«Прогнозирование количества заражений COVID-19 в Швейцарии»

Выпускная аттестационная работа

Работу подготовила: Гурова М.В.

В конце 2019 года было сообщено о заражениях в провинции Хубэй (Китай) коронавирусной инфекцией нового типа. За несколько месяцев этот новый вирус вызвал глобальную пандемию коронавирусной болезни (COVID-19). Несмотря на существование известных эпидемиологических моделей, ни одна из них не может с достаточной долей точности спрогнозировать ситуацию развития эпидемии COVID-19.

Целью данной работы является исследование данных о распространении COVID-19 в Швейцарии и построение краткосрочного прогноза развития пандемии в условиях непредсказуемости поведения исследуемого процесса.

Задачи, которые перед нами поставлены в данной работе:

- проведение анализа данных о распространении COVID-19 в Швейцарии;
- выявление взаимосвязей между переменными, доступными для анализа;
- подбор модели, наиболее точно предсказывающей развитие пандемии в Швейцарии.

4 этапа исследования

1. **Импорт библиотек, ознакомление с данными**
2. **Предобработка данных**
3. **EDA или разведочный анализ данных**
4. **Построение моделей, анализ результатов**

Импорт библиотек, ознакомление с данными

Задача на данном этапе заключалась в определении библиотек, функций и метрик, необходимых для работы, а также в импорте исходного датасета и первоначальном знакомстве с данными.

В работе использовались общедоступные ежедневные данные официальной статистики о развитии пандемии COVID-19 в мире.

В результате мы увидели, что наш исходный датасет включает в себя 67 столбцов, каждый из которых несет определенную информацию, в том числе:

- date - дата наблюдения;
- location - географическое положение (страна);
- total_cases - всего подтвержденных случаев COVID-19;
- new_cases - новые подтвержденные случаи COVID-19;
- total_deaths - общее количество смертей, связанных с COVID-19;
- new_deaths - новые случаи смерти, связанные с COVID-19;
- total_tests - общее количество тестов на COVID-19;
- total_vaccinations - общее количество введенных доз вакцины против COVID-19.

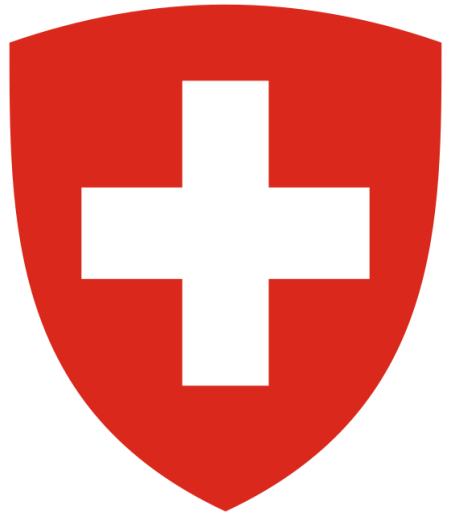
```
## Загружаем pmdarima

!pip install pmdarima
```

...

```
## загружаем fbprophet

!pip install fbprophet
```

```
## Импортируем необходимые библиотеки и функции

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# импорт моделей
from statsmodels.tsa.statespace.sarimax import SARIMAX
from statsmodels.graphics.tsaplots import plot_acf,plot_pacf
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller
from statsmodels.tsa.holtwinters import ExponentialSmoothing
from pmdarima import auto_arima
from fbprophet import Prophet

# метрики
from sklearn.metrics import mean_squared_error
from statsmodels.tools.eval_measures import rmse
from sklearn.metrics import mean_absolute_error

# для игнорирования предупреждений
import warnings
warnings.filterwarnings("ignore")
```

Предобработка данных



Задача на данном этапе заключалась в фильтрации данных, а именно в выборе данных по статистике заболевания в Швейцарии и необходимых признаков в новый датасет.

В качестве объекта для исследования был выбран набор данных о развитии пандемии в Швейцарии за период с 25.02.2020 по 01.12.2021, включающий в себя как общую информацию о заболеваемости, тестировании, смертях и вакцинации, так и информацию о новых выявленных случаях заболевания и смертях, вызванных COVID-19.

В результате мы установили, что наши данные содержат пропуски, которые возникли по причине добавления наблюдения ряда признаков спустя продолжительное время после первого случая заболевания в стране (данные по смертям от COVID-19, тестированию и вакцинации).

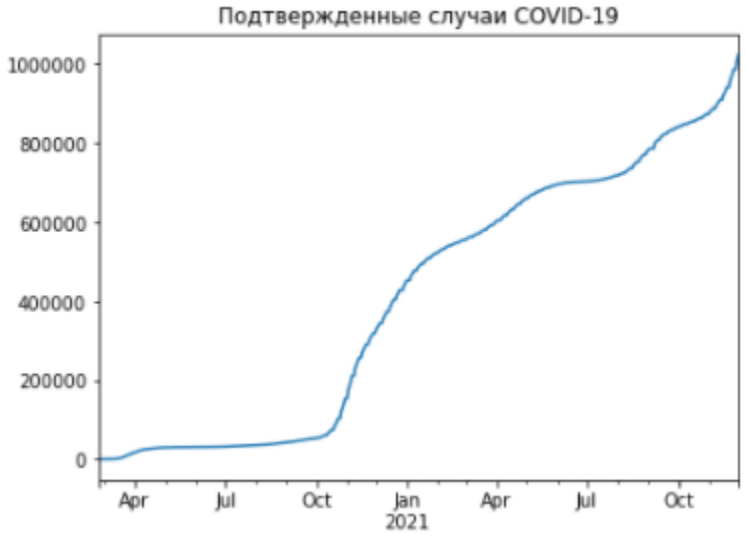
С целью создания непрерывного временного ряда была проведена работа по устранению пропусков в данных, в результате которой данные были подготовлены для дальнейшего исследования.

	date	total_cases	new_cases	total_deaths	new_deaths	total_tests	total_vaccinations
118833	2020-02-25	1.0	1.0	0.0	0.0	0.0	0.0
118834	2020-02-26	1.0	0.0	0.0	0.0	0.0	0.0
118835	2020-02-27	8.0	7.0	0.0	0.0	0.0	0.0
118836	2020-02-28	8.0	0.0	0.0	0.0	0.0	0.0
118837	2020-02-29	18.0	10.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...
119474	2021-11-27	986834.0	0.0	11470.0	0.0	12344102.0	11805235.0
119475	2021-11-28	986834.0	0.0	11471.0	1.0	12363351.0	11809935.0
119476	2021-11-29	1006239.0	19405.0	11500.0	29.0	12403612.0	11861643.0
119477	2021-11-30	1014662.0	8423.0	11522.0	22.0	0.0	11917341.0
119478	2021-12-01	1025129.0	10467.0	11548.0	26.0	0.0	11983616.0

EDA или разведочный анализ данных

Задача на данном этапе заключалась в подробном ознакомлении с исследуемыми данными по заболеваемости в Швейцарии, выявлении взаимосвязей между признаками.

