

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий

Базовая кафедра интеллектуальных систем управления

УТВЕРЖДАЮ

Заведующий кафедрой

_____ Ю. Ю. Якунин

« 11 » июня 2018 г.

БАКАЛАВРСКАЯ РАБОТА

27.03.03 Системный анализ и управление

Повышение точности задачи идентификации по выборке наблюдений с пропусками

Руководитель	_____	канд. техн. наук, доцент	<u>А.А Корнеева</u>
	подпись, дата	должность, ученая степень	инициалы, фамилия
Выпускник	_____		<u>А.А Климова</u>
	подпись, дата		инициалы, фамилия

Красноярск 2018

РЕФЕРАТ

Выпускная квалификационная работа по теме «Повышение точности задачи идентификации по выборке наблюдений с пропусками» содержит 54 страницы текстового документа, 18 использованных источников, 12 рисунков и 7 таблиц.

КЛЮЧЕВЫЕ СЛОВА: НЕПАРАМЕТРИЧЕСКИЙ МЕТОД, DEDUCTOR STUDIO, СРЕДНЕАРИФМЕТИЧЕСКИЙ МЕТОД, ПРОПУСКИ, ЗАПОЛНЕНИЕ ПРОПУСКОВ.

Целью данной работы заключается в повышение точности решения задач идентификации по выборкам наблюдения с пропусками.

Поставленные задачи:

1. Изучить существующие подходы к обработке данных с пропусками;
2. Реализовать и исследовать некоторые известные алгоритмы восстановления пропусков в данных;
3. Оценить влияние данных алгоритмов на точность решения задачи идентификации.

Объектом данной работы является решение задачи идентификации по выборкам наблюдений с пропусками. Предметом работы являются алгоритмы обработки неполных данных.

В ходе работы приведено сравнение между непараметрическим методом восстановления пропусков и методом заполнения среднеарифметическим значением, представленный в программе Deductor Studio.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	4
1 Задача идентификации в условиях многомерной выборки, содержащая пропуски.....	6
1.1 Моделирование	6
1.2 Идентификация	9
1.2.1 Параметрические модели.....	12
1.2.2 Непараметрические модели	12
1.3 Анализ неполных данных	17
1.4 Методы заполнения пропусков	18
Выводы по первой главе.....	25
2 Восстановление пропусков в выборке наблюдений.....	26
2.1 Непараметрический метод восстановления пропусков	26
Выводы по второй главе.....	32
3 Вычислительные эксперименты.....	33
3.1 Результаты исследования метода непараметрического заполнения пропусков.....	33
3.2 Результаты исследований заполнения пропусков, с помощью метода представленного в Deductor studio.....	44
3.3 Сравнительный анализ исследованных алгоритмов	46
Выводы по третьей главе.....	50
ЗАКЛЮЧЕНИЕ.....	51
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	53

ВВЕДЕНИЕ

В разработке научных исследований или технических работ могут возникнуть ситуации, когда данные попросту отсутствуют, являются некомплектными или не совпадает дискретность «входных-выходных» данных. Отсутствие данных и «некомплектность» подразумевает под собой, что исходные данные пропущены по известным или неизвестным причинам.

Причинами появления пропусков могут служить множество факторов, такие как отсутствие возможности получения или обработки данных, искажение или утаивание нужной для работы информации, природные явления или же выход из строя технического оборудования.

Сталкиваясь с проблемой пропусков приходится часто при проведении исследований в разнообразных направлениях, они могут быть социологическими, научными, техническими и т.д. в каждом из них могут присутствовать пропуски, которые в дальнейшем становятся ощутимой проблемой для точности решения поставленной задачи.

На практике многие исследователи, столкнувшиеся с задачами, в которых присутствуют неполные данные, попросту усекают выборку и отбрасывают пропущенные значения, не задумываясь о искажении результатов и сильному различию статистических выводов, при начальном исследовании данных с пропусками и при их отсутствии. Для того чтобы не было потери информации и ошибок вычисления, создано множество методов (заполнение по среднему значению, Resampling метод, восстановление по регрессии и т.д), которые позволяют восстановить данные. Из-за изобилия большого количества методов, встает проблема выбора подходящего алгоритма восстановления пропусков.

В данной дипломной работе предоставлены механизмы формирования пропусков, проведен обзор на наиболее известных методах восстановления данных с пропусками, а так же предложен подход к повышению точности восстанавливаемых пропусков с помощью непараметрической оценки регрессии.

Целью данной работы является повышение точности решения задач идентификации по выборкам наблюдения с пропусками. Для этого необходимо решить следующие задачи:

1. Изучить существующие подходы к обработке данных с пропусками;
2. Реализовать и исследовать некоторые известные алгоритмы восстановления пропусков в данных;
3. Оценить влияние данных алгоритмов на точность решения задачи идентификации.

В данной работе используются такие методы, как: математическое моделирование, теория идентификации, анализ данных и математическая статистика.

Объектом данной работы является решение задачи идентификации по выборкам наблюдений с пропусками. Предметом работы являются алгоритмы обработки неполных данных.

1 Задача идентификации в условиях многомерной выборки, содержащая пропуски

1.1 Моделирование

Моделирование – универсальный научно обоснованный метод получения и использование знаний об окружающей среде. Моделирование в системном анализе занимает одну из самых важных ролей, оно помогает предоставить любую систему с помощью математических процессов. Построение моделей используется для изучения процессов, без потерь и поломок реального объекта.

Как в главе всех открытий лежит фантазия, так и в моделировании основой является теория подобия. Главная цель построение моделей – замена одного реального объекта на подобный. В книге [1] «модели» дается следующее определение «модель – изображение существенных сторон реальной системы (или конструируемой системы), в удобной форме отображающее информацию о системе». Из этого можно сказать, что главным плюсом использованием моделей является то, что исследователь может уделять свое внимание интересующим его процессам и свойствам системы, не думая о том, что может навредить реальному объекту, если введет во входных воздействиях не рабочие данные. В работе невозможно построить реальному объекту идентичную модель, поэтому при моделировании исследователи пытаются достичь максимальную точность исследуемого объекта.

На данный момент существуют изобилия классификаций видов моделирования. Рассмотрим самые распространенные из них.

Полнота модели главный из признаков классификации моделирования. Она определяет и показывает, как хорошо построенная модель описывает стороны реального объекта. Опираясь на выше сказанное в работе [2] модели подразделяют на три способа:

- полное;

- неполное;
- приближенное.

В полном моделировании модели должны быть одинаковы при движении объекта и в формах существования, то есть во времени и в пространстве.

В способе неполном моделировании модели не сохраняют или сохраняют, но частично ту идентичность, которая представлена в полном способе.

Приближенное моделирование можно сравнить с неполным подобием, когда незначительные факторы, которые протекают в изучаемом процессе, моделируются приближенно или совсем опускаются и не придаются значением в данном способе. Нужно заметить, что в приближенном моделировании допускаются погрешности, которые связаны с упрощением системы.

В работе [3] представлены виды моделирования, зависящие от характеристик используемых объектов и разделены на:

- детерминированные или стохастические;
- статические или динамические;
- дискретно-непрерывные, или дискретные или непрерывные.

Детерминированное моделирование предполагает, что в отображаемых процессах нет случайных воздействий. Стохастическое – учитывает все вероятные процессы и события, происходящие в реальном объекте.

Статическое моделирование описывает объект и его состояния в заданный момент времени, а динамическое – описывает развитие системы и ее изменения во времени.

Дискретное моделирование действует тогда, когда поведение системы изменяется в заданный момент времени, а непрерывные модели используются, в тех случаях, когда система во времени изменяется непрерывно.

Классифицируется так же моделирование в зависимости от форм реализации. Клюкина Е.А.[4] предлагает в своей работе следующие классификации:

- реальное моделирование, исследует характеристики на целом объекте или его части.
- мысленное моделирование, применяется в том случае, когда модели не реализуются или отсутствуют условия для их создания;

Мысленное моделирование реализуется в виде наглядной, математической и символической модели.

Символическое моделирование – искусственное создание логического объекта, который может выразить основные свойства рассматриваемого объекта.

При наглядном моделировании исследователем создаются модели, которые показывают процессы, которые протекают в рассматриваемом объекте. Примерами могут послужить плакаты, диаграммы.

В математическом моделировании объект описывается математическим языком. Математическая модель – описание исследуемого процесса с помощью математического языка: алгебраических уравнений, неравенств, дифференциальных, интегральных уравнений и т.д. Построение моделей чаще всего опирается на данные полученные во время наблюдений. Опираясь на [5] можно рассмотреть формирование математических моделей двумя способами:

В первом способе происходит расчленение системы на подсистемы, свойства которых могут быть описаны из известных законов природы и знаний, которые были получены ранее.

Другой способ использует данные полученные с помощью экспериментов. Регистрируются «входные-выходные» данные и формирование модели происходит в соответствии обработки полученных данных. Данный способ называется идентификацией.

1.2 Идентификация

Идентификация – определение по входу и выходу системы из определенного класса систем, которой испытываемая система эквивалентна. [1]

Из выше сказанного, постановку задачу идентификации можно сформулировать следующим образом: по результатам наблюдений входных и выходных данных строится оптимальная модель для данного исследуемого объекта. При этом описывающий объект должен находиться в реальной обстановке, где на него действуют случайные возмущения и помехи. Так же для исследования объекта должна присутствовать априорная информация о случайных возмущениях, ограничениях и требования, которые помогут достичь наилучшего результата.

