PE Population and Economics

18+

Население и экономика 8(2): 133–154
DOI 10.3897/popecon.8.e119577

ОРИГИНАЛЬНАЯ СТАТЬЯ

Наукастинг миграции с использованием Google Trends: применение для разных стран

Георгий Т. Броницкий [1]

1 *Национальный исследовательский университет «Высшая школа экономики», Москва, 101000, Россия*

Получено 26 January 2024 ♦ **Принято в печать** 25 March 2024 ♦ **Опубликовано** 12 September 2024

Цитирование: GT Bronitsky (2024) Migration nowcasting using Google Trends: cross-country application. Population and Economics 8(2):133–154. https://doi.org/10.3897/popecon.8.e119577

Аннотация

Анализ миграционных потоков имеет большое значение для понимания и прогнозирования социально-экономических тенденций в различных странах. В работе описан алгоритм, позволяющий получать оценки миграции с минимальной временной задержкой (наукастинг), для этого используются данные о поисковых запросах Google Trends Index (GTI). Прогнозная сила моделей сравнивается для различных временных периодов, один из которых включает ограничения, связанные с пандемией COVID-19, во время которой возможности для миграции значительно сократились. На примере полученных оценок для миграции из шести различных стран в Германию делаются следующие выводы: во-первых, при отсутствии внешних шоков достаточно использовать только один запрос вида «работа в Германии» на языке страны выбытия и его 12 временных лагов в моделях SARIMAX или моделях распределенных лагов для улучшения точности оценки миграции на помесячных данных по сравнению с SARIMA-моделями; во-вторых, при возникновении шоков в исследуемом периоде наилучшее качество показывает «мультизапросная» модель распределенных лагов, также учитывающая другие поисковые запросы, связанные с намерением мигрировать. Кроме этого, в работе предлагается усовершенствование существующей методологии прогнозирования миграции с помощью GTI, показывающее важность использования именно модели распределенных лагов, а не моделей с отдельными временными лагами GTI. Исследуемые модели с использованием GTI с лагами показывают лучшую предсказательную силу сравнительно с SARIMA-моделями для каждой из рассматриваемых стран как в период шоков, так и вне их.

Ключевые слова

большие данные, Германия, международная миграция, наукастинг, поисковые запросы, прогнозирование, Google Trends, SARIMA

Коды JEL: F22; J61; C53

Введение

Миграция играет существенную роль в формировании демографической, экономической и социальной политики как в стране отправления, так и стране прибытия. Однако миграционные процессы подвержены влиянию внешних шоков, таких как эпидемии, военные конфликты, природные катаклизмы и др. Наличие таких шоков затрудняет прогноз, а также несет риски для государств, связанные с оттоком работоспособной части населения. Важно проводить анализ миграционных потоков, а также намерений о миграции с минимальной задержкой во времени для своевременной реакции на возникающие шоки и уменьшением экономических рисков государства. При этом сложность в сборе статистических данных, задержка в публикации[1], а также неполнота фиксации всех миграционных перемещений создают сложности для проведения таких оценок. С развитием информационных технологий и появлением в сети Интернет «больших данных» о действиях людей появилось новое направление исследований, в основе которых лежит использование альтернативных источников данных для анализа различных экономических показателей [Jun et al., 2018] с минимальной задержкой во времени (наукастинг). В частности, предлагаются различные методы оценки миграционных потоков с использованием данных «цифрового следа» мигрантов [Tjaden, 2021]. К основным источникам такой информации можно отнести данные об активности в социальных сетях, информацию о GPS-положении SIM-карт, данные IP-адресов, а также статистику поисковых запросов, в том числе Google Trends Index (GTI) [Google Trends…], который предоставляет информацию о поисковых запросах в реальном времени.

Цель нашей работы – прогнозирование международной миграции с использованием данных GTI, тестирование применимости моделей распределенных лагов, SARIMA и SARIMAX для ряда стран. На примере нескольких стран проводится оценка качества полученных моделей для различных прогнозных периодов. Сравнивается прогнозная сила моделей для периодов, в которые наблюдались воздействия внешних шоков, а также для периодов вне шоков. При моделировании миграции использовались подходы, описанные в работах [Fantazzini et al., 2021; Цапенко, Юревич, 2022; Броницкий, Вакуленко, 2022], в которых рассматривались модели временных рядов (SARIMAX) с использованием GTI в качестве экзогенных переменных, а также модели распределенных лагов [Броницкий, Вакуленко, 2024]. Особое внимание уделяется работе [Wanner, 2021], в которой проводится сравнение моделей прогноза миграции на основе GTI для оценки потока мигрантов из различных стран в Швейцарию. В данной работе усовершенствуются существующие методы, в том числе описанные в работе [Wanner, 2021], за счет добавления в модель сразу нескольких временных лагов GTI. Кроме этого, описывается алгоритм сбора и обработки данных GTI для оценки миграции, проводится оценка как «однозапросных» моделей с лагами на примере запроса «работа в Германии» на языке мигранта, так и «мультизапросных» моделей с лагами [Броницкий, Вакуленко, 2024] – для них в качестве объясняющих переменных выбирается сразу несколько поисковых запросов, соответствующих теме миграции. В работе приводятся рекомендации по использованию моделей для будущих исследований, описываются преимущества и недостатки каждой из них.

Для прогнозирования миграции на помесячных данных в работе проводится сравнение следующих моделей: модель прогноза сезонных временных рядов SARIMA; модель SARIMAX с использованием экзогенных переменных (в качестве объясняющей переменной используется GTI для поискового запроса «работа в Германии» и его лагов от 1 до 12 месяцев); модель распределенных лагов – модель множественной регрессии для миграции, в качестве объясняющих факторов для которой выбирается GTI-запрос, а также его лаги. На примере миграции из Польши, Италии, Румынии, Испании, Болгарии и России в Германию оцениваются метрики

1 Задержка в подсчете помесячных данных статистическим офисом Германии составляет 3,5 месяца, а годовых данных – 6 месяцев.

качества прогноза для описанных выше моделей. Для выбранных стран фиксировались одни из наибольших значений показателя «прибытия иностранцев» на основе помесячных данных статистического офиса Германии с января 2011 по июнь 2023 г. Прогнозная сила моделей оценивается для двух временных периодов: первый из них, 01.06.2020 – 01.06.2023, включает шок в миграции, связанный с ограничениями в период пандемии COVID-19. Второй период, 01.06.2021 – 01.06.2023, напротив, характеризуется снятием большинства ограничений, связанных с COVID-19. Кроме этого, в период с 2022 г. также возможны шоки на фоне проведения СВО в России.

Статья имеет следующую структуру. Второй раздел посвящен обзору литературы по применению GTI для прогноза различных экономических показателей и миграции, а также моделям, в которых использование GTI улучшает предсказательную силу. В третьем разделе описываются источники данных, а также алгоритм сбора поисковых запросов, используемых в работе. Проводится сравнение различных моделей прогноза миграции, описываются выявленные закономерности. В Заключении приводятся преимущества и недостатки исследуемых моделей, а также их ограничения, описываются направления для будущих исследований в данной области.

Обзор литературы

С развитием информационных технологий и появлением в сети Интернет «больших данных» о действиях людей появилось новое направление исследований, связанных с оценкой экономических показателей с использованием цифрового следа экономических агентов [Cesare et al., 2018]. Применение таких данных позволяет не только быстрее оценивать различные показатели, но также и прогнозировать поведение целевых метрик на несколько периодов вперед [Wanner, 2021]. На примере оценки миграции такие методы становятся актуальными ввиду временно́й задержки между фактом миграции и появлением статистических отчетов [Tjaden, 2021]. Развиваются исследования, использующие новые типы экзогенных данных, способствующие ускорению получения оценок о миграции (наукастингу). Эти данные можно разбить на следующие группы:

- GPS – координаты мобильных устройств [Williams et al., 2015; Bengtsson et al., 2011];
- Данные социальных сетей [Kim et al., 2020; Martín et al., 2020];
- IP-адреса устройств [Zagheni, Weber, 2012];
- Статистика поисковых запросов [Jun et al., 2018];
- Прочие источники (данные об авиаперелетах, новости и др.) [Gabrielli et al., 2019].

