

# РЫНОК ТОВАРОВ И УСЛУГ

**К. В. Украинский**

аспирант Школы вычислительных социальных наук Европейского университета в Санкт-Петербурге

**Ю. В. Раскина**

канд. экон. наук, доцент Школы вычислительных социальных наук Европейского университета в Санкт-Петербурге

## НЕЛИНЕЙНОЕ ПРОГНОЗИРОВАНИЕ МУЗЫКАЛЬНОГО СПРОСА: ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ ДАННЫХ И ЛОКАЛЬНЫЕ НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

### Введение

Цифровые потоковые платформы, такие как Spotify и Яндекс Музыка, не только упростили доступ к музыкальным произведениям, но и трансформировали структуру потребления, превратив факт прослушивания в основу новой модели конвертации пользовательского внимания в доходы. Современный спрос на музыкальные произведения определяется как индивидуальными предпочтениями пользователей, так и особенностями цифровой среды, оказывающими влияние на динамику популярности, распределение доходов авторов (роялти) и стратегию поведения ключевых участников рынка — исполнителей, владельцев прав на произведения (лейблов) и цифровых платформ. Исследования показывают, что доступ к обширному каталогу музыкальных произведений стимулируют общую диверсификацию музыкальной продукции (Anderson et al., 2020). На начальных этапах взаимодействия с платформами пользователи активно исследуют доступные плейлисты и рекомендуемые подборки, однако со временем этот эффект уменьшается, что приводит к увеличению доли «повторных прослушиваний» (Datta, Knox, Bronnenberg, 2018).

В условиях высокой конкуренции и информационной асимметрии, присущих музыкальному рынку, важной становится задача прогнозирования спроса на ранней стадии жизненного цикла — на основе ограниченного объема информации о поведении пользователей в недели или месяцы после выхода композиции. Возможность сделать точный прогноз в этот момент имеет ключевое значение для принятия решений о дальнейшем продвижении, распределении ресурсов и оценке потенциальной рыночной успешности музыкального произведения.

Особенно критичным является начальный период после выхода композиции. В одном из ключевых исследований в данной области (Kaimann, Tanneberg, Cox, 2021) показано, что ранний «органический» успех является основной причиной появления в чартах, а композиции, которые занимают более высокие дебютные позиции, имеют тенденцию дольше удерживаться в рейтинге по сравнению с треками, входящими в чарт медленнее или с более низкой стартовой позицией. Парадокс заключается в том, что произведения, демонстрирующие стремительный рост, часто имеют в последующем резкое снижение популярности, тогда как более плавный старт предвещает сниженный, но более стабильный спрос. Современные произведения либо достигают «мгновенного» успеха, либо не имеют успеха вовсе (Schneider, Gros, 2019).

Исследования, посвященные повторяющимся рекомендациям, показали, что многократное упоминание новых произведений в алгоритмических подборках

может стимулировать более высокие показатели возврата пользователей к этим трекам (Manolovitz, Ogihara, 2020). Анализ прослушиваемого многообразия показал, что пользователи, полагающиеся на рекомендации, получают более узкий спектр прослушиваний по сравнению с самостоятельным выбором контента (Anderson et al., 2020). Подобный результат может отражать как стремление слушать похожую музыку, так и то, что сами рекомендации направляют слушателей к ограниченному набору треков (Datta, Knox, Bronnenberg, 2018). Динамика поведения пользователей подвержена воздействиям как экзогенных факторов, включая возраст и вкусовые предпочтения, так и эндогенных аспектов, связанных с этапами взаимодействия с интерфейсом, контентом и плейлистами. В результате формируется сложная система взаимозависимостей, что подчеркивает необходимость разработки «глубоких» моделей для их комплексной оценки (Mok et al., 2022).

Все эти феномены формируют нелинейную динамику, в которой спрос становится функцией не только качества композиции, но и структуры цифровой среды, в которой она распространяется. В терминах экономической теории это соответствует модели с обратной связью, в которой пользовательский интерес, алгоритмическая дистрибуция и последующая монетизация образуют взаимозависимую систему.

Прогнозирование спроса в таком контексте — не просто задача применения методов анализа данных, а экономическая задача оценки будущих потоков дохода, распределения внимания и управления рисками продвижения. Для платформ решение таких задач — это основа для оптимизации архитектуры рекомендаций и удержания пользователей; для артистов — инструмент принятия решений о моменте выпуска трека, объеме маркетинговой поддержки и стратегии позиционирования; для лейблов и инвесторов — возможность оценки вероятности выхода произведения на траекторию высокой доходности.

Существующие методы анализа временных рядов нередко предполагают линейность или стационарность поведения, что ограничивает их применимость в условиях музыкального рынка, где спрос подвержен резким скачкам, «вирусному» росту и последующим фазам затухания. Для анализа таких процессов требуется методология, способная выявлять нелинейные закономерности поведения на раннем этапе и учитывать сложную динамику цифровой среды.

Для исследования данного феномена в настоящей работе предлагается подход, основанный на функциональном анализе данных (Functional Data Analysis, FDA) и непараметрических регрессионных методах. Это позволяет выявить скрытые шаблоны поведения на ранних этапах взаимодействия пользователей с контентом, что особенно актуально в контексте исследований, подчеркивающих важность «стартового успеха» для долгосрочного удержания в чартах (Luz López García et al., 2015; Ramsay, Silverman, 2005). Использование производных от траекторий прослушиваний вместо их абсолютных значений позволяет анализировать скорость изменения спроса, выявляя сходные фазы роста, спада и нелинейных закономерностей, свойственных динамическим системам.

