

Прикладная эконометрика, 2020, т. 59, с. 71–87.

Applied Econometrics, 2020, v. 59, pp. 71–87.

DOI: 10.22394/1993-7601-2020-59-71-87

Д. А. Петрова, П. В. Трунин¹

Выявление настроений экономических агентов на основе поисковых запросов

В настоящее время поведение пользователей в Интернете становится важным источником информации о предпочтениях и настроении населения. В статье предлагается подход к выявлению настроений экономических агентов с помощью поисковых запросов Google Trends. При этом ключевые слова для поисковых запросов формируются на основе анализа новостей портала РБК. Динамика построенных на основе метода главных компонент индикаторов настроений с января 2010 г. по март 2020 г. в значительной мере коррелирует с индексом финансового стресса АКРА и индексом потребительской уверенности Росстата, что подтверждает возможность их использования при разработке и анализе экономической политики. Анализ предсказательной способности индексов поисковых запросов в прогнозировании индекса потребительской уверенности проводится в рамках модели со смешанной периодичностью данных (MIDAS). Результаты вневыборочного прогнозирования показали, что MIDAS с индикатором настроений на денежном рынке дает наилучший прогноз на следующий квартал и на 1 год, а MIDAS с индикатором настроений на финансовых рынках имеет высокую точность прогноза на 3 квартала.

Ключевые слова: поисковые запросы; Google Trends; тематическое моделирование; анализ новостей; настроения экономических агентов; прогнозирование.

JEL classification: C53; C69; C80; D83.

Введение

Сведения о текущем состоянии экономики являются ценным источником информации при принятии решений экономическими агентами. На основе полученных данных агенты формируют ожидания относительно будущей экономической активности и определяют, сколько они будут потреблять, сберегать и инвестировать, формируя таким образом ключевые макроэкономические показатели. Одним из возможных способов выявления ожиданий выступают опросы населения, предприятий и профессиональных аналитиков. Однако традиционный подход к измерению настроений или уверенности потребителей на основе опросов обладает рядом недостатков (Голощапова, Андреев, 2017; Song, Shin, 2019). Во-первых, для проведения исследования, сбора и агрегирования данных требуется время,

¹ Петрова Диана Абдумуминовна — Российская академия народного хозяйства и государственной службы при Президенте РФ, Москва; petrova-da@РАНХиГС.ru.

Трунин Павел Вячеславович — Российская академия народного хозяйства и государственной службы при Президенте РФ, Институт экономической политики имени Е. Т. Гайдара, Москва; pt@ier.ru.

поэтому опросные индикаторы имеют низкую частотность (ежемесячную или ежеквартальную) и публикуются с запаздыванием. Во-вторых, получаемые результаты чувствительны к формулировкам вопросов, задаваемых респондентам. В-третьих, проведение опросов является дорогостоящим.

В настоящее время повседневная деятельность человека все больше связана с онлайн-сервисами. Согласно данным Росстата, доля пользователей Интернета в России с каждым годом растет, и в 2019 г. она достигла 81.4% среди лиц в возрасте 15–74 лет². Таким образом, с повышением уровня проникновения Интернета появляется возможность анализа предпочтений пользователей. В условиях неопределенности экономические агенты зачастую не хотят принимать решения об объемах потребления, сбережений и инвестиций на основе имеющегося эмпирического опыта и знаний, накопленных в прошлом, и начинают поиск информации во внешних источниках. Неполнота знаний усиливает спрос на информацию и стимулирует поисковую активность. В таком случае спрос на определенную экономическую тему будет тесно связан с рыночными ожиданиями. Чем выше обеспокоенность экономических агентов текущим состоянием экономики, тем больше спрос на информацию в Интернете, где без значительных затрат и в доступной форме можно собрать сведения о происходящих событиях. Получение доступа к новой информации о текущей экономической ситуации позволяет населению лучше оценивать риски, принимать более взвешенные решения, способствует повышению эффективности рынков. Таким образом, данные о поведении пользователей в Интернете становятся одним из ключевых источников информации об их предпочтениях и ожиданиях.

Исследователи используют широкий спектр методов для построения высокочастотных индикаторов ожиданий и настроений экономических агентов на основе комментариев в социальных сетях (Bollen et al., 2011; Голощапова, Андреев, 2017), новостей в СМИ (Яковлева, 2018; Feuerriegel, Gordon, 2019; Larsen, Thorsrud, 2019) или поисковых запросов в Интернете (Dybka, Chojnowski, 2017).

На текущий момент в России существует несколько популярных индикаторов настроений экономических агентов, в том числе индекс потребительской уверенности Росстата, полученный на основе опроса респондентов, а также индекс финансового стресса АКРА, рассчитанный на основе рыночных индикаторов. При этом анализ поведения пользователей Интернета дает возможность строить аналогичные индикаторы, используя высокочастотные данные, независимо от публикации официальной статистики или проведения опросов.

В данной статье поисковые запросы Google Trends будут использованы для построения индекса настроений экономических агентов в России в период с января 2010 г. по март 2020 г. Важной особенностью данного исследования является подход к формированию списка ключевых слов для отбора поисковых запросов, основанный на анализе новостей в СМИ. Поскольку новостные СМИ являются одним из общедоступных источников информации о происходящих событиях, такой подход позволяет определить наиболее важные тематические группы поисковых запросов для выявления обеспокоенности населения экономическими проблемами.

² Доля населения в % от общей численности населения в возрасте 15–74 лет, см. https://www.gks.ru/free_doc/new_site/business/it/fed_nabl-croc/PublishData/Reports/Files/2019/3.1b.xlsx.

