МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра Промысловой океанологии и охраны природных вод

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

На тему Использование Data Mining при решении задач гидрометеорологического прогнозирования

Исполнитель Сантьева Екатерина Константиновна
(фамилия, имя, отчество)

Руководитель доктор географических наук, профессор
(ученая степень, ученое звание)

Малинин Валерий Николаевич
(фамилия, имя, отчество)

«К защите допускаю»
Заведующий кафедрой

(подпись)

кандидат физико-математических наук, доцент
(ученая степень, ученое звание)

Ерёмина Татьяна Рэмовна
(фамилия, имя, отчество)

«19» июня 2017г.

Санкт–Петербург
2017

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра Промысловой океанологии и охраны природных вод

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

На темуИспользование DataMiningпри решении задач гидрометеорологического прогнозирования

Исполнитель Сантьева Екатерина Константиновна
(фамилия, имя, отчество)

Руководитель доктор географических наук, профессор
(ученая степень, ученое звание)

Малинин Валерий Николаевич
(фамилия, имя, отчество)

«К защите допускаю»
Заведующий кафедрой

_____
(подпись)

кандидат физико-математических наук, доцент
(ученая степень, ученое звание)

Ерёмина Татьяна Рэмовна
(фамилия, имя, отчество)

«___»_____20__г.

Санкт–Петербург
2017

Содержание

Введение

Актуальность работы

В данной выпускной квалификационной работе рассматривается использование DataMining для решения гидрометеорологического прогнозирования. Из разнообразных методов, которые составляют различные методы прогнозирования, моделирования и классификации, чья основа базируется на использовании искусственных нейронных сетей, эволюционного программирования, нечёткой логики, деревьев решений, ассоциативной памяти, было решено проверить возможности технологии прогноза дерева решений, а точнее, деревьев классификации.

Для определения эффективности и точности метода деревьев классификации проведено сравнение с классическим методом регрессионного анализа, таким как множественная линейная регрессия.

Методы DataMining широко применяются в различных сферах деятельности, таких как прикладная экономика, социология, при разработке искусственного интеллекта и в прочих областях. В данных областях методы DataMining зарекомендовали себя как эффективные, а полученные результаты достаточно легко интерпретируются.

Актуальность рассматриваемого метода заключается в том, что в настоящее время крайне редко применяется для гидрометеорологических прогнозов. Метод деревьев классификаций является перспективным благодаря простоте реализации и наглядности представляемой информации.

Цель работы

Целью выпускной квалификационной работы является проверка возможности применения метода DataMining для решения гидрометеорологических задач.

Задачи работы

I. Рассмотреть теоретические основы и особенности DataMining и конкретно деревьев классификации.

II. Разбить исходные ряды на две выборки — зависимую и независимую. В качестве независимой выборки взять часть ряда за 5 последних лет для выборок около 30 лет и за последние 9 лет для выборки 59 лет. Значения рядов за предыдущее время будут зависимыми выборками.

III. Построить деревья классификации зависимости межгодового изменения уровня моря в Кронштадте (далее просто уровня моря в Кронштадте) от других гидрометеорологических данных с 1976 г по 2007 г с уровнем моря в Хельсинки и без него, а также с 1950 г по 2007 г без учёта осадков и температуры воздуха по зависимым выборкам.

IV. Построить дерево классификации зависимости стока реки Печоры от осадков в районе водосбора по зависимой выборке.

V. Рассчитать модели множественной линейной регрессии для уровня моря в Кронштадте.

VI. Рассчитать модель множественной линейной регрессии для стока реки Печора.

VII. Вычислить стандартные ошибки для зависимой и независимой выборки по получившимся моделям.

VIII. Сравнить модели деревьев классификации с оптимальной моделью множественной линейной регрессии для уровня моря в Кронштадте.

IX. Сравнить модели деревьев классификации с оптимальной моделью множественной линейной регрессии для стока реки Печора.

X. Построить графики рассчитанных значений по получившимся деревьям классификации и по модели МЛР для зависимой и независимой выборки, совмещённые с фактическими значениями уровня моря в Кронштадте.

XI. Построить графики рассчитанных значений по получившимся деревьям классификации и по модели МЛР для зависимой и независимой выборки, совмещённые с фактическими значениями стока реки Печора.

Исходные данные

Для расчёта уровня моря в Кронштадте.

Температура воздуха над Санкт-Петербургом, осадки над Санкт-Петербургом были взяты из гидрометеорологического архива CDAS-1 (http://iridl.ldeo.columbia.edu) за период с 1976 по 2007 год. Также для расчётов были взяты следующие характеристики за период с 1950-2007 года: зональная составляющая ветра (59.99°с.ш. 30°в.д.), атмосферное давление (61.9° с.ш. 26.25° в.д.), уровень моря в Хельсинки, уровень моря в Кронштадте, сток реки Нева (ст. в Новосаратовке), межгодовые колебания уровня Мирового океана (http://www.cmar.csiro.au, далее уровень Мирового океана) и Северо-Атлантическое колебание (ttps://www.ncdc.noaa.gov).

Для расчёта стока реки Печора.

Данные с 1982 по 2012 года по стоку реки Печора (ст. в Оксино) и данные по осадкам с шести станций (Усть-Уса, Троицко-Печорское, Усть-Цильма, Нарьян-Мар, Ираёль, Петруль) в районе водосбора реки Печора за летний и зимний период (http://aisori.meteo.ru/).

1 Использование DataMining для решения гидрометеорологических задач

1.1 Определение DataMining

Под термином DataMining подразумевают множество методов, применяемые для нахожденияв данных сведений, которые неизвестны, специфичны, легко интерпретируемы и обладают практической пользой. Базовые принципы работы технологии DataMining основаны на модели паттернов (шаблонов), которые представляют собой отражение механизмов многосторонних отношений данных. В подвыборках анализируемых данных можно выявить ряд закономерностей, какие и будут выражены в данных паттернах посредством их предоставления в понятном пользователю виде. Формирование паттернов производится посредством механизмов, не ограниченных априорными законами структурных особенностей выборки и формы, в которой предстают изучаемые показатели.[17]

Datamining позволяет находить оригинальные пути поиска паттернов. Паттерны, что удалось сформировать подобным образом показывают неожиданные закономерности в анализируемых данных. Такие паттерны принято называть скрытым знанием. [3]

Datamining широко используется на пересечение и взаимоинтеграции различных дисциплин: статистики, прикладной экономике, социологии, методов детерминирования образов, достижений в области развития искусственного интеллекта, алгоритмов анализа массивов данных, теории выборок, индетерминантной выборки и др. Из этого следует широкий арсенал применяемых методов и инструментов. [1, 4, 5, 7, 8, 11, 13, 14, 22]

Такое разнообразие превращается систему Datamining в набор техник, включающих в себя различные подходы. Но, в основном, упор делается на какой-то один метод, далее расцениваемый как основной.

Задачи типа Datamining предполагают различные пути их решения, однако деревья решений являются одним из наиболее рациональных путей. Такой путь решения, как правило, имеет вид дерева, построенного по принципу иерархическую структуры с дихотомией логической операции «если… то». [17]

1.2 Деревья решений

Паттерны данных в своём неструктурированном виде сложная для восприятия и анализа пользователями. Для этих ценному они группируются в деревья решений.

Деревья решений — форма иерархического представления данных о структурных шаблонах. Они могут включать в себя деревья классификации и деревья регрессии. Они имеют вид иерархически построенной модели, уровни которой являются собой узлы принимаемых решений. Решения отражают оценку значения выбранных переменных. На их основании предполагаются результирующие значения. [4]

Задачи, которые позволяют выполнять деревья решений, а также механизмы, применяемые для выполнения этих задач, структурные элементные особенности деревьев решений очень сложны. Поэтому перед

рассмотрением данных вопросов следует получить общие сведения об основе модели подобного рода структур данных, т.е. деревьев решений. [4]

Первым элементом построения узла дерева решений являются входные атрибуты. На основе лежащих алгоритмов происходит их оценка с последующим выводом прогнозируемых значений.

Можно провести таксономическое деление деревьев решений на деревья классификации и деревья регрессии. Основанием деления не служат тип используемых данных. Это вызвано тем, что деревья обоих типов могут использовать или символические, или непрерывные значения.[4]

Основанием деления в действительности служат исключительно выходные значения. Так дерево, выходные значения которого непрерывны, называется регрессивным, а дерево, выходные значения которого символические, — деревом классификации.

Дерево создаётся по нисходящему принципу, т.е. сверху вниз. Процесс создания описывается рядом алгоритмов, определяющих этапы «построения» дерева («создания») и его «сокращения». Во время создания решается сумма вопросов: выбор критериев, на основании которых будет производиться процесс расщепления, пути ветвления, а также момент прекращения обучения.

Для получения нужного результата процесса сокращения решаются иначе вопросы: каким будет конечный размер нашего дерева решений, т.е. какое конечное число его ветвей будет удалено. Однако стоит отметить, что такое деление на этапы крайне относительно, ввиду существования алгоритмов, производящих создание и сокращение дерева попеременно (такой метод позволяет предотвратить увеличение объема внутренних узлов), другие же делают это последовательно.

1.3 Характеристика деревьев классификации.

1.3.1 Иерархическая природа деревьев классификации

В основе метода деревьев классификации лежит принцип последовательных вопросов, формирующих, в конечном итоге, иерархическую структуру самого дерева. Конечный итог при этом является результирующей ответов на все выше поставленные вопросы. Для упрощения понимания данного принципа можно привести аналогию: положение листа на дереве по такой же стратегии задаётся через последовательность ветвей, какие к нему ведут: от массивного ствола дерево до последней веточки, к которой и крепится лист.

Из этого следует, что иерархическая структура дерева является его основополагающей характеристикой. [22]

1.3.2 Гибкость метода деревьев классификации

Критерием, заметно отличающим деревья классификации от других форм предоставления и анализа массивов данных, является гибкость. Ранее были затронуты механизмы дифференциации и анализа изменений, вносимых отдельными интегрируемыми переменными. Они отличаются своей

поэтапностью, дающей возможность к последовательному рассмотрению эффекта данных переменных. [22]

Это лишь один из примеров, показывающих более высокую гибкость деревьев классификации в сравнении с иными методами и алгоритмами вычисления и анализа.

При этом в деревьях классификации метод одномерного ветвления по предикторным переменным не является единственным. Измерение таких переменных в интервальной шкале является достаточным основанием для использования деревьев классификации в ветвлениях по линейным комбинациям, так, как это делается в дискретном анализе. В таком линейном дискриминантном анализе максимум количества линейных дискриминантных функций ограничено. Оно составляет на один меньше минимума из множества предикторных переменных и классов зависимой переменной.

Противная ситуация в рекурсивном методе, какой применяется в модуле деревьев классификации. Здесь мы не ограничены подобным образом.

Приведем пример: в модели с двенадцатью предикторными переменными и тремя классами зависимой переменной мы способны поменять сто последовательных ветвлений по линейным комбинациям. [22]

Это предоставляет собой значительное преимущество в сравнении с классическим линейным дискриминантным анализом, лишенным рекурсивных функций. Кроме того, большая доля информации, которая составляет предикторные переменные может даже не применяться.

Предоставим другой пример. Мы имеем широкое числом категорий при малом количестве предикторов. Нам известно, что модели автомобилей имеют различную массу и максимальную скорость. Мы хотим распределить большое число моделей автомобилей, используя лишь эти критерии. В классическом линейном дискриминантном анализе мы добьемся двух дискриминантных

функций, а модели автомобилей будут распределены верно только при их различии не более чем в двух критериях, что мы заранее представили линейными комбинациям максимальной скорости и массы.

Совсем иная картина в тех механизмах, что применяются в модуле деревьев классификации. Мы совершенно не ограничены числом ветвлений по линейным комбинациям, какие можно совершить.

Инструменты ветвления по линейным, используемый в модуле деревьев классификации, с таким же успехом применяется в процессе построения деревьев классификации с одномерным ветвлением.

В действительности стоит отметить, что такой вид ветвления являет собой лишь частную разновидность ветвлениях по линейной комбинации. Это легко понять, если вообразить такой подвид ветвления по линейным комбинациям, в коем при абсолютном числе предикторных переменных весовой коэффициент нулевой. Так как в итоге конечное значение комбинации определяется в действительности значением единственной предикторной переменной (ее коэффициент при этом не должен быть равен нулю), образуемое в итоге ветвление и будет вышеназванным одномерными ветвлением.

1.4 Критерий расщепления

Как уже упоминалось, процесс создания дерева можно назвать нисходящим, ввиду того, что поэтапно он происходит буквально сверху вниз. В рамках данного процесса алгоритмом находит критерий расщепления (также

именуемый, как критерий разбиения), по основанию которого можно было бы разбить некоторое множество на подмножества так, чтобы они ассоциировались с соответствующим данному этапу узлом проверки. При этом каждый узел проверки помечается определенными атрибутом.

Разберём существующее правило выбора атрибута.

Выбранный атрибут, согласно правилу, должен разбить первоначальное множество так, чтобы элементы полученных подмножеств принадлежали к одному логическому классу, либо стремились к максимальному соответствию данного условия. Это значит, что число объектов, которые принадлежат к другим классам в образовавшихся подмножествах (такие объекты принято именовать «примесями») должно быть минимальным.

Можно выделить разные критерии расщепления. Самыми известными являются индекс Gini [6] и мера энтропии.

Брейман предлагает иной критерий, используемый как основание для расщепления — индекс Gini. Он используется в алгоритме CART. Атрибут при использовании этого алгоритма определяется по расстоянию между распределениями классов.

Пусть есть множество T, состоящее из n классов.

Индекс Gini, gini(T) определяется по формуле:

$\text{gini(T)} = 1 - \sum_{j=1}^{n} p_j{}^2$,

где T — текущий узел, p_j — вероятность класса j в узле T, n — количество классов. [16]

Если процессе создания мы получили объемное дерево с большим числом подмножеств, это не значит, что оно может удовлетворять исследовательские цели и быть удобно для работы и анализа. При большем числе частных случаев, которые упоминаются в дереве решений, маленькое

количество объектов окажется в каждом частном случае. Подобного рода деревья решений принято называть «кустистыми» или «ветвистыми». Количество ветвей и узлов в них, как правило, необоснованно велико.

Первоначальное множество в таком случае разбивается на очень большое количество подмножеств, включающих в себя малое количество объектов. Происходит процесс, именуемый «переполнением», при котором возможность обобщения значительно снижается.

Для избегания таких случаев разработан ряд алгоритмов. Благодаря их использованию удается прийти в процессе построения дерева к таковому «оптимального размера» (называемого также «подходящим деревом»).

1.5 Построение деревьев классификации

Основным законом, применяемым для такого (древовидного) построения является логическая операция «если… то». В конечном определении соотнесённости определенной ситуации или объекта его классу, необходимо последовательно отвечать на вопросы, находящиеся в иерархической цепи узлов данного дерева, начиная с корня.

При этом корнем называют первоначальный вопрос, т.н. внутренний узел дерева. Он представляет собой узел определения соответствия действительности какого-либо условия. Затем следует второй вопрос, третий и т.д. Этот процесс завершается достижением в процессе анализа конечного узла дерева, называемым также взлом решения.

Вопросы, как правило, можно представить в следующей конструкции «значение параметра А больше х?». Отсюда происходит бифуркация: при ответе «да» следующем пунктом становится узел правой ветки, если «нет», то узел левой, на которых задаются вопросы, привязанные к соответствующему узлу.

Исходя из того, что деревья имеют иерархическую структуру, вышеназванные ветвления совершаются очерёдно, от корневой вершины. Следующим шагом явления переход к вершинам потомков, т.е. к следующему уровню, цикл повторяется, пока не будет возникать ситуация ветвлениях, и мы не придем к “неразветвленной” вершине, и потомки не станут конечными.

Конечные вершины, именуемые также листьями — это такие узлы дерева классификации, после которых не существует более вопросов, и решения, соответственно, не производятся.

Это графически обозначается в программах, работающих с деревьями классификации.

К примеру программа Statistica выделяет на схематичном изображении дерева конечные вершины красной пунктирной штриховкой, с другие, именуемые решающими вершинами (узлы ветвления) — сплошной синей линией. Основанием такого дерева является самый верхний узел ветвления, именуемые также первой вершиной решения или корнем дерева.

Последовательность создания деревьев решений стоит из двух этапов: созданное дерева и его сокращение.

В рамках каждого из процессов решается ряд вопросов, определяющих конечный итог процесса. В создании дерева эти вопросы определения критерия, по которому будет происходить расщепление, и точку остановки обучения (последний пункт выполняется только в солнечную, если он

предусмотрен заданным алгоритмом). В сокращении определяется, какие ветви будут отсечены.

1.6 Стратегии построения деревьев классификации оптимальных размеров

Рассмотрим две стратегии, позволяющие создать дерево оптимальных размеров.

