1. Осуществляется поиск всех «рёбер перелома» – рёбер триангуляции T, в которых экспозиция (направление) склона меняет своё значение на противоположное (рисунок 5);

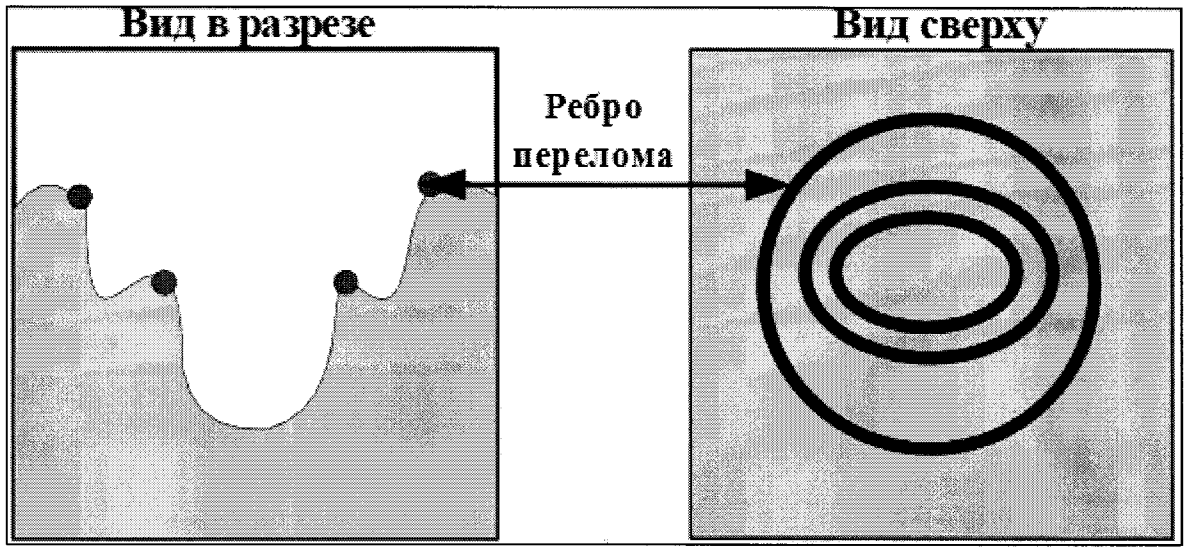

Рисунок 5 – Нахождение ребер перегиба

2. Для каждого найденного ребра перелома находится узел с наименьшей Z-координатой (h). Затем по уровню h строится изолиния исходной триангуляции – геометрическое место точек на поверхности, имеющих высоту h и имеющих в любой своей окрестности другие точки с меньшей высотой. Тем самым моделируется ситуация заполнения поверхности водой до уровня h. При этом контур искомой зоны затопления с максимальным уровнем воды будет соответствовать той части изолинии, которая проходит через ребро перелома.

Если построенная изолиния состоит из нескольких контуров, то необходимо выбросить из рассмотрения те контуры, которые не относятся к текущему ребру перелома. Таким образом, мы будем рассматривать только некоторый локальный участок поверхности, где предположительно может образоваться зона затопления. После этого необходимо проверить содержат

ли контуры изолинии внутри себя граничные узлы триангуляции с меньшей высотой. Если да, то это значит, что вода будет вытекать за границы поверхности. Следовательно, зоны затопления с уровнем h не будет (рисунок 6). В противном случае считаем, что зона затопления найдена и добавляем её в список зон затопления.

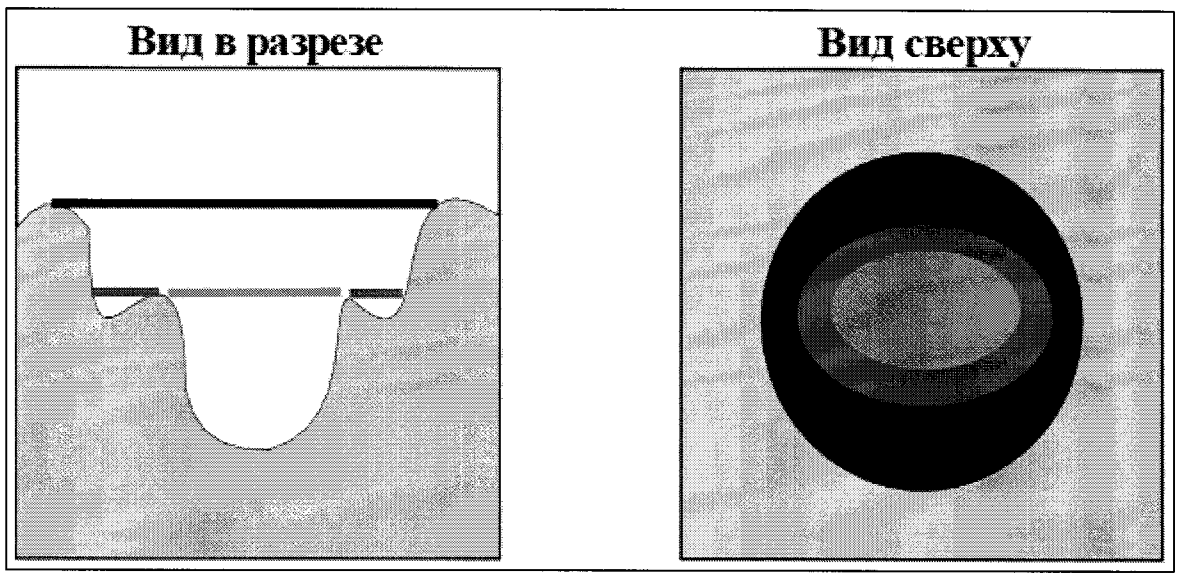

Рисунок 6 – Построение изолиний по ребрам перегиба

3. Строим дерево связей всех найденных зон затопления следующим образом. Если одна зона затопления полностью включает в себя другую (рисунок 7), то считаем, что зона затопления с большим контуром – это родитель, а с меньшим – потомок.

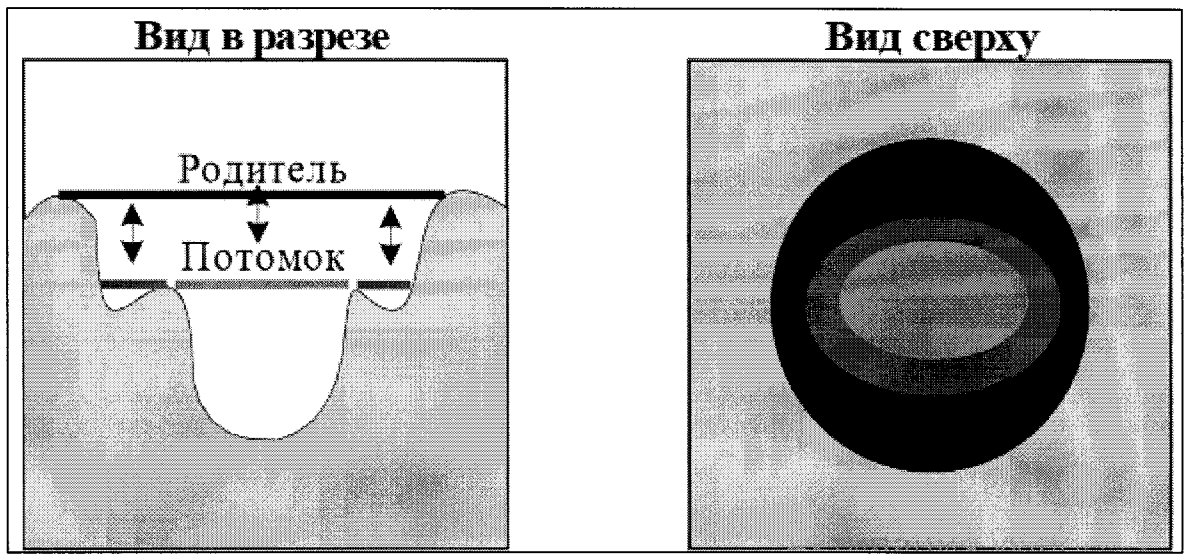

Рисунок 7 – установка связей между изолиниями затопления

4. По каждой зоне затопления строится зона водосбора – список прямоугольников, с которых вода будет стекать в соответствующую зону затопления.

5. Для каждой зоны водосбора рассчитывается объём воды, приходящийся на её площадь по формуле $V_i=S_iV$, где S_i – площадь соответствующей зоны водосбора.

6. Для каждой зоны затопления, начиная с листьев в дереве связей зон затопления, осуществляется проверка на переполнение воды. Если объём воды в зоне водосбора превышает максимальный объём зоны затопления, то осуществляется переход на пункт 7.

7. Перелив воды рассчитывается следующим образом. Если у проверяемой зоны водосбора есть родитель, то это значит, что перелива воды не будет, вода просто поднимется выше в родительскую зону затопления. Поэтому проверять на переполнение следует родительскую зону затопления. Если же у проверяемой зоны затопления нет родителя, то находятся узлы триангуляции T, через которые проходит контур зоны затопления. Именно через эти узлы вода будет переливаться в другие зоны водосбора. Находим для каждого такого узла смежные зоны водосбора и добавляем туда объём перелившейся воды, пропорционально площади прямоугольника, моделирующего поток воды, движущийся в соответствующую зону водосбора. Здесь следует отметить, что часть воды так же будет выливаться за пределы поверхности.

Таким образом, данная процедура продолжается до тех пор, пока не будет осуществлён перелив воды для всех зон затопления. После этого, в результирующий список зон затопления добавляются контуры обработанных на данном шаге родительских зон затопления, соответствующих максимальному уровню заполнения.

Представленный подход не представляет сложности при его программной реализации. Единственным ограничением, препятствующим его широкому внедрению, является отсутствие точных данных по матрице

высот (ограничения зачастую обусловлены необходимостью соблюдения требований, к защите информации содержащей государственную тайну), а также необходимость проведения большого количества вычислительных операций. Количество операций будет прямо пропорционально требованиям к точности проводимых расчетов (чем меньше размеры сетки, тем выше точность, но тем большее число операций необходимо выполнить).

4 Алгоритмы построения моделей прогнозирования уровня подъема паводковых вод, вызванных дождевыми осадками для рек с Северо-Кавказским типом наводнений (на примере реки Туапсе)

4.1 Алгоритмы построения моделей прогнозирования уровня воды с использованием методов Data Mining (на примере реки Туапсе)

4.1.1 Постановка задачи

Пусть задано n метеорологических факторов $X(X_1, X_2, \dots, X_n)$. Пусть X_k – k-й фактор, $k = \overline{(1, n)}$. Пусть имеется m дней. Каждый фактор X_k имеет m наблюдений $X_k(X_{k,1}, X_{k,2}, \dots, X_{k,m})$. Пусть $X_{k,i}$ – наблюдение k фактора в i-ый день, $i = \overline{(1, m)}$ на конкретной метеостанции.

Пусть имеется также m наблюдений уровня подъема воды в реке на конкретном гидропосту в рамках заданной территории $Y(Y_1, Y_2, \dots, Y_m)$. Пусть Y_i – наблюдение уровня подъема воды в i-ый день. Каждому вектору наблюдений погодных условий в i-ый день $W_i(X_{1,i}, X_{2,i}, \dots, X_{n,i})$. поставлено в соответствие значение Y_i (причем таких соответствий m). Пусть $W = \{W_i\}$, $i = \overline{(1, m)}$ – множество всех векторов погодных условий, $Y = \{Y_i\}$, $i = \overline{(1, m)}$ – множество всех наблюдений подъема уровня воды.

Пусть W_i – вход, Y_i – выход, $i = \overline{(1, m)}$. Упорядоченную пару (W_i, Y_i) назовем прецедентом. Множество m прецедентов называется выборкой данных (17):

$$R = \{(W_i, Y_i)\}, \text{ где } i = \overline{(1, m)}. \quad (17)$$

Требуется, основываясь на данных выборки (17) восстановить зависимость между входом и выходом. Иными словами, задача состоит в построении функции

$$T(w) = y, \tag{18}$$

где w – прогнозные значения метеорологических факторов;
y – прогноз подъема уровня воды, выданный моделью;
T – функция, получающая на вход вектор w и определяющая значение выхода y.

Такая задача называется задачей восстановления регрессии. Процесс нахождения функции T называется обучением модели, процесс определения выхода по некоторому входу с помощью уже построенной модели – предсказанием (прогнозированием).

4.1.2 Алгоритм построения модели на основе регрессионного дерева решений

1 Этап. Построение полного дерева решений.

При построении полного дерева решений используются только данные обучающей выборки. Построение полного дерева заключается в рекурсивном разбиении обучающей выборки наилучшим образом до тех пор, пока не будут достигнуты условия остановки. Результатом работы рекурсивного алгоритма разбиения является полное дерево Θ_{max}.

При этом могут быть заданы дополнительные условия остановки: максимальная глубина дерева и минимальное количество элементов, необходимое для разбиения.

1. Задается пустое дерево $\Theta_{max} = \emptyset$, обучающая выборка R.

2. Рассматриваются $k = \overline{(1,n)}$ факторов X. Значения каждого фактора X_k сортируются по возрастанию (19):

$$\acute{X_k} = \left(X_{k,j}, X_{k,j+1}, \dots, X_{k,m_{\text{выб}}}\right), \tag{19}$$

где $\acute{X}_{k,j}$ – j-ое по возрастанию значение X_k.