1. Эмпирические подходы к моделированию макроэкономических показателей на основе интернет-запросов

Д. А. Петрова, П. В. Трунин

Как показывают многие исследования, поисковые запросы в Интернете являются статистически значимыми прокси-факторами ожиданий или настроений экономических агентов. При этом важной исследовательской проблемой является подход к отбору ключевых слов для запросов. Так, существует ряд эмпирических работ, авторы которых основывают свой анализ на небольшом числе поисковых запросов, полученных исходя из экономических предположений об их потенциальной важности для экономических агентов, осуществляющих поиск по той или иной тематике.

Например, в работе (Guzman, 2011) в качестве меры инфляционных ожиданий рассматриваются поисковые запросы по ключевому слову «инфляция» и проводится сравнение данного индикатора с различными показателями, полученными на основе опросов в США.

Одной из первых работ по исследованию поведения пользователей в период финансового кризиса стала статья (Smith, 2012). По мнению автора, Интернет является одним из ключевых источников информации, поэтому повышение интереса пользователей к определенной теме отражает процесс распространения информации и ожидания экономических агентов в условиях кризиса. Автор проводит анализ влияния поисковых запросов по ключевым словам «экономический кризис» и «финансовый кризис», «рецессия» и «инфляция» на волатильность валютных курсов к доллару для семи развитых стран в течение 2004–2010 гг. на еженедельных данных. Выбор ключевых слов обусловлен возможностью проверить, происходит ли в кризисный период переток средств из более рискованных активов в менее рискованные (в кризис инвесторы начинают покупать больше активов с минимальным риском, таких как казначейские облигации США или другие долговые инструменты, номинированные в долларах США), а, следовательно, в кризисные периоды выбранные поисковые запросы могут быть связаны с волатильностью валютных курсов к доллару США. Результаты показали, что интенсивность поискового запроса «экономический кризис + финансовый кризис» отражает обеспокоенность населения кризисными явлениями и оказывает повышательное давление на волатильность валютного курса.

На российских данных также исследовалось поведение пользователей в Интернете и их обеспокоенность экономическими проблемами. В (Борочкин, 2013) рассматривалась динамика интенсивности поискового запроса «кризис» в течение 2005–2012 гг. Результаты свидетельствуют о наличии значимого влияния интернет-запроса «кризис» на отток капитала и безработицу в России.

Также существуют работы, в которых рассматриваются поисковые запросы по широкому кругу экономических вопросов для анализа настроений экономических агентов. В литературе можно выделить два основных эмпирических подхода при работе с большим количеством поисковых запросов — методы отбора и выделения переменных.

Под методами отбора переменных понимаются модели с регуляризацией и байесовские модели. Такие подходы снижают возможность переобучения, решают проблему мультиколлинеарности и способны исключать из анализа менее важные переменные. Например, в работе (Seabold, Coppola, 2015) были использованы поисковые запросы, связанные с потребительскими настроениями, для анализа их влияния на различные индексы потребительских цен в Коста-Рике, Сальвадоре и Гондурасе на протяжении 2004–2014 гг. Авторы показали, что лучше всего отбирать наиболее важные поисковые запросы позволяет эластичная сеть.

В работе (Фантаццини и др., 2018) были продемонстрированы преимущества использования байесовской модели усреднения для отбора наиболее важных категорий интернет-запросов Google Trends для построения индикаторов социального самочувствия российского населения за период 2009–2016 гг. на квартальных данных.

Scott, Varian (2013) для анализа потребительских настроений на основе интернет-запросов Google Trends используют модель байесовских структурных временных рядов для США в период с января 2004 г. по апрель 2012 г. В целом авторы делают вывод, что байесовские методы помогают получить небольшое число ключевых поисковых запросов.

Другим подходом к анализу поисковых запросов является выделение переменных из большого набора данных с помощью факторных моделей. Kholodilin и Podstawski (2009) проводили исследование влияния настроений потребителей, построенных на основе данных Google Trends, на потребление в США с января 2005 г. по декабрь 2009 г. Особенность работы заключалась в том, что на первом этапе отбирались наиболее связанные с тематикой потребления категории, а на втором этапе из выделенных 27 категорий происходил отбор десяти самых популярных ключевых слов. На основе отобранных категорий поисковых запросов осуществлялось построение индикатора потребительских настроений с помощью метода главных компонент и его сравнение с традиционными индексами потребительской уверенности, полученными на основе опросов населения.

В работе (Bulut, 2018) поисковые запросы используются для прогнозирования валютных курсов 11 развитых стран с января 2004 г. по июнь 2014 г. На первом этапе отбираются поисковые запросы из предположений экономических теорий формирования фундаментального валютного курса — монетарного подхода и паритета покупательной способности. На втором этапе для каждой группы строится индикатор ожиданий экономических агентов с помощью метода главных компонент. Результаты эмпирического исследования показали, что модели с поисковыми запросами дают более точный прогноз, чем модель случайного блуждания, только для Австралии, Канады, Дании и Сингапура.

В работе (Dybka, Chojnowski, 2017) предлагается подход к построению настроений экономических агентов на кредитном и финансовых рынках Польши, а также инфляционных ожиданий в рамках анализа поведения пользователей в Интернете в период с января 2004 г. по май 2016 г. Индикаторы настроений строятся с помощью метода главных компонент на основе групп поисковых запросов, выбранных из экономических предположений о возможном влиянии на инфляцию, кредитный и финансовые рынки.

В целом международный и российский опыт говорит о том, что поисковые запросы Google Trends позволяют лучше идентифицировать настроение населения по сравнению с традиционными показателями, полученными на основе опросов, например, индексом потребительской уверенности или инфляционных ожиданий. Это обусловлено тем, что сбор информации в Интернете очень удобен и происходит с минимальными затратами.

