

УДК 334

¹Д. А. Тамбиева, ²М. У. Эркенова

¹Ставропольский государственный аграрный университет, г. Ставрополь,
email: tamjannet@mail.ru

²Северо-Кавказская Государственная академия, г. Черкесск, email: madina033@mail.ru

ОБЗОР КОНЦЕПТУАЛЬНЫХ ПОДХОДОВ В АНАЛИЗЕ ДАННЫХ. ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ: ТЕОРИЯ И ПРАКТИКА

Ключевые слова: анализ данных, временные ряды, статистический анализ, детерминированный/стохастический процесс

В работе рассматриваются концептуальные подходы к анализу данных, в том числе, контексте проблемы детерминированность/стохастичность. Обозначено, что большинство новейших методов интеллектуального анализа данных являются алгоритмической имитацией биологических систем на базе хорошо известных методов статистического анализа данных. Некоторые, ставшие уже классическими методы и техники простейшего статистического анализа данных, остаются весьма эффективными для отдельных классов задач. Представлен статистический анализ реального экономического временного ряда (ВР), соответствующего квазинормальному распределению случайной величины. Продемонстрирована адекватность прогноза, построенного на базе начального момента первого порядка.

¹Д. А. Тамбиева, ²М. У. Эркенова

¹Stavropol State Agrarian University, Stavropol, email: tamjannet@mail.ru

²North Caucasus State Academy, Cherkessk, email: madina033@mail.ru

OVERVIEW OF CONCEPTUAL APPROACHES IN DATA ANALYSIS. ECONOMIC AND MATHEMATICAL METHODS: THEORY AND PRACTICE

Keywords: data analysis, time series, statistical analysis, deterministic/stochastic process

The paper discusses conceptual approaches to analysis, including in the context of the determinism/stochasticity problem. It is indicated that most of the latest methods of data mining are algorithmic imitation of biological systems based on well-known methods of statistical data analysis. Some methods and techniques of the simplest statistical data analysis that have already become classical remain very effective for certain classes of problems. A statistical analysis of a real economic time series (BP) corresponding to a quasi-normal distribution of a random variable is presented. The adequacy of the forecast based on the initial moment of the first order is demonstrated.

Двадцать первый век – это эпоха больших данных. Глобальный связанный мир предлагает бесконечное множество способов генерирования, сбора и хранения данных. Никогда раньше человек не имел доступа к такому количеству информации и только сейчас начинают появляться адекватные методы анализа и прогнозирования поведения исследуемого процесса и/или системы на краткосрочную, среднесрочную и долгосрочную перспективу [1].

Появились новые концепции науки о данных, машинного обучения и глубокому обучению, которые предлагают новый набор методов. Эти методы уже очень активно используют ведущие компании для изучения тенденций в дина-

мике интересов клиентов, поставщиков, конкурентов и др.

Фундаментом для развития современных методов анализа данных являются методы визуализации, математическая статистика, теория вероятностей, методы искусственного интеллекта, машинного обучения и т.д. [2-16].

Теория и практика применения этих методов нацелены на выявление и формализацию связей в переменных данных.

Экспертные оценки рынка программных продуктов позволяют сделать вывод о потенциальной коммерческой успешности проектов разработки программного обеспечения в области анализа данных.

Направления, связанные с анализом данных, весьма востребованы на круп-

ных предприятиях, тогда как средний и малый бизнес до сих пор предпочитает интуитивный подход в анализе доступной информации. Это объясняется тем, что предприятия и организации среднего и малого бизнеса не могут позволить себе приобретение дорогостоящего программного обеспечения от ведущих разработчиков и/или найм узких специалистов, владеющих современными инструментами анализа данных. При этом возникающие как «лавины» и требующие внимания со стороны лица принимающего решение потоки новой информации, вынуждают руководителей предприятий и организаций среднего и малого бизнеса искать новые способы быстрого и эффективного анализа доступной информации и в ближайшем будущем будут создавать спрос на рынке программных продуктов на недорогие, но отвечающие запросам потенциальных клиентов, современные информационные системы анализа данных.

Наш современный информационный век приводит к динамичному и чрезвычайно высокому росту мира интеллектуального анализа данных. Нет сомнения, что она требует адекватных и эффективных различных типов методов анализа данных, техник и инструментов, способных реагировать на постоянно растущие потребности бизнеса в исследованиях.

