ЭКОНОМИЧЕСКАЯ ТЕОРИЯ

Г. С. Куровский[1]
Центральный банк Российской Федерации (Москва, Россия)

ИСПОЛЬЗОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ ДЛЯ ПРОГНОЗИРОВАНИЯ В МАКРОЭКОНОМИКЕ

В работе показано, как текстовая информация может использоваться для прогнозирования в макроэкономике. Рассматривается частный случай прогнозирования — наукастинг, или прогнозирование настоящего на примере безработицы. Суть наукастинга заключается в том, что прогноз строится на период, который уже прошел, но по которому еще не вышли статистические данные. В качестве текстовой информации выступают интернет-запросы. Работа является новой сразу по нескольким направлениям. Впервые в литературе для прогнозирования используется информация сразу двух поисковых систем — Яндекса и Google. Информация, предоставляемая поисковыми системами, дополняет друг друга и позволяет отбирать подходящие слова и словосочетания среди всех интернет-запросов пользователей. Впервые учитывается популярность онлайн-систем как источников информации о наличии рабочих мест. В России популярность интернета как источника информации о доступности рабочих мест выросла более чем в три раза с 2008 по 2018 г. Напрямую использование текстовой информации может отражать переключение населения на другие источники поиска информации, а не изменения в безработице. Большинство предложенных слов и словосочетаний показали значимое улучшение при прогнозировании безработицы на период от 1 до 6 месяцев вперед. Также в работе предлагается методика использования интернет-запросов для прогнозирования в макроэкономике.

Ключевые слова: прогнозирование, наукастинг, счетчик слов Яндекса, Google Trends.

Цитировать статью: *Куровский Г. С.* Использование текстовой информации для прогнозирования в макроэкономике // Вестник Московского университета. Серия 6. Экономика. — 2019. — № 6. — С. 39–58.

G. S. Kurovskiy
Bank of Russia (Moscow, Russia)

USING TEXTUAL INFORMATION TO PREDICT IN MACROECONOMICS

The paper shows how textual information can be used to predict and study cause-effect relationships in macroeconomics. I consider a special case of forecasting - nowcasting

[1] Куровский Глеб Станиславович, экономист, Центральный Банк Российской Федерации, e-mail: gleb.kurovskiy@gmail.com

on the example of unemployment. The key feature of nowcasting is that the forecast is built for a period that has already passed, but which has not yet come out statistics. As textual information, Internet requests are used. The paper is new in several direction. For the first time in the literature, information from two search engines, Yandex and Google, is used at once for forecasting. Information provided by search engines complements each other and allows performing suitable words' selection from the bunch of users' internet-requests. For the first time, the popularity of online systems as sources of information on job availability is taken into account. In Russia, the popularity of the Internet as a source of information on the availability of jobs has more than tripled from 2008 to 2018. If the researcher uses only the dynamics of related internet-requests then the results will show the dynamics of internet-services' popularity rather than unemployment. Most of the models with internet query words show significant quality improvement in fore(now)casting unemployment. The paper proposes the procedure how to use query data for macroeconomic nowcasting.

Key words: forecasting, nowcasting, Yandex wordstat, Google trends.

To cite this document: *Kurovskiy G. S.* (2019). Using textual information to predict in macroeconomics. Moscow University Economic Bulletin, (6), 39–58.

Введение

Поиск все новых и новых источников информации для прогнозирования и изучения причинно-следственных связей в макроэкономике является одним из ключевых направлений ее развития. Данные, публикуемые в официальных изданиях, зачастую выходят с большой задержкой, обладают редкой периодичностью и предоставляются в агрегированном виде. Одним из устоявшихся подходов является использование микроданных по населению, предприятиям. Такие данные обладают невысокой периодичностью, зачастую ежегодной, выходят с большой задержкой, но уже предоставляют детализированную статистику по населению и предприятиям. Другим подходом, набирающим популярность, является использование текстовой информации. Под текстовой информацией может пониматься большой класс данных: информация о пользователях социальных сетей, документы и пресс-релизы компаний, новости, интернет-запросы пользователей. Именно последний класс информации используется в настоящей работе. Интернет-запросы обладают преимуществом перед документами, пресс-релизами и новостями, поскольку отражают массовое поведение населения. Социальные сети также отражают поведение населения, однако такие данные являются трудоемкими для сбора. Периодичность, запаздывание и детализация интернет-запросов зависят от технических возможностей поисковых систем, предоставляющих эти данные. Данные о запросах в поисковых системах Яндекса и Google доступны на ежедневной основе с небольшим запаздыванием и практически по любым словам и словосочетаниям русского языка[1]. В работе по-

[1] Исключением являются слова и словосочетания, которые редко используются пользователями интернет-поисковиков. Данные по таким словам и словосочетаниям не ото-

казывается, как поисковые запросы могут быть использованы для наукастинга безработицы.

Одним из расширений применяемых в литературе подходов к прогнозированию показателей с помощью запросов является использование информации сразу из двух поисковых систем — Яндекса и Google. Данные, предоставляемые Google, обладают более длинной историей, но раскрываются в относительном значении[1]. То есть невозможно узнать, сколько «абсолютных запросов»[2] совершили пользователи. Данные Яндекса представлены за более короткий период, но в абсолютном выражении. Более того, можно узнать не только абсолютное количество запросов, но количество словосочетаний, используемых с данным словом. Совместное использование информации из двух поисковых систем позволит проводить отбор слов. Другим расширением, особенно актуальным для России, является попытка учесть меняющуюся степень проникновения интернета в жизнь населения. В России популярность интернета как источника информации о доступности рабочих мест выросла более чем в три раза с 2008 по 2018 г. В исследовании предложен метод, который позволяет учитывать информацию о способах поиска работы населением.

