

## Прогнозирование миграции из России в Великобританию с использованием Google Trends Index

Михаил Антонович Галкин

([miagalkin@mail.ru](mailto:miagalkin@mail.ru)),

Национальный исследовательский университет

«Высшая школа экономики», Россия.

## Forecasting migration from Russia to the United Kingdom using Google Trends Index

Mikhail Galkin

([miagalkin@mail.ru](mailto:miagalkin@mail.ru)),

HSE University, Russia.

**Резюме:** Официальная статистика международной миграции часто публикуется с существенными задержками, может страдать от неполноты, несопоставимости и различных типов нарушений. Это затрудняет изучение миграционных потоков, их своевременный мониторинг и прогнозирование. Исследование посвящено применению и модификации методов наукастинга международной миграции с использованием данных поисковых интернет-запросов на примере миграции из России в Великобританию за период 2011-2023 гг. Они предполагают: (а) отбор регрессоров методом LASSO (параметр регуляризации подбирается по кросс-валидации), выполнение прогноза целевой переменной линейной регрессией с использованием полученных с помощью моделей SARIMA прогнозов регрессоров; (б) сокращение размерности методом главных компонент с кросс-валидационным выбором числа факторов и их лагов и последующим построением линейной регрессии на этих факторах, чьи будущие значения оцениваются моделями SARIMA. Информационной базой исследования послужили данные индекса Google Trends, отобранные по результатам проведения лингвистического анализа облака слов поисковых запросов, и официальная миграционная статистика британского правительства – представительского органа страны прибытия. По результатам моделирования были сделаны выводы о том, что (а) использование показателей поисковых запросов позволяет осуществлять прогнозирование международной миграции на разных прогнозных горизонтах с системно более низкой ошибкой, нежели при отсутствии учета поискового интереса; (б) объем собираемых для реализации наукастинга данных может быть существенно сокращен ввиду разреженности части оцененных моделей. Исследование продолжает академическую традицию использования данных поискового интереса в качестве альтернативы официальной статистики и вносит вклад в формирование методологической базы для прогнозирования изменяющихся во времени макропроцессов. Результаты работы могут быть использованы при оценке международной миграции между другими странами, а также в рамках статистических исследований в других научных областях.

**Ключевые слова:** международная миграция, миграция между Россией и Великобританией, большие данные, наукастинг, поисковые запросы, Google Trends Index, лингвистический анализ, LASSO, главные компоненты, SARIMA.

**Благодарности:** Выражаю благодарность за научное руководство, помощь и поддержку при подготовке статьи Елене Сергеевне Вакуленко, доктору экономических наук, профессору Национального исследовательского университета «Высшая школа экономики».

**Для цитирования:** Галкин М.А. (2026). Прогнозирование миграции из России в Великобританию с использованием Google Trends Index. Демографическое обозрение, 13(1), 75-101. <https://doi.org/10.17323/demreview.v13i1.33719>

**Abstract:** Official statistics on international migration are often published with substantial delays and may suffer from incompleteness, incomparability, and various forms of bias. This complicates the measurement of migration flows—their timely monitoring and forecasting. The study focuses on the application and modification of nowcasting methods for international migration using web search–query data, taking migration from Russia to the United Kingdom over the period 2011–2023 as a case study. The methods include: (a) selecting predictors via LASSO (with the regularization parameter chosen by cross-validation) and forecasting the target variable using linear regression based on predictor forecasts obtained from SARIMA models; and (b) dimensionality reduction via principal components with cross-validated selection of the number of factors and their lags, followed by linear regression on these factors, whose future values are estimated with SARIMA models. The empirical basis of the study consists of Google Trends Index data—selected through a linguistic analysis of a search–query word cloud—and the official

*migration statistics published by the government of the destination country (the United Kingdom). The modeling results indicate that (a) incorporating search-query indicators enables forecasting of international migration across different horizons with consistently lower errors than forecasts that ignore search interest; and (b) the amount of data required for implementing nowcasts can be substantially reduced due to the sparsity of some estimated models. The study continues the academic tradition of using search-interest data as an alternative to official statistics and contributes to the methodological foundation for forecasting time-varying macroprocesses. The findings can be applied to assessments of international migration between other countries, as well as to statistical research in other scientific fields.*

**Keywords:** *international migration, migration between Russia and Great Britain, big data, nowcasting, search queries, Google Trends Index, linguistic analysis, LASSO, principal components, SARIMA.*

**Acknowledgments:** *I would like to express my gratitude for the scientific guidance, assistance, and support in preparing this article to Elena Sergeevna Vakulenko, Doctor of Economics, Professor at the HSE University.*

**For citation:** *Galkin M. (2026). Forecasting migration from Russia to the United Kingdom using Google Trends Index. Demographic Review, 13(1), 75-101. <https://doi.org/10.17323/demreview.v13i1.33719>*

## Введение

В современной изменяющейся реальности изучение особенностей миграционных потоков в России и за рубежом является одной из важнейших задач. Возможность прогнозирования перемещения масс людей между разными регионами мира помогает учитывать и грамотно распределять экономические ресурсы на микро- и макроуровнях, создавать рабочие места, своевременно купировать образующиеся ресурсные дефициты и так далее. Это создает благоприятные условия для существования и устойчивости различных общественных систем – от корпоративных до государственных. В то же время вопрос учета миграционных потоков стоит очень остро: далеко не всегда он производится корректно. При исследовании международной миграции традиционным подходом является использование официальных статистических служб стран выбытия и прибытия (Aleshkovski 2012; Schenk 2024; Moreh, McGhee, Vlachantoni 2020; McGhee, Heath, Trevena 2012). Тем не менее официальная статистика нередко публикуется с существенными задержками и страдает от неполноты, несопоставимости и других типов нарушений (Avramescu, Wiśniowski 2021; Wanner 2021; Денисенко 2003), что затрудняет своевременный мониторинг и прогнозирование миграции. В таких условиях использование данных поисковых запросов, например индекса Google Trends, отражающего динамику интереса пользователей сети Интернет к конкретно выбранной теме, становится актуальным направлением для оценки и прогнозов миграции.

Объектом исследования является международная миграция. Предметом исследования – прогнозирование миграции из России в Соединенное Королевство Великобритании и Северной Ирландии с использованием данных Google Trends Index. Великобритания была выбрана в качестве страны прибытия как юрисдикция с прозрачной, детализированной и регулярно обновляемой миграционной статистикой, не предлагающая льготные условия для российских граждан, что нивелирует эффект информационных шумов в рассматриваемом интернет-индексе от перемещений в рамках режима Шенгенской зоны, без визы и др.

Цель работы – разработка эконометрической модели прогнозирования миграционных потоков между Россией и Великобританией с разными прогнозными горизонтами.

Для реализации поставленной цели решали следующие задачи:

1. проведение теоретического анализа применения поисковых запросов международной миграции в сети Интернет;
2. изучение особенностей применения и ограничения использования Google Trends в наукастинге<sup>1</sup> миграционных потоков;
3. проведение лингвистического анализа облака слов поисковых запросов по вопросам международной миграции;
4. модификация методов прогнозирования международной миграции;
5. проведение сравнительного анализа эффективности методов прогнозирования международной миграции на разных прогнозных горизонтах.

---

<sup>1</sup> Наукастинг (англ. nowcasting) — предсказание настоящего, ближайшего будущего и недавнего прошлого состояния некоторого индикатора.

В качестве гипотезы исследования взято предположение о том, что индекс поисковых запросов Google Trends по различным тематикам, связанным с миграцией, повышает точность прогнозирования миграционного потока из России в Великобританию по сравнению с прогнозом без его использования на всех прогнозных горизонтах.

Для прогнозирования международных миграционных потоков на основе данных официальной статистики и запросов в сети Интернет в исследовании использовали два альтернативных подхода: (а) отбор объясняющих переменных методом LASSO (Least Absolute Shrinkage and Selection Operator) с последующей линейной регрессией; (б) сокращение размерности методом главных компонент (PCA, Principal Component Analysis) также с реализацией регрессии на выбранных факторах. В обоих случаях будущие значения регрессоров (факторов) предварительно прогнозировали моделями SARIMA и использовали для построения прогноза целевой переменной. В исследовании также были применены методы машинного обучения для поиска облака слов поисковых запросов путем представления словосочетаний в виде векторов. Новизна работы, таким образом, состоит в предложении модификации методов наукастинга международной миграции на разных прогнозных горизонтах, а также модели прогнозирования миграционных потоков из России в Великобританию, предполагающей применение модифицированных методов на эмпирических данных.

## Теоретический обзор

### *Google Trends Index как альтернатива официальным статистическим данным*

Существуют различные способы оценки международной миграции. Традиционно исследователи опираются на данные официальных статистических служб стран выбытия и прибытия, в нашем случае – Российской Федерации (Aleshkovski 2012; Schenk 2024) и Соединенного Королевства Великобритании и Северной Ирландии (Moreh, McGhee, Vlachantoni 2020; McGhee, Heath, Trevena 2012) соответственно. При этом для составления адекватного представления об объемах миграции между странами рекомендуется использовать данные страны прибытия, а не выбытия (De Haas, Castles, Miller 2019; DeWaard, Kim, Raumer 2012; Бронцкий, Вакуленко 2024). В настоящей работе это замечание особенно актуально, учитывая систематическое недооценивание миграционных потоков в России (Денисенко 2003).

Официальные статистические данные о международной миграции, однако, могут быть не всегда надежны. Обычно они публикуются с существенными временными лагами, их качество нередко страдает от неполноты и различных типов смещения, например заниженной оценки миграционных потоков (Avramescu, Wiśniowski 2021; Wanner 2021). Это затрудняет своевременный мониторинг миграции и оперативное прогнозирование изменения ее объемов. В таких условиях использование цифровых следов (данных, оставляемых людьми в сети Интернет) представляет собой актуальное и методологически новое направление для оценки и прогноза миграции. Одним из наиболее доступных и информативных источников цифровых следов являются данные поисковых запросов, в частности Google Trends Index (GTI), отражающий динамику объема запросов по определенным словам и темам.

