

Прикладная эконометрика, 2024, т. 73, с. 78–101.

Applied Econometrics, 2024, v. 73, pp. 78–101.

DOI: 10.22394/1993-7601-2024-73-78-101

Г. Т. Броницкий, Е. С. Вакуленко¹

Применение Google Trends для прогнозирования миграции из России: агрегация поисковых запросов и учет лаговой структуры

В работе предложена методика прогнозирования миграционной статистики с использованием данных поисковых запросов Google Trends Index (GTI). Существующая методология усовершенствована в двух направлениях: во-первых, предложен подход агрегирования ключевых поисковых запросов на основе различных статистических критериев; во-вторых, показана важность включения в модель миграции временной лаговой структуры поисковых запросов в зависимости от целей миграции и связанных с ними GTI. Работоспособность предложенных подходов продемонстрирована на месячных данных статистического офиса Германии об объемах миграции из России в Германию с января 2011 по август 2022 г. Результаты показали, что модели миграции с распределенными лагами GTI лучше прогнозируют миграцию, чем SARIMA модели. Оценки среднего лага, т. е. времени реакции миграционной статистики на поисковые запросы по тематикам «посольство», «работа» и «учеба», оказались равны 5.6, 6.5 и 8 месяцам соответственно. Показано, что для прогнозирования миграции из России в Германию достаточно учитывать только поисковые запросы, связанные с тематикой «посольство».

Ключевые слова: международная миграция; Россия; Германия; Google Trends; поисковые запросы; наукастинг; большие данные; прогнозирование.

JEL classification: J61; C53; C38.

1. Введение

Миграция населения существенным образом влияет на демографические, экономические, социальные и иные показатели стран выбытия и прибытия. Показатели миграции подвержены внешним шокам, таким как эпидемии, военные действия, природные катаклизмы и др. Своевременный анализ миграционных потоков, а также намерений о миграции, может помочь актуализировать меры демографической политики государства. При наступлении таких событий особенно важно иметь оценку миграций без задержки во времени. Такая

¹ Броницкий Георгий Тимурович — Национальный исследовательский университет «Высшая школа экономики», Москва; getibr@gmail.com.

Вакуленко Елена Сергеевна — Национальный исследовательский университет «Высшая школа экономики», Москва; evakulenko@hse.ru.

информация поможет своевременно реагировать на внешние шоки, уменьшая экономические риски страны. Однако сложность сбора статистики, в которой не все перемещения фиксируются, отсутствие достоверных данных, а также предоставление данных о миграции с временной задержкой затрудняет подобные оценки. В литературе развивается целое направление, связанное с оценками миграционных потоков как внутри страны, так и за ее пределы, на основе альтернативных источников данных (Вакуленко, Цимайло, 2011; Чудиновских, 2018, 2020; Tjaden, 2021).

Цель данной работы — краткосрочное прогнозирование внешней миграции (наукастинг) с использованием данных поисковых запросов Google Trends Index² (GTI), которые опережают данные, публикуемые официальными статистическими службами. Такой подход уже ранее применялся для прогнозирования миграции: как внутренней, так и внешней, в том числе и в России (Fantazzini et al., 2021; Цапенко, Юревич, 2022; Броницкий, Вакуленко, 2022). В этих работах рассматривались модели временных рядов (типа SARIMAX), где зависимой переменной был миграционный поток, а в числе объясняющих рассматривались GTI для определенных поисковых запросов, которые, как правило, определяются экспертно. В настоящей работе предлагается усовершенствование методологии прогнозирования миграции с помощью GTI в двух направлениях. Во-первых, разрабатывается подход агрегирования ключевых поисковых запросов на основе различных статистических критериев, демонстрируется их работоспособность и приводятся аргументы в пользу выбора одного из них. Во-вторых, показывается важность включения в модели миграции временной лаговой структуры поисковых запросов в зависимости от целей миграции и связанных с ними GTI.

В качестве примера применения предложенного подхода рассматриваются ежемесячные миграционные потоки из России в Германию с января 2011 по август 2022 г. по данным статистического офиса Германии, т. к. качество учета мигрантов в статистике принимающих стран гораздо выше (Денисенко, 2003; Воробьева, Гребенюк, 2017). Эти потоки не разделены по целям миграции (трудовая, учебная, иная), что усложняет выбор ключевых поисковых запросов и требует учета различных причин перемещений. Для анализа была выбрана миграция в Германию, поскольку по данным Росстата³ эта страна входила в тройку лидеров стран дальнего зарубежья по миграции из России (вместе с Израилем и США) в начале 2000-х (Денисенко, 2003), и лидирует в 2021 г., но уже наравне с Китаем, Индией, Вьетнамом и Грузией. Кроме того, статистический офис Германии предоставляет статистику высокого качества и дает возможность получить данные в открытом доступе либо по запросу.

Статья имеет следующую структуру. Второй раздел содержит обзор литературы по применению альтернативных источников для оценки миграции, в том числе GTI. Затем описываются данные, их особенности и подбор ключевых поисковых запросов GTI, обсуждается методология наукастинга миграции с предложенными усовершенствованиями и приводятся результаты исследования. В заключение обсуждаются преимущества и ограничения разработанных подходов и указываются направления для будущих исследований по этой тематике.

² <https://trends.google.ru/trends/?geo=RU>.

³ <https://www.fedstat.ru/indicator/43513>.

2. Обзор литературы

Как уже упоминалось, данные о миграции предоставляются статистическими службами со значительной временной задержкой, которая составляет от нескольких месяцев до нескольких лет. Кроме того, истинные потоки миграции не регистрируются в полном объеме и не попадают в статистику. Поэтому исследователи прибегают к альтернативным источникам данных, которые позволяют восполнить эти пробелы, таким как данные о GPS положении мобильных устройств (Bengtsson et al., 2011), информация об отметках геолокации в социальных сетях (Chi et al., 2020; State et al., 2014), Twitter (Zagheni et al., 2014; Moise et al., 2016; Kim et al., 2020), расположение IP-адресов пользователей email-аккаунтов (Zagheni, Weber, 2012), Skype (Kikas et al., 2015), библиографическая база Scopus для анализа миграции ученых (Subbotin, Aref, 2021). Сложности применения данных цифрового следа подробно обсуждаются в (Tjaden, 2021). Применение геолокации критикуется в работе (Чудиновских, 2018) из-за невозможности разделить получаемые оценки миграции на постоянные и временные. У перечисленных методов есть и другие недостатки: как правило, эти данные находятся у частных компаний и недоступны для исследователей; при помощи такой информации невозможно заранее оценить намерения о миграции.

В 2006 г. появился сервис представления динамики агрегированных поисковых запросов — Google Trends Index, который открыл новые возможности для исследований в различных областях (ИТ, связь, медицина, здравоохранение, бизнес и экономика). Первым из них считается работа (Ginsberg et al., 2009), в которой прогнозируется распространение гриппа. В (Jun et al., 2018) отмечается прогрессивный рост использования Google Trends за 10 лет их существования на основе метаанализа 657 статей.

Исследования показывают, что мигранты ищут информацию в сети Интернет еще до момента отъезда (Fantazzini et al., 2021; Böhme et al. 2020). Впервые о применимости Google Trends для прогнозирования миграции, в частности, туристических потоков в Гонконг из ряда стран, в том числе во время проведения летней Олимпиады (Beijing Olympics) в августе 2008 г., было написано в работе (Choi, Varian, 2012)⁴. В этой же работе отмечается, что включение поисковых запросов с лагом улучшает прогнозные качества модели. Авторы включали запросы за первые 2 недели месяца для объяснения помесечных данных о туристических потоках. Однако туристические потоки — это перемещения, не связанные с изменением места жительства, нас же интересует миграция, поэтому сконцентрируемся на работах, в которых применяются GTI для прогнозирования миграции с переменной места жительства. Такие работы представлены в табл. 1. При их анализе особое внимание обращается на ключевые поисковые запросы, которые включались в модели миграции, а также на моделирование лаговой структуры запросов. Было найдено всего 8 работ, две из них анализируют миграционные потоки России: внутренние — из Москвы в Санкт-Петербург (Fantazzini et al., 2021), а также внешние — из Таджикистана в Россию (Цапенко, Юревич, 2022). За исключением работы (Wladyk, 2017), все остальные были написаны за последние 3 года.