Часть этих подходов подвергается критике в работе [Чудиновских, 2018] из-за невозможности разделения получаемых оценок на временную и постоянную миграцию. Другим источником данных о миграции могут служить различные опросы и базы данных, как правило, собираемые частными агентствами, однако к недостаткам таких данных можно отнести непрозрачность в методологии, а также невозможность использовать эти данные для краткосрочных прогнозов.

Широкое использование в научной литературе получил сервис Google Trends Index, открывающий новые возможности для исследователей в различных областях медицины, анализе финансового рынка, а также в прогнозировании различных экономических показателей. Одной из первых работ в этом направлении считается [Ginsberg et al., 2009], в которой прогнозируется распространение гриппа на основе поисковых запросов. За последующие 10 лет было найдено более 650 публикаций [Jun et al., 2018] с использованием Google Trends. Использование GTI в задаче оценки миграции впервые было описано в работе [Choi, Varian, 2012], в которой исследовались туристические потоки из различных стран в Гонконг. В этой же работе отмечается, что включение поисковых запросов с лагом улучшает прогнозные качества модели, что также

можно связать с тем, что потенциальные мигранты ищут информацию о стране переезда еще до факта миграции [Fantazzini et al., 2021; Böhme et al., 2020].

Всего было найдено 10 работ, в которых авторы исследовали миграцию с использованием Google Trends. При их анализе можно выделить три основных этапа, необходимых для использования поисковых запросов в моделях: во-первых, важно сформировать множество запросов, по которым будут собираться данные; во-вторых, отобрать запросы, на основе которых будет производиться моделирование; и в-третьих, выбрать модели прогнозирования миграции.

Важную роль при работе с GTI играют способы выбора поисковых запросов, так как от выбранных тематик может зависеть предсказательная сила модели. В основном в качестве поисковых запросов для Google Trends авторы используют такие запросы, как посольство, виза, работа, жилье (на языке мигранта), а также добавляют регион назначения. В работах [Böhme et al., 2020; Golenvaux et al., 2020; Wanner, 2021; Fantazzini et al., 2021; Jurić, 2022; Цапенко, Юревич, 2022] авторы строят модели на основе отдельных поисковых запросов, таких как «переезд в «регион назначения», «работа в «регионе назначения», «зарплата в «регионе назначения» и т.д.; такие запросы, как правило, определяются экспертно. В работе [Avramescu, Wiśniowski, 2021] предложен подход по частичной автоматизации сбора данных, авторы используют WordNet корпус [Fellbaum, 2005] для поиска синонимов ключевых запросов; таким образом, делается шаг в сторону более алгоритмического подхода к формированию множества запросов. В работе [Броницкий, Вакуленко, 2022] авторы используют методы машинного обучения (NLP), при помощи которых удалось сформировать несколько различных множеств поисковых запросов и сравнить их эффективность.

На втором этапе анализа миграции с использованием GTI отбираются поисковые запросы из всего множества запросов, которые затем включаются в модель (используются как факторы моделей). Среди рассматриваемых работ можно встретить следующие основные способы такого отбора: экспертный – авторы самостоятельно отбирают поисковые слова (или не указывают способ их отбора) либо берут все множество поисковых слов, отобранных на предыдущем шаге [Fantazzini et al., 2021; Цапенко, Юревич, 2022]. Наиболее часто можно встретить подход с отбором поисковых запросов по корреляции с объясняемой переменной; так, в работах [Wladyka, 2017; Avramescu, Wiśniowski, 2021] авторы оставляли запросы, соответствующие наибольшей корреляции с миграцией. Кроме этого, в работах [Avramescu, Wiśniowski, 2021; Броницкий, Вакуленко, 2024] предлагается подход по кластеризации исходных запросов (например, метод главных компонент); в этом случае возможно использовать все поисковые слова, получая при этом по одной переменной для каждого из кластеров.

Отдельно рассмотрим основные модели оценки миграции, которые используют авторы при работе с GTI. Как правило, авторы используют базовые ARIMA-модели без информации о GTI в качестве base-line моделей; кроме этого, к базовым моделям можно отнести гравитационную модель оценки миграции [Böhme et al., 2020]. Одной из самых популярных моделей с использованием данных GTI является модель SARIMAX, где в качестве экзогенной переменной X выступают данные о поисковых запросах Google Trends Index [Fantazzini et al., 2021; Avramescu, Wiśniowski, 2021; Цапенко, Юревич, 2022]. Другими часто используемыми в работах моделями являются модели парных и линейных регрессий [Wladyka, 2017; Wanner, 2021; Jurić, 2022; Броницкий, Вакуленко, 2024]. В работе [Wladyka, 2017] авторы оценивают модель парной регрессии в разностях, ввиду того, что исходные переменные обладают годовой частотностью. При оценке миграции с использованием парной регрессии в качестве объясняющих переменных могут выступать как сами поисковые запросы, так и их лаги. Кроме этого, в работах [Böhme et al., 2020; Golenvaux et al., 2020; Fantazzini et al., 2021] при моделировании миграции авторы также включают в модель лаги поисковых запросов. В основном авторы используются GTI с лагом в один период (месяц или год в зависимости от частотности данных) [Böhme et al., 2020; Golenvaux et al., 2020; Fantazzini et al., 2021]. В работе [Броницкий, Вакуленко, 2024] авторы делают шаг в сторону применения «мультизапросной» модели с ла-

гами на примере миграции из России в Германию. Как правило, выбираются лаги в пределах одного года, необходимость исследования которых обусловлена тем, что до фактического переезда потенциальным мигрантам необходимо выполнить ряд действий (выбор жилья, получение визы и т.п.) [Benson-Rea, Rawlinson, 2003], которые, как правило, укладываются в срок до 1 года.

Таким образом, на данный момент в литературе по прогнозированию миграции с помощью Google Trends не найдено работ, где бы производилось сравнение способов прогнозирования международной миграции на примере различных стран, а также делались выводы относительно эффективности применения моделей с использованием только одного поискового запроса сравнительно с моделями с несколькими поисковыми запросами и их лагами. В работе производится попытка обобщения исследуемых в других работах подходов, а также показывается эффективность одновременного добавления в модели лагов поисковых запросов от 1 до 12 месяцев сравнительно с моделями только с каждым из лагов по отдельности.

Методология и данные

Для достижения обозначенной выше цели необходимо решить следующие задачи: во-первых, сравнить и усовершенствовать алгоритмы прогнозирования международной миграции, позволяющие получать оценки количества мигрантов еще до их публикации в официальных статистических службах; во-вторых, произвести оценку качества исследуемых моделей как для периода вне шоков, так и для периодов, содержащих различного рода внешние возмущения. В работе исследуется миграция сразу для нескольких стран, и для возможного сравнения результатов необходимо выбирать одинаковые источники данных для всех приведенных пар стран. Поэтому в работе используется один источник данных о миграции Федерального статистического ведомства Германии (German Federal Statistical Office), а также предлагается единый подход к сбору данных о поисковых запросах, которые будут описаны далее в этой главе.

Были получены и проанализированы данные Федерального статистического ведомства Германии с помесячной частотностью (табл. П1 в Приложении); такие данные не публикуются в открытом доступе, но предоставляются по запросу. На сайте статистического офиса Германии [Database of the Federal…, показатель 12711-0008] имеются годовые данные, задержка в публикации которых составляет примерно 3,5 месяца[2]. В работе исследуется миграция в Германию на примере шести стран (Польша, Италия, Румыния, Испания, Болгария, Россия), из которых фиксировались одни из наибольших значений показателя «прибытия иностранцев» (рис. 1) за период с 01.01.2006 по 01.06.2023. Такой подход называется «зеркальной» статистикой [UN, 1998]; он основан на том, что данные о количестве мигрантов проще агрегировать в стране приезда, чем в стране выезда. Так, международное сообщество часто настаивало на том, чтобы в обмене данными использовались иммиграционные данные по стране назначения, чтобы помочь странам происхождения посчитать свои исходящие миграционные потоки (например, Специальная группа ЕЭК ООН 2009 по «Оценке эмиграции с использованием данных, собранных в принимающей стране», [ЕЭК ООН, 2014]). Показатель собирается на основе данных о первичной регистрации, которая необходима при переезде на постоянное место жительства в Германию, дальнейшие перемещения внутри страны в статистике не учитываются. К ошибкам в методологии подсчета можно отнести тот факт, что люди не снимаются с регистрации при убытии, однако этот факт не влияет на качество оценок, так как для оценок используются только данные о прибытиях.