Однако остается открытым вопрос о способе отбора наиболее важных поисковых запросов. В большинстве работ отбор производится из экономических предположений о механизме влияния поведения пользователя в Интернете на принятие экономических решений. Но при этом возможны потери информации при пропуске важных интернет-запросов, поскольку фактический интерес пользователей к ключевым событиям может не учитываться. Если рассматривать слишком широкий набор ключевых слов для загрузки поисковых запросов, то возникает проблема сложности отбора переменных и получения интерпретируемых показателей настроений.

В данной статье предлагается отбор поисковых запросов на основе анализа новостей в СМИ, который позволяет получить кластеры разделенных по темам ключевых слов, для выявления настроений экономических агентов.

2. Данные и отбор ключевых слов

Сервис Google Trends публикует временные ряды интенсивности поисковых запросов в диапазоне от 0 до 100 баллов по различным ключевым словам в поисковой системе Google для заданного местоположения. Расчет временного ряда проводится следующим образом (Ma et al., 2015):

$$GT_{K_i,t} = \frac{S_{K_i,t}}{V_t} \cdot 100, \quad (1)$$

где $S_{K_i,t}$ — число интернет-запросов в Google Trends по определенному ключевому слову/фразе K_i на момент времени t ; V_t — совокупное число поисковых запросов в момент времени t ; $GT_{K_i,t}$ — интерес к определенной теме.

При использовании рядов интенсивности интернет-запросов в эмпирическом анализе возникает ряд сложностей. Во-первых, временные ряды поисковых запросов доступны только с января 2004 г. Во-вторых, поиск в сети Интернет часто коррелирует с такими демографическими характеристиками, как возраст и доход, поэтому выборка пользователей нерепрезентативна. В-третьих, возникает проблема выбора ключевых слов по причине возможного пропуска важных поисковых запросов в эмпирическом анализе на основе экономических предположений или Google Correlate³.

Для преодоления последнего недостатка в данной работе отбор ключевых слов для получения перечня поисковых запросов основывается на анализе новостей. Применение данного метода позволяет получить разбиение ключевых слов по тематикам, т. к. новостные СМИ публикуют статьи о наиболее важных событиях, на основе которых экономические агенты формируют свое мнение о текущей ситуации в экономике (Nimark, Pitschner, 2019; Thorsrud, 2020).

Как и в (Li и др., 2015), в настоящей статье предполагается, что пользователи формируют список ключевых слов, связанных с важными для них событиями, для поиска информации в Интернете. Такой подход позволяет смоделировать поведение экономических агентов при условии, что ключевые слова удовлетворяют двум предположениям. Во-первых, они содержатся в информации, которую ищут пользователи. Во-вторых, люди при поиске информации руководствуются интересующей их тематикой. На основе полученной информации пользователи Интернета могут принимать экономические решения относительно потребления, сбережений и инвестиций. Когда эти предположения выполнены, ключевые слова могут быть эффективно извлечены из новостей, а значит, будут использоваться при поиске в Интернете.

³ Google закрыл сервис по подбору интернет-запросов с 15 декабря 2019 г.

Выбор новостного источника для формирования ключевых слов основан на следующих принципах. Интернет-издание, во-первых, должно на ежедневной основе освещать широкий круг экономических новостей на русском языке; во-вторых, иметь большую аудиторию пользователей; в-третьих, обладать простой структурой для технической обработки текстов новостей. Подходящим под эти принципы источником информации является интернет-издание РБК (<https://www.rbc.ru/>).

При предобработке 358438 статей РБК по различным тематикам с января 2010 г. по март 2020 г. применяются стандартные процедуры: удаление стоп-слов⁴, лемматизация для русского языка (с использованием Python-библиотеки *py morphology2*). Векторизация новостных статей проводится с помощью мешка слов⁵.

Одним из возможных способов извлечения тем (Воронцов и др., 2015; Яковлева, 2018; Larsen, Thorsrud, 2019), содержащих ключевые слова для выявления настроений экономических агентов, может выступать латентное размещение Дирихле (LDA). В отличие от вероятностного латентного семантического анализа (PLSA), также часто используемого для тематического моделирования корпуса новостей, при использовании данного метода число оцениваемых параметров не увеличивается линейно с ростом числа документов, а также в явном виде задаются распределение тем в тексте и распределение слов в теме. Задаваемое априорное распределение Дирихле в данном случае является байесовской регуляризацией, позволяющей накладывать ограничения на параметры для снижения размерности и упростить получение апостериорных распределений (Воронцов, 2013).

Латентное размещение Дирихле представляет собой вероятностный метод кластеризации текстов по темам, при котором каждое слово из новостной статьи принадлежит с определенной вероятностью скрытой теме. Каждая новостная статья (Blei et al., 2003) рассматривается как смесь заданных тем с вероятностью

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta), \quad (2)$$

где θ — совместное распределение тем; w_n — n -е слово в тексте; z_n — распределение слов в каждой теме; N — число слов в тексте новостной статьи; α, β — гиперпараметры, характеризующие выраженность тем в документе и разреженность распределения слов в темах соответственно.

Для построения латентного размещения Дирихле с помощью пакета *Gensim* в Python задаются 100 тем⁶, по корпусу при обучении модели осуществляется 10 проходов, а для оценки параметров применяется онлайн-алгоритм вариационного байесовского вывода (Hoffman et al., 2010).

В итоге обучения модели была произведена кластеризация текстов новостей по широкому кругу вопросов, включающих политические события, международную тематику,

⁴ Слова, часто встречающиеся и не отражающие смысловую нагрузку в тексте.

⁵ Подсчитывается частота вхождения слов в текст новости.

⁶ Для подбора оптимального числа производился перебор от 50 до 180 тем (T) с шагом в 10. При выборе $T=50$ полученные результаты были плохо интерпретируемы из-за объединения несвязанных тем в одну. В случае $T=180$ происходило повторение тем.

