Прогнозирование новых случаев заболевания COVID –19 в Мексике

А.Н.Петрова

Введение

В марте 2020 г. Всемирная организация здравоохранения объявила пандемию вируса SARS-CoV-2 2019 (COVID-19). Пандемия оказала большое влияние на глобальную экономику, здравоохранение и глобализацию. На глобальном уровне необходимо отслеживать текущую нагрузку и прогнозировать скорость и характер распространения, чтобы направлять осведомленность в области здравоохранения, готовность применить ответные меры. Прогнозирование является одним из ключевых элементов моделирования.

Постановка целей и задач

Цель исследования - анализ и прогноз динамики пандемии COVID-19, на основе статистических данных о заболеваемости в Мексике.

Для достижения поставленной цели необходимо решение следующих задач:

- провести анализ данных о распространении COVID-19;
- подтвердить или опровергнуть зависимость между новыми случаями заболевания и количеством вакцинированных;
- построить прогнозные модели новых случаев заболевания;
- оценить эффективность предложенных моделей прогнозирования.

Импорт библиотек, ознакомление с данными

- Импорт библиотек, моделей и необходимых метрик
- Импорт данных owid-covid-data.csv
- Знакомство с данными
  1. location - географическое положение;
  2. date - дата наблюдения;
  3. new_cases - новые случаи заболевания;
  4. total_cases - случаи заболевания с нарастающим значением;
  5. new_vaccinations - введены новые дозы вакцинации против COVID-19;
  6. people_fully_vaccinated - общее количество людей, получивших все дозы, предписанные протоколом вакцинации.

Предобработка данных

- Фильтрация данных

	location	date	new_cases	total_cases	new_vaccinations	people_fully_vaccinated
88177	Mexico	2020-01-01	NaN	NaN	NaN	NaN
88178	Mexico	2020-01-02	NaN	NaN	NaN	NaN
88179	Mexico	2020-01-03	NaN	NaN	NaN	NaN
88180	Mexico	2020-01-04	NaN	NaN	NaN	NaN
88181	Mexico	2020-01-05	NaN	NaN	NaN	NaN

- Обработка пропусков

```
location                  0
date                      0
new_cases                58
total_cases              58
new_vaccinations        438
people_fully_vaccinated 432
dtype: int64
```

- Проверка типа данных

```
location                  object
date                      object
new_cases                float64
total_cases              float64
new_vaccinations         float64
people_fully_vaccinated  float64
dtype: object
```

EDA (Exploratory Data Analysis)

Разведочный анализ данных (Exploratory Data Analysis) – предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками.

Входная выборка для анализа

date	location	date	new_cases	total_cases	new_vaccinations	people_fully_vaccinated
2020-01-01	Mexico	2020-01-01	0.0	0.0	0.0	0.0
2020-01-02	Mexico	2020-01-02	0.0	0.0	0.0	0.0
2020-01-03	Mexico	2020-01-03	0.0	0.0	0.0	0.0
2020-01-04	Mexico	2020-01-04	0.0	0.0	0.0	0.0
2020-01-05	Mexico	2020-01-05	0.0	0.0	0.0	0.0

