Демографическое обозрение / Demographic Review
2022;9(3)75-92 DOI: https://doi.org/10.17323/demreview.v9i3.16471

Прогнозирование миграции из России в Германию с использованием Google-трендов

Георгий Тимурович Броницкий
(gbronitskiy@hse.ru), Национальный исследовательский университет «Высшая школа экономики», Россия.
Елена Сергеевна Вакуленко
(evakulenko@hse.ru), Национальный исследовательский университет «Высшая школа экономики», Россия.

Using Google Trends for external migration prediction

Georgy T. Bronitsky
(gbronitskiy@hse.ru), HSE University, Russia.
Elena S. Vakulenko
(evakulenko@hse.ru), HSE University, Russia.

Резюме: *Международная миграционная статистика публикуется с большой задержкой, которая может достигать нескольких лет. Эта проблема не позволяет исследователям осуществлять своевременный анализ миграционных потоков. В статье рассматривается метод прогнозирования международной миграции на основе поисковых запросов в сети Интернет на примере потоков из России в Германию в период 2011-2020 гг. Для анализа применяли показатели Росстата, статистического офиса Германии и ОЭСР. Предложенный в работе подход позволяет получать оценки миграционной динамики фактически без задержки во времени. Более того, в некоторых случаях возможно предсказывать миграционные события до фактического переезда, что может быть также использовано для прогнозирования других экономических индикаторов. Для построения необходимых оценок в работе были разработаны и применены методы, позволяющие увеличить частотность исходных наблюдений, а также получить краткосрочные ежемесячные прогнозы. Для получения множества поисковых запросов по миграционной тематике использовали NLP- подходы. Были оценены параметры линейной регрессии, построенной на основе данных о частоте использования поисковых запросов Google Trends, связанных с миграционными намерениями. В отличие от модели сезонных авторегрессионных интегрированных скользящих средних (SARIMA), предложенный подход позволяет учитывать структурные сдвиги и шоки в текущих процессах, отраженные в поисковых запросах в Интернете, и дает возможность получать краткосрочные прогнозы миграции в режиме реального времени (наукастинг). Описанные методы можно использовать как при исследовании других пар стран, так и для оценки других статистических показателей.*

Ключевые слова: *международная миграция, миграционная статистика, Росстат, Россия, Google Trends, поисковые запросы, наукастинг, SARIMA, большие данные.*

Для цитирования: *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*

Abstract: *International migration statistics are published with a delay of up to several years. This prevents researchers from making timely analyses of migration flows. The article reviews a method for forecasting international migration flows based on search queries on the Internet using the example of flows from Russia to Germany during 2011-2020. Rosstat, German and OECD data were used to analyze migration. The approach proposed in the paper makes it possible to solve this problem by obtaining an estimate of migration trends with virtually no time delay. Moreover, in some cases it is possible to predict migration events before the actual relocation, which can also be used to evaluate other statistical indicators. To construct the necessary estimates, we employ methods for increasing the data frequency, making it possible to obtain monthly forecasts.*
NLP approaches were used to obtain many search queries on migration topics. As a result, the parameters of a linear regression based on Google Trends search query data were evaluated, which made it possible to make a forecast of migration statistics before the publication of Rosstat statistics. The proposed models, in contrast to the model of seasonal autoregressive integrated moving averages (SARIMA), make it possible to take into account structural shifts and shocks in current processes reflected in Internet search queries, providing the opportunity to obtain short-term migration forecasts in real time (nowcasting). The described methods can be used both in the study of other pairs of countries and for the evaluation of other statistical data.

Keywords: *external migration, migration statistics, Russia, Rosstat, Google Trends, search queries, nowcasting, SARIMA, big data.*

For citation: *Bronitsky G., & Vakulenko E. (2022). Using Google Trends for external migration prediction. Demographic Review, 9(3), 75-92. https://doi.org/10.17323/demreview.v9i3.16471*

Введение

Международная миграция оказывает влияние на различные сферы экономики и общества. Мигранты прибывают с многообразными профессиональными навыками, позволяющими вносить значимый вклад в обогащение человеческого капитала принимающих стран. С другой стороны, отток высококлассных специалистов может отрицательно влиять на потенциал развития значимых отраслей. Таким образом, миграционная политика может существенным образом влиять на развитие государств. Тем не менее недостаточность и несопоставимость статистики во времени и между странами затрудняет оценку влияния международной миграции на экономические показатели. Труднодоступность качественных данных становится причиной высоких расходов на сбор и предоставление достоверной статистики. Различия в методологии учета мигрантов в разных странах сильно затрудняет анализ миграции на мировом уровне. Другой серьезной проблемой, препятствующей анализу миграционных потоков, является временная задержка в размещении информации со стороны статистических служб. Лаг может достигать 5 и более лет, особенно явно он наблюдается в источниках национальной статистики развивающихся стран и стран с переходной экономикой, в которых искомые показатели зачастую основаны на опросах либо просто недоступны. Необходимо отметить, что один инструмент однородных данных наиболее широкого списка стран все же существует. Это результаты Всемирного опроса Гэллапа (GWP) [1]. Опрос проводится среди 1000 взрослых граждан среди 160 стран, в него включены более 100 вопросов, позволяющих построить различные индексы, в том числе и миграционные. Однако результаты международного опроса обладают некоторыми недостатками: во-первых, информация не в свободном доступе и, как правило, платная; во-вторых, миграционные показатели собраны в определенный момент времени, что не позволяет оценить их динамику.

Цель работы – разработать метод прогнозирования международных миграционных потоков с использованием имеющейся официальной статистики стран, а также поисковых запросов в сети Интернет.

Для реализации поставленной цели решались следующие задачи:

1. изучение методологии учета международной миграции в России и Германии;
2. анализ интернет-сервисов по предоставлению агрегированной статистики поисковых запросов;
3. определение наиболее эффективного способа выделения множества слов для анализа запросов в сети Интернет, употребляемых в контексте миграции;
4. разработка подхода к краткосрочному прогнозированию оперативных показателей миграции (наукастинг) с использованием информации о поисковых запросах.

В работе применена модель SARIMA для выделения годовой сезонности и прогнозирования временного ряда. Также использованы методы машинного обучения для работы с текстом, а именно NLP-подходы для выявления наиболее близких к заданному слову поискового запроса. Решается задача оценки параметров множественной регрессии для прогноза международной миграции с использованием данных о поисковых запросах.