В первой стратегии дерево наращивается до размеров, которые были заданы параметрами, определяемыми заранее пользователем. Выявление этих параметров строится на опыте использования алгоритмов построения деревьев решения конкретным пользователем. Кроме того, некоторые программы, использующие такие вычислительные алгоритмы, обладают функцией “диагностических сообщений”, при помощи которых упрощается принятие решения о размере дерева аналитиком.

Во второй стратегии определение подходящего размера дерева производится при помощи ряда разработанных алгоритмов. Они были впервые представлены в работах Бримана, Куиленда и др. в 1984 году. Сами же авторы в комментариях к своим работам утверждают, что освоение этих алгоритмов очень тяжело для начинающего пользователя. Эти алгоритмы включает в себя процесс отсечения ветвей, применение правил прекращения обучения. Обе операции ведут к сокращению размера дерева.

Однако стоит заметить, что такие алгоритмы крайне разнообразны, и не работают по единственной схеме. Ряд алгоритмов слились из двух самостоятельных последовательных этапов: сначала идёт построение дерева,

затем его сокращение. Иные алгоритмы же попеременно применяют эти процессы в своей работе. Это создано, как правило, для снижения вероятности необоснованного наращивания внутренних узлов.

1.7 Остановка построения дерева.

Разберем правило остановки. Его сущность заключается в определении природы рассматриваемого узла: является ли он внутренним узлом, который должен быть разбит далее на подмножества, либо это конечный узел, который также называется узлом решения.

Остановкой называют момент в создания дерева, после которого необходимо прекратить образования новых узлов (ветвление).

Варианты правил остановки:

Ранняя остановка — она определяет, как правило, необходимость дальнейшего ветвления. Главное достоинство такого механизма — значительное снижение времени, которое требуется для обучения самой модели. Главный недостаток — повышается вероятность неточности классификации. По этой причине более эффективным является метод отсечения.

Ограничение глубины дерева. После достижения заранее определенной глубины построение прекращается.

Определение минимального числа примеров, из которых будут состоять конечные узлы дерева. Этот метод предполагает, что прекращение ветвления будет происходить после наступления момента, при котором все конечные

узлы будут чистыми, либо если число объектов, содержащееся в них не будет превышать таковое заданное изначально.

Число методик, при помощи которых достигается оптимальный размер дерево не ограничивается тремя вышеназванными, но, стоит выделить, что их утилитарная ценность круге низкая, а некоторые могут быть использованы лишь в ряде частных случаев.

1.8 Алгоритмы ветвления деревьев решений. Особенности алгоритма CART

Сегодня мы можем назвать широкое число различных алгоритмов, использующих методы деревьев решений: CART, CN2, C4.5, NEWId, CHAID, ITrule и т.д. [19]

Рассмотрим алгоритмы ветвления, используемые в программе Statistica. [12]

Всего в программе представлено три алгоритма.

Дискриминантное одномерное ветвление для категориальных и порядковых предикаторов (QUEST). Данный алгоритм применяется при условии, что все независимые переменные категориальные либо порядковые, либо являют собой сочетание этих двух типов. Порядковыми переменными в таком случае называют все возможные количественные переменные.

Дискриминантное ветвление по линейным комбинациям порядковых предикторов. Такой алгоритм применим в случае, если для анализа в качестве переменных выбраны только порядковые.

Полный перебор для одномерных ветвлений методом CART. Он применяется для порядковых, категориальных независимых переменных, либо

в ситуации, когда используются оба типа. В отличие от вышеназванных методов здесь при поиске наиболее успешного варианта последовательно производится перебор всех возможных вариантов комбинаций независимых переменных. Как правило, количество этих вариантов оказывается очень большим, что значительно увеличивает продолжительность самой операции, а дерево решений в таком случае получится очень сложным для понимания и восприятия. [22]

Алгоритм CART создаёт по два потомка в каждом узле ветвления дерева решений.

В каждом шаге происходить ветвления по принципу построенного правила, сформированного в узле. Множество объектов, рассматриваемое эти правилом, делится на две ветви подмножеств — правую ветвь, в которой выполняется правило (потомок — right), и левую ветвь, в которой пропилом, соответственно, не выполняется (потомок — left). Для определения оптимального правила применяется функция, при помощи которой производится оценка качества разбиения.

Обучение дерева решений можно назвать обучением с учителем. Отсюда следует, что как обучающая, так и тестовая выборки состоят из классифицированного набора примеров.

Интуитивный принцип снижения объема нечистоты в узле лежит в основе оценочной функции, применяемой алгоритмом CART.

Кроме базовых алгоритмов, описанных выше, есть также из доработанные версии — алгоритмы DB-CART и IndCART.

Рассмотрим каждый из них.

DB-CART — это такой алгоритм, в основу которого положена концепция использования обучающего набора данных не для определения разбиений, а для оценки распределения входных и выходных значений. И уже

полученную информацию применяется в определении разбиения. DB расшифровывается как «distributionbased». Предполагается, что такой усовершенствованный алгоритм позволяет значительно уменьшить ошибку классификации, в сравнении с классическими алгоритмами построения дерева решений.

Алгоритм IndCART, включенный в пакет Ind, отличается от алгоритма CART применением иного метода обработки опущенных данных, не производить регрессионный этап алгоритма CART, обладает другими параметрами отсечения.

Главные отличительные черты алгоритма CART от алгоритмов семейств ID3:

— бинарное представление дерева решений;

— инструмент оценки качества разбиения;

— применение алгоритмов отсечения;

— алгоритм анализа пропущенной информации;

— создание деревьев регрессии.

1.9 Преимущества деревьев классификации

Интуитивность. Полученная в итоге обработки данных модель в виде дерева решений интуитивно понятна и проста в понимании решения поставленной задачи. Пользователи легко осознают и обобщает полученное знание при анализе такого дерева, в отличие, например, от метода нейронных

сетей, где результаты, как правило, требуют дополнительной обработки для упрощения понимания их исследователем. Данное преимущество не только позволяет быстро и точно определить к какому классу объектов относится тот или иной новый объект исследования, но выгодно для интерпретации и осознания сформировавшейся модели в целом. Задача объяснения по каким критериям тот или иной объект исследования принадлежит именно к определенному классу становится интуитивной и легко объяснимой. [3]

Деревья классификации позволяют формировать правила из массива данных на естественном языке. [18]

Деревья классификации высокоэффективны в тех сферах исследовательской деятельности, где весьма сложным является аналитическая формализация знания.

Сам алгоритм деревьев решений не требует от исследователя самостоятельного определения входных атрибутов. Таким образом пользователь может задействовать все имеющиеся атрибуты, а алгоритм уже самостоятельно определит, какие из них наиболее значимы, и на их основе строить дерево.[3]

Модели, созданные с использованием метода деревьев решений, обладают точностью, сопоставимой с иными механизмами моделей классификации (нейронные сними, статистические модели).

Создано множество масштабируемых алгоритмов, какие используются для построения деревьев решения на массивных базах данных. Масштабируемость в данном случае стоит понимать, как наличие линейной зависимости числа примеров или записей определенной базы данных и времени, за которое данная информация будет обработана. [3]

Быстрое обучение. Время, затрачиваемое на создание моделей при помощи деревьев классификации, значительно меньше, черт таковое с использованием алгоритмов нейронных сетей. [3]

Большая часть всех применяемых сегодня деревьев решений имеет функцию специального анализа пропущенных значений.

При работе с деревьями классификации можно использовать как числовые, так и категориальные данные. В традиционных статистических методах можно использовать только числовые данные.

Большая часть используемых ныне классических статистических методов предполагает, что исследователь изначально знает определенную информацию о гипотезе, виде рассматриваемой модели, обладать предположениями о ток форме зависимости, какая возникнет между данными. Деревья классификации, напротив, создают непараметрические модели. Отсюда следует, что деревья классификации являются эффективным инструментом решения задач, в которых исследователь не обладает априорными знаниями о характере выявляемой зависимости исследуемых объектов.

Резюмируя можно перечислить основные преимущества деревьев классификации [10]:

— простота обучения работы с деревьями классификации;

— простота интерпретации полученных результатов с помощью деревьев решений;

— точность расчётов, сопоставимая с другими методами (статистические модели, нейронные сети);

— можно использовать на сверхбольших массивах данных;

— строят непараметрические модели, т.е. возможно решать такие задачи, где нет априорной информации о виде зависимости между исследуемыми данными;

— имеется возможность специальной обработки данных, с отсутствующими значениями.

2 Модель межгодовых колебаний уровня моря в Кронштадте на основе алгоритма деревьев решений

2.1 Построение деревьев классификации

Исходные данные для прогнозирования уровня моря в Кронштадте были поделены на зависимую и независимую выборку, причём независимая выборка для периода с 1976 по 2007 год составляла 5 лет, а для периода с 1950 по 2007 год — 9 лет. С помощью деревьев классификации был рассчитан уровень моря в Кронштадте в зависимости от других гидрометеорологических характеристик за период с 1976 по 2002 год и за период с 1950 по 1998 год без учёта температуры воздуха и осадков. Результаты представлены на Рис. 1 и 2, соответственно.

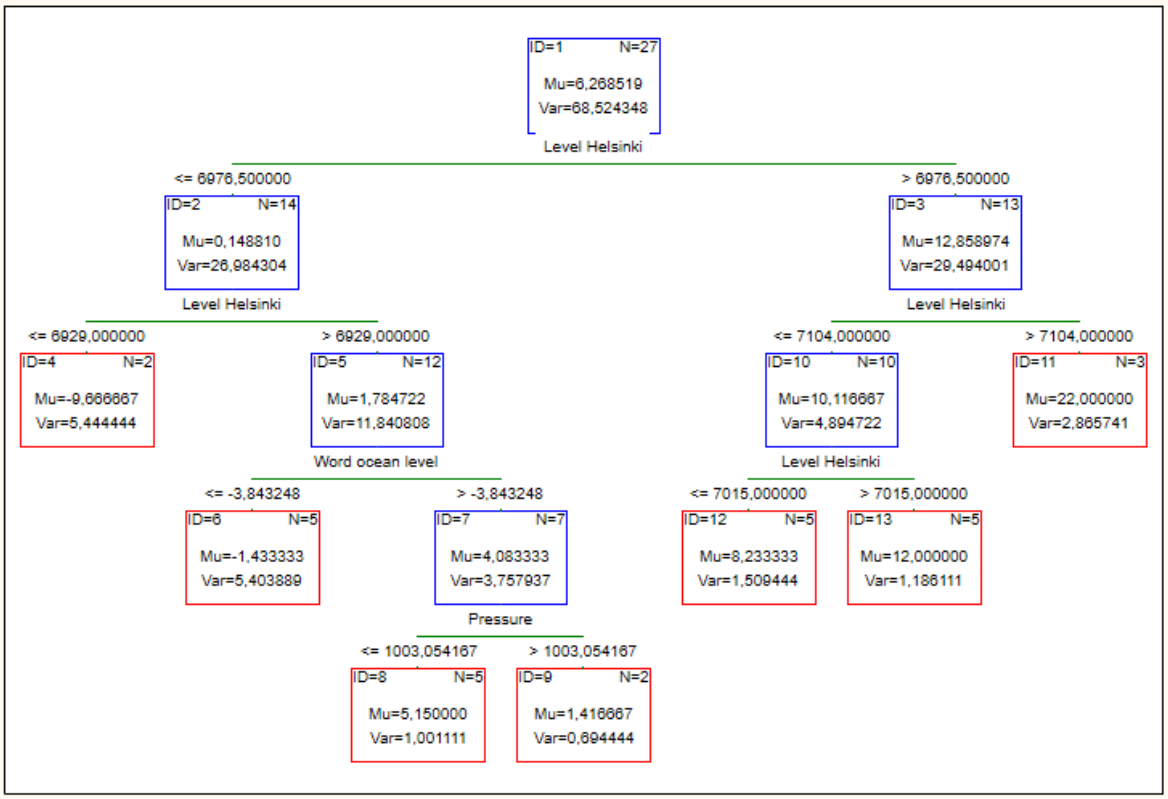


Рис. 1 Дерево классификации связи уровня моря в Кронштадте с другими гидрометеорологическими характеристиками (с учётом уровня моря в Хельсинки)за период с 1976 по 2002 год, зависимая выборка

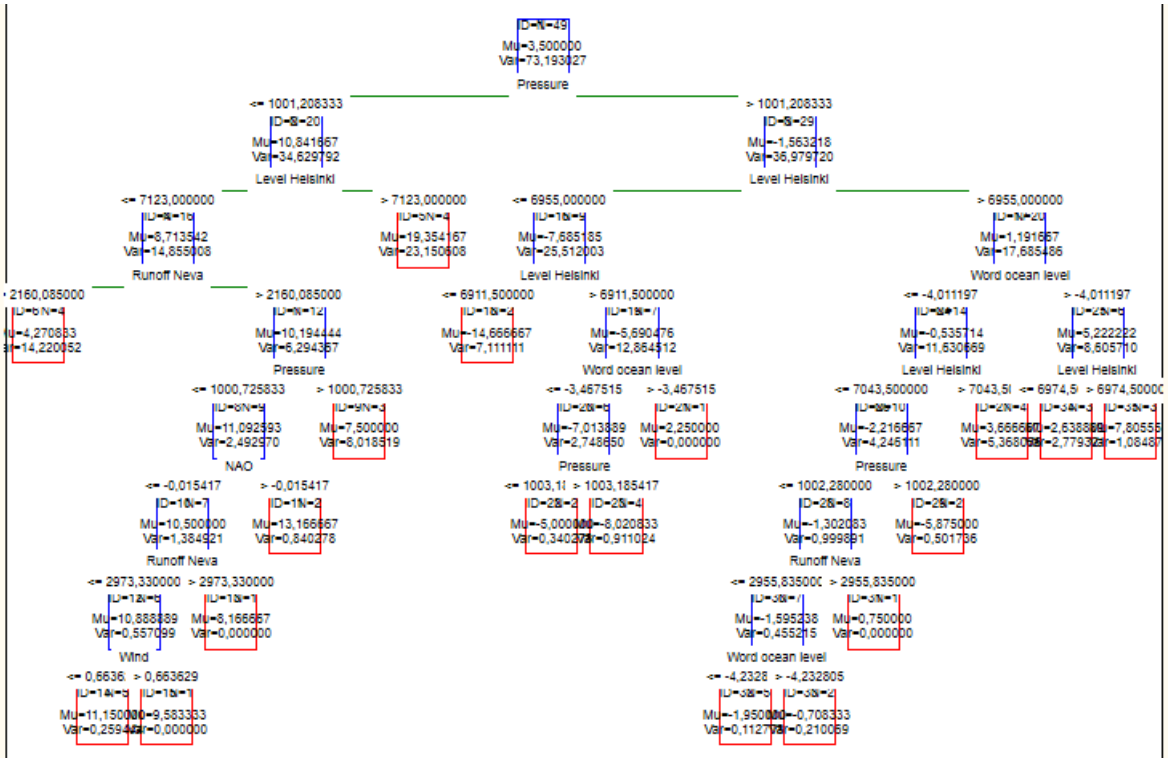


Рис. 2 Дерево классификации связи уровня моря в Кронштадте с другими гидрометеорологическими характеристиками (без учёта температуры воздуха и осадков) за период с 1950 по 1998 год, зависимая выборка

Как видно из Рис. 1, дерево классификации имеет семь терминальных вершин, шесть нетерминальных вершин, то есть всего тринадцать. На Рис. 2 видно, что дерево классификации для большего временного промежутка увеличилось более чем в два раза (восемнадцать терминальных вершин, семнадцать нетерминальных, всего – тридцать пять). Можно сделать вывод, что увеличение длины временного ряда значительно увеличит и число ветвлений, то есть структура дерева классификации усложняется.

В первом дереве классификации в четырёх ветвлениях из шести в качестве характеристики, по которой происходит ветвление, присутствует уровень моря в Хельсинки. Также ветвление происходит по уровню

Мирового океана и атмосферному давлению.Это указывает на то, что при построении прогностической модели по данному дереву решений будет использоваться три переменных из восьми. На самом деле в этом нет ничего удивительного, так как изменения уровня моря в Хельсинки и Кронштадте высоко коррелируются, так как географически они расположены близко друг другу (в пределах одного моря). Уровень моря в Кронштадте, несомненно, напрямую зависит от уровня Мирового океана, но так как он подвержен сильному осреднению, то и корреляция между ними меньше чем между уровнем моря в Кронштадте и Хельсинки. Изменение давления также сильно влияет на исследуемую характеристику.

Совершенно иная картина для второго дерева классификации — все используемые характеристики для построения дерева присутствуют хотя бы по одному разу в ветвлениях. Несмотря на то, что уровень моря в Хельсинки присутствует в наибольшем количестве ветвлений (пяти из семнадцати), но в соотношении данная характеристика используется реже чем в первом дереве. Высоко влияние атмосферного давления и уровня Мирового океана, которые есть в четырёх и трёх ветвлениях. Также в трёх ветвлениях имеется сток Невы.