чрезвычайные происшествия, судебные решения, бюджетно-налоговую политику, макроэкономическую ситуацию, финансовую тематику и т. д. В связи с тем, что задачей авторов является построение индикаторов, учитывающих реакцию экономических агентов на экономические шоки, из всего спектра вопросов было выбрано подмножество тем, связанных с событиями в экономике. В частности, тематическая группа, условно названная «денежный рынок», включает в себя новости, содержащие информацию о крупнейших банках, денежно-кредитной политике, банковских услугах или операциях на долговом рынке. Таким образом, данная тема отражает интерес экономических агентов к банковскому сектору и долговому рынку, подверженных влиянию изменения процентной ставки Банком России. Тематическая группа «экономические условия» содержит ключевые слова, характеризующие состояние экономики. В новостях по тематике «финансовые рынки» обсуждаются вопросы, связанные с фондовым, валютным и нефтяным рынками.

Всего были получены 4 темы для категории «финансовые рынки», 3 темы для «денежного рынка» и 1 тема для «экономических условий». Для каждой темы были отобраны только топ-20 ключевых слов с наибольшей вероятностью распределения слов в теме. Облака слов позволяют показать наиболее важные слова, характеризующие тематику новостей, а также наглядно продемонстрировать разделение тематических групп для последующего отбора поисковых запросов Google Trends.

На следующем этапе с учетом полученных ключевых слов по тематикам финансовых рынков, денежного рынка и экономических условий отбирались только наиболее релевантные поисковые запросы для построения индикаторов настроений на ежемесячной основе. В анализ не включались следующие типы поисковых запросов: характеризующие направление изменения показателей (вырасти, подняться, опуститься и т. д.), имеющие несколько значений при поиске в Интернете (решение, доля, срок и т. д.), а также поисковые запросы небольшой популярности или с изменчивостью во времени. Финальный список ключевых слов можно увидеть в табл. 1.

Таблица 1. Перечень поисковых запросов для построения настроений пользователей

Тематические группы	Поисковые запросы
Финансовые рынки	капитализация, ценная бумага, акция, фондовый индекс, котировка, московская биржа, валютная интервенция, цена на нефть, баррель, рынок нефти, опек, рубль, курс доллара, валюта, евро, дивиденд, экспорт, акционер, инвестор, цб
Денежный рынок	долг, ключевая ставка, процентная ставка, цб, банк, сбербанк, кредит, банковский счет, втб, кредитование, вклад, платеж, актив, центробанк, кредитор, заемщик, банкротство, облигация, аукцион, доходность, ипотека
Экономические условия	экономика, экономический рост, экономическое развитие, ввп, прогноз, инвестиция, инфляция, кризис, доход, цена

По результатам графического анализа было выявлено, что большая часть поисковых запросов имеет сезонный характер. Для устранения зашумленности в данных произведена корректировка сезонности с помощью процедуры X-13-ARIMA-SEATS⁷ в пакете *seasonal* на языке R.

⁷ <https://cran.r-project.org/web/packages/seasonal/seasonal.pdf>.

3. Построение индикаторов настроений на основе поисковых запросов

Агрегированные показатели настроений на финансовых рынках, на денежном рынке и относительно экономических условий строятся с помощью метода главных компонент⁸, в рамках которого по каждой группе поисковых запросов выбирается число главных компонент согласно правилу Кайзера⁹. В данном случае метод главных компонент дает возможность снизить размерность переменных и извлечь настроения пользователей в Интернете из поисковых запросов с минимальными потерями в информации. В результате рассматриваются 2 главные компоненты для индикатора экономических условий¹⁰ и по 5 главных компонент для индексов настроений на финансовых и денежном рынках¹¹. Агрегированные показатели строятся как комбинация главных компонент с весами, соответствующими их вкладу в дисперсию.

Для проверки адекватности полученных результатов оцененные показатели настроений сравниваются с индексом потребительской уверенности¹² и индексом финансового стресса АКРА¹³. Как показано в табл. 2, настроения на финансовых рынках положительно коррелируют с индексом финансового стресса (0.735) и отрицательно — с индексом потребительской уверенности (–0.536). Аналогичный результат был получен для настроений на денежном рынке, который отрицательно коррелирует с индексом потребительской уверенности (–0.523) и положительно — с индексом финансового стресса (0.603).

Индикатор экономических условий, по мнению авторов, недостаточно сильно коррелирует с индексом потребительской уверенности и индексом финансового стресса. По этой причине в дальнейшем остановимся на сравнительном анализе динамики настроений на финансовых и денежном рынках с индексами потребительской уверенности и финансового стресса, а также обсудим реакцию экономических агентов на кризисные явления и экономические шоки.

Таблица 2. Корреляция расчетных индикаторов с официальными показателями финансового стресса и потребительской уверенности

	Индекс финансового стресса АКРА	Индекс потребительской уверенности
Настроения на денежном рынке	0.603	–0.523
Настроения на финансовых рынках	0.735	–0.536
Настроения относительно экономических условий	–0.211	0.431

⁸ Данные по интенсивностям поисковых запросов Google Trends были стандартизированы.

⁹ Выбор числа главных компонент основан на предположении об их информативности, определяемой за счет превышения собственным значением значения 1.

¹⁰ Две главные компоненты объясняют около 67% дисперсии.

¹¹ Пять главных компонент объясняют 74.2 и 79.1% дисперсии поисковых запросов по тематике финансовых и денежного рынков соответственно.

¹² Данные публикуются Росстатом с 2006 г. на квартальной основе. Индекс потребительской уверенности рассчитывается как среднее арифметическое значение балансов оценок благоприятности условий для покупок, произошедших и ожидаемых изменений экономической ситуации в стране и личного материального положения.

