DOI: 10.38050/2078-3809-2025-17-4-9-34

*Вопросы теории*

# ИННОВАЦИОННЫЕ ПОДХОДЫ К ГЕНЕРАЦИИ ОБУЧАЮЩИХ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ СПРОСА НА НЕФТЬ

**Манахова Ирина Викторовна**
***доктор экономических наук, профессор,***
***МГУ имени М.В. Ломоносова, экономический факультет***
***(г. Москва, Россия)***

**Матыцын Александр Владимирович**
***магистр,***
***Высшая школа внешней торговли***
***(г. Париж, Франция);***
***соискатель ученой степени кандидата экономических наук,***
***МГУ имени М.В. Ломоносова, экономический факультет***
***(г. Москва, Россия)***

## Аннотация

*В данной работе исследуются методы генерации обучающих данных для повышения точности прогнозирования спроса на рынке нефти. Рассматриваются ограничения традиционных подходов и обосновывается применение генеративно-состязательных сетей, в частности модели TimeGAN (Time-series Generative Adversarial Network), для создания синтетических временных рядов. Результаты показывают, что TimeGAN позволяет генерировать реалистичные данные, приближенные к реальным, с сохранением волатильности и структурных особенностей рынка. Также выявлены ограничения модели, требующие дальнейшего исследования для повышения эффективности и точности прогнозирования спроса на нефть в условиях рыночной нестабильности.*

**Ключевые слова:** глубокое обучение, генеративно-состязательные сети, обучающие данные, модель TimeGAN.

**JEL коды:** E17, C53.

**Для цитирования:** Манахова И.В., Матыцын А.В. Инновационные подходы к генерации обучающих данных для прогнозирования спроса на нефть // Научные исследования экономического факультета. Электронный журнал. 2025. Том 17. Выпуск 4. С. 9-34. DOI: 10.38050/2078-3809-2025-17-4-9-34.

Манахова И.В., e-mail: manakhovaiv@mail.ru
Матыцын А.В., e-mail: avmatytsyn@gmail.com

## Введение и актуальность

В современном мире высокоуровневое прогнозирование спроса играет важнейшую роль в управлении цепочками поставок, оптимизации уровня запасов и повышении общей операционной эффективности. Несмотря на то, что традиционные методы прогнозирования являются полезными, их использование ограничивается изменчивостью современной динамики рынка (Розенцвайг, 2014). Поскольку компании продолжают бороться с проблемами быстро развивающегося рынка, интеграция методов глубокого обучения (далее ГО) является перспективным решением для повышения точности прогнозирования спроса (Каукин и др., 2023; Копытин, 2024; Farzana, Prakash, 2020). Однако, специфика прогнозирования спроса на рынке нефти указывает на сравнительно небольшой объем обучающих данных, что вносит ряд ограничений в использование моделей ГО. По мнению авторов, возможным решением данной проблемы являются методы генерации обучающих данных.

На сегодняшний день существует значительное количество методологических подходов к генерации обучающих данных, включая такие методы, как генерация на основе геометрического броуновского движения (далее GBM), авторегрессионная интегрированная скользящая средняя (далее ARIMA), бутстрэппинг на основе перемешивания блоков (далее MBB), а также добавление гауссовского шума к простому скользящему среднему (далее GN+SMA). Несмотря на очевидные преимущества указанных методов, с точки зрения авторов, они не обеспечивают должного уровня эффективности при формировании обучающих выборок для моделирования спроса на рынке нефти из-за ряда ограничений, вызванных структурными особенностями архитектуры моделей. Одним из возможных решений данной проблемы могут стать генеративно-состязательные сети (GAN). Применение данного подхода к генерации обучающих данных представляет перспективу значительного повышения точности прогнозных моделей, особенно в контексте задач прогнозирования спроса на нефтяном рынке.

В рамках данного исследования авторы выдвигают следующие гипотезы.

1. Использование модели TimeGAN для генерации синтетических временных рядов позволяет создать обучающую выборку, сохраняющую ключевые статистические и динамические свойства реальных данных.

2. Дополнение обучающей выборки синтетическими данными, сгенерированными с помощью TimeGAN, приводит к повышению точности прогноза модели MLP по сравнению с использованием только реальных данных.

3. TimeGAN обеспечивает более высокое качество синтетических данных для задач прогнозирования спроса на нефть по сравнению с классическими методами генерации (GBM, ARIMA, MBB, GN+SMA).

Для формирования убедительной доказательной базы, необходимой для подтверждения или опровержения выдвинутых гипотез, в рамках данного исследования авторами были поставлены следующие научные задачи.

1. Проанализировать характеристики временного ряда, отражающего динамику цен на нефть марки «Brent», с акцентом на выявление ключевых структурных особенностей, включая сезонные колебания, волатильность и факторы, определяющие долгосрочные и краткосрочные тренды.

2. Осуществить сравнительный анализ различных методов генерации синтетических временных рядов, включая TimeGAN, ARIMA, стохастические и эмпирические подходы, с целью оценки их применимости к задачам расширения обучающих выборок в условиях ограниченности исходных данных.

3. Реализовать и обучить модель генеративно-состязательной сети, адаптированной к особенностям нефтяного рынка, а также провести эмпирическую оценку качества сгенерированных данных с использованием статистических и прогнозных метрик.

Оценить влияние синтетических данных на точность прогностических моделей путем обучения на выборках с реальными и дополненными данными, а также провести количественный анализ прироста прогностической точности.

## 1. Научный обзор релевантной литературы

В рамках исследования выделен ряд актуальных работ, которые посвящены прогнозированию на основе методов глубокого обучения, с акцентом на применение генеративно-состязательных сетей.

Одним из таких исследований является работа «Volatility and Irregularity Capturing in Stock Price Indices Using Time Series Generative Adversarial Networks (TimeGAN)» (Mushunje et al., 2023), в которой авторы прогнозируют динамику цен на акции с учетом волатильности и различных нелинейных движений в индексах цен, особенно при наличии непредвиденных событий, таких как пандемия COVID-19.

Исследователи выдвинули предположение о том, что в таком непредсказуемом условии, как пандемический шок, цены на акции подвержены различным скачкам и колебаниям, что приводит к появлению неоднородных тенденций в данных временного ряда. Традиционные модели прогнозирования не всегда способны учитывать подобные особенности, поэтому их использование может приводить к значительным ошибкам прогнозирования при работе с временными рядами, подверженным резким изменениям. Таким образом, авторы считают, что для прогнозирования цен на акции в условиях пандемического шока необходимы эффективные и надежные инструменты, такие как генеративно-состязательные сети (GAN).

Генеративная модель TimeGAN (Time-series Generative Adversarial Network), разработанная для временных рядов, которые имеют зависимость данных от времени, была использована авторами для создания синтетических временных рядов. Данные синтетические ряды приближены к реальным данным о ценах на акции и учитывают их волатильное поведение. Проведя обучение модели TimeGAN на основании исторических данных, включающих колебания фондового индекса DAX с 2010 по 2022 г., исследователям удалось установить основные закономерности временных рядов. Обучение модели состояло из таких элементов, как генератор, дискриминатор, сеть встраивания и сеть восстановления, которые необходимы для изучения зависимостей, временной структуры и волатильности временных рядов. Было выявлено, что модель TimeGAN способна улавливать резкие скачки и нелинейность, которые свойственны финансовым данным, благодаря чему TimeGAN является эффективным и подходящим инструментом для создания синтетических данных, которые отражают реальную динамику рынка с учетом волатильности. Авторы заключили, что благодаря использованию модели TimeGAN минимизируются ошибки прогнозирования, а также улучшается устойчивость

других прогнозных моделей, таких как LSTM и GRU. Таким образом, данная исследовательская работа выявила эффективность модели TimeGAN в воспроизводстве волатильности временных рядов, что делает ее более точной в сравнении с традиционными прогностическими моделями.

Позже тема использования модели TimeGAN для более точных прогнозов представлена в научной статье «Enhancing Short-Term Power Load Forecasting for Industrial and Commercial Buildings: A Hybrid Approach Using TimeGAN, CNN, and LSTM» (Liu et al., 2023). В данном исследовании авторами используется гибридная модель, сочетающая TimeGAN, сверточные нейронные сети (CNN), а также долгую краткосрочную память (LSTM) для повышения точности прогнозирования электрической нагрузки для промышленной и коммерческой инфраструктуры. Эта работа направлена на решение проблем прогнозирования краткосрочного спроса на электроэнергию вследствие изменения характера нагрузки из-за колебаний спроса, сезонных тенденций и других факторов.

В контексте данного исследования модель TimeGAN используется для создания синтетических временных рядов, которые позволяют дополнить исходный набор данных и устранить пробелы в существующих данных. Для проведения исследования авторы использовали исходные данные о нагрузке на электросети на четырех различных типах промышленных и коммерческих зданий в рамках двухмесячного периода. Однако для эффективного прогнозирования модели глубокого обучения требуют обширных наборов данных, охватывающих период в несколько лет. Для того, чтобы устранить дефицит данных и расширить исходный набор данных, была использована модель TimeGAN. Данная модель была синхронно обучена авторами с помощью трех функций потерь, что позволило сгенерировать синтетические данные для дополнения ограниченного набора исходных данных. Далее сгенерированные данные подвергаются фильтрации через модель CNN, что позволяет оптимизировать извлечение информации и ускорить работу сети прогнозирования. Завершающим этапом является передача извлеченной информации в сеть LSTM для прогнозирования нагрузки на электросеть.

Данное исследование продемонстрировало, что использование гибридного подхода, сочетающего TimeGAN, CNN и LSTM, позволяет выявлять сложные нелинейные взаимосвязи и значительно повышать точность прогнозирования с минимизацией ошибок прогноза по сравнению с применением традиционных моделей LSTM и CNN-LSTM. Это демонстрирует эффективность дополнения данных TimeGAN в повышении точности прогнозирования.

Еще одной работой, в которой применяется модель TimeGAN, является исследование «Multi-Scale Price Forecasting Based on Data Augmentation» (Yue, Liu, 2024). Данное исследование направлено на решение проблем, связанных с малым объемом выборки при прогнозировании цен на сельскохозяйственную продукцию. Для исторических данных о сделках с сельскохозяйственными товарами характерны длительные интервалы выборки или разреженность данных, которые часто приводят к появлению небольших выборок. Обучение на небольших выборках может привести к переобучению и усложнить учет мелкомасштабных колебаний, что существенно снижает точность прогнозирования. Таким образом, для дополнения и рас-

ширения данных путем создания реалистичных синтетических временных рядов в данном исследовании используется модель TimeGAN, что позволяет повысить устойчивость модели для прогнозирования изменения цен в различных временных масштабах.

