

Прогнозирование текущей динамики ВВП на основе данных поисковых запросов

Самвел Сергеевич Лазарян (e-mail: lazaryan@nifi.ru), руководитель Центра макроэкономических исследований Научно-исследовательского финансового института (г. Москва)

Никита Евгеньевич Герман (e-mail: nichita.gherman@gmail.com), магистр 1-го курса МИЭФ (г. Москва)

Аннотация

Оперативная оценка динамики основных макроэкономических переменных, в том числе ВВП, является необходимым условием для проведения эффективной экономической политики. В настоящей работе обосновывается рассмотрение данных, отражающих частоту поисковых запросов, в качестве полезного предиктора текущей динамики ВВП наряду с данными из официальной статистики. Мы проверяем, действительно ли эти данные помогают увеличить качество прогнозных моделей текущего ВВП России в рамках эксперимента в реальном времени. Для этого строятся конкурирующие динамические факторные модели двух классов: включающие поисковые данные и нет. Было получено, что добавление в модель данных о частоте поисковых запросов почти не меняет прогнозы факторных моделей, построенных на основе официальной экономической статистики. В то же время обе модели показали лучшее качество прогнозов в сравнении с AR(1)-моделью. Наконец, нами была предпринята попытка объяснить, с чем связана полученная на данных нерелевантность поисковых запросов для прогнозирования ВВП. Была выделена как фундаментальная причина, так и методологические проблемы, которые могли привести к подобному результату.

Ключевые слова:

прогнозирование, факторные модели, ВВП, новкасты, поисковые запросы, частотность данных

JEL: C32, C53

Наличие оперативной информации о динамике основных макроэкономических показателей — необходимое условие для осуществления эффективной экономической политики. Тем не менее ввиду особенностей статистического учета таких показателей, как, например, динамика ВВП, данные выходят с существенным лагом. Более того, официальная статистика по ВВП представляется в квартальном виде, что также ограничивает возможности непрерывного мониторинга. Таким образом, вопрос оперативного получения качественных со статистической точки зрения оценок динамики основных макроэкономических показателей является актуальным для регуляторов [1–3]. В рамках текущего исследования будет изучен вопрос получения оценок ВВП за текущий квартал (новкастинг) на основе всей доступной информации, в т. ч. из неофициальных источников.

Поскольку ВВП является комплексным индикатором деловой активности, в прогнозные модели ВВП желательно включать широкий набор оперативных рядов, способных в совокупности уловить различные составляющие экономической динамики. Популярный в литературе подход для решения задачи такого ряда — использование динамических факторных моделей. С их помощью из большого объема данных выделяется несколько релевантных индикаторов, позволяющих получить качественные и оперативные новкасты ВВП в рамках формализованных эконометрических моделей.

Основным препятствием для построения качественных новкастов российского ВВП выступают ограничения, накладываемые особенностями официальной статистики. Так, статистика по большей части оперативных показателей публикуется в России с задержкой примерно в один месяц, снижая актуальность этих рядов на момент осуществления прогноза. Также публикуемые на оперативной основе данные подвержены существенным пересмотрам, в связи с чем они могут отражать неправильную информацию об экономической динамике на прогнозный период. Наличие этих ограничений мотивирует при построении модели обращаться к косвенным индикаторам, способным восполнить отсутствующую в статистике информацию.

В настоящей работе в качестве таких индикаторов предлагается рассмотреть данные о частоте поисковых запросов по определенным темам. Выделяется несколько аргументов в пользу рассмотрения этих данных на предмет их включения в прогнозные модели российского ВВП. Во-первых, они обновляются еженедельно, предоставляя более актуальную по сравнению с традиционной статистикой информацию об изменении нескольких ключевых составляющих экономического поведения домохозяйств: потребительского поведения и предложения труда. Во-вторых, в отличие от официальной статистики они не пересматриваются, а значит, им можно больше доверять при построении прогнозов. Наконец, простота сбора и предобработки этих данных позволяет без труда встроить их в используемые прогнозные модели. В рамках данного исследования были построены и оценены две факторные модели: использующие и не использующие поисковые данные наряду с официальной статистикой в прогнозной модели текущего ВВП.

Настоящая работа построена следующим образом. Сначала будет дан краткий обзор практики использования динамических факторных моделей для целей новкаста ВВП, а также работ, посвященных прогнозированию макропоказателей на основе данных поисковых запросов. Далее будут описаны предлагаемая методология построения новкастов и использованные данные, после чего будут представлены полученные результаты. Работа завершается обсуждением полученных результатов.

ОБЗОР ЛИТЕРАТУРЫ

В современной литературе новкастинг ВВП обычно осуществляется в рамках моделей, включающих в себя информацию о динамике большого числа различных оперативных показателей. Наиболее разработанным классом таких прогнозных моделей являются динамические факторные модели (далее — ДФМ). Они предполагают, что динамику рядов можно разложить на две независимые составляющие: общую, свойственную всем показателям, и идиосинкратическую, свойственную конкретному ряду. Данная методология позволяет обойти «проклятие размерности» больших прогнозных моделей за счет определения динамики общей компоненты-фактора и выявления ее связи с изучаемой переменной. Как показано в работах [5; 6], такие модели обладают лучшими прогностическими характеристиками ВВП в сравнении с моделями, построенными на небольшом числе оперативных индикаторов.