Использование поисковых запросов Google в качестве опережающих индикаторов макроэкономических показателей берет начало в работе [H. Varian, 2012], где автор показал, что данные поисковой системы помогают в прогнозировании автомобильных продаж. Основными критериями качества прогноза выступают среднеквадратичные (RMSE[3]) и средние абсолютные (MAE[4]) ошибки. На основе критериев проводится сравнение прогнозной силы базовой модели и модели, включающей опережающие индикаторы в качестве факторов. В работах [Amari, 2017; Parlicek, 2015] в качестве базовой прогнозной модели использовалась авторегрессионная модель. Ошибки прогноза рассчитывались для базовой модели и для модели, содержащей в себе данные поисковых запросов. После чего на основе сравнения ошибок прогноза делался вывод о качестве моделей. Дополнительный критерий качества прогноза в рамках

бражаются поисковиками, поскольку они считают их непопулярными, при этом критерий популярности ими не раскрывается.

[1] Рассчитывается доля запросов по слову в общем числе интернет-запросов за рассматриваемый период, после чего динамический ряд нормируется к максимальному значению самого ряда.

[2] Количество интернет-запросов по словам и словосочетаниям.

[3] $RMSE = \frac{1}{n}\sum (y_j - \hat{y}_j)^2 \hat{y}_j$ — прогнозное значение по модели, y_j — фактическое значение по модели, прогнозируемая переменная.

[4] $MAE = \frac{1}{n}\sum | y_j - \hat{y}_j | \hat{y}_j$ — прогнозное значение по модели, y_j — фактическое значение по модели, прогнозируемая переменная.

опережающих индикаторов был использован в работе [Kholodin, 2010]. Автор применял тест [Deibold—Mariano, 1995][1]. Счет слов Яндекса еще не использовался для задач наукастинга. Однако в работе [Verzilin D. et al., 2017] на основе данных Яндекса изучалась региональная деловая активность в России.

В качестве опережающего индикатора выбран показатель безработицы как наиболее распространенный в литературе. Наличие исследований по другим странам с его использованием позволит сопоставить результаты и сделать выводы о возможностях использования интернет-запросов для прогнозирования в макроэкономике. Большинство существующих исследований проведены по развитым странам, тогда как страны с переходной экономикой остаются за рамками обсуждения. Данная работа призвана провести исследование по России и учесть особенности, которые могут касаться и других стран, например, популярность источников информации о поиске работы. В развитых странах источники информации уже могли сформироваться и оставаться неизменными на протяжении длительного периода времени. В России источники информации по поиску работы существенно менялись на протяжении 2008–2018 гг.

Суть прогнозирования настоящего, или наукастинга, заключается в том, чтобы до выхода официальной статистики получить значения по показателям на основе альтернативных источников. Количество запросов, связанных с безработицей, является хорошим примером, поскольку сегодняшние поиски работы в интернете отражают людей, которые скорее всего являются безработными. При этом данные по запросам выходят с запаздыванием в несколько дней. В табл. 1 приведено сравнение подходов к моделированию безработицы на основе данных поисковых запросов. Работа [Foudeur, 2013] изучает безработицу в конкретной возрастной категории — молодежная безработица. Во всех работах использование поисковых запросов улучшает качество прогноза.

Таблица 1

Сравнение подходов к моделированию безработицы с помощью Google trends

Источники	Базовая модель прогноза для сравнения с Google-трендами	Меры качества прогноза	Набор слов в исследованиях	Страны
Amuri Marcarri, 2017	Авторегрессия	RMSE и Diebold—Mariano тест	Jobs за исключением Steve Jobs	США

[1] Тест Diebold—Mariano, модификация [Harvey, Leybourne, Newbold, 1997]: сравнивается прогнозная сила двух моделей. Для сравнения рассчитывается ряд остатков двух моделей $\hat{e}_{1j}$ и $\hat{e}_{2j}$. Нулевая гипотеза состоит в том, что их прогнозная сила совпадет.

Окончание табл. 1

Источники	**Базовая модель прогноза для сравнения с Google-трендами**	**Меры качества прогноза**	**Набор слов в исследованиях**	**Страны**
Pavlicek Kristoufek, 2015	Авторегрессия	Diebold—Mariano тест	Job or work эквивалент в Чехии, Венгрии, Польше, Словакии	Чехия, Венгрия, Польша, Словакия
Foudeur, 2013	Базовая модель для сравнения отсутствует	RMSE и MAE	Emploi во Франции	Франция
Askitas Zimmerman, 2009	Базовая модель для сравнения отсутствует		Четыре группы слов: unemployment office or agency; unemployment rate, personnel consultant, job related cites	Германия

Источник: составлено автором.

Большинство работ используют данные по макроэкономическим показателям и поисковым запросам с единой периодичностью. Однако существует ряд работ [Fondeur, 2013; Giannone, 2008], которые используют данные разной периодичности. Одним примером является использование поисковых данных о запросах на еженедельной основе для прогнозирования ежемесячной безработицы. Основная идея заключается в том, что внутримесячные данные уже несут в себе значимую для прогноза информацию о ежемесячном показателе. В результате чего возможно дать оценку показателя, не дожидаясь завершения месяца. В работе исследуется взаимосвязь только показателей с одинаковой периодичностью, вопрос об информативности более частых поисковых запросов по сравнению с макроэкономическими показателями остается за рамками обсуждения. Счетчики слов также используются и для прогнозирования других экономических показателей. В работе [Preis et al., 2013] Google Trends позволили получить индикаторы раннего предупреждения финансового кризиса. Одним из исследований по России, которое использует Google интернет-запросы, является работа [Лазарян, Герман, 2018], где на основе факторной модели показано, что запросы Google trends улучшают качество прогноза ВВП по сравнению с авторегрессионными моделями.