На рисунке 1 представлена общая схема процесса идентификации:

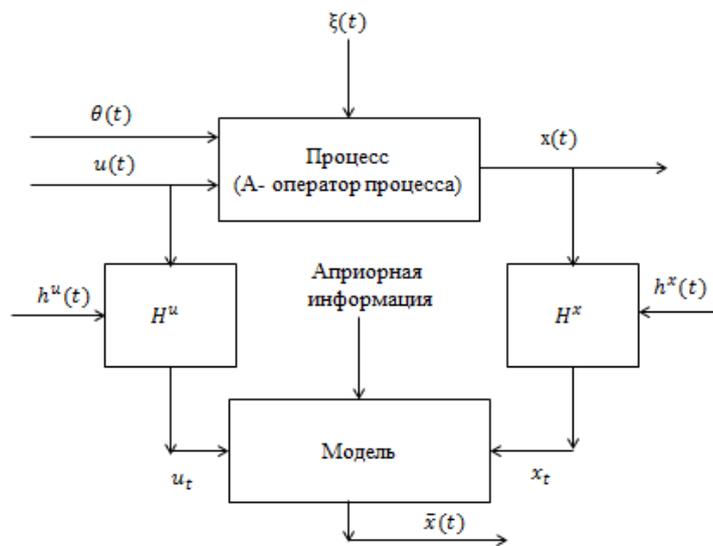


Рисунок 1 - Общая схема задачи идентификации

Введем следующие обозначения: A – оператор процесса, H – каналы измерения переменных процесса, $u(t)$ – контролируемый вектор входных переменных, $x(t)$ – вектор выходных переменных, $h^x(t)$ и $h^u(t)$ – помехи полученные случайно при измерении переменных, $\theta(t)$ – переменные, не поддающиеся контролю, $\xi(t)$ – случайные помехи, действующие на сам

процесс, $u^h(t)$ и $x^h(t)$ – полученные наблюдения в дискретные моменты времени Δt , $\bar{x}(t)$ – выход полученной модели. Для дальнейшего использования информации и простоты записи будем обозначать входные и выходные данные $u^h(t)$ и $x^h(t)$ через $\{u_i, x_i, i = \overline{1, s}\}$, где s – объем выборки заданный исследователем. Соответствие исследуемому процессу математическое обозначение представляется в виде

$$x(t) = A [u(t), \theta(t), \xi(t), t], \quad (1)$$

где $u(t)$ – входные данные, поступающие на объект;

$\theta(t)$ - переменные, не поддающиеся контролю;

$\xi(t)$ – случайные помехи, действующие на объект.

Модель данного процесса можно представить в виде

$$\bar{x}(t) = \bar{A} [u(t + 1), t], \quad (2)$$

где \bar{A} – является оператором построенной модели;

$u(t + 1)$ – входное наблюдение через промежуток времени.

Все случайные факторы, которые действуют на исследуемый объект, имеют нулевое математическое ожидание и ограничение по дисперсии. Стоит заметить, что на блок «Модель» так же действует априорная информация об исследуемом объекте. Для повышения точности задачи идентификации нужно брать во внимание ее полноту.

Основывается априорная информация на различных процессах или результатах научных наблюдений. Априорная информация тесно связана со временем, на заданный промежуток времени может быть достоверной одна априорная информация, а на другой совершенно другая. Так что достоверность априорной информации может утрачиваться со временем. Исследователь может работать только с не полной информацией об объекте.

Существуют следующие уровни априорной информации:

- системы с известной информацией. В данном уровне известна параметрическая модель исследуемого объекта и некоторая область входных допустимых значений, а помехи – отсутствуют;
- уровень параметрической неопределенности [6], у исследуемой параметрической модели известны параметры, которые необходимо оценить. Присутствуют некоторые случайные помехи, которые действуют на объект;
- уровень непараметрической неопределенности, когда у исследуемого объекта неизвестны ни законы распределения помех, ни структура самой модели, но известны качественные характеристики объекта.

Хороший результат идентификации системы зависит от двух факторов: объема полученной измерительной информации и объемом информации о структуре объекта. Делая выводы из уровня априорной информации, ставятся задачи идентификации для решения их в «узком» и «широком» смысле.

Из двух приведенных задач в моделировании дискретно-непрерывных процессов наиболее развита идентификация в «узком» смысле. Смысл теории состоит в том, что на начальном этапе, основываясь на заданной априорной информации, определяется класс оператора A_s , например

$$\bar{x}_s(t) = A_s(u(t), \bar{x}_s, \bar{u}_s), \quad (3)$$

где A_s – параметрическая структура модели;

$$\bar{x}_s = (x_1, x_2, \dots, x_s), \bar{u}_s = (u_1, u_2, \dots, u_s), \quad \text{– временные вектора.}$$

На следующем этапе многое зависит от того, как был определен оператор (3). Находится оценка параметров x_s , которые основываются по заданной выборке $\{u_i, x_i, i = \overline{1, s}\}$.

В идентификации «широком» смысле отсутствует параметрический класс выбор оператора (3). В случае, когда не хватает априорной информации,

необходимо решить ряд задач. Такие как, нахождение структуры модели, оценивание линейности (нелинейности) исследуемого объекта, его однозначность или не однозначность[7]. Примером методов решения задач идентификации в «широком» смысле является оценка функции регрессии.

1.2.1 Параметрические модели

Параметрической идентификацией является определение параметров модели по ее заданной структуре. [8]. Настраивание модели происходит по полученным с помощью наблюдения входных воздействий и выходных величинах, которые обеспечивают экстремум исследуемого критерия. Так же при параметрической идентификации считается, что параметры структуры исследуемого объекта являются известными.

Настраивание параметров описывается во множестве работ [1, 9, 10] в которых для получения правдоподобных результатов решения задач идентификации играет главную роль выбор математического объекта

Примером может послужить метод стохастической аппроксимации [11] и метод наименьших квадратов, который минимизирует для дискретных значений сумму квадратичной невязки.

На практике качество полученной модели зависит от того, как хорошо была подобрана параметрическая структура.

1.2.2 Непараметрические модели

На практике редко данные, с которыми приходится работать могут содержать большое количество априорной информации, которая так необходима для структуры исследуемого объекта и зачастую не возможно определить у исследуемого объекта требуемую для работы структуру. В условиях, когда происходит нехватка априорной информации разумно использовать методы непараметрической идентификации.

Идентификация в «широком» смысле предполагает оценку класс операторов на основе выборки $\{u_i, x_i, i = \overline{1, s}\}$, так как отсутствует этап выбора параметрического класса оператора.

В работах [6, 11] непараметрическая оценка кривой регрессии для моделирования в непараметрической неопределенности рассчитывается по формуле функции регрессии Надарая-Ватсона

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

где s - объем выборки;

x_i - выход i - ого наблюдения;

$\Phi\left(\frac{1}{c_s} u^j - u_i^j\right)$ - ядерная колокообразная функция;

u^j - входное воздействие наблюдение;

u_i^j - входное воздействие i - ого наблюдения;

c_s - коэффициент размытости ядра.

Ядерная функция и коэффициент размытости, опираясь на работу [11] должны удовлетворять представленным ниже условиям сходимости

$$\begin{aligned} c_s > 0; & \quad \Phi\left(\frac{u^j - u_i^j}{c_s}\right) \geq 0; \\ \lim_{s \rightarrow \infty} c_s = 0; & \quad c_s^{-1} \int_{\mathcal{C}(u)} \Phi\left(\frac{u^i - u_j^i}{c_s}\right) du^j = 1; \\ \lim_{s \rightarrow \infty} s c_s^m = \infty; & \quad \lim_{s \rightarrow \infty} \Phi\left(\frac{u^i - u_j^i}{c_s}\right) = \delta(u^i - u_j^i), \end{aligned} \quad (5)$$

где $\delta(u^i - u_j^i)$ - дельта – функция Дикара.

В качестве колокообразной функции используют одно из трех представленных ядер, оно может быть трех видов, которые рассмотрены будут далее.

Треугольное ядро представлено в формуле (6)

$$\Phi\left(\frac{u_i - u_j^i}{c_s}\right) = \begin{cases} 1 - |c_s^{-1}(u_i - u_j^i)|, & |c_s^{-1}(u_i - u_j^i)| \leq 1; \\ 0, & |c_s^{-1}(u_i - u_j^i)| > 1. \end{cases} \quad (6)$$

где u^j - входное воздействие наблюдение;

u_i^i - входное воздействие i - ого наблюдения;

c_s - коэффициент размытости ядра.

На рисунке 2 показана графическая интерпретация треугольного ядра.

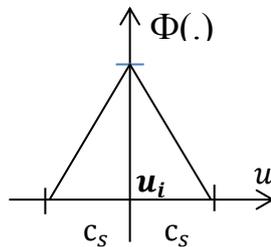


Рисунок 2 – Треугольная ядерная функция

Параболическое представлено в формуле (7)

$$\Phi\left(\frac{u_i - u_j^i}{c_s}\right) = \begin{cases} 0,75(1 - c_s^{-1}(u_i - u_j^i))^2, & |c_s^{-1}(u_i - u_j^i)| \leq 1; \\ 0, & |c_s^{-1}(u_i - u_j^i)| > 1. \end{cases} \quad (7)$$

Рисунок 3 представляет графическое представление вышеописанного ядра.

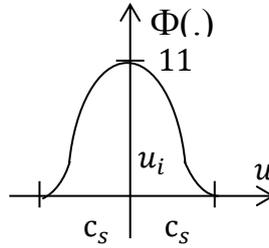


Рисунок 3 – Параболическое ядро

Прямоугольное ядро

$$\Phi\left(\frac{u_i - u_j^i}{c_s}\right) = \begin{cases} 0,5 & , \left|c_s^{-1}(u_i - u_j^i)\right| \leq 1; \\ 0, & \left|c_s^{-1}(u_i - u_j^i)\right| > 1. \end{cases} \quad (8)$$

На рисунке 4 представлена колокообразная прямоугольная ядерная функция.

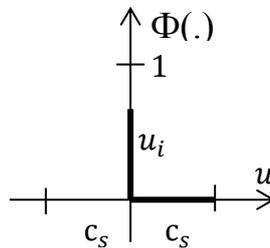


Рисунок 4 – Прямоугольная ядерная функция

Большую значимость на качество функции оказывает c_s – коэффициент размытости ядра, постоянная величина, от которой зависит «размытость» дельта – функции Дикара в окрестностях каждой точки.

Коэффициент размытости c_s находится с помощью задачи минимизации квадратичного показателя, находящий соответствие выхода модели к выходу объекта, по формуле скользящего экзамена, приведенной ниже.

$$R(c_s) = \sum_{k=1}^s (x_k - x_s(u_k, c_s))^2 = \min, k \neq i, \quad (9)$$

где x_k - выход объекта;

x_s - выход модели.

т.е. по индексу i (фигурирующий в формуле 4) исключаются k -е наблюдение переменной.

Каждой компоненте вектора u соответствует компонент вектора c_s . Следуя из этого, формулу (4) можно представить в виде

$$x_s(u) = \sum_{i=1}^s x_i \varphi(u, u_i), \quad (10)$$

$$\varphi(u, u_i) = \frac{\sum_{i=1}^s x_i \prod_{j=1}^m \Phi\left(\frac{u^j - u_i^j}{c_{sj}}\right)}{\sum_{i=1}^s \prod_{j=1}^m \Phi\left(\frac{u^j - u_i^j}{c_{sj}}\right)}, \quad (11)$$

где x_i - выход объекта;

c_{sj} - оптимальный коэффициент размытости, найденный по скользящему экзамену по формуле (9);

u^j - входное воздействие наблюдение;

u_i^j - входное воздействие i – ого наблюдения.