Методология ДФМ получила свое изначальное развитие в работах [6; 7] в рамках класса моделей, названных впоследствии точными (*strict*) факторными моделями. Существенный недостаток точных моделей — необходимость оптимизации функции правдоподобия, что является сложной с алгоритмической точки зрения задачей при наличии большого количества данных. Для улучшения вычислительных способностей модели в работах [8; 9] была представлена методологическая основа для приближенных (*approximate*) факторных моделей. В отличие от точных ДФМ приближенные модели используют различные вариации метода главных компонент для получения состоятельных оценок факторов. Подобный подход позволяет избежать «проклятия размерности», делая возможным включение в прогнозную модель большого массива данных.

Приближенные факторные модели нашли широкое применение, показав хорошие прогностические результаты на реальных данных. Так, в работе [10] ДФМ, построенная на 149 макроэкономических показателях, показала существенно лучшие прогнозы ВВП США в сравнении со стандартными авторегрессионными моделями. Похожие результаты на американских данных с использованием несколько отличающейся методологии были получены в работе [8]. Более полный обзор подходов к построению ДФМ можно найти в работе [5].

Для осуществления новкастов квартального ВВП в режиме реального времени чаще всего используются квартальные факторные модели [11]. В процедуре построения прогнозов можно выделить несколько этапов: на основе данных оперативных показателей оцениваются общие на момент прогнозной итерации факторы, которые затем экстраполируются с целью получения их ожидаемой динамики на оставшиеся периоды прогнозного квартала. Далее, применяя некоторый способ агрегирования данных, частотность факторов меняется на квартальную. Наконец, полученные квартальные значения факторов используются для построения прогнозов в рамках динамической линейной модели.

Основной методологической проблемой, возникающей при попытке построить прогноз в реальном времени, является наличие рваного края — ситуация, когда у показателей различается время последних наблюдений. Экстраполяция рядов в таком массиве данных требует особой методологии для эффективного использования информации, содержащейся в совместной динамике показателей. В литературе в основном используется два подхода: EM-алгоритм [10] и фильтр Калмана [4].

Квартальные факторные модели в целом демонстрируют высокое качество прогнозирования ВВП в реальном времени. Так, в работе [12] показано, что ДФМ выдает более точные прогнозы текущего ВВП по сравнению с AR-моделью на американских данных. В работе [13] ДФМ, построенная на данных еврозоны, показала лучшее качество прогнозов по сравнению с AR-моделью и связующими уравнениями. Аналогичные результаты были также получены в работе [14], в которой ДФМ была построена на основе данных Португалии.

Литература, посвященная прогнозированию ВВП на основе большого набора данных, развивалась в основном в методологическом направлении. При этом мало внимания уделялось расширению числа индикаторов, внедряемых в данные модели: на текущий момент исследователи ограничиваются рассмотрением в ДФМ только официальных экономических индикаторов. В то же время и другие данные, например данные о частоте определенных поисковых запросов, могут быть полезными при прогнозировании макроэкономических индикаторов.

Так, в работах [15–17] показано, что изменение частоты по таким Google-запросам, как «поиск работы», помогает оперативно отслеживать уровень безработицы как в развитых, так и в развивающихся странах. В работах [18; 19] авторы используют данные о частоте запросов, связанных с покупками различных товаров, чтобы показать, что они хорошо прогнозируют текущую динамику частного потребления в Германии и США. В целом в литературе отмечается, что данные интернет-запросов действительно помогают оперативно отслеживать динамику некоторых макроэкономических индикаторов.

ДАННЫЕ И МЕТОДОЛОГИЯ

Данные из официальной статистики

В работе [20] на российских данных были выделены оперативные экономические показатели, которые могут быть использованы для прогнозирования российского ВВП. Этот массив данных представляет собой адаптированный под реалии российской статистики, широко используемый в литературе набор индикаторов, предложенный в работе [10] для прогнозирования ВВП США. В настоящей работе используется схожий набор данных из

официальной статистики. Всего было собрано 78 оперативных индикаторов на месячной основе с января 2004 г. до декабря 2016 г., которые отражают различные составляющие российской экономической динамики. В качестве прогнозируемых переменных были взяты ряды реальных, сезонно скорректированных темпов квартального прироста ВВП и его компонент по элементам использования, предоставляемых Росстатом.

