

Прикладная эконометрика, 2018, т. 50, с. 43–66.
Applied Econometrics, 2018, v. 50, pp. 43–66.

Д. Фантаццини, М. В. Шаклеина, Н. А. Юрас¹

Big Data в определении социального самочувствия населения России

В статье строятся индексы поисковых запросов на основе Google Trends Data с целью возможности предсказания динамики российских индексов социального самочувствия ВЦИОМ. Индексы Google рассчитаны с помощью факторного анализа на основе набора данных Google Trends Data за 2006–2016 гг., который содержит результаты 512 поисковых запросов, касающихся жилищных условий, дохода, образования и др. Модель байесовского усреднения (Bayesian model averaging) была использована для отбора индексов (категорий) поисковых запросов Google, сильно коррелированных с индикаторами социального самочувствия ВЦИОМ. Дополнительные модели множественной регрессии и построенные прогнозы подтвердили результаты модели байесовского усреднения. Построенные на основе данных поисковых запросов индексы субъективного благосостояния Google являются статистически надежными, о чем свидетельствует сильная корреляция между наблюдаемыми и прогнозными значениями индексов ВЦИОМ.

Ключевые слова: индексы социального самочувствия ВЦИОМ; большие данные; факторный анализ; модель байесовского усреднения.

JEL classification: C52; I32.

1. Введение

В последнее время растет интерес к измерению благосостояния населения и социального самочувствия в силу того, что многие экономические макропоказатели не могут измерить нравственные ценности, дружбу, счастье и пр. В этой связи больше внимание уделяется субъективным способам оценки.

С учетом наличия и широкой доступности данных из социальных сетей и онлайн-поисковых систем, растет число исследований, показывающих, что с помощью социальных медиа можно лучше оценить мнения и поведение людей, чем посредством использования традиционных опросов, в которых респонденты искажают ответы (O'Connor et al., 2010; Ceron et al., 2014; Mavragani, Tsagarakis, 2016; Isotalo et al., 2016; Oliveira et al., 2017). Существует обширная литература по социологии, в которой рассматривается проблема искажения респондентом ответа на поставленный вопрос, начиная с известной работы (Human, 1944)

¹ **Фантаццини Деан** — Московская школа экономики МГУ имени М. В. Ломоносова; Fantazzini@mse-msu.ru.
Шаклеина Марина Владиславовна — Московская школа экономики МГУ имени М. В. Ломоносова; m.leshaykina@gmail.com.

Юрас Наталья Анатольевна — Московская школа экономики МГУ имени М. В. Ломоносова; yurasnatasha@gmail.com.

до недавних работ (Preisendörfer, Wolter, 2014) и (Whiteley, 2016). Все эти работы подчеркивают, что предсказательная способность данных социальных сетей и онлайн-поиска усиливается по мере того, как число граждан, выражающих свое мнение через интернет, увеличивается с течением времени. Более того, Cody et al. (2016) обнаружили, что «не только данные Twitter могут предсказывать ответы на опросы, но мы можем использовать отдельные слова в твиттах, чтобы определить, почему один период времени более счастливый, чем другой, что невозможно сделать в традиционных опросах». Polykalas et al. (2013) утверждают, что «прогнозы на основе веб-поиска могут вскоре конкурировать с традиционными опросами, несмотря на то что большие данные содержат меньше демографической информации по сравнению с традиционными опросами». Тем не менее, они также подчеркнули, что последствия социальных событий, не связанных с темой исследования, должны быть отфильтрованы как шум, чтобы получить надежную оценку параметров модели. Аналогичным образом, Beauchamp (2017) подчеркнул необходимость фильтрации данных и последующей корректировки модели «...для обеспечения корректности в измерении, экстраполяции, интерполяции репрезентативной вариации опроса как между штатами, так и во времени». Поэтому предлагаемая в настоящей работе методология и выбор эмпирических данных соответствует современному исследовательскому подходу.

В России для оценки благосостояния широко используются индексы социального самочувствия, которые получает Всероссийский центр изучения общественного мнения (ВЦИОМ). Данные индексы отображают ответы респондентов на вопросы касательно личной и общественной жизни и являются характеристикой адаптации населения к условиям жизни в России, что помогает в построении общего вектора настроений, направление которого указывает на поведение населения в различных сферах жизни общества.

Расширить и существенно улучшить понимание благосостояния населения могут данные Google, т. к. обладают рядом важных преимуществ по сравнению со стандартными методами, используемыми для измерения благосостояния. Во-первых, они позволяют исследовать поведение людей, что более информативно по сравнению с ответами, получаемыми из социологического опроса: анализ поисковых запросов в интернете позволяет лучше раскрыть интересы и предпочтения населения. Во-вторых, данные Google очень своевременны, и это представляет особый интерес для органов власти, участвующих в разработке и проведении социально-экономической политики. Кроме того, они доступны на местном уровне и могут охватывать чрезвычайно широкий круг вопросов, что не относится к стандартным показателям благосостояния. И последнее, но не менее важное: они доступны абсолютно для всех.

Цель данной статьи — показать возможности использования данных Google для объяснения и предсказания динамики российских индексов социального самочувствия ВЦИОМ.

2. Информационное обеспечение исследования

2.1. Индексы социального самочувствия ВЦИОМ

В России для характеристики благосостояния населения с субъективной стороны применяют индикаторы общественного мнения — индексы социального самочувствия ВЦИОМ. Социальное самочувствие отслеживается агентством с 2005 г. на ежемесячной основе через проведение экспресс-опросов. Опросы проводятся по репрезентативной общероссийской

выборке: учитываются квоты по полу, возрасту, образованию и населенному пункту. Территориальный охват составляет 42 субъекта РФ и 130 населенных пунктов России. Число респондентов в каждом опросе составляет 1600 человек.

К индексам социального самочувствия ВЦИОМ относят шесть индексов:

- индекс удовлетворенности жизнью;
- индекс социального оптимизма;
- индекс материального положения;
- индекс экономического положения страны;
- индекс политической обстановки;
- индекс общего вектора развития страны.

Рассмотрим на примере индекса удовлетворенности жизнью технику построения индексов социального самочувствия. Для оценки данного индекса респондентам задают один вопрос: «Если говорить в целом, то в какой мере Вас устраивает или не устраивает сейчас жизнь, которую Вы ведете?». Варианты ответов: «Вполне устраивает, по большей части устраивает», «Отчасти устраивает, отчасти нет», «Совершенно не устраивает, по большей части не устраивает», «Затрудняюсь ответить». Далее суммируется число положительных и отрицательных ответов и находится их разность:

$$J_{\text{ч}} = \sum_{\text{положительных_и_средних_оценок}} - \sum_{\text{отрицательных_оценок}} \quad (1)$$

где Σ — сумма соответствующих оценок.

Диапазон значений каждого индекса колеблется от –100 до 100 пунктов. При положительном результате позитивные оценки доминируют над негативными. Например, значение индекса, равное 60, показывает, что позитивные оценки доминировали над негативными на 60 пунктов, а значит, в обществе преобладает позитивный настрой касательно исследуемого аспекта социального самочувствия.

К индексам общественных настроений, показывающим восприятие россиянами личной жизни, относятся следующие (см. рис. 1).

1. *Индекс удовлетворенности жизнью («В какой мере Вас устраивает сейчас жизнь, которую Вы ведете?»)* показывает, насколько россияне удовлетворены жизнью, которую ведут. Высокое значение индекса указывает, что респонденты удовлетворены образом жизни.

2. *Индекс социального оптимизма («Как Вы считаете, через год Вы (Ваша семья) будете жить лучше или хуже, чем сейчас?»)* показывает, насколько оптимистично россияне смотрят в будущее. Чем выше значение индекса, тем более оптимистично настроены респонденты.

3. *Индекс самооценок материального положения («Как бы Вы оценили в настоящее время материальное положение Вашей семьи?»)* показывает, как россияне оценивают материальное положение своей семьи. Чем выше значение индекса, тем больше респонденты довольны материальным положением семьи.

К индексам положения дел в стране относятся такие (см. рис. 2).

1. *Индекс оценок экономической ситуации («А как бы вы оценили нынешнее экономическое положение России в целом?»)* показывает, как россияне оценивают экономическую ситуацию в стране. Чем выше значение индекса, тем лучше респонденты оценивают экономику.

2. *Индекс оценок политической обстановки («А как бы вы оценили нынешнюю политическую ситуацию в России?»)* показывает, как россияне оценивают политическую обстановку



Рис. 1. Индексы общественных настроений 2006–2016 годы (в пунктах)

Источник: ВЦИОМ.