В основном в качестве поисковых запросов для Google Trends берутся такие, как посольство, виза, работа, жилье, и добавляется регион назначения. Чаще всего авторы концентрируются на отдельных запросах, которые выбирают на основании экспертных суждений, и включают в модель только их (Böhme et al., 2020; Golenvaux et al., 2020; Wanner,

⁴ Первая версия этой работы в виде препринта появилась в 2009 году (Choi, Varian, 2009).

2021; Fantazzini et al., 2021; Jurić, 2022; Цапенко, Юревич, 2022). Лишь в нескольких работах изучается корреляция между миграционными потоками и различными поисковыми запросами (Wladyk, 2017; Avramescu, Wiśniowski, 2021), а затем рассматриваются модели для отдельных поисковых запросов.

Согласно теории миграции, непосредственный переезд происходит только на третьем из пяти этапов миграции на новое место жительства (Benson-Rea, Rawlinson, 2003). До этого следует этап поиска информации о месте переезда. Поэтому важно исследовать глубину лага поисковых запросов на миграцию. В основном исследователи включают GTI с лагом в один период (месяц или год в зависимости от периодичности данных) (Böhme et al., 2020; Golenvaux et al., 2020; Fantazzini et al., 2021). Есть работы, в которых лаговая структура изучалась вплоть до трех лет на помесечных данных (Wanner, 2021). В большинстве случаев значимыми оказываются лаги в пределах года, но, например, для миграции из Франции в Швейцарию значимым оказался лаг GTI в 24 месяца (Wanner, 2021).

Таким образом, на данный момент в литературе по прогнозированию миграции с помощью Google Trends не удалось найти работ, где бы предлагалось решение по агрегированию поисковых запросов и изучению их лаговой структуры. В данной работе предлагается методика агрегации ключевых поисковых запросов в некоторые индексы, которые наилучшим образом будут описывать миграцию. Тем самым, рассматривается сразу группа поисковых запросов, которая характеризует различные цели миграции (трудовая, учебная и т.д.), что позволяет уменьшить размерность при одновременном включении разных типов запросов и изучении лаговой структуры по ним.

3. Методология и данные

Основная цель работы — прогнозирование объемов миграции населения до появления данных в официальных источниках (наукастинг). Таким образом, построив модели на GTI, можно в режиме реального времени оценивать миграцию, поскольку данные о поисковых запросах публикуются без задержки во времени. Кроме того, проверяется гипотеза о том, что миграция происходит не в момент поискового запроса, а с некоторым лагом, необходимым на подготовку документов и принятие решения, как об этом уже было сказано выше.

Для решения поставленной задачи необходимо собрать и подготовить данные о миграции, подобрать оптимальный набор поисковых запросов, выбрать наилучшие модели миграции. Можно выделить три основных этапа в данном исследовании.

1. При помощи методов машинного обучения были отобраны слова и словосочетания, наиболее близкие к слову «миграция». Такие методы позволяют отойти от экспертной оценки поисковых запросов к алгоритмическому определению множества поисковых слов. В результате было сформировано множество поисковых слов, используемых для оценки миграции (Приложение 1).

2. К полученным индексам GTI применена операция стандартизации (1), а также методы снижения размерности данных, что позволило агрегировать исходные GTI в группы на основе статистических критериев. В результате можно в несколько раз сократить число регрессоров и в дальнейшем добавить в модель их лаги. Исследовались различные вариации снижения размерности, включая метод главных компонент (PCA), декомпозицию R^2

Таблица 1. Обзор исследований с использованием Google Trends для прогнозирования миграции

Авторы	Страны	Годы	Поисковые запросы	Лаги, выбор лага	Выбор поисковых запросов	Модели
Władysław (2017)	Из Латинской Америки (Аргентина, Перу, Колумбия) в Испанию	2005–2010	Посольство Испании, виза в Испанию, работа в Испании	Колумбия: работа и посольство (9-й лаг); Аргентина: работа (4-й лаг), посольство (8-й лаг); Перу: посольство (9-й лаг); Кросс-корреляция с миграцией	Кросс-корреляция с миграцией, три отдельных запроса	Парная регрессия на переменных в разностях
Böhme et al. (2020)	В 101 стране происхождения мигрантов в 35 принимающих стран ОЭСР	2004–2015	Миграция и экономика (работа, зарплата, виза) + страна	Головные данные, 1 лаг	Отдельные запросы	FE, панель стран, гравитационная модель
Golenvaux et al. (2020)	Из 101 страны происхождения мигрантов в 35 принимающих стран ОЭСР (из Böhme, Gröger, Stöhr, 2020)	2004–2015	Миграция и экономика (работа, зарплата, виза) + страна	Головные данные, 1 лаг	Отдельные запросы	Нейронные сети (LSTM) лучше прогнозируют, чем ANN, линейная гравитационная модель как в (Böhme, Gröger, Stöhr, 2020)
Wagner (2021)	В Швейцарию из Франции, Италии, Германии и Испании	2006–2016	Работа в Швейцарии	До 3-х лет. По критерию R^2	Экспертно	Линейная регрессия с лагами GTP
Fantazzini et al. (2021)	Россия, Москва и Санкт-Петербург	2009–2018	Поездка в «название региона», работа в «название региона», жилье в «название региона»	1 месяц	Экспертно	ARIMAX, SARIMAX, VECM
Avramescu, Wisniewski (2021)	Из Румынии в Великобританию	2012–2019	Кластеры запросов: занятость, образование, валюта, жилье. WordNet corpus для поиска синонимов ключевых запросов.	Нет лагов, скользящее среднее за 12 месяцев	Корреляция с миграцией, усреднение по кластеру	ARIMAX
Jugić (2022)	Из Хорватии в Австрию и Германию	2004–2020	Работа, жилье, образование, общие (паспорт, виза, гражданство)	Нет лагов	Самый частый запрос, отдельные запросы	Парная регрессия
Цапенко, Юревич (2022)	Из Таджикистана, Киргизии, Узбекистана в Россию	2015–2020	(работа ИЛИ вакансии) И (Москва ИЛИ Россия)	7 или 11 месяцев. Информационные критерии, значимость коэффициентов, алгоритм обратного исключения	Экспертно	ARIMAX

по методу Шепли, применение РСА отдельно по группам поисковых слов для наиболее популярных тематик («посольство», «учеба» и «работа»⁵);

3. Оценивались модели в сезонных разностях с включенными лагами от 1 до 12 месяцев. Проводилось исследование качества прогнозов, оценивался средний лаг в исследуемых группах, делались выводы о скорости влияния регрессоров модели на миграцию.

3.1. Данные о миграции

Официальная статистическая информация о миграции из России в другие страны содержится в Единой межведомственной информационно-статистической системе (ЕМИСС)⁶. Однако данные показателя «Число выбывших»⁷ публикуются с задержкой, которая на момент публикации достигает полугода. Кроме того, в 2011 г. изменилась методология подсчета этого показателя, что также влияет на качество данных (Чудиновских, Степанова, 2020). Так, по данным Росстата, число мигрантов из России в Германию за последние 20 лет сократилось более чем в 7 раз, а по данным организации экономического сотрудничества и развития (ОЭСР⁸) — только в 1.5 раза. Перечисленные выше проблемы затрудняют использование официальных статистических показателей, заставляя прибегать к альтернативным источникам информации. Например, можно пользоваться показателями статистических служб стран прибытия мигрантов, в частности, статистическим офисом Германии⁹, который публикует ежегодные данные о числе мигрантов из других стран. Однако у внешних данных есть и свои недостатки: показатели обладают годовой частотностью, что затрудняет принятие оперативных решений, а данные также имеют задержку, которая на момент написания статьи составляет 2 года.

Для настоящей работы по запросу были получены данные статистического офиса Германии (рис. 1) с помесечной частотностью и с задержкой в 3 месяца. Эта статистика не публикуется в общедоступных источниках, а агрегируется в официальные годовые значения. Подобные данные позволяют получить более точные оценки построенных моделей, что важно на этапе сравнения моделей. Кроме того, такая помесечная статистика может быть полезна для оценки эффективности алгоритмов увеличения частотности, которые предложены, например, в работе (Броницкий, Вакуленко, 2022). В Германии иностранные граждане фиксируются статистическими службами, если они получили разрешение на жительство и собираются остаться в Германии на 3 и более месяцев, а также граждане Германии и лица немецкого происхождения (Aussiedler), которые возвращаются на свою историческую родину. По данным статистической службы Германии, в 90-е годы миграционный отток из России в Германию в 1.2 раза превосходил данные о зарегистрированных мигрантах в России (Денисенко, 2003). И это различие характерно для всех миграционных потоков из России в другие страны. По самым скромным подсчетам, оценки Росстата по миграции

⁵ Поисковые слова для каждой тематики представлены в Приложении 1.