На самом деле интеллектуальный анализ данных не имеет собственных методов анализа данных. Он использует методологии и методы других смежных областей науки.

Среди методов, используемых в анализе малых и больших данных, можно выделить:

- Математические и статистические методы
- Методы, основанные на искусственном интеллекте, машинном обучении
- Визуализация и графический метод и инструменты

Рассмотрим наиболее известные классические и современные типы методов и моделей анализа данных.

Математико-статистические методы анализа данных

Математические и статистические науки могут многое дать управлению и анализу интеллектуального анализа данных.

Большинство методов интеллектуального анализа данных являются инструментами статистического анализа данных. Некоторые методы и техники хорошо известны и очень эффективны.

1. Описательный анализ

Описательный анализ – это проникновение в прошлое. Этот статистический метод делает именно то, что следует из названия – «Описывает». Он смотрит на данные и анализирует прошлые события и ситуации, чтобы получить представление о том, как подойти к будущему [2 – 4].

Описательная аналитика рассматривает прошлые/исторические показатели, чтобы понять причины прошлых неудач или успехов.

Это позволяет нам извлекать уроки из прошлого поведения и выяснять, как оно может повлиять на будущие результаты.

2. Регрессионный анализ

Регрессия – это один из наиболее популярных видов методов анализа данных, который позволяет выявлять связь между зависимой переменной и одной и/или несколькими независимыми переменными.

Данный метод используется в прогнозировании, в интеллектуальном анализе данных, для расчета значений, заданных конкретным набором данных. Существует огромное количество различных типов регрессионных моделей, таких как линейные регрессионные модели, множественная регрессия, логистическая регрессия, гребневая регрессия, нелинейная регрессия, регрессия жизненных данных и многие-многие другие. Например, регрессия может быть использована для прогнозирования цены продукта с учетом других переменных.

Существует огромное количество различных типов регрессионных моделей, таких как линейные регрессионные модели, множественная регрессия, логистическая регрессия, гребневая регрессия, нелинейная регрессия, регрессия жизненных данных и многие-многие другие [5-7].

3. Факторный анализ

Факторный анализ – это метод регрессионного анализа данных, исполь-

зуемый для поиска базовой структуры в наборе переменных. Это связано с поиском новых независимых факторов (переменных), описывающих закономерности и модели отношений между исходными зависимыми переменными [8-11].

Факторный анализ является очень популярным инструментом для исследования переменных взаимосвязей для таких сложных тем, как психологические шкалы и социально-экономический статус.

Факторный анализ – это основной шаг на пути к эффективным процедурам кластеризации и классификации.

4. Дисперсионный анализ

Дисперсионный анализ не является столь распространенным методом, используемым в интеллектуальном анализе данных, но все же играет в нем определенную роль. Дисперсия – это разброс, на который растягивается набор данных. Это метод описания того, насколько расширен набор данных [8-11].

Мера дисперсии помогает ученым данных изучать изменчивость вещей.

Как правило, дисперсия имеет два значения: во-первых, она представляет собой вариацию вещей между собой, а во-вторых, она представляет вариацию вокруг среднего значения. Если разница между значением и средним значительна, то дисперсия высока. В противном случае она низкая.

5. Дискриминантный анализ

Дискриминантный анализ – один из самых мощных методов классификации в интеллектуальном анализе данных. Дискриминантный анализ использует переменные измерения для различных групп элементов, чтобы выделить точки, которые различают эти группы.

Эти измерения используются для классификации новых элементов.

Типичными примерами использования этого метода являются: классификация заявок на кредитные карты по категориям низкого и высокого риска, классификация клиентов новых продуктов на различные группы, медицинские исследования и т. д.

6. Анализ Временных Рядов

Как известно, почти во всех научных областях измерения проводятся с течением

времени. Эти наблюдения приводят к сбору организованных данных, известных как временные ряды.

Хорошим примером временных рядов является дневное значение индекса фондового рынка.

Анализ данных временных рядов – это процесс моделирования и объяснения временных рядов точек данных, зависящих от времени. Цель состоит в том, чтобы извлечь всю значимую информацию (статистику, правила и паттерны) из формы данных.

