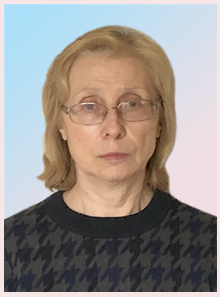
DOI: 10.15838/esc.2022.1.79.4
УДК 314.748, ББК 60.7

Статистика онлайн-запросов в наукастинге миграции



Ирина Павловна
ЦАПЕНКО
Национальный исследовательский институт мировой экономики
и международных отношений имени Е.М. Примакова РАН
Москва, Российская Федерация
e-mail: tsapenko@bk.ru
ORCID: 0000-0001-6065-790X; ResearcherID: B-1993-2017

Максим Андреевич
ЮРЕВИЧ
Финансовый университет при Правительстве РФ
Москва, Российская Федерация
e-mail: maksjuve@gmail.com
ORCID: 0000-0003-2986-4825; ResearcherID: J-9698-2014

Аннотация. Рост значимости международных миграций в жизни современных государств повышает востребованность надежных и релевантных прогнозов этого процесса, особенно в нынешнем турбулентном мире. Однако устоявшиеся процедуры прогнозирования миграции имеют немало ограничений, на фоне которых открываются перспективы использования инновационных подходов, основанных на больших данных, в частности поисковых запросах потенциальных мигрантов в интернете. Объясняющие и предиктивные свойства подобного инструментария в силу его новизны пока еще мало раскрыты. Работа нацелена на изучение возможностей применения таких средств для предвидения потоков населения на постсоветском пространстве. Выдвинута гипотеза о наличии связи между онлайн-запросами по поводу миграции в Россию,

Для цитирования: Цапенко И.П., Юревич М.А. (2022). Статистика онлайн-запросов в наукастинге миграции // Экономические и социальные проблемы: факты, тенденции, прогноз. Т. 15. № 1. С. 74–89. DOI: 10.15838/esc.2022.1.79.4

For citation: Tsapenko I.P., Yurevich M.A. (2022). Nowcasting migration using statistics of online queries. *Economic and Social Changes: Facts, Trends, Forecast*, 15(1), 74–89. DOI: 10.15838/esc.2022.1.79.4

поступающими от жителей Киргизии, Таджикистана и Узбекистана, и последующими людскими потоками из указанных стран в РФ. Проверка гипотезы выполнена на материале миграционной статистики Росстата, данных Google Trends об интенсивности запросов и сервиса Яндекс «Подбор слов», используемого для валидации поисковых образов. Статистические взаимосвязи обнаружены с помощью корреляционного и регрессионного анализа. В результате установлена умеренная зависимость динамики людских потоков от изменений количества предшествующих запросов, которая проявляется с наибольшей силой при лаге в 6–9 месяцев и при нулевом лаге. Получению более точных результатов в этом и подобных исследованиях препятствует изначальная ограниченная предсказуемость миграционного поведения в силу его контекстуальности, подчас ситуативности и иррациональности, а также «зашумленность» статистики по запросам, а нередко – и потокам. В качестве магистрального направления исследований в данной области видится поиск универсальных алгоритмов определения связей между запросами и миграционными потоками.

Ключевые слова: миграция, прогнозирование, большие данные, онлайн-запросы, поисковые образы, моделирование, Россия, Центральная Азия.

Благодарность

Статья подготовлена и опубликована в рамках проекта «Посткризисное мироустройство: вызовы и технологии, конкуренция и сотрудничество» по гранту Министерства науки и высшего образования РФ на проведение крупных научных проектов по приоритетным направлениям научно-технологического развития (Соглашение № 075-15-2020-783).

Рост масштабов[1] и значимости трансграничной миграции населения (сопряженной с временной или постоянной сменой страны обычного проживания) в жизни обществ повышает потребность последних в релевантных, надежных и реалистичных прогнозах людских передвижений, особенно в условиях политической и экологической (в том числе эпидемической) нестабильности на планете. Такие прогнозы необходимы для дальновидного, проактивного, «умного» миграционного управления; повышения готовности систем быстрого реагирования к участившимся миграционным кризисам и разработки эффективных решений возникающих проблем; улучшения планирования деятельности разных сфер и институтов общества, например здравоохранения, образования и т. п.

В то же время прогнозирование миграции сталкивается с серьезными трудностями в силу специфики данного поведенческого феномена, а именно сопряженной с ним высокой неопределенности. Некоторые исследователи, в частности Я. Биек и М. Чайка, категорично утверждают, что «миграция не предсказуема в строгом смысле слова» (Bijak, Czaika, 2020, с. 14). Тем не менее, использование более широкого круга статистических источников, согласование разных видов информации, разработка методологии, соответствующей новым типам и комбинациям данных, сулят продвижение к получению релевантных результатов прогнозов.

Широкие возможности в сфере прогнозирования миграции открываются с появлением инновационных, основанных на больших данных подходов к изучению этого процесса. Весьма перспективным с точки зрения наукастинга и краткосрочного прогнозирования людских потоков представляется зарождающееся в мире направление миграционных исследований, которые опираются на статистику онлайн-запросов потенциальных мигрантов.

Подобный инструментарий анализа и прогнозирования миграции пока еще не получил развития в России. Цель нашей работы состоит в изучении возможностей применения новой методологии для предвидения миграционных потоков на постсоветском пространстве, на материале которого прежде подобные исследования не проводились. Сформулирована и проверяется гипотеза о наличии связи между

[1] По данным ООН, в 2020 году в мире число лиц, проживающих за пределами страны происхождения, достигло беспрецедентного уровня – 281 млн чел. (International Migrant Stock 2020. UNDESA, PD. POP/DB/MIG/Stock/Rev.2020).

поисковыми запросами по поводу миграции в Россию, осуществляемыми в интернете жителями Киргизии, Таджикистана и Узбекистана, и последующими миграционными потоками из указанных стран в РФ. Для достижения цели задействованы средства валидации поисковых образов, корреляционного, регрессионного и теоретического анализа.

Работа была сосредоточена на решении следующих задач: на основе обзора общепринятых методов прогнозирования и источников данных показать ограничения и проблемы в применении конвенциональных процедур; проанализировать зарубежный опыт использования разных видов больших данных, в том числе онлайн-запросов, для изучения и прогнозирования миграции, охарактеризовать их достоинства и недостатки; адаптировать подобный инструментарий к условиям миграции на постсоветском пространстве и апробировать методику предвидения миграции из Киргизии, Таджикистана и Узбекистана в Россию на статистическом материале онлайн-запросов населения стран Центральной Азии (ЦА) и помесячных данных Росстата о числе прибывших оттуда в Россию мигрантов; представить и попытаться объяснить полученные результаты; оценить возможности их практического использования и в целом перспективы данного подхода к прогнозированию миграции в российских условиях.

Ограничения устоявшихся процедур прогнозирования миграции

В XXI веке в мире отмечается резкая активизация прогностической деятельности в сфере миграции. Сформировалась большая группа ученых и институтов, занимающихся прогнозированием миграции. Стремительно растет число работ, в том числе крупных, посвященных количественным оценкам и форсайту будущей миграции, вопросам методологии и техники прогнозирования (см., например, (Szczepanikova, Van Criekinge, 2018; Acostamadiedo et al., 2020; Sohst et al., 2020; Carammia, Dumont, 2018; Bijak, 2016; Лифшиц, 2016; Ткаченко, Гиноян, 2018; Малышева, 2017) и др.[2]).