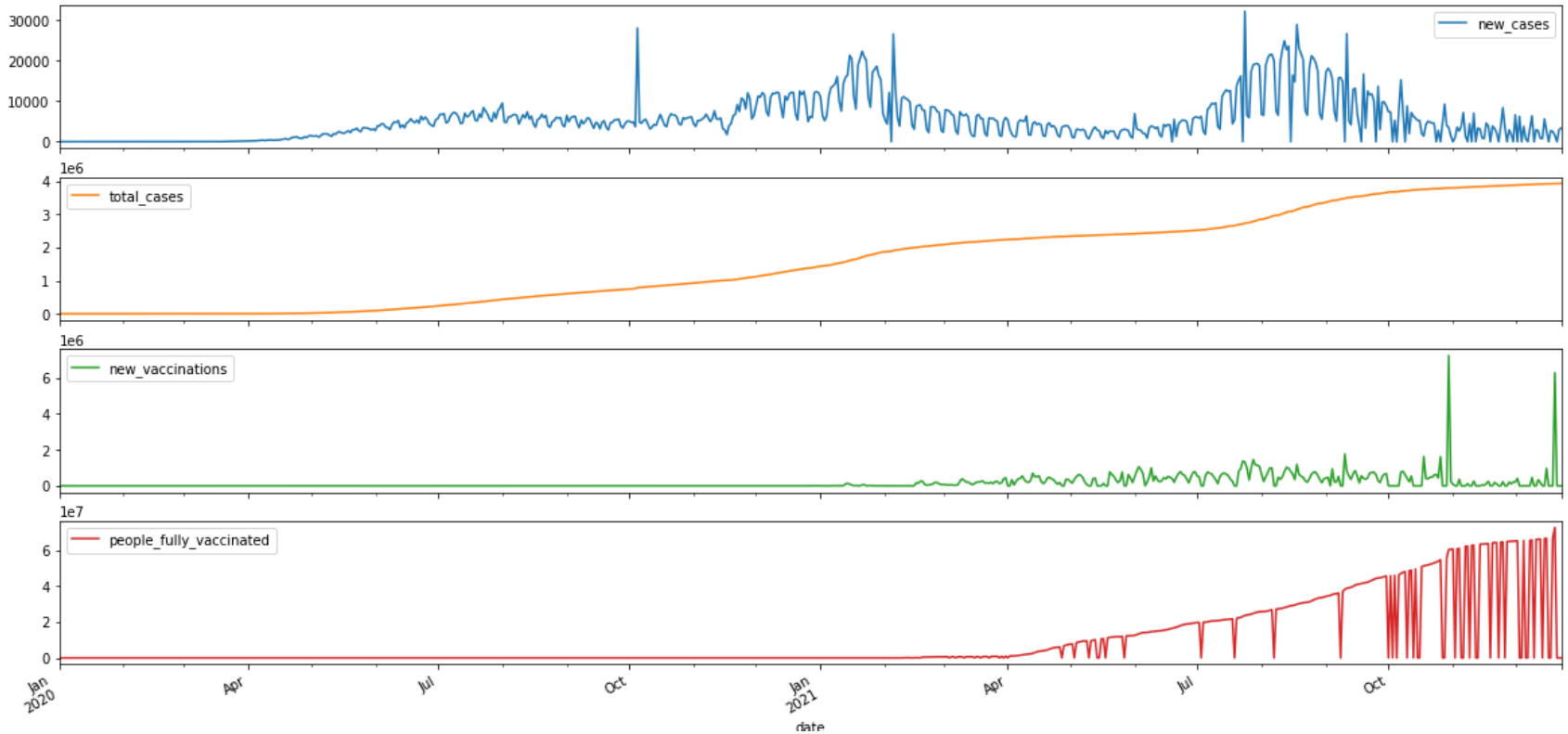
Основные статистические метрики

index	new_cases	total_cases	new_vaccinations	people_fully_vaccinated
count	723.0	723.0	723.0	723.0
mean	5450.070539419087	1584523.0719225449	169250.47579529736	9429359.590594744
std	5381.233439095743	1340547.9146147636	446688.8762389309	17871542.521691795
min	0.0	0.0	0.0	0.0
25%	1442.5	223373.0	0.0	0.0
50%	4448.0	1383434.0	0.0	0.0
75%	6763.0	2500882.5	208000.5	11752341.0
max	32244.0	3940401.0	7246123.0	72649923.0

Размер выборки 723. Общее число заболевших в Мексике составляет 3 940 401 чел. Максимальная численность выявленных заболевших 32 244 чел.

Прошедших все этапы вакцинирования за время пандемии составило – 7 246 123 чел., что составляет 55% от общей численности населения страны.

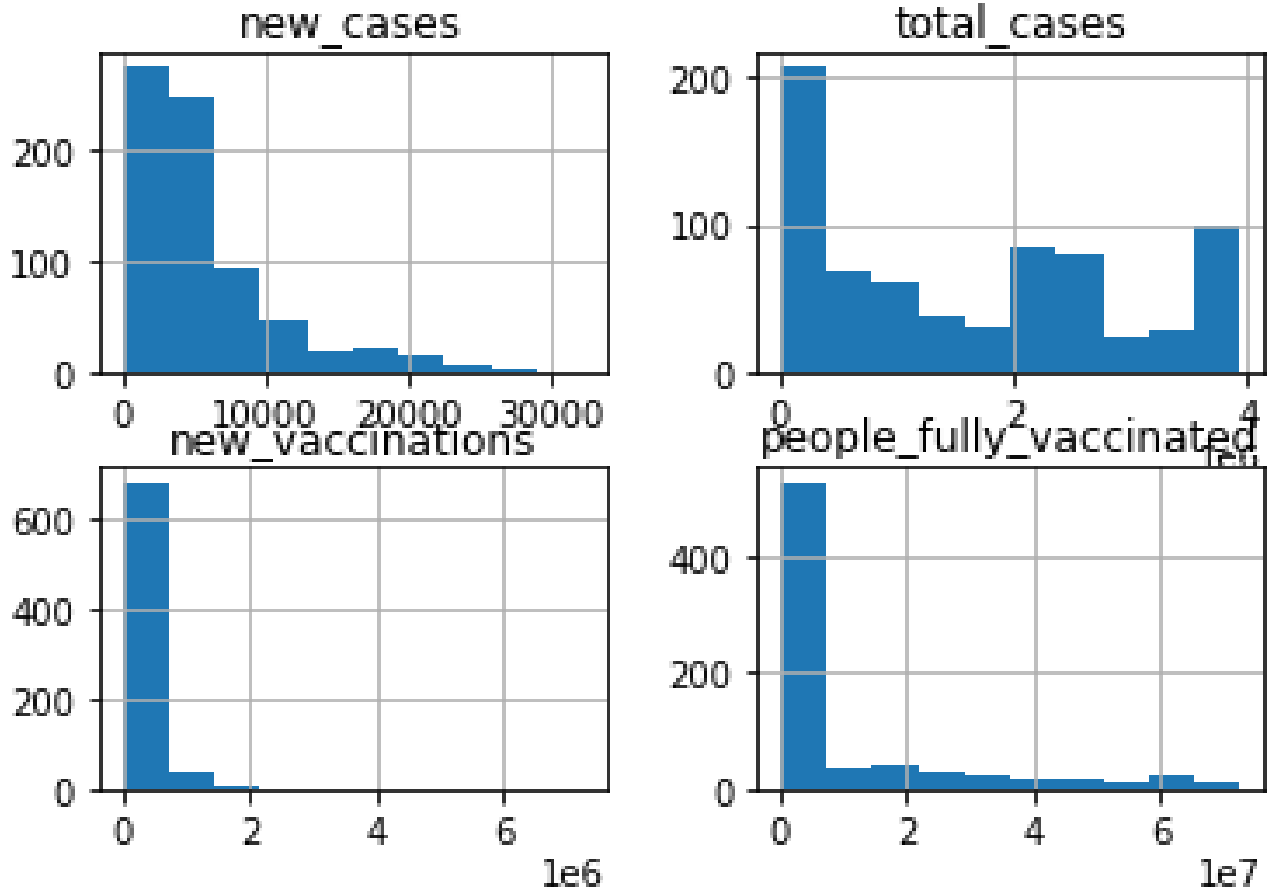
Новые случаи заболевания и вакцинации в Мексике



Анализируемый период с 2020-01-01 по 2021-12-23 .
С октября 2021 г. отмечается снижение уровня заболеваемости, в тоже время отмечается пиковый выброс вакцинации населения приходящийся на данный период. Общее количество заболевших выходит на плато.

Новые случаи заболевания и вакцинации в Мексике

Отсутствие нормального распределения позволяет сделать вывод о пиковых нагрузках в период пандемии. Это отчетливо видно на общем графике.



Матрица корреляции признаков

	new_cases	total_cases	new_vaccinations	people_fully_vaccinated
new_cases	1.000000	0.295854	0.187410	0.137811
total_cases	0.295854	1.000000	0.402635	0.716544
new_vaccinations	0.187410	0.402635	1.000000	0.437865
people_fully_vaccinated	0.137811	0.716544	0.437865	1.000000

На коэффициент корреляции оказывает влияние нулевых показателей, напомним что по исследуемым данным выявлено множество отсутствующих значений. По данным матрицы величина корреляции между новыми случаями заболевания и общим числом вакцинированных составляет 0,1. Оба показателя линейно независимы друг от друга, что так же можно отметить на графике. Гипотеза о зависимости новых случаев заболевания и общим числом вакцинированных не подтверждена.

