Банк России

ДЕКАБРЬ 2020

СЕЗОННОЕ СГЛАЖИВАНИЕ ДАННЫХ ФИНАНСОВЫХ ПОТОКОВ ПЛАТЕЖНОЙ СИСТЕМЫ БАНКА РОССИИ

Серия докладов об экономических исследованиях, №65

С. Селезнев, Н. Турдыева
Р. Хабибуллин, А. Цветкова

Сергей Селезнев
Департамент исследований и прогнозирования, Банк России.
E-mail: SeleznevSM@mail.cbr.ru

Наталья Турдыева
Департамент исследований и прогнозирования, Банк России.
E-mail: TurdyevaNA@mail.cbr.ru

Рамис Хабибуллин
Департамент исследований и прогнозирования, Банк России.
E-mail: KhabibullinRA@mail.cbr.ru

Анна Цветкова
Департамент исследований и прогнозирования, Банк России.
E-mail: TsvetkovaAN@mail.cbr.ru

Серия докладов об экономических исследованиях Банка России проходит процедуру анонимного рецензирования членами Консультативного совета Банка России и внешними рецензентами.

Фото на обложке: Shutterstock/FOTODOM

Адрес: 107016, Москва, ул. Неглинная, 12
Телефон: +7 495 771-91-00, +7 495 621-64-65 (факс)
Официальный сайт Банка России: www.cbr.ru

Резюме

Данная работа описывает алгоритм сезонной корректировки, которая применяется в Банке России для очистки высокочастотных данных платежной системы Банка России. Необходимость работы с дневными данными и отсутствие операций в выходные и праздничные дни осложнило использование известных пакетов типа Prophet (Taylor and Letham (2017). Используя идеи Taylor and Letham (2017), специфику данных финансовых потоков и принимая во внимание ограничения на скорость работы, нами была разработана простая и быстрая процедура, основанная на наборе тригонометрических функций и фиктивных переменных, которая показывает хорошие результаты с точки зрения различных метрик качества и может быть легко модифицирована для работы с более гибкими спецификациями модели.

Разработанная нами процедура сглаживания высокочастотных данных может быть использована для решения большого класса прикладных задач, в том числе, как это было сделано в Банке России, для разработки индикаторов, отражающих изменение экономической активности в режиме реального времени, что особенно важно для принятия информированных решений в области экономической политики в условиях кризиса.

JEL-классификация: C11, C22, E32, E37.

Ключевые слова: дневная сезонная корректировка, временные ряды, отраслевые финансовые потоки, байесовская оценка.

Содержание

1. Введение

В связи с ухудшением экономической ситуации, вызванным распространением коронавирусной инфекции, экономисты и статистики во всем мире столкнулись с необходимостью мониторинга состояния экономики с недельной и даже дневной частотой. Происходившее беспрецедентными темпами снижение экономической активности требовало таких же активных и информированных ответов от правительств всех стран.

Нетрадиционные высокочастотные данные привлекали внимание экономистов, работающих в области экономической политики, уже несколько лет подряд (см., например, Hueng ed. (2020). Основываясь на этих наработках, в ситуации экономического кризиса, вызванного пандемией, было проведено множество исследований, направленных на использование больших объемов высокочастотных данных для анализа складывающегося экономического положения (см., например, Carvalho et al. (2020), Chetty et al. (2020), Lewis et al. (2020).

В этот период Банком России был разработан набор индикаторов, построенных на данных платежной системы Банка России (ПС БР), который позволил практически в режиме реального времени наблюдать динамику отраслевых финансовых потоков. Аналитика этих показателей поставила ряд вызовов, одним из которых стало удаление сезонных эффектов, так как они сильно осложняют понимание причин изменения динамики индикаторов (см. Рисунок 1 в Приложении А). Описанию методологии выделения сезонной компоненты из данных финансовых потоков и посвящена эта работа.

Задача выделения сезонной компоненты, по сути, является частным примером задачи декомпозиции временного ряда на различные составляющие и широко изучена в литературе. Однако нужно понимать, что в простейшей постановке эта задача является не до конца определенной. При всей кажущейся очевидности разбиение временного ряда будет зависеть от того, что исследователь вкладывает в понятия «тренд», «цикл» и «сезонность», и способа их моделирования. К сожалению, в отрыве от контекста не существует объективного критерия для оценки качества разбиения (то, что одному человеку кажется правильной декомпозицией, другой может не считать таковой), и методология должна ориентироваться на конкретную проблему, в процессе решения которой необходимо выделение одной из компонент, и, именно руководствуясь этим, мы выбирали способ сезонной корректировки. В нашем случае целью декомпозиции является очистка данных финансовых потоков от набора паттернов, которые мы хотим исключить ввиду их неинформативности для последующего анализа. По сути, этот набор паттернов и есть определение сезонности, применяемое в данной работе, и качество его выделения служит конечным ориентиром для выбора процедуры сезонной корректировки.

На практике мы столкнулись с рядом дополнительных ограничений, которые также оказали влияние на применяемую в настоящий момент процедуру сглаживания данных финансовых потоков. В их числе тип выделяемых паттернов, особенности данных финансовых потоков, а также скорость работы алгоритмов и возможность быстрого внесения изменений в набор исключаемых паттернов. Необходимость работы с дневными данными, а также несоответствие нашему определению сезонности

исключили возможность работы с широко применяемыми в экономическом анализе алгоритмами, а отсутствие операций в выходные и праздничные дни осложнило использование известных пакетов типа Prophet (Taylor and Letham (2017). Используя идеи Taylor and Letham (2017), специфику данных финансовых потоков и принимая во внимание ограничения на скорость работы и возможность быстрого внесения изменений в набор исключаемых паттернов, мы разработали базовую процедуру для сезонной корректировки, которая в текущий момент применяется для очистки данных в публикуемом еженедельно на сайте Банка России отчете[1].

Базовая процедура, потенциальные расширения и обсуждение методологии представлены в разделе 2. Раздел 3 обсуждает критерии, которые помогают оценить качество сезонной корректировки. В разделе 4 дается описание данных отраслевых финансовых потоков. Раздел 5 описывает результаты. В разделе 6 обсуждаются связанные с сезонной корректировкой работы. Заключение представлено в разделе 7.

2. Сезонная корректировка

2.1. Базовая процедура

В рамках базовой процедуры выделяется набор мультипликативных сезонных паттернов (s_t), который подобно Facebook Prophet состоит из внутринедельной (s_t^w) и внутригодовой (s_t^y) компонент. Для того чтобы явно учесть сезонность внутри месяца, присутствующую в данных (см. Рисунок 1), в дополнение к ним также добавляется внутримесячная компонента (s_t^m). Таким образом, сезонная компонента моделируется как[2]

$$s_t = s_t^w + s_t^m + s_t^y. \quad (1)$$

Мы предполагаем, что оставшаяся часть является суммой нестационарной (tr_t) и стационарной компонент (e_t), которые не содержат в себе сезонности[3] и которые далее будем называть трендом и остатками. С учетом наличия дней с отсутствием платежей (выходные и праздники) наша модель записывается в следующем виде:

$$Y_t = \begin{cases} e^{s_t^w + s_t^m + s_t^y + tr_t + e_t}, \text{если платежи есть} \\ 0, \text{иначе} \end{cases}, \quad (2)$$

[1] «Мониторинг отраслевых финансовых потоков» – еженедельный аналитический материал Банка России, в котором представлена обобщенная информация по платежам, прошедшим через платежную систему Банка России.

[2] Мы не учитываем эффект праздников, так как он не сильно влияет на оценку внутримесячной, дневной и внутригодовой компонент, но он легко может быть включен в модель при необходимости в рамках ее расширения.

[3] Формально мы предполагаем, что

$$\frac{\sum_{t=1}^{T} S_t}{T} \to 0$$

$$\frac{\sum_{t=1}^{T} S_t E_t}{T} \to 0 \; a.s$$

$$\exists K: \forall k \geq K \; \frac{\sum_{t=1}^{T} \Delta^k S_t \Delta^k T_t}{T} \to 0 \; a.s.$$

где Y_t – данные после предобработки. В дальнейшем для простоты записи мы вводим обозначение x_t для логарифма переменной X_t и без потери общности концентрируемся только на ненулевой части.

С точки зрения асимптотической теории выделение нестационарной компоненты в моделях типа (2) должно играть ключевую роль в том смысле, что неверная спецификация тренда практически всегда приводит к несостоятельной оценке, в то время как ошибки в спецификации остатков часто не создают проблем в асимптотике. Учитывая этот факт, стационарная компонента моделируется как нормальное распределение с нулевым средним и оцениваемой дисперсией (σ_e^2):

$$e_t \sim N(0, \sigma_e^2). \quad (3)$$

Предварительный визуальный анализ не выявил каких-либо строгих нелинейностей в данных в период до пандемии, и так как конечной целью построения модели не является прогнозирование в последующий период, 2020 год исключался из оценки, а трендовая компонента моделировалась в виде линейной зависимости от времени:

$$tr_t = \theta^{tr} t\ , \quad (4)$$

где θ^{tr} – оцениваемый параметр.