$m_{\text{выб}} = |R|$ – количество элементов в разделяемой выборке R, на первом шаге алгоритма $m_{\text{выб}}$ - количество элементов в обучающей выборке ($m_{\text{выб}} = m$), а R – вся обучающая выборка;

$j = \overline{(1, m_{\text{выб}})}$;

$\acute{X}_{k,j} \leq \acute{X}_{k,j+1}$.

3. Для каждого фактора X_k вычисляется $(m_{\text{выб}} - 1)$ возможных значений разбиений $S_{k,j}$, $j = \overline{(1, m_{\text{выб}} - 1)}$. Они вычисляются как среднее между 2-мя соседними значениями $\acute{X}_k$ (21):

$$S_{k,j} = \frac{\acute{X}_{k,j} + \acute{X}_{k,j+1}}{2}. \tag{20}$$

Каждому значению $S_{k,j}$, соответствует правило вида (21):

$$X_k \leq S_{k,j}, \tag{21}$$

Данное правило позволяет разделить выборку R на две части: R_1 и R_2. Правило $X_k \leq S_{k,j}$ и множества R_1 и R_2 определяют возможное разбиение выборки t.

4. Для каждого возможного разбиения подмножества R_1 и R_2 равны (22):

$$\begin{aligned} R_1 &= \{(W_j, Y_j)\}, \text{ где } j = \overline{(1, m_{\text{выб}})}, (W_j, Y_j) \in R, X_{k,j} \leq S_{k,j}, \\ R_2 &= \{(W_j, Y_j)\}, \text{ где } j = \overline{(1, m_{\text{выб}})}, (W_j, Y_j) \in R, X_{k,j} > S_{k,j}. \end{aligned} \tag{3}$$

R_1 содержит элементы, в которых $X_k \leq S_{k,j}$, R_2 содержит все остальные элементы. R_1 и R_2 не пересекаются.

5. Для каждого возможного разбиения вычисляется ошибка разбиения, которая определяет качество разбиения (23):

$$E_{k,j} = \sum_{i=1}^{|R_1|}(\bar{Y}_1 - Y_{1_i})^2 + \sum_{i=1}^{|R_2|}(\bar{Y}_2 - Y_{2_i})^2, \quad (23)$$

где $|R_1|$ - количество элементов в R_1;
$\bar{Y}_1$ – среднее значение уровня подъема воды в R_1;
Y_{1_i}– значение уровня подъема воды в i-ом элементе R_1;
$|R_2|$ - количество элементов в R_2;
$\bar{Y}_2$ – среднее значение уровня подъема воды в R_2;
Y_{2_i} – значение уровня подъема воды в i-ом элементе R_2.

6. Определяется разбиение $\hat{s}_{k,j}$ с минимальной ошибкой (24):

$$\hat{s}_{k,j} = \underset{s_{k,j}}{\operatorname{argmin}} E_{k,j} \forall\, k \in \overline{(1,n)}, j \in \overline{(1, m_{\text{выб}})}. \quad (24)$$

Правило $X_k \leq \hat{s}_{k,j}$ является искомым наилучшим разделением выборки, которое

разделяет R, на две ветки дерева: R_1 и R_2.

7. Проверяются условия остановки.

Если дисперсия значений уровня подъема воды в узле $D(R) = \sum_{i=1}^{|R|}(\bar{Y} - Y_i)^2$ равна нулю или выполняется хотя бы одно из заданных условий ранней остановки:

количество элементов $|R|$ меньше заданного минимального количества, необходимого для разбиения;

глубина дерева больше заданной максимальной глубины дерева;

Тогда полученный узел становится листом и добавляется к множеству Θ_{max} (25):

$$\Theta_{max} = \Theta_{max} \cup \{(R, \varphi)\}, \quad (25)$$

где φ - ответ дерева в данном узле, $\varphi = \frac{1}{|R|}\sum_{i=1}^{|R|} Y_i$.

Например, если после выполнения алгоритма было произведено только одно разбиение, то есть дерево состоит из корня и двух листов, то дерево будет иметь вид (26):

$$\Theta_{max} = \{(R_1, \varphi_1), (R_2, \varphi_2)\}. \quad (26)$$

8. Если условия остановки выполняются, и узел стал листом, выполняются следующие действия:

ведется поиск ближайшего родительского узла, у которого возможно разбиение правой ветки;

если такой узел существует, то алгоритм возвращается к этому узлу и выполняет процедуру разбиения его правой ветки, задается $R = R_2$(родительского узла) и происходит переход к шагу 1.2;

если такого узла не существует построение дерева заканчивается, полученное дерево является полным деревом Θ_{max}.

9. Если условия остановки не выполняются проводится разбиение левой ветки, задается $R = R_1$ и происходит переход к шагу 1.2.

2 Этап. Отсечение дерева

В результате первого этапа строится дерево, дающее наименьшую ошибку прогноза на обучающих данных. Как правило, такое дерево обладает излишней сложностью и не является оптимальным для прогнозирования на новых, не входящих в обучающую выборку, данных. Получить оптимальное дерево можно с помощью отсечения некоторых ветвей. Под отсечением подразумевается превращение узла в лист, после чего ответом дерева для данного листа становится среднее значение уровня подъема воды по всем примерам в данном листе.

В алгоритме CART для отсечения применяется механизм minimal cost-complexity tree pruning. По этому механизму ведется поиск дерева, являющегося компромиссом между сложностью дерева и ошибкой прогноза.

Среднеквадратическая ошибка дерева Θ на некоторой выборке R определяется как (27):

$$E(\Theta, \mathrm{R}) = \frac{1}{|R|}\sum_{i=1}^{|R|}(T(W_i;\Theta) - Y_i)^2, \quad (27)$$

где |R| - количество элементов в выборке;
Y_i – фактическое значение уровня подъема воды в i-ом элементе, $Y_i \in R$;
$T(W_j;\Theta)$ – ответ дерева Θ для вектора метеорологических факторов $W_j \in R$.

Стоимость дерева Θ определяется как (28):

$$C_a(\Theta) = E(\Theta, R) + a|\Theta|, \quad (28)$$

где a – коэффициент, изменяющийся от 0 до ∞;
$E(\Theta, R)$ – ошибка дерева на обучающей выборке;
$|\Theta|$ - число листьев дерева.

$C_a(\Theta)$ состоит из двух компонент – ошибки дерева и штрафа за его сложность. Коэффициент a позволяет определить штраф за сложность дерева. Так, при увеличении a, меньшую стоимость будут иметь деревья с меньшим количеством листьев, но дающие большую ошибку на тренировочных данных. Оптимальным для некоторого значения a является дерево с наименьшим значением $C_a(\Theta)$.

На этапе I было получено некоторое полное дерево Θ_{max}. Для получения оптимального дерева для $a = 0$, производится отсечение тех узлов, которые могут быть отсечены без увеличения ошибки.

1. Отсечение заведомо ненужных узлов

1.1. Ведется поиск таких пар листов $(R_1, \varphi_1), (R_2, \varphi_2)$ с общим предком, которые могут быть отсечены в родительский узел $R_{\text{объед}}$ без увеличения ошибки. Для таких R_1, R_2 выполняется неравенство (29):

$$\sum_{i=1}^{|R_{\text{объед}}|}(\varphi_{\text{объед}} - Y_i)^2 \leq \sum_{i=1}^{|R_1|}(\varphi_1 - Y_{1_i})^2 + \sum_{i=1}^{|R_2|}(\varphi_2 - Y_{2_i})^2, \quad (29)$$

где $R_{\text{объед}} = R_1 \cup R_2$ – множество элементов, созданное объединением R_1 и R_2;
$|R_{\text{объед}}|$ - количество элементов в $R_{\text{объед}}$;
Y_i – значение уровня подъема воды в i-ом примере $R_{\text{объед}}$;
$\varphi_{\text{объед}}$ – среднее значение уровня подъема воды в $R_{\text{объед}}$;
$\varphi_{\text{объед}} = \frac{1}{|R_{\text{объед}}|}\sum_{i=1}^{|R_{\text{объед}}|} Y_i$.

1.2. Если такая пара найдена, происходит отсечение R_1, R_2 в родительский узел. Узлы (R_1, φ_1) и (R_2, φ_2) заменяются узлом $(R_{\text{объед}}, \varphi_{\text{объед}})$.

1.3. В результате получается дерево, имеющее такую же или меньшую стоимость, как и Θ_{max}, но менее ветвистое. Это оптимальное дерево при $a = 0$. Обозначим это дерево как $\Theta_1 = \Theta(a = 0)$.

2. Получение последовательности поддеревьев

Возможно построить последовательность уменьшающихся поддеревьев (30):

$$\Theta_1 > \Theta_2 > \Theta_3 > \cdots > \{t\}, \tag{30}$$

где t – корневой узел дерева Θ_1.

Пусть k_{max} – количество поддеревьев в цепочке, $k = \overline{(1, k_{max})}$.

В последовательности Θ_{k+1} может быть получено путем отсечения Θ_k и является поддеревом Θ_k. Каждому дереву Θ_k соответствует значение a_k и Θ_k является оптимальным деревом для $a \in [a_k, a_{k+1})$.

Таким образом, мы можем получить последовательность оптимальных поддеревьев применяя процедуру отсечения начиная с Θ_1. Формирование поддеревьев заканчивается, когда в результате отсечения получается дерево, состоящее только из корня дерева Θ_1. Затем из последовательности поддеревьев выбирается такое поддерево, которое обеспечивает минимальную ошибку прогнозирования на тестовых данных.

2.1. Задаются начальные условия. На первом шаге алгоритма $k = 1$, $\Theta_k = \Theta_1$, $a_k = 0$.

2.2. Каждый узел дерева Θ_k, не являющийся листом, является разбиением и обозначается t_j, $j = \overline{(1, J_{max})}$, J_{max} – количество нелистовых узлов Θ_k. Пусть R_{t_j} это множество элементов, которые узел t_j разбивает на два подмножества R_1 и R_2, $R_{t_j} = R_1 \cup R_2$.

Для каждого узла t_j вычисляется величина (31):

$$g(t_j) = \frac{D(t_j) - E(\Theta_{k,t_j})}{\left|\Theta_{k,t_j}\right| - 1}, \quad (31)$$

где $D(t_j)$ – дисперсия узла t_j вычисляемая как $D(t_j) = \frac{1}{|R_{t_j}|}\sum_{i=1}^{|R_{t_j}|}(\bar{Y} - Y_i)^2$;

Θ_{k,t_j} – дерево с корнем в узле t_j;

$\left|\Theta_{k,t_j}\right|$ – количество листьев дерева Θ_{k,t_j}.

$E(\Theta_{k,t_j})$ – ошибка поддерева Θ_{k,t_j}, определяемая по формуле (32) на выборке R_{t_j}.

$$Q_0 = V_0 \cdot S_0, \quad (4)$$

где Q_0 – расход воды в реке до наступления паводка реки, м^3/с;
V_0 – скорость воды в реке до наступления паводка, м/с;
S_0 – площадь сечения русла реки, м^2;

$$S_0 = 0{,}5 \cdot b_0 \cdot h_0, \quad (5)$$
$$S_0 = 0{,}5 \cdot (a_0 + b_0) \cdot h_0, \quad (6)$$

(33) – для треугольного сечения;
(34) – для трапецеидального сечения.

2.3. Определяется g_{min} - минимальное значение $g(t_j)$ по всем узлам дерева Θ_k.

2.4. Все узлы обходятся сверху вниз и те, для которых $g(t_j) = g_{min}$, отсекаются. Таких узлов может быть больше одного. В результате отсечения

получается дерево Θ_{k+1}. Дереву Θ_{k+1} соответствует значение $a_{k+1} = g_{min}$, то есть $\Theta_{k+1} = \Theta(a = a_{k+1})$. Θ_{k+1} является оптимальным деревом для a_{k+1}.

2.5. Если дерево Θ_{k+1} состоит только из корня, построение цепочки поддеревьев завершается и алгоритм прекращает работу.

В противном случае устанавливается $k = k + 1$, процедура повторяется с шага 2.

3. Выбор финального дерева

Когда сформирована цепочка поддеревьев (1) требуется выбрать оптимальное поддерево.

Пусть k_{max} – количество поддеревьев в цепочке, R_{test} – тестовая выборка.

3.1. Для каждого поддерева вычисляется ошибка дерева Θ_k на тестовой выборке (35).

$$E(\Theta_k, R_{test}) = \frac{1}{|R_{test}|} \sum_{i=1}^{|R_{test}|} (T(W_i; \Theta_k) - Y_i)^2. \quad (35)$$

3.2. Определяется искомое оптимальное дерево решений. Им является поддерево с минимальной ошибкой (36) на тестовой выборке:

$$\Theta_{\text{опт}} = \underset{\Theta_k}{\operatorname{argmin}} E(\Theta_k, R_{test}), k = \overline{(1, k_{max})}. \quad (36)$$

Дерево $\Theta_{\text{опт}}$ в дальнейшем используется для прогнозирования. При поступлении вектора *w*, не обязательно принадлежащего *W*, такое дерево будет давать ответ *y*, не обязательно принадлежащий *Y*, и этот ответ будет являться прогнозом значения уровня подъема воды.

3 Этап. Оценка достоверности модели прогнозирования уровня подъема воды:

Для оценки достоверности модели прогнозирования в работе проводится апробация модели на фактических данных по значениям уровня подъема воды за один паводкоопасный сезон. Например, имеется обучающая

выборка по метеорологическим показателям и уровню подъема воды за 2007-2017 год. Для построения модели используются данные за 2007-2016 год. Затем производится прогнозирование значения уровня подъема воды по модели.