Однако при сопоставлении прогнозных оценок и реальных показателей миграции оказывается, что значительная часть прогнозов не сбылась. Зачастую динамика миграции недооценивалась. Австралийский демограф Т. Вилсон констатирует, что «неаккуратность» и большие величины ошибок фактически стали нормой в прогнозировании как долгосрочных потоков, так и краткосрочных колебаний (Wilson, 2017).

Главная причина промахов коренится в объективных трудностях прогнозирования миграции. В их числе – множественность взаимодействующих факторов миграции и изменчивость их влияния, особенно в условиях современной общественной и экологической (в том числе эпидемической) нестабильности, внешних шоков, глубоких, спрессованных во времени и многоплановых трансформаций, затрудняющих распознавание сигналов будущего[3]; сложная природа миграционного поведения, которое зачастую носит контекстуальный и ситуационный характер, несет печать места и времени при формировании миграционных намерений и принятии решения о переезде, подчас бывает иррациональным. Таким образом, прогнозирование миграции в настоящий момент имеет стохастический характер из-за объективной непредсказуемости «черных лебедей», присущих данному процессу, и большой вариативности конечных исходов моделируемого объекта. На повышение доли неопределенности также влияют ограниченность и неполнота информации, недостаток знаний о миграции как таковой (Bijak, Czaika, 2020).

Статистические данные, традиционно разрабатываемые и собираемые национальными и международными организациями (переписи и выборочные обследования населения, административные данные о въезде и выезде из страны, разрешениях на проживание, работу и т. п.), страдают такими существенными недостатками, как ресурсоемкость, неполнота, изъяны в качестве, существенная задержка в представлении, ограничения в доступности, сопоставимости и дезагрегации и др. Данные

[2] World Population Prospects (2019). Vol. I, 2. NY: UN; Tomorrow's World of Migration (2017). Geneva: FES, Global Future, IOM.

[3] Например, внезапный отток представителей обеспеченных социальных слоев населения из конкретных регионов может быть предвестником назревающих катаклизмов и последующих массовых переселений населения с пострадавших территорий.

социологических служб, в частности результаты опросов о миграционных намерениях населения, осуществляемых в рамках Gallup World Poll, доступны и сопоставимы на международном уровне. Однако такая информация затратна в получении, подвержена рискам нерепрезентативности выборки, зависимости их предиктовной силы от сроков и места проведения опроса, формулировки вопроса (Tjaden et al., 2019) и др.

При том что методология постоянно совершенствуется, ни один из используемых методов не может быть однозначно предпочтительным во всех отношениях (Sohst, Tjaden, 2020). Сценарии, обрисовывающие возможные варианты долгосрочных перспектив миграции, имеют повышенную неопределенность. А ввиду того, что горизонты таких прогнозов выходят за рамки электорального цикла, их результаты непросто транслировать в политические решения. Метод Delphi, несмотря на множественные раунды обследования отобранных групп экспертов, нередко не способен сгладить существенные разногласия в суждениях экспертов (Acostamadiedo et al., 2020).

Не лишены дефектов и эконометрические методы. Они уязвимы к ненадежности или неполноте статистики, национальным различиям в источниках и методах сбора данных, не могут учесть всего разнообразия объясняющих факторов, порой опираются на сомнительные допущения, неподходящие исторические или страновые аналогии, сталкиваются со сложностями при операционализации драйверов, содержащих элементы неопределенности. Кроме того, для обработки и анализа больших массивов информации требуется дополнение регрессионного инструментария другими более сложными методами машинного обучения, включая нейросети.

Турбулентность мира, резко усложняющая прогнозирование миграции, побуждает к поиску нестандартных источников информации и инновационных приемов ее применения. Возможности для этого открываются с развитием цифровых технологий. В частности, такие перспективы сулит использование больших данных и соответствующих им новых методологических подходов к наукастингу и прогнозированию миграции.

Зарубежный опыт использования больших данных в прогнозировании миграции

Альтернативные источники получения информации, принимающей вид больших данных, и новые способы их анализа, в том числе с помощью машинного обучения, все активнее задействуются в прогнозировании миграции (Sîrbu et al., 2021). Этому благоприятствует рост использования мигрантами мобильных телефонов и подключенных к интернету цифровых устройств в процессе планирования и осуществления миграции и т. п. Цифровой след, оставляемый мигрантами, может применяться для определения паттернов и тенденций миграционного поведения.

Большие данные обладают целым рядом несомненных достоинств: оперативность и своевременность, относительная простота и дешевизна получения информации, актуальность сведений, отражение реальных, текущих процессов, охват огромных массивов населения и территорий.

Выполнено немало исследований, основанных на использовании геолокализованных данных пользователей социальных сетей Twitter и Facebook, IP логинов при входе на веб-сайты, сообщений, посланных по электронной почте с серверов Yahoo, детализаций звонков с мобильных телефонов (*call detail record*). Эти работы продемонстрировали серьезный потенциал больших данных для прогнозирования масштабов и маршрутов передвижений населения, выявления типовых схем перемещений в чрезвычайных ситуациях, изменения численности мигрантов, степени их ассимиляции по языку, музыкальным и литературным предпочтениям пользователей сетей и т. п. (Zagheni et al., 2017; Zagheni, Weber, 2012; Hawelka et al., 2014). Применение таких данных позволило предсказать рост числа мигрантов из Венесуэлы в Колумбии и Испании (Spyratos et al., 2019), оценить культурную интеграцию выходцев из Мексики в США (Stewart et al., 2019). С их помощью удалось отследить людские потоки после стихийных бедствий на Гаити и предвидеть их маршруты в Новой Зеландии: в знакомые пострадавшим места (откуда прежде исходило много звонков) и крупные города (Bengtsson et al., 2011), а также контролировать мобильность в условиях эпидемий, включая COVID-19.

Полученные результаты согласовывались с официальными статистическими данными, опубликованными существенно позднее материалами исследований. В то же время применение подобных больших данных обнаруживало ограничения и порождало серьезные вызовы. На представительность выборки и соответственно точность результатов существенно влиял уровень проникновения интернета и мобильной связи. Различалась и интенсивность использования цифровых сервисов разными социальными группами в зависимости от их возрастных и гендерных характеристик, уровня социально-экономического развития территории, типа населенного пункта проживания и т. п. Такие особенности, как, например, молодость пользователей Twitter или принадлежность к старшим группам аудитории Facebook, ограничивают возможности генерализации подобных результатов для выявления общих миграционных паттернов. Сказывается и нестабильность пользования социальными сетями, ненадежность информации, подчас предоставляемой о себе пользователями, существование фейковых и двойных аккаунтов. Возникают проблемы с доступом к данным сетей и сервисов и обеспечением непрерывности потока информации. Из сетей трудно получить сведения о сроках пребывания приезжего в стране назначения, применяемых в официальной статистике в качестве критерия определения мигранта.

При использовании таких данных для отслеживания места нахождения индивида возникают риски нарушений прав человека в области конфиденциальности и безопасности персональной информации, а также этических норм. В худшем случае это может привести к возможному усилению репрессий в отношении подвергающихся преследованиям людей, созданию препятствий для их беженства за границу или же их высылке из страны искомого убежища и т. п. (Beduschi, 2018).