Практика применения Google Trends в статистических исследованиях, в том числе в области демографии, формируется со сравнительно недавнего времени. В работе (Choi, Varian 2012) была предпринята попытка спрогнозировать объем продаж

автомобилей, количество первоначальных заявлений на пособие по безработице и показатели потребительской уверенности с помощью еженедельных значений GTI. С опорой на временные ряды поисковых запросов ученые также работали над прогнозированием объемов торговли акциями (Preis, Moat, Stanley 2013), предсказанием объемов индекса NASDAQ-100 (Bordino et al. 2012), проведением исследования кросс-корреляции волатильности и объема промышленного индекса Доу-Джонса (Kristoufek 2015), эпидемиологическим контролем гриппа (Carneiro, Mylonakis 2009), прогнозами заболеваемости простудой в госпиталях (Dugas et al. 2013), наукастингом эпидемий (Yang, Santillana, Kou 2015) и др.

В работе (D'Amuri, Marcucci 2017) авторы построили индекс интенсивности поиска работы и продемонстрировали, что при включении в ежемесячные модели авторегрессии и распределенного лага он приводит к более точным вневыборочным прогнозам уровня безработицы, чем модели, основанные исключительно на первоначальных заявлениях или на данных опросов. Исследование (Pavlicek, Kristoufek 2014) фокусировалось на использовании временных рядов поисковых запросов в браузере Google, связанных с работой, с целью прогнозирования уровня безработицы в Чехии, Венгрии, Польше и Словакии, в результате чего было обнаружено, что включение таких данных значительно улучшает эффективность прогнозирования на несколько месяцев вперед.

За последнее десятилетие был проведен ряд исследований с использованием GTI и в интересующей нас области миграции населения. В работе (Golenvaux et al. 2020) индексы Google Trends были интегрированы для обработки ключевых слов, связанных с миграцией, в нейронную сеть с долгой краткосрочной памятью для прогнозирования притока миграции на год вперед в 35 странах-членах Организации экономического сотрудничества и развития (ОЭСР), что позволило достигнуть снижения ошибок в 4-5 раз по сравнению с базовыми показателями гравитационной модели. В исследовании (Böhme, Gröger, Stöhr 2020) авторы с помощью модели фиксированных эффектов также успешно прогнозировали экономически обусловленную миграцию населения из 101 страны выбытия в те же 35 стран-членов ОЭСР. Оба исследования (Golenvaux et al. 2020; Böhme, Gröger, Stöhr 2020) использовали запросы, связанные с работой, заработными платами и визами.

В работе (Wladyka 2017), в которой была предпринята попытка спрогнозировать миграцию из стран Латинской Америки в Королевство Испания с помощью парной регрессии, поисковый интерес о заработной плате не рассматривался, но учитывались запросы о посольстве; факт поиска работы при этом также был учтен. В исследовании (Wanner 2021), посвященном прогнозированию трудовой миграции из Франции, Италии, Германии и Испании в Швейцарию с помощью линейной регрессии с лагами, поисковой запрос «работа в Швейцарии» в принципе стал ключевым. Аналогичным путем пошли авторы в работе (Цапенко, Юревич 2022), предполагая слова «работа» и «вакансия» взаимозаменяемыми при прогнозировании трудовой миграции из Средней Азии в Россию с помощью моделей ARIMAX. В (Броницкий 2024) автор, прогнозируя миграционные потоки из Польши, Румынии, Италии, Испании и России в Германию с помощью SARIMA и SARIMAX остановился на различных вариантах поисковых запросов на разных языках, выявляющих интерес мигрантов к трудоустройству в Германии.

GTI по тематикам «переезд», «работа», «жилье» использовались в (Fantazzini et al. 2021), где авторы предприняли попытку спрогнозировать внутреннюю

(междугороднюю) миграцию для Российской Федерации моделями типа ARIMAX, SARIMAX и VECM. В исследовании (Juríć 2022) использовался более широкий пул запросов («работа», «жилье», «образование», «паспорт», «виза» и «гражданство») для исследования миграционных потоков в Германию и Австрию из Хорватии с помощью парной регрессии. Несколько иной набор («виза», «переезд», «учеба», «посольство» и «работа») был в работе (Leysen, Verhaeghe 2023) при рассмотрении миграции из Японии в Европу с применением множественной регрессии. В этом контексте отметим, что набор запросов для GTI может меняться в зависимости от этногеополитической ситуации, связанной с регионом выбытия или прибытия. В работе (Qi, Bircan 2023), например, рассматривался исключительно поисковой интерес по запросам «паспорт», «виза» и «убежище» ввиду прогнозирования с помощью гравитационной модели и множественной регрессии миграции из Афганистана, Армении, Грузии, Ирака, Сирии и Турции в более экономически развитые<sup>2</sup> страны Европейского союза.

В ряде других исследований (Avramescu, Wiśniowski 2021; Броницкий, Вакуленко 2022; 2024) рассматривались целые кластеры поисковых запросов. В (Avramescu, Wiśniowski 2021) была предпринята попытка спрогнозировать миграцию из Румынии в Великобританию моделями ARIMAX без лагов, со скользящим средним за 12 месяцев; в (Броницкий, Вакуленко 2022; 2024) – миграцию из России в Германию с помощью SARIMAX. В (Броницкий, Вакуленко 2022) лаги не рассматривались, бралась помесечная частотность данных, в (Броницкий, Вакуленко 2024) учитывались лаги от 1 до 12 месяцев включительно.

Использование Google Trends является столь популярным, так как, в отличие от традиционных для демографии данных, фиксирующих уже совершившиеся миграции, поисковые запросы позволяют уловить интерес к миграции до ее реализации (Qi, Bircan 2023). Люди, планирующие переезд, часто предварительно ищут информацию о визах, работе, образовании, жизни за границей и др. Поэтому рост числа поисковых запросов по этим темам может служить ранним сигналом увеличения миграционного потока в будущем, а анализ поисковых запросов предоставляет возможность предсказывать те же миграционные потоки в реальном времени, опережая публикацию официальной статистики, т. е. осуществлять наукастинг данных. Наша работа предлагает модель наукастинга миграции из России в Великобританию (раннее не рассматриваемых совместно в подобном контексте суверенных юрисдикций) с использованием широкого спектра поисковых запросов, а также с модификацией существующих методов наукастинга международных миграционных потоков.

### ***Ограничения использования Google Trends Index***

Эффективность Google Trends как миграционного индикатора, однако, не является однозначной и может зависеть от исследовательского контекста. В работе (Qi, Bircan 2023) отмечается, что выгода от GTI заметна не всегда и чувствительна к широкому набору факторов: выбранным ключевым словам, странам выбытия и прибытия, характеристикам конкретной группы пользователей сети Интернет и доле пользователей Google в данном регионе. Есть и более общие проблемы, связанные с использованием временных рядов,

---

<sup>2</sup> Согласно классификации Международного валютного фонда:  
<https://www.imf.org/en/Publications/WEO/weo-database/2023/April/groups-and-aggregates>

составленных на основе Google Trends, при прогнозировании:

- во-первых, механизм выборки вносит существенную изменчивость в выгрузки данных: каждое извлечение может давать разные значения для одного и того же запроса, что приводит к изменению выбора модели и снижению точности прогноза (Medeiros, Pires 2021);
- во-вторых, обновления алгоритма и сбора данных Google могут приводить к ошибкам измерения, проявляющимся в виде ложных ненулевых значений для поисковых запросов до их появления или резких сдвигов индекса после методологических изменений (Liu 2024);
- в-третьих, нормализация объемов поиска, принятая компанией Google, может исказить сравнения по временным периодам и географическим регионам, что вызывает опасения относительно концептуальной валидности данных (Rovetta 2021);
- в-четвертых, чрезмерная зависимость исследования от поисковых данных без строгой проверки может привести к неудаче в прогнозировании, например из-за «высокомерия больших данных» («big data hubris») и алгоритмической переподгонки (Lazer et al. 2014);
- наконец, есть и другие, менее значительные вызовы, связанные с влиянием социальных движений в обществе на точность прогнозирования (Ormerod, Nyman, Bentley 2014), подбором ключевых слов (Challet, Ayed 2014) и др.

Однако, несмотря на существующие ограничения, возможность использования больших данных поисковых запросов в сочетании с традиционной статистикой открывает новое направление в демографических и экономических исследованиях миграции. Данный подход вписывается в более широкую парадигму цифровой демографии, предлагая новые инструменты для отслеживания мобильности населения в режиме, близком к реальному времени.

