ОРИГИНАЛЬНАЯ СТАТЬЯ

CC BY 4.0

DOI: 10.26794/2587-5671-2021-25-5-215-234
УДК 330.59(045)
JEL I10, I12, C53, C55

Возможности интеграции Google Trends и официальной статистики при оценке социальной комфортности и прогнозировании финансового положения населения

М.В. Шаклеина[a] ✉, М.И. Волкова[b], К.И. Шаклеин[c], С.Р. Якиро[d]
[a] Московская школа экономики МГУ имени М.В. Ломоносова, Москва, Россия;
[b] РЭУ им. Г.В. Плеханова, Москва, Россия;
[c] Открытое акционерное общество «Российские железные дороги», Москва, Россия;
[d] Акционерное общество «Страховое общество газовой промышленности», Москва, Россия
[a] https://orcid.org/0000-0002-1947-8640; [b] https://orcid.org/0000-0001-8941-0548;
[c] https://orcid.org/0000-0003-3508-7372; [d] https://orcid.org/0000-0003-2365-8043
✉ Автор для корреспонденции

АННОТАЦИЯ

Целью исследования является развитие теории статистического наблюдения в части научно-методологических подходов к обработке больших данных и определение возможностей интеграции информационных ресурсов различного типа в отношении измерения сложных латентных категорий (на примере социальной комфортности), применение данного опыта на практике посредством использования в прогнозировании индикаторов финансового состояния. Авторами построена модель социальной комфортности, в которой выбор весов для ее компонентов осуществляется на основе **метода** модифицированной главной компоненты. Оценка произведена с использованием данных Google Trends и официальной статистики. Методы анализа данных Google Trends базируются на разработке комплексного подхода к семантическому поиску информации о компонентах социальной комфортности, снижающего долю авторского субъективизма; методологии первичной обработки с учетом принципов сопоставимости, однородности, согласуемости, релевантности, описание функций и моделей, необходимых для отбора и корректировки поисковых запросов. Предложенный алгоритм работы с большими данными позволил определить компоненты социальной комфортности («Образование и обучение», «Безопасность», «Отдых и свободное время»), по которым необходимо произвести непосредственное встраивание больших данных в систему первичного статистического учета с дальнейшей обработкой данных и получением композитных показателей. Сделан **вывод**, что по компоненте «Финансовое положение» найдена устойчивая значимая корреляция, что позволяет использовать ее для дальнейших расчетов и экстраполяции показателей финансового состояния. **Научная новизна** состоит в разработке принципов и направлений интеграции двух альтернативных источников данных в оценке сложных латентных категорий. Полученные выводы и результаты интегральной оценки социальной комфортности могут быть использованы органами государственной статистики по формированию нового вида непрерывного статистического наблюдения на основе использования больших данных, а также исполнительной власти на федеральном, региональном и муниципальном уровнях в части определения приоритетов разрабатываемой социально-экономической политики.

Ключевые слова: социальная комфортность; благосостояние населения; гармонизация информационных ресурсов; официальная статистика; Google Trends; интегральный индикатор

Для цитирования: **Шаклеина М.В., Волкова М.И., Шаклеин К.И., Якиро С.Р. Возможности интеграции Google Trends и официальной статистики при оценке социальной комфортности и прогнозировании финансового положения населения. *Финансы: теория и практика.* 2021;25(5):215-234. DOI: 10.26794/2587-5671-2021-25-5-215-234**

© Шаклеина М.В., Волкова М.И., Шаклеин К.И., Якиро С.Р., 2021

ORIGINAL PAPER

Prospects for the Integration of Google Trends Data and Official Statistics to Assess Social Comfort and Predict the Financial Situation of the Population

M.V. Shakleina[a] ✉, M.I. Volkova[b], K.I. Shaklein[c], S.R. Yakiro[d]
[a] Lomonosov Moscow State University, Moscow, Russia; [b] Plekhanov Russian University of Economics, Moscow, Russia;
[c] OJSC "Russian Railways", Moscow, Russia; [d] JSC "SOGAZ", Moscow, Russia
[a] https://orcid.org/0000-0002-1947-8640; [b] https://orcid.org/0000-0001-8941-0548;
[c] https://orcid.org/0000-0003-3508-7372; [d] https://orcid.org/0000-0003-2365-8043
✉ Corresponding author

ABSTRACT

This paper aims to develop a theory of statistical observation in terms of scientific and methodological approaches to processing big data and to determine the possibilities of integrating information resources of various types to measure complex latent categories (using the example of social comfort) and to apply this experience in practice through the use of the financial situation indicators in forecasting. The authors have built a social comfort model in which the choice of weights for its components is based on a modified **principal component analysis**. The assessment is based on Google Trends data and official statistics. Google Trends data analysis methods are based on the development of an integrated approach to the semantic search for information about the components of social comfort, which reduces the share of author's subjectivity; methodology of primary processing, considering the principles of comparability, homogeneity, consistency, relevance, description of functions and models necessary for the selection and adjustment of search queries. The proposed algorithm for working with big data allowed to determine the components of social comfort ("Education and Training", "Safety", "Leisure and free time"), for which it is necessary to directly integrate big data in the system of primary statistical accounting with further data processing and obtaining composite indicators. The authors **conclude** that a stable significant correlation has been found for the "Financial Situation" component, which makes it possible to use it for further calculations and extrapolation of financial indicators. The **scientific novelty** lies in the development of principles and directions for the integration of two alternative data sources when assessing complex latent categories. The findings and the results of the integral assessment of social comfort can be used by state statistics authorities to form a new type of continuous statistical observation based on the use of big data, as well as by executive authorities at the federal, regional and municipal levels in terms of determining the priorities of socio-economic policy development.

Keywords: social comfort; well-being; harmonization of information resources; official statistics; Google Trends; integral indicator

For citation: **Shakleina M.V., Volkova M.I., Shaklein K.I., Yakiro S.R. Prospects for the integration of Google Trends data and official statistics to assess social comfort and predict the financial situation of the population.** ***Finance: Theory and Practice.*** **2021;25(5):215-234. DOI: 10.26794/2587-5671-2021-25-5-215-234**

ВВЕДЕНИЕ

За последние два десятилетия популяризация использования сети Интернет значительно увеличилась, что способствовало росту объема хранимой информации о действиях пользователей в сети Интернет. Примерами больших данных являются данные социальных сетей, записи телефонных звонков, веб-сайтов, данные поисковых сетей [1]. Складывающиеся тенденции привлекли академический интерес к использованию больших данных в исследованиях.

Большую популярность Big Data приобретают в измерении благосостояния людей, а также в прогнозировании финансовых показателей. На основе анализа зарубежной и отечественной периодики можно выделить следующие основные источники больших данных.

1. **Google Trends (GT).** Y. Algan и др. [2] используют поисковые запросы для построения индекса благосостояния Google в США, далее апробируется методика для построения индекса социального самочувствия россиян [3]. В статье Y. Algan и др. [2] выделяется несколько аспектов благосостояния: материальные условия (финансовое благополучие), социальные аспекты и здоровье.

2. **Статьи в газетах.** В работе E. Carlquist и др. [4] исследовали встречаемость за 1992–2014 гг. в норвежской прессе 39 слов, описывающих финансовое положение населения. Всего было ото-

брано четыре газеты, чтобы охватить социокультурные и региональные различия. В первую очередь авторы стремились исследовать слова и выражения, относящиеся к финансовому положению в повседневной норвежской лексике.

3. **Twitter.** Данные также могут быть использованы для изучения эмоционального благополучия [5–9]. Исследования, опирающиеся на данные Twitter, на основе анализа публикуемых твитов строят либо индекс настроений, либо степень удовлетворения/неудовлетворения жизнью.

4. **Facebook status.** Авторы [10–12] предлагают методику прогнозирования благополучия с использованием социальных сетей. На основе анализа статусов, открытых сообщений определяется семантическая соотнесенность ключевых слов в сообщениях, далее строится агрегированный индекс благополучия.

Большие данные предоставляют огромные возможности для понимания человеческих взаимодействий в масштабе общества с богатой пространственной и временной динамикой, а также для идентификации сложных взаимодействий и нелинейностей между переменными. Одним из наиболее популярных ресурсов больших данных является Google Trends.

Согласно исследованию [13] отмечаются следующие преимущества данных Google Trends:

1. Высокая частота/периодичность. Можно наблюдать, как меняются настроения, предпочтения пользователей каждый день.

2. Поисковые запросы лучше выявляют установки индивида по сравнению с традиционными опросами. Многие респонденты отвечают на вопросы анкеты по альтруистическим причинам, так как нет мотива отвечать на вопросы откровенно и глубоко [14]. Поисковые запросы могут больше раскрыть личной информации. Например, тема потери работы может быть весьма чувствительной для респондента, по которой он может не иметь желания общаться.

С другой стороны, объем поиска по словам «найти работу», «поиск работы» показывает озабоченность этой проблемой для человека. Отсюда делается вывод, что полученные из поисковых запросов данные являются более объективными.

