Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Базовая кафедра «Интеллектуальные системы управления»

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Якунин Ю.Ю.
подпись инициалы, фамилия
« ___» _____ 20 __ г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Непараметрический алгоритм исключения выбросов из выборки наблюдений переменных процесса

27.04.03 Системный анализ и управление
27.04.03.02 Системный анализ данных и технологий принятия решений

Научный руководитель	_____ подпись, дата	доцент, к. т. н.	А. А. Корнеева инициалы, фамилия
Выпускник	_____ подпись, дата		А. А. Молошаг инициалы, фамилия
Рецензент	_____ подпись, дата	доцент, к. т. н.	Н. В. Кононова инициалы, фамилия
Нормоконтроллер	_____ подпись, дата		Н. Б. Позолотина инициалы, фамилия

Красноярск 2019

РЕФЕРАТ

Магистерская диссертация по теме «Непараметрический алгоритм исключения выбросов из выборки наблюдений переменных процесса» содержит 51 страницу текстового документа, 26 используемых источников, 13 иллюстраций, 22 формулы, 6 таблиц.

НЕПАРАМЕТРИЧЕСКАЯ ИДЕНТИФИКАЦИЯ, ПАРАМЕТРИЧЕСКАЯ ИДЕНТИФИКАЦЯ, ВЫБРОСЫ, АЛГОРИТМ, МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, АНАЛИЗ ДАННЫХ.

Целью данной работы является повышение качества решения задач идентификации дискретно-непрерывных процессов.

Объектом исследования является непараметрический алгоритм исключения выбросов из выборки наблюдений.

В результате данной работы был исследован, модифицирован и протестирован на модельных данных непараметрический алгоритм исключения выбросов из выборки наблюдений.

По итогу был проведен анализ данных практической задачи, но применить модифицированный алгоритм к этим данным не удалось.

Содержание

ВВЕДЕНИЕ

Актуальность работы. С развитием информационных технологий и совершенствованием электронно-вычислительных машин математическое моделирование стало играть все более весомую роль в анализе данных.

Работа не с самим объектом, а с его моделью, позволяет безболезненно, относительно быстро и без существенных затрат исследовать его свойства и поведение в любых мыслимых ситуациях [6].

При формировании математических моделей используют такой инструмент, как идентификация. Под идентификацией объектов понимается построение оптимальных в некотором смысле математических моделей по реализации их входных и выходных сигналов [4]. В задаче идентификации недоопределенной является модель, в то время как объект считается в достаточной мере изученным.

В зависимости от объема априорной информации различают задачи идентификации в «узком» и «широком» смысле. В настоящее время наиболее полно развита теория идентификации в «узком» смысле, когда параметры модели определяются по уже известной структуре объекта [18]. При идентификации в «широком» смысле априорной информации об объекте очень мало, поэтому сначала требуется определить его структуру (структурная идентификация), а затем уже оцениваются параметры модели. При этом решаются задачи оценивания числа переменных, линейности, стационарности и других свойств идентифицируемого объекта [19].

Точность решения задачи идентификации дискретно-непрерывного процесса зависит от многих факторов, в том числе и от характеристик исходных данных, таких как размерность и объем выборки наблюдений, наличие в данных пропущенных значений, выбросов и т.д. Под выбросами мы будем понимать данные, которые в силу каких-либо случайных причин, нехарактерных для рассматриваемой предметной области, значительно отличаются по своим

параметрам от другой статистической информации [25]. Выбросы негативно влияют на задачи моделирования и классификации.

Цель работы заключается в повышении точности решения задач идентификации по выборкам наблюдений с выбросами.

Задачи, возникающие в ходе реализации поставленной цели:

- поиск и изучение наиболее значимой литературы по выбранной области исследования;
- реализация и модификация непараметрического алгоритма исключения выбросов из выборки наблюдений;
- анализ практической задачи на возможность применения непараметрического алгоритма исключения выбросов.

Научная новизна магистерской диссертации состоит в модификации существующего непараметрического алгоритма исключения выбросов из выборки наблюдений, предложенного Корнеевой А. А. и Медведевым А. В. [12], а именно:

- автоматическая настройка параметра размытости ядра;
- автоматическая настройка параметра допустимой границы рассогласований значений объекта и модели.

1 Идентификация многомерных дискретно-непрерывных процессов

1.1 Задача моделирования

Замещение одного объекта другим с целью получения информации о важнейших свойствах объекта-оригинала с помощью объекта-модели называется моделированием [21].

Технические, экологические, экономические и иные системы, изучаемые современной наукой, больше не поддаются исследованию (в нужной полноте и точности) обычными теоретическими методами. Прямой натурный эксперимент над ними долог, дорог, часто либо опасен, либо попросту невозможен, так как многие из этих систем существуют в «единственном экземпляре».

Цена ошибок и просчетов в обращении с ними недопустимо высока. Поэтому математическое моделирование является неизбежной составляющей научно-технического прогресса. Сама постановка вопроса о математическом моделировании какого-либо объекта порождает четкий план действий. Его можно условно разбить на три этапа: модель – алгоритм – программа [20].

Сначала строится (или выбирается) модель объекта, ее эквивалент, выраженный в математической форме и отражающий его важнейшие свойства: законы, связи и т.д. На этом этапе получаются предварительные знания об объекте.

На втором этапе разрабатывается (или выбирается) алгоритм для реализации модели. Ее приводят к форме, удобной для применения численных методов, строится план необходимых логических операций для поиска искомых величин с заданной точностью. Вычислительные алгоритмы должны не искажать основные свойства модели (и, следовательно, исходного объекта), быть экономичными и адаптирующимися к особенностям решаемых задач и используемых вычислительных машин.

И завершающим этапом является написание программы, переводящей построенный ранее алгоритм на компьютерный язык. К данной программе так

же предъявляются требования адаптивности и экономичности. В данном виде модель становится пригодной для испытаний на компьютере. Теперь ее можно отлаживать и тестировать с помощью пробных вычислительных экспериментах.

Модель сопровождается улучшением и уточнением на каждом из этапов ее построения.

1.2 Задача идентификации

Существует два способа (а также их комбинации) формирования математических моделей. Первый способ состоит в том, чтобы «расщепить» систему на такие подсистемы, свойства которых очевидны из ранее накопленного опыта. Формальное математическое объединение этих подсистем становится моделью всей системы. Такой подход называется моделированием и в его рамках проведение натуральных экспериментов не обязательно.

В другом способе построения математических моделей используются экспериментальные данные. В этом случае ведется регистрация входных и выходных сигналов системы, и модель формируется в результате обработки соответствующих данных. Этот способ называется идентификацией [16].

Конструирование моделей по данным наблюдений включает три основных компонента:

- данные (учитывая возможные ограничения, необходимо выбрать максимально информативные данные о сигналах системы);
- множество моделей-кандидатов (множество моделей-кандидатов устанавливается посредством фиксации той группы моделей, в пределах которой мы собираемся искать наиболее подходящую);
- правило оценки степени соответствия испытываемой модели данных наблюдений (оценка качества модели связана, как правило, с изучением моделей в процессе их использования для воспроизведения данных измерений).

Существует несколько причин несовершенства моделей:

- численный метод не позволяет найти лучшую по выбранному критерию модели;
- критерий выбран неудачно;
- множество моделей оказалось неполноценным в том смысле, что в этом множестве нет «достаточно хорошего» описания системы;
- множество данных наблюдений не было достаточно информативным для того, чтобы обеспечить выбор хороших моделей.

По существу, главным в приложениях идентификации является итеративное решение всех этих вопросов, особенно третьего, на основе априорной информации и результатов предыдущих попыток, как показано на рисунке 1.

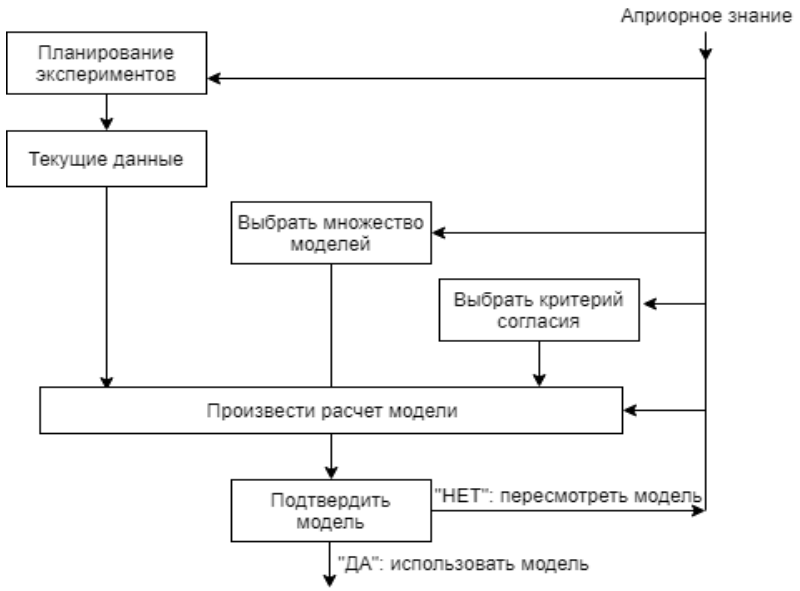

Рисунок 1 – Контур идентификации системы

1.3 Параметрическая идентификация

При параметрическом моделировании (или идентификации в узком смысле) предполагают, что структура объекта каким-то образом выбрана на основании имеющейся априорной информации с точностью до параметров.

Следующий основной этап состоит в оценке этих параметров на основании текущей информации.

Итак, построение параметрической модели состоит из двух основных этапов: определения класса параметрических структур с точностью до параметров и их последующего оценивания по результатам наблюдения входных-выходных переменных исследуемого объекта. Однако на практике объема априорной информации часто недостаточно для обоснованного выбора класса моделей с точностью до параметров. Тем самым неизбежно возникает большая или меньшая неточность на стадии формулировки задачи идентификации [13].

При построении моделей разнообразных дискретно-непрерывных процессов в настоящее время доминирует теория параметрической идентификации, или идентификация в узком смысле [26; 24]. Ее содержание, состоит в том, что на первом этапе на основании имеющейся априорной информации определяется параметрический класс операторов A^{α}, например:

$$\tilde{x}_{\alpha}(t) = A^{\alpha}(u(t), \alpha), \quad (1)$$

где A^{α} – параметрическая структура модели;

α – вектор параметров.

На втором этапе осуществляется оценка параметров α на основе имеющейся выборки $\{x_i,\ u_i,\ i = \overline{1,s}\}$, где s – объем выборки.

Существует большое количество методов получения оценок параметров, например, метод стохастических аппроксимаций [26]:

$$\alpha_s^l = \alpha_{s-1}^l + \gamma_s^l \left(x_s - \sum_{l=1}^{N} \alpha_{s-1}^l \varphi_l(u_s)\right) \varphi_l(u_s),\ l = \overline{1,N}, \quad (2)$$

где γ_s^l – структура стохастической модели;

$l = \overline{1,N}$ – коэффициенты Роббинса-Монро, удовлетворяющие определенным условиям сходимости [24].

На практике обычно основное внимание уделяется задаче определения параметров объекта при заданной или принятой структуре [24]. Качество полученной модели зависит от того, насколько хорошо угадана параметрическая структура. Однако структура зависимости априорно не известна, т. е. она подбирается экспериментальным путем.

1.4 Непараметрическая идентификация

Априорная информация об объекте при непараметрической идентификации, или идентификации в широком смысле, отсутствует или очень бедная, поэтому приходится предварительно решать большое число дополнительных задач.