⁶ <https://rosstat.gov.ru/emiss>.

⁷ <https://www.fedstat.ru/indicator/43513>.

⁸ <https://stats.oecd.org/Index.aspx?DataSetCode=MIG#>.

⁹ <https://www-genesis.destatis.de/genesis/online>, показатель 12711–0011.

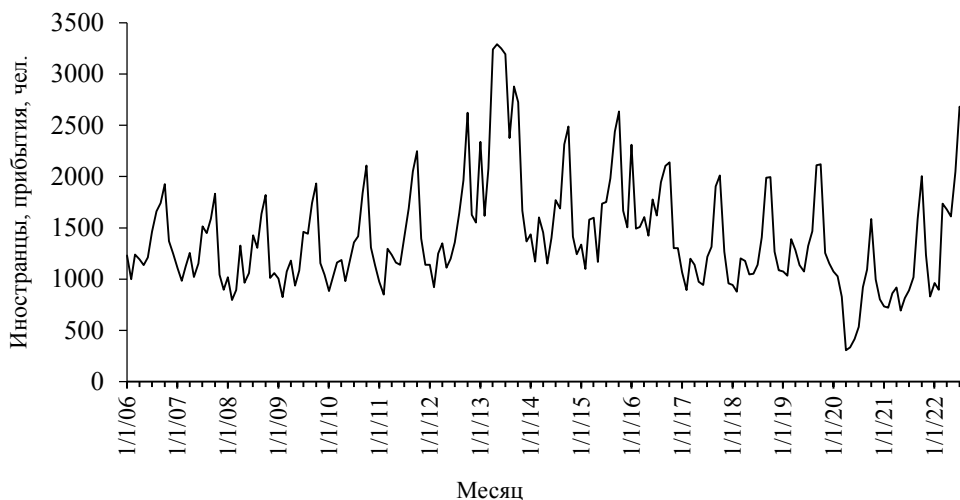


Рис. 1. Число прибытий иностранцев из РФ в Германию по данным статистического офиса Германии (2006–2022 гг.), чел.

из России нужно увеличивать в 3–4 раза (Воробьева, Гребенюк, 2017). По нашим оценкам, данные России и Германии в 2021 г. различались в 3.9 раза. Сравнивать эти данные можно с некоторыми оговорками и допущениями, т.к. статистика Германии включает как краткосрочные, так и долгосрочные перемещения. Поскольку данные интернет-запросов сложно разделить по целям миграции, данные немецкой статистики вполне подходят для наших целей.

Для оценки качества предлагаемых моделей исходные данные разбиваются на две группы: «тестовую» на периоде оценки параметров, и «контрольную», для которой рассчитываются метрики качества моделей (MAE, MAPE). Для «тестирования» выбран промежуток времени 01.01.2011–01.07.2019, начало которого обусловлено ограничением объясняющих переменных (GTI изменил методологию подсчета в начале 2011 г.), а для контрольной выборки взят период 01.08.2019–01.08.2022. Качество моделей оценивается несколькими способами:

1) для различных по времени прогнозных интервалов считаются значения метрик MAE, MAPE для исходных данных о миграции на 1–3 года вперед от начала контрольной выборки 01.08.2019;

2) методом кросс-валидации делается прогноз на 1 месяц вперед, при этом на каждом шаге тестовый интервал расширяется на 1 месяц, а контрольный — сужается¹⁰.

При этом можно сделать следующие наблюдения (рис. 1). Первый год контрольной выборки не имеет характерных шоков, способствующих существенному изменению миграционной активности. Второй год совпадает с периодом ограничений на передвижения из-за пандемии Covid-19¹¹. В третий год также возможны шоки на фоне проведения СВО в России. Кроме этого, из-за малых различий в моделях, обученных на данных 01.01.2011–01.07.2019 и 01.01.2011–01.07.2022, значения прогноза на 3 года фактически совпадают с прогнозом методом кросс-валидации (см. табл. 3 в разделе 3.3).

¹⁰ Также проводилась кросс-валидация с прогнозом на 3 и 6 месяцев, результаты оказались схожими.

¹¹ <https://www.who.int/europe/emergencies/situations/covid-19>.

3.2. Google Trends Index

В качестве фактора, объясняющего наблюдаемые потоки миграции, рассмотрим поисковые запросы. Одним из инструментов, представляющих информацию о популярности поисковых запросов в заданных тематиках, является Google Trends Index (GTI). Количество поисковых запросов, связанных с миграционной тематикой, можно использовать для прогноза динамики потоков миграции в перспективе нескольких месяцев.

Google Trends Index представляет собой индекс популярности интернет-запросов в заданной тематике, на рис. 2 представлен индекс по запросу «работа в Германии» в период с января 2011 по август 2022 г. среди поисковых запросов, осуществленных с территории Российской Федерации.

Основной особенностью индекса является относительный характер данных в рассматриваемом временном интервале, а не количество поисковых запросов за период. Фактически временные ряды нормируются на максимальное значение за выбранный период. Такой подход затрудняет возможность сравнения различных поисковых запросов в одинаковые промежутки времени, т.к. они нормированы на разные максимальные значения. Одним из способов решения описанной проблемы является стандартизация данных Google Trends (Fantazzini et al., 2021):

$$Z = \frac{X - E(X)}{\sigma(X)}, \quad (1)$$

где $E(X)$, $\sigma(X)$ — среднее значение и стандартное отклонение случайной величины X соответственно.

Стандартизация данных дает возможность проводить сравнения поисковых запросов и оценивать относительную популярность различных тем с течением времени (Приложение 2). Кроме того, такой подход позволит применять к преобразованным временным рядам статистические инструменты, такие как метод главных компонент (РСА), поскольку все ряды имеют одинаковый масштаб — нулевое среднее и единичную дисперсию.

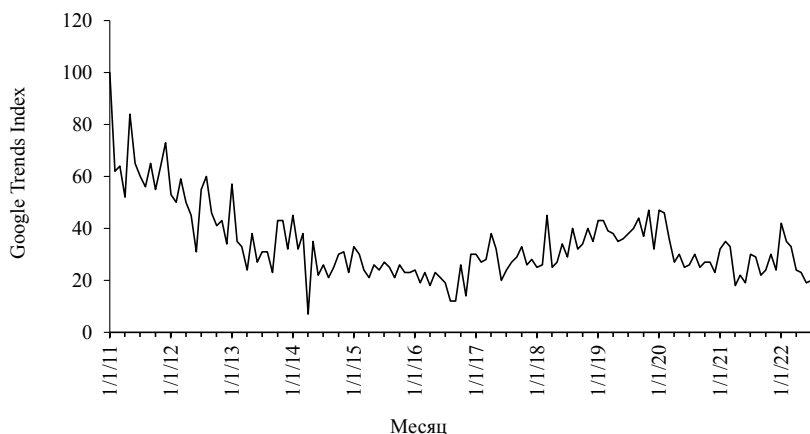


Рис. 2. Google Trends Index по запросу «Работа в Германии»
01.01.2011–01.08.2022

Статистику поисковых запросов также представляет Yandex Wordstat. В отличие от индекса GTI, статистика Яндекса содержит абсолютные показатели поисковых запросов, что упрощает их использование для прогноза. Однако такие данные доступны только за последний год. Также важно отметить, что недостатком статистики Yandex Wordstat является географическое смещение (доля Яндекса) — Яндекс имеет низкую долю среди поисковых систем за пределами стран СНГ, что не позволяет масштабировать предложенный алгоритм.