Цель исследования заключается в анализе устойчивых закономерностей спроса и жизненного цикла произведения на цифровых музыкальных платформах и оценке того, влияет ли способ привлечения пользователя (через алгоритмические рекомендации или самостоятельный выбор) на динамику спроса. Основная гипотеза состоит в том, что поведение слушателей характеризуется относительной стабильностью, а алгоритмы лишь модифицируют степень выраженности существующих тенденций, не изменяя их базового характера. Если обнаруженные структурные паттерны окажутся схожими для всех композиций, то это может свидетельствовать

в пользу того, что алгоритмические рекомендации влияют исключительно на усиление или ослабление уже заложенных динамических особенностей, не внося принципиальных изменений в их структуру. Такой комплексный метод позволяет детально охарактеризовать эволюцию спроса и уточнить, что базовая динамика прослушиваний формируется внутренними особенностями спроса, модифицируемыми, но не трансформируемыми алгоритмическими механизмами. Кроме того, зависимость поведения от начального состояния демонстрирует наличие нелинейного детерминизма в исследуемом процессе, что позволяет рассматривать цифровые платформы как динамические системы с обратной связью (Barnett, 1990; Decoster, Mitchell, 1991).

### Методология исследования

Экономические временные ряды часто демонстрируют специфические свойства, такие как смена состояний и изменение волатильности, что затрудняет выявление нелинейных детерминированных и нелинейных стохастических процессов. Современные работы показывают, что использование локальной непараметрической аппроксимации временного ряда может быть эффективным подходом для выявления и использования нелинейных динамических паттернов, опираясь на идею, что краткосрочная динамика вблизи некоторого состояния может быть аппроксимирована близкими («соседними») точками из обучающей выборки (Jaditz, Sayers, 1993). Такие методы используются в ситуации, когда временные ряды проявляют высокую волатильность или смену динамики, например исследователи ищут похожие паттерны в прошлом, ожидая, что временной ряд будет вести себя схожим образом в будущем (Agnon, Golan, Shearer, 1999; Álvarez-Díaz, 2020; Wang et al., 2024).

Для оценки эффективности локального непараметрического подхода проводится сравнение результатов прогнозирования с фактическими значениями из прогнозной выборки. В качестве инструмента прогнозирования выбран метод регрессии «ближайших соседей» (*k* Nearest Neighbors regression, *k*-NN). Данный метод не требует строгих предположений о распределении данных или наличии трендов, что делает его особенно полезным для анализа временных рядов с нелинейной структурой (Sugihara, May, 1990; Tang, Pan, Yao, 2018).

Обзорные работы, посвященные современным методам прогнозирования функциональных временных рядов, отмечают, что, несмотря на существование более продвинутых методов оценки (например, Гауссова ядерная регрессия, Gaussian Kernel), метод ближайших соседей во многих случаях не уступает им в точности. При этом он выигрывает за счет простоты реализации и отсутствия необходимости в тонкой настройке параметров (Kärnä, Lendasse, 2007; Zhang, Parnell, 2023).

В рамках данного исследования анализ динамики спроса на музыкальные композиции был проведен по следующему алгоритму:

1. Разделение данных на выборки: выделена выборка подгонки, используемая для построения модели, и прогнозная выборка, предназначенная для оценки качества прогнозов.
2. Сглаживание временного ряда: временные ряды предварительно сглаживались сплайнами, а затем дифференцировались, чтобы устранить случайные колебания и выявить основные тенденции поведения.
3. Снижение размерности: для снижения размерности данных применялся функциональный анализ главных компонент (FPCA), позволяющий выделить наиболее значимые паттерны прослушиваний.

4. Кластеризация: полученные компоненты были использованы для группировки композиций с похожей динамикой с помощью алгоритма  $k$ -средних.

Прогнозирование осуществлялось методом ближайших соседей ( $k$ -NN), который основывается на предположении, что схожие траектории в прошлом позволяют предсказывать будущее поведение спроса. Эффективность предложенного подхода проверялась сравнением с авторегрессионной моделью первого порядка, AR (1).

Временные ряды в исследовании представлены в виде дискретных месячных наблюдений  $y_{it}$  и аппроксимируются в гильбертовом пространстве  $H$  гладкими функциями  $\hat{f}_i(t)$ , ( $t \in [0, T]$ ) с помощью сплайнов. Функциональное представление  $\hat{f}_i(t)$  используется в процедурах сглаживания, дифференцирования и FPCA, в то время как для прогноза и расчета расстояний в методе  $k$ -NN применяются дискретные значения  $y_{it}$ , соответствующие наблюдаемым моментам времени. Производные вычисляются как приближенные конечные разности, а расстояния между временными рядами определяются по координатам в пространстве главных компонент, рассчитанным по дискретным точкам.

### Описание данных

Настоящее исследование охватывает 802 музыкальных произведения, представленных в виде временных рядов с месячными наблюдениями за период с июня 2021 г. по сентябрь 2024 г. Временной ряд  $y_{it}$  отражает общее количество прослушиваний композиции  $i$  в месяц  $t$ . Фактические наблюдения нормализованы с помощью логарифмической трансформации для уменьшения асимметрии данных. Данные были собраны из открытых источников путем автоматизированного сбора информации с сайтов музыкальных платформ, таких как Spotify, Pandora. В связи с тем, что большое количество исполнителей генерирует крайне мало прослушиваний (эффект «длинного хвоста») (Dewan, Ramaprasad, 2012) в выборку включены только те композиции, общее количество прослушиваний которых превышает 100 000 за весь период наблюдений. Это позволяет сосредоточиться на наиболее популярных треках и избежать искажений, вызванных наличием большого числа малоизвестных произведений.

При сравнении данных исключаются произведения, вышедшие в ту же неделю, а также произведения того же исполнителя или композитора. Это позволяет исключить влияние произведений с одинаковыми или очень близкими датами выхода, удостовериться, что поведение не зависит от одинаковых факторов.

### Разделение данных на выборки

С практической точки зрения выборка состоит из  $n$  временных рядов, каждый из которых рассматривается как реализация стохастического процесса  $f_i(t)$ , где  $t = 1, 2, \dots, T$ ,  $i = 1, 2, \dots, n$ . Как отмечено выше, наблюдения разбиваются на две выборки: прогнозную выборку  $P$ , используемую для оценки качества прогноза, и выборку «подгонки» (fitting)  $F$ , на основе которой оценивается модель.