Идентификация в широком смысле предполагает отсутствие этапа выбора параметрического класса оператора. В этом случае задача идентификации состоит в оценивании класса операторов на основе выборки $\{x_i, u_i, i = \overline{1,s}\}$ в форме [13]:

$$\tilde{x}_s(t) = A_s\,(u(t), \overrightarrow{x_s}, \overrightarrow{u_s}), \qquad (3)$$

где $\overrightarrow{x_s} = (x_1, x_2, \ldots x_s)$, $\overrightarrow{u_s} = (u_1, u_2, \ldots u_s)$ – временные векторы.

Оценка оператора A_s в некоторых случаях может быть осуществлена средствами непараметрической статистики [17].

В качестве оценки (3) можно использовать непараметрическую оценку Надарая-Ватсона [17]:

$$x_s(u) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)}, \qquad (4)$$

где $\Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)$ – ядерная колоколообразная функция;

$i = \overline{1,s}$;

j = $\overline{1,m}$;

c_s – коэффициент размытости ядра, который удовлетворяет следующим условиям сходимости [17]:

- $c_s > 0$;
- $\lim\limits_{s\to\infty} c_s = 0$;
- $\lim\limits_{s\to\infty} sc_s^m = \infty$;
- $\Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right) \geq 0$;
- $\int_{\Omega(u)} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right) du^j = 1$;
- $\lim\limits_{s\to\infty} c_s \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right) = \delta\left(u^j - u_i^j\right)$.

В качестве колоколообразной функции $\Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)$ могут быть использованы ядра различного вида: треугольное, параболическое или кубическое, представленные на рисунке 2 [12].

Формула треугольного ядра:

$$\Phi\left(\frac{u-u_i}{c_s}\right) = \begin{cases} 1 - |c_s^{-1}(u-u_i)|, & |c_s^{-1}(u-u_i)| \leq 1; \\ 0, & |c_s^{-1}(u-u_i)| > 1; \end{cases} \quad (5)$$

Формула параболического ядра:

$$\Phi\left(\frac{u-u_i}{c_s}\right) = \begin{cases} 0.75(1 - (c_s^{-1}(u-u_i))^2), & |c_s^{-1}(u-u_i)| \leq 1 \\ 0, & |c_s^{-1}(u-u_i)| > 1 \end{cases} \quad (6)$$

Формула кубического ядра:

$$\Phi\left(\frac{u-u_i}{c_s}\right) = \begin{cases} (1+2|c_s^{-1}(x-x_i)|)(1-\ (c_s^{-1}(x-x_i))^2), & |c_s^{-1}(x-x_i)| \le 1 \\ 0, & |c_s^{-1}(x-x_i)| > 1 \end{cases} \quad (7)$$

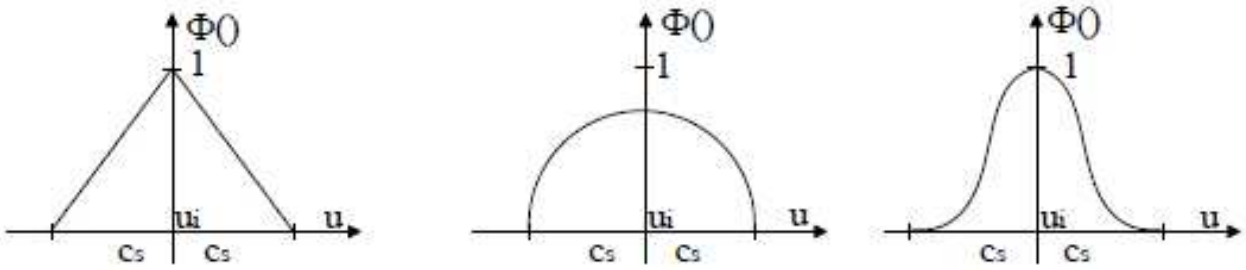

Рисунок 2 – Виды ядер колоколообразной функции.

Параметр размытости c_s при наличии обучающей выборки $\{x_i, u_i, i = \overline{1,s}\}$ находится из задачи минимизации квадратичного показателя соответствия выхода объекта и выхода модели, основанного на методе скользящего экзамена, когда в модели (4) по индексу i исключается k-e наблюдение переменой, предъявляемой для экзамена:

$$R(c_s) = \sum_{k=1}^{s}(x_k - x_s(u_k, c_s))^2 = \min_{c_s},\ k \neq i \quad (8)$$

то есть $i = k$, где индекс i фигурирует в формуле (4).

1.5 Анализ данных. Выбросы. Робастный анализ

Общая логика анализа исходных данных состоит из следующих этапов [5]:

Этап 1: Исходный (предварительный) анализ исследуемой реальной системы. В результате этого анализа определяются:

- основные цели исследования на неформализованном, содержательном уровне;

- совокупность единиц, представляющих предмет статистического исследования;

- перечень ($x^{(1)}$, $x^{(2)}$, … $x^{(p)}$) отобранных из представленного специалистами априорного набора показателей, характеризующих состояние (поведение) каждого из обследуемых объектов, который предполагается использовать в данном исследовании;

- степень формализации соответствующих записей при сборе данных;

- общее время и трудозатраты, отведенные на планируемые работы;

- моменты, требующие предварительной проверки перед составлением детального плана исследования;

- формализованная постановка задачи;

- формы, используемые для сбора первичной информации и для введения ее в ЭВМ.

Этап 2: составление детального плана сбора исходной статистической информации. При составлении этого плана необходимо, по возможности, учитывать полную схему дальнейшего статистического анализа, о чем часто забывают. Априорное представление о том, как и для чего данные будут анализироваться, может оказать существенное влияние на их сбор.

Этап 3: Сбор исходных статистических данных и их введение в ЭВМ. Одновременно в ЭВМ вносятся полные и краткие определения используемых терминов. В пакете должны быть предусмотрены специальные меры, исключающие или резко уменьшающие возможность появления расчетов не с тем подмножеством данных или не для той подгруппы объектов.

Таким образом, к моменту определения основного инструментария статистического исследования исследователь в общем случае располагает в качестве массива исходных статистических данных временной последовательностью матриц вида:

$$X(t) = \begin{pmatrix} x_1^{(1)}(t), x_2^{(1)}(t), \dots, x_n^{(1)}(t) \\ x_1^{(2)}(t), x_2^{(21)}(t), \dots, x_n^{(2)}(t) \\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots\ \dots \\ x_1^{(p)}(t), x_2^{(p)}(t), \dots, x_n^{(p)}(t) \end{pmatrix} (t = t_1, \dots t_N), \quad (10)$$

где $x_i^{(k)}(t)$ – значение *k*-ого признака, характеризующего состояние *i*-го объекта в момент времени *t*.

4 Этап: Первичная статистическая обработка данных.

В ходе первичной статистической обработки данных обычно решаются следующие задачи:

- отображение переменных, описанных текстом, в нормальную или ординальную шкалу;
- статистическое описание исходных совокупностей с определением пределов варьирования переменных;
- анализ резко выделяющихся наблюдений;
- восстановление пропущенных наблюдений;
- проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных;
- унификация типов переменных;
- экспериментальный анализ закона распределения исследуемой генеральной совокупности и параметризации сведений о природе изучаемых распределений.

На практике часто встречается ситуация, когда среди большинства наблюдений X_t, хорошо описываемых каким-либо вероятностным распределением, например, нормальным, встречается небольшая доля нетипичных наблюдений, называемых выбросами. Основными причинами выбросов являются небрежность при сборе данных и неточность математической модели. В этом случае классические методы максимального правдоподобия и наименьших квадратов теряют свою эффективность, более того, они могут выдавать абсурдные результаты [7].

Тем не менее, существую оценки, устойчивые к выбросам, они называются робастные. Основная идея робастности – это построение статистических процедур, устойчивых к возможным отклонениям от принятых вероятностных моделей распределений данных.

Применение робастных алгоритмов является значимым, если исследователь стремится достигнуть высокой точности результатов моделирования. Например, для медицинской отрасли точность является одним из наиболее важных факторов. В статье [23] описан широкий класс задач, решаемых с помощью методов робастной статистики. С помощью разработанных компьютерных систем повышается точность диагностики различных заболеваний, решаются задачи управления банковскими рисками, в частности, эконометрическое прогнозирование и многое другое. Помимо этого, развиваются и совершенствуются уже известные методы робастного оценивания [7].

Однако от выбросов необязательно всегда избавляться. В. И. Тихонов в своей монографии [22] приводит практические примеры, основанные на изучении характеристик и свойств выбросов. Например, выбросы измеряются в большинстве сейсмических или медицинских приборах, в частности, медицинские приборы регистрируют биотоки сердца и мозга, базирующиеся на измерении высоты и длительности выбросов электроосциллограмм, а также интервалы между выбросами. Некоторые резко отличающиеся измерения, полученные при помощи электроэнцефалограмм, могут свидетельствовать о патологиях здоровья человека.

Выводы по главе 1

В первой главе магистерской диссертации были рассмотрены теоретические аспекты данного исследования. Были представлены такие понятия как моделирование, идентификация, а так же краткое описание этапов анализа данных. Изучив эти понятия можно построить обобщенно структурную

схему моделирования: выбор метода моделирования, основанный на количестве и качестве априорной информации, а также на требованиях, поставленных перед аналитиком. Если в данных присутствуют отклоняющиеся от нормы значения (например, пропуски или выбросы), то решаются вопросы по их устранению.

Из всех рассмотренных в главе понятий стоит выделить непараметрическую идентификацию (так как в следующих параграфах будут рассмотрены методы и алгоритмы, основанные на ней) и анализ данных.

2 Алгоритмы обработки данных с выбросами

2.1 Классификация методов обработки данных с выбросами

Методы обработки данных с выбросами будут различаться в зависимости от типа аномальных наблюдений. Каждый метод имеет свои границы применения, особенности и формат представления результата. Ряд алгоритмов может использоваться в случае больших выборок необработанных или минимально обработанных данных, к которым может применяться закон больших чисел.

Среди методов обработки данных с выбросами можно выделить:

а) Методы обнаружения выбросов в данных [9]:

- статистические тесты (как правило, используются для выявления экстремальных значений);
- модельные тесты (сравнение точек модели и объекта, и те точки, которые сильно отклоняются от объекта и считаются аномалией);
- итерационные тесты (методы, которые состоят из итераций, на каждой из которых удаляется группа «особо подозрительных объектов»);
- метрические методы (в них постулируется существование некоторой метрики в пространстве объектов, которая и помогает найти аномалии);
- методы машинного обучения.

б) Робастные методы. Как уже упоминалось в главе 1, робастные методы способны обрабатывать данные с аномальными значениями, при этом нивелируя их влияние на всю выборку. В качестве примера можно выделить следующие алгоритмы:

- алгоритм RIPPER, предложенный Вильямом Кохеном [1], как модификация его же алгоритма IREP. Класс таких задач реализует правило пропозиционального (относящегося к высказываниям или предложениям) обучения. Он основан на ассоциативных правилах с REP (reduced error pruning – сокращение обрезки ошибок). Данная методика была достаточно

распространенной и эффективной для алгоритмов, построенных с помощью дерева решений. Для REP исходная выборка разбивается на два набора: «растущий» набор и набор «подрезки». На первом этапе формируется «растущий» набор правил, с использованием какого-либо эвристического правила. Данный избыточный набор правил затем многократно упрощается, с помощью одного из множества операторов «обрезки». Таким оператором, к примеру, может служить удаление любого единственного условия или любого единственного правила. На каждом этапе упрощения выбирается такой оператор «обрезки», который дает наибольшее уменьшение ошибки на наборе «обрезки». Упрощение заканчивается, когда применение любого оператора «обрезки» приводит к увеличению ошибки в наборе обрезки.

