University of Novi Sad
Faculty of Sciences
Department of Mathematics and Informatics

Student
Dragana Šorak

Optimization of Crop Planting Schedule Based on the Different Growing Degree Units Forecasting Models

Master Thesis

Mentor
Dr Nikola Obrenović

Novi Sad, October 2022

I would like to express my sincere gratitude to my mentor Dr Nikola Obrenović for his advices, help, support, and suggestions during the research and writing of this thesis.

I could not have undertaken this journey without the support of my family, partner, and friends.

Their belief in me has kept my spirits and motivation high during this process, and for that I am forever thankful.

Contents

List of Figures and Tables

1. Introduction

The world's population is growing drastically, and an increase of about 30 percent in the next 30 years is predicted [1]. Population growth requires greater and faster production of food, which must be stored in the right way.

Figure 1: Corn crop

With 1.1 billion tons produced, corn represents the most produced crop in the world. It is followed by wheat (760.9 million tons) and rice (756.7 million tons) [2]. The latest technologies have succeeded in creating new corn hybrids. By doing so, they significantly increased the yield. [3]. On the other hand, a higher yield means a larger surplus later. Poor harvesting and storage organization results in an average of 40% of food being thrown away annually [1]. Therefore a planning tools to overcome these challenges must be created [4].

Figure 2: Corn harvesting and storing

The Crop Planting Schedule Problem (CPSP) [4, 5] algorithm, created for solving the optimization problem, make sure that an ideal planting schedule has almost constant weekly harvested quantities within storage capacity and the planting dates are within the preferred planting timeframe.

Production is not an easy task; it consists of many different stages. It is important to determine each of them as precisely as possible in order to improve the yield. Each of these stages can be precisely determined by calculating the GDU. The GDU actually measures how much heat the plant has accumulated during the day, and therefore, daily temperatures are used to calculate this unit. With the help of GDU, we can determine every stage of the plant, from germination to maturation[1]. If a set of seeds is planted on the same day, in the same field, and later harvested together, we call it a population. Each population has its own amount of GDU which needs to be accumulated. So, when the plant accumulates enough GDU, we can say that it is ready for harvesting. All of this implies that an accurate prediction of GDU is crucial for solving this problem.

GDUs prediction presents the input for the optimization problem, therefore we wanted to investigate how sensitive the model is to different GDUs predictions. For GDUs prediction in this thesis, two different approaches and three different models are used. In the first approach, a time series of daily accumulated GDUs from the year 2009 to the year 2019, is used for the prediction of GDUs for the next two years. Accumulated GDUs are given for two different planning sites, which are denoted as site 0 and site 1.

Since GDUs highly depend on daily temperatures, in the second approach historical data of daily minimum and maximum temperatures for the years 2009-2019 from two different sites, are used. The given time series helped in predicting daily minimum and maximum temperatures for the next two years, and then using a formula for GDUs calculation, daily accumulated GDUs were calculated.

Predictions, in both approaches, are made using the following models for time series prediction: *Moving Average* (MA), *Auto Regressive Integrated Moving Average (ARIMA)* and *Holt-Winters* model. The GDUs predictions are input for the optimization model, which can also be divided into two different scenarios. In the first scenario, both sites have pre-defined capacity values for each week. On the other hand, in the second scenario, pre-defined capacity does not exist. Therefore, the goal is to find which capacity will minimize the harvesting period and maximize the total harvest quantity.

The task of this thesis is to investigate how different GDUs predictions impact the optimized planting schedule, expressed through the number of harvest weeks and quantities.

This thesis is organized into 6 chapters, where the first and sixth chapters provide an introduction and a conclusion. Chapter 2 makes a review of a topic, going through the literature that solves the same or similar problems. Methodologies, models, and algorithms used in this thesis are described in Chapter 3 and Chapter 4, respectively. The planting schedule optimization is applied to each combination of site and scenario, and both approaches, independently. Each combination of results is presented and discussed in detail in Chapter 5.

2. Related work

In the literature, several works tackling similar or related problems have been found. Different machine learning and deep learning algorithms can be successfully trained to predict GDUs, some of them are more successful than others. In the GDUs prediction stage, the selection of the prediction model is essential.

TBATS is an acronym for key features of the model: T: *Trigonometric seasonality* B: *Box-Cox transformation* A: *ARIMA errors* T: *Trend* S: *Season*. In [1] TBATS and 1D-Convolutional Neural Network were compared. 1D-CNN outperforms the TBATS, by reducing the number of weeks required for harvest. They give an advantage to 1D-CNN since it can deal with variance in data, and TBATS tends to smooth the predictions. Additionally, TBATS does not perform well for long-term predictions.

In [6] a Gaussian process mode and a deep recurrent neural network are compared in the same manner. The authors demonstrated the model's capacity to capture the overall trend of historical GDUs as well as fluctuations in daily temperatures. Therefore, models managed to reduce the deviation from capacity, in some cases by 70%. Additionally, they successfully created the periods with almost constant harvest quantities.

Besides Machine learning and Deep Learning models, there are some other approaches for the estimation of future GDUs. The following approach is one of them. Instead of building a prediction model, daily temperatures are replaced by certain statistical values. Minimum temperatures are modeled by the first quartile, median, or third quartile, while maximum temperatures are modeled by the 80th or 90th percentile and maximum. The optimization model showed great dependence on the selection of the specified values [7].

The literature once more offers a variety of problem-solving techniques for CPSP. For instance, in [7] the authors attempted to discover a mixed-integer linear programming solution to the problem. However, in [8], authors used evolutionary algorithms to find a solution. A type of meta-heuristic optimization algorithm called evolutionary algorithms draws its inspiration from biological mechanisms. As in biology, there are parents who give birth to children, and the "survival rate" depends on the environment in which they live.