Многие исследователи, руководители международных организаций видят в перспективе объективную замену данных дорогостоящей официальной статистики, которые доходят с большим опозданием до пользователей, большими данными в целях проведения различных оперативных мониторингов условий жизни населения, достижения целей устойчивого развития ООН и др. Одним из успешных примеров использования больших данных для целей мониторинга является исследование J. Ginsberg и др. [15], в котором показано, что на основе Google Trends отслеживать и предсказывать распространение гриппа можно раньше центров по контролю и профилактике заболеваний.

Вместе с тем существует ряд методологических трудностей использования больших данных. Автор исследования [16] отмечает несопоставимость больших данных с данными официальной статистики вследствие применения разных методологий и классификаций. Однако приведение больших данных в соответствие с требованиями национальных и международных рекомендаций уменьшит их преимущества в части оперативности использования, своевременности и актуальности, что в настоящее время обеспечивает им высокую экономическую эффективность.

Необходимым условием использования больших данных является широкий доступ населения к сети Интернет. Несмотря на стремительное развитие интернета в последнее десятилетие, возможности больших данных в развитых странах выше, чем в развивающихся [16]. Так, согласно данным официальной статистики в 2019 г. в России доля населения, использовавшая сеть Интернет каждый день, составила 73%, в Москве — 82%. Этот факт может привести к смещенным оценкам исследуемых переменных, так как достоверность результатов гарантируется не только большим количеством наблюдений, а в первую очередь — репрезентативностью выборочной совокупности.

Серьезная методологическая работа и высокие риски использования больших данных формируют преграды для их успешной интеграции в официальную статистику. В России существуют некоторые пилотные проекты по использованию интернет-ресурсов для улучшения статистики потребительских цен, данных мобильных операторов для статистики туризма, наблюдения спутниковой связи в целях развития статистики окружающей среды [17]. Однако следует отметить относительно узкую сферу их применения и отсутствие опыта практического использования (официальных публикаций не представлено).

В то же время существуют примеры успешного внедрения больших данных в официальную статистику в некоторых развитых странах. Статистические службы Нидерландов в 2015 г. расширили транспортную статистику публикацией показателей, которые рассчитывались на основе информа-

ции, полученной с датчиков на автомагистралях страны[1]. Оперативные данные позволили принять своевременные решения, когда в северной части страны образовался гололед. Примером подобной интеграции является опыт Канады: прогнозирование урожайности сельскохозяйственных культур производится не только на основе результатов опросов фермеров, но и информации о состоянии земельных угодий, климате, полученной по спутниковой связи [18]. Также ведется активная работа по привлечению больших данных в качестве альтернативного источника информации о потребительских ценах в Дании, Нидерландах, Италии, Норвегии, Австралии, Швейцарии, Бельгии, Новой Зеландии, Швеции.

Так как опыт внедрения больших данных в практику органов государственной статистики уже имеется, необходимо продолжать исследования в области принципов интеграции двух информационных источников, а не ограничиваться лишь общими проектами по изучению потенциала больших данных.

В этой связи целью проводимого исследования является развитие теории статистического наблюдения в части научно-методологических подходов к обработке больших данных и определение возможностей интеграции информационных ресурсов различного типа в отношении измерения сложных латентных категорий (на примере социальной комфортности).

Введение новой экономической категории «социальная комфортность» необходимо для «*определения реального уровня благосостояния населения в динамике, оценки истинного качества жизни людей*» [19]. Несмотря на новизну исследования данного процесса в российской практике, в зарубежных исследованиях приводится анализ комфортных условий для индивида в следующих областях: географии, социологии, медицине, психологии, экономике, финансах. В работе [19] подробно рассматривается аксиоматика и состав вводимой категории. Входящие категории социальной комфортности: здоровье и медицинское обслуживание, образование и обучение, социальная поддержка и пенсионное обеспечение, финансовое положение, трудовая деятельность, жилье и условия проживания, этические нормы и ценности, безопасность, политическая стабильность, отдых и свободное время, экология и окружающая среда, инфраструктура.

В настоящей работе на основе использования ресурсов Федеральной службы государственной статистики (ФСГС) и больших данных предлагается оценить социальную комфортность, определить степень согласованности ее компонентов, построенных по двум источникам данных и выявить возможные направления интеграции.

ЭМПИРИЧЕСКИЕ ПОДХОДЫ К СЕМАНТИЧЕСКОМУ ПОИСКУ ИНФОРМАЦИИ О ФИНАНСОВО-ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЯХ НА ОСНОВЕ ИНТЕРНЕТ-ЗАПРОСОВ

Большинство проводимых исследований продемонстрировали перспективность использования больших данных. Однако до сих пор остается актуальной проблема отбора ключевых слов для определения пользовательских запросов. Во многих работах формирование ключевых слов идет на основе небольшого числа пользовательских запросов вокруг интуитивных предположений авторов исследования о важности того или иного запроса для индивида. Как было замечено ранее, первые работы по внедрению больших данных в статистическую практику и расчеты были посвящены финансовым аспектам.

Так, в работе [20] проводится прогнозирование инфляции на основе интернет-запросов. Автор рассматривает 75 поисковых запросов, которые связаны с финансовыми рынками, а также заинтересованностью населения, с экономическими и финансовыми явлениями и процессами. Данные запросы были отобраны из числа самых популярных на основе анализа корреляционных зависимостей с инфляцией.

Автор исследования [21] инфляционные ожидания анализирует с помощью ключевого слова «инфляция», на основе которого строятся всевозможные поисковые запросы. Далее автор проводит сравнительную оценку с результатами инфляционных ожиданий населения на основе данных социологических опросов.

В работе [2] для каждого индекса субъективного благосостояния, который представлен индексом положительных и отрицательных эмоций, обоснован набор прокси переменных из поисковых запросов Google (таких как счастье, уважение, стресс, беспокойство и т.д.). Выбор в первую очередь определяется направлением вызываемых эмоций: положительным или отрицательным.

Автор исследования [22] сервис Google Trends использует как прокси для предсказания вола-

[1] A13 busiest national motorway in the Netherlands. URL: https://www.cbs.nl/-/media/imported/documents/2015/31/a13-busiest-national-motorway-in-the-netherlands.pdf?la=en-gb (дата обращения: 02.09.2021).

тильности цены энергоносителей. Авторы начинают исследование с набора из 90 терминов, которые используются в энергетическом секторе. К отобранным на основе профессиональной литературы терминам добавляются самые популярные слова по данной тематике из Google. Фильтрация подобного набора слов происходит посредством построения различных моделей, которые лучше всего прогнозируют волатильность цены сырой нефти, мазута, бензина, природного газа.

Автор исследования [13] строит список поисковых терминов, раскрывающих отношение к экономическим условиям. Для того чтобы максимально объективно определить поисковые слова, автор начинает работу со словаря по финансам и текстовой аналитике [23] и подбирает слова, которые классифицируются как «экономические» слова, имеющие положительное или отрицательное влияние на настроение индивида. В исследованиях [13, 23, 24] в качестве обоснования выбора слов для создания агрегированного индекса настроения инвесторов используют гарвардский психосоциальный словарь (Harvard IV-4 Dictionaries), который имеет несколько редакций. Данный словарь был разработан Декстером Данфи и его коллегами [25–27].

В результате формируется список из 149 слов, таких как «инфляция», «рецессия», «безопасность» и др. Далее каждое из 149 слов вводится в Google Trends, который для каждого слова подбирает десять лучших (наиболее популярных) поисковых запросов, Например: взятое из словаря слово «дефицит» приводит к следующим поисковым запросам: «дефицит бюджета», «дефицит внимания», «торговый дефицит» и т.д. В результате 149 слов увеличилось до 1490. На последнем этапе исключаются те слова/словосочетания, предоставленные Google, которые не имеют отношения к экономическим условиям или финансам и обладают нулевым объемом поиска.

После процедуры статистической обработки полученных временных рядов поисковых запросов строится финансово-экономический индекс как новая мера определения и прогнозирования настроения инвесторов. Замечена высокая прогностическая сила индекса и определены возможные перспективы практического применения.

Стоит отметить несколько работ, которые на начальном этапе берут произвольный набор слов-поисковых запросов, а в дальнейшем отбирают наиболее информативные с помощью байесовской модели усреднения [3, 28].

В исследовании [29] авторы с целью выявления детерминант сберегательного поведения населения в странах ЕС используют информацию трех различных типов: макроэкономическую статистику (номинальный эффективный обменный курс, номинальный ВВП, показатели инфляции и пр.), поисковые слова Google (42 слова), которые отражают настроения экономических агентов, поведенческие, психологические факторы, влияющие на предпочтения, а также данные социологических опросов, отражающие ожидания о настоящей и будущей финансово-экономической ситуации. Выбор всех переменных для анализа основан на экономической интуиции авторов. В качестве модельного инструментария используется модель байесовского усреднения. Обоснование ее применения авторы объясняют отсутствием стратегии отбора ключевых слов Google.