Альтернативным подходом к использованию Google Trends и счетчика слов Яндекса могло бы являться использование подхода tf-idf (term frequency — inverse document frequency). В целом подход предназначен для выделения существенных слов из массива текста. Часть tf отвечает за частоту повторений слов и была предложена [HP Luhn, 1957], часть

idf отвечает за уникальность слова в тексте, т.е., например, некоторый термин может однозначно определять суть текста. Часть idf была предложена в работе [Jones, 2004]. Метод tf-idf используется в том числе для изучения тональности текста. Более подробное описание применения метода tf-idf представлено в работе [Gentzkow et al., 2019]. В нашем случае мы знаем ключевые слова, определяющие безработицу, поэтому можем использовать счетчики слов напрямую без применения подхода tf-idf. В целом успешное с точки зрения прогнозирования использование внешней информации отмечается в работе [Khadjeh Nassirtoussi et al., 2011]. Авторы показали, что внешнеэкономическая информация может быть полезна при прогнозировании ценовых движений валютной пары USD/GBP.

В настоящей работе использованы устоявшиеся подходы к прогнозированию безработицы на основе интернет-запросов. Вслед за [Amuri, 2017; Parlicek, 2015] в качестве базовый модели используется авторегрессия. Для анализа качества вневыборочного прогноза используются аналогичные работам [Amuri, 2017; Parlicek, 2015; Fondeur, 2013] критерии: RMSE, MAE и тест Диболда—Марьяно.

Дальнейшие разделы статьи организованы следующим образом. Второй раздел содержит в себе описание моделей прогнозирования безработицы. В третьем разделе описаны данные, в заключительном разделе представлены результаты.

Модель

Для того чтобы проверить качество прогнозирования опережающего индикатора безработицы, необходимо выбрать базовую модель и критерии сравнения. В качестве базовой выбраны две модели — AR(1) и ARIMA (p, d, q) для безработицы. Где p — порядок авторегрессионной компоненты, d — порядок интегрирования, q — порядок скользящей средней. Порядки p, d, q выбираются на основе информационных критериев и в условиях стационарности исходного ряда.

В качестве основной модели (1) выступает расширение авторегрессионных моделей до включения в них в качестве регрессоров текущих и лагированных значений опережающего индикатора по безработице.

$$U_t = \sum_{i=0}^{L} \beta_i U_{t-i} + \sum_{j=0}^{P} \gamma_j Gin_{t-j} + \varepsilon_i. \qquad (1)$$

Безработица U_t зависит от своих прошлых значений вплоть до лага L и от значений опережающего индикатора Gin_t, β_i, γ_j — оцениваемые коэффициенты влияния лагированных значений безработицы и опережающих индикаторов, соответственно $\varepsilon_t \sim iid\ N(0, 1)$. В данной работе рассматривается широкий набор индексов, которые различаются как по набору слов, так и по способам индексирования на популярность интернета как источника поиска информации о свободных вакансиях.

Пусть GY_t означает количество запросов, т.е. слов или словосочетаний, пользователей в поисковой системе Google или Яндекс. Если не учитывать популярность интернета в качестве источника, то связь между количеством запросов и индексом выглядит следующим образом (2):

$$Gin_t = \frac{GY_t}{\max\{GY_t\}} * 100\%. \quad (2)$$

Количество запросов в поисковых системах нормируется к своему максимальному значению в рассматриваемом периоде. Для того чтобы учесть возрастающую популярность интернета в качестве источника информации о свободных позициях, предлагается две дополнительные корректировки. Первой корректировкой является индексирование количества запросов на индекс популярности поиска работы в интернете Ip_t, полученной на основе данных опросов населения. Однако сам по себе показатель может содержать информацию по безработице. Например, даже при возрастающем тренде популярности поиска работы в интернете динамика могла бы быть более гладкой в отсутствие колебаний безработицы. Для того чтобы учесть данный эффект, предлагается вторая корректировка (3). Корректировка замещения рассчитывается как индекс популярности остальных источников поиска работы Is_t. Относительное значение корректировок поможет уменьшить проблему эффекта безработицы, который находится внутри описанных корректировок. С учетом двойного индексирования на прямой эффект и эффект замещения индекс опережающего индикатора рассчитывается следующим образом:

$$Gin_t = \frac{Is_t}{Ip_t} * \frac{GY_t}{\max\{GY_t\}} * 100\%. \quad (3)$$

В случае одновременного использования большого количества различных слов вместе с их лагированными значениями число наблюдений и число переменных могут отличаться незначительно. В такой ситуации следует использовать модели регуляризации (4). В данной работе использована модель LASSO следующего вида:

$$\gamma_{ij}^{lasso} = \arg\min_{\gamma} \sum_{t=1}^{T} (U_t - \sum_{i=1}^{N}\sum_{j=0}^{P} \gamma_{ij} Gin_{it-j})^2 + \lambda \sum_{i=1}^{N}\sum_{j=0}^{P} |\gamma_{ij}| + \varepsilon_t. \quad (4)$$

Параметр λ отбирается методом кросс-валидации на основе среднеквадратичных RMSE-ошибок, γ_{ij}^{lasso} является вектором оценок коэффициентов. В качестве регрессоров по-прежнему выступают запросы, нормированные к своему максимальному значению с учетом индексирования на популярность интернета как источника поиска работы.

Данные

Данные по поисковым запросам пользователей, которые предоставляют поисковые сервисы Яндекс (Яндекс wordstat) и Google (Google Trends), взаимно дополняют друг друга. Данные Google обладают более продолжительной историей. Динамические ряды доступны с 2004 г. на ежемесячной основе. Данные Яндекса предоставляются только за два года, но зато Яндекс раскрывает словосочетания, связанные с поисковым словом.