Большинства указанных недостатков лишены исследования миграции и других форм мобильности на основе статистики онлайн-запросов. Подобные изыскания приобретают все бо́льшую популярность. Поскольку Google является главным поисковиком населения планеты, который используют более 1 млрд человек, запросы в этом браузере, замеряемые с помощью Google Trends[4], можно считать в целом репрезентативными для интернет-аудитории и применять в качестве инструмента прогнозирования. В основе применения этого вида больших данных лежит представление о том, что агрегированная интенсивность запросов по связанным с миграцией поисковым словам служит непосредственным измерителем миграционных намерений (*direct* measure of migration intentions) (Bohme et al., 2020). Как показывают эмпирические исследования, потенциальные мигранты собирают информацию о возможностях переезда, в том числе в сети, поэтому колебания в числе запросов могут указывать на варьирование интереса к миграции и при прочих равных условиях подходят в качестве прокси (аналогового показателя) изменений в числе потенциальных мигрантов и привлекательности для них тех или иных стран, что позволяет применять такие данные для предсказания динамики людских потоков[5].

[4] Сервис Google Trends позволяет получить агрегированную статистику из базы данных запросов в Google по географическим областям, временным интервалам и др., отражающую коллективные поведенческие паттерны. Данные запросов в интернете, получаемые с помощью Google Trends, уже широко используются в разных областях экономических и социальных исследований: для прогнозирования совокупного спроса и частного потребления, уровня безработицы и инфляции, объема продаж конкретных товаров и услуг, распространения заболеваний, таких как грипп, сальмонеллез, ожирение, и др. (Юревич и др., 2020).

[5] Под миграционными намерениями могут пониматься абстрактное желание, конкретное планирование и реальная подготовка к миграции. Согласно исследованию на основе данных Gallup World Poll, лишь 1% взрослых жителей планеты хотели бы переехать в другую страну, из их числа только 10% строят планы отъезда; лишь треть планирующих мигрировать по-настоящему готовится к переезду и всего треть готовящихся реально выезжает (Tjaden et al., 2019). Соответственно, соотношение потенциальных и реальных мигрантов составляет 1 к 100. Хотя увеличению на 1 п. п. доли жителей определенной страны, выражающих интерес к эмиграции в конкретную страну, соответствует рост потока по данному маршруту на 0,75 п. п., подобные связи слабее для жителей развивающихся стран. Это объясняется наличием больших препятствий для их миграции: ограничительной иммиграционной политикой государств назначения, отсутствием ресурсов, дальностью расстояния и т. п. Претворению миграционных планов в жизнь могут помешать изменения жизненной ситуации и положения в стране, состояния здоровья, занятости, семейного статуса, возникновение непредвиденных расходов в связи с эмиграцией (Carling, 2017).

Начало прогнозированию трансграничных передвижений населения на основе подобного подхода было положено аналитиками Google в 2000-е годы. Х. Чой и Х. Вэриан, проанализировав помесячные показатели интенсивности поисковых запросов по слову «Гонконг», осуществлявшихся в определенных странах за 2005–2011 гг., и численности туристов, прибывающих из этих государств в Гонконг, пришли к выводу, что данные Google Trends отражают планирование поездок и позволяют предвидеть будущие туристические потоки (Choi, Varian, 2012). Использование помесячных данных, к которым была обращена названная пионерная работа, стало наиболее распространенной статистической практикой дальнейших исследований.

Изучение собственно международной миграции (критерием которой служит временная или постоянная смена страны обычного проживания) в таком инновационном ключе началось лишь в 2010-е годы. Для более четкой идентификации миграционных потоков и их отграничения от туризма и командировок поисковые образы строились по ключевым словам, связанным с работой, учебой, убежищем. Сотрудники «Глобального пульса»[6] и Фонда народонаселения ООН исходили из того, что люди, интересующиеся миграцией, делают перед отъездом запросы о возможностях занятости за границей. При сопоставлении статистики прибытий иностранцев в Австралию, в том числе ее отдельные города, с агрегированными запросами из разных стран за 2008–2013 гг. по ключевым англоязычным словам «рабочие места в Мельбурне», «работа в Австралии», «рабочая виза» – связь между этими показателями была установлена. Особенно тесной оказалась корреляция между потоками из Италии в Австралию и запросами «работа в Австралии»[7].

Важным результатом последующих исследований стало выявление временного лага между выражением и воплощением намерений. Подобный тайминговый эффект в миграции обусловлен необходимостью подготовки к поездке. Это показало исследование статистики испаноязычных запросов из Перу, Колумбии и Аргентины в 2005–2010 гг. по ключевым словам «работа в Испании», «посольство Испании» и «Испания». Удалось установить тесную связь с лагом в 7–8 месяцев таких запросов в Перу и Колумбии с числом выходцев из этих стран, зарегистрированных в качестве мигрантов в Испании. В отношении Аргентины, потоки из которой на 40–50% состояли из граждан европейских стран, результаты оказались неоднозначными (Wladyka, 2017).

Совпадение языка, используемого в странах происхождения и назначения мигрантов, облегчает работу с поисковыми запросами. Однако более типичны лингвистические расхождения. Исследование миграции в Швейцарию опиралось на поисковые данные на четырех языках: немецком, французском, итальянском и испанском. Были обработаны запросы по ключевому слову «работа в Швейцарии», сделанные соответственно в Германии, Франции, Италии и Испании в 2004–2018 гг. Установлено, что среди потенциальных мигрантов из четырех указанных стран уроженцы Испании и Италии проявляют наибольший интерес к работе в Швейцарии и демонстрируют самую тесную связь между запросами и последующим приездом, что позволяет прогнозировать в краткосрочном плане потоки взрослых мигрантов из этих государств. Напротив, соответствующие показатели по Франции обнаруживают слабую связь, проявляющуюся с лагом в два года. Этот факт объясняется превалированием в передвижениях из Франции семейной иммиграции, а также наличием в потоках большого числа мигрантов, уже работавших прежде в Швейцарии, то есть не нуждающихся в поиске такой информации (Wanner, 2021).

Предприняты первые шаги к краткосрочному прогнозированию вынужденной миграции на основе поисковых запросов. Исследование Центра Pew показало тесную связь между интенсивностью запросов с территории Турции со словом «Греция» на арабском языке с колебаниями в числе иракских и сирийских беженцев, пересекавших Эгейское море в сторону Греции в 2015–2016 гг. (Connor, 2017).

[6] Инициатива ООН по использованию больших данных для прогнозирования в режиме реального времени.

[7] Estimating Migration Flows Using Online Search Data (2014). Global Pulse Project Series. No. 4. Available at: http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Search_Migration_2014_0.pdf

По сравнению со всеми вышеупомянутыми исследованиями, ограничивавшимися использованием небольшого числа поисковых слов применительно к отдельным странам, работу немецкого ученого М. Боме и его коллег отличает масштабность и новый алгоритм анализа. Для построения поисковых образов авторы отобрали 67 слов, относящихся к экономике и миграции, на трех языках: английском, французском и испанском, и собрали статистику соответствующих запросов, сделанных в 101 стране происхождения мигрантов в отношении 35 принимающих стран ОЭСР. Выстроенные временные ряды агрегированных годовых данных за 2004–2015 гг. по каждому поисковому слову из каждой отдающей страны позволяли судить об уровне и динамике эмиграционного потенциала населения, а также ориентации эмиграционных намерений на конкретные страны назначения. Значимость результатов повышалась при ограничении выборки государствами с высоким уровнем проникновения интернета и более распространенными языками (Bohme et al., 2020).

Приведенные результаты исследований свидетельствуют, что данные запросов, связанных с миграцией, могут применяться как прокси миграционных намерений и дополнять официальную информацию, компенсируя недостаток актуальной и сопоставимой статистики и позволяя использовать новые комбинации данных, повышающие ценность каждого их типа (Struijs et al., 2014). При этом включение в такие поисковые образы слов, связанных с учебой, убежищем и т. п., может высвечивать намерения в отношении учебной, гуманитарной миграции и т. д. соответственно.