Мы предлагаем способ оценки миграционной статистики с минимальной задержкой во времени. Хотя подобная идея и описывалась в более ранних работах

[1] https://www.gallup.com/analytics/318875/global-research.aspx

(Fantazzini et al. 2021; Böhme, Gröger, Stoehr 2020; Wanner 2021), она неприменима в явном виде к показателям Росстата. Новизна нашей работы состоит в предложении использовать метод повышения частотности статистических данных, а также разработке подхода определения множества веб-запросов, которые могут судить о намерении эмигрировать. Описанные методы развивают методологию прогнозирования статистических показателей, а также расширяют возможности использования поисковых запросов для оценки экономических индикаторов.

Обзор литературы

Существует немалое количество литературы, посвященной анализу международных миграционных потоков, влиянию на них демографических факторов и уровня жизни населения, различий в доходах разных слоев населения, миграционной политики. С распространением информационных технологий и появлением возможности отслеживать геопозицию поисковых запросов открываются альтернативные источники для статистического учета перемещений. Изучение поведения пользователей в сети Интернет позволяет опередить публикационный период национальных статистических данных (Hauzenberger, Huber, Klieber 2022), а также спрогнозировать действия населения в ближайшей перспективе (Varian 2014: 3-28).

В своих работах авторы используют различные источники информации, позволяющие определить перемещение людей. Так, в (Bengtsson et al. 2011) при помощи информации о положении SIM-карт авторы делают оценки миграции населения во время землетрясения, что существенно опережает статистику со стороны федеральных служб. В (Subbotin, Aref 2021: 7875–7900) исследуется миграция ученых из России и в Россию на основании библиометрических описаний базы данных (БД) Scopus путем изучения изменений их аффилиаций с учетом областей знаний. Миграционные потоки также пытаются исследовать при помощи метаданных сети Интернет. В работах (Zagheni et al. 2014: 439-444; Moise et al. 2016: 663–670; Kim et al. 2020: 274–286) авторы использовали информацию пользователей Twitter с географической привязкой, в (Zagheni, Weber 2012: 348-351) оценивали международную миграцию, опираясь на IP-адреса десятков миллионов пользователей почты Yahoo, в (Kikas, Dumas, Saabas 2015: 17-22) использовали информацию пользователей Skype, в (Chi et al. 2020: 451-465; State et al. 2014: 531-543) рассматривали географическую привязку профилей социальных сетей. Эти работы направлены на выявление подхода для исследования международной миграции. Однако данные, представленные в них, собраны через специальные онлайн-сервисы, их нельзя интерпретировать для обобщения закономерностей миграции. Более того, база пользователей в большей степени собиралась самостоятельно с использованием информации частных компаний, которую сложно получить другим исследователям.

В работе (Tjaden 2021) проводится обзор работ по исследованию миграции с помощью цифрового следа в сети Интернет, обсуждаются трудности и риски использования альтернативных источников информации для оценки миграционных показателей. В (Чудиновских 2018: 48-56) критикуется использование социальных сетей и данных мобильных операторов для анализа миграций. Автор объясняет это невозможностью отделить собственно миграцию от многообразных форм пространственной мобильности населения: смену места жительства от сезонной мобильности, туризма, транзитных поездок, что существенно влияет на интерпретацию результатов.

Кроме описанных выше методов, появляется все больше литературы, в которой ученые исследуют поисковые запросы и информационный след в социальных сетях для краткосрочного прогнозирования поведения людей и их влияния на экономические показатели. Так, авторы работы (Varian, Choi 2009: 3-28) предполагают, что исследование статистики онлайн-поиска имеет большой потенциал в качестве инструмента измерения желаний пользователей в различных направлениях экономической деятельности как в реальном времени, так и для ближайших прогнозов. В работе (Goel et al. 2010: 17486-17490) описана возможность использования такой информации для прогноза продаж жилья, автомобилей, распространения заболеваний. В (Ginsberg et al. 2008: 1012-1014) было показано, как индикаторы GTI (Google Trends Index 2022) [2] можно использовать для предсказания скорости распространения гриппа с задержкой оценок всего в 1 день. На основе поисковых запросов ученые прогнозируют волатильность фондового рынка (Bazhenov, Fantazzini 2019: 79–88), цены на товары потребления (Fantazzini 2014), информационное поле и потенциальные темы новостей (Radinsky, Davidovich, Markovitch 2008: 363-367) и другие явления частого потребления и поведения (Wu, Brynjolfsson 2013). Авторы отмечают, что такие исследования стали возможны благодаря внедрению машинного обучения (Mullainathan, Spiess 2017: 87-106; Celbis 2022) и эконометрических моделей, которые легли в основу проведения практических исследований миграционных потоков.

Возвращаясь к теме миграции, следует отметить, что существует ряд работ, использующих показатели Google Trends Index для оценки миграционных потоков. GTI является инструментом для анализа информационного следа пользователей различных онлайн-ресурсов. Эмпирические данные указывают, что мигранты все чаще ищут актуальную информацию о переезде в сети Интернет в стране их происхождения до отъезда (Fantazzini et al. 2021; Böhme, Gröger, Stoehr 2020). Интенсивность поисковых запросов можно использовать для прогноза динамики потоков миграции в перспективе нескольких месяцев. Такой подход может быть актуальным на частном и национальном уровнях для политических и экономических изменений. Среди недостатков применения Google Trends для анализа миграции называются сложность разделения типов миграции, агрегированность показателей и невозможность отслеживать характеристики потенциальных мигрантов, а также их реальные действия, так как индекс показывает намерения. В работе (Wanner 2021: 1181-1202) Google Trends использовались для краткосрочного прогнозирования международной трудовой миграции в Швейцарию. Авторы решают проблему отделения разных типов миграции подбором более точных веб-запросов, в частности связанных с поиском работы.

Рассматриваемые методы являются продолжением идеи исследования поисковых запросов с целью прогнозирования миграционных потоков. Нами предложены способы формирования множества поисковых запросов на основе лингвистических моделей машинного обучения, что не применялось в предыдущих работах схожей тематики. Кроме этого, мы применяли метод повышения частотности данных, который позволяет перейти от использования годовой статистики к месячным показателям.

[2] https://trends.google.ru/trends/?geo=RU

Статистические данные

Для анализа корреляции между поисковыми запросами населения и фактической международной миграцией мы опираемся на общедоступную статистику о миграционных потоках из Российской Федерации. Статистические офисы стран по-разному учитывают мигрантов, поэтому необходимо изучить их методологию и особенности подсчета. Отметим важные для моделирования и прогнозирования характеристики, которыми должны обладать данные:

- высокая частотность (важно использовать источники с высокой частотой для предсказывания движения населения с меньшей задержкой);
- полнота (в случае пропусков в наблюдениях модель может работать некорректно).

При выборе страны назначения важно учитывать, что мы планируем прогнозировать миграцию по запросам в сети Интернет. Это накладывает ограничение на предлагаемую методологию, связанное с доступностью Интернета для потенциальных мигрантов. Существует значимое смещение в доле населения, имеющей доступ к Интернету в развивающихся и развитых странах, поэтому мы хотели бы сфокусироваться на миграции населения из РФ в развитые страны.

