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
«____» _____ 2018 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Непараметрический алгоритм восстановление пропусков «входных-выходных» переменных процесса

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

Научный руководитель	_____	к.т.н., доцент	Корнеева А.А.
	подпись, дата	должность, ученая степень	инициалы, фамилия
Выпускник	_____		Романова Т.С.
	подпись, дата		инициалы, фамилия
Рецензент	_____	к.т.н., доцент	_____
	подпись, дата	должность, ученая степень	инициалы, фамилия

Красноярск 2018

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Большинство методов многомерного статистического анализа данных, такие как факторный, кластерный, регрессионный анализ и многие другие, требуют отсутствия пропусков в анализируемых данных. Однако в реальности в практических задачах анализа данных приходится сталкиваться с проблемой, когда часть данных отсутствует, что усложняет дальнейшую обработку и анализ данных. Причины могут быть различными, например, неответ респондента на какой-либо вопрос, отсутствие значение по причине выхода из строя оборудования. Перед исследователем возникаем вопрос: отсеять данные с пропуском или заполнить их на основе имеющейся информации.

В первом случае теряется полезная информация по заполненным столбцам, либо в результате удаления данных для анализа, может остаться слишком мало информации. В связи с этим особую важность приобретает задача заполнения пропусков в данных, содержащих пропущенные значения.

Существуют различные подходы к решению данной задачи, которые различаются по своей природе, области применимости и вычислительной сложности. В данной работе рассматриваются алгоритм Zet, основывающийся на оценивании закономерностей взаимосвязи между строками и столбцами на локальном участке анализируемого элемента и непараметрический алгоритм восстановления пропусков.

Целью дипломной работы является повышение точности решения задачи идентификации по выборкам наблюдений с пропусками.

Для достижения поставленной цели необходимо решить следующие задачи:

- Реализовать и исследовать алгоритм непараметрического заполнения пропусков;
- Реализовать и исследовать алгоритм заполнения пропусков Zet;
- Провести сравнительный анализ исследуемых алгоритмов.

1. Задача идентификации дискретно-непрерывных процессов

1.1 Идентификация в системном анализе

1.1.1 Основные понятия системного анализа

Системный анализ — научный метод, который отличается междисциплинарным подходом к решению сложных проблем. Объектом системного анализа выступают практические проблемы, которые связанны с созданием новых и модернизацией существующих систем. Это организационные, экономические, технические, информационные, военные и другие системы.

Цели анализа систем:

- детальное изучение систем для более эффективного использования и принятия решения по дальнейшему совершенствованию или замене;
- исследование альтернативных вариантов вновь создаваемой системы с целью выбора наилучшего варианта.

Задачами системного анализа являются:

- определение объекта анализа;
- определение функциональных особенностей системы управления;
- определение количественных и качественных показателей системы управления;
- оценка эффективности систем управления;
- обобщение и оформление результатов анализа.

Основные этапы системного анализа:

1 Фиксация проблемы.

Постановка проблемы – отправной момент исследования. Проблемой называется ситуация, характеризующаяся различием между необходимым (желаемым) выходом и существующим выходом. Выход является

необходимым, если его отсутствие создает угрозу существованию или развитию системы. Существующий выход обеспечивается существующей системой. Желаемый выход обеспечивается желаемой системой. Проблема есть разница между существующей и желаемой системой. Проблема может заключаться в предотвращении уменьшения выхода или же в увеличении выхода. Система, заполняющая промежуток, является объектом конструирования и называется решением проблемы.

2 Выявление целей.

Цели указывают направление, в котором надо двигаться, чтобы поэтапно решить проблему.

3 Определение конфигуратора.

Необходимым условием успешного решения проблемы является наличие адекватной модели проблемной ситуации, с ее помощью можно будет испытывать и сравнивать варианты предполагаемых действий. Эта модель (или совокупность моделей) неизбежно должна строиться средствами некоторого языка (или языков). Встает вопрос о том, сколько и какие именно языки нужны для работы над данной проблемой и как их выбирать. Например, если произошло дорожно-транспортное происшествие, то для разрешения возникшей проблемы могут потребоваться языки: правовой (кто за что отвечает), медицинский (состояние участников ДТП до и после), технический (состояние дороги и техники), административный (организация ликвидации всех последствий), экономический (финансовое обеспечение) и т.д.

Важно подчеркнуть, что проблемы реальной жизни не бывают однодисциплинарными, т.е. описываемыми на языке какой-нибудь одной специальности.

Конфигуратором называется минимальный набор профессиональных языков, позволяющий дать полное (адекватное) описание проблемной ситуации и ее преобразований. Вся работа в ходе решения проблемы будет происходить на языках конфигуратора. И только на них.

4 Определение критериев.

В ходе решения проблемы будет необходимо сравнивать предлагаемые варианты, оценивать степень достижения цели или отклонения от нее, осуществлять контроль за ходом событий. Это достигается путем выделения некоторых признаков рассматриваемых объектов и процессов. Данные признаки должны быть связаны с интересующими нас особенностями рассматриваемых объектов или процессов, должны быть доступными для наблюдения и измерения. Тогда по полученным результатам измерений мы сможем осуществить необходимый контроль. Такие характеристики называют критериями.

Очевидно, что чем меньше критериев понадобится, тем проще будет проводить сравнение. Но, к сожалению, чаще одним единственным критерием не удается удовлетворительно оценить качество рассматриваемого объекта. Тогда приходится вводить еще какое-то количество критериев, по-разному описывающих объект и дополняющих друг друга.

При выборе критериев иногда можно воспользоваться опытом ранее проведенных работ. Например, при анализе и проектировании технических систем обычно используются такие критерии, как финансовые (стоимость, прибыль и т.д.), инвентарные (количество продукта, ассортимент и т.д.), эксплуатационные (эффективность функционирования, надежность и пр.), живучесть (совместимость с существующими системами, адаптивность к среде, скорость морального устаревания, безопасность и пр.), экологичность, эргономичность и ряд других.

5 Построение и усовершенствование моделей.

Модели значительно облегчают понимание системы, позволяют проводить исследования в абстрактном плане, прогнозировать поведение системы в интересующих нас условиях, упрощать задачи, анализировать и синтезировать совершенно различные системы одними методами. Основная задача и в то же время преимущество модели - выделение частных, но наиболее важных факторов реальной системы, которые подлежат изучению в данном конкретном исследовании. Эти факторы должны быть отражены в модели с

наибольшей полнотой и детализацией, их характеристики в модели должны совпадать с реальными с точностью, определяемой требованиями данного исследования. Остальные, несущественные факторы могут быть либо отражены с меньшей точностью, либо вовсе отсутствовать в модели.

6 Генерирование альтернатив.

На данном этапе необходимо сгенерировать множество альтернатив, из которых затем будет осуществляться выбор наилучшего пути развития системы. Данный этап системного анализа является очень важным и трудным. Важность его заключается в том, что конечная цель системного анализа состоит в выборе наилучшей альтернативы на заданном множестве и в обосновании этого выбора. Если в сформированное множество альтернатив не попала наилучшая, то никакие самые совершенные методы анализа не помогут ее вычислить. Трудность этапа обусловлена необходимостью генерации достаточно полного множества альтернатив, включающего в себя, на первый взгляд, даже самые нереализуемые.

7 Выбор или принятие решений.

Процесс принятия решения состоит в выборе рационального решения из некоторого множества альтернативных решений с учетом системы предпочтений.

8 Реализация улучшающего вмешательства.

После принятия решения о том, какое именно из улучшающих вмешательств следует осуществить (это итог предыдущего этапа), предстоит работа по реализации этого решения

1.1.2 Модель и моделирование

.

Модель – это объект или описание объекта, системы для замещения (при определенных условиях предложениях, гипотезах) одной системы (т.е. оригинала) другой системой для изучения оригинала или воспроизведения его каких-либо свойств.

Моделирование – это замещение некоторого объекта А другим объектом Б. Замещаемый объект А называется оригиналом или объектом моделирования, а замещающий Б - моделью. Другими словами, модель - это объект-заменитель объекта-оригинала, обеспечивающий изучение некоторых свойств оригинала.

Почему же мы прибегаем к моделированию вместо того, чтобы попытаться напрямую взаимодействовать с реальным миром? Можно назвать три основных причины.

1. Сложность реальных объектов.

Учёт всех факторов, относящихся к решаемой проблеме, обычно превосходит человеческие возможности. Поэтому часто единственным выходом является упрощение ситуации с помощью неполной копии, в результате чего уменьшается разнообразие учитываемых факторов до уровня восприимчивости человека, решающего проблему.

2. Сложности с проведением прямого взаимодействия (эксперимента).

На практике часто экспериментальное исследование ограничено из-за высокой стоимости либо вообще невозможно (опасно, либо современная техника эксперимента ещё не доросла до требуемого уровня).

3. Необходимость прогнозирования.

Как правило, любая деятельность редко осуществляется по жесткой программе, только с априорным учетом событий, происходящих на промежуточных этапах. Чаще приходится оценивать текущий результат и выбирать следующий шаг из числа возможных. Это означает, что необходимо сравнивать последствия всех вариантов действия, не выполняя их реально, а проигрывая на модели.

Среди прочих причин можно назвать следующие:

- исследуемый объект слишком велик (Солнечная система), либо слишком мал (атом);

- исследуемый процесс протекает очень быстро (взрыв), либо очень медленно (геологические процессы);

- непосредственное экспериментирование с объектом может привести к его разрушению (испытания самолёта, автомобиля).

Чтобы соответствовать своему предназначению и обеспечивать свою практическую полезность, модели должны отвечать ряду требований, таких как:

- Адекватность – достаточно точное отображение свойств объекта;
- Полнота – предоставление получателю всей необходимой информации об объекте;
- Гибкость – возможность воспроизведения различных ситуаций во всем диапазоне изменения условий и параметров;
- Трудоемкость разработки должна быть приемлемой для имеющегося времени и программных средств;
- Экономичность - определяется затратами ресурсов ЭВМ памяти и времени на ее реализацию и эксплуатацию.

Для того, чтобы построить модель можно выделить следующие этапы:

1. Постановка задачи.

Определение цели анализа и пути ее достижения и выработки общего подхода к исследуемой проблеме. На этом этапе требуется глубокое понимание существа поставленной задачи. Иногда, правильно поставить задачу не менее сложно, чем ее решить.

2. Идеализация объекта

Отбрасываются все факторы и эффекты, которые представляются не самыми существенными для его поведения. Например, при составлении баланса материи не учитывался, ввиду его малости, дефект масс, которым сопровождается радиоактивный распад. По возможности идеализирующие предположения записываются в математической форме, с тем, чтобы их справедливость поддавалась количественному контролю.

3. Формализация.

Заключается в выборе системы условных обозначений и с их помощью записывать отношения между составляющими объекта в виде математических выражений. Устанавливается класс задач, к которым может быть отнесена полученная математическая модель объекта.

4. Выбор метода решения.

На этом этапе устанавливаются окончательные параметры моделей с учетом условия функционирования объекта. Для полученной математической задачи выбирается какой-либо метод решения или разрабатывается специальный метод. При выборе метода учитываются знания пользователя, его предпочтения, а также предпочтения разработчика.

5. Реализация модели.

Разработав алгоритм, пишется программа, которая отлаживается, тестируется и получается решение нужной задачи.

6. Анализ полученной информации.

Сопоставляется полученное и предполагаемое решение, проводится контроль погрешности моделирования.

7. Проверка адекватности реальному объекту.

Процесс моделирования является итеративным. В случае неудовлетворительных результатов этапов 6 или 7 осуществляется возврат к одному из ранних этапов, который мог привести к разработке неудачной модели. Этот этап и все последующие уточняются и такое уточнение модели происходит до тех пор, пока не будут получены приемлемые результаты.

Остановимся на одном из наиболее универсальных видов моделирования - математическом, ставящим в соответствие моделируемому физическому процессу систему математических соотношений, решение которой позволяет получить ответ на вопрос о поведении объекта без создания физической модели, часто оказывающейся дорогостоящей и неэффективной.

Математическое моделирование – это средство изучения реального объекта, процесса или системы путем их замены математической моделью, более удобной для экспериментального исследования с помощью ЭВМ.

Математическая модель является приближенным представлением реальных объектов, процессов или систем, выраженным в математических терминах и сохраняющим существенные черты оригинала. Математические модели в количественной форме, с помощью логико-математических конструкций, описывают основные свойства объекта, процесса или системы, его параметры, внутренние и внешние связи.

В общем случае математическая модель реального объекта, процесса или системы представляется в виде системы функционалов $\Phi_i(X, Y, Z, t) = 0$, где

X – вектор входных переменных, $X=[x_1, x_2, x_3, \dots, x_N]^t$,

Y – вектор выходных переменных, $Y=[y_1, y_2, y_3, \dots, y_N]^t$,

Z – вектор внешних воздействий, $Z=[z_1, z_2, z_3, \dots, z_N]^t$,

t – координата времени.

При построении математической модели перед исследованием возникает задача выявить и исключить из рассмотрения факторы, несущественно влияющие на конечный результат (математическая модель обычно включает значительно меньшее число факторов, чем в реальной действительности). На основе данных эксперимента выдвигаются гипотезы о связи между величинами, выражающими конечный результат, и факторами, введенными в математическую модель.

Конечной целью этого этапа является формулирование математической задачи, решение которой с необходимой точностью выражает результаты, интересующие специалиста.

Для построения математической модели необходимо:

- тщательно проанализировать реальный объект или процесс;
- выделить его наиболее существенные черты и свойства;
- определить переменные, т.е. параметры, значения которых влияют на основные черты и свойства объекта;

- описать зависимость основных свойств объекта, процесса или системы от значения переменных с помощью логико-математических соотношений (уравнения, равенства, неравенства, логико-математические конструкций);
- выделить внутренние связи объекта, процесса или системы с помощью ограничений, уравнений, равенств, неравенств, логико-математических конструкций;
- определить внешние связи и описать их с помощью ограничений, уравнений, равенств, неравенств, логико-математических конструкций.

Математическое моделирование, кроме исследования объекта, процесса или системы и составления их математического описания, также включает:

- построение алгоритма, моделирующего поведение объекта, процесса или системы;
- проверка адекватности модели и объекта, процесса или системы на основе вычислительного и натурного эксперимента;
- корректировка модели;
- использование модели.

Построение математической модели обычно начинается с построения и анализа простейшей, наиболее грубой математической модели рассматриваемого объекта, процесса или системы. В дальнейшем, в случае необходимости, модель уточняется, делается ее соответствие объекту более полным.

Выбор той или иной модели определяется требованием точности. С повышением точности модель приходится усложнять, учитывая новые и новые особенности изучаемого объекта, процесса или системы. Однако, чем выше требования к точности результатов решения задачи, тем больше необходимость учитывать при построении математической модели особенности изучаемого объекта, процесса или системы. Но здесь важно во время остановиться, так как сложная математическая модель может превратиться в трудно разрешимую задачу.

1.1.3 Идентификация

Однако не всегда удаётся построить модель, которая бы полностью соответствовала всем протекающим процессам реального объекта, математическое описание которых известно неточно, либо отсутствует полностью. Для успешного управления такими объектами их необходимо идентифицировать. Что же понимается под задачей идентификации?

В зависимости от уровня априорной информации об объекте различают идентификацию в «узком» и «широком» смыслах.

Под идентификацией в широком смысле понимается определение структуры и параметров динамических объектов по наблюдаемым данным: входному воздействию и выходным величинам. В этом случае исследуемый объект представляет собой «чёрный ящик», структура и параметры внутри которого полностью неизвестны и должны быть определены.

Под идентификацией в узком смысле понимается задача определения параметров элементов в известной структуре математической модели объекта. В этом случае исследуемый объект представляет собой «серый ящик», параметры которого следует определить.

1.1.4 Постановка задачи идентификации

Наиболее общую схему исследования процесса можно представить в следующем виде:

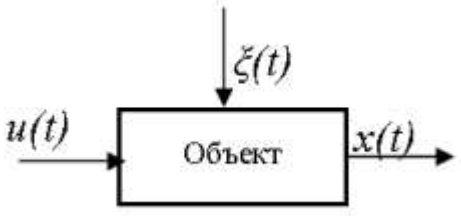


Рисунок 1 – Схема объекта

На рисунке 1 приняты следующие обозначения:

– $u(t)$ – входное воздействие;

– $x(t)$ – выходная переменная процесса;

– (t) – непрерывное время;

– $\xi\,(t)$ – векторное случайное воздействие;

Воздействия внешней среды на объект в обобщенном виде изображены стрелками, направленными к объекту. Объект, в свою очередь, воздействует на окружающую среду. Это воздействие показано стрелкой, направленной от объекта и обозначенной через *x*.