## **Данные поискового интереса и международной миграции**

### ***Сбор данных***

Мы исследовали взаимосвязи между временными рядами показателей поискового интереса к миграции и миграционного потока, с последующим прогнозированием их динамики. Данные поисковых запросов были собраны с помощью Google Trends Index<sup>3</sup> (GTI), который измеряет интенсивность поиска выбираемого набора ключевых слов, показывая общий объем поиска по запросу в браузере Google в выбранном географическом регионе, разделенный на общее количество запросов в этом регионе в определенный момент времени. География региона была ограничена Российской Федерацией – страной выбытия, откуда делались интернет-запросы. Исследуемый период (I квартал 2011 г. – IV квартал 2023 г.) был снизу ограничен изменением методики расчета GTI: в 2011 г., как отмечается в работе Think with Google (Rennie et al. 2020) – официального подразделения компании, изменился подход к определению региона поиска<sup>4</sup>,

---

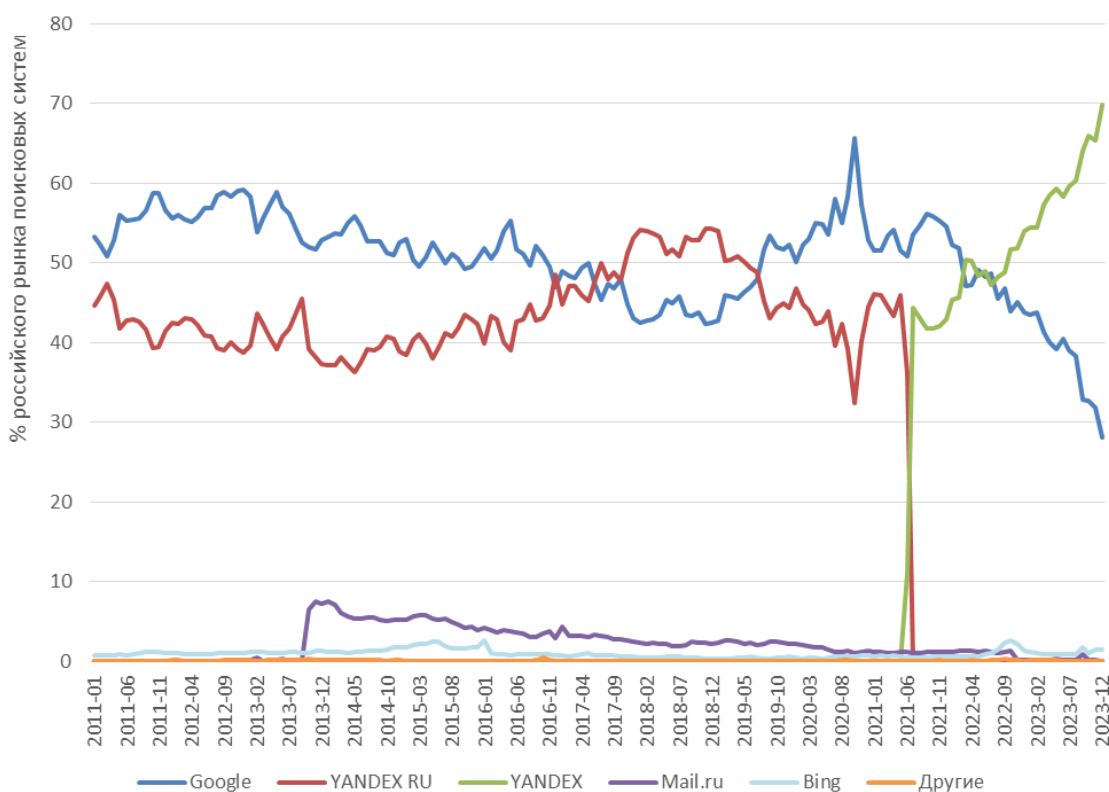
<sup>3</sup> <https://trends.google.com/trends/>

<sup>4</sup> Англ.: geographical assignment.

что повлияло на подсчет индекса. Верхнее ограничение было обусловлено отсутствием официальных данных по динамике миграционного потока.

Стоит отметить, что выбранный временной период характеризуется существенной долей Яндекса на российском рынке поисковых систем (рисунок 1): она варьируется от 44,65% в январе 2011 г. (YANDEX RU) до 69,79% в декабре 2023 г. (YANDEX). Доля же Google в поисковом трафике России, хотя и остается значительной, снижается с 53,23% в начале 2011 г. до 28,05% в конце 2023 г. Подобная динамика может ограничивать возможность экстраполяции данных GTI на весь поисковый трафик России в соответствии с рассмотренными ранее замечаниями (Qi, Bircan 2023).

**Рисунок 1. Доля рынка поисковых систем в России (помесячные данные)**



Источник: Данные портала StatCounter<sup>5</sup>.

Поисковые запросы рассматривали на русском и английских языках ввиду принятия предпосылки о том, что мигранты обычно знают язык страны прибытия достаточно хорошо, чтобы иметь возможность проявлять на нем интерес к процессу миграции (Dustmann, Fabbri 2003). Релевантное нашей работе облако слов искали путем представления словосочетаний в виде векторов. Для нахождения русскоязычных словосочетаний была использована лингвистическая модель-корпус Национальный корпус русского языка<sup>6</sup>, англоязычных – перевод отобранных русских словосочетаний на английский язык с применением Кембриджского англо-русского словаря<sup>7</sup>. Такой алгоритм

<sup>5</sup> <https://gs.statcounter.com/search-engine-market-share/all/russian-federation/#monthly-201101-202312>

<sup>6</sup> <https://ruscorpora.ru/>

<sup>7</sup> <https://dictionary.cambridge.org/ru/%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D1%8C/%D0%B0%D0%BD%D0%B3%D0%BB%D0%BE-%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9/>

подтверждал логику вторичности английского языка для носителя языка русского, отправляющего поисковые запросы из России. Полноту множества отобранных слов оценивали по аналогии с работой (Броницкий, Вакуленко 2022) с использованием RusVectōrēs<sup>8</sup> – сервиса для работы с семантическими моделями для русского языка.

Реализация данного подхода, однако, не всегда приводила к нахождению словосочетаний, реально употребляемых при поиске: за синоним слова «миграция» алгоритмом принимались, например, «расселение», «отток», «перемещение» и др. Таким образом, была осуществлена дополнительная проверка популярности находимых словосочетаний и соответственно полноты множества отобранных слов также через ресурс Google Trends, после чего данные были приведены к квартальной размерности с тем, чтобы сделать их сопоставимыми миграционному потоку, путем взятия среднего арифметического. При этом данные поисковых запросов (GTI) имели одинаковый масштаб и в рамках работы были собраны по отдельности, что позволило не проводить их дополнительную нормализацию. Получившиеся временные ряды были условно названы  $gti^*$ , где \* - соответствует цифрам от 1 до 23 (таблица 1).

**Таблица 1. Расшифровка переменных**

Переменная	Расшифровка
gti1	виза Англия
gti2	виза Великобритания
gti3	виза в Англию
gti4	виза в Великобританию
gti5	британская виза
gti6	английская виза
gti7	английское посольство
gti8	гражданство Великобритании
gti9	английский паспорт
gti10	работа в Англии
gti11	работа Англия
gti12	работа в Великобритании
gti13	работа Великобритания
gti14	жизнь в Англии
gti15	British life
gti16	UK visa
gti17	law UK
gti18	British council
gti19	British embassy
gti20	UK embassy
gti21	UK passport
gti22	job UK
gti23	life UK
flow	миграционный поток

*Источник: Составлено автором.*

Полученные данные характеризуются еще двумя артефактами. Во-первых, наличием падежей и предлогов в русскоязычных словосочетаниях, что, как выяснилось после проверки интереса пользователей через Google Trends, характеризует реальные формулировки запросов. Во-вторых, чередованием топонимов («Англия» и «Великобритания» в русскоязычных запросах, «UK» и «British»

<sup>8</sup> <https://rusvectors.org/ru/about/>

– в англоязычных и др.); это связано с попыткой включить в облако слов все возможные топонимы, характерные при демонстрации мигрантами интереса к стране прибытия, в том числе «Шотландия», «Уэльс», «Соединенное Королевство» и др. Как видно из полного перечня отобранных запросов, представленного в таблице 1, только некоторые из них оказались достаточно популярны среди мигрантов.

Данные миграционного потока были собраны на основе официальной миграционной статистики страны прибытия<sup>9</sup> – Соединенного Королевства Великобритании и Северной Ирландии. Для этого был использован официальный интернет-сайт британского правительства. Квартальные данные использовали в работе, так как на момент их сбора (осень 2024 г.) статистические службы Соединенного Королевства агрегировали миграционные данные исключительно по кварталам, а тренд на подведение месячных итогов только формировался. За миграционный поток были приняты данные по выдаче и отказу в выдаче документов на проживание (за исключением программы Европейского союза по предоставлению вида на жительство) гражданам Европейской экономической зоны (ЕЭЗ) и членам их семей по стране гражданства<sup>10</sup> – квартальная статистика по выданным разрешениям на проживание на территории страны<sup>11</sup>. Данные отбирали по региону (region) «другая Европа» («Europe other») и национальности (nationality) «русский» («Russian»). Впоследствии данные суммировали для каждого квартала без учета категории решения («decision category»).

### **Подготовка и анализ данных**

По причине того, что собранные данные имели квартальную периодичность, максимальное количество лагов для каждого временного ряда было принято равным 4, а общее – 5 (ввиду учета нулевого лага) в соответствии с наработками, предлагаемыми в работах (Canova, Hansen 1995; Burridge, Wallis 1984; Rodrigues, Taylor 2004). Таким образом, набор переменных  $X_t^{(n)}$  был расширен до  $\{X_{t-n}^*\}^4$ , где  $X_t^*$  – значение \*-й переменной в квартале  $t$ ,  $n$  – «шаг» лага.

В качестве первичного был проведен корреляционный анализ полученного набора переменных с расчетом коэффициента корреляции по формуле Пирсона (Pearson 1896). Показатели, относящиеся к тематике виз и документов, оказались сильно коррелированы между собой. При этом поисковые запросы, связанные с жизнью в Великобритании ( $gti14$ ,  $gti15$ ,  $gti23$ ), а также с получением паспортов ( $gti9$ ,  $gti21$ ), показали заметную положительную корреляцию с миграционным потоком при нулевом и небольшом лаге, что может указывать на потенциальную прогностическую значимость данных индикаторов.

Далее был определен порядок интегрированности временных рядов, так как нестационарные ряды могут приводить к ложным регрессиям, когда высокая корреляция обусловлена общим трендом, а не истинной причинно-следственной связью. Проверка рядов на наличие единичных корней была проведена с помощью расширенного теста Дики-Фуллера (Dickey, Fuller 1979), включающего тренд, свободный член и максимальное количество лагов, в нашем случае равное 4, определенное по BIC

<sup>9</sup> <https://www.gov.uk/government/collections/migration-statistics>

<sup>10</sup> Англ.: Issue and Refusal of Residence Documentation (Excluding EU Settlement Scheme) to EEA Nationals and Their Family Members, by Country of Nationality.