Для оценки сходства между сгенерированными синтетическими временными рядами и исходными данными используется метод стохастического вложения соседей с t-распределением – t-SNE. Авторы используют метод t-SNE в сочетании с расхождением Кульбака–Лейблера (KL). Метод t-SNE является методом уменьшения размерности, созданный для улучшения визуального восприятия многмерных рядов данных. Согласно «Обзору методов и систем генерации синтетических обучающих данных», этот метод позволяет наглядно оценить, насколько хорошо TimeGAN моделирует структуру и зависимость временных рядов, а также помогает проверить качество и реалистичность сгенерированных синтетических данных (Рабчевский, 2023). Так, метод t-SNE позволяет авторам визуализировать многомерные данные, сводя их к двумерному пространству, благодаря чему исследователи могут сравнивать нелинейные зависимости, тренды, аномалии, неоднородные колебания как в синтетических, так и в реальных исходных данных. Далее, благодаря мере расхождения Кульбака–Лейблера, становится возможным количественно оценить различия в распределении между этими наборами данных и определить, насколько эффективно синтетические данные воспроизводят характеристики реальных исходных данных.

Таким образом, предложенный авторами подход при работе с ограниченным набором данных повышает эффективность обучения модели. Использование модели TimeGAN позволяет эффективно дополнять исходный набор данных, дает возможность улавливать мелкомасштабные колебания и критические зависимости даже при условии малого объема исходной выборки, а также обеспечивает повышение точности прогнозирования.

## 2. Методология исследования

Исследование направлено на разработку и сравнительный анализ подходов, в рамках которых синтетические временные ряды используются для расширения обучающих выборок при прогнозировании спроса на нефть. Основное внимание уделено модели TimeGAN как современному генеративному методу, способному воспроизводить как статистические свойства, так и временные зависимости оригинального ряда. В то же время, для обеспечения комплексной оценки была сформирована наивная модель и были протестированы классические методы генерации синтетических данных (GBM, ARIMA, MBB, GN+SMA), которые были улучшены для получения максимально «точных» результатов.

Ключевая идея заключается в том, чтобы дополнить ограниченные исторические наблюдения синтетическими временными последовательностями, обладающими высоким уровнем реалистичности, но при этом обеспечивающими вариативность и обучающее разнообразие. Эффективность таких расширений оценивается по двум направлениям:

1) статистическая близость синтетических данных к оригиналу;
2) практическая полезность в задаче прогнозирования.

Результаты обобщены в виде сравнительного анализа всех шести подходов. Ниже представлена структура методологии, отражающая общую логику исследования (табл. 1).

*Таблица 1*

Основные этапы исследования

| № | Этап исследования | Детальное описание |
|---|---|---|
| 1 | Подготовка данных | Загружаются исходные данные (динамика цен на нефть Brent). Временной ряд разбивается на 5 фолдов, которые в свою очередь разбиваются на обучающую и тестовую выборки. В каждом фолде нормализация обучающих и тестовых данных проводится отдельно, min-max вычисляются только по обучающей выборке |
| 2 | Формирование обучающих окон и разметка целевой переменной | Каждый временной ряд разбивается на перекрывающиеся окна длиной 30 наблюдений. Для каждого окна вычисляется вектор относительных изменений. Целевая переменная маркируется по максимальному абсолютному изменению в горизонте 5 шагов: если оно превышает 5%, y=1 (экстремум), иначе y=0 (отсутствие экстремума) |
| 3 | Формирование наивной модели | Наивная модель строится по принципу «если экстремум был в прошлом окне – будет и в следующем». Оцениваются Accuracy, F1-score, ROC AUC. Эта стратегия служит нижней границей для сравнения эффективности машинного обучения и синтетических подходов |
| 4 | Обучение базовой модели на реальных данных (далее РД). Получение метрик базовой модели | Цель модели прогнозирования – по окну из 30 наблюдений прогнозировать, произойдет ли экстремальное изменение (больше 5%) в горизонте 5 шагов вперед. Классификатор получает на вход матрицу относительных изменений за 30 дней, целевая переменная – это факт экстремума на следующих 5 днях. Основная функция потерь – binary_crossentropy.<br>Обучение MLP только на окнах из обучающего ряда реальных данных. Для повышения статистической значимости используется кросс-валидация по пяти фолдам |
| 5 | Создание синтетических рядов и оценка качества | Используются методы GBM, ARIMA, MBB, GN+SMA для создания синтетических данных для каждого из пяти фолдов. Объем синтетических данных всегда равен объему данных в обучающей выборке. Для каждого типа синтетики оценивается степень статистического сходства с реальными рядами на тренировочных сегментах. Используются метрики KL-дивергенции, статистика KS, Wasserstein Distance |
| 6 | Обучение с предобучением на синтетических данных | Для каждого генератора синтетики проводится претрейнинг нейросети на синтетических окнах (20 эпох), далее осуществляется дообучение (40 эпох) на окнах из реальных данных. Веса сети между этапами не сбрасываются |
| 7 | Тестирование моделей | Для всех стратегий тестирование производится исключительно на окнах из реальных тестовых рядов, чтобы избежать смещения и переоценки результатов за счет синтетики. Используются Accuracy, F1-score, ROC AUC |
| 8 | Сравнительный анализ результатов. | Оцениваются и сравниваются наивная стратегия, базовая модель (РД) и все комбинации с предобучением на синтетике. Ключевой критерий – величина прироста метрик по сравнению с обучением |

| № | Этап исследования | Детальное описание |
|---|---|---|
| | Определение причин успеха/провала моделей | только на реальных данных, при обязательном тестировании на идентичных сегментах<br>Для определения причин провала или успеха модели дополнительно производится визуализация фрагментов, автокорреляционные функции, спектральный анализ, t-SNE и статистика экстремумов |

*Источник:* составлено авторами.

На первом этапе формируется исходная экспериментальная база, обеспечивающая корректность и воспроизводимость последующих процедур моделирования и анализа. В качестве первичных данных используются временные ряды цен на нефть марки «Brent», собранные за определенный исторический период с фиксированным шагом дискретизации (дневные значения в период с 2015 по 2020 г.). Далее проводится последовательная организация данных для кросс-валидации. Весь исходный временной ряд разбивается на 5 фолдов. Далее для каждого фолда выделяется собственный тренировочный и тестовый поднабор (Рисунок 1). Тестовые выборки строго не пересекаются с обучающими для предотвращения утечек информации. На финальном шаге данного этапа все временные ряды независимо подвергаются нормализации методом «min-max», что приводит значения в диапазон [0, 1].

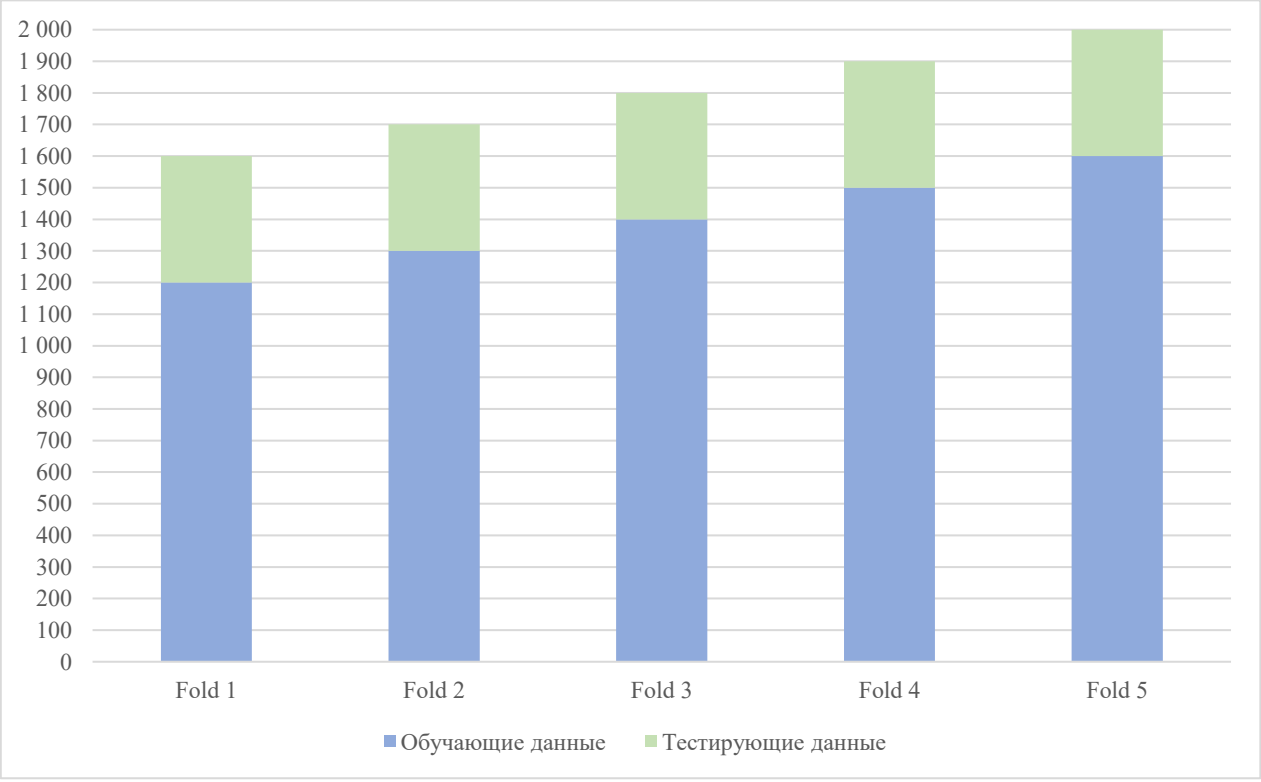


*Рисунок 1.* Обучающие и тестирующие данные (составлено авторами)

На втором этапе из исходных временных рядов формируются входные объекты и соответствующие им целевые значения, используемые далее для обучения моделей классификации экстремальных изменений. Каждый временной ряд преобразуется в совокупность перекрывающихся окон фиксированной длины. Для этого реализуется «скользящее окно»: начиная с первой точки, формируется подмассив длины 30, затем окно сдвигается на 1 шаг, и процедура

повторяется до конца ряда (рис. 2). В результате для каждого ряда получается матрица признаков. Далее для каждого окна рассчитывается вектор относительных изменений между соседними значениями.

