Федеральное государственное автономное образовательное

учреждение высшего образования

«Национальный исследовательский университет

«Высшая школа экономики»

На правах рукописи

Броницкий Георгий Тимурович

МОДЕЛИРОВАНИЕ И ПРОГНОЗИРОВАНИЕ МИГРАЦИИ НАСЕЛЕНИЯ С ИСПОЛЬЗОВАНИЕМ ДАННЫХ GOOGLE TRENDS

РЕЗЮМЕ
диссертации на соискание ученой степени
кандидата экономических наук

Научный руководитель:
д.э.н., доцент
Вакуленко Елена Сергеевна

JEL: J61, C53, C81

Москва – 2024

Миграция населения существенным образом влияет на демографические, экономические, социальные и иные показатели страны выбытия и прибытия. Показатели миграции подвержены внешним шокам, таким как эпидемии, военные действия, природные катаклизмы и пр. Своевременный анализ миграционных потоков, а также намерений о миграции может помочь актуализировать меры демографической политики государства. При наступлении таких событий особенно важно иметь оценку миграций без задержки во времени. Такая информация поможет своевременно реагировать на внешние шоки, минимизируя экономические риски страны. Однако сложность сбора такой статистики (не все перемещения фиксируются в статистике), изменение в методологии сбора (Чудиновских, Степанова, 2020), отсутствие достоверных данных, а также предоставление данных о миграции с временной задержкой (в некоторых случаях задержка составляет более года) затрудняют такие оценки.

По мере развития информационных технологий, в литературе появилось новое направление, связанное с анализом миграционных потоков различных стран на основе цифрового следа мигрантов в сети Интернет. Такие данные открывают новые возможности для исследователей, позволяя получать оценки различных экономических показателей с меньшей задержкой во времени, а иногда и в режиме реального времени (наукастинг). В 2006 году появился сервис, предоставляющий данные о поисковых запросах в сети Интернет – Google Trends Index (GTI), позволяющий получать статистику об интересах пользователей в различных областях (медицина, связь, бизнес и экономика). Одной из первых работ в этой области является (Choi, Varian, 2012[1], в которой авторы отмечают большой потенциал использования данных о поисковых запросах Google Trends Index для измерения интересов экономических агентов в режиме реального времени (спроса на жилье, автомобили и т. д.). В (Jun et al., 2018) отмечается прогрессивный рост использования Google Trends за 10 лет их существования на основе метаанализа 657 статей.

Было найдено всего 10 работ, основные параметры которых приведены в таблице 1. Авторы приведенных исследований производили оценки миграции при помощи Google Trends Index, две из них анализируют миграционные потоки России: внутренние — из Москвы в Санкт-Петербург (Fantazzini et al., 2021), а также внешние — из Таджикистана в

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

Россию (Цапенко, Юревич, 2022). За исключением одной работы (Wladyk, 2017), все остальные были написаны за последние 4 года, т.е. с 2020 года.

Согласно теории миграции, непосредственный переезд происходит только на третьем из пяти этапов миграции на новое место жительство (Benson-Rea, Rawlinson, 2003). Так, в исследованиях (Wladyka, 2017; Böhme, 2020; Цапенко, Юревич, 2022) в качестве объясняющих переменных используют не только фактическое значение индекса GTI, но также и его временные лаги, чаще всего используются лаги в пределах 1 года. Однако в приведенных выше исследованиях авторы также обозначают и ряд проблем, возникающих при оценке миграции с использованием GTI, которые, как правило, связаны с отсутствием доступных данных в помесячной частотности, а также с ограничениями получаемых моделей. Так, в работе (Wladyka, 2017) автор использует альтернативные источники данных (такие как муниципальный реестр резидентов) из-за сложностей исследования лаговой структуры запросов на годовых данных, представляемых официальными статистическими службами. Кроме этого, в большинстве рассматриваемых работ отбор поисковых запросов происходит экспертно, что также может влиять на качество получаемых оценок (Ormerod et al., 2014). Большая размерность множества поисковых запросов мешает одновременному включению нескольких лагов, так, в работах (Wladyka, 2017; Wanner, 2021) оцениваются модели с использованием каждого лага по отдельности. В рамках данного диссертационного исследования делается вклад в области исследования миграции с использованием данных Google Trends. Разрабатываются методы, позволяющие продвинуться в решении проблем, с которыми сталкиваются исследователи при наукастинге миграции, а также новые подходы в оценке миграции с использованием данных Google Trends Index.

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

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

Источник: (Броницкий, Вакуленко, 2024).

Объект, предмет исследования

Объектом исследования является международная миграция. **Предметом** исследования является прогнозирование миграции из различных стран в Германию с использованием данных Google Trends.