Вводится гипотеза о том, что регрессионная модель статистически значима. Оценка значимости модели проводится путем сравнения расчетного и табличного значения F-критерия Фишера-Снедекора (37-38):

$$F_{\text{р}} = \frac{Q_r(n-m)}{Q_e(m-1)} > F_{\text{т}\alpha;k_1;k_2}, \tag{37}$$

$$Q_r = \sum_{i=1}^{n}(T_i - \overline{T_i})^2 ; Q_e = \sum_{i=1}^{n}(T_i - Y_i)^2. \tag{38}$$

Если расчетное значение больше табличного значения при соответствующих степенях свободы, то гипотеза о значимости регрессионной модели принимается, если нет, то отклоняется.

4.1.3 Алгоритм построения модели на основе нейронной сети

Алгоритм обучения нейронной сети может быть представлен в виде последовательного выполнения следующих шагов:

Шаг 0. Задание весов всех связей случайными значениями.

Шаг 1. Выполнение шагов 2-9 до момента выполнения условия завершения алгоритма.

Шаг 2. Выполнение шагов 3-8 для каждой пары из входного набора данных.

Шаг 3. Каждый входной нейрон x_i отправляет сигнал всем нейронам следующего (скрытого) слоя.

Шаг 4. Каждый скрытый нейрон z_j суммирует входящие сигналы (39)

$$z_{in,j} = v_{0,j} + \sum_{i=1}^{n} x_i \cdot v_{i,j}, \quad (39)$$

где $v_{0,j}$ – смещение скрытого нейрона j и применяет функцию активации (40)

$$z_j = f(z_{in,j}). \quad (40)$$

Далее результат посылается всем нейронам следующего (выходного) слоя.

Шаг 5. Каждый выходной нейрон y_k суммирует входящие сигналы (41)

$$y_{in,k} = w_{0,k} + \sum_{j=1}^{m} z_j \cdot w_{j,k}, \quad (41)$$

где $w_{0,k}$ – смещение нейрона на выходе.
и применяет функцию активации

$$y_k = f(y_{in,k}), \quad (42)$$

тем самым получая выходной сигнал.

Шаг 6. Каждый выходной нейрон y_k получает выходное значение из обучающего набора данных и вычисляет ошибку

$$\sigma_k = (t_k - y_k) \cdot f'(y_{in,k}), \quad (43)$$

величину, на которую изменится вес связи

$$\Delta w_{j,k} = a \cdot \sigma_k \cdot z_j, \quad (44)$$

и величину корректировки смещения

$$\Delta w_{0,k} = a \cdot \sigma_k. \quad (45)$$

Далее каждый нейрон посылает σ_k нейронам в предыдущем слое.

Шаг 7. Каждый скрытый нейрон z_j суммирует входящие ошибки

$$\sigma_{in,j} = \sum_{k=1}^{n} \sigma_k \cdot w_{j,k}, \quad (46)$$

затем вычисляет величину ошибки

$$\sigma_j = \sigma_{in,j} \cdot f'(z_{in,j}), \tag{47}$$

величину, на которую изменится вес связи

$$\Delta v_{i,j} = a \cdot \sigma_j \cdot x_i, \tag{48}$$

и величину корректировки смещения

$$v_{0,j} = a \cdot \sigma_j. \tag{49}$$

Шаг 8. На данном шаге происходит изменение весов: каждый выходной нейрон y_k изменяет веса своих связей с элементом смещения и скрытыми нейронами:

$$w_{j,k}(\text{new}) = w_{j,k}(\text{old}) + \Delta w_{j,k}. \tag{50}$$

Каждый скрытый нейрон z_j изменяет веса своих связей с элементом смещения и выходными нейронами:

$$v_{i,j}(\text{new}) = v_{i,j}(\text{old}) + \Delta v_{i,j}. \tag{51}$$

Шаг 9. На данном шаге происходит проверка условия прекращения работы алгоритма. Данным условием является достижение заданной суммарной квадратической ошибки результата на выходе нейронной сети.

Вышеописанный алгоритм отображен на рисунке 8.

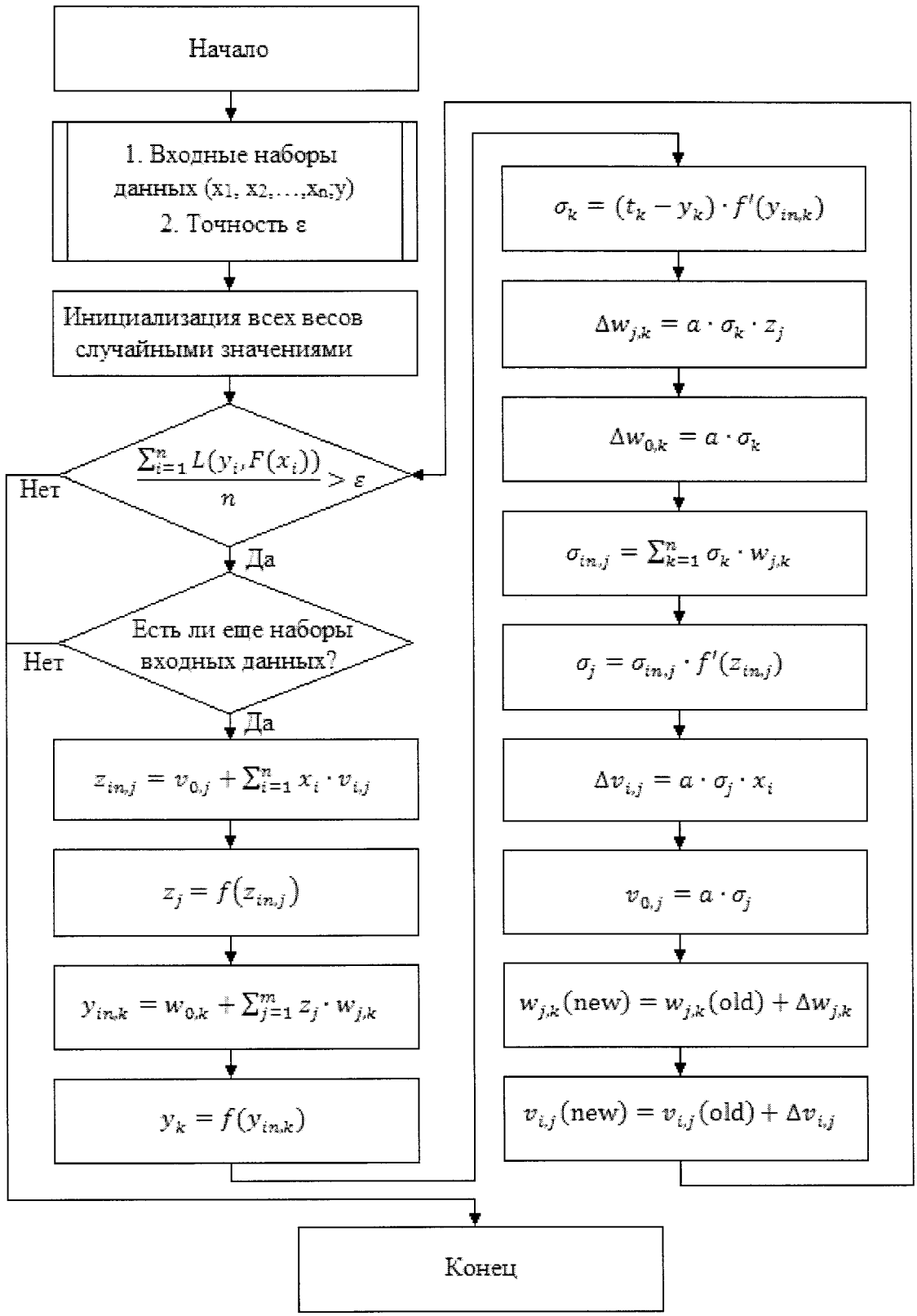

Рисунок 8 – Схема обучения нейронной сети

В основе приведенного метода построения лежит метод градиентного спуска. Градиент функции (в данном случае значение функции – это ошибка,

а параметры – это веса связей) показывает направление наибольшего возрастания функции.

4.1.4 Применение математических моделей прогнозирования на основе методов Data Mining для прогноза подъема уровня воды на реке Туапсе

1. Получение исходных данных

Для построения обучающей выборки был проведен анализ доступных исходных статистических данных по метеорологическим показателям и по уровню подъема воды в реке Туапсе.

В качестве данных по погодным условиям был выбран «Расписание Погоды» (rp5.ru). Данный источник обладает статистическими данными по зафиксированным метеорологическим показателям на метеостанциях мира с 2005 г с измерениями каждые 4 часа. Также данный ресурс располагает информацией о прогнозе метеоусловий на метеостанции с дальностью до 6 суток. В качестве исходных данных по метеорологическим показателям взяты значения наблюдений на метеостанциях по следующим показателям:

1. Температура воздуха;
2. Атмосферное давление;
3. Относительная влажность воздуха;
4. Скорость ветра;
5. Преобладающее направление ветра;
6. Облачность;
7. Количество осадков.

Этим метеорологические показатели выбраны потому, что по ним имеются достаточно полные и непротиворечивые данные.

В качестве статистических данных по подъему уровня воды был использован Интернет-ресурс «Информационная система по водным ресурсам и водному хозяйству бассейнов рек России» gis.vodinfo.ru (гидропост в г. Туапсе). Данные имеются по измерению уровня воды в реке

посуточно с 2001 г, но поскольку данные по погоде только с 2008 г, были также использованы данные с 2008 г. по наст. время.

2. Обработка исходных данных

После того, как данные получены необходимо произвести их обработку (очистку, корректировку, агрегирование). Обработка данных является важнейшей частью при работе с методами Data Mining. От того насколько «чистой» и полной получится обучающая выборка зависит качество построенной модели прогнозирования.

Без применения процедуры очистки невозможно применение некоторых аналитических алгоритмов Data Mining. Критериями качества данных являются:

отсутствие аномальных значений;

отсутствие противоречия в данных;

отсутствие несоответствия данных формату;

отсутствие пропусков в данных;

отсутствие дубликатов в данных.

Проблема отсутствия значений в таблице данных является существенной проблемой при использовании моделей Data Mining. В работе при обработке данных наблюдалось большое количетсво пропущенных значений как погодных данных, так и значений уровня воды. Данные могут отсутствовать за целые периоды, такие строки были удалены.

При использовании моделей Data Mining, данные должны подаваться на вход в числовом формате. Однако некоторые их столбцов могут иметь нечисловые (категориальные) значения. Например, направление ветра – «Ветер, дующий с запада», общая облачность – «20–30%».

Была проведены процедура кодирования категориальных значений. Для по каждому значению были выбраны уникальные записи и производилась кодирование их в числовые значения.

Далее производилось агрегирование данных с целью приведения их к одному формату измерений. Представленные исходные данные по погодным

условиям имеют слишком обширную детализацию (имеются измерения каждые 4 часа), а данные по уровню воды всего одно измерение в сутки. Для этого производилось усреднение данных.

При построении обучающей выборки необходимо также добавить столбцы значений зависимых переменных с лагом на несколько периодов в прошлом, от изменения которых может зависеть результирующая переменная. В качестве таких лагов были выбраны следующие: значение показателя один 1 назад, 2 дня назад, 3 дня назад и среднее значение за неделю до текущего момента времени).

В результате анализа временного ряда уровня воды на гидропосту в г. Туапсе было обнаружено, что с 2008 по 2010 г. уровень воды значительно занижен по сравнению с последующим периодом (рисунок 9). Это может негативно сказаться на качестве модели, поскольку данный период будет смещать в меньшую сторону прогнозное значение уровня. Поэтому указанный период был исключен из выборки.

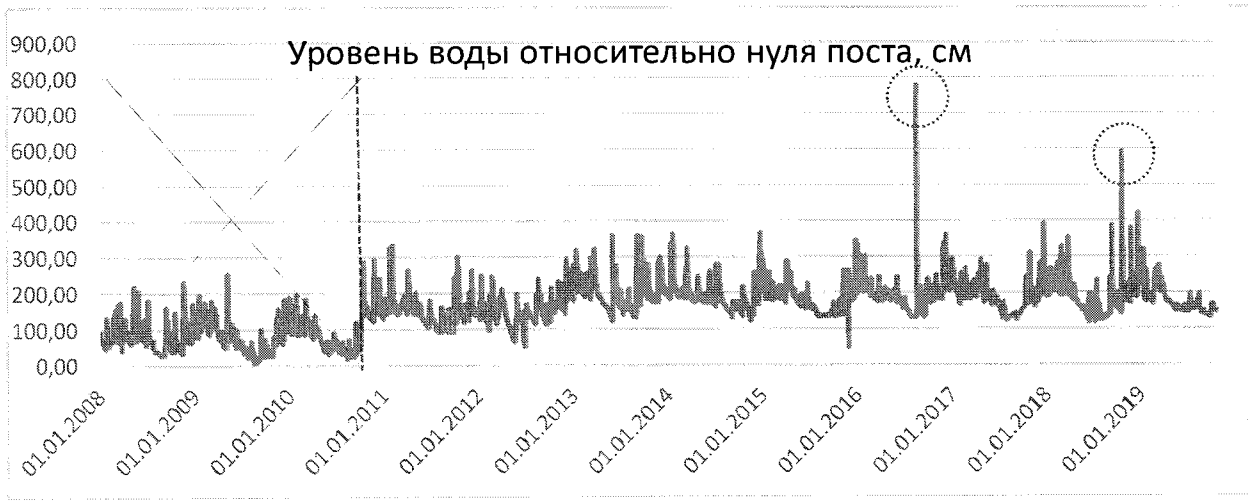

Рисунок 9 – Ход уровня воды на гидропосту в г. Туапсе

Таким образом исходные данные после обработки представляют собой выборку с 2010 по 2019 г. Независимой (выходной) переменной является уровень подъема воды, зависимой (входной) – наблюдения метеопоказателей (таблица 1).

