Открытый лекторий Экономической школы ФЭН

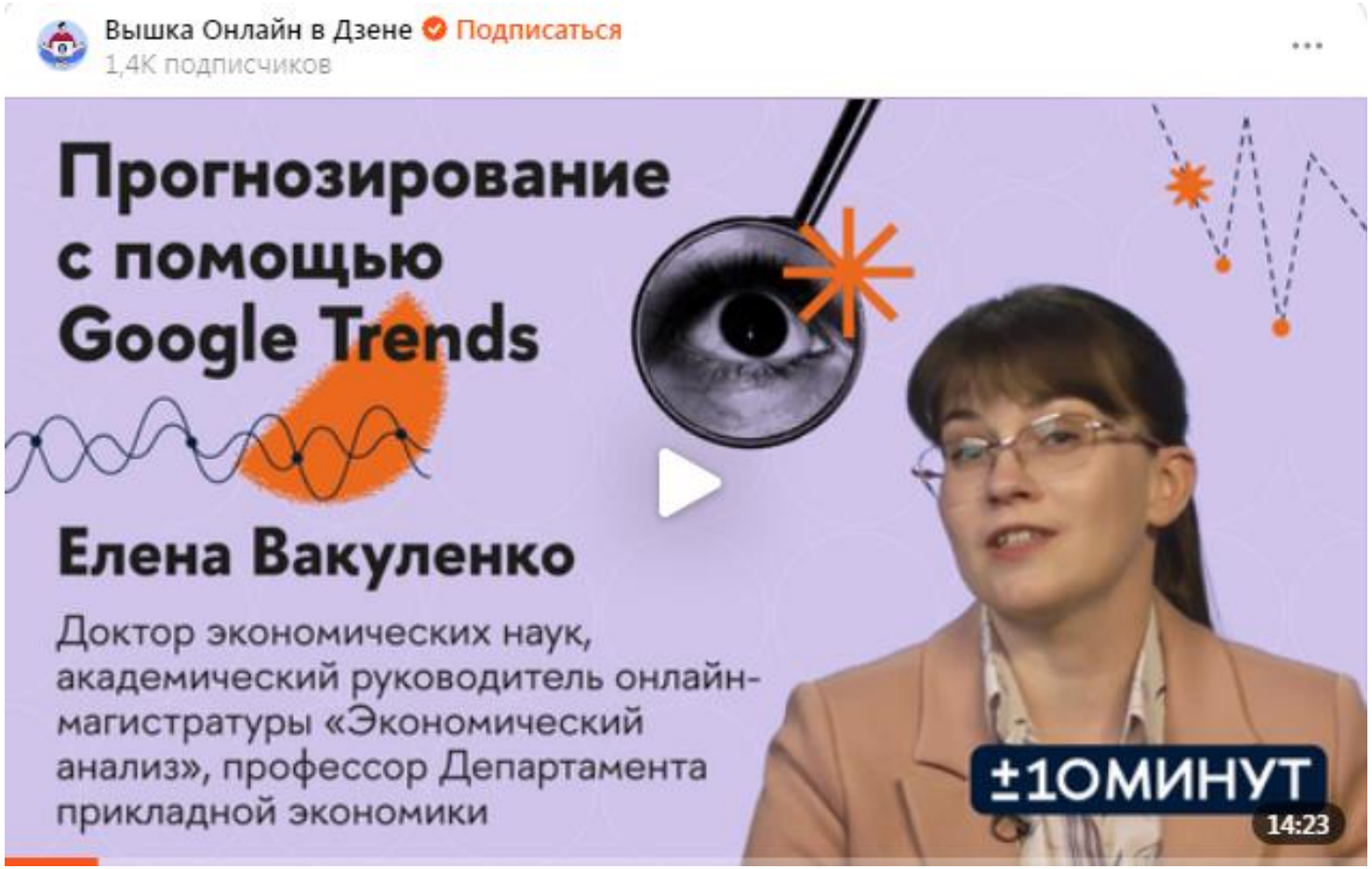
Прогнозирование с помощью гугл-трендов

Вакуленко Е.С., д.э.н., профессор

г. Москва, 12 декабря 2023

План

- Что такое статистика поисковых запросов в сети Интернет?
- Примеры применения Google Trends Index для наукастинга и прогнозирования.
- Простейшие модели для прогнозирования.
- Основные проблемы, которые возникают у исследователей. при работе с поисковыми запросами.
- Модельные лайфхаки на примере прогнозирования миграционных потоков с помощью Google Trends.