ПОДХОДЫ К ОБРАБОТКЕ БОЛЬШИХ ДАННЫХ

Большие данные генерируются самими пользователями. В отличие от данных официальной статистики они не собираются по специально разработанной и утвержденной методологии. В этой связи для их адекватного использования и интеграции в официальную статистику необходимо разработать специальную методологию по сбору и обработке. Далее проанализируем существующий опыт по обработке больших данных в зарубежных исследованиях.

В исследовании [30] авторы анализируют частоту поисковых запросов, связанных с туризмом в Германии, на основе Google Trends и предлагают способы очистки данных с целью исключения ложных прогнозов.

Согласно источнику [31] Google Trends генерируются следующим образом: находится отношение числа интернет-запросов по определенному ключевому слову на момент времени t в регионе r *к общему числу запросов* на момент времени t в регионе r, и найденное отношение умножается на 100 для стандартизации.

Авторы предлагают следующую модификацию исходных данных: найти отношение числа интернет-запросов по определенному ключевому слову на момент времени t в регионе r *к среднему значению* объема интернет-запросов ключевых слов на момент времени t в регионе r. Полученные модифицированные данные называют средними, деленными на анализируемые категории. Однако подобное преобразование вызывает ряд проблем:

- временной ряд имеет более сильную сезонность в результате своеобразных сезонных вариаций каждой временной серии в отдельности;
- временной ряд имеет большее количество выбросов вследствие возрастания роли отдельных временных рядов в знаменателе.

Для преодоления озвученных выше проблем авторы предлагают находить отношение числа интернет-запросов по определенному ключевому слову на момент времени *t* в регионе *r* на средний тренд интернет-запросов ключевых слов на момент времени *t* в регионе *r*. Тренд определяется согласно декомпозиции временного ряда [32]. Модифицированные временные ряды называют делением на средний тренд.

Исследование [24] ориентировано на подтверждение возможности поисковых запросов населения, основанных на настроении пользователей, оказывать влияние на фондовый рынок Португалии. В качестве способов обработки запросов автор предлагает логарифмировать поисковый запрос по конкретному слову за неделю, а затем найти первые разности между объемами запросов по конкретному слову за два промежутка времени. Чтобы обеспечить сопоставимость и согласуемость данных, автор предлагает сделать поправку на сезонность и гетероскедастичность. Для очистки от выбросов выборка поиска делается цензурированной, т.е. отсекается 5% выборки с наименьшим и наивысшим объемом поиска. Для проверки сезонности используется метод дисперсионного анализа, где тестируется гипотеза о равенстве средних за 12 месяцев. Запросы с ярко выраженной сезонностью были очищены путем построения регрессионных моделей с 12 фиктивными переменными (за каждый месяц), в которых были найдены остатки и использованы на последующих итерациях анализа. Для устранения гетероскедастичности используется процедура стандартизации с использованием среднего квадратического отклонения.

В работе [33] предлагается унификация поисковых запросов на стандартное отклонение и процедура десезонолизация для очистки от сезонных колебаний, осуществляемая через пакет seasonal на языке R.

Автор исследования [34] предлагает для приведения временных рядов поисковых запросов к стационарному виду использовать логарифмирование с дальнейшим взятием первых разностей.

Проводятся следующие процедуры по очистке поисковых запросов Google Trends [2]:

1. Стандартизация при помощи Z-оценки, в результате чего распределение данных по различным поисковым запросам приводится к одному масштабу, что позволяет проводить сопоставления.
2. Устранение резких скачков популярности посредством скользящей средней.
3. Удаление поисковых запросов с продолжительным нулевым объемом поиска.
4. Очищение временного ряда от тренда при помощи построения регрессий с временным трендом и удаление поисковых запросов с коэффициентом детерминации выше 0,6.
5. Очищение временного ряда от сезонности при помощи построения регрессий с месячными фиктивными переменными и использование в дальнейшем остатков регрессионных моделей.

Данный методологический подход к обработке больших данных видится наиболее комплексным и последовательным. Основные этапы описанного подхода будут использованы в рамках проводимого исследования.

ОТБОР КЛЮЧЕВЫХ СЛОВ И ОБРАБОТКА ПОИСКОВЫХ ЗАПРОСОВ GOOGLE TRENDS

Основная проблема большинства исследований, посвященных Google Trends, заключается в том, что подбор ключевых слов для семантического раскрытия того или иного социально-экономического процесса или явления подбирается интуитивно, опираясь на опыт автора. Далее используется ряд эконометрических методов (модель байесовского усреднения, факторный анализ, корреляционный анализ и пр.) с целью выбора наиболее информативных поисковых запросов, отражающих анализируемый индекс или другой индикатор. Существенным отличием проводимого исследования является обоснованный подход к семантическому поиску информации (на основе поисковых запросов Google Trends) о компонентах социальной комфортности, который заключается в использовании гарвардского психосоциального словаря (Harvard IV-4 Dictionaries). Особенность данного словаря состоит в том, что он помогает решить проблему неоднозначности отнесения слов к тем или иным категориям. К примеру, в словаре присутствуют группировки слов по следующим категориям: слова позитивного мировоззрения, негативного мировоззрения, слова радости, боли, добродетели, порока; слова, характеризующие социальные категории (образование, финансы, труд и т.д.); мотивационные

слова и др[2]. В этой связи в проводимом исследовании каждая из двенадцати компонент социальной комфортности наполняется ключевыми словами из гарвардского словаря, затем проводится анализ степени соответствия реалиям российской действительности (такие слова, как каноэ, ковбой, День благодарения, День независимости, Конституционный Конвент, Биль о правах, жюри присяжных — были удалены).

Для того чтобы приблизить набор ключевых слов-запросов пользователей к реальным условиям повседневной российской жизни были добавлены следующие слова: ТСЖ, МРОТ, ЕГЭ, поправки в Конституцию, единый день голосования, ЖКХ, ОМС, ДМС. Безусловно, доля авторского субъективизма не исключена, но она составляет менее 26,5% от общего числа слов. Кроме того, стоит отметить, в набор ключевых слов не включались запросы-глаголы: болеть, искать, проигрывать, молиться и пр., а также имеющие несколько лексических значений: Вена (город) и вена (сосуд), вера (доверие) и Вера (имя), железо (витамин) и железо (металл) и др.

Пример наполнения каждого блока социальной комфортности поисковыми запросами можно увидеть в *табл. 1*, где представлена лишь некоторая часть сформированного набора ключевых слов.

Обработка больших данных начинается с их приведения к сопоставимому виду.

На первом этапе анализа больших данных, как показывает опыт проведенных исследований, необходимо проводить масштабирование для обеспечения сопоставимости и согласуемости исходных данных. Одним из самых распространенных способов стандартизации является поправка на стандартное отклонение [2, 33], а другие исследователи [24, 34] используют логарифмирование. В настоящей работе мы будем опираться на вариант нормировки, предложенный С.А. Айвазяном [35] при построении сложных синтетических латентных категорий качества жизни населения:

$$\tilde{x}_{j,t} = \frac{x_{j,t} - x_{j,min}}{x_{j,max} - x_{j,min}} N, \qquad (1)$$

где $\tilde{x}_{j,t}$ — унифицированное значение поискового запроса ($j = 1, 2, \ldots, p; t = 2010, 2011, \ldots, 2021$);

$x_{j,min}, x_{j,max}$ — максимальные и минимальные значения частных индикаторов; $N = 10$.

$$\tilde{x}_{j,t} = \frac{x_{j,max} - x_{j,t}}{x_{j,max} - x_{j,min}} N. \qquad (2)$$

Первый вариант (1) нормировки используется в случае положительного восприятия поискового запроса индивидом, второй вариант (2) нормировки — в случае негативного восприятия. Информация о принадлежности слова к позитивному или негативному мировоззрению имеется в гарвардском психосоциальном словаре (Harvard IV-4 Dictionaries). Слова, которые не представлены в словаре Harvard IV-4 Dictionaries, были разделены авторами по группам самостоятельно.

Стоит отметить следующие закономерности между ростом интереса к тому или иному запросу и его позитивной/негативной оценкой (*табл. 2*).

На втором шаге посредством скользящей средней снижается риск переобучения модели, вызванный резкими скачками. Например: «Конституция» (резкие скачки вызваны голосованием за поправки в Конституцию в 2020 г.) или «футбол» (скачки вызваны проведением чемпионата мира по футболу в 2018 г.) и пр.

Скользящая средняя (MA) порядка q определяется следующим образом:

$$x_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}. \qquad (3)$$

При этом порядок q рассчитывается с учетом числа предыдущих значений случайных отклонений $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$. В проводимом исследовании окно сглаживания будет равно трем месяцам, соответственно, порядок q скользящей средней равен трем [MA(3)].