По графикам отчетливо видно, что резкий рост подтвержденных случаев и смертей, связанных с COVID-19, произошел в конце октября 2020 года. Очевидно, что на рост этих показателей повлияло увеличение охвата тестирования населения на COVID-19, а также, возможно, это связано с увеличением туристического потока, поскольку Швейцария является традиционной страной туризма. Кроме того, по графикам можно увидеть, что в связи с началом вакцинации в Швейцарии с января 2021 года замедлился рост смертей, связанных с COVID-19, что может говорить о том, что вакцина облегчала течение болезни и уменьшала количество смертельных случаев от вирусной инфекции.



По **матрице корреляции** признаков видно, что между рядом признаков имеется тесная зависимость. В парах «общее количество тестов - общее количество смертей», «общее количество тестов - общее количество заболеваний», «общее количество заболеваний - общее количество смертей» наблюдается **очень высокая сила связи (свыше 0,9)** и можно сказать о том, что увеличение одного признака влечет за собой увеличение другого. Например, чем больше тестов проводится, тем больше выявляется в результате этого общих случаев заболевания COVID-19 и, соответственно, смертей от COVID-19.

	total_cases	new_cases	total_deaths	new_deaths	total_tests	total_vaccinations
total_cases	1.00	0.20	0.98	-0.10	0.96	0.84
new_cases	0.20	1.00	0.14	0.44	0.14	0.09
total_deaths	0.98	0.14	1.00	-0.10	0.91	0.74
new_deaths	-0.10	0.44	-0.10	1.00	-0.21	-0.32
total_tests	0.96	0.14	0.91	-0.21	1.00	0.92
total_vaccinations	0.84	0.09	0.74	-0.32	0.92	1.00

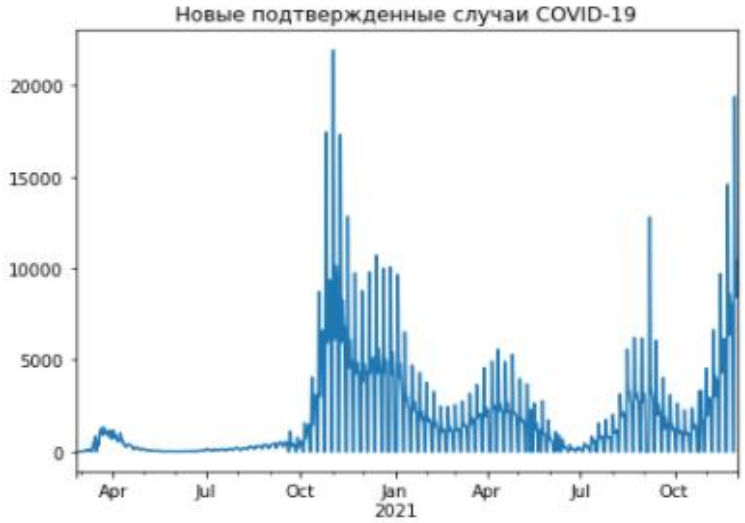
В паре «общее количество заболеваний - общее количество введенных доз вакцины» также наблюдается **высокая взаимосвязь**, что можно объяснить тем, что при введении вакцины снижается иммунитет человека и он становится восприимчивее к инфекциям, в том числе и к COVID-19. А можно интерпретировать и в обратную сторону: увеличение числа заболевших мотивирует людей на необходимость уберечь себя от заболевания путем вакцинирования.

Кроме того, интересно заметить, что между признаками «общее количество введенных доз вакцины» и «новые случаи смерти, связанные с COVID-19» имеется **слабая отрицательная корреляция** (-0.32), что может говорить о том, что вакцинация все же положительно влияет на иммунитет человека и уменьшает количество смертельных случаев.

Опять же стоит отметить, что наличие корреляции между признаками не всегда говорит об их причинной взаимообусловленности, то есть когда за счет эффектов одновременного влияния неучтенных факторов смысл истинной связи может искажаться.

По **статистике** видно, что количество новых случаев заболевания COVID-19 варьировалось от 0 до 21926 человек в день, тогда как максимальное количество смертей достигало 131 человека в день. Минимальное число смертей в день получилось -87 - отрицательное значение показателя, предположительно, связано с тем, что причина данных смертей ранее была связана с COVID-19, а в дальнейшем была отнесена к смерти по иным причинам, в связи с чем была произведена корректировка. В среднем число новых выявленных заболеваний и число смертей составило 1587 и 18 человек в день соответственно.

	new_cases	new_deaths
count	646.0	646.0
mean	1587.0	18.0
std	2680.0	29.0
min	0.0	-87.0
25%	17.0	1.0
50%	380.0	5.0
75%	2088.0	17.0
max	21926.0	131.0



На дневном графике новых подтвержденных случаев COVID-19 отчетливо видно, что максимальное количество выявленных заболевших в день приходилось на конец октября 2020 года, минимальное количество после значительного роста - в июле 2021 года. С конца ноября 2021 года вновь видим значительное увеличение количества новых случаев заболевания в день, что опять же, предположительно, можно объяснить возросшим туристическим потоком. Таким образом, на графике вырисовывается некая закономерность, выявить которую мы попробуем в дальнейшем.

На дневном графике новых случаев смерти, связанных с COVID-19, видно, что в феврале 2021 года были выбросы в данных - как в сторону увеличения, так и в сторону уменьшения. Предположительно это связано с тем, что ряд случаев смерти первоначально был засчитан как смерть, связанная с COVID-19 (что объясняет выброс в сторону увеличения), а затем отнесен к смерти по иным причинам (что объясняет выброс в -87, который мы установили ранее по статистике). Какой-либо закономерности в смертях в данном случае на графике мы не наблюдаем.

На данном этапе мы подробно ознакомились с исследуемыми данными по заболеваемости COVID-19 в Швейцарии, сделали предположения о возможной взаимосвязи между признаками, а также посмотрели статистику по количеству новых случаев заболевания COVID-19 и о новых случаях смерти, связанной с этим заболеванием.

Построение моделей, анализ результатов

Задача на данном этапе заключалась в построении модели, наиболее точно прогнозирующей возникновение новых случаев COVID-19 в Швейцарии.

Одним из важных этапов в работе с временными рядами является их проверка на **стационарность**.

Стационарный временной ряд - это ряд, для которого свойства (а именно среднее значение, дисперсия и ковариация) не зависят от времени. Прогнозирование по нестационарному временному ряду значительно усложняется и результат может не оправдать ожиданий, поэтому как правило перед построением прогноза ряд приводится к стационарному.