¹³ Данные публикуются Аналитическим кредитным агентством с 2006 г. на ежедневной основе. Индекс финансового стресса характеризует состояние финансовой системы России. Задается граница перехода в состояние кризиса — 2.5 пункта.

Динамика временных рядов различных индикаторов настроений, построенных на основе поисковых запросов, и индекса финансового стресса АКРА представлена на рис. 1.



Рис. 1. Сравнение динамики индикаторов настроений и индекса финансового стресса

Настроения на финансовом и денежном рынках имеют схожую динамику с индексом финансового стресса. В качестве ключевых событий можно выделить конец 2014 г., когда в России произошел валютный кризис. В этот период Банк России ужесточил денежно-кредитную политику, подняв ключевую ставку с 10.5 до 17%. Колебания курса рубля, санкции и снижение цен на нефть повысили тревожность на финансовом рынке. В декабре 2014 г. индикатор настроений на денежном рынке достигает своего максимума. Это также согласуется с данными по индексу финансового стресса, который показывает, что в течение периода 2010–2014 гг. индекс впервые превысил порог в 2.5 пункта в декабре 2014 г., что характеризует переход финансовой системы России в состояние кризиса. Еще одним значимым событием являлось снижение цен на нефть до 13-летнего минимума в январе 2016 г. Это вызвало повышение уровня неопределенности на финансовом рынке и обесценение курса рубля из-за пересмотра ожиданий будущих перспектив. В этот период оба индикатора выросли, оставаясь при этом ниже кризисного периода 2014–2015 гг. Третьим стрессовым периодом стал март 2020 г., когда в условиях пандемии коронавируса страны — экспортеры нефти долго не могли договориться о снижении добычи нефти, а все страны столкнулись с параличом экономической активности вследствие введения карантинных мер.

Если сравнивать динамику настроений экономических агентов на финансовом и денежном рынках с индексом потребительской уверенности на квартальной основе (рис. 2), то можно увидеть, что при повышении обеспокоенности относительно финансовой тематики происходит снижение потребительской уверенности.

Стоит отметить, что настроения на финансовом и денежном рынках являются опережающим на один квартал индикатором для индекса потребительской уверенности. Например, повышение тревожности в 4-м квартале 2014 г. вызвало снижение уровня потребительской уверенности в 1-м квартале 2015 г. Снижение интереса к тематике финансовых рынков в течение 3-го квартала 2016 г. — 1-го квартала 2018 г. сопровождалось повышением уровня потребительской уверенности.

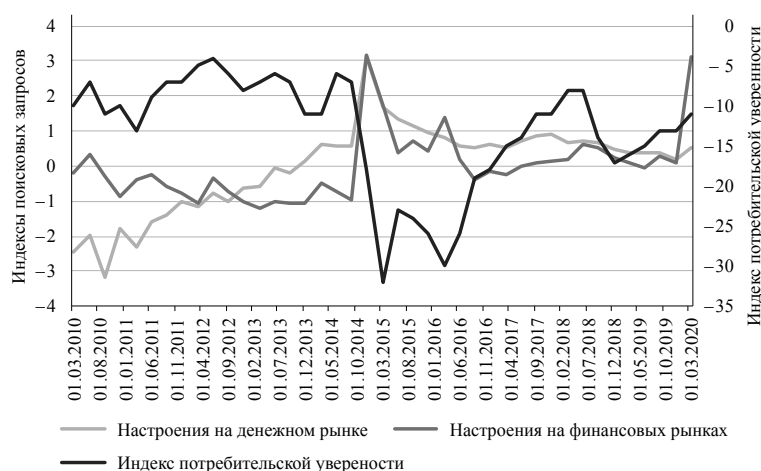


Рис. 2. Сравнение динамики индикаторов настроений и индекса потребительской уверенности

Для проверки устойчивости полученных результатов проведем сравнительный анализ прогностической способности индексов поисковых запросов и индекса финансового стресса АКРА при прогнозировании индекса потребительской уверенности.

4. Прогнозирование индекса потребительской уверенности в рамках модели MIDAS-AR

При прогнозировании макроэкономических показателей часто возникает проблема различной частоты объясняемой переменной и предикторов. Высокочастотные переменные могут характеризовать состояние экономики в режиме реального времени и обеспечивать улучшение качества прогноза низкочастотной переменной. Важно отметить, что простое усреднение высокочастотных временных рядов может отрицательно сказаться на качестве оцениваемой модели при прогнозировании. Для учета различной частоты используемых показателей часто применяется модель со смешанной периодичностью данных (Mixed Data Sampling).

В связи с доступностью статистики по индексу потребительской уверенности только на ежеквартальной основе далее будут использованы два подхода для сравнения предсказательной способности построенных индикаторов при прогнозировании данного показателя модели с усреднением ежемесячных индикаторов до квартальной частоты и модели со смешанной периодичностью данных.

Однофакторная MIDAS, предложенная в (Ghysels et al., 2007), позволяет строить прогноз переменной y_t на один квартал на основе лагов объясняющих факторов x_t^m , где m — периодичность высокочастотной переменной. В нашем случае, как и в работе (Clements, Galvão, 2009), объясняющие переменные имеют месячную частоту, поэтому $m = 3$:

$$y_t = \beta_0 + \lambda y_{t-1} + \beta_1 B(L^{1/3}, \theta)(1 - \lambda L)x_{t-1}^{(3)} + \varepsilon_t. \quad (3)$$

Прогноз на h шагов строится как

$$y_t = \beta_0 + \lambda y_{t-h} + \beta_1 B(L^{1/3}, \theta) (1 - \lambda L^h) x_{t-h}^{(3)} + \varepsilon_t, \quad (4)$$

где y_t — низкочастотная переменная; $x_t^{(3)}$ — высокочастотная переменная: настроения на финансовом (fin) и денежном (money) рынках, построенные с помощью поисковых запросов, и индекс финансового стресса (finstress) с 3 лагами, и каждая из них доступна в 3 раза чаще, чем y_t ; функция $B(L^{1/3}, \theta)$ определяет веса для временного агрегирования; L — оператор запаздываний.