На третьем шаге производится очищение временного ряда от тренда. Необходимость данной операции вызвана тем, что сильный временной тренд способен привести к неадекватному прогнозу социальной комфортности, а также при нахождении корреляций в процессе группировки слов в блоке может возникнуть проблема составления ненадежных блочных индикаторов социальной комфортности, так как общим у слов будет являться одинаковый временной тренд, а не смысловая нагрузка. В этой связи строится уравнение регрессии зависимости поискового запроса от тренда вида:

$$y_t = \theta_1 t + \varepsilon_t, \qquad (4)$$

где $\theta_1 t$ — тренд; ε_t — случайная величина, характеризующая отклонение уровня от тренда.

[2] URL: http://www.wjh.harvard.edu/~inquirer/homecat.htm (дата обращения: 02.09.2021).

Таблица 1 / Table 1

Поисковые запросы Google Trends, характеризующие компоненты социальной комфортности / Google Trends searches describing social comfort components

	Название блока / Block	Примеры поисковых запросов / Examples of search queries	Количество слов в блоке / Number of words in a block
1	Здоровье и медицинское обслуживание	Близорукость, депрессия, ЗОЖ, нервы, правильное питание, скорая помощь, диета и др.	38
2	Образование и обучение	Выпускник, ЕГЭ, бакалавриат, математика, наука, аспирантура, тайм-менеджмент, дистанционные курсы, репетиторы и др.	42
3	Социальная поддержка и пенсионное обеспечение	Пенсия, социальная защита, трудовой стаж, пособие, социальная реабилитация, дом престарелых, социальная поддержка и др.	23
4	Финансовое положение	Кредит, заработная плата, банкротство, вклад, инфляция, собственность, кризис, обязательство, бедность, валюта и др.	37
5	Трудовая деятельность и условия труда	HEADHUNTER, без работы, высокооплачиваемая работа, карьера, повышение квалификации, рынок труда, трудовые отношения, ярмарка вакансий, увольнение и др.	53
6	Жилье, условия проживания	Квитанция, ТСЖ, аренда квартиры, рынок недвижимости, цены на квартиры, соседи, управляющая компания и др.	13
7	Этические нормы и ценности	Атеизм, библия, патриотизм, мораль, репутация, расизм, свобода слова, смысл жизни, христианство, феминизм, честь и др.	52
8	Безопасность	Атака, безопасность, дискриминация, воры, издевательство, кража, мошенничество, наказание, несчастный случай, самоубийство, хищения и др.	60
9	Политическая стабильность	Бюрократия, демократия, гражданство, голодовка, конституция, парламент, чиновники, голос, легитимность, взятки, экстремизм и др.	53
10	Отдых и свободное время	Рыбалка, волейбол, отпуск, отдых на море, художественная литература, пикник, пансионат, зоопарк, музеи, отель и др.	54
11	Экология и окружающая среда	Атмосфера, окружающая среда, переработка мусора, экология, качество воды, природа, климат, парниковый эффект, отходы производства, очистные сооружения, экологическое воспитание, лесные ресурсы и др.	22
12	Инфраструктура (транспорт, связь, интернет)	сотовый оператор, онлайн-магазин, 5G, самокат, велосипед, трасса, транспортное средство, поезд, дорога, мост, аэропорт, автомобиль и др.	28

Источник / Source: составлено авторами на основе Google Trends / compiled by the authors based on Google Trends.

На основе построенного уравнения рассчитывается скорректированный коэффициент детерминации. Модели поисковых запросов со значением данного коэффициента выше 0,6 будут удалены [2].

На четвертом шаге происходит десезонализация поисковых запросов. Наличие экстремальной сезонности в запросах может привести к сильной корреляции между собой, вызванной соответствием одной и той же сезонной схеме. Для десезонализации используется модуль statsmodels языка программирования Python. Данный модуль включает в себя множество классов и функций для оценки различных статистических моделей, а так-

Таблица 2 / Table 2

Влияние на социальную комфортность снижения интереса к словам позитивного и негативного мировоззрения / Impact on social comfort of a decrease in interest in words of a positive and negative worldview

	Рост интереса (увеличение поисковых запросов) / Growth of interest (increase in search queries)	Снижение интереса (уменьшение поисковых запросов) / Decrease in interest (decrease in search queries)
Слова позитивного мировоззрения	**Увеличивает** социальную комфортность	**Увеличивает** социальную комфортность. Интерес к словам позитивного мировоззрения может быть снижен в силу достижения определенного уровня комфорта, который устраивает индивида в отношении данного процесса, характеризующегося поисковым запросом. ***Пример:*** *снижение запросов по слову «интернет» связано с доступностью и хорошим качеством работы интернета в последние годы*
Слова негативного мировоззрения	**Снижает** социальную комфортность	**Увеличивает** социальную комфортность

Источник / Source: составлено авторами на основе Google Trends / compiled by the authors based on Google Trends.

же для проведения статистических тестов и исследования статистических данных. В частности, в данный модуль входит метод Seasonal-Trend Decomposition Procedure Based on Loess (STL) — способ разложения временного ряда на сезонную и трендовую составляющие и на остатки через метод локальных регрессий.

Метод STL раскладывает временной ряд на составляющие аддитивной модели:

$$Y_t = T_t + S_t + E_t, \quad (5)$$

где T_t — трендовая составляющая; S_t — сезонная составляющая; E_t — остаток.

Исключение сезонности происходит путем вычитания сезонной компоненты S_t из временного ряда.

В результате последовательной реализации перечисленных итераций анализа и обработки данных было отобрано 475 слов (из 574 слов до обработки) за период январь 2010 — январь 2021 г. (периодичность — 1 месяц).

Предложенная методология поиска ключевых слов и обработки Google Trends была реализована в виде программного кода Python. Она позволяет построить статистические ряды с соблюдением базовых принципов, обеспечивающих качество статистики — сопоставимости, согласуемости, точности и однородности данных.

ПОСТРОЕНИЕ ИНТЕГРАЛЬНЫХ ИНДИКАТОРОВ СОЦИАЛЬНОЙ КОМФОРТНОСТИ: ДАННЫЕ ОФИЦИАЛЬНОЙ СТАТИСТИКИ И GOOGLE TRENDS

Большие объемы используемой информации вызывают необходимость использования специальных методов агрегирования и снижения размерности. Наиболее популярными являются факторный анализ и метод главных компонент. В нашем исследовании мы будем использовать метод модифицированной главной компоненты, более подробно представленный в работе С.А. Айвазяна [35]. Согласно данному методу, показатели внутри каждого из 12 блоков[3] социальной комфортности агрегируются в блочные индикаторы, которые впоследствии объединяются в сводный интегральный индикатор социальной комфортности. Поскольку в исследовании были использованы два типа данных, соответственно, на выходе мы получили блочные индикаторы и сводный интегральный индикатор социальной комфортности, построенные по данным Google Trends и официальной статистики.

В целях гармонизации и взаимоувязки информационных ресурсов различных типов: GT

[3] Обоснование блочных индикаторов социальной комфортности более подробно приведено в исследовании [19].

Рис. 1 / Fig. 1. **Блочные индикаторы социальной комфортности Москвы за 2010–2019 гг. / Block indicators of social comfort in Moscow for 2010–2019**

Источник / Source: составлено авторами / compiled by the authors.

и данных официальной статистики, представленной на портале ФСГС, предлагается ко всем данным использовать нормировку, осуществляемую по формулам (1) и (2), а для GT данных применить описанные выше процедуры обработки по сглаживанию, исключению тренда и т.д.

Далее обсудим результаты моделирования социальной комфортности, полученные по разным типам данных, и оценим перспективность использования GT в отношении сложных латентных категорий (на примере социальной комфортности).

Результаты моделирования по данным ФСГС

Информационной основой для наполнения блоков социальной комфортности стали показатели социально-экономических условий регионов России за 2010–2019 гг., взятые с сайта Росстата[4]. При выборе индикаторов мы руководствовались подходом к анализу контекстуальных условий РФ и ее регионов (подробнее изложенного в исследовании [19]), а также требованиями, предъявляемыми к набору частных критериев синтетической латентной категории [35]. Большое внимание уделялось степени соответствия социально-экономического содержания показателя непосредственно измеряемой латентной категории («Безопасность», «Жилье и медицинское обслуживание» и пр.), валидности, доступности в официальном источнике. В этой связи такие блоки социальной комфортности, как «Этические нормы и ценности», «Политическая стабильность», остались пустыми в связи с высокой долей субъективизма входящих в них индикаторов и отсутствием информации на официальном ресурсе государственной статистики. В эмпирическую базу исследования вошла панель из 100 показателей за 2010–2019 гг. Собранные индикаторы измеряются по количественной шкале в соответствии с единой методологией и базовыми принципами статистического наблюдения, которые обеспечивают согласуемость и сопоставимость объектов наблюдения.

Поскольку нормировка входящих в панель показателей была осуществлена по 10-балльной шкале, то в результате расчетов по методу модифицированной главной компоненты на выходе значения блочных индикаторов социальной комфортности также будут принадлежать отрезку от 0 до 10.

Рассмотрим в качестве примера результаты моделирования социальной комфортности Москвы за 2010–2019 гг. (*рис. 1*).