В данной работе для проверки временного ряда на стационарность использовался тест Дики-Фуллера, а именно выдвигалась гипотеза H_0 о наличии единичного корня, который как раз и свидетельствует о нестационарности ряда. Если получаемое в результате теста $t_{расч.} > t_{крит.}$ на всех уровнях значимости, то гипотеза H_0 подтверждается.

```
## Запускаем тест Дики-Фуллера.

adfuller(df.new_cases)
```

```
(-2.3422986780966286,
 0.1586883516201179,
 14,
 631,
 {'1%': -3.440755866431696,
  '5%': -2.86613130039063,
  '10%': -2.569215089800357},
 10522.238633996434)
```

В результате выполнения получили следующие значения:

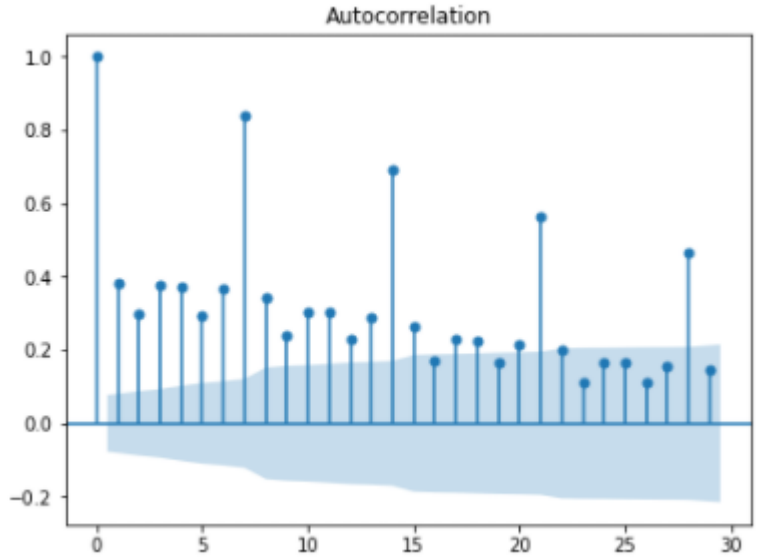
- -2.34 - значение $t_{расч.}$;
- 0.16 - значение pvalue;
- 14 - количество использованных лагов;
- 631 - количество наблюдений, используемых для регрессии ADF и расчета критических значений;
- -3.44, -2.87, -2.57 - критические значения для статистики теста на уровнях 1%, 5% и 10% соответственно;
- 10522.24 - максимальный информационный критерий.

Поскольку $t_{расч.} > t_{крит.}$ на всех уровнях значимости (-2.34 > -2.57 > > -2.87 > -3.44), то рассматриваемый временной ряд является нестационарным и гипотеза H_0 подтверждается.

При анализе и прогнозировании временных рядов широко используются графики автокорреляции и частичной автокорреляции.

График **автокорреляции** показывает связь между текущими и прошлыми значениями временного ряда.

График **частичной автокорреляции** показывает связь между текущим значением ряда и его предыдущими значениями, когда влияние всех промежуточных лагов устранено.



По графику автокорреляции можно заметить, что элементы коррелограммы убывают линейно. Это говорит о том, что временной ряд имеет "долговременную память" и присущие ему свойства и закономерности могут быть с успехом экстраполированы в будущем. Значимый коэффициент корреляции для лагов 7, 14, 21 и 28 намекает на недельную цикличность.

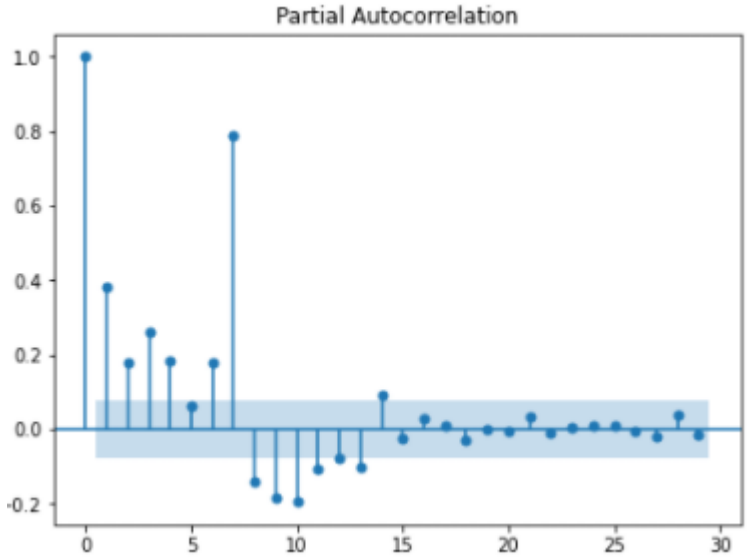
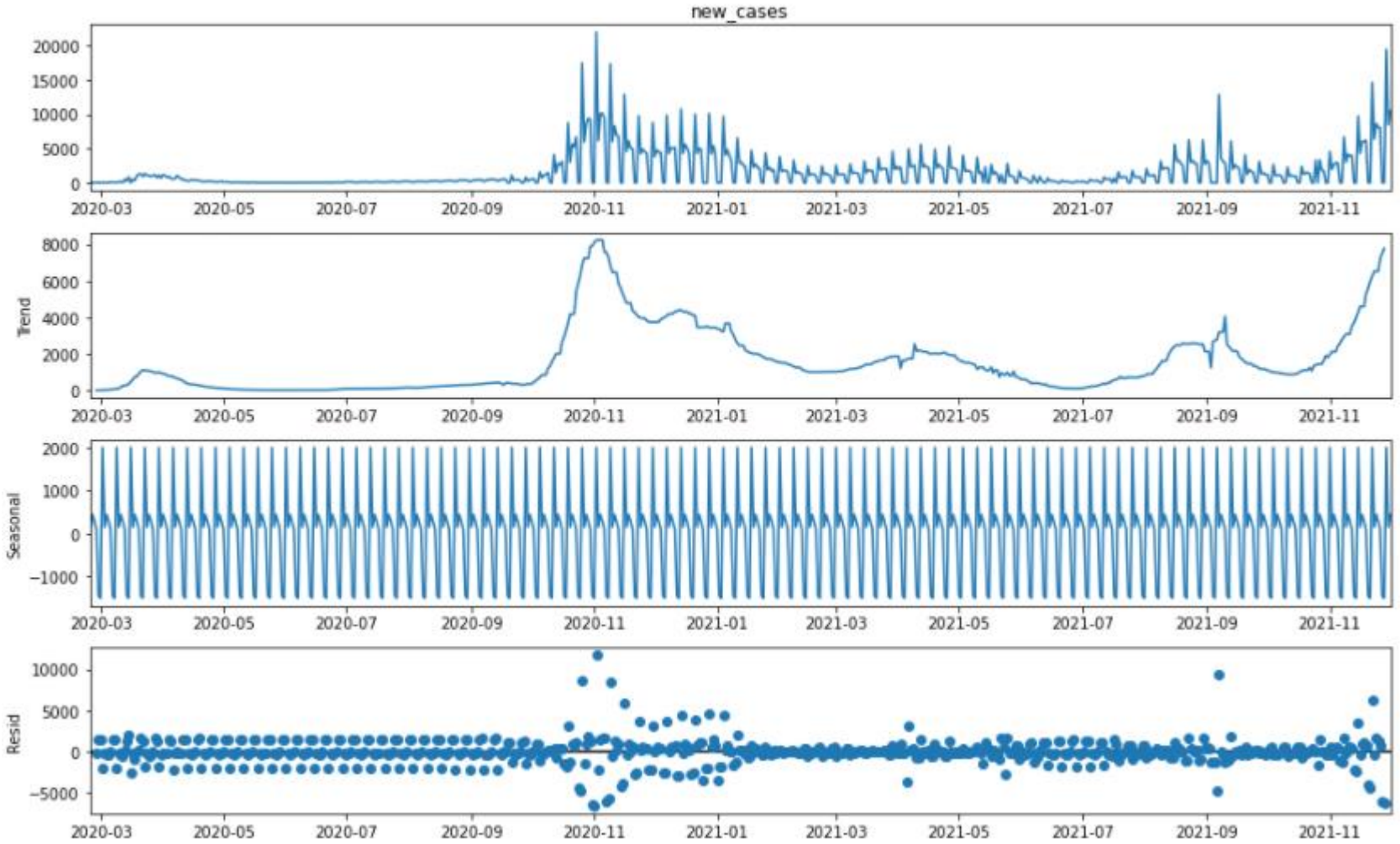