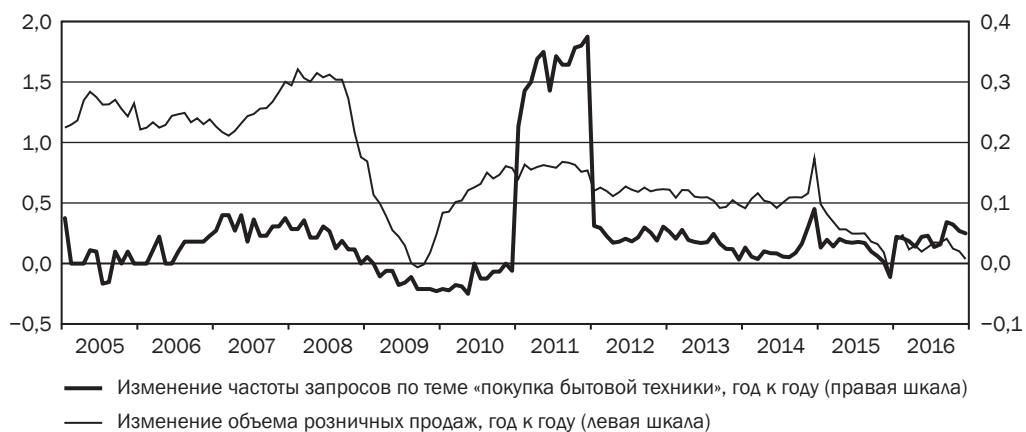
Данные о частоте поисковых запросов

Данные по большинству ключевых оперативных индикаторов поступают с лагом в один месяц. Настоящая работа предлагает заполнить отсутствующую релевантную информацию об изменении экономической динамики более оперативными данными о частоте поисковых запросов. Источником данных о частоте поисковых запросов стала информация из поисковой системы Google, представленная в приложении Google Trends. Это приложение позволяет отследить, как менялась частота запросов, осуществляемых в поисковой системе Google, с течением времени в определенной стране начиная с 2004 г., предоставляя наиболее длинные данные среди конкурирующих поисковых систем. Так, результаты более популярной в России поисковой системы «Яндекс» не были использованы в анализе ввиду коротких временных рядов. Частота запросов рассчитывается Google на еженедельной и ежемесячной основе как отношение числа запросов, содержащих заданные слова, к общему числу запросов, сделанных за эту неделю. С помощью машинных алгоритмов похожие запросы группируются Google Trends в темы.

В настоящей работе мы попытались выделить темы, изменение частоты запросов по которым может оперативнее официальной статистики отражать колебания в деловой активности. Так, по аналогии с рассмотренными выше исследованиями мы предполагаем, что увеличение частоты запросов по таким темам, как «рестораны», «продовольственные магазины», «покупка бытовой техники», «покупка одежды» и др., указывает на рост потребления домохозяйств [18; 19]. Для примера, на рис. 1 отражается связь между ежемесячной динамикой розничных продаж в России и частотой запросов на тему «покупка бытовой техники». На промежутке начиная с 2007 г. (кроме выбросов в 2011–2012 гг.) отчетливо прослеживается положительная корреляция между этими двумя запросами.

Рисунок 1

Связь объема розничных продаж и частоты запросов на тему «покупка бытовой техники»



Источники: Росстат, Google Trends.

В качестве оперативного индикатора динамики на рынке труда мы использовали данные о частоте запросов по теме «работа»: увеличение частоты запросов указывает на рост

безработицы, что видно из рис. 2. Здесь важно указать, что в то время как данные о частоте запросов обновляются еженедельно, данные по розничным продажам и уровню безработицы публикуются с лагом примерно в один месяц после отчетного периода. Всего нами были собраны данные по 27 таким темам.

Рисунок 2



Источники: Росстат, Google Trends.

Методология исследования

Для построения новкастов ВВП в режиме реального времени используется квартальная факторная модель, предложенная в работе [4]. Схожая методология была применена в том числе и на российских данных в работе [20], а также для построения опережающего индикатора РенКап-РЭШ. В рамках нашего исследования модель имеет следующий вид:

$$X_t = \Lambda F_t + \xi_t, \tag{1}$$

$$F_t = BF_{t-1} + Cu_t, \tag{2}$$

$$y_t = \mu + D(L)y_t + M(L)F_t^Q + z_t, \tag{3}$$

где X_t — матрица наблюдаемых оперативных показателей размера $N \times T_1$; N — количество показателей, T_1 — число периодов; F_t — матрица ненаблюдаемых факторов размерности $r \times T_1$, где $r \ll T_1$. В уравнении (2) предполагается, что F_t следует стационарному VAR(1) процессу; $\chi_t = \Lambda F_t$ — общая компонента, отражающая часть динамики ряда, общую для всех рядов из X_t ; ξ_t — идиосинкратическая компонента, отражающая часть динамики ряда из X_t , свойственную только ему; y_t — вектор сезонно скорректированных квартальных темпов прироста реального ВВП размером $T_2 \times 1$, где $T_2 < T_1$; F_t^Q — матрица ненаблюдаемых факторов квартальной частотности, полученная с помощью некоторой схемы агрегирования из F_t ; $D(L)$, $M(L)$ — лаговые операторы; ξ_t , z_t , u_t — стационарные гауссовские процессы ($\dim(u_t) = r$); Λ , B , C — матрицы коэффициентов.

Также вводятся следующие предположения: χ_t и ξ_t — два независимых ненаблюдаемых процесса; ξ_t^i слабо скоррелированы между собой и не обладают выраженной автокорреляцией; $\mathbb{E}[\xi_t u_{t-s}] = 0 \forall s$.

В данной модели есть несколько гиперпараметров, которые подбираются следующим образом. Количество ненаблюдаемых факторов r оценивается на данных с помощью информационного критерия ВИС аналогично работе [11]. Количество лагов ВВП в урав-

нении (3) — $D(L)$ — определяется исходя из результатов тестов на автокорреляцию модели (3): если автокорреляция остатков в модели без лагов y_t не будет обнаружена, положим $D(L) \equiv 0$; в противном случае в модель добавляются лаги ВВП, пока автокорреляция не исчезнет. Наконец, оператор $M(L)$ определяется исходя из информационного критерия BIC по аналогии с работой [11].