Сводный интегральный индикатор по Москве в 2019 г. составил 6,815, что является мак-

[4] Федеральная служба государственной статистики Российской Федерации. URL: https://rosstat.gov.ru/ (дата обращения: 02.09.2021).

симальным значением в 2019 г. среди прочих регионов России и на 13% выше уровня 2010 г. Данная динамика объясняется при незначительном изменении весовых коэффициентов ростом блочных интегральных индикаторов.

За анализируемый промежуток времени 2010–2019 гг. Москва улучшила позиции по восьми из десяти представленных компонент социальной комфортности. Самые значительные изменения произошли по компоненте «Финансовое положение» (+1,3 балла), «Инфраструктура» (+2,1 баллов), «Отдых и свободное время» (+4 балла), «Социальная поддержка и пенсионное обеспечение» (+5,2 балла). Отсутствует динамика по блоку «Здоровье» и «Безопасность».

Причины произошедших изменений видятся в следующем:

- по блоку «Финансовое положение»: валовый региональный продукт на душу населения в Москве является одним из максимальных в абсолютном выражении и за анализируемый период увеличился в примерно 2 раза. Кроме того, значительно снизилась доля численности населения с денежными доходами ниже величины прожиточного минимума (с 10 до 4%). При этом структура ВРП Москвы более 50% формируется за счет сферы услуг;
- по блоку «Инфраструктура»: Москва является первым регионом в России, где использование мобильного высокоскоростного интернета LTE и 4G получило широкое применение среди населения города, что в целом способствовало росту цифровизации социально-экономических процессов. В частности, доля населения, ежедневно использовавшая интернет, увеличилась с 60% в 2014 г. до 82% в 2019 г.;
- по блоку «Отдых и свободное время»: в Москве функционирует более 18 тыс. спортивных объектов, что в 2 раза выше уровня 2010 г., а общероссийский уровень увеличился на 11%;
- по блоку «Социальная поддержка и пенсионное обеспечение»: в Москве проживает 8,5% населения России, данный показатель в период 2010–2019 гг. практически не изменился, при этом доли исполнения бюджетов Пенсионного фонда и Фонда социального страхования по статье «расходы» Москвы в структуре России увеличились на 1,5 и 4,2% соответственно.

Согласно используемому подходу свертки многоаспектных категорий [35] значение сводного интегрального индикатора социальной комфортности будет определяться по формуле:

$$Y_{t,\text{св.}} = 10 - \sqrt{\sum_{j=1}^{12} v_j (y_{j,t} - 10)^2}\ , \qquad (6)$$

где v_j — это нормированные неотрицательные веса, определенные долей объясненной дисперсии первой главной компоненты каждого из 12 блоков[5]; $y_{j,t}$ — блочный индикатор социальной комфортности в год t.

Чем выше вес блочного индикатора, тем большее влияние он имеет на сводный индикатор социальной комфортности. На основе данного метода можно определять приоритеты социально-экономической политики в целях улучшения социальной комфортности.

Далее проанализируем работоспособность данного метода на основе сравнения значений блочных индикаторов и их веса в сводном индикаторе для двух регионов (Республика Бурятия и Тульская область).

Рассмотрим темпы прироста значений блочных индикаторов социальной комфортности за 2010–2019 гг. (*рис. 2*), а также вес каждой компоненты в сводном индикаторе[6]. Можно заметить, что приоритетными направлениями повышения социальной комфортности являются инфраструктура (вес в сводном индикаторе — 22%); здоровье и медицинское обслуживание (21,2%); экология и окружающая среда (13%); жилье, условия проживания (12%). При этом сводный индикатор социальной комфортности в Тульской области вырос на 8,4%, а в Республике Бурятия — лишь на 0,5% вследствие опережающего роста блочных индикаторов социальной комфортности Тульской области, которые являются приоритетными (инфраструктура, жилье). Полученные выводы могут стать основой для проведения мониторинга социальной комфортности и последующей корректировки проводимой социально-экономической политики региона в рамках приоритетных факторов.

Сравнение результатов моделирования по данным GT и официальной статистики

Согласно результатам моделирования по данным Google Trends сводный интегральный индикатор по России с 2010 по 2019 г. увеличился на 54%

[5] По данным официальной статистики было построено 10 блоков в связи с отсутствием информации по блокам «Этические нормы и ценности», «Политическая стабильность», а по данным Google Trends — 12 блоков.

[6] Вес блочных индикаторов является усредненной величиной за 2010–2019 гг. и постоянной для каждого объекта наблюдений.

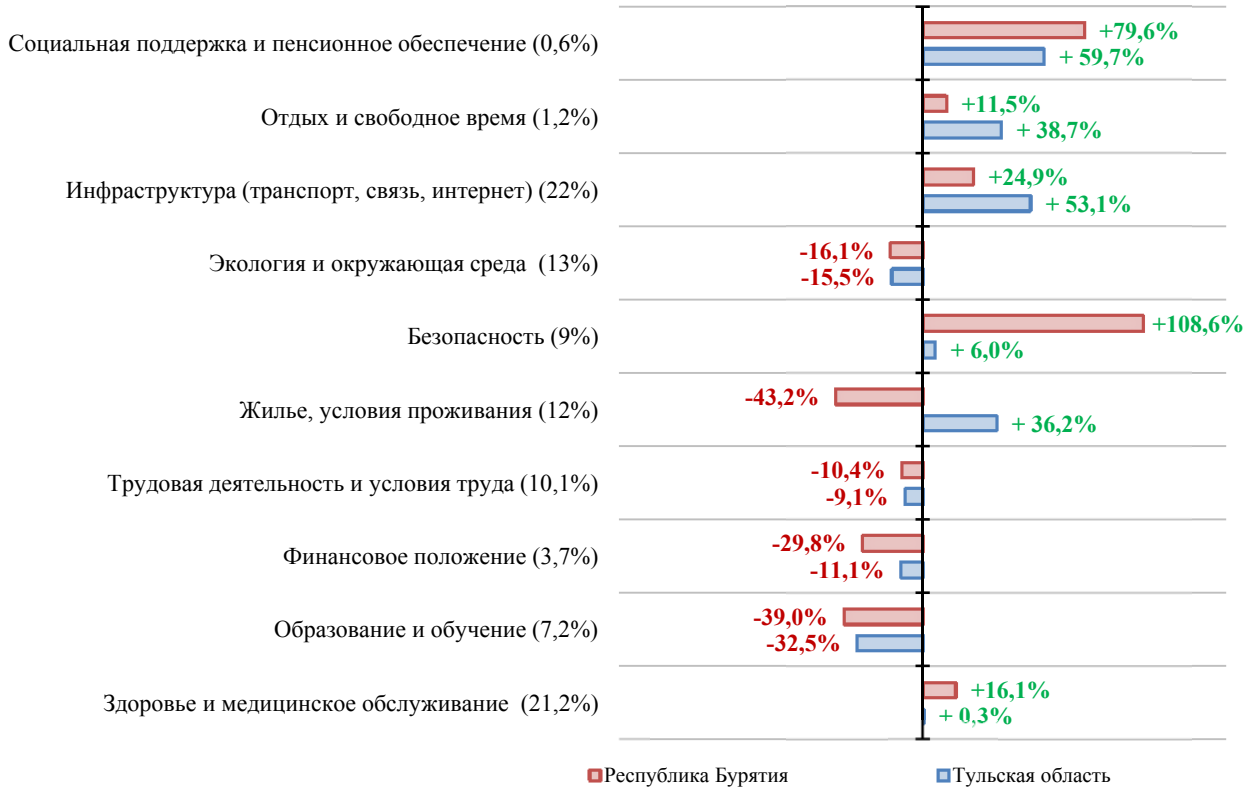

Рис. 2 / Fig. 2. **Сводный индикатор социальной комфортности в разрезе отдельных регионов / Composite indicator of social comfort in regions**

Источник / Source: составлено авторами / compiled by the authors.

и составил 4,631 балла. Субъективная оценка населения, агрегированная на основе поисковых запросов, ниже оценки, полученной на основе данных ФСГС (5,371).

Посредством применения разработанной методологии по обработке Google Trends (1)–(5) были рассчитаны значения блочных индикаторов социальной комфортности и осуществлен сравнительный анализ с аналогичными индикаторами по данным официальной статистики. В *табл. 3* представлены корреляции по блочным индикаторам социальной комфортности.

По данным анализируемой таблицы очевидна значимая корреляция всех блочных индикаторов социальной комфортности. При этом устойчивая положительная линейная зависимость наблюдается по семи из десяти сравниваемых блоков социальной комфортности. **Сильная** положительная корреляция отмечается по блочным индикаторам «Социальная поддержка и пенсионное обеспечение», «Экология и окружающая среда»; **умеренная** положительная корреляция — «Здоровье и медицинское обслуживание», «Трудовая деятельность и условия труда», «Инфраструктура», «Жилье, условия проживания», «Финансовое положение».

