Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Компьютерных сетей и систем

Кафедра Информатики

ЛАБОРАТОРНАЯ РАБОТА №8

«Рекуррентные нейронные сети для анализа

временных рядов»

Магистрант: Проверил:

гр. 956241 Заливако С. С.

Шуба И.А.

Минск, 2020

**ХОД РАБОТЫ**

**Задание.**

**Данные:** Набор данных для прогнозирования временных рядов, который состоит из среднемесячного числа пятен на солнце, наблюдаемых с января 1749 по август 2017. Данные в виде csv-файла можно скачать на сайте Kaggle -><https://www.kaggle.com/robervalt/sunspots/>

**Задание 1.**

Загрузите данные. Изобразите ряд в виде графика. Вычислите основные характеристики временного ряда (сезонность, тренд, автокорреляцию).

**Задание 2.**

Для прогнозирования разделите временной ряд на обучающую, валидационную и контрольную выборки.

**Задание 3.**

Примените модель ARIMA для прогнозирования значений данного временного ряда.

**Задание 4.**

Повторите эксперимент по прогнозированию, реализовав рекуррентную нейронную сеть (с как минимум 2 рекуррентными слоями).

**Задание 5.**

Сравните качество прогноза моделей. Какой максимальный результат удалось получить на контрольной выборке?

**Результат выполнения:**

**Задание 1.** Загрузите данные. Изобразите ряд в виде графика. Вычислите основные характеристики временного ряда (сезонность, тренд, автокорреляцию).

Укажем путь и загрузим датасет.

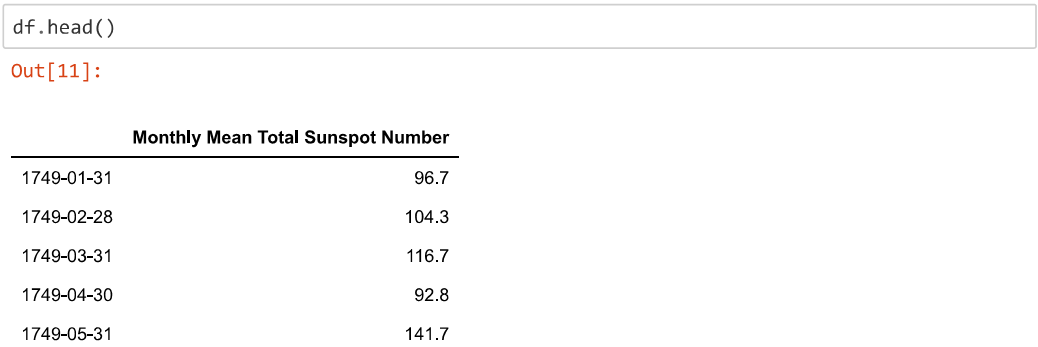
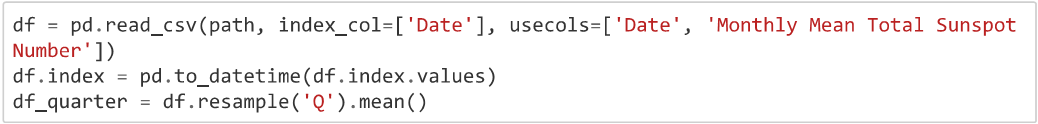
 

Рисунок 1 – Загрузка данных

Отобразим временной ряд в виде графика.