Прогнозная выборка  $P$  определяется как

$$P = \{y_{it} | t = N_f + 1, \dots, T\},$$

где  $y_{it}$  — фактическое (наблюдаемое) значение прослушиваний композиции  $i$  в момент времени  $t$ ;  $N_f$  — момент времени, разделяющий прогнозную выборку и выборку подгонки;  $T$  — конечный момент времени, до которого доступны данные.

Выборка подгонки  $F$  включает наблюдения до момента времени  $N_f$ :

$$F = \{y_{it} \mid t = 1, \dots, N_f\}.$$

В работе используются фиксированные значения  $N_f$  (2, 4, 8 месяцев) для оценки влияния разной длины окна на качество прогноза.

### Сглаживание временного ряда

Выбор сплайнового сглаживания обусловлен несколькими причинами. Во-первых, сплайны обеспечивают хорошую гибкость и способны эффективно описывать нелинейную динамику спроса без необходимости заранее специфицировать форму функциональной зависимости. Во-вторых, сплайновые функции позволяют явно контролировать степень гладкости с помощью параметра регуляризации, что важно для устранения случайных колебаний и выделения общих трендов. Сплайны широко признаны в эконометрической и статистической литературе как надежный и проверенный инструмент функционального анализа временных рядов (Ramsay, Silverman, 2005; Silverman, 1985).

Для сглаживания временного ряда наблюдаемые значения функции  $y_{it}$  аппроксимируются линейной комбинацией базисных функций:

$$\hat{f}_i(t) = \sum_{j=1}^{K+d} c_{ij} \cdot \Phi_j(t),$$

где  $\Phi_j(t)$  — базисная функция;  $c_{ij}$  — коэффициенты базисной функции  $j$  для ряда  $i$ ;  $K$  — количество внутренних узлов;  $d$  — степень сплайна.

Коэффициенты базисной функции находятся решением следующей оптимизационной задачи (Silverman, 1985):

$$\min_{c_i} \sum_t \left( y_{it} - \sum_{j=1}^{K+d} c_{ij} \cdot \Phi_j(t) \right)^2 + \lambda \cdot c_i \cdot R c_i,$$

где  $\lambda$  — параметр регуляризации, чем больше значение параметра, тем более гладкой будет полученная функция;  $R$  — матрица штрафов, элементы которой задаются интегралами от произведений производных базисных функций.

Далее полученная функция  $\hat{f}_i(t)$  дифференцируется для вычисления производной порядка  $p$ :

$$\hat{f}_i^{(p)}(t) = \sum_{j=1}^{K+d} c_{ij} \cdot \Phi_j^{(p)}(t).$$

В данной работе степень сплайна  $d$  и тип базисных функций были выбраны исходя из общепринятой практики функционального анализа временных рядов. В частности, для анализа экономических и социальных данных чаще всего используются кубические сплайны ( $d = 3$ ), поскольку они обеспечивают баланс между гладкостью и гибкостью функции. Кубические сплайны предпочтительны также потому, что обладают непрерывностью и гладкостью первых двух производных. В настоящем исследовании непрерывность и гладкость первой производной имеют особое значение, поскольку она позволяет корректно и без искажений оценивать скорости изменения спроса на музыкальные композиции, что используется для последующего анализа нелинейной динамики спроса.

В качестве базисных функций использованы В-сплайны, широко применяемые в задачах сглаживания и аппроксимации из-за их численной устойчивости и простоты реализации (Ramsay, Silverman, 2005).

Вид и гладкость функции напрямую зависят от параметра регуляризации  $\lambda$ . С ростом параметра функция сходится в простую линейную регрессию, а при приближении к нулю функция становится линейно-кусочной со средним значением в каждой временной точке. Обычно этот параметр выбирается так, чтобы обеспечить оптимальный баланс между гладкостью кривой и точностью аппроксимации исходных данных, основные подходы — обобщенная кросс-валидация или эвристический.

Рис. 1 отображает результат сглаживания функций для временного ряда.

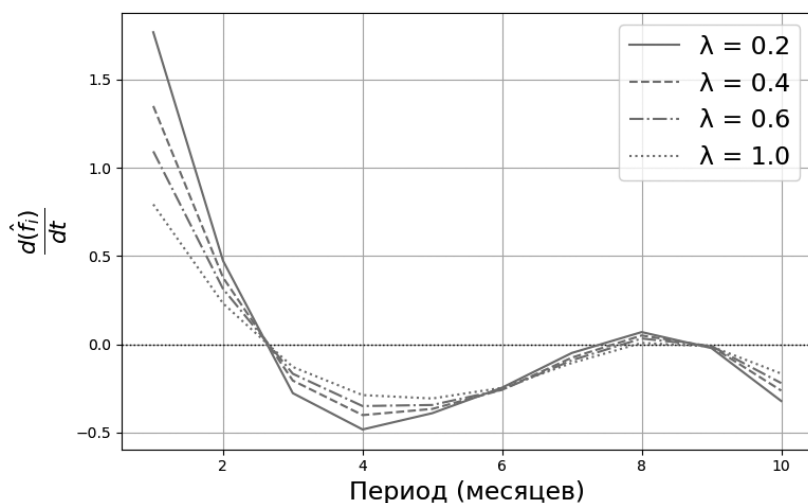


Рис. 1. Первая производная сглаженной кривой для различных  $\lambda$  для первых 10 точек одного (случайного) ряда

В данной работе для определения оптимального параметра регуляризации используется метод обобщенной кросс-валидации (General Cross Validation, GCV). Оптимальное значение  $\lambda$  определяется минимизацией критерия  $GCV(\lambda)$ , отражающего среднюю обобщенную ошибку прогнозирования при разных параметрах регуляризации.

Результаты работы метода представлены в табл. 1, значения всех  $GCV(\lambda)$  усреднены для всех временных рядов, оптимальное значение  $\lambda$  — 0,1.

Таблица 1

Значение параметра GCV при разных  $\lambda$

Параметр $\lambda$	GCV( $\lambda$ )
0,1	1,4159
0,2	1,4627
0,4	1,5078
0,6	1,5498
0,8	1,5882
1,0	1,6269

### Снижение размерности

Для приведения данных к общему масштабу и сокращения их размерности проводится функциональный анализ главных компонент (Functional principal component analysis, FPCA), задачей которого является замена наблюдаемых сглаженных значений  $f_i(t)$  на расчетный параметр главной компоненты  $\xi_i$  (Zhou, Wei, Yao, 2022).