Можно отметить средний уровень корреляции между общим количеством заболевших и общим количеством вакцинированных 0,71, данная взаимосвязь не значимая, что говорит о разных факторах влияния на два показателя.

Построение моделей и анализ результатов

- Обучающая выборка
  train = df_new.iloc[:len(df_new)-10]

train

date	location	date	new_cases	total_cases	new_vaccinations	people_fully_vaccinated
2020-01-01	Mexico	2020-01-01	0.0	0.0	0.0	0.0
2020-01-02	Mexico	2020-01-02	0.0	0.0	0.0	0.0
2020-01-03	Mexico	2020-01-03	0.0	0.0	0.0	0.0
2020-01-04	Mexico	2020-01-04	0.0	0.0	0.0	0.0
2020-01-05	Mexico	2020-01-05	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...
2021-12-09	Mexico	2021-12-09	6395.0	3911714.0	459229.0	65630611.0
2021-12-10	Mexico	2021-12-10	0.0	3911714.0	0.0	0.0
2021-12-11	Mexico	2021-12-11	2992.0	3914706.0	0.0	66003384.0
2021-12-12	Mexico	2021-12-12	2655.0	3917361.0	340364.0	66150375.0
2021-12-13	Mexico	2021-12-13	855.0	3918216.0	187521.0	66225140.0

713 rows × 6 columns

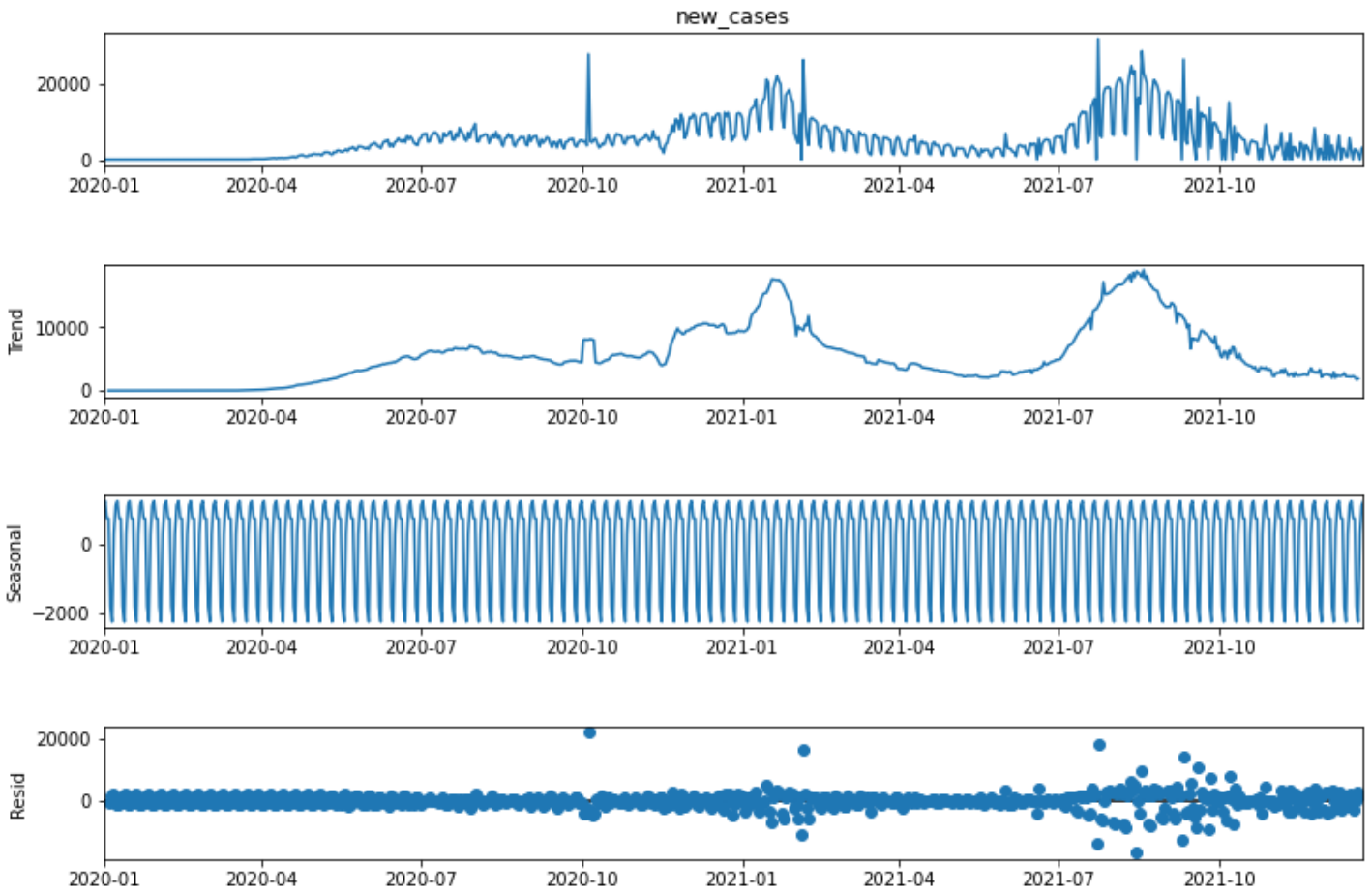
- Тестовая выборка test = df_new.iloc[len(df_new)-10:]

test

date	location	date	new_cases	total_cases	new_vaccinations	people_fully_vaccinated
2021-12-14	Mexico	2021-12-14	771.0	3918987.0	0.0	0.0
2021-12-15	Mexico	2021-12-15	5651.0	3924638.0	0.0	66459570.0
2021-12-16	Mexico	2021-12-16	2627.0	3927265.0	968736.0	66586509.0
2021-12-17	Mexico	2021-12-17	0.0	3927265.0	0.0	0.0
2021-12-18	Mexico	2021-12-18	2750.0	3930015.0	0.0	0.0
2021-12-19	Mexico	2021-12-19	2530.0	3932545.0	0.0	66740075.0
2021-12-20	Mexico	2021-12-20	1557.0	3934102.0	6287242.0	72649923.0
2021-12-21	Mexico	2021-12-21	0.0	3934102.0	0.0	0.0
2021-12-22	Mexico	2021-12-22	2980.0	3937082.0	0.0	0.0
2021-12-23	Mexico	2021-12-23	3319.0	3940401.0	0.0	0.0

ETS декомпозиция

ETS расшифровывается как Error-Trend-Seasonality и представляет собой модель, используемую для декомпозиции временных рядов.