В то же время использование подобных альтернативных источников информации сопряжено с определенными ограничениями. Огромный объем, многосложность и «зашумленность» данных порождают проблемы методологического и аналитического плана (Rango, 2015). Отсутствует единый универсальный подход к применению запросов для прогнозирования миграции. Существуют различия в тесноте связи статистики типовых запросов с последующими передвижениями населения в зависимости от маршрутов, массовости и состава потоков. Возможно, каждому миграционному коридору (миграции между двумя конкретными странами) присущ уникальный запрос (Tjaden et al., 2021). Данные запросов нередко мало информативны в случаях немассовых потоков, передвижений из стран с редкими языками, ограниченным доступом интернету и т. п. (Wladyka, 2017). Искажения могут возникать из-за того, что люди вкладывают разный смысл в одни и те же слова и используют их в разных, в том числе не связанных с миграцией, поисковых целях. Кроме того, при проверке точности прогнозных оценок, сделанных на основе альтернативных источников информации, путем их соотнесения с официальной миграционной статистикой, необходимо принимать во внимание несовершенство последней (Tjaden et al., 2021).

Очевидно, что пока еще малая изученность объясняющих и предиктивных свойств статистики запросов осложняет корректное использование больших данных для прогнозирования миграции, что, в свою очередь, порождает необходимость дальнейших изысканий в этой области.

Методология исследования и калибровка переменных

Выбор для исследования миграционных потоков из Киргизии, Таджикистана и Узбекистана в Россию обусловлен массовостью и устойчивостью маршрутов таких передвижений, порождаемых социально-экономическими различиями и миграционной взаимозависимостью этих стран. Передвижениям благоприятствуют общность исторического прошлого, географическая близость, экономическая интеграция, культурные связи и др.

Указанные государства ЦА относятся к числу основных доноров населения и рабочей силы на постсоветском пространстве. В 2020 году они обеспечили 32% притока мигрантов в РФ и 43% миграционного прироста ее населения; в свою очередь на Россию как главного реципиента людских потоков в 2020 году приходилось 76% всех проживавших за рубежом уроженцев Киргизии (в 2000 г. 81%), 79% – Таджикистана (70%) и 57% – Узбекистана (58%)[8]. Приведенные показатели свидетельствуют о том, что Таджикистан – единственная в данной группе страна, которая четко демонстрирует долго-

[8] Рассчитано по: Численность и миграция населения Российской Федерации в 2020 году (2021). М.: Росстат; International Migrant Stock (2020). UNDESA, PD. POP/DB/MIG/Stock/Rev.2020

срочное, наименее подверженное негативным внешним влияниям усиление ориентации миграционных потоков на Россию.

При определении наиболее подходящей для исследования референтной статистической информации о миграционных потоках рассматривались два типа данных из разных источников. Во-первых, это квартальные данные МВД России о количестве фактов постановки на миграционный учет иностранных граждан, в том числе прибывших с целью работы. Позволяя четко выделить трудовых мигрантов, эти сведения, имеющиеся в открытом доступе лишь за период с конца 2016 года, не обеспечивают достаточной глубины временного ряда.

Во-вторых, помесячные данные Росстата, доступные с 2011 года, о числе прибывающих из-за границы в Россию и регистрируемых по месту жительства или пребывания на срок от 9 месяцев. Однако эти сведения суммарно охватывают не только трудовых мигрантов (не позволяя их выделить), но и студентов, воссоединяющихся членов семей и др. В то же время, учитывая, что современная миграция – в первую очередь миграция рабочей силы и работой интересуются (ищут о ней информацию в сети) не только трудовые, но и другие категории мигрантов, указанными и некоторыми другими недостатками (Чудиновских, Степанова, 2020) за неимением лучших данных можно пренебречь.

При этом сведения Росстата обладают несомненными достоинствами: 1) внушительная глубина временного ряда одновременно с помесячной детализацией данных, обеспечивающая большое число наблюдений и позволяющая выстроить более рельефную траекторию развития миграционных процессов по сравнению с квартальными, более сглаженными данными; 2) соответствие наиболее распространенному формату данных, используемых за рубежом в подобных исследованиях; 3) отражение более устойчивых потоков мигрантов, приезжающих временно на более длительный срок или на постоянное место жительства, в отличие от краткосрочных волатильных циркуляций.

Для последующего количественного анализа были сформированы временны*е* ряды данных миграционного притока из указанных государств ЦА в Россию за период с января 2015 по декабрь 2020 года, включавшие 72 точки (наблюдения) по каждой стране (обозначение показателя прибытий мигрантов – *M*).

В ходе формирования статистики запросов потенциальных мигрантов был составлен перечень возможных поисковых слов или поисковых образов. При формировании списка авторы учли опыт зарубежных исследований, в которых наилучшие результаты показывали поисковые образы, связанные с работой на конкретных территориях.

Для проверки релевантности и точности соответствия изучаемому объекту выбранных поисковых образов использовался сервис «Подбор слов», предоставляемый российской компанией Яндекс. Отечественное web-приложение позволяет получить информацию об абсолютном количестве запросов, включающих поисковый термин, а также контекст его включения (Юревич, 2021). Тогда как американский аналог Google Trends выявляет лишь динамику запросов, но не показывает их абсолютное количество, соответственно, он менее информативен для валидации поисковых образов – определения частоты употребления конкретных терминов в сети и их ассоциативных связей, что затрудняет доказательство истинных взаимосвязей и повышает вероятность обнаружения ложных зависимостей.

В то же время сервис Яндекса не позволяет проводить анализ запросов за большой отрезок времени, ограничивая его всего двумя годами доступных данных, в связи с чем для построения моделей анализа временных рядов в среднесрочном и долгосрочном периодах использовалось приложение Google Trends, гораздо лучше подходящее для этого в силу глубокой ретроспективы данных.

Правомерность обращения к двум указанным поисковым системам основывается на их популярности среди пользователей ЦА, хотя Яндекс заметно уступает Google. По сведениям аналитической службы StatCounter, в 2020 году в Киргизии на долю Google приходилось 89% рынка браузеров, а на Яндекс – 10%, в Таджикистане – соответственно 82 и 15%, в Узбекистане – 84 и 14%[9]. Географическая привязка запросов определялась по IP-адресу пользователя или с помощью соответствующих настроек браузера.

[9] Statcounter GlobalStats. Available at: https://gs.statcounter.com

Согласно данным сервиса «Подбор слов», поисковые запросы, содержащие слово «работа», весьма популярны *(табл. 1)*. Термин «вакансии» используется несколько реже, но тоже довольно распространен. Запросы более специфических слов вроде «жительство» (например, запрос «получить вид на жительство») и «миграция» имеют частоту менее 1000 в месяц. Слово «патент» применяется несколько чаще, но контекст его использования указывает на сильную «зашумленность» термина, его частую ассоциацию с документом, охраняющим интеллектуальную собственность. В то же время запросы, содержащие слова «работа» или «вакансии», по аналогичной причине не подходят для наукастинга или прогнозирования миграционных потоков: во-первых, с помощью этих слов люди часто ищут работу в своей стране, не интересуясь возможностями трудоустройства за рубежом; во-вторых, в интернете встречается много упоминаний слова в образовательном контексте – как «домашняя» и «классная» работа. Не результативна и попытка построить поисковый образ с использованием слов на национальных языках стран ЦА. Видимо, жители ищут с помощью таких запросов работу на местных рынках, причем их усилия не слишком активны. Например, в Узбекистане запрос «ish o rinlari» (рабочие места) имеет 1,3 тыс. упоминаний в месяц; в Таджикистане «кор» (работа) – 5,5 тыс.; в Киргизии «жумуш» (работа) – 2,6 тыс.