Поскольку в анализе используются данные двух поисковых систем как взаимно дополняющие, то необходимо их сопоставить. Из рис. 1 делается вывод, что динамика интернет-запросов в Яндексе и Google отличается незначимо, поэтому результаты сервисов можно использовать как взаимодополняющие. Итоговые отобранные слова — в табл. 2.

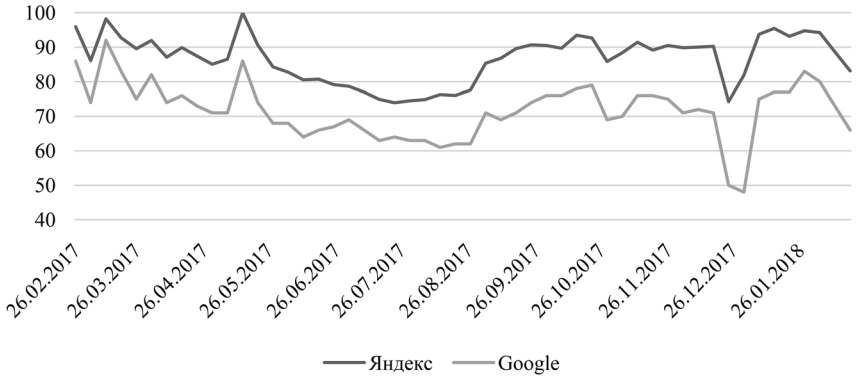

Рис. 1. Количество интернет-запросов по слову «работа» в Яндексе и Google

Примечание: поскольку используются данные компании Яндекс и Google как взаимодополняющие, то необходимо проверить сходство выдаваемых результатов и сопоставимости выборок населения. Для этого было взято относительное количество запросов Яндекса, после чего весь ряд был нормирован к максимальному значению ряда. Данные в аналогичном формате сразу предоставляются компанией Google. Динамические ряды оказались коинтегрированы на 1%-ном уровне значимости, что говорит о сопоставимости поисковых результатов.

Для того чтобы составить представление о том, какие источники информации по поиску работы наиболее популярны в России, использовались данные опросов населения, проводимых НИУ ВШЭ (RLMS). Информация, полученная из этих данных, позволила в дальнейшем учесть (1) популярность интернета в качестве источника поиска работы, а также (2) относительное замещение альтернативных источников интернетом.

Таблица 2

Абсолютные значения числа запросов по ключевым словам в Яндексе в течение трех лет за март 2016, 2017, 2018 гг.

Ключевое слово	Условное обозначение	Число запросов		
		2016	2017	2018
Вакансии	vakansii	17 760 785	23 873 206	23 394 562
Rabota ru	rabota	58 052	63 337	60 729
hh	hh	1 645 460	2 079 815	2 468 249
Career ru	career	6 619	6 065	10 636
Авито вакансии	avito_vakansii	1 369 057	1 653 267	1 546 576
Superjob	superjob	144 663	129 713	96 333
Без опыта работы	wiout_exp	468 651	545 178	551 908
Работа свежая	rabota_sveg	2 840 701	3 640 909	3 855 266
Ищу работу	look_for_job	699 405	935 994	846 863
Заработать деньги	zarabotat_dengi	312 686	401 234	475 728

Примечание: серым выделены те слова, которые имеют достаточно большой (более 100 тыс. в месяц) охват. Число 100 тыс. было выбрано экспертным путем на основе словосочетаний, предоставляемых Яндексом. Словосочетания с повторениями ниже 100 тыс. в месяц представляют из себя составные (из трех и более слов) комбинации. Однако очень специфичные словосочетания брать не рекомендуется, поскольку, например, количество запросов «ищу работу» и «ищу ищущих работу» совпадает. Яндекс распознает данные словосочетания как одинаковые и не передает верный смысл словосочетания.

Источник: построено автором на основе данных Яндекса.

В ходе анализа проводилось два ключевых способа корректировки количества запросов: на популярность интернета как источника поиска вакансий по сравнению с прочими источниками, на популярность интернета как источника поиска вакансий по сравнению со всеми другими источниками, а также с обращениями к друзьям и родственникам (наиболее популярный вариант, согласно результатам опросов).

На рис. 2 представлена разбивка доли людей по способам поиска информации на основе данных опроса населения НИУ ВШЭ (RLMS). Опрос проводится раз в год, тогда как данные по запросам предоставляются на ежемесячной основе. В рамках работы предполагается, что внутригодовая динамика популярности не поддается серьезным изменениям. Как видно из рисунка, популярность интернета как источника информации о наличии работы выросла в 3 раза с 2008 по 2016 г. Если использовать данные запросов населения в поисковых запросах в исходном виде, то можно сделать ошибочный вывод о резком росте безработицы в период с 2008 по 2016 г.

Самым популярным источником поиска работы на протяжении всего периода остаются друзья и родственники. Во всех представленных долях наблюдаются схожие сдвиги, говорящие о росте или снижении безработицы, поэтому двойное индексирование, которые было описано в разделе модели, следует использовать для корректировки динамики поисковых запросов.