https://youtu.be/vEbp_uajcTQ

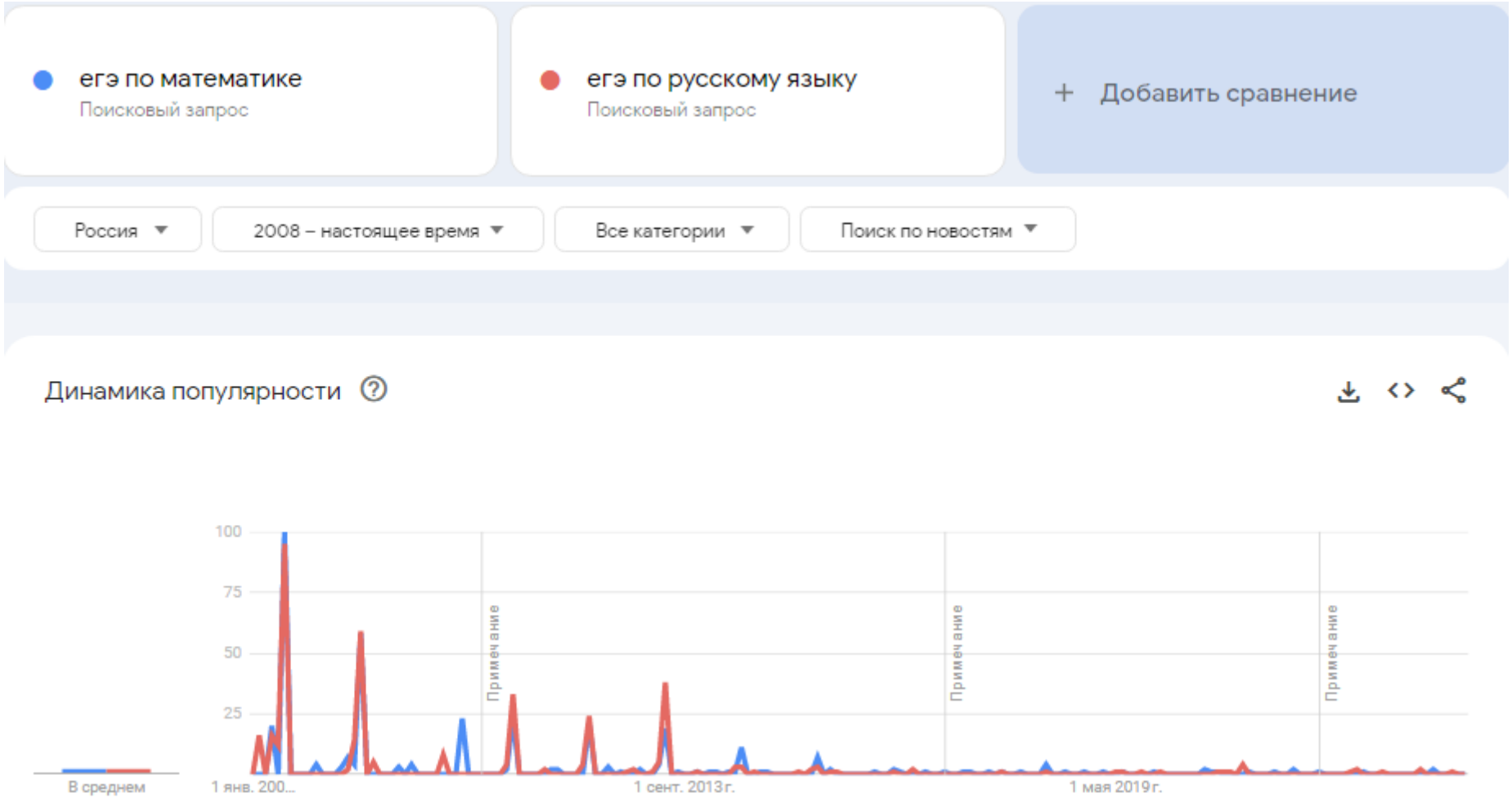
Статистика поисковых запросов

- Агрегированная информация о поиске в сети Интернет, приведенная по ключевым словам запросов.
- Отличный инструмент для анализа трендов, сезонности и прогнозирования спроса.
- Примеры:
  - Google Trends Index: https://trends.google.ru/trends/
  - Яндекс Wordstat: https://wordstat.yandex.ru/

Google Trends

- Числа обозначают уровень интереса к теме по отношению к наиболее высокому показателю в таблице для определенного региона и периода времени.
- 100 баллов означают наивысший уровень популярности запроса, 50 – уровень популярности запроса, вдвое меньший по сравнению с первым случаем.
- 0 баллов означает местоположение, по которому недостаточно данных о рассматриваемом запросе.

Google Trends



Google Trends: ЕГЭ по математике

Популярность по субрегионам

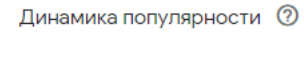
Субрегион

1	Республика Тува	100
2	Республика Дагестан	90
3	Чеченская Республика	84
4	Республика Ингушетия	83
5	Республика Калмыкия	73

Google Trends: ЕГЭ по математике

Похожие запросы  В тренде 

1	математика	Сверхпопулярность
2	решу егэ по математике	Сверхпопулярность
3	егэ 2015 по математике	Сверхпопулярность
4	решу егэ	Сверхпопулярность
5	егэ 2015	Сверхпопулярность

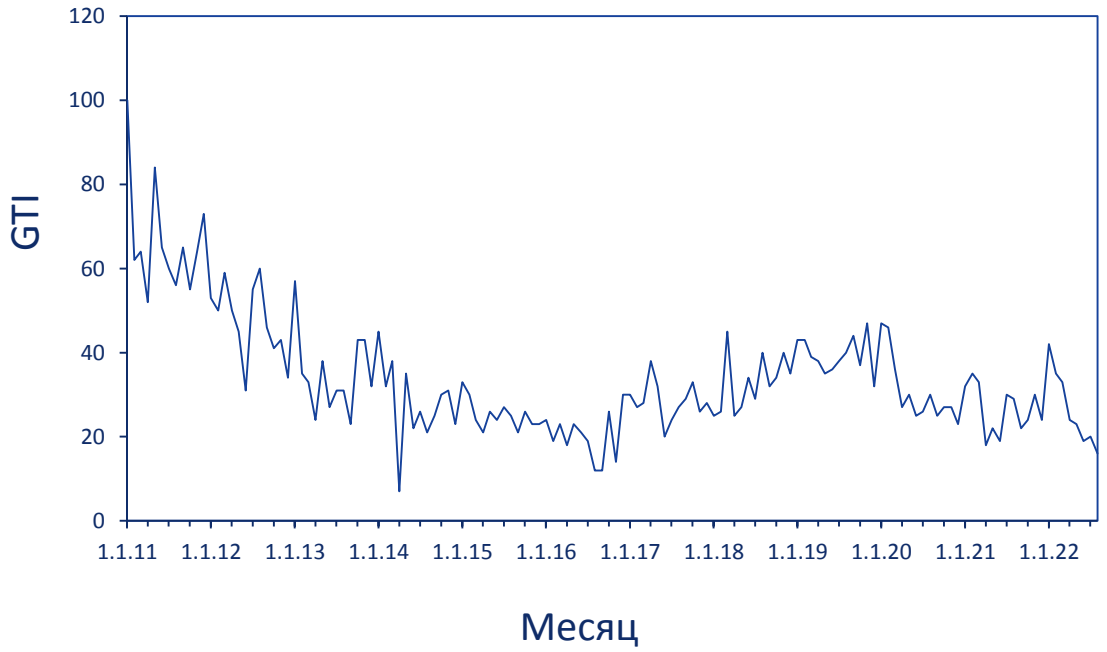
Динамика популярности 

Google Trends Index

Особенности GTI:

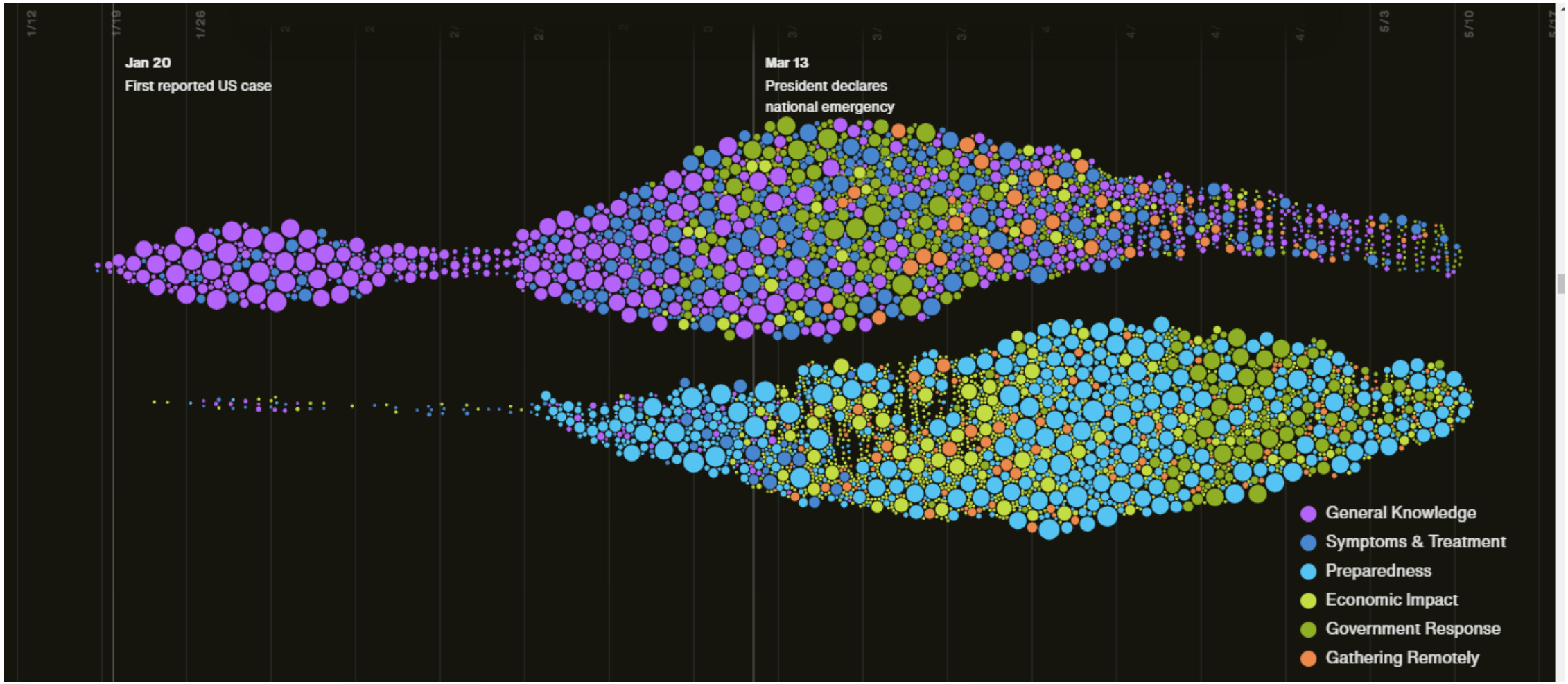
- Необходимо самостоятельно составить список запросов;
- Индекс формируется в заданной тематике, регионе;
- Нормирован в выбранном временном окне;
- Изменилась методология сбора в 2011 году

Google Trends Index по запросу «Работа в Германии» 01.01.2021 — 01.08.2022 гг.



Searching COVID-19: https://searchingcovid19.com/

Представлена визуализация наиболее популярных поисковых запросов, связанных с коронавирусом, в США в период с 20 января по 24 апреля 2020 года.



Яндекс Wordstat

Яндекс
подбор слов

Директ Справочник Метрика Рекламная сеть Маркет ещё

егэ по математике | Подобрать

По словам | По регионам | История запросов | Все регионы

Все | Десктопы | Мобильные | Только телефоны | Только планшеты

Последнее обновление: 10.12.2023

Что искали со словом «егэ по математике» — 1 438 892 показа в месяц

Статистика по словам	Показов в месяц
егэ математик	1 438 892
егэ математика	1 438 892
егэ математика профиль	504 670
математик профиль егэ	504 670
решу егэ математик	454 796
решу егэ математика	454 796
егэ математика база	337 835
егэ математика 2024	330 660
вариант егэ математика	185 130
решу егэ математик профиль	167 296
решу егэ математика профиль	167 296

Запросы, похожие на «егэ по математике»

Статистика по словам	Показов в месяц
математика база	376 178
задача +по математике	681 735
ребусы +по математике	17 308
+как найти вероятность +в математике	2 047
+как обозначается площадь +в математике	1 709
задание +по математике	916 433
многочлен +в математике	2 194
интересные задания +по математике	6 665
решуогэ математике	1 268
статград математика	46 771
решить задачу +по математике	96 461