Впоследствии эта информация используется для создания и моделирования прогнозов, способных предсказывать будущие эволюции.

Методы, основанные на искусственном интеллекте, машинном обучении и эвристических алгоритмах

Современные методы анализа данных, основанные на Искусственном интеллекте, Машинном обучении и Эвристических алгоритмах и др. привлекают внимание исследователей своими расширенными возможностями и способностью решать нетрадиционные задачи. Кроме того, они могут быть легко и эффективно реализованы и выполнены специальными программными системами и инструментами.

Перечислим наиболее популярных из этих типов методов анализа данных:

1. Искусственные Нейронные Сети

Без сомнения, это один из самых популярных новых и современных методов анализа данных.

«Нейтральные сети – это красивая биологически вдохновленная парадигма программирования, которая позволяет компьютеру учиться на наблюдательных данных».

Искусственные нейронные сети (ИНС), часто называемые просто «нейронными сетями» концептуально – это процесс обработки информации в ИНС имитирующий работу взаимосвязанной группы искусственных нейронов. По существу, это организация вычислительного процесса по аналогии с работой головного мозга человека. Важным свойством передовые программные решения ИНС является встроенная процедура

обучения / самообучения нейронной сети. Последнее, зачастую реализуется в виде адаптивной системы, способной легко менять свою структуру на основе информации, протекающей через сеть [12-16].

Применение нейронных сетей в интеллектуальном анализе данных очень широко. Они обладают высокой приемной способностью для зашумленных данных и высокой точностью. Подробно исследуется интеллектуальный анализ данных на основе нейронных сетей. Было показано, что нейронные сети являются очень перспективными системами во многих приложениях прогнозирования и бизнес-классификации.

2. Деревья решений

Дерево решений – это древовидная диаграмма, представляющая классификационную или регрессионную модель. Это еще один очень популярный и современный алгоритм классификации в интеллектуальном анализе данных и машинном обучении.

Данный подход делит набор данных на все меньшие и меньшие поднаборы данных (которые содержат экземпляры с аналогичными значениями), в то время как постоянно развивается связанное дерево решений. Дерево построено, чтобы показать, как и почему один выбор может привести к следующему с помощью ветвей.

Среди преимуществ использования деревьев решений можно отметить отсутствие необходимости знания предметной области; они (деревья) просты для понимания; этапы классификации дерева решений также очень просты и быстры [12-16].

3. Эволюционное программирование

Эволюционное программирование в интеллектуальном анализе данных – это общая концепция, которая объединяет множество различных типов анализа данных с использованием эволюционных алгоритмов. Наиболее популярными из них являются: генетические алгоритмы, генетическое программирование и коэволюционные алгоритмы.

На самом деле многие агентства по управлению данными применяют эволюционные алгоритмы для решения не-

которых из самых больших в мире проблем, связанных с большими данными.

К числу преимуществ эволюционных методов относятся:

- это независимые от предметной области техники
- они обладают способностью исследовать большие пространства поиска и находить хорошие решения
- они относительно нечувствительны к шуму
- может отлично управлять взаимодействием атрибутов.

4. Нечеткая логика

Математический аппарат нечеткой логики применяется для решения проблем с неопределенностями в задачах интеллектуального анализа данных. Нечеткое логическое моделирование – один из вероятностных методов и приемов анализа данных.

Это относительно новая область, но имеет большой потенциал для извлечения ценной информации из различных наборов данных.

Нечеткая логика применима, когда модель содержит параметры, значения которых не могут быть точно определены или эти значения содержат слишком высокий уровень шума.

Нечеткая логика является одним из инновационных типов многозначной логики решения задач с неопределенностями. В отличие от традиционной математической логики, где возможных значений только 2 – это 0 и 1, в нечеткой логике истинностные значения переменных являются вещественным числом и могут принимать различные значения в диапазоне между 0 и 1.

В этом термине значение истинности может варьироваться в соответствии со значением функции принадлежности от полнотой истинного до полностью ложного.

Для реализации адекватных математических моделей прогнозирования необходимо провести предварительный статистический анализ данных, цель которого – идентификация структуры данных в контексте классификационных признаков: марковский и/или детерминированный процесс. Проверка соответствия функции распределения случайной величины нормальному или квазинормальному закону.

