На правах рукописи

Демин Виктор Андреевич

ВЫБОР ПАРАМЕТРА РАЗМЫТОСТИ В НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКЕ УСЛОВНОЙ ФУНКЦИИ НАДЁЖНОСТИ И ЕЁ ПРИМЕНЕНИЕ В КРИТЕРИЯХ СОГЛАСИЯ

Специальность 05.13.17 – Теоретические основы информатики

Автореферат

диссертации на соискание ученой степени

кандидата технических наук

Новосибирск – 2017

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Новосибирский государственный технический университет»

Научный руководитель: кандидат технических наук, доцент **Чимитова Екатерина Владимировна**

Официальные оппоненты: **Кошкин Геннадий Михайлович**, доктор физико-математических наук, профессор, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет», кафедра теоретической кибернетики, профессор;

Каргаполова Нина Александровна, кандидат физико-математических наук, Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, научный сотрудник

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнева», г. Красноярск

Защита состоится «01» июня 2017 г. в 13^{00} часов на заседании диссертационного совета Д 212.173.06 при Федеральном государственном бюджетном образовательном учреждении высшего образования «Новосибирский государственный технический университет» по адресу: 630073, Новосибирск, пр. К. Маркса, 20.

С диссертацией можно ознакомиться в библиотеке Новосибирского государственного технического университета и на сайте http://www.nstu.ru.

Автореферат разослан «___» _____ 2017 г.

Ученый секретарь диссертационного совета Фаддеенков Андрей Владимирович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Современное состояние и актуальность темы исследования. В прикладной математической статистике важное место занимают методы анализа надёжности и выживаемости. Теория надёжности изучает закономерности появления отказов технических устройств, причинами и моделями их возникновения. Методы теории надёжности могут применяться не только в технических приложениях, а также в медицине, экономике – в любой области знаний, где предметом исследования является продолжительность жизни объектов до возникновения некоторого системного события, которое принято называть отказом. В техническом приложении отказом может служить выход из строя изделия, в медицине отказом может считаться смерть пациента, в экономике – банкротство компании.

Функция надёжности является одним из основных понятий и определяет вероятность безотказной работы объекта за некоторое время наработки. Для построения функции надёжности используются как непараметрические методы, так и параметрические вероятностные модели, основанные на предположении о принадлежности времени наработки до отказа некоторому распределению. При построении вероятностных моделей надёжности должна учитываться зависимость вероятности безотказной работы от значений объясняющих переменных (ковариат). В качестве ковариат могут выступать как характеристики объекта (например, материал изделия в промышленности, возраст пациента в медицине и т.п.), так и факторы, воздействующие на объект в процессе наблюдения, которые оказывают влияние на его надёжность (например, температура и давление, при котором эксплуатируется изделие, тип лечения больных и т.д.).

Качество построенной параметрической модели (её адекватность и корректность сопровождаемых выводов) зависит от того, выполняются ли предположения, необходимые для возможности её использования. Поэтому проверка выполнения такого рода предположений является обязательным этапом построения любой параметрической модели. Для проверки соответствия модели исходным данным (при выполнении соответствующих предположений) используют различные критерии согласия. Основной подход к проверке гипотез о виде параметрических регрессионных моделей надёжности основан на применении классических критериев согласия к выборкам остатков. Такой подход позволяет проверить предположение о распределении времени наработки до отказа. Однако при проверке гипотезы относительно вида регрессионной зависимости мощность критериев согласия оказывается очень низкой. В связи с этим актуальной задачей оказывается разработка критерия согласия, имеющего более высокую мощность при проверке гипотез относительно вида регрессионной зависимости.

Непараметрические методы построения моделей функции надёжности рассматривались в работах многих авторов. В частности, можно выделить работы И. Ван Кейлегома (I. Van Keilegom), М. Аркитаса (M. Akritas), Н.

Веравербеке (N. Veraverbeke), В. Хардле (V. Hardle), среди отечественных авторов – Ф.П. Тарасенко, Г.М. Кошкина, А.В. Медведева. Непараметрические методы отличаются своей «неприхотливостью». Они не требуют априорных сведений о виде модели или иных специальных условий для данных.

Наиболее распространённой непараметрической оценкой функции надёжности является оценка Каплана–Мейера, основным преимуществом которой является то, что она позволяет учитывать цензурированные наблюдения. Благодаря этому оценка Каплана-Мейера получила широкое распространение, и с её использованием построен ряд критериев согласия. Однако оценка Каплана–Мейера не позволяют учесть влияние наличия в функции надёжности соответствующих объясняющих переменных. В этом смысле представляется актуальным разработка и развитие более широкого класса непараметрических методов, учитывающих влияние ковариат. К таким методам относится предложенное Р. Бераном обобщение оценки Каплана-Мейера на случай построения регрессионных моделей надежности.

Важнейшую роль при построении непараметрических оценок играет выбор параметра размытости, от которого существенно зависит точность получаемых оценок. Для случая построения непараметрических регрессионных моделей с аддитивной ошибкой разработано множество методов выбора оптимальных значений параметра размытости. Однако смысловая интерпретация этого параметра в оценке Берана существенно отличается от его интерпретации в классических моделях регрессии. В оценке Берана на основе параметра размытости могут быть получены лишь весовые коэффициенты, которые впоследствии используются при построении оценки. Поэтому классические методы выбора параметра размытости не подходят для подбора параметра размытости для оценки Берана. Таким образом, актуальной задачей оказывается разработка метода выбора оптимального параметра размытости, используемого при построении оценки Берана.

Цель и задачи исследования. Целью данной диссертационной работы является разработка и исследование адаптивного алгоритма выбора оптимального параметра размытости для оценки Берана, а также разработка на основе оценки Берана критерия согласия для проверки гипотезы о виде параметрической модели надежности.

В соответствии с поставленной целью предусмотрено решение следующих задач:

1. Разработка адаптивного алгоритма выбора оптимального параметра размытости для оценки Берана.
2. Исследование статистических свойств оценок Берана, построенных с использованием предложенного алгоритма выбора оптимального параметра размытости.
3. Разработка критериев согласия на основе оценки Берана для проверки сложных гипотез о виде параметрических регрессионных моделей надежности.

4. Исследование распределений статистик и мощности предложенных критериев согласия.
5. Разработка программного обеспечения построения оценки Берана на основе оптимального параметра размытости.
6. Решение практических задач с использованием оценки Берана.

Методы исследования. Для решения поставленных задач использовались методы математической статистики, теории вероятностей, математического программирования и статистического моделирования.