Другой важной особенностью данных GTI являются изменения в методологии построения индекса. Последнее существенное изменение произошло в 2011 г., когда Google изменил подход к определению региона поиска, что существенно образом повлияло на представляемую статистику, затруднив совместное использование данных до и после изменений. Во избежание ошибок в моделях в данной работе используются данные за период 01.01.2011–01.08.2022. Одним из шагов при работе с Google Trends является подбор ключевых слов (поисковых запросов), связанных с миграцией. В работе (Броницкий, Вакуленко, 2022) описан подход выделения ключевых слов с использованием методов машинного обучения. Изучаются два основных подхода для поиска множества поисковых слов. Первый — использование заготовленных NLP моделей, называемых «корпус». Используя корпус слов, можно найти наиболее близкие слова к произвольному слову русского языка (в работе используется слово «миграция»). Такие корпуса могут быть получены при анализе текстов Википедии¹², русскоязычных толковых словарей¹³ и художественной литературы. Второй подход — использование сервиса Yandex Wordstat¹⁴, который позволяет найти множество поисковых слов, наиболее близких к заданному на основе поисковой истории. Этот подход оказался более точным, поскольку множество слов учитывает лексику, используемую именно при поиске в сети Интернет. Дальнейшее исследование проводится с использованием поисковых запросов из указанной работы, полученных с использованием сервиса Yandex Wordstat (Приложение 1).

3.3. Методология и результаты

В качестве зависимой переменной в модели миграции берется показатель «прибытия иностранцев» из России в Германию Y_t , $t = 1, \dots, T$ — номер месяца, $X_1, \dots, X_K$ — объясняющие переменные (поисковые запросы GTI), приведенные в табл. П1 в Приложении 1, $\varepsilon_t \sim iid(0, \sigma^2)$ — ошибки регрессии. Было отобрано 36 GTI в качестве объясняющих переменных. Анализ автокорреляционной функции (ACF), частичной автокорреляционной функции (PACF), а также тестов на автокорреляцию показывает наличие годовой сезонности в исходных данных миграции и GTI. Изучаемые ряды являются нестационарными (по результатам ADF теста). Для решения проблемы нестационарности исходных рядов в работе применяется переход к сезонным разностям с периодом в 12 месяцев для всех переменных модели (Приложение 2).

¹² <http://vectors.nlpl.eu/repository/20/182.zip>.

¹³ <https://ruscorpora.ru/>.

¹⁴ <https://wordstat.yandex.ru>.

Для оценки лаговой структуры поисковых запросов добавляются лаги GTI, т.е. оценивается модель распределенных лагов $l=1, \dots, 12$ месяцев для переменных в сезонных разностях¹⁵:

$$Y_t - Y_{t-12} = \beta_0 + \sum_{k=1}^K \sum_{l=0}^{12} \beta_{k,l} (X_{k,t-l} - X_{k,t-l-12}) + \varepsilon_t. \quad (2)$$

Для оценивания модели с большим числом лагов для каждой независимой переменной необходимо существенным образом уменьшить размерность исходных данных. Такая потребность возникает, с одной стороны, из-за ограниченного числа наблюдений в тестовом периоде (примерно 100 значений) и желанием включить лаги GTI от 1 до 12 месяцев, что в 12 раз увеличит число объясняющих переменных. Можно выделить два метода уменьшения размерности данных, использованных в работе и приведенных в табл. 2 и Приложении 3.

1. Декомпозиция R^2 по методу Шепли (Israeli, 2007). Подход позволяет отобрать регрессоры, вносящие наибольший вклад в объясненную долю дисперсии R^2 модели миграции (2) с GTI.

2. Метод главных компонент (PCA) по основным тематикам. Выбираются факторы, объединенные в группы (учеба, работа, посольство)¹⁶ — основные тематики, вносящие наибольший вклад в объясненную долю дисперсии (п. 1)). Выбираются такие главные компоненты по каждой группе, при которых коэффициенты имеют наименьшие p -значения в моделях миграции (Айвазян, 2012).

В таблицах 2 и 3 представлены оцененные модели в сезонных разностях. В базовой модели в качестве регрессоров используются PCA-векторы¹⁷, объединенные по тематикам «учеба», «работа», «посольство» (список используемых запросов приведен в Приложении 1). Собственные значения и векторы оцениваются по данным только тестовой выборки, а затем применяются ко всем доступным данным для оценки значений главных компонент. При исследовании результатов базовой модели была выдвинута гипотеза о том, что экономические и социальные шоки влияют на число запросов миграционных тематик несимметрично, т.е. в моменты роста и падения поисковых запросов миграционные потоки реагируют с разной интенсивностью. Одним из способов смоделировать влияние шоков стало добавление фиктивных переменных, характеризующих рост и падение GTI. Такой подход не дал значимых улучшений в прогнозных качествах относительно исходного, однако исследование поведения мигрантов в момент шоков является интересной темой для дальнейших исследований.

¹⁵ Большее число лагов не рассматривается по нескольким причинам. С одной стороны, сильно растет размерность модели, что усложняет ее оценку, т.к. имеются ограничения по длине исследуемых рядов, а с другой стороны, миграционные намерения, как правило, реализуются в течение года.

¹⁶ Также был сформирован список «остальные», в него попали поисковые запросы либо с низким значением объясняющей доли R^2 , либо не соответствующие ни одной из тематик выше. В моделях не использовались тематики «виза» и «авиабилеты», т.к. они относятся не только к миграции с переменной места жительства, но и к туризму.

¹⁷ Здесь и далее под PCA-векторами имеются в виду главные компоненты.

Таблица 2. Описание моделей в сезонных разностях

Модель	Исходные данные	Описание
РСА по тематикам без лагов	Стандартизированные GPI индексы в разностях	Выделены 3 тематики: «учеба», «работа», «посольство». Для каждой определены РСА-векторы. Используется только первая главная компонента
РСА по тематикам + фиктивные переменные	РСА по тематикам	Используются 2 фиктивные переменные в модели (на константу и угловой коэффициент): 1) бинарная переменная, соответствующая 5%-ным верхним и нижним перцентилям для GPI «посольство»; 2) произведение данной переменной на РСА-вектор «посольство». Отвечают за быстрое изменение поисковой активности на фоне шоков
РСА-вектор «учеба» /«работа»/«посольство» с лагами	РСА по тематикам	Отдельно для каждой тематики взята первая главная компонента с лагом от 1 до 12 месяцев. Перебираются все возможные комбинации лагов, наилучшая модель определяется по AIC-критерию
РСА-векторы «учеба», «работа», «посольство» с лагами	РСА по тематикам	Для всех тематик одновременно взяты РСА-векторы с лагами, определенные в моделях по отдельным тематикам. По AIC-критерию определяются наилучшие лаги и модель

Таблица 3. Критерии качества прогноза миграции в моделях с сезонными разностями

Модель	Число регрессоров	Критерии и прогнозный период							
		MAE				MAPE			
		1 год	2 года	3 года	1 мес. ^{а)}	1 год	2 года	3 года	1 мес.
РСА по тематикам без лагов	4	374.1	348.9	416.1	412.6	0.75	0.55	0.5	0.49
РСА по тематикам + фиктивная переменная	6	379.3	355.5	423.1	418.7	0.77	0.56	0.51	0.5
РСА-вектор «учеба» с лагами	3	471.4	471.7	492.7	484.1	0.87	0.7	0.59	0.58
РСА-вектор «работа» с лагами	8	696.3	689.6	667	654.8	1.38	1.05	0.85	0.83
РСА-вектор «посольство» с лагами	9	257.2	196.9	310.1	305.4	0.44	0.30	0.32	0.31
РСА-векторы «учеба», «работа», «посольство» с лагами	13	177.8	216.1	305.0	295.2	0.35	0.32	0.32	0.31
SARIMA		311.3	349.4	345.9	345.9	0.68	0.58	0.47	0.47

Примечание. ^{а)} Прогноз на 1 месяц вперед в промежутке 01.08.2019–01.08.2022. Тестовая выборка увеличивается на 1 месяц на каждом этапе. Полужирным выделены наилучшие результаты среди рассматриваемых моделей.