Цель и задачи исследования

Основная **цель работы** – прогнозирование миграции населения до появления данных в официальных источниках (наукастинг). В качестве источника таких данных для прогноза миграции возможно использовать информацию о поисковых запросах Google Trends Index [2] (GTI). Использование таких моделей позволяет получать оценки, опережающие данные, публикуемые официальными статистическими службами. Такой подход уже ранее применялся для прогнозирования миграции как внутренней, так и внешней (Böhme et al., 2020), в том числе и в России (Fantazzini et al., 2021; Цапенко, Юревич, 2022). В этих работах рассматриваются модели временных рядов (типа SARIMAX) с зависимой переменной — миграционный поток, а в числе объясняющих рассматриваются GTI для определенных поисковых запросов, которые, как правило, определяются экспертно. В настоящей работе предложено усовершенствование описанных ранее моделей и методов определения множества поисковых слов. Кроме этого, проверяется гипотеза о том, что миграция происходит не в момент поискового запроса, а с некоторым лагом, необходимым на подготовку документов и принятие решения.

Можно выделить **основные задачи** в исследовании:

1. Разработка способа отбора ключевых слов и словосочетаний, наилучшим образом предсказывающих факт миграции. Создание алгоритма, позволяющего отойти от экспертной оценки поисковых запросов в пользу автоматического определения множества поисковых слов;
2. Разработка и тестирование метода повышения частотности исходных данных, позволяющего получить оценку миграции в помесячной частотности, а также использовать пропущенные значения в моделях;
3. Исследование различных способов снижения размерности множества поисковых запросов с целью снизить количество объясняющих переменных и их лагов;

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

4. Оценка моделей миграции с включенными лагами GTI от 1 до 12 месяцев. Исследование качества прогнозов на случай миграции из России в Германию, оценка величины среднего лага в тематиках «посольство», «учеба» и «работа», сравнение полученных значений с результатами других исследований в предметной области.
5. Тестирование асимметрии реакции миграционных потоков на рост и падение динамики поисковых запросов, а также исследование глубины лагов в периоды шоков.
6. Проверка устойчивости полученных выводов для оценки миграции на случай других стран (для оценки миграции из России, Польши, Италии, Румынии, Испании, Болгарии в Германию), разработка алгоритмов сбора поисковых запросов для мигрантов, использующих другие, отличные от русского языки для поиска информации в сети Интернет.

Информационная база исследования

Основой для исследования послужили данные о поисковых запросах, представляемые базой данных Google Trends Index. Для формирования множества поисковых запросов была взята база готовых эмбеддингов (предобученных моделей машинного обучения), использующие в качестве изначальной словарной базы Национальный словарь, а также базу текстовых файлов Википедии.

В работе также использовались данные миграционной статистики: при анализе миграции из России в Германию исследовались данные официальных статистических служб как страны выбытия (Росстат), так и страны прибытия (статистический офис Германии). Кроме этого, использовались данные организации экономического сотрудничества и развития (ОЭСР), которая публикует ежегодные данные о числе мигрантов из стран Евросоюза, а также данные службы Eurostat. Кроме официальных статистических источников, для работы также использовались помесячные данные, полученные по запросу у статистического офиса Германии. При агрегации до годовых значений такие показатели совпадают с открытыми данными.

Теоретическая и методологическая основа исследования

Анализ миграционных потоков из России в различные страны представлен в работах (Чудиновских, Степанова, 2020; Денисенко, 2012), в которых авторы описывают историю изменений миграционных потоков, а также преимущества и недостатки официальных

данных о миграции в России. Из-за существенной задержки в публикации данных (которые иногда достигают пяти лет для развивающихся стран), различиях в методологиях подсчета исследователи все чаще прибегают к альтернативным источникам данных (Cesare et al., 2018), которые позволяют получить оценки миграции, зачастую раньше выхода официальных статистических данных. В литературе развиваются исследования, использующие новые типы экзогенных данных, способствующие ускорению получения оценок о миграции (наукастинг), которые можно разбить на следующие группы (Tjaden, 2021):

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

В работе в качестве источника таких данных выбраны поисковые запросы GTI, ввиду доступности данных в открытых источниках, а также сравнительно меньшей смещенности среди пользователей, попадающих в статистику (Tjaden, 2021). При анализе работ, связанных с оценкой миграции с применением GTI, определим наиболее важные шаги в существующих исследованиях: (1) *Выбор множества поисковых запросов* – какие ключевые поисковые запросы авторы включают в модели миграции, а также *методы их подготовки и отбора*, (2) *Лаговая структура* – наличие лагов поисковых запросов, а также их глубина, (3) *Модели оценки миграции* – какие эконометрические модели используются для оценки миграции, а также качества представленных моделей.

