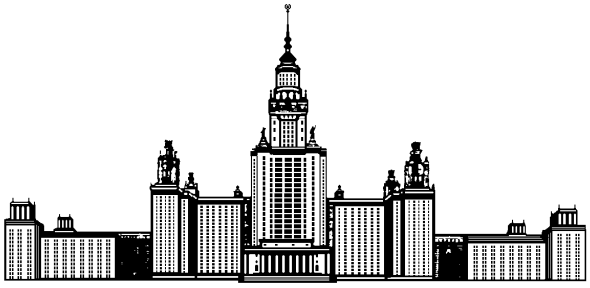


Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра информационной безопасности

Мирпулатов Исломбек Пулат-угли

Анализ и обработка временных рядов для многомасштабного агромоделирования

КУРСОВАЯ РАБОТА

Научный руководитель:

д.ф.-м.н., профессор

К. К. Абгарян

Научный консультант:

к.ф.-м.н., доцент

С. А. Матвеев

Москва, 2022

Содержание

1 Введение

С каждым годом мы все больше сталкиваемся с проблемами нехватки еды. Эти проблемы обусловлены различными факторами: увеличением популяции, загрязнением окружающей среды и нехватки воды для орошения сельскохозяйственных культур. Поэтому область агромоделирования активно изучается учеными по всему миру. Сегодня существует большое количество математических агромоделей. В рамках данной работы мы будем рассматривать многомасштабную модель WOFOST (WOrld FOod STudies). WOFOST принадлежит к семейству моделей, разработанных в Вагенингене школой C.T. De Wit (Боуман и др., 1996). В настоящее время модель WOFOST поддерживается и дорабатывается Вагенинген-ским исследовательским центром окружающей среды (WENR - Wageningen Universiteit en Researchcentrum) в сотрудничестве с группой систем производства растений (PPS - Plant Production Systems) Вагенингенского университета и подразделением по пищевой безопасности объединенного исследовательского центра в Италии.

WOFOST - это имитационная модель для количественного анализа роста и производства однолетних полевых культур. С помощью WOFOST, имея данные о почве, параметрах культуры, погоде и управлении посевом, для местности можно рассчитать достижимую урожайность, общую биомассу, необходимое количество воды.

Чаще всего в агромоделях на вход поступают почвенные профили, климатические и другие данные связанные с севооборотом. Для годовой оценки урожайности в качестве климатических данных подаются различные сценарии погодных условий. Мы же будем прогнозировать погоду, как временной ряд и сравним полученные результаты с результатами работы модели на реальных данных погоды.

Для оценки эффективности метода были выбраны сельскохозяйственные поля в селе Кшень, Курская область, Россия. Для опыта мы выбрали сахарную свеклу (Sugar-beet, *Beta vulgaris*) и озимую пшеницу (Winter wheat, *Triticum aestivum*). Опыты проводились на примере 2015 и 2017 годов.

Модель WOFOST используется более 25 лет и имеет различные реализации на Fortran, Python и R. Мы использовали реализацию модели PCSE/Wofost71_WLP_FD на Python.

2 Анализ и прогнозирование временных рядов

Анализ временных рядов интересный инструмент, имеющий множество приложений от предсказания цен на акции, прогнозов погоды, планирования бизнеса до распределения ресурсов. В рамках нашей работы будем рассматривать задачу прогнозирования погоды, как временного ряда.

Введем некоторые понятия, которые нам потребуются в дальнейшем.

Временной ряд: $y_1, \dots, y_T, \dots, y_t \in R$ - значения признака, измеренные через постоянные временные интервалы.

Любой временной ряд можно разложить на компоненты (STL - decomposition):

- Тренд - плавное долгосрочное изменение уровня ряда.
- Сезонность - циклические изменения уровня ряда с постоянным периодом.
- Цикл - изменения уровня ряда с переменным периодом.
- Ошибка - непрогнозируемая случайная компонента ряда.

2.1 Данные

В качестве датасета будем использовать климатические данные полученные с сервиса NASA POWER.