- алгоритм, представленный в работе Кирик Е. С. [11], основанный на непараметрической оценке Розенблата-Парзена и предусматривающий не только робастную оценку регрессии, но так же «ремонт» данных-выбросов.

2.2 Непараметрический алгоритм исключения выбросов из выборки наблюдений

Приведем подробное описание непараметрического алгоритма по исключению выбросов, исследуемого в ходе данной работы.

На рисунке 3 представлена общая схема исследуемого процесса, принятая в теории идентификации [26], где $u(t)$ – вектор входного воздействия; $x(t)$ – вектор выходных переменных; A – неизвестный оператор; G^x, G^u – блоки контроля переменных, подверженные воздействию случайных помех $g^u(t)$ и $g^u(t)$; u_t и u_x – измерения переменных и в дискретные моменты времени; $\xi(t)$ – векторная случайная помеха. Измерения «входных-выходных» переменных объекта поступают на блок «Модель», где на основании заданного алгоритма находятся значения выхода модели x_{st}.

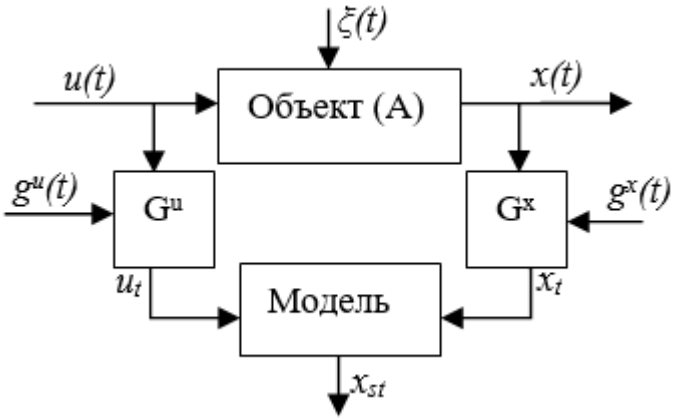

Рисунок 3 – Общая схема идентификации дискретно-непрерывного процесса

Алгоритм основан на сравнении значений выхода объекта x и выхода модели x_s. Если $|x - x_s| > \varepsilon$, где ε – параметр алгоритма, то точка становится претендентом на выброс. Из всех претендентов удаляется точка с наибольшим значением отклонения. Процесс повторяется до тех пора, пока в выборке не останется точек, удовлетворяющих условию $|x - x_s| > \varepsilon$.

Для построения модели была использована непараметрическая оценка Надарая-Ватсона (4), в качестве колоколообразной функции было использовано ядро треугольного вида (5).

Оптимальный параметр размытости здесь может быть определен по формуле (8).

Ошибка моделирования рассчитывалась по формуле:

$$W = \frac{1}{S}\sum_{i=1}^{S} | x^i - x_s^i |, \qquad (11)$$

где S – объем выборки;

x^i – объект;

x_s^i – модель.

Далее рассчитывался параметр Δ между объектом и моделью по формуле:

$$\Delta_i = | x_s^i - x^i |, \qquad (12)$$

где x^i – объект;

x_s^i – модель объекта.

Значение отклонения модели от объекта сравнивалось со значением допустимой границы (ε), и те значения $\hat{x}_i$ и x_i $\mathrm{x_i}$, при которых Δ_i был максимальным, считались выбросами и исключались из выборки.

Так алгоритм итеративно применяется к выборке, пока массив из Δ_i не становился равным нулю.

Предлагаемый алгоритм позволяет обнаруживать и исключать выбросы из выборки наблюдений, что в свою очередь позволяет повысить точность решения задачи идентификации.

2.3 Модификация непараметрического алгоритма исключения выбросов из выборки наблюдений

Главным недостатком описанного ранее алгоритма исключения выбросов из выборки наблюдений является ручная настройка параметров размытости ядра и допустимой границы.

Для автоматической настройки параметра допустимой границы (ε) был использован метод золотого сечения одномерной оптимизации [2].

Смысл применения данного метода к алгоритму по исключению выбросов заключается в следующем:

1) На первой итерации заданный отрезок (вектор ошибок моделирования) делится двумя симметричными относительно его центра точками и рассчитываются значения в этих точках.

2) После чего тот из концов отрезка, к которому среди двух вновь поставленных точек ближе оказалась та, значение в которой максимально (для случая поиска минимума), отбрасывают.

3) На следующей итерации в силу показанного выше свойства золотого сечения уже надо искать всего одну новую точку.

4) Процедура продолжается до тех пор, пока не будет достигнута заданная точность.

Для автоматической настройки параметра размытости ядра (c_s) был выбран метод скользящего экзамена [15]. Данный метод является разновидностью метода перекрёстной проверки, в котором исходная выборка делится на k частей, причём $k = S$, где S – объём исходной выборки. То есть на каждой итерации поочередно каждый клиент удаляется из исходной выборки, а на оставшейся части строятся классификаторы, с помощью которых затем выполняется прогноз удаляемого объекта. После сравнения прогнозируемого объекта кредитоспособности с реальным, удаляемый объект возвращается в исходную выборку. Таким образом, данная процедура повторяется с каждым объектом исходной выборки.

Автоматическая настройка параметров непараметрического алгоритма позволила сократить время его работы.

Блок-схема модифицированного алгоритма выглядит следующим образом (рисунок 4):

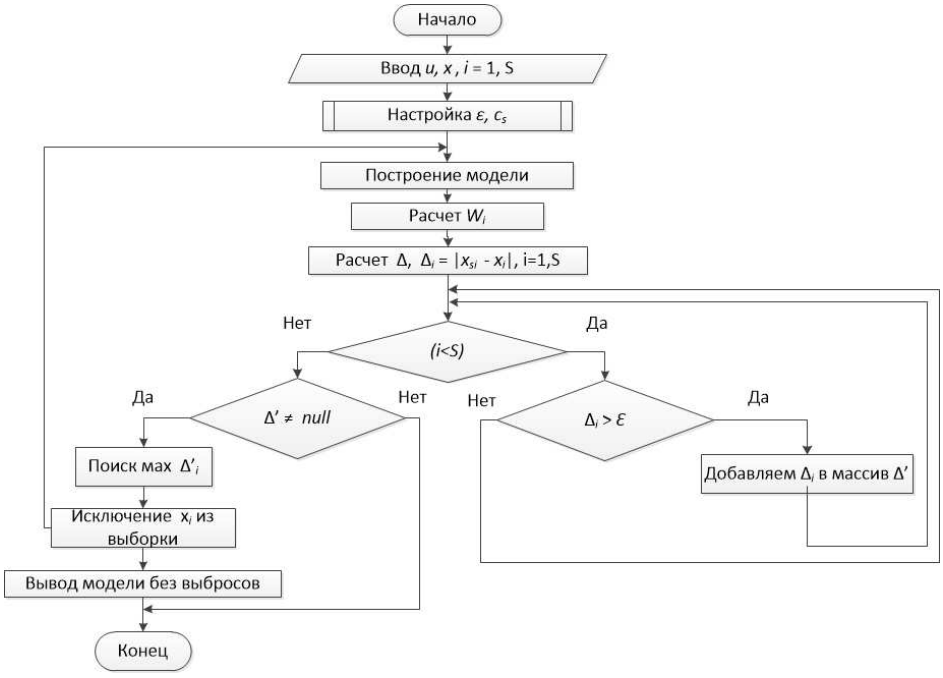

Рисунок 4 – Блок-схема модифицированного непараметрического алгоритма исключения выбросов

На блок-схеме введены следующие обозначения:

- u – входной вектор;
- x – выходной вектор;
- c_s – степень размытости ядра;
- ε – допустимая граница рассогласований значений модели и объекта;
- Δ – массив рассогласований модели и объекта;
- Δ’ – массив рассогласований модели и объекта, превышающих ε.

Модифицированный алгоритм был протестирован на модельных данных. Была взята выборка из 100 наблюдений (S = 100), входная переменная u_i, $i = \overline{1,s}$ – равномерная сетка с шагом 1, начальные значения допустимой границы были приняты в диапазоне $\varepsilon = \overline{0.1,0.9}$ изменялись с шагом 0.1, начальные значения параметра размытости ядра были приняты в диапазоне $c_s = \overline{0.1,10}$ и так же изменялись с шагом 0.1.

Эксперименты проводились при трех различных зависимостях:

- $x_i = 0{,}5 * \sin(u_i * 0.1)$;
- $x_i = \log u_i$;
- $x_i = u_i^3$.

После вычисления значений выходной переменной (x_i) в нее были введены аномальные значения, так называемые «выбросы». Затем к выходной переменной был применен алгоритм.

Результаты применения алгоритма можно посмотреть на рисунках 5, 6, 7. Черной линией представлен график выходных переменных, красной – вектор входных переменных.

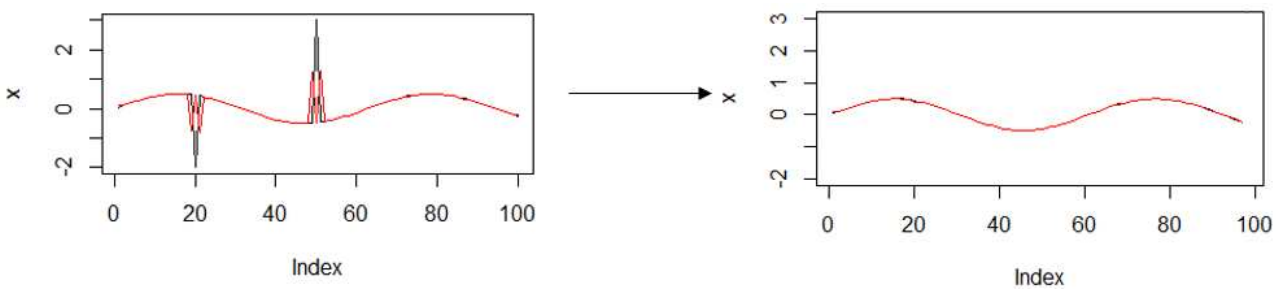

Рисунок 5 – Результат применения модифицированного алгоритма исключения выбросов при $x_i = 0{,}5 * \sin(u_i * 0.1)$

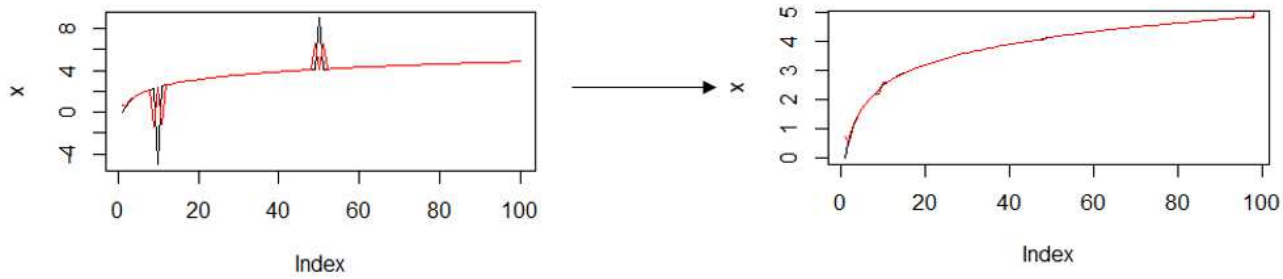

Рисунок 6 – Результат применения модифицированного алгоритма исключения выбросов при $x_i = \log u_i$

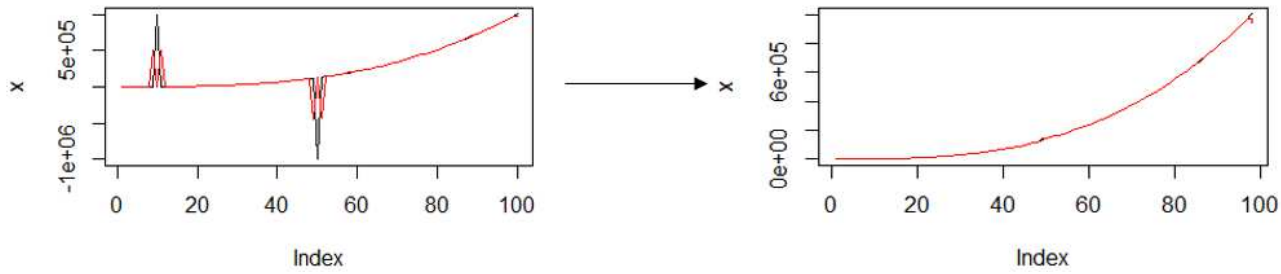

Рисунок 7 – Результат применения модифицированного алгоритма исключения выбросов при $x_i = u_i^3$

В каждом представленном случае ошибка моделирования (W) значительно уменьшалась. Изменения ошибок моделирования после применения алгоритма представлены в таблице 1.