Внутринедельная компонента оценивается с помощью фиктивных переменных на каждый из дней недели, что позволяет убрать любой вид периодичности на дневной основе:

$$s_t^w = \sum_{k=1}^{5} \theta_k^w I_t^k\ , \quad (5)$$

где I_t^k – индикатор, принимающий значение 1 для рабочего для под номером k, θ_k^w – оцениваемые параметры.

Подобно многим работам по сезонной корректировке (см., например, Taylor and Letham (2017), периодически повторяемые паттерны, где количество дней в периоде велико, убираются не фиктивными переменными, а набором тригонометрических функций, которые при достаточном их количестве могут аппроксимировать любую периодическую функцию:

$$s_t^r = \sum_{j=1}^{N_r} \left[\theta_j^{sin.r} \sin\left(\frac{2\pi j}{P_r} t\right) + \theta_j^{cos.r} \cos\left(\frac{2\pi j}{P_r} t\right)\right], \quad (6)$$

где $r \in \{m, y\}$ – индикатор компоненты (месячной, m, или годовой, y), P_r – максимальная длина цикла для компоненты r, рассчитываемая как число дней в данном месяце для $r = m$ и в данном году для $r = y$, $\theta_j^{sin.r}$ и $\theta_j^{cos.r}$ – оцениваемые параметры, N_r – максимальное число используемых циклов для компоненты r.

Одной из проблем моделирования сезонности в рамках параметрического подхода является выбор релевантных циклических компонент. С одной стороны, их недовключение приводит к просачиванию циклических колебаний в очищенный ряд, с другой стороны, слишком большое число может привести к переобучению. Для решения этой проблемы в модель включается большое число компонент ($N_m = 10, N_y = 20$), а для предотвращения переобучения на этапе оценки используется байесовская регрессия с автоматическим выбором гиперпараметров (см., Tipping (2001), которая

откидывает нерелевантные компоненты путем максимизации предельного правдоподобия. Несмотря на то что для байесовской регрессии предельное правдоподобие может быть посчитано напрямую, мы следуем процедуре вариационной байесовской оценки с MF приближением, которая описана в Khabibullin and Seleznev (2020), так как она практически без модификаций позволяет в дальнейшем перейти к расчету моделей с нелинейностями и недетерминистическими трендами.

2.2. Расширения базовой процедуры

В некоторых ситуациях базовая модель может быть недостаточно гибкой для адекватного выделения сезонной компоненты, как, например, в случае, если 2020 год включен в выборку, либо исключать недостаточное количество паттернов для последующей аналитики. В этих случаях она может быть легко модифицирована добавлением дополнительных паттернов. Модель, которая описана в предыдущем подразделе, была выбрана в рамках конкретной задачи по очистке финансовых потоков от конкретного набора паттернов, однако методология (вариационная байесовская оценка с использованием MF приближения), используемая нами для оценки, достаточно гибка и может быть легко применена без особых изменений для любой модели, где функция правдоподобия может быть выписана при условии ненаблюдаемых компонент и параметров модели.

2.3. Обсуждение свойств

Как описывалось во введении, выбор процедуры сезонного сглаживания был обусловлен рядом факторов, в числе которых качество удаления паттернов, обсуждаемое в следующем разделе, а также некоторые особенности данных и необходимость решения нескольких технических сложностей. В этом подразделе мы чуть подробнее обсудим то, как предложенные алгоритмы справляются с этими трудностями.

Несмотря на то что многие алгоритмы (например, Facebook Prophet) могут в оригинальной форме либо с небольшими изменениями, как и базовая процедура, справляться с решением задачи удаления внутринедельных, внутримесячных и внутригодовых периодических колебаний, насколько нам известно, их готовые реализации не предусматривают возможности работы с пропущенными данными. Базовая процедура, которая, по сути, является байесовской регрессией, легко работает с пропущенными данными, поскольку позволяет просто не включать их в выборку. В случае если модификации содержат недетерминистические компоненты, процедура записывается в виде модели пространства состояний, где отсутствуют наблюдения в указанные периоды времени.

Как было сказано выше, вариационная байесовская оценка с использованием MF приближения может быть легко применена к широкому кругу моделей, а также требует буквально несколько строк дополнительного кода при внесении изменений в модель[4], что с легкостью позволяет модифицировать базовую процедуру. Изменение же

[4] Для реализации мы используем библиотеку Tensorflow (Abadi et al. (2016).

других алгоритмов требует большого количества времени и усилий либо на разработку новых алгоритмов оценки, либо на внесение правок в программный код уже готовых библиотек, что зачастую достаточно трудоемко.

Наконец, для четырех лет данных (более 1000 точек) время работы[5] базовой процедуры не превосходит 2 минут, а для расширений, использующих нелинейные модели пространства состояний, не превосходит 10 минут, что при необходимости позволяет на ежедневной основе переоценивать модели для всех отраслей[6] и за разумное время тестировать новые спецификации.

3. Критерии качества

Выделение сезонной компоненты является одним из примеров задачи, в которой правильный ответ неизвестен и которая не содержит обучающих примеров. Для таких задач невозможно классическое измерение качества их решения путем расчета функции потерь на отложенной (тестовой) выборке, поэтому необходим набор косвенных критериев качества, которые позволяют оценить, насколько хорошо выполнена задача. В рамках задачи выделения сезонности из данных финансовых потоков мы основываемся на качестве прогноза, на метриках, позволяющих оценить наличие/отсутствие переобучения и недообучения, а также на визуальной оценке.

Качество прогноза. Несмотря на то что прогнозные свойства не всегда хорошо коррелируют с процедурой очистки данных от сезонной компоненты, сильное различие между моделью сезонной корректировки и альтернативами может быть знаком того, что модель некорректно специфицирована и оценки далеки от желаемого результата.

Отсутствие переобучения. Переобучение приводит к ситуациям, когда модель, ввиду того что обладает излишней вариативностью, помимо тех паттернов, которые она должна была удалить, также удаляет набор паттернов, которые не существуют. Чтобы протестировать такое поведение, мы проводим стандартную диагностику «тест-обучение» (см. Ng (2019).

Отсутствие недообучения. К обратной ситуации приводит недообучение. Такое происходит, когда модель недостаточно гибка в своей спецификации либо излишне регуляризирована. Чтобы детектировать это, строится ряд дополнительных регрессий остатков модели на циклические компоненты, описывающие колебания различной частоты.

Визуальная оценка. Человеческие суждения относительно того, насколько хорошо модель решает поставленную перед ней проблему, являются важным крите-

[5] Расчеты произведены на CPU на компьютере со следующими характеристиками: Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz 2.21GHz, RAM 16 GB.

[6] Для практических целей отчета коэффициенты сезонной компоненты переоцениваются не чаще раза в неделю.

рием во многих задачах, где прямое измерение качества невозможно либо стандартные метрики не полностью могут отловить все нюансы[7]. В текущей работе мы проводим визуальное сравнение выделенной сезонности и исходного ряда, а также анализ очищенного ряда.

4. Данные отраслевых финансовых потоков

Данные представляют собой входящие[8] дневные платежи фирм, которые прошли через ПС БР, агрегированные в отрасли согласно основному коду вида деятельности ОКВЭД 2[9]. Дополнительно к отраслевым платежам в данных также выделены платежи, поступающие физическим лицам, и прочие платежи, которые в дальнейшем обозначены кодами 0 и 100 соответственно. Ряды доступны в период с 1 января 2016 года по настоящее время[10] и имеют особенности, которые, как упомянуто во введении, повлияли на выбор методологии сезонной корректировки. Главная из них – это наличие дней, в которые платежи отсутствуют. Такое происходит в праздничные и выходные дни, но из-за неточностей заполнения платежных поручений существует несколько платежей, которые все-таки попадают на выходные. Эти платежи были исключены нами из рассмотрения для обеспечения робастности процедуры очистки данных.

5. Результаты

Результаты прогнозов базовой процедуры были сопоставлены с несколькими широко используемыми для сезонной корректировки моделями: Facebook Prophet (Taylor and Letham (2017), Seasonal and Trend decomposition using Loess (STL, Cleveland et al. (1990) и Trigonometric Box-Cox transformation ARMA Trend Seasonality (TBATS, Livera et al. (2011)[11]. Ни один из известных нам программных пакетов, реализующих эти модели, не способен работать с нулевыми/пропущенными данными, поэтому мы дозаполняем их логарифмическими скользящими средними за 5 ненулевых дней. В Таблице 1 представлены средние по отраслям RMSFE (относительно модели линейного тренда) для 2019 года на разные горизонты (процедура прогнозирования полностью описана в Приложении Б). SBL-модель работает лучше или сравнимо с другими моделями[12], что является одним из признаков того, что сезонная компонента удалена адекватно. В Таблице 1 также показаны RMSFE для линейной регрессии (без какой-либо регуляризации) с теми же частотами, что и в SBL-модели (Regression), и с

[7] Такая практика часто встречается при построении генеративных моделей картинок (см. например, Arjovsky et al (2017) либо языковых моделей (см. например, Brown et al. (2020).