Яндекс Wordstat

Яндекс
подбор слов

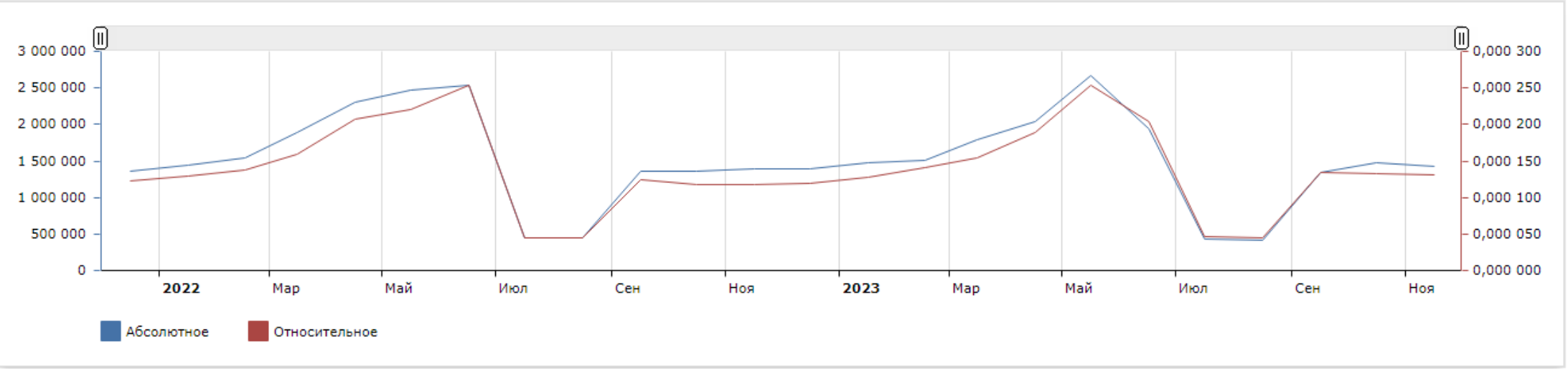
Директ Справочник Метрика Рекламная сеть Маркет ещё

егэ по математике Подобрать

По словам По регионам История запросов Россия

История показов по фразе «егэ по математике»

Группировать по: месяц неделя | Все Десктопы Мобильные Только телефоны Только планшеты



Период	Абсолютное	Относительное	Период	Абсолютное	Относительное

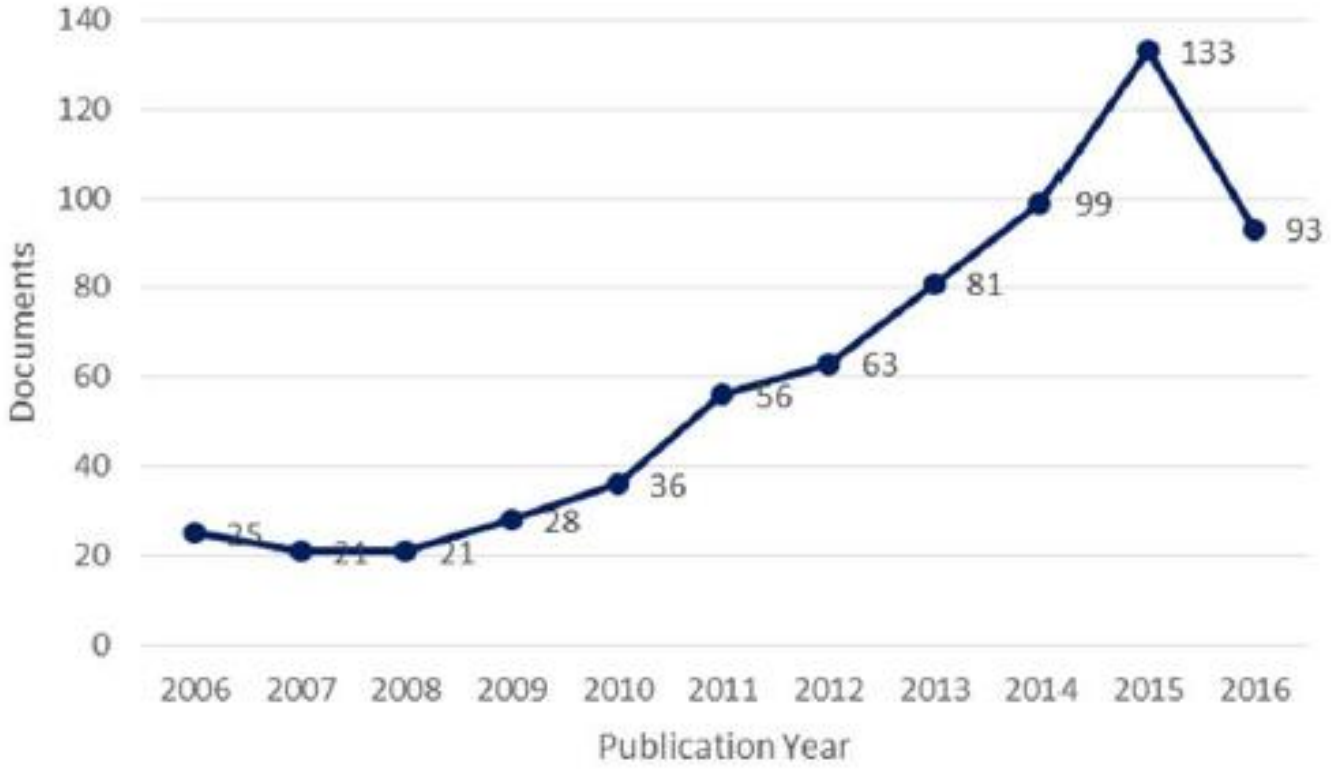
Сравнение поисковой статистики

Google Trends	Яндекс Wordstat
Ежемесячная статистика с 2004 года. Ежедневная или ежеминутная, если за день, неделю или месяц.	Ежемесячная за последний год
Индекс (100 балльная шкала)	Абсолютные количество
Не различаются устройства	Различаются устройства (телефон, планшет и тд)
Доступна по странам, регионам и городам	Доступна по странам, регионам и городам
Поиск по словам, картинкам, новостям, YouTube	Поиск по словам
Показаны похожие запросы	Показаны похожие запросы

Примеры применения Google Trends Index для наукастинга и прогнозирования

- Статистика Google Trends (GTI) Index доступна с 2004 года. Ее применят в прогнозированияи наукастинга в различных областях (IT, связь, медицина, здравоохранение, бизнес и экономика).
- Первая работа (Ettredge et al., 2005) по применению поисковых запросов для прогнозирования **уровня безработицы** в США.
- Одной из первых называют работу (Ginsberg et al., 2009), в которой прогнозируется **распространение гриппа** с помощью GTI.
- В (Jun et al., 2018) отмечается прогрессивный рост использования Google Trends за 10 лет их существования на основе метаанализа 657 статей.
- Впервые о применимости GTI для прогнозирования **миграции**, в частности туристических потоков в Гонконг из ряда стран, в том числе во время проведения летней Олимпиады (Beijing Olympics) в августе 2008 года, было написано в работе (Choi, Varian, 2009).