Таблица 1 – Изменение ошибок моделирования после применения модифицированного алгоритма исключения выбросов

Вид зависимости	*W* до применения алгоритма	*W* после применения алгоритма
$x_i = 0{,}5 * \sin(u_i * 0.1)$	0,44	0,07
$x_i = \log u_i$	0,33	0,19
$x_i = u_i^3$	0,21	0,08

2.4 Непараметрический алгоритм восстановления пропусков в данных

Непараметрический алгоритм исключения выбросов, рассмотренный ранее, удаляет аномальные значения из выборки наблюдений, однако на месте этих удаленных значений останутся пропуски, которые так же отрицательно влияют на точность решения задачи идентификации. Поэтому после удаления необходимо эти пропуски как-то восстановить, для этого рассмотрим непараметрический алгоритм по восстановлению пропусков в данных, предложенный Корнеевой А. А. и Сергеевой Н. А. [14].

Непараметрический алгоритм восстановления пропусков основывается на непараметрической оценке (4) и состоит из трех этапов: настройка коэффициента размытости ядра, заполнение пустых ячеек выборки наблюдений, построение непараметрической оценки по заполненной выборке наблюдений.

Рассмотрим подробнее работу алгоритма на примере. У нас имеется вектор входного воздействия *u(t)* с пропусками, вектор выходного воздействия с пропусками *x(t),* объем выборки *S*.

На первом этапе алгоритма по заполненным значениям выборки наблюдений восстанавливается непараметрическая оценка x_s' вида:

$$x'_s(u) = \frac{\sum_{i=1}^{s'} x_i \prod_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)}{\sum_{i=1}^{s'} \prod_{j=1}^{m} \Phi\left(c_s^{-1}\left(u^j - u_i^j\right)\right)}, \quad (13)$$

На данном этапе работа ведется только по заполненным значениям выборки наблюдений «входных-выходных» переменных процесса (т.е. ячейки с пропусками не учитываются). Так как объем всей выборки равен *S*, то объем выборки «полных» наблюдений обозначим *S′ (S′ < S)*.

После восстановления оценки (13) согласно критерию (8) настраивается значение коэффициента c_s' по выборке, объемом S'.

На втором этапе заполняются пустые ячейки выборки наблюдений с использованием модели x_s' и оптимального значения коэффициента размытости c_s', полученных на первом этапе. В тех ячейках, где наблюдения переменной *x* пропущены, в оценку $x_s'(u_1, u_2, \dots, u_m)$ в ядерную функцию $\prod_{j=1}^{m} \Phi\left(c_s'^{-1}\left(u^j - u_i^j\right)\right)$ вместо текущих значений u_j подставляем значения измеренных $u = (u_1, u_2, \dots, u_m)$ и вычисляем соответствующую оценку x_s, которой восполняем недостающие наблюдения *x*.

На заключительном этапе строится непараметрическая оценка по всей имеющейся выборке с заполненными значениями объема *S*. При этом коэффициент размытости настраивается по всей имеющейся выборке объема *S* еще раз, так же по критерию (8).

Особый интерес задача заполнения матрицы наблюдений представляет в тех случаях, когда количество пропусков достаточно велико, располагаются они не равномерно, то есть в хаотичном порядке (как по входным, так и по выходным переменным), а объем исходной выборки не велик.

Принцип работы предлагаемой методики сохраняется. В случае различной дискретности контроля «входных-выходных» переменных процесса, мы знаем переменные, по которым присутствуют пропуски. Например, в зависимости от того, по какой из переменных мы имеем пропуск, оценка (13) может принимать вид:

$$u_{ks}(x,u) = \frac{\sum_{i=1}^{S} u_{ki}\Phi\left(\frac{x-x_i}{c_S}\right)\prod_{j=1,j\neq k}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_S}\right)}{\sum_{i=1}^{S} u_{ki}\Phi\left(\frac{x-x_i}{c_S}\right)\prod_{j=1,j\neq k}^{m}\Phi\left(\frac{u_j-u_{ji}}{c_S}\right)} \quad (14)$$

Выводы по главе 2

Во второй главе магистерской диссертации были представлены методы работы с данными с выбросами. Были описаны общие методы обнаружения выбросов в данных, а так же два метода робастной статистики.

Далее подробно были описаны алгоритмы, которые планируется применить к практической задаче. Это непараметрические алгоритмы по восстановлению пропусков в данных и по обнаружению и исключению выбросов.

Был выбран именно этот непараметрический алгоритм по обнаружению и исключению выбросов так как практическая задача отвечает условиям малой априорной информации, а следовательно речь идет о непараметрической идентификации, тогда как перечисленные выше методы обнаружения выбросов в данных работают в условиях параметрической идентификации.

Однако этот алгоритм имел недостаток в виде ручной настройки параметра размытости ядра и параметра допустимой границы. Поэтому была произведена модификация данного алгоритма, которая заключалась в автоматизации настройки этих параметров. Таким образом было сокращено время работы этого алгоритма.

Модифицированный непараметрический алгоритм исключения выбросов был протестирован на модельных данных и показал хорошие результаты, поэтому можно сказать, что алгоритм готов к применению на практической задаче.

3 Применение исследуемых алгоритмов к решению практической задачи

3.1 Постановка практической задачи

Задача медицинской диагностики состоит в определении возможных диагнозов больного на основе знаний предметной области и данных его обследования, к которым относят значения признаков (в моменты их наблюдения), значения анатомо-физиологических особенностей (постоянные во времени) и значения произошедших событий (в моменты, когда они происходили).

Исходные данные в практической задаче магистерской диссертации представляют из себя анализы на панкреатит по 31 признаку у 130 человек. Панкреатит классифицируется по трем степеням тяжести. Предполагается, что некоторые признаки содержат сильно отклоняющиеся от среднего значения. Таким образом, практическая задача заключается в нахождении этих отклоняющихся значений, проверка этих значений на выбросы и исключение выбросов из выборки наблюдений.

При решении практических задач анализа медицинских и биологических наблюдений для последующего принятия решения о диагностике и прогнозировании состояния исследователю приходится иметь дело с совокупностью одновременно зафиксированных на объекте исследования количественных и категориальных признаков ($x_1, x_2, \ldots, x_p$).

В практической задаче признаки распределены по трем шкалам, представленных на рисунке 8:

- количественные: 505, 510, 511, 518, 523, 524, 526, 527, 534, 535, 539, 540, 541, 542, 545, 548, age, 512, 529, 530_532, mms;
- булевы: 513, 514, 516, 557, liquid, lpu, real_us;
- категориальные: 515, 567, heavy, nans.

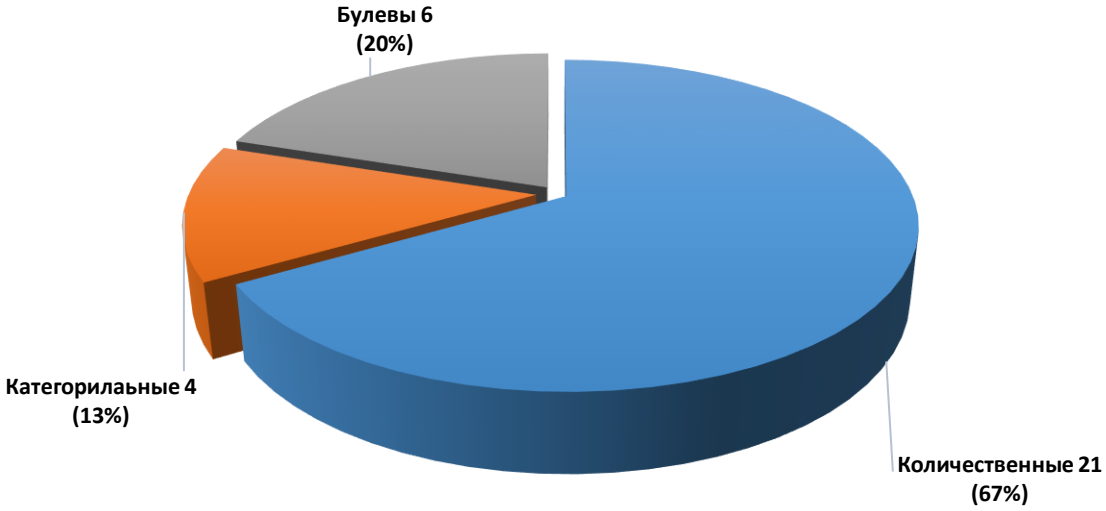

Рисунок 8 – Распределение признаков по шкалам

3.2 Анализ практической задачи на возможность применения исследуемого алгоритма

На первом этапе анализа данных было решено выявить признаки, имеющие отклоняющиеся от нормы значения. Для этого были построены гистограммы по всем признакам, с помощью которых были определены 8 признаков с подозрением на выбросы. Гистограммы с этими признаками представлены на рисунке 9.

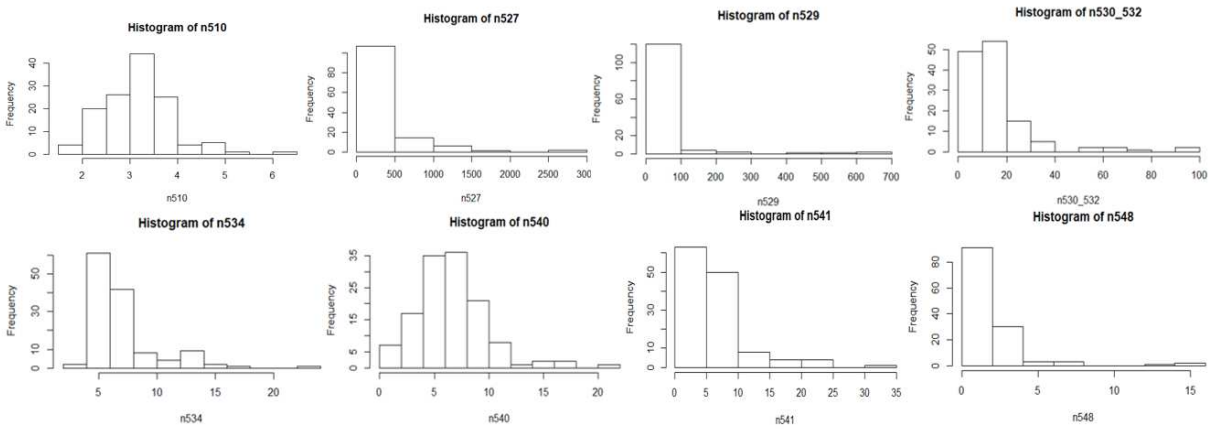

Рисунок 9 – Признаки с подозрением на выбросы

Для того, чтобы применить модифицированный алгоритм исключения выбросов к этим признакам необходимо построить их непараметрические модели. В свою очередь для того, чтобы построить их модели – необходимо подробнее исследовать данные практической задачи.