Так как в первом дереве преобладают ветвления с уровнем моря в Хельсинки, то необходимо построить дерево без учёта уровня моря в Хельсинки.

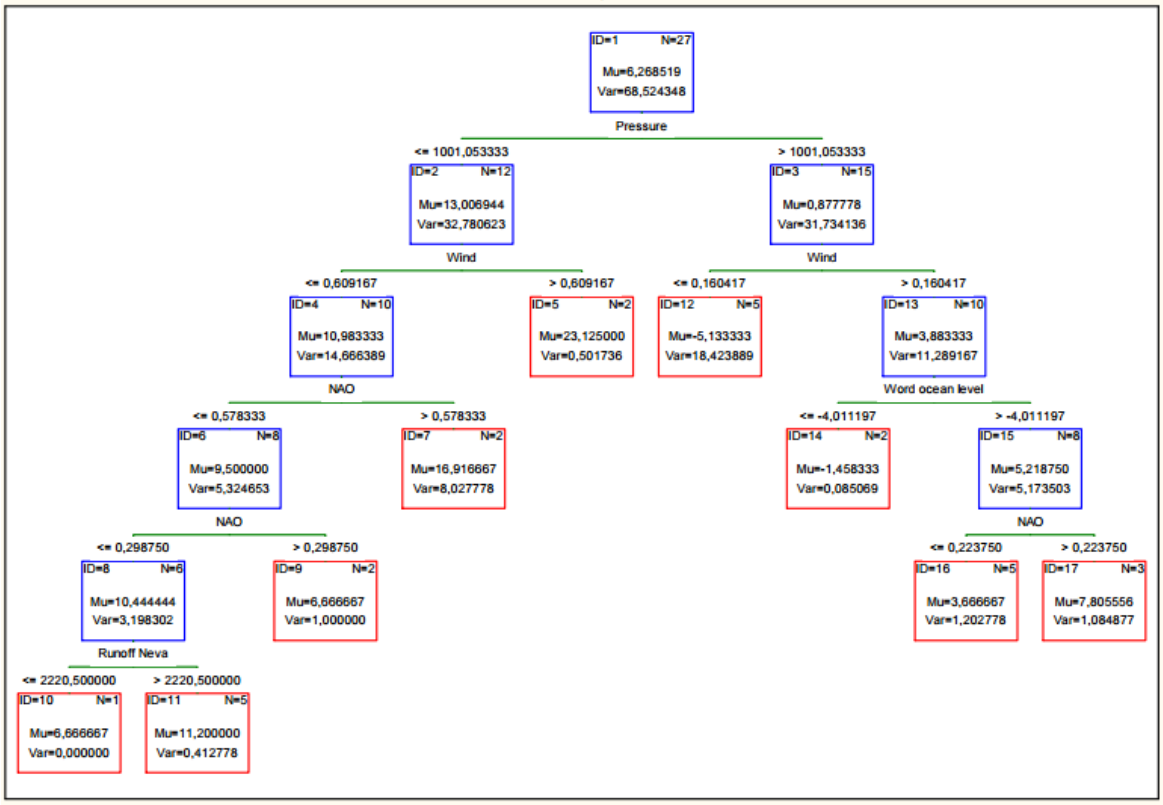


Рис. 3 Дерево классификации связи уровня моря в Кронштадте с другими гидрометеорологическими характеристиками (без учёта уровня моря в Хельсинки) за период с 1976 по 2002 год, зависимая выборка

Как видно из Рис. 3, данное дерево классификации имеет девять терминальных вершин, восемь нетерминальных, всего семнадцать вершин. Это больше чем в первом дереве. Как и в первом дереве, в ветвлении отсутствуют температура воздуха и осадки, что указывает на то, что их внутренняя структура слабо связана с уровнем моря в Кронштадте.

В трёх узлах ветвления из восьми присутствует Северо-Атлантическое колебание, в двух — зональный ветер, а в оставшихся трёх — давление, уровень Мирового океана и сток Невы.

На примере последнего дерева рассмотрим, как происходит ветвление в дереве классификации. Как видно на Рис. 3, на первом ветвлении разделителем является давление атмосферы. Если Давление больше 1001.05 гПа, то прогнозируемый уровень моря в Кронштадте выше среднего (6.27 см)

равен 0.88 см, если меньше — 13.01 см. Здесь и далее «Mu» — среднее значение, «Var»—дисперсия прогнозируемого ряда уровня моря в Кронштадте. На втором ветвлении происходит уточнение 12 значений за счёт зонального ветра: если скорость ветра больше 0.61 м/с см, то уровень моря в Кронштадте будет равен 23.13 см, что больше среднего значения уровня моря в Кронштадте (13.01 см), а если скорость зонального ветра меньше 0.61 см, то уровень моря в Кронштадте будет равен —10,98 см, что, соответственно, ниже среднего значения. Далее, по тому же принципу, происходит дальнейшее ветвление.

Как видно, все деревья достаточно большие, поэтому необходимо найти оптимальный вариант. В программе Statistica [12] используется алгоритм CART, где для поиска оптимального дерева осуществляется следующим способом. Из таблицы выбираем минимальную цену кросс-проверки (CV-cost), прибавляем соответствующую стандартную ошибку (CVstd. error).

Данная сумма используется как пороговую величину CVкрит. После этого перебираются деревья в сторону уменьшения трудности, а оптимальное дерево то крайнее дерево, у которого значение CV-cost меньше CVкрит. Так достигается оптимальность дерева по соотношению «сложность/цена».

По данному алгоритму найдено оптимальное дерево для каждого варианта.

Таблица 1. Оценки кросс-проверки деревьев классификации для изменения уровня моря в Кронштадте с 1976 по 2002 г, зависимая выборка (с учётом уровня моря в Хельсинки)

Показатель	Число терминальных вершин	Цена кросс-проверки (CV-cross)	Стандартная ошибка кросс-проверки (CV std. error)	Цена проверки на обучающейся выборке	Сложность узла
Дерево 1	7	27.85	9.08	2.46	0.00
Дерево 2	6	30.24	11.12	3.20	0.74
Дерево 3	5	34.71	11.40	4.51	1.31
Дерево 4	4	35.68	9.84	7.80	3.29
Дерево 5	3	42.91	8.86	16.12	8.33
Дерево 6	2	45.54	9.93	28.19	12.07
Дерево 7	1	71.02	19.03	68.52	40.33

Как видно из Таблицы 1, для первого варианта дерева классификации минимальная цена кросс-проверки у дерева 1 с 7 терминальными вершинами, то есть шестью ветвлениями, и составляет 27.85, следовательно, CVкрит=36.93. Последнее дерево, у которого CV-cross<CVкрит, - это дерево 4, с тремя ветвлениями.

Таблица 2. Оценки кросс-проверки деревьев классификации для изменения уровня моря в Кронштадте с 1950 по 1998 г, зависимая выборка (без учёта температуры воздуха и осадков)

Показатель	Число терминал-ных вершин	Цена кросс-проверки (CV-cross)	Стандартная ошибка кросс-проверки (CV std. error)	Цена проверки на обучающейся выборке	Сложность узла
Дерево 1	18	35.30	6.95	4.70	0.00
Дерево 2	17	34.96	6.76	4.74	0.04
Дерево 3	16	35.03	6.72	4.78	0.04
Дерево 4	15	35.36	6.74	4.88	0.10
Дерево 5	14	35.21	6.62	5.01	0.13
Дерево 6	13	34.87	6.46	5.24	0.23
Дерево 7	12	36.13	6.90	5.48	0.25
Дерево 8	11	37.55	7.10	6.08	0.59
Дерево 9	10	37.05	7.09	6.76	0.68
Дерево 10	9	37.66	6.82	7.58	0.82
Дерево 11	8	38.44	6.93	9.08	1.50
Дерево 12	7	39.92	6.89	11.10	2.02
Дерево 13	6	42.17	7.00	13.25	2.15
Дерево 14	5	41.97	7.02	15.80	2.56
Дерево 15	4	43.69	7.13	18.64	2.84
Дерево 16	3	56.97	9.58	26.04	7.39
Дерево 17	2	54.11	9.25	36.02	9.98
Дерево 18	1	76.37	15.16	73.19	37.17

Как можно увидеть из Таблицы 2, для второго варианта дерева классификации минимальная цена кросс-проверки у дерева 6 с двенадцатью ветвлениями, и составляет 34.87, следовательно, CVкрит=41.33. Последнее дерево, у которого CV-cross<CVкрит, - это дерево 12, с шестью ветвлениями.

Таблица 3. Оценки кросс-проверки деревьев классификации для изменения уровня моря в Кронштадте с 1976 по 2002 г, зависимая выборка (без учёта уровня моря в Хельсинки)

Показатель	Число терминальных вершин	Цена кросс-проверки (CV-cross)	Стандартная ошибка кросс-проверки (CV std. error)	Цена проверки на обучающейся выборке	Сложность узла
Дерево 1	9	48.04	12.65	4.54	0.00
Дерево 2	8	46.47	12.05	5.18	0.63
Дерево 3	7	46.01	12.71	5.97	0.79
Дерево 4	6	50.53	13.86	7.16	1.19
Дерево 5	5	57.69	13.89	9.80	2.64
Дерево 6	4	58.71	14.21	13.06	3.26
Дерево 7	3	58.99	11.29	22.16	9.10
Дерево 8	2	56.81	10.84	32.20	10.04
Дерево 9	1	71.02	19.03	68.52	36.33

Как можно увидеть из Таблицы 3, для третьего варианта дерева классификации минимальная цена кросс-проверки у дерева 3 с шесть ветвлениями, и составляет 46.01, следовательно, CVкрит=58.72. Последнее дерево, у которого CV-cross<CVкрит, - это дерево 8, с одним ветвлением.

В дальнейшем сравним оптимальные деревья, найденные по кросс-проверке и с помощью стандартных ошибок.

2.2 Расчёт множественной линейной регрессии

Для каждого дерева классификации была рассчитана соответствующие ему модели множественная линейная регрессии от полной до одной переменной и определена оптимальная модель.[2]

Таблица 4. Характеристики качества моделей для изменения уровня моря в Кронштадте с 1976 по 2002 г, зависимая выборка(с учётом уровня моря в Хельсинки)

Номер шага	Кол-во пред-ов	R^2	F*	σ_E	t^*_{min}	$F_{кр}$	σ_y	$t_{кр}$	0.67*σ_y
1	8	0.96	54.92	2.01	0.28	2.48	8.44	2.09	5.65
2	7	0.96	65.95	1.96	0.68	2.51		2.09	
3	6	0.96	78.96	1.94	1.26	2.57		2.08	
4	5	0.96	91.90	1.96	1.44	2.66		2.07	
5	4	0.95	109.02	2.01	1.14	2.80		2.07	
6	**3**	**0.95**	**143.12**	**2.02**	**3.07**	**3.01**		**2.06**	
7	2	0.93	155.40	2.35	5.59	3.39		2.06	
8	1	0.84	126.63	3.49	11.25	4.23		2.06	

Чтобы определить модель, где все коэффициенты значимы, необходимо сравнить t^*_{min} с$t_{кр}$, если$|t^*_{min}|$>$t_{кр}$, то нулевая гипотеза отвергается и коэффициент регрессии значим. [20] А так как t^*_{min} относится к той независимой переменной, которая у которой наименьший критерий Стьюдента, то все остальные переменные будут иметь больший критерий Стьюдента, а значит, все переменные значимы. Данному критерию соответствует модель с тремя (3.07>2.06), двумя (5.59>2.06) и одним (11.25>2.06) предиктором.

Помимо этого, в качественной модели коэффициент детерминации должен быть больше 0.7, что соответствует всем трём выше указанным моделям, наибольший коэффициент детерминации в модели с тремя переменными (R^2= 0.95).

Все из трёх моделей адекватны, так как F*>$F_{кр}$ (модель с тремя переменными 143.12>3.01, с двумя — 155.4>3.39, с одной — 126.63>4.23).

Также для всех моделей стандартная ошибка (σ_E) меньше 0.67*σ_y стандартного отклонения зависимой переменной (для модели с тремя переменными 2.02<5.65 см, с двумя — 2.35<5.65см, с одной — 3.49<5.65 см).

Из всего выше описанного можно сделать вывод, что модели с тремя, двумя и одной переменной являются качественными, так как соответствуют всем необходимым параметрам.

Из этих трёх моделей в качестве оптимальной выбрана модель с тремя переменными, так как у неё наибольший коэффициент корреляции и наименьшая стандартная ошибка модели.

Для оптимальной модели ниже записано уравнение и дан график рассчитанного по модели уровня моря в Кронштадте, совмещённый с фактическими значениями.

В Приложении 2 аналогично дано уравнение и соответствующий график для остальных моделей. Условные обозначения для переменных даны там же.

$H_{Кр}$= — 644.39 + 0.09$H_{Х}$ + 6.22$H_{МО}$ + 1.89W_{zon}

Рис. 4 Совместный график фактического уровня моря в Кронштадте и рассчитанного по оптимальной модели МЛР за период с 1976 по 2007 год (с учётом уровня моря в Хельсинки), зависимая и независимая выборки отделены вертикальной линией

Как видно из Рис. 4, рассчитанные значения по оптимальной модели полностью повторяют ход фактических значений уровня моря в Кронштадте, расходясь лишь незначительно в некоторых пиках. Причём, следует отметить, что это характерно как для зависимой, так и для независимой выборки. Таким образом, можно сделать вывод, что данная модель хорошо подходит для прогнозирования уровня моря в Кронштадте.

Таблица 5. Характеристики качества моделей для изменения уровня моря в Кронштадте с 1950 по 1998 г, зависимая выборка (без учёта температуры воздуха и осадков)

Номер шага	Кол-во пред-ов	R^2	F*	σ_E	t^*_{min}	$F_{\text{кр}}$	σ_y	$t_{\text{кр}}$	0.67*σ_y
1	6	0.96	159.24	1.90	-0.03	2.32	8.64	2.02	5.79
2	5	0.96	193.63	1.87	-1.27	2.43		2.02	
3	**4**	**0.96**	**240.85**	**1.89**	**2.72**	**2.58**		**2.01**	
4	3	0.95	279.01	2.02	4.43	2.81		2.01	
5	2	0.93	291.13	2.39	13.32	3.20		2.01	

6	1	0.64	85.17	5.21	9.23	4.04		2.01	

Оптимальная модель определяется аналогично методом, описанным выше.

Для того, чтобы определить модель, где все коэффициенты значимы, необходимо сравнитькритерий Стьюдентаt^*_{min} с$t_{кр}$, где$|t^*_{min}|$должно быть больше$t_{кр}$. Данному критерию соответствует модель с четырьмя (2.72>2.01), тремя (4.43>2.01), двумя (13.32>2.01) и одним (9.23>2.01) предиктором.

Помимо этого, в качественной модели коэффициент детерминации должен быть больше 0.7, что соответствует первым трём выше указанным моделям, наибольший коэффициент детерминации в модели с четырьмя переменными (R^2= 0.96). В последней модели с одним коэффициентов коэффициент детерминации равен 0.64, то есть данная модель недостаточно описывает дисперсию исходного ряда.

Все из четырёх моделей адекватны, так как F*>$F_{кр}$ (модель с четырьмя переменными 240.85>2.58, с тремя — 279.01>2.81, с двумя — 291.13>3.20, с одной — 85.17>4.04).

Также для всех моделей стандартная ошибка (σ_E) меньше 0.67*σ_y стандартного отклонения зависимой переменной (для модели с четырьмя переменными 1.89<5.79 см, с тремя — 2.02 <5.79 см, с двумя — 2.39<5.79 см, с одной — 5.21<5.79 см).

Из всего выше описанного можно сделать вывод, что модели с четырьмя, тремя и двумя переменными являются качественными, так как соответствуют всем необходимым параметрам.

Из этих трёх моделей в качестве оптимальной выбрана модель с четырьмя переменными, так как у неё наибольший коэффициент корреляции и наименьшая стандартная ошибка модели.

Для оптимальной модели ниже записано уравнение и дан график рассчитанного по модели уровня моря в Кронштадте, совмещённый с фактическими значениями.

$$H_{Кр} = — 666.58 + 0.1H_{Х} + 0.002Run + 1.65H_{МО} + 5.3W_{zon}$$

Рис. 5 Совместный график фактического уровня моря в Кронштадте и рассчитанного по оптимальной модели МЛР за период с 1950 по 2007 год (без учёта температуры воздуха и осадков), зависимая и независимая выборки отделена вертикальной линией

Как видно из Рис. 5, рассчитанные значения по оптимальной модели полностью повторяют ход фактических значений уровня моря в Кронштадте, расходясь лишь незначительно в некоторых пиках. Причём, следует отметить, что для независимой выборки отклонения от фактических значений почти нет. Таким образом, можно сделать вывод, что данная модель хорошо подходит для прогнозирования уровня моря в Кронштадте.