Популярность GTI в научных статьях



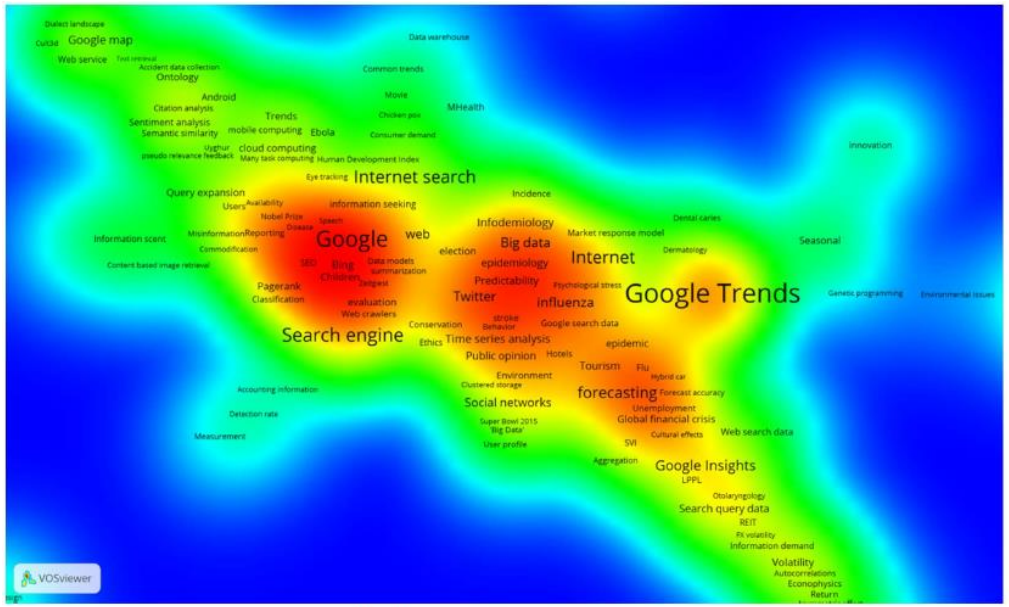


Fig. 2. Density visualization of author keywords.

Рис. Динамика количества публикаций с применением Google Trends

Тематики статей с применением GTI

Table 3
Comparison of the number of documents by ASJC code, First Author Country and Affiliation.

Rank	ASJC Code		First author country		Affiliation	
	Description	Documents	Description	Documents	Description	Documents
1	1705: Computer networks and communications	42	United States	192	Harvard University | USA	14
2	1700: Computer science(all)	40	China	50	Google Inc. | USA	12
3	1710: Information systems	35	Germany	38	Chinese Academy of Sciences | CHN	10
4	1706: Computer science applications	30	United Kingdom	34	Children's Hospital Boston | USA	9
5	1100: Agricultural and biological sciences(all)	28	Italy	24	Johns Hopkins University | USA	8
6	2002: Economics and econometrics	27	Australia	22	University of Pennsylvania | USA	7
7	2700: Medicine(all)	26	Spain	21	University of Sydney | AUS	
8	1712: Software	24	Canada	19	Carnegie Mellon University | USA	
9	2739: Public health, environmental and occupational health	20	South Korea		George Washington University | USA	6
10	1702: Artificial intelligence	15	Taiwan		National University of Singapore | SGP	
11	2718: Health informatics		India	15	University of Michigan | USA	
12	1000: General		Greece	12	University of Melbourne | AUS	
13	2725: Infectious diseases	12	France	11	Virginia Tech | USA	
14	1403: Business and international management	11	Netherlands	10	University of California San Diego | USA	
15	1709: Human-computer interaction		Japan	9	University of Regensburg | DEU	
16	1703: Computational theory and mathematics	10	Ireland	8	University di Verona | ITA	
17	2200: Engineering(all)		Austria	7	City University of Hong Kong | HKG	5
18	2728: Clinical neurology	9	Czech Republic		Vanderbilt University Hospital | USA	
19	1704: Computer graphics and computer-aided design	8	Thailand		University of Washington | USA	
20	2741: Radiology nuclear medicine and imaging		Turkey	6	San Diego State University | USA	
	2746: Surgery				University of Warwick | GBR	
	3315: Communication				University of Genoa | ITA	
					University of Alberta | CAN	
					Santa Fe Institute | USA	

Источник: (Jun et al., 2018). Всего 657 статей.

Основные проблемы использования GTI для прогнозирования

- Подбор ключевых слов. Какие запросы брать?
- Что делать, если ключевых запросов много, а данных мало? Как агрегировать данные запросов?
- Какой временной лаг выбрать? Как быстро поисковые запросы отражаются в моделируемых показателях?
- Асимметрия реакции. Одинаково ли связаны всплески и падения запросов с интересуемым показателем?