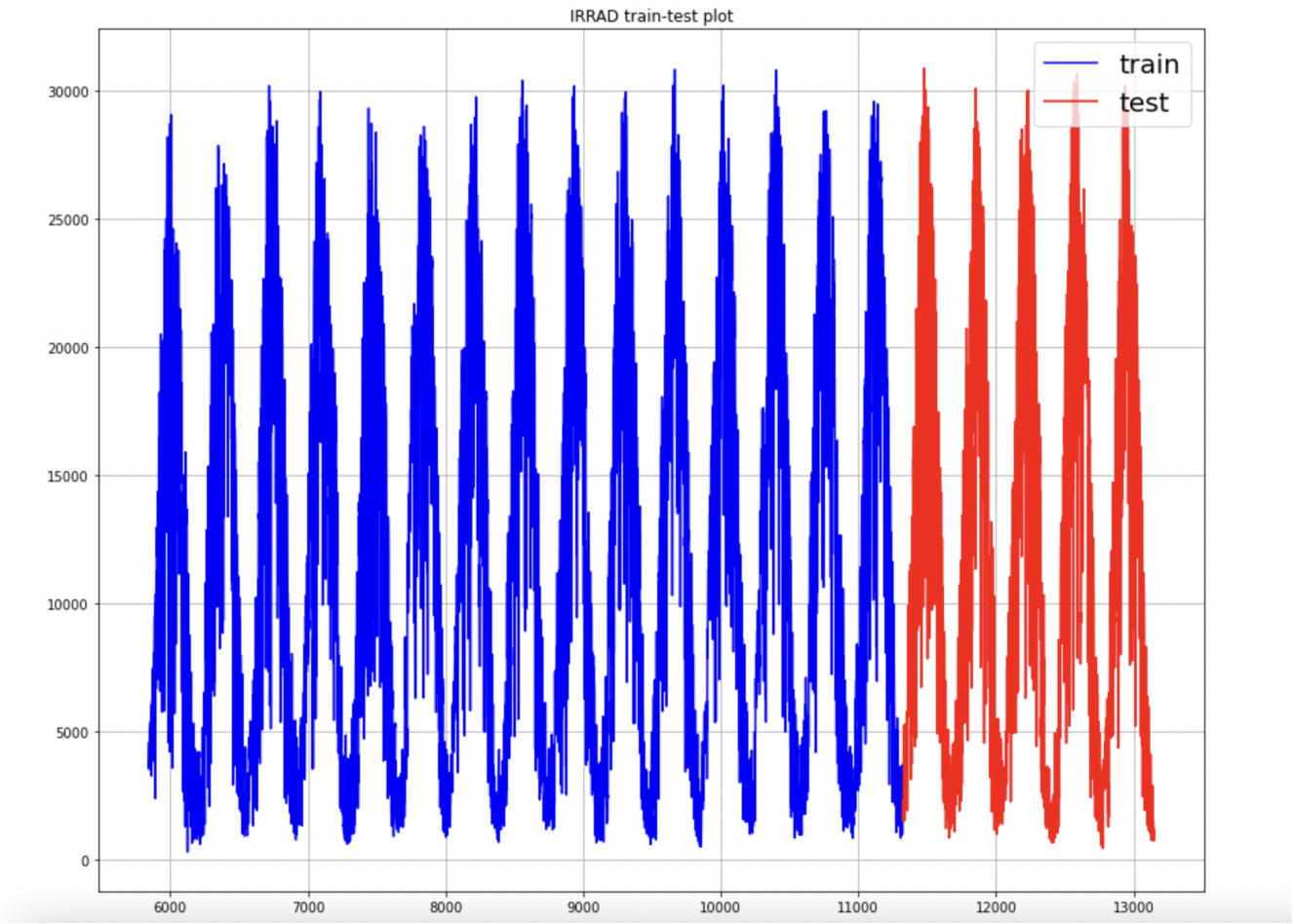
Данные получены в 2000-2020 годах для села Кшень, Курская область (широта = 51.43, долгота = 37.25) в формате CSV.

Описание данных:

Название столбца	Описание	Единица измерения
IRRAD	Солнечное излучение за сутки	$kJ.m^{-2}.day^{-1}$
TMIN	Минимальная температура за сутки	°C
TMAX	Максимальная температура за сутки	°C
VAP	Среднесуточное давление пара	hPa
WIND	Среднесуточная скорость ветра на высоте 2 м над уровнем земли	$m.sec^{-1}$
RAIN	Осадки (дождь или водный эквивалент в случае снега или града)	$cm.day^{-1}$

Каждый столбец можно рассматривать, как временной ряд. Разделим наш ряд на train (2000-2015) и test (2015-2020).