в стране. Чем выше значение индекса, тем лучше респонденты оценивают политическую ситуацию.

3. *Индекс оценок общего вектора развития страны («Насколько Вы согласны с тем, что дела в стране идут в правильном направлении?»)* показывает, как россияне оценивают общий курс развития страны. Чем выше значение индекса, тем больше респонденты одобряют направление развития страны.

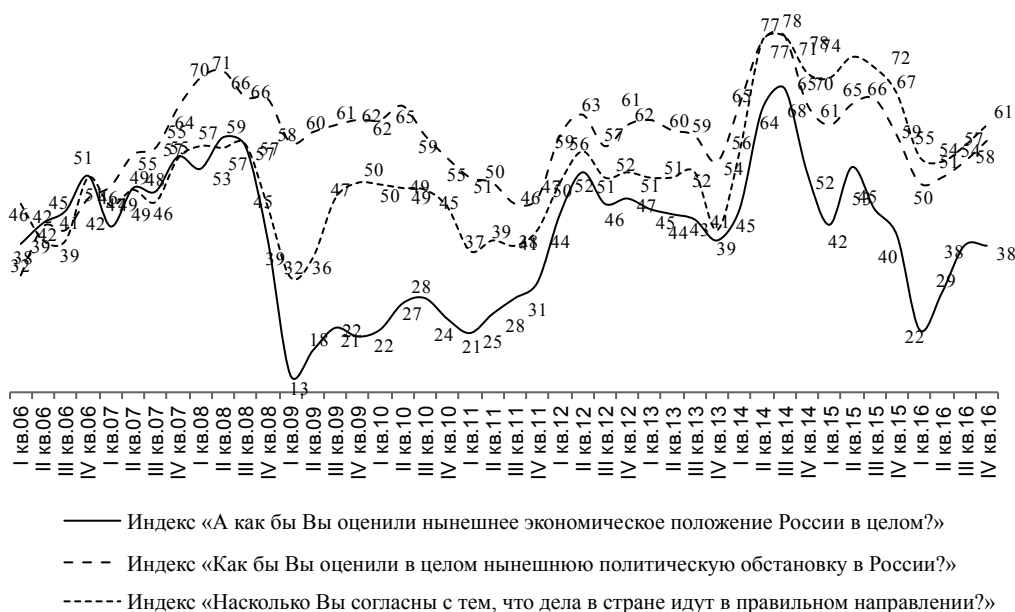


Рис. 2. Индексы положения дел в стране за 2006–2016 гг. (в пунктах)

Источник: ВЦИОМ.

Данные индексы являются отображением ответов респондентов на вопросы касательно личной и общественной жизни, что помогает в построении общего вектора настроений, а его направление указывает на поведение населения в различных сферах жизни общества.

Графики свидетельствуют о том, что социальное самочувствие имеет тренд к снижению в периоды экономического спада (Орлова, 1998; Тощенко, Харченко, 1996). В наибольшей степени отрицательные оценки преобладают в ответах респондентов на вопрос об экономическом положении России. Политическую обстановку респонденты оценивали выше, чем экономическое положение, но после событий 2014 г. началась тенденция к сближению данных индексов. В начале 2016 г. произошло снижение всех индексов, но потом наметился тренд к повышению показателей как для индексов положения дел в стране, так и для индексов социальных настроений. Уровень удовлетворенности жизнью оценивается россиянами положительно, но результаты показывают, что данный уровень намного ниже, чем 2 года назад. В целом это характерно и для оценок материального положения, и для индекса социального оптимизма. По итогам 2016 г. можно сделать вывод о том, что для россиян характерно оптимистическое отношение к жизни, положительная оценка своего самочувствия и положения дел в стране.

2.2. Google Trends Data

Google Trends — сервис, который отображает в виде индекса поисковой активности популярность какого-либо поискового запроса за определенный промежуток времени. Он представляет уникальный взгляд на то, что в данный момент интересует людей, и позволяет отслеживать изменения в предпочтениях и интересах.

Google Trends был создан для отображения фразы или ключевого слова с помощью визуальных и динамических инструментов, которые показывают прошлое, настоящее и потенциально (насколько это можно предсказать) будущее данного запроса (Askitas, 2015).

При выборе слов авторы руководствовались методологией отбора ключевых слов, представленной в работе (Algan et al., 2016). В качестве индексов субъективного благосостояния в ней использовались индексы положительных и отрицательных эмоций из опросов Gallup, которые привязаны к конкретным словам (смех, счастье, уважение, стресс, беспокойство и т. д.). Набор объясняющих индикаторов (ключевых слов) подбирался в поисковой среде Google Trends Data для анализируемых индексов исходя из анализа различных международных опросников: OECD (посвященный нахождению индекса лучшей жизни) и American Time Use Survey (посвященный исследованию использования времени населением). Во внимание принимались ключевые слова, которые потенциально связаны с благосостоянием населения. В результате в (Algan et al., 2016) были выделены следующие аспекты благосостояния: материальные условия, физическое и ментальное здоровье, социальные условия и окружающая среда. В настоящем исследовании в роли индикаторов благосостояния населения выступают индексы социального самочувствия ВЦИОМ. Для исследования авторы отобрали только пять индексов ВЦИОМ, отказавшись от индекса политической обстановки. Прокси-переменными для индексов выступают категории поисковых запросов Google, которые отбирались согласно методологии, предложенной Algan et al. (2016). Ключевые слова для поиска были отобраны из вопросов, которые задавались в рамках исследований, проводимых международными организациями: Gallup World Poll, OECD Better Life Index и NEF. В то же время используемые ключевые слова из опросников международных баз данных

могут быть далеки от повседневной жизни человека, если не характеризуют реальные условия жизни, которые отражаются непосредственно в поисковых запросах, идущих от каждого человека (Algan et al., 2016). В этой связи были добавлены такие ключевые слова, как «поиск работы», «стресс», «цены на жилье», «воспитание ребенка», «репетиторы» и пр. Данный этап отбора слов имеет ряд недостатков, т. к. сопряжен с высокой долей субъективизма.

В результате среди множества поисковых запросов, характеризующих тот или иной аспект социального самочувствия, отбирались только те, которые имеют динамику (видимые изменения) за последние 10 лет. Исходная база из поисковых запросов содержала 512 слов.

Используемый набор данных (поисковых запросов) состоит из квартальных данных, охватывающих промежутки времени 2009–2016 гг. (8 лет), в то время как Algan et al. (2016) использовали набор данных 300 еженедельных наблюдений с 6 января 2008 г. по 4 января 2014 г. (6 лет). Наши оценки могут быть менее эффективными, чем оценки Algan et al. (2016), но для данной задачи продолжительность процесса важнее, чем частота выборки. Например, рассмотрим геометрическое броуновское движение (GBM) и оценку параметра сноса μ с n дискретными данными за временной интервал $[t-h, t]$. Отклонение оценки максимального правдоподобия (ML) сноса μ (т. е. среднего значения выборки дискретных данных) зависит только от длины временного интервала h , а не от частоты дискретизации, в то время как только смещение второго порядка, присутствующее в оценке дисперсии, зависит от n и есть $O(n^{-1})$. Таким образом, широкий диапазон времени с меньшей частотой может быть более эффективным, чем узкий временной интервал с более высокой частотой (Asmussen, Glynn, 2007; Iacus, 2009). Учитывая, что временной диапазон нашего набора данных шире, чем в (Algan et al., 2016), этот факт должен хотя бы частично компенсировать потери эффективности из-за более низкой частоты нашего набора данных.

2.3. Основные проблемы больших данных Google Trends

Данные по поисковому запросу из Google Trends не являются точной характеристикой только данного слова, скорее, это доля от общего числа поисков за определенный период, включающих это ключевое слово, нормированная так, что наибольший объем за период равен 100. В этой связи значения ряда в любой взятый день нельзя сравнивать между поисковыми запросами, т. к. они нормированы к максимальному значению, которое прикреплено к каждому слову. Для решения данной проблемы применяется стандартизованная Z-оценка:

$$Z = \frac{X - E[X]}{\sigma(X)}, \quad (2)$$

где $E[X]$ — среднее значение случайной величины X , $\sigma(X)$ — ее стандартное отклонение. В результате Z-оценки имеется распределение с нулевым средним и единичной дисперсией: $\mu = 0$, $\sigma = 1$.

Преимущество стандартизации заключается в том, что Z-оценки адаптируют распределения из различных наборов данных к одному масштабу, что в дальнейшем позволяет напрямую сравнивать ранее несопоставимые переменные.