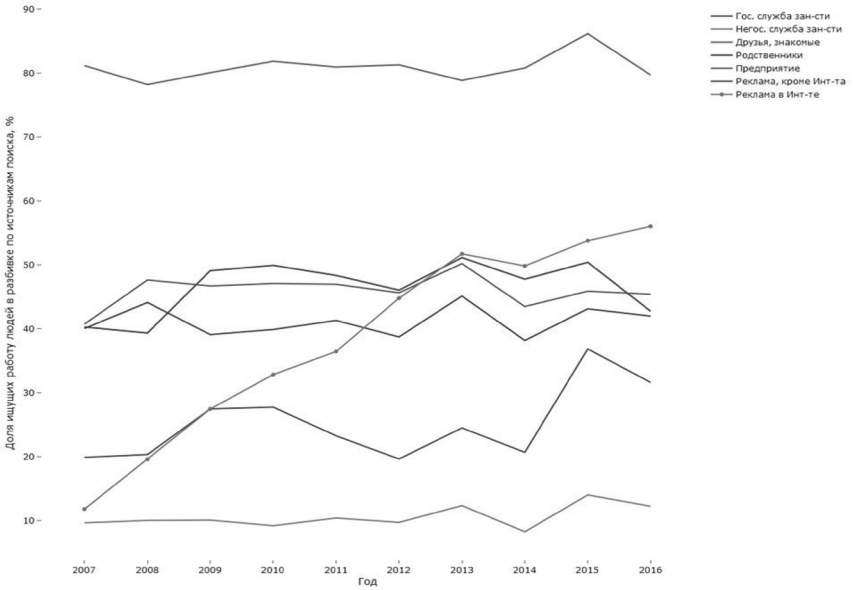

Рис. 2. Доля людей, использующих заданный источник поиска работы, среди тех, кто ищет работу
Примечание: сумма долей не обязательно должна составлять 100%, поскольку один человек мог указать в анкете более одного источника.
Источник: рассчитано автором на основе данных RLMS, построено с помощью пакета plotly.

Для того чтобы определить, какое количество лагированных значений опережающих индикаторов следует использовать, а также максимальный горизонт прогнозирования, использованы данные официальной статистики Росстата и данные RLMS по среднему времени поиска работы.

В качестве верхней границы времени работы выступает вопрос о том, сколько индивиды уже ищут работу. Скорее всего в поисках работы находятся менее квалифицированные работники, которым необходимо больше времени на поиск работы. Нижняя граница основана на опросе работающего населения о сроке, за который индивиды считают возможным найти новую работу. Согласно данным за 2007–2016 гг., период поиска работы составляет от 40 до 140 дней. Стоит отметить, что опрос индивидов о вре-

Таблица 2

Комбинации слов и вариантов их индексирования

Google показатель	Вакансии	Сайт Head Hunter	Сайт Superjob	Сайт Avito rabota	Без опыта работы	Свежая работа	Ищу работу	Заработать деньги
Без индексирования	vakansii	hh	superjob	avito_rabota	wiout_exp	rabota_sveg	look_job	zarabot_ dengi
Индексирование на популярность интернета как источника поиска	vakansii_i	hh_i	superjob_i	avito_ rabota_i	wiout_exp_i	rabota_ sveg_i	look_job_i	zarabot_ dengi_i
Двойное индексирование с учетом замещения всех остальных источников	vakansii_iall	hh_iall	superjob_iall	avito_ rabota_iall	wiout_exp_ iall	rabota_ sveg_iall	look_job_ iall	zarabot_ dengi_iall
Двойное индексирование с учетом обращений к друзьям и родственникам	vakansii_irfr	hh_irfr	superjob_irfr	avito_ rabota_irfr	wiout_exp_ irfr	rabota_ sveg_irfr	look_job_ irfr	zarabot_ dengi_irfr

Источник: составлено автором.

мени поиска может занижать время поиска работы в среднем в два раза (в предположении равномерности распределения) реальное время поиска (индивида могли опросить сразу после потери работы или только перед тем, как он найдет ее). Поэтому в дальнейшем качество вневыборочного прогноза модели оценивалось на горизонт не более 280 дней.

В табл. 2 представлен список обозначений слов и словосочетаний, которые в дальнейшем используются в Приложении 1 для изучения качества прогноза за период с марта 2012 г. по март 2018 г. Дополнительно каждое слово и словосочетание рассматриваются с учетом как простого индексирования, так и двойного индексирования.

Результаты

Эксперимент по изучению прогнозной силы моделей с использованием слов и словосочетаний, связанных с безработицей, заключается в сравнении вневыборочных ошибок моделей. Для этого выборка наблюдений разделена на две части: обучающая и тестовая. Первые две трети наблюдений используются для обучения модели, тогда как остальные — для тестирования.

Основные результаты прогнозного эксперимента представлены в табл. 3. Для односложных и составных запросов рассчитывались RMSE, MAE и статистика теста Diebold—Mariano. В таблице не представлены MAE для составных запросов, поскольку результаты несущественно отличаются от RMSE. Прогнозная сила показателей изучалась на период вплоть до шести месяцев вперед. Для каждого слова и словосочетания предполагался максимальный период опережения восемь месяцев, что связано со средним временем поиска работы, которое составило менее 280 дней.

На основе RMSE и MAE можно сделать вывод, что достаточно большой ряд спецификаций демонстрирует улучшение качества прогноза по сравнению с базовой моделью (ARIMA и AR(1), такие спецификации отмечены в табл. 3. Улучшение качества нельзя назвать случайным, поскольку в 29% случаев наблюдается улучшение качества прогноза, что превышает статистическую погрешность в 5 или 10%. Улучшение в качестве прогноза демонстрируют модели с участием показателей hh, superjob, look_for_job, zarabotat_dengi, vakansii, wiout_exp. Наибольшее количество улучшений в качестве прогноза модели демонстрируют на период 1–3 месяца.

Отобранные слова с наибольшим улучшением качества прогноза, а также их лагированные значения можно использовать для наукастинга безработицы. Средний период запаздывания выхода статистики по безработице составляет два месяца. На рис. 3 изображен пример наукастинга безработицы за последние шесть доступных точек[1] на основе моделей, включающих опережающие индикаторы в качестве регрессоров.

[1] Последние доступные точки в период осуществления данной части исследования.