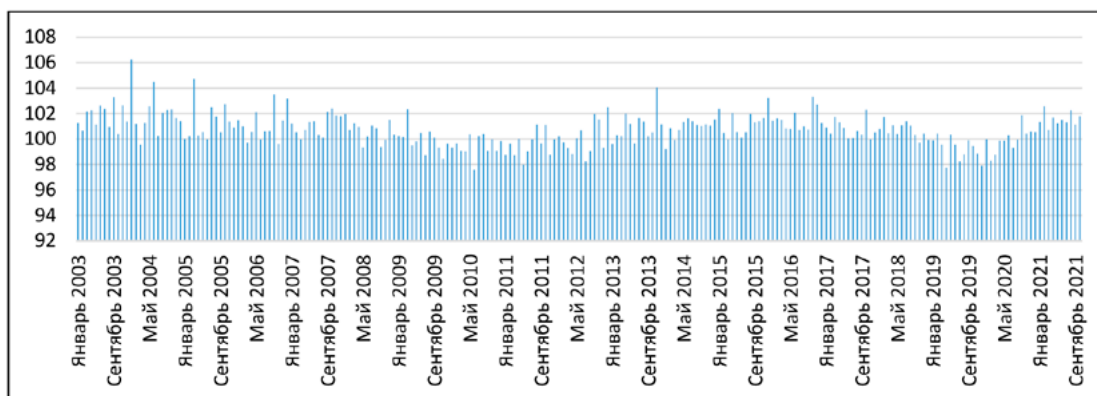


Рис. 1. Временной ряд индексов потребительских цен на продовольственные товары одного из регионов РФ (Карачаево-Черкесской республике) – BP DPI