В данных Google Trends могут быть резкие скачки в популярности поискового запроса (см. Приложение 1). Это создает сложность для оценки, т. к. имеется риск переобучения

(overfitting) при построении. Для того чтобы устранить эту проблему, применяется скользящее среднее. Порядок скользящего среднего определяется числом учитываемых в модели предыдущих значений случайных отклонений.

В настоящем исследовании данные будут сглажены путем трехпериодного скользящего среднего, где период — месяц. Таким образом, каждая точка ряда будет сглажена с учетом среднего значения и с помощью создания составных показателей категорий, чтобы ослабить влияние шока.

Также имеются поисковые запросы, которые в течение продолжительного времени имеют нулевой объем поиска. Такие периоды с большим количеством нулей по своей сути имеют проблемы, схожие с резкими скачками в популярности слова. В результате слова с большим количеством нулей за период исключаются.

Многие поисковые термины имеют сильный временной тренд. Наличие временного тренда в данных является преградой для качественной оценки: во-первых, устойчивый временной тренд может привести к неправильному и нереалистичному прогнозу субъективного благосостояния в будущем, во-вторых, цель данного исследования состоит в составлении категорий слов, которые являются статистически надежными, и поэтому имеется риск в группировке слов, имеющих общий временной тренд, но не отвечающих общей смысловой нагрузке.

Для решения данной проблемы используется скорректированный коэффициент детерминации при построении регрессий с временным трендом. С помощью включения в модель трендовой составляющей можно оценить, насколько поисковый запрос зависит от тренда. В нашем случае все слова, имеющие R_{adj}^2 больше, чем 0.6, удаляются. Такой подход использован потому, что невозможно понять, что движет человеком при выполнении запроса: необходимость или влияние внешних факторов, т. к. в лучшем случае важно учитывать первый вариант, но его невозможно отличить от второго. На данном этапе может быть потеряна важная информация, но в действительности путаница в способе использовании интернета является достаточно серьезной.

Многие поисковые запросы демонстрируют экстремальную сезонность. Это может привести к проблемам, поскольку поисковые запросы будут коррелировать только из-за соответствия одной и той же сезонной схеме. Для решения данной проблемы и устранения сезонных эффектов используются месячные фиктивные переменные.

3. Методологические основы исследования

3.1. Построение индексов благосостояния на основе Google Trends Data

Построение индексов (категорий) благосостояния осуществляется на основе группировки поисковых запросов, которая должна соответствовать логической схеме и иметь схожую смысловую нагрузку. Например, невозможно объединить такие слова, как «стресс» и «музеи», в одну категорию, поэтому для упрощения слова были заранее поделены на общие категории, такие как «работа», «семья» и т. д.

Статистическое обоснование полученных категорий поисковых запросов осуществляется посредством проведения факторного анализа, главной целью которого является объединение поисковых запросов в соответствующие категории (индексы) Google, характеризующие

тот или иной аспект социального самочувствия на основе факторных нагрузок. Факторный анализ используется как метод обнаружения взаимосвязей между значениями переменных посредством исследования структуры ковариационных и корреляционных матриц. С помощью выявления факторов (латентных переменных), показывающих наличие линейных связей между переменными, можно построить категории поисковых слов и обнаружить запросы, которые не согласуются ни с одним построенным фактором. Структура факторов достаточно простая и наглядная, т. к. при анализе переменные, сильно коррелирующие между собой, объединяются в один фактор, и затем с помощью перераспределения дисперсии между компонентами получаются факторные нагрузки, на основе которых судят о согласованности поисковых запросов. Использование факторного анализа необходимо для создания составных категорий поисковых слов, что существенно сократит число потенциально объясняющих переменных.

Для факторного анализа используются слова, которые были нормализованы с помощью Z-оценки и в которых были устранены сезонные колебания.

При проведении факторного анализа поисковые запросы исключаются, если коэффициент факторной нагрузки отрицательный или менее 0.3 (Algan et al., 2016). Многие поисковые запросы не используются, поскольку они не вписываются в какую-либо категорию слов. Неподходящие слова могут разделяться на отдельные категории слов. Так, например, «Здоровье» делится на две составляющие подкатегории: «Состояние здоровья» и «Поддержка тонуса и здоровые привычки». Важно отметить, что если два слова сгруппированы в одну категорию, то это не ведет к тому, что они означают одно и то же, а только то, что им свойственна общая тенденция запросов на данном промежутке времени. В результате после применения скользящего среднего и удаления тренда используются 382 слова. Факторные нагрузки для слов, прошедших проверку, были сгруппированы в категории, которые охватывают три аспекта жизни (табл. 1).

Компоненты слов позволяют наглядно оценить, что из себя представляет каждая категория субъективного благосостояния. Так, например, *экономика* страны представлена отдельно от *финансовой безопасности*, хотя они имеют схожие характеристики, но группы поисковых запросов свидетельствуют о том, что экономика страны представляет из себя все то, что является инструментами фискальной и монетарной политики в вопросах регулирования экономики, которые имеют прямое отношение к жизни людей и отражает состояние экономики. Финансовая безопасность выражена в том, какие шоки может испытывать экономика, какие факторы влияют на финансовую безопасность, как это отражается на населении, и чем они обеспокоены при таком положении дел в стране. На графиках в Приложении 1 представлено изменение во времени категорий поисковых запросов Google.

3.2. Модель байесовского усреднения

В настоящем исследовании используется байесовская модель для определения категорий поисковых запросов Google, которые в наибольшей степени коррелируют с индексами социального самочувствия ВЦИОМ и впоследствии могут стать надежными предикторами для объяснения и прогнозирования индексов социального самочувствия ВЦИОМ.

Модель байесовского усреднения, Bayesian model averaging (BMA), направлена на отбор наиболее надежных (робастных) объясняющих переменных (Scott, Varian, 2015).

Таблица 1. Компоненты категорий субъективного благосостояния

Аспекты жизни	Категории	Группы поисковых запросов
Материальные условия	Экономика страны	Цены на жилье, подоходный налог, аренда квартиры, уровень инфляции, экономический рост, налоги, ставка рефинансирования, государственный бюджет, финансовые риски, глобализация, налог на прибыль, свой бизнес, прибыль
	Финансовая безопасность	Экономический кризис, дефолт, банкротство, цены на нефть, курс валюты, девальвация, как заработать деньги, пособие по безработице, служба занятости, увольнение, фриланс
	Рынок труда и поиск работы	Рынок труда, биржа труда, работа, ярмарка вакансий, профессия, производственная практика, условия труда, гибкий график, трудовые отношения, МРОТ, уровень безработицы
Физическое и ментальное здоровье	Состояние здоровья	Стресс, астма, кровяное давление, медицинские услуги, гипертония, ожирение, хроническая усталость, плохое здоровье
	Поддержка тонуса и здоровые привычки	Диета, диета для похудения, здоровый образ жизни, гимнастика, бег, прыжки, фитнес клуб, медитация, база отдыха, отпуск, волейбол
	Образование	Наука, учеба, аспирантура, гуманитарные науки, естественные науки, скорость чтения, грамоты, культура речи, репетиторы, дистанционные курсы, дошкольное образование, стажировки, тайм-менеджмент
	Идеалы и ценности	Этика, религия, расизм, эстетика, свобода слова, человеческий капитал, духовность, свобода, атеизм, материализм, смысл жизни
	Отдых и времяпрепровождение	Рыбалка, активный отдых, экскурсии, отдых на море, спорт, музеи, боулинг, пансионат, квесты, концерт, спектакли, пикник
Социальные условия и окружающая среда	Семья	Детское питание, воспитание ребенка, частные детские сады, уход за ребенком, уход за новорожденным, детские развлечения, творчество, художественная литература, аквапарки, грудное вскармливание
	Семейные проблемы	Как развестись, опека, насилие, поддержка семьи, нападения
	Личная безопасность	Оружие, агрессия, пожарная безопасность, техника безопасности, безопасность жизнедеятельности, кражи, уровень преступности, самозащита, самооборона, самоубийство, социальная защита
	Политическая обстановка и действия государства	Социальная поддержка, демократия, конституция, нормативные акты, транспорт, терроризм, партия, бездомные, иностранцы, взятки, бюрократия, уровень жизни, социальное страхование, социальное обеспечение
	Экология	Охрана природы, природные ресурсы, устойчивое развитие, качество воды, загрязнение окружающей среды, экологические проблемы, загрязнение атмосферы, природа, климат, парниковый эффект, переработка мусора, экологическое воспитание, лесные ресурсы

Пусть X — это множество доступных категорий (регрессоров) и $X_\gamma \in X$ — рассматриваемое подмножество. ВМА подход состоит из оценки большого числа моделей M_γ по следующему принципу:

$$y = \alpha_\gamma + \beta_\gamma X_\gamma + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \quad (3)$$

где y — зависимая переменная (категории социального самочувствия), α_γ — константа, β_γ — коэффициент, X_γ — рассматриваемое подмножество регрессоров.