Таблица 1 – Пример выборки данных

Дата	Уровень воды	T	T1	T2	T3	T7	P	P1	P2	…
01.10.2010	80	18,99	21,08	21,41	23,18	22,66	760,4	755,13	758,8	
02.10.2010	67	17,41	18,99	21,08	21,41	22,33	762,6	760,4	755,13	
03.10.2010	64	13,55	17,41	18,99	21,08	21,36	761,31	762,6	760,4	
04.10.2010	60	14,61	13,55	17,41	18,99	19,87	764,79	761,31	762,6	
05.10.2010	57	13,33	14,61	13,55	17,41	18,6	765,23	764,79	761,31	
…	…	…	…	…	…	…	…	…	…	…

3. Построение моделей прогнозирования на период развития паводка (октябрь 2018 г.)

Для прогнозирования подъема уровня воды на гидропосту на октябрь 2018 г. обучающая выборка была отсечена: для обучения взят период с 01.2010 г. по 09.2018 г.

Были построены 2 модели: дерево решений (рис. 10) и нейронная сеть (рисунок 11). Построение и настройка моделей производились с использованием программного обеспечения Deductor Studio Academic.

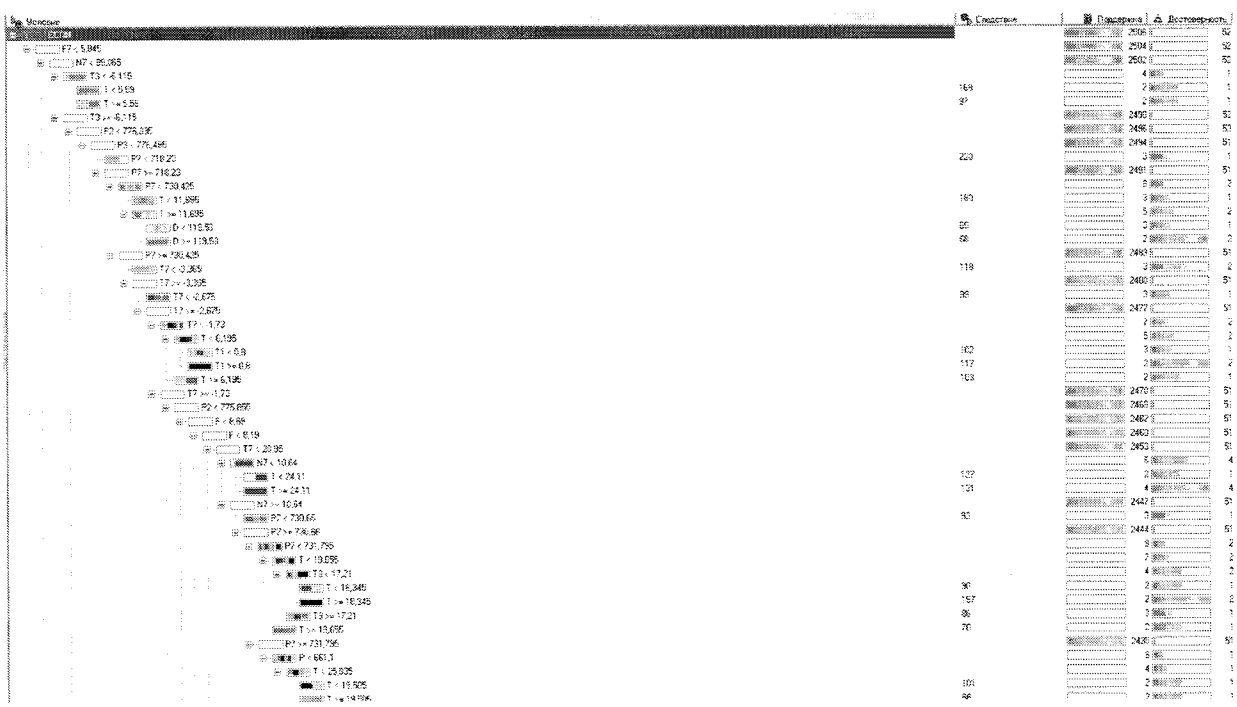

Рисунок 100 – Фрагмент дерева решений

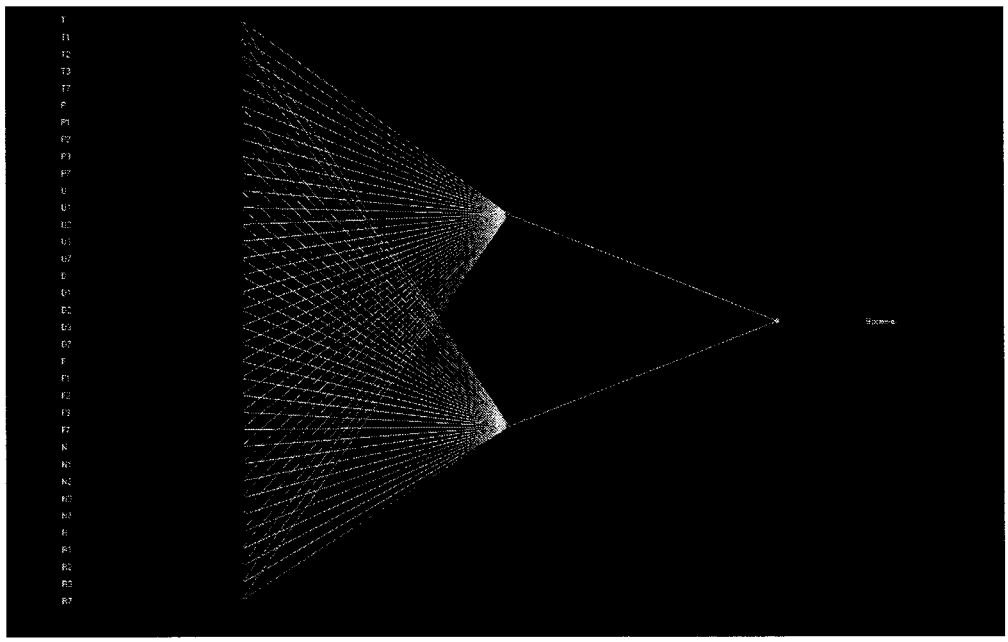

Рисунок 11 – Структура нейросети

Фактические и прогнозные значения уровня воды, полученные с помощью обеих моделей представлены в табл. 2 и на рисунке 12.

Анализ графика показывает, что прогноз полученный с помощью нейросети более сглаженный. Дерево решений дает больший разброс, это свидетельствует о том, что модель дерева решений получилась более чувствительной. Катастрофический подъем уровня воды 25.10 до отметки 590 см от уровня поста обе модели предсказали слабо, хотя попытка в обеих моделях сделана (наблюдается небольшой рост с последующим медленным падением, особенно характерно выражено у модели дерева решений).

Плохое качество работы моделей при прогнозе подъема уровня воды 25.10 легко объяснить тем, что в обучающей выборке попросту даже не было прецедентов, когда уровень воды принимал такое значение (кроме случая 21.08.2016 г.). Поэтому данный прецедент является аномалией для моделей и модели прогнозируют значение выходной переменной только в рамках того коридор значений, которые были заложены при их обучении.

Таблица 2 – Результаты прогноза

Дата	Фактический уровень воды	Прогноз дерево решений	Прогноз нейросеть	Ошибка дерево решений	Ошибка нейросеть
01.10.2018	157	177	149	20	8
02.10.2018	156	149	149	7	7
03.10.2018	156	149	152	7	4
04.10.2018	177	130	164	47	13
05.10.2018	198	192	182	6	16
06.10.2018	184	209	170	25	14
07.10.2018	178	157	160	21	18
08.10.2018	198	192	154	6	44
09.10.2018	170	145	154	25	16
10.10.2018	168	144	157	24	11
11.10.2018	165	140	152	25	13
12.10.2018	164	140	144	24	20
13.10.2018	161	132	147	29	14
14.10.2018	161	194	145	33	16
15.10.2018	159	192	142	33	17
16.10.2018	156	180	137	24	19
17.10.2018	152	129	136	23	16
18.10.2018	149	129	137	20	12
19.10.2018	145	129	137	16	8
20.10.2018	145	107	139	38	6
21.10.2018	142	137	142	5	0
22.10.2018	157	92	145	65	12
23.10.2018	160	53	150	107	10
24.10.2018	157	160	241	3	84
25.10.2018	590	224	195	366	395
26.10.2018	370	207	182	163	188
27.10.2018	312	207	172	105	140
28.10.2018	267	207	162	60	105
29.10.2018	232	207	162	25	70
30.10.2018	224	207	164	17	60
31.10.2018	202	207	146	5	56
Сумма				**1374**	**1412**
Среднее				**85,88**	**88,25**

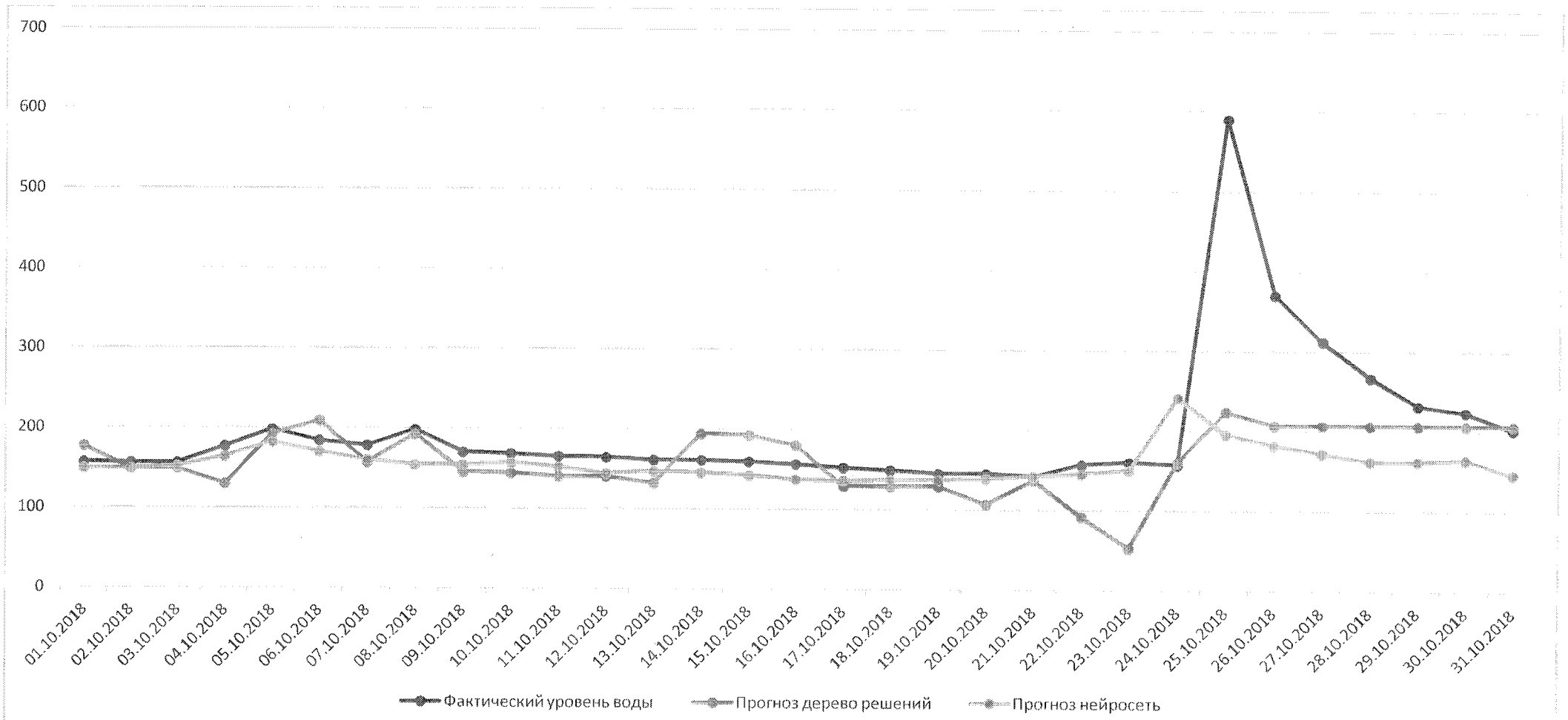

Рисунок 12 – Фактические и прогнозные значения уровня воды

4.2 Алгоритм прогнозирования уровня воды на основе метода регрессионного анализа (на примере реки Туапсе)

Для водосбора реки Туапсе использовались данные одной метеостанции.

Для створа р. Туапсе – г. Туапсе использовались данные гидрометрических наблюдений за период с 1949 г. по 2010 г. На основе этих данных была построена и уточнена кривая расходов, представленная на рисунке 13.