<sup>11</sup> <https://www.gov.uk/government/statistical-data-sets/immigration-system-statistics-data-tables>

(Schwarz 1978). В результате было выявлено, что миграционный поток  $flow$  не стационарен в уровнях, но становится стационарным при взятии первых разностей. Аналогично несколько регрессоров требовали дифференцирования:  $gti2$ ,  $gti5$  и  $gti13$  оказались интегрированными рядами первого порядка,  $gti18$  – второго; остальные регрессоры не имели единичного корня. Тот факт, что  $flow$  оказался рядом, стационарным в первых разностях, могло вызвать осложнения при моделировании его взаимосвязи с другими переменными: в частности, в работе (Granger, Newbold 1974) было показано, что объединение нестационарных рядов в регрессии может часто приводить к ложным значимостям.

В связи с этим мы проверили, не образуют ли нестационарные переменные долгосрочного равновесия, т.е. имеют ли они коинтегрированные отношения. Для нестационарных рядов нами была проведена проверка на коинтеграцию с помощью теста Йохансена (Johansen 1988) с использованием  $trace$ -статистики теста. Для всех временных рядов статистика оказалась меньше порогового значения<sup>12</sup>. Следовательно, напрямую моделировать миграционный поток на уровне вместе с нестационарными регрессорами не представлялось возможным, в противном случае регрессия была бы несостоятельна. Вместо этого был сделан переход к модельным уравнениям в первых разностях с проведением дальнейшего анализа на стационарных преобразованиях с рассмотрением  $\Delta flow$  и  $\Delta gti^*$ .

Как отмечалось выше, в качестве потенциальных регрессоров прогностических моделей мы рассматривали одновременные изменения и лагированные значения показателей поискового запроса. Добавление лагов позволило среди прочего учесть распределенный лаговый эффект причинного фактора: например, рост запросов о трудоустройстве за рубежом мог реализоваться в реальных переездах только спустя несколько месяцев (Броницкий, Вакуленко 2022; Böhme, Gröger, Stöhr 2020; Avramescu, Wiśniowski 2021). Однако это численно усложнило задачу, так как число потенциальных предикторов увеличилось до  $23 \times 5 = 115$ , где 23 – число исходных поисковых индикаторов, 5 – число их лагов (Canova, Hansen 1995; Burrige, Wallis 1984; Rodrigues, Taylor 2004). Таким образом, далее задачу решали, используя два различных алгоритма:

- регуляризацию методом LASSO (Least Absolute Shrinkage and Selection Operator), так как такой подход позволяет осуществить встроенный отбор регрессоров (Tibshirani 1996);
- и агрегирование регрессоров методом главных компонент (PCA, Principal Component Analysis), так как данный подход позволяет создавать крупные тематические группы регрессоров, что было реализовано в других научных работах в области наукастинга миграции (Avramescu, Wiśniowski 2021; Броницкий, Вакуленко 2022; 2024).

<sup>12</sup> Результаты теста могут быть предоставлены по запросу.

## Методология исследования

### Метод LASSO

LASSO накладывает  $L_1$ -штраф на коэффициенты регрессии, обнуляя часть из них и тем самым осуществляя встроенный отбор признаков (Tibshirani 1996). Таким образом, метод подходит для решения нашей задачи, когда число потенциальных регрессоров велико и заранее не известно, какие из них действительно информативны. При этом LASSO нередко улучшает точность прогноза по сравнению с моделями, оцененными без регуляризации (Zou 2006).

Поиск оптимального значения  $\alpha$  – параметра регуляризации LASSO – был осуществлен через процедуру  $k$ -блочной кросс-валидации,  $k=5$ . С целью рассмотрения широкого спектра значений  $\alpha$ , было взято 200 значений данного параметра, равномерно покрывающих широкий диапазон:  $[-1\ 000\ 000,000; 1\ 000\ 000,000]$ . Обучаемые модели сортировали по усредненной по 5 блокам кросс-валидационной (вневыборочной) среднеквадратичной ошибке (CV MSE) на валидационных подвыборках по возрастанию показателя. При этом для сохранения точности процедуры кросс-валидации для каждого рассматриваемого  $\alpha$  модель LASSO заново обучали на всей выборке. Если два разных значения  $\alpha$  соответствовали одинаковому поднабору переменных, то далее рассматривали только ту модель, у которой CV MSE была меньше, что позволяло обеспечить более удачную регуляризацию для данного набора. В итоге было получено множество моделей, отличающихся набором признаков и сложностью (таблица 2).

**Таблица 2. Отбор моделей LASSO**

Модель	CV MSE	Регрессоры	$\alpha$
0	2 068,759	—	1 000 000,000
1	2 118,850	$gti15_{t-1}$ («British life», лаг = 1)	276,829
2	2 175,928	$gti15_{t-1}$ («British life», лаг = 1) $gti21_t$ («UK passport», лаг = 0)	209,705
3	2 210,284	$gti15_{t-1}$ («British life», лаг = 1) $gti21_t$ («UK passport», лаг = 0) $gti23_{t-3}$ («life UK», лаг = 3)	182,518
4	2 262,774	$gti15_{t-1}$ («British life», лаг = 1) $gti15_{t-4}$ («British life», лаг = 4) $gti21_t$ («UK passport», лаг = 0) $gti23_{t-3}$ («life UK», лаг = 3)	158,857
5	2 440,413	$gti15_{t-1}$ («British life», лаг = 1) $gti15_{t-4}$ («British life», лаг = 4) $gti21_t$ («UK passport», лаг = 0) $gti23_{t-1}$ («life UK», лаг = 1) $gti23_{t-3}$ («life UK», лаг = 3)	120,338

Источник: Расчеты автора.

Модель 0 оказалась непригодна для дальнейшего использования при прогнозировании ввиду отсутствия регрессоров, поэтому далее рассматривали регрессионные модели (1-5), имеющие наименьшее CV MSE:

**Модель 1:**

$$\widehat{\Delta flow} = -2,702 + 0,050 \Delta gti15_{t-1} \quad (1)$$

**Модель 2:**

$$\widehat{\Delta flow} = -2,667 - 0,036 \Delta gti15_{t-1} + 0,170 \Delta gti21_t \quad (2)$$

**Модель 3:**

$$\widehat{\Delta flow} = -2,682 - 0,085 \Delta gti15_{t-1} + 0,218 \Delta gti21_t + 0,012 \Delta gti23_{t-3} \quad (3)$$

**Модель 4:**

$$\widehat{\Delta flow} = -2,700 - 0,143 \Delta gti15_{t-1} + 0,271 \Delta gti15_{t-4} - 0,025 \Delta gti21_t + 0,095 \Delta gti23_{t-3} \quad (4)$$

**Модель 5:**

$$\widehat{\Delta flow} = -2,703 - 0,240 \Delta gti15_{t-1} + 0,358 \Delta gti15_{t-4} - 0,097 \Delta gti21_t - 0,004 \Delta gti23_{t-1} + 0,226 \Delta gti23_{t-3} \quad (5)$$

где  $\widehat{\Delta flow}$  – оценка первых разностей временного ряда миграционного потока  $flow$ ;  $\Delta gti_{t-n}^*$  – первые разности поисковых запросов  $gti^*$  в квартале  $t$ ,  $n$  – «шаг» лага.

Отобранные модели (1-5) были обучены на всем объеме данных. Подход к прогнозированию миграционного потока предполагал рекурсивный (динамический) метод и сочетал регрессии LASSO с независимым прогнозированием факторов моделями типа SARIMA (Вох, Jenkins 1970). Дело в том, что значения поисковых индексов (регрессоров) были известны только до текущего периода и если бы динамический прогноз реализовывался исключительно с опорой на лаги  $\Delta flow$ , то:

- во-первых, вся информация из  $gti^*$  была бы утрачена, хотя она ранее оказалась значимой при отборе признаков (таблица 2);
- во-вторых, при многозвенном рекурсивном прогнозе известные лаги  $\Delta gti^*$  быстро бы иссякли.

Модели  $SARIMA(p, d, q) \times (P, D, Q)_s$  для каждого временного ряда (таблица 1) подбирали по  $p, P \in [0; 4]$  (по числу лагов),  $d, D \in [0; 1]$  (по глубине интегрируемости в результате проверки на стационарность с учетом уже реализованного перехода в первые разности),  $q, Q \in [0; 4]$  (также по числу лагов) на полном рассматриваемом историческом интервале. Для этого для перебираемых моделей минимизировали информационный критерий AIC (Akaike 1974). Подобранная конфигурация модели для  $\Delta flow$  –  $SARIMA(3, 1, 3) \times (4, 1, 3)_4$  была построена в качестве базовой модели, прогноза миграционного потока без регрессоров<sup>13</sup>.

Прогнозирование осуществляли на горизонте четырех кварталов, так как ранее мы лагировали переменные до четырех кварталов включительно (Вох, Jenkins 1970). Формально для решения прогностической задачи использовали формулу (6):

<sup>13</sup> Результаты проверки адекватности модели могут быть предоставлены по запросу.

$$\widehat{\Delta y}_{T+h|T} = \sum_{s \in S} \widehat{\beta}_s \widehat{\Delta x}_{s,T+h-n_s}, \widehat{\Delta x}_{s,t} = \begin{cases} \Delta x_{s,t}, & t \leq T \\ \widehat{\Delta x}_{s,t|T}^{SARIMA}, & t > T \end{cases} \quad (6)$$

где  $\widehat{\Delta y}_{T+h|T}$  – точечный прогноз первой разности  $y_{T+h}$ , построенный по информации, доступной на момент  $T$ ;  $T$  – последний наблюдаемый момент времени;  $h = 1, \dots, 4$  – шаг прогноза;  $S$  – множество предикторов, отобранных LASSO (GTI-индикаторы и их лаги);  $n_s$  – лаг для фактора  $s \in S$ ;  $\widehat{\beta}_0, \widehat{\beta}_s$  – оценки коэффициентов линейной регрессии, полученные методом LASSO (параметр регуляризации выбран по кросс-валидации) на данных  $t \leq T$ ;  $\widehat{\Delta x}_{s,t}$  – используемая в прогнозе первая разность предиктора  $s$  в момент  $t$ ;  $\Delta x_{s,t}$  – наблюдаемое значение первой разности предиктора  $s$  на момент  $t$ ;  $\widehat{\Delta x}_{s,t|T}^{SARIMA}$  – прогноз значения первой разности данного предиктора на момент  $t > T$ , полученный моделью SARIMA, оцененной по данным до момента  $T$ .