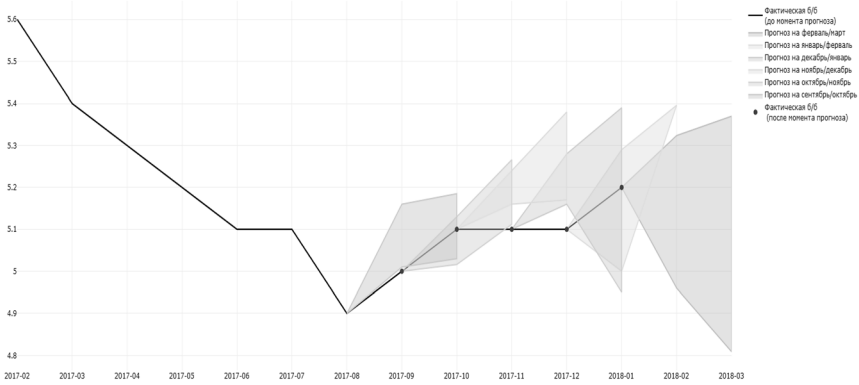

Рис. 3. Изучение качества прогноза безработицы на основе исторических данных

Примечание: на графике построено шесть прогнозов безработицы на два месяца вперед, что соответствует периоду запаздывания выхода помесячной официальной статистики по безработице. Каждый из шести прогнозов является наукастом по трем моделям (AR(1), ARIMA(p, d, q) и LASSO со словами и словосочетаниями по безработице), представленным в виде интервала с минимальным и максимальным значением безработицы по моделям. Если рассмотреть август 2017 г., например, то наукаст на сентябрь и октябрь строится в начале ноября. С выходом официальной статистики фактические значения добавляются на рисунок в виде красных точек. Еще одна особенность заключается в том, что прогноз строится на два месяца вперед каждый месяц, поэтому для точек можно видеть пересекающиеся области в виде двух прогнозов, сделанных в разный период. Из сделанных винтажей только октябрьский прогноз демонстрирует отклонение вверх.

Источник: построено автором с помощью пакета plotly.

Таким образом, использование запросов, связанных с безработицей, помогает улучшить качество прогноза. Использование индексирования позволяет найти взаимосвязь для тех словосочетаний, для которых без индексирования улучшение в прогнозе не было обнаружено. Примерами таких слов и словосочетаний являются запросы «вакансии» и «без опыта работы». Также показано, как можно использовать запросы, чтобы совершать наукастинг с учетом особенностей статистики (например, запаздывания статистики по безработице на два месяца).

Опыт использования интернет-запросов для прогнозирования безработицы можно перенести и на другие макроэкономические показатели. Для этого необходимо осуществить следующие шаги.

Шаг 1. Выбрать макроэкономический показатель, который интересует исследователя, и изучить лаг выхода статистики. Например, данные по инфляции в России выходят с задержкой менее двух недель, тогда как данные по безработице с задержкой около двух месяцев.

Шаг 2. Придумать набор слов, словосочетаний и сайтов, которые может использовать население при осуществлении запроса на данную тематику макроэкономического показателя.

Шаг 3. На основе сервиса Яндекса изучить абсолютное количество запросов по отобранным словам, подобрать возможные синонимы. Если количество запросов по словам, словосочетаниям и/или сайтам небольшое или слишком сложное, то их следует исключить из дальнейшего анализа.

Шаг 4. На примере одного или нескольких слов проверить сопоставимость динамики запросов в Яндексе и Google.

Шаг 5. Изучить третьи факторы, которые могут влиять на динамику запросов, и по возможности учесть их. В случае с безработицей таким третьим фактором является рост популярности интернета как источника поиска работы.

Шаг 6. Выбрать базовую модель для сравнения. Рекомендуется использовать AR(1) и/или ARIMA (p,d,q).

Шаг 7. На основе критериев RMSE, MAE и Диболда—Марьяно сопоставить результаты в моделях, использующих интернет-запросы, с базовыми моделями.

Шаг 8. Те запросы, которые улучшают качество прогноза, использовать для наукастинга значения макроэкономической переменной.

Заключение

В данной работе выявлены свидетельства в пользу использования интернет-запросов для прогнозирования, в частности наукастинга в макроэкономике.

На примере безработицы показано, по какому принципу можно отбирать слова, связанные с интересующей тематикой, на основе сразу двух поисковиков — Яндекса и Google. Изучение вневыборочных свойств моделей позволило сделать вывод, что в 29% (по критерию RMSE) случаев использование запросов помогает при прогнозировании безработицы. Также показано, как можно учитывать популярность интернет-запросов с помощью правильного индексирования данных[1].

В работе предложена универсальная методика, которая может использоваться для наукастинга произвольного макроэкономического показателя. Методика включает в себе следующие этапы: выбор макроэкономического показателя, условия использования наукастинга, отбор подходящих интернет-запросов, выбор базовой и составной модели, проверка вневыборочного качества составной модели.

---

[1] Автор выражает благодарность Анастасии Могилат.