Таблица 6. Характеристики качества моделей для изменения уровня моря в Кронштадте с 1976 по 2002 г, зависимая выборка (без учёта уровня моря в Хельсинки)

Номер шага	Кол-во пред-ов	R^2	F*	σ_E	t^*_{min}	$F_{кр}$	σ_y	$t_{кр}$	0.67*σ_y
1	7	0.92	31.77	2.77	0.78	2.51	8.44	2.09	5.65
2	6	0.92	37.69	2.74	0.72	2.57		2.08	
3	**5**	**0.92**	**46.19**	**2.71**	**2.30**	**2.66**		**2.07**	
4	4	0.90	47.19	2.96	2.48	2.80		2.07	
5	3	0.87	49.70	3.28	3.06	3.01		2.06	
6	2	0.81	51.87	3.81	3.72	3.39		2.06	
7	1	0.70	59.35	4.68	-7.70	4.23		2.06	

Для того, чтобы определить модель, где все коэффициенты значимы, необходимо сравнить критерий Стьюдента t^*_{min} с$t_{кр}$, где$|t^*_{min}|$ должно быть больше$t_{кр}$. Данному критерию соответствует модели с пятью (2.30>2.07), четырьмя (2.48>2.07), тремя (3.06>2.06), двумя (3.72>2.06) и одним (|-7.7|>2.06) предиктором.

Помимо этого, в качественной модели коэффициент детерминации должен быть больше 0.7, что соответствует всем выше указанным моделям, наибольший коэффициент детерминации в модели с пятью переменными (R^2= 0.92).

Все из четырёх моделей адекватны, так как F*>$F_{кр}$ (модель с пятью переменными 46.19>2.66, с четырьмя — 47.19>2.8,с тремя — 49.7>3.01, с двумя — 51.87>3.39, с одной — 59.35>4.23).

Также для всех моделей стандартная ошибка (σ_E) меньше 0.67*σ_y стандартного отклонения зависимой переменной (для модели с пятью переменными 2.71<5.65 см,с четырьмя — 2.96<5.65 см,с тремя — 3.28<5.65 см, с двумя — 3.81<5.65 см, с одной — 4.68<5.65 см).

Из всего выше описанного можно сделать вывод, что модели с пятью, четырьмя, тремя, двумя и одной переменной являются качественными, так как соответствуют всем необходимым параметрам.

Из этих пяти моделей в качестве оптимальной выбрана модель с пятью переменными, так как у неё наибольший коэффициент корреляции и наименьшая стандартная ошибка модели.

Для оптимальной модели ниже записано уравнение и дан график рассчитанного по модели уровня моря в Кронштадте, совмещённый с фактическими значениями.

$$H_{Кр} = 2810.75 — 2.84P + 0.009Run + 1.66H_{МО} + 5.09NAO + 0.03Pr$$

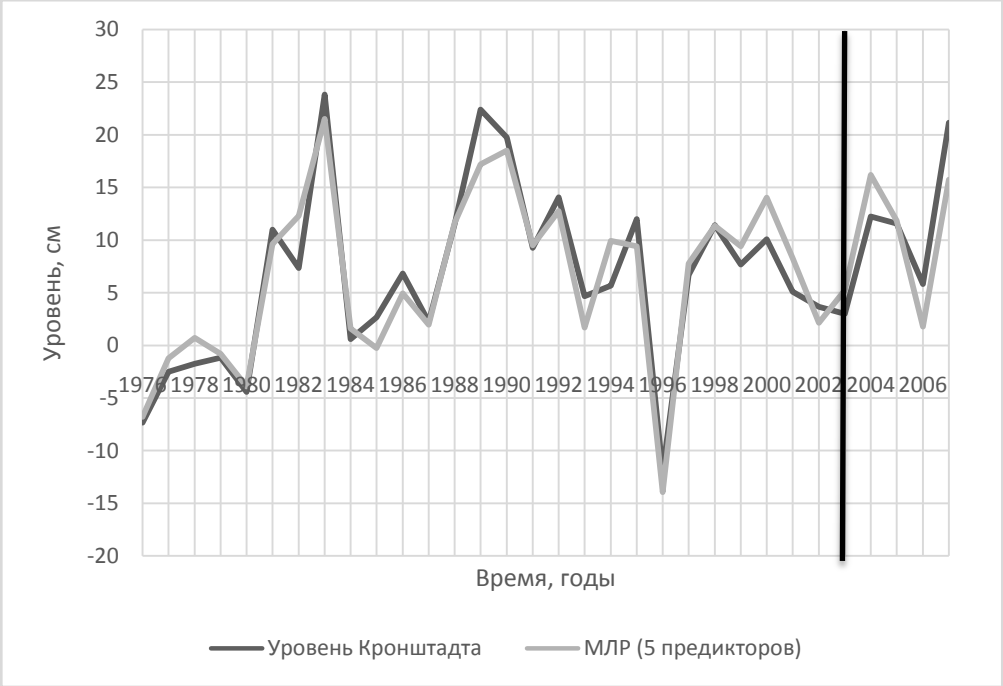


Рис. 6 Совместный график фактического уровня моря в Кронштадте и рассчитанного по оптимальной модели МЛР за период с 1976 по 2007 год (без учёта уровня моря в Хельсинки), зависимая и независимая выборки отделена вертикальной линией

Как видно из Рис. 6, рассчитанные значения по оптимальной модели полностью повторяют ход фактических значений уровня моря в Кронштадте,

расходясь лишь пиковых значениях. Причём, следует отметить, что это характерно как для зависимой, так и для независимой выборки. Однако, если сравнить с графиками на Рис. 4, то видно, что в модели, где не учитывается уровень моря в Хельсинки, разница между рассчитанными и фактическими значениями больше чем при модели, где уровень моря в Хельсинки учитывается. Это было неизбежно, так как уровень моря в Кронштадте и Хельсинки сильно взаимосвязаны, следственно модель, где учитывается уровень моря в Хельсинки, несомненно будет лучше модели, где данный параметр не рассматривается. И хоть эта модель хуже описывает изменения уровня моря в Кронштадте чем две предыдущие модели, она пригодна для прогнозирования уровня моря в Кронштадте, так как является качественной, как и две предыдущие.

2.3 Стандартные ошибки. Сравнение результатов расчёта уровня моря в Кронштадте по методу деревьев классификации и по методу множественной линейной регрессии

Чтобы найти оптимальную модель в деревьях классификации, необходимо рассчитать модели от первого до последнего ветвления, а после это рассчитать их статистические оценки. Также статистические оценки рассчитаны для моделей МЛР, так как необходимо сравнить по какому методу модели лучше описывают изменение уровня моря в Кронштадте.

В Приложении 3 представлена таблица со статистическими оценками уровня моря в Кронштадте за период с 1976 по 2007 год по зависимой и независимой выборке (с учётом уровня моря в Хельсинки) для всех моделей.

Из таблицы видно, что для полного дерева стандартная ошибка рассчитанного уровня по зависимой выборке равна 1.57 см, среднеквадратичное отклонение по зависимой выборке равно 0.19, акоэффициент детерминации между исходными и рассчитанными значениями уровня по зависимой выборке R^2= 0.96. Как можно видеть стандартная ошибка по зависимой выборке с увеличением количества ветвлений уменьшается, а по независимой — сначала возрастает, а потом незначительно уменьшается. Так как стандартная ошибка по зависимой выборке всегда меньше чем по независимой, исключая первое ветвление, а также то, что по независимой выборке только в первом ветвлении СКО меньше 0.67, что является важным критерием качества модели,то в качестве оптимальной модели следовало бы выбрать дерево с одним ветвлением. Но при этом коэффициент детерминации по зависимой выборке в модели с одним ветвлением равен 0.59, т.е. меньше 0.7, следовательно, не подходит для прогноза.

Строго говоря, ни одна из данных моделей в полной мере не удовлетворяет требований для качественной и оптимальной модели. Поэтому рассмотрим модель с одним, тремя и шестью ветвлениями и сравним их. Также, следует помнить, что при кросс-проверке в качестве оптимального дерева была выбрано с тремя ветвлениями.

Рис. 7 Совместный график фактического уровня моря в Кронштадте и рассчитанного по дереву классификации с одним ветвлением за период с 1976

по 2007 год (с учётом уровня моря в Хельсинки), зависимая и независимая выборки отделены вертикальной линией

Как видно из Рис. 7, рассчитанный график по дереву решений с одним ветвлением в области со значениями из независимой выборки действительно неплохо повторяет ход значений фактического уровня. Это же действительно и для зависимой выборки. Однако модель, рассчитанная по дереву решений с одним ветвлением, сильно сглаживает ход уровня, а разница между фактическими и рассчитанными значениями может составлять более десяти сантиметров. Учитывая, что стандартное отклонение исходного ряда по зависимой выборке составляет 8.44 см, данная модель недостаточно хорошо подходит для прогнозирования уровня моря в Кронштадте.

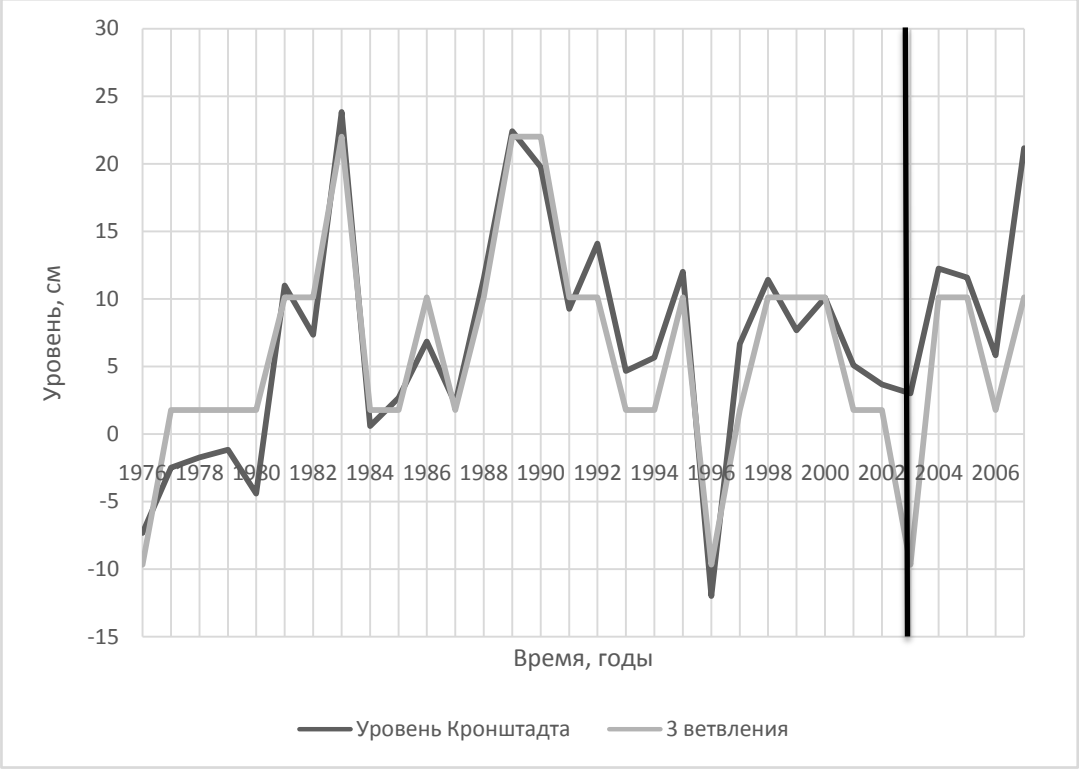


Рис. 8 Совместный график фактического уровня моря в Кронштадте и рассчитанного по дереву классификации с тремя ветвлениями за период с 1976 по 2007 год (с учётом уровня моря в Хельсинки), зависимая и независимая выборки отделены вертикальной линией

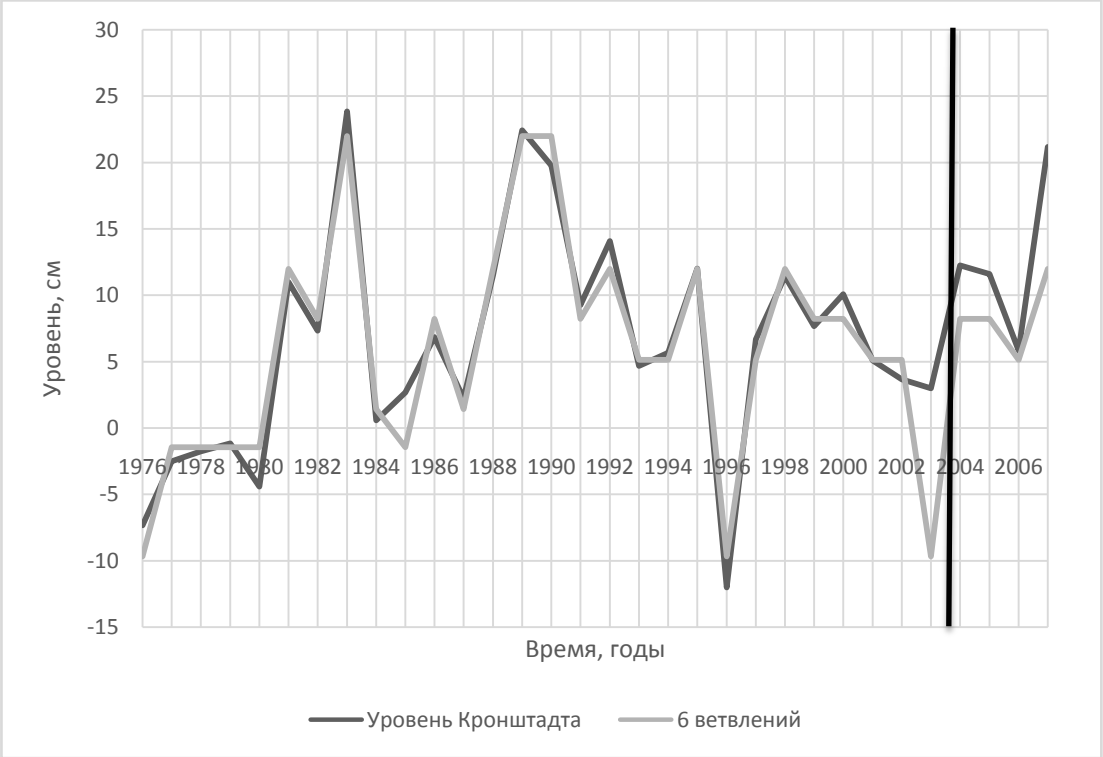


Рис. 9 Совместный график фактического уровня моря в Кронштадте и рассчитанного по дереву классификации с шестью ветвлениями за период с 1976 по 2007 год (с учётом уровня моря в Хельсинки), зависимая и независимая выборки отделены вертикальной линией

Как видно из Рис.8 и 9, в области зависимой выборки рассчитанные уровни достаточно хорошо соответствует фактическому уровню моря в Кронштадте, причём в модели с шестью ветвлениями различие между рассчитанными и фактическими значениями незначительно. Однако, для независимой выборки разница между рассчитанными и фактическими значениями намного больше, особенно велика разница в год, который следует сразу после зависимой выборки.

Таким образом, можно сказать, что в данном случае ни одна модель, рассчитанная помощью дерева решений, не будет являться качественной и оптимальной, а сам прогноз будет обладать низкой точностью. Но в целом,

если при прогнозировании важен сам ход характеристика, а не точные значения, то модели с тремя ветвлениями и более подойдут.

В Приложение 3 также представлены стандартные ошибки, СКО и коэффициент детерминации для моделей МЛР с разным количеством предикторов. Стандартные ошибки по зависимой и независимой выборке с увеличением числа переменных уменьшаются, они всегда меньше стандартного отклонения исходного уровня моря в Кронштадте (8.44 см). Стандартные ошибки по независимой выборке меньше чем по зависимой во всех моделях кроме первых двух. Коэффициент детерминации для всех моделей больше 0.7 (от 0.84 до 0.96). Из этого следует, что все модели МЛР в данном случае подходят для прогнозирования.

В Приложении 4 представлена таблица со статистическими оценками уровня моря в Кронштадте за период с 1976 по 2007 год по зависимой и независимой выборке (без учёта уровня моря в Хельсинки) для всех моделей.

Из таблицы видно, что для полного дерева стандартная ошибка рассчитанного уровня по зависимой выборке равна 2.13 см, среднеквадратичное отклонение по зависимой выборке равно 0.25, а коэффициент детерминации между исходными и рассчитанными значениями уровня по зависимой выборкеR^2= 0.93. Как можно видеть стандартная ошибка по зависимой выборке с увеличением количества ветвлений уменьшается, а по независимой изменяется вариативно. Стандартная ошибка по зависимой выборке всегда меньше чем по независимой. СКО по зависимой выборке меньше 0.67 и уменьшается с увеличением числа ветвлений. При трёх, шести и восьми ветвлениях СКО по независимой выборке близки к 0.67 (0.69, 0.68 и 0.69, соответственно). Коэффициент детерминации по независимой выборке для моделей с тремя, шестью и восьмью ветвлениями равен 0.81, 0.91 и 0.93, соответственно. В качестве оптимальной модели выбрано дерево с

тремяветвлениями. Данное дерево соответствует оптимальному дереву, найденному с помощью кросс-проверки.