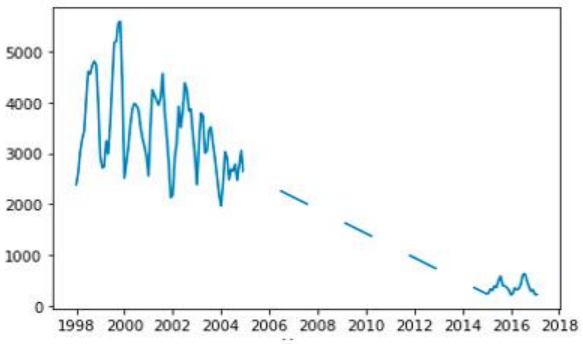
Для оценки полноты и числа мигрантов мы использовали данные Единой межведомственной информационно-статистической системы (ЕМИСС) [3], которая содержит официальную статистическую информацию, формируемую субъектами официального статистического учета (Росстат 2022). Для поиска информации о миграции из РФ в периоды с 1998 г. по настоящее время был выбран показатель «Число выбывших» [4], в фильтре «потоки миграции» было выбрано значение «Международная». Опираясь на эти показатели и перечисленные выше критерии, мы определили список потенциально интересных для исследования стран, которые исторически связаны с Россией наиболее значимыми миграционными связями: Германия, Израиль и США. Объединение нескольких стран может привести к увеличению ошибки в прогнозе, поэтому было решено выбрать одну страну для исследования. Мы остановимся на исследовании миграции из России в Германию, однако предложенные методы можно применять и для других пар стран.

Для оценки миграции из России в Германию были изучены показатели, публикуемые Росстатом. В работах (Чудиновских, Степанова 2020: 54–82; Чудиновских 2010: 8-16; 2016: 32-46; 2020: 68-90; Чудиновских, Донец 2018: 11-26) исследуется методология, используемая для сбора миграционной статистики в России. Колебания статистических данных зачастую объясняется не столько изменением миграционных потоков, сколько изменением методологии сбора статистики. Наиболее существенным было изменение в 2011 г., согласно которому в России начали вести учет получающих регистрацию по месту пребывания на 9 месяцев и более. Такое изменение повлияло на показатели миграционных потоков, в основном из-за временных/трудовых мигрантов, имеющих временную регистрацию, а также стала учитываться студенческая миграция (Чудиновских 2020: 68-90). Однако в случае миграции граждан РФ необходимо сняться с регистрации по месту жительства для попадания в миграционную статистику. Поэтому такой показатель не позволяет в полной мере оценить миграционный поток в

[3] https://rosstat.gov.ru/emiss
[4] https://www.fedstat.ru/indicator/43513

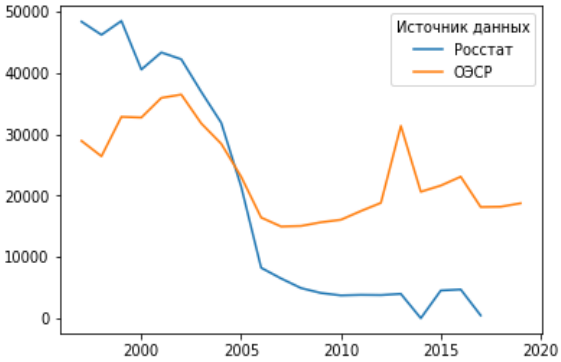
выбранную страну. Среди преимуществ можно отметить, что статистика по международной миграции доступна с 1998 г. с месячной частотностью. Однако есть и ряд недостатков: в данных ЕМИСС отсутствует статистика за период 2005-2015 гг. (рисунок 1), также можно заметить, что миграционные показатели публикуются с большой задержкой. На момент написания статьи она достигает 5 лет.

Рисунок 1. Число мигрантов из РФ в Германию по месяцам



Источник: Данные Росстата (1998-2018 гг.).

Рисунок 2. Число мигрантов из РФ в Германию по годам



Источник: Данные Росстата, ОЭСР.

Учитывая приведенные выше факты, мы не можем опираться только на показатели Росстата, поэтому были изучены альтернативные источники статистики. Согласно (Денисенко 2003: 157-169) в 90-е годы эмиграционный отток из России в Германию оценивался статистической службой Германии в 1,2 раза выше, чем аналогичные показатели о зарегистрированных мигрантах со стороны статистической службы России. В статистике Германии иностранные граждане регистрируются как иммигранты, если они получили разрешение на жительство и собираются остаться в Германии на 3 и более месяцев. Эмигрантами также считаются «граждане Германии и лица немецкого происхождения (Aussiedler), которые возвращаются на свою историческую родину» и практически автоматически получают гражданство. Таким образом, сравнивать немецкие и российские показатели можно с некоторыми оговорками и допущениями, поскольку статистика Германии включает как краткосрочные миграции, так и долгосрочные перемещения. Но для нас важны оба вида миграции, так как по статистике интернет-запросов сложно их разделить.

Указанное превышение в объемах миграции характерно не только для потоков в Германию, но и в другие страны, которые рассматривались в работе (Денисенко 2003). Исходя из этого делается вывод о систематическом недооценивании эмиграционных потоков в России. Воспользуемся показателями статистической службы Германии [5], а также Организации экономического сотрудничества и развития (ОЭСР) (OECD 2020) [6], в которой указываются ежегодные миграционные потоки среди стран-участниц организации. Поскольку эта база данных пополняется регистрациями населения по месту жительства и занятости из стран-членов ОЭСР, она охватывает только легальную иммиграцию, т. е. миграцию, связанную с работой, лиц, ищущих убежища, и другие типы легальных

[5] https://www-genesis.destatis.de/genesis/online?operation=statistic&levelindex=0&levelid=1668599374655&code=12711#abreadcrumb
[6] https://stats.oecd.org/Index.aspx?DataSetCode=MIG#

иммигрантов. Критерием для пополнения статистики является наличие вида на жительство и нахождение в стране более 1 недели [7]. В статистике ОЭСР учитываются показатели статистической службы Германии, при этом этнические немцы, о которых говорилось ранее, в ОЭСР не учитываются. В работе использованы данные ОЭСР, а не первоисточника, так как эта организация приводит статистику миграционных служб разных стран к единому виду, что позволяет их сравнивать друг с другом, а также применять описанные далее методы к другим парам стран. Также можно получить миграционную статистику в базе Eurostat [8] (Eurostat 2020), однако срез данных по гражданству иммигранта недоступен для выбранной пары стран. Одной из проблем при использовании такой информации о миграционных потоках является частотность: это только годовые показатели.