Список литературы

1. *Лазарян С. С., Герман, Н. Е.* Прогнозирование текущей динамики ВВП на основе данных поисковых запросов // Финансовый журнал. — 2018. — № 6. — С. 83–94.
2. Подбор слов Яндекса. URL: https://wordstat.yandex.ru
3. Федеральная служба государственной статистики. URL: http://www.gks.ru
4. *D'Amuri F., Marcucci J.* The predictive power of Google searches in forecasting US unemployment // International Journal of Forecasting. — 2017. — 33(4). — 801–816.
5. *Askitas N., Zimmermann K. F.* Google econometrics and unemployment forecasting // Applied Economics Quarterly. — 2009. — 55(2). — 107–120.
6. *Choi H., Varian H.* Predicting the present with Google Trends // Economic Record. — 2012. — 88. — 2–9.
7. *Diebold F. X., Mariano R. S.* Comparing predictive accuracy // Journal of Business & economic statistics. — 2002. — 20(1). — 134–144.
8. *Fondeur Y., Karamé F.* Can Google data help predict French youth unemployment? // Economic Modelling. — 2013. — 30. — 117–125.
9. *Gentzkow Matthew, Bryan Kelly, Matt Taddy.* Text as Data // Journal of Economic Literature. — 2019. — 57 (3). — 535–74.
10. *Giannone D., Reichlin L., Small D.* Nowcasting: The real-time informational content of macroeconomic data // Journal of Monetary Economics. — 2008. — 55(4). — 665–676.
11. Google Trends. URL: https://www.google.com/trends
12. *Harvey D., Leybourne S., Newbold P.* Testing the equality of prediction mean squared errors // International Journal of forecasting. — 1997. — 13(2). — 281–291.
13. *Jones K. S.* A statistical interpretation of term specificity and its application in retrieval // Journal of documentation. — 2004.
14. *Kholodilin K. A., Podstawski M., & Siliverstovs B.* Do Google searches help in nowcasting private consumption? // A real-time evidence for the US. 2010.
15. *Luhn H. P.* A statistical approach to mechanized encoding and searching of literary information // IBM Journal of research and development. — 1957. — 1(4). — 309–317.
16. *Nassirtoussi A. K., Wah T. Y., Ling D. N. C.* A novel FOREX prediction methodology based on fundamental data // African Journal of Business Management. — 2011. — 5(20). — 8322.
17. *Pavlicek J., Kristoufek L.* Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries // PloS one. — 2015. — 10(5), e0127084.
18. Plotly Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC Date of publication: 2015. URL: https://plot.ly
19. *Preis T., Moat H. S., Stanley H. E.* Quantifying trading behavior in financial markets using Google Trends // Scientific reports. — 2013. — 3. — 1684.
20. Russia Longitudinal Monitoring survey, RLMS-HSE, conducted by National Research University "Higher School of Economics" and OOO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology of the Federal Center of Theoretical and Applied

Sociology of the Russian Academy of Sciences. RLMS-HSE web sites: http://www.cpc.unc.edu/projects/rlms-hse, http://www.hse.ru/org/hse/rlms
21. *Verzilin D., Maximova T., Sokolova I.* Online socioeconomic activity in Russia: patterns of dynamics and regional diversity // International Conference on Digital Transformation and Global Society. — 2017, June. — P. 55–69. Springer, Cham.

The List of References in Cyrillic Transliterated into Latin Alphabet

1. Federal'naya sluzhba gosudarstvennoj statistiki. URL: http://www.gks.ru
2. *Lazaryan S. S., German, N. E.* Prognozirovanie tekushchej dinamiki VVP na osnove dannyh poiskovyh zaprosov // Finansovyj zhurnal. — 2018. — № 6. — S. 83–94.
3. Podbor slov Yandeksa. URL: https://wordstat.yandex.ru

Приложение 1. **Сравнение качества прогноза безработицы на основе интернет-запросов в Яндексе и Google**

Таблица 3

Сравнение качества прогноза (на h месяцев вперед) в различных спецификациях модели (обозначения переменных — в табл. 2)

RMSE	лучше, чем AR и ARIMA
MAE	лучше, чем AR и ARIMA
* 10%, **5% ***1%	Тест Diebold—Mariano

Модель	Тест	h =1	h =2	h =3	h =4	h =6
ar(1)	RMSE	2,66	5,76	7,41	9,18	10,56
	MAE	13,42	22,98	23,42	24,94	28,76
arima	RMSE	2,69	6,11	6,64	6,57	6,53
	MAE	12,48	20,82	20,7	21,08	21,31
vakansii	RMSE	2,81	6,83	10,22	12,53	13,54
	MAE	14,54	21,61	26,28	28,38	30,29
vakansii (8 lags)	RMSE	3,19	7,99	13,31	19,15	26,78
	MAE	14,68	24,57	30,98	37,14	45,14
vakansii_i	RMSE	2,76	6,27	6,84	6,79	6,7
	MAE	12,68	21,16	21,06	21,4	21,53
vakansii_i (8 lags)	RMSE	3,48	5,82	6,67	9,79	13,49
	MAE	15,78	22,3	22,38	27,55	33,04
vakansii_iall	RMSE	2,56	5,38	7,86	10,26	10,49
	MAE	12,56	18,96	22,72	25,64	25,19
vakansii_iall (8 lags)	RMSE	3,22	7,49	11,41	17,14	23,4
	MAE	14,51	23,64	29,63	35,67	41,38
vakansii_irfr	RMSE	2,51	5,37	6,91	9,36	10,47
	MAE	12,15	18,59	21,81	24,36	26,74
vakansii_irfr (8 lags)	RMSE	2,93	6,75	10,52	16,01	22,45
	MAE	13,59	22,31	29,01	35,2	40,54
hh	RMSE	2,12	4,28	5,04	5,68	6,74
	MAE	10,94	16,18	18,99	21,53	22,51
hh (8 lags)	RMSE	2,14	4,16	5,49	7,19	8,28
	MAE	12,39	17,78	21,21	22,59	24,28