Вне выборки прогноз индекса потребительской уверенности в псевдореальном времени строится с 1-го квартала 2019 г. по 1-й квартал 2020 г. Как и в работе (Günaş, 2020), рассматривались различные функции взвешивания $B(L^{1/3}, \theta)$: полиномиальный распределенный лаг Алмона (Almon), бета-функция (beta), шаговая функция (step), экспоненциальное взвешивание Алмона (expalmon) и неограниченная модель со смешанной периодичностью данных (umidas). Шаговая функция взвешивания присваивает одинаковые веса выбранному числу лагов на каждом шаге¹⁴. Полиномиальный распределенный лаг Алмона строится как

$$\beta w_i(\theta_0, \dots, \theta_p) = \sum_{k=0}^p \theta_k i^k. \quad (5)$$

В неограниченной U-MIDAS модели оцениваются коэффициенты для каждого лага высокочастотной переменной:

$$y_t = \beta_0 + \lambda y_{t-h} + B(L) x_{t-h} + \varepsilon_t. \quad (6)$$

В отличие от трех предыдущих подходов к временному агрегированию, модели с экспоненциальным лагом Алмона и бета-функцией оцениваются нелинейным методом наименьших квадратов. Экспоненциальный лаг Алмона имеет следующий вид:

$$b(i; \theta) = \frac{\exp(\theta_1 i + \theta_2 i^2)}{\sum_{j=0}^k \exp(\theta_1 j + \theta_2 j^2)}. \quad (7)$$

Бета-функция взвешивания:

$$w_i(\theta_1, \theta_2, \theta_3) = \frac{\omega_i^{\theta_1-1} (1 - \omega_i)^{\theta_2-1}}{\sum_{j=0}^k \omega_j^{\theta_1-1} (1 - \omega_j)^{\theta_2-1}} + \theta_3. \quad (8)$$

В качестве эталонной прогнозной модели для индекса потребительской уверенности рассматривается авторегрессионная (AR) модель. Также оцениваются ARX модели, включающие усредненные объясняющие факторы, для сравнения схем с простым усреднением и временным агрегированием в модели со смешанной периодичностью данных. Качество

¹⁴ Например, устанавливая шаг, равный 3, получим одинаковый вес θ_0 для первых трех запаздываний высокочастотной переменной, θ_1 для следующих трех лагов и т. д.

прогноза сравнивается по показателю отношения корня из среднеквадратичной ошибки прогноза (RMSFE) моделей с индикаторами настроений к RMSFE эталонной модели¹⁵. Модели обучаются на периоде с 1-го квартала 2010 г. по 4-й квартал 2018 г.

В таблице 3 представлены основные результаты прогнозирования индекса потребительской уверенности на тестовой выборке 1-й квартал 2019 г. — 1-й квартал 2020 г. Точность прогноза на следующий квартал в целом выше у расширенных моделей, включающих индексы поисковых запросов или финансового стресса АКРА. Стоит отметить, что на данном горизонте прогноза модели со смешанной периодичностью данных, включающие индексы поисковых запросов или индекс финансового стресса, имеют более высокую точность прогноза, чем ARX и AR модели. Это говорит о том, что в лагах объясняющих факторов содержится важная информация для прогнозирования индекса потребительской уверенности, которая теряется при простом усреднении индикаторов настроений. Наилучший прогноз дают модели MIDAS с настройками на денежном рынке.

Таблица 3. Отношение RMSFE для различных моделей и горизонтов прогнозирования

Модели	$h = 1$	$h = 2$	$h = 3$	$h = 4$
ar	1.000	1.000	1.000	1.000
ar-fin	0.976	1.088	0.700	0.794
ar-finstress	0.813	1.398	1.061	1.086
ar-money	1.183	1.251	0.858	0.858
midas-fin-almon	0.898	1.050	0.678	0.848
midas-fin-beta	0.889	1.100	0.664	0.852
midas-fin-step	0.858	1.026	0.702	0.855
midas-fin-expalmon	0.898	1.092	0.818	0.852
midas-fin-umidas	0.889	1.100	0.818	0.852
midas-finstress-almon	0.624	1.397	1.036	1.123
midas-finstress-beta	0.726	1.455	1.081	1.146
midas-finstress-step	0.624	3.246	2.378	1.991
midas-finstress-expalmon	0.610	1.762	1.109	1.172
midas-finstress-umidas	0.448	1.475	1.081	1.153
midas-money-almon	0.501	1.202	0.865	0.828
midas-money-beta	0.508	1.196	0.915	0.890
midas-money-step	0.980	1.081	0.757	0.715
midas-money-expalmon	0.873	1.161	0.890	0.890
midas-money-umidas	0.508	1.195	0.890	0.890

При прогнозировании на следующие два квартала лучшей моделью оказалась эталонная. Однако, если сравнивать качество расширенных моделей, то модели с индексами поисковых запросов имеют более высокую точность прогноза, чем модели с индексом финансового стресса. Таким образом, с увеличением горизонта прогнозирования уменьшается предсказательная сила такого рыночного индикатора настроений экономических агентов, как индекс финансового стресса.

¹⁵ Расширенная модель лучше эталонной при отношении $RMSFE < 1$.