Разберем на примере, как в настоящей работе строились новкасты ВВП. Предположим, новкаст необходимо построить в момент времени j . За X^j и y^j примем известные на момент осуществления прогноза данные о динамике наблюдаемых оперативных показателей и ВВП соответственно. Выполним последовательно следующие шаги:

1) преобразование рядов из X^j , осуществляемое таким образом, чтобы итоговые ряды отражали темпы прироста в логарифмах соответствующих показателей по отношению к тому же месяцу предыдущего квартала;

2) детрендинг полученных рядов и их десезонализация с помощью процедуры X-13-ARIMA-SEATS;

3) проверка стационарности рядов. Результаты показали, что после выполнения предыдущих шагов ряды не нуждались в дополнительной трансформации для обеспечения стационарности;

4) стандартизация X^j и выделение первых r главных компонент на сбалансированной панели из X^j ;

5) применение метода главных компонент для параметризации уравнения (1) и получения предварительной оценки динамики факторов F_t ;

6) параметризация уравнения (2);

7) применение сглаживания Калмана в рамках оцененной модели (1)–(2) для получения окончательной оценки факторов F_t ;

8) усреднение месячных значений F_t за соответствующий квартал для получения ряда F_t^Q . Подобная агрегация факторов совместно с трансформацией, примененной на шаге 1), хорошо зарекомендовала себя на данных [14];

9) оценивание уравнения (3);

10) применение фильтра Калмана к оцененной на шаге 6 модели (1)–(2) для прогнозирования динамики факторов на текущий квартал и использование ее для получения новкастов ВВП из оцененного уравнения (3).

Методологические особенности прогнозирования на основе данных о частоте поисковых запросов

Описанная выше методология хорошо зарекомендовала себя в литературе на данных из официальной статистики. Однако особенности данных Google Trends требуют осторожности при попытке включить их в модель наряду с официальными экономическими рядами. Так, предположение о том, что общая компонента в динамике поисковых данных совпадает с таковой рядов из официальной статистики, кажется нереалистичным. Более правильным будет разделить динамику каждого запроса на три независимые компоненты: общую для данных официальной экономической статистики и поисковых запросов; свойственную только данным поисковых запросов; идиосинкратическую. Очевидно, для прогнозирования ВВП нам интересна именно общая с экономическими рядами компонента. Мы выделяем ее следующим образом. После предобработки данных поисковых запросов на шаге 1) и 2) полученные ряды очищаются от общей компоненты, выделенной на рядах поисковых запросов по макротемам, мало отражающим динамику деловой активности: «интернет и телекоммуникации», «книги и литература», «красота и фитнес», «спорт», «здравоохранение», «искусство и развлечения» и «люди и общество». Полученные ряды затем включаются в модель (1)–(3) наряду с данными из официальной статистики в X^j для осуществления шагов 3–10. Поскольку ряды по вышеперечисленным темам,

вероятнее всего, не отражают колебания в деловой активности, но при этом включают в себя общую для данных поисковых запросов динамику, в ходе данной процедуры мы убираем из исходных рядов компоненту, нерелевантную для отслеживания экономической динамики. Так, например, подобным образом мы очищаем релевантные ряды от существенного сдвига в уровнях, произошедшего вслед за изменением методологии сбора данных в январе 2011 г.

Процедура построения прогнозов

Для того чтобы выяснить, помогает ли информация о частоте поисковых запросов лучше предсказывать текущую экономическую динамику, нами сравнивалась точность прогнозирования изменения текущего ВВП двух конкурирующих ДФМ, отличающихся наборами данных. Назовем их модель 1 и модель 2: модель 1 построена только на основе данных из официальной статистики, модель 2 использует помимо них данные о частоте поисковых запросов. Обе модели применяют методологию, описанную выше.

Сравнение качества прогнозных моделей ВВП осуществлялось нами на основании прогнозов, построенных моделями в режиме реального времени с расширяющимся временным окном на отложенной выборке с января 2013 г. по декабрь 2016 г. Прогноз темпа роста ВВП за текущий квартал строился раз в месяц сразу после выхода доклада Росстата «Социально-экономическое положение России», отражающего информацию об изменении основных экономических показателей за предыдущий месяц.

Остановимся подробнее на процедуре прогнозирования на отложенной выборке. Эксперимент в реальном времени¹ начинается с построения в конце января 2013 г. прогноза ВВП за I кв. 2013 г. Примерно в это время выходит доклад Росстата «Социально-экономическое положение России» за декабрь 2012 г. В рамках предложенной методологии выделяется информация ($X^{2013:1}$, $y^{2013:1}$), где $X^{2013:1}$ и $y^{2013:1}$ — наиболее актуальные данные о динамике оперативных индикаторов и ВВП на январь 2013-го. Важно отметить, что панель $X^{2013:1}$ является несбалансированной. Так, последние данные о розничных продажах в России доступны лишь на предварительной основе за декабрь 2012 г. При этом данные о частоте поисковых запросов доступны также и за январь 2013 г. Что касается $y^{2013:1}$, последние данные о квартальном ВВП доступны лишь за III кв. 2012 г. В связи с этим для осуществления прогноза ВВП за I кв. 2013 г. используется рекурсивный подход: сначала осуществляется прогнозирование темпов роста ВВП за IV кв. 2012 г., а затем на его основе прогнозируется ВВП за I кв. 2013 г.