Пример деления для столбца IRRAD:



2.2 Техники прогнозирования временных рядов

Для оценки качества прогнозирования временного ряда мы будем использовать коэффициент детерминации R^2. Суть его работы заключается в измерении количества отклонений в прогнозах, объясненных набором данных, т.е. это разница между выборками в наборе данных и прогнозами, сделанными моделью.

$$R^2(y_{true}, y_{pred}) = 1 - \frac{\sum_{i=1}^{l}(y_{true}^i - y_{pred}^i)^2}{\sum_{i=1}^{l}(y_{true}^i - \overline{y_{pred}})^2}$$

Далее описаны некоторые подходы к прогнозированию временных рядов с погодными данными.

2.2.1 Группировка по среднему значению

Самый простой подход к прогнозированию временного ряда - это периодическое продление, но так как этот подход не давал нужных результатов, было решено использовать другой

подход в качестве baseline'а для оценки последующих методов.

Данные в обучающей выборке были сгруппированы по датам и усреднены по годам. На выходе мы получили кривую за один год и ее зациклили. Полученные результаты мы проверили на тестовой выборке и получили следующие результаты:

Data	R^2	Data	R^2
IRRAD	0.7609	VAP	0.7784
TMIN	0.8061	WIND	0.0985
TMAX	0.8712	RAIN	-0.0569

2.2.2 ARIMA и SARIMA

Перед тем как приступить к изучению данного класса моделей для них должно соблюдаться необходимое условние - **стационарность**.

Понятие стационарного временного ряда означает, что его среднее значение не изменяется во времени, т.е. временной ряд не имеет тренда. Кроме того, ковариация между разными элементами временного ряда (как между случайными величинами) зависит только от того, насколько сильно они отдалены друг от друга во времени.

Существует множество техник для достижения стационарности ряда.

Дифференцирование ряда — переход к попарным разностям его соседних значений:

$$y_1, ..., y_T \to y'_2, ..., y'_T$$

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

Дифференцированием можно стабилизировать среднее значение ряда и избавиться от тренда и сезонности. Также может применяться неоднократное дифференцирование.

Сезонное дифференцирование ряда — переход к попарным разностям его значений в соседних сезонах:

$$y_1, ..., y_T \to y'_{s+1}, ..., y'_T$$

$$y'_t = y_t - y_{t-s}$$

Авторегрессия AR

$$AR(p): \quad y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

где y_t - стационарный ряд с нулевым средним, $\phi_1, \ldots, \phi_p$– константы $(\phi_p \neq 0), \varepsilon_t$– гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2. Если среднее равно μ, модель принимает вид

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

где $\alpha = \mu\left(1 - \phi_1 - \cdots - \phi_p\right)$. Другой способ записи:

$$\phi(B)y_t = \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) y_t = \varepsilon_t$$

где B - разностный оператор ($By_t = y_{t-1}$). Линейная комбинация p подряд идущих членов ряда даёт белый шум.

Чтобы ряд AR(p) был стационарным, должны выполняться ограничения на коэффициенты. Например,

в AR(1) необходимо $-1 < \phi_1 < 1$;

в AR(2) необходимо $-1 < \phi_2 < 1, \phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1$.

С ростом p вид ограничений усложняется.

Скользящее среднее MA

$$MA(q): \quad y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q},$$

где y_t - стационарный ряд с нулевым средним, $\theta_1, \ldots, \theta_q-$ константы $\left(\theta_q \neq 0\right), \varepsilon_t-$ гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2. Если среднее равно μ, модель принимает вид

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}.$$

Другой способ записи:

$$y_t = \theta(B)\varepsilon_t = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right) \varepsilon_t$$

где B - разностный оператор. Линейная комбинация q подряд идущих компонент белого шума ε_t даёт элемент ряда.

Чтобы модель MA(q) была обратимой, должны выполняться ограничения на коэффициенты. Например,

в MA(1) необходимо $-1 < \theta_1 < 1$;

в MA(2) необходимо $-1 < \theta_2 < 1, \theta_1 + \theta_2 > -1, \theta_1 - \theta_2 < 1$.

С ростом q вид ограничений усложняется.