Предполагая, что аппроксимированная функция  $\hat{f}_j^{(p)}(t)$  воспроизводит структуру ковариации, ковариационную функцию можно найти как

$$C(s, t) = \text{Cov}(\hat{f}_i(s), \hat{f}_i(t)),$$

где  $s$  и  $t$  — две независимые точки из области определения функции.

Обозначим эмпирическую оценку ковариационной матрицы коэффициентов как:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n c_i c_i^T,$$

где  $c_i \in \mathbb{R}^{K+d}$  — вектор коэффициентов базисных функций для  $i$ -го ряда.

Тогда оценка ковариационной функции принимает вид:

$$\hat{C}(s, t) = \varphi^{(p)}(s)^T \hat{\Sigma} \varphi^{(p)}(t),$$

где  $\varphi^{(p)}(t) = [\varphi_1^{(p)}(t), \dots, \varphi_{K+d}^{(p)}(t)]$  — вектор производных базисных функций.

Для получения функциональных главных компонент необходимо провести спектральное разложение оцененной матрицы  $\hat{\Sigma}$ .

В результате анализа были выбраны три главные компоненты, которые позволяют «сжать» исходную информацию об изменениях спроса в несколько ключевых паттернов. Рис. 2 демонстрирует график доли объясненной дисперсии для трех главных компонент (при  $\lambda = 0, 1$  и 5 временных точках), демонстрируя вклад каждой компоненты в общее разнообразие данных. При этом первая компонента объясняет около 90% вариации, что указывает на высокую концентрацию признаков вокруг одного направления и упрощает дальнейший анализ.

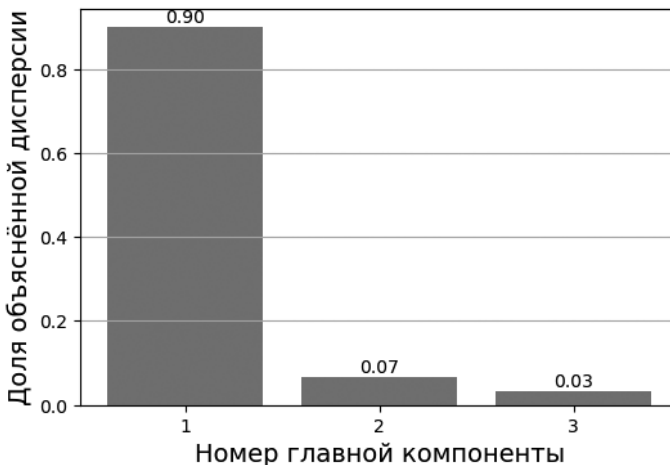


Рис. 2. Доля объясненной дисперсии по компонентам

## Кластеризация

В данной работе коэффициенты главных компонент используются для кластеризации временных рядов при помощи алгоритма  $k$ -средних ( $k$ -means) (Luz López García, García-Ródenas, González Gómez, 2015). Алгоритм определяется следующим образом:

Пусть выбрано  $M \leq K + d$  наибольших собственных значений  $\hat{\lambda}_k$ . Тогда для  $i$ -го временного ряда его координаты в пространстве  $M$  главных компонент примут вид:

$$\hat{\xi}_{ik} = c_i^T \hat{q}_k,$$

где  $k = 1, \dots, M$  — номер главной компоненты;  $\hat{q}_k$  —  $k$ -й собственный вектор эмпирической ковариационной матрицы коэффициентов базисного разложения  $i$ .

Алгоритм кластеризации выглядит следующим образом. Пусть задано желаемое количество кластеров  $K^*$ .

1. Случайно инициализируем центры кластеров:  $\mu_1^{(0)}, \dots, \mu_{K^*}^{(0)}$
2. На каждом шаге  $l + 1$ :
  - а) назначаем каждому  $\xi_i$  метку кластера, решая задачу:

$$r_i^{(l+1)} = \arg \min_{1 \leq u \leq K^*} \|\xi_i - \mu_u^{(l)}\|,$$

где:  $r_i$  — метка кластера для  $i$ -го ряда,  $u$  — индекс кластера,  $u = 1, \dots, K^*$ .

б) обновляем центры кластеров:

$$\mu_u^{(l+1)} = \frac{1}{\left| \{i : r_i^{(l+1)} = u\} \right|} \sum_{i: r_i^{(l+1)} = u} \hat{\xi}_i.$$

3. Продолжаем итерации до сходимости.

Таким образом метод  $k$ -средних, рассчитываемый по найденным параметрам  $\hat{\xi}_i$  позволяет сгруппировать временные ряды исходя из их проекций на главные компоненты, что дает удобную интерпретацию сходства или различий между динамикой поведения временного ряда.

Количество кластеров ( $K^* = 3$ ) было определено при помощи силуэтного анализа (Silhouette analysis), который показал, что максимальное значение метрики (0,615) достигается при трех кластерах. Это свидетельствует о наиболее четком разделении данных именно на три группы (Rousseeuw, 1987).

На рис. 3 показаны собственные векторы, полученные в результате разложения временных рядов, а также примеры сглаженных кривых производных, демонстрирующие характерные паттерны динамики.

На нижних графиках видно, что композиции разделились на три кластера со схожей динамикой. В кластере 1 большинство кривых демонстрируют рост популярности, в кластере 2 наблюдается скорее убывающий или волнообразный тренд, кластер 3 можно считать «смешанным» вариантом, где кривые имеют умеренный подъем и более плавную динамику.

Таким образом, три главные компоненты позволили «сжать» исходную информацию об изменениях спроса в несколько ключевых паттернов, а кластеризация по этим паттернам выявила три основные группы музыкальных произведений, отличающиеся формой траектории популярности.