Переходим к построению прогноза ВВП в следующем месяце. Выборка располагаемых данных ($X^{2013:2}$, $y^{2013:2}$) теперь включает больше актуальной информации о деловой активности за текущий квартал: так, становится доступной предварительная статистика промышленного сектора за январь 2013 г., а также ряд показателей за февраль 2013 г. Кроме того, примерно в середине февраля выходят предварительные данные о темпах роста ВВП за IV кв. 2012 г. Ожидается, что новая информация поможет построить в феврале более точный прогноз, чем в январе.

Наконец, на основе данных, располагаемых к концу марта 2013 г., строится последний новкаст ВВП, который должен увеличить точность в свою очередь февральских прогнозов. Схожим образом нами строятся прогнозы для последующих кварталов вплоть до IV кв. 2016 г.

¹ Важно отметить, что поскольку мы не располагаем фактическими данными, доступными на прогнозный момент, а лишь их исторической ретроспективой, наш эксперимент нельзя считать в полной мере осуществляемым в реальном времени. Впрочем, в научной литературе использованный нами подход к прогнозированию широко распространен ввиду отсутствия релевантной информации. Он получил особое название: прогнозирование в псевдореальном времени (*pseudo real-time forecasting*).

РЕЗУЛЬТАТЫ

На основе динамики 78 рядов, представленных в официальной статистике, и 27 показателей частоты поисковых запросов нами были параметризованы модели 1 и 2 и построены прогнозы квартальных темпов прироста текущего ВВП в режиме псевдореального времени. Качество прогнозов ДФМ было также сопоставлено с точностью новкастов, выдаваемых AR(1)-моделью. Использование AR(1)-моделей является популярным бенчмарком для оценки качества полученных моделей — если модель работает лучше, чем AR(1), то ее дальнейшее развитие является перспективным.

В связи с различиями в фактической динамике ВВП внутри прогнозного периода было произведено разделение псевдореальной выборки на две подвыборки: 01.2013–12.2014 и 01.2015–12.2016. Так, первая подвыборка характеризуется в большинстве положительными, но замедляющимися темпами роста, а вторая — кризисными явлениями в экономике.

Спецификация гиперпараметров уравнений (1)–(3) осуществлялась отдельно для каждой подвыборки. Для обеих моделей и прогнозных выборок критерий BIC и тесты на автокорреляцию отдали предпочтение двухфакторной модели с одним лагом ВВП, или, в рамках предложенных обозначений: $r = 2$; $D(L) = L$; $M(L) \equiv 1$. Однако для проверки на устойчивость нами было также исследовано качество прогнозных моделей с тремя и четырьмя факторами, а также различные спецификации $D(L)$ и $M(L)$. Помимо этого, в уравнение (3) для модели 2 была добавлена дамми-переменная на январь 2011 г. для того, чтобы уловить последствия изменения метода сбора данных Google Trends, которые могла не уловить обозначенная выше процедура выделения из них общей неэкономической составляющей.

Табл. 1 и 2 отражают основные результаты настоящего исследования. В них представлены отношения RMSE и MAE прогнозов квартальных темпов роста сезонно скорректированного реального ВВП ДФМ к прогнозам AR(1)-моделей. Таким образом, значения меньше единицы указывают на более высокую точность прогнозов ДФМ в сравнении с AR(1)-моделями, а значения больше единицы — наоборот, меньшую. Для удобства значения меньше единицы выделены жирным шрифтом. Также в таблицах представлены соответствующие значения метрик качества AR(1)-моделей в абсолютных величинах. Прогнозные характеристики конкурирующих моделей дезагрегированы по двум направлениям. Во-первых, метрики качества посчитаны отдельно на всей выборке и отдельно для двух обозначенных ранее подвыборок. Во-вторых, внутри каждой подвыборки качество прогнозных моделей сравнивается для новкастов ВВП, осуществленных в первый, второй и третий месяц квартала.

Таблица 1

Относительные RMSE прогнозов моделей в зависимости от периода и номера месяца квартала, в котором строился прогноз

Тип модели	01.2013–12.2014				01.2015–12.2016				01.2013–12.2016			
	1	2	3	1–3	1	2	3	1–3	1	2	3	1–3
Модель 1, $r = 2^*$	1,11	0,85	0,85	0,93	1,07	0,96	0,93	0,99	1,09	0,91	0,90	0,96
Модель 1, $r = 3$	1,11	0,85	0,85	0,93	1,09	0,94	0,98	1,00	1,10	0,90	0,92	0,97
Модель 2, $r = 2^*$	1,16	0,86	0,85	0,95	1,11	0,96	0,94	1,00	1,13	0,92	0,90	0,98
Модель 2, $r = 3$	1,13	0,86	0,83	0,93	1,18	0,94	1,00	1,04	1,16	0,91	0,93	0,99
AR(1)	0,61	0,70	0,70	0,67	0,72	0,77	0,77	0,75	0,67	0,73	0,73	0,71

Примечания: звездочкой отмечены спецификации моделей, которым было отдано предпочтение на основании критерия BIC. Значения 1, 2, 3 и 1–3 в названии столбцов соответствуют номерам месяцев квартала, в которых строились прогнозы. Ошибки AR(1)-модели умножены на 100.