Для оценки миграционных потоков из России в Германию предпочтительнее использовать показатели принимающей стороны. Так как статистика Росстата применяет меняющиеся со временем методики учета мигрантов, а также может не учитывать отток, если при выезде из страны эмигранты не снимаются с регистрационного учета, они не видны в статистике. Стоит отметить, что в данных принимающей стороны мы можем видеть не только долгосрочную, но и краткосрочную миграцию, но с видом на жительство и со сменой места регистрации. Несмотря на то, что в немецкой статистике указаны годовые показатели миграции, в своем исследовании мы использовали комбинацию нескольких источников: Росстата (выделив сезонную компоненту по месячным показателям) и ОЭСР. Подробнее о методологии вычисления сезонности будет рассказано в следующем разделе.

При сравнении показателей Росстата и ОЭСР (рисунок 2) видно, что они имеют схожие тренды в период 1998-2004 гг. Снижение числа фиксируемых Росстатом переселений совпадает с началом периода отсутствия месячной статистики. Годовые показатели миграции из РФ в Германию в исследуемых источниках начали существенно различаться. Такие различия приводят к необходимости отдельного изучения миграционных данных, разработке алгоритмов повышения частотности.

Несмотря на то, что информация о миграционных потоках доступна за более ранние периоды, мы сосредоточимся на периоде, начинающемся с 2004 г., когда доступна статистика используемых нами поисковых запросов, и до 2020 г. – последнего года публикации ОЭСР показателей международных миграционных потоков.

Данные Интернет-запросов

Раздел посвящен обзору открытых источников поисковых запросов, таких как Google Trends и Yandex Wordstat [9]. Индекс Google Trends (GTI) предоставляет собой временной ряд интенсивности поиска выбранных пользователем ключевых слов. Индекс запросов показывает долю обращений: общий объем поиска по данному запросу в выбранном географическом регионе, разделенный на общее количество запросов в этом регионе в определенный момент времени.

GTI может быть ограничен по географическому региону, дате, набору общих категорий поиска, таких как «Работа и образование» или «Путешествия», а также по типу

[7] Важно отменить, что в статистику миграции ОЭСР туристы не попадают, так как важно выполнение сразу двух условий: наличие вида на жительство и нахождение в стране.
[8] https://ec.europa.eu/eurostat/databrowser/view/migr_imm2ctz
[9] https://wordstat.yandex.ru/

поиска. Мы используем первые два ограничения, основанные на веб-поиске, для создания ежемесячных временны*х* рядов интенсивности онлайн-поиска по выбранным странам. К недостаткам можно отнести то, что индекс показывает нормированный показатель доли запросов в выбранной тематике. Это приводит к несопоставимости индексов для разных временны*х* промежутков (разные показатели одного запроса при разном временно*м* окне) и отсутствию возможности рассчитать абсолютные показатели. Другой проблемой может стать изменение способов подсчета статистики со стороны Google. Так, в 2011 г. изменился подход к определению региона поиска, что повлияло на подсчет индекса. Из-за этого мы не можем использовать веб-запросы до 2011 г. Еще одной проблемой использования более ранних поисковых запросов является низкая доля домохозяйств, имеющих доступ к сети Интернет в РФ: в 2000 г. – 2% домохозяйств, в то время как в 2011 г. – 56,8%. Учитывая описанные ограничения, было решено применять индексы интернет-запросов, начиная с 2011 г. При том мы все же использовали статистику более ранних миграционных показателей 1998-2011 гг. для вычисления месячной сезонности и увеличения частотности исходных данных.

Альтернативным подходом можно считать Yandex Wordstat, в отличие от индекса GTI они содержат абсолютные показатели поисковых запросов, что упрощает их использование для прогноза. Однако такая статистика доступна только за ограниченный период времени (за последний год). Еще одним минусом применения Yandex Wordstat является его географическое смещение (доля Яндекса): Яндекс имеет низкую долю пользователей среди поисковых систем за пределами стран СНГ, что не позволит масштабировать используемый алгоритм. Несмотря на это, добавление показателей Yandex Wordstat может увеличить точность алгоритма за счет добавления абсолютных величин. В своей работе мы не проводим сравнение моделей, построенных только на индексах GTI, с моделями, дополненными статистикой Yandex Wordstat за последний год, однако это является интересной темой для последующих исследований.

Для использования Google Trends или Yandex Wordstat необходимо выбрать множество поисковых запросов, которые будут коррелировать с фактической миграцией. Ввиду многомерности миграционных процессов и мотивов, эта задача является более сложной, чем в других приложениях, где набор потенциальных ключевых слов довольно узок, например, в случае с продажами автомобилей, ценами на нефть и реестрами безработных. В работе применены 2 подхода к поиску ключевых слов. Первый – определение семантического ядра с использованием сервиса Yandex Wordstat. Такой подход позволяет найти популярные запросы со схожей тематикой. Так, например, для запроса «эмиграция» Yandex Wordstat находит похожим запрос «гражданство». К преимуществам этого подхода можно отнести высокое качество поисковых запросов, так как фактически с их помощью пользователи ищут информацию в Интернете. Но также есть и недостатки, основным из которых является невозможность оценить полноту полученного множества. Для оценки полноты множества полученных слов использовали ресурс RusVectores (NLPL 2022) [10,11], а также мы провели сравнение с множеством, полученным в работе (Böhme, Gröger, Stoehr 2020). Важно отметить, что для формирования множества популярных поисковых запросов производили поиск по ключевым словам с использованием информации о регионе (например, «переезд в Германию», «работа в

[10] https://rusvectores.org/
[11] http://vectors.nlpl.eu/repository/

Германии»). Второй способ определения множества слов – использование эмбеддингов (представление слов или словосочетаний в виде векторов). Модели в качестве обучающих выборок могут использовать статьи *Википедии*, языковые словари и другие источники, в которых можно найти качественную взаимосвязь между словами. Такой подход позволяет найти слова и словосочетания, которые наиболее часто употребляются в литературе в контексте эмиграции. Так, модель, обученная на текстах русскоязычной Википедии, определяет наиболее близкими к слову «эмиграция» слова «эмигрант», «заграница», «беженство», «диаспора». Модель, построенная для работы с текстом, называется *корпус*, такие модели могут брать за основу различные источники для поиска взаимосвязей между словами. Мы использовали наиболее популярные модели: Российский национальный корпус [12], корпус Тайга [13], а также корпус российской Википедии [14]. К преимуществам такого подхода можно отнести полноту полученных данных, есть возможность определять порог качества, для которого мы считаем словосочетания близкими друг к другу. Противоположно первому подходу не всегда получаются словосочетания, которыми пользуются при поиске в Интернете (например, «эвакуация в Германию»; хотя слово «эвакуация» находится близко к слову «эмиграция» с точки зрения модели, поисковых запросов с таким сочетанием крайне мало). В случае малого количества запросов словосочетаний для них нет истории GTI, поэтому мы не можем получить часть индексов. В результате поиска по всем 150 запросам была найдена история по 20-35 запросам в каждом из множеств.

