

*Московский государственный университет имени М.В. Ломоносова*

*Факультет вычислительной математики и кибернетики*

*Кафедра математической статистики*

Птицын Евгений Генрихович

**Алгоритм Скользящего Разделения Смесей и его применение в анализе реальных данных.**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Научный руководитель:**

д.ф.-м.н., заслуженный профессор

В.Ю. Королёв

Оглавление

[1 Введение 3](#_Toc6928461)

[1.1 Актуальность темы исследования 3](#_Toc6928462)

[1.1.1 Анализ данных в современном мире 3](#_Toc6928463)

[1.2 Используемые подходы и методы 5](#_Toc6928464)

[1.3 Цель 5](#_Toc6928465)

[1.4 Краткое содержание работы 5](#_Toc6928466)

[1.5 Основные результаты 5](#_Toc6928467)

[2 Вводные понятия 5](#_Toc6928468)

[2.1 Смесь нормальных законов распределения 5](#_Toc6928469)

[3 ЕМ-алгоритм 5](#_Toc6928470)

[4 Скользящее разделение смесей 5](#_Toc6928471)

[4.1 Описание метода 5](#_Toc6928472)

[4.2 Декомпозиция волатильности финансовых индексов 5](#_Toc6928473)

[4.2.1 Волатильность 5](#_Toc6928474)

[4.3 Изучение турбулентности 5](#_Toc6928475)

[5 Заключение 5](#_Toc6928476)

[6 Список литературы 5](#_Toc6928477)

[7 Приложение А 5](#_Toc6928478)

# Введение

## Актуальность темы исследования

### Анализ данных в современном мире

#### Наука о данных

Наука о данных (англ. data science; иногда даталогия — datalogy) — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных.

Началом формирования выделенной дисциплины считается 1966 год, когда был учреждён Комитет по данным для науки и техники (CODATA)[6], а первое введение термина data science относится к книге Петера Наура 1974 года, в которой он явно определил науку о данных как дисциплину, изучающую жизненный цикл цифровых данных — от появления до преобразования для представления в других областях знаний[7] (существует мнение, что Наур употреблял термин «data science» ещё в конце 1960-х[8]).

Однако, только в 1990-е годы термин, обозначающий дисциплину, получил широкое употребление[9][6], и только в начале 2000-х стал общепризнанным, прежде всего, благодаря статье статистика Bell Labs Уильяма Кливленда (по состоянию на 2012 год — профессор статистики в Университете Пердью), в которой он опубликовал план развития технических аспектов статистических исследований и выделил науку о данных как отдельную академическую дисциплину, в которой эти технические аспекты должны быть сконцентрированы[10][11].

В 2002 году Комитетом по данным для науки и техники начат выпуск журнала CODATA Data Science Journal, содержащего в названии наименование дисциплины, а в январе 2003 года вышел первый номер The Journal of Data Science Колумбийского университета.

Очередной взлёт широкого интереса к науке о данных относится к появлению парадигмы «больших данных», которая фокусируется на новых технологических возможностях обработки данных больших объёмов и разнообразия, в том числе, за счёт применения методов, разрабатываемых в 2000-е годы в науке о данных. С 2011 года O’Reilly проводит серию крупных конференций по науке о данных — Strata[12], корпорация EMC начиная с 2011 года проводит ежегодной саммит по науке о данных[13]. McKinsey в 2011 году спрогнозировал спрос в США на 440—490 тыс. новых специалистов с «глубокими аналитическими навыками по работе с большими данными» к 2018 году и дефицит в 50 % — 60 % в таких специалистах при сохранении образовательных трендов[14], в связи с этим прогнозом во многом был подогрет интерес к созданию учебных программ[15].

В 2012 году профессия data scientist неоднократно отмечается как одна из самых привлекательных (англ. sexy) и перспективных в современном мире, утверждается, что такие специалисты будут играть ключевую роль в организациях, за счёт возможностей получения конкурентных преимуществ благодаря анализу, быстрой обработке и извлечению закономерностей в данных, прежде всего, в технологических отраслях[16][5].