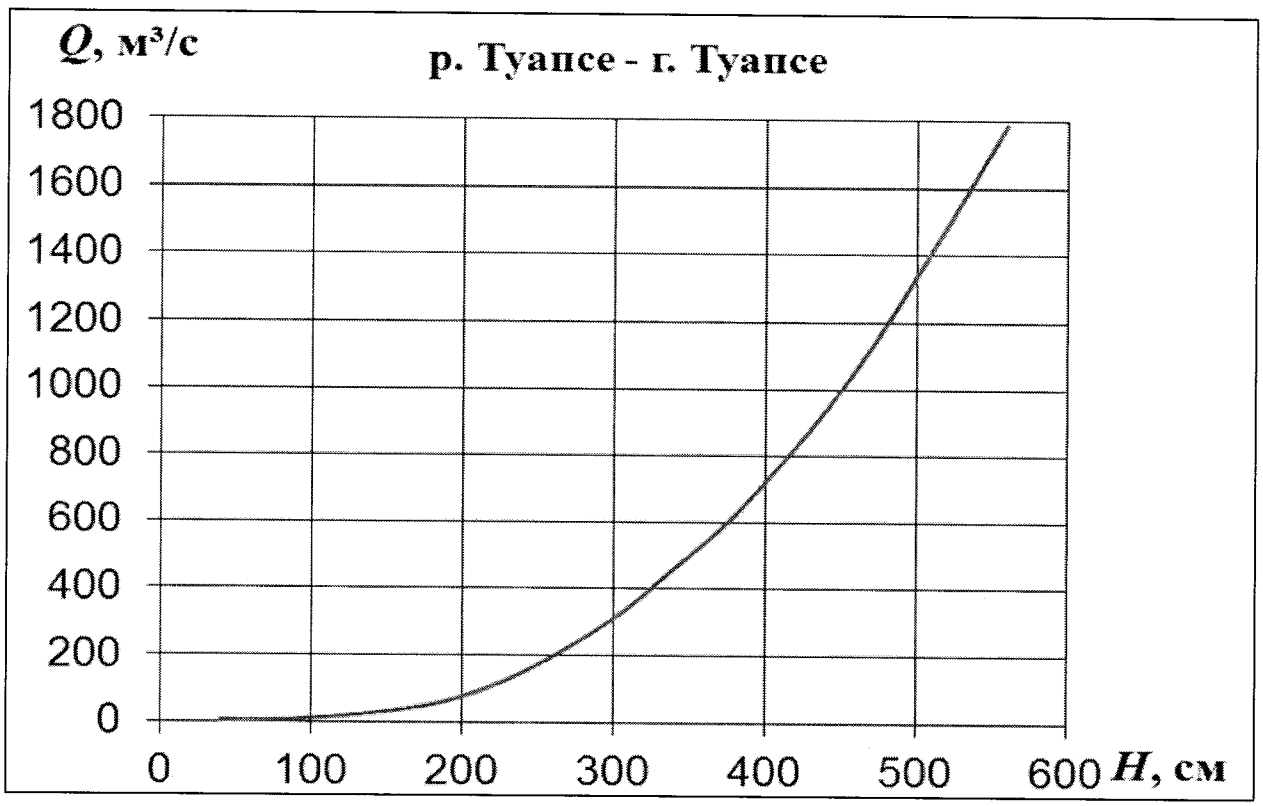

Рисунок 13 – Кривая зависимости между расходом и уровнем воды в створе р. Туапсе – г. Туапсе.

В распоряжении имелись только данные о среднесуточных значениях уровней и расходов воды. Использовались данные метеорологических наблюдений за период с 1984 по 1987 г. и с 1990 по 2005 г. на метеостанции г. Туапсе, расположенной недалеко от замыкающего створа в самой нижней части водосбора на высоте z_0 = 60 м. Базовый период за 9 лет включает 3189 ежесуточных совместных гидрологических и метеорологических наблюдений.

Для данного створа приняты следующие критические значения среднесуточной температуры приземного слоя воздуха на метеостанции: минимальная температура, при которой происходит снеготаяние $T_{min,S}$ = 0°C; минимальная температура, при которой осадки выпадают в жидком виде $T_{min,P}$ = 2°C; максимальное значение учитываемой температуры T_{max} = 25°C.

Прогноз среднесуточного расхода воды предлагается получать по формуле (51):

$$\begin{aligned}\tilde{Q}(t+1) = {} & a_0 + a_1 Q(t) + a_2 Q(t-1) + a_3[T(t+1)] - T_{minS}]^2 \\ & + a_4[T(t+1)] - T_{minS}] + a_5 P(t+1)[T(t+1)] - T_{minP}]^2 \\ & + a_6 P(t+1)[T(t+1)] - T_{minP}] + a_7 P(t+1) \\ & + a_8[T(t)] - T_{minS}]^2 + a_9[T(t)] - T_{minS}] \\ & + a_{10} P(t)[T(t)] - T_{minP}]^2 + a_{11} P(t)[T(t)] - T_{minP}] \\ & + a_{12} P(t).\end{aligned} \tag{51}$$

Параметры которой для каждого месяца помещены в таблице 3.

Таблица 3 – Характеристика потерь населения по зонам затопления в зависимости от скорости течения потока воды.

	янв	фев	март	апр	май	июнь	июль	авг	сен	окт	нояб	дек
a_0	1,4	2,4	2,4	-14,5	-58	22,6	78,5	-6,9	15	3,99	0,5	6,5
a_1	0,48	0,73	0,49	0,45	0,19	0,58	0,38	0,11	0,15	0,29	0,35	0,23
a_2	-0,02	-0,11	0,09	0,08	0,13	-0,03	0,18	0,16	0,06	0,13	0,1	0,1
a_3	-0,04	0,15	0.004	-0,16	-0,217	-0,03	0,169	0,544	-0,101	0,051	-0,02	-0,074
a_4	0,99	-0,94	0,16	5,18	8,66	1,59	-7,89	-24,19	5,46	-1,4	1,11	-1,1
a_5	0,021	-0,007	-0,073	-0,007	-0,052	-0,03	-0,01	0,25	-0,05	-0,012	0,0007	-0,049
a_6	0,391	0,465	0,703	-0,219	0,911	1,01	0,42	-9,23	1,674	0,112	-0,038	0,737
a_7	0,27	-0,21	0,05	3,56	-1,53	-7,67	-4,24	84,84	-12,21	1,29	2,17	-0,74
a_8	-0,12	-0,11	-0,038	0,108	0,065	0,083	-0,025	-0,56	0,141	-0,054	0,021	0,0515
a_9	0,98	0,93	0,48	-3,26	-2,56	-3,64	1,17	24,9	-7,05	1,21	-1,095	1,115
a_{10}	0,032	-0,05	0,056	-0,021	-0,103	-0,001	0,024	-0,01	-0,008	0,0007	0,0021	-0,033
a_{11}	-0,01	0,552	-0,352	0,285	2,63	-0,04	-0,89	0,497	0,072	0,005	-0,024	0,487
a_{12}	-0,51	-1,62	0,38	-0,24	-15,2	0,76	8,36	-5,12	1,615	-0,05	0,29	-0,445

Погрешность методики прогноза среднесуточных расходов воды оценивалась на независимом материале. Для этого исключались данные за один год, производилась переоценка параметров, а данные за исключенный год использовались для сравнения прогноза среднесуточных расходов воды с их

фактическими значениями. Для оценки погрешности прогноза эта процедура производилась для всех лет базового периода. На рисунке 14 приведены совмещенные графики колебаний фактических и спрогнозированных среднесуточных расходов воды в 1995 г.

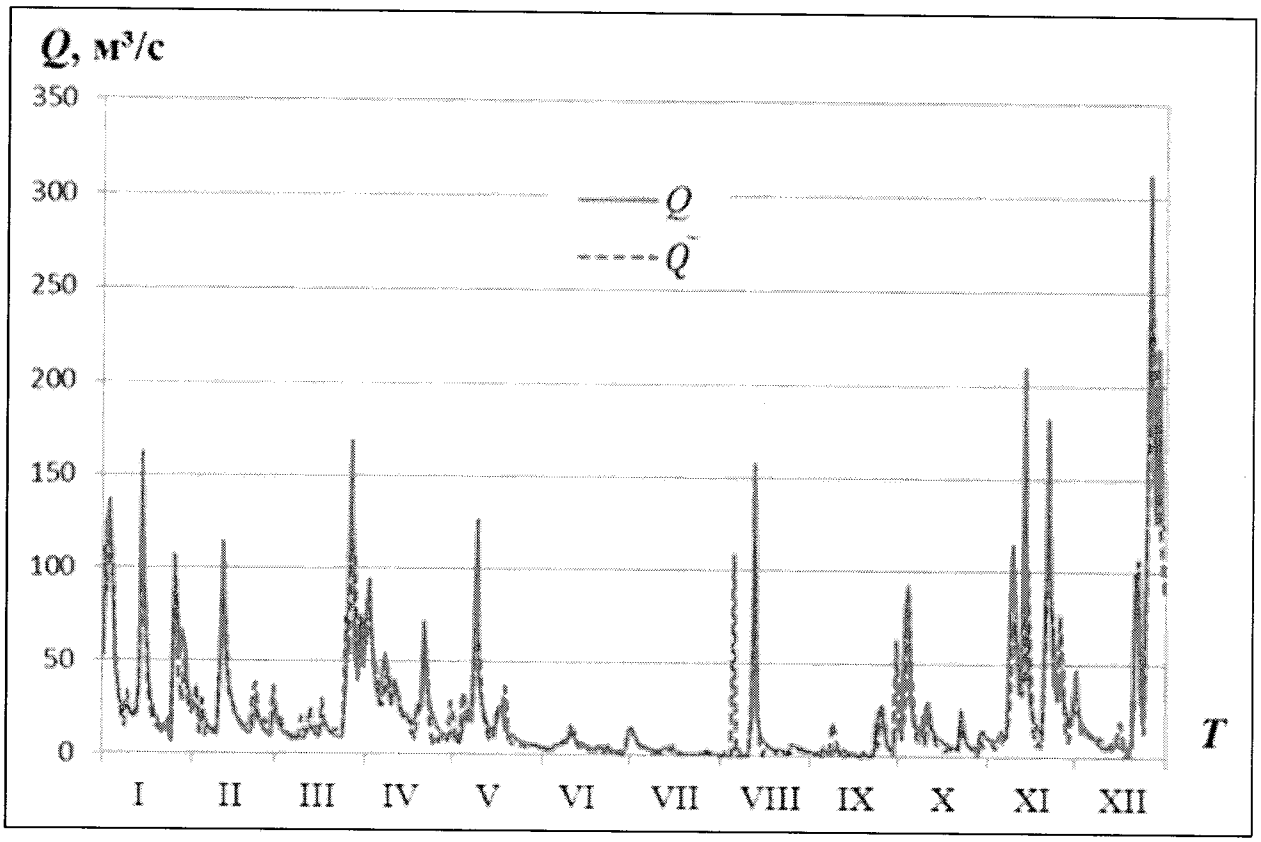

Рисунок 1411 – График колебаний фактических (синий цвет) и спрогнозированных (красный цвет) среднесуточных расходов на проверочный период (1995 г.).

В целом предлагаемая методика прогноза среднесуточных расходов воды в створе р. Туапсе – г. Туапсе имеет коэффициент корреляции между прогнозируемыми и фактическими значениями среднесуточных расходов воды R = 0,86; среднюю квадратическую погрешность ошибки прогноза $\sigma_{пр}$=14,2 м^3/с и показатель эффективности прогноза $\sigma_{пр}$ / σ_{Δ} = 0,53. Таким образом, при условии точного метеорологического прогноза предлагаемая методика достаточно точна и эффективна по сравнению с инерционным прогнозом.

С целью проверки предлагаемой схемы получения прогноза для случаев экстремально большого увеличения расходов воды в течение суток из всего

набора 3189 фактических и прогнозируемых среднесуточных расходов воды была выбрана 31 пара с наибольшими значениями суточного роста расходов. Таким образом, были выделены случаи увеличения расхода воды за сутки обеспеченности не выше 1%. Для них была получена оценка средней квадратической погрешности прогноза 79,8 м^3/с и показателя эффективности 0,44.

Для получения прогноза суточных максимумов расхода воды $\tilde{Q}_{max}(t+1)$ (в случае ожидания паводков) прогнозируемые значения $\tilde{Q}(t+1)$ среднесуточных расходов следует умножать на коэффициент k'. Для подстраховки средние для каждого месяца значения k' можно заменять их значением (5%) k, соответствующим обеспеченности 5%. Для каждого месяца значения k' и (5%) k помещены в таблице 4.

Таблица 4 – Значения переходных коэффициентов k' и (5%) k.

Месяц	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
k'	1,9	2,3	1,68	1,54	2,22	2,98	2,87	2,41	2,87	2,86	2,21	2
(5%) k	3,32	6,4	2,61	2,39	7,16	6,9	6,64	4,69	8,35	7,95	4,3	3,5

Оценки средней квадратической погрешности и показателя эффективности прогноза максимальных расходов воды в створе р. Туапсе – г. Туапсе равны: $\sigma_{пр}$ = 52,5 м^3/с, $\sigma_{пр} / \sigma_{\Delta}$ = 0,57.

Прогноз среднесуточных H(t+1) и максимальных H_{max}(t+1) уровней воды в створе р. Туапсе – г. Туапсе следует получать на основе соответствующих прогнозов среднесуточного $\tilde{Q}(t+1)$ и максимального расхода воды $\tilde{Q}_{max}(t+1)$, используя обратное преобразование H(Q) с помощью графика на рисунке 13. При выпуске прогноза максимальных расходов и уровней воды в створе р. Туапсе – г. Туапсе в 19^{00} по местному времени, его заблаговременность составляет 14 - 26 часов.

Методика вероятностного прогнозирования максимальных расходов и уровней воды в створе р. Туапсе – г. Туапсе с заблаговременностью одни сутки основана на теоретических положениях и результатах статистического анализа.

Для данного речного створа соответствующие различным уровням опасности критические значения уровня и расхода воды помещены в таблице 5.

Таблица 5 – Значения переходных коэффициентов k' и (5%) k.