Входные и выходные воздействия могут описываться определенными функциями (обычно функциями времени). Математически соответствие между входной и выходной функциями можно записать в виде выражения

$$x(t) = f\,(u(t), \xi\,(t)\,), \qquad (1.1)$$

где f – неизвестный оператор объекта.

Рассмотрим процесс, который относится к классу дискретно-непрерывных, при котором сам процесс непрерывен, но его «входные-выходные» данные фиксируются в дискретные промежутки времени.

Переменные $u(t), \xi\,(t)$ контролируются через интервал времени Δt. Для измерения выходной переменной $x(t)$ требуется время ΔT. При этом $\Delta t \leq \Delta T$. Различие дискретности измерения переменных может быть обусловлено средствами контроля. Например, измерения некоторых переменных может осуществляться электрическими средствами, измерения же других переменных может быть проведено в результате длительных физико-механических испытаний или химического анализа. В результате переменные процесса контролируются с различной дискретностью, что приводит к появлению пропусков. Матрица наблюдений "входных-выходных" в этом случае имеет следующий вид (Таблица 1).

Таблица 1 - Матрица наблюдений "входных-выходных" переменных с пропусками

u				x
u_1	u_2	...	u_m	
u_{11}	u_{21}	...	u_{m1}	-
u_{12}	u_{22}	...	u_{m2}	x_2
u_{13}	u_{23}	...	u_{m3}	x_3
u_{14}	u_{24}	...	u_{m4}	-
...	...	...	...	...
u_{1s}	u_{2s}	...	u_{ms}	x_s

В таблице 1 столбцы представляют собой входные (u_m) и выходные (x_s) переменные процесса, строки – наблюдения, «-» - пропуски.

Пропуски в данных значительно усложняют процесс моделирования и снижают точность задачи идентификации. В связи с чем, необходимо заполнить пропуски в данных с целью повышения точности задачи идентификации. Для решения данной задачи можно воспользоваться методами параметрической и непараметрической идентификации.

1.1.4.1 Параметрическая и непараметрическая идентификация

Непараметрические методы ориентированы на случай, когда априорная информация о структуре модели объекта отсутствует или игнорируется, т.е. когда объект рассматривается как «черный ящик». Если имеется априорная информация об уравнениях модели объекта, заданных с точностью до неизвестных параметров, то задача идентификации сводится к оценке этих параметров. Это случай параметрической идентификации.

Общих рекомендаций, когда следует использовать методы параметрической, а когда – непараметрической идентификации не существует – все определяется конкретной задачей исследования. Однако во всех случаях идентификации приходится считаться с неточностями в задании модели, неточностями в измерениях сигналов шумами и вычислительными погрешностями. Как результат малые погрешности в эмпирических данных и в задании модели могут привести к значительным ошибкам в результатах идентификации.

1.1.4.2 Постановка задачи параметрической идентификации

Если априорная информация об идентифицируемом объекте отсутствует или недостаточна, приходится предварительно осуществлять выбор структуры системы и класса моделей, то речь идет об идентификации в «узком» смысле. На основании имеющейся априорной информации, определяется параметрический класс операторов A^{α}

$$\tilde{x}_{\alpha}(t) = A^{\alpha}(u(t), \alpha), \quad (1.2)$$

где A^{α} – параметрическая структура модели;

α – вектор параметров.

Задача сводится к определению параметров объекта α по результатам наблюдений за входными и выходными сигналами.

На первом этапе определяется структура модели с точностью до параметров, например

$$x_{\alpha}(t) = f(u(t), \alpha), \quad (1.3)$$

где f – некоторая функция;

α – вектор параметров;

На следующем этапе осуществляется оценка параметров α на основе результатов измерений значений его входного и выходного сигналов $\{u_i, x_i, i = \overline{1, S}\}$. Для этого чаще всего используется метод наименьших квадратов. Данный

метод основан на принципе минимизации суммы квадратов отклонений некоторой функций от искомых значений

$$I(\alpha) = \frac{1}{S}\sum_{i=1}^{S}(x_i - \hat{x}_i)^2 \to \min_{\alpha}, \quad (1.4)$$

где x_i – выход объект;

$\hat{x}_i$ – выход модели.

Затем находятся частные переменные по коэффициентам α и приравниваются к нулю

$$\frac{\partial I}{\partial \alpha} = 0 \quad (1.5)$$

В результате получаем систему уравнений, которую решаем любым удобным способом, например методом Крамера, Гаусса. После чего получаем коэффициенты α и строим модель.

1.1.4.3 Постановка задачи непараметрической идентификации

.

Если априорная информация об объекте достаточно велика, внутренняя структура объекта известна и задан класс моделей, к которому можно отнести данный объект, то речь идет об идентификации в «широком» смысле. В данном случае отсутствует этап выбора параметрического класса оператора (1.2). Часто оказывается значительно проще определить класс операторов на основе сведений качественного характера, например линейности процесса ли типа нелинейности, однозначности либо неоднозначности и др. В этом случае задача идентификации состоит в оценивании этого оператора на основе выборки $\{x_i, u_i, i = \overline{1,S}\}$ в форме

$$\tilde{x}_s(t) = A_s(u(t), \vec{x}_s, \vec{u}_s), \quad (1.6)$$

где $\vec{x}_s = (x_{1,}x_2, \dots, x_s)$ – выходной вектор;

$\vec{u}_s = (u_{1,}u_2, \dots, u_s)$ – входной вектор.

В качестве модели объекта примем непараметрическую оценку функции регрессии по наблюдениям:

$$x_s(u) = \frac{\sum_{i=1}^{S} x_i \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{Cs}\right)}{\sum_{i=1}^{S} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{Cs}\right)}, \quad (1.7)$$

где $\Phi(\cdot)$ – ядерная колокообразная функция;

Cs – коэффициент размытости ядра.

Параметр размытости Cs определяется путем решения задачи минимизации квадратичного критерия выхода объекта и выхода модели, основанного на скользящем экзамене, когда в модели (1.7) исключается i-я переменная, предъявляемая для экзамена:

$$R(C_s) = \sum_{k=1}^{S} (x_k - x_s(u_k, c_s))^2 = \min \text{Cs}, \text{k} \neq i \quad (1.8)$$

Существует множество видов ядер, например, треугольное (1.9), параболическое (1.10) и кубическое (1.11).

$$\Phi(\cdot) = \begin{cases} 1 - \frac{u-u_i}{Cs}, \frac{u-u_i}{Cs} \leq 1 \\ 0{,}1 < \frac{u-u_i}{Cs} \end{cases}, \quad (1.9)$$

$$\Phi(\cdot) = \begin{cases} 0{,}75 \cdot (1 - (\frac{u-u_i}{Cs})^2), \frac{u-u_i}{Cs} \leq 1 \\ 0{,}1 < \frac{u-u_i}{Cs} \end{cases}, \quad (1.10)$$

$$\Phi(\cdot) = \begin{cases} (1 + 2 \cdot \left|\frac{u-u_i}{Cs}\right|)(1 - (\frac{u-u_i}{Cs})^2), \frac{u-u_i}{Cs} \leq 1 \\ 0{,}1 < \frac{u-u_i}{Cs} \end{cases}. \quad (1.11)$$

Графики данных ядер представлены на рисунке 2.

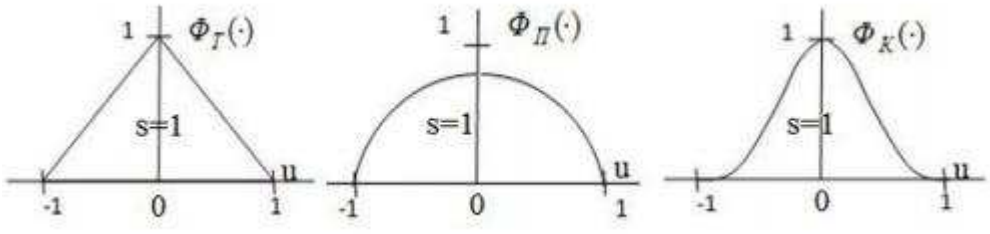


Рисунок 2 – Виды ядерных функций

Для расчета каждого значения x_s используются только несколько ближайших точек к x_s, т.е. те точки, которые попали в ядро. Очевидным является тот факт, что от количества элементов, по которым вычисляется оценка, зависит ее точность. В случае, если в Cs – окрестности точки u' нет ни одного элемента обучающей выборки, то оценку будет дать невозможно. В этом случае оценка (1.7) будет равна неопределенности вида (0/0). Одни из возможных способов получения оценки на этом пути является увеличение параметра размытости *Cs*. В некоторых случаях это позволит получить прогноз (уйти от неопределенности), но точность прогноза $x_s(u')$ может оказаться неудовлетворительной.

1.2 Анализ данных при решении задачи идентификации

При моделировании и идентификации объектов различной природы имеются некоторые начальные сведения об объекте. Эти сведенья обычно понимают как априорную или начальную информацию. Эта информация включает в себя различные требования к процессу, его уравнения и параметры, а также некоторые воздействия. Априорная информация с течением некоторого времени может утратить свою достоверность.

Следует не путать априорную и апостериорную информацию. Апостериорная или текущая информация может быть получена в результате выполнения эксперимента и является выборкой «входных-выходных» переменных процесса. Апостериорная информация обновляется в каждый момент времени выполнения эксперимента. Иногда она используется для накопления априорной информации для данного объекта, но в основном текущая информация предназначена для компенсации недостатка априорной информации.

Анализируемая выборка должна отвечать критериям качественности и полноты. Степень полноты априорной информации напрямую влияет на

точность решения задачи идентификации. Однако, на практике приходится сталкиваться с ситуацией, когда некоторые из признаков одного или нескольких объектов отсутствуют - возникает ситуация данных с пропусками, что значительно осложняет дальнейшую статистическую обработку. Наиболее простой способ устранения «некачественных» данных – это исключение из рассмотрения наблюдения, имеющего пропуск/шумы/выбросы. Часто, процессы в сложных системах имеют довольно большую размерность, но небольшой объём выборки, удаление же наблюдений с пропусками приведет к уменьшению и так небольшой выборки. Так же есть вероятность удаления важной и полезной информации, отсутствие которой может привести к сильным смещениям основных статистических характеристик, таких как математическое ожидание или дисперсия, что в последствии, приведет к неверному принятию решений.

Поэтому такие данные требуют предварительной обработки. Более подробно о методах заполнения пропусков в данных рассмотрим в следующей главе.

2 Заполнение пропусков в матрице наблюдений процесса

С проблемой обработки пропусков в массивах данных можно столкнуться в самых разнообразных задачах статистического анализа. Самым простым способом решения данной проблемы является исключение наблюдений, содержащих пропуски, и проведение дальнейших анализов только на "полных" данных. Как говорилось ранее, такой подход приводит к сильному различию статистических выводов. Поэтому более перспективным является другой путь – предварительная обработка данных – заполнение пропусков. Однако данный подход так же имеет ряд недостатков, таких как:

- смещение полученных результатов от реальных;
- искажение структуры данных.

Очевидно, что такие модели так же будут менее точными по сравнению с моделями, построенными на полных наблюдениях. Поэтому очень важно выбрать «правильный» алгоритм заполнения пропусков. Выбор алгоритма зависит от множества факторов, однако, ключевыми факторами можно назвать знание механизма появления пропусков, а также структура входных-выходных данных. Сегодня создано множество методов восстановления пропусков, однако единая методология обработки подобных данных отсутствует, несмотря на ее необходимость, в данной главе приведем обзор существующих методов.

2.1 Механизмы возникновения пропусков в данных

Для того чтобы понять, как правильно обработать пропуски, необходимо определить механизмы их формирования.

Различают следующие 3 механизма формирования пропусков: MCAR, MAR, MNAR.

- MCAR (Missing Completely At Random) — механизм формирования пропусков, при котором вероятность пропуска для каждой записи набора

одинакова. Например, если проводился социологический опрос, в котором каждому десятому респонденту один случайно выбранный вопрос не задавался, причем на все остальные заданные вопросы респонденты отвечали, то имеет место механизм MCAR. В таком случае игнорирование/исключение записей содержащих пропущенные данные не ведет к искажению результатов.

- MAR (Missing At Random) — на практике данные обычно пропущены не случайно, а ввиду некоторых закономерностей. Пропуски относят к MAR, если вероятность пропуска может быть определена на основе другой имеющейся в наборе данных информации (пол, возраст, занимаемая должность, образование…), не содержащей пропуски. В таком случае удаление или замена пропусков на значение «Пропуск», как и в случае MCAR, не приведет к существенному искажению результатов.
- MNAR (Missing Not At Random) — механизм формирования пропусков, при котором данные отсутствуют в зависимости от неизвестных факторов. MNAR предполагает, что вероятность пропуска могла бы быть описана на основе других атрибутов, но информация по этим атрибутам в наборе данных отсутствует. Как следствие, вероятность пропуска невозможно выразить на основе информации, содержащейся в наборе данных.

Рассмотрим различия между механизмами MAR и MNAR на примере.

Люди, занимающие руководящие должности и/или получившие образование в престижном вузе чаще, чем другие респонденты, не отвечают на вопрос о своих доходах. Поскольку занимаемая должность и образование сильно коррелируют с доходами, то в таком случае пропуски в поле доходы уже нельзя считать совершенно случайными, то есть говорить о случае MCAR не представляется возможным.

Если в наборе данных есть информация об образовании и должности респондентов, то зависимость между повышенной вероятностью пропуска в графе доходов и этой информацией может быть выражена математически,

следовательно, выполняется гипотеза MAR. В случае MAR исключение пропусков вполне приемлемо.

Однако если информация о занимаемой должности и образовании у нас отсутствует, то тогда имеет место случай MNAR. При MNAR просто игнорировать или исключить пропуски уже нельзя, так как это приведет к значительному искажению распределения статистических свойств выборки.

2.2 Алгоритмы заполнения

Все методы заполнения пропусков можно классифицировать на простые и сложные (Рисунок 3).