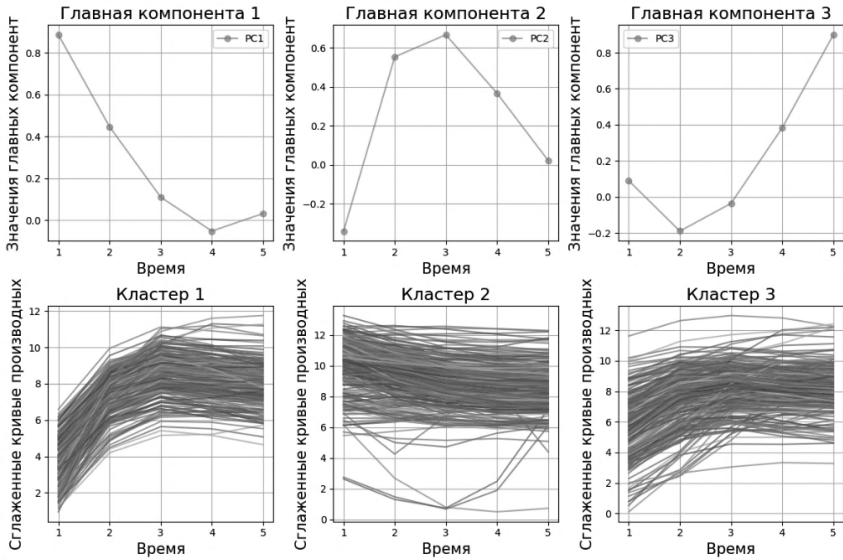


Рис. 3. Кластеризация по частичному сплайну  $\lambda$  на примере недельных данных по первым 5 точкам

### Прогнозирование методом «ближайших соседей»

В основе  $k$ -NN-прогнозирования лежит предположение, что значения целевой переменной  $y_{it}$  в будущий момент времени будут похожи на значения аналогичных периодов с «близкой» динамикой. Для поиска таких «близких» периодов для каждого наблюдения  $i$  формируется вектор компонент  $\hat{\xi}_i = (\hat{\xi}_{i1}, \dots, \hat{\xi}_{im})$ , где  $\hat{\xi}_{iz}$  — координата  $i$ -го ряда по  $z$ -й главной компоненте. Наиболее распространенным подходом является поиск Евклидова расстояния между координатами, но в статье (Kidron, Klein, 2007) предложен улучшенный подход, чтобы скорректировать влияние разномасштабных показателей, где каждая компонента нормируется по среднему и стандартному отклонению  $\sigma_s$  в соответствующем кластере. Расстояние между вектором  $\xi_i$  и вектором  $\xi_s$  из кластера  $k$  вычисляется следующим образом:

$$d(\xi_i, \xi_s) = \sqrt{\sum_{z=1}^m \left( \frac{\hat{\xi}_{iz} - \hat{\xi}_{sz}}{\sigma_z} \right)^2}.$$

После того, как найдены все расстояния до всех точек вектора компонент  $\xi_i$ , выбирается  $k$  ближайших соседей. Для учета «вклада» каждого вектора применяется средневзвешенный подход:

$$\hat{y}_i(t) = \frac{\sum_{s=1}^k w_s y_s(t)}{\sum_{s=1}^k w_s},$$

где  $w_s = \frac{1}{d(\hat{\xi}_i, \hat{\xi}_s)}$  — вес, убывающий с ростом расстояния между  $i$  и  $s$ .

Оптимальный  $k$  подбирается перебором по наименьшей средней абсолютной ошибке (MAE). Для данной задачи подобранное оптимальное  $k = 8$ .

Наряду с методом  $k$ -NN, в качестве проверочной (базовой) модели берется авторегрессионная модель первого порядка AR (1), формулируемая как:

$$y_{it} = \alpha + \beta \cdot y_{it-1} + \varepsilon_{it},$$

где  $\alpha, \beta$  — параметры такой модели;  $\varepsilon_{it}$  — ошибка модели.

Модель AR(1) обучается на всей «известной» выборке  $F$ , и дальнейшие предсказания осуществляются по принципу «скользящего окна», используя предыдущее предсказанное значение для оценки следующего. Предполагается, что сравнительный анализ  $k$ -NN и AR(1) продемонстрирует, насколько существенен вклад нелинейных факторов в динамику временных рядов и действительно ли предложенный метод превосходит типичные линейные подходы.

## Результаты

Табл. 1 показывает результаты сравнительного прогноза при разных горизонтах, иллюстрируя типичное поведение временных рядов, в которых присутствуют элементы нелинейной динамики. Для оценки точности прогнозов использовались RMSE (Root Mean Square Error) — среднеквадратичная ошибка, которая более чувствительна к большим отклонениям (чем ниже значение, тем лучше модель) и MAE (Mean Absolute Error) — средняя абсолютная ошибка, показывающая среднюю величину отклонения прогноза от фактических значений. Эти метрики демонстрируют, что при малой начальной выборке  $N_f = \{2, 4\}$  ошибки AR(1) быстро возрастают уже на горизонте в 2–3 шага, что согласуется с гипотезой о наличии в данных нелинейных паттернов. В такой ситуации локальные непараметрические методы, используя «ближайшие» аналоги по траектории прослушиваний, лучше воспроизводит быстро меняющуюся динамику, тогда как AR(1) не способна отразить резкие и непредсказуемые скачки. Ухудшение прогноза AR(1) с увеличением горизонта может отражать чувствительность линейной модели к накоплению ошибок при существенной нелинейности во временном ряду. Напротив,  $k$ -NN, находя «схожие» отрезки динамики, способно успешно справляться с возмущениями в системе даже на больших отрезках, что согласуется с предположением о наличии локального детерминизма и сложных взаимодействий.