График частичной автокорреляции характеризуется резко выделяющимся значением на седьмом лаге, что также говорит о возможной недельной цикличности.

Чтобы лучше понять данные, с которыми мы работаем, также была проведена **ETS декомпозиция**, а именно разложение исходного ряда данных на составляющие: тренд, сезонная компонента и случайная компонента (остаток).



Сезонная компонента, представленная на графике, подтверждает ранее выдвинутое предположение о наличии недельной сезонности.

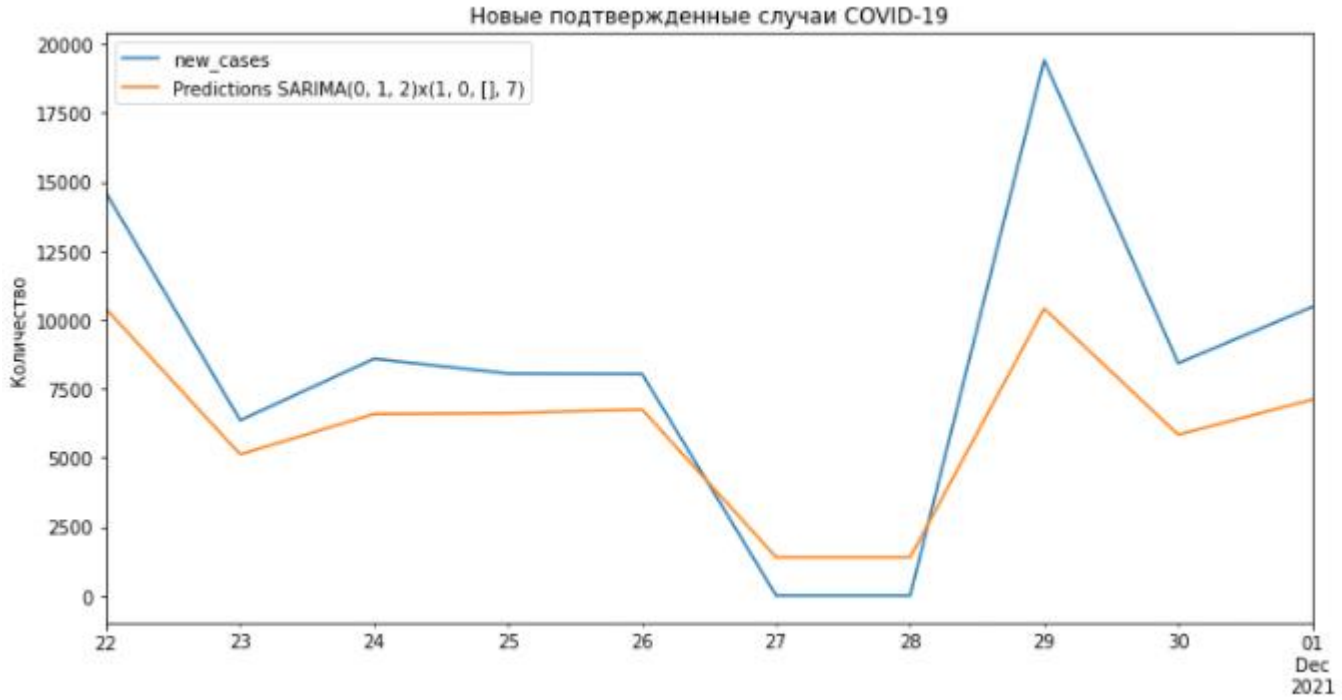
В связи с наличием четко выраженной сезонности для прогнозирования возникновения новых случаев заражения COVID-19 мы будем строить три модели, которые предусматривают наличие сезонной компоненты:

- SARIMA;
- Хольта-Винтерса;
- PROPHET.

Модель SARIMA

В качестве первой модели для прогнозирования данных была построена модель **авторегрессионного интегрированного скользящего среднего с учетом сезонности (SARIMA)**. Построение модели производилось с помощью функции auto_arima, которая в качестве входных данных принимает одномерный временной ряд и использует алгоритм, который объединяет тесты для подбора лучшей модели ARIMA. Для получения модели SARIMA мы указывали в параметрах учет сезонности (в нашем случае недельной сезонности).

Получили следующие оптимальные параметры модели:
(0, 1, 2)x(1, 0, [], 7).