Научная новизна диссертационной работы заключается в следующем:

– впервые предложен адаптивный алгоритм выбора оптимального параметра размытости для непараметрической оценки Берана условной функции надёжности;

– методами компьютерного моделирования показано, что применение предложенного алгоритма выбора оптимального параметра размытости позволяет существенно повысить точность оценки Берана при различных планах эксперимента;

– предложены новые критерии согласия на основе оценки Берана, позволяющие проверять простые и сложные гипотезы о виде параметрических регрессионных моделей надежности;

– на основе результатов исследования распределений статистик и мощности предложенных критериев согласия сформулированы рекомендации по их использованию для различных планов эксперимента.

Положения, выносимые на защиту:

1. Результаты исследования статистических свойств оценки Берана в зависимости: от вида выбираемых ядерных функций, от объёма выборок, от числа опорных точек плана эксперимента, от вида регрессионной зависимости.

2. Предложенный и реализованный адаптивный алгоритм выбора оптимального значения параметра размытости, позволяющий строить непараметрическую оценку Берана, более точно описывающую условную функцию надежности по результатам экспериментальных наблюдений.

3. Предложенные критерии согласия, статистики которых представляют собой меры отклонения непараметрической оценки Берана от параметрической регрессионной модели надежности.

4. Результаты исследования распределений статистик предложенных критериев согласия, результаты сравнительного анализа мощности критериев при проверке близких конкурирующих гипотез, методика применения предложенных критериев.

Личный творческий вклад автора заключается:

– в разработке адаптивного алгоритма выбора оптимального параметра размытости и проведении исследований свойств оценки Берана с использованием предложенного алгоритма;

– в разработке статистического критерия согласия на основе оценки Берана и проведении исследований распределений статистик и мощности предложенных критериев;

– в разработке программного обеспечения, реализующего предложенные алгоритмы построения оценки Берана и проверку гипотез о виде параметрических регрессионных моделей с использованием предложенных критериев согласия;

– в решении задач анализа реальных данных с применением разработанных алгоритмов и программного обеспечения.

Практическая ценность результатов заключается в разработке адаптивного алгоритма выбора параметра размытости для оценки Берана, в формировании рекомендаций по использованию предложенного метода и оценки Берана для построения непараметрической оценки условной функции надёжности. Предложен универсальный критерий согласия на основе оценки Берана, предложено несколько статистик критерия, сформированы рекомендации по использованию критерия.

Разработанный алгоритм выбора параметра размытости и построение оценок Берана по цензурированным выборкам реализованы в программной системе статистического анализа данных типа времени жизни «LiTiS» (Свидетельство о государственной регистрации программы для ЭВМ № 2014661905, 2015 г. – М.: Федеральная служба по интеллектуальной собственности (Роспатент)).

Результаты внедрены в практику деятельности ООО «ДГ-Софт», филиала ПАО «Электросетьсервис ЕНЭС», а также нашли практическое применение в учебном процессе на факультете прикладной математики и информатики ФГБОУ ВО «Новосибирский государственный технический университет», что подтверждается соответствующими актами о внедрении.

Исследования и разработка программного обеспечения проводились при поддержке Министерства образования и науки РФ в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» (соглашение №14.B37.21.0860 от 6 сентября 2012 г.) и в рамках проектной части государственного задания (проект № 2.541.2014/К).

Соответствие диссертации паспорту научной специальности. Содержание диссертации соответствует п. 5 области исследований «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений» паспорта специальности научных работников 05.13.17 – «Теоретические основы информатики» по техническим наукам.

Апробация результатов диссертации. Результаты работы докладывались на Международном семинаре “Applied methods of statistical analysis” (Новосибирск, 2013, 2015 г.); Международном семинаре “International Workshop on Simulation” (Римини, Италия, 2013 г.; Вена, Австрия, 2015 г.); международной научно-технической конференции “Актуальные проблемы электронного приборостроения” (Новосибирск, 2012, 2014 г.); Российской научно-технической конференции “Обработка информационных сигналов и математическое моделирование” (Новосибирск, 2012 г.); Российской научно-технической конференции “Информатика и

проблемы телекоммуникаций", (Новосибирск, 2010, 2011г.); всероссийской научной конференции молодых ученых "Наука. Технология. Инновации", (Новосибирск, 2011, 2015 г.); всероссийском научном симпозиуме " Непараметрика-XIV" (Томск, 2012 г.); Международном форуме по стратегическим технологиям IFOST-2016 (Новосибирск, 2016 г.).

Публикации. По результатам диссертационных исследований опубликовано 13 печатных работ, в том числе четыре статьи в научных журналах и изданиях, рекомендуемых ВАК РФ, восемь публикаций в материалах Международных и Российских конференций, получено свидетельство о государственной регистрации программы для ЭВМ.

Структура и объем диссертации. Общий объем диссертационной работы составляет 127 страниц, основная часть изложена на 123 страницах и состоит из введения, четырех глав основного содержания, включая 20 таблиц и 34 рисунка, заключения, списка использованных источников из 113 наименований и приложения.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертации, сформулированы цель и задачи исследования, определены научная новизна и практическая ценность работы, дано краткое содержание работы по разделам.

В первой главе диссертационной работы рассматривается непараметрическая регрессионная модель зависимости вероятности дожития до времени t от некоторого фактора, влияющего на продолжительность жизни. Исследуются свойства получаемых непараметрических оценок условной функции надежности.

В задачах анализа данных типа времени жизни объектом исследования является группа объектов, для каждого из которых определено некоторое системное событие, часто называемое отказом. Данные любого случайного эксперимента, в результате которого получены отказы объектов, можно считать данными типа времени жизни.

Одной из особенностей данных типа времени жизни является их неполнота. Во время испытаний может выйти из строя лишь некоторый процент исследуемых объектов, ряд объектов по каким-то причинам может быть снят с испытаний, время испытаний может быть ограничено. Таким образом, к концу эксперимента часть объектов может остаться в работоспособном состоянии. Такие данные называют цензурированными.

Цензурированной справа называют выборку вида:

$$(Y_1,\delta_1),(Y_2,\delta_2),...,(Y_n,\delta_n), \quad (1)$$

где Y_i – значение наблюдения;

$$Y_i = \min(T_i, C_i),$$

где T_i – время наступления отказа;

C_i – время цензурирования (время завершения наблюдения за i-м объектом), δ_i – индикатор события, который содержит информацию о причине прекращения наблюдения, $i = 1,2,...,n$.

Если в ходе эксперимента было зафиксировано время отказа, то $Y_i = T_i, \delta_i = 1$, и данное наблюдение называется *полным*. Если же нам неизвестно T_i по причине окончания наблюдения в момент $C_i \le T_i$, то $Y_i = C_i, \delta_i = 0$, и наблюдение называется *цензурированным справа*.