Метод прогнозирования - SARIMAX

SARIMAX - сезонная авторегрессионная интегрированная скользящая средняя с экзогенными регрессорами.
Основой данной модели является авторегрессионное интегрированное скользящее среднее или ARIMA - метод прогнозирования для одномерных данных временных рядов, поддерживает элементы авторегрессии и скользящего среднего.

- Запускаем pmdarima.auto_arima чтобы получить набор параметров для нашей модели

  auto_arima(df_new['new_cases'],seasonal=True, m=7).summary()

  auto_arima SARIMAX(1, 1, 2)x(1, 0, 1, 7)

- создаем модель с подобранными параметрами

  model_sarimax = SARIMAX(train['new_cases'],order=(1, 1, 2), seasonal_order=(1, 0, 1, 7))

- обучаем модель на обучающей выборке данных

  results_sarimax = model_sarimax.fit()

- получаем результаты

  results_sarimax.summary()

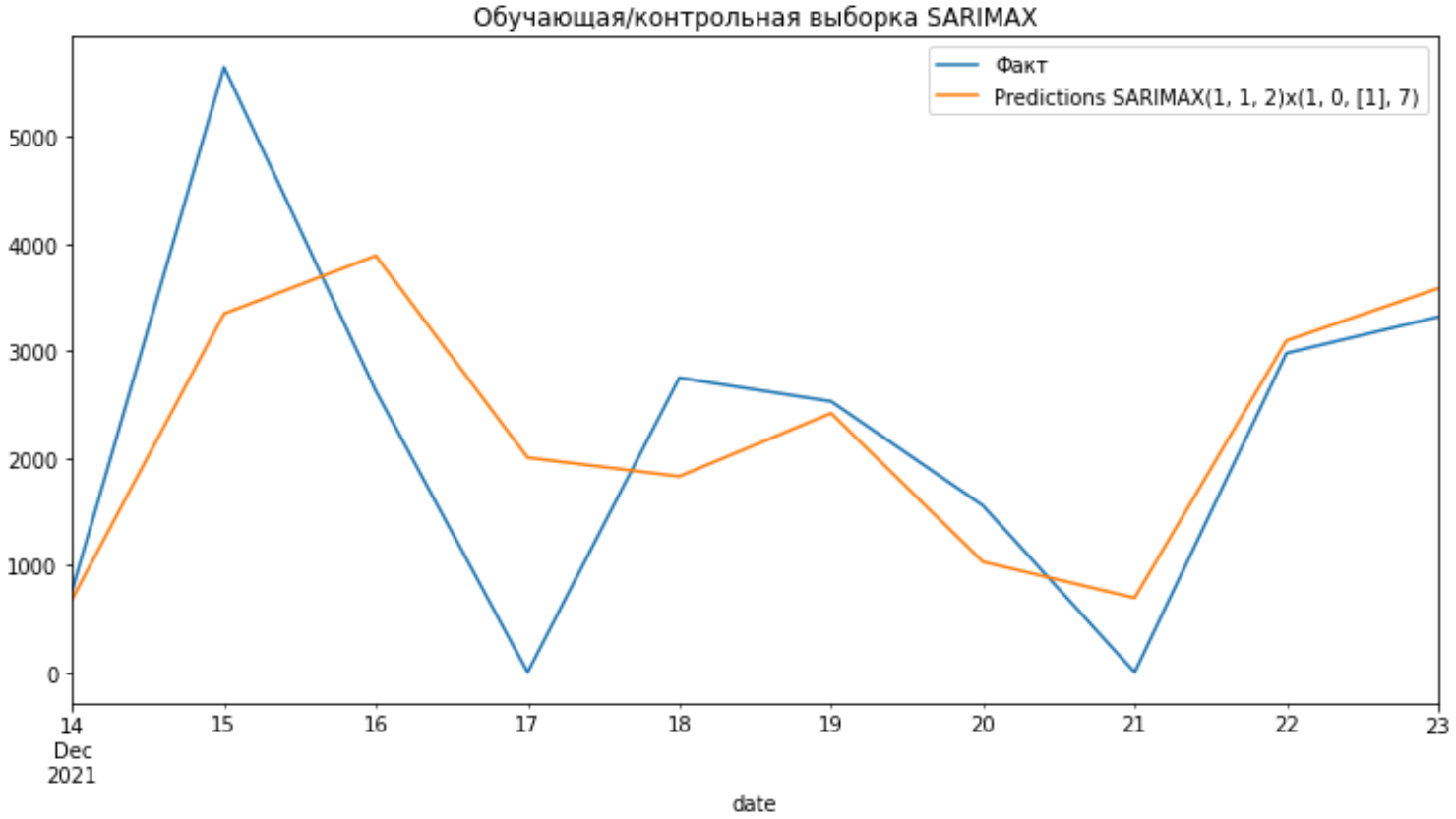
Метод прогнозирования - SARIMAX

- предсказываем значения, передав модели results точку начала и окончания

prediction_sarimax = results_sarimax.predict(start=len(train), end=len(train)+len(test)-1, dynamic=False, typ='levels').rename(' Predictions SARIMAX(1, 1, 2)x(1, 0, [1], 7)