[8] В данной работе мы описываем методологию на основе входящих платежей. К исходящим платежам применяется та же корректировка.

[9] Из данных исключены платежи внутри одного ИНН.

[10] В этой работе выборка заканчивается 23 октября 2020 года.

[11] Дополнительно была оценена модель сезонной корректировки Бундесбанка (Ollech (2018), однако ее результаты оказались значительно хуже других альтернатив, поэтому мы не включили их в Таблицу 1.

[12] Возможно, это следствие того, что дозаполнение не совсем хорошо влияет на поведение этих процедур либо недостаточно потраченного времени на подбор гиперпараметров, однако дополнительные способы корректировки данных и процедура валидации позволяют надеяться, что это не так.

частотами, которые обусловлены экономическими соображениями (Cut Regression)[13]. Экономически обусловленные частоты демонстрируют чуть худшие результаты, что и базовая процедура, а регрессия – примерно те же, что косвенно свидетельствует об отсутствии переобучения.

О наличии переобучения, однако, сигнализирует отношение среднеквадратичных ошибок на тестовой и обучающей выборках. Рисунок 2 демонстрирует, что для многих серий это отношение значительно больше единицы. Причиной этого может быть как переобучение, так и отличие тестовой и обучающей выборок. Чтобы понять, что именно является причиной, мы дополнительно сравнили отношение ошибок за те же периоды на модели, которая оценена на данных с 2016 по 2019 год. Рисунок 3 показывает небольшую разницу между соотношениями детрендированных ошибок для двух различных периодов обучения и позволяет сделать вывод об отличии тестовой и обучающей выборок, а не переобучении. Мы также не нашли каких-либо признаков недообучения, оценив дополнительные регрессии остатков SBL-модели на тот же набор тригонометрических функций, что и в исходной SBL-модели.

На Рисунок 4 – Рисунок 9 продемонстрированы сезонные компоненты данных входящих финансовых потоков для шести различных отраслей, которые отражают типичные результаты нашей базовой процедуры. Рисунок 4 – Рисунок 6 демонстрируют серии с доминирующими дневными, внутримесячными и внутригодовыми паттернами соответственно. Во всех трех случаях SBL-модель улавливает необходимые частоты и является достаточной с точки зрения визуального анализа и описанных выше тестов для удаления сезонной компоненты.

Несмотря на то что для большинства серий свойства выделенной сезонной компоненты похожи на свойства рядов на Рисунок 4 – Рисунок 6, мы также заметили несколько ситуаций, где базовая процедура демонстрирует неудовлетворительное для последующей аналитики поведение и требует дополнительной корректировки[14]. Такие ситуации показаны на Рисунок 7 – Рисунок 9 и в большинстве своем могут быть учтены путем различных модификаций. Первая из них связана с изменившимся поведением ряда и, как следствие, изменившейся сезонностью. Этот ряд не исследуется дополнительно ввиду явного структурного сдвига в поведении сезонности, который никак не подпадает под определение периодически повторяющегося события и требует отдельного изучения.

Вторая ситуация связана с плавающими во времени событиями. Она наглядно изображена на Рисунок 8, где пики, сконцентрированные около 25 числа каждого месяца, не полностью объясняются сезонной компонентой. Так происходит из-за того, что день, на который приходится пик в разные месяцы, расположен немного в отличающейся фазе цикла[15], а это приводит к усреднению в сезонной компоненте между пиковыми и обычными днями. В месяцах, где 25 и 26 числа является рабочими днями, мы сравнили остатки и обнаружили, что из 29 случаев только в одном в оба дня

[13] Квартальные колебания, месячные колебания и дневная компонента.

[14] В настоящий момент эти корректировки не применяются в публикуемых отчетах, но недостатки принимаются во внимание при проведении аналитики.

[15] Как из-за того, что пик расположен в разные дни месяца, так и из-за то, что месяцы имеют разное количество дней.

ошибка положительная, в остальных же она разная по знаку (25 – положительная, 26 – отрицательная), при этом средние значения ошибок близки по модулю: 0,68 и -0,7. Несмотря на то что такие закономерности не являются строго периодическими, их исключение может быть востребовано при последующей аналитике. В каждой конкретной ситуации модификация базовой процедуры должна быть разной и зависеть от задачи[16]. Это может быть добавление набора фиктивных переменных, мультиномиальное распределение или, в случае когда суммарный платеж разбит по нескольким дням, распределение Дирихле. Для ряда, изображенного на Рисунок 8, мы добавляем набор переменных, которые равны единице для 25 числа или первого последующего рабочего дня, если 25 является выходным. Рисунок 10 показывает, что удаленная таким образом сезонность удачно справляется с такого типа паттернами.

Третья ситуация возникает, когда амплитуда ряда меняется во времени как на Рисунок 9. Для дополнительного удаления таких закономерностей мы включаем в регрессию (2) дополнительные признаки в виде произведения периодических признаков и времени[17]. Результаты после добавления новых признаков показаны на Рисунок 11.

Как было описано в разделе 2, мы используем только данные с 2016 по 2019 год. Это связано с наличием аномального периода в данных после введения ограничений, вызванных распространением коронавирусной инфекции. Чтобы учесть такое поведение и не ограничиваться докоронавирусным периодом, базовая процедура может быть модифицирована путем добавления гибкой спецификации, например модели локального линейного тренда или стохастического тренда со стохастической волатильностью в инновациях тренда. Для иллюстрации возможности применения модели с сильной нелинейностью мы демонстрируем второй вариант. На Рисунок 12 представлены данные, включающие дополнительный период до 23 октября 2020 года, а также очищенные от сезонности ряды с использованием базовой процедуры и расширенной процедуры с гибким трендом. Можно увидеть, что первая из них не способна учесть адекватно резкое уменьшение входящих потоков и выделяет ложные пики в апреле. Это происходит из-за того, что для компенсации низких значений 2020 года модель чуть завышает значения в предыдущие годы. В модели же с гибким трендом такого не происходит, так как апрельское снижение приходится на тренд, что приводит к визуально более приемлемым результатам.

Стоит дополнительно отметить, что добавление стохастических трендов увеличивает время работы алгоритма примерно в 5 раз (с 2 до 10 минут), однако ряд предварительных расчетов для модели локального линейного тренда показал, что, несмотря на чуть лучшие прогнозные результаты, это не приводит к значительным изменениям в сезонной компоненте на данных до 2020 года.

[16] Автоматическую процедуру для добавления таких паттернов мы оставляем для будущих исследований. Сейчас это делается вручную для отдельных серий. Мы также заметили, что регрессии модулей 1–5-дневных скользящих остатков на набор синусов и косинусов полезны при детектировании необходимости этих паттернов.

[17] В качестве альтернативы могут быть использованы различные трансформации данных, однако, попробовав трансформацию Бокса-Кокса, на предварительном этапе мы выяснили, что стохастические алгоритмы оптимизации таких моделей часто сходятся к визуально худшим результатам, чем простое добавление новых признаков.

6. Релевантные исследования

Оценке сезонной компоненты посвящено большое количество работ, однако многие из них, такие как X-12, X-13 TRAMO-SEATS и JDemetra+, не применимы к нашем задаче, так как разработаны для очистки рядов месячной и квартальной частот.

Работы, которые описывают алгоритмы сезонной корректировки для дневных данных, можно разделить на две категории: с переменной и постоянной во времени сезонностью. Несмотря на то что часть работ с изменяющейся во времени сезонностью, как и мы, используют тригонометрические функции при построении моделей (см., например, Livera et al. (2011), Ollech (2018), они, как и другие модели из этого класса, основанные на локальных регрессиях и низкочастотных фильтрах (см., например, Cleveland et al. (1990), Verbesselt et al. (2010) и Wen et al. (2020) или моделях пространства состояний (см., например, Koopman et al. (2009), Koopman and Ooms (2006), не подходят под наше определение сезонности. Ко второй группе относятся, например, алгоритмы Prophet (Taylor and Letham (2017) и STR[18] (Dokumentov and Hyndman (2015), а также метод, основанный на регрессии с дисперсией остатков в форме GARCH (Campbel, et al. (2005). Эти работы наиболее близки к нашей и во многом служат отправной точкой для построения процедуры сезонной корректировки данных финансовых потоков, однако ни одна из известных нам реализаций не способна принимать во внимание все особенности наших данных (включая пропущенные значения и гибкие тренды). Также в рамках процедур, включающих в себя RegARIMA модели (Ghysels, et al. (2001), Ollech (2018), сезонность может моделироваться с помощью сезонных лаговых полиномов. Однако это требует интенсивных вычислений при учете большого числа периодических паттернов.