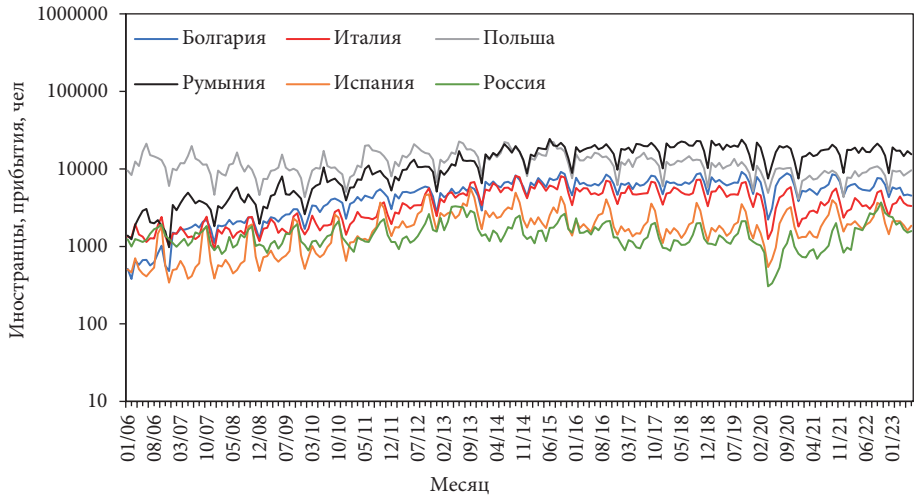

Рисунок 1. Количество мигрантов (показатель «прибытия иностранцев») в Германию из различных стран по данным статистического офиса Германии 2006–2022 гг., чел.
Источник: данные Федерального статистического ведомства Германии.

Для оценки и прогнозирования миграции в качестве объясняющей переменной используются данные о поисковых запросах Google Trends Index; такие данные общедоступны и публикуются в реальном времени. Индекс GTI отражает динамику интенсивности запросов $S_{d,r}$ пользователей по определенным ключевым словам во времени (d) в рамках выбранного региона (r). Однако индекс $S_{d,r}$ характеризует не абсолютное количество запросов по выбранному поисковому запросу $V_{d,r}$, а его отношение ко всем поисковым запросам в данном регионе в данный день $T_{d,r}$. Таким образом, индекс $S_{d,r} = \frac{V_{d,r}}{T_{d,r}}$ показывает долю запросов, связанных с определенным поисковым запросом, относительно общего объема запросов в выбранном географическом регионе в определенный момент времени. Индекс популярности поискового запроса собирается для страны, из которой совершают эмиграцию, а в качестве языка используется официальный язык страны. При работе с GTI следует учитывать следующие особенности:

1. Индекс отображает нормированное (и приведенное к 100%) значение в выбранном периоде и тематике, а не фактическое количество запросов по выбранному поисковому запросу. Это затрудняет сравнение различных поисковых запросов вследствие того, что они нормированы на разные максимальные значения. Одним из способов решения проблемы является стандартизация данных Google Trends [Fantazzini et al., 2021] (1), которая позволяет проводить сравнение индексов по разным поисковым запросам:

$$Z = \frac{X - \overline{X}}{\hat{\sigma}_X}, \quad (1)$$

где $\overline{X}, \hat{\sigma}_X$ – среднее значение случайной величины X и ее стандартное отклонение соответственно.

2. Google вносил изменения в алгоритм сбора данных (такие изменения проводились 01.01.2011, 01.01.2016 и 01.01.2022), что затрудняет использование временных рядов за весь доступный период (01.01.2004 – по н.вр.). Самым существенным стало изменение 1 января 2011,

после которого поменялся алгоритм определения региона запросов, что не позволяет объединять при анализе периоды до и после описанного изменения. Поэтому в этой и других работах при работе с GTI рекомендуется использовать период с 01.01.2011.

Существуют также альтернативные источники истории поисковых запросов пользователей; например, такую статистику представляет Yandex Wordstat. В отличие от индекса GTI, такая статистика представляется в абсолютных значениях, что упрощает ее использование для дальнейших прогнозов. Однако существенным недостатком является ограниченность выборки данных в помесячной частотности (на момент написания работы доступны помесячные данные начиная с 01.01.2018). Кроме этого, существует географическое смещение в использовании Яндекса – за пределами стран СНГ Яндекс имеет низкую долю поисковых запросов, что затрудняет использование предложенных алгоритмов для анализа международной миграции. Однако при накоплении достаточной для анализа истории запросов в базе Yandex Wordstat такие данные могут стать альтернативой данным GTI при оценке миграции из стран СНГ.

Другой задачей становится определение множества поисковых запросов, которые в дальнейшем будут использоваться в качестве объясняющих факторов при моделировании миграции. В работе [Броницкий, Вакуленко, 2024] при оценке миграции из России в Германию используется «мультизапросная» модель, содержащая поисковые запросы в следующих тематиках: «работа в Германии», «учеба в Германии», «посольство Германии» и их лаги. В работе используются лингвистические, а также инструменты машинного обучения (ML-инструменты) для поиска множества поисковых запросов, которые учитывают особенности поиска в России, а также необходимость для россиян посещать посольство перед миграцией в Германию. О необходимости лингвистических исследований также говорится в работе [Tjaden, 2021]. Так, например, жители Сирии при поиске информации о миграции на английском языке будут использовать не те же самые термины, что жители Канады, также ищущие информацию о миграции на английском языке. В настоящей работе с целью упрощения исследования был выбран только один поисковый запрос «работа в Германии» на языке страны выбытия. Возможно, необходимо также использовать тематику «посольство Германии», но все рассматриваемые в работе страны (кроме России) являются странами – членами ЕС, для их граждан нет необходимости в посещении посольства при переезде в Германию.

Ниже представлен алгоритм сбора и обработки данных на примере Польши (аналогичный алгоритм используется для остальных стан):

1. Необходимо определить официальный язык в стране эмигранта (польский).
2. Также важно перевести запрос «работа в Германии» на польский язык, в результате формируется поисковый запрос на официальном языке мигранта «Praca w Niemczech».
3. С использованием сервиса Google Trends возможно составить запрос для получения тренда поисковых запросов «Praca w Niemczech» с 01.01.2011 по 01.11.2023 в Польше. Также можно форматировать url-ссылку для запроса, подставляя в нее необходимые параметры. Такая ссылка имеет следующий формат: https://trends.google.com/trends/explore?date=2011-01-01%202023-11-01&geo=PL&q=Praca%20w%20Niemczech, где date – промежуток дат для получения тренда, geo – регион, в котором происходил поиск, q (query) – поисковый запрос, для которого необходимо получить тренд; вместо пробела используется знак «%20».
4. Кроме этого, возможно получить похожие популярные запросы по тематике; для этого во вкладке «похожие запросы» необходимо выбрать «лидеры», где отобразятся другие запросы, которые использовали мигранты до или после поиска «Praca w Niemczech». Среди таких запросов отбираются те, что связаны с выбранной темой «работа в Германии». Такой подход позволяет решить сразу несколько проблем: во-первых, в случае нескольких официальных языков в стране, представленным алгоритмом можно выявить

все используемые транслитерации искомого поискового запроса; во-вторых, задействуется лингвистическая составляющая, включаются в том числе сленговые запросы, которыми пользуются потенциальные мигранты.

5. Для дальнейшего использования и сравнения друг с другом индексов по разным запросам применяется процедура стандартизации (1), в результате которой все ряды имеют одинаковый масштаб: нулевое среднее значение и единичную дисперсию.