«Финансовое положение», согласно расчетам, формируется преимущественно словами «кредит», «вклад», «прибыль», «инфляция», «накопление», «наличные» и пр. Таким образом, блок «Финансовое положение» определяется преимущественно словами, характеризующими процессы поступления/получения денежных средств населением.

При этом аналогичная компонента социальной комфортности в официальной статистике оценивается показателями: «потребительские расходы», «коэффициент фондов», ИПЦ, доля расходов на питание и пр. В целом официальная статистика преимущественно рассматривает население с позиции формирования расходной части ассортимента товаров и услуг.

Повысить корреляцию можно посредством дополнения перечня индикаторов официальной статистики, характеризующих финансовое положение населения в части формирования доходной его части, например: средняя доходность по банковским вкладам, доходы, полученные от сделок на финансовых рынках, структура формирования располагаемых денежных доходов населения и пр.

Дискуссионным вопросом является сильная отрицательная корреляция по блокам «Образование и обучение», «Безопасность», «Отдых и свободное время». Далее обсудим возможные причины сложившихся зависимостей.

Показатели блока «**Образование**» по данным официальной статистики образованы количественными индикаторами охвата средним, средним специальным и высшим образованием. Большинство показателей данного блока (примерами могут быть «Доля учащихся в высших учебных заведениях в составе численности населения трудоспособного возраста», «Численность аспирантов» и пр.) демонстрируют отрицательную динамику. Подобная тенденция, начавшаяся с 2010 г., характерна для большинства российских регионов [36] и связана с демографическими проблемами в стране. В этой связи можем наблюдать отрицательную динамику блочного индикатора «Образование», что противоречит динамике индикатора «Образование» по данным GT. Стоит заметить, что существенным недостатком данных, публикуемых на портале ФСГС в части данного блока, является то, что они не отражают в полной мере уровень интеллектуального развития региона/страны, который обеспечивает конкурентоспособность, меняет стандарты жизни и уровень социальной комфортности. В этой связи показатели охвата образованием должны быть дополнены показателями, характеризующими успеваемость школьников, студентов, качеством сданных выпускных/вступительных испытаний, международных тестов (GMAT, IELTS и др.); числом победителей олимпиад; показателями привлекательности/ открытости/престижности университета; доступности получения образования (дошкольного, среднего, высшего).

В работе [37] подробно излагаются преимущества и объективность использования нового подхода к оценке уровня интеллектуального капитала страны на основе использования новых показателей оценки уровня образования в стране. Использование других источников информации: опросы, данные GT позволяют расширить показатели охвата образованием в России качественными характеристиками.

Пример сильной обратной корреляции указывает на необходимость расширения статистики образования качественными индикаторами, в частности рассмотрения возможностей использования альтернативных источников — данных GT, что увеличит объективность и качество предоставляемой информации. Так, проведенный анализ показал, что наиболее популярными и весомыми в блоке «образование» поисковыми запросами являются слова «дистанционные курсы», «английский язык», «математика», «успеваемость». Растущий интерес населения по данным тематикам свидетельствует о росте интеллектуального капитала населения.

Таблица 3 / Table 3

Коэффициенты корреляции Пирсона блочных индикаторов социальной комфортности, построенных по данным официальной статистики и Google Trends / Pearson correlation coefficients of block indicators of social comfort based on official statistics and Google Trends data

	Коэффициент корреляции Пирсона / Pearson correlation coefficient	Значимость (p-value) / P-value
Здоровье и медицинское обслуживание	0,59	0,0700
Образование и обучение	–0,90	0,0003
Социальная поддержка и пенсионное обеспечение	0,88	0,0007
Финансовое положение	0,46	0,0176
Трудовая деятельность и условия труда	0,58	0,0813
Жилье, условия проживания	0,35	0,0317
Безопасность	–0,77	0,0100
Отдых и свободное время	–0,76	0,0104
Экология и окружающая среда	0,70	0,0250
Инфраструктура (транспорт, связь, интернет)	0,41	0,0208

Источник / Source: составлено авторами / compiled by the authors.

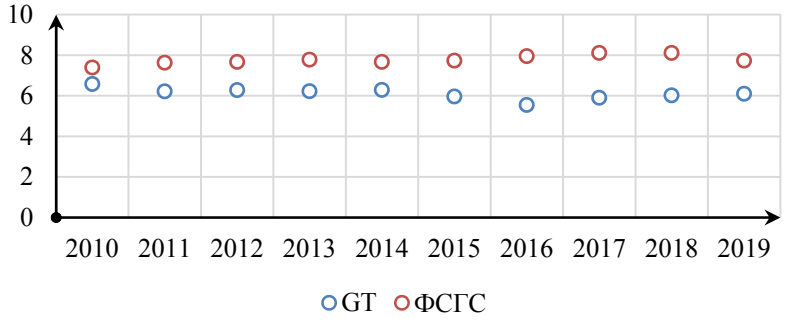

Рис. 3 / Fig. 3. **Сравнительная динамика блочных индикаторов «Безопасность», построенных по данным официальной статистики и Google Trends / Comparative dynamics of the "Safety" block indicators based on official statistics and Google Trends data**

Источник / Source: составлено авторами / compiled by the authors.

Рис. 4 / Fig. 4. **Динамика компоненты «Отдых и свободное время» по данным Google Trends / Dynamics of the component "Leisure and free time" based on Google Trends data**

Источник / Source: составлено авторами / compiled by the authors.

По блоку «**Безопасность**» корреляция составляет –0,77. Показатели данного блока социальной комфортности должны характеризовать уровень риска для жизни и здоровья населения, а не успешность работы системы правоохранительных органов. Отсюда показатели безопасности, предоставляемые ФСГС, нельзя рассматривать как достаточно валидные с точки зрения характеристики социальной комфортности, поскольку они представлены числом совершенных преступлений различного характера, грабежей, причинения тяжкого вреда здоровью.

В работе [38] отмечается, что на самоощущение безопасности граждан оказывают влияние личные установки граждан и повседневные практики (приходится ли носить оружие, газовые баллончики и пр.), а не количество и раскрытие преступлений. Кроме того, в работе [39] показано, что преступления, которые попадают под статью «Убийства», имеют коэффициент латентности в России 2,3. Это означает, что реальное число преступлений в 2,3 раза превышает показатели официальной статистики. Данный факт подтверждается результатами проведенного исследования. Численная оценка безопасности по данным GT намного ниже оценки по данным официальной статистики: 6 баллов против 8 баллов (*рис. 3*).

Если проанализировать динамику двух блочных индикаторов на *рис. 3*, то можно заметить, что в период неблагоприятной экономической ситуации в 2014–2016 гг., вызванной введением санкций, эскалацией геополитического конфликта и, как следствие, удешевлением отечественной валюты, наблюдается прогиб блочного индикатора «Безопасность» по данным GT. Это объясняется беспокойством граждан, снижением чувства безопасности, комфорта. Но график, построенный по данным официальной статистики, в период 2014–2016 гг. демонстрирует уверенный рост, что является контринтуитивным и противоречивым. Есть основание полагать, что для адекватной оценки уровня безопасности населения в статистику правонарушений также целесообразно интегрировать индикаторы безопасности, построенные по данным GT.

Блочные индикаторы «Отдых и свободное время» также демонстрируют обратную линейную зависимость. Статистика данного блока представлена лишь тремя показателями: числом спортивных учреждений; количеством путевок, реализованных через турфирмы, и числом российских туристов, обслуженных туристскими фирмами. В то же время в анализируемый блок по данным GT вошло около 54 показателей, отражающих различные аспекты свободного времени и отдыха индивида (*рис. 4*).

Динамика анализируемой компоненты GT вполне адекватна реалиям экономической жизни: последствия пандемии COVID-19 негативно отразились на отдыхе россиян, так как пришлось изменить свои туристические предпочтения. Следствие — снижение значений данной компоненты на 31% в течение одного года. Поскольку в сводном показателе социальной комфортности «отдых» имеет существенный вклад — 10%, следует большее внимание уделять развитию новых туристических направлений, досуга населения. Значительная роль блочного индикатора «Отдых и свободное время» в формировании социального комфорта, а также слабая представленность в официальной статистике обосновывают необходимость использования Google Trends в развитии методологии учета статистики туризма и отдыха.

Таким образом, компоненты социальной комфортности, построенные по данным GT и имеющие значимую положительную корреляцию с компонентами официальной статистики, могут быть использованы в целях проведения оперативного мониторинга условий жизни населения. По компонентам «Образование и обучение», «Безопасность», «Отдых и свободное время» требуется серьезная методологическая работа по вариантам интеграции альтернативных источников информации в данные официальной статистики в силу несогласованности результатов и слабого отражения уровня социальной комфортности проживания населения.

ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ BIG DATA В ПРОГНОЗИРОВАНИИ ИНДИКАТОРОВ ФИНАНСОВОГО СОСТОЯНИЯ

Расчеты, представленные ранее, демонстрируют хорошие возможности использования больших данных к прогнозированию по 7 из 10 компонентов социальной комфортности, в том числе по компоненте «Финансовое положение». В силу ухудшения финансового положения населения в России остаются крайне актуальными перспективы использования поисковых запросов, отражающих финансовое настроение экономических агентов для прогнозирования экономических показателей.