- сравниваем прогноз и тестовую выборку

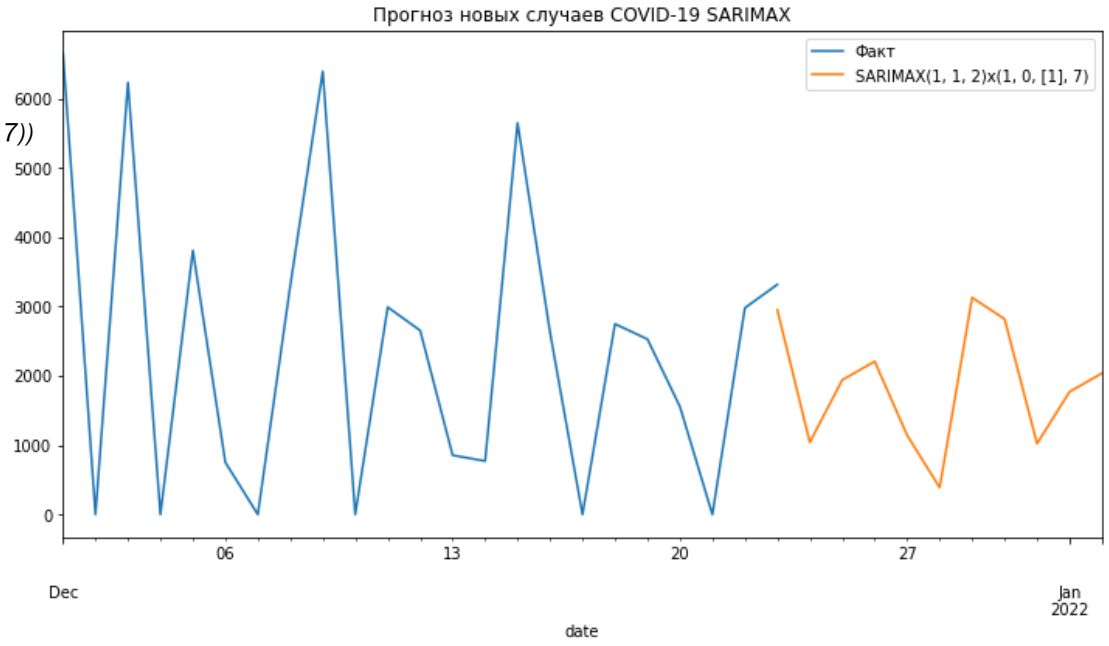
results_sarimax = model_sarimax.fit()



Метод прогнозирования - SARIMAX

- оцениваем качество модели методом MSE, RMSE, MAE, MAPE
  - MAPE Error: inf
  - MAE Error: 828.9008889
  - MSE Error: 1262404.078
  - RMSE Error: 1123.567567
- прогноз на будущее - 10 дней вперед

```
fut_model_sarimax = SARIMAX(df_new['new_cases'],order=(1, 1, 2), seasonal_order=(1, 0, [1], 7))
fut_results_sarimax = fut_model_sarimax.fit()
fut_results_sarimax.summary()
```



Метод прогнозирования - PROPHET

Prophet - это процедура для прогнозирования данных временных рядов на основе аддитивной модели, в которой нелинейные тенденции соответствуют годовой, еженедельной и дневной сезонности, а также праздничным эффектам. Он лучше всего работает с временными рядами, которые имеют сильные сезонные эффекты и несколько сезонов исторических данных. Prophet устойчив к отсутствию данных и сдвигам в тренде и, как правило, хорошо справляется с выбросами.

- Переименуем столбцы в обучающем, тестовом и входящем датасетах, чтобы они подходили для использования методов Prophet

```
train_prophet.columns = ['ds', 'y']

test_prophet.columns = ['ds', 'y']

df_new_prophet.columns = ['ds', 'y']
```

- создаем модель с подобранными параметрами

```
model_prophet = Prophet(seasonality_mode='multiplicative')
```

- обучаем модель на обучающей выборке данных

```
model_prophet.fit(train_prophet)
```

- создаем датафрейм на 10 дн. вперед

```
future_prophet = model_prophet.make_future_dataframe(periods=10)
```

Метод прогнозирования - PROPHET

- предсказываем значения по модели

  prediction_prophet = model_prophet.predict(future_prophet)

  prediction_prophet.head()

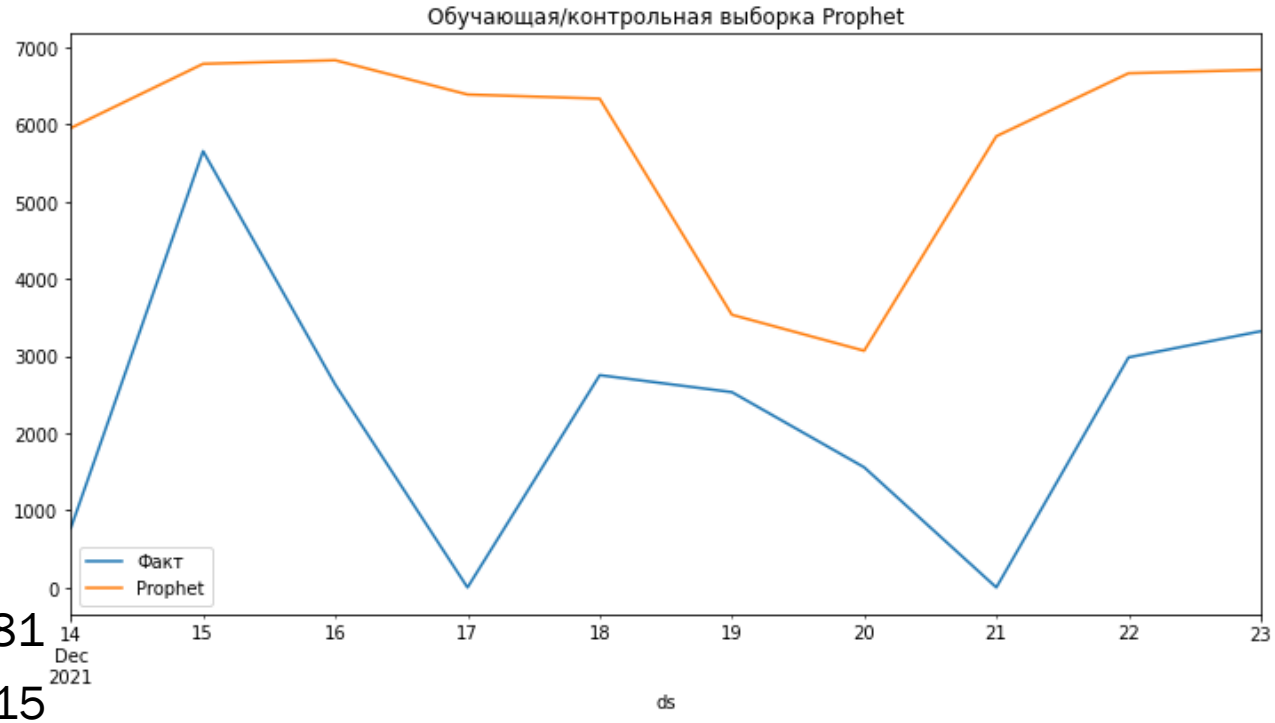
- устанавливаем индекс, сортируем необходимые поля

  prediction_prophet.index = prediction_prophet.ds

  prediction_prophet.head()

  prediction_prophet[['ds', 'yhat']]