Опираясь на описанные выше параметры, в работе (Броницкий, Вакуленко, 2024) была составлена таблица 1, в которой представлен сравнительный анализ работ, наиболее близких к теме данного диссертационного исследования, в работе также отмечалось, что «в основном в качестве поисковых запросов для Google Trends авторы берут такие как посольство, виза, работа, жилье и добавляют регион назначения. Чаще всего авторы концентрируются на отдельных запросах, которые выбирают на основании экспертных суждений, и включают в модель только их (Böhme et al., 2020; Golenvaux et al., 2020; Wanner, 2021; Fantazzini et al., 2021; Jurić, 2022; Цапенко, Юревич, 2022). Лишь в нескольких работах изучается корреляция между миграционными потоками и различными поисковыми запросами (Wladyk, 2017; Avramescu, Wiśniowski, 2021), а затем рассматриваются модели для отдельных поисковых запросов. Важно отметить, что приведенные выше исследования,

за исключением работы (Wladyk, 2017), написаны за последние 4 года, что характеризует актуальность выбранного в исследовании направления.

Выбор множества поисковых запросов. Для использования GTI в качестве объясняющих переменных необходимо определение множества поисковых слов. Возможны различные подходы для формирования такого множества:

- экспертно, в том числе опираясь на исследования других авторов (Golenvaux et al., 2020; Wanner, 2021; Jurić, 2022)
- при помощи методов машинного обучения, позволяющих найти слова и словосочетания, наиболее близкие к слову «миграция» (Avramescu, Wiśniowski, 2021)
- опираясь на статистику использования поисковых слов в сети Интернет (например, по данным Yandex Wordstat)

В диссертационном исследовании используется комбинация второго и третьего способов: при помощи методов машинного обучения были получены изначальные словосочетания русского языка, близкие к слову «миграция». Затем эти словосочетания были использованы для поиска в системе Yandex Wordstat. Такой подход имеет наилучшее качество предсказания миграции среди схожих моделей с другими алгоритмами сбора множества поисковых запросов. При формировании множества поисковых запросов также была выполнена кластеризация запросов по наиболее популярным тематикам. К основным тематикам можно отнести запросы, связанные с поиском работы, учебы, а также посольства (Böhme et al., 2020; Fantazzini et al., 2021). Такой подход предлагается называть «мультизапросным», он используется для оценки миграции из России в Германию. Кроме этого, в работе предложен «однозапросный» алгоритм, используемый для оценки миграции из Польши, Италии, Румынии, Испании, Болгарии, России в Германию. Такой подход основан на использовании только одного поискового запроса «работа в Германии» на языке страны выбытия мигрантов и призван упростить процесс сбора данных для оценки миграции.

Подготовка данных о поисковых запросах GTI. Индекс Google Trends (GTI) отражает динамику интенсивности запросов $S_{d,r}$ пользователей по определенным ключевым словам во времени (d) в рамках выбранного региона (r). Однако индекс $S_{d,r}$ характеризует не абсолютное количество запросов по выбранному поисковому запросу $V_{d,r}$,

а его отношение ко всем поисковым запросам в данном регионе в данный день $T_{d,r}$. Таким образом, индекс $S_{d,r} = \frac{V_{d,r}}{T_{d,r}}$ показывает долю запросов, связанных с определенным поисковым запросом, относительно общего объема запросов в выбранном географическом регионе в определенный момент времени.

При работе с GTI возникает две основные проблемы: во-первых, данные являются относительными, что затрудняет их использования в моделях без преобразований; во-вторых, данные могут различаться при выборе разных временных окон. Одним из способов решения описанной проблемы является стандартизация данных Google Trends (Fantazzini et al., 2021):

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

, где $\bar{X}, \hat{\sigma}_X$ - выборочное среднее значение и стандартное отклонение случайной величины X соответственно.

Стандартизация (1) данных дает возможность проводить сравнения поисковых запросов и оценивать относительную популярность различных тем с течением времени. Кроме того, такой подход позволит применять к преобразованным временным рядам статистические инструменты, такие как метод главных компонент (PCA), поскольку все ряды имеют одинаковый масштаб — нулевое среднее и единичную дисперсию.