SARIMAX Results

Dep. Variable:	y	No. Observations:	646
Model:	SARIMAX(0, 1, 2)x(1, 0, [], 7)	Log Likelihood	-5433.079
Date:	Wed, 22 Dec 2021	AIC	10874.157
Time:	19:20:48	BIC	10892.034
Sample:	0	HQIC	10881.094
	- 646		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-1.0675	0.018	-58.518	0.000	-1.103	-1.032
ma.L2	0.2522	0.018	14.167	0.000	0.217	0.287
ar.S.L7	0.9048	0.008	119.840	0.000	0.890	0.920
sigma2	1.193e+06	2.5e+04	47.805	0.000	1.14e+06	1.24e+06

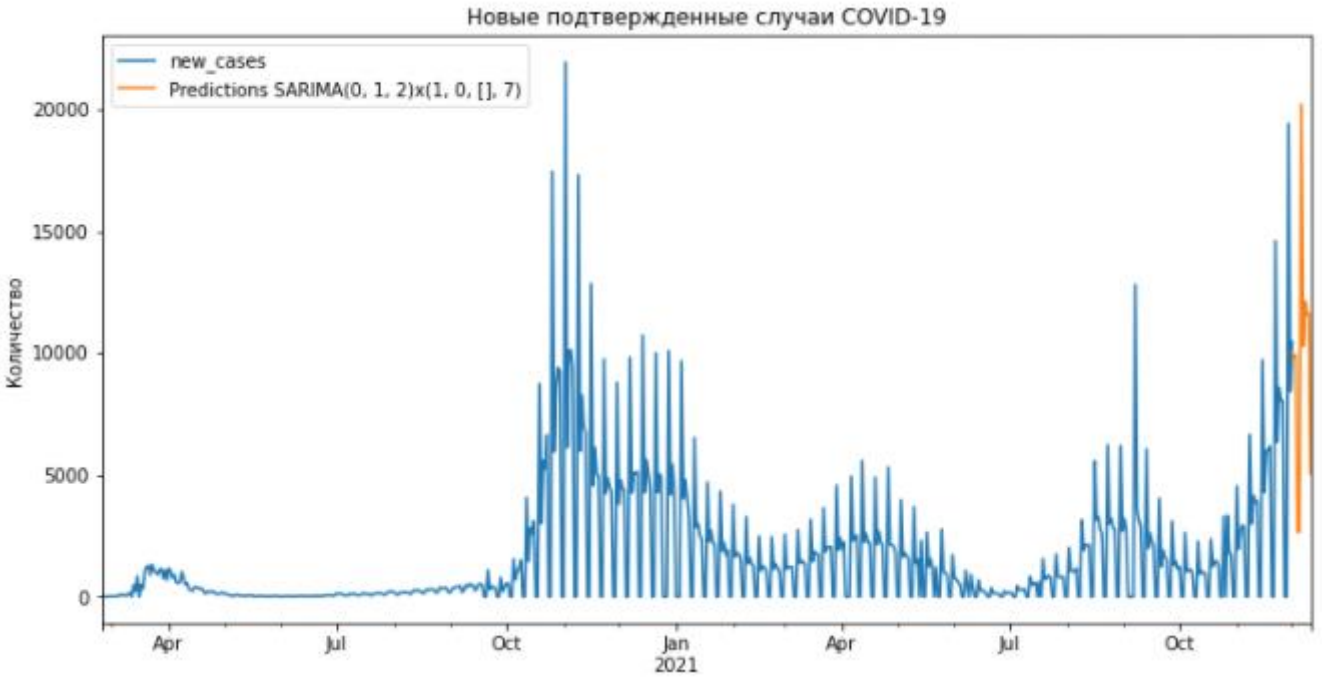
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	24327.44
Prob(Q):	0.91	Prob(JB):	0.00
Heteroskedasticity (H):	58.27	Skew:	0.19
Prob(H) (two-sided):	0.00	Kurtosis:	33.08

```
SARIMAX(0, 1, 2)x(1, 0, [], 7) MSE Error: 12975138.24
SARIMAX(0, 1, 2)x(1, 0, [], 7) RMSE Error:  3602.10192
```

Среднеквадратичная ошибка, полученная по модели SARIMA, составила 3602, таким образом прогноз примерно на 3600 случаев отклоняется от истинных значений заболеваемости в день.

Модель SARIMA

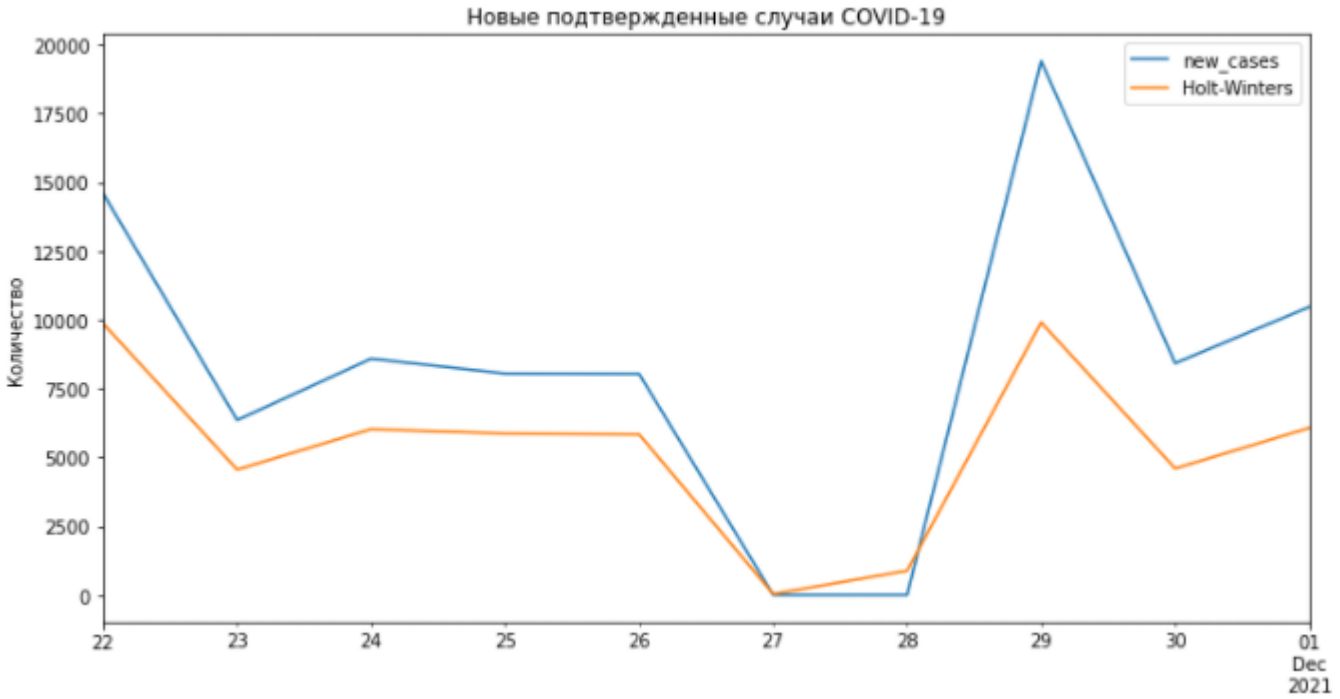
Модель SARIMA достаточно неплохо показала себя на наших данных, дав предсказания на тестовой выборке, в ряде случаев близкие к реальным данным, однако имелись и значительные расхождения, о чем говорит значение среднеквадратичной ошибки в количестве 3602 случаев. Причиной в данном случае можно назвать наличие ряда других признаков, которые оказывают влияние на количество возникновения новых случаев заболевания и не учитываются в нашей модели (зависимость скорости распространения инфекции от плотности населения, индекса самоизоляции, соблюдения мер безопасности и прочих факторов).