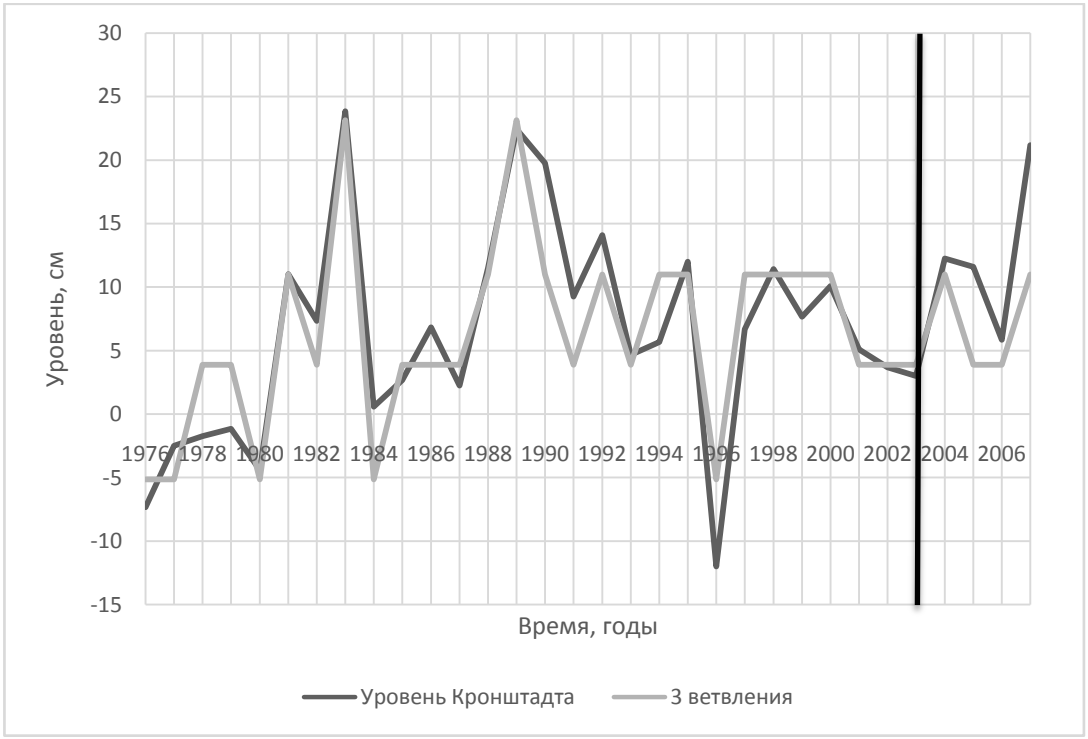


Рис. 10 Совместный график фактического уровня моря в Кронштадте и рассчитанного по дереву классификации с трём ветвлениями за период с 1976 по 2007 год (без учёта уровня моря в Хельсинки), зависимая и независимая выборки отделены вертикальной линией

Как можно увидеть на Рис. 10, рассчитанные значения уровня, в целом, достаточно хорошо повторяют фактические, если не считать пиковых значений. Хоть рассчитанные значения в зависимой выборке лучше соответствуют фактическимчем в независимой, однако, даже в ней рассчитанные значения достаточно хорошо повторяют ход изменений фактических значений, а разница между ними меньше стандартного отклонения исходного ряда (8.44 см). Эта модель, определённо, лучше прогнозирует уровень моря в Кронштадте чем модели деревьев решений, где учитывался уровень моря в Хельсинки. Однако больший интерес представляет сравнение с моделью МЛР, рассчитанная по тем же значениям, что и это дерево классификации.

Из сравнения Приложения 3 и 4 видно, что в моделях МЛР, где уровень моря в Хельсинки не учитывается, коэффициент детерминации нижечем в моделях, где этот предиктор учитывается. Но даже в этом случае все значения коэффициента не ниже 0.7 (от 0.7 до 0.92). Нельзя не отметить, что стандартная ошибка относительно прошлой модели увеличилась, впрочем, СКО больше 0.67 только в модели с двумя предикторами по независимой выборке (0.82). Также, в отличие от модели МЛР, учитывающая уровень моря в Хельсинки, стандартная ошибка по независимой выборке больше чем по зависимой.

Оптимальная модель МЛР, не учитывающая уровень моря в Хельсинки, лучше предсказывает изменения уровня моря в Кронштадте чем оптимальная модель дерева классификации, что хорошо видно из сравнения Рис. 6 и 10. Но разница между этими двумя моделями не так велика, как в случае с моделями, где уровень Хельсинки учитывался.

В Приложении 5 представлена таблица со статистическими оценками уровня моря в Кронштадте за период с 1950 по 2007 год по зависимой и независимой выборке (без учёта температуры воздуха и осадков) для всех моделей.

Из таблицы видно, что для полного дерева стандартная ошибка рассчитанного уровня по зависимой выборке равна 2.17 см, среднеквадратичное отклонение по зависимой выборке равно 0.25, а коэффициент детерминации между исходными и рассчитанными значениями уровня по зависимой выборке $R^2 =$ 0.94. Как можно видеть стандартная ошибка по зависимой выборке с увеличением количества ветвлений уменьшается, а по независимой возрастает первых три модели, после чего также уменьшается. Стандартная ошибка по зависимой выборке всегда меньше чем по независимой. СКО по зависимой выборке при 1 ветвлении равно 0.69, после чего уменьшается и всегда меньше 0.67. По независимой

выборке СКО в первых семи моделях больше 0.67 (от 0.81 до 1.03), а уже при 8 ветвлениях СКО равно 0.5и с увеличением ветвлений уменьшается до 0.46. Коэффициент детерминации при 1 и 2 ветвлениях меньше 0.7 (0.51 и 0.61, соответственно), далее коэффициент детерминации по зависимой выборке с увеличением числа ветвлений больше 0.7.

При выборе оптимальной модели, помимо вышеизложенных факторов, необходимо учитывать число ветвлений — их не должно быть слишком много. В качестве оптимальной модели выбрана модель с 11 ветвлениями, так как у неё первой стандартная ошибка независимойвыборки равна 3.95 см, и она является наименьшей. Также здесь маленькая стандартная ошибка по зависимой выборке (2.44 см) и большой коэффициент детерминации (0.92).

Рис. 11 Совместный график фактического уровня моря в Кронштадте и рассчитанного по дереву классификации с одиннадцатью ветвлениями за период с 1950 по 2007 год (без учёта температуры воздуха и осадков), зависимая и независимая выборки отделены вертикальной линией

Как видно из Рис. 11, рассчитанные значения по оптимальной модели дерева классификации очень хорошо соответствуют фактическим значениям уровня моря в Кронштадте по зависимой выборке, лишь незначительно

расходясь в пиковых значениях. Для независимой выборки, в целом, рассчитанные значения хорошо соответствуют фактическим.

Сравним с оптимальным деревом по кросс-проверке

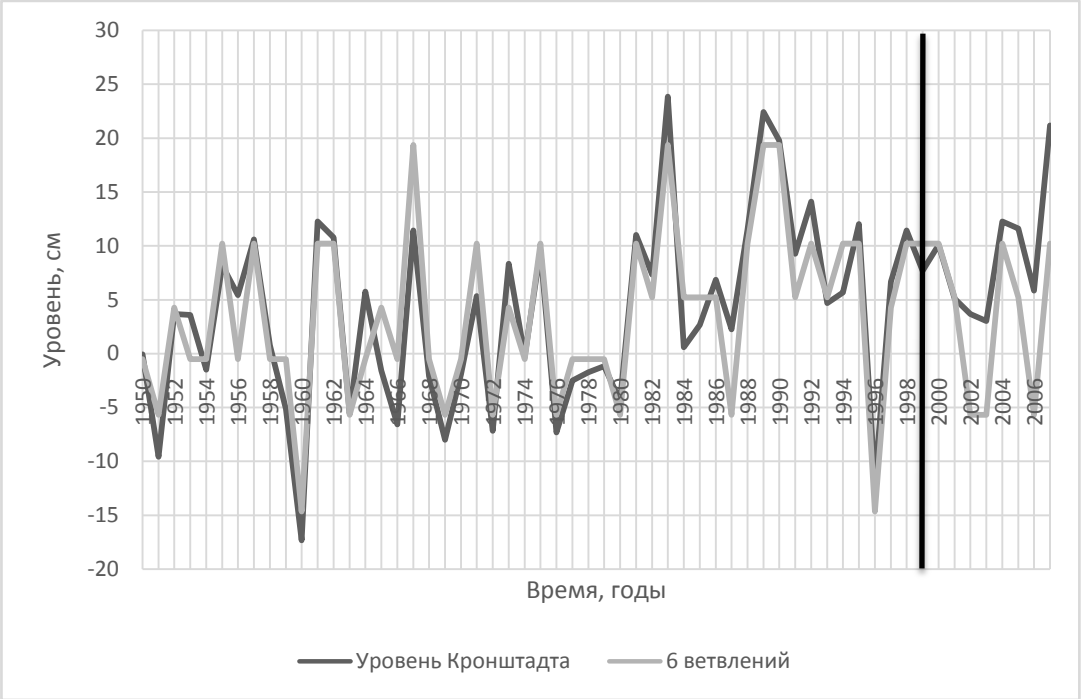


Рис. 12 Совместный график фактического уровня моря в Кронштадте и рассчитанного по дереву классификации с шестью ветвлениями за период с 1950 по 2007 год (без учёта температуры воздуха и осадков), зависимая и независимая выборки отделены вертикальной линией

Как видно из сравнения Рис. 11 и 12, оптимальное дерево, определённое по стандартным ошибкам, лучше прогнозирует уровень Кронштадта чем определённое по кросс-проверке.

Определённо, данная модель дерева классификации наилучшим образом прогнозирует уровень моря в Кронштадте относительно рассмотренных выше моделей деревьев классификации. Однако, необходимо отметить, что помимо увеличившейся точности прогноза уровня, возросло и количество ветвлений, причём, более чем в двое. Это привело к тому, что оптимальная модель имеет большое число ветвлений, следовательно, она сложнее в использовании.

Из Приложения 5 видно, что модели МЛР, не учитывающие температуру воздуха и осадки, обладают хорошими статистическими оценками, если не считать модели с одним предиктором. Коэффициент детерминации для всех моделей больше 0.9, кроме первой (0.64). Стандартная ошибка по зависимой выборке с увеличением числа предикторов уменьшается, для модели с одним предиктором составляет 5.1 см, а далее от 2.31 до 1.76 см. Стандартная ошибка по независимой выборке также с увеличением количества предикторов уменьшается, для модели с одним предиктором равна 8.58 см, после чего уменьшается от 1.14 до 0.42 см. Стандартная ошибка по зависимой выборке больше чем по независимой, если не считать модели с одним предиктором. СКО и для зависимой, и для независимой выборки уменьшаются с увеличением предикторов и всегда меньше 0.67, если не считать СКО по независимой выборке для модели с одним предиктором (0.99). Кроме модели с одним предиктором, все модели МЛР в данном случае качественные.

Если сравнить Рис. 5 и Рис. 11, то можно увидеть, что при данной длине временного ряда и выбранных г/м характеристиках оптимальная модель дерева классификации получилась вполне пригодной для прогнозирования уровня, она всё равно хуже оптимальной модели МЛР, рассчитанной по тем же данным.

Из всего вышеописанного можно сделать вывод, что рассчитанные значения по оптимальным моделям МЛР лучше прогнозируют изменения уровня моря в Кронштадте чем оптимальные модели деревьев классификации. Однако оптимальные модели деревьев классификации, где не учитывается уровень моря в Хельсинки или температура воздуха и осадки, хорошо соответствуют фактическим значениям уровня моря в Кронштадте по зависимой выборке и достаточно неплохо прогнозируют уровень. Так что данные модели вполне могут применяться для прогноза.

Как и в случаи с моделями МЛР, графики, рассчитанные по моделям деревьев решений, не представленные выше, вынесены в приложение (Приложение 1).

3 Модель межгодовых колебаний стока Печоры на основе алгоритма деревьев решений

3.1 Построение деревьев классификаций. Расчёт множественной линейной регрессии

Построение модели прогноза межгодовых колебаний стока Печоры является сложной задачей, потому что область водосбора реки плохо покрыта гидрометеорологическими станциями (шесть станций), следовательно, мало данных. В качестве исходных данных для построения модели стока были взяты осадки за летний и зимний период, за i-1 год и i-2 год, где i — год стока. Так как изменчивость суммарного испарения много меньше чем осадков, то испарение в модели не учитывается. [16]

Прежде чем построить дерево классификации по этим данным, следует рассчитать модель МЛР. Получена оптимальная модель МЛР, при использовании пошагового алгоритма (см. Приложение 6). Её статистические параметры: количество переменных — 1, коэффициент детерминации — 0.38, стандартная ошибка зависимой выборки — 354.26 км^3/год, критерий Фишера — 14.57. Как можно видеть, только при одной переменной, все они значимые. Также рассчитаны статистические оценки для моделей МЛР, и для оптимальной модели они следующие: СКО зависимой выборки — 0.77, стандартная ошибка независимой выборки —436.02 км^3/год, СКО независимой выборки — 0.63. Хоть СКО независимой выборки и меньше 0.67, но при этом СКО зависимой выборки больше 0.67, а коэффициент детерминации меньше 0.7, следовательно, модель некачественная. Нельзя не отметить, что с большим количеством предикторов, хоть и высокий

коэффициент детерминации зависимой выборки, и мала стандартная ошибка зависимой выборки, но стандартная ошибка независимой выборки моделей во много раз превышает стандартное отклонение фактических значений, то есть они не пригодны для прогноза (см. Приложение 7).

Построено дерево классификации для прогнозирования стока Печоры.

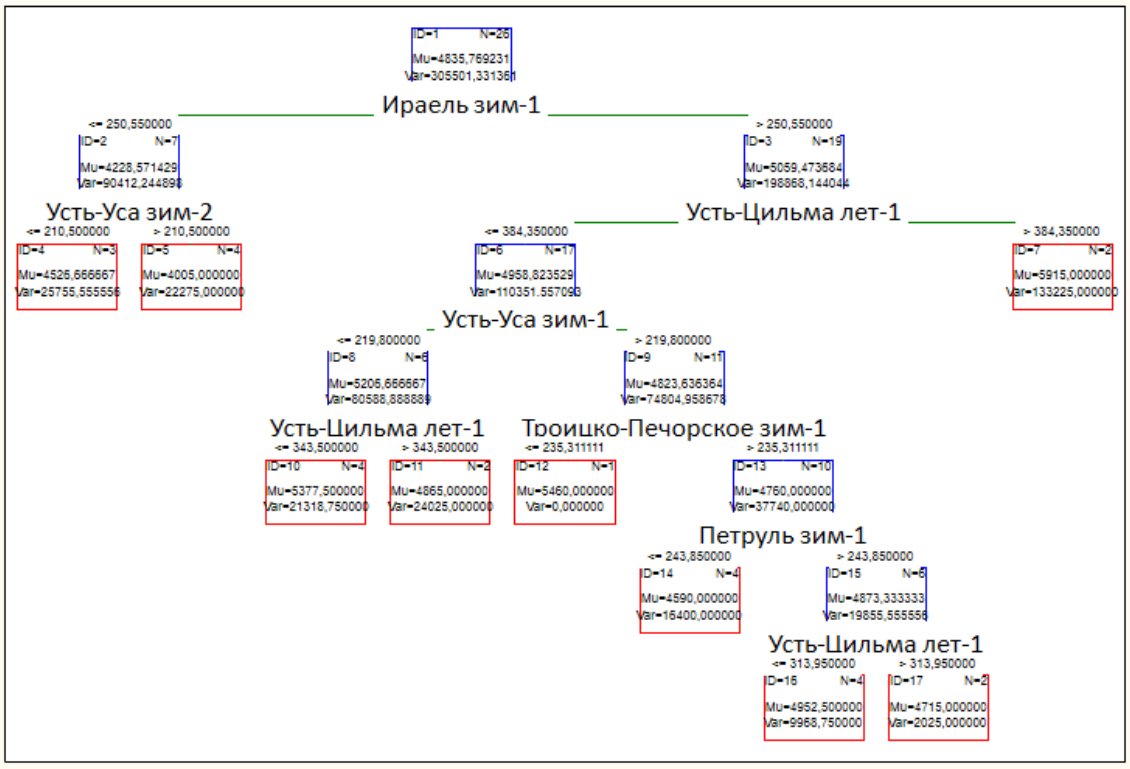


Рис. 13 Дерево классификации связи стока Печоры с осадками, измеренными на шести станциях, за период с 1982 по 2007 год, зависимая выборка

Как можно увидеть на Рис. 13, дерево классификации состоит из девяти терминальных вершин, 8 нетерминальных вершин, то есть всего 17. Из двадцати четырёх переменных только шесть задействованы в построение дерева классификации: Ираёль зим-1, Троицко-Печорское зим-1, Усть-Уса зим-1, Петрунь зим-1, Усть-Уса зим-2, Усть-Цильма зим-1. Причём две последних в дереве классификации являются критерием ветвления не единожды, а два и три раза, соответственно.