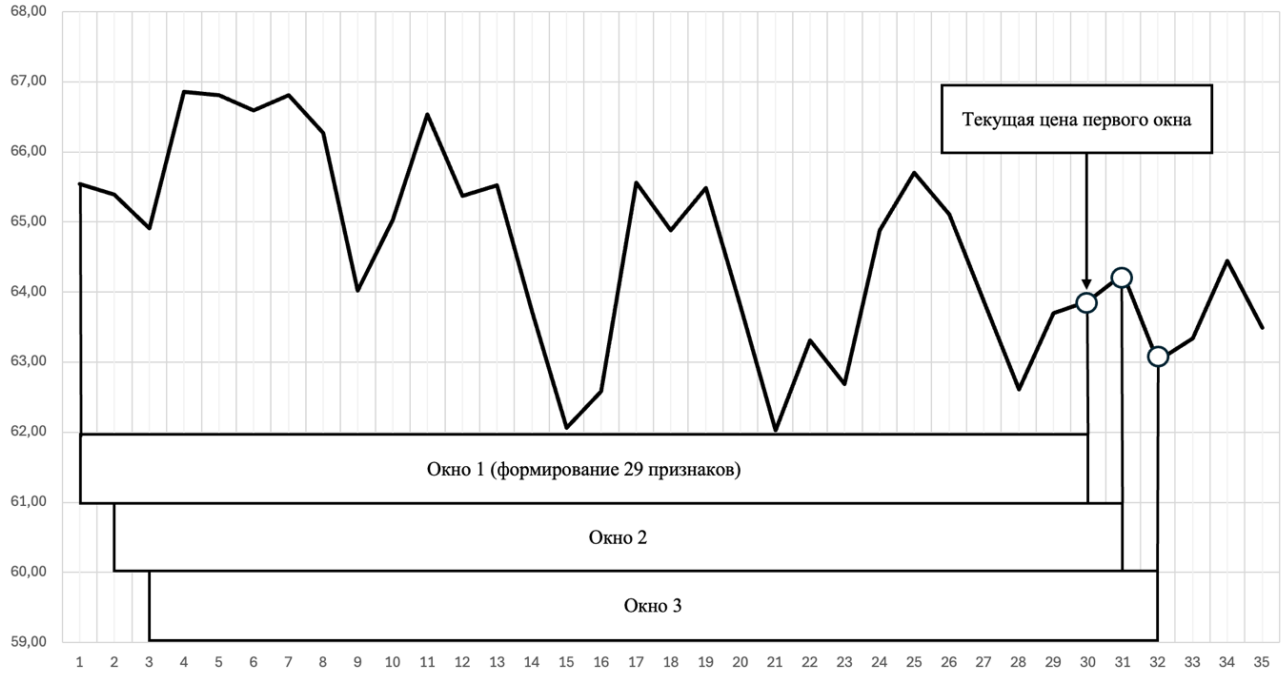


*Рисунок 2.* Схема формирования обучающих окон и признаков методом скользящего окна на временном ряду (составлено авторами)

Для каждого окна, определяемого как последовательность из 30 точек, выделяется текущая цена – последний элемент окна. Затем анализируется поведение ряда на фиксированном горизонте в пять точек. Если в течение этого интервала максимальное абсолютное отклонение от базовой цены превышает заданный порог, то окну присваивается метка 1 (произошел экстремум). В противном случае метка равна 0 (экстремальное изменение отсутствует) (рис. 3). Выбор порога в 5% для определения экстремальных изменений цен на нефть «Brent» обоснован статистическим и методологическим подходом. Расчеты показали, что 5% изменений происходят лишь в 5,2% случаев нашего эмпирического ряда, что соответствует верхнему 94–96-квантильному диапазону. Такой подход согласуется с практикой порогового анализа и регрессией по экстремальным квантилям, широко используемой в современных исследованиях рынка нефти. Например, ряд ученых (Reboredo, Ugolini, 2016) применяют квантильные зависимости для оценки влияния экстремальных изменений цен на нефть на финансовые рынки, анализируя при этом, как «самые волатильные 5% движений» воздействуют на акции. Это подтверждает обоснованность применения фиксированных порогов, аналогичных нашему уровню, в задачах выделения значимых и редких рыночных шоков.

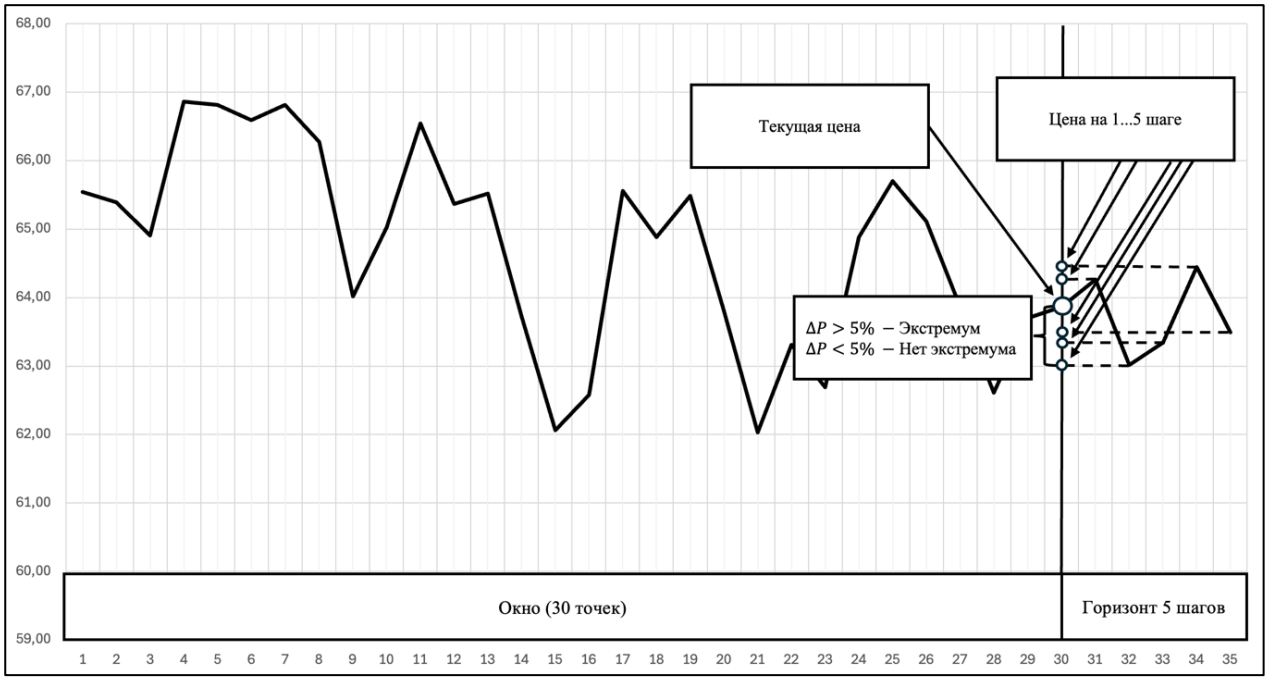


*Рисунок 3.* Схема формирования бинарной метки экстремального изменения по скользящему окну и прогнозному горизонту (составлено авторами)

Таким образом, для каждого фолда и для каждого временного ряда (реального или синтетического) формируются две матрицы:

1) X – матрица признаков (окна относительных изменений);

2) Y – вектор целевых меток (0/1 по факту наступления или отсутствия экстремума).

Подход с перекрывающимися окнами и локальными относительными изменениями позволяет максимально использовать информацию временного ряда, повышая статистическую мощность анализа даже при ограниченном объеме данных. Бинарная целевая переменная ориентирована на задачу детекции экстремальных событий, что соответствует практическим целям мониторинга и прогноза на финансовых рынках. Этап завершается получением наборов признаков и целевых переменных, пригодных для последующего машинного обучения и сопоставления реальных и синтетических сценариев. Стоит отметить, что после получения «новых» синтетических данных процесс формирования входных данных и целевых значений для дальнейшего тестирования на MLP модели, повторяется.

Третий этап исследования посвящен построению и анализу наивной модели, которая служит нижней границей для сравнения всех последующих методов прогнозирования. Задача наивной модели заключается в том, чтобы оценить минимальный уровень качества, который можно достичь, не используя никаких сложных признаков или обучающих моделей, а опираясь лишь на простейшее правило: для каждого окна модель просто повторяет метку экстремума из предыдущего окна – т. е., если экстремум (существенное изменение цены) был в прошлом горизонте, то он предсказывается и для текущего, что имитирует правило «если событие было, значит оно и будет». Такой подход исходит из предположения, что экстремальные события на рынке обладают определенной автокорреляцией – если резкое движение уже произошло, возможно, оно продолжится или повторится. Данная стратегия не использует никаких дополнительных данных или сложных признаков, что делает ее удобным «бейслайном» для сравнения: если продвинутые методы, включая нейронные сети, не способны показать явного

преимущества по сравнению с наивной моделью, это свидетельствует о бессмысленности построенных «сложных» моделей.

На четвертом этапе исследования проводится обучение базовой модели машинного обучения на окнах, сформированных исключительно из обучающего временного ряда РД. В качестве основной прогностической модели была выбрана полносвязная нейронная сеть. Этот выбор был сделан с учетом особенностей архитектуры MLP, которая отличается универсальностью и способностью эффективно решать задачи классификации на основе табличных признаков. В отличие от более сложных рекуррентных, или сверточных, моделей, полносвязная нейронная сеть демонстрирует высокую устойчивость при работе с относительно небольшими и потенциально несбалансированными выборками, что характерно для временных рядов нефтяного рынка. Благодаря своей структуре MLP не требует строгой предподготовки или нормализации входных данных и способна работать с различными типами признаков, полученными из временных окон – например, относительными изменениями цен. Особенно важно отметить, что MLP, обладая достаточной глубиной, хорошо аппроксимирует нелинейные зависимости, которые могут возникать в сложных экономических процессах, и не теряет производительность при наличии редких, но важных событий – таких как экстремальные скачки цен. Гибкая архитектура сети также позволяет легко адаптироваться под задачу детекции экстремумов, эффективно интегрируя как реальные, так и синтетические обучающие примеры.

Обучение проводится на базе полносвязной нейронной сети, состоящей из входного слоя, двух скрытых слоев с активацией ReLU, нормализацией и дропаутом для предотвращения переобучения, а также выходного слоя с сигмоидальной активацией для предсказания вероятности экстремума. Оптимизация осуществляется с помощью Adam, используется функция потерь binary_crossentropy. Для борьбы с дисбалансом классов реализуется схема взвешивания классов, что предотвращает доминирование основной динамики рынка над редкими экстремумами. Для оценки качества обучения реальная выборка разбивается на обучающую и валидационную части в пропорции 80/20, при этом сохраняется пропорция экстремумов и обычных событий. Такой подход позволяет объективно оценить обобщающую способность модели и избежать переобучения на специфических паттернах обучающей выборки. Таким образом, данный этап формирует определенный уровень качества для всех последующих экспериментов и позволяет объективно оценить вклад синтетических данных.

Перейдем к анализу пятого этапа и рассмотрим ключевые особенности использования традиционных генеративных моделей в представленном исследовании (табл. 2).

*Таблица 2*

Основные особенности генеративных моделей

| **Метод** | **Архитектура и алгоритм** | **Особенности подбора параметров** |
|---|---|---|
| ARIMA | Классическая модель авторегрессии и скользящего среднего с интегрированием. Для каждого фолда перебираются параметры (p, d, q), модель подбирается индивидуально по минимальному AIC. Прогноз строится на всю длину исходного (обучающего) ряда, результат – синтетический | Перебор (grid search) p,d,q из диапазона (p=0-3, d=0-2, q=0-3); отбор по минимуму AIC для каждого фолда |

| Метод | Архитектура и алгоритм | Особенности подбора параметров |
|---|---|---|
| | временной ряд, максимально приближенный к реальному тренду | |
| MBB | Формирование перекрывающихся блоков длины от 5 до 20 с последующей случайной перестановкой. Генерация синтетики с сохранением локальных временных зависимостей. Подбор оптимального блока на каждом фолде для минимизации средней статистической дистанции по трем метрикам | Автоматический перебор размеров блока, скоринг по средней KL+KS+WD. Для каждого фолда выбирается лучший блок |
| GN+SMA | Сначала к каждому ряду добавляется нормальный шум (5% от std), затем применяется сглаживание скользящим средним (окно 5). Алгоритм реализован минималистично, без дополнительной адаптации | Фиксированные параметры (масштаб шума и длина окна) |
| GBM | Генерация ряда путем последовательного моделирования экспоненциального роста и случайных блужданий на основе эмпирических лог-доходностей реального ряда | Оценка μ и σ на обучающем ряду, однопроходная генерация с фиксированными параметрами для каждого фолда |

*Источник:* составлено авторами.