1.3 Анализ неполных данных

В реальном мире приходится работать с большим количеством информации. И эту информацию не всегда можно собрать полностью, может по вине исследователя, который невнимательно собирал нужные данные или же отсутствие возможности получения этой информации. Таблицы с данными могут содержать от малого количества пропущенных данных до существенного. В результате на вход подаются данные с пустотами. Большинство методов анализа данных не могут работать с «некомплектными» таблицами. Возникшая проблема с обработкой неполных данных повлекла за собой множество решений. Самым простым решением было удалить наблюдения содержащие пропуски, но такой радикальный метод может привести к безвозвратному удалению данных и в дальнейшем приведет к неточным выводам и искажению результатов.

Поэтому главной целью стало не только заполнить пропущенные значения, но и повысить точность и максимально приблизиться к истинным.

Для того, чтоб подобрать необходимый метод, чтобы восстановить пропущенные значения, нужно понимать механизм их формирования. В работе [12] классифицируются три механизма формирования пропусков:

- MAR (Missing At Random) – механизм пропуска, когда данные пропущены случайно[13];
- MCAR (Missing Completely At Random) — механизм получения условия пропуска строже, чем MAR. Тип «полностью независимый пропуск» рассматривается как абсолютно случайный;
- MNAR (Missing Not At Random) – самый неудобный для исследования пропуск. Неслучайные пропуски сложнее для восстановления и анализа.

На практике будут рассмотрены пропуски (MAR и MCAR), поскольку с ними после сбора информации есть возможность эффективно бороться, с помощью восстановления пропущенных данных методами которые мы рассмотрим далее.

1.4 Методы заполнения пропусков

На данный момент существуют множество методов заполняющие пропуски, но у каждого метода есть свои недостатки и достоинства. Но общими недостатками являются искажение результирующих данных и смещение реального результата от полученного. Восстановления пропусков используется не только для заполнения пропущенных значений, но и для сохранения уже имеющейся информации.

Модели содержащие восстановленные данные, скорее всего будут менее точными, чем те модели, которые построены на полных наблюдениях. Точность зависит от выбранного метода и от качества восстанавливаемых значений. Выбор метода, который послужит для заполнения пропусков, должен зависеть от метода анализа данных.

Как говорилось ранее, знание механизма, который порождает пропуск, сильно влияет на выбор метода интерпретации результатов. Классифицировал наиболее распространенные методы статист Литтл Р. Д. [12] как показано на рисунке 2.

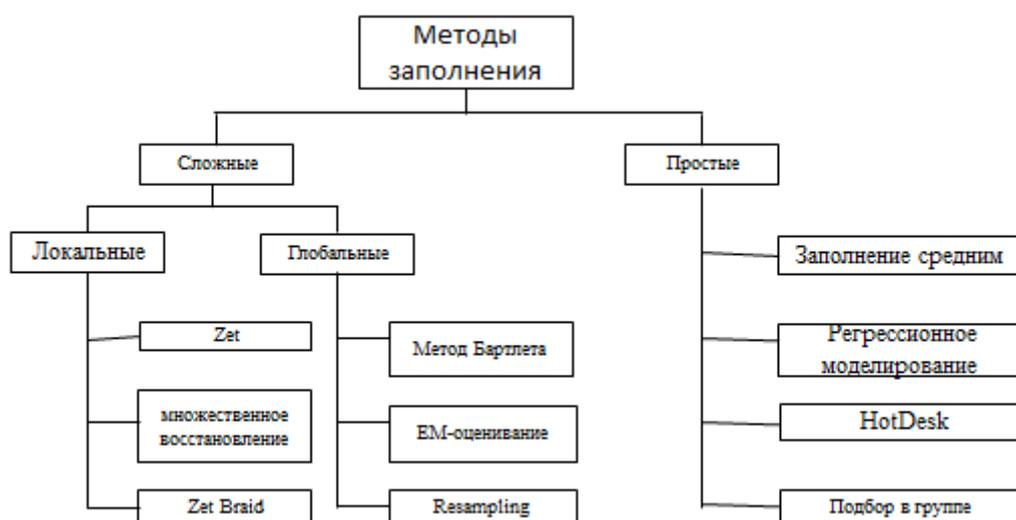


Рисунок 2 – Классификация методов заполнения пропусков

Методы заполнения делятся на два типа: сложные и простые. Простые не итеративные методы основываются на простейших арифметических операциях, а сложные итеративные методы основаны на оценки точности значения, которое подставляется вместо пропуска. Так же сложные методы подразделяются на локальные и глобальные.

Локальные методы – методы, когда происходит оценивание любого из пропущенных значений, в которых участвуют наблюдения без пропусков, находящиеся на расстоянии рассматриваемого наблюдения.

Глобальные методы – методы, когда для оценивания любого из пропущенных значений, участвуют вся данная выборка.

Рассмотрим вышеперечисленные методы по отдельности и более подробно.

Методы восстановления пропусков описаны подробно в работах [14, 17].

Заполнение пропуска средним значением – является одним из самых простых методов восстановления данных. Суть метода заключается в том, чтобы заполнить пропущенные данные среднеарифметическим значением, вычисленным из доступной выборки.

Данный метод очень прост в реализации, но имеет множество недостатков, такие как:

- а) Происходит искажение результатов, что является недопустимым для исследователя, который делает выводы, опираясь на графики или гистограммы;
- б) Из-за снижения объема выборки, происходит искажение дисперсии и стандартное отклонение

Метод Hot Desk. Данный метод описан подробно в работе [16], чаще называется этот метод «ближайшего соседа». Смысл метода заключается в том, что на место пропущенного значения ставится тот объект, который наиболее близко расположен к пропуску с полной информацией. Подбор объекта с полной информацией может осуществляться как по всей выборке, так и из небольших подгрупп связанных по какому – либо признаку.

При подборе ближайшего наблюдения от рассматриваемого объекта с пропуском находится расстояние d до каждого объекта с полной информацией. Значение признаков должно быть известно как для наблюдений с пропусками, так и для объектов с полными наблюдениями. Исследователь, выбирает сам ту меру расстояния, ссылаясь на тип используемых данных и характер связи между переменными. Существует множество мер расстояние, но самое распространенное - используют Эвклидово расстояние

$$d_{u_1 u_2} = \sqrt{\sum_{i=1}^n (x_{u_1 i} - x_{u_2 i})^2}, \quad (12)$$

где i – порядковый номер признака;

u_1 и u_2 – рассматриваемые точки в n -мерном пространстве;

$x_{u_1 i}$ и $x_{u_2 i}$ - координаты точек u_1 и u_2 по признаку i .

Недостатком этого метода является, что после восстановления появляется зависимость между значениями, которые восстановили и заниженной оценкой дисперсии. Снижение уровня зависимости происходит в случае, когда

возрастает число подгрупп, что подразумевает под собой большой объем выборки.

Несмотря на недостатки метод Hot Desk пользуется спросом в статистических организациях.

Метод подбора внутри групп. В данном методе происходит деление объектов на группы по определенному признаку для каждого пропуска. Для заполнения выбирается различные значения из полных данных. Данный метод очень распространен и имеет множество аналогов, использующие сложные схемы для отбора на подгруппы.

Недостатком является искажение распределения обрабатываемой выборки.

В сложных глобальных алгоритмах рассматриваются такие методы как: EM – алгоритм, метод Бартлетта и resampling. Каждый из трех методов мы рассмотрим далее по отдельности.

Метод Бартлетта для заполнения пропусков. В своей работе Злоба Е. и Яцкив [17] считают, что этот метод состоит из двух этапов:

- происходит подстановка вместо пропусков начальных значений;
- проведение ковариационного анализа неизвестной переменной.

На последнем этапе используется индикатор полноты наблюдений, он нужен для того, чтобы показать, где в матрице есть пропуск. Если индикатор равен 0, то значение не пропущено, если равен 1, то значение является пропущенным.

Метод обладает преимуществами:

- если структура пропусков является вырожденной, то метод «предупреждает» исследователя, что ответ, возможно, будет некорректным;
- метод является неинтеративным;
- метод дает правильные оценки и приемлемые стандартные ошибки.

У этого метода множество плюсов, но его часто невозможно реализовать непосредственно, потому что многие программы не могут вести обработку данных при многих сопутствующих переменных.

Метод максимального правдоподобия EM - алгоритм. Этот метод проводит двухступенчатую итерацию, состоящий из двух шагов (E – шаг и M – шаг), где E обозначается ожидание, а M – является максимизацией. Метод предназначен не только для восстановления пропусков, но и для оценивания средних значений, корреляционных и ковариационных матриц.

Как сказано ранее, алгоритм разбит на два этапа.

На первом этапе (этап E) для каждого пропущенного значения рассчитывается предполагаемое значение целевой переменной, опираясь на полные наблюдения. После того, как пропущенные значения были восстановлены, происходит оценка основных статических параметров: коэффициент взаимной корреляции и ковариации, показатель разброса.

На следующем этапе M, сравниваются ожидаемые значения с теми, которые были восстановлены, а также происходит соответствие структуры заполненных данных структуре данных полных наблюдений.

Недостатками данного метода является то, что если в представленной работе имеется множество пропусков, то работа требует больших вычислительных ресурсов.

Resampling метод. [17] Главное преимущество этого метода – это то, что он не нуждается в априорной информации о знании вероятностного закона распределения исходных данных, а значит это, что он способен работать с любыми представленными данными. Отличие Resampling метода от других представленных, заключается в том, что он многократно обрабатывает различные части одних и тех же данных, тем самым дает более точный результат.

Работа данного метода заключается в том, что он заменяет пропущенные значения случайными строками из рассматриваемой матрицы. После чего, отсутствующие значения предсказывает по регрессионному уравнению.

Построение модели повторяется заданное количество раз. Полученные значения регрессионных коэффициентов усредняются и на выход получают значения для заполнения пропуска.

Недостатками метода является то, что отсутствует возможность оптимизации метода.

Переходя к локальным сложным методам, начнем с метода **множественного восстановления**. Этот метод разработал Дональд Рубиним. В настоящее время метод становится популярным, но реализация его происходит в основном в коммерческом программном обеспечении.

Метод импутирования заключается в том, что пропущенному объекту приписывается несколько возможных значений. Существует разброс между каждым выбранным значением, это доказывает о неопределенности выбранной модели.

Данные каждого набора восстановленных значений хранится в отдельном массиве, который в дальнейшем анализируется как восстановленная матрица не имеющая пропусков.