С 2013 учебного года Университет Данди, Оклендский университет, Университет Южной Калифорнии запустили магистерские программы по науке о данных, а бизнес-школа Имперского колледжа Лондона — программу подготовки «магистров наук по науке о данных и менеджменту» (англ. MSc Data Science & Management)[17]. В том же году Вашингтонский университет, Университет Калифорнии в Беркли и Нью-Йоркский университет получили грант в размере $37,8 млн на развитие науки о данных, в рамках которого в течение пяти лет должны будут, в том числе, выстроить учебные программы и создать возможности для академической карьеры в данной области[18].

Основная практическая цель профессиональной деятельности в науке о данных — обнаружение закономерностей в данных[19], извлечение знаний из данных в обобщённой форме[20]. Для объяснения навыков, необходимых для деятельности в этой области, часто используется диаграмма Венна[21], на которой навыки, требуемые специалисту, отражены на пересечении сфер общепредметного опыта (англ. substantive expertise), практического опыта в информационных технологиях (hacking skills) и знания математической статистики[22].

В качестве эпистемологической особенности дисциплины указывается приоритет практической применимости результатов, то есть, успешности предсказаний, перед их причинностью, тогда как в традиционных исследовательских областях существенно объяснение природы явления[23]. В сравнении с классической статистикой, на методах которой во многом основывается и наука о данных, в ней подразумевается исследование сверхбольших разнородных массивов цифровой информации и неразрывная связь с информационными технологиями, обеспечивающими их обработку[24]. В сравнении с деятельностью в области проектирования и работы с базами данных, где предполагается предварительное проектирование модели данных, отражающей взаимосвязи предметной области и последующее исследование загруженных данных относительно простыми (арифметическими) методами, в науке о данных предполагается опора на аппарат математической статистики, искусственного интеллекта, машинного обучения, зачастую без предварительной загрузки данных в модели. В сравнении с профессией аналитика, основная цель деятельности которого в описании явлений на основе накопленных данных относительно простыми пользовательскими средствами (вроде электронных таблиц или средств класса Business Intelligence), профиль специалиста по науке о данных в меньшей степени требует концентрации на содержании предметных областей, но требует более глубоких знаний в математической статистике, машинном обучении, программировании, и в целом более высокого образовательного уровня (магистры, кандидаты наук, Ph.D в сравнении с бакалаврами и специалистами)[25].

[Наука о данных, ]

#### Data mining

Data mining (рус. добыча данных, интеллектуальный анализ данных, глубинный анализ данных) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Термин введён Григорием Пятецким-Шапиро в 1989 году[1][2][3].

Английское словосочетание «data mining» пока не имеет устоявшегося перевода на русский язык. При передаче на русском языке используются следующие словосочетания[4]: просев информации, добыча данных, извлечение данных, а также интеллектуальный анализ данных[5][6][7]. Более полным и точным является словосочетание «обнаружение знаний в базах данных» (англ. knowledge discovery in databases, KDD).

Основу методов data mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечёткой логики. К методам data mining нередко относят статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов, анализ выживаемости, анализ связей). Такие методы, однако, предполагают некоторые априорные представления об анализируемых данных, что несколько расходится с целями data mining (обнаружение ранее неизвестных нетривиальных и практически полезных знаний).

Одно из важнейших назначений методов data mining состоит в наглядном представлении результатов вычислений (визуализация), что позволяет использовать инструментарий data mining людьми, не имеющими специальной математической подготовки.

Применение статистических методов анализа данных требует хорошего владения теорией вероятностей и математической статистикой.

## Используемые подходы и методы

## Цель

## Краткое содержание работы

## Основные результаты

# Вводные понятия

## Смесь нормальных законов распределения

# ЕМ-алгоритм

[Королёв, 2011] [Gentle, Härdle, Mori, 2012]

# Скользящее разделение смесей

## Описание метода

## Влияние выбора ширины окна

## Декомпозиция волатильности финансовых индексов

### Волатильность

[Королёв, 2011]

## Изучение турбулентности

# Заключение

В данной работе было показано бла бла бла.

# Список литературы

1. Gentle J.E., Härdle W.K., Mori Y. Springer Handbooks of Computational Statistics. : Springer, 2012. 1192 с.

2. Королёв В.Ю. Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. Москва: Издательство Московского университета, 2011. 512 с.

3. Наука о данных [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Наука\_о\_данных (дата обращения: 23.04.2019).

# Приложение А

Рисунки схемы таблицы