Также рассмотрим особенности авторской TimeGAN модели (табл. 3).

*Таблица 3*

Основные особенности TimeGAN модели

| Аспект | Описание |
|---|---|
| Тип базовых блоков | Все модули построены на слоях GRU, что обеспечивает способность улавливать временные зависимости в данных |
| Модульная структура | Сеть включает пять отдельных модулей: Embedder, Recovery, Generator, Supervisor, Discriminator. Каждый модуль – это отдельная RNN (GRU) |
| Embedder + Recovery | Служат для кодирования исходных временных рядов в скрытое пространство и их восстановления обратно |
| Generator | Принимает случайный шум и генерирует скрытые последовательности, имитирующие динамику реальных данных |
| Supervisor | Особый модуль для прогнозирования следующего шага в скрытом пространстве – обучается удлинять латентную последовательность, что позволяет моделировать длинные временные зависимости |
| Discriminator | Отличает реальные скрытые представления от сгенерированных |
| Два этапа обучения | 1. Предобучение: обучение Embedder и Supervisor на задаче восстановления данных<br>2. Совместное обучение Generator, Supervisor, Discriminator, Embedder, Recovery |

| Аспект | Описание |
|---|---|
| Разделение оптимизаторов | Для каждого модуля используется отдельный оптимизатор (Adam), что позволяет гибко настраивать скорость и стратегию обучения разных частей сети. |
| Мини-батч и шум | На вход генератора подается нормальный шум размерности (batch, seq_len, z_dim); обучение ведется с батчами даже для малых выборок |
| Восстановление | Восстановление временного ряда в оригинальное пространство происходит только после завершения adversarial-обучения. Это предотвращает утечку информации и повышает реалистичность синтетических данных |
| Совместимость размеров | Все модули согласованы по размерности скрытого состояния и длине входной последовательности. |

*Источник:* составлено авторами.

Результатом генерации синтетических рядов являются временные ряды размерности от 1200 до 1600 элементов, которые в дальнейшем были использованы для предобучения модели MLP описанной в четвертом этапе.

Далее реализуется статистическая оценка качества синтетических временных рядов, полученных с помощью различных методов. Синтетические данные были включены в процесс предобучения основной модели прогнозирования. Для количественного анализа результатов использовались стандартные метрики классификации: Accuracy, F1-score, ROC AUC. Такой подход обеспечивает комплексную и воспроизводимую оценку вклада синтетических данных в решение прикладных задач прогноза. Благодаря прямому сравнительному анализу моделей, обученных только на реальных данных, и моделей с предобучением на синтетике, удалось выявить условия, при которых генерация синтетических рядов позволяет достичь существенного улучшения прогностических характеристик. Также был проведен дополнительный анализ, включающий исследования автокорреляции, спектров, количество экстремумов и распределение признаков у синтетических данных для определения причин полученных результатов.

Таким образом, применяемая методология в полной мере охватывает ключевые этапы – от генерации и валидации синтетических данных до их интеграции в прикладные задачи прогнозирования, а последующий анализ на независимой тестовой выборке дает объективную картину эффективности синтетических подходов в экономических задачах с дефицитом исторических наблюдений[1].

## 3. Результаты проведенного исследования

Рассмотрим количественную оценку качества синтетических данных по основным статистическим метрикам (табл. 4).

[1] Для обеспечения воспроизводимости исследования полный программный код для запуска экспериментов опубликован на GitHub по адресу: https://github.com/AleksandrM27/Synthetic-data.MLP-experiment/tree/c34468f6091bb377bd3f423c70784ea9ecdf36dc.

*Таблица 4*

Метрики синтетических данных

| Метод | KL-дивергенция | KS-статистика | Wasserstein-дист. |
|---|---|---|---|
| MBB | 0,0622 | 0,031 | 0,5736 |
| ARIMA | 0,0014 | 0,00434 | 0,05766 |
| GBM | 2,288 | 0,638 | 2,88 |
| GN+SMA | 0,045 | 0,0148 | 0,1488 |
| TimeGAN | 0,1844 | 0,089 | 1,7382 |

*Источник:* составлено авторами на основе Приложения 1.

Из приведенных данных можно сделать следующие выводы.

1. ARIMA демонстрирует наилучшую способность к воспроизведению глобальных статистических характеристик исходных данных. Минимальные значения всех трех метрик свидетельствуют о высокой идентичности синтетических и реальных временных рядов как с точки зрения распределения значений, так и по общей форме динамики. Это указывает на исключительную пригодность ARIMA для задач, где критически важно сохранение тренда и «больших» закономерностей. Однако, как показывает анализ реальных графиков, столь высокая статистическая близость достигается за счет сильного сглаживания: модель подавляет спонтанные краткосрочные флуктуации, что может быть критичным недостатком для задач поиска экстремумов и волатильных эпизодов.

2. GN+SMA и MBB показывают достаточно хорошее совпадение с исходными данными. Для GN+SMA малые значения KL и KS свидетельствуют о том, что метод хорошо имитирует эмпирическое распределение, но его простота (добавление шума с последующим сглаживанием) ограничивает воспроизведение сложных зависимостей: динамика становится менее естественной, а автокорреляции зачастую размываются. MBB, напротив, отлично воспроизводит внутри-блочные структуры и частично сохраняет короткосрочные паттерны, однако искусственно разрушает длительные зависимости на стыках блоков. Таким образом, оба метода подходят для задач, где требуется «быстрая» генерация реалистичных, но не идеально динамических сценариев.

3. GBM демонстрирует наихудшие показатели соответствия: высокие значения всех метрик указывают на значительные отклонения синтетики от реальных данных. Это обусловлено тем, что модель, ориентированная на чисто стохастическую динамику с постоянными параметрами дрейфа и волатильности, не способна уловить ни реальных трендов, ни характерных автокорреляций и асимметрий, присущих экономическим временным рядам. На практике это приводит к синтетике с чрезмерной волатильностью и низкой прогностической ценностью для прикладных задач.

4. TimeGAN демонстрирует двойственный характер результатов: низкая KL-дивергенция и KS-статистика указывают на успешное воспроизведение общей формы распределения, однако по сравнению с некоторыми классическими методами данные показатели сравнительно высокие. Данный результат указывает на то, что модель может искажать отдельные детали структуры распределения и динамики. Такой эффект может объясняться

сложностью архитектуры TimeGAN и ее склонностью к генерации разнообразных, но не всегда идеально «выверенных» сэмплов.

Таким образом, если целью является максимальное воспроизведение статистических свойств исходного ряда, наиболее точный результат обеспечивает ARIMA. Однако для задач, где важно сохранять не только распределение, но и динамические особенности или сложную структуру, целесообразно рассматривать комбинирование классических и нейросетевых методов.

Перейдем к результатам моделирования MLP модели (табл. 5).

*Таблица 5*

Метрики MLP модели

| **Подход** | **Точность** | **F1-score** | **ROC AUC** |
|---|---|---|---|
| Бейслайн | 0,586 | 0,404 | 0,542 |
| РД | 0,686 | 0,494 | 0,628 |
| РД + MBB | 0,652 | 0,454 | 0,636 |
| РД + ARIMA | 0,674 | 0,456 | 0,638 |
| РД + GBM | 0,658 | 0,456 | 0,608 |
| РД + GN+SMA | 0,712 | 0,542 | 0,672 |
| РД + TimeGAN | 0,706 | 0,536 | 0,676 |

*Источник:* составлено авторами на основе Приложения 2.

Как показывают результаты табл. 5, включение различных типов синтетических данных по-разному влияет на качество прогнозирования экстремальных событий на временных рядах. Обучение только на реальных данных позволяет достичь сбалансированных показателей по всем метрикам. Добавление бутстрэппинга и ARIMA в обучающую выборку приводит к незначительному снижению точности и F1-score по сравнению с использованием только реальных данных, несмотря на небольшое улучшение ROC AUC. Это говорит о том, что подобные виды синтетики не привносят достаточного разнообразия или сложности, а иногда даже могут ослаблять способность модели выявлять экстремумы за счет нарушения временных паттернов. GBM также не демонстрирует значимых улучшений, показатели точности и F1-score остаются на уровне, близком к реальным данным, а ROC AUC даже немного снижается. Таким образом, генерация данных по GBM не увеличивает прогностическую ценность для рассматриваемой задачи.

GN+SMA и TimeGAN выделяются на фоне других подходов. Оба метода обеспечивают максимальные значения по основным метрикам. Особенно важно, что у TimeGAN наблюдается наибольший прирост ROC AUC, что отражает рост чувствительности модели к аномальным случаям без потери общего качества. Таким образом, для рассматриваемой задачи синтетические ряды, созданные GN+SMA и TimeGAN, оказываются наиболее эффективными для повышения качества прогноза. Эти методы позволяют не только расширить обучающее пространство, но и улучшить способность модели обнаруживать редкие и значимые события, что принципиально важно для анализа сложных временных рядов.

Перейдем к анализу синтетических данных для определения причин провала и успеха генеративных методов (рис. 4).

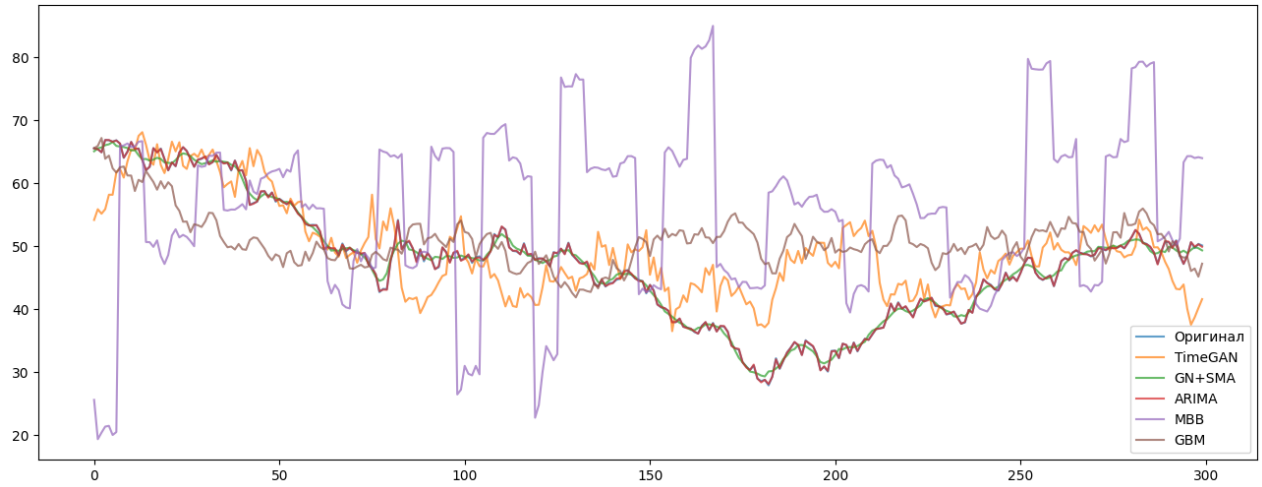


*Рисунок 4.* Визуальное сравнение сгенерированных временных рядов с оригиналом (первые 300 точек) (составлено авторами)

Уже на этапе визуального сравнения сгенерированных рядов можно наблюдать фундаментальные различия в поведении моделей. Модели GN+SMA и ARIMA демонстрирует наибольшую близость к оригиналу: они сохраняют форму сигнала, локальные тренды и колебания, не внося лишней волатильности. TimeGAN, как и ожидалось, обеспечивает достаточно реалистичную динамику, но временами уходит в более сильные флуктуации, что может указывать на переобучение на локальных паттернах. GBM демонстрирует чрезмерно сглаженное поведение, теряя микроструктуру оригинального ряда – вероятно, из-за стохастического характера модели и отсутствия обучающего механизма. Наихудший результат наблюдается у MBB, где видно множество неестественных скачков и артефактов, что может быть связано с тем, что модель просто добавляет куски оригинального ряда, не заботясь о согласованности границ. Перейдем к анализу автокорреляционных функций (рис. 5).