Источник: составлено авторами.

Таблица 2

Относительные MAE прогнозов моделей в зависимости от периода и номера месяца квартала, в котором строился прогноз

Тип модели	01.2013–12.2014				01.2015–12.2016				01.2013–12.2016			
	1	2	3	1–3	1	2	3	1–3	1	2	3	1–3
Модель 1, $r = 2^*$	1,08	0,66	0,76	0,82	1,12	0,96	0,97	1,02	1,10	0,82	0,87	0,93
Модель 1, $r = 3$	1,11	0,69	0,83	0,86	1,13	0,99	1,02	1,05	1,12	0,85	0,93	0,97
Модель 2, $r = 2^*$	1,15	0,70	0,73	0,84	1,16	0,96	1,00	1,04	1,16	0,84	0,87	0,95
Модель 2, $r = 3$	1,14	0,69	0,75	0,84	1,25	0,96	1,04	1,08	1,20	0,83	0,91	0,98
AR(1)	0,43	0,51	0,51	0,48	0,59	0,60	0,60	0,60	0,51	0,55	0,55	0,54

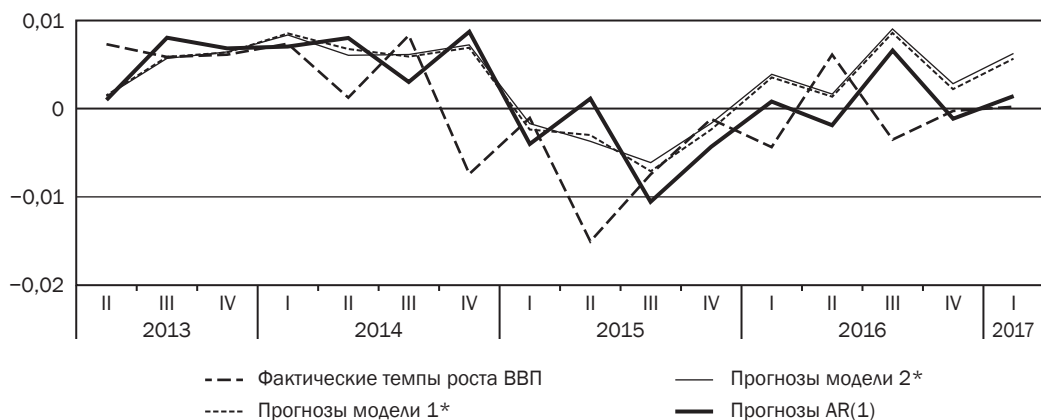
Примечания: звездочкой отмечены спецификации моделей, которым было отдано предпочтение на основании критерия BIC. Значения 1, 2, 3 и 1–3 в названии столбцов соответствуют номерам месяцев квартала, в которых строились прогнозы. Ошибки AR(1)-модели умножены на 100.

Источник: составлено авторами.

Как видно из обеих таблиц, модели 1 и 2 показывают очень схожее качество прогнозов: значения их относительных метрик качества очень близки. На рис. 3 отображены фактические квартальные темпы роста ВВП за 01.2013–12.2016, а также его динамика, спрогнозированная двухфакторными моделями 1 и 2, а также AR(1).

Рисунок 3

Сравнение фактической и спрогнозированной в третьем месяце квартала динамики текущих квартальных темпов роста реального ВВП



Источник: составлено авторами.

Заметим, что на всей отложенной выборке модели 1 и 2 выдают очень похожие прогнозы в третьем месяце квартала. Аналогичный вывод верен и для прогнозов, осуществленных в первом и втором месяце квартала, а также для трехфакторной модели. Таким образом, в рамках предложенной нами методологии и данных добавление в модель информации о частоте поисковых запросов не увеличивает точность прогнозов текущего ВВП.

Остановимся также на сравнении качества прогнозов ДФМ и AR(1)-модели. В целом на подвыборке 01.2012–12.2014 ДФМ показывает более высокую точность прогнозирования в терминах как RMSE, так и MAE. В свою очередь, на подвыборке за 1.2015–12.2016 ДФМ не выигрывает в качестве у AR(1)-модели. Вероятнее всего, факторная модель оказалась не способной из имеющейся официальной статистики по оперативным индикаторам выделить фундаментальные общие факторы, которые движут экономикой как в период кризиса, так и в моменты подъема.

Что касается различий в точности прогнозов, осуществляемых в различные месяцы кварталов, заметим, что лучшей прогнозной моделью для первого месяца оказалась AR(1). Впрочем, данный результат не выглядит неожиданным, поскольку, как уже отмечалось, большая часть оперативных индикаторов доступна за предыдущий месяц, то есть в данном случае за последний месяц предыдущего квартала. Отсутствие релевантной квартальной статистики и необходимость прогнозировать динамику факторов на три месяца вперед в рамках методологии ДФМ определяют ее слабые прогнозные характеристики. Относительно второго месяца прогнозы ДФМ становятся точнее, чем прогнозы AR(1)-модели. Однако в третьем месяце качество прогнозов ДФМ не изменяется или немного снижается в сравнении с прогнозами, осуществленными в конце второго месяца.

Нерелевантность данных поисковых запросов для прогнозирования текущего ВВП в рамках моделей, включающих традиционную экономическую статистику, является несколько неожиданным результатом с учетом отмеченных в литературе успехов, достигнутых с их помощью в прогнозировании макроиндикаторов и отмеченной неактуальностью ряда показателей из официальной статистики на момент построения прогнозов.