Одной из важнейших задач статистического анализа надёжности является изучение зависимости времени до наступления отказа от объясняющих переменных, которые также называются ковариатами. *Ковариатой* является величина, описывающая какое-либо свойство, характеристику объекта, либо степень внешнего воздействия на объект, условия проведения эксперимента. Область значений ковариаты x определяется условиями проведения эксперимента и представляет собой отрезок числовой прямой. В настоящем диссертационном исследовании рассматривается случай скалярной ковариаты, и предполагается, что ковариата является постоянной по времени величиной. Все объекты выборки разделяются на m групп, соответствующих различным значениям ковариаты. Обычно для удобства значения ковариат в плане эксперимента линейным преобразованием приводят к отрезку $[0, 1]$.

Если значения ковариаты выбираются случайно, в соответствии с некоторым законом распределения, то такой план эксперимента называют *случайным.* При этом если распределение ковариаты является непрерывным (например, равномерное на заданном отрезке), то число групп m будет равно объёму выборки n.

Данные об отказах с ковариатами представляются в следующем виде:

$$(Y_1, x_1, \delta_1), (Y_2, x_2, \delta_2), ..., (Y_n, x_n, \delta_n), \quad (2)$$

где Y_i – время безотказной работы или момент цензурирования *i*-го объекта; δ_i – индикатор цензурирования *i*-го объекта, который принимает значение 1, если наблюдение полное, и 0, если цензурированное; x_i – значение ковариаты, при котором наблюдался *i*-й объект, $i = 1,2,...,n$.

Обозначим через T_x длительность жизни (время наработки до отказа), которая является неотрицательной случайной величиной с непрерывным законом распределения. Поскольку время безотказной работы объектов может зависеть от их свойств или условий проведения эксперимента (от значений ковариаты x), будем рассматривать условную *функцию надёжности*:

$$S(t \mid x) = P(T_x \ge t) = 1 - F(t \mid x), \quad (3)$$

где $F(t \mid x)$ – условная функция распределения случайной величины T_x.

В теории надёжности регрессионную модель принято записывать через функцию надёжности следующим образом:

$$S(t \mid x) = H\left(S_0(t); x\right), \quad (4)$$

где $S_0(t) = S(t \mid x = 0)$ – базовая функция надежности; оператор $H(\cdot) : [0,1] \to [0,1]$ определяет изменение базовой функции надёжности для различных значений объясняющих переменных.

На основе данных об отказах объектов, полученных в результате исследования (эксперимента), можно построить вероятностные модели для оценки показателей надёжности. В зависимости от характера привлекаемой априорной информации различают два основных подхода к построению оценки вероятностной модели надёжности: *параметрический* и *непараметрический.*

В рамках параметрического подхода предполагается, что распределение времени безотказной работы принадлежит какому-либо семейству распределений, например, семейству распределений Вейбулла, экспоненциальному, логнормальному и др. Помимо параметризации функции распределения, также предполагается некоторый вид зависимости базовой функции надежности от ковариат. Для этого вводится функция от ковариат с неизвестными параметрами. Функция от ковариат и базовая функция распределения могут по-разному учитываться в модели. Поэтому существует множество параметрических моделей надёжности.

Одной из наиболее популярных вероятностных моделей, описывающих зависимость надёжности от независимых ковариат, является модель пропорциональных интенсивностей, которая также называется «пропорциональной моделью Кокса». Функция надёжности для модели пропорциональных интенсивностей имеет вид:

$$S_x(t;\beta) = \left(S_0(t)\right)^{r(x;\beta)}, \tag{5}$$

где $r(x;\beta)$ – функция от ковариаты, $S_0(\cdot)$ – базовая функция надёжности.

В основе данной модели лежит предположение, что отношение функций интенсивности при разных значениях ковариаты x^1 и x^2 не зависит от времени:

$$\frac{\lambda_{x^2}(t)}{\lambda_{x^1}(t)} = \frac{\exp\left(\beta \cdot x^2\right)}{\exp\left(\beta \cdot x^1\right)} = \text{const}. \tag{6}$$

Не менее распространённой моделью является модель ускоренных испытаний (Accelerated Failure Time). AFT-модель предназначена для оценивания функции надёжности изделий (систем), функционирующих в нормальных условиях эксплуатации (при воздействии нормальных нагрузок), по данным об отказах, полученным в результате ускоренных испытаний (при использовании повышенных нагрузок). В общем виде AFT-модель ускоренных испытаний выглядит следующим образом:

$$S_x(t) = S_0\left(\frac{t}{r(x;\beta)}\right). \tag{7}$$

При параметрическом подходе всегда делаются некоторые предположения о виде зависимости функции надёжности от ковариат.

Важнейшим этапом построения параметрической регрессионной модели является проверка того, согласуются ли введенные предположения с данными об отказах. Другими словами требуется проверить сложную гипотезу о виде модели

$$H_0 : S(t \mid x) = H\left(S_0(t;\theta);x,\beta\right), \tag{8}$$

где $S_0(t;\theta)$ – базовая функция надежности, β – вектор регрессионных параметров.

Для проверки гипотезы (8) используются различные критерии согласия, в основе которых чаще всего лежит оценка расстояния между непараметрической оценкой и функцией надёжности, соответствующей проверяемой гипотезе. В работах Чимитовой Е.В. предлагается подход к проверке данной гипотезы, основанный на применении классических критериев согласия к выборкам остатков. Такой подход позволяет проверить предположение о базовом распределении отказов, однако при проверке гипотезы относительно вида регрессионной зависимости мощность критериев согласия оказывается очень низкой. Таким образом, актуальной задачей оказывается разработка критерия согласия, имеющего более высокую мощность при проверке гипотез относительно вида регрессионной зависимости.

На практике не всегда существуют априорные предположения о виде распределения отказов или о функциональной зависимости функции надёжности от ковариат. Непараметрические методы оценивания используют только выборку и не требуют никаких априорных предположений или каких-либо специальных условий, что является несомненным плюсом непараметрического подхода перед параметрическим. Пожалуй, самой распространённой непараметрической оценкой функции надёжности является оценка Каплана-Мейера:

$$\hat{S}(t) = \prod_{j:\delta_j=1, Y_j \le t} \left(1 - \frac{d_j}{r_j}\right), \tag{9}$$

где δ_i – индикатор цензурирования, r_j – число объектов, наблюдаемых в момент t, d_j – число объектов, отказавших в момент t.