Простейшие модели для прогнозирования с помощью GTI

1) Авторегрессии

AR(1): $y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t$

AR(p): $y_t = \alpha_0 + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \varepsilon_t$

2) Модель распределенных лагов $y_t = \beta_0 + \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + \varepsilon_t$

Исследование лаговой структуры индексов Google Trends в задаче прогнозирования миграции из России

Вакуленко Е.С., д.э.н., профессор
Броницкий Г.Т., аспирант

Будет опубликована в 1 номере 2024 года журнала «Прикладная эконометрика»

г. Москва, 2023

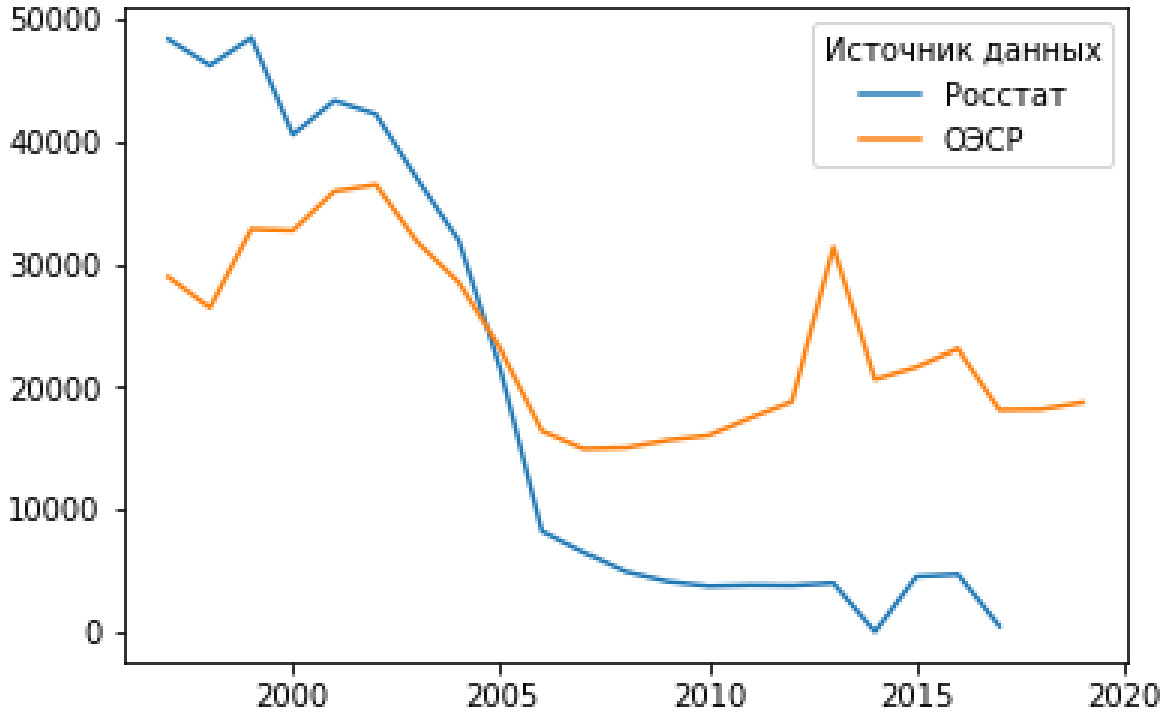
Исследовательский вопрос

- задержка в публикации миграционных данных →
- изменение методологий подсчета мигрантов →
- годовая частотность данных

Как оценивать миграцию с меньшей задержкой во времени?

Источники данных о миграции

- Росстат
- Миграционные офисы принимающих стран
- Отчет ОЭСР[1]
- Опрос Гэллапа (GWP)



1. Организации экономического сотрудничества и развития

Альтернативные источники данных

- Положение SIM-карт
- Данные о местоположении в социальных сетях
- IP-адреса почтовых сервисов
- Активность поиска в сети Интернет

Критика таких подходов (Чудиновских, 2018; Чудиновских, 2020; Tjaden, 2021)

Google Trends Index в задачах оценки миграции

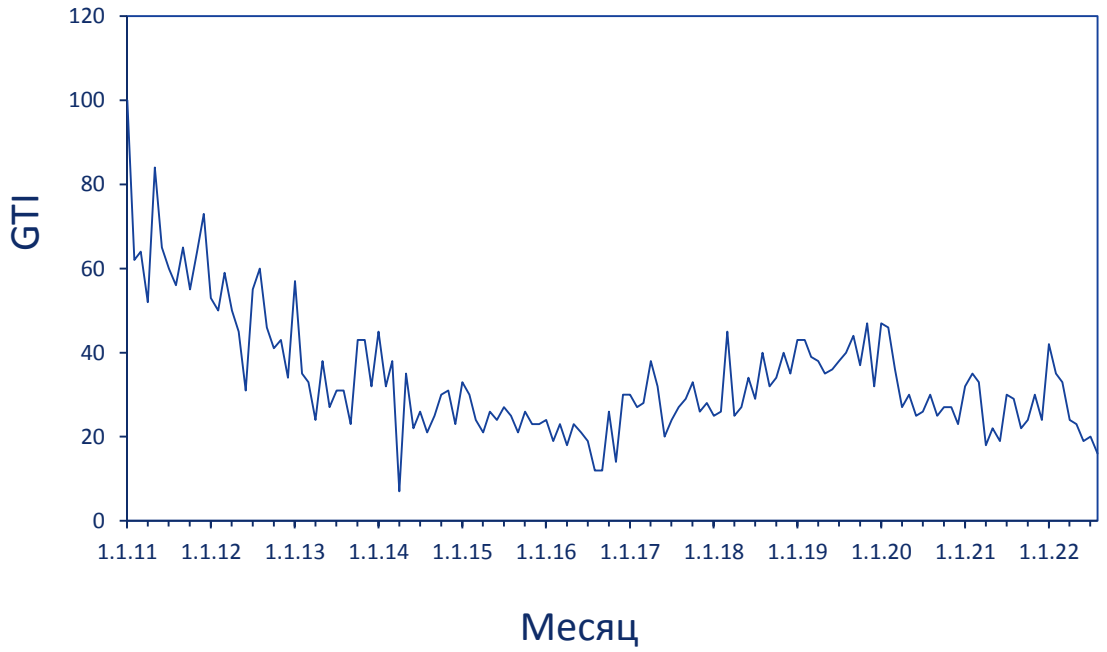
Автор	Страны	Года	Лаги	Модели
Wladyk (2017)	Из Аргентины, Перу, Колумбии в Испанию	2005-2010	Колумбия: 9-ый лаг работа и посольство; Аргентина: 4-ый лаг работа, 8-ой лаг посольство; Перу: 9-ый лаг посольство.	Парная регрессия на переменных в разностях
Böhme et al. (2020)	Из 101 страны в 35 стран ОЭСР	2004–2015	Годовые данные, 1 лаг	FE, панель стран, гравитационная модель
Golenvaux et al. (2020)	Из работы Böhme, Gröger, Stöhr (2020)	2004-2015	Годовые данные, 1 лаг	Нейронные сети (LSTM) лучше прогнозируют, чем, ANN, линейная гравитационная модель
Wanner (2020)	В Швейцарию из Франции, Италии, Германии и Испании	2006-2016	До 3-х лет	Линейная регрессия с лагами GTI
Fantazzini et al. (2021)	Россия, Москва-Санкт-Петербург	2009-2018	1 месяц	ARIMAX, SARIMAX, VECM
Avramescu, Wisniowski (2021)	Из Румынии в Великобританию	2012-2019	Нет лагов, скользящее среднее за 12 месяцев	ARIMAX
Jurić (2022)	Из Хорватии в Австрию и Германию	2004-2020	-	Парная регрессия
Цапенко, Юревич (2022)	Из Таджикистана, Киргизии, Узбекистана в Россию	2015-2020	7 или 11 месяцев	ARIMAX