Поэтому следующим шагом была проверка на линейные зависимости между признаками. Был проведен корреляционный анализ, при котором связь между признаками рассчитывалась по формуле:

$$r_{xy} = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}}, \quad (15)$$

где x_i – значение перемнной *X*;

y_i – значение переменной *Y*;

$\bar{x}$ – среднее арифметическое для переменной *X*;

$\bar{y}$ – среднее арифметическое для переменной *Y*.

Связи между признаками могут быть слабыми и сильными (тесными). Их критерии оцениваются по шкале Чеддока [10]:

- $0.1 < r_{xy} < 0.3$: слабая;
- $0.3 < r_{xy} < 0.5$: умеренная;
- $0.5 < r_{xy} < 0.7$: заметная;
- $0.7 < r_{xy} < 0.9$: высокая;
- $0.9 < r_{xy} < 1$: весьма высокая.

В данном исследовании корреляционный анализ в большинстве случаев показал умеренные и слабые линейные связи между признаками, заметной связью выделились всего 6 пар признаков из 420, показанных в таблице 2. Полную таблицу с корреляционной связью можно посмотреть в приложении А.

Таблица 2 – Пары признаков с заметной линейной связью

Признаки	Кор. связь
511, 513	0.631
Heavy, lpu	0.619
Heavy, 545	0.592
Heavy, liquid	0.584
Lpu,505	0.540
Lpu, 545	0.517

Соотношение всех пар признаков с различными корреляционными связями показаны на рисунке 10.

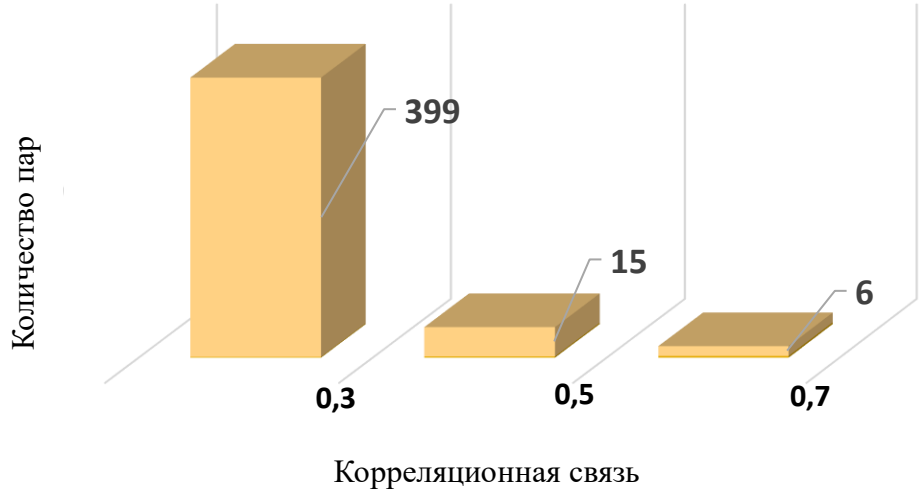

Рисунок 10 – Соотношение пар с различными корреляционными связями.

Так как корреляционный анализ показал слабую линейную связь между признаками, далее мы начали искать нелинейные зависимости. Для этого были решено строить непараметрические прогнозы признаков.

Для построения прогнозов были выбраны те комбинации признаков, которые позволяли рассчитать наибольшее количество значений прогнозируемых признаков. Такие признаки представлены в таблице 3.

Таблица 3 – Комбинации признаков, позволяющие прогнозировать наибольшее количество значений

Комбинации признаков	Количество прогнозируемых значений признака, %
513, 514, real_us, lpu	100
516, liquid, real_us, lpu	98,46
513, 516, 557, real_us	97,69
557, liquid, real_us, lpu	96,92
514, 557, liquid, lpu	96,92
…	…
513, 514, real_us, lpu, 515, 516, 557, 567, liquid	70
…	….

Так как данная оценка (4) является методом локальной аппроксимации, то при расчете модели используется произведение колоколов, который в случае с булевыми и категориальными признаками должны возвращать 1, если значения признаков совпадают и 0 иначе:

$$\Phi(x_{si} - x_i) = \begin{cases} 1, x_{si} = x_i \\ 0, x_{si} \neq x_i \end{cases} \quad (16)$$

Таким образом, при прогнозировании по булевым и категориальным признаком параметр размытости ядра не учитывался.

В следующем вычислительном эксперименте были построены непараметрические модели качественных признаков исследуемого объекта на основе лишь категориальных и булевых. Для этого были выбраны следующие сочетания булевых и категориальных признаков, так как именно они позволяют получить большее число спрогнозированных объектов:

- при комбинации булевых признаков, который позволяют посчитать 100% (130 из 130) значений (513, 514, real_us, lpu);

- при комбинации категориальных и булевых признаков, которые прогнозировали 70% (91 из 130) значений (513, 514, real_us, lpu, 515, 516, 557, 567, liquid).

Относительная ошибка прогнозирования считалась по следующей формуле:

$$W = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{x_i - x'_i}{x'_i}\right)*100\%, \quad (17)$$

где n – количество спрогнозированных значений признака;

x_i – значения объекта;

x'_i – значение прогноза.

Дисперсия считалась по следующей формуле:

$$D = \sum_{i=1}^{n}\frac{(x_i - x_{\text{ср}})^2}{n-1}, \quad (18)$$

где $x_{ср}$ – среднее значение признака.

Среднеквадратичное отклонение считалось по следующей формуле:

$$\sigma = \sqrt{D} \quad (19)$$

Значения прогнозов рассчитывались по формуле (4) с колоколом (16):

$$x_s(u) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi\left(u^j - u_i^j\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(u^j - u_i^j\right)} \quad (20)$$

Результаты прогнозирования представлены в таблице 4.

Таблица 4 – Результаты прогнозирования количественных признаков

Признак	Относительная ошибка по 9 признакам (513, 514, real_us, lpu, 515, 516, 557, 567, liquid), %	Относительная ошибка по 4 признакам (513, 514, real_us, lpu), %	Дисперсия	СКО
505	0,54	0,64	0,38	0,61
510	18,01	17,55	0,71	0,84
511	25,17	22,53	0,6	0,77
518	16,68	14,51	22,21	4,78
523	38,52	39,87	112,9	10,62
524	17,75	15,71	0,78	0,88
526	54,26	55,81	6,39	2,52
527	342,82	311,72	447,66	21,15
534	36,35	27,88	3,01	1,73
535	12,87	14,04	0,7	0,84
539	56,51	61,13	12,04	3,47
540	65,81	49,4	3,2	1,79
541	61,8	58,19	4,72	2,17
542	2,49	2,22	4,36	2,09
545	92,56	100,91	18,82	4,34
548	104,99	138,01	2,52	1,59
512	18,56	15,33	0,54	0,74
529	167,49	126,21	102,63	10,13
530_532	105,43	90,67	15,55	3,94

Судя по таблице 4 можно сказать, что значения признаков прогнозировались с большой ошибкой. Особенно большая ошибка была у тех признаков, которые имели большое среднеквадратичное отклонение.

Следующей попыткой прогноза признаков было построение моделей признаков по всей выборке наблюдений.

Так как прогнозы строятся с помощью непараметрической оценки Надарая-Ватсона (4), была необходимость для каждого количественного

признака настраивать соответствующий параметр размытости ядра (c_s). Чтобы упростить этот процесс была произведена нормировка количественных признаков по формуле:

$$y_i(x_i) = \frac{x_i - \hat{m}}{\sigma}, \quad (21)$$

где y_i – нормированное значение признака;

x_i – исходное значение признака;

$\hat{m}$ – математическое ожидание признака.

Таким образом появилась возможность использовать одномерные значения параметра размытости ядра при его настройке.

Результаты прогнозирования по всей выборке наблюдения с нормированными количественными признаками приведен в таблице 5.

Таблица 5 – Относительная ошибка прогноза по всей выборке наблюдений

Признак	Относительная ошибка прогноза, %
510	431,1
518	127,8
505	71,1
511	330,9
523	104,3
524	243,5
526	140,9
527	107,1
534	200,1
535	168,9
539	290,7

Окончание таблицы 5

Признак	Относительная ошибка прогноза, %
540	706,5
541	198,2
542	214,1
545	123,3
548	997,3
512	4596,2
529	613,64
530_532	164,6

Проанализировав таблицу 5 можно сказать, что по всей матрице наблюдений модели строятся с большей ошибкой, чем только по булевым и категориальным признакам.

Так же были попытки прогнозировать по другим сочетаниям признаков. Например, к наборам булевых и категориальных признаков добавляли нормированные количественные признаки, которые имели наиболее сильную корреляционную связь с прогнозируемым признаком, однако и эти сочетания не показали хороших результатов. Таким образом, непараметрическую модель по наблюдениям практической задачи построить не удалось.

Так как предлагаемый в данном исследовании модифицированный непараметрический алгоритм исключения выбросов применяется именно к непараметрической модели объекта, то можно сделать вывод, что к данным поставленной практической задачи его применить не получится.

Следующим этапом анализа данных было построение параметрических моделей в студии машинного обучения Azure.

Microsoft Azure Machiny Learning Studio – это служба прогнозной аналитики, позволяющей за короткое время создавать, тестировать и управлять прогнозными моделями для решения задач аналитики. Исследователю

предлагаются готовые библиотеки алгоритмов для быстрого развёртывания прогнозных решений.

Azure представляет собой визуальное рабочее пространство, необходимое для создания, тестирования, обучения модели прогнозной аналитики. Исследователь перемещает наборы данных и модули анализа на рабочее пространство и связывает их вместе.

Модуль – это какой-либо алгоритм для работы с данными. Модули необходимы для ввода данных, оценки и проверки. В ходе построения эксперимента исследователь выбирает из каких модулей будет состоять его проект.

Данными манипуляциями он создаёт эксперимент, который будет выполнен в студии машинного обучения. Когда эксперимент будет готов, происходит его преобразование в прогнозный и исследователь решает, стоит ли опубликовать его как веб-службу, чтобы другие пользователи могли использовать созданный проект в своих целях.

Рассмотрим построение параметрической модели на примере 545 признака, схема представлена на рисунке 12.