Оценка Каплана–Мейера не учитывает влияния ковариат и не позволяет оценить условную функцию надёжности. В 1981 году Бераном была предложена оценка, обобщающая оценку Каплана-Мейера и имеющая вид:

$$\tilde{S}_{b_n}\left(t \mid x\right) = \prod_{Y_{(i)} \le t} \left\{1 - \frac{W_{n(i)}\left(x;b_n\right)}{1 - \sum_{j=1}^{i-1} W_{n(j)}\left(x;b_n\right)}\right\}^{\delta_i}, \tag{10}$$

где x – значение ковариаты, для которой оценивается функция надёжности; $W_{n(i)}\left(x;b_n\right),\ i = 1,\ldots,n-$ веса Надарая–Ватсона, которые вычисляются по формуле:

$$W_{n(i)}\left(x;b_n\right)=K\left(\frac{x-x_i}{b_n}\right)\Big/\sum_{j=1}^{n}K\left(\frac{x-x_j}{b_n}\right), \qquad (11)$$

где $K(\cdot)$ – ядерная функция, удовлетворяющая условиям регулярности: $K(y)=K(-y)$, $0\le K(y)<\infty$, $\int_{-\infty}^{\infty}K(y)dy=1$, b_n – параметр размытости такой, что $\lim\limits_{n\to\infty}b_n=0$, $\lim\limits_{n\to\infty}nb_n=\infty$.

Такая непараметрическая модель способна служить оценкой функции надёжности для наблюдений с ковариатами. Основная идея оценки Берана заключается в том, что каждому наблюдению присваивается свой вес в зависимости от близости значения ковариаты этого наблюдения к значению ковариаты, для которого строится оценка Берана. Чем это расстояние меньше, тем больше вес. При значениях весов Надарая–Ватсона $W_{n(i)}\left(x;b_n\right)=n^{-1}$, то есть когда всем наблюдениям присваивается одинаковый вес, оценка Берана сводится к оценке Каплана-Мейера. Для расчёта весов Надарая–Ватсона используются ядерные функции.

Как известно, в ядерном сглаживании ключевую роль играет параметр размытости, от которого зависит то, по каким элементам выборки будет происходить сглаживание. При построении оценки Берана от выбора этого параметра, главным образом, зависит качество полученной модели.

Для исследования статистических свойств непараметрических оценок и исследования распределений статистик и мощности критериев согласия в работе использовалась методика статистического моделирования.

Показано, что точность оценки Берана существенно зависит от выбора параметра размытости. На основе проведённых исследований был сделан вывод о существовании оптимального значения параметра размытости, при котором отклонение оценки Берана от истинной функции надёжности минимально. Было показано, что оптимальный параметр размытости зависит от нескольких факторов: от плана эксперимента, от числа групп, от объёма выборки, от степени влияния ковариаты на функцию надёжности, от выбора ядерной функции, от того, для какого значения ковариаты строится условная функция надёжности.

В ядерном сглаживании существует множество различных методов для выбора оптимального значения параметра размытости. Однако параметр размытости для оценки Берана определяет только вес наблюдения. По сути, при построении оценки Берана процедура ядерного сглаживания отсутствует как таковая, т.к. не производится сглаживание (восстановление) регрессионной зависимости отклика от ковариаты. Поэтому, пожалуй, единственным существующим методом, который можно применить для выбора параметра размытости оценки Берана, является референтный метод:

$$b_{n,opt}=Cn^{-1/5}.$$

Очевидно, что он обладает рядом минусов. Во-первых, отсутствует какой-либо алгоритм для подбора константы C. Поэтому, в качестве константы в

исследованиях, как правило, принимается единица. Понятно, что данная формула взята из ядерного сглаживания, где число одинаковых наблюдений обычно крайне мало. Поэтому второй минус заключается в том, что данный метод предназначен только для случайного плана. В-третьих, данный подход не учитывает числа групп и степень влияния значения ковариаты на функцию надёжности. В-четвертых, данный метод не позволяет выбрать значение параметра размытости для конкретного значения ковариаты, т.е. не является адаптивным.

Таким образом, актуальной задачей является разработка алгоритма выбора оптимального значения параметра размытости для оценки Берана, который бы удовлетворял следующим требованиям:

– алгоритм должен быть адаптивным, и способным находить оценку оптимального значения параметра размытости для любого значения ковариаты;

– алгоритм должен работать для любого числа опорных точек плана эксперимента, в том числе для случайного плана;

– алгоритм должен учитывать особенности используемой модели, в том числе степень влияния ковариаты на функцию надёжности;

– алгоритм должен учитывать объём выборки.

Во второй главе предлагается адаптивный алгоритм выбора оптимального параметра размытости, который учитывает особенности плана эксперимента и степень влияния ковариаты на функцию надёжности. Методами статистического моделирования исследуется точность оценки Берана, строящейся с использованием предложенного алгоритма для выбора оптимального значения параметра размытости.

Предложенный алгоритм выбора оптимального значения параметра размытости b_n основан на минимизации среднего отклонения времен отказов $Y_1, Y_2, ..., Y_n$ от непараметрической оценки обратной функции надежности $S_x^{-1}(p)$. Разработанный метод позволяет учесть объём выборки, число групп, степень влияния ковариаты на функцию надёжности. Кроме того, метод является адаптивным, то есть позволяет подобрать оценку параметра размытости для произвольного значения ковариаты.

Алгоритм выбора оптимального параметра размытости можно сформулировать следующим образом.

1. Нормировать значения ковариат:

$$x_i = \frac{x_i}{x_{\max}},\ i = \overline{1,n},$$

где $x_{\max}$ – наибольшее значение ковариаты в выборке.

2. Задать $k = 1$.
3. Решить задачу одномерной оптимизации на интервале:

$$b_n^k = \arg \min_{b_n \in \left(\frac{k-1}{m}, \frac{k}{m}\right]} \sum_{i=1}^{n} \delta_i \cdot \left| \hat{g}\left(\hat{p}_i \mid x_i\right) - Y_i \right|,$$

где параметр размытости b_n входит в оценку Берана, которая является параметром в ядерной оценке обратной функции надёжности:

$$\hat{g}\left(\hat{p}_i \mid x_i\right) = \sum_{j=1}^{n} \omega_j\left(\hat{p}_i\right) \cdot Y_j$$

.