Уровень опасности	$H_{кр}$ (см)	$Q_{кр}$ ($м^3/с$)
Отметка для подачи штормовых телеграмм	350	600
Отметка неблагоприятных явлений	400	880
Отметка опасного явления	420	980

Оценки вероятностей превышения критических расходов и уровней воды в течение всего года и каждого месяца, полученные по многолетним рядам соответствующих максимумов расхода воды, помещены в таблице 6.

Таблица 6 – Вероятность превышения критических значений $H_{кр}$ (%).

Месяц	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Год
$Q_{кр}$=600	0,72	1,2	0,8	0,45	1,1	1,2	1,8	3,2	2,4	6,8	10	4,3	18
$Q_{кр}$=880	0,11	0,29	0,19	0,12	0,47	0,49	1	2	1,3	4,3	6,2	1,07	5,4
$Q_{кр}$=980	0,06	0,19	0,12	0,08	0,36	0,38	0,82	1,8	1,1	3,8	5,3	0,69	3,6

При известном прогнозе максимального расхода $\tilde{Q}_{max}(t+1)$ и уровня воды $\tilde{H}_{max}(t+1)$ на сутки t+1 вероятность того, что в течение этих суток фактический максимальный расход $Q_{max}(t+1)$ превысит критическое значение $Q_{кр}$, а фактический максимальный уровень воды $H_{max}(t+1)$ превысит $H_{кр}$, определяется формулой (52).

$$p_{кр}\tilde{Q}_{max} = 1 - Ф\left\{\frac{lnQ_{кр} - ln[\widetilde{Q}_{max}(t+1)] - m_{ln\varepsilon}}{\sigma_{ln\varepsilon}}\right\}. \qquad (52)$$

Параметры которой для каждого месяца помещены в таблице 7.

Таблица 7 – Значения переходных коэффициентов k' и (5%) k.

Месяц	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
$m_{ln\varepsilon}$	-0,03	-0,08	-0,09	0,01	0,02	-0,15	-0,08	-0,66	-0,16	-0,25	-0,21	-0,16
$\sigma_{ln\varepsilon}$	0,47	0,41	0,45	0,74	0,73	0,69	0,7	1,14	1,57	1,25	1,23	0,79

На рисунке 15 для двух месяцев – февраля (сезон дождевых паводков) и августа (летняя межень) приведены графики функции $p_{\text{кр}}\tilde{Q}_{max}$, определяющей вероятностный прогноз максимальных расходов и уровней воды на одни сутки.

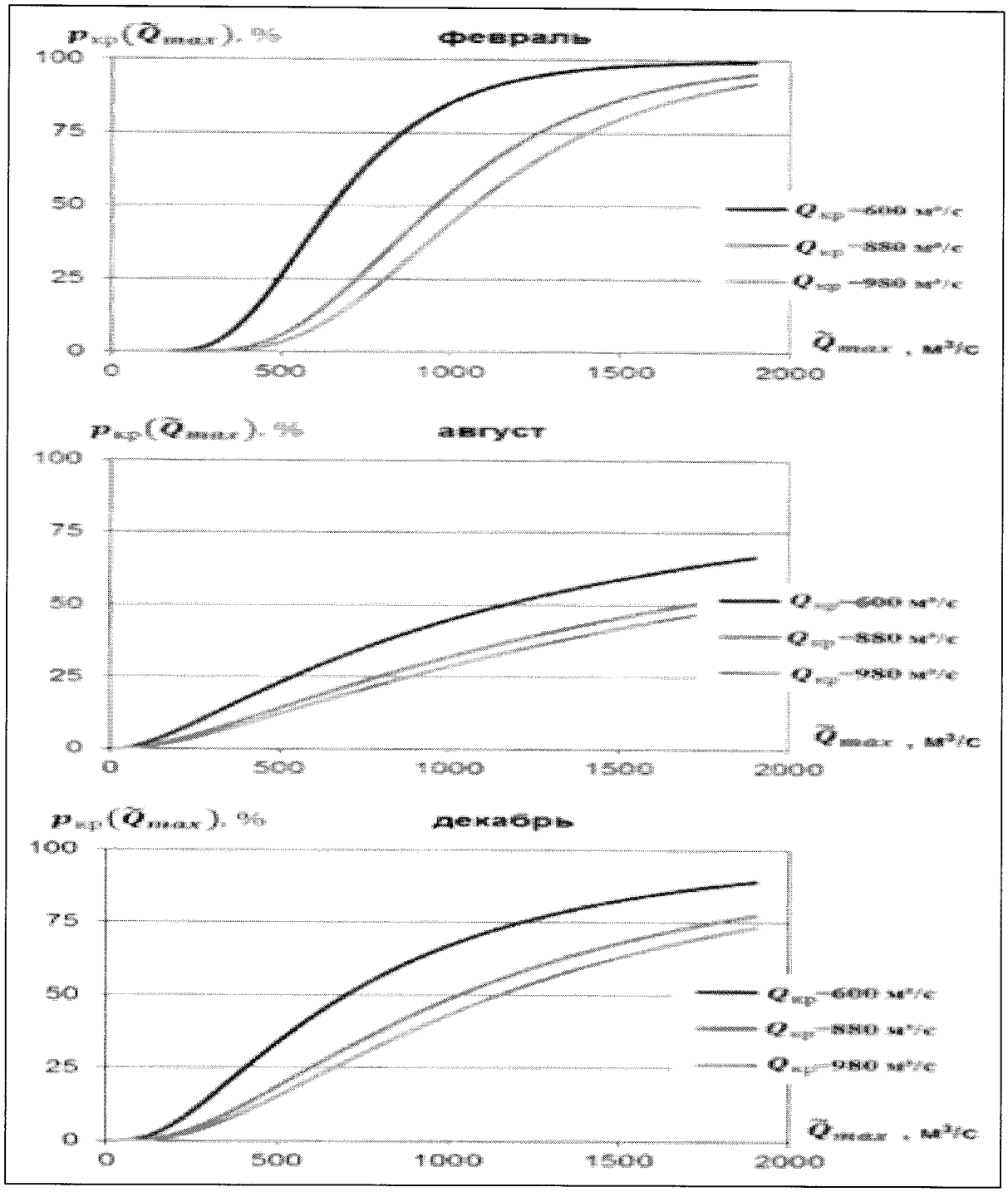

Рисунок 1512 – Графики функций прогностической вероятности $p_{\text{кр}}\tilde{Q}_{max}$ для февраля, августа и декабря в створе р. Туапсе – г. Туапсе

При известных на дату составления прогноза t значениях $Q_{max}(t)$, $Q_{max}(t-1)$, P(t), T(t) и прогнозе T(t+1) среднесуточной приземной температуры воздуха критический слой осадков $P_{кр}(q)$ определяется, исходя из заданной вероятности q превышения расходом и уровнем воды в створе р. Туапсе – г. Туапсе своих критических значений.

Для предложенной регрессионной модели был осуществлен расчет прогнозных значений расхода и уровня воды для 24, 25 октября 2018 г. (превышение опасного явления, вызванное дождевыми осадками).

В качестве исходных данных выступали сведения полученные из информационных ресурсов «Информационной системы по водным ресурсам и водному хозяйству бассейнов рек России» gis.vodinfo.ru (рисунок 16).

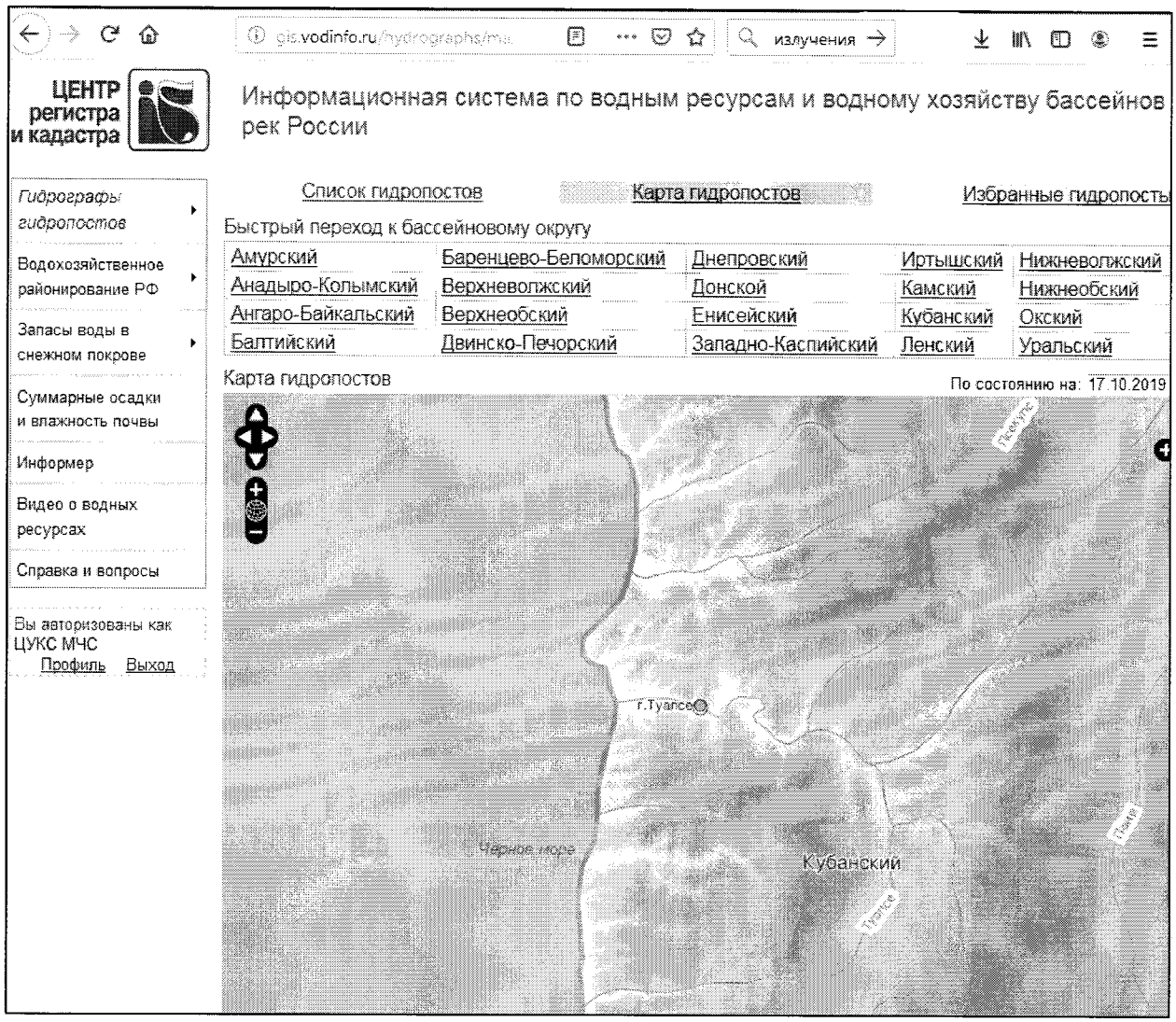

Рисунок 16 – Интерфейс сайта gis.vodinfo.ru

А также информационных ресурсов Федеральной службы по гидрометеорологии и мониторингу окружающей среды meteoinfo.ru (рисунок 17).

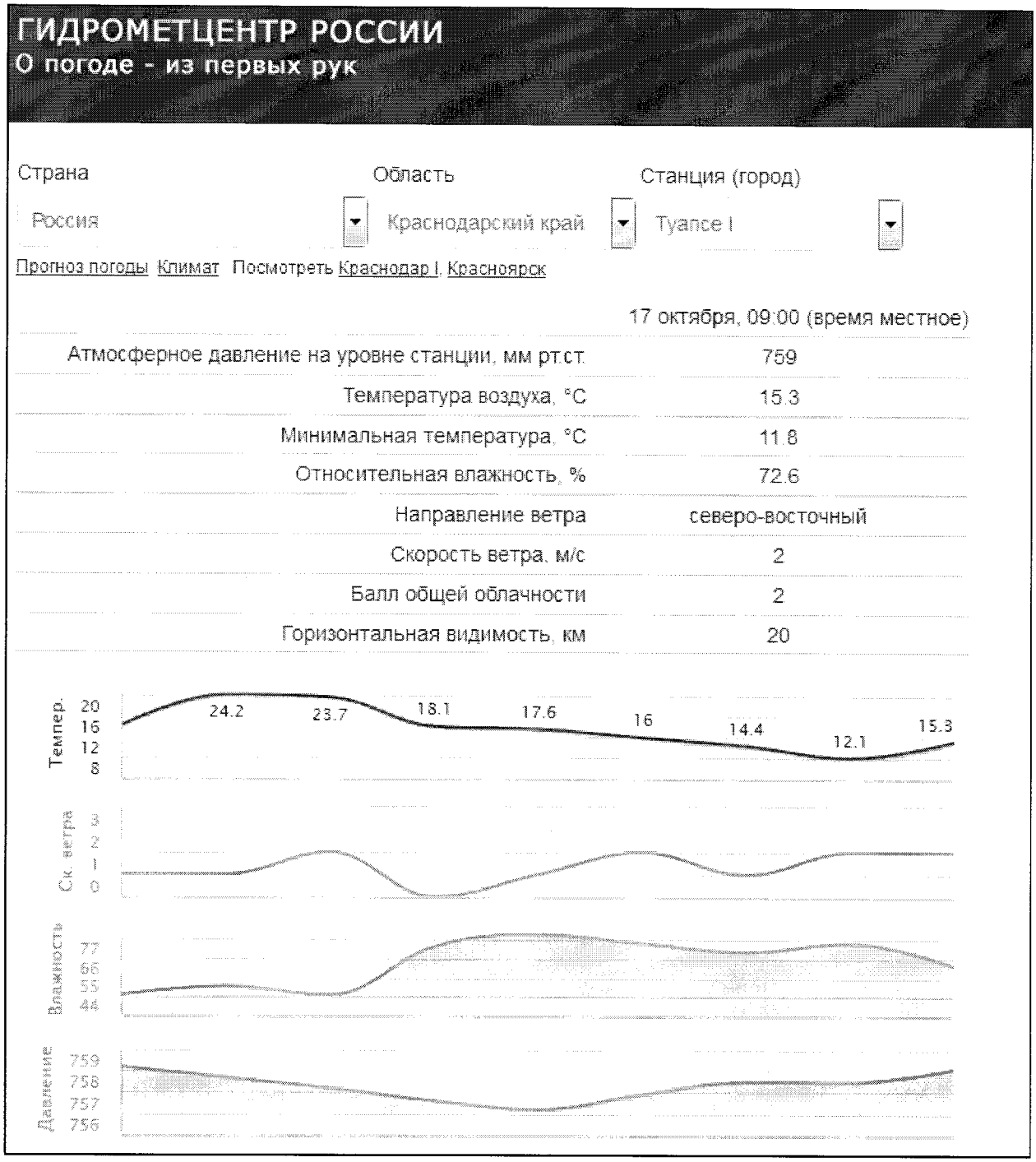

Рисунок 17 – Интерфейс сайта meteoinfo.ru

И информационных ресурсов открытых источников сети Internet pogoda-service.ru (рисунок 18).