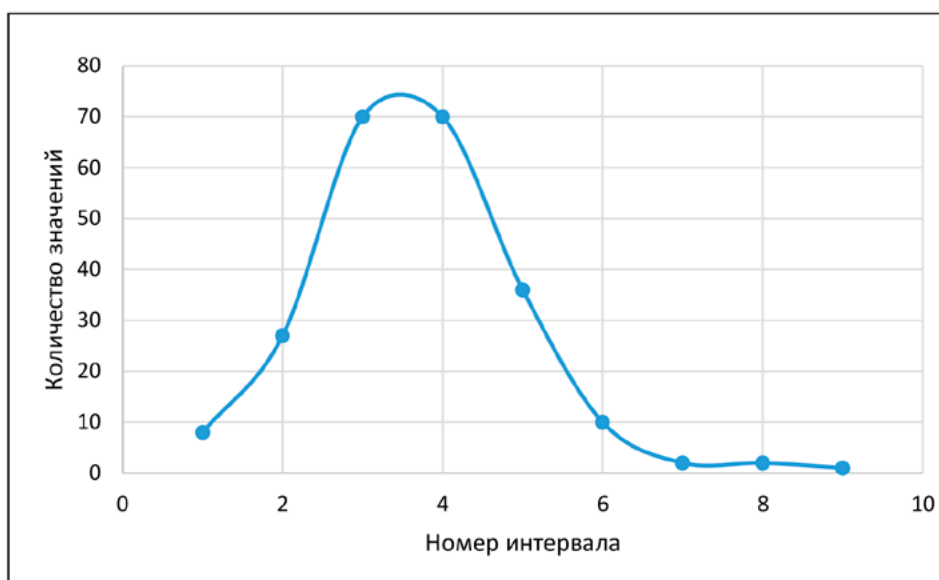


Рис. 2. Функция распределения случайной величины BP DPI

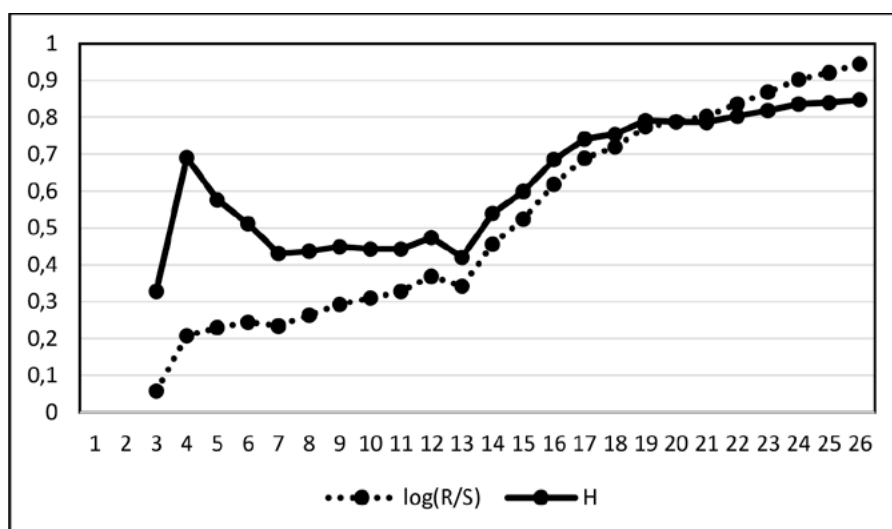


Рис. 3. Результат реализации фрактального анализа (R/S-анализа).

Рассмотрим ВР индексов потребительских цен на продовольственные товары одного из регионов РФ (Карачаево-Черкесской республике), который обозначим как ВР DP1 (см. Рис. 1).

В результате проведенной расчетно-графической работы были получены следующие результаты.

Таблица 1

Статистические показатели временного ряда (ВР) DP1

Статистический показатель	Значение
Минимум (min)	97,59
Максимум (max)	106,26
Размах (R)	8,67
Матожидание (MX)	100,73
Дисперсия (DX)	1,59
СКО (SX)	1,264
Коэффициент вариации (VX)	1%

Для построения функции распределения случайной величины была использована формула Стерджеса [17], в соответствии с которой было произведено разбиение временного ряда DP1 на интервалы. Отметим, что количество наблюдений во временном ряде DP1 равно 226 ($n=226$).

Количество интервалов согласно формулы Стерджеса:

$$s \approx 1 + 3,322 \cdot \lg n \quad (1)$$

Для ВР DP1 равно 8,8 с шагом 0,98.

Полученному разбиению случайной величины X ВР DP1 на интервалы соответствует распределение, представленное в таблице 2 и рисунке 2.

Асимметрия и эксцесс функции распределения случайной величины X ВР DP1 соответственно принимают значения 0,51 и 1,56.

В дополнение ко всему вышеизложенному отметим, что для целей распознавания исследуемого процесса и/или явления в контексте понятий детерминированность/стохастичность, специалисты в области анализа данных активно используют фрактальный анализ [18-21]. На рис. 3 представлены графики R/S- и H-траекторий – результат реализации одного из алгоритмов фрактального анализа (R/S-анализа).

Таблица 2

Разбиение временного ряда DP1 на интервалы, согласно формуле Стерджеса (1)

№ интервала	Диапазон интервала	Количество значений
1	97,59-98,57	8
2	98,57-99,56	27
3	99,56-100,54	70
4	100,54-101,52	70
5	101,52-102,50	36
6	102,5-103,49	10
7	103,49-104,47	2
8	104,47-105,45	2
9	105,45-106,44	1

В целом R/S-анализ позволяет диагностировать состояние наблюдаемого процесса и/или явления, в том числе предсказывать нестабильные состояния, периодичность и фрактальность ВР.

R/S- и H – траектории ВР DP1

Проведение R/S-анализа предусматривает наблюдение за значениями контролируемой величины и оценивание отличий этих значений от появляющихся случайно (от «случайных блужданий», например, гауссовских значений контролируемого показателя) [20 стр.86].

$$\frac{R(i)}{S(i)} = A i^H(i) \quad (2)$$

Здесь i – номер наблюдения ВР DP1 $i=3, 4, \dots, n$ из общего числа членов n и введены следующие обозначения:

– A – константа, которую можно оценить как свободный элемент уравнения регрессии,

– $S(i)$ – стандартное отклонение:

$$S(i) = \sqrt{\frac{1}{i-1} \sum_{k=1}^i (X_k - \bar{X}_i)^2}$$

($\lambda(t) = \lambda$ – текущая среднеарифметическая концентрация загрязняющего вещества в воде, $k = 3, 4 \dots i$),

– $R(i)$ – нормированный размах вариации:

$$R(i) = \max_{1 \leq k \leq i} (X_{i,1}) - \min_{1 \leq k \leq i} (X_{i,1}),$$

представляющий собой разность между максимальным и минимальным накопленными отклонениями:

$$\overline{X}_{i,1} = \sum_{k=1}^i (X_k - \overline{X}_j),$$

а $j = \overline{1, i}$ – среднее между 1-м и i -м членами ВР,

– H – показатель Херста, важнейшая характеристика выражения (2), определяющая поведение ВР. При $H > 0,5$ исследуемый ряд является персистентным, и для него характерно наличие долговременной памяти ВР [17-20]. Что указывает на его (ВР) фрактальность.