ARMA (autoregressive moving average)

$ARMA(p,q)$: $y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$ где y_t - стационарный ряд с нулевым средним, $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q-$ константы $\left(\phi_p \neq 0, \theta_q \neq 0\right), \varepsilon_t-$ гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2.

ARIMA (autoregressive integrated moving average)

$ARIMA(p,d,q) - ARMA(p,q)$ для временного ряда, который был продифференцирован d раз.

Seasonal multiplicative ARMA/ARIMA

Предположим, что ряд имеет сезонность порядка S. Возьмём модель ARMA (p,q):

$$y_t = \alpha + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Добавим к ней P последних сезонных авторегрессионных компонент:

$$+\varphi_S y_{t-s} + \varphi_{2S} y_{t-2S} + \cdots + \varphi_{PS} y_{t-PS}$$

Добавим также Q последних сезонных компонент скользящего среднего:

$$+\theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \cdots + \theta_Q S_{\varepsilon t-QS}$$

Мы получили модель SARMA$(p,q) \times (P,Q)$. SARIMA $(p,d,q) \times (P,D,Q)$ есть SARMA $(p,q) \times (P,Q)$ для временного ряда, которыи был d раз продифференцирован и D раз сезонно продифференцирован.

В качестве модели для наших данных использовалась SARIMA из библиотеки алгоритмов машинного обучения Scikit-learn, но даже после подбора параметров по сетке (GRID Search), не удалось получить результаты лучше baseline'a. Впоследствии планируется подробнее изучить данный класс моделей и получить качество лучше baseline'a.

2.2.3 Prophet

Prophet — это программное обеспечение с открытым исходным кодом, выпущенное командой Facebook Core Data Science. Реализовано в виде библиотеки для Python и R.

Prophet предлагает нам процедуру прогнозирования временных рядов, при наличии нескольких сезонов исторических данных. Библиотека позволяет взаимодействовать даже с временными рядами с несколькими сезонностями.

Prophet считает, что временной ряд может быть разложен следующим образом:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

$g(t) -$ тренд,

$s(t) -$ сезонность,

$h(t) -$ каникулы, т.е аномальные данные,

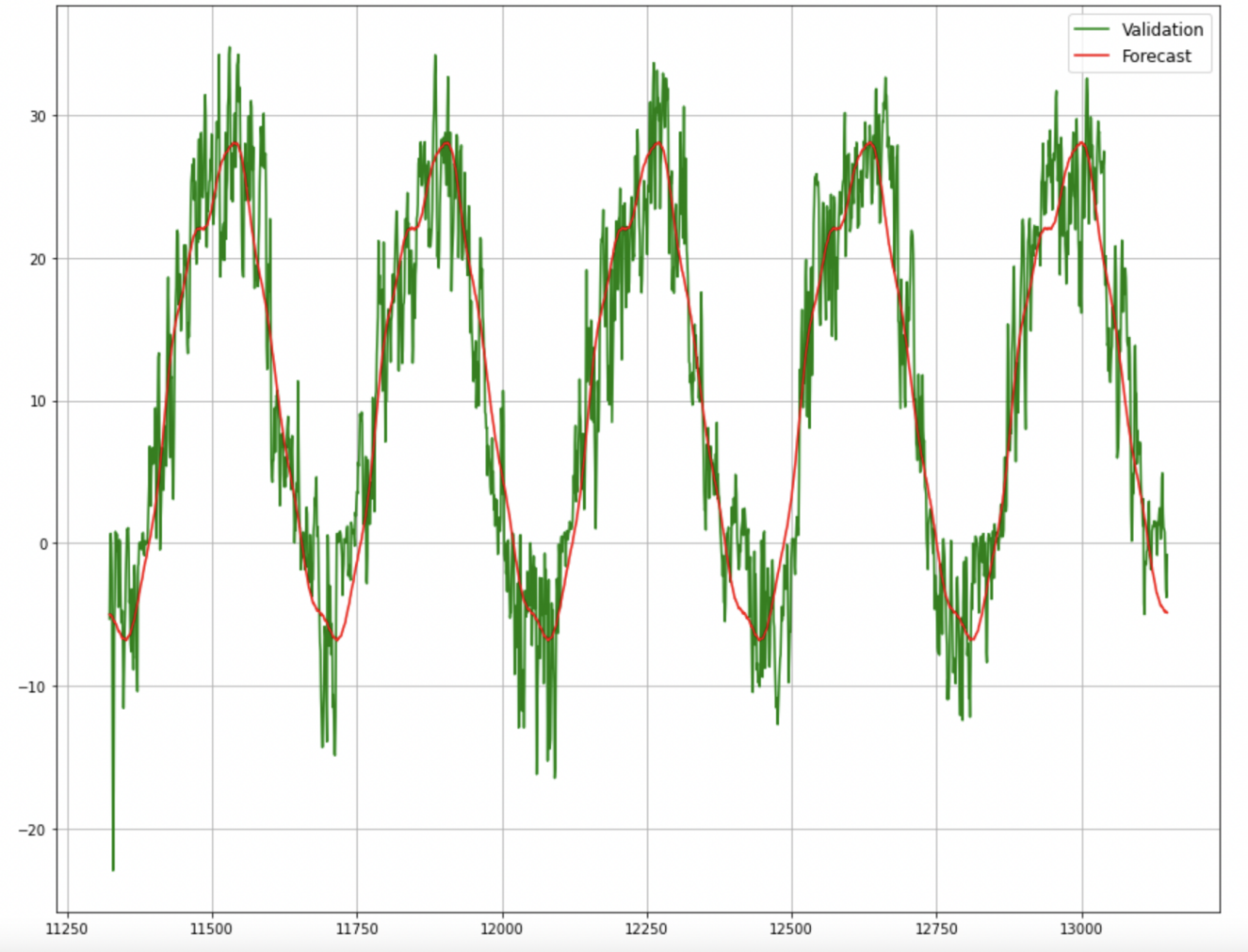
$\varepsilon(t) -$ ошибки.