В модели РСА с лагами были применены следующие подходы: по отдельности для каждой из тематик «учеба», «работа», «посольство» оценивались модели с включением лагов от 1 до 12. Для определения числа лагов выбиралась модель, соответствующая наименьшему значению информационного критерия (AIC) среди всех возможных моделей с лагами с одинаковым числом наблюдений (Приложение 3). Среди оцененных таким образом

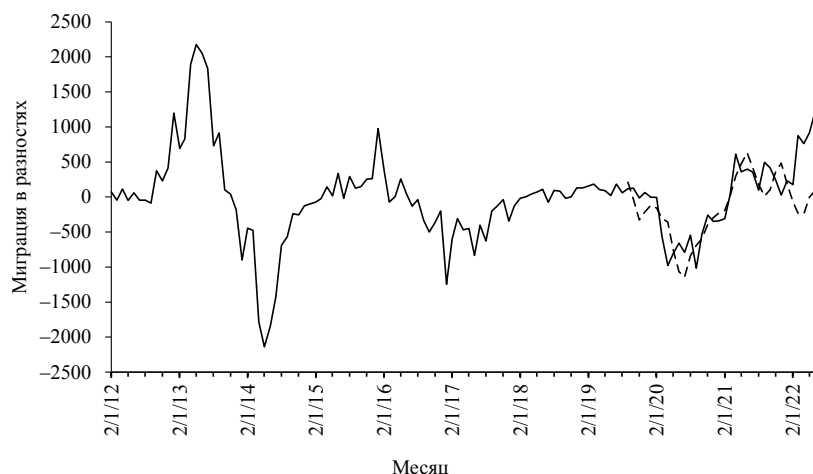


Рис. 3. Изменение числа миграций из РФ в Германию 2011–2022 гг.
Примечание. Пунктирной линией отмечено предсказание изменения миграций с использованием модели РСА-вектор «посольство» с лагами для контрольного периода 01.08.2019–01.08.2022.

моделей наименьшим значением АИС-критерия обладает модель РСА-вектор «посольство» с лагами, которая также обладает хорошим качеством для периодов без шоков, и наилучшим качеством для постковидного периода (см. рисунки 3 и 4) среди оцененных моделей.

Кроме этого, была оценена объединенная модель (РСА-векторы «учеба», «работа», «посольство» с лагами) с использованием значимых лагов, определенных в моделях по тематикам. Для нахождения наилучшей модели также использовался АИС-критерий.

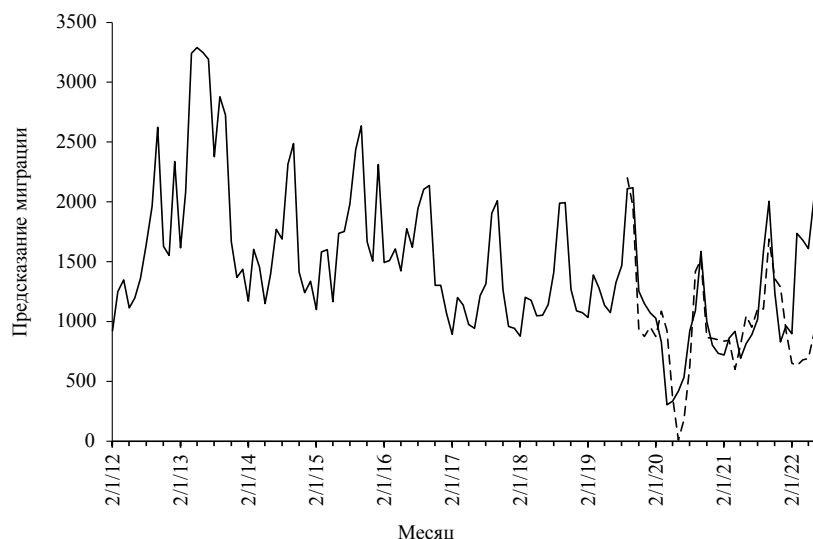


Рис. 4. Миграции из РФ в Германию 2011–2022 гг., чел.
Примечание. Пунктирной линией отмечено предсказание количества миграций с использованием модели РСА-вектор «посольство» с лагами для контрольного периода 01.08.2019–01.08.2022.

Полученная модель обладает схожим качеством с моделью РСА-вектор «посольство» с лагами, что позволяет использовать в модели миграции только вектор «посольство» и существенно упрощает модель. Для сравнения с моделями с GTI индексами была также оценена SARIMA(0,1,1)×(0,1,1,12), которая оказалась хуже по прогнозным качествам, поскольку SARIMA модели не учитывают внешние изменения.

Для модели с распределенными лагами (2) можно измерить вклад лагов при помощи оценки среднего лага L_k в рамках заданной тематики поисковых запросов k :

$$L_k = \sum_{l=0}^{12} l\beta_{k,l}^2 / \sum_{l=0}^{12} \beta_{k,l}^2. \quad (3)$$

Для вычисления значения среднего лага L_k берутся квадраты коэффициентов при лагах $l = 1, \dots, 12$ из модели (2) для того, чтобы корректно рассматривать случаи отрицательных коэффициентов $\beta_{k,l}$. Средний лаг считается по отдельности для каждой тематики k («работа», «учеба», «посольство»). Значения коэффициентов в оцененной модели являются случайными величинами, как и величина среднего лага (3). Для оценки доверительного интервала сначала при помощи метода Монте-Карло оценивается распределение среднего лага, а затем определяется среднее для полученного распределения при помощи процедуры бутстрэп (Приложение 3). Большие значения среднего лага L характеризуют больший вклад лагов, близких к 12 месяцам, что может означать большую задержку факта миграции во времени от момента поиска. В то же время малые значения среднего лага L соответствуют более быстрой реакции миграции на изменение GTI. Оказалось, что для тематики «посольство» $L = 5.6$, а 95%-ный доверительный интервал есть [3.64; 7.92]¹⁸, в то время как для тематик «учеба» и «работа» $L = 8.0$ [5.36; 10.8] и $L = 6.5$ [4.72; 8.21] соответственно. Такие значения средних лагов можно объяснить временем, необходимым для поиска учебы или работы, и необходимостью оформлять визу для граждан России. Потенциальные мигранты начинают заранее узнавать о возможности учиться или работать в другой стране, затем обращаются в посольство для получения визы или вида на жительство. Данный результат может быть справедливым и для миграции из России в другие страны, где требуется получение визы. В таких случаях для прогнозирования миграции достаточно будет информации о поисковых запросах, связанных с тематикой «посольство».

Авторы также оценили величину среднего лага в работе (Цапенко, Юревич, 2022), где такой показатель не рассчитывался. В этой работе исследовалась миграция из Киргизии, Таджикистана и Узбекистана в Россию с использованием GTI с лагами по тематике «работа». Оцененная величина среднего лага для этих стран равна $L_k = 9.67$, $L_T = 9$, $L_V = 11$ месяцам соответственно, что несколько выше оценок из настоящей статьи. В работах (Wladyka, 2017; Wanner, 2021) нет возможности оценить величину среднего лага, т.к. авторы оценивали модели с каждым лагом по отдельности, а не единую модель с распределенными лагами.

¹⁸ Представлены оценки средних лагов из моделей, где поисковые запросы по тематикам включались по отдельности. В Приложении 3 также есть оценки средних лагов для объединенной модели со всеми тематиками одновременно. В этом случае средний лаг для тематики «посольство» немного ниже (4.3), а для «работы» (7.9) и «учебы» (11) — выше, однако доверительные интервалы пересекаются.

Таким образом, полученные оценки хорошо интерпретируемы и подтверждают необходимость включать лаги при моделировании миграции, позволяя учитывать факт задержки во времени между поиском и миграцией.

4. Заключение

В данной работе предложена методика прогнозирования миграционной статистики с минимальной задержкой во времени, так называемый наукастинг статистики миграции. Хотя подобная идея и описывалась в более ранних работах, но она неприменима в явном виде к показателям Росстата и статистических служб широкого круга стран. За счет того, что статистику поисковых веб-запросов можно формировать в реальном времени, потоки миграции можно оценивать, исходя из связи между ними в прошлом. Работоспособность предложенной методики продемонстрирована на примере миграционных потоков из России в Германию. В отличие от работы (Броницкий, Вакуленко, 2022), в данном исследовании изучается глубина лага поисковых запросов в зависимости от их тематики (посольство, работа, учеба), снижается размерность модели, предложены подходы к выбору ключевых поисковых запросов и построению на их основе агрегированных индексов.

Полученные результаты показали, что модели миграции с GTI индексами лучше прогнозируют миграцию, чем SARIMA модели. Для оценки миграции из РФ можно использовать только поисковые слова, связанные с тематикой «посольство», т.к. практически всем мигрантам необходимо оформить разрешение на пребывание в другой стране. Модели, оцененные с использованием всех рассматриваемых ключевых запросов («работа», «учеба», «посольство») не показывают значимого улучшения прогнозной силы по сравнению с запросами только о посольстве.