Добавление в поиск названия страны ЦА позволяет уточнить зону поиска и характеристики рабочих мест, однако приводит к слиянию в общей статистике запросов типа «работа в Таджикистане» и «работа для граждан Таджикистана». И наоборот, при использовании названий этих стран в качестве стоп-слов из поля наблюдения выпадают все запросы, содержащие слова «Таджикистан», и т. п.

Напротив, включение в запрос названий принимающей страны, ее городов и т. п. оправдывает себя. Высокий уровень соответствия поставленной задаче демонстрирует поисковый образ «(работа ИЛИ вакансии) И (Москва ИЛИ Россия)», который в силу конкретизации географии поиска почти не зашумлен. В то же время другие города и субъекты РФ, согласно сервису «Подбор слов», привлекают гораздо меньше интереса граждан каждой из трех стран.

Составленный поисковый образ запросов «(работа ИЛИ вакансии) И (Москва ИЛИ Россия)» был введен в сервис Google Trends для формирования длительного временного ряда (обозначение показателя – GT). Полученная статистика автоматически нормализована по отношению к максимальному значению за рассматриваемый период с января 2014 по декабрь 2020 года, т. е. показатель приобрел форму индекса. Для наглядности временные ряды были нормализированы от 0 до 1 по процедуре МИНМАКС ($y = (x - min)/(max - min)$).

Таблица 1. Характеристика популярности поисковых терминов, связанных с потенциальной трудовой миграцией, в месяц

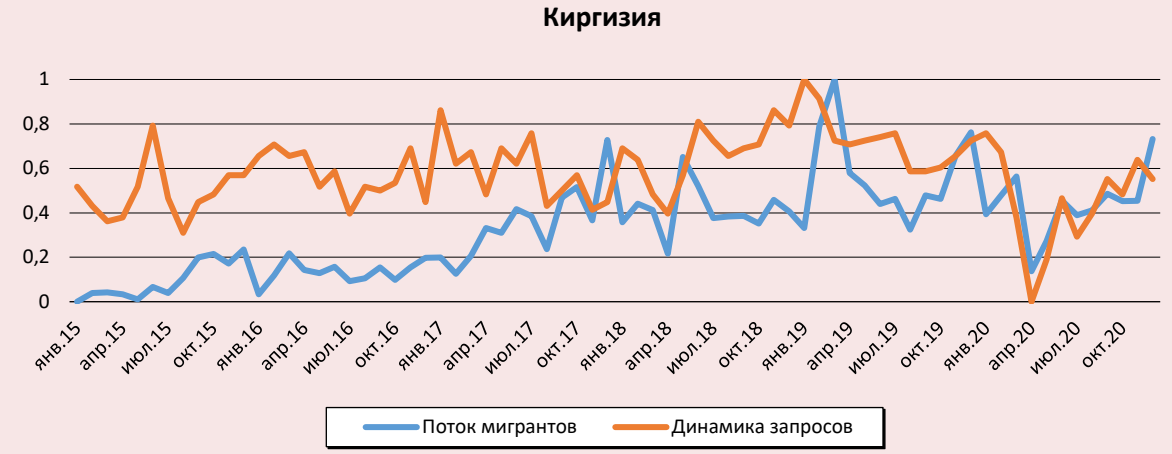
Поисковый запрос с учетом морфологии	Количество запросов в месяц, ед.		
	Киргизия	Таджикистан	Узбекистан
работа	63945	41460	67731
вакансии	11686	7599	12835
работа|вакансии	72387	46901	77481
работа|вакансии - Бишкек - Киргизия - домашняя	59937		
работа|вакансии - Душанбе - Сомон*- Таджикистан		41379	
работа|вакансии - Ташкент - Узбекистан - Самарканд			62832
работа|вакансии Москва|Россия	6842	6400	4537
патент	1118	3032	1741

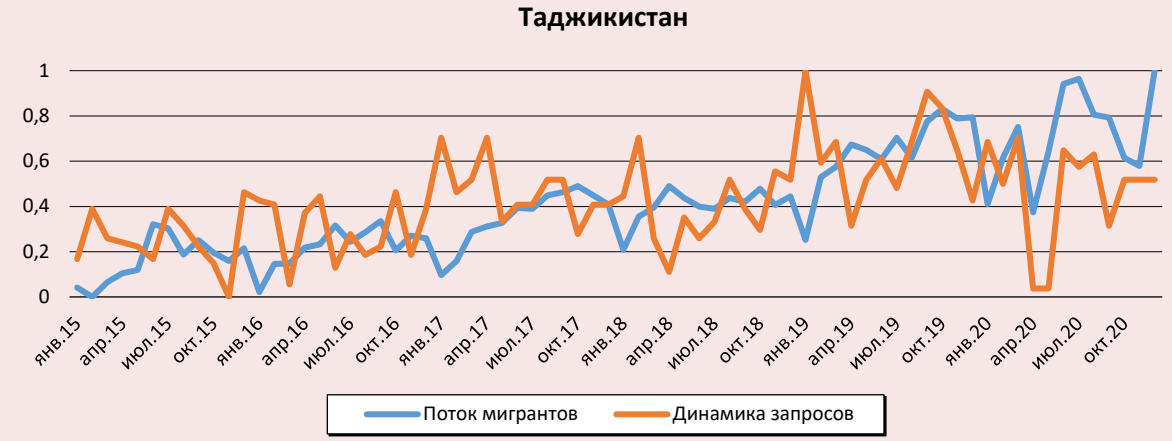
* https://somon.tj – аналог авито, вакансии только в Таджикистане.

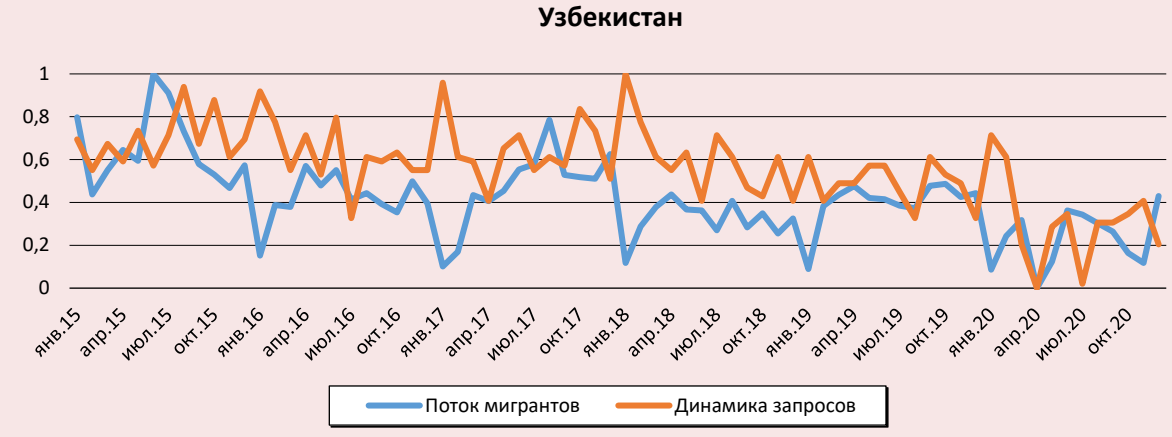
Примечание. Символ «|» выполняет функцию оператора «или»; символ «-» исключает последующее слово из возможных вариантов запроса. В разряд исключенных попали те слова, которые, очевидно, не соответствуют целям идентификации миграционных настроений.