Если $H < 0,5$, то ВР антиперсистентный, т.е. склонен к «возврату к среднему».

Отметим, что ВР индексов потребительских цен на продовольственные товары Карачаево-Черкесской республике (см. Рис. 1) согласно фрактальному анализу можно идентифицировать как марковский процесс, при этом простейший статистический анализ данных ВР DP1 позволяет получить адекватный прогноз на краткосрочную перспективу.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-37-90102

Библиографический список

1. Форман Дж. Много цифр: Анализ больших данных при помощи Excel. М.: Альпина Паблишер, 2019. 461 с.
2. Миркин Б. Г. Введение в анализ данных. М.: Юрайт. 2020. 175 с.
3. Козлов А.Ю., Мхитарян В.С., Шишов В.Ф. Статистический анализ данных в MS Excel: Учебное пособие. М.: Инфра-М, 2018. 80 с.
4. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере: Учебное пособие. Науч. ред. В.Э. Фигурнов. М.: ИД ФОРУМ, 2017. 368 с.
5. Халафян А.А., Боровиков В.П., Калайдина Г.В. Теория вероятностей, математическая статистика и анализ данных. Основы теории и практика на компьютере. Statistica. Excel. Более 150 примеров решения задач. Учебное пособие. М.: Ленанд. 2017. 320 с.
6. Чашкин Ю.Р. Математическая статистика. Анализ и обработка данных: Учебное пособие / Под ред. С.Н. Смоленский. Рн/Д: Феникс, 2017. 236 с.
7. Айзек М.П. Графика, формулы, анализ данных в Excel. Пошаговые примеры. СПб.: Наука и техника, 2019. 384 с.
8. Мхитарян В.С. Теория планирования эксперимента и анализ статистических данных. М.: Юрайт. 2020. 491 с.
9. Сидняев Н.И. Теория планирования эксперимента и анализ статистических данных. М.: Юрайт. 2020. 496 с.
10. Ниворожкина Л.И., Арженовский С.В., Рудяга А.А. Статистические методы анализа данных: Учебник. М.: Риор, 2018. 320 с.
11. Симчера В.М. Методы многомерного анализа статистических данных. М.: Финансы и статистика, 2018. 400 с.
12. Гашев С.Н. Математические методы в биологии: анализ биологических данных в системе Statistica. М.: Юрайт. 2020. 208 с.
13. Кулаичев А.П. Методы и средства комплексного анализа данных: Учебное пособие. М.: Форум, 2018. 160 с.
14. Макшанов А.В., А.Е. Журавлев Технологии интеллектуального анализа данных: Учебное пособие. СПб.: Лань, 2018. 212 с.
15. Макшанов А.В. Технологии интеллектуального анализа данных. М.: Лань. 2019. 212 с.
16. Рафалович В. Data mining, или интеллектуальный анализ данных для занятых. Практический курс. М.: SmartBook, 2018. 352 с.
17. Sturges H. The choice of a class-interval. J. Amer. Statist. Assoc. 1926. Vol. 21. P. 65-66.
18. Голубев С.Н. R/S – анализ стабильности запаздывающего временного ряда (недоступная ссылка) // Лабораторный журнал: электрон. научн.-практич. журн. 2013. № 1 (1).
19. Петерс Э. Фрактальный анализ финансовых рынков. Применение теории хаоса в инвестициях и экономике. М.: Интернет-трейдинг, 2004. 304 с.
20. Петерс Э. Хаос и порядок на рынках капитала. М.: Мир, 2000. 333 с.
21. Ширяев А.Н. Основы стохастической финансовой математики. М.: Фазис, 1998. Т. 1. 512 с.