Модель Хольта-Винтерса

В качестве второй модели для прогнозирования данных была построена модель Хольта-Винтерса.

Данная модель предусматривает тройное экспоненциальное сглаживание - это расширение экспоненциального сглаживания, которое явно добавляет поддержку сезонности в одномерный временной ряд. Тройное экспоненциальное сглаживание реализовывается в Python с помощью класса ExponentialSmoothing Statsmodels.



ExponentialSmoothing Model Results

Dep. Variable:	new_cases	No. Observations:	636
Model:	ExponentialSmoothing	SSE	772106912.596
Optimized:	True	AIC	8932.001
Trend:	Additive	BIC	8981.008
Seasonal:	Additive	AICC	8932.586
Seasonal Periods:	7	Date:	Wed, 22 Dec 2021
Box-Cox:	False	Time:	17:51:47
Box-Cox Coeff.:	None		

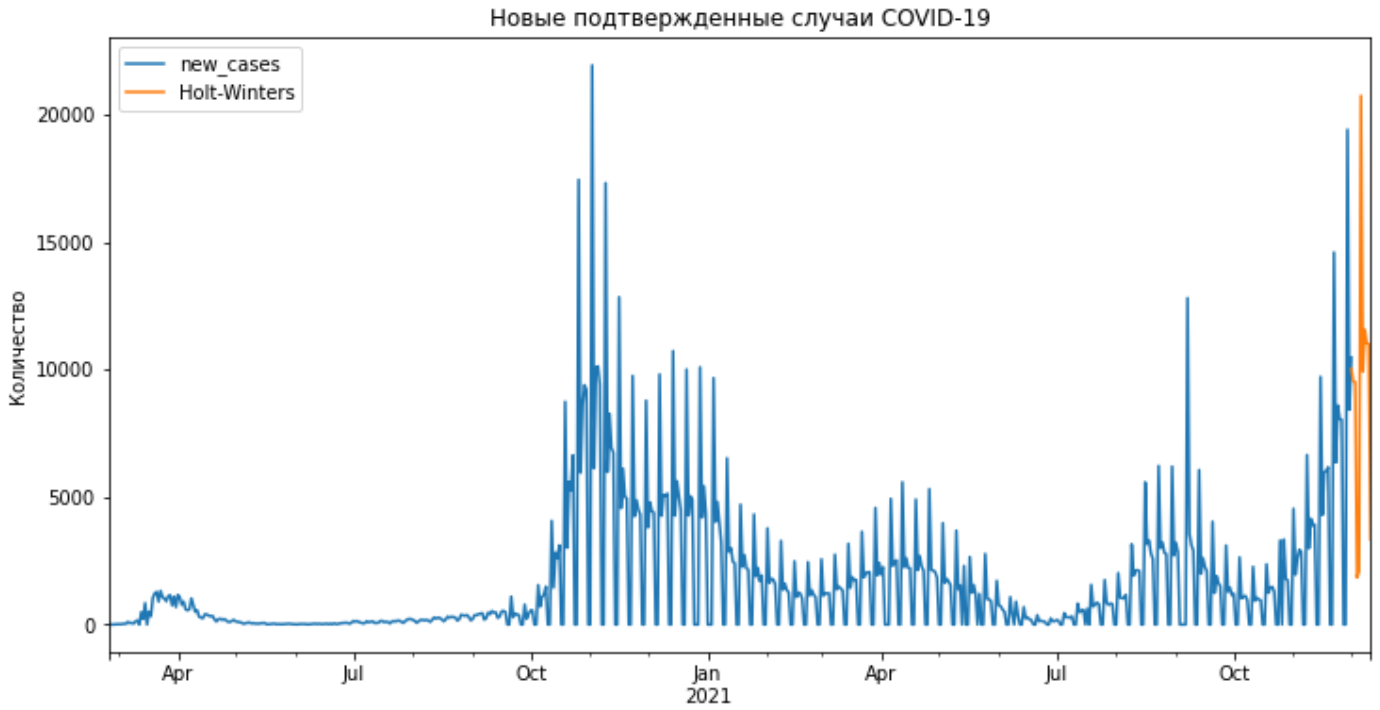
	coeff	code	optimized
smoothing_level	0.1110714	alpha	True
smoothing_trend	0.0001	beta	True
smoothing_seasonal	0.7901587	gamma	True
initial_level	-9.7619048	l.0	True
initial_trend	7.9748918	b.0	True
initial_seasons.0	-29.378571	s.0	True
initial_seasons.1	-112.52143	s.1	True
initial_seasons.2	-28.592857	s.2	True
initial_seasons.3	133.48571	s.3	True
initial_seasons.4	42.335714	s.4	True
initial_seasons.5	69.157143	s.5	True
initial_seasons.6	-74.485714	s.6	True

```
Holt-Winters MSE Error: 16568926.11
Holt-Winters RMSE Error: 4070.494578
```

Среднеквадратичная ошибка, полученная по модели Хольта-Винтерса, составила 4070, таким образом прогноз примерно на 4070 случаев отклоняется от истинных значений заболеваемости в день.