В результате применения алгоритма (6) к моделям (1-5) были получены прогнозы, для количественного анализа качества которых на каждом прогнозном горизонте (квартале) была вычислена по формуле (7)  $MSE_{OOS}$  – вневыборочная среднеквадратичная ошибка, рассчитанная на отложенной выборке<sup>14</sup> (Hyndman, Athanasopoulos 2018):

$$MSE_{OOS} = \frac{1}{H} \sum_{h=1}^H (y_{T+h} - \widehat{y}_{T+h})^2 \quad (7)$$

где  $y_{T+h}$  – истинное значение целевой переменной на  $h$ -м шаге за концом обучающей выборки;  $\widehat{y}_{T+h}$  – прогноз, полученный моделью для того же шага;  $T$  – индекс последнего наблюдения обучающей выборки;  $H = 4$  – длина тестового отрезка.

### **Метод главных компонент**

Альтернативный подход к прогнозированию миграционного потока, рассмотренный в исследовании, основан на агрегировании исходных признаков с помощью метода главных компонент (Hotelling 1933). Применение метода позволяет сократить размерность исходных данных и устранить мультиколлинеарности путем построения нескольких интегральных факторов, которые затем используются в регрессии. Данный метод ранее широко применялся в демографических и эконометрических исследованиях (Броницкий, Вакуленко 2022; Stock, Watson 2002; Hyndman, Ullah 2007; Bai 2003) для выявления скрытых драйверов в высокомерных данных. Более того, PCA содержательно актуален для нашей работы:

<sup>14</sup> Для оценки качества прогноза использован только этот критерий, так как на последних восьми кварталах (I квартал 2022 г. – IV квартал 2023 г.) реальный миграционный поток принимает нулевые значения. Это, в частности, видно на рисунках 2 и 3 (см. «Базовая модель (факт)») при взятии первых разностей и может быть связано с изменившейся геополитической ситуацией. Соответственно расчет любых метрик качества, не зависящих от единиц измерения показателей, представляется проблематичным. Метрики же, дублирующие оценку качества прогноза (например, MAE или RMSE) демонстрируют значения, пропорциональные  $MSE_{OOS}$ . В связи с этим они не приведены в рамках дальнейших рассуждений, но могут быть предоставлены по запросу.

- во-первых, корреляционный анализ показал значимую связь групп переменных между собой;
- во-вторых, все рассматриваемые переменные  $gti^*$  могут быть агрегированы по тематике смысловых запросов с опорой на опыт перечисленных выше работ.

В связи с этим для каждой группы регрессоров, определенных согласно тематике включаемых в них переменных, был выполнен PCA на стандартизированных первых разностях входящих в нее серий, которые предлагается интерпретировать как индексы общего интереса к той или иной тематике (таблица 3).

**Таблица 3. Группы компонент**

Главные компоненты	Включенные переменные
виза	gti1 («виза Англия») gti2 («виза Великобритания») gti3 («виза в Англию») gti4 («виза в Великобританию») gti5 («британская виза») gti6 («английская виза») gti16 («UK visa»)
посольство	gti7 («английское посольство») gti18 («British council») gti19 («British embassy») gti20 («UK embassy»)
гражданство	gti8 («гражданство Великобритании») gti9 («английский паспорт») gti21 («UK passport»)
работа	gti10 («работа в Англии») gti11 («работа Англия») gti12 («работа в Великобритании») gti13 («работа Великобритания») gti22 («job UK»)
жизнь	gti14 («жизнь в Англии») gti15 («British life») gti17 («law UK») gti23 («life UK»)

Источник: Составлено автором.

При составлении компонент в них не были включены лаги  $gti^*$ , так как в рамках дальнейшего анализа набор переменных был расширен путем создания лагов непосредственно для компонент аналогично тому, как ранее формировались лаги регрессоров для  $gti^*$ . По реализации данной процедуры было получено всего  $5 \times 5 = 25$  (где 5 – число компонент, 5 – число их лагов (Canova, Hansen 1995; Burr ridge, Wallis 1984; Rodrigues, Taylor 2004)) таких признаков, что свидетельствует о сокращении размерности в 4,600 раза. Так задача свелась к выбору наилучшего поднабора из 25 кандидатных переменных, включая лаги регрессоров.

В нашей работе были рассмотрены все возможные комбинации признаков размерностью от 1 до 6 регрессоров в модели. Такое ограничение было введено, чтобы избежать переобучения моделей и сократить пространство перебора, так как теоретически количество комбинаций без вводимого ограничения составляло свыше 33 млн возможных вариантов. Каждое подмножество признаков заданной размерности оценивали с помощью 5-блочной кросс-валидации, используемой в том числе для отбора лагов.

Данную процедуру мы задумывали методологически схожей с использованной для LASSO, однако она имела специфическую корректировку: PCA выполнялся заново на каждом блоке только на обучающей части данных, что предотвращало утечку информации. Факторные веса вычислялись без учета тестовых наблюдений, благодаря чему усредненная по 5 блокам вневыборочная CV MSE, по которой модели сортировали в порядке возрастания показателя (от лучшей к худшей), оставалась релевантной. Модели обучались линейной регрессией методом наименьших квадратов (Aitken 1935). В итоге было получено множество моделей, отличающихся набором признаков, включая их лаговую вариативность, и соответственно сложностью (таблица 4).

**Таблица 4. Отбор моделей PCA**

Модель	CV MSE	Главные компоненты	Доля объясненной дисперсии
1	1 854,626	гражданство <sub>t</sub> (лаг = 0)	0,989
		работа <sub>t</sub> (лаг = 0)	0,809
		виза <sub>t-1</sub> (лаг = 1)	0,853
2	1 858,154	посольство <sub>t</sub> (лаг = 0)	0,641
		гражданство <sub>t</sub> (лаг = 0)	0,989
		работа <sub>t</sub> (лаг = 0)	0,809
3	2 034,180	виза <sub>t-1</sub> (лаг = 1)	0,853
		гражданство <sub>t</sub> (лаг = 0)	0,641
		работа <sub>t</sub> (лаг = 0)	0,809
4	2 191,649	виза <sub>t-1</sub> (лаг = 1)	0,853
5	2 203,904	посольство <sub>t-1</sub> (лаг = 1)	0,641

*Источник: Расчеты автора.*

Ни в одной из выбранных в качестве лучших по CV MSE моделей не наблюдается больше 5 регрессоров, что свидетельствует о корректно выставленном ограничении на максимальное количество регрессоров, равном 6. Агрегирование регрессоров по компонентам оказалось успешным и в смысле высокой доли объясненной дисперсии (Hotelling 1933) (таблица 4).

Далее были рассмотрены регрессионные модели, имеющие наименьшую CV MSE:

**Модель 1:**

$$\widehat{\Delta flow} = -2,719 - 0,474 \Delta \text{гражданство}_t - 0,280 \Delta \text{работа}_t + 0,085 \Delta \text{виза}_{t-1} \quad (8)$$

**Модель 2:**

$$\widehat{\Delta flow} = -2,721 - 0,067 \Delta \text{посольство}_t - 0,473 \Delta \text{гражданство}_t - 0,279 \Delta \text{работа}_t + 0,082 \Delta \text{виза}_{t-1} \quad (9)$$

**Модель 3:**

$$\widehat{\Delta flow} = -3,215 - 0,461 \Delta \text{виза}_t - 0,384 \Delta \text{гражданство}_t - 0,317 \Delta \text{работа}_t - 0,045 \Delta \text{виза}_{t-1} \quad (10)$$

**Модель 4:**

$$\widehat{\Delta flow} = -2,703 + 0,148 \Delta \text{виза}_{t-1} \quad (11)$$

**Модель 5:**

$$\widehat{\Delta flow} = -2,783 - 4,850 \Delta \text{посольство}_{t-1} \quad (12)$$

где  $\widehat{\Delta flow}$  – оценка первых разностей временного ряда миграционного потока  $flow$ ;  $\Delta \text{компонента}_{t-n}^*$  – первые разности главных компоненты в квартале  $t$ ,  $n$  – «шаг» лага.

Отобранные регрессионные модели (8-12) были обучены на всем объеме данных. Миграционный поток при реализации метода главных компонент также прогнозировался по двухэтапной рекурсивной схеме (13): коэффициенты множественной линейной регрессии были оценены на факторах PCA, а будущие значения этих факторов получены из отдельных моделей SARIMA и далее подставлены в регрессию:

$$\widehat{\Delta y}_{T+h|T} = \sum_{s \in S} \widehat{\beta}_s \widehat{\Delta z}_{s, T+h-n_s}, \widehat{\Delta z}_{s,t} = \begin{cases} \Delta z_{s,t}, & t \leq T \\ \widehat{\Delta z}_{s,t|T}^{SARIMA}, & t > T \end{cases} \quad (13)$$

где  $\widehat{\Delta y}_{T+h|T}$  – точечный прогноз первой разности  $y_{T+h}$ , построенный по информации, доступной на момент  $T$ ;  $T$  – последний наблюдаемый момент времени;  $h = 1, \dots, 4$  – шаг прогноза;  $S$  – множество отобранных факторов (компонент PCA);  $n_s$  – лаг для фактора  $s \in S$ ;  $\widehat{\beta}_0, \widehat{\beta}_s$  – коэффициенты множественной линейной регрессии, оцененные по данным  $t \leq T$ ;  $\widehat{\Delta z}_{s,t}$  – используемая в прогнозе первая разность значения фактора PCA блока  $s$  в момент  $t$ ;  $\Delta z_{s,t}$  – наблюдаемое значение первой разности фактора PCA для блока  $s$  на момент  $t$ ;  $\widehat{\Delta z}_{s,t|T}^{SARIMA}$  – прогноз значения первой разности данного фактора на момент  $t > T$ , полученный моделью SARIMA, оцененной по данным до момента  $T$ .