Дата	Максимальная температура	Минимальная температура	Средняя температура	Атмосферное давление	Скорость ветра	Осадки	Эффективная температура
01.01.2018	10.1	7.8	8.8	1006.3	2	11	7.0
02.01.2018	11.9	6.3	8.4	1007.8	2	0	6.2
03.01.2018	11.3	4.4	7.4	1006.4	2	0	5.1
04.01.2018	13.2	5.7	11.6	1001.2	2	0	9.3
05.01.2018	17.3	9.8	12.1	1002.4	2	0	10.1
06.01.2018	14.7	7.0	9.3	1013.9	1	0	8.1
07.01.2018	11.9	5.5	8.9	1021.8	2	0	6.9
08.01.2018	11.3	8.2	10.1	1022.2	3	0	7.7
09.01.2018	10.4	-0.2	5.3	1021.3	2	10	2.7
10.01.2018	8.2	-1.7	2.7	1022.6	2	0	-0.4
11.01.2018	10.8	1.9	6.1	1010.3	3	0	2.4
12.01.2018	8.3	2.1	5.4	1008.7	5	3	0.7
13.01.2018	1.4	-1.8	-0.3	1014.3	6	0	-6.2
14.01.2018	2.4	-2.1	0.3	1012.3	4	0	-4.3
15.01.2018	3.8	0.4	2.7	1020.1	3	1	-0.8
16.01.2018	7.9	2.0	5.4	1015.4	3	3	2.2
17.01.2018	8.7	5.5	7.5	1004.2	5	13	3.1
18.01.2018	13.9	6.9	10.0	993.3	3	0	6.9
19.01.2018	8.8	5.1	7.4	1000.8	4	19	4.2
20.01.2018	10.1	4.9	8.3	1010.9	4	6	5.1
21.01.2018	12.3	8.5	9.6	1007.1	4	2	6.4
22.01.2018	13.0	7.3	10.7	998.8	3	0	7.8
23.01.2018	13.1	4.8	9.3	1001.2	3	8	6.7
24.01.2018	3.5	-1.1	1.9	1015.8	3	14	-1.7
25.01.2018	2.3	-2.7	-0.4	1017.6	4	0	-5.0
26.01.2018	6.0	-0.1	2.9	1016.6	4	0	-1.4
27.01.2018	9.1	1.5	3.4	1021.3	2	0	0.5
28.01.2018	10.1	1.0	5.1	1024.1	1	0	3.2
29.01.2018	6.6	3.0	4.6	1017.5	3	12	1.5
30.01.2018	7.3	3.8	6.0	1013.7	6	26	1.2
31.01.2018	8.9	2.3	5.7	1012.2	3	22	2.6

Рисунок 1813 – Интерфейс сайта pogoda-service.ru

Итоги проведенных по предлагаемой модели расчетов приведены на рисунке 19.

	24-е		25-е	
	Фактически	Прогноз	Фактически	Прогноз
H (mm)	157	168	590	250
Q (m^3/s)	40	44,1038	2000	148,316

	22	23	24	25
h (mm)	157	160	157	590
q(m^3/s)	40	41	40	2000
t(°C)		16	17	14
p (mm)		4	84	125

a_0	3,99
a_1	0,29
a_2	0,13
a_3	0,051
a_4	-1,4
a_5	-0,012
a_6	0,112
a_7	1,29
a_8	-0,054
a_9	1,21
a_{10}	0,0007
a_{11}	0,005
a_{12}	-0,05

Рисунок 19 – Результаты прогнозных расчетов на основе регрессионных моделей и их сравнение с фактическими значениями

Как видно из полученных данных для первого дня расхождения между прогнозируемым уровнем подъема воды и фактическим уровнем подъема составили порядка 7% в то время, как для второго дня (когда и наблюдался основной пик паводка) расхождения составили уже более 100%.

Данное обстоятельство позволяет сделать следующий вывод: предлагаемая регрессионная модель обеспечивает удовлетворительную сходимость прогнозных и фактических значений подъемов уровня воды (вызванных выпадением осадков в виде дождя) в случае незначительных величин осадков (доверительный интервал нуждается в дополнительном уточнении). При этом для «залповых» случаев выпадения осадков модель в представленном виде применяться не может вследствие большой величины ошибки.

В качестве пути совершенствования модели может выступать направление, заключающееся в уточнении параметров модели конкретно для случаев «залпового» выпадения осадков.

4.3 Алгоритм прогнозирования последствий подъема уровня воды на основе триангуляцинного метода (на примере реки Туапсе)

С целью осуществления прогноза оценки последствий воздействий паводка на объекты необходимо осуществить ряд операций:

1. Определение участков местности, для которых необходимо осуществить расчет (рисунок 20). Как правило, предпочтение отдается нескольким относительно небольшим участкам русла реки. Выбор участков обосновывается наличием объектов жилой застройки с учетом специфики рельефа местности, наличием объектов инфраструктуры, критически важных и потенциально опасных объектов;

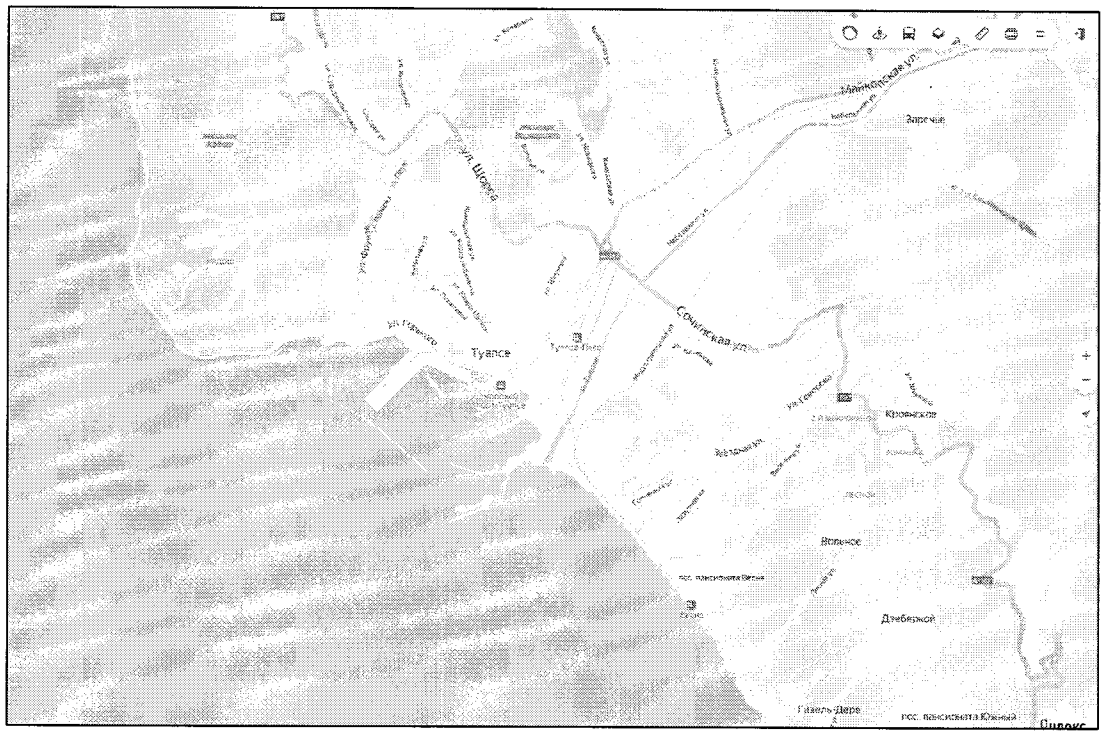

Рисунок 2014 – Пример выбора участка в устье реки Туапсе для проведения расчетов (н.п. Туапсе)

2. Определение «матрицы высот» для рассматриваемого участка. Масштаб шага сетки и высот рекомендуется выбирать в зависимости от рельефа местности выбранного участка. Следует обращать внимание на то, что при задании слишком малого шага повышается точность проводимых расчетов, но увеличивается потребность в машинных ресурсах, что при их ограниченности может привести к снижению оперативности прогноза последствий.

3. Внесение матрицы высот в виде массива данных (рисунок 21), в базы данных геоинформационной подложки. Значения абсолютной высоты на местности, в задаваемом массиве определяется как высота «центра масс» рассматриваемых конечных элементов.

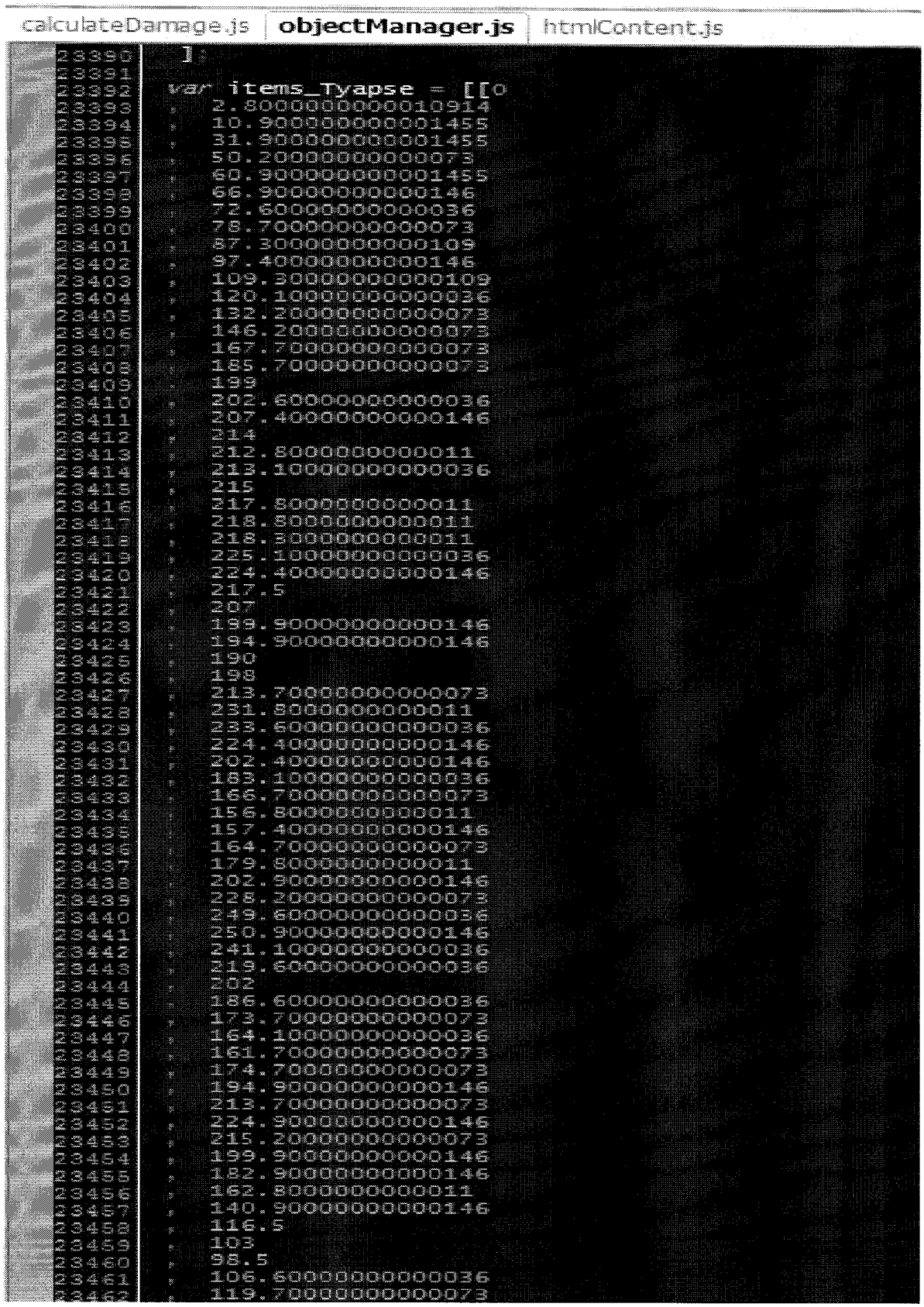

Рисунок 151 – Массив исходных данных, содержащий сведения о «матрице высот» рассматриваемого участка

4. Внесение в базы данных сведениях о объектах (рисунок 22), находящихся в зоне возможного затопления. При этом в базу должны вноситься сведения о основных архитектурных решениях, применяемых при создании объектов, численности проживающих/работающих (рекомендуется разделить базу данных на две составляющих день/ночь с целью большей детализации при проведении расчетов по определению числа пострадавших). Также при занесении сведений о объекте, необходимо четко определить его координаты (либо адрес), с целью обеспечения корректной работы «счетчика», определяющего число объектов, попавших в зону воздействия поражающих факторов паводка.