Когда имеется большое число потенциальных объясняющих переменных в матрице X , сложно оценить, какие переменные $X_y \in X$ должны быть включены в модель и насколько эти переменные важны.

ВМА решает проблему с помощью оценки моделей со всеми возможными комбинациями: если X содержит K потенциальных переменных, то это означает оценку 2^K комбинаций переменных, и в итоге мы получаем 2^K моделей.

Оценка модели M_y с регрессорами X_y , которые будут выражать категории поисковых запросов, может быть сделана с помощью апостериорной модели вероятностей (posterior model probability, PMP), вытекающей из байесовской теории, где по мере поступления исходных статистических данных распределение уточняется и переходит от априорного распределения к апостериорному:

$$PMP_y = p(M_y | y, X) = \frac{p(y | M_y, X) p(M_y)}{p(y | X)}, \quad (4)$$

где $p(y | X)$ не зависит от рассматриваемой модели и является константой, $p(M_y)$ — априорное распределение, $p(y | M_y, X)$ — предельное правдоподобие.

Из данного выражения необходимо найти распределение вероятностей параметров модели после того, как приняты во внимание данные — это PMP_y , которое выражается через произведение предельного правдоподобия $p(y | M_y, X)$ и априорной вероятности модели $p(M_y)$, а $p(y | X)$ является константой. Предельное правдоподобие — это вероятность данных при условии фиксированных параметров модели. Априорная вероятность является формализацией нашей интуиции о предмете, т. е. знания, которое имелось еще до проведения анализа, эксперимента. Она определяется исследователем и самый приемлемый вариант — взять равномерную априорную вероятность для каждой модели $p(M_y) \propto 1$, поскольку имеется недостаток в априорной информации (Zeugner, 2011).

Далее, полученные PMP всех моделей, куда была включена какая-либо переменная (в нашем случае категории поисковых запросов), суммируются и используются для вывода апостериорных включений вероятностей (posterior inclusion probability, PIP), анализ которых необходим, чтобы оценить, насколько переменные важны для описания данных, и отразить корреляционную связь.

В работе (Algan, 2016) наиболее надежные категории были отобраны, если апостериорные включения вероятностей (PIP) больше, чем 0.7. В настоящем исследовании будет использоваться такой же критерий. Категории поисковых запросов Google будут являться детерминантами социального самочувствия, если PIP входит в промежуток (0.7, 1).

Результаты, полученные после применения байесовского метода, представлены в Приложении 3.

Стоит отметить, что ВМА не подсчитывает предельную вероятность и апостериорное распределение для каждой возможной модели, напротив, она идентифицирует подмножество моделей, где сосредоточен наибольший вес апостериорной вероятности. Это достигается с помощью алгоритма Метрополиса–Гастингса, который, просматривая участки модели с помощью случайного блуждания, находит наиболее важные части. Этот алгоритм был предложен одновременно с методом Монте-Карло (Metropolis, Ulam, 1949), а затем развит в работах (Metropolis et. al., 1953; Hastings, 1970).

На практике был использован «birth-death» процесс, который случайно выбирает одну из K представленных переменных на каждом шагу алгоритма и на следующем шаге отбрасывает переменные из модели, если они уже были включены.

Тем не менее, анализ ВМА может быть неэффективен, когда имеется большое количество переменных, сильно коррелирующих друг с другом, в том смысле, что в результате моделирования учитываются надежные регрессоры, но отводится в сторону рассмотрение переменных, не сильно коррелирующих с зависимой переменной. Эта известная проблема связана с проблемой «независимости нерелевантных альтернатив» (independence of irrelevant alternatives) в моделях дискретного выбора и может быть устранена с помощью предварительного факторного анализа для уменьшения числа потенциальных объясняющих переменных.

3.3. Тестирование модели

Модель байесовского усреднения строится на промежутке времени, охватывающем 2009–2016 гг., т. к. по поведению категорий поисковых запросов Google можно было заметить, что некоторые из них становились наиболее важными в кризисные периоды, а скачки в траектории могут привести к риску переобучения в ситуациях, если выбранные категории будут применяться на других выборках. В результате модель строится на данных режима обучения (training set), которые чаще всего используются для нахождения потенциальных связей (т. е. 2009–2016 гг.).

На заключительном этапе моделирования результаты байесовской модели уточняются с помощью регрессионного анализа. В качестве зависимой переменной выступает какой-либо индекс социального самочувствия, а независимых переменных — отобранные с помощью байесовской модели усреднения категории поисковых запросов Google (Приложение 4). Однако следует отметить, что использование байесовской модели усреднения (ВМА) в нашем случае является более надежным инструментом моделирования по сравнению с множественной регрессией, учитывая малый размер набора данных (Fernandez et al., 2001; Liang et al., 2008; Claeskens, Hjort, 2008).

Статистическая проверка надежности построенной модели осуществляется путем разбиения заданного набора данных на внутривыборочный период, который используется для первоначальной оценки параметров модели, и вневыборочный период, используемый для оценки качества прогнозирования. Algan et al. (2016) используют центральную часть своего набора данных для оценки модели, а первый и последний год — в качестве вневыборочного периода для проверки качества модели. В нашем исследовании внутривыборочным периодом для построения и оценки модели являются данные с 2010 по 2015 г., в то время как данные за 2009 и 2016 гг. используются как два отдельных вневыборочных периода, на которых тестируется построенная модель.

Обоснованием для определения двух вневыборочных периодов является проверка того, как работают наши индексы Google в период кризиса (например в 2009 г., когда российский ВВП упал почти на 10%), а также в период относительной стабильности (например в 2016 г., когда ВВП упал незначительно — на 0.2%). Поэтому сначала выбираются самые важные категории Google с использованием байесовской модели усреднения, затем строятся регрессии индексов ВЦИОМ по выбранным категориям, а далее удаляется любая категория, которая является статистически незначимой на уровне 5%. Для прогнозирования индексов ВЦИОМ используется только модель со значимыми коэффициентами за период с 2009 по 2016 г.

4. Эмпирические результаты исследования

4.1. Детерминанты индексов социального самочувствия ВЦИОМ, выявленные из корреляций с категориями поисковых запросов Google

На основе байесовской модели усреднения для каждого индекса социального самочувствия ВЦИОМ были определены следующие детерминанты. Курсивом выделены категории поисковых запросов Google, включенные в модель регрессии (PIR больше 0.7).

Индексы положения дел в стране

1. Для оценки экономической ситуации в России важными являются: *рынок труда и поиск работы, отдых и времяпрепровождение, личная безопасность*, образование, политическая обстановка и действия государства, поддержка тонуса и здоровые привычки, семья (Приложение 3).

2. Для оценки общего вектора развития страны: *образование, состояние здоровья, семья, личная безопасность*, отдых и времяпрепровождение, поддержка тонуса и здоровые привычки, политическая обстановка и действия государства.

Индексы социальных настроений

1. Для оценки материального положения россиянам важны: *рынок труда и поиск работы, семья, отдых и времяпрепровождение, личная безопасность*, образование, поддержка тонуса и здоровые привычки, идеалы и ценности.

2. Для оценки удовлетворенности жизнью: *семья, рынок труда и поиск работы, образование*, личная безопасность, политическая обстановка и действия государства, отдых и времяпрепровождение, поддержка тонуса и здоровые привычки.

3. Для оценки социального оптимизма: *рынок труда и поиск работы, семья, поддержка тонуса и здоровые привычки*, семейные проблемы, финансовая безопасность.

Интересным остается тот факт, что ни одна из таких анализируемых категорий, как «Экология» и «Экономика страны», не оказывают никакого влияния на индексы социального самочувствия.

Результаты байесовской модели проверялись с помощью множественной регрессии. В итоге было построено шесть уравнений регрессии, в каждом из них в роли зависимой переменной выступал один из индексов социального самочувствия ВЦИОМ, а в качестве независимых — категории поисковых запросов Google, которые были отобраны на предыдущем шаге с помощью байесовской модели усреднения (они представлены в Приложении 3). Регрессионная модель построена на квартальных данных с 2009 по 2016 г. и включает три фиктивных переменные, соответствующие первым трем кварталам. Результаты регрессионного моделирования представлены в Приложении 4. Перейдем к интерпретации полученных результатов.