Недостаток данного метода является то, что вся второстепенная информация хранится и является избыточной.

Zet и ZetBraid метод [18]. Сутью алгоритма Zet является то, что при подборе пропуска используют только часть наблюдений, которая называется компонентной матрицей. Эта матрица состоит из компонентных строк и столбцов. Некоторая величина, которая представляет компетентность объекта, равна обратно пропорциональному декартовому расстоянию до строки с неполным наблюдением.

На следующем этапе строится функциональная зависимость прогнозируемого значения от значения, взятого из компонентной матрицы. На данной зависимости происходит импутирование.

Метод ZetBraid является доработкой метода Zet. Главное их отличие и достоинством модернизированного метода является возможность изменение размерности компонентной матрицы.

Недостаток метода *Zet* является то, что компетентная матрица является фиксированной и то, что в эту матрицу могут попасть неинформативные данные создающие помехи. У метода *ZetBraid* есть весомый недостаток – это что только при корреляционно – регрессивном анализе возможен расчет статистических оценок неизвестного значения.

Выводы по первой главе

В результате написания первой главы можно сделать следующие выводы:

- получены необходимые для дальнейшей работы теоретические знания, так же приведены основные понятия;
- Представлен главный этап построение модели – идентификация, она была описана в «узком» и «широком» смысле;
- Рассмотрены задачи идентификации при параметрическом и непараметрическом моделировании;
- Описана задача неполных данных и рассмотрены самые распространенные методы для повышения точности восстанавливаемых пропусков.

2 Восстановление пропусков в выборке наблюдений

2.1 Непараметрический метод восстановления пропусков

Непараметрические модели рассматривались ранее (см. 1.2.2 Непараметрические модели). На данном этапе работы будут рассмотрены этапы восстановления пропусков в выборке наблюдений.

Непараметрическое восстановление отличается от других тем, что отсутствует априорная информация и больше требуется качественная информация об уровне объекта (статический или динамический объект, линейный или нелинейный и т.д.). Работа основывается на непараметрической оценке функции регрессии по наблюдениям.

На рисунке 1 представлена общая схема исследуемого процесса в данной работе.

Контроль входной переменной $u(t)$ осуществляется через незначительный промежуток времени Δt , а для измерения выходной переменной $x(t)$ требуется немного больше времени - ΔT . При этом $\Delta t \leq \Delta T$. В результате входные – выходные данные контролируются с разной дискретностью, что приводит к проблеме появлению пропусков.

Матрица наблюдений «входных–выходных» переменных процесса в данной работе имеет вид (таблица 1):

Таблица 1 – Матрица наблюдений «входных–выходных» переменных процесса с пропусками

<i>i</i>	<i>u</i>				<i>x</i>
1	<i>u</i>₁₁	<i>u</i>₂₁	...	<i>u</i>_{m1}	<i>x</i>₁
2	<i>u</i> ₁₂	<i>u</i> ₂₂	...	<i>u</i> _{m2}	—
3	<i>u</i> ₁₃	<i>u</i> ₂₃	...	<i>u</i> _{m3}	—
4	<i>u</i>₁₄	<i>u</i>₂₄	...	<i>u</i>_{m4}	<i>x</i>₄
5	<i>u</i> ₁₅	<i>u</i> ₂₅	...	<i>u</i> _{m5}	—
6	<i>u</i> ₁₆	<i>u</i> ₂₆	...	<i>u</i> _{m6}	—
7	<i>u</i>₁₇	<i>u</i>₂₇	...	<i>u</i>_{m7}	<i>x</i>₇
8	<i>u</i> ₁₈	<i>u</i> ₂₈	...	<i>u</i> _{m8}	—
9	<i>u</i> ₁₉	<i>u</i> ₂₉	...	<i>u</i> _{m9}	—
...
<i>s</i>	<i>u</i>_{1s}	<i>u</i>_{2s}	...	<i>u</i>_{ms}	<i>x</i>_s

В таблице 1 представлены столбцы описывающие переменные процесса, а строки – наблюдения. Визуально видно, что дискретность переменных разная, $x(t)$ в три раза больше, чем изменение входной переменной $u(t)$ ($\Delta T=3\Delta t$), s – заданный объем исходной выборки. В качестве приближения $x(t)$ используется математическое ожидание $x(u)=M\{x/u\}$, а непараметрическая оценка используется для его оценки.

В процессе работы пропуски сильно усложняют процесс моделирования, и происходит снижение точности решения задач идентификации. Для повышения точности и качества моделирования представляет интерес задача заполнения пропусков.

Математическая постановка задачи для данной работы можно обозначить как: даны наблюдения $\{u_i, x_i, i = \overline{1, s}\}$ случайных величин $u(t), x(t)$, которые распределены с неизвестной плотностью вероятности $p(x, u)$. Необходимо восстановить пропуски $x(u)=M\{x/u\}$ и заполнить их значениями повышенной

точностью, полученные непараметрической оценкой x_s по управляемой заданной функции Надарая-Ватноса. Она приведена в формуле (4).

Рассматриваемый процесс в данной работе является непрерывным, а его «входные-выходные» переменные контролируются.

Для проведения эксперимента будут применяться функции прямоугольного, треугольного ядра или параболического. В основе эксперимента будет выбрано одно ядро, которое лучше подходит для решения поставленной задачи. Коэффициент размытости ядра будет рассмотрен в пределах $[0;1]$ и минимизация производится с помощью метода скользящего экзамена приведенный в формуле (9).

Для оценки качества полученной модели $x_s(u)$ для каждой матрицы наблюдений производится нахождения ошибки моделирования по ниже представленной формуле

$$W = \sqrt{\frac{1}{s} \sum_{i=1}^s (x_i - x_s(u_i))^2 / \frac{1}{s-1} \sum_{i=1}^s (m - x_i)^2}, \quad (12)$$

где x_i - выход объекта;

$x_s(u_i)$ - выход модели по i – ому входному наблюдению;

s - объем выборки;

m – оценка математического ожидания выхода объекта, которая находится по формуле (13)

$$m = \frac{1}{s} \sum_{i=1}^s x_i, \quad (13)$$

где x_i - выход объекта;

s - объем выборки.

Алгоритм восстановления пропусков можно разделить на три этапа [21].

На первом этапе алгоритма производится настройка коэффициента размытости C_s (9).

Чтобы заполнить пропущенные значения производится обучение непараметрической модели, для этого производится разделение исходной выборки на две части: обучающую и экзаменационную. Для решения задачи прогнозирования будет изменена немного формула Надарая-Ватсона (4). Измененная формула будет иметь вид (14)

$$x'_s = \frac{\sum_{i=1}^{s'} x_i \prod_{j=1}^m \Phi\left(\frac{u^j - u_i^j}{c_{s'}}\right)}{\sum_{i=1}^{s'} \prod_{j=1}^m \Phi\left(\frac{u^j - u_i^j}{c_{s'}}\right)}, \quad (14)$$

где $u = (u_1, u_2, \dots, u_m)$, - m - мерный вектор входных данных;

x'_s - прогнозируемое наблюдение;

x_i – выходное наблюдение обучающей выборки;

$\Phi\left(\frac{1}{c_s} u^j - u_i^j\right)$ - ядерная колокообразная функция;

s' - объем обучающей выборки;

u^j - входное воздействие прогнозируемого наблюдения;

u_i^j - входное воздействие i – ого наблюдения обучающей выборки;

c_s - коэффициент размытости ядра.

В данном этапе участвуют только данные с полностью заполненными строками «входных-выходных» переменных процесса. В таблице 1 для наглядности полные строки выделены жирным шрифтом. Если представление всей выборки равно s , то объем выборки «полных» строк с наблюдениями обозначим s' ($3s' = s$).

После восстановления оценки по формуле (14) согласно критерию (3) происходит подбирание оптимального коэффициента размытости c_s по выборке s' .

На этапе II происходит заполнение пустых ячеек в матрице наблюдений с помощью оценки (14) и найденного коэффициента размытости c_s , который получен на предыдущем этапе. Для заполнения пропусков в столбце x подставляем в оценку (14) в выбранную ядерную функцию $\Phi\left(\frac{u^j - u_i^j}{c_s}\right)$ вместо u_i^j значения измеренных данных $u = (u_1, u_2, \dots, u_m)$ и вычисляем нужную оценку x_s , которая восполняет пропущенное значение. В итоге получаем заполненную матрицу, которая представлена в таблице 2.

Таблица 2 – Восстановленная матрица наблюдений «входных-выходных» переменных

i	u				x
1	u_{11}	u_{21}	...	u_{m1}	x_1
2	u_{12}	u_{22}	...	u_{m2}	x_{s2}
3	u_{13}	u_{23}	...	u_{m3}	x_{s3}
4	u_{14}	u_{24}	...	u_{m4}	x_4
5	u_{15}	u_{25}	...	u_{m5}	x_{s5}
6	u_{16}	u_{26}	...	u_{m6}	x_{s6}
7	u_{17}	u_{27}	...	u_{m7}	x_7
8	u_{18}	u_{28}	...	u_{m8}	x_{s8}
9	u_{19}	u_{29}	...	u_{m9}	x_{s9}
...
s	u_{1s}	u_{2s}	...	u_{ms}	x_s

Этап III. На заключительном этапе восстановления $u = (u_1, u_2, \dots, u_m)$ происходит построение непараметрической оценки по всем восстановленной

матрице наблюдений объема s (таблица 2). При этом настраивание коэффициента размытости происходит по всей выборки еще раз, также по критерию (9).

Выводы по второй главе

В представленной главе можно сделать вывод, что при непараметрическом методе восстановления пропусков «входных-выходных» переменных необходимо иметь выборку наблюдения вида $\{u_i, x_i, i = \overline{1, s}\}$ и обладать сведениями о качественном характере процесса.

Был изложен алгоритм непараметрического заполнения пропусков для повышения точности в матрице наблюдений с различной дискретностью «входных-выходных» данных. На основе приведённого алгоритма будет получено решение задачи идентификации по выборке наблюдений с пропусками для повышения точности полученных данных.

3 Вычислительные эксперименты

3.1 Результаты исследования метода непараметрического заполнения пропусков

Для проведения численного эксперимента рассмотрим многомерный стохастический объект, известный только в рамках вычислительного эксперимента

$$x(u) = 0.5 * \sin(u_1) + 0.5 * \cos(u_2) + 0.5 * (u_3)^2, \quad (15)$$

где $u_i \in [0,4], i = \overline{1,3}$.