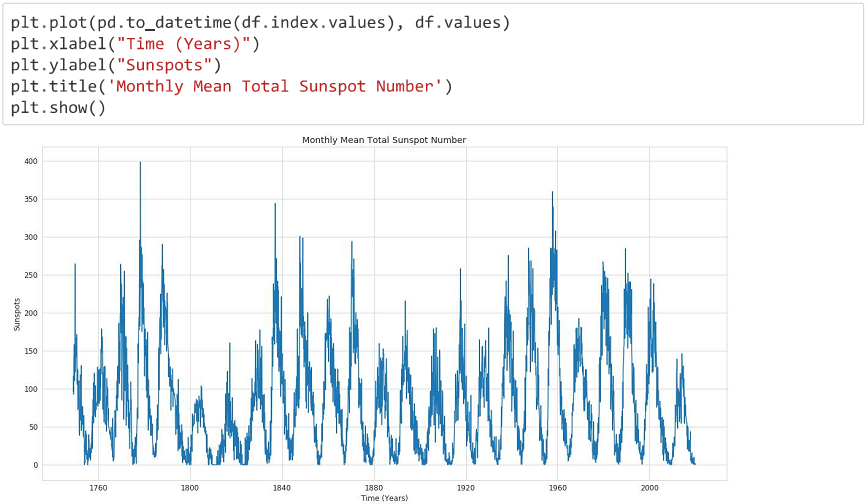


Рисунок 2 – Временной ряд в виде графика

Заметим, что явных пиков на рисунке 2 примерно 25, разделив весь период наблюдения на количество пиков можно приблизительно узнать период повторения. Таким образом 11х12 месяцев = 132 – получим период повторения пика в месяцах.



Рисунок 3 – Расчет сезонности

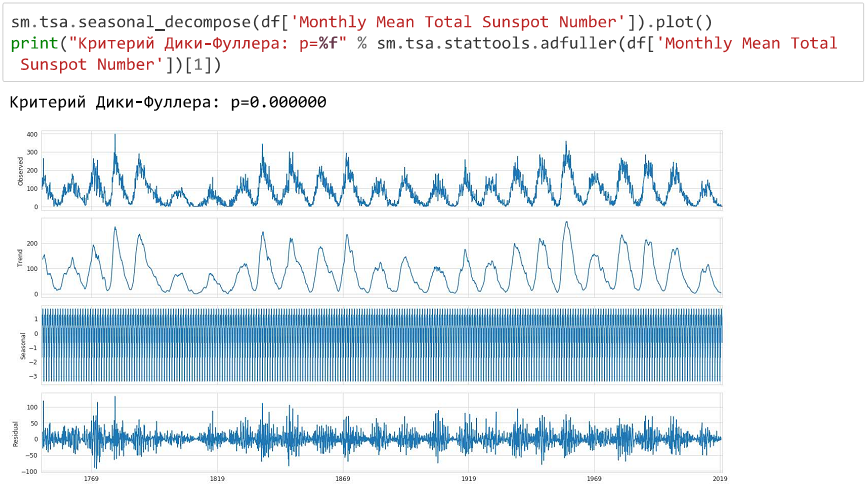


Рисунок 4 – Декомпозиция временного ряда

Построим графики автокорреляции и частичной автокорреляции. Укажем количество лагов 132, равное периоду повторения характерного пика, вычисленного выше.