1. **Индекс самооценок материального положения.** Больше число поисковых запросов, связанных с *рынком труда*, сигнализирует об ухудшении материального положения (что вполне понятно). Увеличение запросов о *семье и личной безопасности* положительно связано с материальным положением. Следует отметить неоднозначную интерпретацию индекса поисковых запросов «личная безопасность». Данный индекс составлен из запросов, касающихся проблем безопасности и имеющих негативную коннотацию, однако в него вошли и запросы (социальная защита, безопасность жизнедеятельности), рост интереса к которым не означает ухудшение индекса личной безопасности Google. В то же время исследуемый индекс самооценок материального положения имеет тенденцию быть немного выше

в течение первых двух кварталов года, что, вероятно, является следствием большого числа праздников в этот период. Затруднительна интерпретация у коэффициента при переменной «отдых и времяпрепровождение», который имеет отрицательное значение. Данная категория постоянно снижается с 2009 до 2014 г., а затем стабилизируется в течение последних трех лет. Возможно, проблемы чартерных авиакомпаний России в последние годы, а также теракты в некоторых крупных курортах (Турции и Египта) могут объяснить отрицательную корреляцию между этой категорией и индексом самооценок материального положения.

2. Индекс удовлетворенности жизнью. Увеличение запросов, связанных с *семьей*, отражает повышение удовлетворенности жизнью, в то время как большее число запросов, связанных с *рынком труда*, отражает ее снижение. Сезонных эффектов не наблюдается.

3. Индекс социального оптимизма. Рост запросов, связанных с *семьей*, положительно коррелирует с данным индексом. Однако увеличение времени на запросы, посвященные *рынку труда и поиску работы*, отражает снижение социального оптимизма. Аналогично динамике индекса самооценок материального положения, в первые два квартала значения индекса социального оптимизма выше, что объясняется большим числом праздников в этот период.

4. Индекс оценок экономической ситуации. Высокая *личная безопасность* предполагает улучшение текущей экономической ситуации, в то время как озабоченность *поиском работы* сигнализирует об ухудшении экономической обстановки. Переменная «отдых и времяпрепровождение» имеет отрицательный коэффициент, который, вероятно, подтверждается предыдущей интерпретацией индекса самооценок материального положения.

5. Индекс оценок общего вектора развития страны. Отсутствие проблем со *здоровьем* и гармония в *семье* сигнализируют о большей уверенности в направлении развития страны. Интересно, что с увеличением запросов, связанных с *образованием*, уменьшается уверенность в направлении развития страны. Это интересная, но достаточно спорная проблема (см., например, (Karlin, 2017; Kirkegaard, Grigoriev, 2017)).

4.2. Прогнозирование индексов социального самочувствия ВЦИОМ и надежность построенной модели

В таблице 2 представлены корреляции между фактическими значениями индексов ВЦИОМ и прогнозируемыми значениями на основе данных Google, как для внутривыборочного, так и для вневыборочного (тестового) периодов. На рисунке 3 показаны прогнозируемые и наблюдаемые значения индексов ВЦИОМ. Корреляции относительно внутривыборочного периода очень высоки, как и ожидалось, и всегда выше 0.9. Эти корреляции остаются довольно высокими также в течение первого вневыборочного периода в 2009 г. со значениями в диапазоне от 0.11 до 0.97, тогда как индексы ВЦИОМ и индексы на основе Google показывают близкие и аналогичные изменения (рис. 3). Другие результаты имеют место для второго периода вне выборки в 2016 г.: корреляции между прогнозируемыми и наблюдаемыми индексами ВЦИОМ всегда отрицательны в диапазоне от -0.29 до -0.93, но есть единственное исключение с положительной корреляцией 0.65. В целом, все индексы благосостояния, основанные на Google, выше индексов ВЦИОМ в 2016 г. В этой связи хочется отметить, что использование месячных данных может улучшить результаты прогнозирования. ВЦИОМ с июля 2017 г. начал публиковать индексы социального самочувствия по месяцам, что может расширить рамки будущих исследований.

Таблица 2. Корреляции между фактическими и прогнозными значениями индексов социального самочувствия ВЦИОМ

Период	Индекс самооценок материального положения	Индекс удовлетворенности жизнью
Внутривыборочный	0.93	0.92
2009	0.66	0.11
2016	-0.29	-0.42
	Индекс оценок экономической ситуации	Индекс оценок общего вектора развития страны
Внутривыборочный	0.95	0.96
2009	0.97	0.43
2016	-0.58	-0.93
	Индекс социального оптимизма	
Внутривыборочный	0.91	
2009	0.89	
2016	0.65	

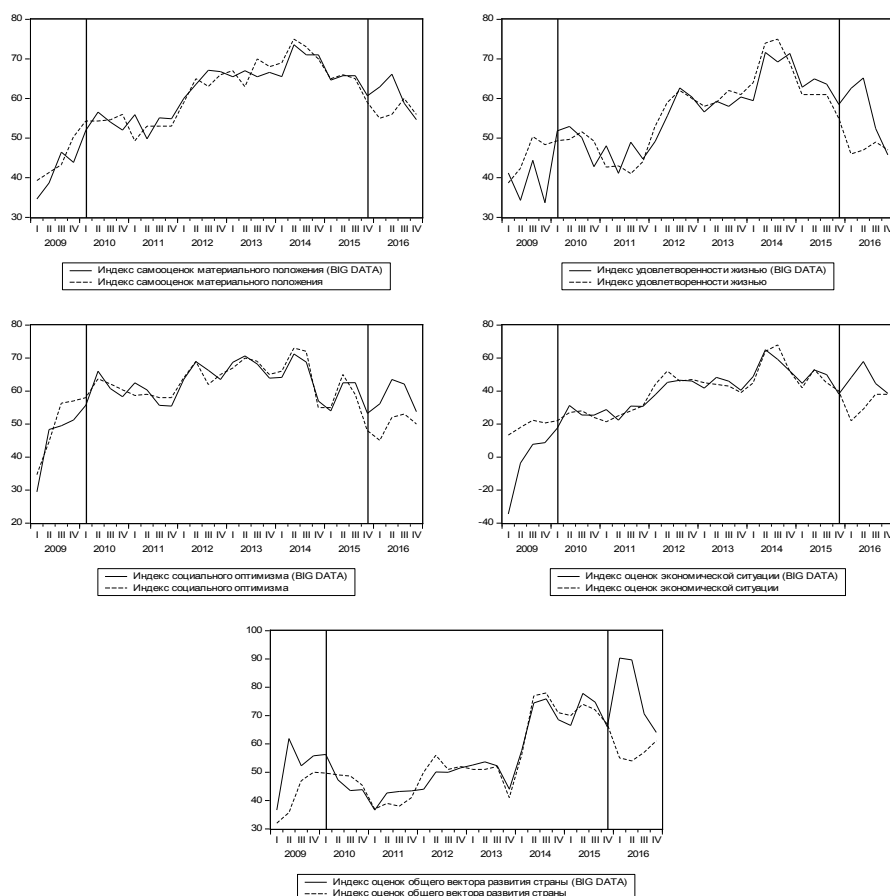


Рис. 3. Прогнозные и фактические значения индексов социального самочувствия ВЦИОМ (внутривыборочный период находится в пределах вертикальных границ)

Плохая точность прогноза для двух последних индексов может быть связана с влиянием экономического кризиса и небольшой выборкой данных, которая была доступна. Вполне возможно, что в действительности социальное самочувствие опрашиваемых было намного выше, чем следует из их ответов интервьюерам ВЦИОМ. Как ранее обсуждалось, расширяются исследования, в которых демонстрируется, что социальные медиа инструменты могут лучше представлять интересы и мнения людей, чем традиционные опросы, которые искажаются из-за ложных ответов респондентов. Это подчеркивает в своей работе Нуман (1944). Однако для проверки данного предположения в исследовании необходимы гораздо более длительные периоды наблюдения. В этом смысле новые публикации ВЦИОМ, представляющие ежемесячные индексы социального самочувствия, могут, бесспорно, улучшить качество построенных моделей.

5. Заключение

Категории поисковых запросов Google, являющиеся отражением различных аспектов жизни, были использованы как детерминанты социального самочувствия россиян. Для построения категорий этих запросов была реализована следующая последовательность шагов. Во-первых, проанализированы поисковые запросы на наличие скачков в объеме поиска, трендов и нулевого объема поиска; во-вторых, у всех слов устранены сезонные эффекты, после чего они были приведены к Z-оценке; в-третьих, на основе полученных данных проведен факторный анализ для проверки согласованности слов.