При помощи описанного выше алгоритма для поискового запроса «работа в Германии» были собраны данные с 01.01.2011 по 01.11.2023 г. для исследуемых стран (табл. П2), всего от 1 до 6 поисковых запросов в зависимости от выбранной страны. Однако при оценке моделей нет возможности рассматривать все запросы одновременно вследствие роста размерности: для каждого из запросов в модель включается 12 лагов. Из-за ограниченного количества наблюдений (155 наблюдений с 01.01.2011 г.) и необходимости разбить данные на тестовую и контрольную группы количество объясняющих переменных может получиться больше, чем количество наблюдений. Анализ автокорреляционной функции (ACF) и частично автокорреляционной функции (PACF) для данных о миграции, а также стандартизированных GTI-индексов показал наличие годовой сезонности в данных. Наличие нестационарности (на основе теста Дикки – Фуллера) затруднит оценку модели распределенных лагов, поэтому при использовании этой модели для зависимой и объясняющих переменных одновременно выполняется переход к сезонным разностям с периодом 12 месяцев**.** Для уменьшения размерности выбирается по одному запросу для каждой страны, на основе наибольшей корреляции поисковых запросов в сезонной разности с зависимой переменной, взятой также в сезонной разности. Выбранные таким образом поисковые запросы выделены курсивом в таблице П2, однако можно заметить, что для некоторых из них корреляция менее 20%, а в случае Испании и России – имеет отрицательные значения. Это может быть связано в том числе с тем, что миграция происходит с некоторой задержкой во времени от момента поиска информации в сети Интернет до фактической миграции в другую страну. Для проверки этой гипотезы для отобранных поисковых запросов (в сезонных разностях) была составлена корреляция их лагов от 1 до 12 месяцев с данными о миграции, также в сезонной разности, которые представлены в таблице П3. Исходя из нее можно сделать выводы о том, что на этапе отбора поисковых запросов важно смотреть не только на корреляцию текущих значений, но также и их лагов. Так, на примере Испании видна корреляция в 25-26% для 6, 9, 10 лагов.

Оценка качества моделей производится на основе разбиения исходных данных на «тестовую» и «контрольную» выборки, что является важным методологическим компонентом, позволяющим получить вневыборочные оценки качества моделей (в работе производятся оценки средней абсолютной процентной ошибки прогноза – MAPE, а также средней абсолютной ошибки прогноза – MAE). Для всех приведенных оценок тестовая и контрольные группы не пересекаются. Оценка моделей, а также прогнозы производятся с использованием переменных в сезонных разностях. Далее для сравнения оценок качества моделей с целью лучшей интерпретируемости результатов производится переход к исходным переменным; таким образом, оцениваются именно ошибки прогноза миграции, а не их сезонных разностей. В работе используются три пары тестовых и контрольных групп, соответствующие второму и третьему прогнозным годам. Первая пара групп предназначена для оценки качества за 2-летний прогнозный период 01.06.2021 – 01.06.2023 (соответствующая тестовая выборка 01.01.2011 – 01.06.2021). Важно отметить, что в данном периоде большинство ограничений, связанных с эпидемией COVID-19, было снято, что дает возможность сделать выводы относительно поведения моделей в условиях, свободных от значительных внешних шоков. Вторая пара временных периодов используется для исследования 3-летнего прогноза 01.06.2020 – 01.06.2023 (соответствующая тестовая выборка 01.01.2011 – 01.06.2020). На этот период приходится ряд ограничений, связанных с эпидемией COVID-19: из-за сложностей

в перемещениях для всех стран наблюдается снижение миграционной активности в начале этого периода. Исследование указанного периода дает возможность оценить, насколько хорошо модели могут справляться с шоками, вызванными эпидемией. Данные выводы обобщаются на случай учета и других видов внешних воздействий, таких как военные действия, природные катаклизмы и пр. Третья пара групп является 2-летней подвыборкой для 3-летней группы 01.06.2020 – 01.06.2022, которая исследуется с целью проверки гипотезы относительно изменения качества прогноза из-за внешних шоков, а не увеличения величины прогнозного периода.

Модели прогнозирования миграции

В данном разделе описываются модели, используемые для оценки миграции из различных стран в Германию. Приводится алгоритм оценки моделей, а также выбора наилучших параметров модели с использованием информационного критерия (AIC). Производится сравнение предсказательной силы моделей для 2-летнего и 3-летнего периодов прогноза, на основе которого делаются выводы относительно целесообразности использования той или иной модели при наличии различных внешних возмущений. Кроме этого, на примере миграции из России в Германию сравниваются «однозапросные» и «мультизапросные» модели, а также проверяется гипотеза о необходимости включения в модель сразу нескольких лагов.

Одной из исследуемых моделей является модель распределенных лагов, в которой в качестве зависимой переменной используется показатель «прибытие иностранцев» из рассматриваемой страны в Германию, а в качестве объясняющих переменных выступают лаги (от 1 до 12 месяцев) временного ряда по запросу «работа в Германии». Это необходимо из-за того, что существует некоторый лаг между поиском информации в Интернете и фактической миграцией. Для разных стран величина лага может варьировать в зависимости от различных факторов, таких как миграционная политика или сложность в логистике. Модель распределенных лагов $l = 1…12$ месяцев для переменных в сезонных разностях можно выразить следующим образом:

$$Y_t - Y_{t-12} = \beta_0 + \sum_{l=0}^{12} \beta_l \left(X_{t-l} - X_{t-12-l} \right) + \varepsilon_t, \qquad (2)$$

где

- Y_t – зависимая переменная, показатель «прибытие иностранцев» в Германию
- X_{t-1} – объясняющие переменные (поисковый запрос GTI с лагами $l = 1…12$)
- ε_t – ошибки регрессии, $\varepsilon_t \sim iid(0, \sigma^2)$
- $t = 1…T$ – рассматриваемый год

Для определения необходимого числа лагов в модели распределенных лагов разработан алгоритм, который перебирает все возможные модели, включающие лаги от 1 до 12 месяцев (всего оценивается 8192 модели для каждой из стран). Для выбора оптимальной модели используется информационный критерий AIC, при помощи которого выбирается наилучшая модель. Одним из условий, необходимых для сравнения моделей с применением AIC-критерия, является использование одинакового числа наблюдений при оценке параметров моделей. Для SARIMAX-моделей также используется AIC-критерий для определения параметров модели p, d, q, где p – порядок авторегрессии, d – порядок интеграции, q – порядок скользящего среднего, параметр s = 12 выбирается на основе ACF. Данные параметры также выбираются при помощи информационного критерия AIC путем перебора параметров на данных «тестовой» группы. Для SARIMAX в качестве экзогенных переменных используются сам поисковый запрос и его лаги, подобранные для модели распределенных лагов. Повторный поиск наилуч-

ших лагов не производится из-за возрастания количества параметров модели для перебора и ограниченности в вычислительных ресурсах.

В работе сравниваются различные модели прогнозирования миграции, в том числе проводится сравнение с SARIMA-моделью без использования каких-либо данных о поисковых запросах GTI. Тестируются следующие модели:

- SARIMA – сезонная авторегрессионная модель скользящего среднего с сезонной компонентой (на основе анализа ACF данные о миграции имеют сезонность с периодом в 12 месяцев). Модель применяется для прогнозирования миграции без использования экзогенных данных;
- SARIMAX – разновидность SARIMA-моделей, в которой в числе объясняющих переменных используются GTI-индекс и его лаги от 1 до 12;
- Модель распределенных лагов – множественная регрессия, где в качестве объясняемой переменной выступает количество мигрантов, а в качестве объясняющих переменных – GTI-индексы и их лаги от 1 до 12 месяцев.