The ALNS meta-heuristic described in [4] is the foundation of this thesis. The algorithm is successful in solving the Crop Planting Scheduling Problem. However, the optimization algorithm's sensitivity to the GDUs forecast was discovered. The inspiration for this thesis lies in that information.

3. Methodologies

In this Chapter, the models for time series prediction together with the *Adaptive Large Neighborhood Search* (ALNS) meta-heuristic [4], will be presented. For the evaluation of the model's performance, Akaike's *Information Criterion* (AIC) is used.

3.1. Time series models

For the GDU time series forecasting three different models are used: *Moving Average* (MA), *Auto Regressive Integrated Moving Average* (ARIMA), and *Holt-Winters* model for exponential smoothing. Each model is defined and explained in the following subsections.

3.1.1. Moving Average model

The Moving-Average model (MA), is a method for modeling and forecasting univariate time series in time series analysis. According to the Moving-Average model, the relationship between the current value and the present and previous error terms is linear. A moving average model is denoted as ***MA(q)*** where ***q*** is the order. Using a coefficient denoted with theta (θ), the impact of historical errors on the present value is evaluated. For a moving average model, the prediction formula is:

$$\overline{X_t} = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (1)$$

where ε_t is white noise.

The number of past error terms that have an impact on the current value is defined by the Moving Average model's order ***q***. To fit the suitable model, it is crucial to establish its order. The parameter ***q*** can be determined if we follow the next steps. Firstly, test stationarity. Each time series can be decomposed into a level, trend, and seasonal component. The average value for a given time period represents the *level component* of the series. The movement of a series to relatively higher or lower values over an extended period of time is represented by a pattern in data known as a *trend component*. In other words, a trend can be seen when the time series has a rising or decreasing slope. The tendency often lasts for a while before disappearing; it does not continue. Seasonal variation is a repeating pattern that occurs throughout each year or any other fixed period. The term "*seasonal component*" refers to the repeating pattern [9]. If time series has a trend or seasonal component then it is not stationary. If our series is not stationary, we apply transformations, such as differencing, until the series is stationary. Time series must be stationary because we want to reduce or eliminate the dependence between the data. Denote the time series as $X = \{X_1, X_2, \ldots, X_t, \ldots, X_n\}$, where X_t is value at time t. Discrete differentiation is defined as follows:

$$\text{If } d=0: \; y_t = X_t \quad (2)$$
$$\text{If } d=1: \; y_t = X_t - X_{t-1} \quad (3)$$
$$\text{If } d=2: \; y_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2} \quad (4)$$

The second difference of X is the *first-difference-of-the-first difference*, which is the discrete equivalent of a second derivative, not the difference from two periods ago. After the first differentiation, if the time series becomes stationary, the parameter d is 1, then it is 2, and so on.

The sample autocorrelation function can also indicate deviations from stationarity. *Autocorrelation function* (ACF) and *Partial Autocorrelation function* (PACF) have graphs that slowly decay and are almost periodic, respectively. We once more use differentiation to remove trends and seasonality. For the stationarity test, in this case, we plot the Autocorrelation Function (ACF) graph and check it for outliers. The maximum value of ***q*** is the number of outliers. The MA algorithm sometimes does not perform well while predicting future values. In that case, it must be applied together with the Auto Regressive model to obtain better results [10].

3.1.2. Auto-Regressive Integrated Moving Average model

Auto Regressive Integrated Moving Average (**ARIMA)** is a model for forecasting future outcomes of time series based on its own past values, that is, its own lags and lagged forecast errors. The ARIMA model is characterized by three parameters:

- ***p***-order of Auto Regressive term
- ***d***-number of differencing required to make time series stationery
- ***p***- order of Moving Average term

Below each parameter of the ARIMA model has been explained as well as the way of finding each one of them. Parameter ***d*** stands for differentiation of time series. We perform differentiation to make time series stationary, and the reason for this is explained in subsection 3.1.1. Parameter ***p*** is referred to as the Auto Regressive part of the model. The term *auto-regression* indicates that it is a regression of the variable against itself. To determine the value of ***p*** we will look at the graph of PACF. The number of outliers in PACF provides the value for the parameter ***p*** of the model.

If the $\overline{X_t}$ is the value of time series X at time t that we want to predict, and ***p is*** the number of previous values of time series that we want to include in prediction, formula of an autoregressive model can be written as:

$$\overline{X_t} = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (5)$$

where ε_t is white noise and $y_{t-1}, \ldots y_{t-p}$ are dth derivative of past values.

A Moving Average model creates a regression-like model which uses previous prediction mistakes rather than past values of the forecast variable. The Moving Average model of order **q** has the following prediction formula:

$$\overline{X_t} = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (6)$$

where ε_t is white noise. The way of finding value for parameter q is described in section above. If we sum up all this together, the formula for ARIMA(p,d,q) is:

$$\overline{X_t} = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \tag{7}$$

where $y_{t-1}, \ldots y_{t-p}$ are dth derivative of past values, and $\varepsilon_1, \ldots, \varepsilon_{t-q}$ are past forecast errors [11], [17].

3.1.3. Holt Winters model

Holt-Winters or Triple Exponential Smoothing model belongs to a class of Exponential Smoothing models. The output of an ARIMA model is a linear weighted sum of previous observations. Weights in that sum do not exhibit any patterns. On the other hand, the output of the Holt-Winters model is again a linear combination of previous observations, but weights decrease exponentially as observations get older. That implies the following- *Recent observations have bigger weights than older observations*. ARIMA model cannot directly deal with time series with trend or seasonality, but Exponential Smoothing algorithms can.