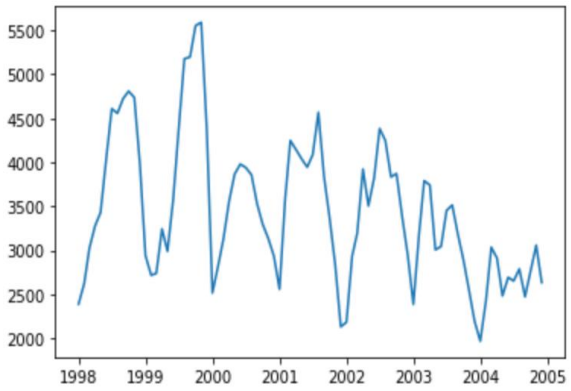
Поскольку во время подготовки и сбора данных было трудно оценить наилучший способ составления множества поисковых запросов, мы решили использовать оба подхода и построить модель для каждого из них. В результате для каждого из подходов определили множество из примерно 150 запросов, для которых в дальнейшем производили поиск в истории Google Trends. Стоит отметить, что мы рассматривали запросы только на русском языке. Как обсуждается в работе (Wanner 2021: 1181–1202), эмигранты в основном используют язык страны выбытия для поиска.

Методология и анализ результатов

Для подготовки миграционной статистики и повышения ее частотности было решено выделить сезонную компоненту, используя показатели Росстата в периоды 1997-2005 гг. На рисунке 3 можно заметить годовую сезонность (12 месяцев), при этом нет ярко выраженного тренда. Рассмотрим также ряд первых разностей исходного ряда. Для этого из каждого текущего значения ряда вычтем предыдущее (рисунок 4). Кроме этого, посчитаем скользящее среднее ряда окном за последние 12 периодов. По графику скользящего среднего можно отметить отсутствие тренда в данных. Используя тест Дики-Фуллера (спецификация с константой), проверим стационарность указанных рядов. При уровне значимости в 1% мы можем отвергнуть гипотезу, что ряд первых разностей не стационарен, получив значение статистики (ADF = -3,48; p_value = 0,0085).

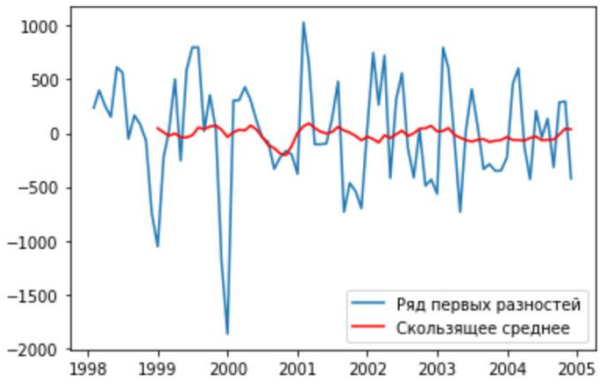
[12] https://ruscorpora.ru/
[13] https://tatianashavrina.github.io/taiga_site/
[14] http://vectors.nlpl.eu/repository/20/182.zip

Рисунок 3. Число мигрантов из РФ в Германию по месяцам, 1998-2005



Источник: Данные Росстата.

Рисунок 4. Скользящее среднее и ряд первых разностей для числа мигрантов из РФ в Германию по месяцам, 1998-2005



Источник: Расчеты авторов по данным Росстата.

После проведения предварительного анализа используем информационный критерий AIC (Akaike 1974: 716–723) для определения наилучших параметров модели $SARIMA\ (p,d,q)(P,D,Q)_s$.

Используя выбранный критерий, определили параметры модели, наилучшим образом описывающие временной ряд, представленный на рисунке 3. Согласно AIC критерию, исходный временной ряд можно описать процессом со следующими параметрами: $SARIMA\ (0,1,1)(0,1,1)_{12}$, такой процесс будет иметь следующий вид:

$$\left(1-\frac{X_{t-1}}{X_t}\right)\left(1-\frac{X_{t-12}}{X_t}\right)X_t=\left(1-\theta\frac{X_{t-1}}{X_t}\right)\left(1-\Theta\frac{X_{t-12}}{X_t}\right)\epsilon_t \quad (1)$$

где θ,Θ – оценки параметров модели, ϵ_t – остатки модели, из статистики Льюнга-Бокса (Q-stat = 0,11, p_value = 0,13) следует, что они соответствуют модели «белого шума» и независимы.

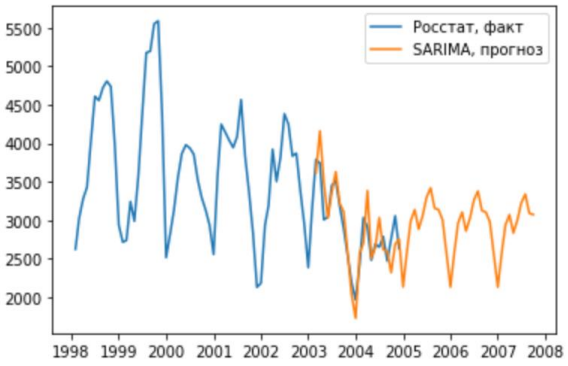
Составим уравнение временного ряда, используя полученные параметры [15]:

$$X_t=X_{t-1}+X_{t-12}-X_{t-13}+\ \epsilon_t+\underset{(0,130)}{0.433}\epsilon_{t-1}-\underset{(1,978)}{0.967}\epsilon_{t-12}-\underset{(0,257)}{0.419}\epsilon_{t-13} \quad (2)$$

Сделаем прогноз временного ряда, используя полученное уравнение (2) для интервала 2003-2008 гг. (рисунок 5). В таком случае для этого периода мы можем сравнить фактические значения показателя с прогнозными. Кроме этого, мы строим прогноз значений в интервале 2005-2008 гг., для которых нет ретроспективной информации с месячной частотностью. Оцененный ряд можно продлить и далее, но из-за отсутствия тренда он не будет видоизменяться от года к году.

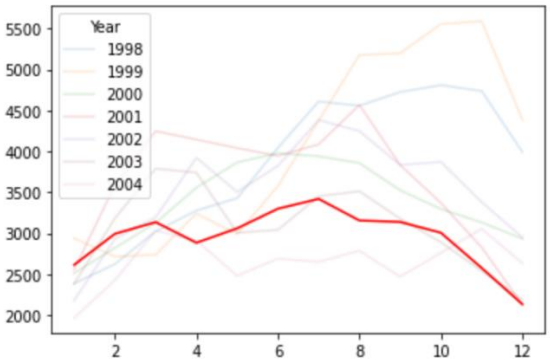
[15] В круглых скобках представлены стандартные ошибки коэффициентов модели.