Данные интернет-запросов неплохо связаны с отдельными составляющими деловой активности, однако не способны уловить общую экономическую динамику, моделируемую факторами. Это может объясняться тем, что те составляющие деловой активности, которые мы собрались дополнять информацией из поисковых запросов, на самом деле не определяют совместную динамику экономических рядов (факторов). Однако собранная статистика по частоте запросов гораздо сильнее смещена в сторону прогнозирования потребительского поведения домохозяйств, которое, как оказалось, вносит достаточно ощутимый вес в итоговую главную компоненту: 13-е место для ряда «оборот розничной торговли» и 16-е — для ряда «оборот оптовой торговли». При этом даже несмотря на то что, например, показатель «уровень безработицы» почти не определяет динамику факторов, если бы какие-то ряды поисковых запросов хорошо улавливали именно общую компоненту данного показателя, их R2 оказался бы все равно высоким. Однако этого мы не наблюдаем: скорее поисковые данные лучше отражают именно идиосинкратическую компоненту экономической динамики. Таким образом, мы считаем следующее объяснение полученных в настоящей работе результатов основным: данные поисковых запросов улавливают в основном идиосинкратическую компоненту динамики рядов, а не общую.

Стоит отметить, что такое объяснение получившихся результатов не является единственным. Так, отметим слабость предложенной нами методологии вычленения общей компоненты, не связанной с экономической динамикой, из поисковых данных. Возможно, явное моделирование этой компоненты в рамках модифицированной методологии помогло бы лучше оценить экономическую составляющую рядов поисковых запросов. Кроме того, в силу того что Google Trends представляет данные о частоте запросов в целочисленном формате, идиосинкратическая компонента этих рядов подвержена явной гетероскедастичности. Если представить, например, что ненаблюдаемый индекс частоты для некоторого запроса равен 1,3 в некоторый момент времени, а затем снижается в следующем периоде до 0,8, данные Google Trends не отразят это изменение: целочисленный индекс не изменится и останется равным единице, хотя относительное изменение в частоте запросов по данному показателю на самом деле было значительным. Воздействие данной гетероскедастичности можно победить, например, взвешивая данные в рамках метода главных компонент. Подобный подход показал хорошие результаты даже на рядах, представленных в официальной статистике [21].

ЗАКЛЮЧЕНИЕ

В настоящей работе построена прогнозная модель текущего ВВП России с включением в нее в дополнение к макроэкономическим данным данных о частоте поисковых запросов.

Опираясь на соответствующую литературу и на особенности российской статистики, мы объясняем, почему эти данные могут помочь в прогнозировании.

Нами были построены две ДФМ: включающие и не включающие данные о частоте поисковых запросов наряду с данными из официальных источников. Сравнение их качества на отложенной выборке на временном промежутке 01.2013–12.2016 показало почти полную идентичность прогнозов, построенных на основе обеих моделей. Вместе с тем обе ДФМ показали более высокую точность прогнозирования текущего ВВП в сравнении с AR(1)-моделью.

Наконец, мы попытались объяснить полученный результат. Нами была выделена следующая вероятная причина нерелевантности поисковых данных для прогнозирования ВВП: несмотря на то что данные неплохо объясняют динамику отдельных составляющих экономической динамики, они скорее связаны именно с их идиосинкратической составляющей. Впрочем, нами также были отмечены некоторые недостатки предложенного нами способа моделирования данных поисковых запросов, которые могли снизить прогнозные характеристики этих данных. Возможно, усовершенствование предложенной в настоящей работе методологии поможет в дальнейшем более эффективно встраивать данные поисковых систем в существующие прогнозные модели ВВП.