Снижение размерности данных. При оценке моделей миграции, появляются ограничения на количество используемых переменных. Это связано с ограниченным количеством наблюдений в данных о миграции. Доступно всего 132 наблюдения о миграции из России в Германию с начала 2011 года (статистику до этого невозможно использовать из-за изменения методологии сбора GTI). Кроме этого, необходимо выделить «контрольную» группу для оценки прогнозной силы оцениваемых моделей, а также перейти к сезонным разностям с периодом в 12 месяцев. В работе проверяется гипотеза относительно необходимости включения лагов поисковых запросов (от 1 до 12 лагов) в модели, что обусловлено «некоторым временем», прошедшим от момента поиска до самой миграции (особый интерес представляет оценка этого периода времени для разных поисковых запросов), включение лагов в 12 раз увеличит число объясняющих переменных. Приведенные ограничения в большей мере характерны для «мультизапросных» моделей при оценке миграции из России в Германию, ввиду большого количества объясняющих

переменных. Такой подход требует дополнительного исследования с выбором объясняющих переменных и уменьшении размерности данных перед их использованием в моделях:

1. На первом шаге производится декомпозиция R^2 по методу Шепли (Israeli, 2007) для модели множественной регрессии с включением всех поисковых запросов. Подход позволяет отобрать регрессоры, вносящие наибольший вклад в объясненную долю дисперсии R^2 модели миграции с GTI.
2. На втором шаге применяем метод главных компонент (PCA) только к тем поисковым запросам, которые были выделены в п.1, которые объединяются в группы (учеба, работа, посольство — основные тематики, вносящие наибольший вклад в объясненную долю дисперсии (п. 1)). Выбираются такие главные компоненты по каждой группе, при которых коэффициенты имеют наименьшие *p*-значения в моделях миграции (Айвазян, 2012).

Таким образом, в результате последовательного применения метода Шепли, а также метода главных компонент получены 3 вектора запросов, агрегирующие запросы в следующих тематиках: «работа», «учеба», «посольство». Далее производятся оценки моделей с использованием этих данных, а также включением лагов от 1 до 12 по каждой из тематик (таблица 2). Важно отметить, что при дальнейшем использовании описанного алгоритма, собственные значения и векторы оцениваются по данным только тестовой выборки.

Модели распределенных лагов. Для моделирования миграции между различными странами исследуются следующие модели: SARIMA - сезонная авторегрессионная модель скользящего среднего с сезонной компонентой, модель применяется для прогнозирования миграции без использования экзогенных данных и является «базовой» оценкой миграции; SARIMAX - разновидность SARIMA моделей, в которой числе объясняющих переменных используются GTI индекс и его лаги от 1 до 12; а также предложена модель распределенных лагов - множественная регрессия, где в качестве объясняемой переменной выступает количество мигрантов, а в качестве объясняющих переменных выступают GTI индексы и их лаги от 1 до 12. Для оценки лаговой структуры поисковых запросов добавляются лаги

GTI, т. е. оценивается модель распределенных лагов $l = 1 \ldots 12$ месяцев для переменных в сезонных разностях[3]:

$$Y_t - Y_{t-12} = \beta_0 + \sum_{k=1}^{m} \sum_{l=0}^{12} \beta_{k,l} \left(X_{k,l-1} - X_{k,l-12-1} \right) + \varepsilon_t \qquad (2)$$

где в качестве зависимой переменной в модели миграции используется показатель «прибытия иностранцев» в Германию Y_t , $t = 1,...,T$ — номер года, $X_1,...,X_k$ — объясняющие переменные (поисковые запросы GTI), $\varepsilon_t \sim iid\left(0,\sigma^2\right)$ — ошибки регрессии.

При анализе миграции из России в Германию всего было отобрано 36 GTI в качестве объясняющих переменных, при преобразовании которых с использованием PCA разложения было получено 3 вектора в тематиках «учеба», «работа», «посольство».

Таблица 2. Описание моделей в сезонных разностях.

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

Источник: (Броницкий, Вакуленко, 2024)

Разработан алгоритм определения необходимого числа лагов в модели распределенных лагов, перебирающий все возможные модели, включающие лаги от 1 до 12 месяцев (всего оценивается 8192 модели для каждой из стран). Для выбора оптимальной

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

модели используется информационный критерий AIC (Akaike, 1974: 716–723), при помощи которого выбирается наилучшая модель. Одним из условий, необходимых для сравнения моделей с применением AIC критерия, является использование одинакового числа наблюдений при оценке параметров моделей. Для SARIMAX моделей также используется AIC критерий для определения параметров модели p, d, q, где p - порядок авторегрессии, d - порядок интеграции, q - порядок скользящего среднего, параметр s=12 выбирается на основе ACF. Данные параметры также выбираются при помощи информационного критерия AIC путем перебора параметров на данных «тестовой» группы. Для SARIMAX в качестве экзогенных переменных используется сам поисковый запрос и его лаги, подобранные для модели распределенных лагов, повторный поиск наилучших лагов не производится из-за возрастания количества параметров модели для перебора и ограниченности в вычислительных ресурсах.