Кроме моделей, построенных на основе одного запроса «работа в Германии», к сравнению добавлена «мультизапросная» модель, содержащая в качестве объясняющих переменных сразу несколько поисковых запросов из различных тематик (запросы и тематики представлены в таблице П4). Однако из-за ограниченности количества наблюдений в тестовой выборке возможно снизить размерность данных при помощи метода главных компонент (PCA) для разных поисковых тематик («учеба», «работа», «посольство») [Броницкий, Вакуленко, 2024]. В мультизапросной модели используются сразу три PCA-вектора, соответствующих каждой из представленных выше тематик, а также их лаги от 1 до 12 месяцев (итого 36 переменных). Особенностью данного подхода является более трудоемкая работа со множеством поисковых запросов, для определения которого в том числе используются NLP-подходы машинного обучения, а также проводится анализ лингвистических особенностей при поиске информации в интернете на русском языке. В данной же работе применяется упрощенный подход в сборе данных для различных стран. Однако на примере миграции из России в Германию появляется возможность сравнить прогнозные силы этих двух подходов. Модель распределенных лагов (2) для «мультизапросной» модели включает в себя уже несколько поисковых запросов (PCA-сверток) и их лагов и выглядит следующим образом:

$$Y_t - Y_{t-12} = \beta_0 + \sum_k \sum_{l=0}^{12} \beta_{k,l} \left(X_{k,t-l} - X_{k,t-12-l} \right) + \varepsilon_t, \quad (3)$$

где

- Y_t – зависимая переменная, показатель «прибытие иностранцев» в Германию
- $X_1 \ldots X_k$ – объясняющие переменные (PCA-свертки поисковых запросы GTI с лагами)

Используя описанную выше методологию, было оценено три типа моделей (модель распределенных лагов, SARIMAX и SARIMA) для шести стран. Также в качестве базовой модели использовались предсказания средним, однако такие оценки для большинства рассматриваемых стран оказались хуже SARIMA-моделей, поэтому они не приводятся в сравнении. Для каждого из 2-летних и 3-летних прогнозных периодов по отдельности оценивались параметры моделей, а также рассчитывались метрики прогнозного качества моделей (MAPE, MAE), приведенные в таблице 2. На рисунках 2–7 приведены результаты прогноза моделей распределенных лагов для 2-летнего и 3-летнего периодов. Кроме этого, представлены оценки параметров модели распределенных лагов и SARIMAX для 2-летнего контрольного периода (таблицы 1, П5). Также представлены метрики качества моделей для 2-летних периодов (с шоками и без) с целью отвержения гипотезы о различиях в качестве из-за увеличенного прогнозного окна. В работе не представлены оценки моделей для 3-летнего периода ввиду того, что они практически не отличаются от оценок моделей для 2-летнего периода.

Таблица 1. Оценки моделей распределенных лагов (2) для 2-летнего контрольного периода

	Лаг, месяц	Польша	Италия	Румыния	Испания	Болгария	Россия
Константа		-521.53*** (158.1)		967.73*** (301.78)	-35.11 (46.65)	413.33*** (103.36)	-105.37 (68.41)
GTI «Работа в Германии»	0	863.55*** (285.8)		1 489.25*** (471.77)		485.81*** (147.45)	
	1						
	2	502.82* (307.42)				592.14*** (148.07)	
	3						
	4				264.80** (119.62)		
	5	426.61 (293.97)					
	6						
	7					184.40 (140.45)	
	8		306.58*** (156.9)		163.31 (108.46)		-180.54** (86.14)
	9				193.46* (104.02)		
	10	603.75** (237.36)	375.62*** (153.2)		186.55* (98.47)		
	11						
	12			957.20** (447.19)			
Число наблюдений		102	102	102	102	102	102
AIC		1 791	1 661	1 922	1 517	1 687	1 619
BIC		1 804	1 669	1 930	1 530	1 697	1 624
R^3		0.29	0.39	0.10	0.14	0.26	0.04
F-статистика[4]		9.92***	31.9***	5.67***	4.09***	11.3***	4.39**
Средний лаг (95% доверительный интервал в скобках)		6.2 [2.9;9.4]	8.8 [8.1;9.8]	12.04	7.1 [4.9;9.1]	2.65 [2.01;4.4]	8.0

Источник: оценки автора.

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

3 Проверка гипотезы об адекватности регрессии. Основная гипотеза формулируется следующим образом: выбранный набор независимых переменных не оказывает линейного влияния на объясняемую переменную.

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

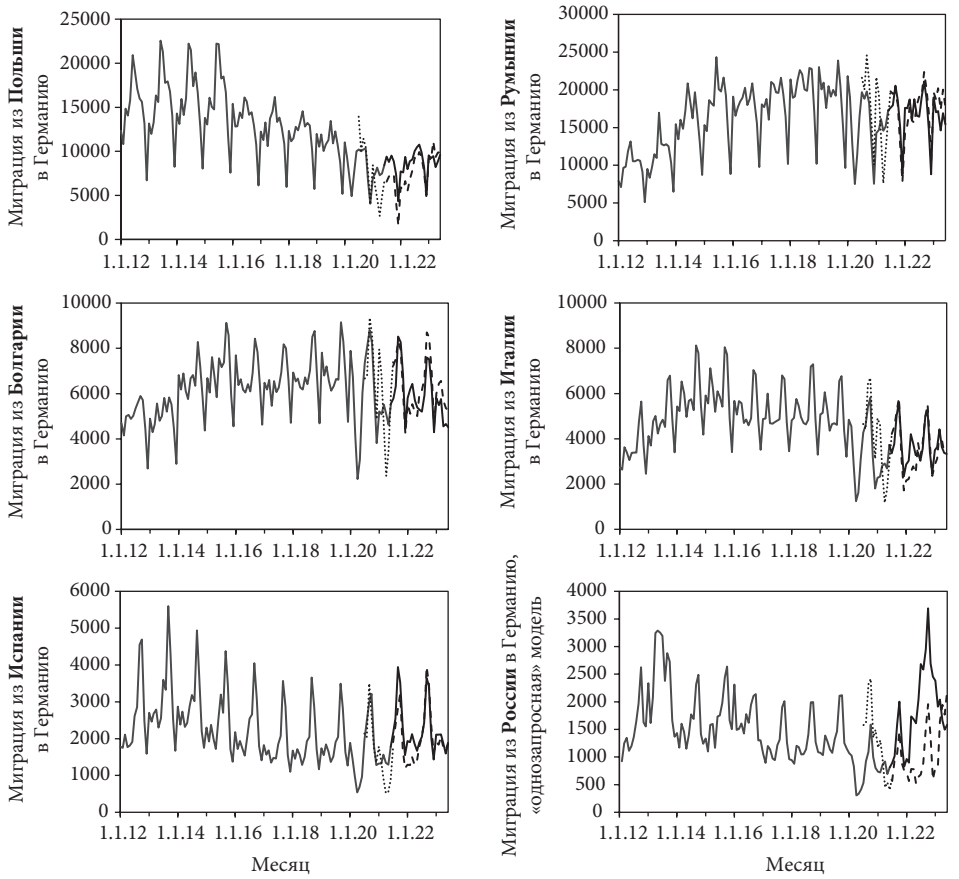

Рисунок 2-7. Оценка миграции из различных стран в Германию с использованием модели распределенных лагов для 2- и 3-летних прогнозных периодов.

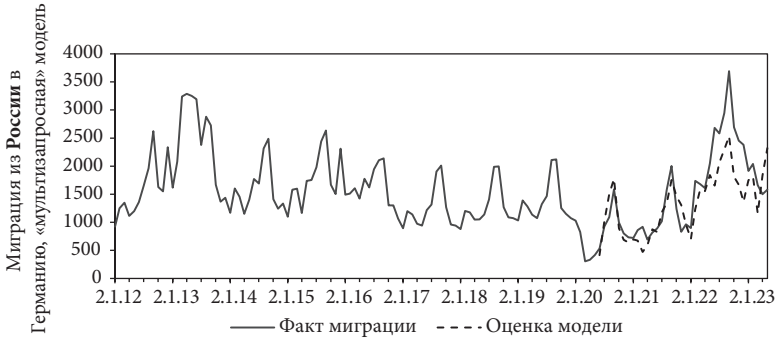

Рисунок 8. Оценка миграции из России в Германию с использованием «мультизапросной» модели распределенных лагов для 3-летнего прогнозного периода. *Источник:* расчеты автора на основе данных Федерального статистического ведомства Германии.

