Федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет экономических наук

КУРСОВАЯ РАБОТА

Прогнозирование макроэкономических переменных на основе интернет запросов пользователей

Название темы

по направлению подготовки Экономика
образовательная программа «Экономика и статистика»

Выполнил:
Студент группы БСТ 196
Мурадов Анар Рафикович
Ф.И.О.

Руководитель:
к.э.н., доц.
Мамедли Мариам Октаевна
степень, звание, должность Ф.И.О.

Москва 2021

Содержание

Введение

Прогнозирование макроэкономических переменных является неотъемлемой частью функционирования любого государства, поскольку достоверное представление о возможных изменениях во внутренней и внешней экономике страны дает большие преимущества для корректировки политик и ведет к реализации основных целей государства. В данной работе рассматривается возможность прогнозирования макропеременных при помощи запросов пользователей в интернете. С развитием технологий, облегчением доступа в интернет практически во всех уголках страны, становится доступным и отслеживание поведения населения с еще большей точностью. По данным ресурса LiveInternet наиболее популярными поисковыми системами на территории Российской Федерации являются Google, Яндекс, Search Mail.ru, Rambler, Bing. В список не вошли системы с долей запросов менее 0.1% от общего количества, такие как Яндекс (картинки), Ukr.net, Yahoo, tut.by, Baidu. При этом лидером является Google с долей в среднем 58.25% за 4 квартала 2020 г. Отставание отечественного сервиса Yandex является небольшим, и доля составляет приблизительно 40% от всех запросов. По этой причине для исследования был выбран сервис Google.

Количество пользователей сети интернет в России стремительно растет начиная с 2010 г., когда их совокупная доля составляла около 42% от всего населения и достигло рекордных 81.1% в 2020 г.[1] Это говорит о масштабной популяризации интернета. Учитывая среднее количество запросов равное приблизительно 125 млн. в месяц за 2020 г. Можно сделать вывод, что чем больше людей пользуются интернетом, тем больше они «гуглят». На основе данной гипотезы была выдвинута идея проверить влияние настроений населения, основанных на их Google запросах, на то, что в действительности происходит с макропеременными. Положительное влияние может говорить о

[1] Статистика приведена на основе исследования Digital 2020.

том, что Google запросы пользователей могут служить настоящими сигналами о возможных движениях в статистике и могут давать более быструю картинку происходящего, нежели официальные статистики, выпускаемые с задержкой в месяц, квартал либо год. Иными словами, отслеживая текущую динамику запросов, можно сделать вывод о статистике, которая будет предоставлена в рамках отчетного периода, и о возможных методах регулирования, которые могут возникнуть после появления отчетов. Так, полагаясь на Google, можно значительно ускорить реагирование на возникающие шоки на рынке и оперативно предпринимать меры по их устранению. Для подробного анализа мы выбрали влияние запросов на уровень безработицы в Российской Федерации. Выдвинутую нами гипотезу уже использовали западные коллеги и был написан ряд статей учеными, сотрудниками институтов о том, как сильно коррелируют запросы пользователей Google на фактический уровень безработицы и какие прогнозы можно из этого составить. Для прояснения картины мы ознакомились с основными работами в этой области.

ГЛАВА 1. Методология исследования

1.1 Данные

Для проведения анализа используются данные по релевантным запросам пользователей за период с 01.01.2010 по 31.12.2020. Для этого на платформе Google Trends был задан нужны десятилетний период и выбрана тема запросов Безработица. Особенностью платформы является возможность воспользоваться выбором целиком темы запросов, которая включает все наиболее частые запросы так или иначе связанные с безработицей. Основными и наиболее частыми запросами по теме являются «Пособие по безработице», «Встать на учет по безработице», «Центр занятости населения» и т. п. Данный метод отбора запросов является предпочтительнее, чем собирать отдельные конкретные запросы, поскольку имеет гораздо широкую базу запросов. На платформе динамика представлена в виде относительного значения количества запросов по теме разделенных на общее количество запросов в регионе за период, именуемый далее Google Index[2].(Формула индекса предложена в разделе 1.1 Главы 1). Далее получены данные по уровню безработицы в РФ за отчетный период, которая ввиду особенностей регламента источника[3] до 2015 года представлялась поквартально, далее ежемесячно. Данные по безработице не масштабированы относительно данных запросов пользователей, поскольку в исследовании важна динамика, а не конкретные значения за конкретный период. Графическое представление полученных данных выведено на Рис.1 и Рис.2. Визуализация данных позволяет определить существующие закономерности на гипотетическом уровне и выдвинуть предварительные гипотезы для дальнейшего анализа, такие как наличие корреляции, сезонности, тренда, аномальных явлений и т. п.

[2] Впервые название Google Index было употреблено в Chloi & Varian (2008)
[3] Росстат, Федеральная служба государственной статистики.

Рис. 1. Динамика уровня безработицы в России за 2010–2020 гг.

Источник: Росстат, Федеральная служба государственной статистики.