Модель Хольта-Винтерса

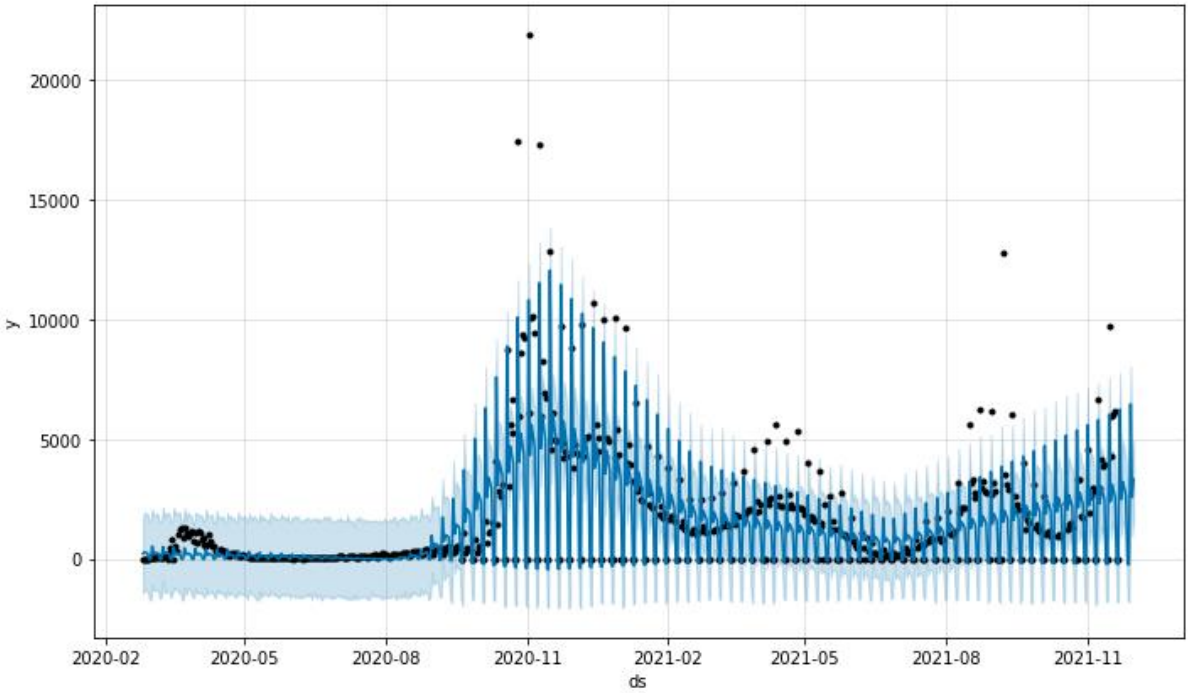
Модель Хольта-Винтерса также достаточно неплохо показала себя на наших данных, дав предсказания на тестовой выборке, в ряде случаев близкие к реальным данным, однако имелись и значительные расхождения, о чем говорит значение среднеквадратичной ошибки в количестве 4070 случаев. Причиной в данном случае можно назвать наличие ряда других признаков, которые оказывают влияние на количество возникновения новых случаев заболевания и не учитываются в нашей модели, а также то, что мы не приводили ряд к стационарному перед построением модели, что могло также сказаться на ухудшении прогнозирования.

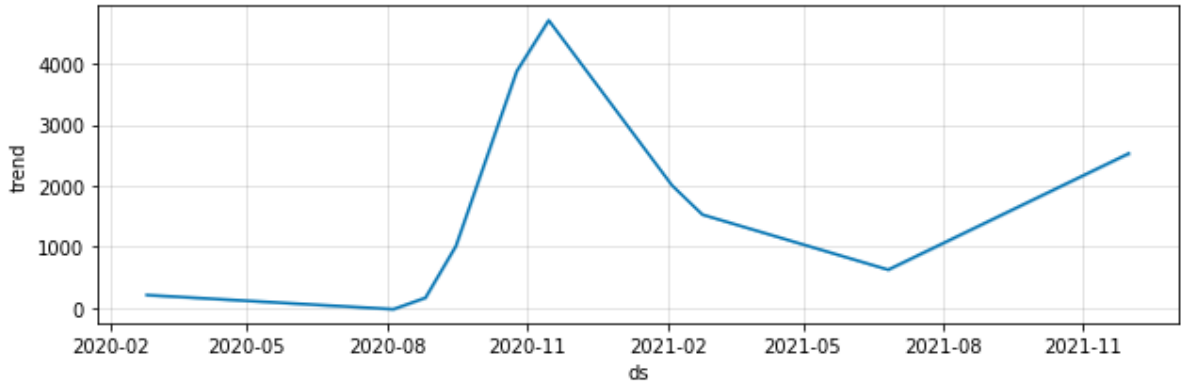


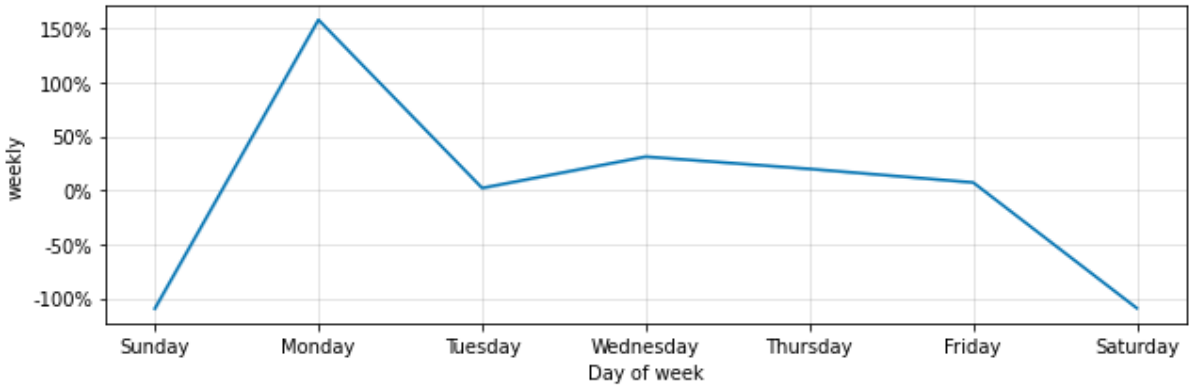
Модель PROPHET

В качестве третьей модели для прогнозирования данных была построена модель Prophet.

В основе данной модели лежит процедура подгонки аддитивных регрессионных моделей со следующими четырьмя основными компонентами: тренд, сезонность, аномальные дни (праздники) и ошибки (содержит информацию, которая не учтена моделью). При построении модели не была найдена годовая и ежедневная сезонность, зато была установлена еженедельная сезонность, что подтверждает выводы, полученные ранее.