Simple Exponential Smoothing and Double Exponential Smoothing cannot be applied in our case, since our time series has both trend and a seasonal component. That is why in this thesis the Holt-Winters model is applied. Let us start by explaining what Simple Exponential smoothing is made for, then the model will be upgraded for cases when time series express seasonality or have trends.

Let an observed time series be X= {X_1, X_2, ..., X_t, …, X_n}, with the only level component. Formally, the simple exponential smoothing predicting equation takes the following form:

$$\overline{X}_{i+1} = \alpha X_i + (1-\alpha)\overline{X}_i \tag{8}$$

where X_i is the actual, known series value at the time i, $\overline{X}_i$ is the forecast value of the X at the time i, $\overline{X}_{i+1}$ is the forecast value at time $i+1$ and α is the smoothing constant. Smoothing constant α is a selected number between zero and one, $0 < \alpha < 1$. Rewriting the model (8) to see one of the neat things about how the model

$$\overline{X}_{i+1} - \overline{X}_i = \alpha\left(X_i - \overline{X}_i\right) \tag{9}$$

changes in forecasting value is proportionate to the forecast error. That is

$$\overline{X}_{i+1} = \overline{X}_i + \alpha e_i \tag{10}$$

where $e_i = X_i - \overline{X}_i$ is forecast error for a time *i*. So, the exponential smoothing forecast is the old forecast plus an adjustment for the error that occurred in the last forecast. By continuing to substitute previous forecasting values back to the starting point of the data in model (8) we receive:

$$\overline{X}_{i+1} = \alpha X_i + (1-\alpha)\left(\alpha \overline{X}_{i-1} + (1-\alpha)\overline{X}_{i-1}\right) = \alpha X_i + \alpha(1-\alpha)X_{i-1} + (1-\alpha)^2\, \overline{X}_{i-1} \quad (11)$$

$$\overline{X}_{i+1} = \alpha X_i + \alpha(1-\alpha)X_{i-1} + \alpha(1-\alpha)X_{i-2} + (1-\alpha)^3\, \overline{X}_{i-2} \quad (12)$$

$$\overline{X}_{i+1} = \alpha X_i + \alpha(1-\alpha)X_{i-1} + \alpha(1-\alpha)^2\, X_{i-2} + \alpha(1-\alpha)^3\, X_{i-3} + (1-\alpha)^4\, \overline{X}_{i-4} \quad (13)$$

If we continue substituting, the forecast equation in general form is:

$$\overline{X}_{i+1} = \alpha X_i + \alpha(1-\alpha)X_{i-1} + \alpha(1-\alpha)^2\, X_{i-2} + \cdots + \alpha(1-\alpha)^{i-2}\, X_2 + \alpha(1-\alpha)^{i-1}\, X_1 \quad (14)$$

[10, 11].

There are two types of time series, *additive*-time series is represented as sum of its components, and *multiplicative*-time series is represented as a product of its components. If our time series has trend component, then that component is added to equations for level update and we get Double Exponential Smoothing model. The equation for trend update is:

$$T_i = \beta(L_t - L_{t-1}) + (1-\beta)T_t \quad (15)$$

where T_t is trend update at the moment t and L_t is level at moment t.

If our time series has an additive trend, forecasting values are obtained as follows:

$$\overline{X}_{i+k} = L_i + kT_i \quad (16)$$

where number k is a number of predictions into the future. On the other hand, if our data has a multiplicative trend, the forecasting equation is:

$$\overline{X}_{i+k} = L_i * T_i^k \quad (17)$$

Parameter β is from the interval [0,1] and it is an additional parameter that we need to take into consideration while finding the best model.

Holt-Winters or Triple Exponential smoothing model is an extension of previous models for data with seasonal component. Seasonality components can be, as trend component, additive and multiplicative, which lead us to two equations for seasonal update:

$$\overline{S}_i = \gamma(X_i - L_i) + (1-\gamma)S_{i-m} \quad (18)$$

$$\overline{S}_i = \gamma\frac{X_i}{L_i} + (1-\gamma)S_{i-m} \quad (19)$$

where ***m*** is the period-the length of that fixed period, and S_{i-m} is seasonal estimation at time ***i***. γ is the seasonality component parameter from the interval [0,1], the third to be tuned while finding the best model. In this case, our time series has both trend and seasonal component, therefore four different forecast equations will be listed below [11]:

- Both components are additive: $\overline{X}_{i+k} = L_i + kT_i + S_{i+k-m}$ (20)
- Trend component is additive, seasonal component is multiplicative:
  $$\overline{X}_{i+k} = (L_i + kT_i) * S_{i+k-m} \quad (21)$$
- Trend component is multiplicative, seasonal component is additive:
  $$\overline{X}_{i+k} = L_i + T_i^k + S_{i+k-m} \quad (22)$$
- Both components are multiplicative: $\overline{X}_{i+k} = L_i * T_i^k * S_{i+k-m}$ (23)

3.1.4. Akaike's Information Criterion

For the evaluation of the performance of the models the *Akaike's Information Criterion* (AIC) was used. AIC is a single number score, and it is used to determine which of multiple models is the best one for a given dataset. It estimates models relatively, meaning that AIC scores are only useful in comparison with other AIC scores for the same dataset. A lower AIC score is better. AIC is especially useful for time series analysis. Akaike's Information Criterion is calculated using the formula defined as:

$$\text{AIC} = \text{-2(log-likelihood)} + 2\text{K} \quad (24)$$

where K is the number of model parameters (the number of variables in the model plus the intercept) and Log-likelihood is a measure of model fit.