Как и для уровня моря в Кронштадт следует найти оптимальное дерево с помощью кросс-проверки.

Таблица 7. Оценки кросс-проверки деревьев классификации для изменения стока реки Печоры с 1982 по 2007 г, зависимая выборка (с учётом уровня моря в Хельсинки)

Показатель	Число терминал-ных вершин	Цена кросс-проверки (CV-cross)	Стандартная ошибка кросс-проверки (CV std. error)	Цена проверки на обучающейся выборке	Сложность узла
Дерево 1	9	4.6E+05	7.5E+04	2.6E+04	0.0E+00
Дерево 2	8	4.6E+05	7.5E+04	2.9E+04	2.9E+03
Дерево 3	7	4.3E+05	7.6E+04	3.6E+04	7.4E+03
Дерево 4	6	3.9E+05	6.9E+04	5.0E+04	1.3E+04
Дерево 5	5	3.8E+05	6.9E+04	6.7E+04	1.7E+04
Дерево 6	4	4.3E+05	9.3E+04	8.5E+04	1.8E+04
Дерево 7	3	3.1E+05	8.4E+04	1.1E+05	2.2E+04
Дерево 8	2	2.6E+05	7.3E+04	1.7E+05	6.3E+04
Дерево 9	1	3.2E+05	9.5E+04	3.1E+05	1.4E+05

Как можно увидеть из Таблицы 7, для данного варианта дерева классификации минимальная цена кросс-проверки у дерева 8 с одним ветвлением, и составляет 2.6E+05, следовательно, CVкрит=3.4E+05. Последнее дерево, у которого CV-cross<CVкрит, - это дерево 9, где нет ветвления. Понятное дело, что использование данного дерева для прогноза невозможно.

3.3 Стандартные ошибки.

Чтобы найти оптимальную модель в деревьях классификации, необходимо рассчитать модели от первого до последнего ветвления, а после это рассчитать их статистические оценки.

Таблица 8. Статистические оценки стока Печоры с 1982 по 2012 г по зависимой и независимой выборке для всех моделей

	Ошибка зависимой выборки, км^3/год	Ошибка независимой выборки, км^3/год	СКО зависимой выборки	СКО независимой выборки	Коэффициент детерминации зависимой выборки
1 ветвление	411.91	423.82	0.73	0.75	0.44
2 ветвления	389.52	423.82	0.69	0.75	0.50
3 ветвления	297.99	398.60	0.53	0.71	0.71
4 ветвления	258.64	496.19	0.46	0.88	0.78
5 ветвлений	231.13	520.25	0.41	0.92	0.83
6 ветвлений	190.50	517.96	0.34	0.92	0.88
7 ветвлений	169.94	513.09	0.30	0.91	0.91
8 ветвлений	161.21	514.72	0.29	0.91	0.91

Из Таблицы 8 видно, что для полного дерева стандартная ошибка рассчитанного уровня по зависимой выборке равна 161.21 км^3/год, СКО по зависимой выборке равно 0.29, а коэффициент детерминации между исходными и рассчитанными значениями уровня по зависимой выборкеR^2=0.91. Как можно видеть стандартная ошибка по зависимой выборке с увеличением количества ветвлений уменьшается, а по независимой изменяется вариативно.

Стандартная ошибка по зависимой выборке всегда меньше чем по независимой. СКО по зависимой выборке при первом ветвлении равно 0.73 и уменьшается с увеличением числа ветвлений. При трёх ветвлениях СКО по независимой выборке близок к 0.67 (0.71). Коэффициент детерминации по независимой выборке для моделей с тремя ветвлениями равен 0.71. Таким образом, в качестве оптимальной модели выбрано дерево с тремя ветвлениями. Графики, рассчитанные по моделям с другим количеством ветвления представлены в Приложении 8.

Рис. 14 Совместный график фактического стока Печоры и рассчитанного по дереву классификации с тремя ветвлениями за период с 1982 по 2012 год, зависимая и независимая выборки отделены вертикальной линией

Как можно увидеть на Рис. 14, график, рассчитанный по оптимальной модели дерева решений для стока Печоры, достаточно неплохо повторяет ход фактических значений, но сглаживает незначительные пики. Сравним эту модель с оптимальной моделью по МЛР.

Рис. 15 Совместный график фактического стока Печоры и рассчитанного по МЛР с одним предиктором за период с 1982 по 2012 год, зависимая и независимая выборки отделены вертикальной линией

Из Рис. 15 видно, что рассчитанные значения стока Печоры по зависимой выборке довольно плохо повторяет ход стока Печоры, при этом амплитуда колебаний рассчитанного стока значительно меньше фактического, что определённо искажает прогнозируемые результаты.

Таким образом, можно сделать вывод, что в данном случае значительной разницы между оптимальными моделями МЛР и дерева классификации нет, они одинаково годятся для прогноза стока Печоры.

Заключение

В данной работе было рассмотрено применение к гидрометеорологическим задачам одного из метода DataMining, а именно деревьев классификации. В качестве гидрометеорологических задач были использованы прогнозирование уровня моря в Кронштадте и стока Печоры. Для прогнозирования уровня моря в Кронштадте были построено три дерева классификации, которые имели либо различный набор гидрометеорологических характеристик, по которым рассчитывался уровень, либо различной длины временной ряд. Также, для сравнения, были рассчитана МЛР по тем же характеристикам, что и деревья классификации.

В процессе выполнения были выявлены недостатки в применении деревьев классификации к гидрометеорологическим задачам. Во-первых, сложность и нетривиальность выбора оптимального размера дерева классификации. Во-вторых, значительное увеличение размеров дерева с ростом временного ряда — при возросшем меньше чем в два раза временном ряду, дерево увеличилось более чем в два раза. В-третьих, стандартная ошибка по независимой выборке практически всегда больше чем по зависимой, что означает, возможные сильные ошибки в прогнозе.

В сравнении метода МЛР и деревьев классификации, модели, рассчитанные по последнему методу, проигрывают в точности прогноза. Однако, нельзя не отметить, что и с помощью метода дерева классификации можно получить модель, которая будет достаточно хорошо прогнозировать гидрометеорологическую характеристику. При том что ветвление в дереве классификации происходит чисто формально, без учёта физических взаимосвязей. Также отсутствие какой-либо важнойсоставляющей для

прогнозаисходной характеристики не сказывается сильно отрицательно на моделях деревьев классификации.

Увеличение числа переменных, используемых для прогноза, не усложняет структуру дерева классификации, в отличие от МЛР. Следовательно, оптимальную модель можно быстрее и проще выбрать, и рассчитать по ней прогноз.

Таким образом, использование метода дерева классификации для решения различных гидрометеорологических задач представляется перспективным направлением. Однако следует найти такие гидрометеорологические задачи, для решения которых наиболее оптимально будет использования деревья классификации и которые сложно и/или не всегда эффективно решать классическими физико-статистическими методами. Также для наиболее рационального использования метода дерева решений необходимо решить проблему с выбором оптимального размера дерева.

Список литературы

1. *Андреев И.* Деревья решений — CART: математический аппарат [Электронный ресурс] // BaseGroupLabs: технологии анализа данных. — URL: Часть 1: https://basegroup.ru/community/articles/math-cartpart1; Часть 2: https://basegroup.ru/community/articles/math-cart-part2.

2. *Малинин В.Н.* Статистические методы анализа гидрометеорологической информации. — СПб.: РГГМУ, 2008. — 407 с.

3. *Чубукова И.А.*DataMining. — М.: Интернет-университет информационных технологий; Бином, лаборатория знаний, 2008. — 384 с.

4. *Шампандар А.Е.* Деревья классификации и регрессии // Искусственный интеллект в компьютерных играх. — М.: ИД «Вильямс», 2007. — С. 385–401.

5. *Bramer M.* Principles of Data Mining. — Springer, 2007. — 344 p.– DOI:10.1007/978-1-84628-766-4.

6. *Breiman L., Friedman J., Olshen R., Stone C.* Classification and Regression Trees. — Wadsworth, Belmont, CA, 1984. — 358 p.

7. *Classification* and Regression Trees: textbook [Electronic resource]. — Carnegie Mellon University, Statistics Department. — URL: http://www.stat.cmu.edu/~cshalizi/350 /lectures/ 22/lecture-22.pdf

8. *Data* Mining / Википедия-свободнаяэнциклопедия [Электронныйресурс]. — URL: https://ru.wikipedia. org/wiki/Data_mining#cite_note-comp-0

9. *Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.* Advances in knowledge discovery & data mining. — Cambridge, MA: MIT Press, 1996.

10. *Hunt E.B., Marin J., Stone P.J.* Experiments in induction. — N.Y., Academic Press, 1966.

11. *Hand D.J., Mannila H., Smith P.* Principles of Data Mining. — The MIT Press, 2001. — 546 p.

12. *Interactive* Trees (C&RT, CHAID): Statistica Help / StatSoftinc. [Electronic resource]. — URL: http://documentation.statsoft.com/STATISTICAHelp.aspx?path= Gxx/Indices/InteractiveTreesCRTCHAID_HIndex

13. *Murthy S.* Automatic construction of decision trees from data: A multidisciplinary survey // Data Mining and Knowledge Discovery, 1998, vol. 2, iss. 4, p. 345–389. — DOI:10.1023/A:1009744630224.

14. *Popular* Decision Tree: Classification and Regression Trees (C&RT) / DELL Software [Electronic resource]. — URL: http://documents.software.dell.com/Statistics/ Textbook/Classification-and-Regression-Trees

15. *Pregibon D.* Data Mining // Statistical Computing and Graphics, 1997, vol. 7, p. 8.

16. *Гордеева С.М., Малинин В.Н.* Использование DataMining в задаче

гидрометеорологического прогнозирования // Учёные записки РГГМУ №44— СПб.: РГГМУ, 2016

17. *Методы* построения деревьев решений в задачах классификации в DataMining— URL: https://www.ami.nstu.ru/~vms/lecture/data_mining/trees.htm

18. *Методы* классификации и прогнозирования. Деревья решений— URL: http://www.intuit.ru/studies/courses/6/6/lecture/174

19. *Сидоров А.В., Миронова Ю.Н.* Алгоритмы создания дерева принятия решений— URL:http://econf.rae.ru/pdf/2014/03/3245.pdf

20. *Гордеева С.М.* ПРАКТИКУМпо курсуСтатистические методы обработки и анализагидрометеорологической информации
(электронная версия)

21. *Гордеева С.М., Малинин В.Н.*Изменчивость морского уровня Финского залива— СПб.: РГГМУ, 2014

22. *Деревья* классификации // StatSoft. Электронный учебник по статистике— URL: http://www.statsoft.ru/home/textbook/default.htm

Приложение 1

Графики уровня моря в Кронштадте, рассчитанные по деревьям классификации

Совместные графики фактического уровня моря в Кронштадте и рассчитанных по дереву классификации с различным числом ветвлений за период с 1976 по 2007 год (с учётом уровня моря в Хельсинки), зависимая и независимая выборки отделена вертикальной линией

Совместные графики фактического уровня моря в Кронштадте и рассчитанных по дереву классификации с различным числом ветвлений за период с 1976 по 2007 год (без учёта уровня моря в Хельсинки), зависимая и независимая выборки отделена вертикальной линией


30
25
20
15
10
5
0
-5
-10
-15
Уровень, см
1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006
Время, годы
Уровень Кронштадта
2 ветвления


30
25
20
15
10
5
0
-5
-10
-15
Уровень, см
1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006
Время, годы
Уровень Кронштадта
4 ветвления


30
25
20
15
10
5
0
-5
-10
-15
Уровень, см
1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006
Время, годы
Уровень Кронштадта
5 ветвлений


30
25
20
15
10
5
0
-5
-10
-15
Уровень, см
1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006
Время, годы
Уровень Кронштадта
6 ветвлений


30
25
20
15
10
5
0
-5
-10
-15
Уровень, см
1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006
Время, годы
Уровень Кронштадта
7 ветвлений


30
25
20
15
10
5
0
-5
-10
-15
Уровень, см
1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006
Время, годы
Уровень Кронштадта
7 ветвлений

Совместные графики фактического уровня моря в Кронштадте и рассчитанных по дереву классификации с различным числом ветвлений за период с 1950 по 2007 год (без учёта температуры воздуха и осадков), зависимая и независимая выборки отделена вертикальной линией


Уровень, см
Время, годы
Уровень Кронштадта
3 ветвления


Уровень, см
Время, годы
Уровень Кронштадта
4 ветвления


Уровень, см
Время, годы
Уровень Кронштадта
5 ветвлений


Уровень, см
Время, годы
Уровень Кронштадта
6 ветвлений


Уровень, см
Время, годы
Уровень Кронштадта
7 ветвлений


Уровень, см
Время, годы
Уровень Кронштадта
8 ветвлений


Уровень, см
Время, годы
Уровень Кронштадта
9 ветвлений


Уровень, см
Время, годы
Уровень Кронштадта
10 ветвлений


Уровень, см
Время, годы
Уровень Кронштадта
12 ветвлений


Уровень, см
30
25
20
15
10
5
0
-5
-10
-15
-20
1950
1953
1956
1959
1962
1965
1968
1971
1974
1977
1980
1983
1986
1989
1992
1995
1998
2001
2004
2007
Время, годы
Уровень Кронштадта
13 ветвлений


Уровень, см
30
25
20
15
10
5
0
-5
-10
-15
-20
1950
1953
1956
1959
1962
1965
1968
1971
1974
1977
1980
1983
1986
1989
1992
1995
1998
2001
2004
2007
Время, годы
Уровень Кронштадта
14 ветвлений


Уровень, см
30
25
20
15
10
5
0
-5
-10
-15
-20
1950
1953
1956
1959
1962
1965
1968
1971
1974
1977
1980
1983
1986
1989
1992
1995
1998
2001
2004
2007
Время, годы
Уровень Кронштадта
15 ветвлений


30
25
20
15
10
5
0
-5
-10
-15
-20
Уровень, см
1950
1953
1956
1959
1962
1965
1968
1971
1974
1977
1980
1983
1986
1989
1992
1995
1998
2001
2004
2007
Время, годы
Уровень Кронштадта
16 ветвлений


30
25
20
15
10
5
0
-5
-10
-15
-20
Уровень, см
1950
1953
1956
1959
1962
1965
1968
1971
1974
1977
1980
1983
1986
1989
1992
1995
1998
2001
2004
2007
Время, годы
Уровень Кронштадта
17 ветвлений

Приложение 2

Графики уровня моря в Кронштадте, рассчитанные по моделям МЛР, и уравнения к ним

Условные обозначения:

$H_{Кр}$ — уровень моря в Кронштадте;

$H_{Х}$ — уровень моря в Хельсинки;

$H_{МО}$ — уровень Мирового океана;

W_{zon}—ср. зональный ветер;

Run — сток Невы;

P — атмосферное давление;

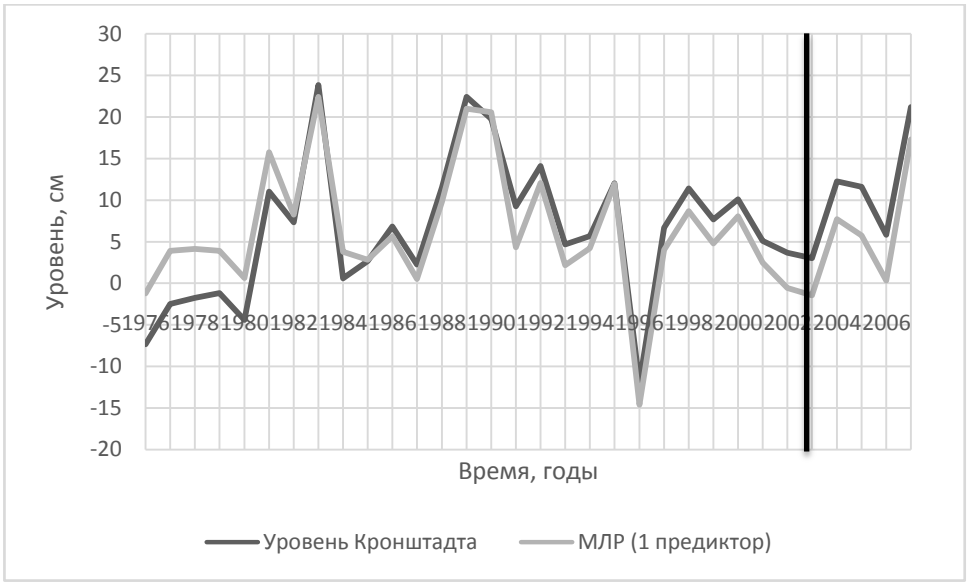
NAO — Северо-Атлантическое колебание;

Pr — осадки;

Ta — температура воздуха.