Наша работа также связана с направлением исследований, занимающимся определением релевантных частот. Наиболее популярным является выбор спецификации, основанный на информационных критериях (см, например, BAYSEA (Akaike (1980), Akaike and Ishiguro (1983), а также процедуры Taylor and Letham (2017), Ollech (2018). Альтернативой является регуляризация с помощью Ridge или LASSO регрессий, как в Dokumentov and Hyndman (2015). Но эти подходы, в отличие от SBL-процедуры, используемой в этой работе, требуют переоценки модели для каждой спецификации, что может быть вычислительно сложно и не подходит в случаях, когда сама модель оценивается длительное время.

Наша базовая процедура использует линейный тренд и нормально распределенную иррегулярную компоненту, что близко к идеям, применяемым в библиотеке Prophet (Taylor and Letham (2017), которая использует кусочно-линейную спецификацию. Однако расширения базовой процедуры могут легко включать любую детерминистическую либо стохастическую модель как для тренда, так и для иррегулярной компоненты, что близко по духу к моделям пространства состояний (см., например, Koopman et al. (2009), Koopman and Ooms (2006).

[18] STR также может быть оценен и с изменяющейся сезонностью.

7. Заключение

В данной работе мы описали методологию сезонной корректировки дневных данных, которая используется в Банке России для предварительной очистки данных отраслевых финансовых потов. Простая базовая процедура, описанная в разделе 4, хорошо справляется с задачей удаления периодически повторяющихся событий, а при необходимости может быть легко модифицирована для добавления новых паттернов и использования более гибких моделей.

Литература

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. a Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. and X. Zheng (2016). Tensorflow: a System for Largescale Machine Learning, OSDI, 16, 265–283.
2. Akaike, H. (1980). Seasonal adjustment by a Bayesian modeling. Journal of time series analysis, 1(1), 1-13.
3. Akaike, H., & Ishiguro, M. (1983). Comparative study of the x-ll and baysea procedures of seasonal adjustment. Applied Time Series Analysis of Economic Data, 17-50.
4. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875.
5. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
6. Carvalho, Vasco M. and Garcia, Juan R. and Hansen, Stephen and Ortiz, Alvaro and Rodrigo, Tomasa and Rodriguez Mora, Jose V. and Ruiz, Pep, Tracking the COVID-19 Crisis with High-Resolution Transaction Data, NBER Conference «Economic Fluctuations and Growth Summer 2020» paper, July 2020, Available at: http://conference.nber.org/conf_papers/f143494.pdf
7. Campbell, S. D., & Diebold, F. X. (2005). Weather forecasting for weather derivatives. Journal of the American Statistical Association, 100(469), 6-16.
8. Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. Journal of official statistics, 6(1), 3-73.
9. Chetty, R., Friedman, J., Hendren, N., & Stepner, M. (2020). The economic impacts of COVID-19: Evidence from a new public database built from private sector data. Opportunity Insights.

10. De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American statistical association, 106(496), 1513-1527.

11. Dokumentov, A., & Hyndman, R. J. (2020). STR: A seasonal-trend decomposition procedure based on regression. arXiv preprint arXiv:2009.05894.

12. Ghysels, E., Osborn, D. R., & Sargent, T. J. (2001). The econometric analysis of seasonal time series. Cambridge University Press.

13. Hueng, C. James, ed. (2020). Alternative Economic Indicators. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

14. Khabibullin, R., & Seleznev, S. (2020) Stochastic Gradient Variational Bayes and Normalizing Flows for Estimating Macroeconomic Models. Central Bank of the Russian Federation Working Papers (Paper: wps61)

15. Koopman, S. J., & Ooms, M. (2006). Forecasting daily time series using periodic unobserved components time series models. Computational Statistics & Data Analysis, 51(2), 885-903.

16. Koopman, S. J., Ooms, M., & Hindrayanto, I. (2009). Periodic unobserved cycles in seasonal time series with an application to US unemployment. Oxford Bulletin of Economics and Statistics, 71(5), 683-713.

17. Lewis, Daniel J. and Mertens, Karel and Stock, James H. and Trivedi, Mihir (2020). Measuring Real Activity Using a Weekly Economic Index. FRB of New York Staff Report No. 920, Revised September 2020, Available at: https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr920.pdf

18. Ng, A. (2019). Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. Retrieved online at https://www.mlyearning.org

19. Ollech, D. (2018). Seasonal adjustment of daily time series (No. 41/2018). Deutsche Bundesbank.

20. Taylor, S. J., & Letham, B. (2017). Forecasting at scale. PeerJ Preprints 5: e3190v2

21. Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1(Jun), 211-244.

22. Verbesselt, J., Hyndman, R., Newnham, G., & Culvenor, D. (2010). Detecting trend and seasonal changes in satellite image time series. Remote sensing of Environment, 114(1), 106-115.

23. Wen, Q., Zhang, Z., Li, Y., & Sun, L. (2020, August). Fast RobustSTL: Efficient and Robust Seasonal-Trend Decomposition for Time Series with Complex Patterns. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2203-2213).nd X. Zheng (2016). Tensorflow: a System for Largescale Machine Learning, OSDI, 16, 265–283.

24. Akaike, H. (1980). Seasonal adjustment by a Bayesian modeling. Journal of time series analysis, 1(1), 1-13.

25. Akaike, H., & Ishiguro, M. (1983). Comparative study of the x-ll and baysea procedures of seasonal adjustment. Applied Time Series Analysis of Economic Data, 17-50.

26. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875.

27. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

28. Carvalho, Vasco M. and Garcia, Juan R. and Hansen, Stephen and Ortiz, Alvaro and Rodrigo, Tomasa and Rodriguez Mora, Jose V. and Ruiz, Pep, Tracking the COVID-19 Crisis with High-Resolution Transaction Data, NBER Conference «Economic Fluctuations and Growth Summer 2020» paper, July 2020, Available at: http://conference.nber.org/conf_papers/f143494.pdf

29. Campbell, S. D., & Diebold, F. X. (2005). Weather forecasting for weather derivatives. Journal of the American Statistical Association, 100(469), 6-16.

30. Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. Journal of official statistics, 6(1), 3-73.

31. Chetty, R., Friedman, J., Hendren, N., & Stepner, M. (2020). The economic impacts of COVID-19: Evidence from a new public database built from private sector data. Opportunity Insights.

32. De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American statistical association, 106(496), 1513-1527.

33. Dokumentov, A., & Hyndman, R. J. (2020). STR: A seasonal-trend decomposition procedure based on regression. arXiv preprint arXiv:2009.05894.

34. Ghysels, E., Osborn, D. R., & Sargent, T. J. (2001). The econometric analysis of seasonal time series. Cambridge University Press.

35. Hueng, C. James, ed. (2020). Alternative Economic Indicators. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

36. Khabibullin, R., & Seleznev, S. (2020) Stochastic Gradient Variational Bayes and Normalizing Flows for Estimating Macroeconomic Models. Central Bank of the Russian Federation Working Papers (Paper: wps61)

37. Koopman, S. J., & Ooms, M. (2006). Forecasting daily time series using periodic unobserved components time series models. Computational Statistics & Data Analysis, 51(2), 885-903.

38. Koopman, S. J., Ooms, M., & Hindrayanto, I. (2009). Periodic unobserved cycles in seasonal time series with an application to US unemployment. Oxford Bulletin of Economics and Statistics, 71(5), 683-713.

39. Lewis, Daniel J. and Mertens, Karel and Stock, James H. and Trivedi, Mihir (2020). Measuring Real Activity Using a Weekly Economic Index. FRB of New York Staff

Report No. 920, Revised September 2020, Available at: https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr920.pdf

40. Ng, A. (2019). Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. Retrieved online at https://www.mlyearning.org
41. Ollech, D. (2018). Seasonal adjustment of daily time series (No. 41/2018). Deutsche Bundesbank.
42. Taylor, S. J., & Letham, B. (2017). Forecasting at scale. PeerJ Preprints 5: e3190v2
43. Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1(Jun), 211-244.
44. Verbesselt, J., Hyndman, R., Newnham, G., & Culvenor, D. (2010). Detecting trend and seasonal changes in satellite image time series. Remote sensing of Environment, 114(1), 106-115.
45. Wen, Q., Zhang, Z., Li, Y., & Sun, L. (2020, August). Fast RobustSTL: Efficient and Robust Seasonal-Trend Decomposition for Time Series with Complex Patterns. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2203-2213).