Таблица 2. Критерии качества прогноза миграции из различных стран в Германию для 2-летнего и 3-летнего прогнозных периодов

Страна	Метрики качества моделей MAPE (MAE) на тестовой выборке					
	Модель распределенных лагов		SARIMAX		SARIMA	
	2 года	3 года	2 года	3 года	2 года	3 года
Польша	0.18 (1 466.5)	**0.22 (1 817.7)**	**0.08 (638.3)**	0.67 (5 376.1)	0.29 (2 386.2)	0.48 (3 866.6)
Италия	**0.13 (479.7)**	**0.23 (708.6)**	0.17 (616.6)	0.48 (1 700.5)	0.20 (688.3)	0.65 (2 260.6)
Румыния	**0.11 (1 782.6)**	**0.15 (2 427.3)**	0.11 (1 844.9)	0.52 (8 255.1)	0.14 (2 320.7)	0.47 (7 454.8)
Испания	0.12 (304.8)	**0.19 (376.7)**	**0.08 (178.3)**	0.65 (1 304.9)	0.09 (184.4)	0.70 (1 405.7)
Болгария	0.13 (723.5)	**0.16 (946.9)**	**0.09 (518.1)**	0.18 (1 111.6)	0.10 (574.2)	0.16 (952.7)
Россия, однозапросная модель	**0.38 (826.5)**	**0.42 (706.7)**	0.45 (953.4)	0.42 (909.7)	0.79 (1 258.7)	0.79 (1 300.2)
Россия, мультизапросная модель	**0.22 (436.9)**	**0.21 (355.1)**	-	-		

Источник: расчеты автора.

Примечание. В скобках указано значение MAE, рассчитанное для вневыборочных данных. Полужирным выделены наилучшие результаты среди рассматриваемых моделей для 2-летнего и 3-летнего периодов по отдельности.

Анализируя результаты, можно заметить, что добавление GTI при оценке миграции приводит к повышению прогнозной силы моделей (таблица 2). Так, для четырех из шести стран модель распределенных лагов и модель SARIMAX для всех стран позволяют получить меньшую ошибку на вневыборочных данных сравнительно с результатами SARIMA-моделей при прогнозе на 2-летний период. Полученные результаты указывают на эффективность использования GTI в качестве экзогенной переменной при прогнозировании миграции.

При сравнении 2-летнего и 3-летнего прогнозов SARIMA-моделей было обнаружено, что для всех стран прогноз существенно ухудшается при включении в модель прогнозного периода, связанного с пандемией COVID-19 (таблица 2). Это может быть объяснено зависимостью от тренда временных рядов и отсутствием экзогенных данных, способных учесть внешние шоки, вызванные пандемией. Однако такое расхождение в качестве может быть вызвано не только шоками, но также и ухудшением качества предсказания из-за увеличения прогнозного периода от двух лет до трех. С целью проверки этой гипотезы было также оценено качество моделей для сопоставимых по времени 2-летних периодов (01.06.2021 – 01.06.2023, а также 01.06.2020 – 01.06.2022), с наличием внешних шоков и без них (таблица П6). В результате сравнения этих периодов также наблюдается расхождение в периоды с шоками и без них, при

сравнении этих периодов представленные выше выводы (в таблице 2) не меняются. Полученные результаты указывают на ухудшение предсказательной силы SARIMA-моделей в период шоков, а также на сравнительно лучшее качество моделей с экзогенными переменными в такие периоды.

При анализе 3-летнего периода наблюдается, что модель распределенных лагов во всех случаях показывает лучшие результаты по сравнению с SARIMAX- и SARIMA-моделями. Это подтверждает важность учета внешних факторов при оценке и прогнозировании миграции, особенно в периоды шоков. Однако для моделей распределенных лагов также наблюдается ухудшение прогнозной силы моделей при добавлении в рассматриваемый период временного шока, связанного с пандемией COVID-19.

Также для модели с распределенными лагами (2) можно измерить вклад лагов при помощи оценки среднего лага L_k для каждой страны k по отдельности:

$$L_k = \sum_{l=0}^{12} l\beta_{k,l}^2 \Big/ \sum_{l=0}^{12} \beta_{k,l}^2. \quad (4)$$

Большая величина среднего лага L означает, что факт миграции происходит с большой задержкой во времени от момента поиска. Кроме этого, так как средний лаг является случайной величиной, в работе при помощи метода Монте-Карло оцениваются величина доверительного интервала для среднего лага. В работе [Wanner, 2021] оценивается 12 отдельных моделей для каждой из стран с использованием только одного лага (модель только с первым лагом, со вторым и т.д.). В настоящей работе было произведено сравнение таких моделей с моделями, в которые были включены сразу несколько лагов одновременно (модель распределенных лагов). На основе AIC-критерия, а также метрик качества прогноза было установлено, что для пяти из шести стран модели распределенных лагов имеют лучшую предсказательную силу. Полученные оценки величины среднего лага для пяти стран из шести также показывают, что величина среднего лага превышает 6 месяцев, что еще раз подтверждает гипотезу о том, что при анализе миграции необходимо учитывать не только текущее значение GTI. При этом на примере Болгарии величина среднего лага значительно отличается от других стран (2.65 месяца); это может быть связано как со спецификой миграционной политики страны, так и с особенностями трудоспособного населения, готового в более короткие сроки сменить страну работы.

Также можно отметить, что для миграции из России в Германию коэффициенты модели отрицательные (таблица 1). Это может объясняться следующим: во-первых, миграция из России в Германию имеет тренд на снижение числа мигрантов (кроме шоков после начала специальной военной операции (СВО)), поисковые запросы при этом уменьшаются медленнее; во-вторых, видна низкая корреляция поисковых запросов (и всех их лагов) с данными о миграции (таблица П3), что может говорить о неоптимальном поисковом запросе и необходимости также включать смежные тематики. Так, при анализе «однозапросной» и «мультизапросной» моделей (рис. 7, 8) на примере России видно значительное превосходство по прогнозной силе последней сравнительно с моделью, построенной с использованием одного запроса «работа в Германии» и его лагов. Кроме этого, можно заметить, что «мультизапросная» модель показывает сравнительно лучшее качество при работе с шоками. Это единственная модель, где прогнозное качество улучшается при включении в анализ периода пандемии. Также стоит отметить, что модель прогнозирует увеличение миграции с начала проведения СВО, однако реальная миграция оказывается еще больше. Одна из гипотез возникновения такого расхождения заключается в том, что в моменты шоков уменьшается величина среднего лага (3) (т.е. потенциальные мигранты быстрее совершают миграцию от момента поиска информации до переезда) по сравнению с относительно стабильными периодами без внешних шоков.

Заключение

В работе исследуются модели прогнозирования международной миграции населения, позволяющие получать оценки количества мигрантов без задержки во времени, сравнительно с официальными статистическими источниками. Производится сравнение прогнозной силы как для моделей с использованием данных Google Trend Index на примере запроса «работа в Германии», так и для моделей без них. Были получены оценки прогнозного качества моделей на примере миграции из шести стран в Германию. В качестве прогнозных периодов были выбраны 2-летний и 3-летний периоды, второй из которых характеризуется наличием шоков, связанных с ограничениями из-за эпидемии COVID-19. Кроме этого, производится сравнение эффективности «однозапросных» моделей с «мультизапросными» на примере миграции из России в Германию [Броницкий, Вакуленко, 2024]. В отличие от работы [Wanner, 2021], в данной работе исследуются модели с одновременно включенными лагами GTI от 1 до 12 месяцев, а также приводится алгоритм выбора поискового запроса на языке страны выбытия мигранта.

Полученные результаты демонстрируют увеличение предсказательной силы моделей миграции при использовании GTI и их лагов в качестве объясняющих переменных. В результате сравнения прогнозных качеств моделей делается вывод, что в относительно спокойные периоды, без внешних шоков, SARIMAX-модели и модель распределенных лагов показывают схожие результаты и оказываются лучше SARIMA-моделей для всех рассматриваемых в работе стран. Однако в случае прогноза для периодов с внешними воздействиями модель распределенных лагов показывает лучшую предсказательную силу по сравнению с другими моделями. На примере оценки миграции из России показано, что использование «мультизапросных» моделей вдвое повышает точность прогноза для 3-летнего прогнозного периода по сравнению с «однозапросной» моделью. Для пяти стран из шести модель распределенных лагов с одновременно включенными в нее сразу несколькими лагами оказывается эффективнее моделей только с одним отдельным лагом (модели только с первым лагом, только со вторым лагом и т.д.).