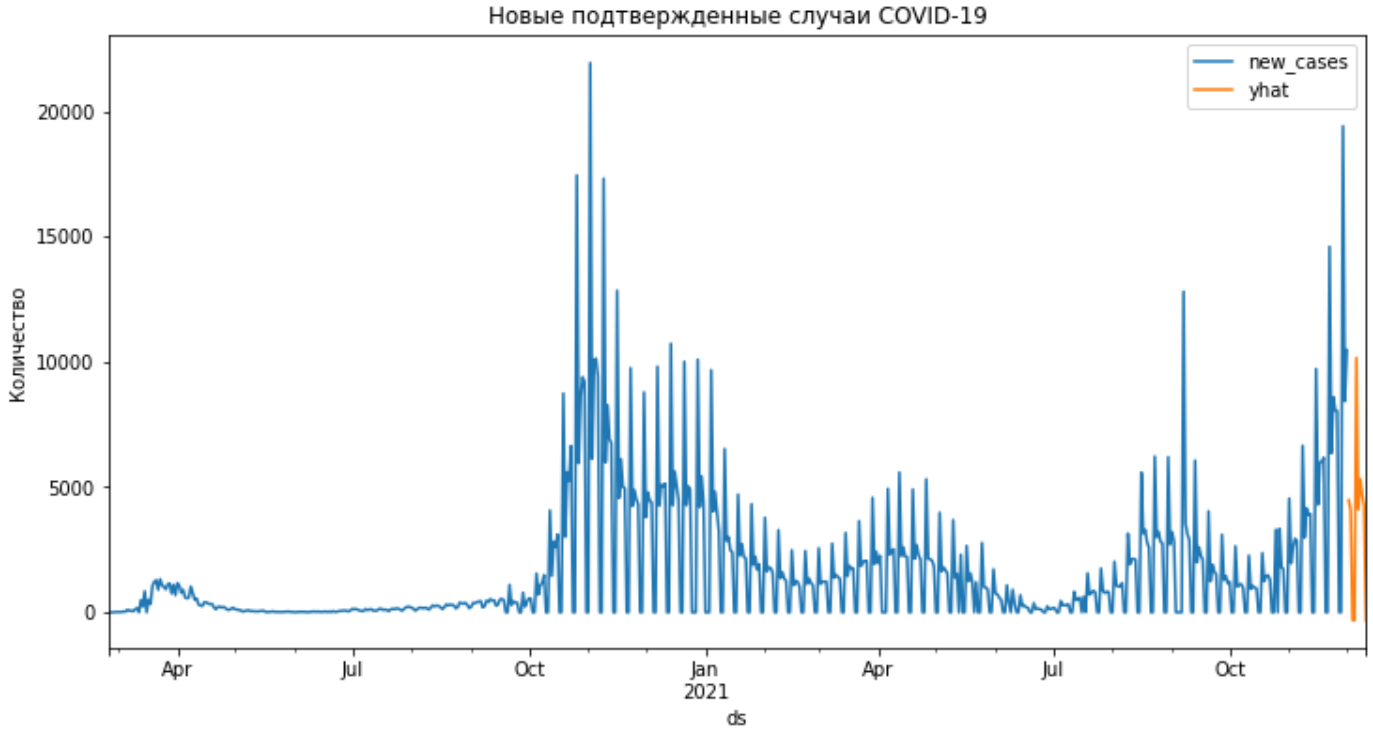

```
Prophet MSE Error: 42162058.85
Prophet RMSE Error: 6493.231772
```

Среднеквадратичная ошибка, полученная по модели Prophet, составила 6493, таким образом прогноз примерно на 6493 случаев отклоняется от истинных значений заболеваемости в день.

Модель PROPHET

Модель PROPHET показала себя на наших данных хуже всех (значение среднеквадратичной ошибки в количестве 6493 случаев), однако дала более подробное представление о еженедельной сезонности.

По графику еженедельной сезонности на предыдущем слайде видно, что повышение количества выявленных заболеваний приходится на понедельник и снижается на выходных, что вполне объяснимо тем, что заболевшие обращались в медицинские учреждения в будние дни больше, чем в выходные.

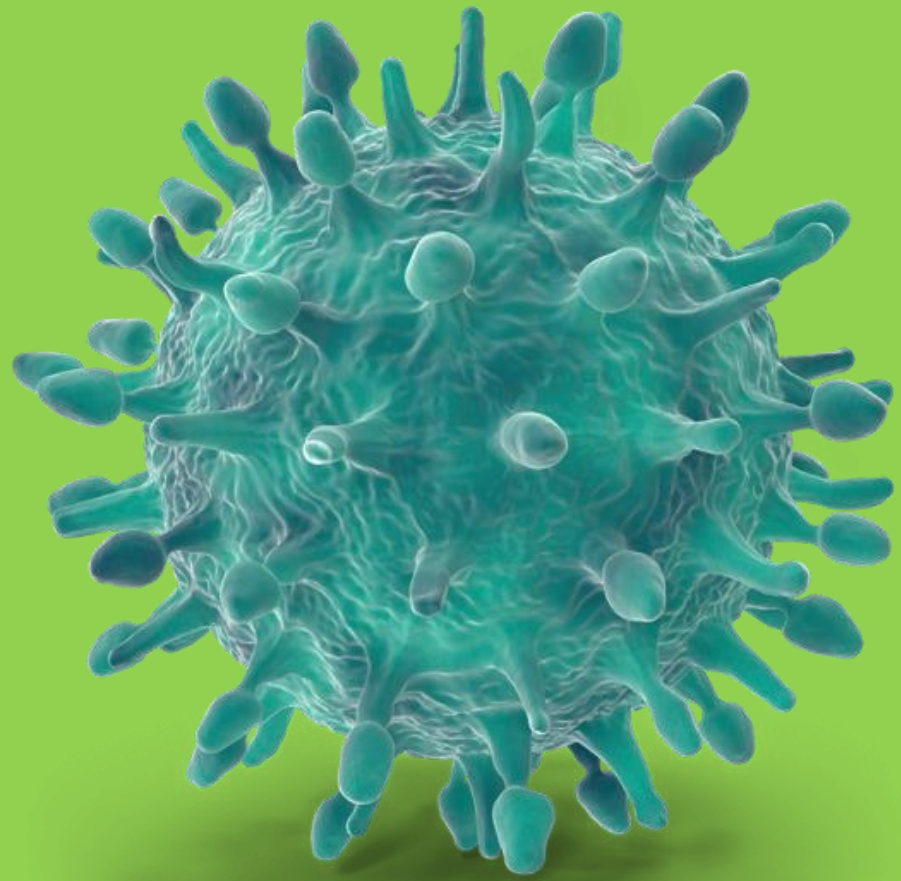


Вывод

Исходя из проведенного исследования можно сделать вывод, что требуется разработка новых подходов к моделированию и прогнозированию временных рядов в условиях непредсказуемости поведения исследуемого процесса. Говорить о построении модели, которая адекватно описывала бы процесс распространения коронавирусной инфекции, в настоящий момент преждевременно, поскольку прошло мало времени с начала пандемии, нет возможности оценить возможный сезонный характер данных (например, годовую сезонность). Однако на примере данных Швейцарии мы получили результат, который говорит о том, что из трех рассмотренных моделей наилучшим образом для прогнозирования подошла модель SARIMA с установленным параметром еженедельной сезонности.

В прогнозах, представленных на 10 дней, по выявлению новых случаев COVID-19 можно усмотреть тенденцию на возрастание, таким образом можно сказать что повторяется ситуация прошлого года, когда был зафиксирован резкий скачок числа заболеваний, приходящийся на рождественские праздники и рост числа туристов в Швейцарии. При этом он будет не такой значительный, поскольку активно проводится иммунизация населения.

В качестве направлений дальнейших исследований можно рассматривать изучение и выявление особенностей других моделей прогнозирования временных рядов, в том числе учитывающих какие-либо дополнительные параметры, которые влияют на распространение заболевания на территории Швейцарии.



Спасибо за внимание!