Рисунок 5. Фактические и прогнозные значения числа мигрантов из РФ в Германию по месяцам



Источник: Данные Росстата.

Рисунок 6. Число мигрантов из РФ в Германию по месяцам в году; сезонность, выявленная в модели SARIMA, 1998-2004



Источник: Данные Росстата.
Примечание: Жирная линия – сезонность, выявленная в модели SARIMA.

Получившийся временной ряд можно использовать для обучения модели, основанной на поисковых запросах. Однако тот факт, что он не учитывает миграционный тренд, существенно затрудняет его использование на данных 2011-2022 гг., для которых есть статистика по запросам GTI. В связи с этим было решено частично использовать выводы, которые мы получили из модели SARIMA, выделив сезонную компоненту в исследуемых показателях.

Анализ временного ряда, основанного на данных Росстата за 1997-2005 гг., позволяет выявить годовую сезонность (12 месяцев). Так как статистика Росстата за указанные периоды представлена с помесячной частотностью, определим среднюю сезонность на периоде 1997-2005 гг. (рисунок 6). Полученную сезонность применим к годовым показателям ОЭСР. Таким образом получим временные ряды, учитывающие тренд, с нужной для исследования частотностью. Результат повышения частотности данных отображен на рисунке 7.

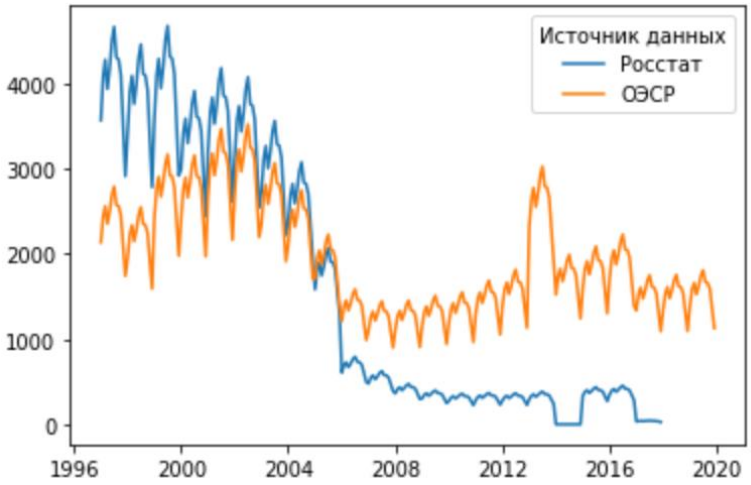
Используя ретроспективные показатели миграции между двумя странами, можно оценить корреляцию между поисковыми запросами и фактической эмиграцией. Ранее мы писали о методах формирования множества поисковых запросов, соответствующих миграционным намерениям. Для каждого множества поисковых запросов мы оценили параметры множественной линейной регрессии:

$$\gamma_t = \sum_i \beta_i x_{i,t} + u_t \qquad (3)$$

где в качестве объясняющих переменных x_i выбраны значения индексов Google Trends для соответствующего запроса i. В качестве объясняемой переменной Y_t взяты показатели миграции из России в Германию по статистике ОЭСР с увеличенной нами частотностью (рисунок 7). Для оценки коэффициентов модели β_i применяли метод наименьших квадратов на данных за 2011-2018 гг. Для сравнения точностей прогнозов моделей (3) с различными множествами поисковых запросов использовали период 2019-2020 гг. (таблица и рисунки 8-11). Мы также делаем прогноз миграции из России в Германию с

2020 г. по настоящее время.

Рисунок 7. Годовые данные о миграции с примененным внутригодовым трендом, 1996-2020



Источник: Данные Росстата, ОЭСР.

Примечание: Статистика миграции из РФ в Германию.

Модель, основанная на множестве поисковых запросов, определенных с помощью Yandex Wordstat, показала наименьшую ошибку (MAE и MAPE). Это объясняется тем, что в такой модели используются словосочетания, которые чаще всего применяются для поиска в Интернете. Этот подход подбора ключевых слов позволил найти больше информации по поисковым запросам в GTI.

Таблица. Сравнение моделей (3) для прогноза миграции из РФ в Германию с помощью различных поисковых множеств

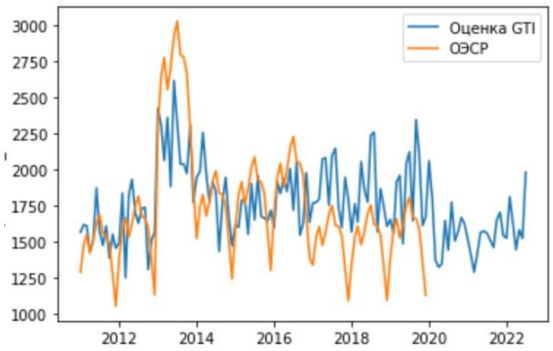
Источник поисковых запросов	Найденные запросы	MAPE	MAE	Значимые запросы*
Национальный корпус	20	14,7	271,3	визовый Германия право Германия гражданство Германия
Википедия	25	14,4	167,4	виза Германия билет Германия посольство Германия
Тайга корпус	25	14,4	265,8	виза Германия ВНЖ Германия иммиграция Германия
Yandex Wordstat	32	10,4	183,5	шенгенская виза в Германию посольство Германии в Москве работа в Германии

*Примечание: * – Значимые запросы на уровне значимости 5%. Приведены примеры, соответствующие наибольшим значениям параметров β_i в модели линейной регрессии (4). ВНЖ – вид на жительство.*

Мы можем сравнить поведение модели, основанной на годовой сезонности (показатели ОЭСР на рисунках 8-11), с прогнозами миграции на основе запросов Google Trends. Прогнозы, полученные на основе моделей временны*х* рядов (SARIMA), не способны отобразить шоков в миграционных показателях, возможных структурных сдвигов, так как строятся на основе ретроспективной информации, это хорошо заметно на

рисунках 8 и 10, где видно падение миграционного потока в период Covid-19 в начале 2020 г. по моделям Google Trends. Этот факт объясняется тем, что индекс GTI отражает реальный интерес к миграции в каждый момент времени, в то время как прогнозирование временных рядов строится на ретроспективных данных и в большей степени отражает предшествующий тренд и сезонность.