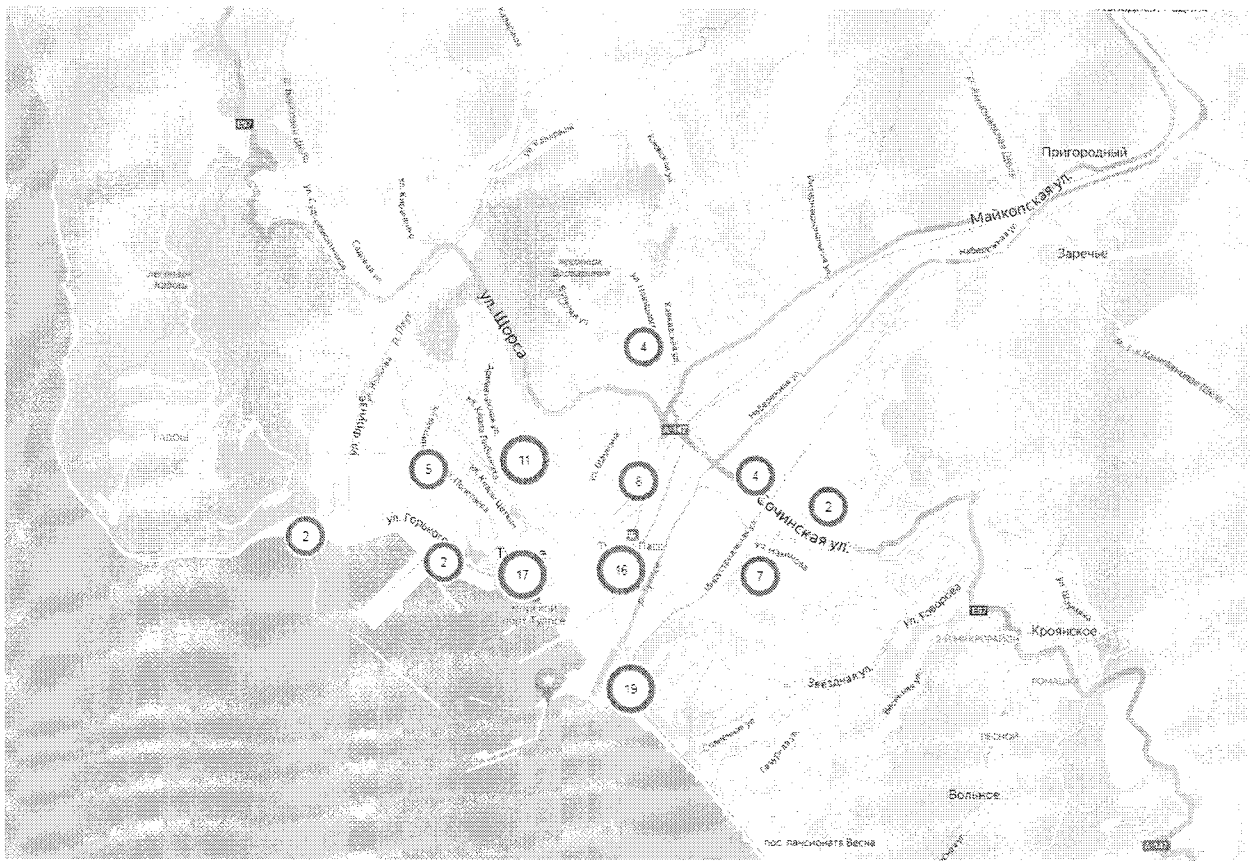

Рисунок 2216 – Вид рассматриваемого участка с внесенными в базу данных объектами

5. Ввод исходных данных (рисунок 23) для моделирования последствий воздействия паводка на объект. Для этого необходимо задать

исходные данные, необходимые для расчета параметров поражающих факторов паводка, рассчитываемые по моделям, представленным в пунктах 2.4.1 и 2.4.2. В настоящее время идет работа по автоматизации процесса получения исходных данных из материалов внешних источников (сайт meteoinfo.ru, gis.vodinfo.ru). Реализация процесса автоматизированной загрузки данных позволит проводить оперативные расчеты в режиме реального времени, единственным ограничением будет являться регламент обновления информации на сайтах внешних источников.

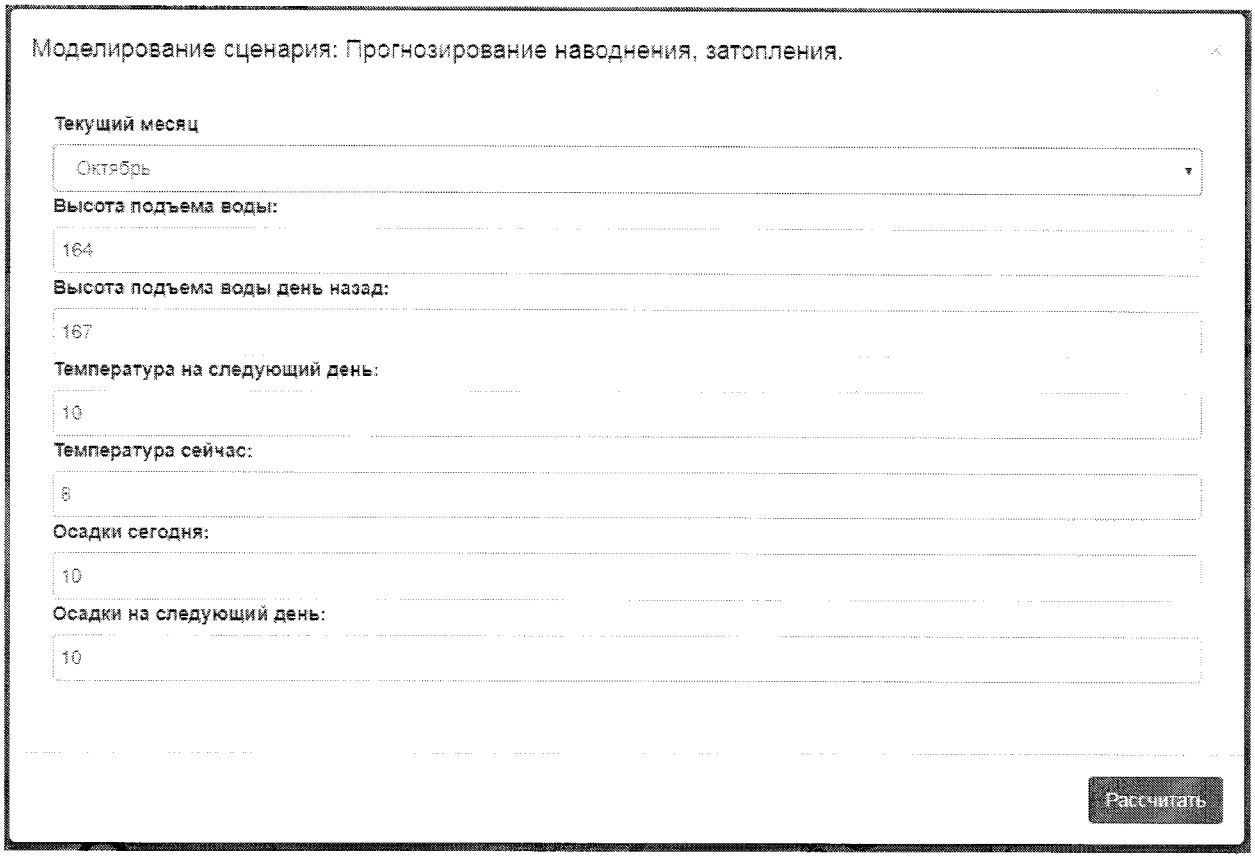

Рисунок 23 – Вид окна для ввода исходных данных

6. Расчет числа объектов (и численности населения) попавшего в прогнозируемую зону затопления (рисунок 24). Данная операция осуществляется автоматически с учетом сведений, подгруженных в базы данных.

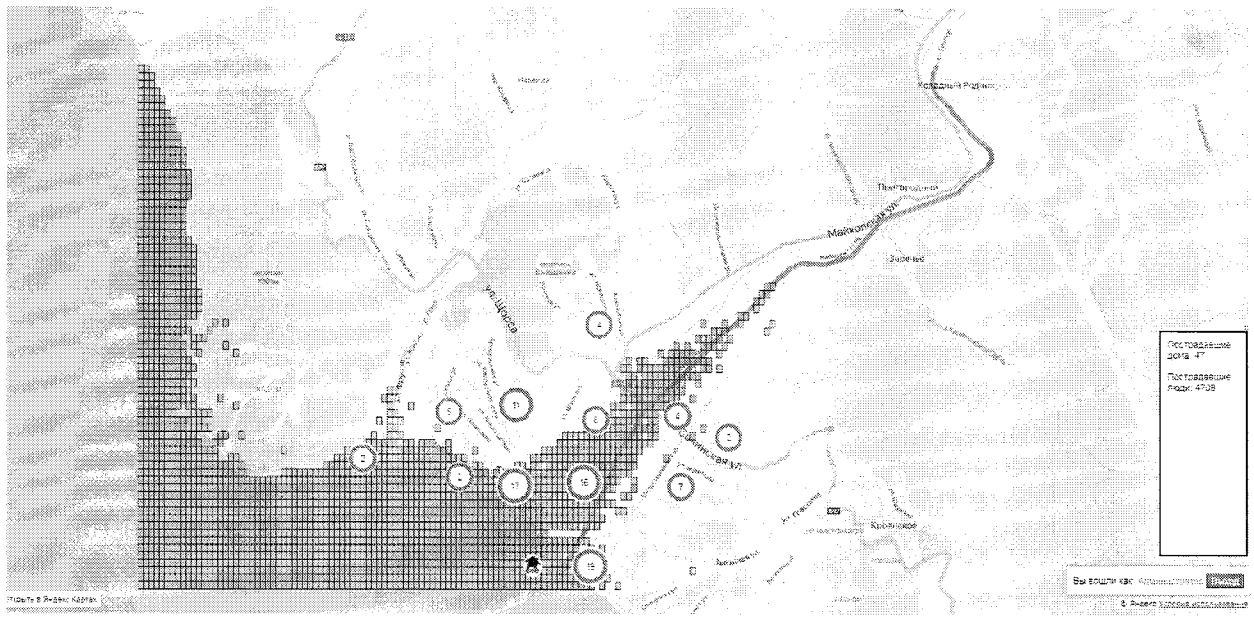

Рисунок 24 – Вид рассматриваемого участка с рассчитанным количеством объектов, попадающих в прогнозируемую зону затопления

Представленный алгоритм является существенно упрощенным и содержит сведения по основным операциям, осуществляемым для одного участка местности и для одной реки.

Данный подход в целом универсален для всех источников паводковой опасности, однако требует значительных затрат (временных, машинных) для реализации на большом количестве участков. Кроме того необходимость существования больших баз данных будет сопряжена с отвлечением значительных серверных мощностей для их формирования, хранения и применения. Базы данных (применительно к жилым объектам и объектам инфраструктуры) будут нуждаться в периодической корректировке, уточнении.

Предлагаемая модель в целом позволит осуществлять прогнозные расчеты по оценке числа объектов попадающих в зону возможного затопления, но без сопряжения с моделями прогноза подъема уровня воды будут представлять интерес только при оценке фактических последствий.

ЗАКЛЮЧЕНИЕ

В настоящих Методических рекомендациях приведены принципы построения существующих моделей и подходов к прогнозированию подъема уровня паводковых вод, вызванных дождевыми осадками. На основе приведенных моделей предложен научно-методический аппарат прогноза подъема уровня паводковых вод, вызванных дождевыми осадками и ущерба, вызванных паводками.

Предложена программная реализация предложенного научно-методического аппарата в виде модуля «Интерактивной информационной системы моделирования последствий чрезвычайных ситуаций». Показаны основные этапы программной реализации и проблемные вопросы, связанные с возникновением ошибок в моделировании и прогнозировании, которые неизбежны и обусловлены неточностью и неполнотой данных и спецификацией модели.

Одним из возможных способов для принятия оперативных решений могут служить так называемые дэшборды. Такой подход позволяет представить совокупность интерпретированной информации в виде графиков и таблиц.

Необходимо системно подходить к организации прогнозирования и предупреждения ЧС. Основные усилия сосредоточить на:

1. Обеспечение доступа к данным. Доступ к данным позволит научному сообществу находить зависимости и закономерности развития чрезвычайных ситуаций.

2. Дать возможность на «местах» видеть аналитику.

3. Моделировании и прогнозировании, в том числе, на основе применения искусственного интеллекта.

Предложенный в работе подход позволил выполнить поставленные частные задачи. Однако, в целях развития и уточнения предложенного научно-методического аппарата, необходима дальнейшая проработка вопросов,

касающихся проблем обеспечения достоверности и точности прогнозов, осуществляющихся при применении разных моделей.