Приложение А. Графики и таблицы

Рисунок 1. Обеспечение электрической энергией, газом и паром; кондиционирование воздуха, входящие потоки, нормировано на выборочное стандартное отклонение

Таблица 1. Средняя среди серий относительная RMSFE за 2019 год

	SBL	Regression	Cut Regression	Prophet no chngepoints	Prophet with changepoints	STL	TBATS
1 день	0,81	0,80	0,84	0,90	0,88	1,05	0,91
2 дня	0,81	0,81	0,85	0,90	0,89	1,05	0,97
3 дня	0,81	0,81	0,85	0,90	0,89	1,05	1,01
4 дня	0,81	0,81	0,85	0,91	0,89	1,05	1,05
5 дней	0,81	0,82	0,85	0,91	0,90	1,05	1,08
1 неделя	0,81	0,82	0,85	0,90	0,89	1,06	1,01
2 недели	0,81	0,81	0,85	0,91	0,90	1,05	1,13
1 месяц	0,81	0,81	0,85	0,90	0,89	1,05	1,30
1 квартал	0,82	0,82	0,86	0,90	0,92	1,03	2,13

Рисунок 2. Отношение среднеквадратичных ошибок (1 день) на тестовой и обучающей выборках по отраслям, входящие потоки

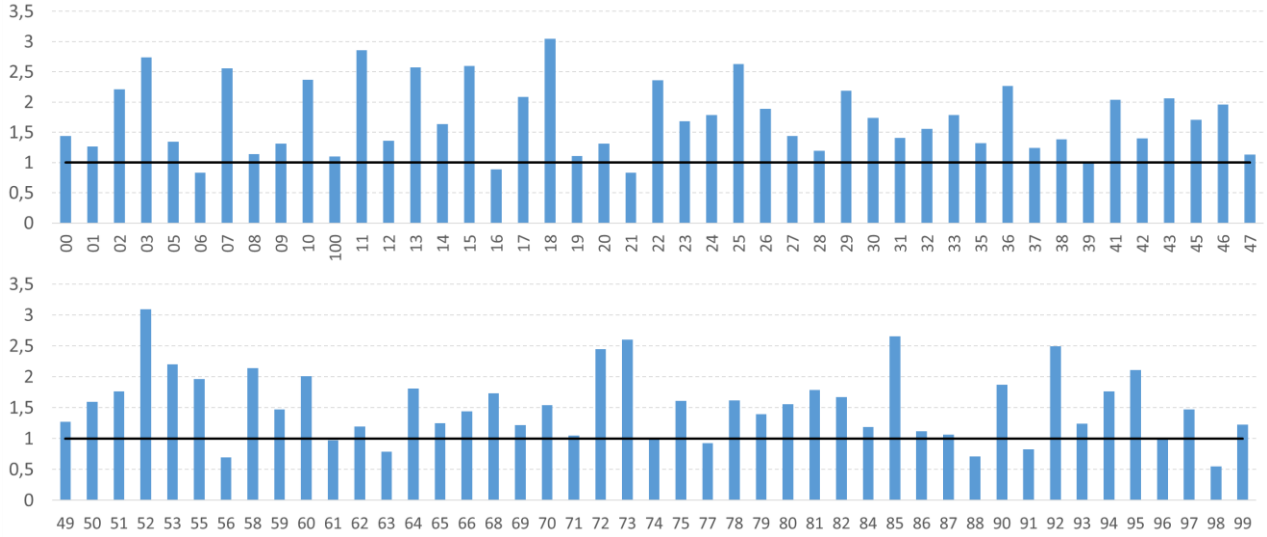


Рисунок 3. Отношение среднеквадратичных детрендированных ошибок для двух различных обучающих периодов, входящие потоки

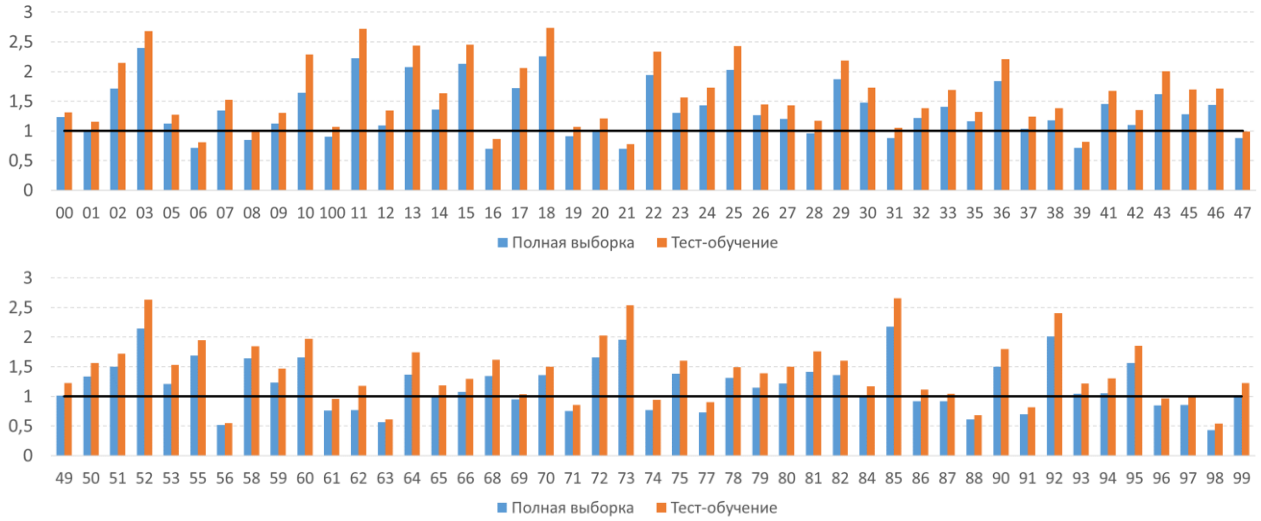


Рисунок 4. Сезонная компонента и остатки для деятельности по предоставлению продуктов питания и напитков, нормировано на выборочное стандартное отклонение

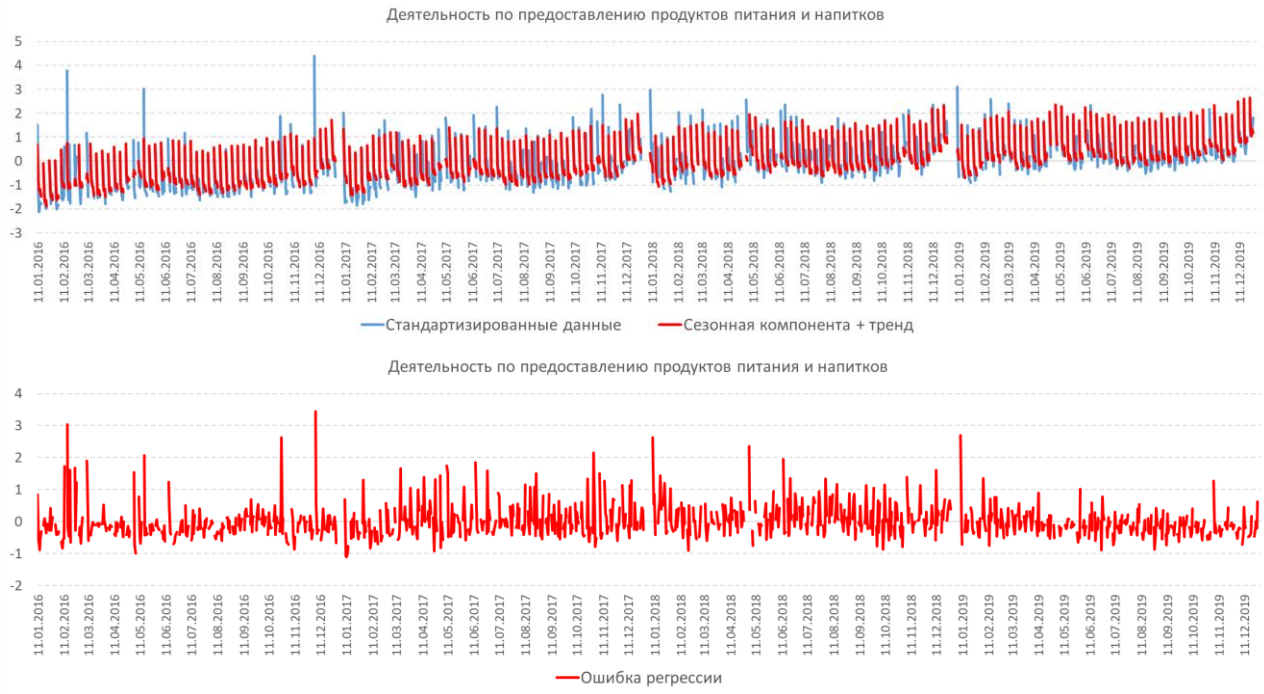


Рисунок 5. Сезонная компонента и остатки для торговли оптовой и розничной автотранспортными средствами и мотоциклами и их ремонта, нормировано на выборочное стандартное отклонение