В качестве $\omega_j\left(\hat{p}_i\right)$ рекомендуется использовать веса Пристли-Чао:

$$\omega_j^{(2)}\left(\hat{p}_i\right) = n\left(\hat{p}_{(i)} - \hat{p}_{(i-1)}\right) K\left(\frac{\hat{p}_i - \hat{p}_j}{h_{NS}}\right),$$

где для оценки параметра размытости h_{NS} используется метод минимума средней интегральной ошибки:

$$h_{NS} = \left[\frac{8\pi^{1/2} R(K)}{3\mu_2(K)^2 n}\right]^{1/5} \hat{\sigma}_{robust}$$

с робастной оценкой среднеквадратического отклонения вида:

$$\hat{\sigma}_{robust} = 1.4826 \underset{i=1..n}{med} \left| \hat{p}_i - \underset{j=1..n, k=j+1..n}{med} \left(\frac{\hat{p}_j + \hat{p}_k}{2}\right) \right|.$$

4. Если $k < m$, то $k = k + 1$ и перейти к пункту 3, иначе перейти к пункту 5.
5. Вычислить оптимальный параметр размытости:

$$b_n^{opt} = \arg \min_{b_n \in \left\{b_n^1, b_n^2, \ldots, b_n^m\right\}} \sum_{i=1}^{n} \delta_i \cdot \left| \hat{g}\left(\hat{p}_i \mid x_i\right) - Y_i \right|. \qquad (12)$$

Выбор весовой функции Пристли-Чао, а также оценки параметра размытости h_{NS} обусловлен тем, что проведённые исследования показали, что наилучшие результаты достигаются при использовании весовой функции Пристли-Чао с параметром сглаживания h_{NS}, вычисляемого на основе медианы абсолютных отклонений $\hat{\sigma}_{robust}$. Также среди рассмотренных вариантов ядерных функций предпочтительней использовать ядро Епанечникова и квартическую функцию, поскольку они позволяют получать более точные оценки Берана.

Таким образом, при изменении параметра размытости b_n изменяется и оценка Берана, которая в свою очередь является параметром для ядерной оценки обратной функции надёжности. Минимизируя разницу между оценкой обратной функции надёжности и временами отказов, которые нам известны, получаем оптимальное значение параметра размытости.

В диссертации методами статистического моделирования проведены исследования статистических свойств оценки Берана с использованием предложенного алгоритма выбора оптимального параметра размытости в зависимости: от плана эксперимента, от числа групп, от объёма выборки, от степени влияния ковариаты на функцию надёжности, от выбора ядерной

функции, от того, для какого значения ковариаты строится условная функция надёжности. Показано, что оценки Берана, получаемые на основе предложенного алгоритма выбора оптимального значения параметра размытости, оказываются точнее, чем при использовании референтного правила выбора параметра размытости, который применим только при случайном плане эксперимента. Таким образом, предложенный алгоритм удовлетворят всем требованиям, сформулированным в главе 1.

В третьей главе для проверки гипотез о виде параметрических регрессионных моделей надёжности предлагаются непараметрические критерии согласия, в статистиках которых используются оценки Берана. Методами статистического моделирования исследуются распределения статистик и мощность предложенных критериев согласия, проводится сравнение с классическими критериями согласия, применяемыми к выборкам остатков.

Существующие критерии проверки гипотезы о виде параметрической вероятностной модели пропорциональных интенсивностей основаны на построении остатков Кокса-Снелла, которые можно вычислить следующим образом:

$$R_i = \Lambda_{x_i}\left(Y_i;\hat{\beta}\right),\ i = 1,2,...,n\ , \tag{13}$$

где $\Lambda_{x_i}(\cdot)$ – кумулятивная функция риска, $\hat{\beta}$ – ОМП параметров предполагаемой модели.

Если гипотеза о виде модели верна, то полученная выборка остатков $\mathbf{R}_n = \left\{\left(R_1,\delta_1\right),...,\left(R_n,\delta_n\right)\right\}$ принадлежит стандартному экспоненциальному распределению.

Подобный подход применяется и для моделей ускоренных испытаний. Для проверки гипотезы о виде построенной параметрической AFT-модели также анализируют выборку остатков. Остатки для параметрической AFT-модели вычисляются иначе, чем остатки для моделей пропорциональных интенсивностей. Кроме того, для AFT моделей ковариата может быть – как постоянной, так и ступенчатой. В данной работе исследуется только случай с постоянной по времени ковариатой, поэтому рассмотрим остатки только для этого случая:

$$R_i = \frac{Y_i}{r\left(x_i;\hat{\beta}\right)},\ i = \overline{1,n}\ . \tag{14}$$

Если данные хорошо описываются построенной AFT-моделью, остатки распределены в соответствии с базовым законом распределения отказов $F_0\left(t;\hat{\theta}\right)$, стандартизованным по параметру масштаба (при параметре масштаба $\theta_1 = 1$).

Для проверки гипотезы о принадлежности выборки остатков экспоненциальному распределению в случае моделей пропорциональных интенсивностей с базовым распределением $F_0\left(t;\theta\right)$ в случае AFT-модели

можно использовать критерии согласия типа Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга.

Следует заметить, что при построении моделей с учетом ковариат, проверяемая гипотеза о согласии выборки остатков со стандартным экспоненциальным законом или с базовым распределением отказов, является сложной, так как проверка осуществляется по той же выборке, по которой были оценены параметры модели.

Исследования показали, что рассматриваемые критерии на основе выборок остатков обладают высокой мощностью при проверке гипотез о виде базового распределения отказов. Однако при проверке гипотезы о виде регрессионной зависимости высокая мощность данных критериев достигается лишь при очень больших объемах выборок.

Основная идея предлагаемого критерия проверки гипотезы о виде параметрической регрессионной модели надёжности на основе оценки Берана заключается в том, чтобы оценить расстояние между теоретической условной функцией надёжности, соответствующей проверяемой гипотезе, и непараметрической оценкой Берана при различных значениях ковариаты. Идея схожа с классическим подходом к проверке гипотез о виде моделей с помощью непараметрических критериев согласия. Главное отличие разработанного подхода заключается в том, что не возникает необходимости перехода к выборке остатков. Отсутствие такого перехода делает предлагаемый метод более гибким и универсальным.

Для того чтобы учесть все значения ковариаты в выборке, в работе предлагается использовать статистику вида:

$$S_B^1 = \sup_{i=1..m}\left[\left(\frac{nb_n}{\ln n}\right)^{0.5}\sup_{t>0}\left|S_{x_i}(t;\theta,\beta)-\tilde{S}_{b_n}(t\mid x_i)\right|\right], \qquad (15)$$

в которой берется наибольшее отклонение оценки Берана от предполагаемой функции надёжности, либо среднее отклонение вида:

$$S_B^2 = \left(\frac{1}{m}\right)\sum_{i=1}^{m}\left(\frac{nb_n}{\ln n}\right)^{0.5}\sup_{t>0}\left|S_{x_i}(t;\theta,\beta)-\tilde{S}_{b_n}(t\mid x_i)\right|, \qquad (16)$$

где $\tilde{S}_{b_n}(t\mid x_i)$ – оценка Берана при значении ковариаты x_i.