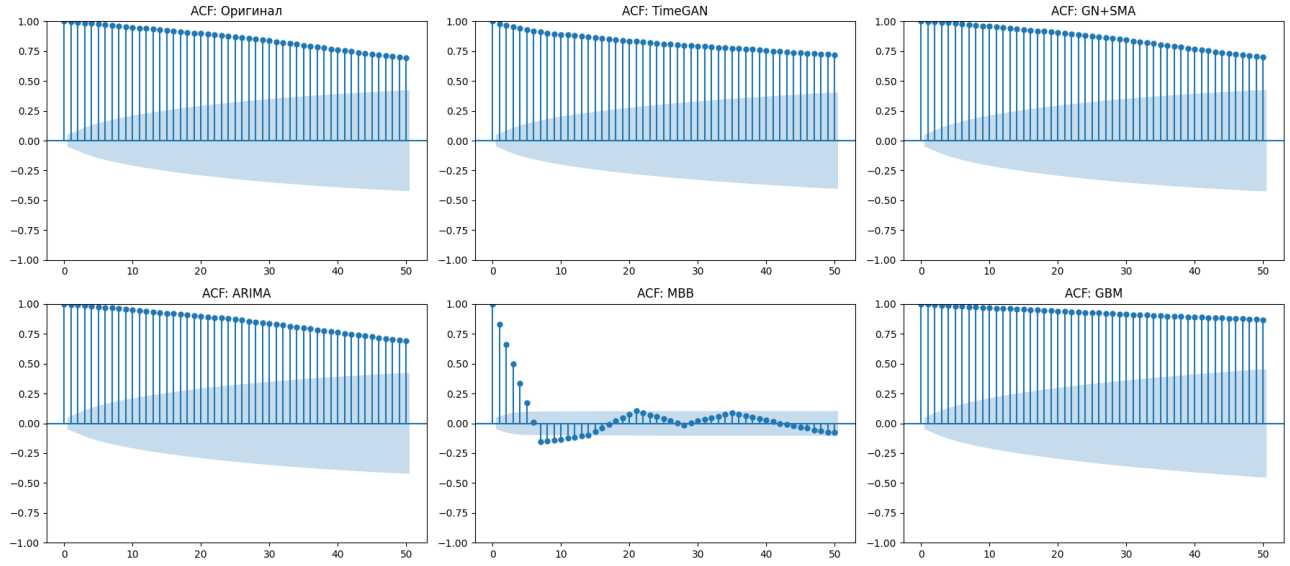


*Рисунок 5.* Автокорреляционные функции (ACF) временных рядов: оригинал и генерации (составлено авторами)

Оригинальный ряд демонстрирует устойчивую и медленно убывающую автокорреляцию, характерную для процессов с сильной временной зависимостью. GN+SMA, ARIMA и

TimeGAN показывают очень схожую ACF-структуру: медленный спад и устойчивые положительные корреляции на лаге до 50. Это указывает на то, что эти модели успешно захватывают внутреннюю зависимость между временными шагами. GBM, напротив, демонстрирует почти плоскую ACF, близкую к случайному шуму – в точности как и ожидалось от геометрического броуновского движения. MBB снова проваливает задачу: ACF у него не просто быстро убывает, но еще и ведет себя нестабильно, что указывает на плохую стыковку блоков при бутстрепировании. Это подтверждает, что MBB не может воспроизвести глобальную временную структуру.

Следующим этапом анализа является амплитудный спектр (рис. 6).

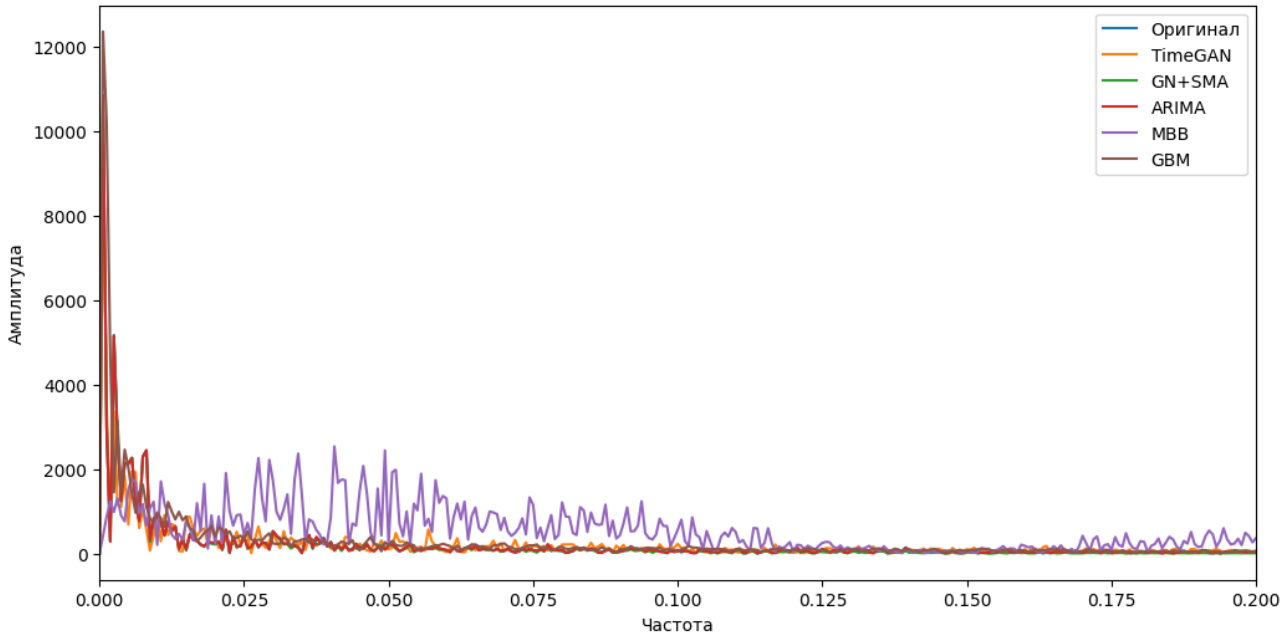


*Рисунок 6.* Сравнение спектральных характеристик временных рядов (БПФ-анализ) (составлено авторами)

Важно отметить, что оригинальные данные имеют четкое доминирование в области низких частот, указывающее на наличие трендов и плавных циклов. GN+SMA, ARIMA и TimeGAN в целом следуют спектру оригинала, что означает, что они сохраняют основные компоненты частотного спектра. MBB, напротив, проявляет всплески в среднем диапазоне частот – это соответствует визуально наблюдаемым скачкам и шумам. GBM остается близким к шуму: энергия спектра распределена относительно равномерно и быстро затухает, что также говорит о недостатке внутренней структуры.

Рассмотрим сравнение количества экстремумов (рис. 7).

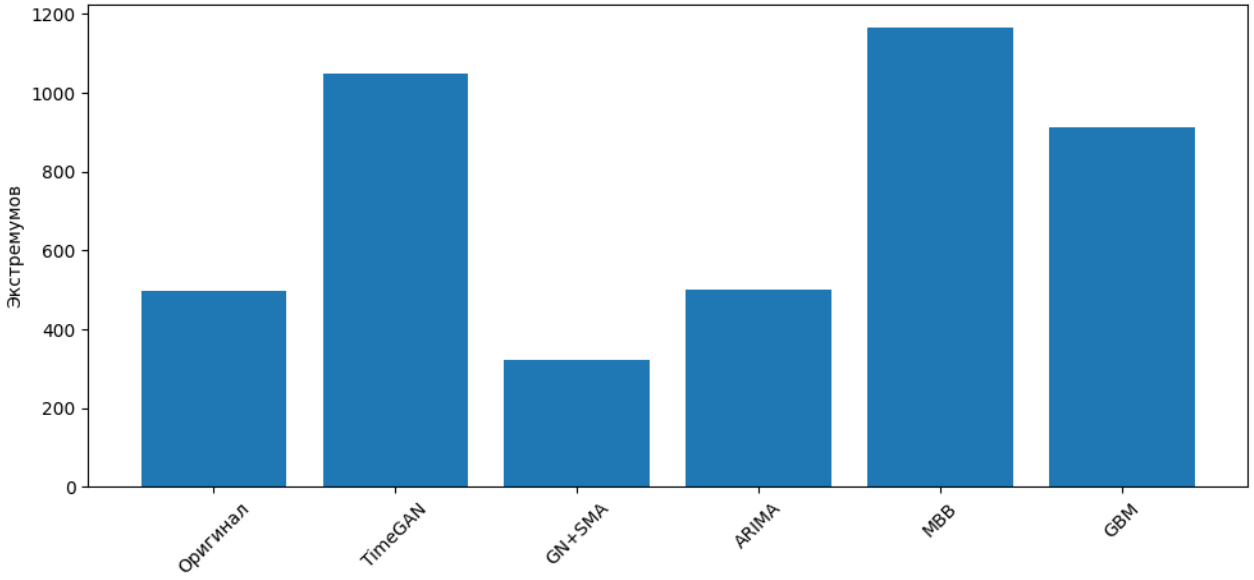


*Рисунок 7.* Сравнение количества экстремумов (составлено авторами)

На представленном рисунке 7 можно заметить, что GN+SMA и ARIMA воспроизводят количество экстремумов, близкое к оригинальному ряду. TimeGAN существенно переоценивает экстремумы, что согласуется с визуальными наблюдениями об избыточной волатильности. MBB – снова антилидер: количество экстремумов почти в 2,5 раза выше, чем у оригинала, что указывает на грубые скачки. GBM также имеет существенные отличия в количестве экстремумов в отличие от оригинальных данных.

Перейдем к заключительной, t-SNE визуализации «окон» (рис. 8).