The maximum likelihood estimation (log-likelihood) of a model is used by AIC as a fitness metric. The model that "fits" the data the best is the one with the highest likelihood. For computational simplicity, the likelihood's natural log is used. AIC is low for models with high log-likelihoods (the model fits the data better, which is what we want), but it adds a penalty term for models with higher parameter complexity because a model is more likely to overfit the training data when it has more parameters. When deciding between several distinct model types without access to out-of-sample data, AIC is frequently used [12].

3.2. Adaptive Large Neiborhood Search

A heuristic approach based on the Adaptive Large Neighborhood Search meta-heuristic is developed [4], as the CPSP is an NP-hard issue. “*A heuristic algorithm is a method that uses the structure of the problem to quickly identify good feasible solutions while exploring the set of feasible solutions to an optimization problem in an intuitive manner* “ [13].

The Neighborhood Search heuristic will now be discussed. A possible optimization problem solution x and a set N(x) of further feasible solutions are connected by a neighborhood structure N, which is a function. The neighbors of x are made by modifying the x. Set of all possible modifications represent *the neighborhood of x*, and the nature of it depends on the problem. Usually, the problem solving is easy to solve in the vicinity of the feasible solution, and that is how we can recognize the good neighborhood structure. The concept of a local search is: “*Until the current feasible solution is the best in the neighborhood, the current iteration is replaced by its best feasible neighbor at each iteration.”* [13].

ALNS is a general framework into which problem-specific information can be inserted. One operator *o* from a group of neighborhood operators *O* can be used to create a new solution from an old one. In a way similar to simulated annealing, the new solution is evaluated and approved or refused. In metallurgy, annealing is a process that involves heating a metal and then gradually cooling it to enhance its qualities. In a similar manner, simulated annealing is a local search extension that permits iterations with larger objective function values. The chance of accepting such iterations is controlled by a parameter called "temperature"[13]. The likelihood of accepting them increases with temperature. In particular, if the new solution to a multi-objective optimization problem, like the CPSP, is not dominated by one of the previously obtained solutions, then it is always accepted. On the other hand, if the new solution is dominated, the likelihood of accepting it is influenced by the solution's worsening and by the temperature T, which is a function of the algorithm's cooling schedule. The likelihood of selecting a certain operator at each iteration is based on the weights given to each operator, which are a function of the operator's performance. This method allows the algorithm to choose the most promising operators throughout the meta-heuristic [4]. In Algorithm 1, the ALNS's pseudocode is presented. A heuristic algorithm is obtained when the operators related to the problem are defined and inserted into the ALNS framework. The pseudocodes for 8 problem specific operators are presented in subsection 4.5.

***Algorithm 1*: ALNS with simulated annealing**

Initialize ALNS and simulated annealing parameters

Initialize probabilities P^{0}_{o} for each operator o∈O

Start from the initial solution S. Assign $S_{curr} \leftarrow S$

Initialize list of non-dominated solutions L←{S}

Iter ← 1

repeat

    Draw random number to select an operator o∈O according to probabilities P

    Apply operator o to obtain new solution S_{new}

    if S_{new} is not dominated by a solution in L **then**

        $S_{curr} \leftarrow S_{new}$

        Add S_{new} to list L

        Remove from L all solutions dominated by Snew

    else

        Calculate F_{new}, F_{curr} weighted sum of objective function values for S_{new}, S_{curr} respectively

        Draw random probability r

        if $r < e^{\frac{F_{new}-F_{curr}}{T}}$ **then**

            $S_{curr} \leftarrow S_{new}$

    Iter ← Iter+1

    Update probabilities P based on performance of operator o

    Update temperature T

until *Iter > MaxIter*

4. Models and Algorithms

4.1. Data

Datasets used in this thesis are obtained from two sources. For the first approach, synthetic datasets have been used. The datasets were created to be suitable for solving the CPSP, inspired by the original datasets used for Syngenta Crop Challenge in Analytics 2021. Dataset 1 describes the input variables for an optimization model. For each corn population, three following dates are given: actual planting date, earliest date the population could have been planted and latest date the population could have been planted, the number of Growing Degree Units (GDUs) in Celsius needed for harvest and finally, scenario 1 and scenario 2 harvest quantities. Dataset 2 gives the GDUs in Celsius accumulated for each day for site 0 and site 1, over the last 10 years.

For the second approach, dataset that contains daily for the past 10 years are obtained from the website *www.meteostat.net* [14]. Temperatures for site 0 are taken from meteorological station *"Novi Sad-Rimski Šančevi"*, and for site1 from meteorological station "*Batajnica*". Now Dataset 2 has minimum and maximum daily temperatures instead of GDUs. The task is to predict daily temperatures for each site and then to calculate GDUs, using the formula for GDU calculation:

$$GDU = \frac{T_{min} - T_{max}}{2} - 10°C \qquad [16] \qquad (25)$$

4.2. Data Preprocessing

Each population is described by the following attributes: harvest quantity, required GDUs for harvesting, initial planting date, the earliest and the latest date that population could have been planted. The predictions made with time series models stated in the previous section are the input of the optimization model, while the output is estimated harvest week. That information helps to define the feasible set F of population ***p*** and harvest week ***w*** pairs. Then, we can backtrack and choose a possible planting date, the one that minimizes the objectives. Therefore, the optimization model is based on harvest weeks [5].

For the second approach, more realistic datasets are made. Instead of daily accumulated GDUs, for each site, now dataset has daily minimum and maximum temperatures for last 10 years. Each population is described in the same manner as in the first approach.

4.3. Time series forecasting models

All models are trained using Google Collaboratory [15]. As mentioned above, three different models for predictions are used: MA, ARIMA and Holt-Winters.