В работе описывается алгоритм сбора, подготовки и использования GTI при прогнозировании миграции. Опираясь на полученные результаты, рекомендуется использовать «однозапросную» модель с включенными лагами от 1 до 12 месяцев во всех рассматриваемых случаях, что улучшает предсказательную силу, получаемую при применении SARIMA-моделей. В случае необходимости получения наиболее точной оценки, а также при прогнозе в периоды шоков рекомендуется использовать «мультизапросную» модель, детально изучая особенности миграционной политики исследуемой страны, а также лингвистические особенности при поиске информации в сети Интернет населением разных стран.

Стоит отметить имеющиеся ограничения рассматриваемых методов.

Во-первых, в работе исследуются общие потоки мигрантов, без выделения причин миграции (учеба, работа, запрос на убежище и пр.);

во-вторых, при использовании данных Google Trends Index для оценки миграции возникает смещение в сторону мигрантов, использующих сеть Интернет для поиска информации; кроме этого, есть ряд стран, где рассматриваемая поисковая система не используется (например, Китай);

в-третьих, с целью унификации алгоритма в работе используется только один поисковый запрос «работа в Германии», однако возможно также исследовать второй тип запросов «посольство Германии» для тех стран, в которых необходимо получение визы для миграции;

в-четвертых, при работе с шоками возможно уменьшение величины среднего лага за счет необходимости быстрее принимать решение о миграции; в текущих моделях делается предположение о том, что глубина лагов для разных периодов одинакова.

Направлением для дальнейших исследований могут стать методы работы с шоками при использовании «однозапросных» моделей, что позволит существенно увеличить область при-

ложения моделей и качество оценок. Одной из гипотез для проверки может служить добавление в модель dummy-переменных, соответствующих периодам с шоками. Также возможно использование переменной, отвечающей за скорость роста или снижения тренда в сравнении с предыдущими периодами, т.е. тестирование асимметрии влияния шоков.

Финансирование

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

Благодарности

Автор выражает благодарность своему научному руководителю Е.С. Вакуленко за ценные комментарии и замечания.

Литература

Броницкий Г.Т., Вакуленко Е.С. (2022) Прогнозирование миграции из России в Германию с использованием Google-трендов // Демографическое обозрение: 9(3): 75–92. https://doi.org/10.17323/demreview.v9i3.16471

Броницкий Г.Т., Вакуленко Е.С. (2024) Применение Google Trends для прогнозирования миграции из России: агрегация поисковых запросов и учет лаговой структуры // Прикладная эконометрика: (73): 78–101. https://doi.org/10.22394/1993-7601-2024-73-78-101

Цапенко И.П., Юревич М.А. (2022) Статистика онлайн-запросов в наукастинге миграции // Экономические и социальные перемены: факты, тенденции, прогноз: 15(1): 74–89. https://doi.org/10.15838/esc.2022.1.79.4

Чудиновских О.С. (2018) Большие данные и статистика миграции // Вопросы статистики: 25(2): 48–56. URL: https://voprstat.elpub.ru/jour/article/view/629

Avramescu A., Wiśniowski A. (2021) Now-casting Romanian migration into the United Kingdom by using Google Search engine data // Demographic Research: (45): 1219–54. https://doi.org/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). https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.jdeveco.2019.04.002

Cesare N., Lee H., McCormick T., Spiro E., Zagheni E. (2018) Promises and Pitfalls of Using Digital Traces for Demographic Research. Demography: 55(5): 1979-99. https://dx.doi.org/10.1007/s13524-018-0715-2.

Choi H., Varian H. (2012) Predicting the present with Google Trends // Economic record: (88): 2–9. https://doi.org/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. https://doi.org/10.3390/forecast3040048

Fellbaum C. (2005) WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition. Elsevier, Oxford, 665–70.

Gabrielli L., Deutschmann E., Natale F., Recchi E., Vespe M. (2019) Dissecting global air traffic data to discern different types and trends of transnational human mobility // EPJ Data Science: 8(1): 26. https://doi.org/10.1140/epjds/s13688-019-0204-x

Ginsberg J., Mohebbi M., Patel R., Brammer L., Smolinski M., Brilliant L. (2009) Detecting Influenza Epidemics Using Search Engine Query Data // Nature: (457): 1012–14. https://doi.org/10.1038/nature07634

Golenvaux N., Alvarez P.G., Kiossou H.S., Schaus P. (2020) An LSTM approach to Forecast Migration using Google Trends. https://doi.org/10.48550/arXiv.2005.09902

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. https://doi.org/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. https://doi.org/10.32910/ep.73.2.2

Kim J., Sîrbu A., Giannotti F., Gabrielli L. (2020) Digital Footprints of International Migration on Twitter: 274–86. https://doi.org/10.1007/978-3-030-44584-3_22

Martín Y., Cutter S.L., Li Z., Emrich C.T., Mitchell J.T. (2020) Using geotagged tweets to track population movements to and from Puerto Rico after hurricane Maria // Population and Environment: 42: 4–27. https://doi.org/10.1007/s11111-020-00338-6

Tjaden J. (2021) Measuring migration 2.0: A review of digital data sources // Comparative Migration Studies: 9(1): 59. https://doi.org/10.1186/s40878-021-00273-x

Wanner P. (2021) How well can we estimate immigration trends using Google data? // Quality & Quantity: 55(4): 1181–202. https://doi.org/10.1007/s11135-020-01047-w

Williams N.E., Thomas T.A., Dunbar M., Eagle N., Dobra A. (2015) Measures of human mobility using mobile phone records enhanced with GIS Data // PLoS ONE: 10(7): 1–16. http://doi.org/10.1371/journal.pone.0133630

Wladyka D. (2017) Queries to Google Search as predictors of migration flows from Latin America to Spain. Journal of Population and Social Studies [JPSS]: 25(4): 312–27. https://doi.org/10.25133/JPSSv25n4.002

Zagheni E., Weber I. (2012) You are where you E-mail: Using E-mail data to estimate international migration rates / Proceedings of the 3rd Annual ACM Web Science Conference, WebSci'12. https://doi.org/10.1145/2380718.2380764

Другие источники информации

ЕЭК ООН (2014) Тенденции в двойном гражданстве и их последствия для сбора миграционной статистики. Конференция европейских статистиков. Рабочая сессия по статистике миграции населения. URL: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2014/mtg1/WP_10_UNECE_ru.pdf

Database of the Federal Statistical Office of Germany. URL: https://www-genesis.destatis.de/genesis/online?operation=sprachwechsel&language=en

Google Trends Index. URL: https://trends.google.ru/trends

UN (1998) Recommendations on Statistics of International Migration. Statistical Papers, No. 58, Rev.1. Department of Economic and Social Affairs, United Nations, New York. URL: https://unstats.un.org/unsd/publication/seriesm/seriesm_58rev1e.pdf

Приложение

Таблица П1. Описательная статистика данных о миграции в Германию из других стран, показатель «прибытия иностранцев» за период 2006–2023 гг.[5]

	Страна	Число наблюдений	Min	Max	Std	Mean
Европа	Бельгия	209	-	571	101	283
	Болгария	209	380	9 153	2 213	4 854
	Франция	209	522	2 840	524	1 267
	Греция	209	-	3 959	888	1 872
	Италия	209	904	8 124	1 764	3 680
	Хорватия	209	381	6 911	1 627	2 304
	Мальта	209	-	83	12	12
	Нидерланды	209	562	1 362	149	880
	Австрия	209	542	2 164	239	985
	Польша	209	4 060	22 557	3 971	11 835
	Португалия	209	221	1 410	240	679
	Румыния	209	976	24 341	6 634	12 285
	Швеция	209	-	688	97	254
	Словакия	209	-	1 789	297	933
	Испания	209	-	5 598	1 012	1 805
	Чешская Республика	209	236	1 588	253	786
	Венгрия	209	957	6 349	1 268	3 128
	Босния и Герцеговина	209	-	2 732	695	1 316
	Северная Македония	18	-	2 628	472	119
	Российская Федерация	209	305	3 692	574	1 461
	Швейцария	209	190	945	149	468
	Сербия	209	-	5 342	1 022	1 837
	Турция	209	1 063	9 751	1 457	2 983
	Великобритания	209	-	2 094	390	975
Азия	Ливан	209	-	2 087	246	340
	Пакистан	209	97	4 400	552	531
	Филиппины	209	57	839	146	212
	Сирия	74	-	9 960	1 879	838
	Вьетнам	209	-	1 131	175	419

5 Выбраны страны, для которых было зарегистрировано более 200 миграций в Германию за месяц в период с 01.01.2023 по 01.06.2023.