На выход объекта действует помеха

$$\xi = x \zeta^k, \quad (16)$$

где ζ - случайная величина, распределенная по нормальному закону в интервале $[-1,1]$;

k - процент помехи, задающийся пользователем.

Для решения задачи заполнения пропусков в матрице наблюдений была разработана программа, написанная на языке программирования C# и основанная на составлении прогноза, опираясь на непараметрическую оценку Надарая-Ватсона (14).

Вся первоначальная выборка была разделена на две части: было взято 30% наблюдений из исходной матрицы, в которых не содержатся пропуски и соответственно 70% выборки с пропусками. Наблюдения будут собраны с помощью генератора случайных чисел по нормальному закону распределения в промежутке $[0;4]$. В оценке (14) тестирование будет проходить по трем типам ядерной функции: треугольной, параболической и прямоугольной. Настройка

коэффициента размытости лежит в пределах от $[0;1]$ и подбирается оптимальное значение коэффициента размытости C_s . по формуле (9).

Постановка эксперимента:

1. На первом этапе эксперимента строится модель по заданному объекту (16) с входными данными, сгенерированные по нормальному закону распределения в интервале $[0;4]$. Отметим, что в данном случае пропуски отсутствуют. Так же подбирается оптимальное значение коэффициента размытости C_s . по формуле (9) и находится ошибка моделирования (12).

2. Для создания пропусков в выходных данных каждое третье значение остается неизменным, а остальные удаляются. Тем самым только 30% известных наблюдений из исходной матрицы являются «полными». Построена модель для выборки с наблюдений с пропусками, подобран оптимальный коэффициент размытости C_s . по формуле (9) и найдена ошибка моделирования.

3. Для того чтобы заполнить пропущенные значения, созданные на предыдущем шаге, происходит восстановление данных с помощью оценки Надарая-Ватсона (14). Так же строится модель по восстановленной выборке наблюдений с подобранным оптимальным коэффициентом размытости C_s . по формуле (9) и так же находится ошибка моделирования.

Для того, чтобы достичь цели работы сравниваются ошибки моделирования для каждого из этапов и делается вывод о повышении точности. Исследуется зависимость ошибки моделирования от помехи, объема выборки и от вида ядерной функции.

Для дальнейшей работы введем обозначения: W_1 – ошибка моделирования по полной выборке наблюдений; W_2 – ошибка моделирования по выборке с пропусками; W_3 – ошибка моделирования по восстановленной выборке наблюдений.

Рассмотрим, как будет вести себя модель при разных значениях объема выборки, процента помехи и вида ядерной функции.

Таблица 3 – Зависимость ошибки моделирования от объема выборки при 0% помехи

Объем выборки s	Ядро	Значение ξ	$W1(\%)$	$W2(\%)$	$W3(\%)$
30	прямоугольное	0	22,3	33,9	23,4
	треугольное	0	25,35	39,35	32,4
	параболическое	0	20,1	32,4	22,1
50	прямоугольное	0	15,5	23,1	18,1
	треугольное	0	16,4	21,8	17
	параболическое	0	15,3	20,7	16,6
100	прямоугольное	0	14	16	15,2
	треугольное	0	12,2	14,6	13
	параболическое	0	11,8	14,1	13,2
500	прямоугольное	0	8,9	10,5	9,2
	треугольное	0	9,2	10,9	9,9
	параболическое	0	8,2	9,5	8,8
1000	прямоугольное	0	6,2	7,8	6,7
	треугольное	0	6,5	8,3	6,9
	параболическое	0	6,1	7,6	6,5

По таблице 3 видно, что лучший результат показывает параболическое ядро, а худший – треугольное. Но при большом объеме данных треугольное ядро проявляет самую маленькую ошибку. Видно, что ошибка моделирования уменьшается с повышением объема выборки. Так же стоит отметить, что ошибка моделирования по полной выборке наблюдений является самой маленькой, ошибка с пропусками является самой большой во всех экспериментах, а ошибка моделирования по восстановленной выборке больше $W1$, но меньше чем $W3$. Это означает, что точность наблюдений повышена.

Рассмотрим также зависимость ошибки моделирования от объема выборки при 5% помехи.

Таблица 4 - Зависимость ошибки моделирования от объема выборки при 5% помехи

Объем выборки s	Ядро	Значение ξ	$W1(\%)$	$W2(\%)$	$W3(\%)$
30	прямоугольное	5	26,6	40,5	31,5
	треугольное	5	28,6	44,2	34,9
	параболическое	5	25,8	42,5	33,2
50	прямоугольное	5	24,7	28,3	26,1
	треугольное	5	26,5	30,2	27,3
	параболическое	5	22,1	27,1	23,9
100	прямоугольное	5	18	21,1	18,8
	треугольное	5	17,6	21,3	19
	параболическое	5	17,2	20,5	18,1
500	прямоугольное	5	12,5	16,3	13,2
	треугольное	5	13,1	16,7	14,3
	параболическое	5	12,3	15,5	13,3
1000	прямоугольное	5	8,8	10,5	9,6
	треугольное	5	9,1	10,9	10,3
	параболическое	5	8,7	9,8	9,2

По таблице 4 можно сказать, что при увеличении процента помехи остались те же выводы, что и при меньшем проценте помехи (таблица 3). Так же во многих случаях лучше себя проявляет параболическое ядро, а хуже - ядро треугольное. Можно так же отметить, что при увеличении объема выборки при одинаковой помехе наблюдается уменьшение ошибки моделирования в каждом случае. Видно, что ошибка моделирования по полной выборке наблюдений является самой маленькой, ошибка с пропусками является самой большой во всех экспериментах, а ошибка моделирования по восстановленной выборке больше $W1$, но меньше чем $W3$. Это означает, что точность наблюдений повышена.

В таблице 5 будет рассмотрена зависимость ошибки моделирования от объема выборки при 10% помехи.

Таблица 5 - Зависимость ошибки моделирования от объема выборки при 10% помехи

Объем выборки s	Ядро	Значение ξ	$W1(\%)$	$W2(\%)$	$W3(\%)$
30	прямоугольное	10	36,6	45,2	37,2
	треугольное	10	37	46,1	37,5
	параболическое	10	35,6	44,8	36,3
50	прямоугольное	10	32,9	38,2	33,8
	треугольное	10	33,5	39	34,7
	параболическое	10	30,5	34,2	31,9
100	прямоугольное	10	28,8	33,4	29,9
	треугольное	10	29,5	34,8	30,4
	параболическое	10	27,9	32	29,2
500	прямоугольное	10	20,7	23,5	21,1
	треугольное	10	18,9	22	20,2
	параболическое	10	18,1	21,2	19,5
1000	прямоугольное	10	12,2	14,5	12,6
	треугольное	10	13,4	14,9	13
	параболическое	10	11,9	14	12,2

В таблице 5 видим, что лучше себя показывает все так же ядро параболическое, самые плохие результаты дает прямоугольное ядро. Увеличения объема выборки, снижает ошибку моделирования. Стоит отметить, что минимальный результат ошибки тогда, когда выборка работает без пропусков, а ошибка, когда пропуски уже заполнены, является срединным значением между полной выборкой и с пропусками. Это означает, что точность наблюдений повышена.

Рассмотрим, как ошибка моделирования зависит от коэффициента размытости C_s при разных помехах, используя параболическое ядро и ошибку моделирования при полной выборке наблюдений.

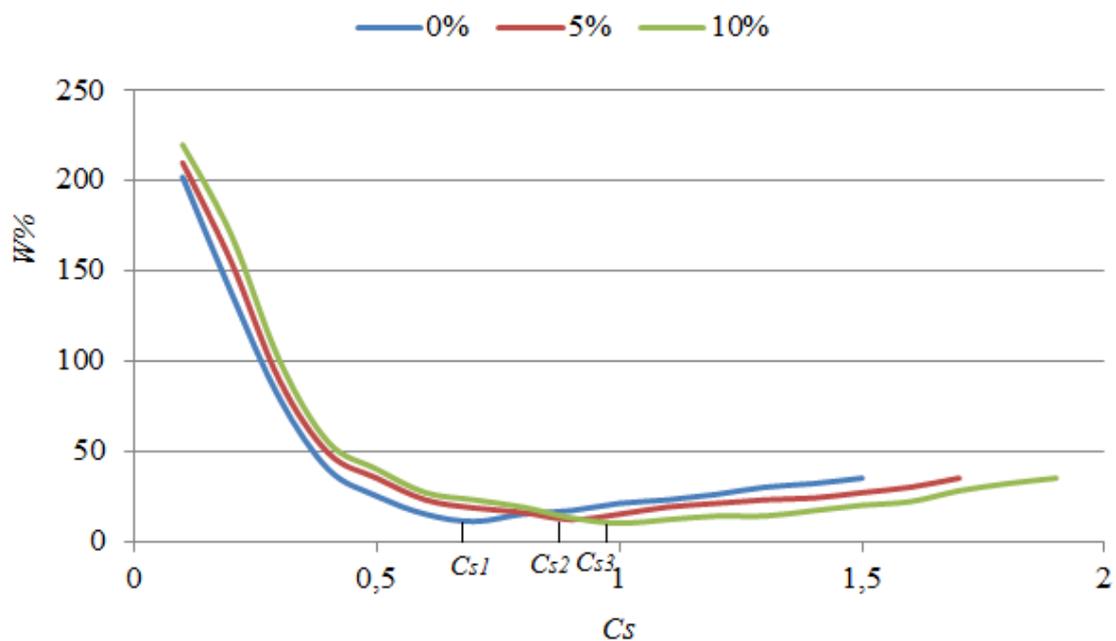


Рисунок 3 – Зависимость W от C_s , при $S = 100$

На рисунке 3 представлена зависимость ошибки моделирования от коэффициента размытости при $S = 100$. Так же представлены оптимальные коэффициенты размытости для каждой ошибки. Как видно по рисунку с повышением ошибки моделирования возрастает оптимальный коэффициент размытости.