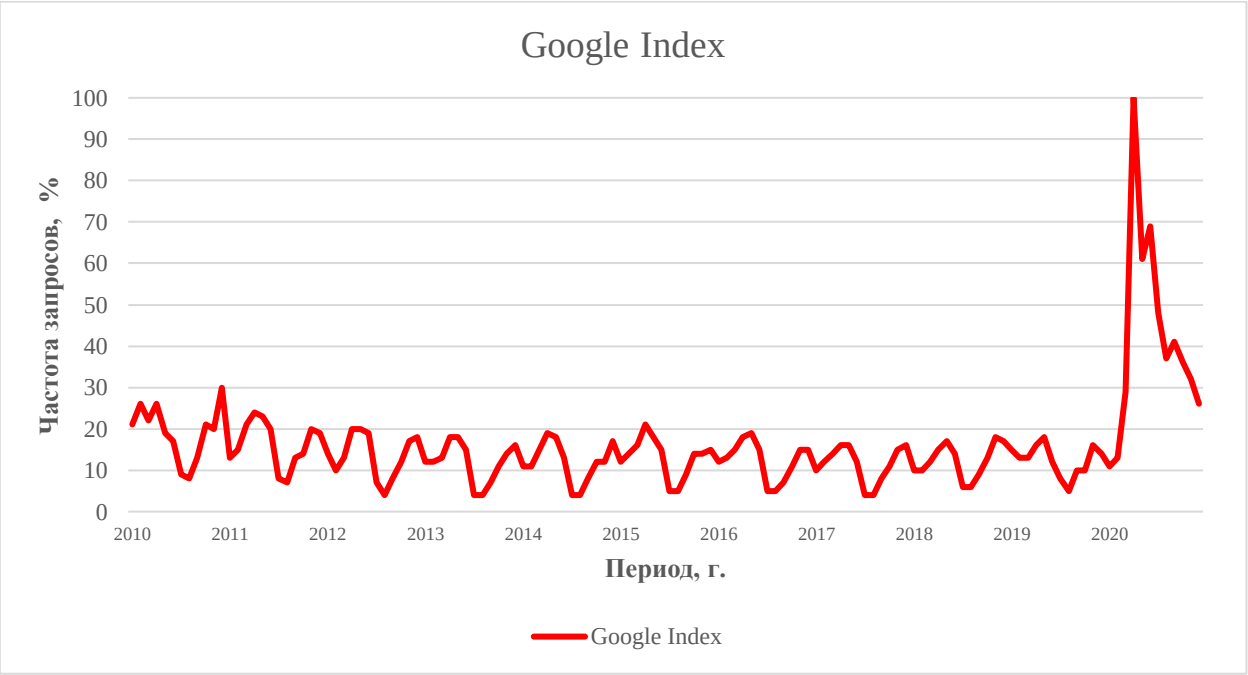

Рис.2. Динамика релевантных запросов пользователей теме «Безработица» в России за 2010–2020 гг.

Источник: Google Trends

Отсюда заметим схожий тренд и очевидную сезонность в данных. На обоих графиках отчетливо виден шок, возникший в последствии введения ограничительных мер связанных с пандемией COVID-19, впервые

зарегистрированный на территории РФ 1 марта 2020 года, после чего начался интенсивный рост заболеваемости. Для предотвращения распространения пандемии государством было принято решение перевести максимальную часть занятого населения на удаленный режим работы. На фоне этого части бизнеса пришлось сокращать штаб сотрудников, а некоторые компании и вовсе были вынуждены закрыться. Это повлияло на уровень безработицы, который достиг своего максимума за последние восемь лет и пробил отметку в 4.5 млн по информации на май 2020 года. Население перешло в стадию активного поиска работы, переквалификацией или рассматрением вариантов трудоустройства в других в организациях, предлагающих удаленный формат работы. Другой фактор подтверждающий схожесть двух графиков – схожая сезонность. Это подтверждает существование связи между предложенными переменными, нетрудоустроенные интернет-пользователи действительно обращаются к поисковому сервису Google для ознакомления с доступными вакансиями в сети.

Ознакомившись с данными и проведя визуальный анализ, перейдем к инструментальному анализу и формированию модели.

1.2 Анализ данных

Полученные данные загружены в виде дата фрейма в рабочую среду языка программирования python. Для работы использованы следующие пакеты: pandas, numpy, matplotlib, sklearn, statsmodels. Первично для исследования проверена гипотеза о существовании корреляции между непосредственно уровнем безработицы за период и Google Index. Данная гипотеза подтвердилась наличием умеренной связи между переменными ($r_{corr} = 0{,}31$), что дает основание на проведение дальнейшего анализа. Поскольку выбранные для работы данные представляют собой временные ряды, то использование линейных и нелинейных моделей машинного обучения не гарантирует корректного прогноза, так как не предоставляется

возможным удаление выбросов методами IQR[4], это может сильно ухудшить достоверность прогноза, во временном ряду следует учитывать каждое наблюдение как произошедший факт в прошлом и оценивать его влияние на будущее. В работе J. Tuhkuri, IFSDA, были использованы авторегрессионные модели (AR(p)), однако для нашего исследования данная модель не подходит, поскольку мы имеем дело одновременно с двумя временными рядами и предполагается их одновременное влияние друг на друга. В связи с этим мы используем модель векторной авторегрессии (VAR(p)), которая позволяет одновременно прогнозировать все временные ряды с влиянием каждой переменной. Первичный анализ имеющихся данных проверкой причинностью Грейнджера показал удовлетворительные результаты, что свидетельствует о существовании взаимосвязи между переменными. Значения менее требуемого уровня значимости 0.05 говорят о том, что переменная X или Y объясняет переменную Y или X.