Предложенные статистики позволяют использовать особенности каждого из видов плана эксперимента. Статистика (15) предназначена для плана эксперимента с большим количеством элементов в группе, так как в этом случае информации о каждом значении ковариаты в плане эксперимента достаточно для того, чтобы использовать наибольшее отклонение. При случайном же плане, либо при планах, в которых количество наблюдений для каждого значения ковариаты невелико, следует использовать статистику (16).

Гипотеза о виде параметрической регрессионной модели надёжности (8) отвергается при больших значениях статистик S_B^1 и S_B^2. Применение предложенных критериев предполагает реализацию интерактивного режима

для моделирования распределений статистик $G_N(y|H_0)$ в ходе проводимого анализа, чтобы использовать $G_N(y|H_0)$ при формировании статистического вывода о результатах проверки гипотезы.

Для проверки гипотезы о виде параметрической регрессионной модели надёжности (16) с помощью предложенных критериев необходимо:

1. В соответствии с построенной моделью $H(S_0(t;\hat{\theta}),x,\hat{\beta})$, где $\hat{\theta},\hat{\beta}$ – ОМП параметров модели по исходной выборке, смоделировать выборку отказов $Y=\{(Y_1,x_1),(Y_2,x_2),...,(Y_n,x_n)\}$.
2. По полученной выборке оценить параметры модели методом максимального правдоподобия.
3. Вычислить значение статистики (15) или (16).
4. Повторить пункты 1-3 N раз, получив в результате эмпирическое распределение статистики критерия $G_N(y|H_0)$.

Проверяемая гипотеза H_0 отклоняется, если достигнутый уровень значимости $\alpha_n=1-G_N(S_B|H_0)$ меньше заданного α, где S_B – значение статистики соответствующего критерия согласия, полученное по исходной выборке.

Использование описанного алгоритма проверки гипотез о виде параметрических регрессионных моделей надёжности позволяет проверить предположение как о виде параметризации базового распределения отказов, так и о виде регрессионной зависимости функции надёжности от ковариат.

Проведенные исследования показали, что предложенный критерий согласия на основе оценки Берана способен различить гипотезы о виде регрессионной зависимости с большей мощностью, чем критерий Колмогорова, основанный на остатках. В то же время предложенные критерии согласия в большинстве случаев уступают критерию, основанному на остатках, в случаях проверки гипотезы о виде базового распределения. Это вполне логично, так как критерии, основанные на остатках, разрабатывались специально для проверки такого рода гипотез. Таким образом, предложенные критерии целесообразно рекомендовать для проверки гипотез о виде параметрической регрессионной модели.

В четвёртой главе представлено описание разработанного программного обеспечения, позволяющего осуществлять построение оценки Берана. Описываются основные этапы статистического анализа данных о частичных разрядах, возникающих в жидких диэлектриках. Приводится решение задачи статистического анализа продолжительности сотрудничества ООО «ДГ-Софт» с рекламодателями.

В очередную версию программной системы «LiTiS 1.1» автором диссертации была добавлена возможность построения оценки Берана с использованием предлагаемого алгоритма выбора оптимального значения параметра размытости. Расширенная версия программной системы была зарегистрирована как программная система статистического анализа данных типа времени жизни «LiTiS 1.2». В данной версии пользователь имеет

возможность построить оценки Берана для интересующих значений ковариаты с использованием предложенного алгоритма выбора оптимального параметра размытости, добавить на график теоретические функции надёжности и визуально оценить близость предполагаемых параметрических моделей надёжности к оценкам.

Анализ данных о частичных разрядах, возникающих в жидких диэлектриках. С целью исследования энерговложения в жидкость А.Л. Бычковым[1] были проведены экспериментальные исследования по определению коэффициента газообразования в диэлектрических жидкостях. В результате была получена выборка частичных разрядов (ЧР) объёмом $n = 626$. Наибольший интерес вызывает зависимость мгновенного напряжения (U) от действующего напряжения (U^W). В данном случае значение мгновенного напряжения можно считать случайной величиной, зависящей от действующего напряжения, которое можно считать ковариатой. Тогда полученные значения можно представить в виде выборки:

$$\mathbf{X}_n = \left\{(U_1, U_1^W), (U_2, U_2^W), ..., (U_n, U_n^W)\right\}.$$

В результате проведенного анализа построена параметрическая вероятностная модель для описания условной функции надежности наблюдаемого напряжения при возникновении ЧР в зависимости от величины действующего напряжения U^W, которая имеет следующий вид:

$$S_x(t) = P\left\{U > t \mid x = U^W\right\} = \exp\left(-\left(\frac{t}{\exp\left(9.0207 + 5.4734 \cdot 10^{-5} x\right)}\right)^{14.3848}\right). \quad (17)$$

На рисунке 1 представлены оценки Берана и соответствующие функции надёжности (17) при различных значениях действующего напряжения. Из рисунка 1 видно, что подобранная модель достаточно хорошо согласуется с оценками Берана, построенными по выборкам при различных значениях действующего напряжения. С помощью предложенного в настоящей диссертации критерия согласия проверена сложная гипотеза о виде модели (17), гипотеза не отклоняется, так как достигнутый уровень значимости (*p*-value) равен 0.87. Таким образом, получена статистическая модель, способная описать распределение частичных разрядов в зависимости от действующего напряжения, что позволяет моделировать процесс деградации изоляционных жидкостей, делать прогноз срока службы изоляционных жидкостей при различных напряжениях. Также было показано, что при больших действующих напряжениях, в жидких диэлектриках начинают возникать частичные разряды двух различных типов, что необходимо учитывать при расчёте коэффициента газообразования.

[1] Бычков, А.Л. Исследование газообразования при частичных разрядах и совершенствование пробоотбора для газового анализа высоковольтного маслонаполненного электрооборудования: дис. … канд. техн. наук: 05.14.12. / Бычков Александр Леонидович. – Новосибирск, 2014. – 156 с.