Источник: https://wordstat.yandex.ru/ (дата обращения 19.10.2021).

Сравнение динамики притока мигрантов и запросов «(работа ИЛИ вакансии) И (Москва ИЛИ Россия)» в поисковой системе Google в 2015–2020 гг.







Источник: расчеты авторов.

Диаграммы *(рисунок)* свидетельствуют в пользу корреляции между рассматриваемыми переменными, хотя и не очень тесной. Ломаная траектория динамики потоков и особенно запросов, колеблющихся с большой амплитудой, отражает нестабильность отечественной экономики в условиях санкций западных государств в отношении России, пандемии т. п. При этом хорошо заметны синхронные реакции переменных на знаковые, особенно экстраординарные, события и экстремальные ситуации. Так, рост запросов и потоков с конца лета 2019 года может быть связан с принятием закона № 257-ФЗ, упрощающего порядок предоставления разрешения на временное проживание и вида на жительство некоторым категориям мигрантов, и более широким психологическим эффектом этой либерализационной меры. Напротив, ввод в России пакета жестких мер против распространения коронавирусной эпидемии спровоцировал стремительное сокращение притока мигрантов и свел к минимуму количество запросов весной 2020 года. Однако с появлением в российском законодательстве положения о возможности неоднократного продления, начиная с мая 2020 года, патентов, а также принятием пакета либеральных правил пребывания в стране мигрантов с июня почти параллельно росли как запросы, так и потоки.

Изменение показателей в противоположных направлениях наблюдается во всех трех странах год от года в январе, когда поток миграции сильно снижается, но интерес к работе на территории РФ заметно возрастает. Это говорит о вероятном присутствии сезонности в обеих переменных.

Однако траектория запросов, в отличие от движения потоков, не демонстрирует подъем в летний период спроса российской экономики на иностранный труд, что объяснимо в типовых условиях наличием лага между запросами и последующими потоками. Также заметно, что потоки из Киргизии с большим отставанием отреагировали на вступление 1 января 2015 года страны в ЕАЭС (членство в котором, похоже, не сказывается в каких-либо явных особенностях сетевого и миграционного поведения граждан), а передвижения из Таджикистана и Узбекистана – на введение с 1 января 2015 года патентов для работы мигрантов из СНГ у юридических лиц.

Кроме того, как видно на графиках, приток мигрантов из Киргизии и особенно Таджикистана имеет выраженный тренд к росту. К повышательному тренду тяготеют и темпы изменения запросов из этих стран. Напротив, динамика запросов, а также приездов мигрантов из Узбекистана демонстрирует тенденцию к ослаблению. Вероятно, в силу инерции затухающей динамики потоков вступление в силу в декабре 2017 года соглашения между правительствами Узбекистана и России об организованном наборе граждан Узбекистана для осуществления временной трудовой деятельности в РФ, вызвав всплеск запросов, не отразилось в последующем подъеме миграции.

Проведенные статистические тесты подтвердили эти гипотезы. Расширенный тест Дики-Фуллера (ADF-тест) и Q-тест Льюнга-Бокса указали на нестационарность временных рядов (расчеты выполнены в приложении RStudio, пакет «tseries»). Кроме того, комплексный тест на сезонность, включающий тесты на сезонные дамми-переменные, тесты Фридмана, Краскела-Уоллиса и др.[10], свидетельствует о наличии сезонности в переменной M по Узбекистану и в переменных GT по всем странам (расчеты выполнены в приложении RStudio, пакет «seastests»). Одним из способов получения стационарности временных рядов является взятие первых разностей, т. е. в дальнейшем будет анализироваться месячное изменение числа прибывших мигрантов и изменение индекса запросов (ΔM и ΔGT). Сглаживание сезонных колебаний в тех переменных, в которых обнаружена сезонность, выполнено при помощи алгоритма X-13ARIMA-SEATS[11]. После этой операции с целью первичного анализа зависимости между переменными был проведен корреляционный анализ с включением лагов от 0 до 12 месяцев. На этом этапе обнаружена крайне низкая степень взаимосвязи между наблюдаемыми переменными по Таджикистану. Несколько лучшие результаты зафиксированы при изменении поискового образа путем удаления слова «вакансии».

[10] Package “seastests”. Available at: https://www.rdocumentation.org/packages/seastests/versions/0.14.2/topics/isSeasonal

[11] R-interface to X-13ARIMA-SEATS. Available at: http://www.seasonal.website

Корреляционный анализ показал умеренную взаимосвязь между переменными. При этом наиболее плотная взаимозависимость наблюдается с лагом в 6–9 месяцев, требующимся для принятия решения и основательной подготовки к отъезду на продолжительный срок, а также при нулевом лаге, что указывает на поиск свежей информации о работе в России непосредственно перед отбытием. Полученные результаты согласуются с похожими выводами некоторых зарубежных работ (Wladyka, 2017; Wanner, 2021), отнюдь не опровергая гипотезы исследования.

Результаты моделирования и их обсуждение

Исследование коэффициентов корреляции продемонстрировало примерно идентичную силу взаимосвязи переменных с разными лагами в рамках одной страны. Но с учетом достаточно умеренных значений этих коэффициентов и с целью повышения общей стабильности моделей в число объясняющих регрессоров был введен показатель месячного притока мигрантов (*M*) также с лагами. Поиск оптимальных спецификаций моделей по каждой из стран выполнен по алгоритму пошаговой регрессии, определяющими параметрами стали величина информационного критерия Акаике и значимость коэффициентов при объясняющих переменных (расчеты выполнены в приложении RStudio, пакет «MASS»). Дополнительно был реализован алгоритм обратного исключения переменных (leapBackward, пакет «caret»), который, помимо прочего, учитывает точность прогнозов. Итоговый набор спецификаций определен исходя из теоретических предпосылок о знаках при коэффициентах; также приняты во внимание величина средней абсолютной ошибки (MAE), значимость коэффициентов и другие показатели качества моделей *(табл. 2)*.

Полученные модели создают предпосылки для подтверждения основной гипотезы исследования: динамика прибытий мигрантов демонстрирует положительную зависимость от изменений в количестве запросов, связанных с поиском жителями ЦА работы на территории РФ. При этом более сильная связь характерна для Узбекистана и Киргизии, в случае Таджикистана она наименее прочная. Величина лагов между зависимой и объясняющими переменными говорит о том, что потенциальные мигранты заранее готовятся к переезду и заблаговременно ищут места трудоустройства.

Таблица 2. Модели прогнозирования притока мигрантов

Страна	Киргизия		Таджикистан		Узбекистан
Спецификация	C1	C2	C3	C4	C5
Константа	930.11** (361.96)	885.60** (345.57)	1117.78** (499.29)	181.12 (462.52)	1477.41*** (516.70)
ΔGT (–7)	41.92** (16.99)		19.57[1] (12.15)	19.55* (10.17)	
ΔGT (–11)		59.71*** (17.08)			18.05** (7.65)
M (–1)	-0.59*** (0.11)	-0.64*** (0.11)	-0.17** (0.08)	-0. 56*** (0.10)	-0.17** (0.07)
M (–6)				0.19* (0.10)	-0.14* (0.07)
M (–9)	0.37*** (0.10)	0.42*** (0.10)		0.41** (0.12)	
Количество наблюдений	59	59	59	59	59
Нормированный R-квадрат	0.37	0.43	0.10	0.37	–
R-квадрат	–	–	–	–	0.20
MAE	477.6	446.2	609.4	539.8	–

[1] P < 0.12

Уровень значимости в соответствии с t-критерием Стьюдента: *p < 0.10; **p < 0.05; ***p < 0.01; в скобках указана величина стандартной ошибки; в первом столбце при переменных в скобках указана величина лага.