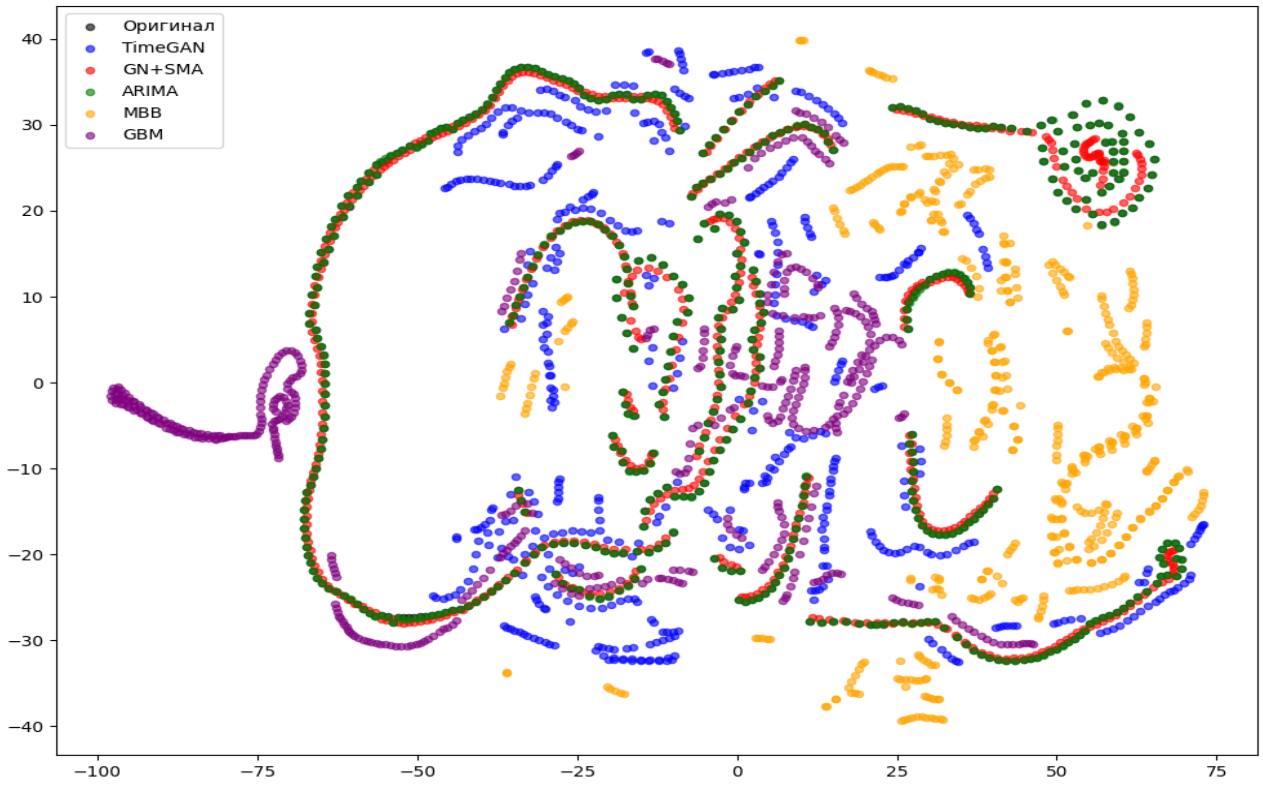


*Рисунок 8.* Сравнение t-SNE «окон» из разных источников (составлено авторами)

GN+SMA и ARIMA демонстрируют наиболее близкое распределение к оригиналу: их окна размещаются вблизи кластеров оригинального ряда, что говорит о высоком уровне воспроизведения локальных зависимостей и статистической структуры. Это согласуется с их поведением на графиках ACF и спектрального анализа. В то же время TimeGAN формирует более обособленные области на проекции, отличающиеся от структуры оригинальных окон. Это говорит о том, что модель не просто копирует структуру обучающего ряда, а генерирует вариативные – и отчасти новые – паттерны поведения. Это может быть как преимуществом, так и источником потенциальных отклонений от реальных данных, особенно если новизна выходит за рамки допустимого. MBB и GBM, как и ожидалось, демонстрируют наибольшее удаление от оригинала. Их окна формируют плотности, которые не пересекаются с оригинальными и указывают на принципиально иную динамику. В случае MBB это связано с грубым копированием блоков и отсутствием глобальной согласованности.

Комплексная визуализация свойств синтетических и оригинальных временных рядов демонстрирует принципиальные различия между подходами GN+SMA и TimeGAN. GN+SMA практически идентичен реальным данным по визуальному рисунку, спектру и автокорреляции, но значительно занижает число экстремальных событий, т. е. сглаживает динамику и не вносит новых сценариев. TimeGAN, напротив, создает большее разнообразие паттернов и формирует избыточное число экстремумов, что проявляется как в t-SNE (отдельные кластеры), так и в спектре и количестве экстремумов. Таким образом, GN+SMA дает максимально «безопасную» синтетику, а TimeGAN, расширяет пространство сценариев, но требует дополнительной настройки для контроля волатильности и экстремумов.

Анализ MLP метрик и визуализаций позволяет сделать ряд выводов о причинах успехов и провалов моделей.

1. GN+SMA – это единственный традиционный метод, который уверенно показывает сильные результаты. Причина, с точки зрения авторов, – это способность сохранять ключевые структурные свойства временного ряда. Он воспроизводит плавность, автокорреляцию и характерный частотный спектр оригинальных данных, без внесения шумов или артефактов. Визуально сигналы, сгенерированные этим методом, неотличимы от реальных, и это подтверждается тем, что в t-SNE их окна частично перекрываются с окнами оригинала. Однако GN+SMA не создает новые сценарии, он сглаживает поведение и снижает количество экстремумов, фактически устраняя крайние события. Тем не менее, это работает: классификатор получает больше стабильных, чистых примеров и выучивает основную структуру сигнала гораздо лучше. Успех GN+SMA не в расширении пространства, а в усилении сигнала и устранении шумов, что критически важно при обучении модели, ориентированной на распознавание форм.

2. TimeGAN также показывает высокие значения по качеству и делает это за счет принципиально другой стратегии. Модель не просто воспроизводит форму сигнала, она генерирует новые паттерны, основанные на скрытых структурах в обучающем ряду. Визуально данные TimeGAN выглядят реалистично, но обладают более выраженной волатильностью, более сложным спектром и большим числом экстремумов. В t-SNE их окна образуют обособленные кластеры, что говорит о реальном расширении сценарного пространства. Это делает модель особенно ценной для задач, где важно научить классификатор справляться с вариативностью и неожиданными случаями, например, для стресс-тестов. Однако та же самая вариатив-

ность может снижать надежность: если разнообразие выходит за границы статистической достоверности, классификатор начинает учиться на нерелевантных паттернах. Успех TimeGAN заключается в новизне и обобщении, но его слабость в потребности калибровки, чтобы не создать избыточную нестабильность в обучении.

3. ARIMA – это модель, строго ограниченная своей природой. Она воспроизводит только предсказуемую, стационарную компоненту ряда. Визуально сигналы выглядят чистыми и корректными: ACF стабильна, спектр сосредоточен в низких частотах, экстремумы соответствуют оригиналу. Однако именно это делает ARIMA структурно бесполезной в контексте задачи синтетического обогащения. Она не вносит ничего нового в обучающую выборку. На t-SNE видно, что ее окна полностью накладываются на оригинальные, не добавляя ни ширины, ни глубины в признаковое пространство. Это объясняет слабый, но стабильный прирост метрик: ARIMA не искажает, но и не обучает. Ее эффект обусловлен лишь увеличением объема выборки, а не качественным расширением данных.

4. MBB, на первый взгляд, представляется логичным методом, так как работает с реальными фрагментами исходного ряда и способен сохранять локальные временные зависимости внутри блока. Однако его применение к нестационарным данным приводит к существенным методологическим ограничениям. Главная проблема заключается в нарушении согласованности на стыках блоков: соединение случайных отрезков порождает неестественные скачки, нестабильную автокорреляционную функцию, резкие искажения спектра и аномальный рост числа экстремумов. Анализ t-SNE дополнительно демонстрирует отчетливое отделение окон, сгенерированных MBB, от облака оригинальных данных, что указывает на структурную деформацию синтетики. В результате классификатор, обучаемый на таких данных, начинает воспринимать шум и артефакты как значимые сигналы, чего нет в реальных рыночных паттернах. Это ведет к размыванию границ классов и снижению обобщающей способности модели. Таким образом, основной недостаток MBB заключается в генерации вредных, искажающих примеров, которые нарушают структуру исходного ряда и затрудняют формирование устойчивых прогностических правил. Для повышения эффективности такого метода требуется либо стационаризация исходного ряда, либо применение более адаптивных генеративных подходов, способных корректно работать с особенностями рыночных данных.

5. GBM – это наиболее оторванный от реальности подход. Он генерирует траектории без памяти, без автокорреляции и без структуры, опираясь на независимые приращения. На визуализациях GBM производит либо упрощенные, либо случайные сигналы, спектр у него быстро затухает, ACF стремится к нулю, а в t-SNE он оказывается на максимальном удалении от оригинальных данных. Такие ряды не несут обучающей ценности: они структурно неинформативны и статистически чужды. Классификатор, сталкиваясь с ними, не учится – он дезориентируется. GBM не просто не помогает, он разрушает обучающее пространство, внося совершенно нерелевантные зависимости. Это объясняет его слабый результат: он добавляет объем, но в нем нет смысла.

Таким образом, результаты свидетельствуют о том, что применение TimeGAN действительно позволяет существенно расширить разнообразие обучающих данных и смоделировать сценарии, не встречавшиеся в историческом ряду. Это повышает потенциал модели для обобщения и тестирования на неожиданных рыночных режимах, что может быть крайне ценно для задач стресс-тестирования и выявления уязвимостей прогностических моделей. Однако отме-

ченное увеличение числа экстремумов и локальных колебаний требует дополнительной калибровки и контроля параметров генерации, чтобы избежать переоценки рисков и поддерживать реалистичность синтетических данных.

Результаты исследования лишь частично подтверждают выдвинутые гипотезы, и по каждой из них следует сделать важные оговорки.

1. Синтетические временные ряды, сгенерированные с помощью TimeGAN, действительно демонстрируют относительно низкую KL-дивергенцию, KS-статистику и Wasserstein-дистанцию (для нейросетевых моделей). Однако данные метрики заметно выше, чем у некоторых классических методов. Это указывает на то, что, несмотря на сохранение общей формы распределения, синтетика TimeGAN не всегда полностью воспроизводит тонкие особенности динамики исходных данных. Следовательно, ключевые статистические и динамические свойства могут быть переданы не в полной мере, особенно если сравнивать с результатами ARIMA по ряду метрик.