Google Trends Index

Особенности GTI:

- Необходимо самостоятельно составить список запросов;
- Индекс формируется в заданной тематике, регионе;
- Нормирован в выбранном временном окне;
- Изменилась методология сбора в 2011 году

Google Trends Index по запросу «Работа в Германии» 01.01.2021 — 01.08.2022 гг.



Стандартизация GTI

Одним из способов решения описанной ранее проблемы является стандартизация данных Google Trends (Fantazzini et al., 2021) (1):

$$Z = \frac{X - E[X]}{\sigma(X)}$$

- Позволяет сравнивать разные тематики
- Применять статистические инструменты, такие как PCA
- Все переменные имеют нулевое среднее и единичную дисперсию

Поисковые запросы

Цель:

Найти поисковые запросы, наилучшим образом описывающие желание мигрировать

Алгоритм выбора поисковых запросов описан в работе (Броницкий, Вакуленко, 2022)

Источник поисковых запросов	Найденные запросы	MAPE	MAE	Значимые запросы
Национальный корпус	20	14,7	271,3	визовый Германия право Германия гражданство Германия
Википедия	25	14,4	267,4	виза Германия билет Германия посольство Германия
Тайга корпус	25	14,4	265,8	виза Германия внж Германия иммиграция Германия
Yandex Wordstat	32	10,4	183,5	шенгенская виза в Германию посольство Германии в Москве работа в Германии

Снижение размерности данных

1. **Метод главных компонент (PCA)** — осуществляет переход к новым переменным, позволяет агрегировать переменные, оставив регрессоры, соответствующие главным компонентам с наибольшей дисперсией, т.е. с минимальной потерей информации.
2. **Декомпозиция R^2 по методу Шепли** (Israeli, 2007) - подход позволяющий отобрать регрессоры, вносящие наибольший вклад в объясненную долю дисперсии, R^2 модели миграции с GTI.
3. **Метод главных компонент** по основным тематикам — комбинированный подход, использующий преимущества первого и второго: берем факторы, объединенные в группы (учеба, работа, посольство — основные тематики, вносящие наибольший вклад в объясненную долю дисперсии (п.2)). Выбираем такие главные компоненты по каждой группе, при которых коэффициенты имеют наименьшие p–value в моделях миграции (Айвазян, 2012).

Модель с распределенными лагами для сезонных разностей

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

- Y_t – объясняемая переменная, показатель «прибытия иностранцев» из России в Германию
- $X_1 \cdots X_k$ – объясняющие переменные (PCA – вектора и их лаги по тематикам «работа», «учеба», «посольство»)
- $\varepsilon_t \sim \mathrm{idd}(0, \sigma^2)$ - ошибка регрессии
- $t = 1 \ldots T$ – номер месяца для объясняемой и объясняющей переменных
- $l = 1 \ldots t$ – номер лага объясняющей переменной

Модель с распределенными лагами для сезонных разностей: тематика «посольство»

1. По отдельности для каждой из тематик «учеба», «работа», «посольство» оценивались модели с включением лагов от 1 до 12
2. Для определения количества лагов, выбиралась модель, соответствующая наименьшему значению информационного критерия (AIC) среди всех возможных моделей с лагами

Переменная		PCA вектор «посольство» с лагами
Константа		-39.62 (55.15)
PCA – вектор «посольство»	T	-177.98*** (56.57)
	T-1	-212.96*** (57.94)
	T-2	-223.17*** (58.19)
	T-3	-102.92* (54.00)
	T-9	104.09* (55.26)
	T-10	165.59*** (50.90)
	T-11	108.38** (49.03)
Observations		80
AIC		1226
R2		0.58
F Statistic		14.15***
Средний лаг		5.6 [3.64; 7.92]