4.3.1. First approach

In the first approach, historical data of accumulated GDUs are separated into train (80% of the data) and test (20% of the data) dataset. Those datasets are used for selecting the best parameters of the models for an out-of-sample prediction. In this subsection the results of different models for GDUs predictions are presented. Predictions are based on historical data, the accumulated GDUs for years 2009 to 2019, for site 0 and site 1, independently. Time series of historical data are presented in Figure 3(a) and Figure 3 (b), for site 0 and site 1, respectively. With the help of stated models, forecasts of the accumulated GDU units per day for a period of 80 weeks are made, starting from 01/01/2020. Models will be compared using AIC criterion, and the best result, model with the lowest AIC, will be the input for the optimization algorithm. We will discuss how predictions made with different models affected the number of harvesting weeks and harvesting quantities.

4.3.1.1. Moving Average model

To train the MA model, firstly we need to determine the value of the parameter $\boldsymbol{q}$. However, to determine the value of parameter $\boldsymbol{q}$, we need to check if our time series is stationary. In Figure 3 and Figure 4, respectively, the decomposition of time series of accumulated GDUs for site 0 and site1, are shown.

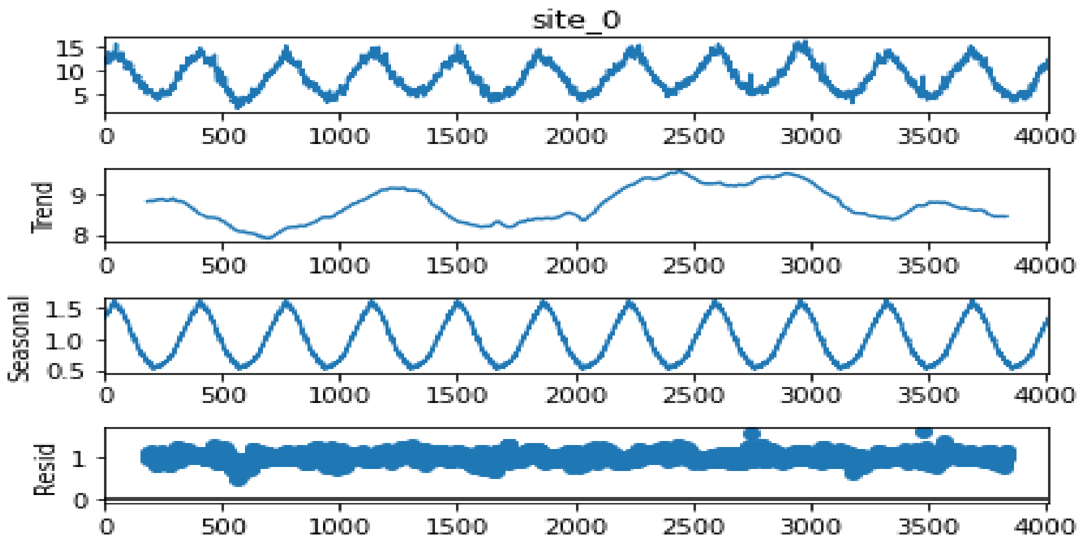

Figure 3: Decomposition of time series for site 0

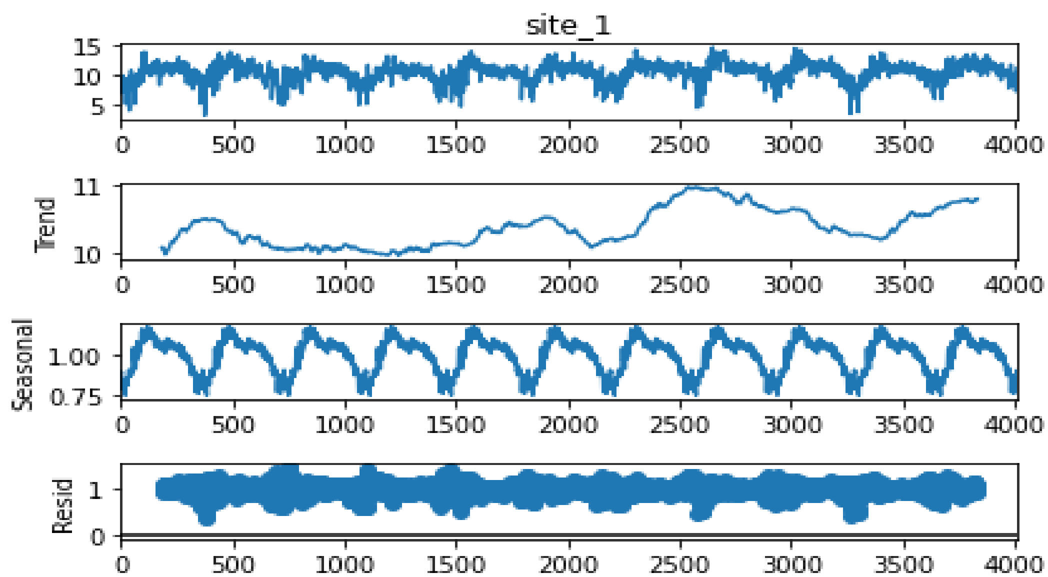

Figure 4: Decomposition of time series for site 1

It can be seen that time series of accumulated GDUs for site 0, has seasonality component, therefore it is not stationary. Same holds for time series of accumulated GDUs for site 1.

The parameter ***q*** can be determined from the ACF of differentiated time series. Figure 5 and Figure 6, show the ACF of site 0 and site 1, respectively.