Таблица 1

Расчет MAE и RMSE для различного горизонта прогноза

Горизонт прогноза, месяцы	Метод $k$ -NN, RMSE	Метод AR(1), RMSE	Метод $k$ -NN, MAE	Метод AR(1), MAE
$N_f = 2$				
1	0,679	7,294	0,625	4,315
2	0,774	29,713	0,680	14,222
3	0,902	111,277	0,693	40,789
4	1,010	436,785	0,721	121,241
$N_f = 4$				
1	0,573	0,601	0,360	0,412
2	0,724	0,886	0,476	0,579
3	0,687	1,507	0,490	0,760
4	0,727	2,800	0,538	1,041
$N_f = 8$				
1	0,532	0,445	0,382	0,333
2	0,616	0,568	0,460	0,420
3	0,646	0,640	0,480	0,473
4	0,712	0,736	0,534	0,542

На горизонте  $N_f = 8$  и далее метод AR(1) несущественно превосходит  $k$ -NN, указывая на то, что часть ряда после начального старта может иметь умеренно линейные фрагменты или автокорреляцию. При коротких горизонтах прогноза линейные модели нередко превосходят «сложные» подходы, но в условиях возрастающей нелинейности и при удаленном горизонте точность линейных прогнозов быстро падает (рис. 4) (Jaditz, Sayers, 1993).

Для проверки значимости результатов применяется тест Diebold–Mariano, в котором нулевая гипотеза формулируется так: «Два рассматриваемых прогноза имеют одинаковую среднюю точность», то есть разница в их ошибках статистически не отлична от нуля. Отрицательное (или положительное) значение статистики DM указывает, какая именно из моделей дает более низкую ошибку.

Результаты DM-теста (табл. 2) подтверждают неоднородную картину сравнительных прогнозных свойств моделей. Для  $N_f = 2$  и  $N_f = 4$  все значения DM-статистики отрицательны и значимы, указывая, что метод  $k$ -NN дает существенно меньшую ошибку, чем AR(1). Это согласуется с предыдущими выводами о нелинейной природе временных рядов и превосходстве локальных методов в улавливании динамики.

Таблица 2

Результаты Diebold-Mariano теста для  $N_f = \{2, 4, 8\}$

Горизонт прогноза, месяцы	DM	$p$ -значение
$N_f = 2$		
1	-6,790	0,0
2	-4,651	0,0
3	-3,434	0,0003
4	-2,539	0,0111
$N_f = 4$		
1	-2,183	0,0291
2	-3,300	0,0031
3	-2,791	0,0053
4	-2,495	0,0126
$N_f = 8$		
1	2,019	0,0435
2	0,254	0,7996
3	-0,776	0,4378
4	-1,644	0,1001
5	-3,141	0,0017
6	-4,099	0,0

В случае  $N_f = 8$  наблюдается иная картина: при первом шаге прогноза значение теста DM оказывается положительным и статистически значимым ( $DM = 2,019$ ,  $p = 0,0435$ ), что говорит о превосходстве AR(1) на совсем коротком отрезке. Однако далее (шаги 2–4) статистика указывает на отсутствие статистически значимых различий, а начиная с шага 5 метод  $k$ -NN снова становится лучше (отрицательные и значимые значения DM). Подобная смена лидерства свидетельствует о том, что при достаточно «гладком» поведении ряда линейная компонента AR(1) может краткосрочно давать преимущества, но по мере нарастания нелинейных эффектов локальный подход  $k$ -NN снова выходит на первое место и показывает более точные результаты.

Рис. 4 показывает коэффициент детерминации  $R^2$  для моделей  $k$ -NN и AR(1) для  $N_f = 8$ . Видно, что изначально обе модели показывают довольно высокую точность, но со временем качество прогноза закономерно снижается. При этом  $k$ -NN всегда сохраняет положительные значения  $R^2$ , в то время как AR(1) быстро переходит в отрицательные значения. Другими словами, даже на отдаленных шагах прогноза сохраняется некоторый общий паттерн, что может указывать на сходство динамики в зависимости от начальных условий.

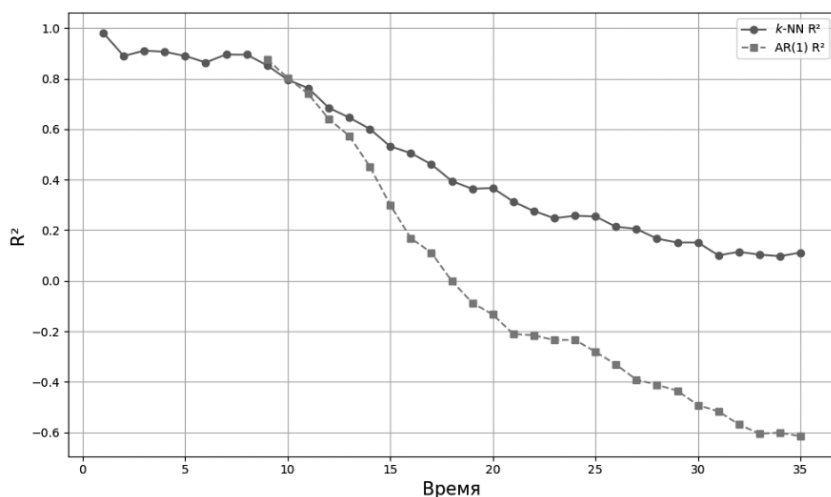


Рис. 4. Коэффициент детерминации  $R^2$   $k$ -NN и AR (1), скользящее окно при  $N_f = 8$

На приведенных графиках (рис. 5) показаны реальные траектории прослушиваний (линия) для четырех случайно выбранных произведений прогнозы двух моделей:  $k$ -NN (пунктирная кривая, тире) и AR(1) (пунктирная линия, точки).