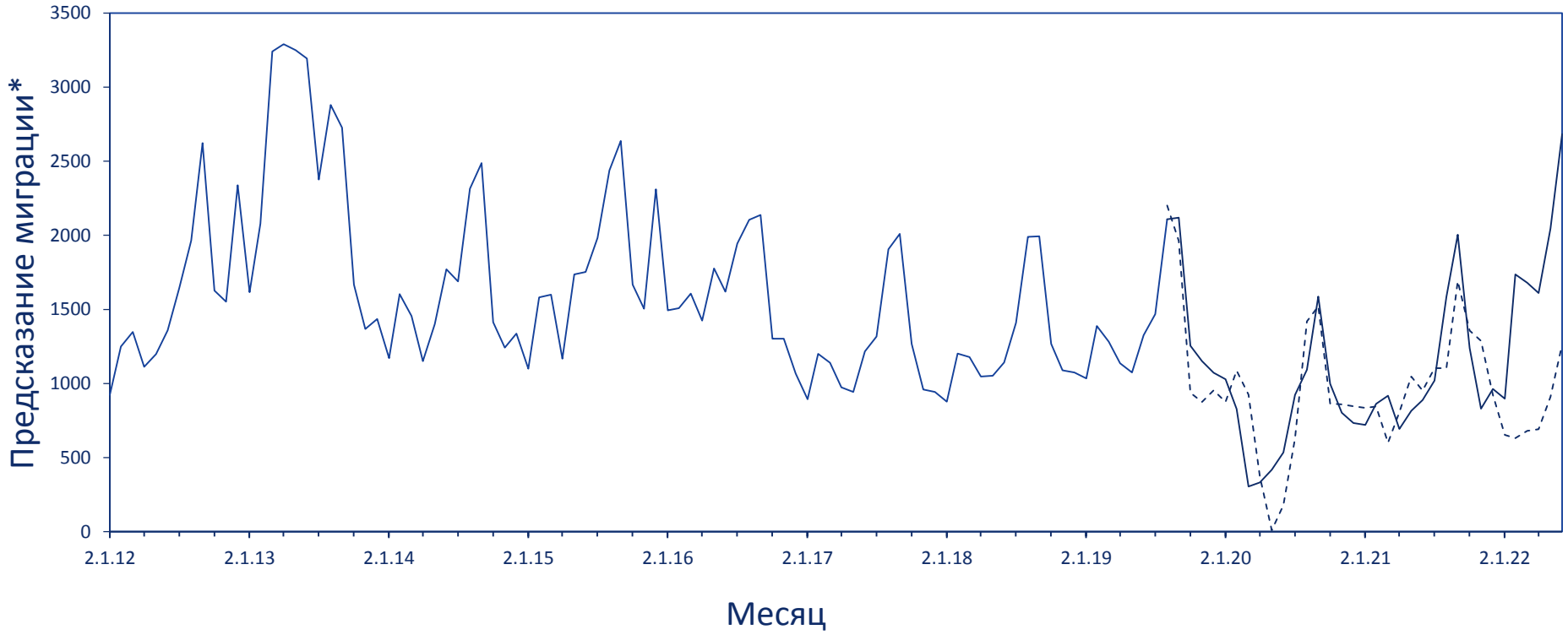
Значимость коэффициентов модели: *** p <0.01, ** p <0.05, * p <0.1. В скобках представлены стандартные ошибки. с

Модели с распределенными лагами для сезонных разностей

Модель	Количество регрессоров	MAE				MAPE				Средний лаг
Прогнозный период		1 год	2 года	3 года	1 месяц*	1 год	2 года	3 года	1 месяц*	
PCA по тематикам без лагов	4	374.1	348.9	416.1	412.6	0.75	0.55	0.5	0.49	
PCA по тематикам + dummy переменная	6	379.3	355.5	423.1	418.7	0.77	0.56	0.51	0.5	
PCA вектор «учеба» с лагами	3	471.4	471.7	492.7	484.1	0.87	0.7	0.59	0.58	8.0 [5.36; 10.8]
PCA вектор «работа» с лагами	8	696.3	689.6	667	654.8	1.38	1.05	0.85	0.83	6.5 [4.72; 8.21]
PCA вектор «посольство» с лагами	9	257.2	**196.9**	310.1	305.4	0.44	**0.30**	**0.32**	**0.31**	5.6 [3.64; 7.92]
PCA векторы «учеба», «работа», «посольство» с лагами	13	**177.8**	216.1	**305.0**	**295.2**	**0.35**	0.32	0.32	0.31	Учеба 11 Работа 7.9 [7.01; 8.98] Пос. 4.3 [2.45;7.18]
SARIMA		311.3	349.4	345.9	345.9	0.68	0.58	0.47	0.47	

*Прогноз кросс-валидацией на 1 месяц вперед в промежутке 01.08.2019–01.08.2022 гг. Тестовая выборка увеличивается на 1 месяц на каждом шаге.

Прогноз миграции из РФ в Германию. 2011-2022 гг., чел.



*Предсказание количества миграций с использованием модели PCA вектор «посольство» с лагами для 01.08.2019–01.08.2022 гг., чел., отмечено пунктирной линией.

Выводы

1. Предложена методика оценки миграционной статистики с минимальной задержкой во времени, т.к. называемый наукастинг статистики миграции.
2. GTI применимы для наукастинга миграции из России в Германию.
3. Мы продолжили исследование (Броницкий, Вакуленко, 2022), дополнив его изучением **глубины лага** поисковых запросов в зависимости от тематики поисковых запросов (посольство, работа, учеба), а также **снижением размерности модели**, предложив подходы к выбору ключевых поисковых запросов и построением агрегированных индексов на основе них.

Выводы

4. **Включение** в модель **лагов** до 12 месяцев **дает** существенное **улучшение предсказательной силы** (MAE 295.2 для модели с включенными лагами против MAE 412.6 для модели без лагов)
5. Различные типы поисковых запросов (посольство, работа, учеба), агрегированные с помощью метода главных компонент по смысловым группам, имеют **различную лаговую структуру в модели**.
6. Миграция реагирует на запросы типа учеба и работа – средние лаги 8 и 6.5 месяцев соответственно. А вот для поисковых запросов, связанных с посольством, средний лаг равен 5.6 месяцев.

Статьи

Броницкий Г. Т., Вакуленко Е. С. (2022). Прогнозирование миграции из России в Германию с использованием Google-трендов. *Демографическое обозрение*, 9 (3), 75-92.

Броницкий Г. Т., Вакуленко Е. С. (2024). Применение Google Trends для прогнозирования миграции из России: агрегация поисковых запросов и учет лаговой структуры. *Прикладная эконометрика (в печати).*

Спасибо за внимание!

evakulenko@hse.ru