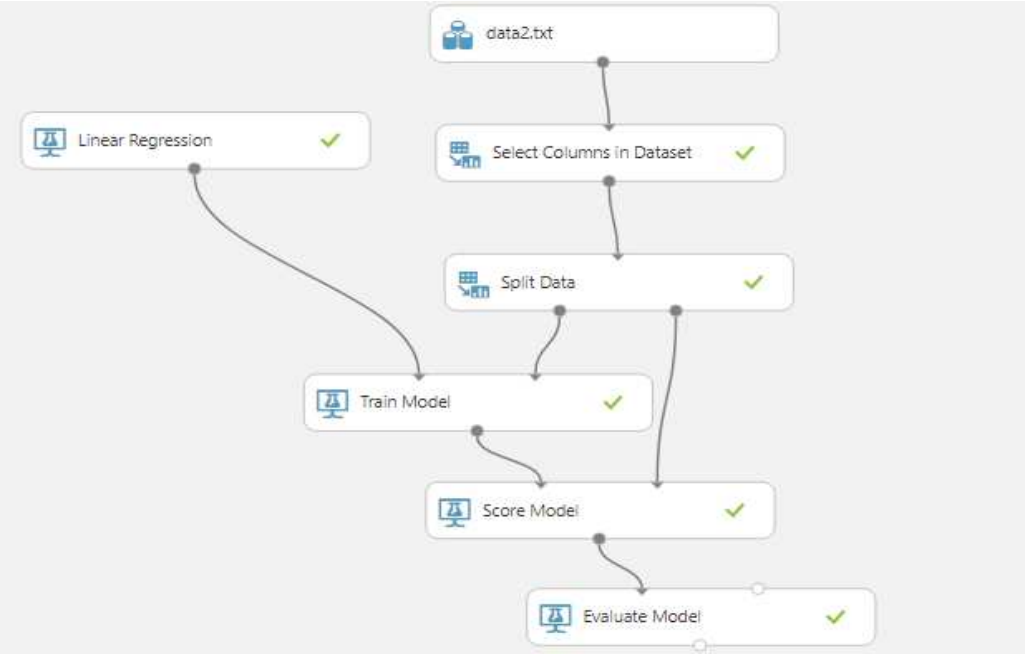

Рисунок 11 – Построение параметрической модели в Azure

Как видно на рисунке 11 наша схема состоит из следующих модулей:

- Dataset – модуль, в который мы загружаем нашу матрицу наблюдений;
- Select Columns In Dataset – модуль, в котором мы выбираем конкретные признаки (сочетание признаков), которые будут учувствовать в построении модели (прогнозируемый признак и признаки, по которым будет строиться прогноз);
- Split Data – модуль, в котором мы делим всю выборку наблюдений на обучающую и тестовую, в нашем примере мы разделили выборку на 4 части и 3 сделали обучающей, 1 – экзаменационной;
- Linear Regression – модуль, который устанавливает правило, по которому будет строиться прогноз (в нашем случае – линейная регрессия);

Линейная регрессия – используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости [8].

$$y(x) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n, \quad (22)$$

где b_i – параметры модели;

x_i – факторы (регрессоров) модели;

n – количество факторов модели.

- Train model – в этом модуле к данным из модуля Split Data применяется установленное правило (Linear Regression);
- Score Model – модуль, в котором считается прогноз;
- Evaluate Model – модуль, который оценивает ошибку построенного прогноза.

Значения прогноза, которые посчитались при данном эксперименте можно посмотреть в модуле Score Model – Visualize. Фрагмент итоговой таблицы представлен на рисунке 12.

rows 32 columns 30

567	liquid	lpu	512	529	530_532	nans	545	Scored Labels
0	1	1	3	24.2	9.4	4	65	48.03672
0	0	0	3.9	92	9.8	1	14	26.096933
0	0	0	1.5	24	12.1	3	4	3.484457
0	1	0	2.9	67	9.4	5	12	22.758986
0	0	0	2.3	24.2	8.9	8	46	9.248457
0	0	0	4.1	38	92.4	5	43	41.815731
0	1	0	3.186469	21	7.4	8	13	8.119812
0	0	0	3.856773	37.3	11.5	1	13	-0.252059

Рисунок 12 – Фрагмент параметрической модели 545 признака.

Последний столбец таблицы – Scored Labels – это и есть параметрическая модель 545 признака. И как видно на рисунке 12 – некоторые значения модели и признака расходятся очень сильно (например, 46 и 9.2).

Для того, что бы оценить качество моделирования в целом, обратимся к модулю Evaluate Model – Visualize (рисунок 13).

Mean Absolute Error	12.851024
Root Mean Squared Error	15.608523
Relative Absolute Error	0.89413
Relative Squared Error	0.819042
Coefficient of Determination	0.180958

Рисунок 13 – Ошибки моделирования 545 признака

Здесь мы видим следующие виды ошибок моделирования:

- Mean Absolute Error (Средняя абсолютная ошибка) – среднее значение арифметических отклонений;
- Root Mean Squared Error (Средняя среднеквадратическая ошибка) - квадратный корень из среднего значения возведенных в квадрат арифметических отклонений спрогнозированных значений тестового набора данных;
- Relative Absolute Error (Относительная абсолютная ошибка) - среднее арифметическое отклонение по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений;
- Relative Squared Error (Относительная квадратичная ошибка) - среднее арифметическое среднеквадратичных отклонений по отношению к абсолютной разнице между фактическими значениями и средним арифметическим всех фактических значений;
- Coefficient of Determination (Коэффициент детерминации) - статистический показатель, который оценивает соответствие модели данным.

Таким образом были построены модели для количественных признаков, имеющих наиболее сильную корреляционную связь, по наборам признаков из таблицы 3 (признаки, прогнозирующие наибольшее количество наблюдений), а так же по всей выборке наблюдений. Результаты построения параметрических моделей представлены в таблице 6.

Таблица 6 – Ошибки моделирования параметрических моделей

	Относительная ошибка модели по 9 признакам (513, 514, real_us, lpu, 515, 516, 557, 567, liquid), %	Относительная ошибка модели по 4 признакам (513, 514, real_us, lpu), %	Относительная ошибка модели по всем признакам, %
505	15	15	21
511	58	57	41
545	1303	1197	1285

По таблице 6 так же можно заметить, что ошибка моделирования достаточно большая, что говорит о том, что параметрическую модель по данным практической задачи так же построить не удалось.

Выводы по главе 3

В главе 3 была рассмотрена практическая задача, относящаяся к медицинской диагностике. Данные, предоставляемые в рамках этой задачи предполагают наличие выбросов, которые нам требовалось обнаружить и исследовать матрицу наблюдений на возможность применения к ней предлагаемого модифицированного непараметрического алгоритма исключения выбросов.

Данные анализировались в четыре этапа:

1) были построены гистограммы, с помощью которых мы выявили признаки с претендентами на выбросы;

2) был проведен корреляционный анализ, с целью поиска линейных зависимостей;

3) были построены различные параметрические модели, которые должны были показать возможность прогноза признаков для дальнейших их исследований;

4) были построены параметрические модели для признаков, имеющих наибольшую корреляционную связь.

Проведенный анализ данных показал слабую линейную зависимость, а большие ошибки при моделировании признаков говорят о том, что параметрические и непараметрические модели построить невозможно.

Так как для применения непараметрического алгоритма по исключению выбросов, исследуемого в данной работе, необходима параметрическая модель, можно сделать вывод, что к данной практической задаче он не применим.

Заключение

В магистерской диссертации рассмотрены теоретические аспекты исследования, такие как: задача моделирования, задача идентификации, а так же такие понятия как параметрическая и непараметрическая идентификация.

Так как задачи, поставленные в исследовании, связаны с алгоритмом по исключению выбросов, были рассмотрены методы обработки данных с выбросами, которые представляют из себя два класса: методы обнаружения и удаления данных с выбросами и методы робастной статистики.

Непараметрический алгоритм исключения выбросов, исследуемый в данной работе относится к классу методов обнаружения выбросов, однако среди рассмотренных методов не было таких, которые применимы в условиях непараметрической неопределенности.

В работе была предложена модификация непараметрического алгоритма, а именно: автоматизация настройки параметра размытости ядра и параметра допустимой границы. После модификации алгоритм был протестирован на экспериментальных данных, которые имели различные функциональные зависимости. Все тесты показали хорошие результаты, ошибка моделирования после применения алгоритма уменьшалась в разы.

Следующим этапом работы являлся анализ практической задачи на возможность применения к ней исследуемого алгоритма. Исходные данные были проанализированы различными способами, однако показывали отрицательны результат: слабую корреляционную связь и большую ошибку моделирования при построении различных параметрических и непараметрических моделей. Таким образом можно сделать вывод, о том что к данной практической задаче исследуемый алгоритм применить не удалось.

Список использованных источников

1. Cohen, W. W. Fast effective rule induction / W. W. Cohen // Proceedings of the twelfth international conference on machine learning. – 1995. – P. 115-123
2. Аббасов, М.Э. Методы оптимизации : учебное пособие / М. Э. Аббасов. – Санкт-Петербург : ВВМ, 2014. – 10 с.
3. Айвазян, С.А. Классификация многомерных наблюдений : учебное пособие / С. А. Айвазян. – Москва : Статистика, 1974. – 237 с.
4. Айвазян, С.А. Прикладная статистика. Классификация и снижение размерности: учебное пособие / С.А. Айвазян, В. М. Бухштабер, И.С. Енюков, Л.Д.Мешакин. – Москва : Финансы и статистика, 1989. – 608 с.
5. Айвазян, С.А. Прикладная статистика. Основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д.Мешакин. – Москва : Финансы и статистика, 1983. – 472 с.
6. Бурулько, Л.К. Математическое моделирование в электротехнике: учебное пособие / Л. К. Бурулько, Е. В. Овчаренко; под общ. ред. В.А. Бейнарович – Томск : Изд-во Томского политехнического университета, 2003. – 100 с.
7. Горяинов, В. Б. Робастное оценивание в авторегрессионной модели со случайным коэффициентом / В. Б. Горяинов, С. Ю. Ермаков // Наука и образование: научное издание МГТУ им. Н. Э. Баумана. – 2016. – №9. – 111–122 с.
8. Демиденко, Е. З. Линейная и нелинейная регрессия : учебное пособие / Е. З. Демиденко. – Москва: Финансы и статистика, 1981. — 302 с.
9. Дьяконов, А. Г. Некоторые технологии решения задач анализа данных. / А. Г. Дьяконов, А. А. Вороненко // Интеллектуализация обработки информации: 9-я международная конференция. – Москва : ТорусПрессМосква, – 2012. – 94-97 с.
10. Елисеева, И. И. Эконометрика: учебник / И. И. Елисеева − 2-е изд., перераб. и доп. – Москва: Финансы и статистика, 2007. − 576 с.

11. Кирик, Е. С. Моделирование и оптимизация робастных оценок функций по наблюдениям / Е. С. Кирик // Вычислительные технологии. – 2001. – Т. 6. – 351 – 355 с.

12. Корнеева, А. А. Непараметрические модели и алгоритмы управления для многомерных систем с запаздыванием: дис. к. т. н: 05.13.01 / Корнеева Анна Анатольевна. – Крансоярск, 2014. – 176 с.

13. Корнеева, А. А. О параметрическом моделировании стохастических объектов / А. А. Корнеева, Е. А. Чжан // Вестник СибГАУ. – 2013. – № 2 (48). – 37–42 с.

14. Корнеева, А.А., Непараметрическая идентификация дискретно-непрерывных процессов «трубчатой» структуры при наличии пропусков в данных / А.А. Корнеева, Н.А. Сергеева // Системы управления и информационные технологии. – 2012. – №4.1 (50). – 155-159 с.

15. Лбов, Г. С. Анализ данных и знаний : учебное пособие / Г. С. Лбов. – Новосибирск: Издательство НГТУ, 2001. − 90 с.

16. Льюнг, Л. Идентификация систем: учебное пособие / Л. Льюнг. – Москва : Наука, 1991. – 432 с.

17. Медведев, А. В. Анализ данных в задаче идентификации // Компьютерный анализ данных моделирования. Минск : Изд-во Белорус. гос. ун-та, 1995. Т. 2. – 201–206 с.