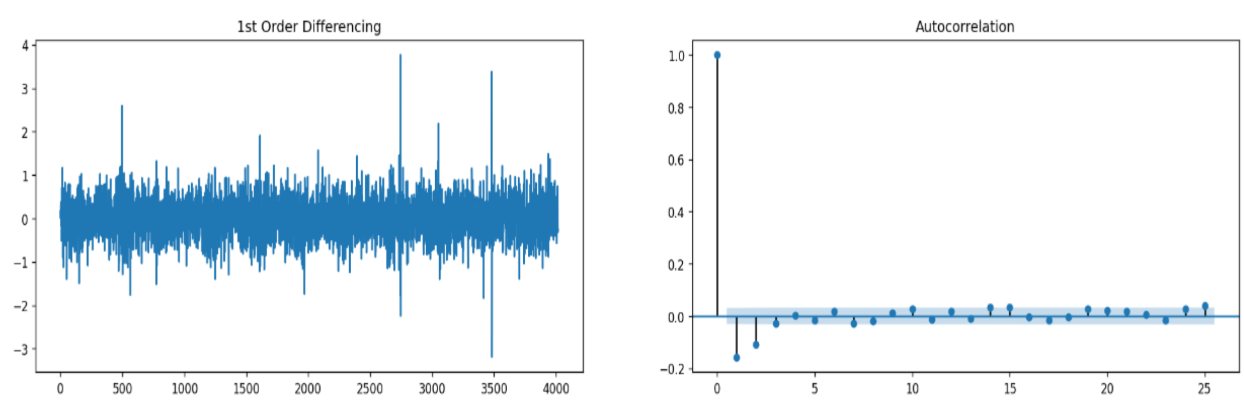

Figure 5: Autocorrelation function of differentiated time series for site 0

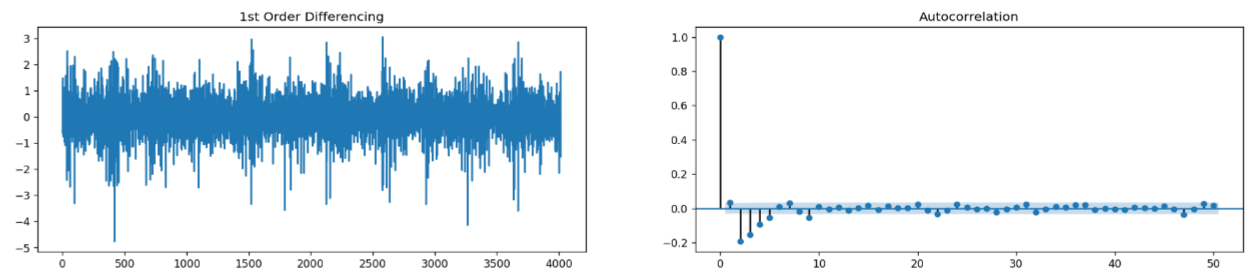

Figure 6: Autocorrelation function of differentiated time series for site 1

The ACF graphs imply that MA parameter should be maximum 3 for site 0, and maximum 8 for site 1. Here, we take advantage of Akaike's Information Criterion (AIC) to find the best parameters among several candidates. The lowest AIC for site 0 is achieved with MA(3), and for site 1, with MA(5). The predictions made with these models are input for the optimization model. It will be discussed how those predictions impacted on the output.

4.3.1.2 ARIMA model

To improve the results, the ARIMA model for prediction of the GDUs is applied. Firstly, let us determine the values of the parameters. We need to check if time series is stationary or not, if we want to determine the value of parameter ***d.*** In the previous subsection, it is already concluded that both time series are non-stationary. Stationarity is achieved after only one differentiation, therefore ***d***=1.

As mentioned above, the parameters ***p*** and ***q*** can be determined by plotting the graphs of autocorrelation functions. In subsection 4.1.1 the maximum values for MA part of the model, have already been determined, now we will find out the values for AR parts of the models. Looking at the PACF graph of site 0 (Figure 7), the AR parameter should be maximum 4.

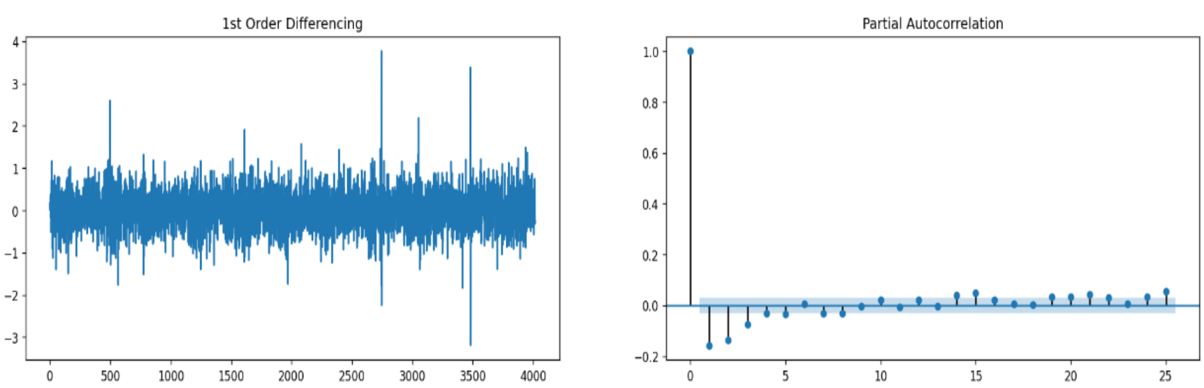

Figure 7: Partial Autocorrelation function for site 0

From Figure 8 we can conclude that maximum p for site 1 is 7.