Метрики оценки качества моделей. Оценка качества моделей производится на основе разбиения исходных данных на «тестовую» и «контрольную» выборки, что является важным методологическим компонентом, позволяющим получить вневыборочные оценки качества моделей (в работе производятся оценки MAPE и MAE). Для всех приведенных оценок тестовая и контрольные группы не пересекаются, при этом размер «тестовой» группы берется с начала рассматриваемого периода (01.01.2011) до 01.06.2020 (а также до 01.06.2021, в зависимости от размера «контрольной» группы). Оценка моделей, а также прогнозы производятся с использованием переменных в сезонных разностях. Далее для сравнения оценок качества моделей с целью лучшей интерпретируемости результатов, производится переход к исходным переменным, таким образом, приведены именно ошибки прогноза миграции, а не их сезонных разностей. В работе используются пары тестовых и контрольных групп, соответствующие 2-м и 3-м прогнозным годам соответственно. Первая пара групп предназначена для оценки качества за 2-летний прогнозный период 01.06.2021 – 01.06.2023 (соответствующая тестовая выборка 01.01.2011 – 01.06.2021). Важно отметить, что в данном периоде большинство ограничений, связанных с эпидемией Covid-19, были сняты, что дает возможность сделать выводы относительно поведения моделей в условиях, свободных от значительных внешних шоков. Вторая пара временных периодов используется для исследования 3-летнего прогноза 01.06.2020 – 01.06.2023 (соответствующая тестовая выборка 01.01.2011 – 01.06.2020). В этот период попадает ряд ограничений, связанных с эпидемией Covid-19, из-за сложностей в перемещениях для всех стран наблюдается снижение миграционной активности в начале этого периода.

Исследование такого периода дает возможность оценить, насколько хорошо модели могут справляться с шоками, вызванными эпидемией. Данные выводы обобщаются на случай работы и с другими видами внешних воздействий, таких как военные действия, природные катаклизмы, и пр.

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

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

Для вычисления значение среднего лага L_k берутся квадраты коэффициентов при лагах $l = 1,...,12$ из модели (2) для того, чтобы корректно рассматривать случаи отрицательных коэффициентов $\beta_{k,l}$. Средний лаг считается по отдельности для каждой тематики k («работа», «учеба», «посольство»). Значения коэффициентов в оцененной модели являются случайными величинами, как и величина среднего лага (3). Для оценки доверительного интервала сначала при помощи метода Монте-Карло оценивается распределение среднего лага, а затем определяется среднее для полученного распределения при помощи процедуры бутстрэп. Большие значения среднего лага L характеризуют бóльший вклад лагов, близких к 12 месяцам, что может означать большую задержку факта миграции во времени от момента поиска. В то же время малые значения среднего лага L соответствуют более быстрой реакции миграции на изменение GTI.

Структура диссертации

Диссертационное исследование общим объемом 105 страниц состоит из введения, 3 глав, заключения, списка использованной литературы, а также приложений. Основой для первой главы является работа (Броницкий, Вакуленко, 2022), в которой описываются основные проблемы, возникающие при использовании данных о миграции из России из официальных источников. К основным из них можно отнести задержку в публикации данных, а также недоучет мигрантов из России в официальных статистических источниках. В главе описываются используемые в диссертации данные о миграции как принимаемой стороны, так и страны убытия мигранта. Приводятся методы, позволяющие подготовить данные для дальнейшего моделирования, в том числе алгоритмы увеличения частотности исходных данных о миграции. Во второй части главы производится обзор различных

источников данных цифрового следа в сети Интернет, которые возможно использовать при моделировании миграции. Приводится анализ работ, связанных с применением данных Google Trends Index при моделировании как внешней, так и внутренней миграции. Исследуются основные преимущества и недостатки моделей оценки миграции с использованием данных цифрового следа в сети Интернет.

В основе второй главы лежит исследование (Броницкий, Вакуленко, 2024). В главе на примере миграции из России в Германию исследуются различные способы моделирования миграции с использованием данных Google Trends Index. Описываются способы понижения размерности исходных данных, позволяющие учитывать не только текущие значения поисковых индексов, но и их лаги от 1 до 12 месяцев. Сравниваются следующие типы моделей: без лагов; с лагами; по различным тематикам (таким как «работа», «учеба», «посольство») по отдельности; по всем тематикам одновременно. Кроме этого, оценивается величина среднего лага, характеризующего время, проходящее от момента поиска информации в сети Интернет до фактической миграции. Делаются выводы относительно эффективности используемых моделей в сравнении с базовыми SARIMA моделями без учета внешних данных о миграции.