Таблица 1

Результаты теста причинности Грейнджера.

	Y_x	**Google_Index_x**
Y_y	1.0000	0.0000
Google_Index_y	0.0263	1.0000

Источник: Вычисления автора.

После получения положительных результатов теста Грейнджера можно проверить статистическую значимость переменных исследования. Для этого существует тест коинтеграции Йохансена, который включен в пакет statsmodels инструментом coint_johansen. В первую очередь требуется знать порядок интеграции временного ряда, а это ни что иное как порядок дифференциации или иными словами количество произведенных конечных разностей над рядом, для приведения его стационарному виду. В нашем случае визуально Google Index имеет практически стационарный вид и незначительный линейный тренд, в то время как в случае с целевой

[4] IQR (Interquartile range) – Межквартильный диапазон.

переменной отчетливо виден линейный тренд. В связи с этими условиями, в параметрах теста за порядок интеграции возьмем число дифференциаций равное двум, позже при приведении рядов к стационарному виду убедимся в правильности решения. Однако при существовании двух и более временных рядов, то возникает понятие «коинтеграции», иными словами одновременной интеграции всех рядов. Поскольку ряды имеют линейный тренд, то отметим это еще одним параметром теста. По результатам теста получаем, что обе переменные имеют статистическую значимость, это позволяет сделать вывод о том что между ними существует долгосрочная статистическая связь, отсюда можно быть уверенным в использовании наших временных рядов для применения в VAR модели.

Следующий этап - деление выборки на тренировочную и тестовую части. В качестве тестовой будут использоваться данные за последние 38 месяцев, в таком случае разбиение приблизительно равно 70% в тренировочной и 30% имеющихся данных в тестовой. Теперь нужно проверить стационарность рядов, это позволит исключить изменение характеристик временного ряда с течением времени и сделать прогноз более точным. Для этого следует провести расширенный тест Дики-Фуллера, который удаляет возможную автокорреляцию во временном ряду и проводит аналогичную процедуре в обычном тесте Дики-Фуллера. Включенный в пакет statsmodels тест adfuller позволяет проверить стационарность временного ряда при помощи проверки размерности корня в авторегрессии модели. Если корень единичный, то модель сведена к линейному случайному процессу и ряд стационарен, если в модели d корней, то ее нужно изменить d раз, чтобы привести ряд к стационарности. В наших временных рядах после первичного проведения теста, целевая переменная оказалась не стационарной, в то время как Google Index уже имел стационарный вид. В связи с этим мы воспользовались методом разностей для приведения к стационарности. Метод заключается в том, что если изменение целевой переменной имеет линейный

или схожий ему характер, то взятие первой разности остатков приводит к стационарности исследуемого ряда. Преобразование имеет следующий вид:

$$y'_t = \Delta y_t = y_t - y_{t-1}$$

При этом если после первого применения разностей ряд все еще не приведен к стационарности, то соответственно ряд все еще не сведен к линейному случайному процессу и данный метод следует применять d раз, как описывалось выше. В случае с исследуемыми нами данными, то временной ряд Google Index был приведен к стационарности после первой разности, и удовлетворяет расширенному тесту Дики-Фуллера, при этом целевая переменная приняла стационарный вид лишь после второй конечной разности остатков. Следует отметить, что при применении метода разностей, количество наблюдений во временном ряду уменьшается, поэтому при применении его по отношению к одному из рядов векторной авторегрессионной модели, то аналогичное нужно произвести и с другими рядами, для сохранения равного количества наблюдений. Мы убедились в том, что выдвинутая гипотеза при проведении теста Йохансена на коинтеграцию, параметром порядка интеграции нужно выбрать число два.

Теперь, когда мы провели проверку причинности по Грейнджеру, получили статистическую значимость и стационарность, можно приступать к непосредственному обучению векторной авторегрессионной модели. Для начала рассмотри систему уравнений для нашей модели:

$$Y_{1,t} = \alpha_1 + \beta_{11,1} Y_{1,t-1} + \beta_{12,1} Y_{2,t-1} + \cdots + \beta_{1n,1} Y_{n,t-1} + \epsilon_{1,t}$$

$$Y_{2,t} = \alpha_2 + \beta_{21,1} Y_{1,t-1} + \beta_{22,1} Y_{2,t-1} + \cdots + \beta_{2n,1} Y_{n,t-1} + \epsilon_{2,t}$$

Имеем систему из двух уравнений VAR(n) порядка, для которой нам нужно определить оптимальный размер порядка лагов. Для этого нужно постепенно подбирать лаги до тех пор, пока не будет максимальный доступный лаг, далее нужно воспользоваться информационным критерием и выбрать лаг с наименьшим уровнем выбранного критерия. Максимальное количество лагов зависит от количества наблюдений и вычисляется по формуле, встроенной в инструмент .select_order пакет statsmodels:

$$12 \times (\text{кол} - \text{во наблюдений } \div 100.) \times\times 1./4$$

Отсюда получаем максимальное количество лагов равное 12 и одновременно благодаря использованному встроенному в statsmodels инструменту .select_order получаем значения сразу четырех основных информационных критериев: Акайке (AIC), Байесовскй (BIC), финальная ошибка предсказания (FPE), Ханнана-Квина (HQIC). Критерий AIC тесно связан с BIC, но в отличие от него включает линейно зависимую от числа параметров функцию штрафов, которая показывает зависимость роста критерия от количества параметров, а не от необъясненной дисперсии ошибки. Также в сравнении есть схожая, но немного с другой формулой нахождения (параметры уравнения идентичны) критерий Ханнана-Квина, а также второй критерий Акайке – финальной ошибки предсказания, который подходит при использовании нескольких методов предсказания, в нашем случае VAR модель будет работать методом наименьших квадратов. В связи с этим основной упор следует сделать на AIC.