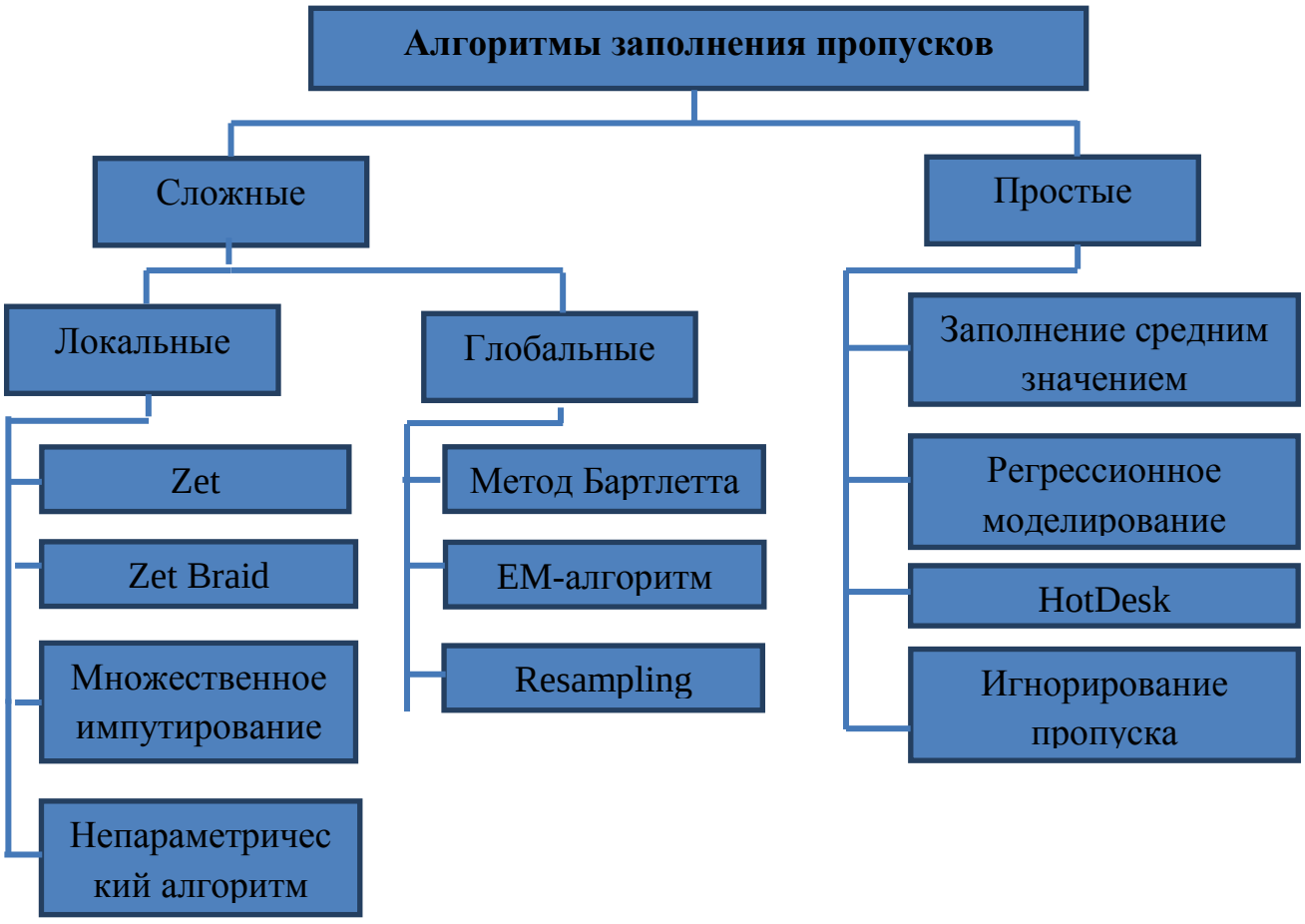


Рисунок 3 – Классификация алгоритмов

Простые методы – это неитеративные методы, основанные на простейших арифметических операциях.

Сложные методы – это итеративные методы, основанные на оценки точности подставляемого вместо пропуска значения. Они подразделяются на локальные и глобальные.

Глобальные методы – это методы, в которых оценивание любого из пропущенных значений участвуют все наблюдения рассматриваемой выборки.

Локальные методы - это методы, в которых оценивание любого из пропущенных значений участвуют наблюдения без пропусков, находящиеся в некоторой окрестности рассматриваемого наблюдения.

Рассмотрим каждый из данных метод подробнее.

Игнорирование объекта с пропущенным значением.

Простейшим методом решения проблемы пропущенных значений в выборке является отбрасывание объектов, имеющих пропуски. Данный метод следует применять только в том случае, когда в выборке присутствует небольшое число пропусков. Преимуществом данного подхода является простота и невозможность испортить данные путем замены пропусков. В случае достаточно большого размера выборки и небольшого количества пропусков метод может показывать хорошие результаты. Альтернативным вариантом в случае наличия пропусков в небольшом количестве признаков является удаление таких признаков из выборки.

Замена самым частым или средним значением.

Суть данного метода заключается в замене пропущенных значений на среднее арифметическое, рассчитанное на основе выборки, учитывая только заполненные наблюдения. Преимуществом этого метода по сравнению с предыдущим методом является то, что при расчете учитываются имеющиеся данные, однако имеется и ряд серьезных недостатков, таких как:

- не учитываются возможные корреляции между переменными;
- искажение распределения данных;

- уменьшении дисперсии.

Подбор внутри группы.

Из предположения о том, что близкие объекты по значениям среди заполненных признаков близки в признаках (гипотеза компактности), происходит выбор группы определенных наблюдений, схожих по определенному признаку.

Метод HotDesk.

По заданной метрике определяется ближайший к заполняемому объект по всем присутствующим значениям переменных, и пропуск заполняется значением соответствующей переменной ближайшего объекта.

Для заполнения пропуска выбирается тот объект матрицы, до которого расстояние от пропущенного значение минимально.

Регрессионное моделирование.

По комплектным данным строится уравнение линейной регрессии и вычисляются пропущенные значения переменной. Регрессионные коэффициенты для каждого из параметров находятся методом наименьших квадратов в массиве с полными данными. Подставляя значения параметров в регрессионное уравнение, получают прогноз пропущенного показателя. Недостатком данного метода является тот факт, что качество восстановления пропусков напрямую зависит от успешного выбора взятой за основу регрессионной модели.

Метод Бартлетта.

Данный метод состоит из двух этапов: подстановке вместо пропусков начальных значений на первом этапе и проведении на втором этапе ковариационного анализа искомой переменной. На 2 этапе используется индикатор полноты наблюдения, который показывает, есть ли в матрице наблюдений пропущенное значение. Индикатор полноты наблюдения равен 0, если значение не является пропусков, 1, если это значение пропущено.

EM – алгоритм.

Алгоритм представляет собой итерационную процедуру, предназначенную для решения задач оптимизации некоторого функционала, через аналитический поиск экстремума целевой функции.

На первом этапе E (expectation) по совокупности имеющихся абсолютно полных или частично (по искомой переменной) полных наблюдений рассчитываются условные ожидаемые значения искомой переменной для каждого неполного наблюдения. Затем после получения массива полных наблюдений, оцениваются основные статистические параметры: меры средней тенденции и разброса, показатели взаимной корреляции и ковариации переменных. В случае работы с неполными данными на E – этапе определяется функция условного математического ожидания логарифма полной функции правдоподобия при известном значении целевой переменной X.

На втором этапе M (maximization), задача алгоритма максимизировать степень взаимного соответствия ожидаемых и реально подставляемых данных, а также соответствия структуры импутированных данных структуре данных полных наблюдений.

Resampling.

Суть алгоритма заключается в том, что значения для пропусков выбираются из имеющихся случайным образом, с возвращением (когда значение может использоваться еще раз после выбора) или без него. После этого на всем массиве строится регрессионная модель, позволяющая предсказать значения для пробелов. Для всех предполагаемых предикторов находятся регрессионные коэффициенты и константа. Затем вычисляются итоговые значения регрессионных коэффициентов, по которым и будут предсказаны окончательные пропущенные значения.

Преимуществом данного метода является повторное использование исходных данных, ведь увеличение числа подвыборок позволяет наиболее полно использовать исходную информацию. С другой стороны, объем новой информации уменьшается для каждой новой подвыборки, так как увеличивается вероятность того, что данные элементы выборки были уже выбраны раньше, – это основной недостаток метода.

Множественное импутирование.

Данный метод предложен в 1970 Д.Рубиным. Его суть заключается в том, что для каждого пропущенного значения подставляется несколько возможных значений, разброс между которыми существенен. Данные каждого набора заполненных значений хранятся в собственном массиве, каждый из которых анализируется, как матрица наблюдений, не имеющая пропусков.

Недостатком является избыточность ненужной информации.

Zet-алгоритм.

Суть алгоритма заключается в подборе для каждого пропущенного значения не из всей совокупности полных наблюдений, а из некоторой ее части, называемой компетентной матрицей. Она состоит из компетентных строк и столбцов. Компетентность некоторой строки или объекта представляет собой величину обратно пропорциональную декартовому расстоянию до целевой строки (неполного наблюдения с пропуском) в пространстве, оси которого заданы переменными – рассматриваемыми характеристиками объектов. По данным компетентной матрицы затем строится функциональная зависимость прогнозируемого значения от соответствующего значения в компетентной матрице, на основе которой, затем, прогнозируется значение пропуска.

Непараметрический алгоритм. Для заполнения пропусков в матрицах наблюдений предлагается дать оценки выходной переменной x в незаполненных строках матрицы наблюдений (табл. 1) при известных значениях входных переменных u. Для восстановления пропусков используется выборка, состоящая только из результатов заполненных

строк матрицы наблюдений. В итоге мы получим заполненную матрицу, и оценки будем осуществлять уже на основании заполненной матрицы наблюдений. Рассмотри подробнее два последних алгоритма.

2.2.1 ZET-алгоритм

Пусть задана таблица входных данных $U = (u_{ij}),\ i = \overline{1,S}, j = \overline{1,m}$ (Таблица 2).

Таблица 2 - Матрица входных переменных процесса с пропусками

u			
u_1	u_2	…	u_m
u_{11}	u_{21}	…	u_{m1}
u_{12}	u_{22}	…	u_{m2}
u_{13}	u_{23}	…	u_{m3}
u_{14}	-	…	u_{m4}
…	…	…	…
u_{1s}	u_{2s}	…	u_{ms}

S – количество строк-объектов, m - количество столбцов-свойств, «-» - пропуск.

Ставится задача восстановления отсутствующих (пропущенных) значений «-» в таблице 2 .

В основе алгоритма лежат три гипотезы:

- гипотеза избыточности: предполагается, что в таблице 3 присутствует избыточность в строках (объекты могут быть похожи между собой) и столбцах (между свойствами могут быть зависимости). При отсутствии избыточности все

строки и столбцы имеют одинаковый вес при прогнозировании и смысл локальности алгоритма теряется;

- гипотеза аналогичности: предполагается, что если два объекта «похожи» по значениям *(m -1)* свойств, то они «похожи» и по *m* -му свойству;
- гипотеза локальной компетентности: предполагается, что избыточность строк и столбцов носит локальный характер, то есть для каждого пропущенного значения имеется только некоторое количество объектов – аналогов объекта с пропуском и свойств – аналогов свойства с пропуском. Поэтому предлагается использовать для прогнозирования только такие «компетентные» объекты и свойства, которые выбираются для каждого пропуска отдельно.

Основные этапы алгоритмы алгоритма ZET для отработки таблицы с *l* пропусками:

1.3 Предварительная обработка начальных данных;

1.4 Прогнозирование пропуска – выполняется *l* раз:

2.1 Формирование компетентной матрицы;

2.2 Подбор параметров модели прогнозирования;

2.3 Прогнозирование пропуска.

Рассмотрим подробнее каждый этап.

1. Вначале столбцы матрицы нормируются по дисперсиям для приведения различных свойств к единой шкале

$$u_{ij} = \frac{u^{ij} - \overline{u_j}}{G_j}, \qquad (1.12)$$

2. Следующие этапы выполняют *l* раз. Пусть координаты текущего элемента с пропуском *x,y;*

2.1 Формирование компетентной матрицы;

2.1.1 Задать размеры компетентной матрицы s_{ij}, $i=\overline{1,p}, j=\overline{1,q}$, $2<p<m$, $2<q<S$;

2.1.2 Выбрать *(p-1)* компетентных строк для строки с пропуском;

Компетентность L строки i по отношению к строке с пропуском определяется по формуле

$$L_{iy} = \frac{t_{iy}}{r_{iy}} \quad (1.13)$$

где t_{iy}- комплектность, то есть число значений известных для обеих строк i и y;

r_{iy}- декартово расстояние между строками (элементы с пропусками не учитываются).

Компетентная строка не должна содержать пропуска на x-й позиции.

2.1.3 Выбрать $(q-1)$ компетентных столбцов для столбца с пропуском.

Компетентность L столбца i по отношению к столбцу с пропуском x определяется по формуле

$$L_{iy} = |k_{ix}| \cdot t_{ix}, \quad (1.14)$$

где t_{ix} – комплектность столбцов i и x;

k_{ix} – коэффициент корреляции между столбцами i и x. При расчете k_{ix} используются только те значения столбцов, которые должны принадлежать к компетентным строкам. Компетентный столбец не должен содержать пропуск на y-й позиции.

2.2 Подбор параметров модели прогнозирования αr (по строкам) и αс (по столбцам) – коэффициенты, регулирующие влияние компетентности на результат предсказания.

2.2.1 Задаём пределы изменения коэффициентов αr и αс и шаг их изменения.

2.2.2 находим оптимальные коэффициенты αr и αс для прогноза пропуска по строкам и по столбцам по следующему алгоритму (одинаков для строк и столбцов). Подавая значения коэффициента α ($\alpha = \alpha r$ для строк, $\alpha = \alpha$с для столбцов) в указанных пределах и с указанным шагом минимизируем функцию

$$\sum_i |a_{ik} - b_{ik}| \to min, i \neq " - " \text{ (i – строка без пропуска)}, \quad (1.15)$$

где a_{ik} – реальное значение элемента *i* строки (столбца) *k* с пропуском;

b_{ik}- прогноз этого элемента с помощью компетентных строк (столбцов).

b_{ik}- рассчитывается по формуле

$$b_{ik} = \frac{\sum_{j=1}^{c-1} bl_{jk} \cdot L_{ij}^{\alpha}}{\sum_{j=1}^{c-1} L_{ij}^{\alpha}},$$

(1.16)

где $c=p$ для строк;

$c=q$ для столбцов;

bl_{jk}- прогноз для известных значений строки (столбца) с пропуском k с помощью i-й строки (столбца), рассчитывается с помощью линейной регрессии вида $y=ax+b$ по МНК.

2.3 Прогнозирование пропуска;

2.3.1 Прогнозирование пропуска по строкам выполняется по формуле

$$b_x = \frac{\sum_{i=1}^{q-1} bl_{ix} \cdot L_{ix}^{\alpha c}}{\sum_{i=1}^{q-1} L_{ix}^{\alpha c}} \quad (1.17)$$

2.3.2 Прогнозирование пропуска по столбцам выполняется по формуле

$$b_y = \frac{\sum_{i=1}^{p-1} bl_{iy} \cdot L_{iy}^{\alpha c}}{\sum_{i=1}^{q-1} L_{iy}^{\alpha r}} \quad (1.18)$$

2.3.3 Общий прогноз получается усреднением прогнозов по строкам и столбцам

$$b_{yx} = \frac{b_y + b_x}{2} \quad (1.19)$$

2.2.2 Непараметрический алгоритм заполнения пропусков

Пусть задана таблица «входных-выходных» переменных процесса.

Таблица 3 - Матрица "входных-выходных" переменных процесса с пропусками

u				x
u_1	u_2	...	u_m	
u_{11}	u_{21}	...	u_{m1}	-
u_{12}	u_{22}	...	u_{m2}	x_2
u_{13}	u_{23}	...	u_{m3}	x_3
u_{14}	u_{24}	...	u_{m4}	-
...	...	...	...	...
u_{1s}	u_{2s}	...	u_{ms}	x_s

В таблице 3 столбцы представляют собой входные (u_m) и выходные (x_s) переменные процесса, строки – наблюдения, «-» - пропуски, *S*-объем выборки. Принято, что дискретность измерения выходной переменной *x(t)* в три раза больше дискретности измерения входной переменной *u(t)* ($\Delta T = 3\Delta t$).

Методику заполнения пропусков матрицы наблюдений можно разделить на 3 этапа.

На 1 этапе восстанавливается функция регрессии x_s (1.20) по наблюдениям *u* по полностью заполненным строкам. Настраивается оптимальное значение коэффициента размытости *Cs*.

$$x_s(u) = \frac{\sum_{i=1}^{S} x_i \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{Cs}\right)}{\sum_{i=1}^{S} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{Cs}\right)}, \quad (1.20)$$

где $\Phi(\cdot)$ – ядерная колокообразная функция;

Cs – коэффициент размытости ядра.

Коэффициент размытости Cs определяется путем минимизации квадратичного критерия выхода объекта x_i и выхода модели $\hat{x}_i$, основанного на скользящем экзамене:

$$I = \frac{1}{S}\sum_{i=1}^{S}(x_i - \hat{x}_i)^2 \to \min_{\alpha} \quad (1.21)$$

На втором этапе происходит заполнение пропусков матрицы с использованием оценки x_s и оптимального параметра размытости Cs, полученных на предыдущем этапе. Там, где наблюдения *x* пропущены, в оценку $x_s(u_1, u_2, \dots, u_m)$ подставляем значения измеренных $u = (u_1, u_2, \dots, u_m)$ и вычисляем соответствующую оценку x_s, которой восполняем недостающее наблюдение *x*. После этого матрица принимает следующий вид (Таблица 4), где $(X_{s1}, X_{s2}, \dots, X_s)$ – восстановленные значения пропусков.

.