Источник: расчеты авторов.

В то же время построенные модели имеют относительно небольшой процент объясненной дисперсии и умеренную среднюю абсолютную ошибку. Причиной этого может служить несколько обстоятельств. Во-первых, объединение в составе зависимой переменной данных не только о трудовых, но и прочих категориях мигрантов изначально понижает точность модели прогнозирования миграции рабочей силы, поскольку запросам разных категорий мигрантов могут быть релевантны разные поисковые образы.

Во-вторых, немалая доля мигрантов в составе потоков из Киргизии, Таджикистана и Узбекистана приходится на граждан России и других государств: порядка 23, 33 и 28% соответственно среди прибывших в 2018 году из указанных стран[12]. У таких мигрантов могут быть иные модели поискового поведения в сети, и их повышенный процент в потоке из Таджикистана, вероятно, ослабил связь между конкретными запросами и потоками, что согласуется с полученными за рубежом результатами.

В-третьих, как известно из зарубежных исследований, невысокий уровень проникновения интернета может искажать связи между запросами и потоками. Согласно материалам, представленным на сайте DataReportal, в начале 2021 года аудитория интернета включала 50,4% всех жителей Киргизстана, 34,5% – Таджикистана, 55,2% – Узбекистана, что ниже среднемирового показателя[13]. Хотя жители ЦА в возрасте 20–34 лет составляют повышенную долю как среди прибывающих в Россию, так и среди интернет-пользователей, не исключена деформация представительности последних по отношению к потенциальным мигрантам. Самый низкий уровень проникновения интернета в Таджикистане по сравнению с двумя другими странами ЦА может сказываться в наименее прочной связи между запросами и потоками населения этого государства.

В-четвертых, учитывая давнюю историю и повторяемость подобных поездок, потребность в поиске информации о миграции в сети снижается у уже побывавших в России мигрантов, равно как и получающих от них сведения соотечественников.

В-пятых, как следует из иностранного опыта, статистика запросов мигрантов из стран с редкими языками демонстрирует более слабые связи с потоками из этих государств. Языки ЦА также не относятся к распространенным, и население региона ищет информацию о России на русском языке. Если учесть, что русский не родной язык для мигрантов и они зачастую владеют им лишь поверхностно при общем невысоком уровне образования, исследователям трудно представить, какими словами мигрант сформулировал запрос. В 2020 году высшее и среднее профессиональное образование имели только 14% мигрантов старше 14 лет из Киргизии, 18% – из Таджикистана и 25% – из Узбекистана[14]. Тем не менее, вероятно, более скрупулезная валидация поискового образа в виде включения стоп-слов могла бы усилить его предиктивную способность.

В-шестых, возможно, при использовании продвинутых и ресурсоемких алгоритмов машинного обучения, в частности нейронных сетей (Blazquez, Domenech, 2018), нередко применяемых для сложного анализа больших массивов данных, удалось бы добиться более точных результатов. Зарубежные миграционные ведомства уже задействуют такие методы в прогностических целях[15]. Тем не менее продвинутые методы искусственного интеллекта из-за их высокой технической ресурсоемкости и необходимости значительного количества времени на процедуру обучения и регуляризации порой оказываются менее подходящими, чем регрессионные подходы, для выявления наличия и характера причинно-следственных связей, которые удалось установить в исследовании.

Более того, полученные величины средней абсолютной ошибки относительно невелики применительно к прогнозируемому помесячному изменению объема потока. Так, если соот-

[12] Рассчитано по: Демографический ежегодник России 2019 (2020). М.: Росстат.

[13] Posts Tagged Central Asia. Available at: https://datareportal.com/reports/?tag=Central+Asia

[14] Рассчитано по: Численность и миграция населения Российской Федерации в 2020 году (2021). М.: Росстат.

[15] В 2012 году Европейское бюро по вопросам предоставления убежища запустило систему раннего предупреждения и готовности, использующую механизмы обмена информацией из разных источников, в том числе больших данных, обрабатываемых с помощью машинного обучения (Albertinelli et al., 2020).

нести их со средней общей численностью прибывших за месяц мигрантов, то погрешность будет находиться в пределах 10% для Таджикистана и Узбекистана и 15% для Киргизии. С учетом «зашумленности» зависимой переменной подобные результаты представляются вполне приемлемыми.

Выводы

В ходе исследования установлена умеренная взаимосвязь между изменением количества онлайн-запросов по поводу миграции в РФ, поступающих от жителей Киргизии, Таджикистана и Узбекистана, и динамикой последующих миграционных потоков оттуда в Россию. Выделены контекстные факторы, которые могли сказаться на релевантности поисковых образов и точности оценок тесноты взаимосвязей: состав мигрантов по категориям, гражданству и уровню образования, знание ими русского языка, уровень проникновения интернета в странах ЦА. Полученные результаты свидетельствуют, что статистика запросов может использоваться в качестве предиктора миграции в Россию из стран ЦА, особенно Таджикистана и Узбекистана. Преимущества новых способов получения информации о мигрантах на основе чтения их «цифрового следа» открывают перспективы интеграции альтернативных данных в изучение и прогнозирование миграционных процессов и использования таких данных в миграционной политике российского государства.

Проведенное исследование может пополнить опыт подобного прогнозирования миграции методикой применения возможностей Яндекс для валидации поисковых образов, а также полученными результатами на материалах ранее не изучавшегося в этом ключе постсоветского региона. В то же время для получения более надежных результатов требуется дальнейшее развитие методологии и выработка общих подходов к проведению подобных исследований с учетом широкого контекста при выборе поисковых слов и языка запросов, а также при определении связи последних с людскими потоками.

Литература

Лифшиц М.Л. (2016). Прогнозирование мировой миграционной ситуации на основе анализа нетто-миграции в странах мира // Прикладная эконометрика. Т. 41. С. 96–122.

Малышева Д.Б. (2017). Миграционные процессы в странах Центральной Азии // Постсоветские государства: 25 лет независимого развития / отв. ред. А.Б. Крылов. Т. 1. М.: ИМЭМО РАН. С. 160–171.

Ткаченко А.А., Гиноян А.Б. (2018). Оценка миграционного потенциала стран СНГ на основе модели международной миграции // Вопросы статистики. № 25 (11). С. 46–56.

Чудиновских О.С., Степанова А.В. (2020). О качестве федерального статистического наблюдения за миграционными процессами // Демографическое обозрение. Т. 7. № 1. С. 54–82.

Юревич М.А., Екимова Н.А., Балацкий Е.В. (2020). Цифровая трансформация экономической науки // Информационное общество. № 2. С. 39–47.

Юревич М.А. (2021). Инфляционные ожидания и инфляция: наукастинг и прогнозирование // Journal of Economic Regulation. Т. 12. № 2. С. 22–35.

Acostamadiedo E. et al. (2020). *Assessing Immigration Scenarios for the European Union in 2030 – Relevant, Realistic and Reliable?* Geneva: IOM and e Hague: NIDI.

Albertinelli A. et al. (2020). Forecasting asylum-related migration to the European Union, and bridging the gap between evidence and policy. *Migration Policy Practice,* X(4), 35–41.

Beduschi A. (2018). The big data of international migration: Opportunities and challenges for states under international human rights law. *Georgetown Journal of International Law,* 49, 982–1017.