18. Медведев, А. В. О компьютерном исследовании К-моделей. / Т. В. Мальцева, А. В. Медведев. – Вестник СибГАУ. – 2013. – № 2 (48). –52 с..

19. Носик, Т.А. О применении методов идентификации в задачах технической диагностики : учебное пособие / Т. А. Носик // Математическое и компьютерное моделирование. – 2012, – 154 с.

20. Самарский, А. А. Математическое моделирование: Идеи. Методы. Примеры : учебное пособие / А. А. Самарский, А. П. Михайлов. – 2-е изд., испр. – Москва: ФИЗМАТЛИТ, 2005. – 7 с.

21. Советов, Б. Я. Моделирование систем: Учебное пособие для вузов / Б. Я. Советов, С. А. Яковлев. – 3-е издание, перераб. и доп. – Москва: Высш. шк., 2001. – 6 с.

22. Тихонов, В. И. Выбросы случайных процессов. / В. И. Тихонов. – Москва : Наука, 1970. – 392 с.

23. Харин, Ю. Робастная статистика и ее применение / Ю. Харин // Наука и инновации. – 2010. – Т. 8. – № 90. – 22-23 с.

24. Цыпкин, Я. З. Основы информационной теории идентификации. / Я. З. Цыпкин. – Москва : Наука, 1984. – 336 с.

25. Черепанов, Ф.М. Нейросетевой фильтр для исключения выбросов в статистической информации / Ф. М. Черепанов. – Пермь: Вестник пермского университета, №4(20), 2008г, – 151 с.

26. Эйкхофф, П. Основы идентификации систем управления / П. Эйкхофф. – Москва : Мир, 1975. – 681 с.

ПРИЛОЖЕНИЕ А – Корреляционные связи между признаками

	505	510	511	513	514	515	516	518	523	524	526	527	534	535	539	540	541	542	545	548	557	567	heavy	age	liquid	lpu	512	529	530_532	mms	nans	real_us
505 Корреляция Пирсона	1	,072	,057	,162	,071	-,159	,094	-,086	,139	-,012	,349**	,048	,023	,321**	,364**	,177*	,114	,294**	,309**	-,113	,457**	,392**	,442**	-,218*	,373**	,540**	,090	,003	,020	,235**	,323**	,014
Знач. (двухсторонняя)		,413	,523	,065	,423	,072	,287	,328	,116	,893	,000	,588	,795	,000	,000	,043	,195	,001	,000	,201	,000	,000	,000	,039	,000	,000	,308	,973	,825	,007	,000	,874
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
510 Корреляция Пирсона	,072	1	,631**	,173*	,279**	,094	,165	-,051	,071	-,076	,289**	,132	,065	-,067	-,187*	-,122	,111	-,048	,233**	-,090	-,103	,151	,175*	-,010	,129	,076	,418**	,063	,014	-,022	,078	,019
Знач. (двухсторонняя)	,413		,000	,049	,001	,287	,061	,562	,422	,388	,001	,134	,460	,449	,033	,165	,208	,591	,008	,310	,243	,086	,046	,925	,142	,388	,000	,477	,870	,807	,378	,830
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
511 Корреляция Пирсона	,057	,631**	1	,121	,271**	-,013	,017	-,103	,051	-,149	,202*	,129	,116	,062	-,084	-,032	,178*	-,053	,283**	,025	-,078	,171	,192*	-,091	,076	,070	,405**	,206*	,069	-,053	,145	,120
Знач. (двухсторонняя)	,523	,000		,171	,002	,885	,846	,242	,568	,092	,021	,144	,187	,486	,340	,722	,043	,547	,001	,774	,376	,052	,029	,393	,391	,431	,000	,019	,434	,547	,100	,173
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
513 Корреляция Пирсона	,162	,173*	,121	1	,187*	-,074	,007	-,019	,068	,036	,152	,020	,047	-,099	-,150	,026	,087	-,058	,205*	-,002	-,077	,055	,167	,172	,285**	,171	,084	,043	-,020	,166	,023	,140
Знач. (двухсторонняя)	,065	,049	,171		,033	,400	,937	,832	,442	,681	,084	,821	,592	,262	,089	,765	,323	,513	,019	,982	,386	,533	,057	,105	,001	,052	,342	,628	,824	,059	,796	,112
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
514 Корреляция Пирсона	,071	,279**	,271**	,187*	1	-,163	,037	-,166	,000	-,150	,116	,112	,150	-,147	,255**	-,052	,093	-,089	,301**	-,009	-,037	,087	,192*	,161	,226**	,192*	,172	,064	,017	,062	,115	,063
Знач. (двухсторонняя)	,423	,001	,002	,033		,063	,674	,059	,997	,089	,188	,207	,088	,095	,003	,556	,293	,312	,001	,921	,674	,326	,028	,129	,010	,029	,050	,470	,849	,482	,192	,476
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
515 Корреляция Пирсона	-,159	,094	-,013	-,074	-,163	1	-,098	,055	-,058	,095	-,127	-,065	-,061	,016	,150	,336**	-,097	,121	,259**	-,072	,329**	,031	,305**	,094	,251**	-,196*	-,084	-,145	-,031	-,105	,245**	-,103
Знач. (двухсторонняя)	,072	,287	,885	,400	,063		,270	,534	,512	,281	,150	,459	,489	,858	,089	,000	,273	,171	,003	,416	,000	,729	,000	,377	,004	,025	,342	,099	,723	,234	,005	,241
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130

Продолжение приложения А

	505	510	511	513	514	515	516	518	523	524	526	527	534	535	539	540	541	542	545	548	557	567	heavy	age	liquid	lpu	512	529	530_532	mms	nans	real_us
516 Корреляция Пирсона	,094	,165	,017	,007	,037	-,098	1	,055	,182*	,061	,044	,101	,026	,008	-,101	,036	,044	-,031	,121	,004	-,035	,093	,192*	-,018	,132	,136	-,053	-,129	-,118	,093	-,199*	,119
Знач. (двухсторонняя)	,287	,061	,846	,937	,674	,270		,531	,038	,493	,618	,251	,770	,927	,253	,686	,617	,728	,169	,961	,697	,293	,029	,866	,135	,123	,548	,144	,182	,294	,023	,177
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
518 Корреляция Пирсона	-,086	-,051	-,103	-,019	-,166	,055	,055	1	-,082	,817**	-,035	-,060	-,036	,333**	,106	,088	-,254**	-,059	-,423**	,211*	,190*	-,186*	-,353**	-,171	-,265**	-,309**	-,003	-,131	,013	-,104	,107	,083
Знач. (двухсторонняя)	,328	,562	,242	,832	,059	,534	,531		,355	,000	,695	,495	,683	,000	,231	,320	,004	,505	,000	,016	,030	,034	,000	,107	,002	,000	,974	,137	,885	,238	,228	,348
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
523 Корреляция Пирсона	,139	,071	,051	,068	,000	-,058	,182*	-,082	1	-,005	,269**	-,004	,021	-,100	-,123	,098	-,226**	-,159	,290**	,071	,083	,061	,235**	,063	,228**	,277**	-,088	-,177*	-,059	,248**	-,055	,074
Знач. (двухсторонняя)	,116	,422	,568	,442	,997	,512	,038	,355		,954	,002	,966	,811	,257	,163	,267	,010	,070	,001	,419	,349	,490	,007	,553	,009	,001	,317	,044	,503	,004	,534	,404
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
524 Корреляция Пирсона	-,012	-,076	-,149	,036	-,150	,095	,061	,817**	-,005	1	,062	-,074	-,077	,254**	,079	,022	-,140	,000	-,444**	,108	,171	-,125	-,291**	,000	-,228**	-,170	-,031	-,107	,017	-,055	,050	,138
Знач. (двухсторонняя)	,893	,388	,092	,681	,089	,281	,493	,000	,954		,487	,404	,383	,004	,369	,802	,113	,996	,000	,220	,052	,158	,001	,996	,009	,053	,728	,226	,849	,532	,574	,118
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
526 Корреляция Пирсона	,349**	,289**	,202*	,152	,116	-,127	,044	-,035	,269**	,062	1	,170	,090	-,267**	-,439**	-,102	,121	-,232**	,317**	-,019	-,231**	,301**	,350**	-,127	,218*	,268**	,138	,023	,089	,248**	-,103	,082
Знач. (двухсторонняя)	,000	,001	,021	,084	,188	,150	,618	,695	,002	,487		,053	,306	,002	,000	,249	,170	,008	,000	,833	,008	,001	,000	,235	,013	,002	,117	,793	,312	,004	,243	,355
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130

Продолжение приложения А

	505	510	511	513	514	515	516	518	523	524	526	527	534	535	539	540	541	542	545	548	557	567	heavy	age	liquid	lpu	512	529	530_532	mms	nans	real_us
527 Корреляция Пирсона	,048	,132	,129	,020	,112	-,065	,101	-,060	-,004	-,074	,170	1	-,012	-,007	-,180*	-,012	,073	-,028	,229**	-,065	-,019	,064	,121	,070	-,071	,038	,176*	,179*	,298**	-,078	-,063	,122
Знач. (двухсторонняя)	,588	,134	,144	,821	,207	,459	,251	,495	,966	,404	,053		,896	,933	,041	,891	,409	,752	,009	,461	,833	,473	,169	,510	,420	,668	,045	,042	,001	,375	,479	,167
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
534 Корреляция Пирсона	,023	,065	,116	,047	,150	-,061	,026	-,036	,021	-,077	,090	-,012	1	-,041	,309**	-,031	,144	,316**	,180*	-,180*	-,071	,141	,245**	,209*	,104	,180*	,119	-,025	-,083	-,050	,005	,109
Знач. (двухсторонняя)	,795	,460	,187	,592	,088	,489	,770	,683	,811	,383	,306	,896		,646	,000	,730	,102	,000	,040	,040	,420	,109	,005	,048	,238	,040	,177	,780	,349	,573	,959	,218
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
535 Корреляция Пирсона	,321**	-,067	,062	-,099	-,147	,016	,008	,333**	-,100	,254**	,267**	-,007	-,041	1	,300**	-,078	-,107	,098	,301**	,205*	,211*	,300**	,301**	-,107	,316**	,376**	-,107	-,077	-,130	-,164	,213*	,030
Знач. (двухсторонняя)	,000	,449	,486	,262	,095	,858	,927	,000	,257	,004	,002	,933	,646		,001	,375	,224	,265	,001	,019	,016	,001	,001	,314	,000	,000	,225	,382	,141	,063	,015	,739
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
539 Корреляция Пирсона	,364**	-,187*	-,084	-,150	,255**	,150	-,101	,106	-,123	,079	,439**	-,180*	,309**	,300**	1	-,060	,374**	,206*	,452**	,216*	,284**	,267**	,465**	-,032	,294**	,496**	-,022	-,024	-,171	,260**	,276**	-,055
Знач. (двухсторонняя)	,000	,033	,340	,089	,003	,089	,253	,231	,163	,369	,000	,041	,000	,001		,499	,000	,019	,000	,013	,001	,002	,000	,766	,001	,000	,806	,785	,051	,003	,002	,534
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
540 Корреляция Пирсона	,177*	-,122	-,032	,026	-,052	,336**	,036	,088	,098	,022	-,102	-,012	-,031	-,078	-,060	1	-,148	-,075	,179*	-,038	-,102	,053	,255**	-,057	,274**	,239**	-,062	,080	,054	,229**	,332**	,213*
Знач. (двухсторонняя)	,043	,165	,722	,765	,556	,000	,686	,320	,267	,802	,249	,891	,730	,375	,499		,092	,398	,042	,667	,248	,553	,003	,593	,002	,006	,484	,367	,543	,009	,000	,015
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130