- сравниваем прогноз и тестовую выборку
- оцениваем качество модели методом MSE, RMSE, MAE, MAPE
  - Prophet MAPE Error: inf
  - Prophet MAE Error: 3590.079981
  - Prophet MSE Error: 16180765.15
  - Prophet RMSE Error: 4022.532181



Метод прогнозирования - PROPHET

- обучаем модель на всем датасете

fut_model_prophet = Prophet(seasonality_mode='multiplicative')

fut_model_prophet.fit(df_new_prophet)

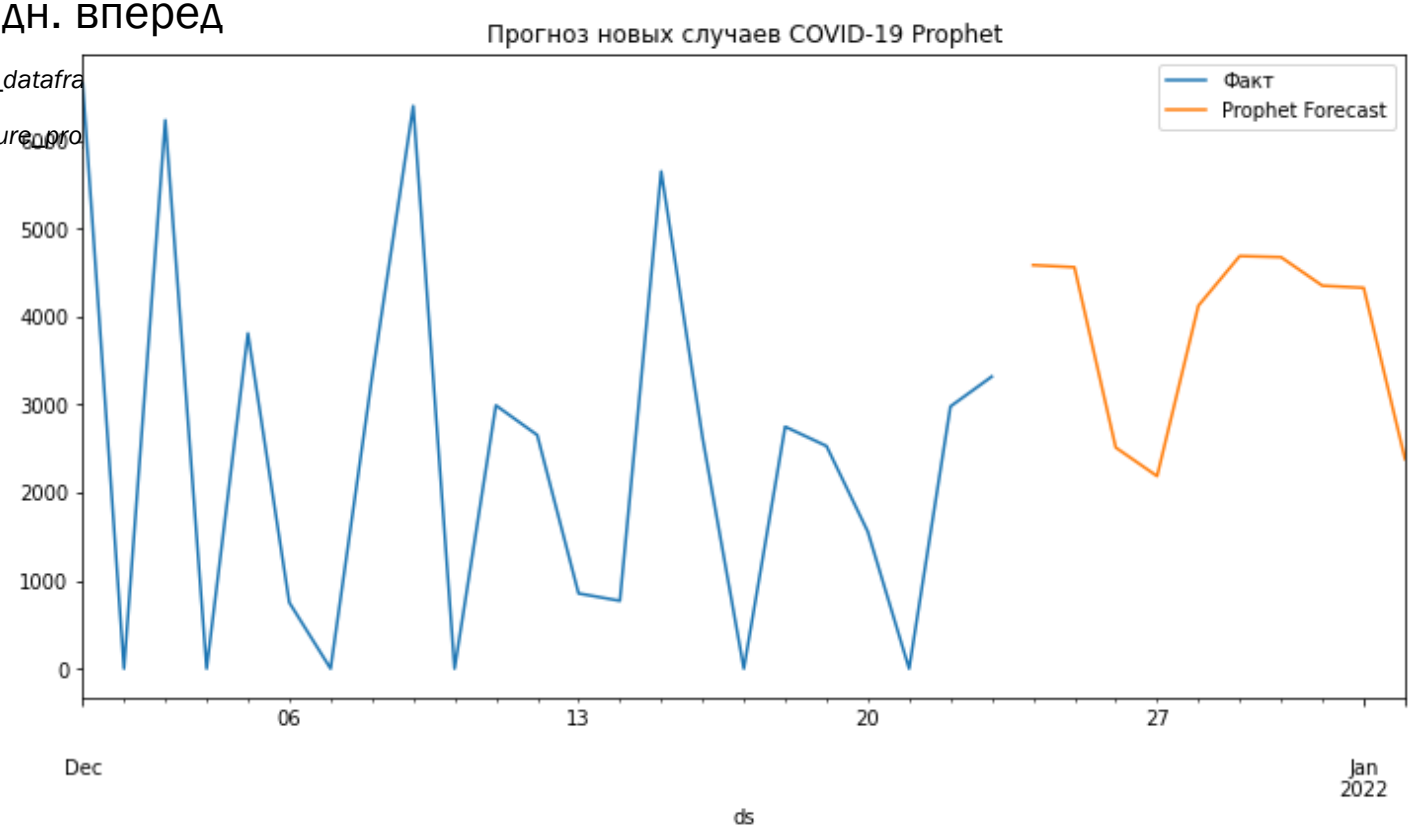
- формируем дата фрейм на 10 дн. вперед

fut_future_prophet = fut_model_prophet.make_future_datafra

fut_fcast_prophet = fut_model_prophet.predict(fut_future_pro

- устанавливаем индекс

fut_fcast_prophet.index = fut_fcast_prophet.ds



Метод прогнозирования Exponential smoothing

Прогнозы, полученные с использованием методов экспоненциального сглаживания, представляют собой средневзвешенные значения прошлых наблюдений, причем веса экспоненциально убывают по мере того, как наблюдения становятся старше.

- Создаем модель с подобранными параметрами

```
model_exps = ExponentialSmoothing(train['new_cases'], seasonal_periods=7, trend = 'add')
```

- обучаем модель на обучающей выборке данных

```
model_exps.fit()
```

- предсказываем значения,передав модели results точку начала и окончания

```
prediction_exps = model_exps.predict(model_exps.params, start=test.index[0], end=test.index[-1])
```