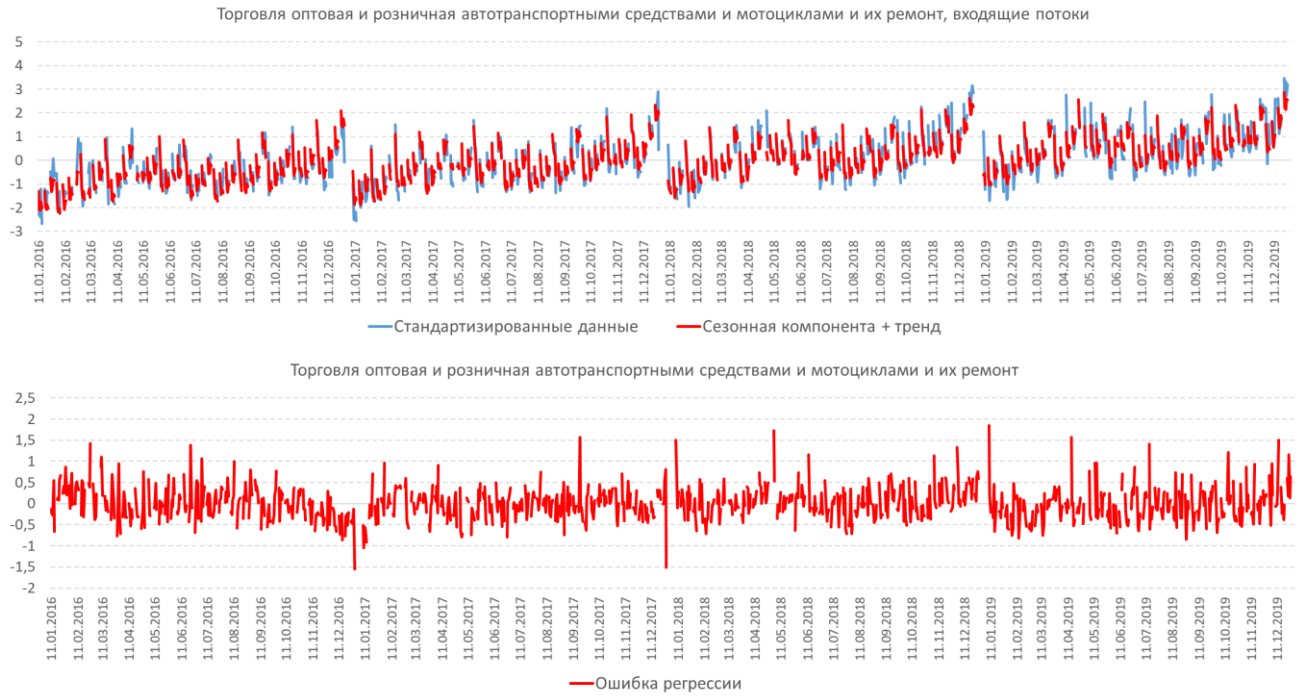


Рисунок 6. Сезонная компонента и остатки для деятельности туристических агентств и прочих организаций, предоставляющих услуги в сфере туризма, нормировано на выборочное стандартное отклонение

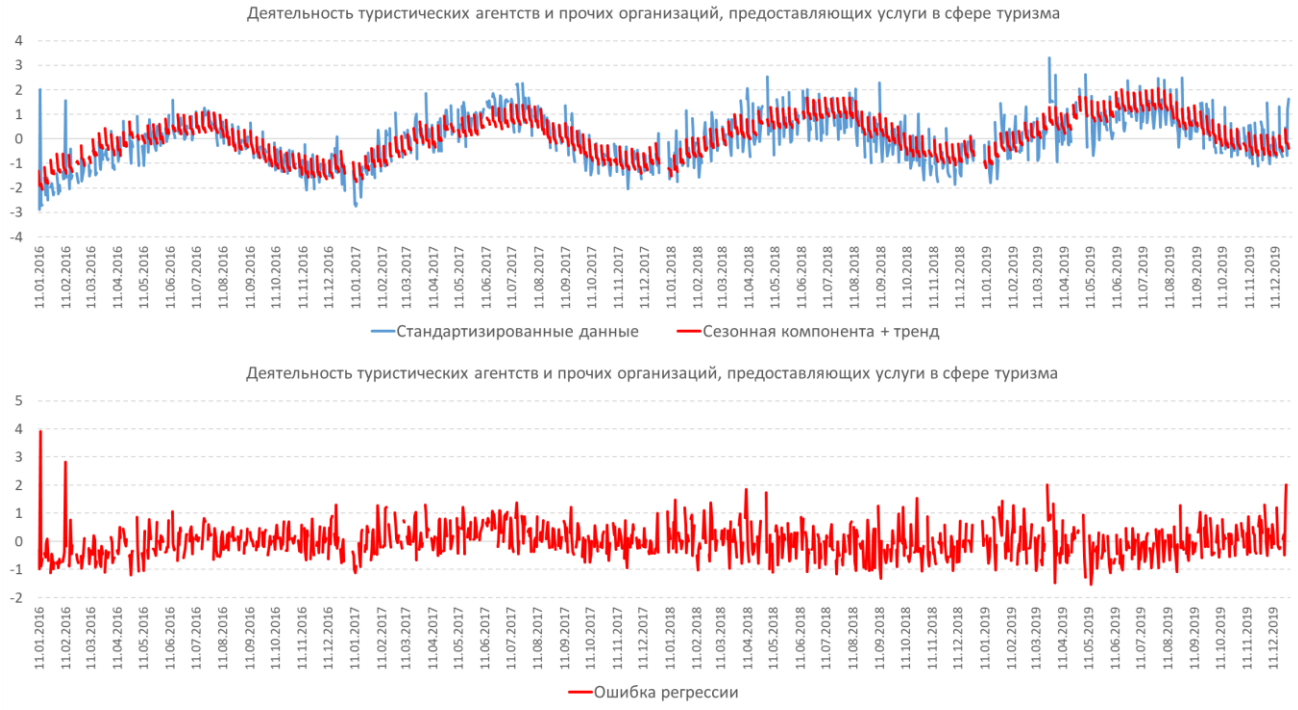


Рисунок 7. Сезонная компонента и остатки для деятельности по предоставлению финансовых услуг, кроме услуг по страхованию и пенсионному обеспечению, нормировано на выборочное стандартное отклонение

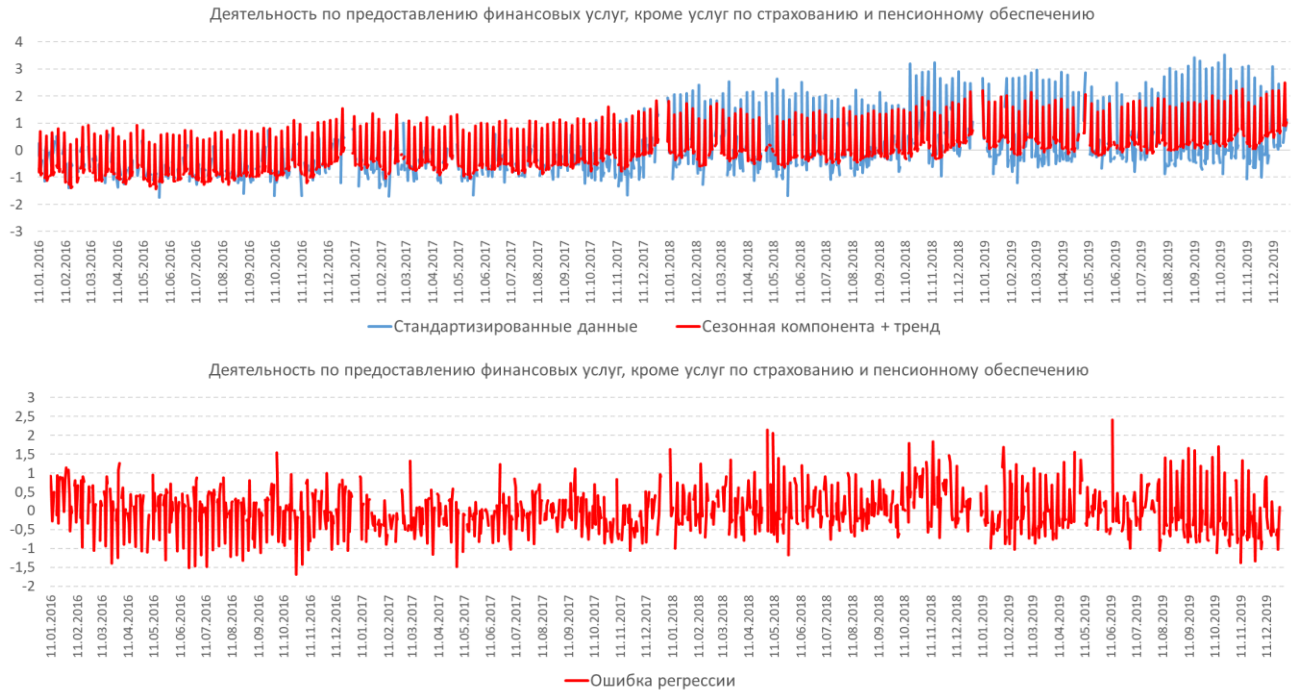


Рисунок 8. Сезонная компонента и остатки для деятельности органов государственного управления по обеспечению военной безопасности, обязательному социальному обеспечению, нормировано на выборочное стандартное отклонение