Bengtsson L. et al. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A postearthquake geospatial study in Haiti. *PLoS Med*, 8(8), e1001083.

Bijak J. (2016). Migration forecasting: Beyond the limits of uncertainty. *IOM's GMDAC Data Briefing Series*, 6, 7. Available at: gmdac.iom.int/gmdac-databriefing-migration-forecasting-beyondlimits-uncertainty

Bijak J., Czaika M. (2020). *Assessing uncertain migration futures: A typology of the unknown*. QuantMig Project Deliverable D1.1. University of Southampton and Danube University Krems. Available at https://www.quantmig.eu/res/files/QuantMig%20D1.1%20Uncertain%20Migration%20Futures%20V1.1%2030Jun2020.pdf

Bijak J., Czaika M. (2020). Black swans and grey rhinos: Migration policy under uncertainty. *Migration Policy Practice,* 2020, X(4), 14–18. Available at: https://publications.iom.int/books/migration-policy-practice-vol-x-number-4-september-december-2020

Blazquez D., Domenech J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change,* 130, 99–113.

Bohme M. et al. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142, 14. DOI:10.1016/j.jdeveco.2019.04.002

Carammia M., Dumont J. (2018) Can we anticipate future migration flows? *OECD/EASO Migration Policy Debate*, 16, 9.

Carling J. (2017). How does migration arise? In: M. McAuliffe and M. Klein Solomon (Conveners) *Ideas to Inform International Cooperation on Safe, Orderly and Regular Migration*. Geneva: IOM, 19–26.

Choi H., Varian H. (2012). Predicting the present with Google trends. Predicting. *The Economic Record*, 88 (June), 2–9. DOI: 10.1111/j.1475-4932.2012.00809.x

Connor P. (2017). *The Digital Footprint of Europe's Refugees*. Pew Research Center. Available at: https://www.pewresearch.org/global/wp-content/uploads/sites/2/2017/06/Pew-Research-Center_Digital-Footprint-of-Europes-Refugees_Full-Report_06.08.2017.pdf

Hawelka B. et al. (2014). Geo-located Twitter was proxy for global mobility patterns. *Cartography and Geographic Information Science,* 41(3), 260–271.

Rango M. (2015). How big data can help migrants, *World Economic Forum*, 2 (October 5, 2015), Available at: https://www.weforum.org/agenda/2015/10/how-big-data-can-help-migrants/

Sîrbu A. et al. (2021). Human migration: The big data perspective. *International Journal of Data Science and Analytics*, 11, 341–360. DOI: 10.1007/s41060-020-00213-5

Sohst R., et al. (2020). *The Future of Migration to Europe: A Systematic Review of the Literature on Migration Scenarios and Forecasts*. Geneva: IOM and Hague: NIDI.

Sohst R., Tjaden J. (2020). Forecasting migration: A policy guide to common approaches and models. *Migration Policy Practice,* 4, 8–13.

Spyratos S. et al. (2019). Quantifying international human mobility patterns using Facebook Network data. *PLoS One*, 14(10), e0224134. https://doi.org/10.1371/journal.pone.0224134

Stewart I. et al. (2019). Rock, rap, or reggaeton? Assessing mexican immigrants' cultural assimilation using Facebook data. In: *WWW '19*. NY: Association for Computing Machinery, 3258–3264. DOI: 10.1145/3308558.3313409

Struijs P. et al. (2014). Official statistics and big data. *Big Data &* Society, April–June, 1–6. DOI: 10.1177/2053951714538417

Szczepanikova A., Van Criekinge T. (2018). *The Future of Migration in the European Union: Future Scenarios and Tools to Stimulate Forward-Looking Discussions*. Luxembourg: Publications Office of the European Union. DOI: 10.2760/000622

Tjaden J. et al. (2021). *Tale of high expectations, promising results and a long road ahead*. Available at: https://medium.com/@UNmigration/using-big-data-to-forecast-migration-8c8e64703559

Tjaden J., Auer D., Laczko F. (2019). Linking Migration Intentions with flows: Evidence and potential use. *International Migration*, 57(1), 36–57. DOI: 10.1111/imig.12502

Wanner P. (2021). How well can we estimate immigration trends using Google data? *Quality & Quantity*, 55, 1181–1202. DOI: 10.1007/s11135-020-01047-w

Wilson T. (2017). Can international migration forecasting be improved? The case of Australia. *Migration Letters*, 14(2), 285–299. DOI: 10.33182/ml.v14i2.333

Wladyka D. (2017). Queries to google search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies,* 2017, 25(4), 312–327. DOI: 10.25133/JPSSv25n4.002

Zagheni E., Weber I., Gummadi K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43, 721–734. Available at: https://doi.org/10.1111/padr.12102

Zagheni E., Weber I. (2012). You are where you e-mail: Using e-mail data to estimate international migration rates. In: *WebSci '12: Proceedings of the 4th Annual ACM Web Science Conference*. NY: Association for Computing Machinery, 348–351. DOI: 10.1145/2380718.2380764

Сведения об авторах

Ирина Павловна Цапенко – доктор экономических наук, заведующий сектором, Национальный исследовательский институт мировой экономики и международных отношений имени Е.М. Примакова Российской академии наук (117997, Российская Федерация, г. Москва, ул. Профсоюзная, д. 23; e-mail: tsapenko@bk.ru)

Максим Андреевич Юревич – научный сотрудник, Финансовый университет при Правительстве РФ (109456, Российская Федерация, г. Москва, 4-й Вешняковский проезд, д. 4; e-mail: maksjuve@gmail.com)

Tsapenko I.P., Yurevich M.A.

Nowcasting Migration Using Statistics of Online Queries

Abstract. Due to international migration's growing importance in modern countries' lives, there is an increasing need for reliable and relevant forecasts of this process, especially in today's turbulent world. However, established migration forecasting procedures suffer from a number of limitations, against which innovative approaches based on big data, notably online searches made by potential migrants, offer many advantages. Because of their novelty, such tools have not yet revealed their full explanatory and predictive properties. The work explores the possibility of using these tools to predict the population flows within the post-Soviet space. We hypothesize that there is a there is a statistical relationship between online queries about migration to Russia made by residents of Kyrgyzstan, Tajikistan and Uzbekistan and subsequent human flows from these countries to Russia. The hypothesis was tested using the migration statistics of Rosstat, the Federal State Statistics Service of the Russian Federation, Google Trends data on search intensity, and Yandex Wordstat service of word matching for validation of search images. As a result of correlation and regression, we found a moderate dependence of the dynamics of human flows on previous queries, which is most evident at a lag of 6–9 months and at zero lag. Obtaining more accurate results in this and similar studies is hindered by the initial limited predictability of migration behavior due to its contextual, sometimes situational and irrational nature, as well as "noisiness" of statistics of queries and often the flows themselves. The search for universal algorithms of determination of relations between queries and migration flows is seen as the main direction of research in this field.

Key words: migration, forecasting, big data, online queries, search images, modeling, Russia, Central Asia.

Information about the Authors

Irina P. Tsapenko – Doctor of Sciences (Economics), head of sector, Primakov National Research Institute of World Economy and International Relations, Russian Academy of Sciences (23, Profsoyuznaya Street, Moscow, 117997, Russian Federation; e-mail: tsapenko@bk.ru)

Maksim A. Yurevich – Researcher, Financial University under the Government of the Russian Federation (4, 4th Veshnyakovsky Proezd, Moscow, 4109456, Russian Federation; e-mail: maksjuve@gmail.com)

Статья поступила 24.11.2021.