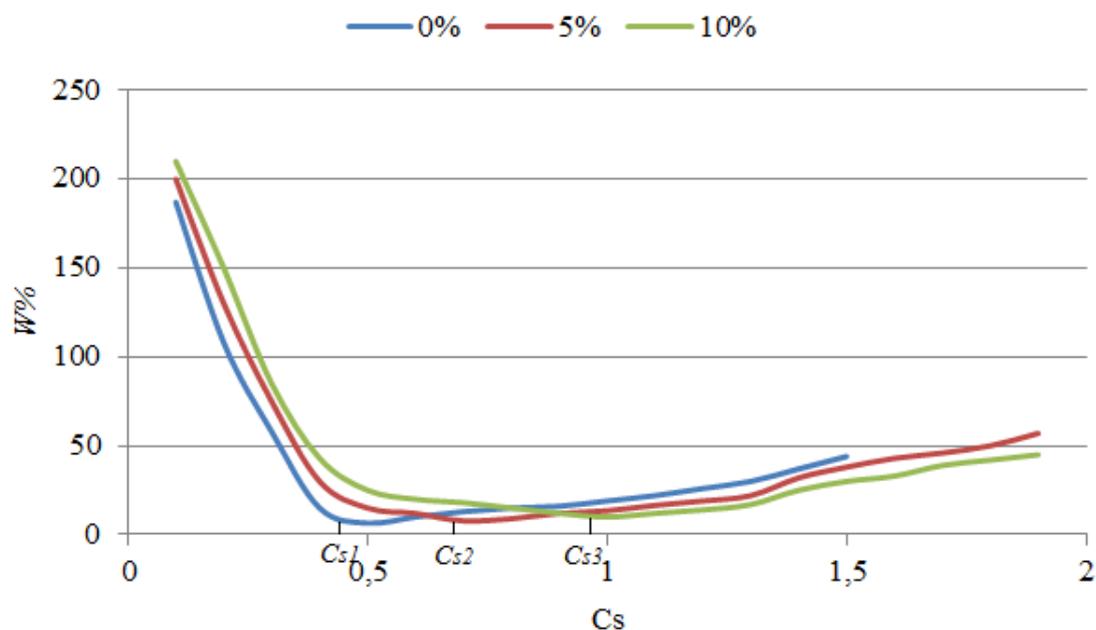


Рисунок 4 – Зависимость W от C_s , при $S = 500$

На рисунке 4 представлена зависимость ошибки моделирования от коэффициента размытости при $S = 500$. Так же представлены оптимальные коэффициенты размытости для каждой ошибки. Как видно по рисунку с повышением ошибки моделирования возрастает оптимальный коэффициент размытости. Если сравнить данный рисунок 4 с рисунком 3 то можно отметить, что при повышении объема выборки понижается оптимальный коэффициент размытости.

Далее представим результаты зависимости ошибок моделирования от объема выборки в виде графиков. Рассмотрим параболическое ядро, так как по экспериментам представленные в таблице 3-5 - это ядро дало лучшие результаты. При коэффициенте размытости 1.

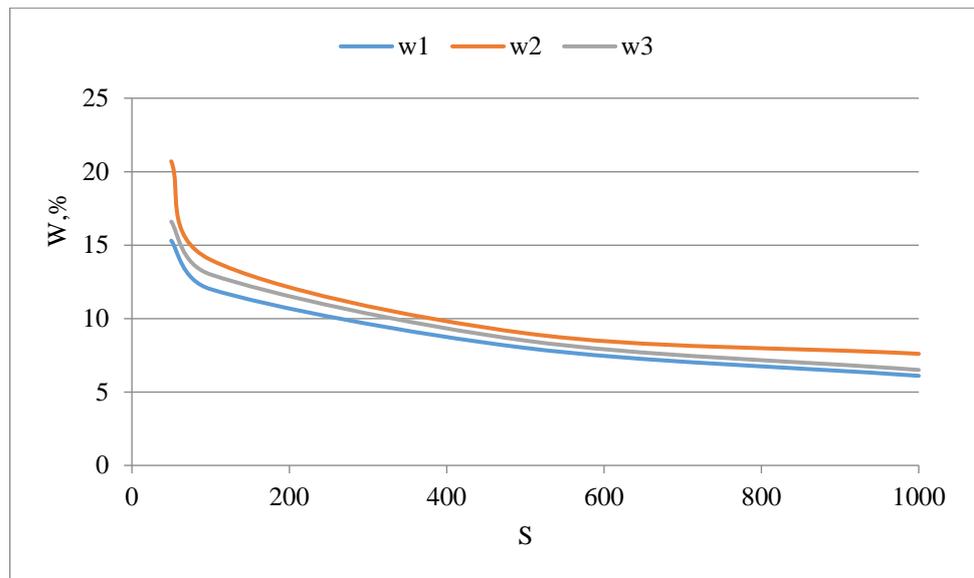


Рисунок 5 – Зависимость W от S , при $\xi = 0 \%$

На рисунке 5 представлена зависимость между ошибкой моделирования и объемом выборки при помехе равной 0%. Видим, что при увеличении объема выборки, ошибка моделирования возрастает – это означает, что точность вычислений увеличивается. Так же стоит отметить, что ошибка моделирования по исходным данным самая маленькая, а при удалении пропусков самая большая. Это означает, что потеря информации влечет за собой ухудшение полученных результатов. Минимальный результат ошибки моделирования соответствует модели построенной по полной выборке наблюдений, а ошибка, когда пропуски уже заполнены, является приближенным значением к ошибки по полной выборке наблюдений. Это означает, что точность наблюдений повышена.

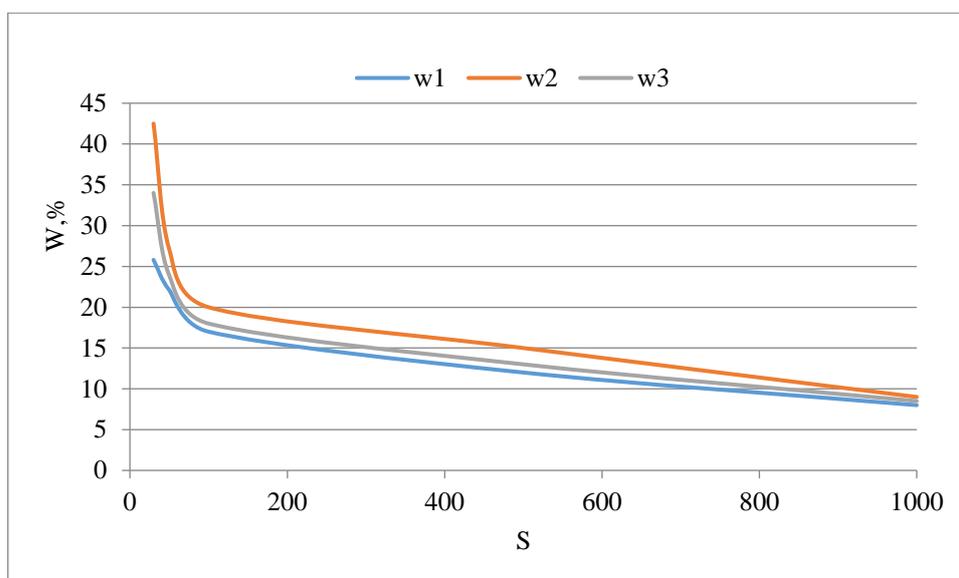


Рисунок 6 - Зависимость W от S , при $\xi = 5\%$

На рисунке 6 представлена зависимость между ошибкой моделирования и объемом выборки при помехе равной 5%. Стоит отметить, что при увеличении S уменьшается ошибка моделирования. Так же, минимальной ошибкой моделирования является значение по полной выборке, а ошибка, когда пропуски уже заполнены, является срединным значением между полной выборкой и выборкой данных с пропусками. Это означает, что точность наблюдений повышена.

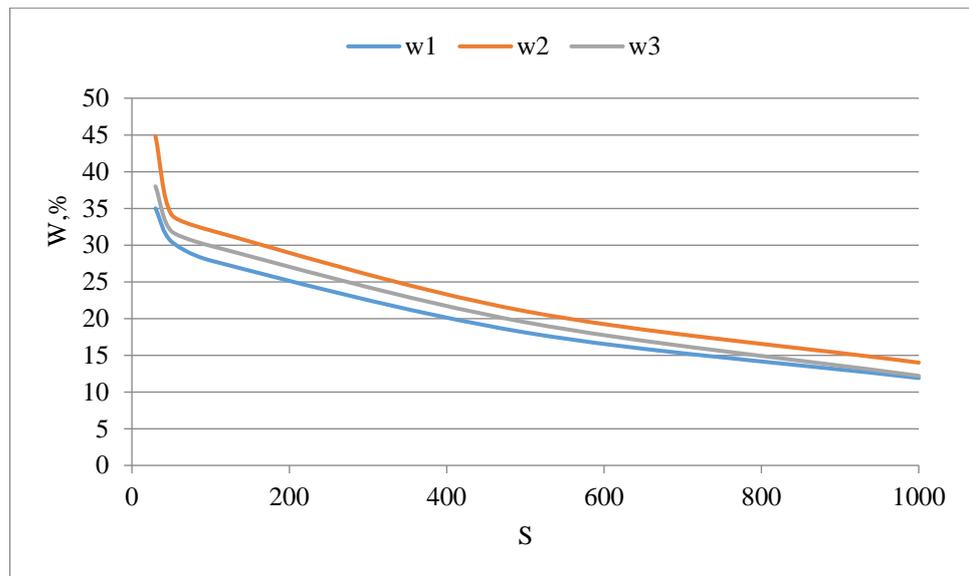


Рисунок 7 - Зависимость W от S , при $\xi = 10\%$

Сравнивая рисунок 6 и 7 можно отметить, что увеличения помехи, дает возрастание каждой из ошибки. На рисунке 7 так же можно отметить, что при увеличении объема выборки, рассчитанные значения становятся приближенными к исходным данным. Так же видно, что самый лучший результат ошибки моделирования строится по полной выборке, а ошибка, когда пропуски уже заполнены, является срединным значением между полной выборкой и выборкой данных с пропусками. Это означает, что точность наблюдений повышена.

Для того чтобы понять точность рассчитанных данных приведем сравнительный анализ для реального объекта и восстановленного по параболическому ядру.

На рисунке 8 представлено сравнение объекта и модели по исходным данным, а так же модель с рассчитанными данными по непараметрическому методу (4). Использовалось параболическое ядро $S = 100$, при $\xi = 0\%$.