Третьей главой является работа (Броницкий, 2024), которая обобщает все предыдущие работы для анализа миграции из различных стран в Германию. В главе описываются методы сбора множества поисковых слов на случай работы с различными языками. Проверяется гипотеза относительно увеличения предсказательной силы моделей оценки миграции при добавлении данных GTI, особенно в периоды внешних шоков, в которые SARIMA модели без экзогенных переменных показывают худшее качество. Также сравнивается качество моделей, с использованием запросов в нескольких тематиках с моделями, построенными по одному поисковому запросу и его лагах.

Научная новизна исследования

На данный момент в литературе по прогнозированию миграции с помощью Google Trends не удалось найти работ, где бы предлагалось решение по автоматизированному методу сбора множества поисковых запросов, а также по агрегированию поисковых запросов и изучению их лаговой структуры:

1. В диссертационной работе предполагается усовершенствование сбора множества поисковых запросов, которое в большинстве исследуемых работ производится экспертно (Wanner, 2021; Jurić, 2022). В работе применяются методы машинного

обучения (NLP), позволяющие отойти от экспертной оценки к автоматическому сбору множества поисковых запросов, характеризующих намерение мигрировать;

2. Разработан алгоритм, позволяющий увеличить частотность исходных данных о миграции от годовых до помесячных значений. В отличие от MIDAS (Mixed Data Sampling) моделей, применяемых для работы с данными смешанной частотности, такой подход позволяет исследовать лаговую структуру поисковых запросов.
3. Предложен метод агрегации ключевых поисковых запросов в индексы по тематикам запросов, которые наилучшим образом будут описывать миграцию. Тем самым, рассматривается сразу группа поисковых запросов, которая характеризует различные цели миграции (трудовая, учебная и т. д.), что позволяет уменьшить размерность при одновременном включении разных типов запросов и изучении лаговой структуры по ним;
4. В диссертационном исследовании производится усовершенствование существующих методов работы с лагами поисковых запросов, в том числе описанных в работе (Wanner, 2021) за счет добавления в модель сразу нескольких временных лагов GTI (модели распределенных лагов). Полученные при сравнении различных стран выводы говорят о преимуществе моделей сразу с несколькими лагами относительно моделей с только одним лагом.
5. Произведена оценка величины среднего лага по тематикам запросов, которая характеризует количество времени, проходящее с момента поиска информации в сети Интернет до отображения в миграционной статистике момента миграции. Также был разработан алгоритм вычисления доверительного интервала для получаемых оценок среднего лага при помощи метода Монте-Карло;
6. Разработан алгоритм сбора поисковых запросов на случай анализа миграции из различных стран (исследуется миграция из шести различных стран в Германию), произведенный анализ указывает на улучшение предсказательной силы в моделях с использованием GTI, кроме этого, показано, что модели распределенных лагов имеют лучшую предсказательную силу сравнительно с моделями, где использовались только отдельно по одному лагу (Wanner, 2021).

Положения, выносимые автором на защиту

В работе предлагается методика оценки миграционной статистики с минимальной задержкой во времени, т. к. называемый наукастинг статистики миграции. Хотя подобная

идея и описывалась в более ранних работах, она неприменима в явном виде к показателям Росстата и статистических служб широкого круга стран.

1. Предложен алгоритм сбора множества поисковых запросов, позволяющий автоматически определять необходимые для дальнейшего прогноза поисковые слова.
2. Предложена методология повышения частотности данных за счет выделения сезонной компоненты в данных о миграции, позволяющий получить оценку помесячных показателей из годовых значений.
3. Исследованы различные эконометрические модели прогноза миграции (SARIMA, SARIMAX, ETS – модель, модели распределенных лагов), показано, что использование экзогенных переменных, таких как GTI по различным тематикам, улучшает предсказательную силу моделей для всех рассматриваемых в работе стран.
4. Показано, что использование сразу нескольких лагов GTI уменьшает ошибки предсказания моделей сравнительно с моделями, в которых каждый лаг взят по отдельности. Так, для 5 из 6 исследуемых стран модель сразу с несколькими лагами демонстрирует наилучшее качество.
5. Предложена методология оценки времени, проходящего от момента изменения динамики поисковых запросов до изменения миграции при помощи величины среднего лага. Средний лаг L для тематики «посольство» $L = 5.6$, 95% доверительный интервал $[3.64; 7.92]$, в то время как для тематик «учеба» и «работа» $L = 8.0\ [5.36; 10.8]$, и $L = 6.5\ [4.72; 8.21]$ соответственно.
6. В результате сравнения прогнозных качеств моделей на примере миграции стран в Германию делается вывод, что в случае прогноза для периодов с внешними воздействиями (такими как эпидемия Covid-19, начало СВО в России) модель распределенных лагов показывает лучшую предсказательную силу по сравнению с SARIMAX моделями для всех рассматриваемых в работе стран. В случае миграции из Польши, Италии, Румынии, Испании в Германию метрика ошибки модели MAPE на 3-летнем интервале оказывается в более чем 2 раза ниже сравнительно с SARIMAX моделями.