- преобразуем в датафрейм с индексами

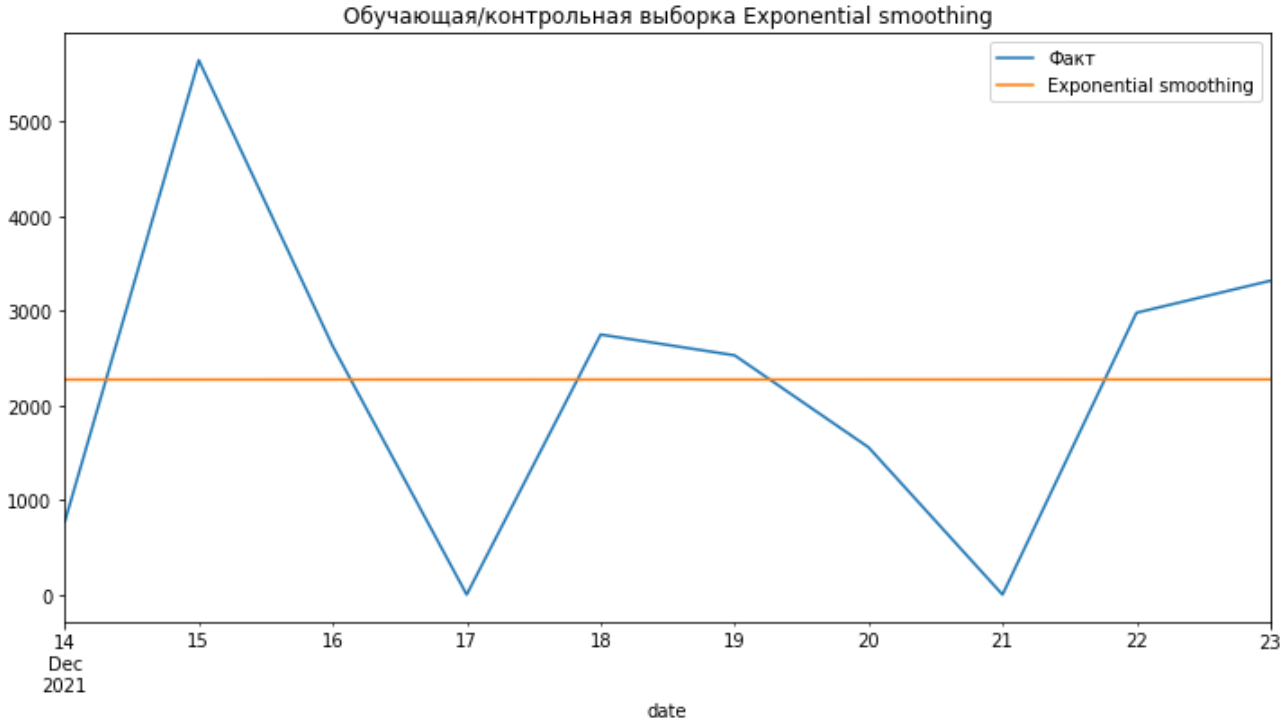
```
prediction_exps = pd.DataFrame(prediction_exps)

prediction_exps.index = pd.date_range("2021-12-14 00:00:00", periods=10, freq="D")

prediction_exps.columns = ['prediction_exps']
```

Метод прогнозирования Exponential smoothing

- сравниваем прогноз и тестовую выборку
- оцениваем качество модели методом MSE, RMSE, MAE, MAPE
  - Exponential smoothing MAPE Error: inf
  - Exponential smoothing MAE Error: 1298.000521
  - Exponential smoothing MSE Error: 2652753.816
  - Exponential smoothing RMSE Error: 1628.727668



Метод прогнозирования Exponential smoothing

- обучаем модель на всем датасете

fut_model_exps = ExponentialSmoothing(df_new['new_cases'],seasonal_periods=7, trend = 'add')

fut_model_exps.fit()

- задаем точки будущего

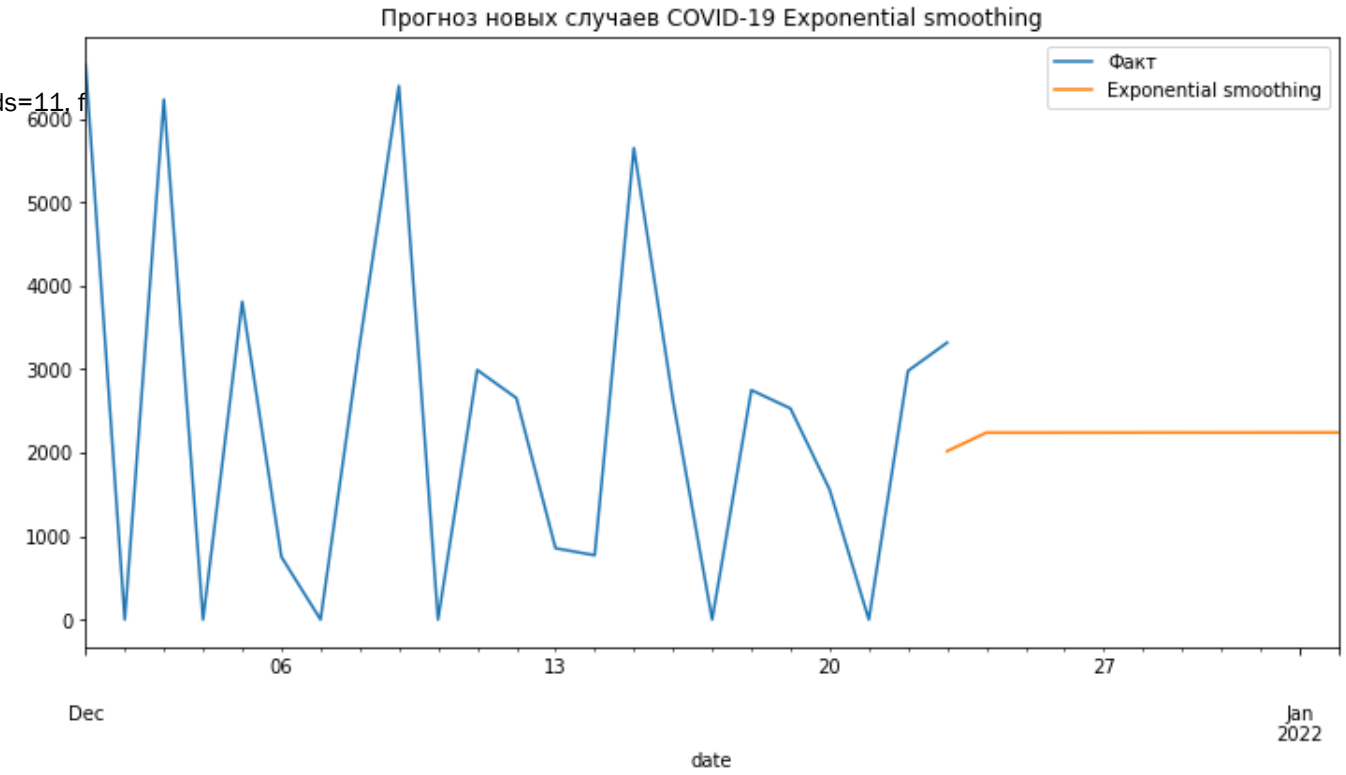
fut_fcast_exps = fut_model_exps.predict(fut_model_exps.params, start=len(df_new)-1, end=len(df_new)+9)