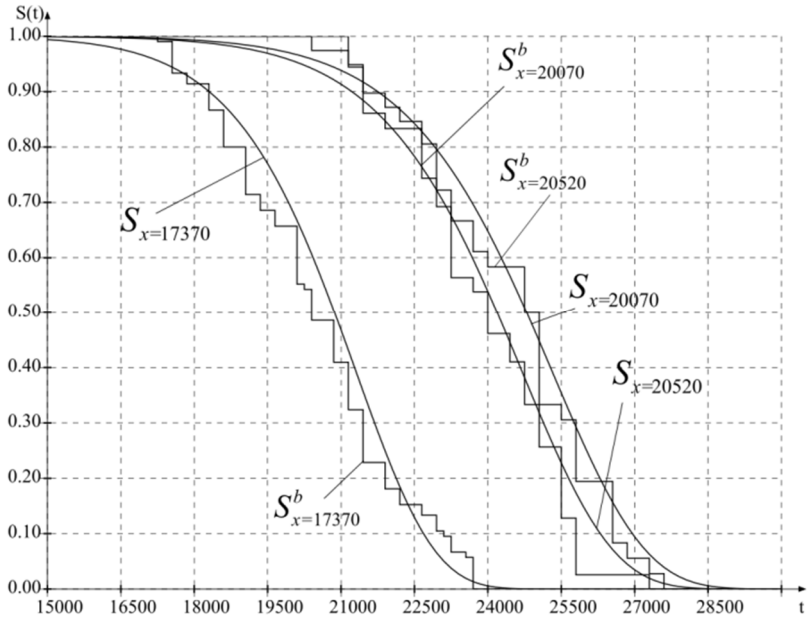

Рисунок 1– Функции надёжности и оценки Берана при различных значениях действующего напряжения

Анализ вероятности продления контракта рекламодателем. Компания ООО «ДГ-Софт» предоставляет бесплатный справочный и картографический продукт «2ГИС». Монетизация (прибыль) компании происходит в большей степени за счёт продажи рекламных позиций (как в справочнике, так и на карте). Таким образом, экономическое благополучие компании зависит от количества фирм-рекламодателей (РД). Основным показателем, характеризующим эффективность работы с текущими РД является доля фирм, которые решили продолжить размещение рекламы после первого периода размещения.

В качестве времени наработки до отказа примем количество месяцев, в течении которых фирма размещала рекламу в «2ГИС» до момента расторжения контракта. Вероятность безотказной работы (функция надёжности) зависит от количества рекламных позиций, поэтому в качестве ковариаты использовалось начальное количество рекламных позиций, купленных фирмой в первый месяц размещения.

В исходную выборку попали данные о фирмах, купивших рекламу в «2ГИС» в апреле 2014 года в городе Казань. Процент продлений в этом городе составлял 28%. Особенность этих данных заключается в том, что часть фирм не расторгла контракт с ООО «ДГ-Софт» в этот период, поэтому такие фирмы рассматриваются как цензурированные справа наблюдения. Выборка в этом случае имеет вид:

$$(Y_1, x_1, \delta_1), (Y_2, x_2, \delta_2), ..., (Y_n, x_n, \delta_n),$$

где Y_i – количество месяцев до расторжения контракта i-й фирмой ($\delta_i = 1$), либо до момента цензурирования ($\delta_i = 0$);

x_i – начальное количество купленных позиций;

δ_i – индикатор цензурирования.

Объём выборки составил $n = 746$, из которых 155 наблюдений являются цензурированными справа. На рисунке 2 представлены оценки Берана при $x = 1, 4, 10$, полученные при использовании предложенного алгоритма выбора оптимального значения параметра размытости.

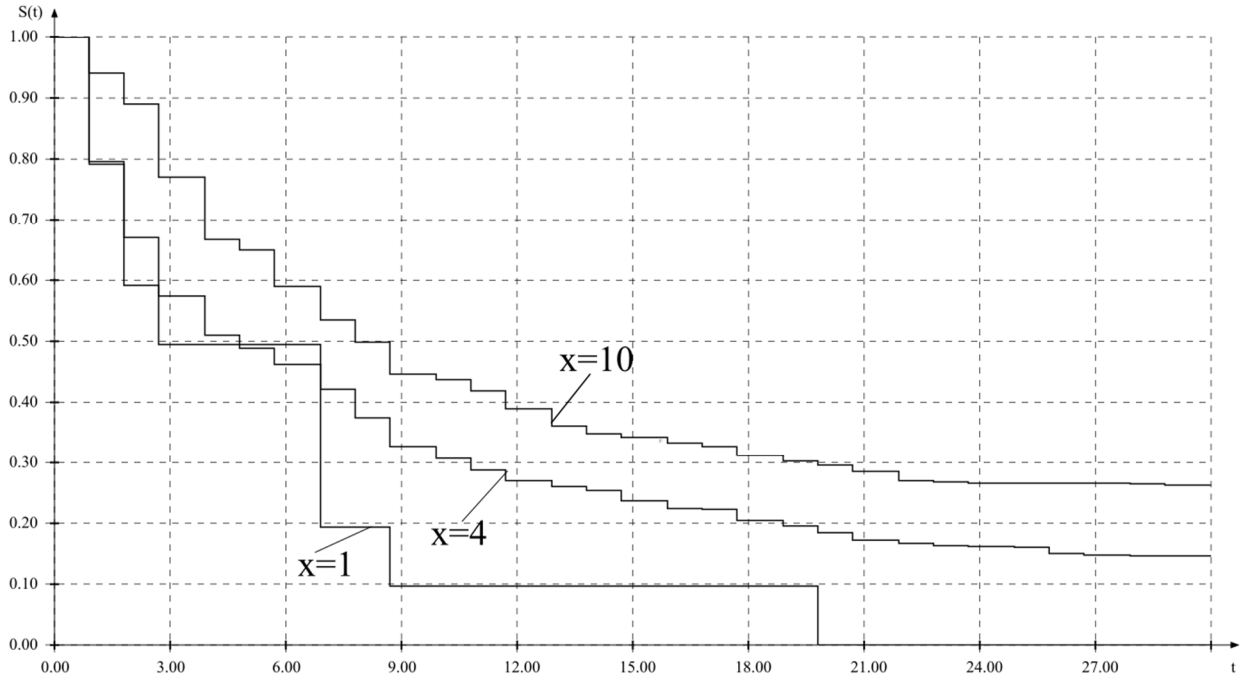

Рисунок 2 – Оценки Берана для фирм, которые купили 1, 4 или 10 позиций

Из рисунка 2 видно, что чем больше позиций купила фирма, тем больше вероятность её долгосрочного сотрудничества с «2ГИС». Это связано с тем, что эффект от рекламы при б*о*льшем количестве позиций выше, чем при малом. Также видно, что для фирм с малым числом позиций критическим является седьмой месяц размещения, после которого произошло множество прерываний контрактов. Следовательно, стоит уделить повышенное внимание группам с малым числом позиций (1, 2) в период с шестого по девятый месяц размещения. После девятого месяца размещения рекламы характер функции надёжности меняется как для группы с четырьмя позициям, так и для группы с десятью позициями: график функции становится более пологим. Другими словами, если РД с большим числом позиций «дожил» до девятого месяца, то вероятность того, что он расторгнет контракт в дальнейшем, падает. Поэтому на такие фирмы не стоит тратить ресурсы по удержанию клиентов. Следует отметить, что для всех рассмотренных функций надёжности наблюдается резкое падение в первые три месяца, это объясняется тем, что фирмы не получили ожидаемый эффект от размещения рекламы и сразу прекратили сотрудничество. Следует добавить, что все фирмы, рассматриваемые как цензурированные справа наблюдения, являются фирмами с большим числом рекламных позиций.