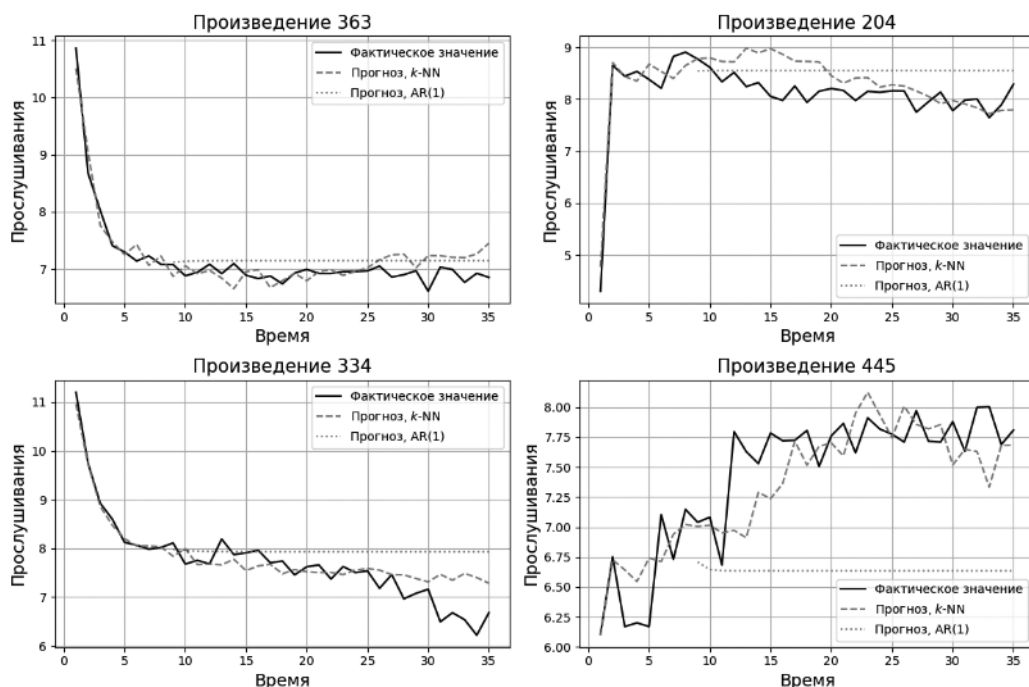


Рис. 5. Прогнозные и фактические траектории временных рядов для четырех случайных композиций

Видно, что когда на ранних этапах динамика похожа (рост или спад), то близкие по поведению произведения имеют схожую динамику в дальнейшем. Произведения ниже демонстрируют именно такой сценарий — раннее схожее поведение приводит к высокому сходству траекторий и в более поздние моменты времени.

### Обсуждение

Анализ показал, что сглаживание сплайнами адекватно восстанавливает ранние динамические особенности прослушиваний. Даже при малой выборке ( $N_f = 2$  и  $N_f = 4$ ) метод  $k$ -NN заметно превосходит AR(1), что подтверждается тестом Diebold–Mariano (отрицательные значения статистики DM при  $p < 0,05$ , табл. 3). Это указывает на выраженную нелинейность временных рядов и чувствительность  $k$ -NN к быстрым изменениям спроса.

Графики фактических и прогнозных траекторий (рис. 5) демонстрируют, что сравнение с ближайшими соседями по первым нескольким точкам временного ряда позволяет получить статистически значимый и более точный прогноз дальнейшей эволюции спроса по сравнению с линейными моделями, указывая на предсказуемость траектории популярности уже на ранних этапах. Наблюдаемая зависимость динамики от начального состояния демонстрирует наличие нелинейного детерминизма, что позволяет рассматривать цифровые платформы как динамические системы с обратной связью.

Полученные результаты свидетельствуют в пользу ключевой гипотезы исследования: динамика спроса на музыкальные произведения определяется устойчивыми внутренними паттернами, которые остаются неизменными по своему базовому характеру, а алгоритмические механизмы лишь усиливают или ослабляют выраженность этих тенденций. Эти выводы подчеркивают сложное взаимодействие пользователей и алгоритмов, а также актуальность дальнейших исследований нелинейной динамики в социальных и культурных процессах.

### Заключение

В рамках исследования разработан и протестирован подход на основе функционального анализа данных (FDA) и локальных непараметрических методов ( $k$ -NN). Сравнение с моделью AR(1) показало, что предложенный метод точнее отражает ранние резкие изменения спроса жизненного цикла произведений. Результаты подчеркивают важность начального этапа и демонстрируют эффект обратной связи в системе. Имеется практическая ценность для музыкальной индустрии, так как точное понимание ранней динамики помогает эффективнее продвигать музыкальные произведения. Перспективы дальнейших исследований включают расширение набора признаков и применение более сложных нелинейных моделей.

### Список литературы

*Agnon Y., Golan A., Shearer M.* Nonparametric, Nonlinear, Short-Term Forecasting: Theory and Evidence for Nonlinearities in the Commodity Markets // *Economics Letters*. 1999. Vol. 65. N 3. P. 293–299.

*Álvarez-Díaz M.* Is it Possible to Accurately Forecast the Evolution of Brent Crude Oil Prices? An Answer Based on Parametric and Nonparametric Forecasting Methods // *Empirical Economics*. 2020. Vol. 59. N 3. P. 1285–1305.

*Anderson A., Maystre L., Anderson I., Mehrotra R., Lalmas M.* Algorithmic Effects on the Diversity of Consumption on Spotify. Proceedings of The Web Conference 2020. Taipei Taiwan, ACM, 2020. P. 55–65.