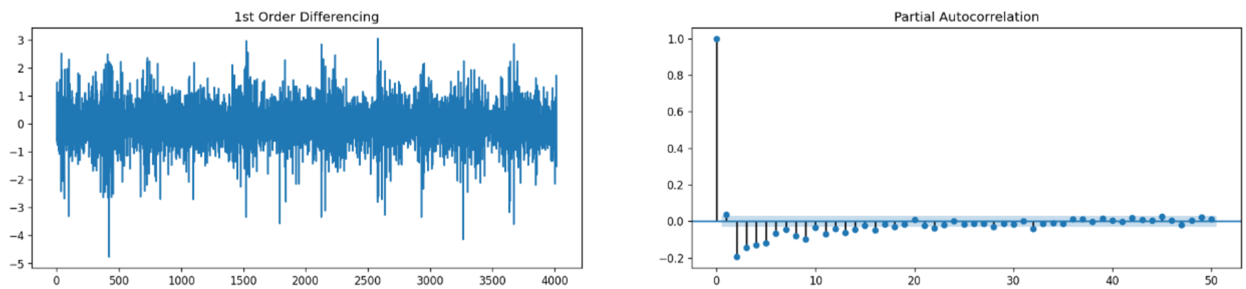

Figure 8: Partial Autocorrelation function for site 1

Again, the AIC criterion determines the best model. Minimum AIC is achieved for ARIMA(2,1,2) for site 0. Similarly, minimum AIC suggests that ARIMA(7,1,8) should be used for forecasting for site 1.

ARIMA models gave better predictions and lower AIC on test dataset than the MA models. Therefore, better results of optimization model could be expected with ARIMA predictions as input.

4.3.1.3. Holt Winters model

The Holt-Winters method has several parameters that could be tuned. Those are *smoothing level* or, α parameter, then β or *smoothing slope* for trend component, and finally, γ or *smoothing seasonal* for seasonal component of time series. If time series has seasonality component, which is already concluded, the period (seasonal frequency) must be specified. Since we have daily predictions, seasonal frequency is set to 365. Additional thing that we need to pay attention to is the nature of trend and seasonal component, namely they can be additive or multiplicative.

Firstly, we trained and fitted the model on train dataset, without specifying the values for parameters, because Holt-Winters model can pick best parameters automatically. However, sometimes models can be improved to give better or more logical results if we tune parameters manually. Secondly, after the model automatically picked the parameters, we tried to make new models by increasing or decreasing some of parameters and tracking the behavior of AIC. In some cases, an additional parameter is helpful. Figures 3 and 4 show us that both site 0 and site 1, time series has seasonality components.

Since ARIMA models cannot directly deal with seasonality, Holt-Winters model for GDUs predictions is chosen, to improve the performance of optimization model. The definition of multiplicative trend and seasonality implies that values of time series must be positive to calculate predictions. Therefore, it could only be stated that time series has additive components, and that is what we applied while training Holt Winters model.

Again, AIC criterion helped to decide which model is the best. That was the one with the lowest AIC. For site 0 GDUs, model showed that optimal parameters are $\alpha \approx 0.94$, $\beta \approx 0.002$, $\gamma \approx 0.002$. However, the following parameters gave us to lowest AIC: $\alpha \approx 0.75$, $\beta \approx 0.00006$, $\gamma \approx 0.002$.

On the other hand, for site 1 time series, we did not improve the results if we changed those that model tuned automatically. Optimal parameters for site 1 out-of-sample GDUs predictions are: $\alpha \approx 0.01$, $\beta \approx 0.001$, $\gamma \approx 0.001$. Recent predictions have a greater impact on predictions for site 0. On the other hand, for site 1 older predictions are more important. Parameter α implies those conclusions.

Stated models gave us the GDUs predictions, inputs for optimization algorithm, in both scenarios. If we compare the AIC of Holt Winters models, with AIC of MA and ARIMA, Holt Winters has the lowest AIC, but the predictions are not the best, since the models did not fit the data very well.

4.3.2. Second approach

The out-of-sample prediction, together with CPSP model optimization, was performed in PyCharm software, for all models and in both approaches.

In this subsection the results of different models for minimum and maximum daily temperature predictions will be presented, separately. Models will, again, be compared using AIC criterion, and the model with the lowest AIC, will be the input for the optimization algorithm. It will be discussed how different models affected the number of harvesting weeks and harvesting quantities. Let us recall the definition for calculating GDUs:

$$GDU = \frac{Maximum\ daily\ temperature - Minumum\ daily\ temperature}{2} - 10°C \qquad (26)$$

With the following constraints:

- *if the daily max is greater than 30*°C then *it is set to be equal to 30*°C
- *if the daily minimum is less than 10*°C *then it is set to be equal to 10*°C
- *If the GDU is negative then it is set to be equal to 0"* [16]

There are two possibilities for predicting temperatures, according to GDU (Growing Degree Units) formula definition. The first possibility is to predict the daily temperatures and then round them following the constraints. On the other hand, we can first round the temperatures and then perform the prediction. This idea arises after noticing that some models have problems with predicting high oscillations in the temperatures, and negative temperatures as well and the results showed that it is better to first round the temperatures and then to make predictions. Both possibilities have been tested in the following way. Firstly, we made the predictions of minimum and maximum daily temperatures, without rounding it, independently, and then calculated the accumulated GDUs for each day. Secondly, we repeat the process but, with rounded temperatures. In the end we compared the calculated GDUs with test dataset, and the lower MSE is obtained with second possibility i.e., with a prediction made with rounded temperatures. In the following subsection both time series will be presented, before and after rounding the temperatures.

4.3.2.1. Moving Average Model

As in the first approach, firstly describe the way of finding parameters of the model will be described. Again, predictions of the model with lowest AIC are used as input for optimization model. Figures 9 (a) and 10 (b) show the time series of minimum daily temperatures for site 0 before and after setting all temperatures lower than 10, to be equal to 10, respectively.