Теоретическая и практическая значимость исследования

Теоретические результаты диссертационного исследования заключаются в развитии научных подходов по оценки миграции с минимальной задержкой во времени (наукастинг) с использованием данных цифрового следа мигрантов в сети Интернет. Такие подходы

могут быть использованы для оценки как миграционных потоков, так и других экономических показателей, в которых необходима минимальная задержка во времени. Они также могут быть использованы при подготовке изменений в методологии оценки миграции, например, при помощи предложенных методов возможно оценить расхождения в данных о миграции за счет нелегальных мигрантов.

Также описываемые в работе подходы могут быть использованы в качестве материалов лекций для студентов, занимающихся наукастингом экономических показателей. Данные лекции помогут студентам получить знания об основных способах сбора и обработки данных, позволяющих проводить дальнейшее моделирование экономических процессов. Среди различных источников, в которых возможно получать данные о цифровом следе в сети Интернет, основной фокус производится на работе с Google Trends Index.

Практическая значимость исследования заключается в следующем: в ходе исследования была выявлена проблема недоучета данных о мигрантах в официальной статистике России, подтверждена необходимость использования «зеркального метода» оценки миграции на основе регистрируемых в странах въезда данных. При оценке количества мигрантов с использованием данных GTI необходимо использовать не только текущие данные о запросах, но одновременно и их лаги. При оценке внешней миграции в различные страны возможно использовать тематики «посольство», «работа» и «учеба».

Результаты исследования могут быть приняты во внимание Главным управлением по вопросам миграции МВД России при разработке новой методологии учета мигрантов, методов сбора такой статистики, а также правительством России и профильными министерствами при разработке мер демографической политики в области миграции.

Ограничения исследования

Стоит отметить о некоторых ограничениях предложенных подходов, связанных с данными о миграции, а также с данными о поисковых запросах Google Trends Index и используемых моделей. Во-первых, в исследовании используются данные о совокупных потоках миграции в различные страны без выделения целей миграции (трудовая, учебная, воссоединение семей и пр.); во-вторых, возможен недоучет мигрантов в официальных статистиках, используемых для построения моделей, так как в данных могут не учитываться мигранты с двойным гражданством, а также нелегальные мигранты; в-третьих, при

использовании данных Google Trends Index для оценки миграции появляется смещение в сторону мигрантов, использующих сеть Интернет для поиска информации, кроме этого, есть ряд стран, где рассматриваемая поисковая система не используется; в-четвертых, в работе исследуется асимметрия реакций на внешние шоки (такие как пандемии и прочие кризисы), однако для большинства моделей не учитывается возможность разной скорости реакции при внешних факторах, и, как следствие, возможное расхождение полученных данных с истинными объемами миграции. Кроме этого, асимметрия может быть также вызвана тем, что экономические агенты могут искать информацию в сети Интернет из-за их широкого обсуждения в СМИ или другого события, подверженного высокой огласке среди населения, а не из-за истинного желания мигрировать; в-пятых, при обобщении полученных алгоритмов на широкий список стран используется запрос «работа в Германии» ввиду отсутствия необходимости в получении визы, однако возможно также исследовать второй тип запросов «посольство Германии» для тех стран, в которых необходимо получения визы для миграции.

Степень достоверности и апробация результатов исследования

Результаты проведенных исследований в рамках работы над диссертацией были представлены к обсуждению на следующих **конференциях**:

- XXIV Ясинская (Апрельская) международная научная конференция по проблемам развития экономики и общества (выступление 04.04.2023). Секция демография и рынки труда. Доклад: Прогнозирование международной миграции с использованием Google трендов.
- Пятый Российский экономический конгресс 2023 (выступление 11.09.2023). Секция прикладная эконометрика. Доклад: Исследование лаговой структуры индексов Google Trends в задаче прогнозирования миграции из России.
- Седьмая конференция консорциума журналов (выступление 25.10.2023). Сессия современные методы и данные в демографическом анализе. Доклад: Применение Google Trends для прогнозирования миграции из России.

- XI международная конференция «Многомерный статистический анализ, эконометрика и моделирование реальных процессов» имени С.А. Айвазяна (выступление 25.06.2024). Секция Многомерный статистический анализ и эконометрика. Доклад: Наукастинг миграции с помощью Google Trends Index.