Финансовое настроение экономических агентов, агрегированное в едином показателе с помощью поисковых запросов Google Trends, становится важным источником информации об их предпочтениях и поведении. С целью определения перспектив использования рассчитанной компоненты как прокси для индикаторов финансового состояния рассчитаем корреляцию с показателем «Рыночная капитализация национальных компаний, акции которых обращаются на фондовой бирже». Согласно публикуемой информации Всемирного банка банковская система и фондовый рынок напрямую связаны с экономическим ростом, что является основным фактором, влияющим на уменьшение бедности, которая, в частности, учитывается в финансовой компоненте социальной комфортности.

Высокая корреляция с рыночной капитализацией показывает, что поисковые запросы могут использоваться для экстраполяции индексов финансового состояния без использования ресурсов финансовой статистики, что упрощает и ускоряет расчет.

Реакция интереса интернет-пользователей по отношению к показателям финансового рынка в ответ на изменение экономического индикатора (ВВП, индекса капитализации ММВБ, инфляции, ставки по депозитам и др.) в дальнейшем может быть использована в модельной спецификации типа VAR (векторная авторегрессия), позволяющей прогнозировать социальную комфортность в зависимости от изменения показателей финансового положения населения и ожидаемых тенденций изменения экономического роста.

ВЫВОДЫ

Большие данные располагают более детализированными статистическими оценками различных явлений и процессов в обществе, что является необходимым аргументом в развитии положений концепции качества жизни населения как одной из важнейших категорий социальной и экономической науки. Введение в научный оборот латентной категории «социальная комфортность» углубляет теорию качества жизни населения в части исследования человека с точки зрения включенности его в социум, расширения субъективного

Рис. 5 / Fig. 5. **Динамика компоненты «Финансовое положение» по данным Google Trends и рыночной капитализации отечественных компаний, котирующихся на бирже / Dynamics of the component "Financial situation" and market capitalization of listed domestic companies**

Источник / Source: составлено авторами на основе Google Trends и данных Всемирного банка / compiled by the authors based on Google Trends and World Bank Datebase.

аспекта измерения, что объясняется необходимостью использования Google Trends.

Предлагаемое исследование проводит интегральную оценку социальной комфортности с использованием двух источников информации: данных официальной статистики и Google Trends. Интегральная оценка социальной комфортности позволяет, в свою очередь, увидеть общую картину развития явления во времени, а также дать оценку проводимой социально-экономической политике.

В целях минимизации авторского субъективизма к оценке социальной комфортности, по данным Google Trends, используется новый подход к семантическому поиску информации о компонентах социальной комфортности на основе применения специализированного словаря, располагающего классификациями различных процессов и явлений, а также анализа валидности каждого поискового запроса в части раскрытия социальной комфортности и соотнесенности с реалиями российской действительности.

В процессе моделирования решается задача по гармонизации и взаимоувязке ресурсов различных типов: использован комплекс эконометрических методов по диагностике данных на наличие временного тренда, резких скачков популярности запроса, наличие нулевого объема поиска, экстремальной сезонности и приведению временных рядов к сопоставимому виду не только между собой внутри одного информационного ресурса. Используемый метод стандартизации поисковых запросов позволяет строить надежные оценки и проводить моделирование составных категорий по данным различного типа.

На основе примененной методологии в анализе социальной комфортности использованы 475 поисковых запросов Google Trends и 100 показателей, взятых из официальной статистики, которые были агрегированы в блочные индикаторы социальной комфортности. Корреляционный анализ блочных индикаторов продемонстрировал устойчивую положительную корреляцию между компонентами социальной комфортности, построенными по GT и данными официальной статистики (7 из 10 компонент), что говорит о хороших перспективах использования альтернативных источников информации (GT, например) в целях оценки социальной комфортности для осуществления оперативного мониторинга, не прибегая к данным официальной статистики. По трем компонентам «Образование и обучение», «Безопасность», «Отдых и свободное время» наблюдается сильная отрицательная линейная зави-

симость, которая объясняется преимущественно слабой валидностью статистических показателей для оценки социальной комфортности и определяет первоочередную необходимость интеграции больших данных в эти направления для совместного использования разных источников информации с целью получения более надежных оценок.

Таким образом, можно заключить, что применение больших данных в оценке латентных категорий дает хорошие результаты, сопоставимые с данными официальной статистики, что открывает возможности их использования в проведении мониторинга и прогнозирования финансового положения. Однако интеграция двух источников данных должна осуществляться последовательно при проведении возможной проверки с другими источниками, например с данными социологических опросов.

БЛАГОДАРНОСТЬ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20–310–70037 «Стабильность». Московская школа экономики МГУ им. М.В. Ломоносова, Москва, Россия.

ACKNOWLEDGEMENTS

The reported study was funded by RFBR, project No. 20–310–70037 “Stability”. Lomonosov Moscow State University, Moscow, Russia.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Liu J., Li J., Li W., Wu J. Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2016;115:134–142. DOI: 10.1016/j.isprsjprs.2015.11.006
2. Algan Y., Beasley E., Guyot F., Higa K., Murtin F., Senik C. Big data measures of well-being: Evidence from a Google well-being index in the United States. OECD Statistics Working Papers. 2016;(03). URL: https://www.oecd-ilibrary.org/docserver/5jlz9hpg0rd1-en.pdf?expires=1629818036&id=id&accname=guest&checksum=7ED 855395D 5B 778D 71E 405ED 1925ECE 3
3. Фантаццини Д., Шаклеина М.В., Юрас Н.А. Big Data в определении социального самочувствия населения России. *Прикладная эконометрика*. 2018;50:43–66.
   Fantazzini D., Shakleina M., Yuras N. Big data for computing social well-being indices of the Russian population. *Prikladnaya ekonometrika = Applied Econometrics*. 2018;50:43–66. (In Russ.).
4. Carlquist E., Nafstad H.E., Blakar R.M., Ulleberg P., Delle Fave A., Phelps J.M. Well-being vocabulary in media language: An analysis of changing word usage in Norwegian newspapers. *The Journal of Positive Psychology*. 2017;12(2):99–109. DOI: 10.1080/17439760.2016.1163411
5. Curini L., Iacus S., Canova L. Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case. *Social Indicators Research*. 2015;121(2):525–542. DOI: 10.1007/s11205–014–0646–2
6. Prata D.N., Soares K.P., Silva M.A., Trevisan D.Q., Letouze P. Social data analysis of Brazilian’s mood from Twitter. *International Journal of Social Science and Humanity*. 2016;6(3):179–183. DOI: 10.7763/IJSSH.2016.V6.640
7. Nguyen Q.C., Kath S., Meng H.-W., Li D., Smith K.R., VanDerslice J.A., Wen M., Li F. Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*. 2016;73(8):77–88. DOI: 10.1016/j.apgeog.2016.06.003
8. Yang C., Srinivasan P. Life satisfaction and the pursuit of happiness on Twitter. *PLoS ONE*. 2016;11(3):e0150881. DOI: 10.1371/journal.pone.0150881
9. Wang W., Hernancez I., Newman D.A., He J., Bian J. Twitter analysis: Studying US weekly trends in work stress and emotion. *Applied Psychology*. 2016;65(2):355–378. DOI: 10.1111/apps.12065
10. Liu P., Tov W., Kosinski M., Stillwell D.J., Qui L. Do Facebook status updates reflect subjective well-being? *Cyberpsychology, Behavior, and Social Networking*. 2015;18(7):373–379. DOI: 10.1089/cyber.2015.0022
11. LiKamWa R., Liu Y., Lane N.D., Zhong L. MoodScope: Building a mood sensor from smartphone usage patterns. In: Proc. 11th Annu. int. conf. on mobile systems, applications, and services (MobiSys). (Taipei, June 25–28). New York: ACM; 2013:389–402. DOI: 10.1145/2462456.2464449
12. Schwartz H.A., Sap M., Kern M.L., Eichstaedt J.C., Kapelner A., Agrawal M., Ungar L.H. et al. Predicting individual well-being through the language of social media. In: Proc. Pacific symp. on biocomputing (PSB). (Big Island of Hawaii, Jan. 4–8, 2016). Singapore: World Scientific Publishing Co.; 2016:516–527.