Библиография / References

1. Pleskachev Y., Ponomarev Y. Short-Term GDP Forecasting: How Can Its Quality Be Improved (December 26, 2017). *Monitoring of Russia's Economic Outlook*, 2017, no. 22, pp. 18–20. Available at: <http://dx.doi.org/10.2139/ssrn.3093226>.
2. Варьяш И. Ю., Швандар К. В., Бурова Т. Ф. Актуализация макроэкономических прогнозов в бюджетном процессе // Научно-исследовательский финансовый институт. Финансовый журнал. 2014. № 4. С. 118–128 [Varjas I., Shvandar K., Burova T. Actualization of Macroeconomic Forecasts for Aim of Fiscal Process. *Finansovyy zhurnal – Financial Journal*, 2014, no. 4, pp. 118–128 (In Russ.)].
3. Шахова Г. Я. Современные тенденции в методологии бюджетного прогнозирования // Научно-исследовательский финансовый институт. Финансовый журнал. 2013. № 4. С. 61–70 [Shakhova G. Modern Trends in the Methodology of Budget Forecasting. *Finansovyy zhurnal – Financial Journal*, 2013, no. 4, pp. 61–70 (In Russ.)].
4. Doz C., Giannone D., Reichlin L. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 2011, vol. 164, iss. 1, pp. 188–205. Available at: <https://doi.org/10.1016/j.jeconom.2011.02.012>.
5. Eickmeier S., Ziegler C. How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach. *Journal of Forecasting*, 2008, vol. 27, iss. 3, pp. 237–265. Available at: <https://doi.org/10.1002/for.1056>.
6. Geweke J. F., Singleton K. J. Maximum Likelihood “Confirmatory” Factor Analysis of Economic Time Series. *International Economic Review*, 1981, vol. 22, no. 1, pp. 37–54. Available at: <https://doi.org/10.2307/2526134>.
7. Watson M. W., Engle R. F. Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 1983, vol. 23, no. 3, pp. 385–400. Available at: [https://doi.org/10.1016/0304-4076\(83\)90066-0](https://doi.org/10.1016/0304-4076(83)90066-0).
8. Forni M. et al. The Generalized Dynamic-Factor Model: Identification and Estimation. *Review of Economics and Statistics*, 2000, vol. 82, no. 4, pp. 540–554. Available at: <https://doi.org/10.1162/003465300559037>.
9. Stock J. H., Watson M. W. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 2002, vol. 97, iss. 460, pp. 1167–1179. Available at: <https://doi.org/10.1198/016214502388618960>.
10. Stock J. H., Watson M. W. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, 2002, vol. 20, iss. 2, pp. 147–162. Available at: <https://doi.org/10.1198/073500102317351921>.
11. Forni C., Marcellino M. A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. *International Journal of Forecasting*, 2014, vol. 30, iss. 3, pp. 554–568. Available at: <https://doi.org/10.1016/j.ijforecast.2013.01.010>.
12. Giannone D., Reichlin L., Small D. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 2008, vol. 55, iss. 4, pp. 665–676. Available at: <https://doi.org/10.1016/j.jmoneco.2008.05.010>.
13. Angelini E. et al. Short-term forecasts of euro area GDP growth. *The Econometrics Journal*, 2011, vol. 14, iss. 1, pp. 25–44. Available at: <https://doi.org/10.1111/j.1368-423X.2010.00328.x>.

14. Dias F., Pinheiro M., Rua A. Forecasting Portuguese GDP with factor models: pre-and post-crisis evidence. *Economic Modelling*, 2015, vol. 44, pp. 266–272. Available at: <https://doi.org/10.1016/j.econmod.2014.10.034>.
15. Askitas N., Zimmermann K. F. Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 2009, vol. 55, iss. 2, pp. 107–120. Available at: <https://doi.org/10.3790/aeq.55.2.107>.
16. D'Amuri F., Marcucci J. "Google it!" Forecasting the US Unemployment Rate with a Google Job Search Index. FEEM Working Paper No. 31.2010. Available at: <http://dx.doi.org/10.2139/ssrn.1594132>.
17. Pavlicek J., Kristoufek L. Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. *PLoS one*. 2015, vol. 10, no. 5, pp. 1–11.
18. Vosen S., Schmidt T. Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, vol. 30, iss. 6, pp. 565–578. Available at: <https://doi.org/10.1002/for.1213>.
19. Vosen S., Schmidt T. A monthly consumption indicator for Germany based on Internet search query data. *Applied Economics Letters*, 2012, vol. 19, iss. 7, pp. 683–687. Available at: <https://doi.org/10.1080/13504851.2011.595673>.
20. Поршаков А., Дерюгина Е., Пономаренко А., Синяков А. Краткосрочное оценивание и прогнозирование ВВП России с помощью динамической факторной модели / Банк России. Серия докладов об экономических исследованиях. 2015. № 2 [Porshakov A., Deryugina E., Ponomarenko A., Sinyakov A. Short-Term Estimation and Forecasting of Russia's GDP Using a Dynamic Factor Model. Bank of Russia. Economic Research Series, 2015, no. 2 (In Russ.)].
21. Boivin J., Ng S. Are more data always better for factor analysis? *Journal of Econometrics*, 2006, vol. 132, iss. 1, pp. 169–194. Available at: <https://doi.org/10.1016/j.jeconom.2005.01.027>.

S. S. Lazaryan, N. E. German

Forecasting Current GDP Dynamics With Google Search Data

Authors' affiliation:

Samvel S. Lazaryan (e-mail: lazaryan@nifi.ru), ORCID 0000-0001-6829-3979, Financial Research Institute, Moscow 127006, Russian Federation

Nikita E. German (e-mail: nichita.gherman@gmail.com), ORCID 0000-0001-9852-6032, National Research University Higher School of Economics, Moscow 101000, Russian Federation

Abstract

In order to conduct a conscious economic policy, timely assessment of the main economic indicators, viz GDP, is needed. In this paper the authors justify the reason why an inclusion of query search data may help to produce better nowcasts compared to the currently used Russian GDP models, which are built on the data from official statistical services. The authors also check, whether this hypothesis holds true in the real-time forecasting experiment. For this purpose the authors suggest two competing dynamic factor models: the one, which includes data on the query search frequency, and the other one, which excludes it. The models show that the inclusion of query search data does not change the forecast performance of the model built only upon official economic indicators. At the same time, both models have produced more accurate nowcasts of Russian GDP than AR(1) model did. Finally, the authors try to explain the resulting irrelevance of query search data in nowcasting GDP. The article discusses both fundamental reasons and the pitfalls of the methodology used in this paper, which could have led to such result.

Keywords:

forecasting, factor models, GDP, nowcasting, search queries, data frequency

JEL: C32, C53

<https://doi.org/10.31107/2075-1990-2018-6-83-94>