На основе модели байесовского усреднения были отобраны категории, которые в наибольшей степени коррелируют с индексами социального самочувствия, а модель множественной регрессии подтвердила полученные результаты.

Индексы Google представляют собой интересный дополнительный инструмент, который может использоваться наряду со стандартными индексами. Найденные детерминанты индексов социального самочувствия ВЦИОМ — категории (индексы) поисковых запросов Google могут быть использованы в определении ключевых направлений в социально-экономической политике, а обеспечение существенно более высокого уровня не только материальных, но и социальных благ позволит улучшить социальное самочувствие и, как следствие, качество жизни.

Построенные на основе данных поисковых запросов категории субъективного благосостояния являются статистически надежными и могут использоваться в дальнейшем анализе, например, для понимания того, что в наибольшей степени определяет социальное самочувствие человека в период экономического кризиса или после реализации каких-либо мер социально-экономической политики.

Список литературы

Орлова Л. А. (1998). О социальном самочувствии учителей Московской области (по результатам социологических опросов). *Социологические исследования*, 8, 89–94.

Тощенко Ж. Т., Харченко С. В. (1996). *Социальное настроение*. М.: Academia.

Algan Y., Beasley E., Guyot F., Higa K., Murtin F., Senik C. (2016). Big data measures of well-being: Evidence from a Google well-being index in the United States. *OECD Statistics Working Papers* 2016/3, 1–38.

Askatas N. (2015). Google search activity data and breaking trends. IZA World of Labor. <https://wol.iza.org/articles/google-search-activity-data-and-breaking-trends/long>.

Asmussen S., Glynn P. (2007). *Stochastic simulation: Algorithms and analysis* (No. 57). Springer Science and Business Media.

Beauchamp N. (2017). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, 61 (2), 490–503.

Claeskens G., Hjort N. (2008). Model selection and model averaging. Cambridge: Cambridge University Press.

Ceron A., Curini L., Iacus S. M., Porro G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media and Society*, 16 (2), 340–358.

Cody E. M., Reagan A. J., Dodds P. S., Danforth C. M. (2016). Public opinion polling with twitter. *arXiv preprint*. arXiv:1608.02024.

Fernandez C., Ley E., Steel M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100 (2), 381–427.

Hastings W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1), 97–109.

Hyman H. (1944). Do they tell the truth? *The Public Opinion Quarterly*, 8 (4), 557–559.

Iacus S. (2009). *Simulation and inference for stochastic differential equations: With R examples*. Springer Science and Business Media.

Isotalo V., Saari P., Paasivaara M., Steineker (2016). Predicting 2016 US presidential election polls with online and media variables. *Designing Networks for Innovation and Improvisation*. Springer International Publishing, 45–53.

Karlin A. (2017). The Moscow municipal elections 2017. <http://www.unz.com/akarlin/the-moscow-municipal-elections-2017/>.

Kirkegaard O., Grigoriev A. (2017). Moscow districts analysis. http://rpubs.com/EmilOWK/Moscow_districts.

Liang F., Paulo R., Molina G., Clyde M., Berger J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103 (481), 410–423.

Mavragani A., Tsagarakis K. (2016). YES or NO: Predicting the 2015 Greek referendum results using Google Trends. *Technological Forecasting and Social Change*, 109, 1–5.

Metropolis N., Rosenbluth A., Rosenbluth M. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6), 1087–1092.

Metropolis N., Ulam S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44 (247), 335–341.

O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 122–129.

Oliveira D., Bermejo S., Santos P. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology and Politics*, 14 (1), 34–45.

Polykalas S. E., Prezerakos G. N., Konidakis A. (2013). An algorithm based on Google Trends' data for future prediction. Case study: German Elections. In: *Signal Processing and Information Technology*, 69–73, IEEE International Symposium.

Preisendörfer P., Wolter F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, 78 (1), 126–146.

Scott S., Varian H. (2015). Bayesian variable selection for nowcasting economic time series. In: *Economic analysis of the digital economy*. Chicago, University of Chicago Press.

Whiteley P. (2016). Why do voters lie to the pollsters? *Political Insight*, 7 (1), 16–19.

Zeugner S. (2011). *Bayesian model averaging with BMS*. Tutorial to the R-package BMS 1e30. <https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf>.

Поступила в редакцию 14.10.2017;
принята в печать 08.06.2018.

Приложение 1

Результаты анализа больших данных и первичной обработки

Для построения надежных категорий субъективного благосостояния необходимо отобрать данные, которые являются релевантными и не будут приводить к риску переобучения.

Первым пунктом анализа данных будет рассмотрение поисковых запросов, имеющих резкие скачки в объеме поиска.

Например, скачок в поиске слова «рак» в январе 2014 г. был связан с известием об обнаружении рака головного мозга у певицы Жанны Фриске, что очень сильно взволновало российскую общественность и привело к увеличению запросов по данному слову (рис. 1). Это показывает, что люди были заинтересованы данным известием, но не отражает, насколько данная категория важна для формирования субъективного благосостояния россиян.

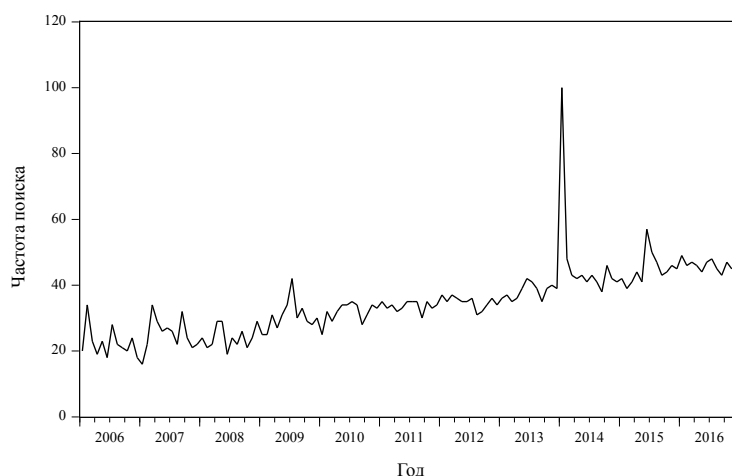


Рис. 1. Частота поиска слова «рак» за 2006–2016 гг. (в пунктах)

Как было отмечено ранее, такие поисковые запросы приводят к риску переобучения, поэтому для решения этой проблемы применяется скользящее среднее третьего порядка — МА(3) (рис. 2).

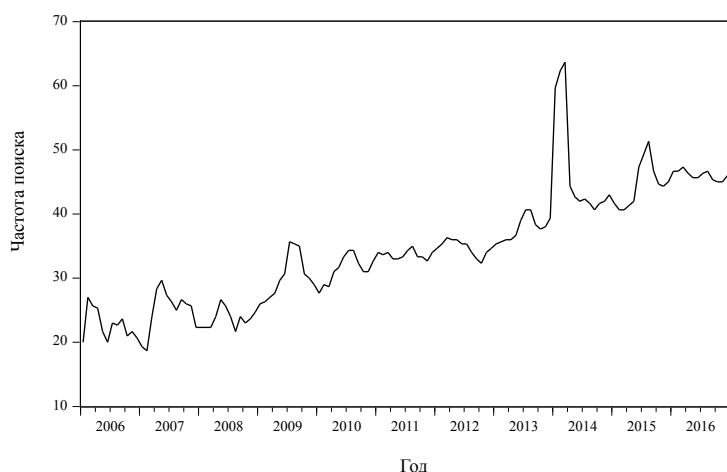


Рис. 2. Частота поиска слова «рак» за 2006–2016 гг. (в пунктах) после применения скользящего среднего

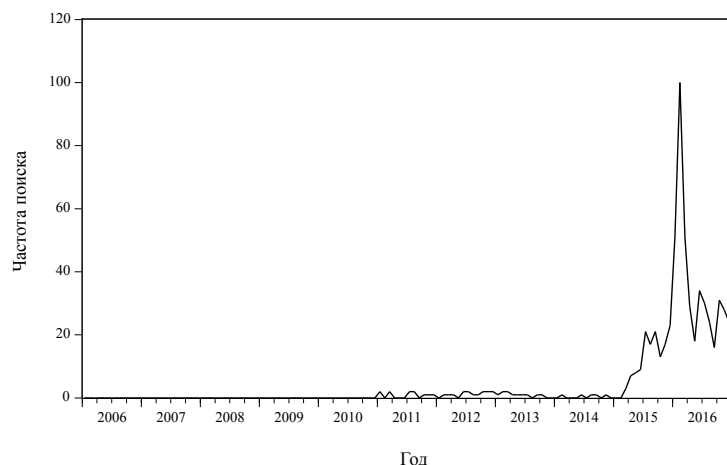


Рис. 3. Частота поиска фразы «ипотека с господдержкой» за 2006–2016 гг. (в пунктах)

В результате после применения скользящего среднего получаем сглаженный временной ряд для поискового запроса. Стоит отметить, что данный запрос имеет временной тренд, который тоже является преградой для построения надежных категорий субъективного благосостояния, что будет отмечено ниже.