	Страна	Число наблюдений	Min	Max	Std	Mean
Африка	Египет	209	97	1 445	255	379
	Алжир	209	-	1 757	229	243
	Марокко	209	161	2 182	258	463
	Нигерия	209	-	1 317	276	340
	Сомали	209	-	1 178	254	207
	Тунис	209	94	929	182	335
Америка	Бразилия	209	286	1 896	313	701
	Колумбия	209	54	651	136	229
	Венесуэла	209	-	399	72	84
	США	209	-	3 808	778	1 583

Источник: данные Федерального статистического ведомства Германии (2006–2023 гг.).

Таблица П2. Описательная статистика стандартизированных GTI-индексов, используемых для оценок миграции в Германию за период 2013–2023 гг.

Страна	Поисковый запрос	min	max	Корреляция с миграцией в сезонных разностях
Польша	Praca w Niemczech	-1.10	4.60	0.38
	Niemcy praca	-2.01	3.57	0.47
Италия	Lavorare in Germania	-1.14	3.62	0.38
	Lavoro in Germania	-1.14	3.57	0.39
	Lavoro Germania	-1.13	3.65	0.38
Румыния	Munca in Germania	-1.36	3.45	0.01
	Locuri munca Germania	-1.70	4.11	0.21
	Locuri de munca in Germania	-1.26	3.85	0.13
Испания	Trabajar en Alemania	-0.74	6.87	-0.10
	Trabajo Alemania	-0.65	6.98	-0.15
	Trabajo en Alemania	-0.73	6.72	-0.12
	Ofertas trabajo Alemania	-0.83	7.05	0.01
	Como trabajar en Alemania	-0.89	6.50	-0.01
Болгария	*Работа в Германия*	-1.70	2.91	0.34
Россия	Работа в Германии вакансии	-1.29	5.87	-0.10
	Работа в Германии	-1.12	5.58	-0.04

Источник: расчеты автора на основе данных Google Trends Index (2013–2023 гг.).

Примечание. Всего 155 наблюдений по времени. У стандартизованных индексов нулевое среднее и единичная дисперсия, поэтому они не приводятся. Курсивом отмечены запросы с наибольшей корреляцией с показателем «прибытия иностранцев» из выбранной страны в Германию.

Таблица П3. Таблица корреляций данных о миграции из различных стран в Германию с данными о поисковых запросах и их лагах от 1 до 12 месяцев (в скобках) за период с 01.01.2013 по 06.01.2023. Данные о миграции, а также данные о поисковых запросах приведены в сезонных разностях с периодом в 12 месяцев.

	Лаг, месяц	**Польша** (*Niemcy praca*)	**Италия** (*Lavoro in Germania*)	**Румыния** (*Locuri munca Germania*)	**Испания** (*Ofertas trabajo Alemania*)	**Болгария** (*Работа в Германия*)	**Россия** (*Работа в Германии*)
GTI «Работа в Германии» на языке страны мигранта	0	0.47	0.39	0.21	0.01	0.34	-0.04
	1	0.36	0.43	0.18	0.14	0.36	-0.07
	2	0.39	0.47	0.12	0.14	0.38	-0.03
	3	0.29	0.46	0.02	0.07	0.25	-0.05
	4	0.32	0.48	-0.02	0.02	0.17	0.01
	5	0.31	0.51	-0.02	0.19	0.08	0.01
	6	0.25	0.53	-0.01	0.27	0.10	-0.10
	7	0.16	0.54	0.01	0.17	0.13	-0.11
	8	0.14	0.57	0.09	0.20	0.08	-0.09
	9	0.12	0.55	0.16	0.26	0.07	-0.05
	10	0.08	0.55	0.12	0.24	0.08	-0.07
	11	0.02	0.54	0.07	0.01	0.10	-0.17
	12	-0.14	0.55	0.08	-0.05	-0.01	-0.19

Источник: расчеты автора на основе данных Google Trends Index (2013–2023 гг.).

Таблица П4. Описательная статистика переменных, используемых для оценки миграции из России в Германию для «мультизапросной» модели в работе [Броницкий, Вакуленко, 2024] за период 2021–2022 гг.

	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 Германии	–0.79	3.73
Посольство	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

Источник: расчеты автора на основе данных Google Trends Index (2013–2023 гг.).

Примечание. Из таблицы исключены 22 запроса, отнесенные к категории «остальные».

Таблица П5. Оценки SARIMAX-моделей для 2-летнего контрольного периода

	Лаг, месяц	Польша	Италия	Румыния	Испания	Болгария	Россия
		Спецификация модели SARIMAX(p, d, q)x(P, D, Q, S)					
		(1, 1, 1)x (0, 1, 1, 12)	(1, 1, 1)x (0, 1, 1, 12)	(1, 1, 1)x (0, 1, 1, 12)	(0, 1, 1)x (0, 1, 1, 12)	(1, 1, 1)x (0, 1, 1, 12)	(1, 1, 1)x (0, 1, 1, 12)
	0	1 187.5*** (287.66)		1 343.5*** (408.77)		414.50** (167.34)	
	1						
	2	555.29 (340.25)				265.34** (131.96)	
	3						
	4				-187.42** (94.46)		
	5	618.99*** (236.81)					
GTI «Работа в Германии»	6						
	7					-1.21 (147.45)	
	8		291.67** (138.59)		16.47 (106.12)		- 55.52 (35.96)
	9				-41.92 (100.58)		
	10	274.13 (279.53)	153.32 (178.92)		13.40 (79.98)		
	11						
	12			281.14 (460.67)			
Число наблюдений		114	114	114	114	114	114
AIC		1 746	1 569	1 834	1 451	1 620	1 438
BIC		1 767	1 575	1 840	1 469	1 638	1 444

Источник: расчеты автора.

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

Таблица П6. Оценки качества моделей распределенных лагов для 2-летних контрольных периодов (01.06.2021 – 01.06.2023, а также 01.06.2020 – 01.06.2022)

Страна	Метрики качества моделей MAPE (MAE) на тестовой выборке					
	Модель распределенных лагов		SARIMAX		SARIMA	
	2021–2023	2020–2022	2021–2023	2020–2022	2021–2023	2020–2022
Польша	0.18 (1 466.5)	**0.28 (2 178.5)**	**0.08 (638.3)**	0.59 (4 570.4)	0.29 (2 386.2)	0.34 (2 549.7)
Италия	**0.13 (479.7)**	**0.30 (904.9)**	0.17 (616.6)	0.37 (1 274.6)	0.20 (688.3)	0.49 (1 666.4)
Румыния	**0.11 (1 782.6)**	**0.16 (2 534.7)**	0.11 (1 844.9)	0.41 (6 492.3)	0.14 (2 320.7)	0.36 (5 667.9)
Испания	0.12 (304.8)	**0.25 (487.9)**	**0.08 (178.3)**	0.55 (1 083.4)	0.09 (184.4)	0.59 (1 158.1)
Болгария	0.13 (723.5)	**0.16 (994.8)**	**0.09 (518.1)**	0.18 (1 134.7)	0.10 (574.2)	0.16 (1 005.3)
Россия, однозапросная модель	**0.38 (826.5)**	**0.52 (555.4)**	0.45 (953.4)	0.63 (748.5)	0.79 (1 258.7)	0.66 (778.7)
Россия, мультизапросная модель	**0.22 (436.9)**	**0.17 (184.1)**	-	-		

Источник: расчеты автора.

Сведения об авторе

- Броницкий Георгий Тимурович – стажер-исследователь научно-учебной лаборатории макроструктурного моделирования экономики России, Национальный исследовательский университет «Высшая школа экономики», Москва, 101000, Россия. Email: getibr@gmail.com