При горизонте прогноза на три квартала наилучшей моделью в среднем является MIDAS с настройками на финансовом рынке. MIDAS с настройками на денежном рынке имеет более высокую предсказательную способность по сравнению с эталонной моделью. ARX и MIDAS, включающие индекс финансового стресса, имеют невысокую точность прогноза. Это свидетельствует о преимуществе использования индексов поисковых запросов при прогнозировании индекса потребительской уверенности.

Наилучшей моделью при прогнозе на один год стала модель со смешанной периодичностью данных, включающая индекс настроений на денежном рынке. Стоит отметить, что все модели с индексами поисковых запросов имеют более высокую предсказательную силу по сравнению с эталонной моделью. Как и на горизонтах в 2–3 квартала, модели с индексом финансового стресса остаются менее точными, чем модели с поисковыми запросами или авторегрессионная модель.

Заключение

Сравнение построенных в данной статье показателей настроений с рыночным и опросным индикаторами показало, что настроения на финансовом и денежном рынках могут отражать обеспокоенность населения экономическими проблемами. Кроме того, настроения на финансовом и денежном рынках являются опережающими индикаторами для индекса потребительских настроений и могут улучшить его прогноз.

Для сравнения предсказательной способности различных индикаторов настроений населения при прогнозировании индекса потребительской уверенности применялись модели со смешанной периодичностью данных. Такой подход позволил решить проблему различия в частотности между квартальным индексом потребительской уверенности и ежемесячными индикаторами настроений, а также учесть влияние лагов объясняющих переменных при оценке небольшого числа параметров.

Модели со смешанной периодичностью данных с поисковыми запросами показали высокую предсказательную способность при прогнозировании индекса потребительской уверенности на следующие один и три квартала по сравнению с эталонной моделью. Это говорит о преимуществах использования индексов поисковых запросов, содержащих информацию о настроениях населения, для прогнозирования опросных индикаторов. Сравнительный анализ качества моделей с индикаторами настроений, построенных на основе поисковых запросов, и индексом финансового стресса на различных горизонтах прогноза позволил получить следующие результаты.

Наилучшей прогнозной моделью на горизонтах один и четыре квартала оказалась MIDAS с настройками на денежном рынке. Независимо от функции взвешивания, более высокую предсказательную способность индекса потребительской уверенности на следующие 3 квартала показала модель MIDAS с настройками на финансовом рынке. При этом ни один из высокочастотных индикаторов не дал улучшения качества прогноза индекса потребительской уверенности на следующие два квартала по сравнению с авторегрессионной моделью.

В целом по результатам работы можно сделать следующие выводы. Во-первых, полученные результаты показывают, что поисковые запросы могут быть использованы для построения сводных индексов настроений пользователей в Интернете, влияющих на принятие экономических решений. Использование интенсивностей поисковых запросов особенно

актуально в периоды экономического кризиса для понимания настроений экономических агентов и возможных последствий изменения их поведения. Во-вторых, индикаторы настроений на финансовом и денежном рынках могут быть построены исходя из активности пользователей в Интернете без использования экономической или финансовой статистики, что упрощает и ускоряет их расчет. В-третьих, на основе интернет-данных представляется возможным получить опережающие индикаторы для индекса потребительской уверенности.

Список литературы

Борочкин А. А. (2013). Использование статистики поисковых запросов в сети Интернет для краткосрочного прогнозирования макроэкономических переменных. *Деньги и кредит*, 8, 27–32.

Воронцов К. В. (2013). Вероятностное тематическое моделирование. <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>.

Воронцов К. В., Фрей А. И., Апишев М. А., Потапенко А. А. (2015). Тематическое моделирование в BigARTM: теория, алгоритмы, приложения. <http://www.machinelearning.ru/wi-ki/images/b/bc/Voron-2015-BigARTM.pdf>.

Голощапова И. О., Андреев М. Л. (2017). Оценка инфляционных ожиданий российского населения методами машинного обучения. *Вопросы экономики*, 6, 71–93.

Фантащини Д., Шаклеина М., Юрас Н. (2018). Big Data в определении социального самочувствия населения России. *Прикладная эконометрика*, 50, 43–66.

Яковлева К. (2018). Оценка экономической активности на основе текстового анализа. *Деньги и кредит*, 77 (4), 26–41.

Blei D., Ng A., Jordan M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Bollen J., Mao H., Zeng X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2 (1), 1–8.

Bulut L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*, 37 (3), 303–315.

Clements M. P., Galvão A. B. (2009). Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics*, 24 (7), 1187–1206.

Dybka P., Chojnowski M. (2017). Is exchange rate moody? Forecasting exchange rate with Google Trends data. *Econometric Research in Finance*, 2, 1–21.

Feuerriegel S., Gordon J. (2019). News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. *European Journal of Operational Research*, 272 (1), 162–175.

Ghysels E., Sinko A., Valkanov R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26 (1), 53–90.

Günay M. (2020). Nowcasting Turkish GDP with MIDAS: Role of functional form of the lag polynomial. *Working Paper No. 20/02*. Central Bank of the Republic of Turkey.

Guzman G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*, 36 (3), 119–167.

Hoffman M., Bach F., Blei D. (2010). Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 23, 856–864.

Larsen V. H., Thorsrud L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210 (1), 203–218.

Li X., Shang W., Wang S., Ma J. (2015). A MIDAS modelling framework for Chinese inflation index forecast incorporating Google search data. *Electronic Commerce Research and Applications*, 14, 112–125.

Ma J., Li X., Wang S., Zhang X. (2015). How does Google search affect trader positions and crude oil prices? *Economic Modelling*, 49, 162–171.

Nimark K. P., Pitschner S. (2019). News media and delegated information choice. *Journal of Economic Theory*, 181, 160–196.