2. Добавление синтетических рядов TimeGAN к обучающей выборке действительно приводит к росту всех целевых метрик прогноза (Accuracy, F1-score, ROC AUC) относительно использования только реальных данных. Однако этот прирост носит умеренный характер: показатели улучшаются не радикально, а различия между TimeGAN и, например, GN+SMA зачастую незначительны, что говорит о том, что эффективность TimeGAN может зависеть от специфики конкретной задачи и выбранной архитектуры модели. Важно также отметить, что некоторые классические методы, несмотря на скромные статистические метрики, могут в ряде случаев обеспечивать сравнимый или даже лучший вклад в итоговое качество модели – в частности, по ROC AUC GN+SMA и TimeGAN показывают близкие результаты.

3. Гипотеза о безусловном превосходстве TimeGAN над классическими генераторами требует уточнения. Хотя TimeGAN действительно обеспечивает лучший баланс между статистическим соответствием и прикладной пользой синтетики, отдельные классические подходы (например, ARIMA) по ряду статистических критериев обеспечивают более высокое совпадение с оригинальным рядом, а GN+SMA по некоторым прикладным метрикам не только не уступает TimeGAN, но и превосходит. Таким образом, применение TimeGAN целесообразно в задачах, где важна генерация структурно разнообразных и вариативных данных, однако его преимущества не всегда оказываются однозначными или максимальными по всем направлениям оценки.

В целом, результаты демонстрируют, что нейросетевая генерация синтетических рядов с помощью TimeGAN способна повысить качество обучения и прогноза в условиях ограниченной исторической информации, однако эффект зависит от сочетания статистических, динамических и прикладных характеристик данных, а преимущества над классическими подходами могут быть не столь однозначны, как предполагалось изначально.

Тем не менее, по мнению авторов, результаты представленного исследования все-таки указывают, что использование модели TimeGAN позволяет эффективно дополнять исходный набор данных. Данный вывод соотносится с результатами исследований «Synthetic Time Series Data Generation Using Time GAN with Synthetic and Real-Time Data Analysis» (Juneja et al., 2023), «Volatility and Irregularity Capturing in Stock Price Indices Using Time Series Generative

Adversarial Networks (TimeGAN)» (Mushunje et al, 2023), «Multi-Scale Price Forecasting Based on Data Augmentation» (Yue, Liu, 2024).

Также подчеркиваются некоторые ограничения модели TimeGAN. Хотя модель способна воспроизводить крупные тренды и резкие скачки, ей может быть сложно воспроизвести мелкие колебания и редкие аномалии с такой же точностью. Это согласуется с выводами других исследователей, которые отмечают, что высокая вычислительная сложность и архитектурные особенности TimeGAN могут быть препятствием для ее широкого применения в задачах, требующих моделирования сложных взаимосвязей.

## Заключение

Проведенный анализ продемонстрировал, что синтетические временные ряды, созданные с помощью TimeGAN, обладают высоким уровнем реалистичности. Это подтверждается количественной оценкой. С учетом результатов анализа специфики динамики стоимости нефти марки «Brent» можно сделать вывод, что модели TimeGAN являются сравнительно эффективными для увеличения обучающей выборки и имеют ряд преимуществ перед традиционными моделями. Однако, важным аспектом данного исследования также является определение проблем моделей TimeGAN, которые указывают на существенные сложности в использовании данного типа моделей в условиях слабых вычислительных мощностей и высокую требовательность к навыкам в области настройки моделей машинного обучения. Это указывает на необходимость дальнейших исследований для оптимизации архитектуры TimeGAN и ее адаптации к специфике нефтяного рынка.

В перспективе требуется более детально изучить риск нестабильности обучения для избегания критических ошибок моделирования, а также сформулировать методологию для проверки синтетических данных на реалистичность. Эти шаги позволят еще более эффективно использовать потенциал синтетических данных и способствовать повышению устойчивости и точности прогнозирования в условиях изменчивой рыночной среды.

## Список литературы

Каукин А.С., Павлов П.Н., Косарев В.С. Краткосрочное прогнозирование цен на электроэнергию с использованием генеративных нейронных сетей // Бизнес-информатика. 2023. № 3 (17). С. 7–23.

Копытин И.А. Трансформация рынка нефти в Европе: тенденции и перспективы // Проблемы прогнозирования. 2024. № 6. С. 217–226. https://doi.org/10.47711/0868-6351-207-217-226.

Рабчевский А.Н. Обзор методов и систем генерации синтетических обучающих данных // Прикладная математика и вопросы управления. 2023. № 4. С. 6–45. https://doi.org/10.15593/2499-9873/2023.4.01.

Розенцвайг А.К. Методы эконометрического моделирования и анализа социально-экономических явлений. Набережные Челны, 2014. 121 с. https://doi.org/10.13140/RG.2.1.3998.5526.

Brajard J., Carrassi A., Bocquet M., Bertino L.. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the

Lorenz 96 model // Journal of Computational Science. 2020. Vol. 44. https://doi.org/101171.10.1016/j.jocs.2020.101171.

Farzana G.S., Prakash N. Machine Learning in Demand Forecasting // A Review. In Proceedings of the 2nd International Conference on IoT, Social, Mobile, Analytics and Cloud in Computational Vision and Bio-Engineering (ISMAC-CVB 2020). 2020. http://dx.doi.org/10.2139/ssrn.3733548.

Juneja T., Bajaj S., Sethi N. Synthetic Time Series Data Generation Using Time GAN with Synthetic and Real-Time Data Analysis // Proceedings of the International Conference on Advances in Computing and Data Sciences. Springer. 2023. P. 567–576. https://doi.org/10.1007/978-981-99-0601-7_51.

Kanagarathinam K. Comprehensive Overview of Optimization Techniques in Machine Learning Training // Computational Science and Optimization Letters. 2024. Vol. 2. No. 1. P. 69–85. https://doi.org/10.59247/csol.v2i1.69.

Li J., Liu Y., Li Q. Generative Adversarial Network and Transfer Learning Based Fault Detection for Rotating Machinery with Imbalance Data Condition // Measurement Science and Technology. 2022. https://doi.org/33.10.1088/1361-6501/ac3945.

Liu Y., Liang Z., Li X. Enhancing Short-Term Power Load Forecasting for Industrial and Commercial Buildings: A Hybrid Approach Using TimeGAN, CNN, and LSTM // IEEE Open Journal of the Industrial Electronics Society. 2023. No. 4. P. 451–462. https://doi.org/10.1109/OJIES.2023.3319040.

Matta C., Bianchesi N.M., Oliveira M., Balestrassi P., Leal F. A comparative study of forecasting methods using real-life econometric series data // Production. 2021. Vol. 31. https://doi.org/10.1590/0103-6513.20210043.

Moroff N., Kurt E., Kamphues J. Machine Learning and Statistics: A Study for Assessing Innovative Demand Forecasting Models // Procedia Computer Science. 2021. No.180. P. 40–49. https://doi.org/10.1016/j.procs.2021.01.127.

Mushunje L., Allen D., Peiris S. Volatility and Irregularity Capturing in Stock Price Indices Using Time Series Generative Adversarial Networks (TimeGAN) // Computational Engineering, Finance, and Science. 2023. https://doi.org/10.48550/arXiv.2311.12987.

Reboredo J.C., Ugolini A. Quantile dependence of oil price movements and stock returns // Energy Economics. 2016. Vol. 54. P. 33–49. https://doi.org/10.1016/j.eneco.2015.11.015.

Sacco M.A., Ruiz J.J., Pulido M., Tandeo P. Evaluation of Machine Learning Techniques for Forecast Uncertainty Quantification // Quarterly Journal of the Royal Meteorological Society. 2021. Vol. 147. No. 34. P. 1234–1251. https://doi.org/10.48550/arXiv.2111.14844.

Yue T, Liu Y. Multi-Scale Price Forecasting Based on Data Augmentation // Applied Sciences. 2024. Vol. 14. No. 19. P. 8737. https://doi.org/10.3390/app14198737.

# Приложение 1

Метрики синтетических данных

| MBB | | | | | |
|---|---|---|---|---|---|
| Fold | Best block size | Средний Score | KL-дивергенция | KS-статистика | Wasserstein-дист. |
| Fold 1 | 5 | 0,315 | 0,112 | 0,045 | 0,787 |
| Fold 2 | 9 | 0,245 | 0,098 | 0,027 | 0,609 |
| Fold 3 | 10 | 0,180 | 0,040 | 0,024 | 0,475 |
| Fold 4 | 15 | 0,201 | 0,034 | 0,033 | 0,536 |
| Fold 5 | 7 | 0,172 | 0,027 | 0,026 | 0,461 |
| ARIMA | | | | | |
| Fold | ARIMA(p,d,q) | | KL-дивергенция | KS-статистика | Wasserstein-дист. |
| Fold 1 | (1, 1, 0) | | 0,0007 | 0,0050 | 0,0644 |
| Fold 2 | (0, 1, 0) | | 0,0000 | 0,0015 | 0,0480 |
| Fold 3 | (2, 1, 2) | | 0,0021 | 0,0043 | 0,0544 |
| Fold 4 | (2, 1, 3) | | 0,0019 | 0,0053 | 0,0612 |
| Fold 5 | (2, 1, 3) | | 0,0021 | 0,0056 | 0,0603 |
| GBM | | | | | |
| Fold | μ | σ | KL-дивергенция | KS-статистика | Wasserstein-дист. |
| Fold 1 | -4,7E-05 | 2,1E-02 | 6,2E-01 | 1,8E-01 | 8,5E+00 |
| Fold 2 | -3,8E-05 | 2,1E-02 | 6,4E+00 | 8,0E-01 | 4,5E+01 |
| Fold 3 | -6,2E-04 | 2,4E-02 | 1,2E+00 | 5,5E-01 | 1,5E+01 |
| Fold 4 | -2,6E-04 | 2,7E-02 | 1,1E+00 | 7,2E-01 | 3,2E+01 |
| Fold 5 | -1,5E-04 | 2,6E-02 | 3,5E+00 | 7,2E-01 | 3,3E+01 |
| GN+SMA | | | | | |
| Fold | Noise scale | SMA window | KL-дивергенция | KS-статистика | Wasserstein-дист. |
| Fold 1 | 0,05 | 5 | 0,033 | 0,014 | 0,154 |
| Fold 2 | 0,05 | 5 | 0,068 | 0,016 | 0,152 |
| Fold 3 | 0,05 | 5 | 0,029 | 0,016 | 0,138 |
| Fold 4 | 0,05 | 5 | 0,044 | 0,015 | 0,147 |
| Fold 5 | 0,05 | 5 | 0,051 | 0,013 | 0,153 |
| TimeGAN | | | | | |
| Fold | | | KL-дивергенция | KS-статистика | Wasserstein-дист. |
| Fold 1 | | | 0,196 | 0,071 | 1,428 |
| Fold 2 | | | 0,171 | 0,085 | 1,491 |
| Fold 3 | | | 0,213 | 0,103 | 1,713 |
| Fold 4 | | | 0,178 | 0,096 | 2,075 |
| Fold 5 | | | 0,164 | 0,090 | 1,984 |

*Источник:* составлено авторами.