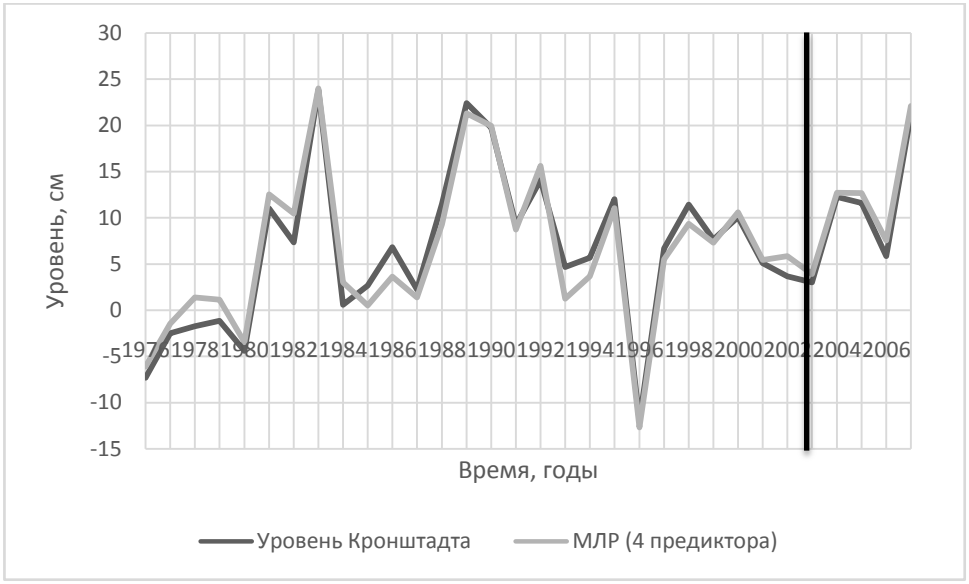
Совместные графики фактического уровня моря в Кронштадте и рассчитанных по моделям МЛР (к каждой прилагается соответствующее уравнение) с различным числом предикторов за период с 1976 по 2007 год (с учётом уровня моря в Хельсинки), зависимая и независимая выборки отделена вертикальной линией.

$H_{Кр}$= — 754.81 +0.11*$H_{Х}$

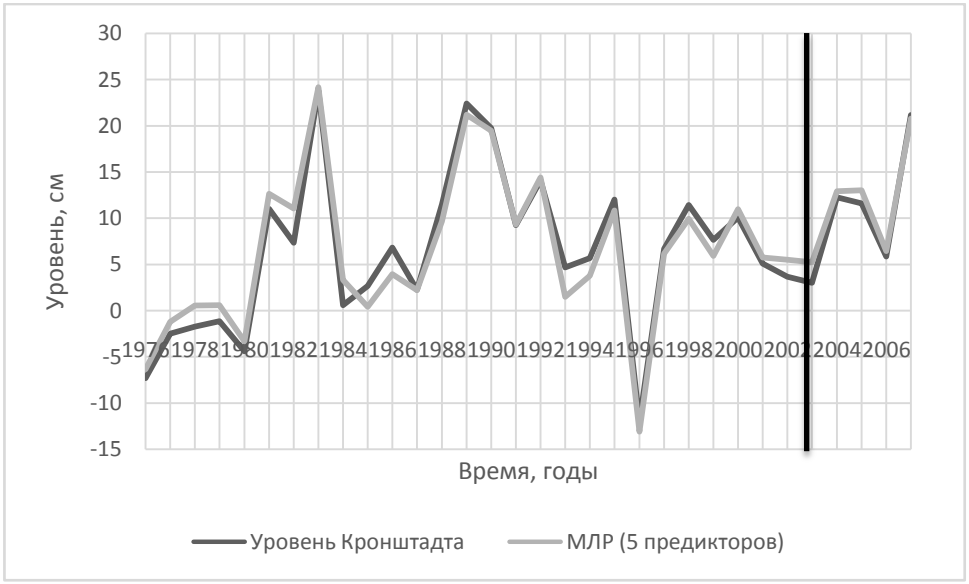


$H_{Кр}$= — 705.64 + 0.1*$H_{Х}$+ 2.28*$H_{МО}$

$$H_{Кр} = — 622.8 + 0.09*H_{Х} + 1.93*H_{МО} + 5.59*W_{zon} + 0.002*Run$$

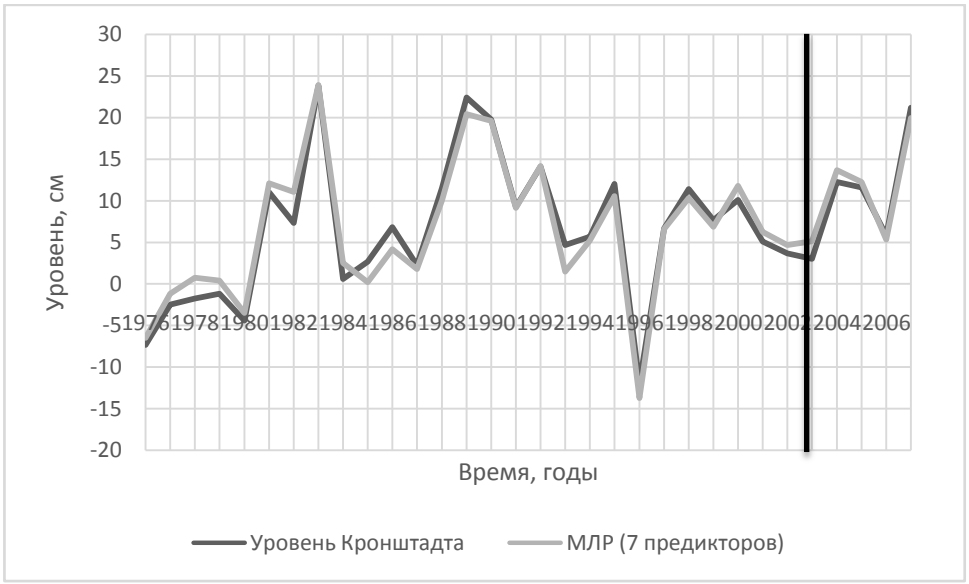


$$H_{Кр} = — 594.18 + 0.08*H_{Х} + 1.87*H_{МО} + 5.91*W_{zon} + 0.003*Run + 0.009*\ Pr$$



$H_{Кр}$= 130.07 + 0.07*$H_{Х}$+ 1.79*$H_{МО}$ + 5.19*W_{zon}+ 0.004*Run + 0.01* Pr— 0.63*P

$H_{Кр}$= 267.15 + 0.07*$H_{Х}$+ 1.79*$H_{МО}$ + 4.5*W_{zon}+ 0.004*Run + 0.01* Pr — 0.74*P +1.28*NAO

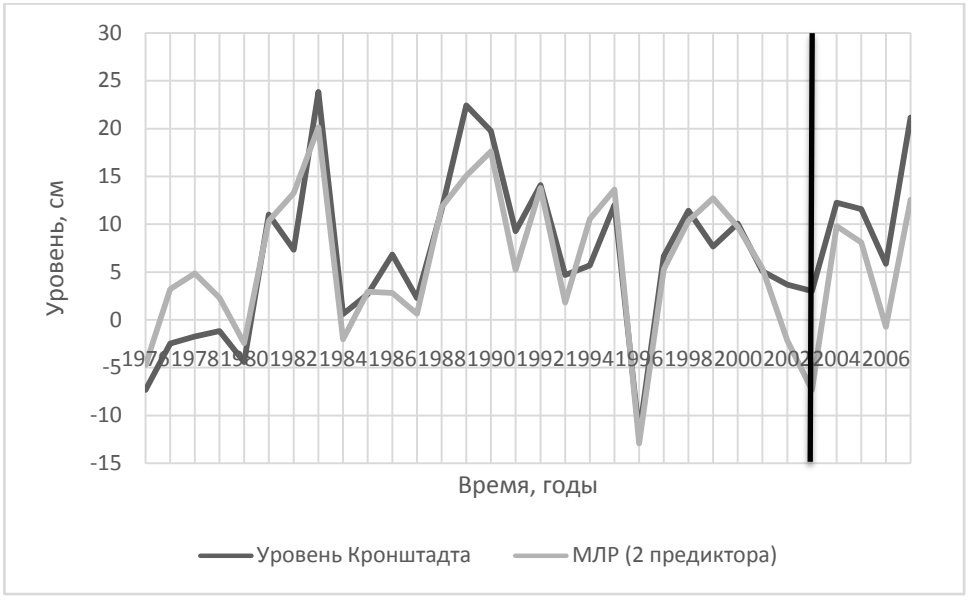


$H_{Кр}$= 264.45 + 0.07*$H_{Х}$+ 1.7*$H_{МО}$ + 4.55*W_{zon}+ 0.004*Run + 0.01* Pr — 0.73*P +1.16*NAO + 0.17*Ta

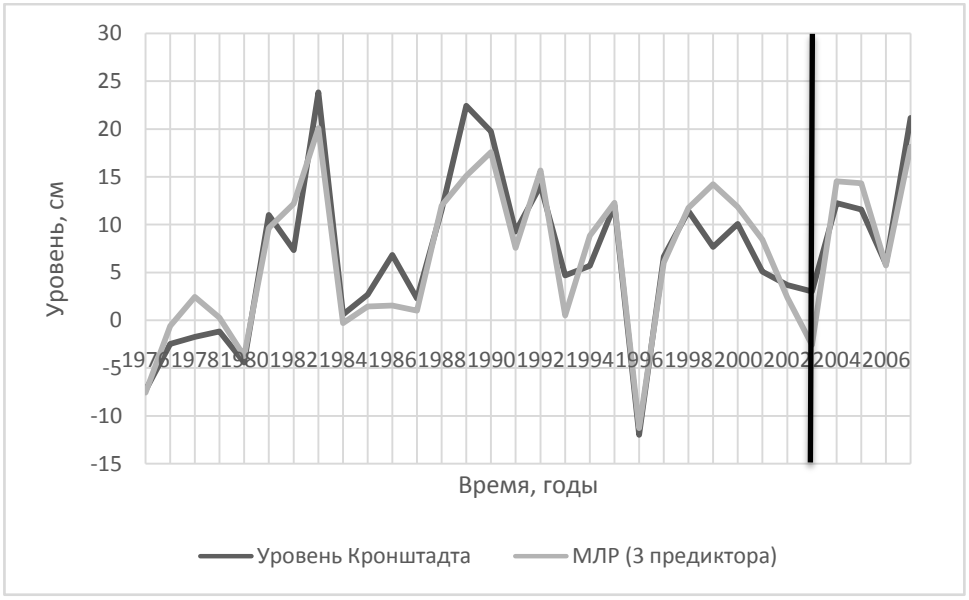
Совместные графики фактического уровня моря в Кронштадте и рассчитанных по моделям МЛР (к каждой прилагается соответствующее уравнение) с различным числом предикторов за период с 1976 по 2007 год (без учёта уровня моря в Хельсинки), зависимая и независимая выборки отделена вертикальной линией.

$H_{Кр}$=4284.79 — 4.27*P

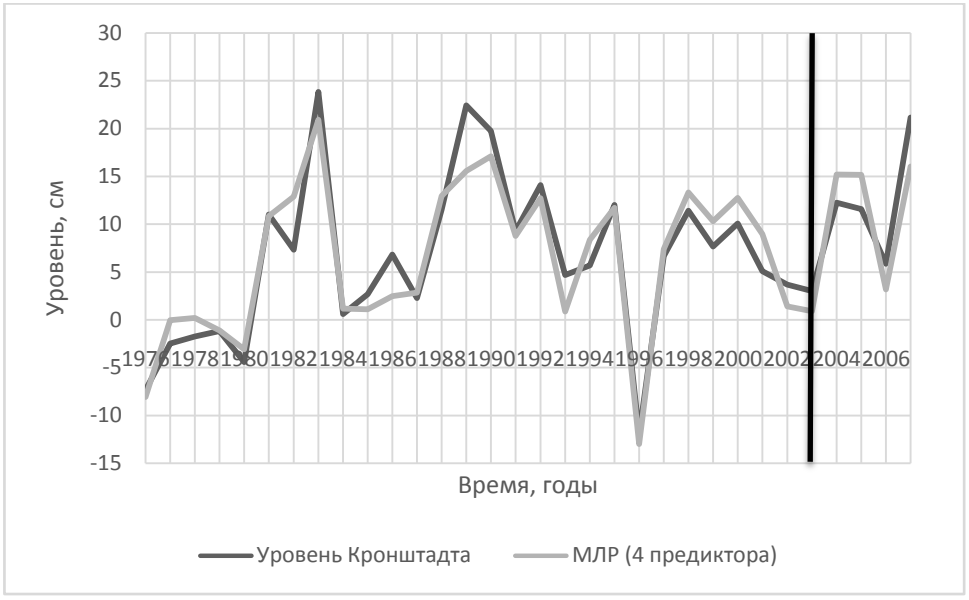
$H_{Кр}$= 3615.18 — 3.63*P + 0.01*Run



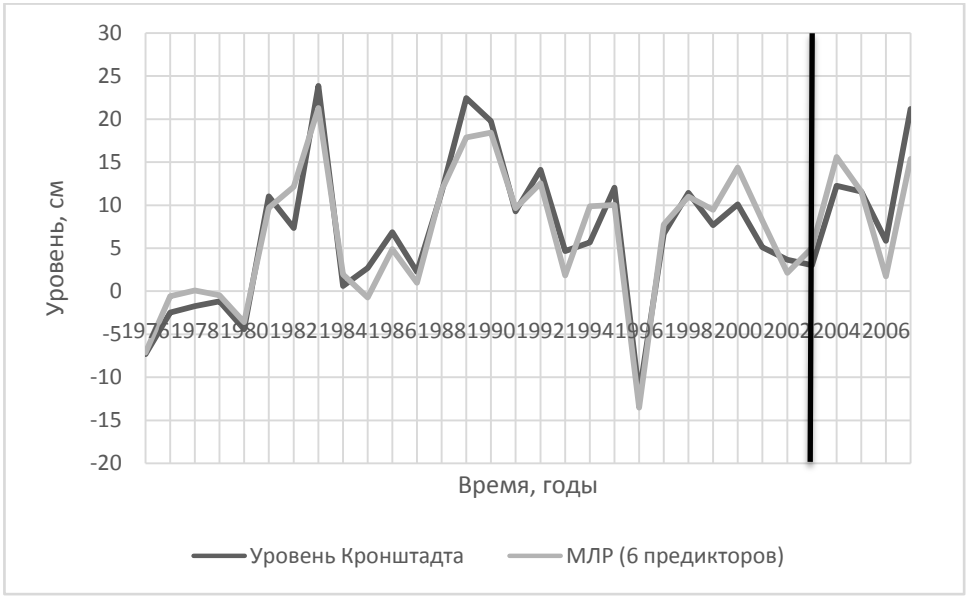
$H_{Кр}$= 3280.83 — 3.29*P + 0.01*Run + 1.78*$H_{МО}$



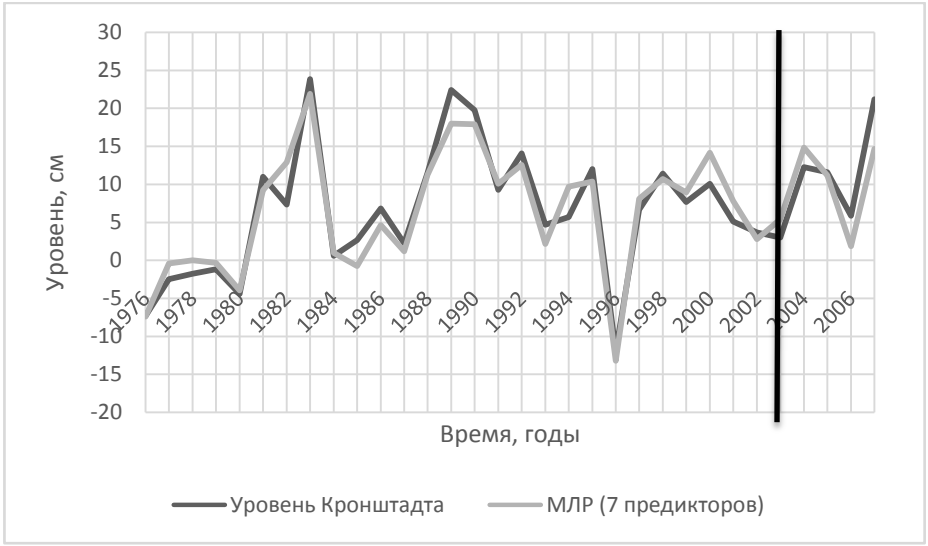
$H_{Кр}$=2964.4 — 2.99*P + 0.01*Run + 1.7*$H_{МО}$+ 0.02*Pr



$H_{Кр}$=2712.35 — 2.74*P + 0.01*Run + 1.37*$H_{МО}$+ 0.03*Pr + 4.62*NAO + 0.57*Ta

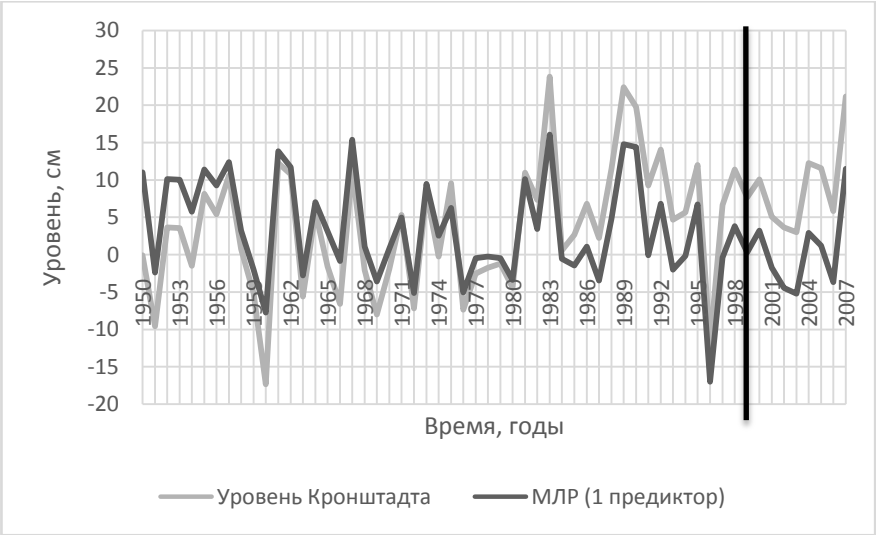


$H_{Кр}$= 2586.84 — 2.61*P + 0.01*Run + 1.21*$H_{МО}$+ 0.03*Pr + 4.62*NAO + 0.64*Ta + 2.56*W_{zon}



Совместные графики фактического уровня моря в Кронштадте и рассчитанных по моделям МЛР (к каждой прилагается соответствующее уравнение) с различным числом предикторов за период с 1950 по 2007 год (без учёта температуры воздуха и осадков), зависимая и независимая выборки отделена вертикальной линией.