Таблица 4 - Матрица наблюдений с заполненными значениями

u				*x*
u_1	u_2	…	u_m	
u_{11}	u_{21}	…	u_{m1}	x_{s1}
u_{12}	u_{22}	…	u_{m2}	x_2
u_{13}	u_{23}	…	u_{m3}	x_3
u_{14}	u_{24}	…	u_{m4}	x_{s4}
…	…	…	…	…
u_{1s}	u_{2s}	…	u_{ms}	x_s

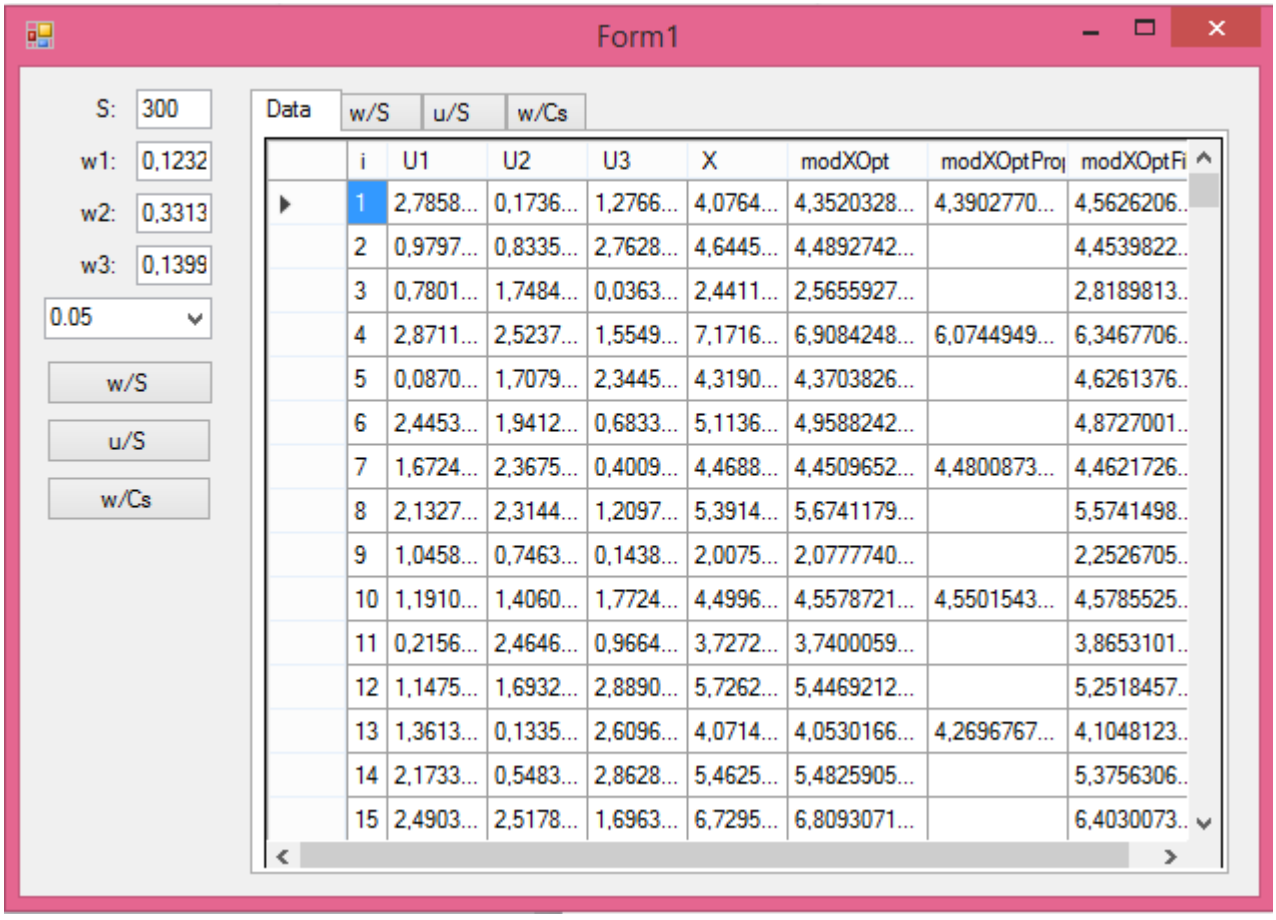
3 этап заполнения пропусков состоит в построении модели. Используется непараметрическая функция оценки регрессии (1.20) по всей заполненной матрице наблюдений (Таблица 4). При этом параметр размытости Cs настраивается по всей выборке еще раз.

3 Результаты и исследование алгоритмов заполнения пропусков

Рассматриваемые алгоритмы были реализованы на языке программирования C# с помощью интерфейса программирования приложений Windows Form в среде разработке Microsoft Visual Studio 2013.

Интерфейс программы для непараметрического алгоритма представлен на рисунках 4,5,6,7. Пользователь задаёт объем выборки S и уровень помехи. В полях w1, w2, w3 отображаются рассчитанные ошибки аппроксимации по полной матрице, по заполненной матрице, исключая наблюдения с пропусками, а так же по восстановленной матрице.

На вкладке «Data» отображаются сгенерированные входные параметры U, выход объекта X, рассчитанное значение модели modXOptProp для каждого 3 пропуска, а так же оценки modXOptFill по восстановленной матрице.



i	U1	U2	U3	X	modXOpt	modXOptProj	modXOptFi
1	2,7858...	0,1736...	1,2766...	4,0764...	4,3520328...	4,3902770...	4,5626206..
2	0,9797...	0,8335...	2,7628...	4,6445...	4,4892742...		4,4539822..
3	0,7801...	1,7484...	0,0363...	2,4411...	2,5655927...		2,8189813..
4	2,8711...	2,5237...	1,5549...	7,1716...	6,9084248...	6,0744949...	6,3467706..
5	0,0870...	1,7079...	2,3445...	4,3190...	4,3703826...		4,6261376..
6	2,4453...	1,9412...	0,6833...	5,1136...	4,9588242...		4,8727001..
7	1,6724...	2,3675...	0,4009...	4,4688...	4,4509652...	4,4800873...	4,4621726..
8	2,1327...	2,3144...	1,2097...	5,3914...	5,6741179...		5,5741498..
9	1,0458...	0,7463...	0,1438...	2,0075...	2,0777740...		2,2526705..
10	1,1910...	1,4060...	1,7724...	4,4996...	4,5578721...	4,5501543...	4,5785525..
11	0,2156...	2,4646...	0,9664...	3,7272...	3,7400059...		3,8653101..
12	1,1475...	1,6932...	2,8890...	5,7262...	5,4469212...		5,2518457..
13	1,3613...	0,1335...	2,6096...	4,0714...	4,0530166...	4,2696767...	4,1048123..
14	2,1733...	0,5483...	2,8628...	5,4625...	5,4825905...		5,3756306..
15	2,4903...	2,5178...	1,6963...	6,7295...	6,8093071...		6,4030073..

Рисунок 4 – Интерфейс программы для непараметрического алгоритма

На вкладке «w/S» строится график зависимости ошибки от объема выборки.

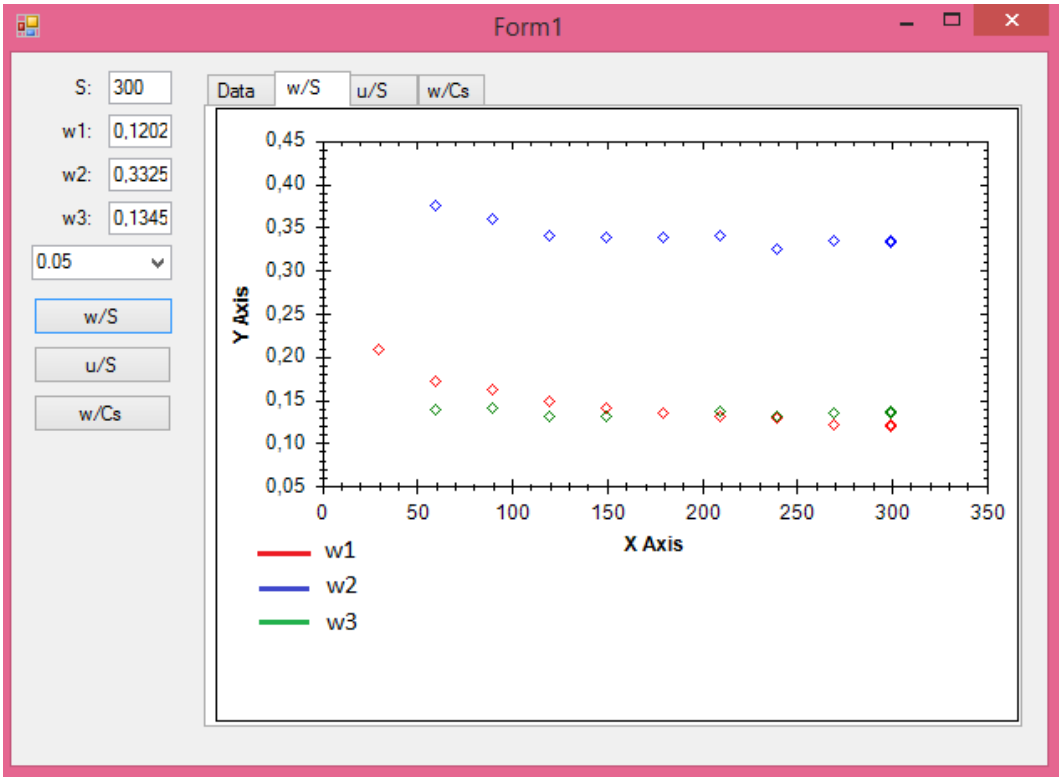


Рисунок 5 – Зависимость ошибки от выборки

На вкладке «u/S» строится график выхода объекта и выхода модели.

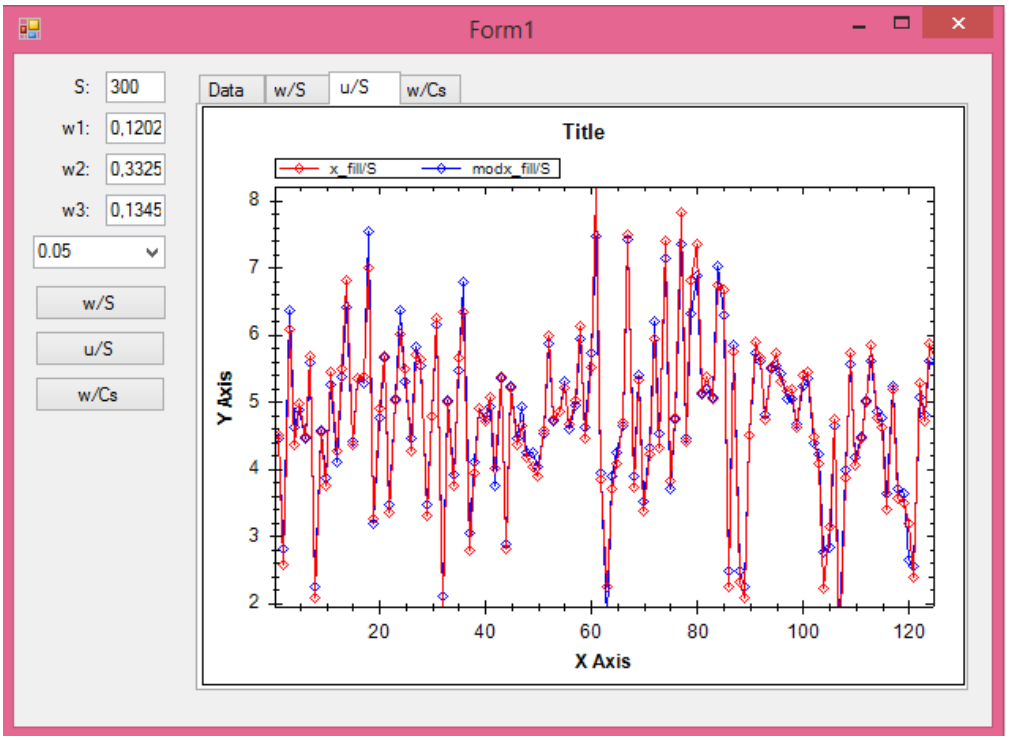


Рисунок 6 – Выход объекта и модели

На вкладке «w/Cs» строится график зависимости ошибки от коэффициента размытости Cs.

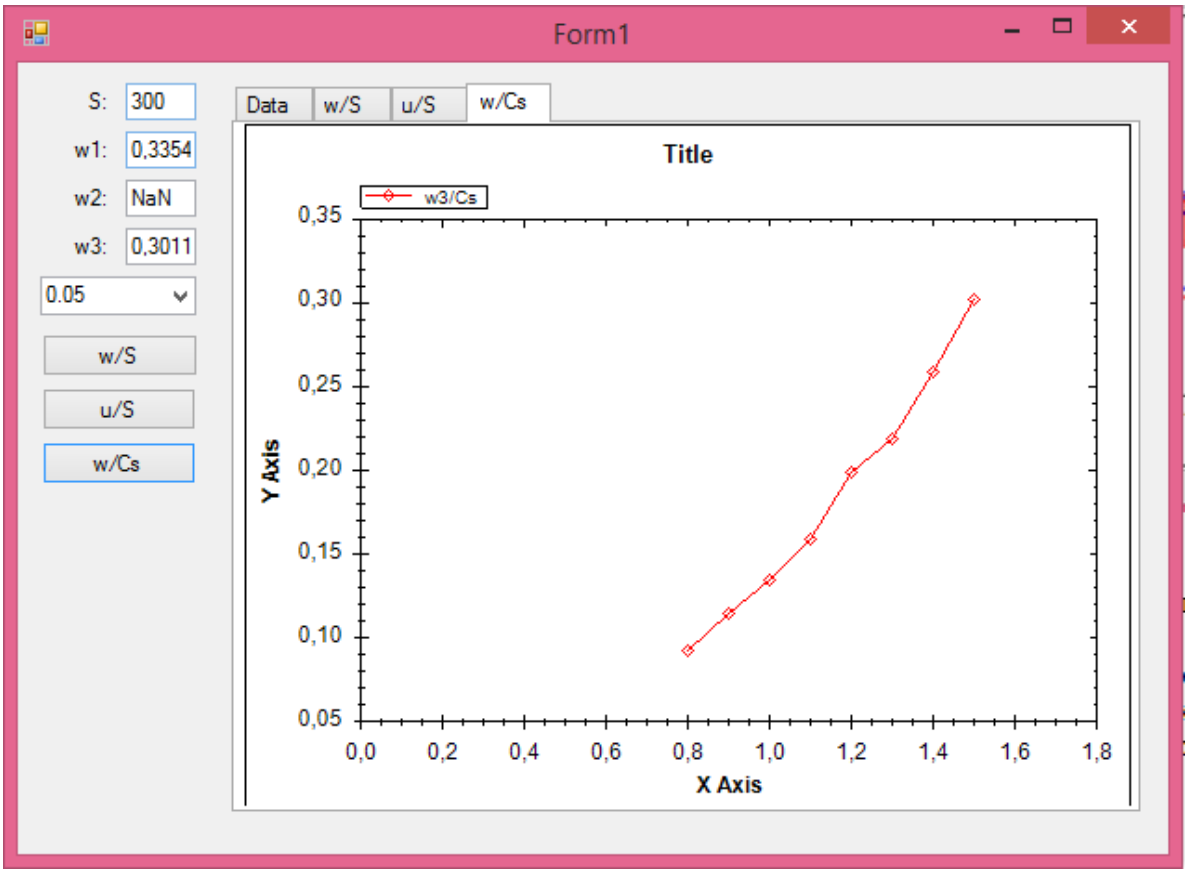


Рисунок 7 – Зависимость ошибки от параметра Cs

Интерфейс программы для Zet-алгоритма представлен на рисунке 8. Пользователь задаёт объем выборки S, количество входных параметров U, а так же размерность компетентной матрицы. В таблице отображены данные по U, в ListBox выводятся найденные алгоритмом компетентные строки и столбцы, а так же прогнозируемое значение пропуска, в поле w отображается рассчитанная ошибка по восстановленному значению.