Продолжение табл. 3

Модель	Тест	h =1	h =2	h =3	h =4	h =6
hh_i	RMSE	3,01	6,98	8,05	8,93	10
	MAE	14,6	24,02	25,09	25,17	29,23
hh_i (8 lags)	RMSE	2,4	3,74	3,95	5,72	7,75
	MAE	11,99	16,59	16,68	19,97	23,34
hh_iall	RMSE	2,18	4,8*	5,35*	5,32	5,37
	MAE	11,32	16,81*	18,03*	19,86	21,52
hh_iall (8 lags)	RMSE	2,33	4,04	4,81	6,94	8,32
	MAE	12,2	16,29	19,75	22,52	24,73
hh_irfr	RMSE	1,94	4,08*	4,73*	5,24	5,17
	MAE	10,33	15,65*	17,68*	19,91	21,66
hh_irfr (8 lags)	RMSE	2,08	3,52	4,15	6,21	8,24
	MAE	11,44	15,28	17,97	21,55	23,74
superjob	RMSE	2,44	4,76	5,82	7,4	7,21
	MAE	11,63	17,27	19,14	22,94	24,03
superjob (8 lags)	RMSE	2,35	4,85	6,57	9,42	12,25
	MAE	11,87	19,64	22,3	25,06	29,37
superjob_i	RMSE	2,84	6,73	7,98	8,78	9,68
	MAE	13,72	21,83	22,19	23,46	24,09
superjob_i (8 lags)	RMSE	2,04	4,07	5,57	8,1	10,91
	MAE	11,17	17,51	19,87	24,89	28,7
superjob_iall	RMSE	2,53	5,6	5,76	6,19	6,3
	MAE	11,35	19,11	18,51	20,37	22,39
superjob_iall (8 lags)	RMSE	2,36	5,41	8,28	12,37	17,12
	MAE	12,04	20,22	24,51	30,1	35,88
superjob_irfr	RMSE	2,72	5,34	5,17	4,78	5,25
	MAE	11,96	18,3	16,9	17,92	21,11
superjob_irfr (8 lags)	RMSE	2,36	5,06	7,41	11,04	15,2
	MAE	11,97	19,54	23,28	28,59	33,43
avito_rabota	RMSE	2,36	5,06	7,41	11,04	15,2
	MAE	12,6	20,71	20,97	21,38	21,07
avito_rabota (8 lags)	RMSE	2,61	6,11	6,82	6,83	6,79
	MAE	12,79	22,42	30,32	39,76	62,93
avito_rabota_i	RMSE	2,96	8,75	18,05	32,45	85,17
	MAE	12,32	20,64	21,14	21,85	22,17

Продолжение табл. 3

Модель	Тест	h =1	h =2	h =3	h =4	h =6
avito_rabota_i (8 lags)	RMSE	2,58	6,12	7,05	6,98	6,79
	MAE	15,25	25,51	30,55	35,59	45,14
avito_rabota_iall	RMSE	3,72	8,69	12,65	18,26	26,49
	MAE	12,57	20,77	21,01	21,22	21,29
avito_rabota_iall (8 lags)	RMSE	3,28	10,01	19,05	32,29	75,55
	MAE	13,66	25,39	33,64	39,94	55,78
avito_rabota_irfr	RMSE	2,66	6,21	6,84	6,78	6,64
	MAE	12,59	20,87	21,01	21,34	21,34
avito_rabota_irfr (8 lags)	RMSE	3,82	11,39	19,61	29,63	60,89
	MAE	14,71	26,15	30,47	34,8	47,07
wiout_exp	RMSE	2,74	5,17	5,6	5,38	5,24
wiout_exp (8 lags)	RMSE	3,84	7,69	10,84	15,11	17,69
wiout_exp_i	RMSE	2,95	6,6	7,25	7,04	6,45
wiout_exp_i (8 lags)	RMSE	2,87	5,04	5,96	8,32	9,79
wiout_exp_iall	RMSE	2,74	5,91	6,4	5,89	5,39
wiout_exp_iall (8 lags)	RMSE	3,38	6,3	8,43	11,93	13,88
wiout_exp_irfr	RMSE	2,7	4,95	5,64	5,44	5,07
wiout_exp_irfr (8 lags)	RMSE	3,27	5,76	7,73	11,21	13,31
rabota_sveg	RMSE	2,7	7,23	9,74	11,43	12,46
rabota_sveg (8 lags)	RMSE	4,19	8,87	10,64	13,43	15,11
rabota_sveg_i	RMSE	2,72	6,81	8,39	8,99	9,75
rabota_sveg_i (8 lags)	RMSE	3,32	6,36	6,55	8,63	10,73
rabota_sveg_iall	RMSE	2,68	7	9,06	10,33	11,1
rabota_sveg_iall (8 lags)	RMSE	3,94	8,11	9,4	12,35	14,81
rabota_sveg_irfr	RMSE	2,63	7,03	9,18	10,57	11,46
rabota_sveg_irfr (8 lags)	RMSE	3,9	7,88	9,07	11,82	15,7
look_job	RMSE	2,22	5,67	7,25	7,14	5,66
look_job (8 lags)	RMSE	1,86	3,27*	3,11**	3,13	4,92
look_job_i	RMSE	2,71	6,41	7,09	6,87	6,49
look_job_i (8 lags)	RMSE	3,17	7,81	12,23	17,98	23,05
look_job_iall	RMSE	2,38	6	6,87	6,73	6,07
look_job_iall (8 lags)	RMSE	2,09	4,3	5,25	7,06	10,32

Окончание табл. 3

Модель	Тест	h =1	h =2	h =3	h =4	h =6
look_job_irfr	RMSE	2,3	5,83	6,7	6,57	5,96
look_job_irfr (8 lags)	RMSE	1,95	4,19	5,17	7,02	9,6
zarabot_dengi	RMSE	2,29	4,94	5,33	5,43	6,01
zarabot_dengi (8 lags)	RMSE	1,95	3,48*	4,83*	6,67	10,21
zarabot_dengi_i	RMSE	2,56	5,49	5,59	5,11	7,96
zarabot_dengi_i (8 lags)	RMSE	2,88	6,82	10,86	15,97	21,98
zarabot_dengi_iall	RMSE	2,68	5,97	6,48	6,32	6,49
zarabot_dengi_iall (8 lags)	RMSE	2,18	4,39	6,22	9,02	13,48
zarabot_dengi_irfr	RMSE	2,65	5,76	6,09	5,32	6,35
zarabot_dengi_irfr (8 lags)	RMSE	2,14	4,64	6,68	9,52	13,85