Воспользовавшись функцией .select_order мы получили следующие результаты:

Таблица 2

Выбор порядка VAR модели, минимальные значения выделены.

	AIC	**BIC**	**FPE**	**HQIC**
0.	1.121	1.181	3.068	1.145
1.	0.9235	1.102	2.518	0.9951
2.	0.8690	1.167	2.385	0.9884
3.	0.5254	0.9422	1.693	0.6925
4.	0.2744	0.8103	1.318	0.4893
5.	-0.4802	0.1749	0.6209	-0.2175
6.	-0.6341	0.1400	0.5335	-0.3238
7.	-0.8063	0.08700	0.4506	-0.4481
8.	-0.9461	0.06624	0.3934	-0.5402
9.	-0.9796	0.1518	0.3825	-0.5260
10.	-1.007	0.2431	0.3744	-0.5061
11.	-1.625	-0.2554	0.2036	-1.076
12.	**-1.897***	**-0.4086***	**0.1566***	**-1.300***

Заметим, все существующие критерии показали минимальные значения на 12 по счету лаге, он же является максимальным для нашей модели. Для дальнейшего прогноза будем использовать VAR(12) модель 12 порядка соответственно.

1.3 Прогнозирование

Наконец переходим непосредственно к прогнозу. Используем функцию model.fit(12) и далее получим результаты регрессии при помощи .summary(). Как и говорилось выше, для прогноза используется метод наименьших квадратов. После получения результатов проверим остатки на множественную автокорреляцию. Наличие автокорреляции в ряду после прогноза говорит о зависимости его переменных друг от друга, что является сигналом того, что данные в выборке распределены неслучайно. Воспользуемся наиболее часто используемым тестом Дарбина-Уотсона, который проверяет нулевую гипотезу о наличии автокорреляции. Значения в результатах теста колеблются от 0 до 4, при этом чем ближе они к 2, тем с большей вероятностью отвергается гипотеза о наличии автокорреляции. Отклонения вверх и вниз от 2, говорят о положительной или отрицательной автокорреляции соответственно. Загрузим встроенный в пакет statsmodels инструмент durbin_watson и проведем тест. Получим следующие результаты:

Таблица 3

Результат теста Дарбина-Уотсона

Y	1.88
Google Index	2.02

Источник: Вычисления автора

Результаты удовлетворяют гипотезе об отсутствии автокорреляции остатков. Можно переходить к непосредственному прогнозу. Для этого используем функцию .forecast, далее создадим дата фрейм из полученных данных. Поскольку полученные данные произошли от дважды дифференцированного исходного ряда, то следует провести де-дифференциацию, иными словами инвертировать данные обратно в исходных

масштаб. Для этого создадим собственную функцию python, предложенный в статье Selva Prabhakaran и код которой будет в приложении.

1.4 Оценка прогноза

Последним шагом перед визуализацией полученного прогноза является выведение основных метрик оценки прогнозирования. В нашем случае имеем следующие результаты:

Таблица 4

Основные метрики оценки результатов прогнозирования

	Y	**Google Index**
MAPE	0.0926	0.3247
MAE	0.5528	11.4799
MPE	-0.0799	-0.3116
RMSE	1.017	22.3306
Correlation coefficient	-0.519	0.1333
Min/Max	0.0924	0.3242

Источник: Вычисления автора

Для оценки используются MAPE (Mean Absolute Percent Error) – позволяет в процентах выразить MAE (Mean Absolute Error). MAE – проверяет разницу между прогнозированными данными и фактическими, деленное на общее количество наблюдений. При помощи его процентной интерпретации возможно установить точность прогноза. MPE (Mean Percent Error) – процентное значение средней ошибки. RMSE (Root Mean Sqaured Error) – среднеквадратическая ошибка, наиболее показательная метрика, которая часто применяется при прогнозировании. По формуле ее вычисления можно понять, что она показывать корень от суммы разности квадрата между прогнозируемым и фактическим значением разделенным на количество наблюдений. При значениях ниже 180, можно говорить об адекватности модели и ее применимости для прогноза. В нашем случае данный показатель для целевой переменной показывает значение около единицы, что является положительным показателем для модели. Далее следует полученный коэффициент корреляции, говорящий наличие линейной связи между

данными, а также представлен показатель Min/Max сравнивающий отклонение экстремальных значений прогноза от фактических

В целом все метрики показывают достаточно приемлемые результаты для целевой переменной. Значение MAPE равное 0.09 говорит о 91% уровне надежности прогноза, высокие MAE говорит о наличии умеренной связи между прогнозом и фактическими значениями. Отклонение экстремумов также остается на низком уровне. Однако опираясь на коэффициент корреляции, наблюдается отрицательная связь между данными, что может сулить наличие расхождений в части прогноза.

Для наглядного обзора полученных результатов и выявления возможных неточностей прогноза получим визуализацию прогноза.