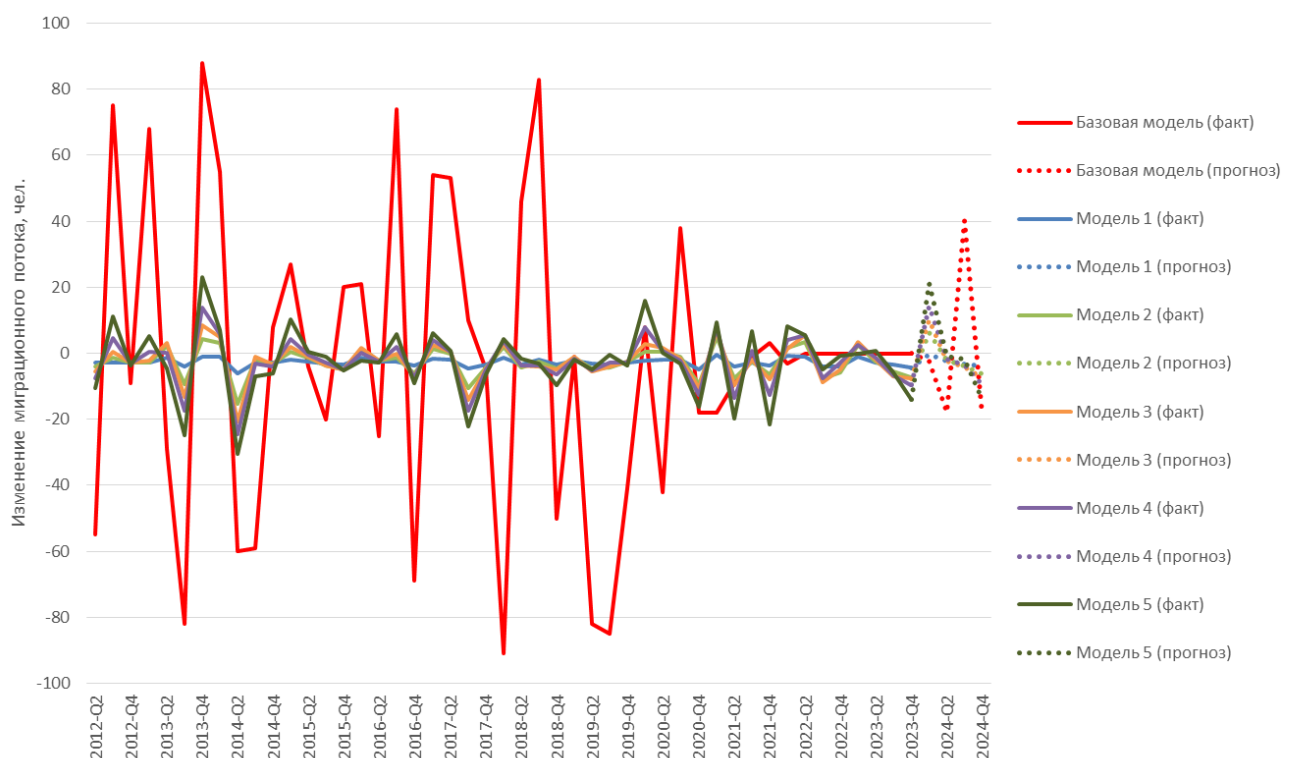
Предпосылки применения подобного сочетания методов уже рассматривались выше. Модели SARIMA для каждой компоненты подбирали по тем же параметрам, что и для LASSO, на полном историческом интервале с минимизацией информационного критерия AIC (Akaike 1974). Базовая модель для PCA полностью совпала с таковой для LASSO ввиду однородности используемой методологии. В результате применения алгоритма (13) к моделям (8-12) были получены прогнозы, качество которых на разных горизонтах было оценено с помощью выбранного критерия ошибки  $MSE_{OOS}$  (7).

## Результаты исследования

Отобранные модели LASSO (1-5) характеризуют простые регрессии, включающие от 1 до 5 признаков, причем, согласно логике регуляризации, чем больше  $\alpha$ , тем меньше регрессоров включается в модель. Таким образом, оптимальные модели оказались разреженными, что согласуется с принципом минимальной достаточности (Lee et al. 2016): не привлекать лишних факторов без необходимости. Модели PCA (8-12) включают от 1 до 4 признаков и также были отобраны согласно данной логике. Максимальный лаг регрессоров в данных моделях оказался равным одному кварталу. Это объяснимо в контексте того, что лагирование касалось компонент «посольство» и «виза» – проверочный интерес по данным темам характерен перед непосредственной реализацией миграционных интенций (Qi, Reed, Bevelander 2025).

Модели LASSO аппроксимируют динамику миграционного потока (изменение относительно предыдущего квартала, чел.), в том числе его ключевые подъемы и спады (рисунок 2). Однако модели (1-2) с меньшим числом регрессоров улавливают его частично, а модели (3-5) с большим числом переменных более точно стремятся к повторению реальных колебаний. Похожая ситуация наблюдается и применительно к моделям PCA с поправкой на то, что динамику миграционного потока (также изменение относительно предыдущего квартала, чел.) они повторяют не столь однонаправленно и несколько более хаотично (рисунок 3).

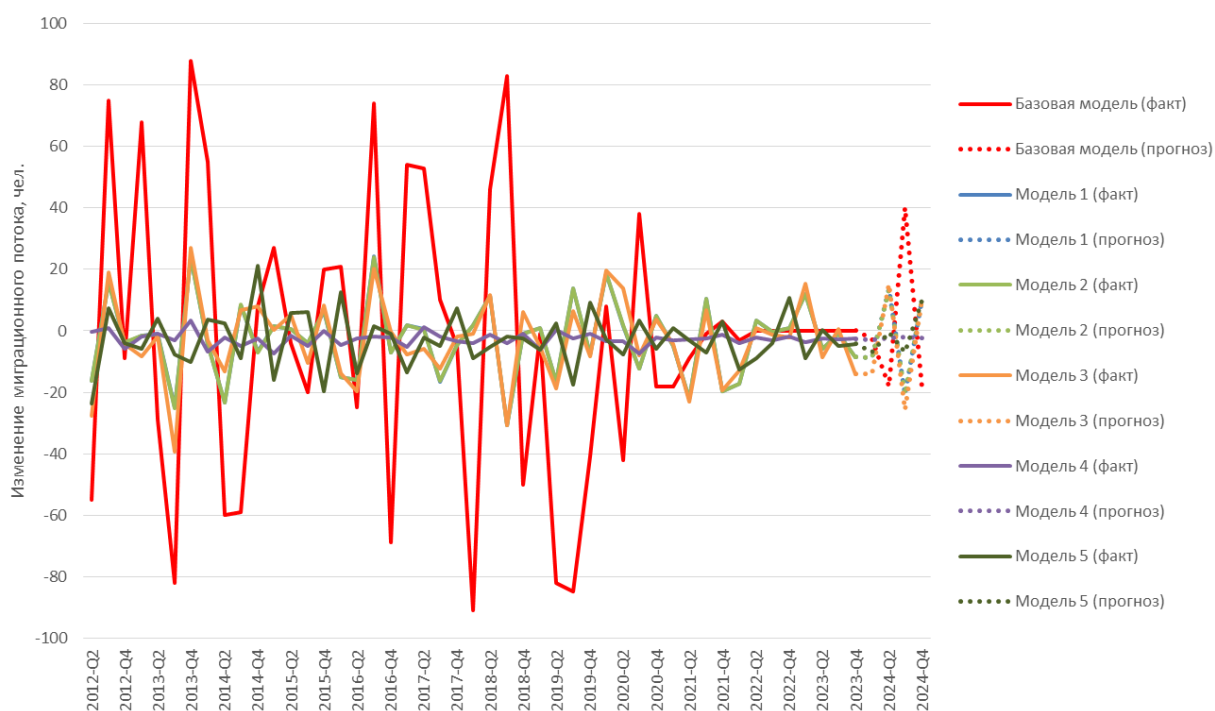
**Рисунок 2. Прогнозирование изменения миграционного потока методом LASSO на 4 квартала, чел.**



Источник: Расчеты автора.

Примечание: Рассматриваемый горизонт начинается со II квартала 2012 г. ввиду взятия первых разностей и лагов.

**Рисунок 3. Прогнозирование изменения миграционного потока методом PCA на 4 квартала, чел.**



Источник: Расчеты автора.

Примечание: Рассматриваемый горизонт начинается со II квартала 2012 г. ввиду взятия первых разностей и лагов.

**Таблица 5. Сравнение результатов моделей LASSO и PCA по  $MSE_{00s}$**

Прогнозный горизонт, кв.	I	II	III	IV
LASSO				
Модель 1	0,178	19,625	12,379	7,695
Модель 2	16,838	39,436	16,861	8,678
Модель 3	21,315	62,354	11,811	13,196
Модель 4	10,487	64,264	1,651	36,393
Модель 5	0,077	59,638	6,796	102,707
PCA				
Модель 1	156,546	61,742	5,949	15,467
Модель 2	158,572	63,053	5,481	16,068
Модель 3	242,851	124,547	7,484	0,057
Модель 4	14,660	6,728	6,932	7,293
Модель 5	88,469	5,335	0,689	59,301
Базовая модель				
SARIMA (3, 1, 3) × (4, 1, 3) <sub>4</sub>	740,053	2357,927	533,995	38,183

Источник: Расчеты автора.