Следующая категория слов, которую необходимо рассмотреть — слова, продолжительное время остающиеся без запроса. Такие запросы хорошо работают на одном массиве данных, но плохо объясняют общие закономерности на другом. Пример такого поискового запроса представлен на рис. 3. Данный график иллюстрирует, что в течение 5 лет поисковый запрос «ипотека с господдержкой» имел низкую популярность, в то же время нельзя исключить тот факт, что данное слово играет важную роль в жизни россиян, но в связи с принятой методологией такие запросы в данной статье были удалены.

Далее, в данных были найдены поисковые запросы, которые имеют сильный временной тренд. Рассмотрим пример такого запроса — «купить квартиру» (рис. 4).

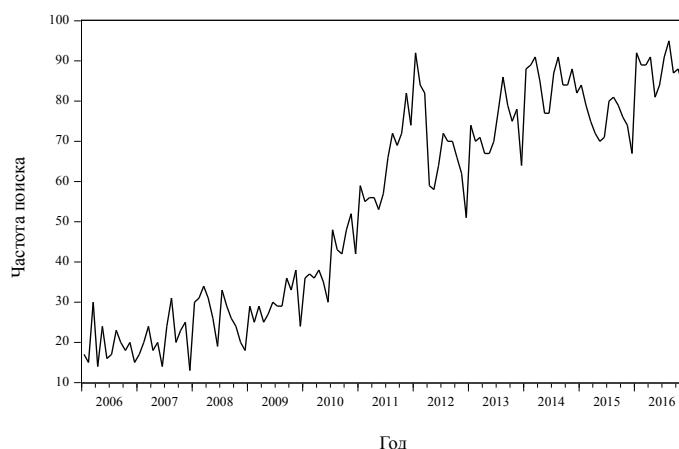


Рис. 4. Частота поиска фразы «купить квартиру» за 2006–2016 гг. (в пунктах)

В первую очередь, может быть так, что покупка квартиры с каждым годом становится более важным и необходимым условием в жизни россиян, вследствие этого наблюдается рост объема поиска в течение 6 лет. Возможен и такой вариант, что люди не чувствуют необходимости покупать квартиру, но обращаются к интернету за советом потому, что произошел какой-либо культурный сдвиг или влияние внешних факторов, как например, рост цен на квартиры в связи с экономическим кризисом, или люди стали пользоваться интернетом намного чаще.

Для устранения этой проблемы проверяем слова на временной тренд с помощью построения регрессии методом наименьших квадратов. Регрессия слова с временным трендом, где $R_{adj}^2 > 0.6$, исключаются из дальнейшего анализа.

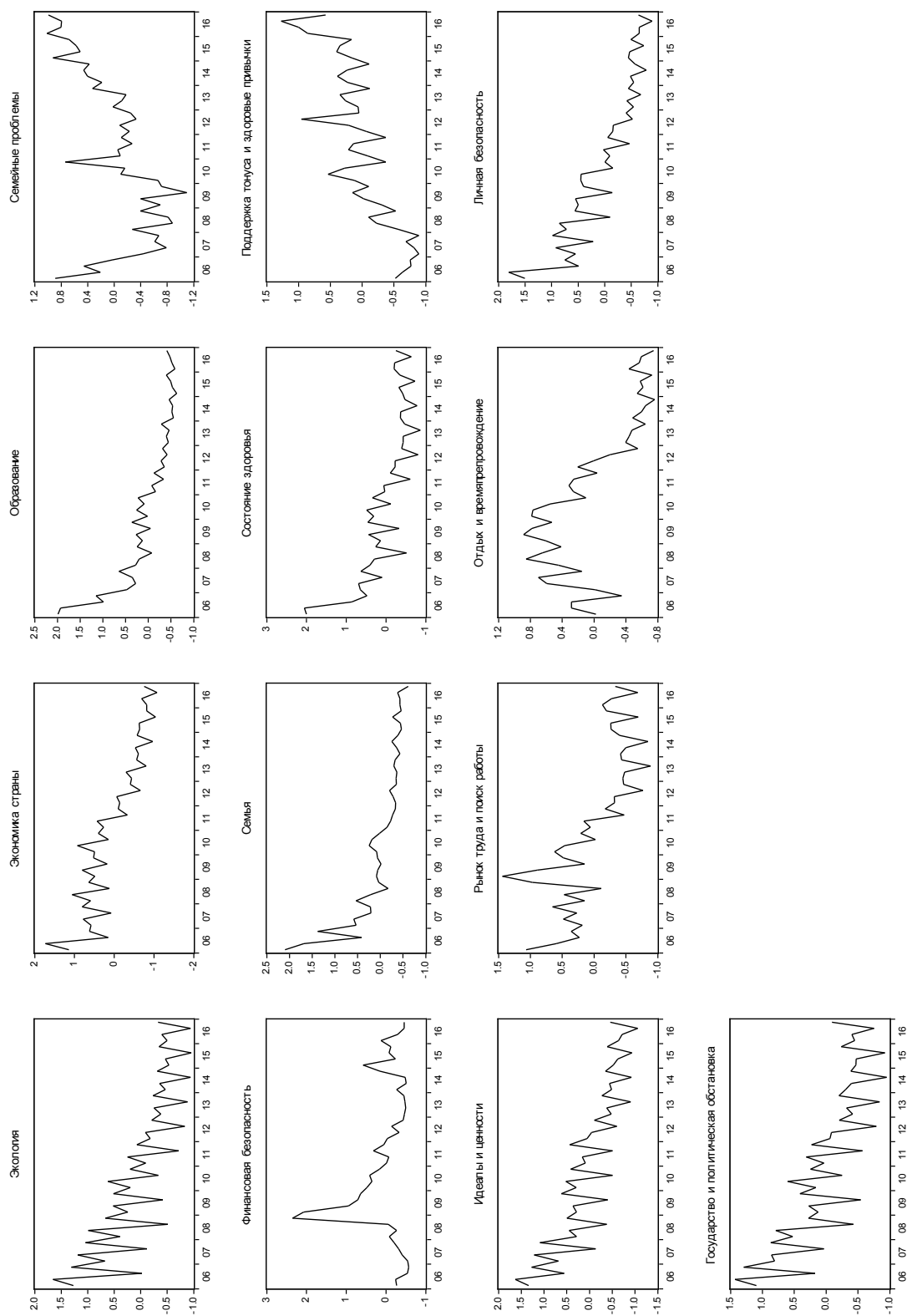
Прежде чем провести факторный анализ, необходимо выполнить два этапа действий: построение Z-оценки и устранение сезонности.

Z-оценка для каждого поискового запроса необходима, чтобы отследить, как изменялся объем поиска именно по данному запросу, а не для оценки того, какие запросы были популярны среди остальных поисковых запросов. Таким образом, все слова были приведены к стандартному нормальному распределению.

Далее, многие поисковые слова имеют экстремальную сезонность, и для того чтобы переменные не коррелировали только потому, что соответствуют одной сезонной схеме, были созданы месячные фиктивные переменные, и с помощью использования месячных остатков регрессий, которые образуют очищенный ряд, устранялась сезонность в данных.

Приложение 2

Графики, иллюстрирующие изменение категорий субъективного благосостояния за 2006–2016 гг.