Рис. 3 Динамика фактического и прогнозируемого уровня безработицы в России за период, включенный в тестовую выборку

Источник: Росстат, Федеральная служба государственной статистики, вычисления автора.

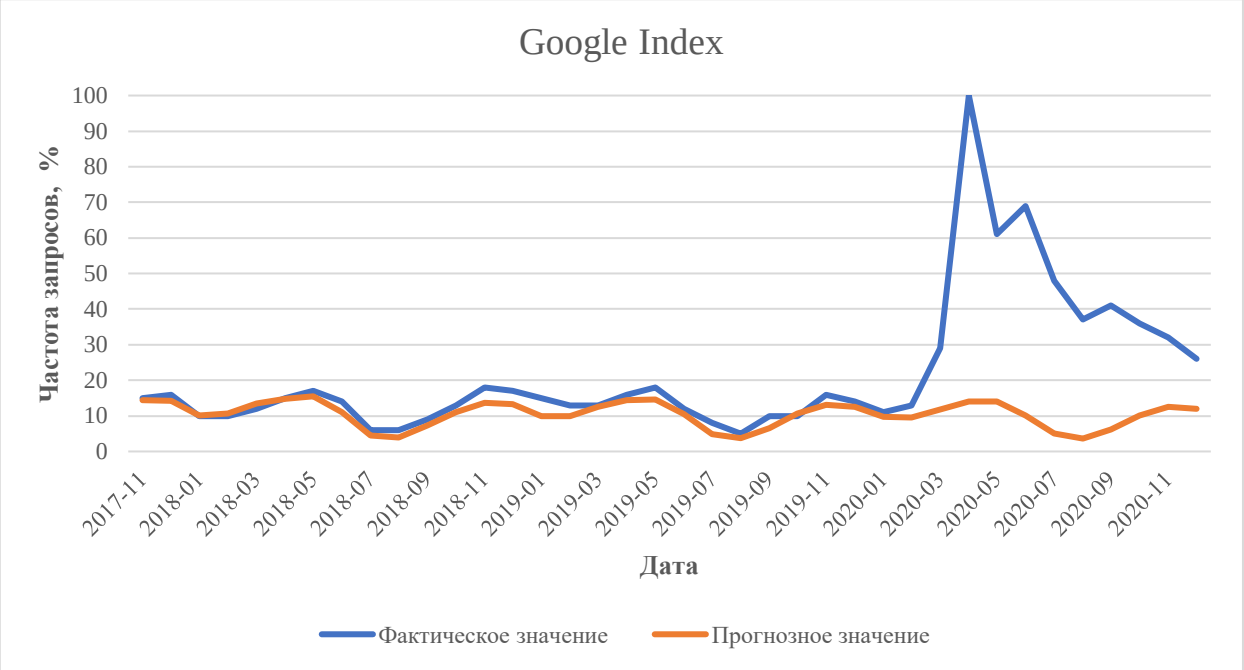

Рис. 4 Динамика фактических и прогнозируемых релевантных запросов пользователей теме «Безработица» за период, включенный в тестовую выборку.

Источник: Google Trends, вычисления автора.

После визуализации отчетливо видно хорошее качество прогноза до шока, возникшего в начале 2020 года. Прогноз получился более сглаженным по сравнению с фактическими данными. В целом прогноз получился достаточно точным, за исключением части с шоком. Не способность модели отреагировать на события 2020 может быть вызвано достаточно большим размером лага, выбранным как оптимальное значение, однако основываясь на базовые правила формирования моделей, то ее способность не реагировать на шоки, которые в линейных регрессионных моделях интерпретируются как выбросы, является скорее положительной чертой, чем отрицательной. Это позволяет рассчитывать на более корректные долгосрочные прогнозы, которые заведомо сглажены на выбросы. Также следует отметить точное восстановление сезонности данных и тренда до возникновения отклонений 2020 года. При исключении пандемии, полагаясь на прогноз можно сделать вывод о том, что в целом уровень безработицы устойчиво шел на спад.

Для улучшения прогноза, следует добавить в исследование дополнительную категориальную переменную, как это рекомендуют исследователи из статьи 1.3 главы 1. Мы внедрили в дата фрейм новую бинарную dummy переменную, обозначив шоковые дни категорией 1, остальные дни оставив равными нулю. Поскольку это не временной ряд, а лишь ряд имеющий вес в прогнозировании, то включать ее в предварительный анализ не требуется. После деления выборки на тестовую и тренировочную было обнаружено, что тренировочная выборка включает лишь нулевые значения dummy переменной. Это значит, что определитель матрицы, используемой бля составления VAR модели будет меньше нуля, что не позволит сделать прогноз. Данное замечание подтвердилось при попытке создания модели и выдало соответствующую ошибку в коде python.

1.5 Дополнительное исследование

Однако, для расширения прогноза мы решили добавить в работу дополнительные переменные и проверить в какую сторону изменятся прогнозы, как это повлияет на качество модели и сделать вывод о том, целесообразно ли использовать в прогнозировании уровня безработицы по запросам пользователей в Google дополнительные связанные с целевой макропеременной регрессоры. Для этого выбраны следующие показатели: темпы роста ВВП в текущих ценах, значения реальной ставки на период полученные стандартным методом (r $= \pi -$ ключевая ставка, где $r -$ реальная ставка, $\pi -$ инфляция), также инфляция, выраженная в ценах за предыдущий аналогичный период наблюдения, что равна динамике индекса потребительских цен за исследуемый период.