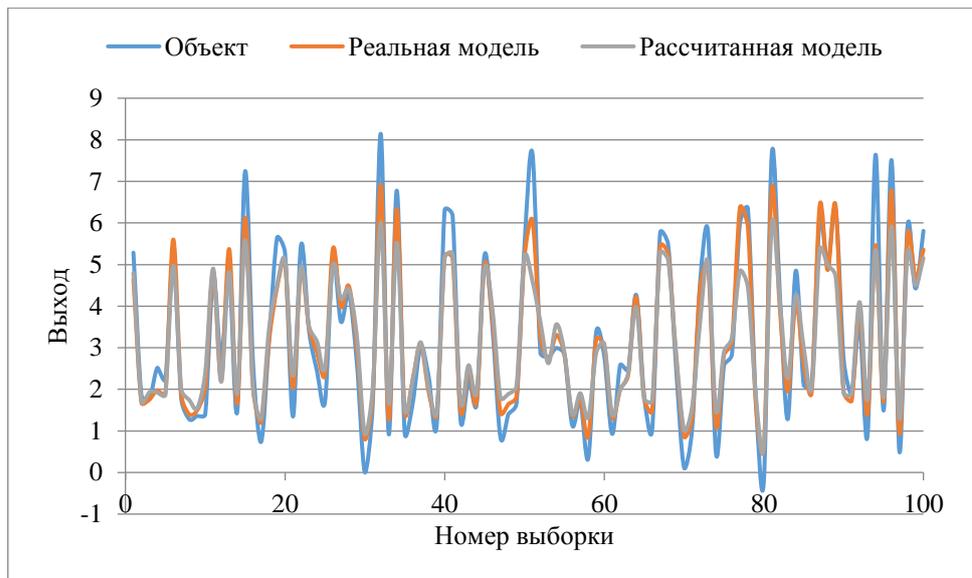


Рисунок 8 – Модель и объект при $S = 100$ и $\xi = 0 \%$

На рисунок 8 видно, что модель по объекту строится почти одинаково, с допустимой погрешностью, что показывает правильность работы программы. Так же видим сравнение реальной модели и рассчитанной. Стоит отметить, что при некоторых данных восстановление происходит почти идентично реальному объекту, что показывает точность восстановление пропущенных наблюдений.

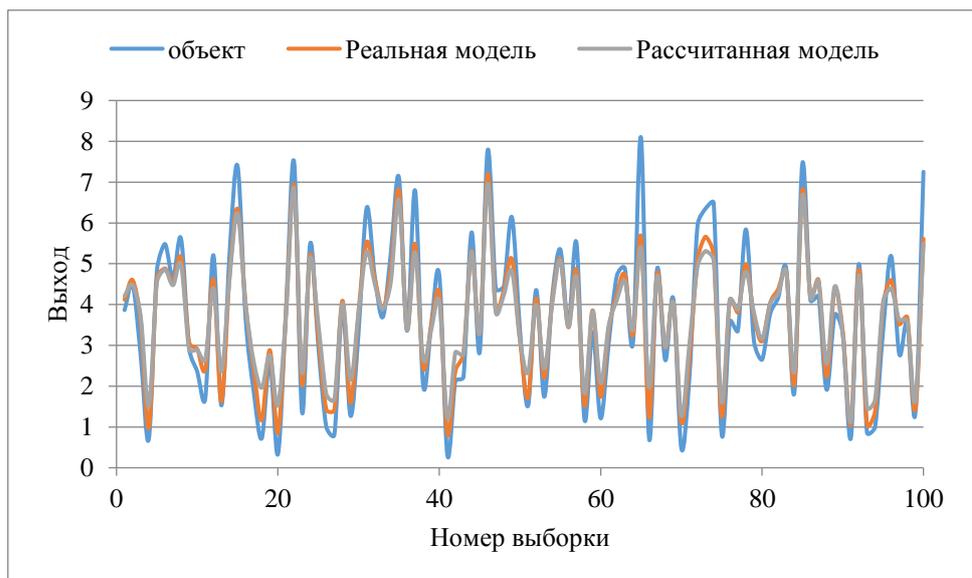


Рисунок 9 – Модель и объект при $S = 100$ и $\xi = 5 \%$

На рисунке 9 представлены объект и модель реальных данных, а так же рассчитанные пропуски модели. Видно, что рассчитанные значения не сильно отклоняются от реальных данных. Стоит отметить, что при некоторых данных восстановление происходит почти идентично реальному объекту, что показывает точность восстановления пропущенных наблюдений.

3.2 Результаты исследований заполнения пропусков, с помощью метода представленного в Deductor studio

Deductor studio – программа, которая содержит механизмы импорта, обработки, визуализации и экспорта данных для быстрого и эффективного анализа и прогнозирования.

Данная программа восстанавливает наблюдения с помощью среднеарифметического значения по известным данным. Для визуализации представим формулу, по которой будут рассчитываться данные.

$$x_s = \frac{1}{s} \sum_{i=1}^s x_i, \quad (17)$$

где s – объем известной выборки;

x_i - выход объекта.

Для того чтобы провести сравнительный анализ возьмем входные данные, сгенерированные по нормальному закону распределения на промежутке $[0;4]$, которые использовались для непараметрического метода, по заданному объекту (15). Использовались данные рассчитанные при параболическом ядре, так как оно лучше всех проявило себя в прошлых экспериментах. Так же будет найдена ошибка моделирования для рассчитанных значений по программе Deductor studio по формуле (18)

$$W = \sqrt{\frac{1}{s} \sum_{i=1}^s (x_{sd} - x_s)^2 / \frac{1}{s-1} \sum_{i=1}^s (m_{x_s} - x_s)^2}, \quad (18)$$

где s - объем выборки;

x_{sd} - рассчитанная модель по Deductor studio;

x_s - реальное значение модели;

m_{x_s} - математическое ожидания по реальным наблюдениям.

В таблице 6 представлены рассчитанные ошибки моделирования по найденным наблюдениям.

Введем обозначения ошибки моделирования полученной с помощью программы Deductor Studio – $W4$.

Таблица 6 - Зависимость объема выборки от помехи.

Объем выборки s	Значение ξ	$W4(\%)$
50	0	46.25
	0.5	50.42
	5	54.3
100	0	40.9
	0.5	43.01
	5	44.65
500	0	39.2
	0.5	41.29
	5	42.36

По таблице 6 можно сказать, что при увеличении объема выборки уменьшается ошибка моделирования. Стоит отметить, что повышая помеху, увеличивается и ошибка моделирования.

В следующем параграфе произведен сравнительный анализ непараметрического метода и метода представленный в программе Deductor studio.

3.3 Сравнительный анализ исследованных алгоритмов

Для сравнения алгоритмов будут использованы одинаковые данные, где входные наблюдения сгенерированы генератором случайных чисел по нормальному закону распределения на промежутке $[0;4]$. Выход объекта рассчитывался по формуле (15). Для построения модели использовано параболическое ядро, так как оно по экспериментам показало лучшие результаты. При коэффициенте размытости 1.

Исследуем зависимость ошибки моделирования от помехи и объема выборки.

Таблица 7 – Сравнение методов восстановления пропусков

Объем выборки s	Значение ξ	W1%	W2%	W3%	W4%
50	0	20,1	32,4	22,1	46,25
	5	22,1	34,1	23,9	50,42
	10	30,5	36,1	31,9	54,3
100	0	15,3	20,7	16,6	40,9
	5	17,2	21,5	18,1	43,01
	10	27,9	32	29,2	44,65
500	0	11,8	14,1	13,2	39,2
	5	12,3	15,5	13,3	41,29
	10	18,1	21,2	19,5	42,36

По таблице 7 можно сделать выводы, что самая большая ошибка моделирования соответствует методом заполнения среднеарифметическим, представленным в программе Deductor studio. При повышении объема выборки

есть ожидаемое уменьшение ошибки. Самую минимальную ошибку можно наблюдать у исходных данных.

Для большей наглядности данные из таблице 7 представим в виде столбиковой диаграммы.

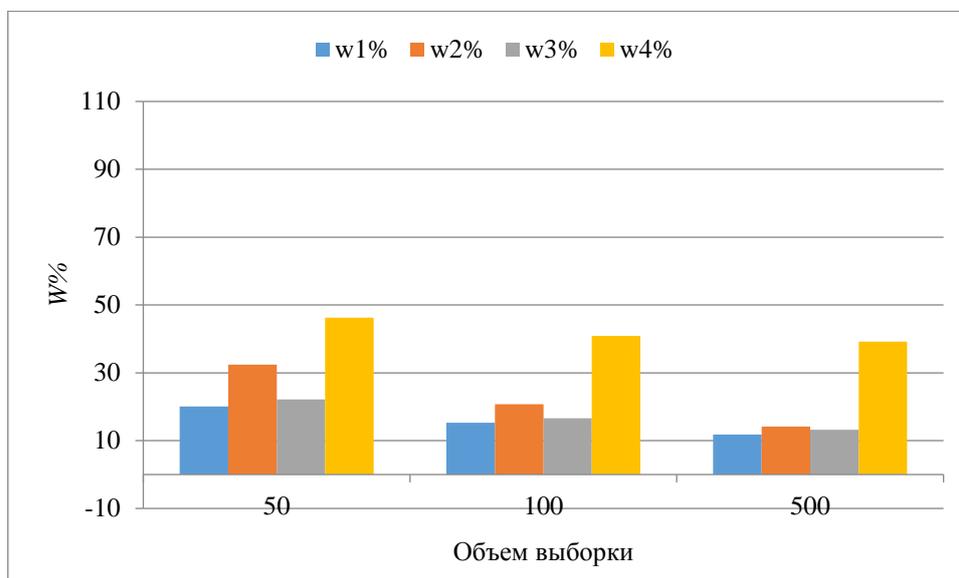


Рисунок 10 – Зависимость W от S, при $\xi = 0\%$

На рисунке 10 – построена зависимость между ошибкой моделирования и объемом выборки при разных методах.

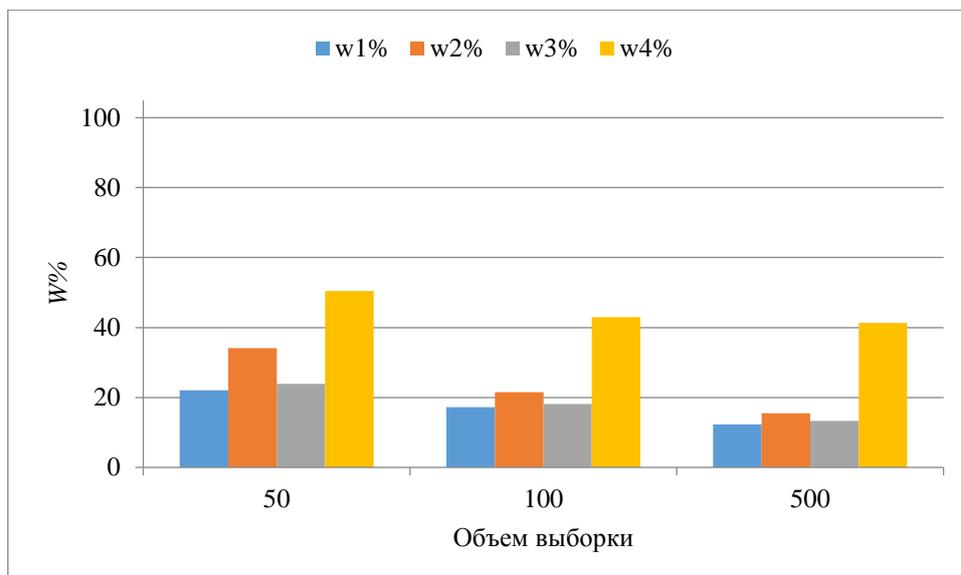


Рисунок 11 – Графическое представление W от S, при $\xi = 5\%$

На рисунке 11 – построена зависимость между ошибкой моделирования и объемом выборки при определенном методе.