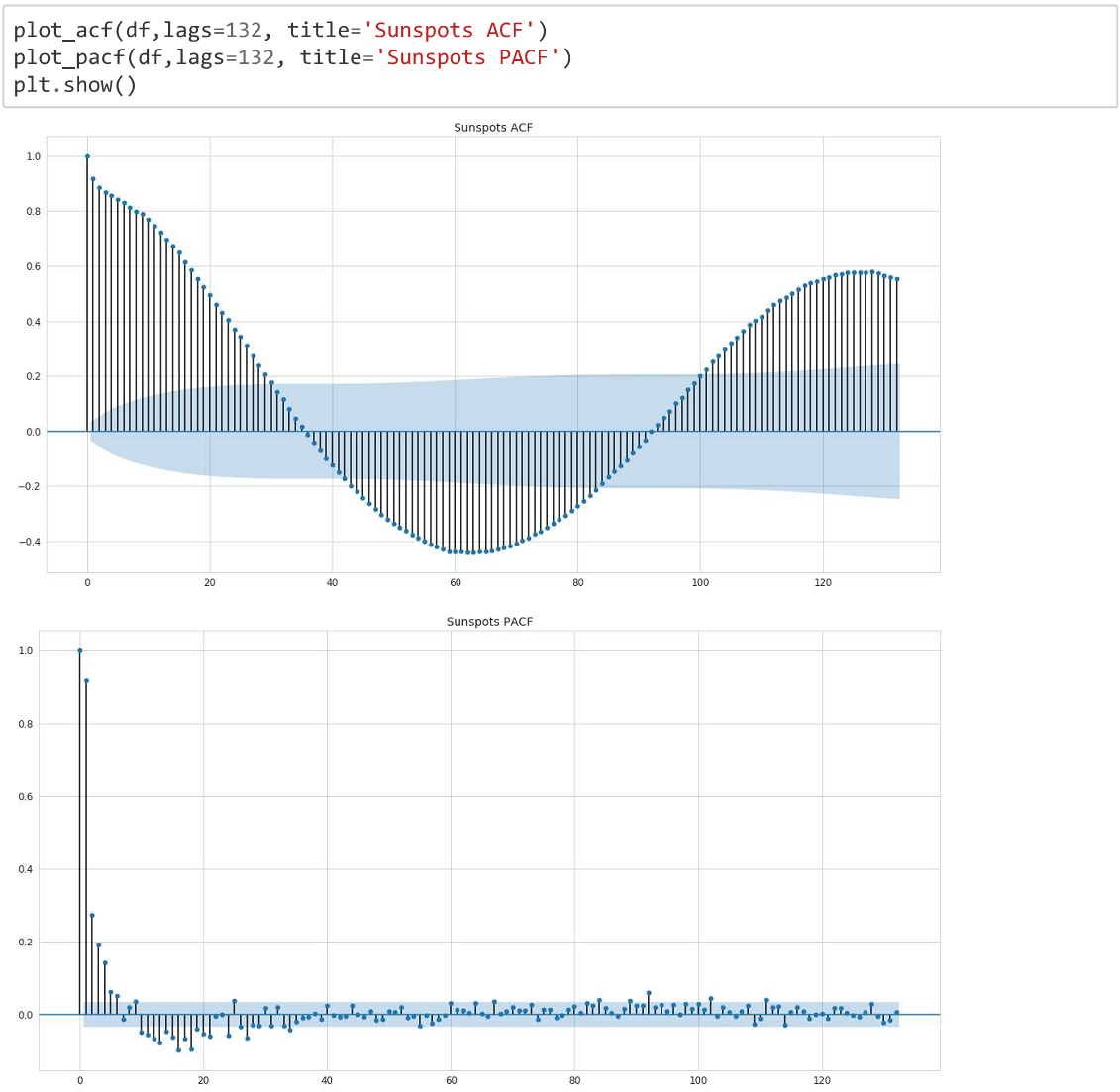


Рисунок 5 – Графики автокорреляции и частичной автокорреляции

**Задание 2.** Для прогнозирования разделите временной ряд на обучающую, валидационную и контрольную выборки.



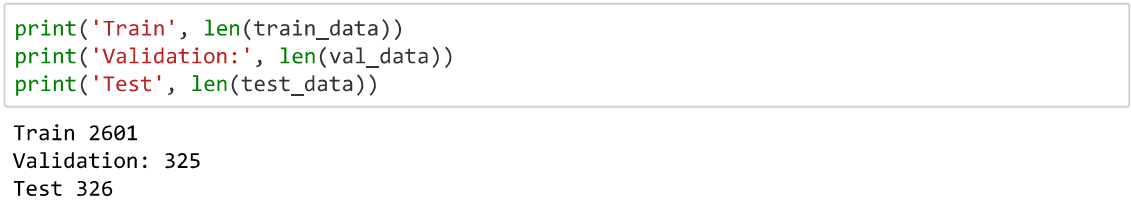
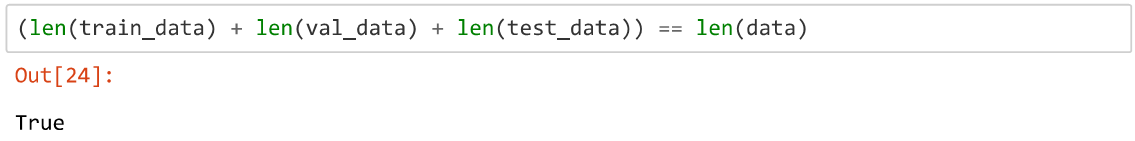


Рисунок 6 – Разделение временного ряда

**Задание 3.** Примените модель ARIMA для прогнозирования значений данного временного ряда.

Многие методы и модели основаны на предположениях о стационарности ряда, но как было замечено ранее наш ряд таковым, скорее всего, не является. Поэтому для проверки стационарности проведем [обобщенный тест Дикки-Фуллера](http://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D1%81%D1%82_%D0%94%D0%B8%D0%BA%D0%B8_%E2%80%94_%D0%A4%D1%83%D0%BB%D0%BB%D0%B5%D1%80%D0%B0) на наличие единичных корней.  Напишем функцию для определения стационарности ряда.