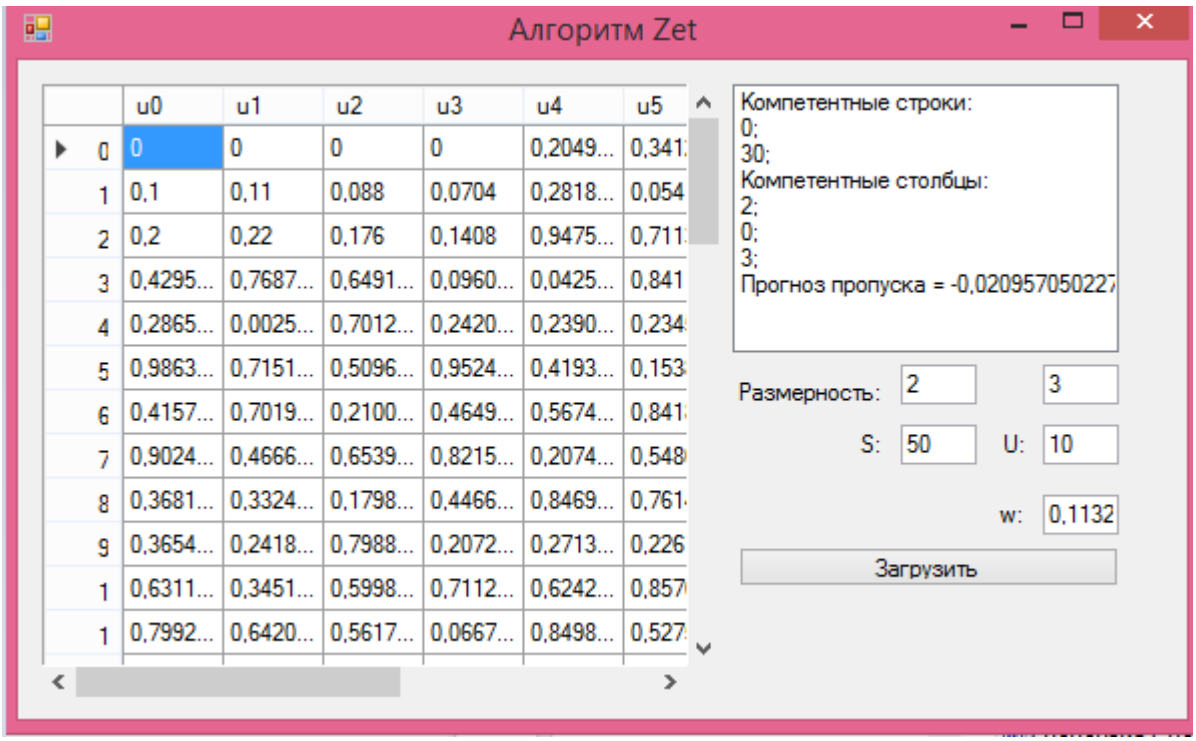


Рисунок 8 - Интерфейс программы для Zet-алгоритма

1.2 Входные данные

В качестве входных данных были сгенерированы различные выборки, описывающиеся 6 признаками $u = (u_1, u_2, u_3, u_4, u_5, u_6)$. Выход объекта *x*.

Непараметрический алгоритм.

Все локальные алгоритмы основаны на гипотезе компактности. Поэтому в случае непараметрики обеспечим похожесть на выходе.

Входные параметры *u* независимы друг от друга и генерируются случайным образом. На объект так же накладывается помеха (10%, 20%). Структура объекта задана следующим уравнением

$$x = u_1 + 0.5u_2 + \sin u_3 + 2\sqrt{u_4} + 0.5u_5^2 + \frac{u_6}{2} \qquad (1.22)$$

Zet-алгоритм.

Случай 1.

Входные параметры *u* генерируются случайным образом

Случай 2.

Между всеми столбцами *u* присутствует следующая зависимость

$$\begin{cases} u_1 = Rand; \\ u_2 = 0.5 * u_1; \\ u_3 = 0.9 * u_2; \\ u_4 = u_3 * 0.9; \\ u_5 = u_4 * 1.1; \\ u_6 = u_5 + 0.5. \end{cases} \quad (1.23)$$

Случай 3.

Между всеми строками присутствует зависимость

$$\begin{cases} u_{i1} = Rand[0,3], i = \overline{1,6} - \text{ первая строка}, \\ u_{ij} = u_{i,j-1} * 0.9, \text{j} = \overline{1,\text{S}} - \text{последующие строки} \end{cases} \quad (1.24)$$

Случай 4.

Матрица входных параметров u имеет ярко выраженную компетентную матрицу (компетентные столбцы - u_1, u_2, u_3, u_4, компетентные строки – 1,2,3). Связь между 1,2 и 3 строками

$$\begin{cases} u_{1i} = 0{,}1 \cdot i \\ u_{2i} = u_1 \cdot 1.1 \\ u_{3i} = u_2 \cdot 1.1 \\ u_{4i} = u_3 \cdot 0.9, \end{cases} \quad (1.25)$$

где $i = \overline{1,3}$.

$$\begin{cases} u_{4i}, u_{5i} = Rand; \\ u_{1j,} u_{2j}, u_{3j}, u_{4j} = Rand' \end{cases} \quad (1.26)$$

где $i = \overline{1,S}$;

$j = \overline{3,S}$.

	u0	u1	u2	u3	u4	u5
0	1,5173...	0,7586...	0,3793...	0,1896...	0,4355...	0,1896...
1	0,3195...	0,1597...	0,0798...	0,0399...	0,1998...	0,0399...
2	1,6587...	0,8293...	0,4146...	0,2073...	0,4553...	0,2073...
3	2,6542...	1,3271...	0,6635...	0,3317...	0,5760...	0,3317...
4	2,4381...	1,2190...	0,6095...	0,3047...	0,5520...	0,3047...
5	1,9151...	0,9575...	0,4787...	0,2393...	0,4892...	0,2393...
6	1,1408...	0,5704...	0,2852...	0,1426...	0,3776...	0,1426...
7	2,1778...	1,0889...	0,5444...	0,2722...	0,5217...	0,2722...
8	2,4906...	1,2453...	0,6226...	0,3113...	0,5579...	0,3113...
9	0,0306...	0,0153...	0,0076...	0,0038...	0,0618...	0,0038...
1	2,0541...	1,0270...	0,5135...	0,2567...	0,5067...	0,2567...
1	0,6944...	0,3472...	0,1736...	0,0868...	0,2946...	0,0868...

Рисунок 9 – Матрица наблюдений

1.3 Результаты исследований

1.3.1 Непараметрический алгоритм

Так как результат работы алгоритма зависит от входных параметров, а именно от задаваемого уровня помехи и объема выборки, рассмотрим результаты данного метода при различных значениях данных параметров.

На рисунке 9 представлена зависимость ошибки моделирования w от объема выборки S, где w_1 – ошибка по полной матрице, w_2 – ошибка по заполненной матрице, исключая пропуски, w_3– ошибка по восстановленной матрице.

Рассмотрим результаты при объеме выборке S=150 и уровне помехи 0%, ошибка по заполненной матрице, исключая пропуски w_2 = 37%, в то время как оценивание по заполненной матрице дает ошибку $w_3 = 11\%$.

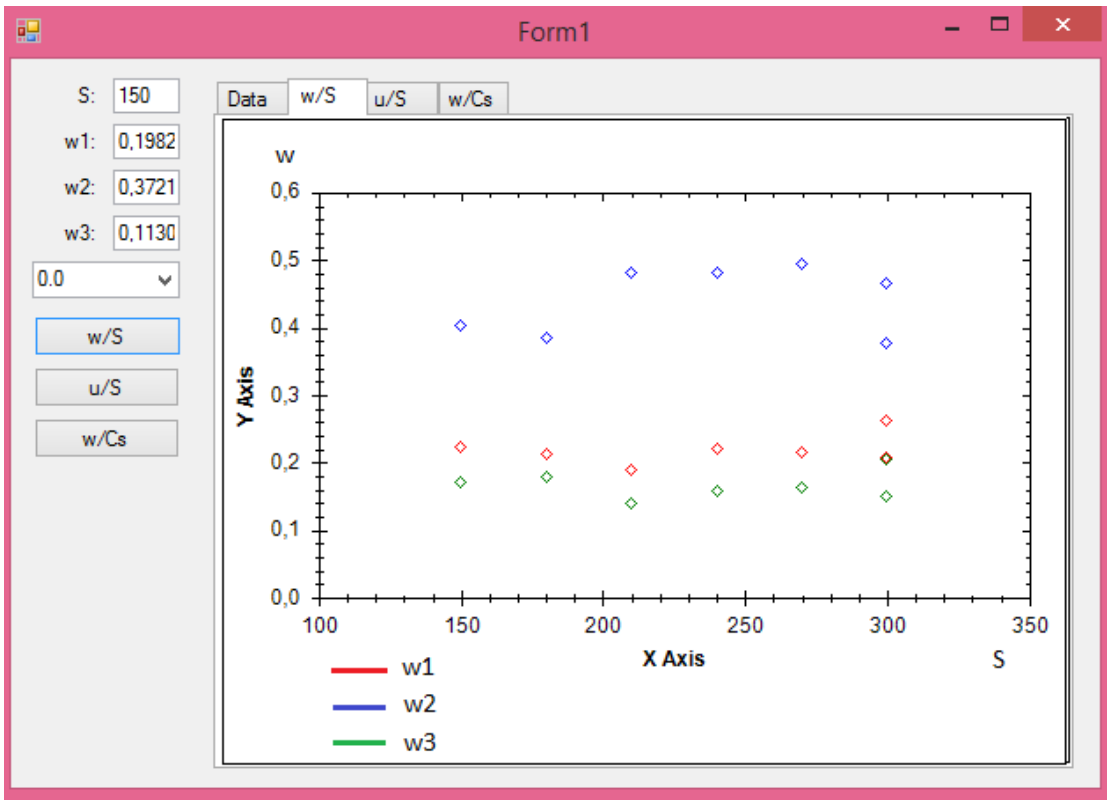


Рисунок 10 – Зависимость ошибки от объема выборки S=150, помеха = 0%

Для объема выборки S=100 и уровне помехи 10%, ошибка исходной матрице w_2 = 56% , по заполненной матрице w_3 = 20%.

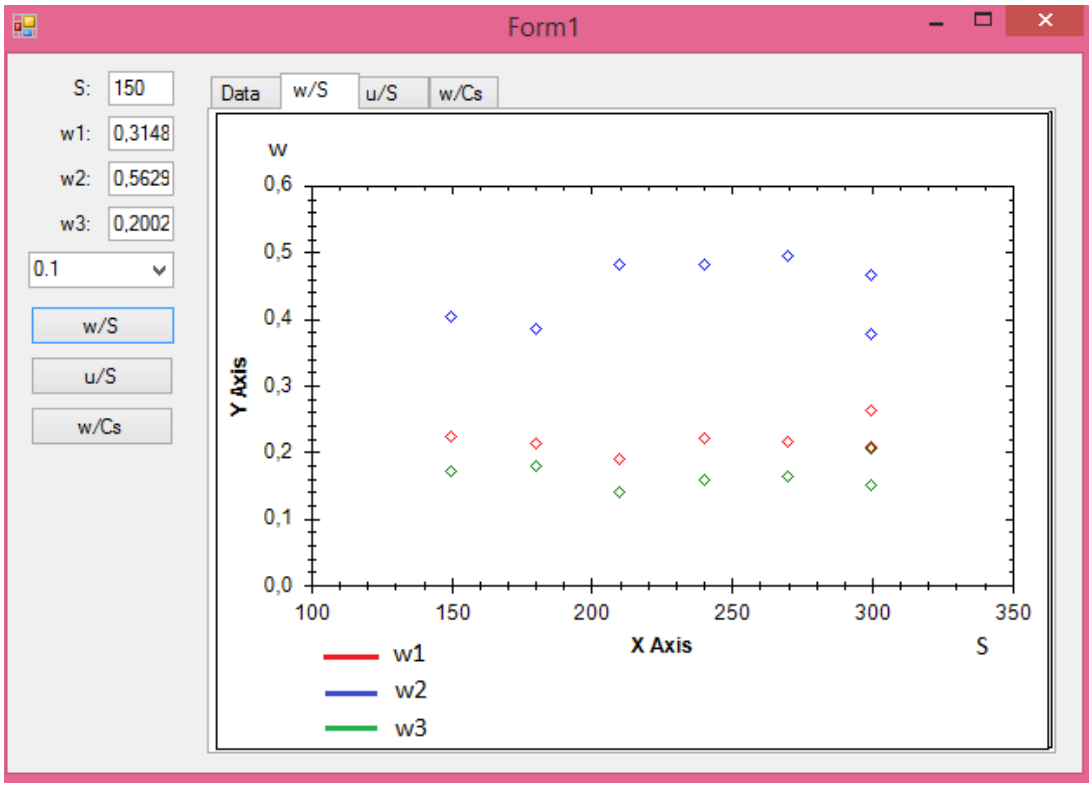


Рисунок 11 – Зависимость ошибки от объема выборки S=150, помеха = 10%

При увеличении уровня помехи до 20% и S=100 ошибка по матрице с пропусками составляет w_2 = 66, в то время как оценивание по восстановленной матрице составляет $w_3 = 21\%$.

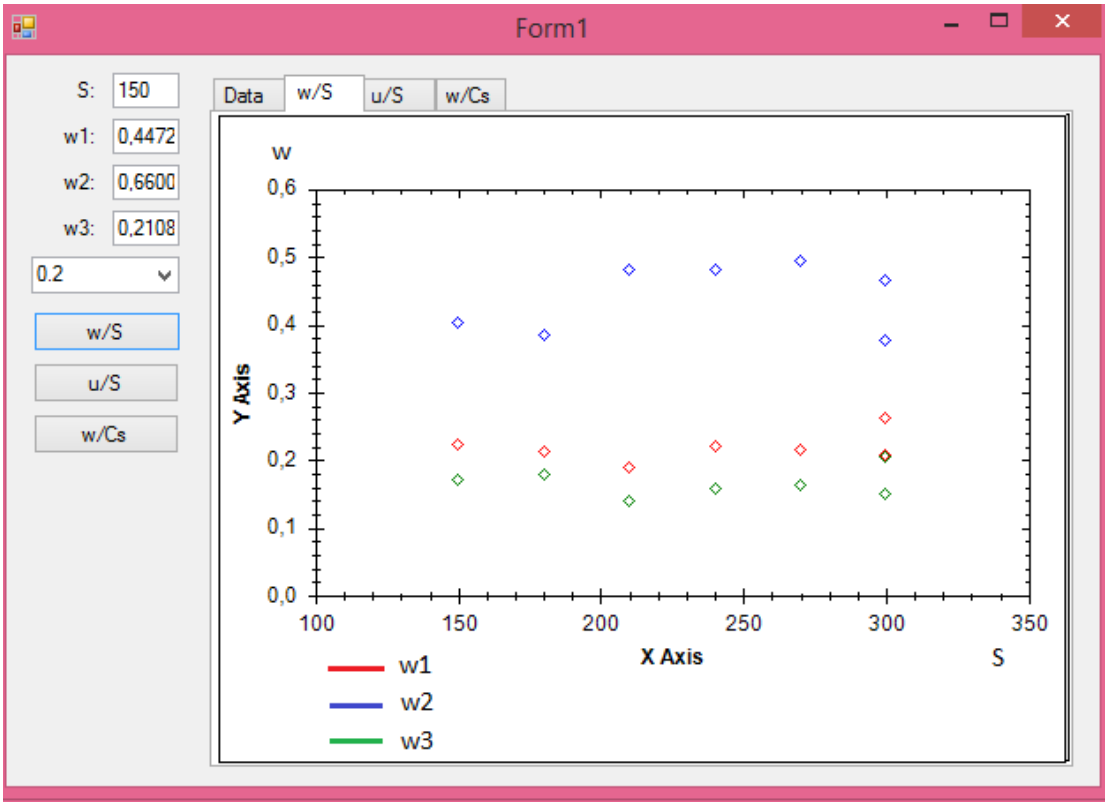


Рисунок 12 – Зависимость ошибки от объема выборки S=150, помеха = 20%

При увеличении выборки S до 300 и уровне помехи 0% при исходной матрице с пропусками ошибка w_2 составляет 37%, в то время как оценивание по заполненной матрице дает ошибку $w_3 = 13\%$.