На разных прогнозных горизонтах модели LASSO различной сложности показали себя лучше с точки зрения  $MSE_{00s}$  (таблица 5). Так, при прогнозировании на I квартал наименьшую ошибку имела модель (5), обладающая наибольшим количеством регрессоров; на II и IV кварталы – существенно более разреженная модель (1); на III квартал – модель (4) также со сравнительно большим числом предикторов. При реализации PCA, как также видно из таблицы 5, при прогнозировании на I квартал наименьшую ошибку

имела модель (11) с 1 регрессором; на II и III кварталы – модель (12), также включающая 1 регрессор; на IV квартал – одна из самых сложных из отобранных модель (10) с 4 регрессорами. Таким образом, при реализации и LASSO, и PCA не менее чем на половину рассматриваемых прогнозных горизонтов лучше прогнозировали разреженные модели (LASSO – на II и IV кварталы; PCA м на I-III кварталы), дополняемые сложными моделями с большим количеством предикторов (LASSO – I и III кварталы; PCA – IV квартал).

Итак, согласно таблице 5, на разных длинах прогнозов среди моделей LASSO лучше других осуществляет наукастинг модель (1); среди моделей PCA – модель (12). При совместном рассмотрении примененных методов, приоритетной моделью для I квартала является модель (5), для II – модель (12), для III – также модель (12), для IV – модель (10). Соответственно, модель (12) лучше других прогнозирует изменение миграционного потока на разных горизонтах. В то же время все полученные после реализации методов модели демонстрируют устойчивое преимущество перед моделью базовой. Таким образом подтвердилась выдвинутая нами гипотеза о том, что индекс поисковых запросов Google Trends по различным тематикам, связанным с миграцией, повышает точность прогнозирования миграционного потока из России в Великобританию на всех прогнозных горизонтах по сравнению с прогнозом без его использования.

Агрегирование итогов применения методов показывает преимущество моделей LASSO при прогнозировании на I квартал и PCA – на II, III и IV кварталы<sup>15</sup>; при совместном их рассмотрении вневыборочная среднеквадратичная ошибка, рассчитанная на отложенной выборке ( $MSE_{OOS}$ ), согласно таблице 5, не поднималась выше 5,335 на рассматриваемых данных и прогнозных горизонтах. При этом модели (5), (12) и (10), имеющие наименьшую ошибку при прогнозировании на I-IV кварталы соответственно, включают запросы из всех тематических блоков: «виза», «посольство», «гражданство», «работа» и «жизнь». Это может отражать как тенденцию к более рациональному и осознанному планированию миграции, когда люди заранее прорабатывают правовые, бытовые и трудовые аспекты переезда, так и растущую зависимость от онлайн-информации и официальных институтов (посольств, визовых центров) при принятии решений о долгосрочной релокации.

## Заключение

Предлагаемая модель наукастинга с использованием показателей запросов в сети Интернет позволяет осуществлять прогнозирование международной миграции из России в Великобританию с системно более низкой ошибкой, чем при отсутствии учета поискового интереса. Работа продолжает формирующуюся академическую традицию использования данных Google Trends Index в качестве альтернативы официальной статистики и вносит вклад в развитие методологической базы для прогнозирования изменяющихся во времени макропроцессов.

Теоретический обзор составляет представление о принципах отбора поисковых запросов при прогнозировании миграционных потоков, статистических методах наукастинга миграции с применением поисковых данных, а также об ограничениях, актуальных при использовании GTI. Лингвистический анализ облака слов в нынешнем виде

<sup>15</sup> Прогнозируемое изменение миграционного потока относительно предыдущего квартала моделями LASSO и PCA см. в таблице П Приложения.

может быть применен в рамках проведения последующих исследований международной миграции, в том числе при дальнейшем наукастинге миграционных потоков из России в Великобританию. Предлагаемые алгоритмы модификации методов прогнозирования временных рядов: (а) использование LASSO с подбором параметра по кросс-валидации для отбора регрессоров и формированием прогноза линейной регрессией с подстановкой будущих значений регрессоров, экстраполированных моделями SARIMA; (б) использование PCA для сокращения размерности (выбор числа компонент и лагов осуществляется кросс-валидацией) с последующей реализацией линейной регрессии на факторах, будущие значения которых оцениваются моделями SARIMA и используются для построения прогноза; – обладают важным преимуществом: объем собираемых для их реализации данных может быть существенно сокращен. Не менее половины итоговых моделей в обоих случаях оказались разреженными, т. е. требующими данных по небольшому числу поисковых запросов. Во многом благодаря данному преимуществу результат настоящего исследования – эконометрическая модель прогнозирования миграционных потоков с разными прогнозными горизонтами – может быть использован в рамках широкого спектра статистических исследований, учитывающих различные поисковые интересы макрогрупп. При этом для сохранения точности прогнозов рекомендуется регулярно переобучать модель на релевантных данных.

Среди возможных ограничений экстраполяции результатов работы можно выделить: (а) снижающуюся долю Google на российском рынке поисковых систем (рисунок 1); (б) наличие нетипичных, шоковых фаз в рамках рассматриваемого периода 2011-2023 гг. (пандемия COVID-19, резкие геополитические изменения, усилившееся после 2022 г. санкционное давление и др.), способных на структурном уровне изменить как динамику международной миграции, так и поисковые интересы пользователей сети Интернет, что затрудняет обобщить результаты на условно «нормальные» годы без подобных шоков; (в) другие общие ограничения использования Google Trends Index, подробно рассмотренные в теоретической части работы.

Дальнейшие перспективы исследования могут быть направлены как на работу с существующими ограничениями (рассмотрение данных других популярных в России поисковых систем (например, Yandex Wordstat), анализ разрыва тренда или изменения структуры модели до и после шоковых событий и др.), так и на несколько иное агрегирование методов LASSO и PCA. Сокращение размерности перебора после тематического формирования главных компонент может, например, достигаться через использование регуляризации вместо ограничения максимального числа регрессоров. Могут иметь место и другие алгоритмические корректировки, включающие декомпозицию временного ряда миграционного потока, применение экспоненциального сглаживания и др.

## Литература

- Броницкий Г.Т. (2024). Наукастинг миграции с использованием Google Trends: применение для разных стран. *Население и экономика*, 8(2), 133–154.  
Bronitsky G. (2024). Migration nowcasting using Google Trends: cross-country application. *Population and Economics*, 8(2), 133–154. (In Russ.).  
<https://doi.org/10.3897/popecon8.e119577>
- Броницкий Г.Т., Вакуленко Е.С. (2022). Прогнозирование миграции из России в Германию с использованием Google-трендов. *Демографическое обозрение*, 9(3), 75–92.  
Bronitsky G., Vakulenko E. (2022). Forecasting Migration from Russia to Germany Using Google Trends. *Demographic Review*, 9(3), 75–92. (In Russ.).  
<https://doi.org/10.17323/demreview.v9i3.16471>
- Броницкий Г.Т., Вакуленко Е.С. (2024). Применение Google Trends для прогнозирования миграции из России: агрегация поисковых запросов и учет лаговой структуры. *Прикладная эконометрика*, 73, 78–101.  
Bronitsky G., Vakulenko E. (2024). Using Google Trends to Forecast Migration from Russia: Aggregation of Search Queries and Accounting for the Lag Structure. *Applied Econometrics*, 73, 78–101. (In Russ.).  
<https://doi.org/10.22394/1993-7601-2024-73-78-101>
- Денисенко М.Б. (2003). Эмиграция из России по данным зарубежной статистики. *Мир России: Социология, этнология*, 12(3), 157–169.  
Denisenko M. (2003). Emigration from Russia According to Foreign Statistics. *Universe of Russia: Sociology, Ethnology*, 12(3), 157–169. (In Russ.).  
<https://mirros.hse.ru/article/view/5283>
- Цапенко И.П., Юревич М.А. (2022). Статистика онлайн-запросов в наукастинге миграции. *Экономические и социальные перемены: факты, тенденции, прогноз*, 15(1), 74–89.  
Tsapenko I., Yurevich M. (2022). Statistics of Online Queries in Migration Nowcasting. *Economic and Social Changes: Facts, Trends, Forecast*, 15(1), 74–89. (In Russ.).  
<https://doi.org/10.15838/esc.2022.1.79.4>
- Aitken A.C. (1935). On Least Squares and Linear Combinations of Observations. *Proceedings of Royal Statistical Society*, 55, 42–48.  
<https://doi.org/10.1017/S0370164600014346>
- 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>
- Aleshkovski I.A. (2012). International Migration, Globalization and Demographic Development of the Russian Federation. *Journal of Globalization Studies*, 3(1), 95–110.
- Avramescu A., Wiśniowski A. (2021). Now-Casting Romanian Migration into the United Kingdom by Using Google Search Engine Data. *Demographic Research*, 45, 1219–1254.  
<https://www.demographic-research.org/articles/volume/45/40>
- Bai J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1), 135–171.  
<https://doi.org/10.1111/1468-0262.00392>