В работе исследовалась лаговая структура для различных типов поисковых запросов (посольство, работа, учеба), агрегированных с помощью метода главных компонент по смысловым группам, которые имеют различную лаговую структуру в модели. Расчеты среднего лага для моделей распределенных лагов показали, что статистика миграции быстрее всего реагирует на поисковые запросы, связанные с посольством, средний лаг для которого равен 5.6 месяцев. А для запросов, связанных с учебой и работой, средние лаги больше (8.0 и 6.5 месяцев соответственно). Это означает, что потенциальные мигранты начинают искать учебное заведение в среднем примерно за 8 месяцев до реального перемещения и фиксации их в статистике, а работу — за 6.5 месяцев.

Стоит отметить имеющиеся ограничения предложенных подходов. Во-первых, при работе с данными о суммарных потоках миграции без выделения целей (трудовая, учебная, воссоединение семей и т.д.) приходится рассматривать все виды поисковых слов, что добавляет дополнительный шум в модели и увеличивает размерность. Во-вторых, при сравнении прогнозных значений вне обучающей выборки с официальной миграционной статистикой необходимо принимать во внимание ее несовершенство и возможные расхождения с истинными объемами миграции (Tjaden, 2021), особенно в периоды шоков (пандемии, экономические и социально-политические кризисы и т.д.), когда объемы миграции значительно отклоняются от тренда.

Тем не менее, предложенные подходы позволяют развить методологию прогнозирования статистических показателей, а также расширить возможности использования поисковых

запросов для оценки не только миграции населения, но и иных экономических индикаторов. Возможным направлением будущих исследований может стать тестирование асимметрии реакции миграционных потоков на изменение динамики поисковых запросов, а также исследование глубины лагов в периоды шоков.

Благодарности. Авторы выражают благодарность О. С. Чудиновских за помощь в получении доступа к данным миграционной статистики статистической службы Германии¹⁹, а также анонимным рецензентам и А. Д. Слостникову за ценные комментарии и предложения.

Литература

- Айвазян С. А. (2012). *Анализ качества и образа жизни населения (эконометрический подход)*. М: Наука.
- Броницкий Г. Т., Вакуленко Е. С. (2022). Прогнозирование миграции из России в Германию с использованием Google-трендов. *Демографическое обозрение*, 9 (3), 75–92. DOI: 10.17323/demreview.v9i3.16471.
- Вакуленко Е. С., Цимайло В. В. (2011). Учет нелегальной миграции населения: методы и оценки. *Демоскоп Weekly*, 479–480.
- Воробьева О. Д., Гребенюк А. А. (2017). Сравнительный анализ отечественной и зарубежной статистической информации об эмиграции граждан России. *Вопросы статистики*, 1 (9), 64–73.
- Денисенко М. Б. (2003). Эмиграция из России по данным зарубежной статистики. *Мир России: Социология, этнология*, 12 (3), 157–169.
- Цапенко И. П., Юревич М. А. (2022). Статистика онлайн-запросов в наукастинге миграции. *Экономические и социальные перемены: факты, тенденции, прогноз*, 15 (1), 74–89. DOI: 10.15838/esc.2022.1.79.4.
- Чудиновских О. С. (2018). Большие данные и статистика миграции. *Вопросы статистики*, 25 (2), 48–56.
- Чудиновских О. С., Степанова А. В. (2020). О качестве федерального статистического наблюдения за миграционными процессами. *Демографическое обозрение*, 7 (1), 54–82. DOI: 10.17323/demreview.v7i1.10820.
- Чудиновских О. С. (2020). Оценки масштабов нелегальной миграции в России: возможные подходы и источники информации. *Вопросы статистики*, 27 (1), 8–28. DOI: 10.34023/2313-6383-2020-27-1-8-28.
- Avramescu A., Wiśniowski A. (2021). Now-casting Romanian migration into the United Kingdom by using Google Search engine data. *Demographic Research*, 45, 1219–1254. DOI: 10.4054/DemRes.2021.45.40.
- Bengtsson L., Lu X., Thorson A., Garfield R., von Schreeb J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Medicine*, 8 (8). DOI: 10.1371/journal.pmed.1001083.

¹⁹ <https://www-genesis.destatis.de/genesis/online>.

- Benson-Rea M., Rawlinson S. (2003). Highly skilled and business migrants: Information processes and settlement outcomes. *International Migration*, 41 (2), 59–79. DOI: 10.1111/1468-2435.00235.
- Böhme M. H., Gröger A., Stöhr T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142, 102347. DOI:10.1016/j.jdeveco.2019.04.002.
- Chi G., State B., Blumenstock J. E., Adamic L. (2020). Who ties the world together? Evidence from a large online social network. In: Cherifi H., Gaito S., Mendes J., Moro E., Rocha L. (eds). *Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence*, 882. DOI: 10.1007/978-3-030-36683-4_37.
- Choi H., Varian H. (2009). Predicting the present with Google Trends. Technical report, Google. http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.
- Choi H., Varian H. (2012). Predicting the present with Google Trends. *Economic Record*, 88, 2–9. DOI: 10.5018/economics-ejournal.ja.2018-34.
- Fantazzini D., Pushchelenko J., Mironenkov A., Kurbatskii A. (2021). Forecasting internal migration in Russia using Google Trends: Evidence from Moscow and Saint Petersburg. *Forecasting*, 3 (4), 774–803. DOI:10.3390/forecast3040048.
- Ginsberg J., Mohebbi M., Patel R., Brammer L., Smolinski M., Brilliant L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1014. DOI: 10.1038/nature07634.
- Golenvaux N., Alvarez P. G., Kiossou H. S., Schaus P. (2020). An LSTM approach to forecast migration using Google Trends. DOI: 10.48550/arXiv.2005.09902.
- Israeli O. (2007). A Shapley-based decomposition of the R-square of a linear regression. *Journal of Economic Inequality*, 5, 199–212. DOI: 10.1007/s10888-006-9036-6.
- Jun S. P., Yoo H. S., Choi S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, 130, 69–87. DOI: 10.1016/j.techfore.2017.11.009.
- Jurić T. (2022). Facebook and Google as an empirical basis for the development of a method for monitoring external migration of Croatian citizens. *Ekonomski Pregled*, 73 (2), 186–214. DOI: 10.32910/ep.73.2.2.
- Kikas R., Dumas M., Saabas A. (2015). Explaining international migration in the Skype network. *SideWayS '15: Proceedings of the 1st ACM Workshop on Social Media World Sensors*, 17–22. DOI: 10.1145/2806655.2806658.
- Kim J., Sîrbu A., Giannotti F., Gabrielli L. (2020). Digital footprints of international migration on Twitter. In: Berthold M., Feelders A., Krempel G. (eds). *Advances in Intelligent Data Analysis XVIII. IDA 2020. Lecture Notes in Computer Science*, vol. 12080, 274–286. Springer. DOI: 10.1007/978-3-030-44584-3_22.
- Moise I., Zurich E., Gaere E., Merz R., Pournaras E. (2016). Tracking language mobility in the Twitter landscape. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 663–670. DOI: 10.1109/ICDMW.2016.0099.
- State B., Rodriguez M., Helbing D., Zagheni E. (2014). Migration of professionals to the U.S. evidence from LinkedIn data. In: L. M. Aiello, D. McFarland (eds). *6th International Conference on Social Informatics*, 531–543. DOI: 10.1007/978-3-319-13734-6_37.

- Subbotin A., Aref S. (2021). Brain drain and brain gain in Russia: Analyzing international migration of researchers by discipline using Scopus bibliometric data 1996–2020. *Scientometrics*, 126 (9), 7875–7900. DOI: 10.1007/s11192-021-04091-x.
- Tjaden J. (2021). Measuring migration 2.0: A review of digital data sources. *Comparative Migration Studies*, 9 (1), 59. DOI: 10.1186/s40878-021-00273-x.
- Wanner P. (2021). How well can we estimate immigration trends using Google data? *Quality & Quantity*, 55 (4), 1181–1202. DOI: 10.1007/s11135-020-01047-w.
- Wladyka D. K. (2017). Queries to Google Search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies*, 25 (4), 312–327.
- Zagheni E., Garimella V., Weber I., State B. (2014). Inferring international and internal migration patterns from Twitter data. In: *WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web*, 439–444. DOI: 10.1145/2567948.2576930.
- Zagheni E., Weber I. (2012). You are where you E-mail: Using E-mail data to estimate international migration rates. In: *Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12*, 348–351. DOI: 10.1145/2380718.2380764.

Поступила в редакцию 14.04.2023;
принята в печать 20.09.2023.

Приложение 1

Таблица П1. Описательная статистика переменных, используемых для оценки миграции из РФ в Германию за период 2011–2022 гг.