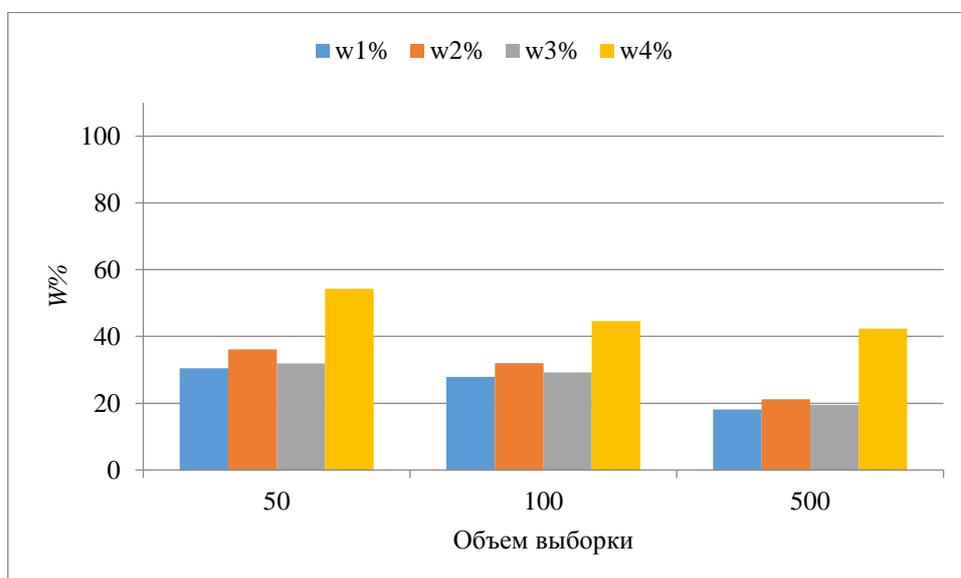


Рисунок 12 – Графическое представление W от S, при $\xi = 10\%$

На рисунке 12 – построена зависимость между ошибкой моделирования и объемом выборки при определенном методе.

На рисунках 10-12 представлены зависимости между объемом выборки и ошибкой моделирования при определенном методе и при разных помехах. На столбиковых диаграммах видно, что ошибка моделирования при удалении строк меньше, чем при заполнении среднеарифметическим значением, не зависимо от процента помехи. Так же видно, что ошибка непараметрического алгоритма намного точнее восстанавливает значения, чем метод представленный программой Deductor studio. Стоит отметить, что на каждой столбиковой диаграмме видно, что при увеличении объема выборки понижается ошибка моделирования.

Выводы по третьей главе

В третьей главе выпускной квалификационной работе были проведены численные исследования по двум методам:

1) На основе первого эксперимента была построена модель прогнозирования, основанная на оценке Надарая-Ватсона. В ходе работы были сделаны следующие выводы:

- Параболическое ядро показывает во всех экспериментах лучшие результаты, чем остальные ядра;
- Худшие результаты экспериментов показало прямоугольное ядро;
- Опыты, у которых происходило сокращение размерности ошибка моделирования значительно больше, чем у тех опытов, у которых рассчитаны пропущенные значения;
- При увеличении объема выборки, во всех случаях снижается ошибка моделирования, показывая этим, что чем больше информации мы имеет, тем лучше строится модель.

2) Были проведены эксперименты в программе Deductor Studio, которая заполняет пропущенные наблюдения среднеарифметическим значением по известным данным.

Так же было произведено сравнение двух выше перечисленных методов, где выяснилось, что заполнение пропусков с помощью непараметрического метода является наиболее точным, чем заполнение среднеарифметическим значением.

ЗАКЛЮЧЕНИЕ

Моделирование играет очень большую роль в самых различных сферах. Его необходимо применять, так как эксперименты и наблюдения над реальным объектом могут быть неприемлемыми, и на то есть различные причины: опасность для здоровья и окружающей среды, невозможность наблюдать за внутренним устройством системы в реальном мире, дороговизна и т.д. В работе приведено несколько этапов задачи идентификации. Рассмотрены определения идентификации в «узком» и «широком» смысле.

Был предложен и исследован метод непараметрического восстановления пропусков, основанный на оценке Надарая-Ватсона. Было экспериментально выявлено, что из трех представленных ядерных функции более точно строит и показывает результаты – параболическое ядро. Так же рассмотрен представленный метод в программе Deductor Studio, который основан на заполнении пропусков с помощью среднеарифметического значения по известным данным.

Выводы во всех экспериментах почти одинаковы. Чем больше выборка - тем меньше ошибка моделирования, отсюда следует вывод, что для того, чтобы получить более правильную картину о построении модели или о работе метода, нужно собрать как можно больше данных для увеличения объема выборки. Такая выборка будет более приемлемой для анализа. Маленький объем выборки не имеет большого смысла, так как моделирование будет плохо отображать объект.

Стоит помнить, что чем меньше помеха, тем точнее вычисления, но применять помеху необходимо, чтобы посмотреть, как ведет себя модель объекта и насколько она устойчива в реальных условиях, где помеха и погрешность присутствуют всегда.

Так же был проведен сравнительный анализ алгоритмов, где у непараметрического алгоритма ошибка моделирования намного меньше, чем в программе Deductor Studio. Это означает, что непараметрический метод более точно заполняет пропущенные значения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1) Эйкхофф, П. Основы идентификации систем управления : учебник / П. Эйкхофф. – Москва : Мир, 1975. – 680 с.
- 2) Веников, В. А. Теория подобия и моделирования : учебное пособие / В. А. Веников. – Москва : Высшая школа, 1976. – 479 с.
- 3) Самарский, А. А. Математическое моделирование: Идеи. Методы. Примеры : монография / А. А. Самарский, А. П. Михайлов – Москва.: Физматлит, 2005 – 320 с.
- 4) Клюкина, Е.А. Общая теория систем : учебное пособие / Е. А. Клюкина. – Петрозавоск : ПетрГУ, 2014. – 86 с.
- 5) Сушкин, И. Н. Вычислительная техника и информационные технологии : учебное пособие / И. Н. Сушкин. – Красноярск, 2007 – 183 с.
- 6) Игнатьев, Д. А, Медведев, А. В., Сергеев, Ц. В., Шестернев, А. И. О непараметрическом моделировании многосвязных процессов / Д. А. Игнатьев, А. В. Медведев, Ц. В. Сергеев // Вестник Сибирского государственного аэрокосмического университета им. академика М. Ф. Решетнева. – 2008. – № 3 (20) – С. 69 – 72.
- 7) Цепкова, М. В., Сергеева, Н. А. О непараметрическом моделировании динамических процессов / М. В. Цепкова, Н. А. Сергеева // Вестник : томского государственного университета управление, вычислительной техники и информатики. – 2013. – № 2 (23). – С. 92 – 101.
- 8) Коновалов, В. И. Идентификация и диагностика систем : учебное пособие / В. И. Коновалов. – Томск : ТПУ, 2006. – 152 с.
- 9) Цыпкин, Я. З. Информационная теория идентификации : монография / Я. З. Цыпкин — Москва.: Наука. Физматлит, 1995. – 336 с.
- 10) Боровков, А.А. Математическая статистика. Оценка параметров. Проверка гипотез : монография / А. А. Боровков.– Москва : Наука, 1984. – 472 с.

- 11) Корнеева А. А., Чжан Е. А. О непараметрическом моделировании стохастических объектов / А. А. Корнеева, Е. А. Чжан // Вестник СибГАУ. – 2013. – № 2 (23). – С. 37 – 42
- 12) Литтл, Р. Дж. А., Рубин Д. Б. Статистический анализ данных с пропусками: Пер. с англ / Р. Дж. А. Литтл, Д. Б. Рубин – Москва : Финансы и статистика. 1990. – 336 с.
- 13) Тихова Г.П. Пропуск данных в выборке: как решать проблему и как ее избежать. / Г. П. Тихова // Регионарная анестезия и лечение острой боли. – 2016. – № 10 (3). – С. 205–209.
- 14) Банникова А. В., Михов Е. Д. О непараметрическом алгоритме управления макрообъектом / А. В. Банников, Е. Д. Михов // Молодой ученый. – 2014. – № 1 (60). – С. 115–119.
- 15) Медведев А. В. Теория непараметрических систем. Процессы / А. В. Медведев // Вестник СибГАУ. – 2010. – № 4(29). – С. 4–9.
- 16) Зангиева, И. К. Проблема пропусков в социологических данных: смысл и подходы к решению / И. К. Зангиева // Социология: методология, методы, математическое моделирование. – 2011. – № 33. – С. 28–56.
- 17) Злоба, Е., Яцкив И. Статистические методы восстановления пропущенных данных / Е. Злоба, И. Яцкив // Computer Modeling & New Technologies. – 2004. – № 6. – С. 51–61.
- 18) Алексеева В. А., Донцова Ю. С., Клячкин В. Н. Восстановление пропущенных наблюдений при классификации объектов / В. А. Алексеева, Ю. С. Донцова, В. Н. Клячкин / Самарский науч.центр РАН. – 2014. – № 6(2). – С. 357 – 359.

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

УТВЕРЖДАЮ

Заведующий кафедрой

 Ю. Ю. Якунин

« 11 » июня 2018 г.

БАКАЛАВРСКАЯ РАБОТА

27.03.03 Системный анализ и управление

Повышение точности задачи идентификации по выборке наблюдений с пропусками

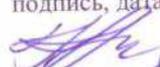
Руководитель


8.06.18
подпись, дата

канд. техн. наук, доцент
должность, ученая степень

А.А Корнеева
инициалы, фамилия

Выпускник


8.06.18
подпись, дата

А.А Климова
инициалы, фамилия

Красноярск 2018