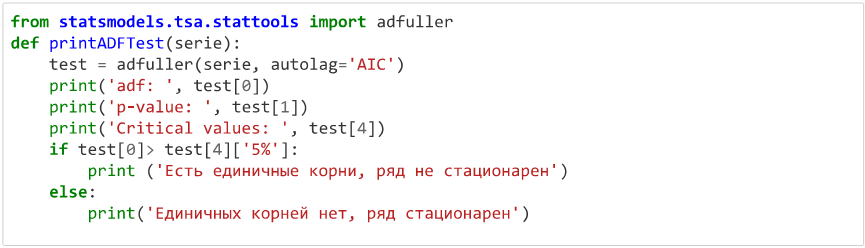


Рисунок 7 – Функция для определения стационарности ряда

Применим реализованную функцию к данному временному ряду.

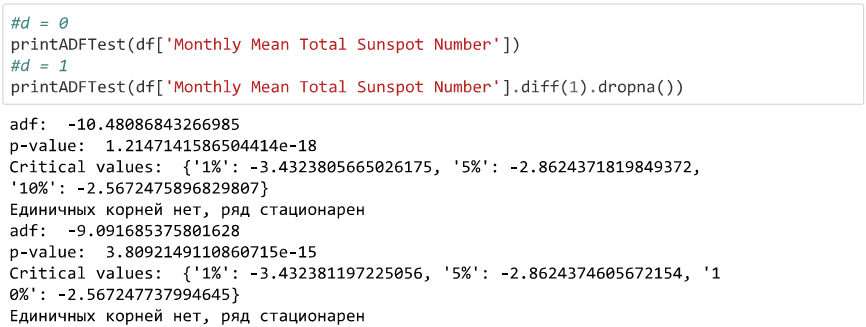


Рисунок 8 – Вывод функции по определению стационарности ряда

В обоих случаях ряд является стационарным, поэтому можно пробовать подбирать параметры для модели ARIMA.

Для прогнозирования временного ряда с очевидной сезонностью будем использовать модель SARIMA(Seasonal Autoregressive Integrated Moving Average).

Определим параметры модели.

**S = 132** – количество точек (месяцев) в периоде

**D = 1** – у ряда есть устойчивая сезонная составляющая во времени.

По графику частичной автокорреляции (PACF) определяем p – первый лаг, когда значение входит в уровень значимости, таким образом, **p = 7**.

По графику автокорреляции (ACF) определяем q – первый лаг, когда значение входит в уровень значимости, таким образом, **q = 26**. Также по этому графику определяем P и Q. Если значение ACF на 132 лаге, то **P = 1**, а **Q = 0**, при этом должно выполняться условие P+Q ≤ 2.

Построим модель SARIMA.

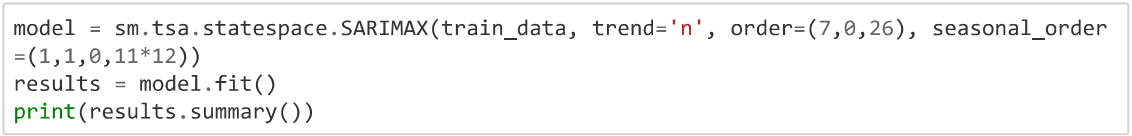


Рисунок 9 – Обучение модели SARIMA

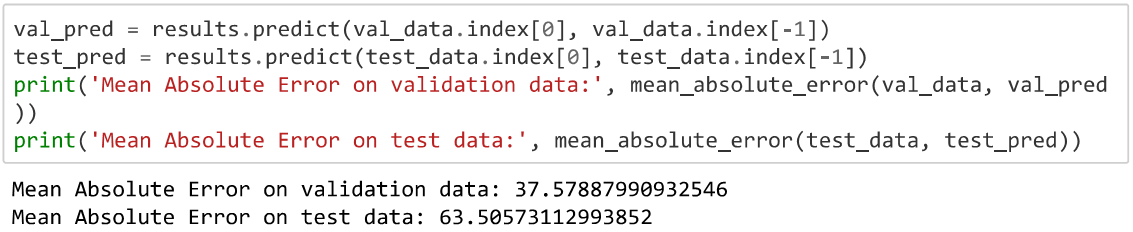


Рисунок 10 – Результаты прогнозирования

Метрика Mean Absolute Error на валидационных данных несколько выше, чем на контрольных. Построим график временного ряда для валидационных и контрольных данных.