Количество прибытий иностранцев из РФ в Германию				
	Минимум	Максимум	Стандартная ошибка	Среднее
	305	3289	598.3	1481.7
Стандартизированные GTI индексы				
	Переменная		Минимум	Максимум
Работа	работа в Германии для русских		−1.16	4.76
	работа в Германии		−1.90	4.66
	работа в Германии вакансии		−0.80	4.77
	работа в Германии без знания языка		−0.55	5.90
	работа в Германии для русских вакансии		−0.39	6.68
	релокация в Германию		−0.40	6.69
Учеба	учеба в Германии		−1.15	5.17
	виза для учебы в Германии		−1.05	4.72
Посольство	посольство Германии		−0.79	3.73
	посольство Германии в Москве		−0.97	4.33
	шенгенская виза в Германию		−1.39	4.27
	посольство России в Германии		−1.06	4.13
	как получить гражданство Германии		−1.91	3.18
	виза центр Германии		−1.24	3.55

Окончание табл. III

Остальные	посольство Германии в Москве официальный сайт	-0.41	6.47
	виза в Германию для россиян	-0.83	3.88
	виза в Германию	-1.95	2.60
	вид на жительство в Германии	-0.98	4.83
	как получить вид на жительство в Германии	-0.61	4.04
	Германия вид на жительство недвижимость	-0.38	6.28
	переезд в Германию на пмж	-0.45	4.74
	оформление визы в Германию	-1.04	4.96
	гостевая виза в Германию	-1.06	4.78
	пмж в Германии	-0.96	4.55
	переезд в Германию	-1.58	4.18
	как получить визу в Германию	-1.51	5.38
	рабочая виза в Германию	-1.20	4.44
	получение визы в Германию	-1.04	4.02
	документы для визы в Германию	-1.34	4.29
	Германия без визы	-0.52	4.38
	жилье в Германии	-0.88	3.97
	авиабилеты в Германию	-1.21	4.77
	открыть визу в Германию	-0.98	4.22
	анкета на визу в Германию	-1.04	4.02
	сколько стоит виза в Германию	-1.42	3.04

Примечание. Всего — 140 наблюдений по времени. У стандартизированных индексов нулевое среднее и единичная дисперсия, поэтому они не приводятся. Курсивом выделены поисковые запросы, объясняющие наибольшую долю R^2 по методу Шепли в заданных тематиках.

Приложение 2

Анализ и обработка временного ряда Google Trends Index

Приведем пример стандартизации GTI для запроса «работа в Германии» в промежутке 2011–2022 гг.

Исходный показатель нормирован на максимальное значение за период (рис. 2), значение GTI = 100 означает максимальную заинтересованность тематикой в соответствующую этому значению дату. Используя Z-оценку (1), можно привести исходные значения индекса (рис. 2) к единому распределению с нулевым средним и единичной дисперсией (рис. П1), что позволит сравнивать популярность различных поисковых запросов за один промежуток времени.

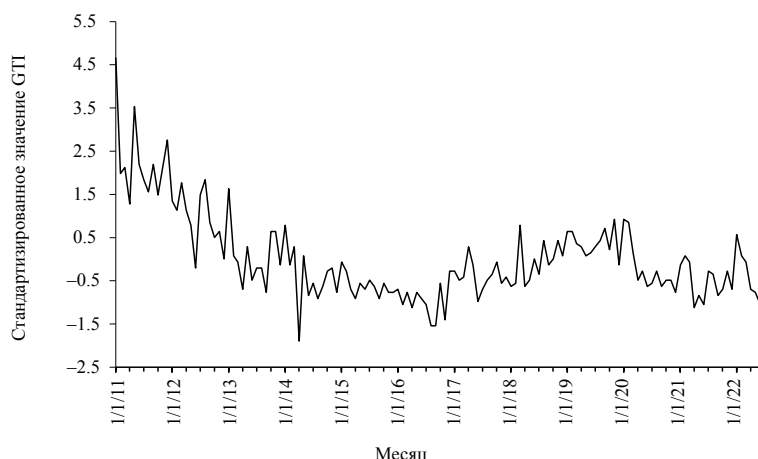


Рис. П1. Стандартизованный GTI для тематики «Работа в Германии» 2011–2022 гг.

Кроме этого, оценивая сезонную разность стандартизованного GTI за 12 периодов, получим разностный ряд для нормированного индекса (рис. П2).

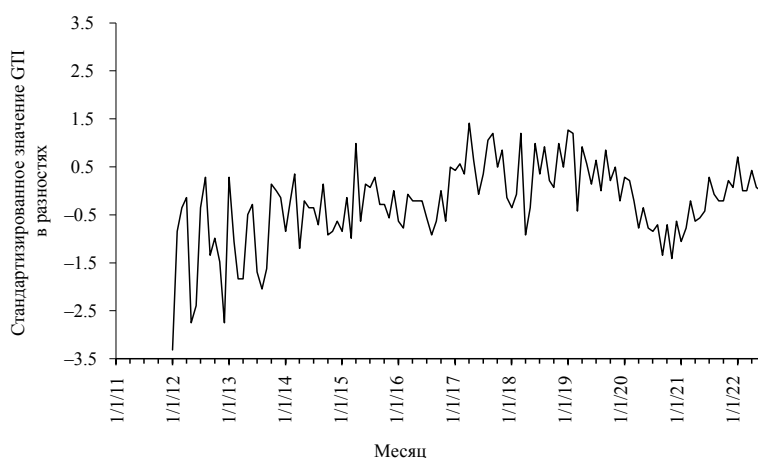


Рис. П2. Сезонная разность стандартизованного GTI для тематики «Работа в Германии» 2011–2022 гг.

Приложение 3

Таблица П2. Оценки моделей распределенных лагов (2) для различных поисковых запросов

Переменная	Лаг, мес.	РСА по тематикам	РСА с фиктивной переменной	РСА-вектор «посольство» с лагами	РСА-вектор «учеба» с лагами	РСА-вектор «работа» с лагами	AIC- критерий «учеба», «работа», «посольство»
Константа		1.49 (65.30)	-9.42 (66.05)	-39.62 (55.15)	-40.18 (76.67)	-41.49 (56.97)	-63.43 (60.68)
РСА-вектор «посольство»	0	-174.33*** (56.48)	-161.21** (61.80)	-177.98*** (56.57)			-286.96*** (56.47)
	1			-212.96*** (57.94)			-286.96*** (56.47)
	2			-223.17*** (58.19)			-138.62** (62.66)
	3			-102.92* (54.00)			-144.19** (58.99)
	9			104.09* (55.26)			124.52** (56.17)
	10			165.59*** (50.90)			
	11			108.38** (49.03)			
РСА-вектор «учеба»	0	39.54 (50.04)	52.34 (53.57)				
	5				137.99** (63.55)		
	11				-139.21*** (52.35)		-123.51*** (42.51)
РСА-вектор «работа»	0	-88.95* (48.49)	-87.46* (47.07)			-113.81** (46.84)	
	1					-128.53*** (45.71)	
	2					-120.33*** (43.49)	
	6					-101.93** (48.22)	
	7					-133.07*** (46.66)	-85.27* (45.76)
	9					93.54** (44.08)	71.91 (43.50)
	10					129.58*** (40.13)	
	11					129.45*** (36.66)	

Г. Т. Броницкий, Е. С. Вакуленко

Переменная	Лаг, мес.	PCA по тематикам	PCA с фиктивной переменной	PCA-вектор «посольство» с лагами	PCA-вектор «учеба» с лагами	PCA-вектор «работа» с лагами	AIC- критерий «учеба», «работа», «посольство»
Фиктивная переменная: GTI «посольство» больше 95%-го персентилья (Фиктивная переменная)× (PCA-вектор «посольство»)			2399.77* (1272.64)				
			778.50* (462.38)				
Наблюдений		92	92	80	80	80	80
AIC		1449	1449	1226	1273	1231	1240
BIC		1459	1464	1245	1281	1253	1259
R ²		0.19	0.22	0.58	0.14	0.56	0.50
F Statistic		6.82***	4.91***	14.15***	6.34***	11.52***	10.43***
Средний лаг [95%-ный доверительный интервал]				5.6 [3.64; 7.92]	8.0 [5.36; 10.8]	6.5 [4.72; 8.21]	Посольство – 4.3 [2.45; 7.18]. Учеба – 11 ^{а)} . Работа – 7.9 [7.01; 8.98]

Примечание. Значимость коэффициентов модели: *** — $p < 0.01$, ** — $p < 0.05$, * — $p < 0.1$. В круглых скобках представлены стандартные ошибки.