На графиках рассмотрим визуальное представление добавленных переменных и сделаем вывод о характере динамики дополнительных данных на Рис. 5-7:

Рис. 5 Темпы роста ВВП в России за период 2010–2020 гг.

Источник: Банк России

Рис. 6 Динамика реальной ставки процента в России за период 2010–2020 гг.

Источник: Росстат, Федеральная служба государственной статистики, вычисления автора.

Рис. 7 Динамика уровня инфляции в России за период 2010–2020 гг.

Источник: Росстат, Федеральная служба статистики.

Отметим наличие схожей сезонности с целевой переменной в динамике темпов роста ВВП, а также шока в марте 2020 г., обвалившего темпы роста до минимального значения за весь исследуемый период. Это было вызвано упомянутой раннее пандемией COVID-19. Схожее поведение демонстрирует график динамики реальный ставки, одна обратив внимание на изменения инфляции, возможно заметить значительное увеличение показателя начиная с конца 2014. Данный шок объясняется аннексией Крыма в 2014 г. и последовавшим введением санкций иностранных стран, что сказалось на курсе рубля и ценах импортных товаров соответственно. Однако данные события не имеют влияния на безработицу за период с 2014 по 2016 гг., поскольку колебания валют и инфляции не сказались на рынке труда, также не было введено масштабных ограничений на деятельность работодателей, которое могло бы спровоцировать рост безработицы.

Первично мы снова проверили наличие корреляции между регрессорами и целевой переменной, а также степень их связи между собой. В результате получены значения, представленные в Таблице 5:

Таблица 5

Корреляция между переменными исследуемых данных

	Y	Google Index	Динамика ВВП	Реальная ставка	Инфляция
Y	1	0,31	0,14	-0,33	0,24
Google Index	0,31	1	-0,51	-0,09	-0,15
Динамика ВВП	0,14	-0,51	1	0,26	-0,26
Реальная ставка	-0,33	-0,09	0,26	1	-0,78
Инфляция	0,24	-0,15	-0,26	-0,78	1

Источник: Вычисления автора

Отчетливо заметно существование корреляции как между целевой переменной и регрессорами, так и между отдельно регрессорами, что позволяет на данном этапе продолжить анализ новых данных.

Далее воспользуемся тестом причинности Грейнджера:

Таблица 6

Результаты теста причинности Грейнджера

	Y_x	Google_Index_x	Динамика_ВВП_x	Реальная_ставка_x	Инфляция_x
Y_y	1.0000	0.0000	0.0067	0.0627	0.2465
Google_Index_y	0.0263	1.0000	0.0001	0.7027	0.2826
Динамика_ВВП_y	0.0529	0.0023	1.0000	0.3064	0.1612
Реальная_ставка_y	0.0018	0.1318	0.0004	1.0000	0.0000
Инфляция_y	0.2116	0.6105	0.2190	0.0101	1.0000

Источник: Вычисления автора

По условию нулевая гипотеза о наличии причинности отвергается при p-value выше уровня надежности $\alpha = 0.05$, отсюда следует, что между инфляцией и безработицей, также, как и между безработицей и инфляцией отсутствует причинность по Грейнджеру. Аналогичное наблюдается между реальной ставкой и целевой переменной, поэтому не соответствующие тесту переменные были удалены из исследования.

Далее проверим коинтеграцию переменных тестом Йохансена, выбрав параметр d равный двум, полагаясь на предыдущее исследование, где минимальное количество дифференциаций целевой переменной для достижения его стационарности равно двум. Все параметры имеют статистическую значимость, соответственно мы можем переходить к делению выборки на тренировочную и тестовую

Деление произведено в аналогичных долях, как и в предыдущем исследовании, для сохранения равных условий тренировки выборки.

Проверим ряды на стационарность. После начального теста Дики-Фуллера, стационарным является только временной ряд «Google Index». Первая разность привела к стационарному виду ряд «Динамика ВВП». Вторая разность позволила добиться стационарности целевой переменной.

После получения стационарности перейдем к подбору оптимального количества лагов модели при помощи функции .select_order и получим следующие результаты основных информационных критериев:

Таблица 7

Выбор порядка второй VAR модели, минимальные значения выделены.

	AIC	**BIC**	**FPE**	**HQIC**
0.	1.382	1.471	3.984	1.418
1.	1.173	1.530	3.232	1.316
2.	0.9797	1.605	2.667	1.230
3.	0.6060	1.499	1.840	0.9642
4.	0.4240	1.585	1.541	0.8895
5.	-0.3926	**1.037***	0.6865	0.1804
6.	-0.4992	1.198	0.6241	0.1813
7.	-0.7608	1.204	0.4881	0.02708
8.	-1.081	1.152	0.3621	-0.1854
9.	-1.236	1.265	0.3187	-0.2334
10.	-1.162	1.607	0.3557	-0.05172
11.	-1.596	1.441	0.2410	-0.3781
12.	**-1.813***	1.492	**0.2049***	**-0.4879***

Источник: Вычисления автора

Оптимальный порядок равен двенадцати, этому свидетельствуют наименьшие значения критериев AIC, FPE, HQIC. Поскольку изначально AIC была выбрана наиболее показательной для нашего исследования, то различие в значениях порядка BIC не меняет выбора оптимального лага равного 12.