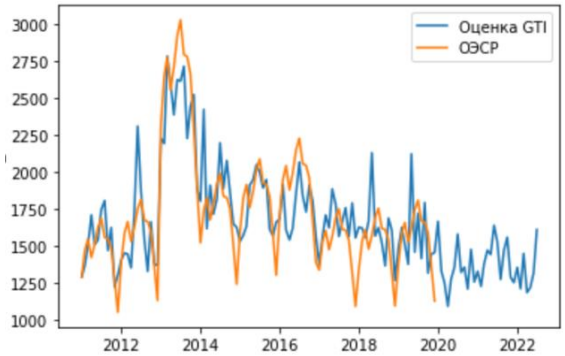
Рисунок 8. Прогноз миграционных показателей (данные Национального корпуса)



Источник: Данные Google Trends Index, ОЭСР

Примечание: Множество поисковых запросов (поисковых слов) сформировано на основе анализа взаимосвязи слов русскоязычной литературы, новостей, словарей (национальный корпус).

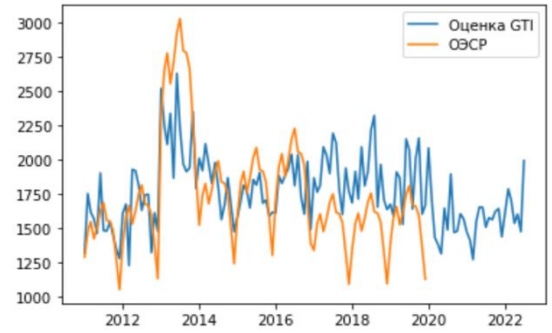
Рисунок 9. Прогноз миграционных показателей (данные Yandex Wordstat)



Источник: Данные Google Trends Index, ОЭСР, Yandex Wordstat

Примечание: Множество поисковых запросов (поисковых слов) сформировано на основе поисковой истории Яндекса за последний год.

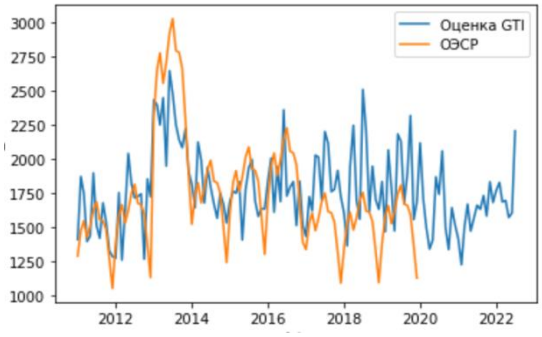
Рисунок 10. Прогноз миграционных показателей (данные Википедии)



Источник: Данные Google Trends Index, ОЭСР.

Примечание: Множество поисковых запросов (поисковых слов) сформировано на основе анализа взаимосвязи слов из русскоязычной Википедии (корпус Википедии).

Рисунок 11. Прогноз миграционных показателей (данные корпуса Тайга)



Источник: Данные Google Trends Index, ОЭСР.

Примечание: Множество поисковых запросов (поисковых слов) сформировано на основе анализа взаимосвязи слов из русскоязычной литературы (корпус Тайга).

Недостатком модели, построенной на основе запросов Google Trends, является то, что в момент шоков возможны изменения в соотношении количества целевых действий

(миграций) и количества соответствующих запросов. Одним из решений обозначенной проблемы может стать добавление переменной, отражающей скорость изменения поисковых запросов заданной тематики. Кроме этого видно, что модель, обученная с использованием данных Yandex Wordstat (рисунок 9), сравнительно лучше повторяет падение с 2017 по 2020 г., это может быть объяснено тем, что она построена на основе слов, которыми люди чаще пользуются при поиске в Интернете, когда принимают решение о миграции.

Заключение

В работе был предложен подход к прогнозированию статистических показателей о миграции из РФ в Германию на основе поисковых запросов в сети Интернет. Подход состоит из нескольких этапов. Первым делом мы анализировали имеющуюся статистику по международной миграции из России в Германию по информации статистических служб каждой из стран, обсуждали их сопоставимость. Для построения краткосрочных прогнозов необходимы высокочастотные данные, хотя бы месячной периодичности. Для этого нами был разработан и применен метод увеличения частотности статистических показателей, который позволил из годовых данных статистического офиса Германии и ОЭСР получить помесячные. На втором этапе проводили работу по определению множества поисковых запросов в сети Интернет, описывающих намерения мигрировать. Для этого нами были применены лингвистические методы машинного обучения. В основе использованных подходов лежит представление слов в виде векторов. Производили поиск наиболее близких слов в контексте миграции, вычисляя расстояние между полученными векторами. Мы сравнивали несколько источников русскоязычного текста (статьи Википедии, словари и русскоязычная литература), таким образом получали различные множества, близкие слову «эмиграция», в зависимости от источника. В результате было отобрано от 20 до 35 поисковых запросов для каждого из исследуемых множеств. На финальном третьем этапе оценивали линейные регрессионные модели, позволяющие предсказывать потоки миграции, опираясь на поисковые запросы по статистике Google Trends. В результате было показано, что модель, обученная с использованием множества поисковых слов Yandex Wordstat, лучше других описывает падение миграции с 2017 по 2020 г. Хотя оцененные параметры модели не могут быть использованы с другим временны*м* промежутком, в работе описан принцип получение прогноза в будущем. Было проведено сравнение модели прогноза на основе сезонного тренда и линейной модели с переменными запроса Google Trends. Преимуществом модели, основанной на индексах Google Trends, сравнительно с моделью SARIMA является возможность учитывать структурные сдвиги в динамике миграции из-за шоков различной природы (Covid-19, изменение социально-экономического положения, кризисы и др.).

Применение описанного выше подхода позволяет получить оперативную информацию о международной миграции раньше, чем она будет опубликована в официальных статистических источниках (наукастинг миграционных потоков). Так, данные Росстата могут иметь задержку более двух лет, что может затруднять оценку экономических показателей, основанных на демографических переменных. Более того, в некоторых случаях описанный подход позволяет предсказать миграционные намерения еще до момента миграции.

Отдельно нужно обсуждать расхождение полученных прогнозных значений по предложенным моделям с поступившими с некоторой задержкой данными официальной

статистики. На самом деле эти различия могут быть вызваны разными причинами: 1) статистической ошибкой прогноза по модели [16]; 2) не воплотившимися в миграцию намерениями о миграции, 3) неучетом в статистике всех возможных краткосрочных перемещений, которые видны в поисковых запросах. Несмотря на имеющиеся ограничения, мы считаем, что предложенный подход имеет право на существование, поскольку дает нам оперативную информацию о миграционных трендах, отраженных в поисковых запросах, которые могут изменяться в моменты сильных структурных шоков (пандемия, войны, землетрясения и др.). Подходы к прогнозированию, основанные на анализе динамики временны*х* рядов и демографической структуры населения, не позволяют получить адекватные оценки в таких ситуациях.