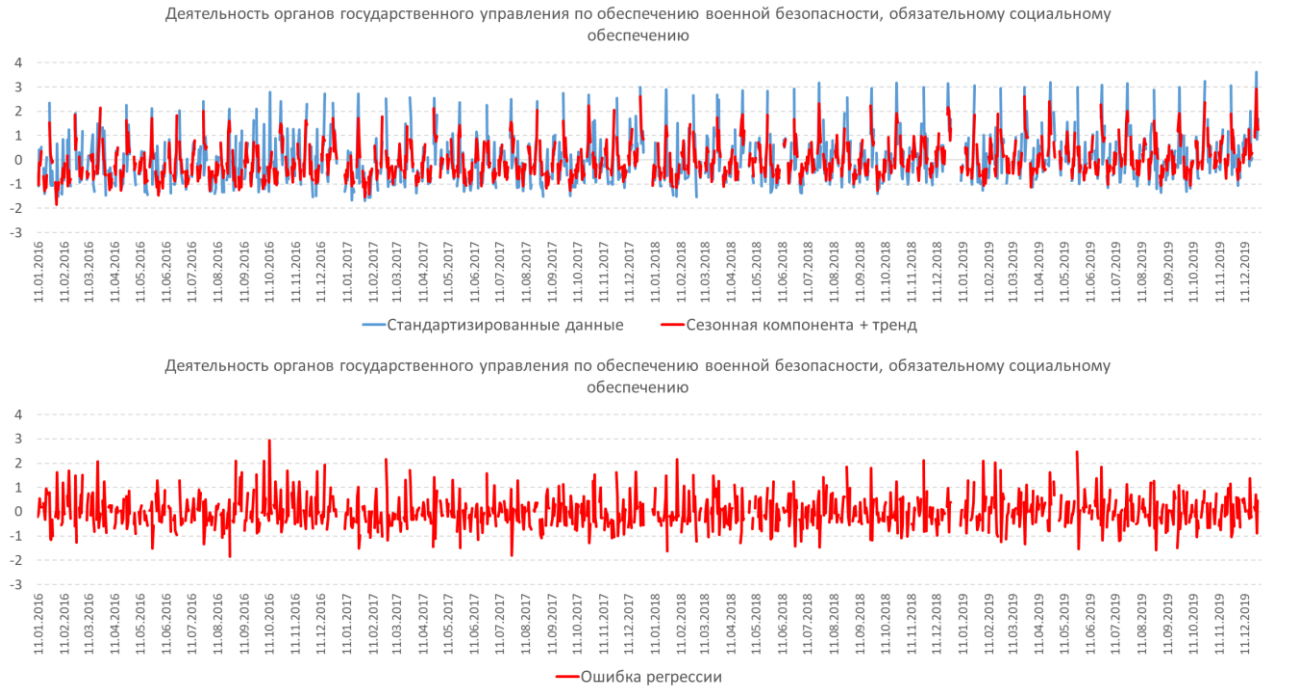


Рисунок 9. Сезонная компонента и остатки для работ строительных специализированных, нормировано на выборочное стандартное отклонение

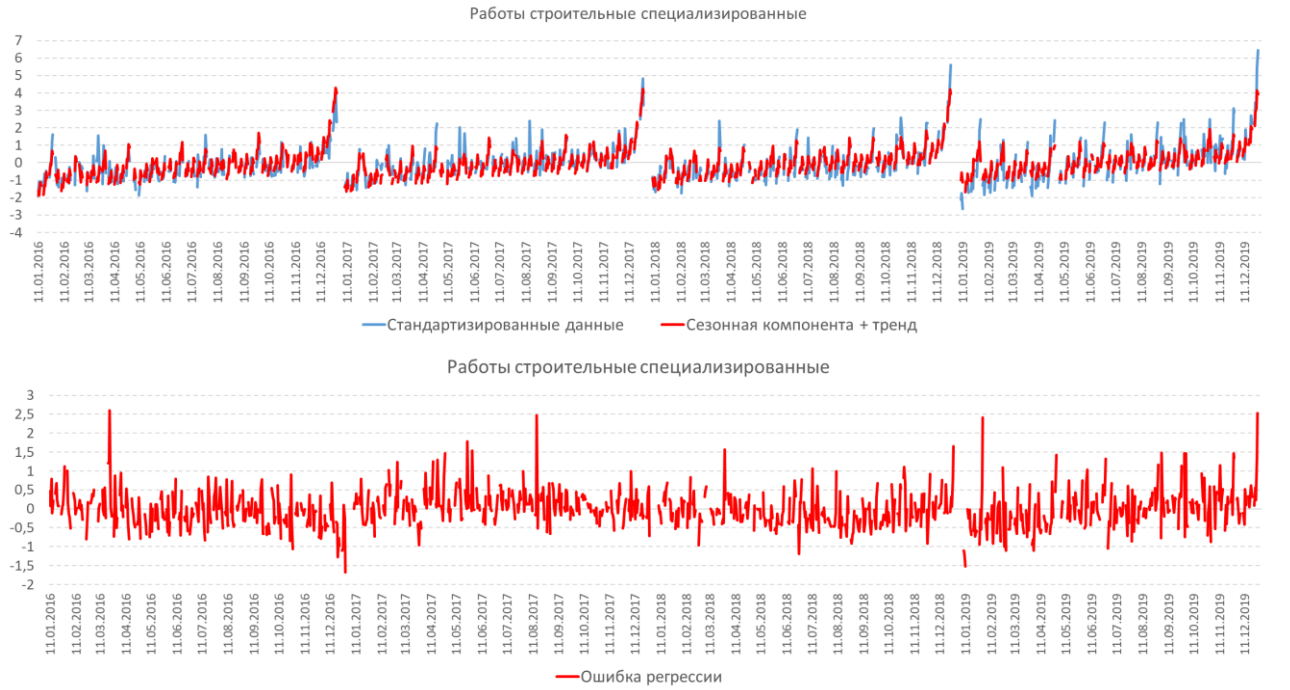


Рисунок 10. Скорректированная сезонная компонента для деятельности органов государственного управления по обеспечению военной безопасности, обязательному социальному обеспечению, нормировано на выборочное стандартное отклонение

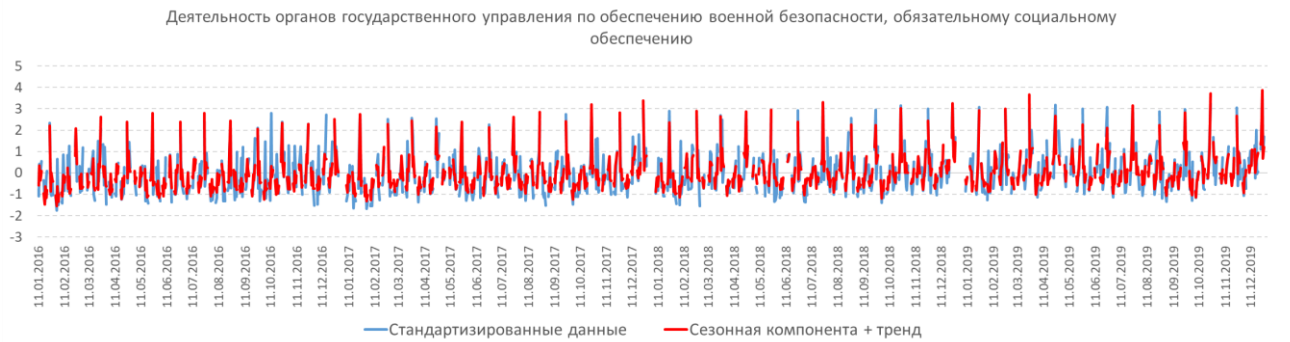


Рисунок 11. Скорректированная сезонная компонента для работ строительных специализированных, нормировано на выборочное стандартное отклонение