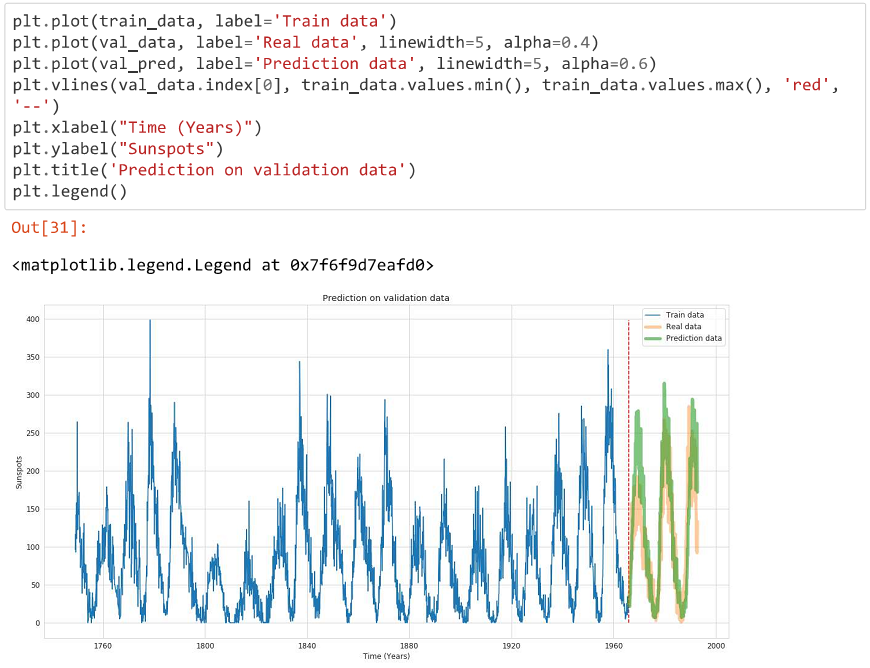


Рисунок 11 – Предсказание на валидационных данных

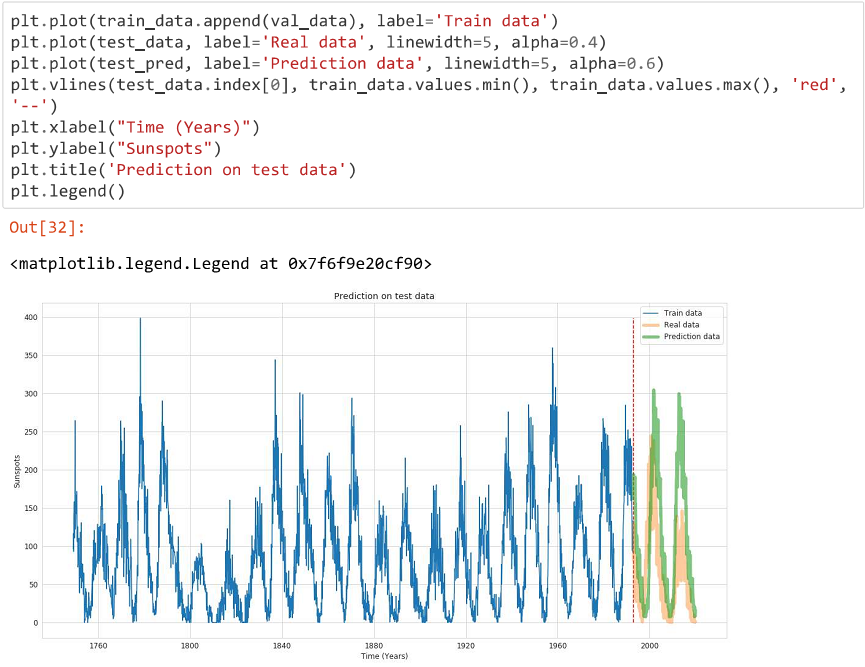


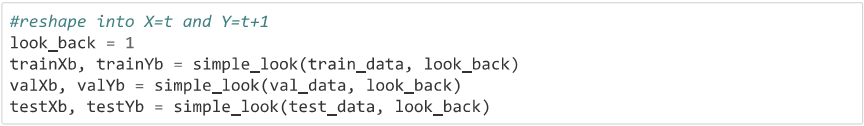
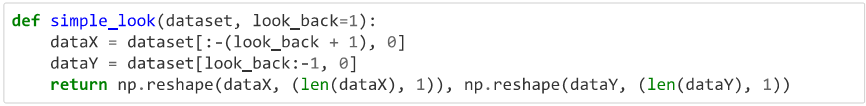
Рисунок 12 – Предсказание на контрольных данных

**Задание 4.** Повторите эксперимент по прогнозированию, реализовав рекуррентную нейронную сеть (с как минимум 2 рекуррентными слоями).

Преобразуем данные для подачи их в нейронную сеть.

Рисунок 13 – Разделение и нормализация исходного датасета



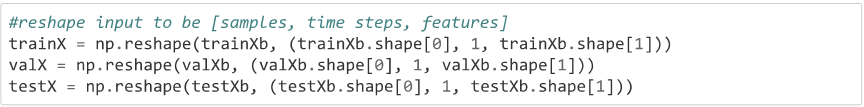


Рисунок 14 – Изменение формы данных



Рисунок 15 – Реализация рекуррентной нейронной сети

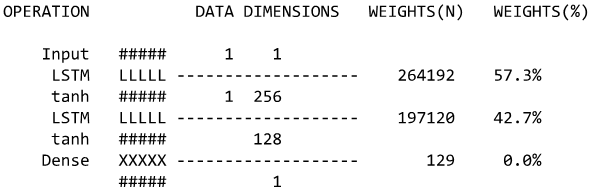


Рисунок 16 – Схема рекуррентной нейронной сети



Рисунок 17 – Обучение рекуррентной нейронной сети

Из графиков видно, что существует достаточно большой разброс в метрике MSE, что присуще рекуррентным сетям, однако еще можно заметить, что ниспадающего тренда при обучении уже, приблизительно, после 10 эпохи нет, что означает, что для получения примерно такого же результата точности модели, хватилось бы приблизительно 10 эпох.