- преобразуем в датафрейм с индексами

fut_fcast_exps = pd.DataFrame(fut_fcast_exps)

fut_fcast_exps.index = pd.date_range("2021-12-23", periods=11, f

fut_fcast_exps.columns = ['fut_fcast_exps']



Вывод

По данным анализа максимальная численность выявленных заболевших 32 244 чел. за сутки, пиковый выброс заболеваемости приходиться на июль - август 2021 года. С октября 2021 г. отмечается снижение уровня заболеваемости, на данный период можно отметить пик вакцинации населения. Опровергнута гипотеза о зависимости новых случав заболевания и общей численности вакцинированных.

Для построения прогнозной модели применены следующие методы:

- SARIMAX;
- Prophet;
- Exponential smoothing.

Оценка качества моделей прогнозирования новых случаев заболеваемости COVID - 19 произвели с помощью ключевых показателей - метрик. Для оценки качества была определены метрики:

- средняя абсолютная ошибка в процентах (MAPE)
- средняя абсолютная разница между предсказаниями и фактическими значениями (MAE);
- среднеквадратичная ошибка (MSE);
- среднеквадратичной ошибкой (RMSE).

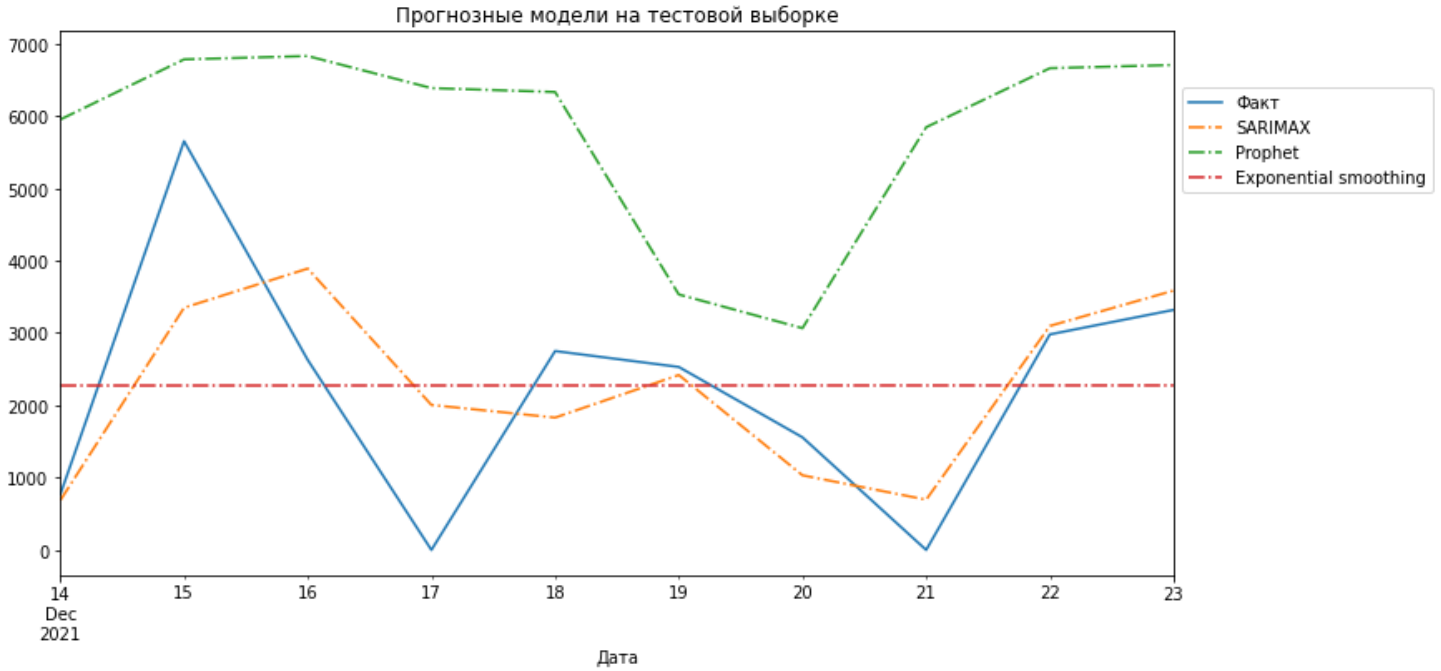
Данная система метрик позволит объективно оценить точность моделей на тестовой выборке и сравнить используемые методы.

Вывод

Анализ контрольной и тестовой выборки производился с временным интервалом обучающей выборкой 01.01.2020 - 13.12.2021 и тестовой выборкой 14.12.2021 - 23.12.2021 (10дн.). Видим что MAPE - некорректное значение inf, это связанно с тем, что фактические значения временного ряда близки к 0 и для оценки необходимо применить другие метрики.

Все три модели хорошо адаптированы к недельной сезонности. Метрики модели PROPHET оказались худшими в тесте, что подтверждается графиком. Показатели метрик SARIMAX и ES близки, но производительность модели SARIMAX выше. По графику видно что SARIMAX хоть и не улавливает пики, но примерно описывает график контрольных данных.

	Метрика	SARIMAX	PROPHET	ES
0	MAPE	inf	inf	inf
1	MAE	828.9	3590.1	1298.0
2	MSE	1262404.1	16180765.1	2652753.8
3	RMSE	1123.6	4022.5	1628.7



Вывод

Прогнозирование временных рядов проводилось на интервале 24.12.2021-02.01.2022 г. (10дн.), с обучающей выборкой 01.01.2020 - 23.12.2021.

Таким образом, на основании проведенного анализа оценки качества моделей прогнозирования наиболее точным прогнозом новых случаев заболеваемости COVID– 19 в Мексике можем считать модель построенную методом SARIMAX.