# Приложение 2

Результаты моделирования

| Бейслайн | | | | | | |
|---|---|---|---|---|---|---|
| Фолд | Точность | F1-score | ROC AUC | Экстр (%) | Кол-во | ConfMat |
| 1 | 0,62 | 0,55 | 0,61 | 41,89 | 74 | [[29, 14], [14, 17]] |
| 2 | 0,55 | 0,49 | 0,55 | 44,59 | 74 | [[25, 16], [17, 16]] |
| 3 | 0,57 | 0,30 | 0,50 | 31,08 | 74 | [[35, 16], [16, 7]] |
| 4 | 0,61 | 0,29 | 0,51 | 28,38 | 74 | [[39, 14], [15, 6]] |
| 5 | 0,58 | 0,39 | 0,54 | 35,14 | 74 | [[33, 15], [16, 10]] |
| Обучение на РД | | | | | | |
| Фолд | Точность | F1-score | ROC AUC | Экстр (%) | Кол-во | ConfMat |
| 1 | 0,65 | 0,55 | 0,66 | 41,89 | 74 | [[32, 11], [15, 16]] |
| 2 | 0,65 | 0,54 | 0,64 | 44,59 | 74 | [[33, 8], [18, 15]] |
| 3 | 0,73 | 0,44 | 0,57 | 31,08 | 74 | [[46, 5], [15, 8]] |
| 4 | 0,72 | 0,46 | 0,65 | 28,38 | 74 | [[44, 9], [12, 9]] |
| 5 | 0,68 | 0,48 | 0,62 | 35,14 | 74 | [[39, 9], [15, 11]] |
| Обучение на РД+MBB | | | | | | |
| Фолд | Точность | F1-score | ROC AUC | Экстр (%) | Кол-во | ConfMat |
| 1 | 0,57 | 0,50 | 0,61 | 41,89 | 74 | [[26, 17], [15, 16]] |
| 2 | 0,65 | 0,54 | 0,65 | 44,59 | 74 | [[33, 8], [18, 15]] |
| 3 | 0,70 | 0,35 | 0,60 | 31,08 | 74 | [[46, 5], [17, 6]] |
| 4 | 0,73 | 0,47 | 0,64 | 28,38 | 74 | [[45, 8], [12, 9]] |
| 5 | 0,61 | 0,41 | 0,68 | 35,14 | 74 | [[35, 13], [16, 10]] |
| Обучение на РД+ARIMA | | | | | | |
| Фолд | Точность | F1-score | ROC AUC | Экстр (%) | Кол-во | ConfMat |
| 1 | 0,72 | 0,66 | 0,71 | 41,89 | 74 | [[33, 10], [11, 20]] |
| 2 | 0,62 | 0,52 | 0,68 | 44,59 | 74 | [[31, 10], [18, 15]] |
| 3 | 0,66 | 0,32 | 0,57 | 31,08 | 74 | [[43, 8], [17, 6]] |
| 4 | 0,73 | 0,41 | 0,66 | 28,38 | 74 | [[47, 6], [14, 7]] |
| 5 | 0,64 | 0,37 | 0,57 | 35,14 | 74 | [[39, 9], [18, 8]] |
| Обучение на РД+GBM | | | | | | |
| Фолд | Точность | F1-score | ROC AUC | Экстр (%) | Кол-во | ConfMat |
| 1 | 0,69 | 0,61 | 0,64 | 41,89 | 74 | [[33, 10], [13, 18]] |
| 2 | 0,61 | 0,43 | 0,59 | 44,59 | 74 | [[34, 7], [22, 11]] |
| 3 | 0,65 | 0,38 | 0,55 | 31,08 | 74 | [[40, 11], [15, 8]] |
| 4 | 0,69 | 0,38 | 0,63 | 28,38 | 74 | [[44, 9], [14, 7]] |
| 5 | 0,65 | 0,48 | 0,63 | 35,14 | 74 | [[36, 12], [14, 12]] |
| Обучение на РД+(GN+SMA) | | | | | | |
| Фолд | Точность | F1-score | ROC AUC | Экстр (%) | Кол-во | ConfMat |
| 1 | 0,72 | 0,60 | 0,72 | 41,89 | 74 | [[37, 6], [15, 16]] |
| 2 | 0,61 | 0,51 | 0,64 | 44,59 | 74 | [[30, 11], [18, 15]] |
| 3 | 0,77 | 0,56 | 0,67 | 31,08 | 74 | [[46, 5], [12, 11]] |
| 4 | 0,70 | 0,39 | 0,56 | 28,38 | 74 | [[45, 8], [14, 7]] |
| 5 | 0,76 | 0,65 | 0,77 | 35,14 | 74 | [[39, 9], [9, 17]] |
| Обучение на РД+TimeGAN | | | | | | |
| Фолд | Точность | F1-score | ROC AUC | Экстр (%) | Кол-во | ConfMat |
| 1 | 0,69 | 0,62 | 0,75 | 41,89 | 74 | [[32, 11], [12, 19]] |
| 2 | 0,64 | 0,58 | 0,67 | 44,59 | 74 | [[30, 11], [15, 18]] |
| 3 | 0,74 | 0,49 | 0,59 | 31,08 | 74 | [[46, 5], [14, 9]] |
| 4 | 0,76 | 0,47 | 0,70 | 28,38 | 74 | [[48, 5], [13, 8]] |
| 5 | 0,70 | 0,52 | 0,67 | 35,14 | 74 | [[40, 8], [14, 12]] |

*Источник:* составлено авторами.

DOI: 10.38050/2078-3809-2025-17-4-9-34

*Theoretical Issues*

# INNOVATIVE APPROACHES TO TRAINING DATA GENERATION FOR OIL DEMAND FORECASTING

**Irina V. Manakhova**
***Doctor of Economics, Professor,***
***Lomonosov Moscow State University, Faculty of Economics;***
***(Moscow, Russia)***

**Aleksandr V. Matytsyn**
***Master's Degree,***
***École Supérieure du Commerce Extérieur (ESCE)***
***(Paris, France);***
***Candidate of Sciences Degree in Economics***
***Lomonosov Moscow State University, Faculty of Economics***
***(Moscow, Russia)***

## Abstract

*This study explores methods for generating training data to improve demand forecasting accuracy in the oil market. The limitations of traditional approaches are examined, and the use of generative adversarial networks, specifically the TimeGAN (Time-series Generative Adversarial Network) model, is proposed for creating synthetic time series data. The results demonstrate that TimeGAN can generate realistic data closely resembling actual data, preserving market volatility and structural characteristics. However, model limitations were identified, suggesting the need for further research to enhance forecast efficiency and accuracy on the oil in volatile market conditions.*

**Keywords:** deep learning, generative adversarial networks, training data, model TimeGAN.

**JEL:** E17, C53.

**For citation:** Manakhova, I.V., Matytsyn, V.M. (2025) Innovative Approaches to Training Data Generation for Oil Demand Forecasting. Scientific Research of Faculty of Economics. Electronic Journal, vol. 17, no. 4, pp. 9-34. DOI: 10.38050/2078-3809-2025-17-4-9-34.

## References

Kaukin A.S., Pavlov P.N., Kosarev V.S. Kratkosrochnoe prognozirovanie tsen na elektroenergiyu s ispol'zovaniem generativnykh neyronnykh setey. Biznes-informatika. 2023. No. 3 (17). P. 7–23. (In Russ.).

Kopytin I.A. Transformatsiya rynka nefti v Evrope: tendentsii i perspektivy. Problemy prognozirovaniya. 2024. No. 6. P. 217–226. https://doi.org/10.47711/0868-6351-207-217-226. (In Russ.).

Rabchevskiy A.N. Obzor metodov i sistem generatsii sinteticheskikh obuchayushchikh dannykh. Prikladnaya matematika i voprosy upravleniya. 2023. No. 4. P. 6–45. https://doi.org/10.15593/2499-9873/2023.4.01. (In Russ.).

Rozentsvayg A.K. Metody ekonometricheskogo modelirovaniya i analiza sotsial'no-ekonomicheskikh yavleniy. Naberezhnye Chelny, 2014. 121 p. https://doi.org/10.13140/RG.2.1.3998.5526. (In Russ.).

Brajard J., Carrassi A., Bocquet M., Bertino L.. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. Journal of Computational Science. 2020. Vol. 44. https://doi.org/101171. 10.1016/j.jocs.2020.101171.

Farzana G.S., Prakash N. Machine Learning in Demand Forecasting. A Review. In Proceedings of the 2nd International Conference on IoT, Social, Mobile, Analytics and Cloud in Computational Vision and Bio-Engineering (ISMAC-CVB 2020). 2020. http://dx.doi.org/10.2139/ssrn.3733548.

Juneja T., Bajaj S., Sethi N. Synthetic Time Series Data Generation Using Time GAN with Synthetic and Real-Time Data Analysis. Proceedings of the International Conference on Advances in Computing and Data Sciences. Springer. 2023. P. 567–576. https://doi.org/10.1007/978-981-99-0601-7_51.

Kanagarathinam K. Comprehensive Overview of Optimization Techniques in Machine Learning Training. Computational Science and Optimization Letters. 2024. Vol. 2. No. 1. P. 69–85. https://doi.org/10.59247/csol.v2i1.69.

Li J., Liu Y., Li Q. Generative Adversarial Network and Transfer Learning Based Fault Detection for Rotating Machinery with Imbalance Data Condition. Measurement Science and Technology. 2022. https://doi.org/33.10.1088/1361-6501/ac3945.

Liu Y., Liang Z., Li X. Enhancing Short-Term Power Load Forecasting for Industrial and Commercial Buildings: A Hybrid Approach Using TimeGAN, CNN, and LSTM. IEEE Open Journal of the Industrial Electronics Society. 2023. No. 4. P. 451–462. https://doi.org/10.1109/OJIES.2023.3319040.

Matta C., Bianchesi N.M., Oliveira M., Balestrassi P., Leal F. A comparative study of forecasting methods using real-life econometric series data. Production. 2021. Vol. 31. https://doi.org/10.1590/0103-6513.20210043.

Moroff N., Kurt E., Kamphues J. Machine Learning and Statistics: A Study for Assessing Innovative Demand Forecasting Models. Procedia Computer Science. 2021. No.180. P. 40–49. https://doi.org/10.1016/j.procs.2021.01.127.

Mushunje L., Allen D., Peiris S. Volatility and Irregularity Capturing in Stock Price Indices Using Time Series Generative Adversarial Networks (TimeGAN). Computational Engineering, Finance, and Science. 2023. https://doi.org/10.48550/arXiv.2311.12987.

Reboredo J.C., Ugolini A. Quantile dependence of oil price movements and stock returns. Energy Economics. 2016. Vol. 54. P. 33–49. https://doi.org/10.1016/j.eneco.2015.11.015.

Sacco M.A., Ruiz J.J., Pulido M., Tandeo P. Evaluation of Machine Learning Techniques for Forecast Uncertainty Quantification. Quarterly Journal of the Royal Meteorological Society. 2021. Vol. 147. No. 34. P. 1234–1251. https://doi.org/10.48550/arXiv.2111.14844.

Yue T, Liu Y. Multi-Scale Price Forecasting Based on Data Augmentation. Applied Sciences. 2024. Vol. 14. No. 19. P. 8737. https://doi.org/10.3390/app14198737.