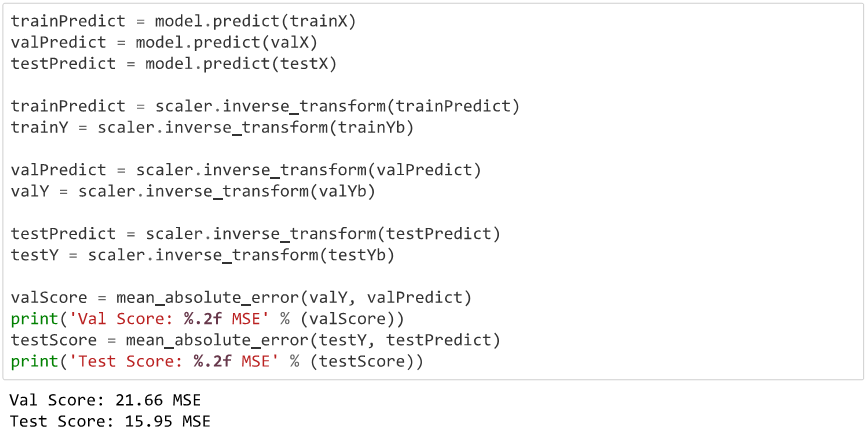


Рисунок 18 – Результаты модели на валидационной и контрольной выборке

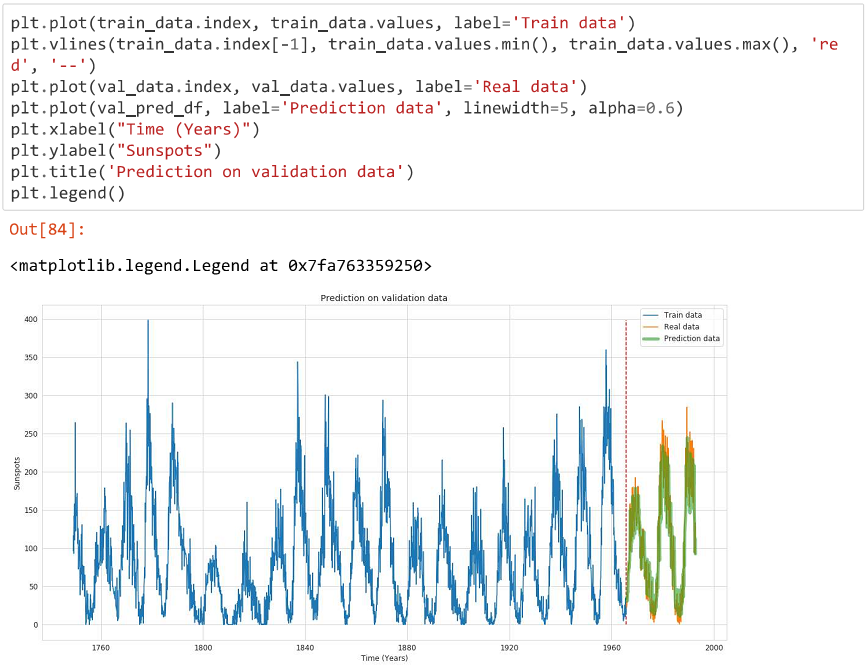


Рисунок 19 – Предсказание на валидационных данных

В целом, с помощью нейронной сети получился результат намного лучше, нежели с подбором параметров для модели SARIMA. Также можно было бы написать автоматический перебор параметров для SARIMA модели. А лучшую модель можно было бы выбрать с помощью наименьшего информационного критерия Акаике (AIC). Однако данный процесс занял бы очень много времени, так как по графику автокорреляции q составило 26.

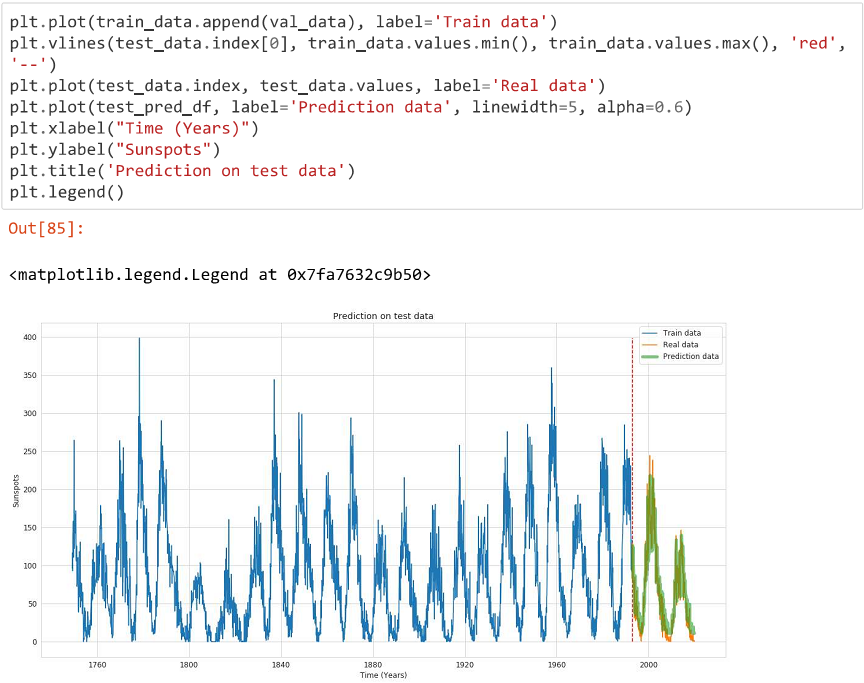


Рисунок 20 – Предсказание на контрольных данных

**Вывод:**

В ходе выполнения лабораторной работы были изучены подходы для работы с временными рядами. Были построены графики декомпозиции, автокорреляции, частичной корреляции для исходного набора данных. Были подобраны оптимальные параметры для модели SARIMA. В результате метрика MAE для контрольной выборки составила 63, что на самом деле еще является оптимальной ошибкой, так как значение было получено суммой числа пятен за 326/12 = 27 лет. Также были изучены рекуррентные нейронные сети. Была написана простейшая архитектура, состоящая из 2 рекуррентных слоев. В результате метрика MAE для контрольной выборки составила 15,95, что практически в 4 раза меньше ошибки полученной с помощью модели SARIMA.