Seabold S., Coppola A. (2015). Nowcasting prices using Google Trends. An application to Central America. *Policy Research Working Paper Series 7398*. The World Bank.

Scott S. L., Varian H. R. (2013). Bayesian variable selection for nowcasting economic time series. *NBER Working Papers*, 19567.

Smith J. P. (2012). Google Internet search activity and volatility prediction in the market for foreign currency. *Finance Research Letters*, 9, 103–110.

Song M., Shin K. (2019). Forecasting economic indicators using a consumer sentiment index: Survey-based versus text-based data. *Journal of Forecasting*, 38 (6), 504–518.

Thorsrud L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business and Economic Statistics*, 38 (2), 393–409.

Поступила в редакцию 13.06.2020;
принята в печать 01.08.2020.

Petrova D., Trunin P. Revealing the mood of economic agents based on search queries. *Applied Econometrics*, 2020, v. 59, pp. 71–87.

DOI: 10.22394/1993-7601-2020-59-71-87

Diana Petrova

Russian Presidential Academy of National Economy and Public Administration. Moscow, Russian Federation; petrova-da@ranepa.ru

Pavel Trunin

Russian Presidential Academy of National Economy and Public Administration, Gaidar Institute for Economic Policy. Moscow, Russian Federation; pt@iep.ru

Revealing the mood of economic agents based on search queries

Currently, user behavior on the Internet is becoming a key source of information about the market sentiments and the public mood. In this regard, it becomes relevant to study the usefulness of such indicators in modeling macroeconomic indicators. This article proposes an approach to assessment the mood of economic agents using Google Trends search queries. A key particularity of the article is the selection of keywords based on the analysis of RBC news from January 2010 to March 2020. The results showed that sentiments in financial and money markets based on a principal component analysis strongly correlated with the financial stress index of ACRA and Rosstat consumer confidence index, which confirms the possibility of use search queries in the development and analysis of economic

policy. We use search query indices for the consumer confidence index forecasting with mixed data sampling (MIDAS). In out-of-sample forecasting, our results show that MIDAS which includes the indicator of sentiment in the money market gives the best forecast performance for the next quarter, and MIDAS with the sentiment in the financial markets has a high accuracy of forecasting the consumer confidence index for 3 quarters.

Keywords: search queries; Google Trend; topic modeling; text analysis; sentiment; forecasting.

JEL classification: C53; C69; C80; D83.

References

- Borochkin A. A. (2013). The use of search statistics for the short-term forecasting of macroeconomic variables. *Money and Credit*, 8, 27–32 (in Russian).
- Vorontsov K. V. (2013). Probabilistic topic modeling. <http://www.machinelearning.ru/wi-ki/images/2/22/Voron-2013-ptm.pdf> (in Russian).
- Vorontsov K. V., Frey A. I., Apishev M. A., Potapenko A. A. (2015). Topic modeling in BigARTM: Theory, algorithms, applications. <http://www.machinelearning.ru/wi-ki/images/b/bc/Voron-2015-BigARTM.pdf> (in Russian).
- Goloshchapova I., Andreev M. (2017). Measuring inflation expectations of the Russian population with the help of machine learning. *Voprosy Ekonomiki*, 6, 71–93 (in Russian).
- Fantazzini D., Shackleina M., Yuras N. (2018). Big Data for computing social well-being indices of the Russian population. *Applied Econometrics*, 50, 43–66 (in Russian).
- Yakovleva K. (2018). Text mining-based economic activity estimation. *Russian Journal of Money and Finance*, 77 (4), 26–41 (in Russian).
- Blei D., Ng A., Jordan M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bollen J., Mao H., Zeng X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2 (1), 1–8.
- Bulut L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*, 37 (3), 303–315.
- Clements M. P., Galvão A. B. (2009). Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics*, 24 (7), 1187–1206.
- Dybka P., Chojnowski M. (2017). Is exchange rate moody? Forecasting exchange rate with Google Trends data. *Econometric Research in Finance*, 2, 1–21.
- Feuerriegel S., Gordon J. (2019). News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. *European Journal of Operational Research*, 272 (1), 162–175.
- Ghysels E., Sinko A., Valkanov R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26 (1), 53–90.
- Günay M. (2020). Nowcasting Turkish GDP with MIDAS: Role of functional form of the lag polynomial. *Working Paper No. 20/02*. Central Bank of the Republic of Turkey.
- Guzman G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*, 36 (3), 119–167.
- Hoffman M., Bach F., Blei D. (2010). Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 23, 856–864.

Larsen V. H., Thorsrud L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210 (1), 203–218.

Li X., Shang W., Wang S., Ma J. (2015). A MIDAS modelling framework for Chinese inflation index forecast incorporating Google search data. *Electronic Commerce Research and Applications*, 14, 112–125.

Ma J., Li X., Wang S., Zhang X. (2015). How does Google search affect trader positions and crude oil prices? *Economic Modelling*, 49, 162–171.

Nimark K. P., Pitschner S. (2019). News media and delegated information choice. *Journal of Economic Theory*, 181, 160–196.

Seabold S., Coppola A. (2015). Nowcasting prices using Google Trends. An application to Central America. *Policy Research Working Paper Series 7398*. The World Bank.

Scott S. L., Varian H. R. (2013). Bayesian variable selection for nowcasting economic time series. *NBER Working Papers*, 19567.

Smith J. P. (2012). Google Internet search activity and volatility prediction in the market for foreign currency. *Finance Research Letters*, 9, 103–110.

Song M., Shin K. (2019). Forecasting economic indicators using a consumer sentiment index: Survey-based versus text-based data. *Journal of Forecasting*, 38 (6), 504–518.

Thorsrud L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business and Economic Statistics*, 38 (2), 393–409.

Received 13.06.2020; accepted 01.08.2020.