Описанные методы можно использовать как при исследовании других пар стран, так и для оперативного прогнозирования других статистических показателей. При применении предложенного подхода важно учитывать особенности методологического учета миграции в разных странах. Можно также проводить дополнительные исследования в области выбора множества поисковых запросов, основываясь на минимизации дисперсии при прогнозировании.

К недостаткам предложенного подхода прогнозирования оперативных значений показателей миграции можно отнести тот факт, что полученные параметры регрессии не являются постоянными. Google Trends Index для одного и того же дня и запроса может отличаться в зависимости от выбранного временного периода (окна). Это связано с тем, что Google нормирует показатель на выбранном периоде (делит каждое из значений на максимальное). Таким образом, значения Google Trends Index в момент времени $T+1$ могут отличаться от тех, которые были в момент написания статьи (кроме случая выбора аналогичного периода). Для решения этой проблемы можно заново оценить параметры регрессии, используя описанный выше алгоритм. Такой подход позволит построить правильный прогноз модели в будущем.

Направлением дальнейших исследований может стать изучение временны*х* лагов между запросами в Интернете и непосредственной миграцией. Также можно попробовать подобрать поисковые запросы так, чтобы отделить миграцию разных видов (смена места жительства, работа, учеба). В своей работе мы долгосрочную и краткосрочную миграцию не разделяли. Стоит также задуматься о поисковых запросах на других языках, в нашем случае на английском и немецком в дополнении к русскому.

Литература

Денисенко М.Б. (2003). Эмиграция из России по данным зарубежной статистики. *Мир России: Социология, этнология,* 12(3), 157-169.

Росстат (2022). База данных. Федеральная служба государственной статистики, показатель «Число выбывших». https://www.fedstat.ru/

Чудиновских О.С. (2010). Современное состояние статистики миграции в России: новые возможности и нерешенные проблемы. *Вопросы статистики*, 6, 8-16.

[16] В данном случае стоит ориентироваться на доверительный интервал для прогноза, а не точечные оценки.

Чудиновских О.С. (2016). Административная статистика международной миграции: источники, проблемы и ситуация в России. *Вопросы статистики*, 2, 32-46.

Чудиновских О.С. (2018). Большие данные и статистика миграции. *Вопросы статистики,* 25(2), 48-56.

Чудиновских О.С. (2020). К вопросу о статистическом обеспечении исследований миграции и миграционной политики в России. *Управление миграцией и модели миграционной политики: возможности и риски*, 68-90.

Чудиновских О.С., Донец Е.В. (2018). О новых технологиях и статистике миграции в России. *Вопросы статистики*, 25(5), 11-26.

Чудиновских О.С., Степанова А.В. (2020). О качестве федерального статистического наблюдения за миграционными процессами. *Демографическое Обозрение*, *7*(1), 54-82. https://doi.org/10.17323/demreview.v7i1.10820

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Bazhenov T., Fantazzini D. (2019). Forecasting Realized Volatility of Russian stocks using Google Trends and Implied Volatility. *Russian Journal of Industrial Economics*, 12(1), 79-88. https://doi.org/10.17073/2072-1633-2019-1-79-88

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

Böhme M. H., Gröger A., Stoehr T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142.

Celbiş M. G. (2022). Unemployment in Rural Europe: A Machine Learning Perspective. *Applied Spatial Analysis and Policy*. https://doi.org/10.1007/s12061-022-09464-0

Fantazzini D. (2014). Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data. *PloS One*, 9. https://doi.org/10.1371/journal.pone.0111894

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

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. https://doi.org/10.1007/978-3-030-36683-4_37

Eurostat (2020). Database of statistical office of the European Union. Immigration Database. https://ec.europa.eu/eurostat/web/main/data/database

Ginsberg J., Mohebbi M., Patel R., Brammer L., Smolinski M., Brilliant L. (2008). Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, 457, 1012–1014. https://doi.org/10.1038/nature07634

Goel S., Hofman J., Lahaie S., Pennock D., Watts D. (2010). Predicting Consumer Behavior with Web Search. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 17486-17490. https://doi.org/10.1073/pnas.1005962107

Google Trends Index (2022). Database. Explore what the world is searching for by entering a keyword or a topic. www.google.com/trends/

Hauzenberger N., Huber F., Klieber K. (2022). Real-time inflation forecasting using non-linear dimension reduction techniques. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2022.03.002

Kim J., Sîrbu A., Giannotti F., Gabrielli L. (2020). Digital Footprints of International Migration on Twitter. In Berthold, M., Feelders, A., Krempl, G. (Eds.), *Advances in Intelligent Data Analysis XVIII* (pp. 274-286). Konstanz: Springer. https://doi.org/10.1007/978-3-030-44584-3_22

Kikas R., Dumas M., Saabas A. (2015). *Explaining International Migration in the Skype Network*. 17-22. https://doi.org/10.1145/2806655.2806658

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* (pp. 663-670). Konstanz: Springer. https://doi.org/10.1109/ICDMW.2016.0099

Mullainathan S., Spiess J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87-106. https://doi.org/10.1257/jep.31.2.87

NLPL (2022). Database. NLP models trained with stated hyperparametes.: http://vectors.nlpl.eu/repository/#

OECD (2020). Database on Immigrants in OECD Countries (DIOC). International Migration Database. https://stats.oecd.org/Index.aspx?DataSetCode=MIG#

Radinsky K., Davidovich S., Markovitch S. (2008). Predicting the News of Tomorrow Using Patterns in Web Sea*r*ch Queries. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 363-367). Sydney: IEEE. https://doi.org/10.1109/WIIAT.2008.215

State B., Rodriguez M., Helbing D., Zagheni E. (2014). Migration of professionals to the U.S.: Evidence from linkedin data. In Aiello L.M., McFarland D. (Eds.), *6th international conference on social informatics* (pp. 531–543). Barcelona: Springer. https://doi.org/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. https://doi.org/10.1007/s11192-021-04091-x

Tjaden J. (2021). Measuring migration 2.0: a review of digital data sources. *CMS* 9, 59. https://doi.org/10.1186/s40878-021-00273-x

Varian H., Choi H. (2009). Predicting the Present with Google Trends. *Economic Record*, 88. https://doi.org/10.2139/ssrn.1659302

Varian H.R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wanner P. (2021). How well can we estimate immigration trends using Google data? *Qual Quant,* 55, 1181-1202. https://doi.org/10.1007/s11135-020-01047-w

Wu L., Brynjolfsson E. (2013). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2022293

Zagheni E., Garimella V., Weber I., State B. (2014). Inferring international and internal migration patterns from Twitter data. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 439-444). New York: Association for Computing Machinery https://doi.org/10.1145/2567948.2576930

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* (pp.348-351). New York: Association for Computing Machinery. https://doi.org/10.1145/2380718.2380764