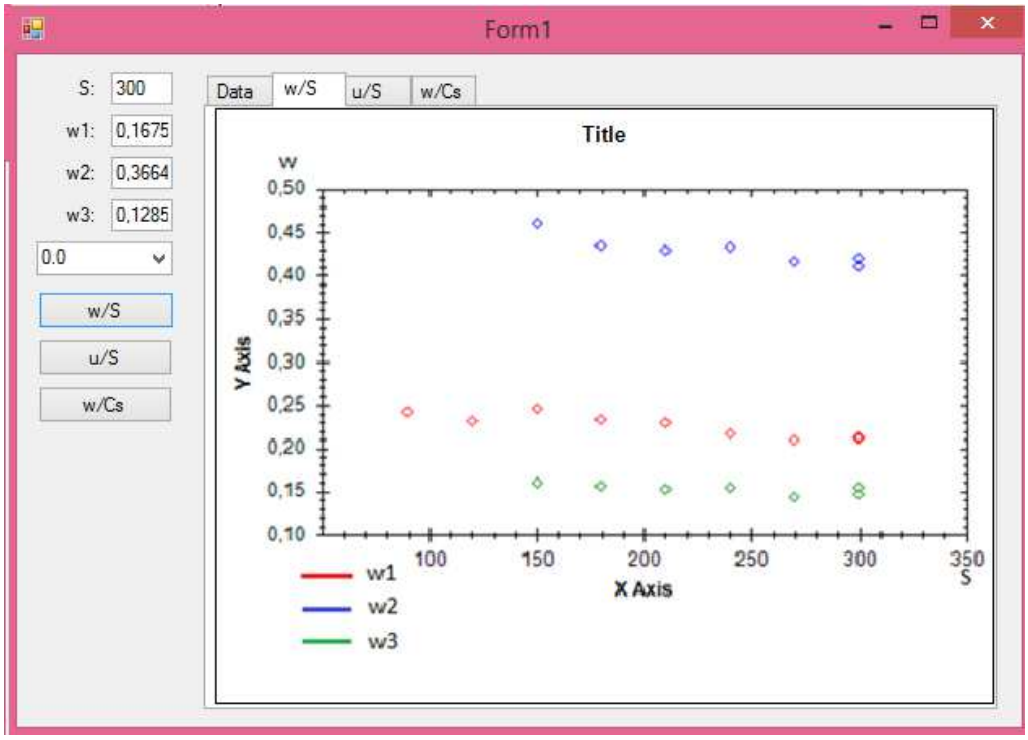


Рисунок 13 – Зависимость ошибки от объема выборки S=300, помеха = 0%

Для объема выборки S = 300 и уровне помехи 10% ошибка по исходной матрице с пропусками w_2 составляет 45%, в то время как оценивание по восстановленной матрице дает ошибку $w_3 = 12\%$.

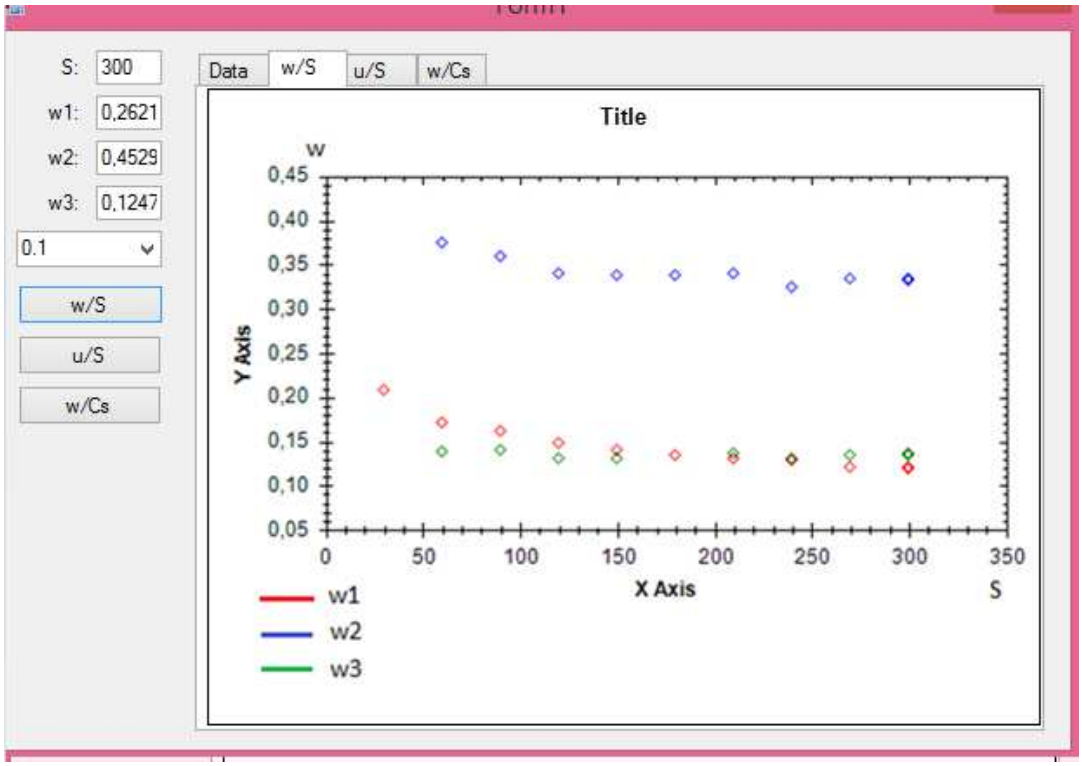


Рисунок 14 – Зависимость ошибки от объема выборки S=300, помеха = 10%

При увеличении уровня помехи до 20% и той же выборке S = 300, получаем следующие результаты, w_2 = 57%, w_3 = 13%.

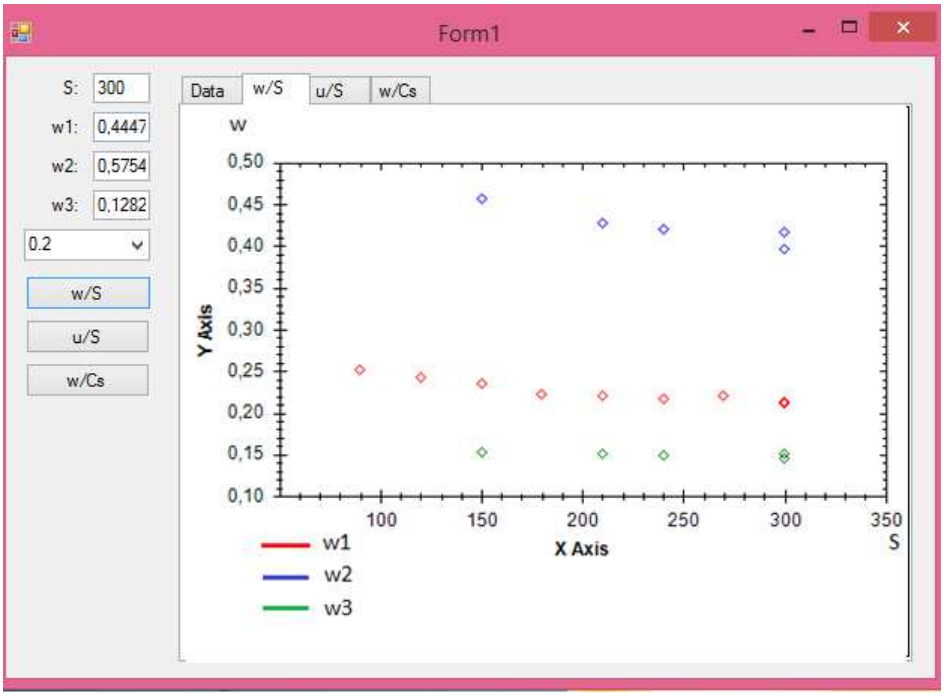


Рисунок 15 – Зависимость ошибки от объема выборки S=300, помеха = 20%

При увеличении выборки S до 450 и уровне помехи 0% при исходной матрице с пропусками ошибка w_2 составляет 38%, в то время ошибка по заполненной матрице w_3 = 15%.

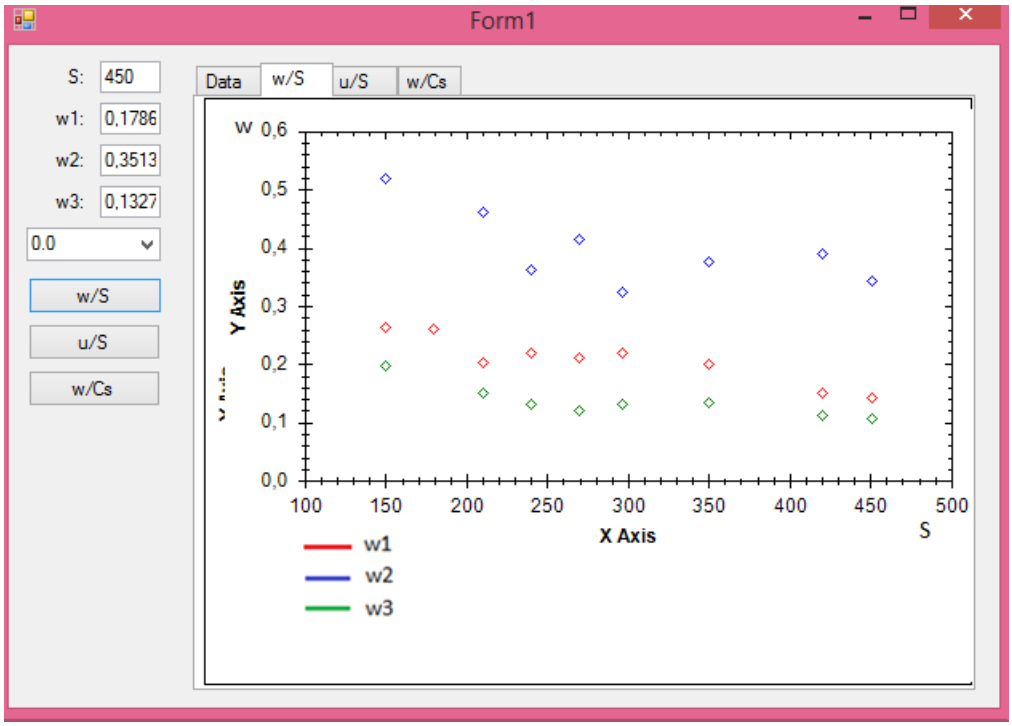


Рисунок 16 – Зависимость ошибки от объема выборки S=450, помеха = 0%

При том же объеме выборки S=450 и уровне помехи 10% получаем $w_2 =$ 23% и $w_3 = 14\%$.

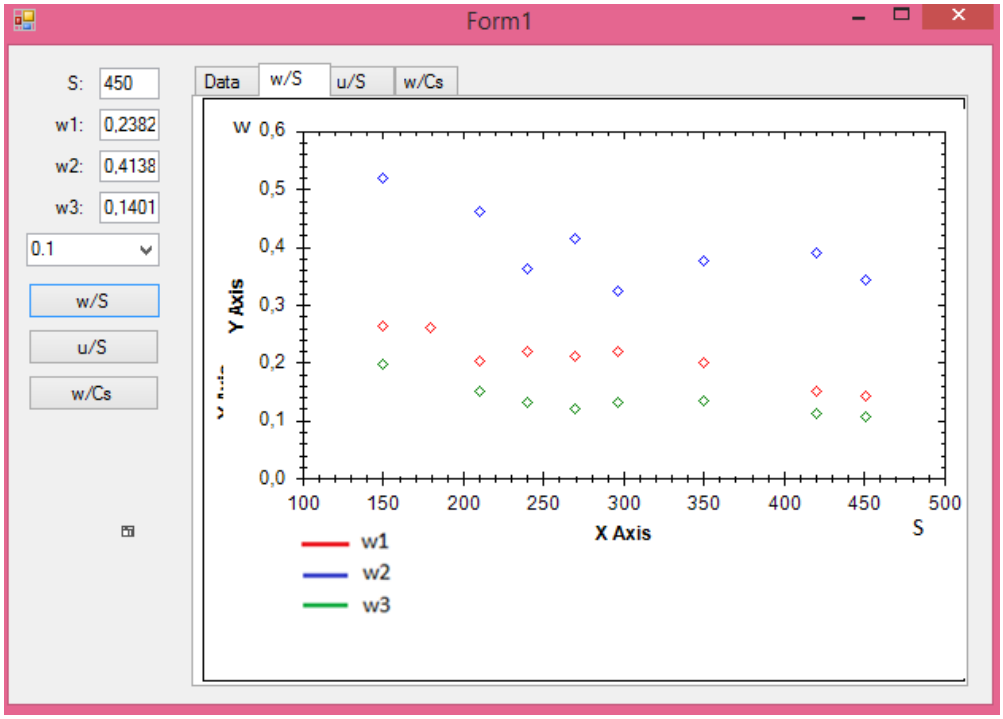


Рисунок 17 – Зависимость ошибки от объема выборки S=450, помеха = 10%

При S=450, уровне помехи 20% ошибка по исходной матрице с пропусками составляет $w_2 =$ 38%, после восстановления пропущенных значений ошибка $w_3 = 14\%$.

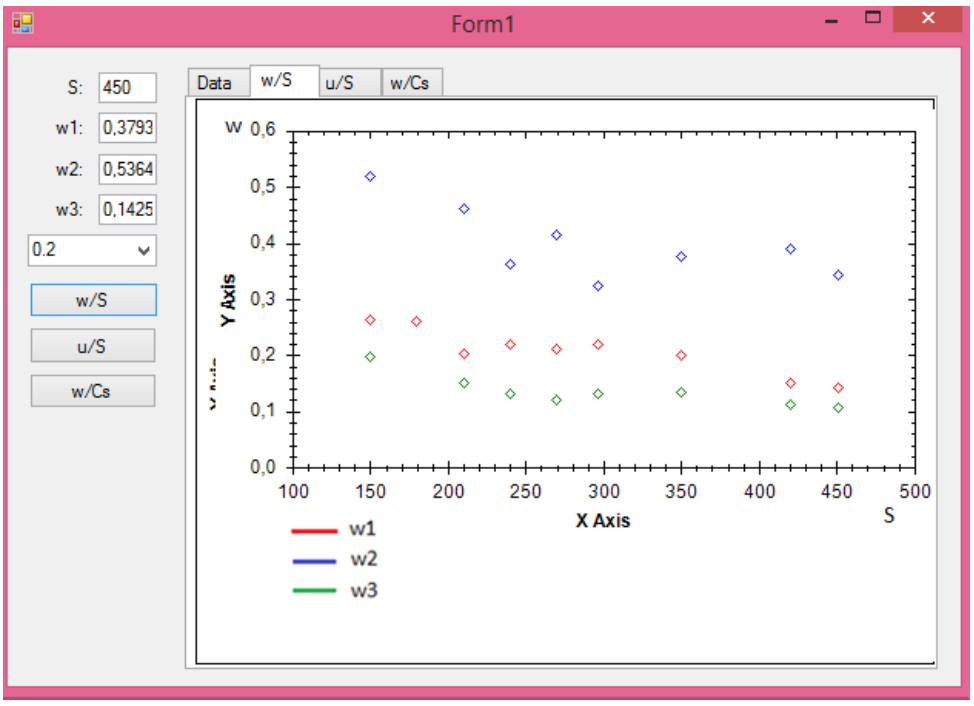


Рисунок 18 – Зависимость ошибки от объема выборки S=450, помеха = 20%

На рисунке 18 представлен график зависимости ошибки от коэффициента размытости Cs. Из графика видно, что при увеличении параметра Cs увеличивается и ошибка.