Таким образом, на основе полученных результатов анализа данных сформулированы соответствующие рекомендации по работе менеджеров в городе Казань.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В соответствии с поставленной целью диссертационного исследования получены следующие основные результаты.

1. Предложен адаптивный алгоритм выбора оптимального параметра размытости для построения непараметрической оценки Берана регрессионной модели надежности, который учитывает количество опорных точек плана эксперимента, вид ядерной функции, степень влияния ковариаты на функцию надёжности и значение ковариаты, для которой строится оценка.

2. Методами статистического моделирования исследованы свойства оценки Берана при использовании оптимальных значений параметра размытости, получаемых предложенным адаптивным алгоритмом. Показано, что по сравнению с ранее известными методами выбора параметра размытости применение его оптимальных значений позволяет получать более точные оценки Берана.

3. Предложены новые критерии согласия для проверки гипотезы о виде параметрической регрессионной модели надёжности, основанные на использовании оценки Берана. Методами статистического моделирования показано, что мощность предложенных критериев относительно конкурирующих гипотез, соответствующих другому виду регрессионной зависимости, превосходят мощность критерия типа Колмогорова, применяемого к выборкам остатков.

4. На основе результатов исследований разработан модуль программной системы статистического анализа данных типа времени жизни «LiTiS». Версия данной системы, в которой реализованы предложенный алгоритм выбора оптимального параметра размытости для оценки Берана и критерии согласия зарегистрирована в виде объекта интеллектуальной собственности как программа для ЭВМ.

Результаты проведенных исследований и разработанное программное обеспечение были внедрены в практику деятельности ООО «ДГ-Софт» и филиала ПАО «Электросетьсервис ЕНЭС», а также нашли практическое применение в учебном процессе на факультете прикладной математики и информатики ФГБОУ ВО «Новосибирский государственный технический университет», что подтверждается соответствующими актами о внедрении.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Издания из Перечня ВАК ведущих рецензируемых научных изданий для опубликования основных научных результатов диссертаций:

1. Демин В.А. Выбор оптимального параметра сглаживания для непараметрической оценки регрессионной модели надежности / В.А. Демин, Е.В. Чимитова // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2013. – № 1. – С. 59-65.

2. Демин В.А. Исследование метода выбора оптимального параметра сглаживания при непараметрическом оценивании регрессионных моделей надежности / В.А. Демин, Е.В. Чимитова, В.Ю. Щеколдин // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2014. – № 2 (27). – С.10-18.

3. Demin V. A method for selection of the optimal bandwidth parameter for Beran's nonparametric estimator / E. Chimitova, V. Demin // Topics in statistical simulation: research papers from the 7th intern. workshop on statistical simulation. - New York, Heidelberg, Dordrecht, London: Springer, 2014. – P. 139-147. – (Book series: Springer Proceedings in Mathematics and Statistics; vol. 114).

4. Демин В.А. Разработка и исследование критериев согласия для параметрических регрессионных моделей надежности на основе оценки Берана // Доклады АН ВШ РФ. – № 2(31) — Новосибирск. – 2016.– С. 43-56.

Сборники научных трудов:

5. Демин В.А. Непараметрические оценки и критерий согласия для регрессионных моделей / В.А. Демин, Е.В. Чимитова // Материалы всероссийской научной конференции молодых учёных «Наука. Технологии. Инновации». – Новосибирск, 2011. – С. 93-95.

6. Демин В.А. Непараметрическая оценка регрессионной модели надёжности. Материалы конференций / В.А. Демин, Е.В. Чимитова // Обработка информационных сигналов и математическое моделирование. – Новосибирск. – 2012. – С. 29-31.

7. Demin, V. Selection of the optimal smoothing parameter for the nonparametric estimation of the regression reliability model / V. Demin, E. Chimitova // Applied methods of statistical analysis. Applications in survival analysis, reliability and quality control – AMSA'2013, Novosibirsk, 25–27 Sept. 2013: proc. of the intern. workshop. – Novosibirsk: NSTU publ., 2013. – P. 83-91.

8. Демин В.А. Критерии согласия в задачах проверки адекватности параметрических моделей надежности и выживаемости / М.А. Семёнова, В.А. Демин, Е.В. Чимитова // Материалы Российской научно-технической конференции «Обработка информационных сигналов и математическое моделирование». – Новосибирск, 2013. – С. 38-40.

9. Демин В.А. Разработка критерия согласия для регрессионной модели надежности на основе оценки Берана / Е.В. Чимитова, В.А. Демин, Т.А. Лисичкина // Актуальные проблемы электронного приборостроения (АПЭП–2014) = Actual problems of electronic instrument engineering (APEIE–2014) : тр. 12 междунар. конф., Новосибирск, 2–4 окт. 2014 г. : в 7 т. – Новосибирск : Изд-во НГТУ, 2014. – Т. 6. – С. 103-107.

10. Demin V. An adaptive method for selecting an optimal bandwidth parameter in nonparametric estimate of the conditional reliability function / V. Demin, E. Chimitova // Applied Methods of Statistical Analysis. Nonparametric Approach - AMSA'2015, Novosibirsk, Russia, 14-19 September, 2015: Proceedings of the International Workshop. - Novosibirsk: NSTU publisher, 2015. – P. 212-219.

11. Демин В.А. Адаптивный метод выбора параметра размытости для оценки Берана // Материалы всероссийской научной конференции молодых учёных «Наука. Технологии. Инновации». – Новосибирск, 2015. – С. 17-19.

12. Demin V. A goodness-of-fit test based on the Beran estimator / V. Demin, E. Chimitova // 11 International forum on strategic technology (IFOST 2016) : proc., Novosibirsk, 1–3 June 2016. – Novosibirsk : NSTU, 2016. – Pt. 1. – P. 483-487.

Свидетельство о государственной регистрации программы для ЭВМ:

13. Чимитова Е.В., Румянцев А.В., Семёнова М.А., Галанова Н.С., Демин В.А. Система статистического анализа данных типа времени жизни «LiTiS 1.2» // Свидетельство о государственной регистрации программы для ЭВМ № 2014661905. – М.: Роспатент. – 2015.

Подписано в печать 20.03.2017 г. Формат 60 х 84 х 1/16
Бумага офсетная. Тираж 100 экз. Печ. л. 1.5.
Заказ № 2039

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр-т К. Маркса, 20