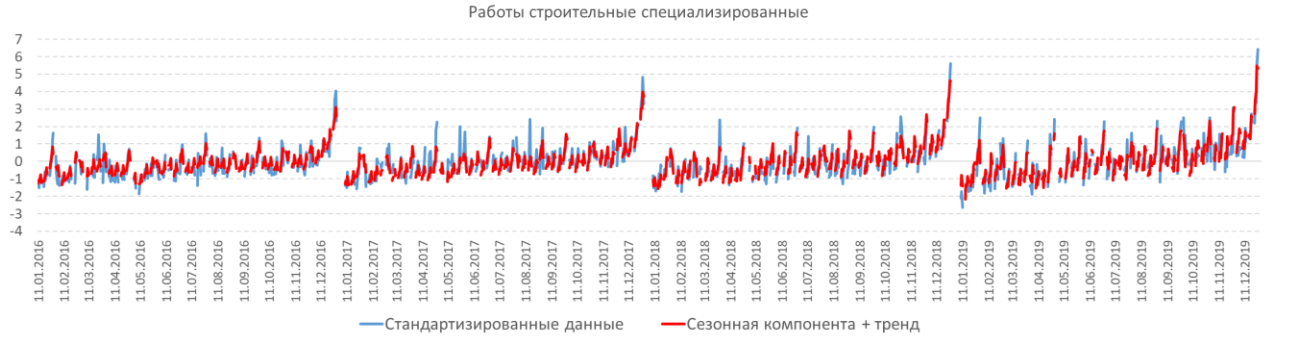
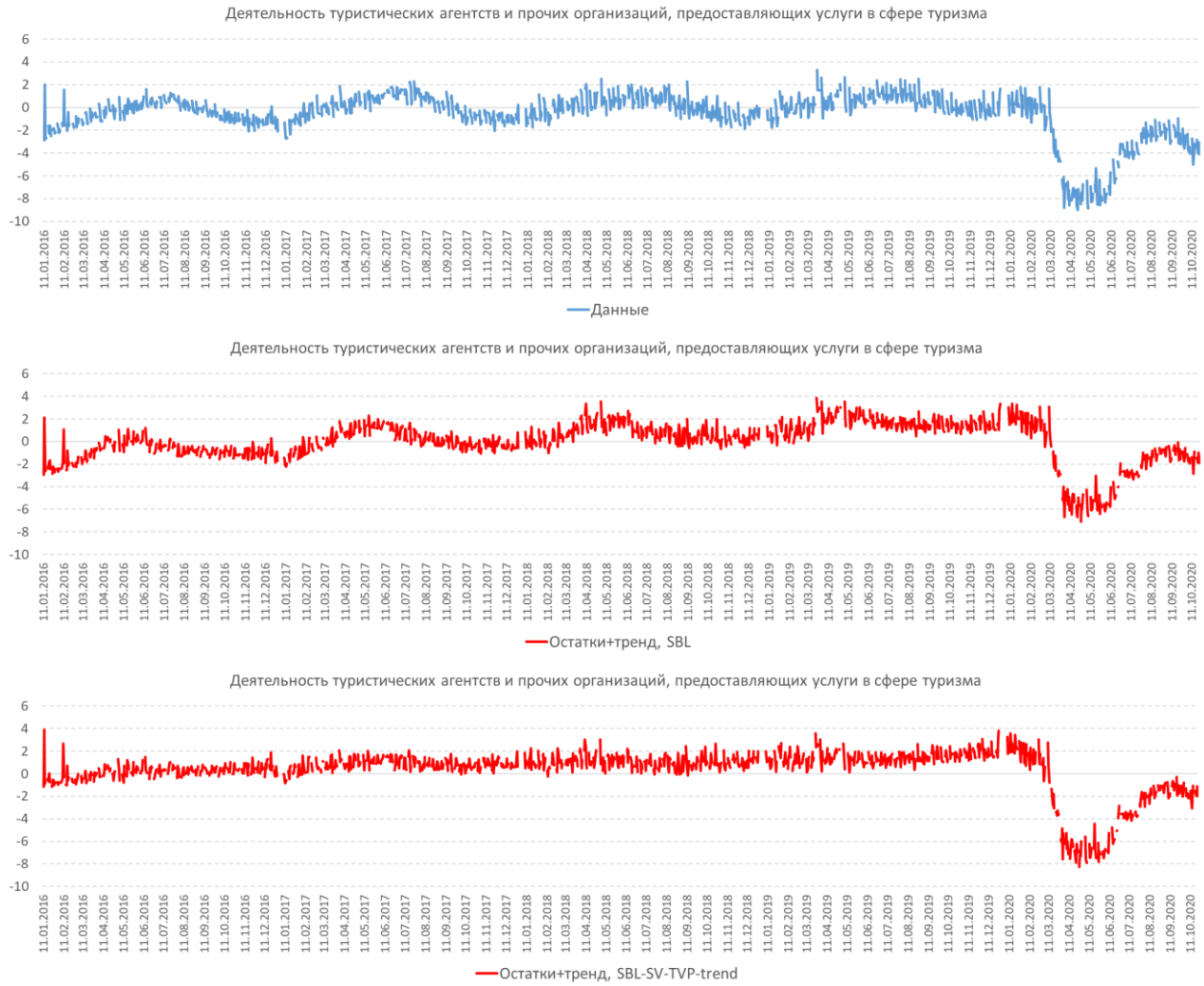


Рисунок 12. Данные и остатки для деятельности туристических агентств и прочих организаций, предоставляющих услуги в сфере туризма, нормировано на выборочное стандартное отклонение за 2016–2019 годы



Приложение Б. Процедура прогнозирования

В рамках данной работы одним из критериев качества выделения сезонной компоненты является точность ее вневыборочного прогноза, рассчитанная с помощью RMSE для логарифма финансовых потоков на нескольких горизонтах прогнозирования. Во-первых, это прогноз для рабочих дней на горизонте 1–5 дней, а также на 7, 14, 28 и 84 календарных дня.

Для этого каждая модель последовательно оценивалась на расширяющемся окне с шагом в один день на выборке, состоящей из всех дней 2019 года, за исключением выходных и праздничных. Первый прогноз строился на основании информации, доступной к концу 2018 года (28 декабря 2018 года), а последний – на основе информации, доступной к предпоследней точке 2019 года (27 декабря 2019 года). Для каждой модели строился прогноз на $h \in \{1, \dots, 5, 7, 14, 28, 84\}$ дней вперед, что формирует тестовую выборку, на основании которой оценивались RMSE. Так, например, для $h = 1$ тестовая выборка состояла из 246 точек с 9 января до 30 декабря 2019 года, а для $h = 2$ с 10 января 2019 года по 9 января 2020 года.

Учет праздничных и выходных дней на тестовой выборке для разных горизонтов производился по-разному:

1. *Прогноз рабочих дней на горизонте с 1 до 5 дней вперед.* Модели переоценивались на подвыборках, заканчивающихся рабочими днями. Это делалось рекурсивно с добавлением одного рабочего дня в выборку, на которой оценивалась модель. При этом прогноз строился на h рабочих дней вперед. Например, если $h = 2$, а последний день оцениваемой выборки оказался пятницей, то прогнозируются логарифмы потоков во вторник. Таким же образом исключаются праздничные дни. Так, если последний день выборки оказался прямо перед началом праздничных дней, то для $h = 2$ прогнозируются логарифмы потоков через два дня после окончания праздников.
2. *Прогноз на 7, 14, 28 дней и 84 календарных дня вперед*, что примерно соответствует периоду прогнозирования в неделю, две недели, месяц и квартал вперед. Модели также переоценивались рекурсивно с добавлением одного рабочего дня в выборку, на которой оценивалась модель. Однако прогноз строился на h календарных дней (включая праздничные и выходные дни) вперед. При этом, если прогнозируемый день оказывался выходным или праздничным днем, данный прогноз исключался из тестовой выборки. В силу того что выбранные горизонты прогнозирования кратны 7, каждый раз прогнозируется тот же день недели через неделю, две недели, примерно через месяц и примерно через квартал. Таким образом, исключаются из выборки только прогнозы на/в праздничные дни.

SBL-регрессия ограничена тем, что она оценивалась только на выборке рабочих дней. Методы Prophet, STL и TBATS оценивались на выборке календарных дней,

включая выходные и праздничные дни. При этом все значения выходных и праздничных дней заменялись на средние значения потоков за рабочую неделю до начала данного выходного или праздничного дня. Однако для всех горизонтов прогнозирования тестовая выборка полностью совпадала с SBL-моделью. Более того, каждое значение тестовой выборки прогнозировалось с той же даты, что и SBL-модель.

Гиперпараметры моделей оценивались на выборке до конца 2018 года, как если бы отсутствовала информация о динамике переменных 2019 года. Однако для каждой модели существуют свои особенности выбора гиперпараметров.

1. Для SBL-регрессии гиперпараметры априорной дисперсии оценивались на выборке с 1 января 2016 года до 31 декабря 2018 года, за исключением праздничных и выходных дней.
2. Для моделей Prophet с поворотными точками и без поворотных точек и STL-модели формировалась валидационная выборка, на которой строился вневыборочный прогноз по типу тестовой выборки и рассчитывался RMSE. На основании рассчитанного RMSE выбираются гиперпараметры моделей. При этом валидационная выборка строилась таким образом, чтобы выбрать гиперпараметры на основании информации вплоть до конца 2018 года (28 декабря 2018 года) и при этом так, чтобы размер выборки был равен ровно одному году. Например, если $h = 7$, первый прогноз строится на основании оценки модели до 22 декабря 2017 года (предпоследняя пятница 2017 года), а последний – на основании модели, построенной на данных до 21 декабря 2018 года (предпоследняя пятница 2018 года).
3. Для модели TBATS гиперпараметры выбирались на обучающей выборке, так как перебор на валидации является вычислительно сложным. Прогнозирование на тестовой выборке реализовывалось аналогично моделям Prophet и STL.

При прогнозировании LLT-модели процедура прогнозирования отличается. Все параметры модели оценивались на данных до конца 2018 года. Эти параметры затем фиксировались на тестовой выборке. Для того чтобы получить прогноз трендовой компоненты на расширяющемся окне, на каждой новой подвыборке оценивался прогноз с помощью фильтра Калмана при условии фиксированных параметров. Так же фиксировались и регрессионные коэффициенты для оценки сезонной компоненты.