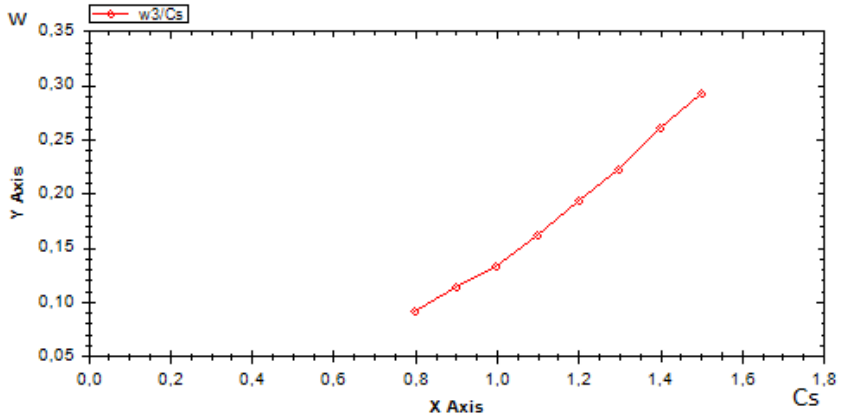


Рисунок 19 – Зависимость ошибки от коэффициента размытости Cs

На следующих рисунках представлены зависимости выхода объекта и модели по восстановленной матрице от выборки при различных уровнях помех (0%, 10%, 20%) и S=300. По графикам видно, что разница между выходом объекта и выходом модели небольшая, с увеличением уровня помехи увеличивается и ошибка.

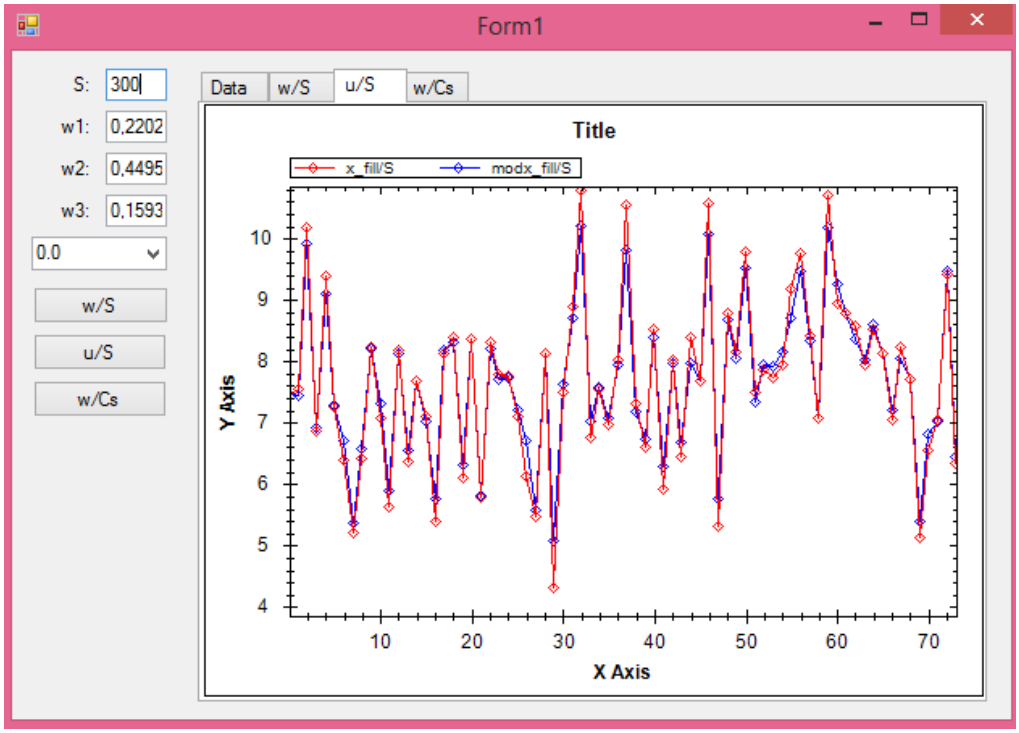


Рисунок 20 – Графики выхода объекта и модели при помехе 0%

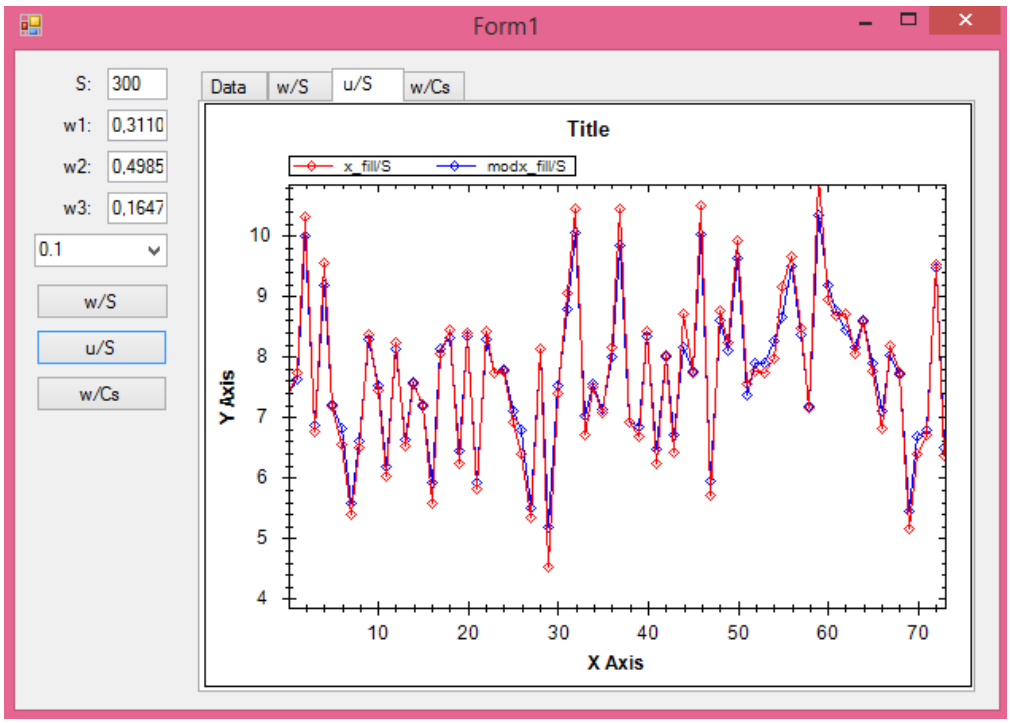


Рисунок 21 – Графики выхода объекта и модели при помехе 10%

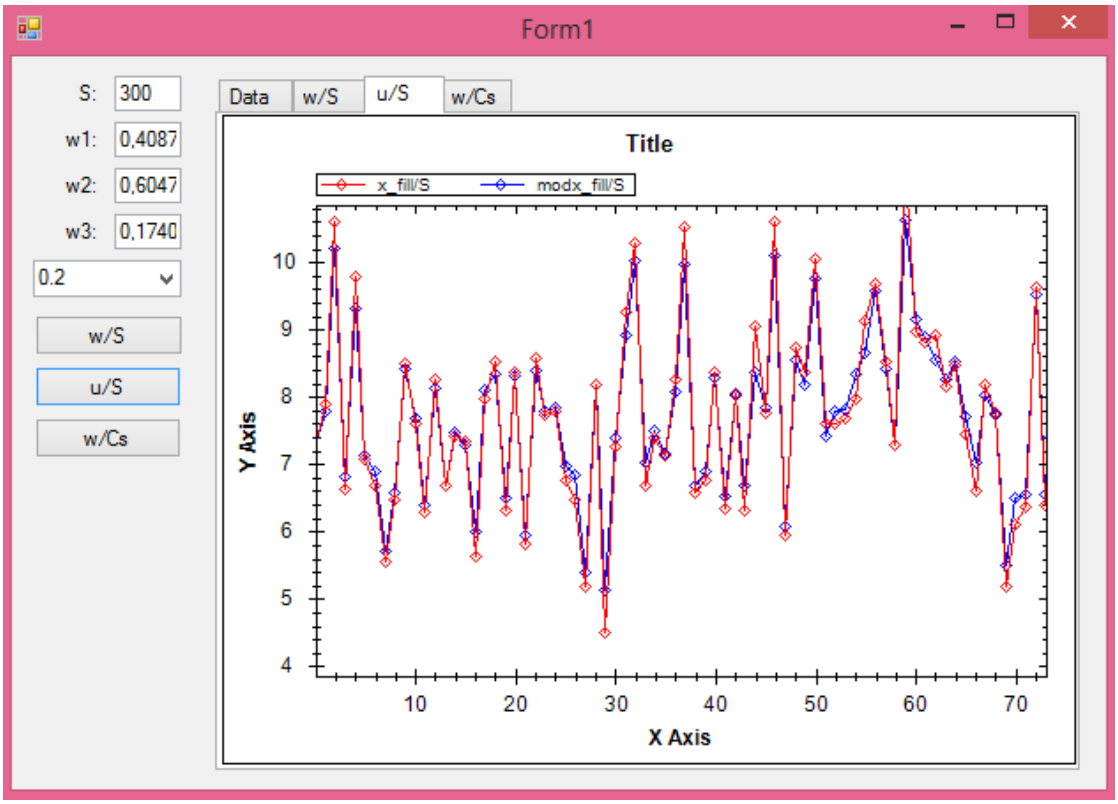


Рисунок 22 – Графики выхода объекта и модели при помехе 20%

Выводы

По представленным выше результатам, была построена сводная таблица результатов работы непараметрического алгоритма заполнения пропусков при

различных входных параметрах, а именно при разном объеме выборки S и уровне помехи. Исходя из таблицы, можно сделать вывод о том, что заполнение пропусков с помощью данного алгоритма приводит к повышению качества работы модели. Ошибка моделирования после заполнения сокращается в среднем на 30%.

Таблица 5 – Результаты работы непараметрического заполнения пропусков

Алгоритм	Объем выборки	Уровень помехи, %	Ошибка,%		
			w_1	w_2	w_3
Непараметрический	150	0	20	37	11
		10	31	56	20
		20	45	66	21
	300	0	16	37	13
		10	26	45	12
		20	44	57	13
	450	0	18	35	13
		10	23	41	14
		20	38	54	14

1.3.2 Zet-алгоритм

Рассмотрим случай 1, когда входные параметры сгенерированы случайным образом от 0 до 1. Объем выборки S = 6. Размер компетентной матрицы зададим 3x3. Красным цветом выделена ячейка, в которой имеется пропуск.

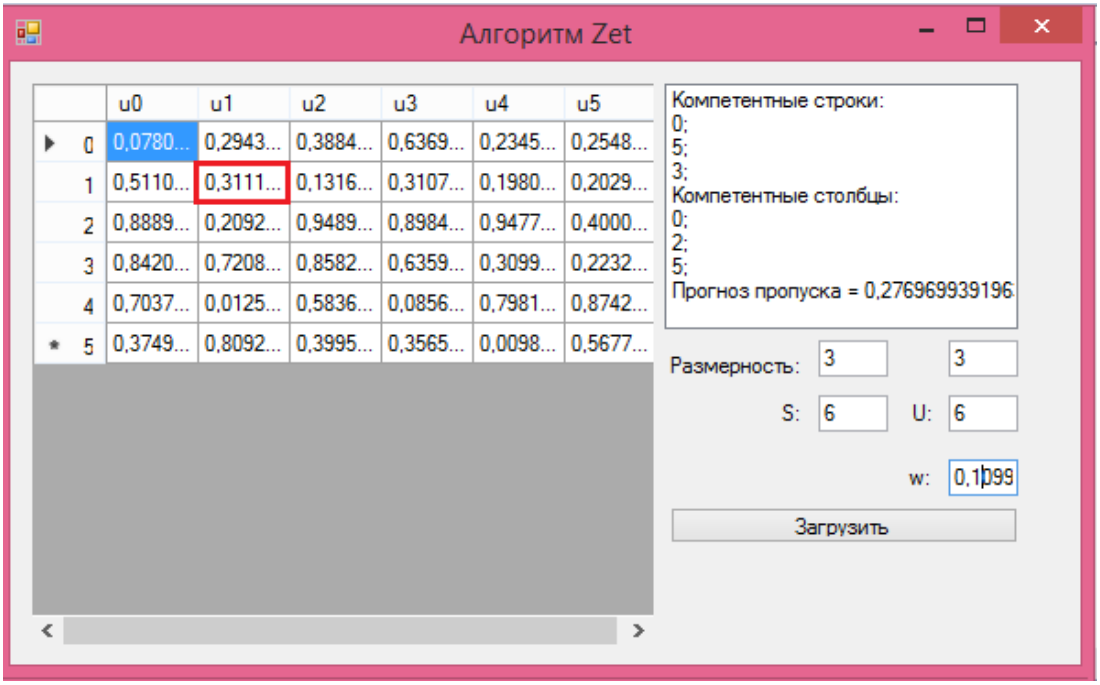


Рисунок 23 – Результат работы Zet-алгоритма в случае 1 при U = Rand[0,1]

В таблицах 5 и 6 представлены найденные алгоритмом компетентные строки столбцы по отношению к строке и столбцу с пропуском.

Таблица 5 - Компетентные строки

строка с пропуском						№
0,51107	0,31120	0,13165	0,31072	0,19800	0,20291	1
компетентные строки						
0,07807	0,29436	0,38843	0,63693	0,23455	0,25488	0
0,84203	0,72088	0,85824	0,63598	0,30994	0,22327	3
0,37497	0,80925	0,39955	0,35652	0,00990	0,56779	5

Таблица 6 - Компетентные столбцы

столбец с пропуском	Компетентные столбцы		
1	0	2	5
0,29436	0,07807	0,38843	0,25488
0,31120	0,51107	0,13165	0,20291
0,20921	0,88891	0,94898	0,40008
0,72088	0,84203	0,85824	0,22327
0,01257	0,70378	0,58370	0,87423
0,80925	0,37497	0,39955	0,56779

Как видно из рисунка 23 алгоритм неплохо справился с нахождением компетентных строк и столбцов, в результате, восстановленное значение $u'_{11} = 0{,}276$, когда искомое значение $u_{11} = 0{,}311$. Ошибка w = 11%.

Однако, если увеличить диапазон генерирования входных параметров случайными числами, например, от 0 до 5, результаты заметно ухудшились. Искомое значение $u_{11} = 4{,}891$, восстановленное значение $u'_{11} = 2{,}624$. Ошибка w = 44%.

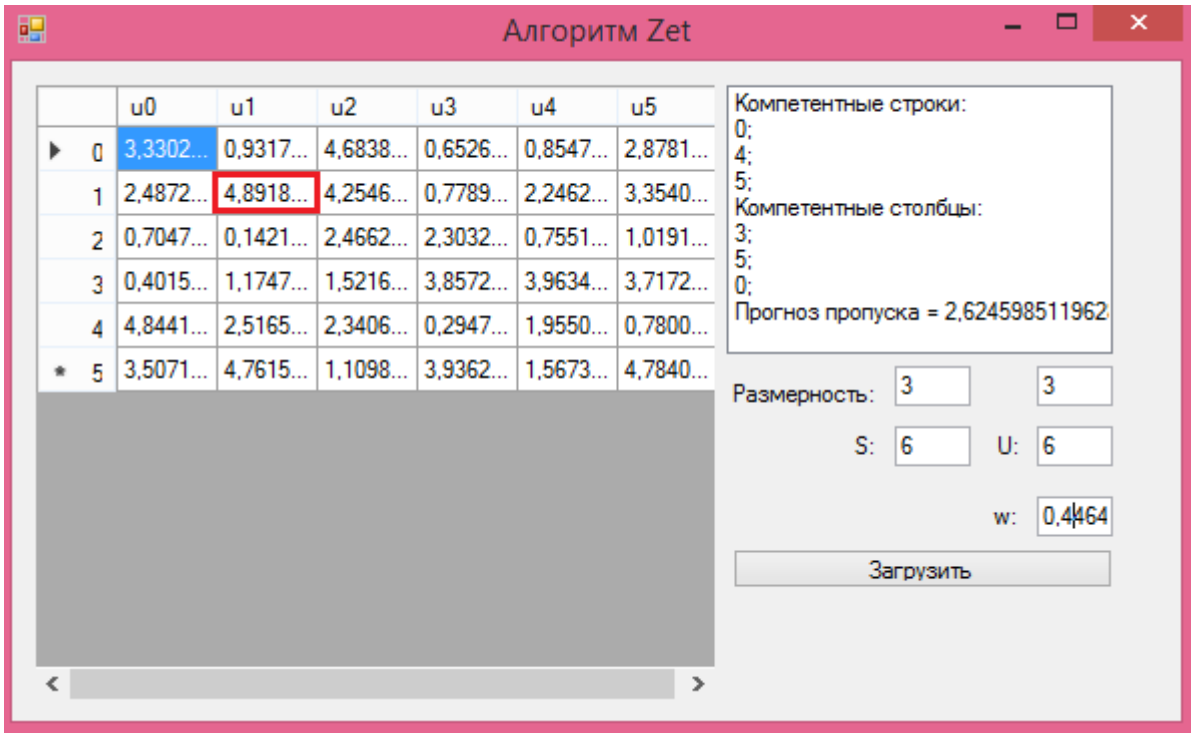


Рисунок 24 – Результат работы Zet-алгоритма в случае 1 при U = Rand[0,5]

Рассмотрим случай 2, в котором обеспечивается похожесть по всем столбцам, когда имеется некоторая зависимость между ними (1.23). $u_{11} = 0{,}576$, восстановленное значение $u'_{11} = 0{,}431$. Ошибка существенно уменьшилась w = 12%.