- Barnett W. Econometrics of Chaos // *Social Science Computer Review*. 1990. Vol. 8. N 4. P. 528–531.
- Datta H., Knox G., Bronnenberg B. J. Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery // *Marketing Science*. 2018. Vol. 37. N 1. P. 5–21.
- Decoster G. P., Mitchell D. W. The Efficacy of the Correlation Dimension Technique in Detecting Determinism in Small Samples // *Journal of Statistical Computation and Simulation*. 1991. Vol. 39. N 4. P. 221–229.
- Dewan S., Ramaprasad J. Research Note — Music Blogging, Online Sampling, and the Long Tail // *Information Systems Research*. 2012. Vol. 23. N 3-part-2. P. 1056–1067.
- Jaditz T., Sayers C. L. Is Chaos Generic in Economic Data? // *International Journal of Bifurcation and Chaos*. 1993. Vol. 03. N 03. P. 745–755.
- Kaimann D., Tanneberg I., Cox J. “I Will Survive”: Online Streaming and the Chart Survival of Music Tracks // *Manage Decis Econ*. 2021. Vol. 42. N 1. P. 3–20.
- Kärnä T., Lendasse A. Comparison of FDA Based Time Series Prediction Methods // *European Symposium on Time Series Prediction*. 2007. P. 77–86.
- Kidron A., Klein S. T. An Information Retrieval Approach to Predicting Meteorological Data // *International Journal of Modelling and Simulation*. 2007. Vol. 27. N 3. P. 218–225.
- Luz López García M., García-Ródenas R., González Gómez A. K-means Algorithms for Functional Data // *Neurocomputing*. 2015. Vol. 151. P. 231–245.
- Manolovitz B., Ogihara M. Practical Evaluation of Repeated Recommendations in Personalized Music Discovery. Zenodo, 2020.
- Mok L., Way S. F., Maystre L., Anderson A. The Dynamics of Exploration on Spotify // *ICWSM*. 2022. Vol. 16. P. 663–674.
- Ramsay J. O., Silverman B. W. *Functional Data Analysis*. New York, 2005.
- Rousseeuw P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis // *Journal of Computational and Applied Mathematics*. 1987. Vol. 20. P. 53–65.
- Schneider L., Gros C. Five decades of US, UK, German and Dutch music charts show that cultural processes are accelerating // *Royal Society Open Science*. 2019. Vol. 6. N 8. P. 19–94.
- Silverman B. W. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting // *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 1985. Vol. 47. N 1. P. 1–21.
- Sugihara G., May R. M. Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series // *Nature*. 1990. Vol. 344. N 6268. P. 734–741.
- Tang L., Pan H., Yao Y. K-Nearest Neighbor Regression with Principal Component Analysis for Financial Time Series Prediction. Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, Chengdu China, ACM, 2018. P. 127–131.
- Wang Y., Xie Y., Wu Y., Yang Y. Improved KNN-based Stock Price Prediction // *AJCIS*. 2024. Vol. 7. N 6.
- Zhang M., Parnell A. Review of Clustering Methods for Functional Data // *ACM Transactions on Knowledge Discovery from Data*. 2023. Vol. 17. N 7. P. 1–34.
- Zhou H., Wei D., Yao F. Theory of Functional Principal Component Analysis for Discretely Observed Data // *arXiv*. 2022.

## References

- Agnon Y., Golan A., Shearer M. Nonparametric, Nonlinear, Short-Term Forecasting: Theory and Evidence for Nonlinearities in the Commodity Markets. *Economics Letters*, 1999, vol. 65, N 3, pp. 293–299.
- Álvarez-Díaz M. Is it Possible to Accurately Forecast the Evolution of Brent Crude Oil Prices? An Answer Based on Parametric and Nonparametric Forecasting Methods. *Empirical Economics*, 2020, vol. 59, N 3, pp. 1285–1305.
- Anderson A., Maystre L., Anderson I., Mehrotra R., Lalmas M. *Algorithmic Effects on the Diversity of Consumption on Spotify*. Proceedings of The Web Conference 2020, Taipei Taiwan, ACM, 2020, pp. 55–65.
- Barnett W. Econometrics of Chaos. *Social Science Computer Review*, 1990, vol. 8, N 4, pp. 528–531.
- Datta H., Knox G., Bronnenberg B. J. Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery. *Marketing Science*, 2018, vol. 37, N 1, pp. 5–21.
- Decoster G. P., Mitchell D. W. The Efficacy of the Correlation Dimension Technique in Detecting Determinism in Small Samples. *Journal of Statistical Computation and Simulation*, 1991, vol. 39, N 4, pp. 221–229.
- Dewan S., Ramaprasad J. Research Note — Music Blogging, Online Sampling, and the Long Tail. *Information Systems Research*, 2012, vol. 23, N 3-part-2, pp. 1056–1067.
- Jaditz T., Sayers C. L. Is Chaos Generic in Economic Data? *International Journal of Bifurcation and Chaos*, 1993, vol. 03, N 03, pp. 745–755.

- Kaimann D., Tanneberg I., Cox J. "I Will Survive": Online Streaming and the Chart Survival of Music Tracks. *Manage Decis Econ*, 2021, vol. 42, N 1, pp. 3–20.
- Kärnä T., Lendasse A. Comparison of FDA Based Time Series Prediction Methods. *European Symposium on Time Series Prediction*, 2007, pp. 77–86.
- Kidron A., Klein S. T. An Information Retrieval Approach to Predicting Meteorological Data. *International Journal of Modelling and Simulation*, 2007, vol. 27, N 3, pp. 218–225.
- Luz López García M., García-Ródenas R., González Gómez A. K-means algorithms for functional data. *Neurocomputing*, 2015, vol. 151, pp. 231–245.
- Manolovitz B., Ogihara M. *Practical evaluation of repeated recommendations in personalized music discovery*. Zenodo, 2020.
- Mok L., Way S. F., Maystre L., Anderson A. The Dynamics of Exploration on Spotify. *ICWSM*, 2022, vol. 16, pp. 663–674.
- Ramsay J. O., Silverman B. W. *Functional Data Analysis*. New York, NY, Springer New York, 2005.
- Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 1987, vol. 20, pp. 53–65.
- Schneider L., Gros C. Five decades of US, UK, German and Dutch music charts show that cultural processes are accelerating. *Royal Society Open Science*, 2019, vol. 6, N 8, pp. 19–94.
- Silverman B. W. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1985, vol. 47, N 1, pp. 1–21.
- Sugihara G., May R. M. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 1990, vol. 344, N 6268, pp. 734–741.
- Tang L., Pan H., Yao Y. *K-Nearest Neighbor Regression with Principal Component Analysis for Financial Time Series Prediction*. Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, Chengdu China, ACM, 2018, pp. 127–131.
- Wang Y., Xie Y., Wu Y., Yang Y. Improved KNN-based Stock Price Prediction. *AJCIS*, 2024, vol. 7, N 6.
- Zhang M., Parnell A. Review of Clustering Methods for Functional Data. *ACM Transactions on Knowledge Discovery from Data*, 2023, vol. 17, N 7, pp. 1–34.
- Zhou H., Wei D., Yao F. Theory of Functional Principal Component Analysis for Discretely Observed Data. *arXiv*, 2022.

*Статья поступила в редакцию 13 мая 2025 г.*

*Статья рекомендована в печать 20 июня 2025 г.*