Продолжение приложения А

	505	510	511	513	514	515	516	518	523	524	526	527	534	535	539	540	541	542	545	548	557	567	heavy	age	liquid	lpu	512	529	530_532	mms	nans	real_us
541 Корреляция Пирсона	,114	,111	,178*	,087	,093	-,097	,044	-,254**	-,226**	-,140	,121	,073	,144	-,107	-,374**	-,148	1	-,030	,209*	-,101	-,246**	,168	,214*	,021	,044	,243**	,197*	,182*	,162	-,018	-,073	-,044
Знач. (двухсторонняя)	,195	,208	,043	,323	,293	,273	,617	,004	,010	,113	,170	,409	,102	,224	,000	,092		,733	,017	,254	,005	,056	,014	,844	,619	,005	,025	,038	,066	,838	,412	,621
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
542 Корреляция Пирсона	-,294**	-,048	-,053	-,058	-,089	,121	-,031	-,059	-,159	,000	-,232**	-,028	-,316**	,098	,206*	-,075	-,030	1	-,281**	,064	,149	-,061	-,249**	,127	-,186*	-,265**	-,086	,107	,004	-,033	,156	-,109
Знач. (двухсторонняя)	,001	,591	,547	,513	,312	,171	,728	,505	,070	,996	,008	,752	,000	,265	,019	,398	,733		,001	,471	,090	,487	,004	,233	,034	,002	,329	,225	,963	,708	,076	,218
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
545 Корреляция Пирсона	,309**	,233**	,283**	,205*	,301**	-,259**	,121	-,423**	,290**	-,444**	,317**	,229**	,180*	-,301**	-,452**	,179*	,209*	-,281**	1	-,091	-,274**	,232**	,592**	,030	,435**	,517**	,135	,113	,126	,302**	-,217*	,119
Знач. (двухсторонняя)	,000	,008	,001	,019	,001	,003	,169	,000	,001	,000	,000	,009	,040	,001	,000	,042	,017	,001		,302	,002	,008	,000	,776	,000	,000	,127	,202	,153	,000	,013	,177
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
548 Корреляция Пирсона	-,113	-,090	,025	-,002	-,009	-,072	,004	,211*	,071	,108	-,019	-,065	-,180*	,205*	,216*	-,038	-,101	,064	-,091	1	,112	-,177*	-,135	-,224*	-,150	-,258**	,017	-,108	-,132	-,155	,141	,060
Знач. (двухсторонняя)	,201	,310	,774	,982	,921	,416	,961	,016	,419	,220	,833	,461	,040	,019	,013	,667	,254	,471	,302		,204	,044	,126	,034	,088	,003	,849	,223	,135	,079	,111	,497
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
557 Корреляция Пирсона	-,457**	-,103	-,078	-,077	-,037	,329**	-,035	,190*	,083	,171	-,231**	-,019	-,071	,211*	,284**	-,102	-,246**	,149	-,274**	,112	1	-,318**	-,411**	,038	-,174*	-,291**	-,119	-,148	-,015	-,176*	,238**	,119
Знач. (двухсторонняя)	,000	,243	,376	,386	,674	,000	,697	,030	,349	,052	,008	,833	,420	,016	,001	,248	,005	,090	,002	,204		,000	,000	,720	,048	,001	,176	,093	,867	,046	,006	,179
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130

Продолжение приложения А

	505	510	511	513	514	515	516	518	523	524	526	527	534	535	539	540	541	542	545	548	557	567	heavy	age	liquid	lpu	512	529	530_532	mms	nans	real_us
567 Корреляция Пирсона	,392**	,151	,171	,055	,087	,031	,093	-,186*	,061	-,125	,301**	,064	,141	-,300**	-,267**	,053	,168	-,061	,232**	-,177*	-,318**	1	,305**	,088	,153	,345**	-,037	-,006	,211*	,167	-,222*	,013
Знач. (двухсторонняя)	,000	,086	,052	,533	,326	,729	,293	,034	,490	,158	,001	,473	,109	,001	,002	,553	,056	,487	,008	,044	,000		,000	,409	,083	,000	,672	,944	,016	,058	,011	,881
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
Heavy Корреляция Пирсона	,442**	,175*	,192*	,167	,192*	-,305**	,192*	-,353**	-,235**	-,291**	-,350**	,121	,245**	-,301**	-,465**	-,255**	-,214*	-,249**	,592**	-,135	-,411**	,305**	1	,074	,584**	,619**	,153	,123	,148	,286**	-,444**	,128
Знач. (двухсторонняя)	,000	,046	,029	,057	,028	,000	,029	,000	,007	,001	,000	,169	,005	,001	,000	,003	,014	,004	,000	,126	,000	,000		,490	,000	,000	,083	,162	,093	,001	,000	,146
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
Age Корреляция Пирсона	-,218*	-,010	-,091	,172	,161	,094	-,018	-,171	,063	,000	-,127	,070	,209*	-,107	-,032	-,057	,021	,127	,030	-,224*	,038	,088	,074	1	-,027	,015	,038	-,065	,042	-,109	-,065	,187
Знач. (двухсторонняя)	,039	,925	,393	,105	,129	,377	,866	,107	,553	,996	,235	,510	,048	,314	,766	,593	,844	,233	,776	,034	,720	,409	,490		,804	,889	,724	,545	,692	,304	,543	,078
N	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
Liquid Корреляция Пирсона	,373**	,129	,076	,285**	,226**	-,251**	,132	-,265**	-,228**	-,228**	-,218*	-,071	,104	-,316**	-,294**	,274**	,044	-,186*	,435**	-,150	-,174*	,153	,584**	-,027	1	,483**	,054	,028	-,038	,297**	-,351**	,093
Знач. (двухсторонняя)	,000	,142	,391	,001	,010	,004	,135	,002	,009	,009	,013	,420	,238	,000	,001	,002	,619	,034	,000	,088	,048	,083	,000	,804		,000	,544	,749	,668	,001	,000	,294
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
Lpu Корреляция Пирсона	,540**	,076	,070	,171	,192*	-,196*	,136	-,309**	-,277**	-,170	,268**	,038	,180*	-,376**	-,496**	,239**	,243**	-,265**	,517**	-,258**	-,291**	,345**	,619**	,015	,483**	1	,006	,038	,119	,318**	-,418**	,084
Знач. (двухсторонняя)	,000	,388	,431	,052	,029	,025	,123	,000	,001	,053	,002	,668	,040	,000	,000	,006	,005	,002	,000	,003	,001	,000	,000	,889	,000		,942	,667	,178	,000	,000	,344
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130

Продолжение приложения А

	505	510	511	513	514	515	516	518	523	524	526	527	534	535	539	540	541	542	545	548	557	567	heavy	age	liquid	lpu	512	529	530_532	mms	nans	real_us
512 Корреляция Пирсона	,090	,418**	,405**	,084	,172	-,084	-,053	-,003	-,088	-,031	,138	,176*	,119	-,107	-,022	-,062	,197*	-,086	,135	,017	-,119	-,037	,153	,038	,054	,006	1	,110	,021	-,151	,104	,085
Знач. (двухсторонняя)	,308	,000	,000	,342	,050	,342	,548	,974	,317	,728	,117	,045	,177	,225	,806	,484	,025	,329	,127	,849	,176	,672	,083	,724	,544	,942		,215	,810	,086	,240	,338
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
529 Корреляция Пирсона	,003	,063	,206*	,043	,064	-,145	-,129	-,131	-,177*	-,107	,023	,179*	-,025	-,077	-,024	,080	,182*	,107	,113	-,108	-,148	-,006	,123	-,065	,028	,038	,110	1	,255**	-,075	-,090	-,062
Знач. (двухсторонняя)	,973	,477	,019	,628	,470	,099	,144	,137	,044	,226	,793	,042	,780	,382	,785	,367	,038	,225	,202	,223	,093	,944	,162	,545	,749	,667	,215		,003	,394	,310	,484
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
530_532 Корреляция Пирсона	,020	,014	,069	-,020	,017	-,031	-,118	,013	-,059	,017	,089	,298**	-,083	-,130	-,171	,054	,162	,004	,126	-,132	-,015	,211*	,148	,042	-,038	,119	,021	,255**	1	-,045	-,045	-,041
Знач. (двухсторонняя)	,825	,870	,434	,824	,849	,723	,182	,885	,503	,849	,312	,001	,349	,141	,051	,543	,066	,963	,153	,135	,867	,016	,093	,692	,668	,178	,810	,003		,609	,612	,642
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
mms Корреляция Пирсона	,235**	-,022	-,053	,166	,062	-,105	,093	-,104	,248**	-,055	,248**	-,078	-,050	-,164	-,260**	,229**	-,018	-,033	,302**	-,155	-,176*	,167	,286**	-,109	,297**	,318**	-,151	-,075	-,045	1	-,302**	,061
Знач. (двухсторонняя)	,007	,807	,547	,059	,482	,234	,294	,238	,004	,532	,004	,375	,573	,063	,003	,009	,838	,708	,000	,079	,046	,058	,001	,304	,001	,000	,086	,394	,609		,000	,489
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130

Окончание приложения А

	505	510	511	513	514	515	516	518	523	524	526	527	534	535	539	540	541	542	545	548	557	567	heavy	age	liquid	lpu	512	529	530_532	mms	nans	real_us
Nans Корреляция Пирсона	-,323**	,078	,145	,023	,115	,245**	-,199*	,107	-,055	,050	-,103	-,063	,005	,213*	,276**	-,332**	-,073	,156	-,217*	,141	,238**	-,222*	-,444**	-,065	-,351**	-,418**	,104	-,090	-,045	-,302**	1	,105
Знач. (двухсторонняя)	,000	,378	,100	,796	,192	,005	,023	,228	,534	,574	,243	,479	,959	,015	,002	,000	,412	,076	,013	,111	,006	,011	,000	,543	,000	,000	,240	,310	,612	,000		,234
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130
real_us Корреляция Пирсона	,014	,019	,120	,140	,063	-,103	,119	,083	,074	,138	,082	,122	,109	,030	-,055	,213*	-,044	-,109	,119	,060	,119	,013	,128	,187	,093	,084	,085	-,062	-,041	,061	,105	1
Знач. (двухсторонняя)	,874	,830	,173	,112	,476	,241	,177	,348	,404	,118	,355	,167	,218	,739	,534	,015	,621	,218	,177	,497	,179	,881	,146	,078	,294	,344	,338	,484	,642	,489	,234	
N	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	130	90	130	130	130	130	130	130	130	130

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Базовая кафедра «Интеллектуальные системы управления»

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Якунин Ю.Ю.
подпись инициалы, фамилия
«01» 07 2019 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Непараметрический алгоритм исключения выбросов из выборки наблюдений переменных процесса

27.04.03 Системный анализ и управление
27.04.03.02 Системный анализ данных и технологий принятия решений

Научный руководитель	1.07.19 подпись, дата	доцент, к. т. н.	А. А. Корнеева инициалы, фамилия
Выпускник	1.07.19 подпись, дата		А. А. Молошаг инициалы, фамилия
Рецензент	1.07.19 подпись, дата	доцент, к. т. н.	Н. В. Кононова инициалы, фамилия
Нормоконтроллер	1.07.19 подпись, дата	ст. преподав.	Н. Б. Позолотина инициалы, фамилия

Красноярск 2019