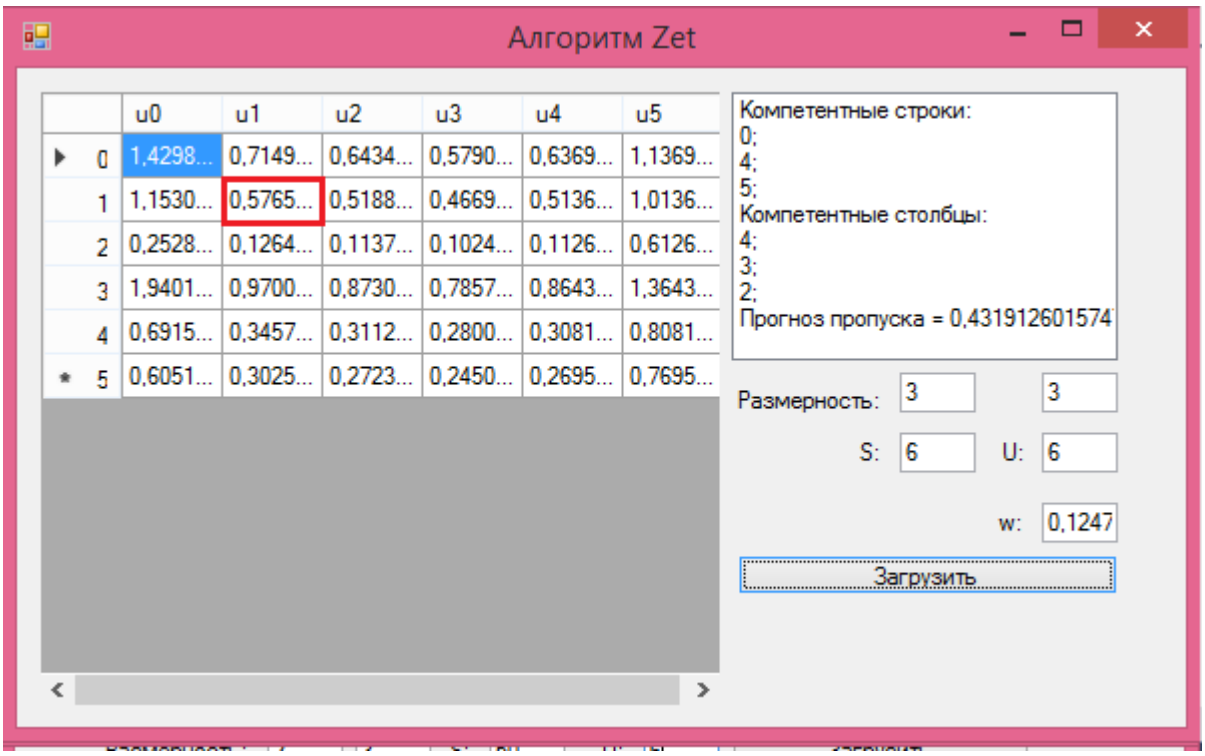


Рисунок 25 – Результат работы Zet-алгоритма в случае 2

В случае 3, когда похожесть обеспечивается по всем строкам зависимостью (1.24). Алгоритм так же показал неплохие результаты $u_{11} = 1{,}2834$, восстановленное значение $u'_{11} = 1{,}025$. Ошибка w = 20%.

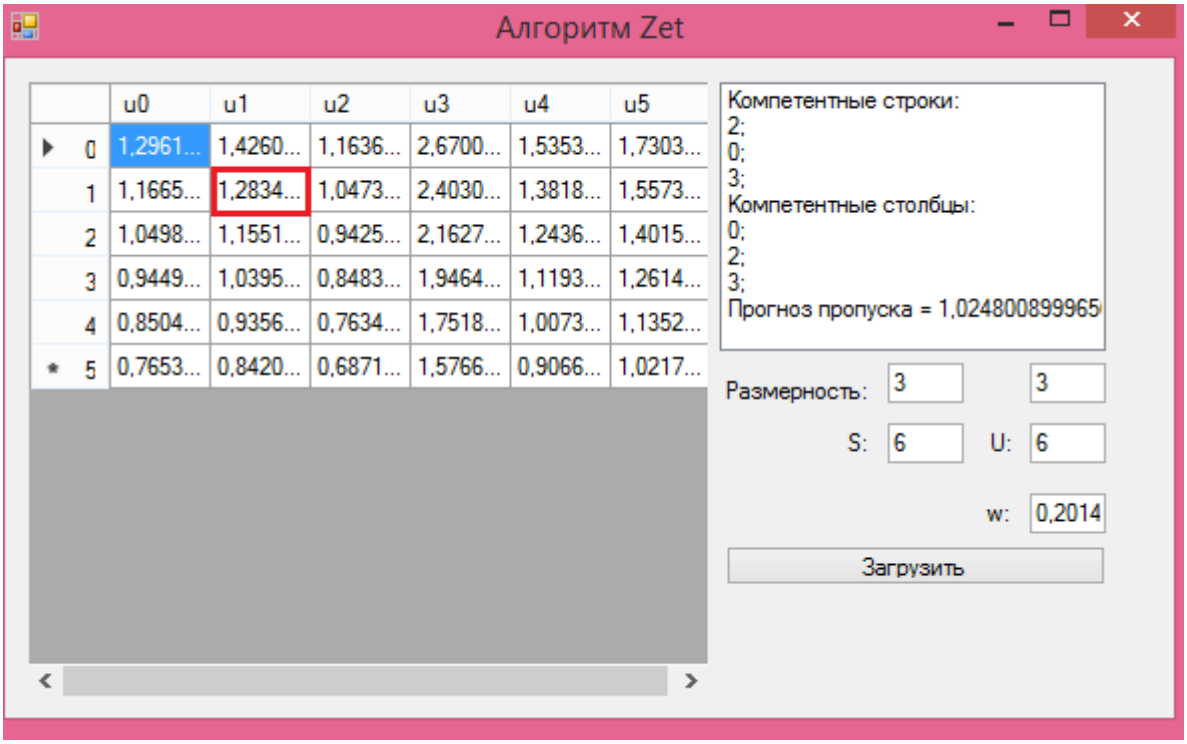


Рисунок 26 – Результат работы Zet-алгоритма в случае 3

Рассмотрим 4 случай, когда во входных данных четко выражена компетентная матрица размера 2x3 и описана зависимостью (1.25) и (1.26).

Таблица 7 - Компетентная матрица для случая 4

№	0	2	3	4
0	0	0	0	0
2	0,2	0,22	0,176	0,1408

Результаты заметно улучшились, $u_{11} = 0{,}11$, восстановленное значение $u'_{11} = 0{,}104$, ошибка равна 6%.

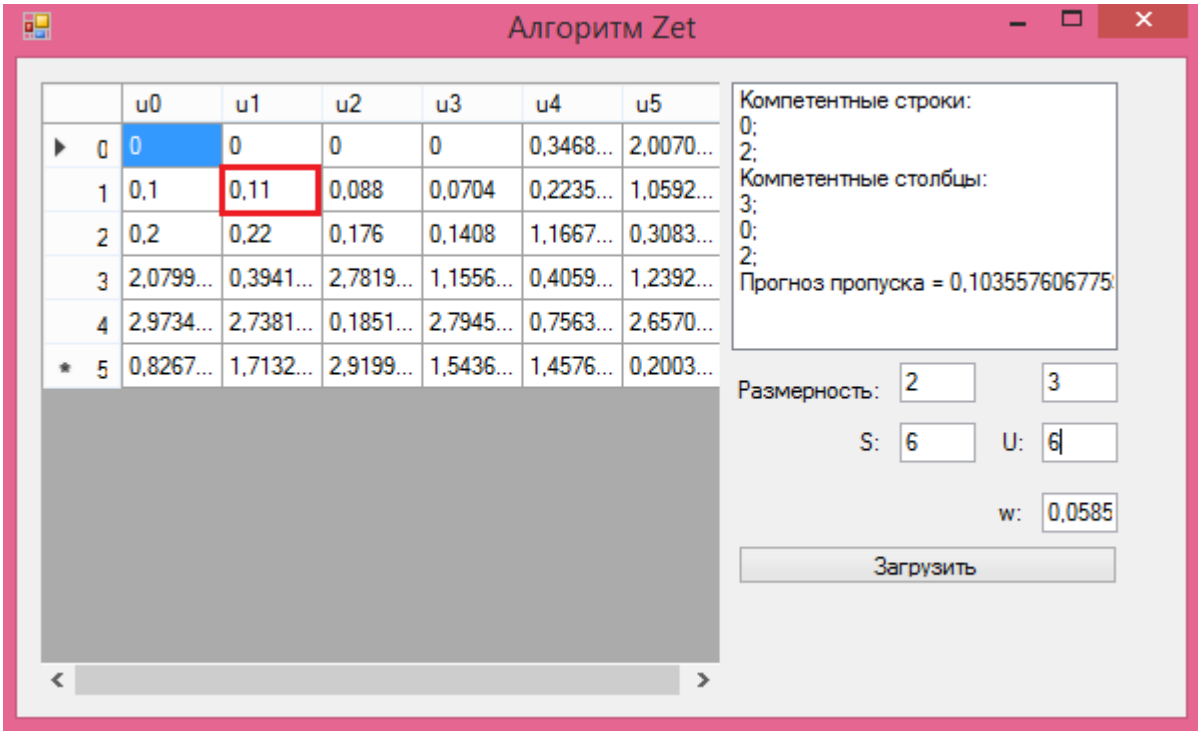


Рисунок 27 – Результат работы Zet-алгоритма в случае 4

Выводы.

В ходе исследования Zet-алгоритма были проведены вычислительные эксперименты на различных наборах данных, в результате чего был выявлен ряд недостатков, значительно влияющие на результат, это попадание в компетентную матрицу неинформативных строк, и, как следствие, неинформативных столбцов. Исходная матрица может содержать столбцы (свойства), не влияющие на целевую характеристику («шумящие» столбцы). Так как компетентность строки-объекта обратно пропорциональна декартовому

расстоянию до целевой строки в пространстве всех свойств таблицы, шумящие свойства могут вносить помехи при подсчете «похожести» строчек. Как следствие, в компетентную матрицу могут попасть неинформативные строчки, у которых значение шумящих характеристик случайно совпало со значениями шумящих характеристик целевой строки.

ЗАКЛЮЧЕНИЕ

Целью дипломной работы являлась повышение точности решения задачи идентификации, для этого были реализованы и исследованы следующие алгоритмы:

- непараметрический алгоритм заполнения пропусков;
- Zet-алгоритм.

Непараметрический алгоритм заполнения пропусков был исследован при различном объеме выборки и уровне помехи. В результате можно сделать вывод о том, что точность оценивания по восстановленной матрице наблюдений выше, чем по матрице с пропусками. Ошибка моделирования после заполнения сокращается в среднем на 30%.

Zet-алгоритм был исследован на различных наборах данных при различных зависимостях между строками и столбцами матрицы наблюдения. По результатам вычислительных экспериментов можно сделать вывод о том, что алгоритм дает наилучшие результаты на данных, для которых выполняется гипотеза избыточности, проявляющаяся в наличии похожих между собой объектов (строк) и зависящих друг от друга свойств (столбцов), причем гипотеза избыточности несет локальный характер.

Таким образом, все поставленные в начале работы задачи выполнены. В результате можно сделать вывод о том, что рассмотренные алгоритмы заполнения пропусков значительно повышают точность решения задачи идентификации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Корнеева, А.А., Медведев А.В. О НЕПАРАМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ ДИСКРЕТНО-НЕПРЕРЫВНЫХ ПРОЦЕССОВ ПРИ РАЗЛИЧНОЙ ДИСКРЕТНОСТИ КОНТРОЛЯ ПЕРЕМЕННЫХ / А.А. Корнеева, А.В. Медведев // Современные проблемы науки и образования. – 2014. – № 2.;
2. Корнеева, А.А., Сергеева Н.А., Чжан Е.А. О НЕПАРАМЕТРИЧЕСКОМ АНАЛИЗЕ ДАННЫХ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ / А.А. Корнеева, Н.А. Сергеева, Е.А. Чжан // Вестник томского государственного университета. – 2013. – № 1(22).;
3. Медведев ,А.В., Чжан, Е.А., “О непараметрическом моделировании многомерных безынерционных систем с запаздыванием”/ А.В. Медведев, Е.А. Чжан // Вестн. ЮУрГУ. Сер. Матем. моделирование и программирование. – 2017. – С.124–136
4. Rubin, D.B. Multiple Imputation for Nonresponse in Surveys : manual / D.B. Rubin. - New Yirk : Willey, 1987;
5. Литтл, Р.Дж.А. Статистический анализ данных с пропусками : учебник / Р.А.Литтл, Д.Б. Рубин. – Москва : Наука, 1991. – 198с;
6. Злоба, Е. Статистические методы восстановления пропущенных данных / Е.Злоба, И.Яцкив. // Computer Modeling & New Technologies.; Vol.6.2004;
7. Алгоритм Zet //Информационные интеллектуальные системы. Вып.40, 2008. – Режим доступа: http://iissvit.narod.ru/rass/vip40.htm;
8. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. - Новосибирск: ИМ СО РАН, 1999, 264с.;
9. Загоруйко, Н.Г. Алгоритм заполнения пропсуков в эмпирических таблицах (алгоритм ZET) : сб. тр. / Загоруйко Н.Г., Елкина В.Н., Тимеркаев В.С. – Новосибирск : Вычислительные системы, 1975. – С.3-27;

10. Синитюк, В.Е. Алгоритм ZetBraid [Электронный ресурс] / В.Е. Синитюк // Информационные интеллектуальные системы Вып.40, 2008. - Режим доступа: http://iissvit.narod.ru/rass/vip40.htm;

11. Протасов, К.В. Статистический анализ данных : учебник / К.В. Протасов. – Москва : Мир, 2005. – 142 с.;

12. Россиев, А.А. Моделирование данных для восстановления пробелов в таблицах / А.А. Россиев // Материалы конференции молодых ученых Института вычислительного моделирования СО РАН. – 1998. - с. 46-61.

13. Россиев, А.А. Моделирование данных при помощи кривых для восстановления пробелов в таблицах / А.А. Россиев // Методы нейроинформатики: сборник научных трудов. - 1998. - С. 6-22;

14. Айвазян, С.А. Прикладная статистика: Основы моделирования и первичная обработка данных. М.: Финансы и статистика / Айвазян С.А., Енюков И.С., Мешалкин Л.Д. -1983. - 471с;

15. Лапко, А.В., Лапко, В.А, Цугленок, Г.И. Синтез и анализ непараметрических моделей стохастических зависимостей и распознавания образов в условиях пропуска данных / А.В. Лапко, В.А. Лапко, Г.И. Цугленок // Вест КрасГАУ. – 2005. - №7. – С.64;

16. Лапко, А.В., Лапко, В.А. Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных / А.В. Лапко, В.А. Лапко // Автометрия. – 2008. – том 44, №3. – С.65-74;

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
«____» _____ 2018 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Непараметрический алгоритм восстановления пропусков «входных-выходных» переменных процесса

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

Научный руководитель	08.06.2018	к.т.н., доцент	Корнеева А.А.
	подпись, дата	должность, ученая степень	инициалы, фамилия
Выпускник	08.06.2018		Романова Т.С.
	подпись, дата		инициалы, фамилия
Рецензент	08.06.18	к.т.н., доцент	Комиссарова И.В.
	подпись, дата	должность, ученая степень	инициалы, фамилия

Красноярск 2018