Поскольку параметры выбраны, построим регрессионную модель и проверим остатки на автокорреляцию. Тест Дарбина-Уотсона показывает следующие результаты:

Таблица 8

Результаты теста Дарбина-Уотсона для второй модели

Y	2.03
Google Index	2.2
Динамика ВВП	1.91

Источник: Вычисления автора.

Результаты показывают незначительную автокорреляцию в ряду переменной динамики ВВП и показателя Google Index. Стоит учесть данное обстоятельство и сделать вывод о том, что качество модели может быть ухудшено.

Далее сделаем прогноз, приведем данные к изначальному виду и сравним полученные метрики точности прогноза:

Таблица 9

Основные метрики оценки результатов прогнозирования

	Y	**Google Index**	**Динамика ВВП**
MAPE	0.1255	0.2587	1.4263
MAE	0.6217	9.3585	3.1192
MPE	0.0783	-0.0769	-1.1791
RMSE	0.6744	19.5236	3.4407
Corr. coef.	0.4436	0.4496	0.6486
Min/Max	0.1117	0.2385	1.0526

Источник: Вычисления автора

Обратим внимание в первую очередь на метрики целевой переменной. RMSE уменьшился практически в два раза, что дает основу полагать об улучшении качества модели, но в то же время незначительно увеличился MAPE это говорит о снижении надежности прогноза. Новый параметр показывает непригодные для прогнозирования результаты. MAPE больше 100 процентов говорит о полном отсутствии точности прогноза. Возникновение таких результатов возможно по причине наличия автокорреляций в остатках исследуемых временных рядов, обнаруженные после теста Дарбина-Уотсона. Для наглядного изучения результатов просмотрим сравнительные динамики фактических и прогнозируемых показателей на графиках:

Рис. 8 Динамика фактического и прогнозируемого по второй модели уровня безработицы в России за период, включенный в тестовую выборку

Источник: Росстат, Федеральная служба государственной статистики, вычисления автора.

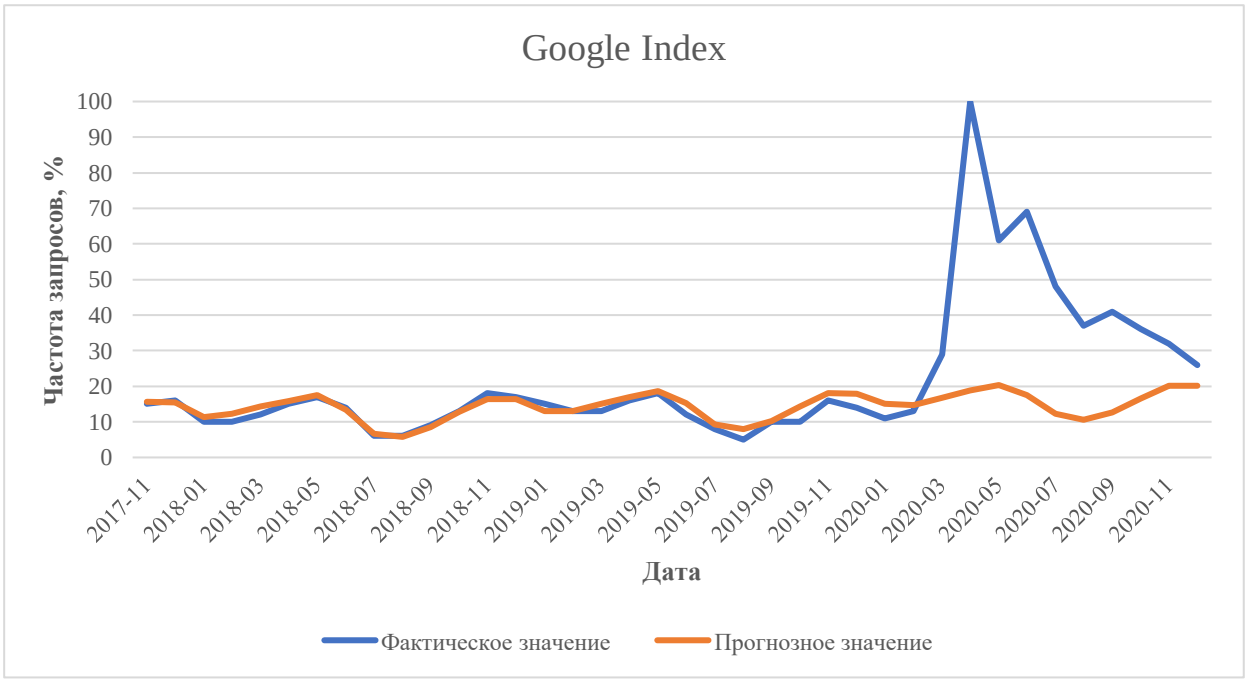

Рис. 9 Динамика фактических и прогнозируемых по второй модели релевантных запросов пользователей теме «Безработица» за период, включенный в тестовую выборку.

Источник: Google Trends, вычисления автора.