Приложение 3

Модель байесовского усреднения

Д. Фантазини, М. В. Шаклеина, Н. А. Юрас

Индекс самооценок материального положения Как бы Вы оценили в настоящее время материальное положение Вашей семьи?		Индекс удовлетворенности жизнью В какой мере Вас устраивает сейчас жизнь, которую Вы ведете?		Индекс социального оптимизма Как Вы считаете, через год Вы (Ваша семья) будете жить лучше или хуже, чем сейчас?	
Категория	PIP	Категория	PIP	Категория	PIP
Рынок труда и поиск работы	0.98²	Семья	0.99	Рынок труда и поиск работы	1.00
Семья	0.95	Рынок труда и поиск работы	0.80	Семья	0.98
Отдых и времяпре- провождение	0.90	Образование	0.68	Поддержка тонуса и здоровые привычки	0.88
Личная безопасность	0.81	Личная безопасность	0.58	Семейные проблемы	0.54
Образование	0.40	Политическая обстановка и действия государства	0.55	Финансовая безопасность	0.40
Поддержка тонуса и здоровые привычки	0.40	Отдых и времяпре- провождение	0.49	Политическая обстановка и действия государства	0.27
Идеалы и ценности	0.30	Поддержка тонуса и здоровые привычки	0.41	Личная безопасность	0.25
Политическая обстановка и действия государства	0.20	Семейные проблемы	0.28	Экология	0.19
Экономика страны	0.20	Идеалы и ценности	0.26	Состояние здоровья	0.19
Состояние здоровья	0.19	Состояние здоровья	0.23	Образование	0.18
Семейные проблемы	0.17	Финансовая безопасность	0.23	Экономика страны	0.16
Финансовая безопасность	0.16	Экономика страны	0.18	Идеалы и ценности	0.15
Экология	0.14	Экология	0.14	Отдых и времяпре- провождение	0.14
Индекс оценок экономической ситуации А как бы вы оценили нынешнее экономическое положение России в целом?		Индекс оценок общего вектора развития страны Насколько Вы согласны с тем, что дела в стране идут в правильном направлении?			
Категория	PIP	Категория	PIP		
Рынок труда и поиск работы	0.91	Образование	0.92		
Отдых и времяпрепровождение	0.85	Состояние здоровья	0.90		
Личная безопасность	0.83	Семья	0.84		
Образование	0.41	Личная безопасность	0.54		
Политическая обстановка и действия государства	0.40	Отдых и времяпрепровождение	0.46		
Поддержка тонуса и здоровые привычки	0.39	Поддержка тонуса и здоровые привычки	0.36		
Семья	0.30	Политическая обстановка и действия государства	0.32		
Экология	0.22	Экономика страны	0.23		
Идеалы и ценности	0.21	Семейные проблемы	0.20		
Финансовая безопасность	0.20	Финансовая безопасность	0.20		
Экономика страны	0.20	Рынок труда и поиск работы	0.20		
Семейные проблемы	0.20	Идеалы и ценности	0.19		
Состояние здоровья	0.17	Экология	0.11		

² Жирным выделены категории поисковых запросов Google, для которых апостериорные включения вероятностей (PIP) больше, чем 0.7.

Приложение 4 Регрессии категорий субъективного благосостояния ВЦИОМ

	Индекс самооценок материального положения	Индекс удовлетворенности жизнью	Индекс социального оптимизма	Индекс оценок экономической ситуации	Индекс оценок общего вектора развития страны
Материальные условия					
Экономика страны					
Финансовая безопасность					
Рынок труда и поиск работы	-19,62****	-20,43****	-27,35****	-25,46****	
Физическое и ментальное здоровье					36,30**
Состояние здоровья			-7,91****		
Поддержка тонуса и здоровые привычки					
Образование					-118,98****
Идеалы и ценности				-32,89****	
Отдых и времяпрепровождение	-21,40****				
Социальные условия	24,93****	19,27	37,09****		72,45****
Семейные проблемы					
Личная безопасность	20,25****			48,60****	
Политическая обстановка и действия государства					
Экология					
Константа	62,10****	56,63****	63,65****	34,16****	60,11****
I квартал	6,42**	1,13	4,45**	8,37*	-24,79****
II квартал	4,21**	1,26	8,60****	9,90**	-13,57**
III квартал	1,11	-7,29*	-5,99**	15,04**	-8,59
R^2	0,90	0,50	0,87	0,78	0,68
R^2_{adj}	0,87	0,40	0,84	0,73	0,60
Число наблюдений	32	32	32	32	32

Примечание. p -значения: **** — $p < 0,001$, *** — $p < 0,01$, ** — $p < 0,05$, * — $p < 0,1$.

Fantazzini D., Shakleina M., Yuras N. Big Data for computing social well-being indices of the Russian population. *Applied Econometrics*, 2018, v. 50, pp. 43–66.

Dean Fantazzini

Moscow School of Economics — Moscow State University;
Fantazzini@mse-msu.ru

Marina Shakleina

Moscow School of Economics — Moscow State University;
m.leshaykina@gmail.com

Natalia Yuras

Moscow School of Economics — Moscow State University;
yurasnatasha@gmail.com

Big Data for computing social well-being indices of the Russian population

The article builds indices of social well-being based on Google Trends Data for predicting VCIOM indices. The Google indices were computed using a Google Trends dataset for 2006–2016 containing 512 search queries relative to housing conditions, income, education, etc., and applying factor analysis. Bayesian Model Averaging was then used to select the indexes of individual social well-being mostly associated with the VCIOM indices which measure the social well-being of the Russian population. Additional regression models and forecasting exercises confirmed the previous results. Based on the Google Trends Data, the indices of the subjective social well-being are statistically reliable, as evidenced by a strong correlation between the observed and predicted values of the VCIOM indices.

Keywords: social well-being indices; Google Trends Data; Factor analysis; Bayesian Model Averaging

JEL classification: C52; I32.

References

- Orlova L. A. (1998). O social'nom samochuvstvii uchitelej Moskovskoj oblasti (po rezul'tatam sociologicheskikh oprosov). *Sociological Studies*, 8, 89–94 (in Russian).
- Toshchenko Zh. T., Harchenko S. V. (1996). *Social'noe nastroyenie*. M.: Academia (in Russian).
- Algan Y., Beasley E., Guyot F., Higa K., Murtin F., Senik C. (2016). Big data measures of well-being: Evidence from a Google well-being index in the United States. *OECD Statistics Working Papers* 2016/3, 1–38.
- Askatas N. (2015). Google search activity data and breaking trends. IZA World of Labor. <https://wol.iza.org/articles/google-search-activity-data-and-breaking-trends/long>.
- Asmussen S., Glynn P. (2007). *Stochastic simulation: Algorithms and analysis* (No. 57). Springer Science and Business Media.
- Beauchamp N. (2017). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, 61 (2), 490–503.
- Claeskens G., Hjort N. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.
- Ceron A., Curini L., Iacus S. M., Porro G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media and Society*, 16 (2), 340–358.

- Cody E. M., Reagan A. J., Dodds P. S., Danforth C. M. (2016). Public opinion polling with twitter. *arXiv preprint*. arXiv:1608.02024.
- Fernandez C., Ley E., Steel M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100 (2), 381–427.
- Hastings W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1), 97–109.
- Hyman H. (1944). Do they tell the truth? *The Public Opinion Quarterly*, 8 (4), 557–559.
- Iacus S. (2009). *Simulation and inference for stochastic differential equations: With R examples*. Springer Science and Business Media.
- Isotalo V., Saari P., Paasivaara M., Steineker (2016). Predicting 2016 US presidential election polls with online and media variables. *Designing Networks for Innovation and Improvisation*. Springer International Publishing, 45–53.
- Karlin A. (2017). The Moscow municipal elections 2017. <http://www.unz.com/akarlin/the-moscow-municipal-elections-2017/>.
- Kirkegaard O., Grigoriev A. (2017). Moscow districts analysis. http://rpubs.com/EmilOWK/Moscow_districts.
- Liang F., Paulo R., Molina G., Clyde M., Berger J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103 (481), 410–423.
- Mavragani A., Tsagarakis K. (2016). YES or NO: Predicting the 2015 Greek referendum results using Google Trends. *Technological Forecasting and Social Change*, 109, 1–5.
- Metropolis N., Rosenbluth A., Rosenbluth M. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6), 1087–1092.
- Metropolis N., Ulam S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44 (247), 335–341.
- O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 122–129.
- Oliveira D., Bermejo S., Santos P. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology and Politics*, 14 (1), 34–45.
- Polykalas S. E., Prezerakos G. N., Konidaris A. (2013). An algorithm based on Google Trends' data for future prediction. Case study: German Elections. In: *Signal Processing and Information Technology*, 69–73, IEEE International Symposium.
- Preisendörfer P., Wolter F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, 78 (1), 126–146.
- Scott S., Varian H. (2015). Bayesian variable selection for nowcasting economic time series. In: *Economic analysis of the digital economy*. Chicago, University of Chicago Press.
- Whiteley P. (2016). Why do voters lie to the pollsters? *Political Insight*, 7 (1), 16–19.
- Zeugner S. (2011). *Bayesian model averaging with BMS*. Tutorial to the R-package BMS 1e30. <https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf>.

Received 14.10.2017; accepted 08.06.2018.