Подгонка модели представляет собой упражнение по подгонке кривой, поэтому она явно не учитывает структуру временной зависимости в данных. Это также позволяет проводить наблюдения с нерегулярным интервалом.

На наших данных даже без подбора параметров модель Prophet на тестовых данных дает улучшения остностельно нашего baseline'a.

Data	R^2	Data	R^2
IRRAD	0.7695	VAP	0.7830
TMIN	0.8125	WIND	0.1671
TMAX	0.8736	RAIN	-0.0077

Пример прогноза максимальной температуры:



По графику видно, что модель в среднем неплохо прогнозирует, но есть отклонения в экстримальных значениях.

Также можно сказать, что даже после улучшения модель плохо предсказывает ветер и дождь (R^2 в районе нуля).

Полученными моделями генерируем три датасета с климатическими данными:

1. Действительные данные о погоде с 2015-2020 года
2. Прогноз погоды с 2015-2020 года
3. Частичный прогноз данных (значения по ветру и дождю взяты из действительных данных)

3 Многомасштабная модель WOFOST

WOFTOST это модель, которая объясняет ежедневный рост урожая на основе таких процессов, как фотосинтез, дыхание и другие, в условиях влияния окружающей среды на эти процессы. Рост урожая рассчитывается с временными шагами в один день на основе знаний о процессах на более низком уровне интеграции (кривая реакции фотосинтеза на свет одного листа). Затем процессы низкого уровня объединяются с другими процессами (фенология, дыхание) для объяснения поведения системы на более высоком уровне интеграции.

Так как это математическая модель, т.е. представляет собой упрощенную реальность, некоторые процессы, которые недостаточно изучены, в ней не представлены, либо заданы статически. Мы будем рассматривать модель в условиях ограниченности воды.

Задаем входные параметры почвы и севооборота полученные из исследуемой местности, а в качестве климатических данных будем использовать три датасета:

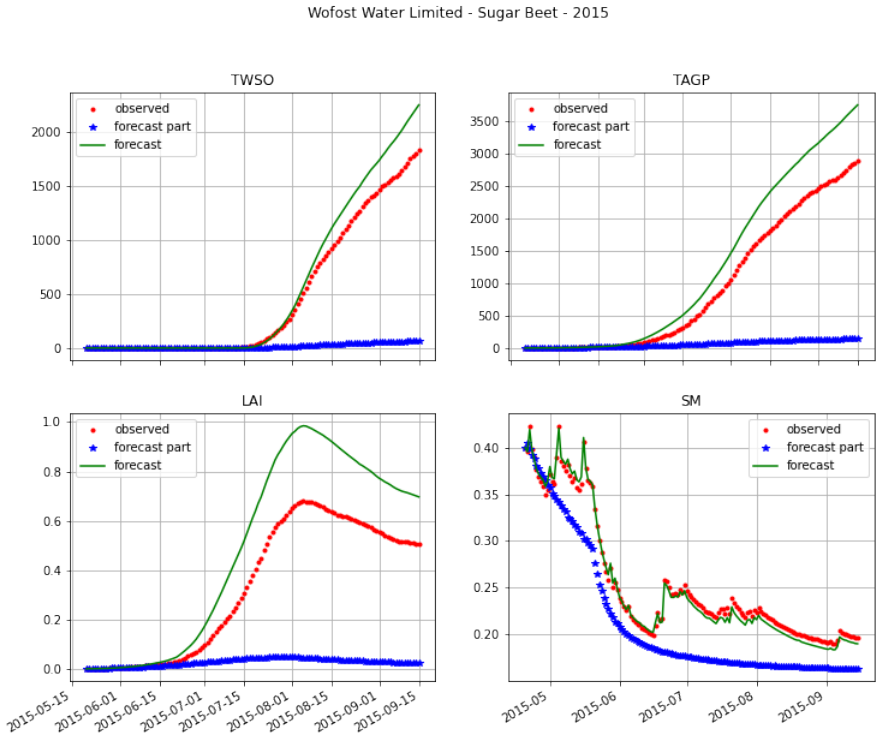
- Действительные данные о погоде с 2015-2020 года
- Прогноз погоды с 2015-2020 года
- Частичный прогноз данных (значения по ветру и дождю взяты из действительных данных)

Получим данные и сравним результаты. На выходе будем оценивать четыре основных параметра:

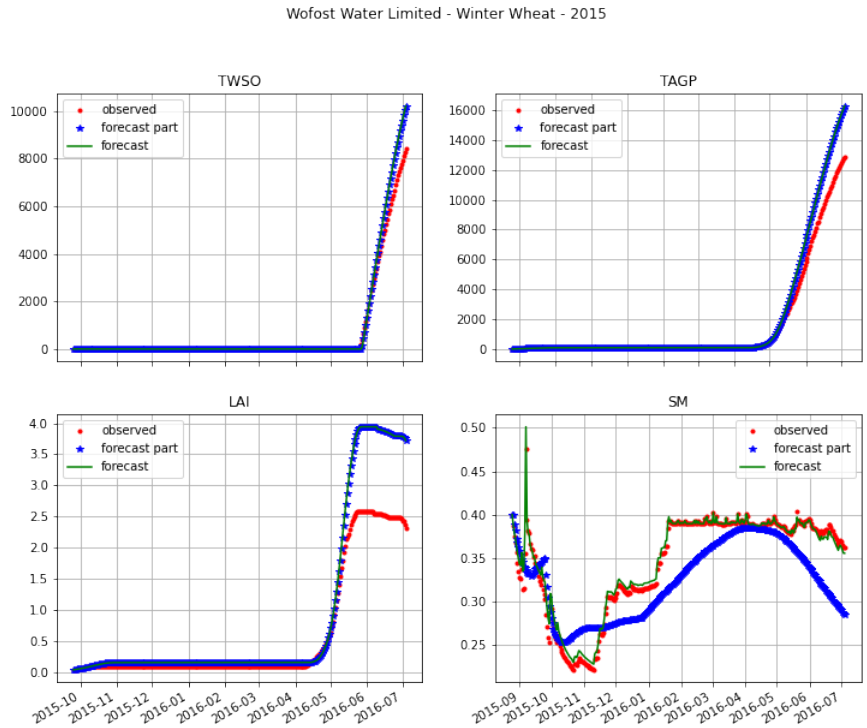
1. **TWSO**: представляет собой продукт сбора урожая (урожайность) в виде сухого веса в кг/га. Для зерновых это зерна, для картофеля — клубни, для сои — бобы и т.д.
2. **TAGP**: представляет собой общую надземную биомассу, произведенную культурой, в виде сухого веса в кг/га. Последнее означает, что будет разница между собранным урожаем и смоделированным урожаем из-за разницы между сухим весом и сырым весом. Величина этой разницы зависит главным образом от содержания воды в конечном урожае. Например, зерновые содержат только 10-15 процентов воды, в то время как картофель и сахарная свекла содержат много воды, и поэтому существует большая разница между сырым и сухим весом, оцененным WOFOST.
3. **LAI**: это безразмерная переменная, определяющая площадь одной стороны живых (зеленых) листьев, приходящуюся на площадь поверхности земли.
4. **SM**: представляет влажность почвы корневой зоны в виде объемной доли. Полностью сухая почва имеет нулевое значение, в то время как максимальное значение относится к полностью насыщенной почве (зависит от почвы, но обычно около 0.4 для минеральных почв).