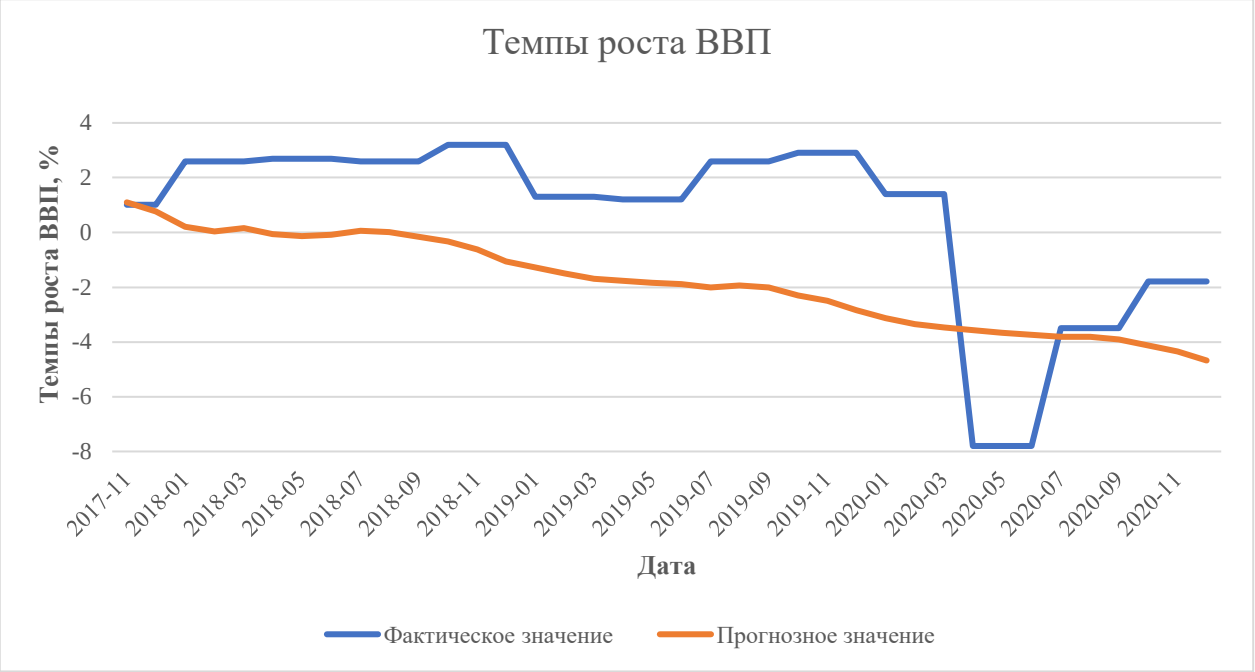

Рис.10 Динамика фактических и прогнозируемых по второй модели темпа роста ВВП за период, включенный в тестовую выборку.

Источник: Банк России, вычисления автора.

Несмотря на низкие метрики нового прогноза, значения для целевой переменной оказались завышены. Прогноз шока также не соответствует фактическим данным, что говорит о том, что даже добавление новой переменной с положительным трендом в период шока не влияет на прогнозирование данной части данных. Можно отметить лишь положительный тренд за весь предсказываемый период, в отличии от первичной модели. При этом точность прогноза запросов пользователей сохраняет высокий уровень точности, за исключением шока 2020 г. Подтвердились сигналы метрик прогноза Темпов роста ВВП. Наблюдается значительное отклонение от фактических данных, что позволяет сделать вывод в целом по всей модели. Модель с используемым нами дополнительным параметром оказалась непригодна для вывода корректных прогнозов. Таким образом первое исследование с использованием лишь индекса Google является наиболее предпочтительным для использования.

Заключение

Использованный нами подход к прогнозированию, с использованием векторной авторегрессии (VAR(12)) показал, что запросы пользователей действительно влияют на динамику уровня безработицы. Использование интернет запросов пользователей как сигналов для предсказания текущего уровня безработицы актуально, если нет сильных изменений, т. е. отсутствуют шоки. Однако выбранный способ моментального формирования прогноза дал некорректные результаты. Опираясь на исследование западных коллег и проделанную работу стоит отметить возможность осуществления «итерируемого» прогноза, посредством постепенного добавления в модель новых данных. В таком случае следует также проверить работоспособность dummy переменных, которые не были корректно реализованы в нашем исследовании. При получении положительных результатов и высокой надежности прогноза, создается возможность создания платформы, которая в реальном времени будет собирать информацию и делать прогноз на короткий период вперед. Такой опыт применен в Европе и может быть стимулом для внедрения в России. Таким образом, прогнозирование макропеременных на основе интернет запросов пользователей реализуемо в России и может быть основой для использования как методологии прогноза безработицы в стране.

Список литературы

1. Big Data: Do Google Searches Predict Unemployment? J. Tuhkuri, University of Helsinki, May 2015
2. ETLAnow: A Model for Forecasting with Big Data – Forecasting Unemployment with Google Searches in Europe, J. Tuhkuri, June 2016
3. Nowcasting the Unemployment Rate in the EU with Seasonal BVAR and Google Search Data, J. Anttonen, The Research Institute of the Finnish Economy, November 2018
4. Nowcasting the Unemployment Rate in Canada Using Google Trends Data, The Institute of Fiscal Studies and Democracy (IFSD), Fall 2017
5. Predicting the Present with Google Trends, Varian & Choi, August 2010
6. Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python, S. Prabhakaran, July 2019
7. https://www.statsmodels.org/
8. https://www.machinelearningplus.com/
9. https://rosstat.gov.ru/statistic/
10. https://cbr.ru/