$H_{Кр}$= — 678.73 +0.1*H_{X}



$H_{Кр}$= — 752.56 + 0.11*$H_{Х}$+ 1.96*$H_{МО}$

$H_{Кр}$= — 675.59 + 0.1*$H_{Х}$+ 1.63*$H_{МО}$ + 5.97*W_{zon}

$$H_{Кр} = — 220.26 + 0.09*H_{Х} + 1.58*H_{МО} + 5.13*W_{zon} + 0.003*Run — 0.39*P$$

$$H_{Кр} = — 226.47 + 0.09*H_{Х} + 1.59*H_{МО} + 5.15*W_{zon} + 0.003*Run — 0.38*P — 0.03*NAO$$

Приложение 3

Таблица статистической оценки уровня моря в Кронштадте

Таблица 9. СтатистическиеоценкиуровняморявКронштадтес 1976 по 2007 г(с учётом уровня моря в Хельсинки) по зависимой и независимой выборке для всех моделей

		Ошибка зависимой выборки, см	Ошибка независимой выборки, см	СКО зависимой выборки	СКО независимой выборки	Коэффициент детерминации зависимой выборки
Дерево классификации	1 ветвление	5.31	4.72	0.63	0.56	0.59
	2 ветвления	4.46	7.04	0.53	0.83	0.71
	3 ветвления	2.79	7.82	0.33	0.93	0.89
	4 ветвления	2.12	7.65	0.25	0.91	0.93
	5 ветвлений	1.79	7.41	0.21	0.88	0.95
	6 ветвлений	1.57	7.38	0.19	0.87	0.96
МЛР	1 предиктор	3.36	4.90	0.40	0.58	0.84
	2 предиктора	2.22	2.38	0.26	0.28	0.93
	3 предиктора	1.87	1.58	0.19	0.22	0.95
	4 предиктора	1.81	1.10	0.22	0.13	0.95
	5 предикторов	1.73	1.26	0.21	0.15	0.96
	6 предикторов	1.67	1.06	0.20	0.13	0.96
	7 предикторов	1.65	1.34	0.20	0.16	0.96
	8 предикторов	1.64	1.32	0.19	0.16	0.96

Приложение 4

Таблица статистической оценки уровня моря в Кронштадте

Таблица 10. Статистические оценки уровня моря в Кронштадте с 1976 по 2007 г (без учёта уровня моря в Хельсинки) по зависимой и независимой выборке для всех моделей

		Ошибка зависимой выборки, см	Ошибка независимой выборки, см	СКО зависимой выборки	СКО независимой выборки	Коэффициент детерминации по зависимой выборке
Дерево классификации	1 ветвление	5.67	6.49	0.67	0.77	0.53
	2 ветвления	4.81	7.06	0.57	0.84	0.66
	3 ветвления	3.61	5.82	0.43	0.69	0.81
	4 ветвления	3.13	6.44	0.37	0.76	0.86
	5 ветвлений	2.68	6.16	0.32	0.73	0.90
	6 ветвлений	2.52	5.73	0.30	0.68	0.91
	7 ветвлений	2.28	6.10	0.27	0.72	0.92
	8 ветвлений	2.13	5.80	0.25	0.69	0.93
МЛР	1 предиктор	4.51	5.26	0.53	0.62	0.70
	2 предиктора	3.59	6.95	0.43	0.82	0.81
	3 предиктора	3.03	3.25	0.36	0.39	0.87
	4 предиктора	2.67	3.44	0.32	0.41	0.90
	5 предикторов	2.39	3.66	0.28	0.43	0.92
	6 предикторов	2.36	3.64	0.28	0.43	0.92
	7 предикторов	2.32	3.76	0.28	0.45	0.92

Приложение 5

Таблица статистической оценки уровня моря в Кронштадте

Таблица 11. Статистические оценки уровня моря в Кронштадте с 1950 по 2007 г (без учёта температуры воздуха и осадков) по зависимой и независимой выборке для всех моделей

		Ошибка зависимой выборки, см	Ошибка независимой выборки, см	СКО зависимой выборки	СКО независимой выборки	Коэффициент детерминации по зависимой выборке
Дерево решений	1 ветвление	6.00	6.99	0.69	0.81	0.51
	2 ветвления	5.35	7.38	0.62	0.85	0.61
	3 ветвления	4.32	8.94	0.50	1.03	0.75
	4 ветвления	4.06	8.69	0.47	1.01	0.77
	5 ветвлений	3.73	7.82	0.43	0.90	0.81
	6 ветвлений	3.33	7.21	0.39	0.83	0.85
	7 ветвлений	3.24	7.09	0.37	0.82	0.86
	8 ветвлений	3.00	4.36	0.35	0.50	0.88
	9 ветвлений	2.64	4.36	0.31	0.50	0.90
	10 ветвлений	2.48	4.10	0.29	0.47	0.92
	11 ветвлений	2.44	3.95	0.28	0.46	0.92
	12 ветвлений	2.39	3.95	0.28	0.46	0.92
	13 ветвлений	2.24	3.95	0.26	0.46	0.93
	14 ветвлений	2.21	3.95	0.26	0.46	0.93
	15 ветвлений	2.19	3.95	0.25	0.46	0.93
	16 ветвлений	2.18	3.95	0.25	0.46	0.94
	17 ветвлений	2.17	3.95	0.25	0.46	0.94
МЛР	1 предиктор	5.10	8.58	0.59	0.99	0.64
	2 предиктора	2.31	1.14	0.27	0.13	0.93
	3 предиктора	1.93	1.00	0.22	0.12	0.95
	4 предиктора	1.79	0.53	0.21	0.06	0.96

5 предикторов	1.76	0.42	0.20	0.05	0.96
6 предикторов	1.76	0.42	0.20	0.05	0.96

Приложение 6

Таблица характеристик моделей МЛР стока Печоры

Таблица 12. Характеристики качества моделей для изменения стока реки Печора с 1982 по 2007 г, зависимая выборка

Номер шага	Кол-во пред-ов	R^2	F*	σ_E	t^*_{min}	$F_{кр}$	σ_y	$t_{кр}$	0.67* σ_y
1	24	0.99	2.16	387.51	0.026	19.45	563.67	4.30	377.66
2	23	0.98	4.51	274.11	-0.14	8.64		3.18	
3	22	0.98	7	224.95	-0.29	5.79		2.78	
4	21	0.98	9.51	197.49	-0.83	4.55		2.57	
5	20	0.98	10.59	191.38	0.96	3.87		2.45	
6	19	0.97	11.26	190.04	1.27	3.46		2.36	
7	18	0.97	10.83	198.29	1.08	3.17		2.31	
8	17	0.92	11.18	257.25	2.23	2.97		2.26	
9	16	0.86	3.61	345.03	-1.96	2.83		2.23	
10	15	0.9	2.79	391.21	1.45	2.72		2.20	
11	14	0.77	2.58	410.27	0.68	2.64		2.18	
12	13	0.76	2.88	400.9	1.17	2.58		2.16	
13	12	0.73	2.92	406.43	-0.84	2.53		2.14	
14	11	0.72	3.19	402.21	0.74	2.51		2.13	
15	10	0.7	3.56	396.2	0.87	2.49		2.12	

				2					
16	9	0.69	3.93	393.2	-0.65	2.49		2.11	
17	8	0.68	4.52	386.43	0.95	2.51		2.10	
18	7	0.66	5.07	385.45	-1.69	2.54		2.09	
19	6	0.61	4.94	404.08	1.15	2.60		2.09	
20	5	0.58	5.58	407.33	-1.21	2.68		2.08	
21	4	0.55	6.46	411.78	-1.63	2.82		2.07	
22	3	0.5	7.19	427.01	1.11	3.03		2.07	
23	2	0.47	10.07	429.1	1.96	3.40		2.06	
24	**1**	**0.38**	**14.57**	**453.83**	**3.82**	**4.24**		**2.06**	

Приложение 7

Таблица статистической оценки уровня моря в Кронштадте

Таблица 13. Статистические оценки стока реки Печора по зависимой и независимой выборке для всех моделей МЛР

	Ошибка зависимой выборки, км^3/год	Ошибка независимой выборки, км^3/год	СКО зависимой выборки	СКО независимой выборки	Коэффициент детерминации зависимой выборки
1 предикторов	436.02	354.26	0.77	0.63	0.03
2 предикторов	403.59	347.74	0.72	0.62	0.47
3 предикторов	392.79	321.93	0.70	0.57	0.49
4 предикторов	370.07	371.52	0.66	0.66	0.55
5 предикторов	357.25	277.36	0.63	0.49	0.58
6 предикторов	345.43	325.98	0.61	0.58	0.61
7 предикторов	320.72	399.18	0.57	0.71	0.66
8 предикторов	312.47	521.26	0.55	0.92	0.68
9 предикторов	308.45	407.74	0.55	0.72	0.69
10 предикторов	300.95	362.90	0.53	0.64	0.70
11 предикторов	295.14	435.97	0.52	0.77	0.71
12 предикторов	287.39	551.52	0.51	0.98	0.73
13 предикторов	272.36	663.31	0.48	1.18	0.76
14 предикторов	266.86	880.96	0.47	1.56	0.77
15 предикторов	242.62	1624.63	0.43	2.88	0.81

	Ошибка зависимой выборки, км^3/год	Ошибка независимой выборки, км^3/год	СКО зависимой выборки	СКО независимой выборки	Коэффициент детерминации зависимой выборки
16 предикторов	203.00	1930.93	0.36	3.43	0.87
17 предикторов	111.08	3839.21	0.20	6.81	0.96
18 предикторов	102.89	3922.08	0.18	6.96	0.97
19 предикторов	91.29	4831.16	0.16	8.57	0.97
20 предикторов	83.93	5042.74	0.15	8.95	0.98
21 предиктор	77.46	5238.14	0.14	9.29	0.98
22 предиктора	76.41	5334.90	0.14	9.46	0.98
23 предиктора	76.02	5137.77	0.13	9.11	0.98
24 предиктора	76.00	5176.54	0.13	9.18	0.98

Приложение 8

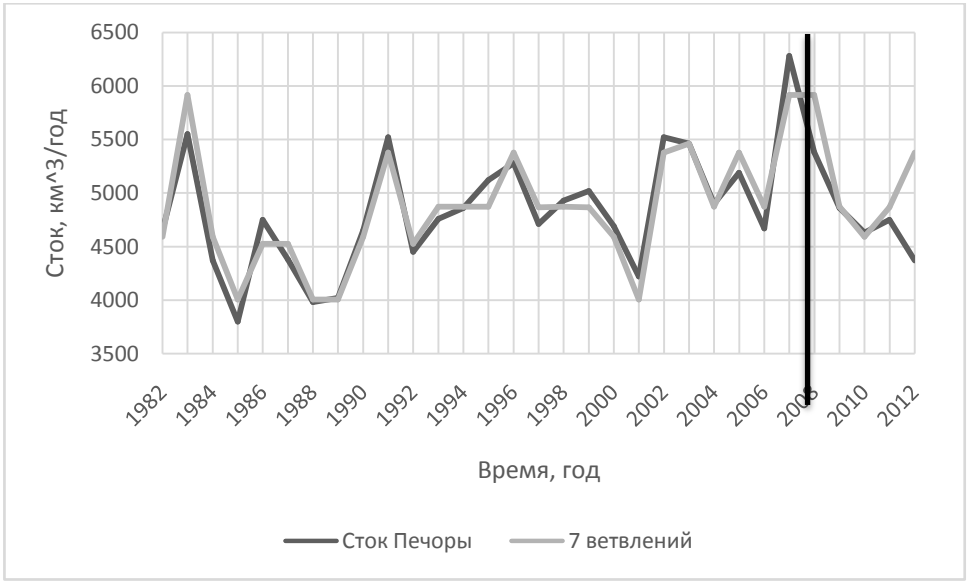
Графики стока Печоры, рассчитанные по моделям деревьев классификации

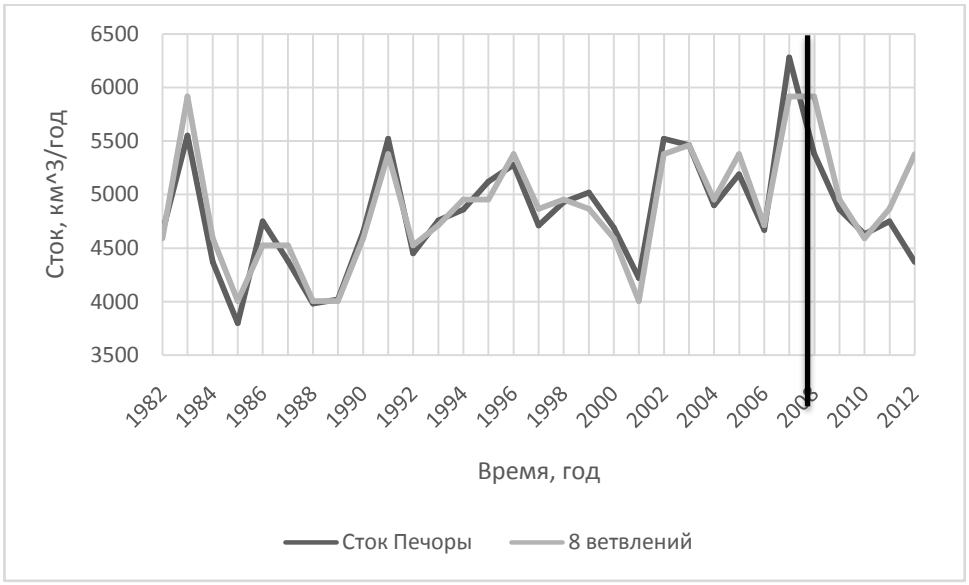
Совместные графики фактического стока Печоры и рассчитанных по дереву классификации с различным числом ветвлений за период с 1982 по 2012 год, зависимая и независимая выборки отделены вертикальной линией


6500
6000
5500
5000
4500
4000
3500
Сток, км^3/год
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Время, год
Сток Печоры
4 ветвления


6500
6000
5500
5000
4500
4000
3500
Сток, км^3/год
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Время, год
Сток Печоры
5 ветвлений


6500
6000
5500
5000
4500
4000
3500
Сток, км^3/год
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Время, год
Сток Печоры
6 ветвлений


6500
6000
5500
5000
4500
4000
3500
Сток, км^3/год
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Время, год
Сток Печоры
7 ветвлений


6500
6000
5500
5000
4500
4000
3500
Сток, км^3/год
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
Время, год
Сток Печоры
8 ветвлений