Сводные графики за 2015 год по трем типам климатических данных:

Сахарная свекла:

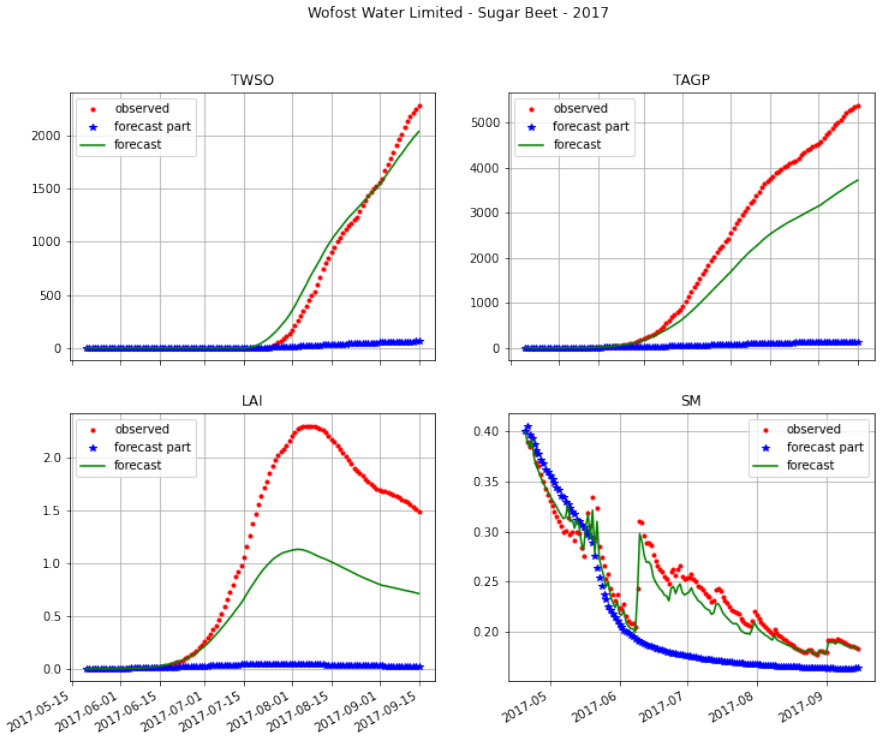


Озимая пшеница:



Выход за 2017 год по трем типам климатических данных:

Сахарная свекла:



Озимая пшеница:

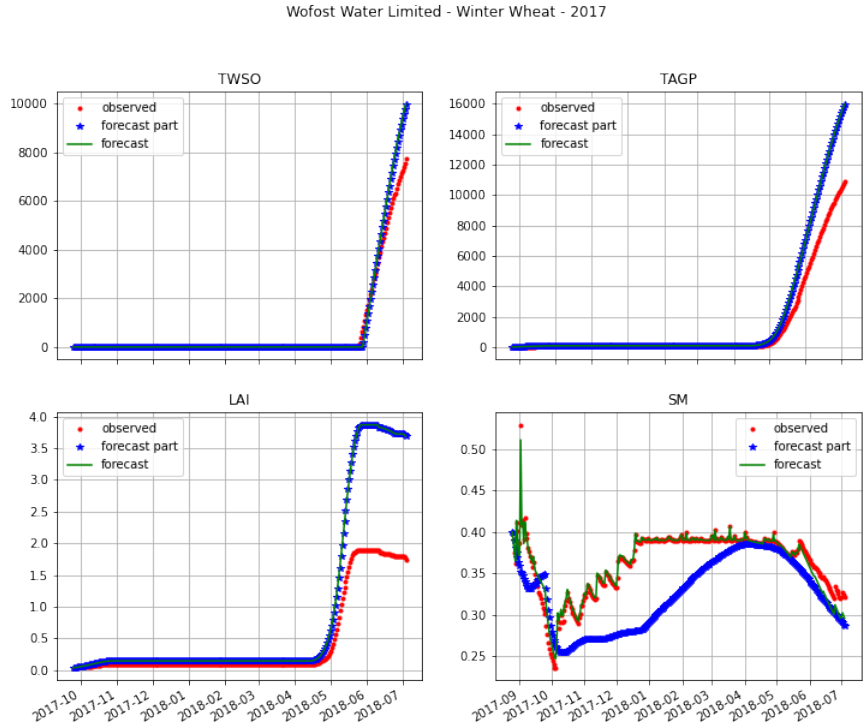


Таблица урожайности кг/га:

	Real	Partitial	Predicted
Сахарная свекла / 2015	1834.94	63.05	2253.0
Озимая пшеница / 2015	8404.01	10212.43	10212.43
Сахарная свекла / 2017	2290.39	62.05	2039.8
Озимая пшеница / 2017	7724.74	9990.28	9990.28

4 Заключение

В рамках данной работы мы изучили различные подходы к прогнозированию временных рядов, а также научились запускать многомасштабную модель WOFOST с интерфейсом на Python.

Исходя из полученных результатов можно сделать несколько выводов. В среднем предсказание урожайности модели на настоящих климатических данных и предсказанных разнится в 20%, причем в положительную сторону относительно предсказанных климатических данных - это связанно с тем, что предсказанный временной ряд глаже и содержит меньше экстремальных значений, следовательно сельскохозяйственная культура подвергается меньшему стрессу и мы получаем результаты лучше.

Также можно заметить, что показатели урожайности для сахарной свеклы на комбинированном климатическом датасете сильно отличаются от реальных. Пока не получилось найти закономерности в следствии которых это может наблюдаться.

Существует целый спектр возможностей для развития данной работы: улучшение параметров текущих моделей прогнозирования временных рядов, использование рекурентных нейронных сетей для прогнозирования временых рядов, интервальные методы прогнозирования временных рядов. Также можно прогнозировать временные ряды, как зависимые друг от друга.

Список литературы

[1] *Project Path* GitHub repository:
`https://github.com/mirpulatov/CS_MSU/tree/main/Course%20Work`

[2] *Н. В. Артамонов, Е. А. Ивин, А. Н. Курбацкий, Д. Фантаццини* Введение в анализ временных рядов. ВолНЦ РАН

[3] *Рябенко Евгений* Анализ временных рядов. Прикладной статистический анализ данных

[4] *Neerc ITMO* Анализ временных рядов.

[5] *MachineLearningMastery* How to Grid Search SARIMA Hyperparameters for Time Series Forecasting

[6] *MachineLearningMastery* Time Series Forecasting With Prophet in Python

[7] *Toward Data Science* Implementing Facebook Prophet efficiently

[8] *Allard de Wit Hendrik Boogaard* A gentle introduction to WOFOST. Wageningen Environmental Research. May 2021

[9] *Wit, Allard de, Hendrik Boogaard, Davide Fumagalli, Sander Janssen, Rob Knapen, Daniel van Kraalingen, Iwan Supit, Raymond van der Wijngaart, and Kees van Diepen.* 25 Years of the WOFOST Cropping Systems Model. Agricultural Systems 168 (January 1, 2019): 154–67.

[10] *Sparks, A.H* a NASA POWER global meteorology, surface solar energy and climatology data client for R. J. Open Source Softw.

[11] *Mikhail Gasanov, Daniil Merkulov, Artyom Nikitin, Sergey Matveev, Nikita Stasenko, Anna Petrovskaia, Mariia Pukalchik and Ivan Oseledets*
Sensitivity Analysis of Soil Parameters in Crop Model Supported with High-Throughput Computing.

[12] *Mikhail Gasanov, Daniil Merkulov, Artyom Nikitin, Sergey Matveev, Nikita Stasenko, Anna Petrovskaia, Mariia Pukalchik and Ivan Oseledets*
A New Multi-objective Approach to Optimize Irrigation Using a Crop Simulation Model and Weather History.