13. Da Z., Engelberg J., Gao P. The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*. 2015;28(1):1–32. DOI: 10.1093/rfs/hhu072
14. Singer E. The use of incentives to reduce nonresponse in household surveys. The University of Michigan. Survey Methodology Program. 2002;(051). URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.611.1597&rep=rep1&type=pdf
15. Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S., Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012–1014. DOI: 10.1038/nature07634
16. Упадхьяя Ш. Большие данные и официальная статистика. *Вопросы статистики*. 2019;26(12):5–14. DOI: 10.34023/2313–6383–2019–26–12–5–14
    Upadhyaya S. Big data and official statistics. *Voprosy statistiki*. 2019;26(12):5–14. (In Russ.). DOI: 10.34023/2313–6383–2019–26–12–5–14
17. Оксенойт Г.К. Цифровая повестка, большие данные и официальная статистика. *Вопросы статистики*. 2018;25(1):3–16.
    Oksenoyt G.K. Digital agenda, big data and official statistics. *Voprosy statistiki*. 2018;25(1):3–16. (In Russ.).
18. Плеханов Д.А. Большие данные и официальная статистика: обзор международной практики внедрения новых источников данных. *Вопросы статистики*. 2017;(12):49–60.
    Plekhanov D.A. Bigdata and official statistics: A review of international experience with integration of new data sources. *Voprosy statistiki*. 2017;(12):49–60. (In Russ.).
19. Шаклеина М.В., Волкова М.И., Шаклеин К.И., Якиро С.Р. Теоретические и методологические проблемы измерения социальной комфортности: результаты эмпирического анализа на российских данных. *Экономические и социальные перемены: факты, тенденции, прогноз*. 2020;13(5):135–152. DOI: 10.15838/esc.2020.5.71.8
    Shakleina M.V., Volkova M.I., Shaklein K.I., Yakiro S.R. Theoretical and methodological problems of measuring social comfort: Results of empirical analysis based on Russian data. *Economic and Social Changes: Facts, Trends, Forecast*. 2020;13(5):135–152. DOI: 10.15838/esc.2020.5.71.8
20. Петрова Д.А. Прогнозирование инфляции на основе интернет-запросов. *Экономическое развитие России*. 2019;26(11):55–62.
    Petrova D.A. Inflation forecasting based on Internet search queries. *Ekonomicheskoe razvitie Rossii = Russian Economic Developments*. 2019;26(11):55–62. (In Russ.).
21. Guzmán G. Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*. 2011;36(3):119–167. DOI: 10.3233/JEM-2011–0342
22. Afkhami M., Cormack L., Ghoddusi H. Google search keywords that best predict energy price volatility. *Energy Economics*. 2017;67:17–27. DOI: 10.1016/j.eneco.2017.07.014
23. Tetlock P.C. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*. 2007;62(3):1139–1168. DOI: 10.1111/j.1540–6261.2007.01232.x
24. Brochado A. Google search-based sentiment indexes. *IIMB Management Review*. 2020;32(3):325–335. DOI: 10.1016/j.iimb.2019.10.015
25. Dunphy D.C., Bullard C.G., Crossing E.E.M. Validation of the general inquirer Harvard IV dictionary. Cambridge, MA: Harvard University Library; 1974. 158 p.
26. Kelly E.F., Stone P.J. Computer recognition of English word senses. Amsterdam: North-Holland; 1975. 269 p.
27. Zuell C., Weber R.P., Mohler P.P. Computer-aided text classification for the social sciences: The General Inquirer III. Mannheim: ZUMA, Center for Surveys, Research and Methodology; 1989.
28. Scott S.L., Varian H.R. Bayesian variable selection for nowcasting economic time series. NBER Working Paper. 2013;(19567). URL: https://www.nber.org/system/files/working_papers/w19567/w19567.pdf
29. Kapounek S., Deltuvaitė V., Koráb P. Determinants of foreign currency savings: Evidence from Google search data. *Procedia — Social and Behavioral Sciences*. 2016;220:166–176. DOI: 10.1016/j.sbspro.2016.05.481
30. Bokelmann B., Lessmann S. Spurious patterns in Google Trends data — An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*. 2019;75:1–12. DOI: 10.1016/j.tourman.2019.04.015
31. Google Trends help — how Trends data is adjusted. Google 2018. URL: https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052 (accessed on 23.04.2018).
32. Cleveland R.B., Cleveland W.S., McRae J.E., Terpenning I. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*. 1990;6(1):3–73. URL: https://www.wessa.net/download/stl.pdf
33. Петрова Д.А., Трунин П.В. Выявление настроений экономических агентов на основе поисковых запросов. *Прикладная эконометрика*. 2020;(3):71–87. DOI: 10.22394/1993–7601–2020–59–71–87

Petrova D.A. Trunin P.V. Revealing the mood of economic agents based on search queries. *Prikladnaya ekonometrika = Applied Econometrics*. 2020;(3):71–87. (In Russ.). DOI: 10.22394/1993–7601–2020–59–71–87

34. Parker J., Cuthbertson C., Loveridge S., Skidmore M., Dyar W. Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google Trends data. *Journal of Affective Disorders*. 2017;213:9–15. DOI: 10.1016/j.jad.2016.10.038
35. Айвазян С.А. Анализ качества и образа жизни населения. М.: Наука; 2012. 432 с.
Ayvazyan S.A. Analysis of the quality and lifestyle of the population. Moscow: Nauka; 2012. 432 p. (In Russ.).
36. Миндели Л.Э., Пашинцева Н.И. Российская система образования и ее отражение в статистике. *Вопросы статистики*. 2016;(11):67–84.
Mindeli L.E., Pashinceva N.I. Russian education system and how it is reflected in statistics. *Voprosy statistiki*. 2016;(11):67–84. (In Russ.).
37. Chan K.L. Intelligence Capital Index. 2017. URL: http://www.kailchan.ca/wp-content/uploads/2017/04/KC_Intelligence-Capital-Index-full-results-and-methodology_Apr-2017_v2.pdf
38. Сатаров Г.А., Благовещенский Ю.Н. Статистическое сравнение России и других стран. Комитет гражданских инициатив. Фонд ИНДЕМ. URL: https://komitetgi.ru/upload/iblock/3cf/3cfcb375eced922f253c446a4b37645b.pdf
Satarov G.A., Blagoveshchenskii Yu.N. Statistical comparison of Russia and other countries. Civil Initiatives Committee. INDEM Foundation. URL: https://komitetgi.ru/upload/iblock/3cf/3cfcb375eced922f253c446a4b37645b.pdf (In Russ.).
39. Иншаков С.М. Теоретические основы исследования и анализа латентной преступности. М.: ЮНИТИ-ДАНА; 2011. 839 с.
Inshakov S.M. Theoretical foundations of research and analysis of latent crime. Moscow: UNITY-DANA; 2011. 839 p. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / ABOUT THE AUTHORS

Марина Владиславовна Шаклеина — кандидат экономических наук, доцент, Московская школа экономики МГУ им. М.В. Ломоносова, Москва, Россия
Marina V. Shakleina — Cand. Sci. (Econ.), Assoc. Prof., Moscow School of Economics of the Lomonosov Moscow State University, Moscow, Russia
shakleina.mv@gmail.com

Мария Игоревна Волкова — кандидат экономических наук, лаборатория «Моделирование социально-экономических систем» ФГБОУ ВО «РЭУ им. Г.В. Плеханова», Москва, Россия
Mariya I. Volkova — Cand. Sci. (Econ.), Head of the Laboratory “Modeling of socio-economic systems”, Plekhanov Russian University of Economics, Moscow, Russia
frauwulf@gmail.com

Константин Игоревич Шаклеин — кандидат экономических наук, главный специалист Департамента экономики, Открытое акционерное общество «Российские железные дороги», Москва, Россия
Konstantin I. Shaklein — Cand. Sci. (Econ.), Chief Specialist of Department of Economics OJSC “Russian Railways”, Moscow, Russia
mrshaklein@gmail.com

Станислав Ростилавович Якиро — главный специалист Дирекции по рискам и анализу экономической эффективности, Акционерное общество «Страховое общество газовой промышленности», Москва, Россия
Stanislav R. Yakiro — Chief Specialist of Department of Risk and Economic Performance Analysis JSC “SOGAZ”, Moscow, Russia
yakirosr@yandex.ru

Заявленный вклад авторов:
Шаклеина М.В. — введение, актуальность темы исследования и постановка задачи, анализ источников литературы, изучение проблематики исследования, разработка методологии исследования, интерпретация полученных результатов.
Волкова М.И. — обоснование выбора показателей для анализа, формирования выводов и рекомендаций по результатам исследования.
Шаклеин К.И. — статистический анализ данных, описание используемой методики расчета, анализ полученных результатов, табличное и графическое представление результатов.
Якиро С.Р. — сбор и обработка больших данных, оценка и прогнозирование.

Authors' declared contribution:
Shakleina M.V. — introduction, relevance of the research topic and problem statement, analysis of literature sources, study of research problems, development of research methodology, interpretation of the results.
Volkova M.I. — justification of the choice of indicators for analysis; formation of conclusions and recommendations based on the results of the study.
Shaklein K.I. — statistical analysis of data, description of the calculation method used, analysis of the results, tabular and graphical representation of the results.
Yakiro S.R. — collection and processing of big data, assessment and forecasting.

Статья поступила в редакцию 13.05.2021; после рецензирования 27.05.2021; принята к публикации 27.08.2021.
Авторы прочитали и одобрили окончательный вариант рукописи.
The article was submitted on 13.05.2021; revised on 27.05.2021 and accepted for publication on 27.08.2021.
The authors read and approved the final version of the manuscript.