Результаты диссертационного исследования **опубликованы в следующих статьях**:

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

- 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. DOI:10.22394/1993-7601-2024-73-78-101. НИУ ВШЭ список B. Scopus Q3.
- Броницкий Г. Т. (2024). Наукастинг миграции с использованием Google Trends: применение для разных стран. *Population and Economics*. (в печати). НИУ ВШЭ список C. Scopus Q2.

Список литературы

Айвазян С. А. (2012). Анализ качества и образа жизни населения (эконометрический подход). М: Наука, 432 с.

Броницкий Г. Т. (2024). Наукастинг миграции с использованием Google Trends: применение для разных стран. *Population and Economics*. (в печати).

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

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

Воробьева О. Д., Гребенюк А. А. (2017). Сравнительный анализ отечественной и зарубежной статистической информации об эмиграции граждан России. *Вопросы статистики*, 1 (9), 64–73.

Денисенко М. Б. (2012). Эмиграция из России в страны дальнего зарубежья. Демоскоп Weekly. No 513- 514.

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

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

Чудиновских О. С., Степанова А. В. (2020). О качестве федерального статистического наблюдения за миграционными процессами. Демографическое обозрение, 7 (1), 54–82. DOI: 10.17323/demreview.v7i1.10820.

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

Avramescu A., Wiśniowski A. (2021). Now-casting Romanian migration into the United Kingdom by using Google Search engine data. *Demographic Research*, *45*, 1219–1254. DOI: 10.4054/DemRes.2021.45.40.

Bengtsson L., Lu X., Thorson A., Garfield R., von Schreeb J. (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. PLoS Medicine, 8 (8). DOI: 10.1371/journal.pmed.1001083.

Benson-Rea M., Rawlinson S. (2003). Highly skilled and business migrants: Information processes and settlement outcomes. International Migration, 41 (2), 59–79. DOI: 10.1111/1468-2435.00235.

Böhme M. H., Gröger, A., Stöhr T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, *142*, 102347. DOI:10.1016/j.jdeveco.2019.04.002.

Chi G., State B., Blumenstock J.E., Adamic L. (2020). Who Ties the World Together? Evidence from a Large Online Social Network. In: Cherifi, H., Gaito, S., Mendes, J., Moro, E., Rocha, L. (eds) Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence, 882. DOI: 10.1007/978-3-030-36683-4_37.

Choi H., Varian H. (2009). Predicting the Present with Google Trends. Technical report, Google. [Cited 1 April 2012.] Available from: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Choi H., Varian H. (2012). Predicting the present with Google Trends. Economic record, 88, 2–9. DOI: 10.5018/economics-ejournal.ja.2018-34.

Fantazzini D., Pushchelenko J., Mironenkov A., Kurbatskii, A. (2021). Forecasting internal migration in Russia using Google Trends: evidence from Moscow and Saint Petersburg. Forecasting, 3 (4), 774–803. DOI:10.3390/forecast3040048.

Gabrielli, L., Deutschmann, E., Natale, F., Recchi, E., Vespe, M. (2019). Dissecting global air traffic data to discern different types and trends of transnational human mobility. EPJ Data Science, 8(1), 26.

Golenvaux N., Alvarez P. G., Kiossou H. S., Schaus P. (2020). An LSTM approach to Forecast Migration using Google Trends. DOI: 10.1145/1122445.1122456.

Israeli O. (2007). A Shapley-based decomposition of the R-square of a linear regression. *Journal of Economic Inequality*, 5, 199 –212. DOI: 10.1007/s10888-006-9036-6.

Jun S. P., Yoo H. S., Choi S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, 130, 69–87. DOI: 10.1016/j.techfore.2017.11.009.

Jurić T. (2022). Facebook and Google as an Empirical Basis for the Development of a Method for Monitoring External Migration of Croatian Citizens. *Ekonomski pregled*, *73* (2), 186–214. DOI: 10.32910/ep.73.2.2.

Kim J., Sîrbu A., Giannotti F., Gabrielli L. (2020). Digital Footprints of International Migration on Twitter (pp. 274–286). DOI: 10.1007/978-3-030-44584-3_22.

Ormerod P., Nyman R., Bentley A. (2014). Nowcasting economic and social data: when and why search engine data fails, an illustration using Google Flu Trends. DOI:10.48550/arXiv.1408.0699

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

Wanner, P. (2021). How well can we estimate immigration trends using Google data?. Quality & Quantity, 55 (4), 1181–1202. DOI: 10.1007/s11135-020-01047-w.

Wladyka D. K. (2017). Queries to Google Search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies [JPSS]*, *25* (4), 312–327.

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