- Böhme M.H., Gröger A., Stöhr T. (2020). Searching for a Better Life: Predicting International Migration with Online Search Keywords. *Journal of Development Economics*, 142, 102347. <https://doi.org/10.1016/j.jdeveco.2019.04.002>
- Bordino I., Battiston S., Caldarelli G., Cristelli M., Ukkonen A., Weber I. (2012). Web Search Queries Can Predict Stock Market Volumes. *PLOS ONE*, 7(7), e40014. <https://doi.org/10.1371/journal.pone.0040014>
- Box G.E.P., Jenkins G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- Burrige P., Wallis K.F. (1984). Unobserved-Components Models for Seasonal Adjustment Filters. *Journal of Business & Economic Statistics*, 2(4), 350–359. <https://doi.org/10.1080/07350015.1984.10509408>
- Canova F., Hansen B.E. (1995). Are Seasonal Patterns Constant Over Time? A Test for Seasonal Stability. *Journal of Business & Economic Statistics*, 13(3), 237–252. <https://doi.org/10.1080/07350015.1995.10524598>
- Carneiro H.A., Mylonakis E. (2009). Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, 49(10), 1557–1564. <https://doi.org/10.1086/630200>
- Challet D., Ayed A.B.H. (2014). Do Google Trend Data Contain More Predictability Than Price Returns? *arXiv preprint arXiv:1403.1715*. <https://doi.org/10.48550/arXiv.1403.1715>
- Choi H., Varian H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88, 2–9. <https://doi.org/10.1111/j.1475-4932.2012.00809.x>
- D’Amuri F., Marcucci J. (2017). The Predictive Power of Google Searches in Forecasting US Unemployment. *International Journal of Forecasting*, 33(4), 801–816. <https://doi.org/10.1016/j.ijforecast.2017.03.004>
- DeWaard J., Kim K., Raymer J. (2012). Migration Systems in Europe: Evidence from Harmonized Flow Data. *Demography*, 49, 1307–1333. <https://doi.org/10.1007/s13524-012-0117-9>
- De Haas H., Castles S., Miller M.J. (2019). *The Age of Migration: International Population Movements in the Modern World*. Bloomsbury Publishing.
- Dickey D.A., Fuller W.A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 74(366a), 427–431. <https://doi.org/10.1080/01621459.1979.10482531>
- Dugas A.F., Jalalpour M., Gel Y., Levin S., Torcaso F., Igusa T., Rothman R.E. (2013). Influenza Forecasting with Google Flu Trends. *PLOS ONE*, 8(2), e56176. <https://doi.org/10.1371/journal.pone.0056176>
- Dustmann C., Fabbri F. (2003). Language Proficiency and Labour Market Performance of Immigrants in the UK. *Economic Journal*, 113(489), 695–717. <https://doi.org/10.1111/1468-0297.t01-1-00151>

- 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>
- Golenvaux N., Alvarez P.G., Kiossou H.S., Schaus P. (2020). An LSTM Approach to Forecast Migration Using Google Trends. *arXiv preprint arXiv:2005.09902*.  
<https://doi.org/10.48550/arXiv.2005.09902>
- Granger C.W.J., Newbold P. (1974). Spurious Regressions in Econometrics. *Journal of Econometrics*, 2(2), 111–120.  
[https://doi.org/10.1016/0304-4076\(74\)90034-7](https://doi.org/10.1016/0304-4076(74)90034-7)
- Hotelling H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6), 417–441.  
<https://doi.org/10.1037/h0071325>
- Hyndman R.J., Athanasopoulos G. (2018). *Forecasting: Principles and Practice*. OTexts.
- Hyndman R.J., Ullah M.S. (2007). Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach. *Computational Statistics & Data Analysis*, 51(10), 4942–4956.  
<https://doi.org/10.1016/j.csda.2006.07.028>
- Johansen S. (1988). Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control*, 12(2–3), 231–254.  
[https://doi.org/10.1016/0165-1889\(88\)90041-3](https://doi.org/10.1016/0165-1889(88)90041-3)
- Jurić T. (2022). Facebook and Google as an Empirical Basis for the Development of a Method for Monitoring External Migration of Croatian Citizens. *Ekonomski Pregled*, 73(2), 186–214.  
<https://doi.org/10.32910/ep.73.2.2>
- Kristoufek L. (2015). Power-Law Correlations in Finance-Related Google Searches, and Their Cross-Correlations with Volatility and Traded Volume: Evidence from the Dow Jones Industrial Components. *Physica A: Statistical Mechanics and its Applications*, 428, 194–205.  
<https://doi.org/10.1016/j.physa.2015.02.057>
- Lazer D., Kennedy R., King G., Vespignani A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176), 1203–1205.  
<https://doi.org/10.1126/science.1248506>
- Lee J.D., Sun D.L., Sun Y., Taylor J.E. (2016). Exact Post-Selection Inference, with Application to the Lasso. *The Annals of Statistics*, 44(3), 907–927.  
<https://doi.org/10.1214/15-AOS1371>
- Leysen B., Verhaeghe P.P. (2023). Searching for Migration: Estimating Japanese Migration to Europe with Google Trends Data. *Quality & Quantity*, 57, 4603–4631.  
<https://doi.org/10.1007/s11135-022-01560-0>
- Liu K. (2024). The Measurement Errors of Google Trends Data. *Discover Data*, 2, 7.  
<https://doi.org/10.1007/s44248-024-00013-3>
- McGhee D., Heath S., Trevena P. (2012). Dignity, Happiness and Being Able to Live a ‘Normal Life’ in the UK—An Examination of Post-Accession Polish Migrants’ Transnational Autobiographical Fields. *Social Identities*, 18(6), 711–727.  
<https://doi.org/10.1080/13504630.2012.709002>

- Medeiros M.C., Pires H.F. (2021). The Proper Use of Google Trends in Forecasting Models. *arXiv preprint arXiv:2104.03065*.  
<https://doi.org/10.48550/arXiv.2104.03065>
- Moreh C., McGhee D., Vlachantoni A. (2020). The Return of Citizenship? An Empirical Assessment of Legal Integration in Times of Radical Sociolegal Transformation. *International Migration Review*, 54(1), 147–176.  
<https://doi.org/10.1177/0197918318809924>
- Ormerod P., Nyman R., Bentley R.A. (2014). Nowcasting Economic and Social Data: When and Why Search Engine Data Fails, an Illustration Using Google Flu Trends. *arXiv:1408.0699*.  
<https://doi.org/10.48550/arXiv.1408.0699>
- Pavlicek J., Kristoufek L. (2014). Can Google Searches Help Nowcast and Forecast Unemployment Rates in the Visegrad Group Countries? *arXiv preprint arXiv:1408.6639*.  
<https://doi.org/10.48550/arXiv.1408.6639>
- Pearson K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, 187, 253–318.  
<https://doi.org/10.1098/rsta.1896.0007>
- Preis T., Moat H., Stanley H. (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports*, 3(1), 1–6.  
<https://doi.org/10.1038/srep01684>
- Qi A.H., Bircan T. (2023). Can Google Trends Predict Asylum-Seekers' Destination Choices? *EPJ Data Science*, 12(1), 41.  
<https://doi.org/10.1140/epjds/s13688-023-00419-0>
- Qi H., Reed H.E., Bevelander P. (2025). Can Internet Search Data Predict Human Migration Intentions? *Comparative Migration Studies*, 13, 1–22.  
<https://doi.org/10.1186/s40878-025-00450-2>
- Rennie A., Protheroe J., Charron C., Breatnach G. (2020). Decoding Decisions. Making Sense of the Messy Middle. *Think with Google*.  
[https://www.thinkwithgoogle.com/\\_qs/documents/9998/Decoding\\_Decisions\\_The\\_Messy\\_Middle\\_of\\_Purchase\\_Behavior.pdf](https://www.thinkwithgoogle.com/_qs/documents/9998/Decoding_Decisions_The_Messy_Middle_of_Purchase_Behavior.pdf)
- Rodrigues P.M.M., Taylor A.M.R. (2004). Alternative Estimators and Unit Root Tests for Seasonal Autoregressive Processes. *Journal of Econometrics*, 120(1), 35–73.  
[https://doi.org/10.1016/S0304-4076\(03\)00206-9](https://doi.org/10.1016/S0304-4076(03)00206-9)
- Rovetta A. (2021). Reliability of Google Trends: Analysis of the Limits and Potential of Web Inveillance During COVID-19 Pandemic and for Future Research. *Frontiers in Research Metrics and Analytics*, 6, 670226.  
<https://doi.org/10.3389/frma.2021.670226>
- Schenk C. (2024). Counting Migrants in Russia: The Human Dimension of Administrative Data Production. *International Migration Review*, 58(2), 936–963.  
<https://doi.org/10.1177/01979183231154565>

- Schwarz G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461–464.  
<https://doi.org/10.1214/aos/1176344136>
- Stock J.H., Watson M.W. (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460), 1167–1179.  
<https://doi.org/10.1198/016214502388618960>
- Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 58(1), 267–288.  
<https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Wanner P. (2021). How Well Can We Estimate Immigration Trends Using Google Data? *Quality & Quantity*, 55(4), 1181–1202.  
<https://doi.org/10.1007/s11135-020-01047-w>
- Wladyka D. (2017). Queries to Google Search as Predictors of Migration Flows from Latin America to Spain. *Journal of Population and Social Studies*, 25(4), 312–327.  
<https://doi.org/10.25133/JPSSv25n4.002>
- Yang S., Santillana M., Kou S.C. (2015). Accurate Estimation of Influenza Epidemics Using Google Search Data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47), 14473–14478.  
<https://doi.org/10.1073/pnas.1515373112>
- Zou H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.  
<https://doi.org/10.1198/016214506000000735>

## Приложение

**Таблица П. Прогнозируемое изменение миграционного потока относительно предыдущего квартала моделями LASSO и PCA, чел.**

Прогнозный горизонт, кв.	I	II	III	IV
LASSO				
Модель 1	-0,360	-2,353	-3,131	-3,760
Модель 2	6,366	-2,007	-3,965	-6,368
Модель 3	10,076	-2,275	-4,136	-7,416
Модель 4	14,202	-1,856	-3,625	-9,408
Модель 5	21,099	-0,796	-1,662	-14,108
PCA				
Модель 1	-8,678	13,533	-20,745	10,252
Модель 2	-8,635	13,430	-20,701	10,329
Модель 3	-14,016	15,308	-25,837	10,427
Модель 4	-2,828	-2,366	-2,120	-2,313
Модель 5	-6,914	-0,624	-6,782	9,998
Базовая модель				
SARIMA (3, 1, 3) × (4, 1, 3) <sub>4</sub>	-2,525	-17,910	40,111	-18,663

*Источник: Расчеты автора.*