^{а)} При включении в модель единственного лага величина среднего лага (3) не является случайной величиной, поэтому доверительный интервал для нее оценить невозможно.

Bronitsky G., Vakulenko E. Using Google Trends to forecast migration from Russia: Search query aggregation and accounting for lag structure. *Applied Econometrics*, 2024, v. 73, pp. 78–101.

DOI: 10.22394/1993-7601-2024-73-78-101.

Georgy Bronitsky

HSE University, Moscow, Russian Federation;
getibr@gmail.com

Elena Vakulenko

HSE University, Moscow, Russian Federation;
evakulenko@hse.ru

Using Google Trends to forecast migration from Russia:
Search query aggregation and accounting for lag structure

This paper proposes an approach for predicting migration statistics using Google Trends Index (GTI) search query data. We improved the existing methodology in two directions: firstly, we proposed an approach

of aggregating key search queries based on various statistical criteria; secondly, we showed the importance of including in the migration model the time lag structure of search queries, depending on migration goals and the associated GTIs. We demonstrate the performance of the proposed approaches on monthly data from the German statistical office on migration volume from Russia to Germany from January 2011 to August 2022. The results show that distributed lag migration models with GTI are better predict migration than SARIMA models. Average lag estimates, i.e. the reaction time of migration statistics to search queries on the topics “embassy”, “work” and “study”, were 5.6, 6.5 and 8 months, respectively. We demonstrate that for forecasting migration from Russia to Germany, it is sufficient to consider only search queries related to the topic “embassy”.

Keywords: international migration; Russia; Germany; Google Trends; search queries; nowcasting; big data; forecasting.

JEL classification: J61; C53; C38.

References

- Aivazian S. A. (2012). *Quality of life and living standards analysis*. M: Nauka (in Russian).
- Bronitsky G., Vakulenko E. (2022). Using Google Trends for external migration prediction. *Demographic Review*, 9 (3), 75–92 (in Russian). DOI: 10.17323/demreview.v9i3.16471.
- Vakulenko E. S., Tsimailo V. V. (2011). Accounting for illegal migration of the population: Methods and estimates. *Demoscope Weekly*, 479–480 (in Russian).
- Vorob'eva O. D., Grebenyuk A. A. (2017). Comparative analysis of domestic and foreign statistics on the statistics on the emigration of Russian citizens, 1 (9), 64–73 (in Russian).
- Denisenko M. B. (2003). Emigration from Russia according to foreign statistics. *World of Russia: Sociology, Ethnology*, 12 (3), 157–169 (in Russian).
- Tsapenko I. P., Yurevich M. A. (2022). Nowcasting migration using statistics of online queries. *Ekonomicheskie i Sotsialnye Peremeny*, 15 (1), 74–89 (in Russian). DOI: 10.15838/esc.2022.1.79.4.
- Chudinovskikh O. S. (2018). Big Data and statistics on migration. *Voprosy statistiki*, 25 (2), 48–56 (in Russian).
- Chudinovskikh O. S. (2020). Assessing the scale of illegal migration in Russia: Possible approaches and sources of information. *Voprosy statistiki*, 27 (1), 8–28 (in Russian). DOI: 10.34023/2313-6383-2020-27-1-8-28.
- Chudinovskikh O. S., Stepanova A. V. (2020). On the quality of the federal statistical observation of migration processes. *Demographic Review*, 7 (1), 54–82 (in Russian). DOI: 10.17323/demreview.v7i1.10820.
- Avramescu A., Wiśniowski A. (2021). Now-casting Romanian migration into the United Kingdom by using Google Search engine data. *Demographic Research*, 45, 1219–1254. DOI: 10.4054/DemRes.2021.45.40.
- Bengtsson L., Lu X., Thorson A., Garfield R., von Schreeb J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Medicine*, 8 (8). DOI: 10.1371/journal.pmed.1001083.
- Benson-Rea M., Rawlinson S. (2003). Highly skilled and business migrants: Information processes and settlement outcomes. *International Migration*, 41 (2), 59–79. DOI: 10.1111/1468-2435.00235.

- Böhme M. H., Gröger A., Stöhr T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142, 102347. DOI:10.1016/j.jdeveco.2019.04.002.
- Chi G., State B., Blumenstock J. E., Adamic L. (2020). Who ties the world together? Evidence from a large online social network. In: Cherifi H., Gaito S., Mendes J., Moro E., Rocha L. (eds). *Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence*, 882. DOI: 10.1007/978-3-030-36683-4_37.
- Choi H., Varian H. (2009). Predicting the present with Google Trends. Technical report, Google. http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.
- Choi H., Varian H. (2012). Predicting the present with Google Trends. *Economic Record*, 88, 2–9. DOI: 10.5018/economics-ejournal.ja.2018-34.
- Fantazzini D., Pushchelenko J., Mironenkov A., Kurbatskii A. (2021). Forecasting internal migration in Russia using Google Trends: Evidence from Moscow and Saint Petersburg. *Forecasting*, 3 (4), 774–803. DOI:10.3390/forecast3040048.
- Ginsberg J., Mohebbi M., Patel R., Brammer L., Smolinski M., Brilliant L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1014. DOI: 10.1038/nature07634.
- Golenvaux N., Alvarez P. G., Kiossou H. S., Schaus P. (2020). An LSTM approach to forecast migration using Google Trends. DOI: 10.48550/arXiv.2005.09902.
- Israeli O. (2007). A Shapley-based decomposition of the R-square of a linear regression. *Journal of Economic Inequality*, 5, 199–212. DOI: 10.1007/s10888-006-9036-6.
- Jun S. P., Yoo H. S., Choi S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, 130, 69–87. DOI: 10.1016/j.techfore.2017.11.009.
- Jurić T. (2022). Facebook and Google as an empirical basis for the development of a method for monitoring external migration of Croatian citizens. *Ekonomski Pregled*, 73 (2), 186–214. DOI: 10.32910/ep.73.2.2.
- Kikas R., Dumas M., Saabas A. (2015). Explaining international migration in the Skype network. *SideWayS '15: Proceedings of the 1st ACM Workshop on Social Media World Sensors*, 17–22. DOI: 10.1145/2806655.2806658.
- Kim J., Sirbu A., Giannotti F., Gabrielli L. (2020). Digital footprints of international migration on Twitter. In: Berthold M., Feelders A., Krempel G. (eds). *Advances in Intelligent Data Analysis XVIII. IDA 2020. Lecture Notes in Computer Science*, vol. 12080, 274–286. Springer. DOI: 10.1007/978-3-030-44584-3_22.
- Moise I., Zurich E., Gaere E., Merz R., Pournaras E. (2016). Tracking language mobility in the Twitter landscape. In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 663–670. DOI: 10.1109/ICDMW.2016.0099.
- State B., Rodriguez M., Helbing D., Zagheni E. (2014). Migration of professionals to the U.S. evidence from LinkedIn data. In: L. M. Aiello, D. McFarland (eds). *6th International Conference on Social Informatics*, 531–543. DOI: 10.1007/978-3-319-13734-6_37.
- Subbotin A., Aref S. (2021). Brain drain and brain gain in Russia: Analyzing international migration of researchers by discipline using Scopus bibliometric data 1996–2020. *Scientometrics*, 126 (9), 7875–7900. DOI: 10.1007/s11192-021-04091-x.

- Tjaden J. (2021). Measuring migration 2.0: A review of digital data sources. *Comparative Migration Studies*, 9 (1), 59. DOI: 10.1186/s40878-021-00273-x.
- Wanner P. (2021). How well can we estimate immigration trends using Google data? *Quality & Quantity*, 55 (4), 1181–1202. DOI: 10.1007/s11135-020-01047-w.
- Wladyka D. K. (2017). Queries to Google Search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies*, 25 (4), 312–327.
- Zagheni E., Garimella V., Weber I., State B. (2014). Inferring international and internal migration patterns from Twitter data. In: *WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web*, 439–444. DOI: 10.1145/2567948.2576930.
- Zagheni E., Weber I. (2012). You are where you E-mail: Using E-mail data to estimate international migration rates. In: *Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12*, 348–351. DOI: 10.1145/2380718.2380764.

Received 14.04.2023; accepted 20.09.2023.