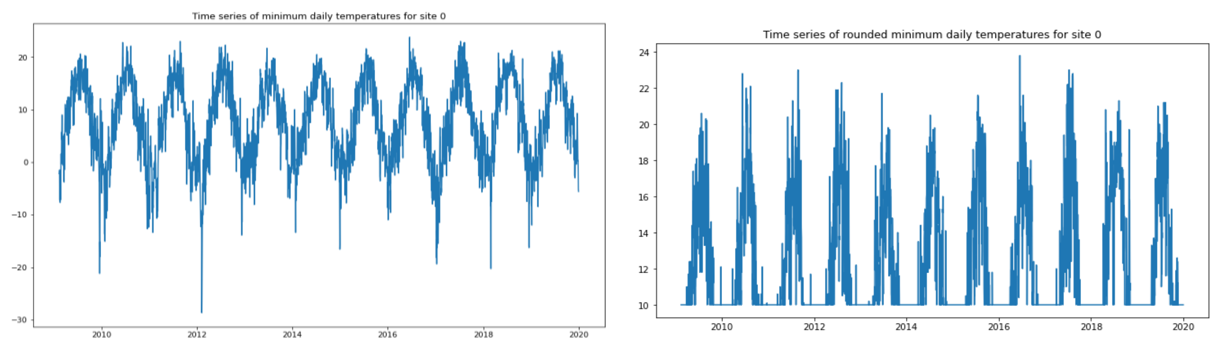

Figure 9: Time series of minimum daily temperatures for site 0 (a) before and (b) after rounding

Analysis and parameter tuning is performed for rounded time series. Let us first check the seasonality decomposition of time series, to determine the stationarity and parameter ***d***. Time series has seasonality component; therefore, it is not stationary, however after first order differentiation, we obtain stationarity, and parameter ***d*** is 1. Figure below implies the stated conclusion.

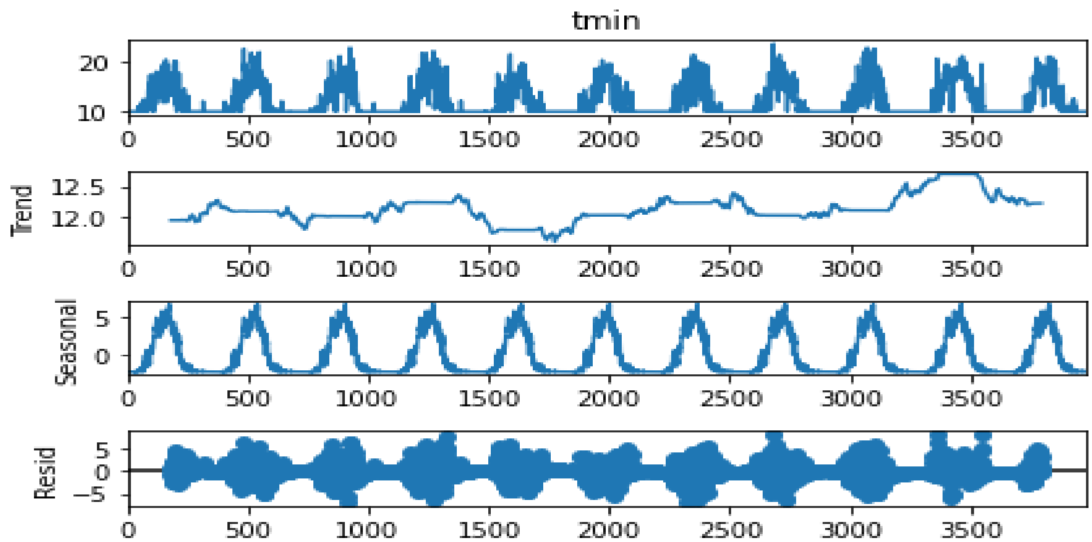

Figure 10: Decomposition of time series of minimum daily temperatures for site 0

Parameter ***q*** is the number of outliers of ACF function of differentiated time series; therefore, ***q*** is maximum 6. However, the lowest AIC is obtained with MA(5).

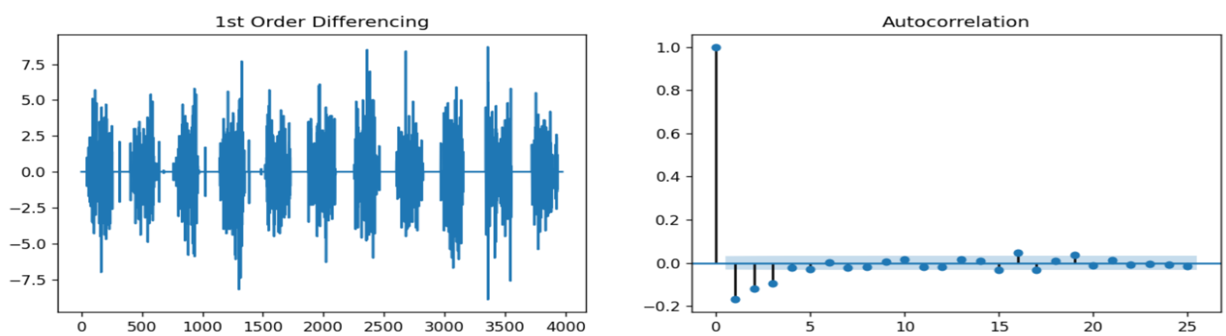

Figure 11: Autocorrelation function of time series of minimum daily temperatures site 0

In the same way we determined the parameters for the maximum daily temperature predictions. Time series before and after rounding all temperatures higher than 30, are shown in the following Figure.

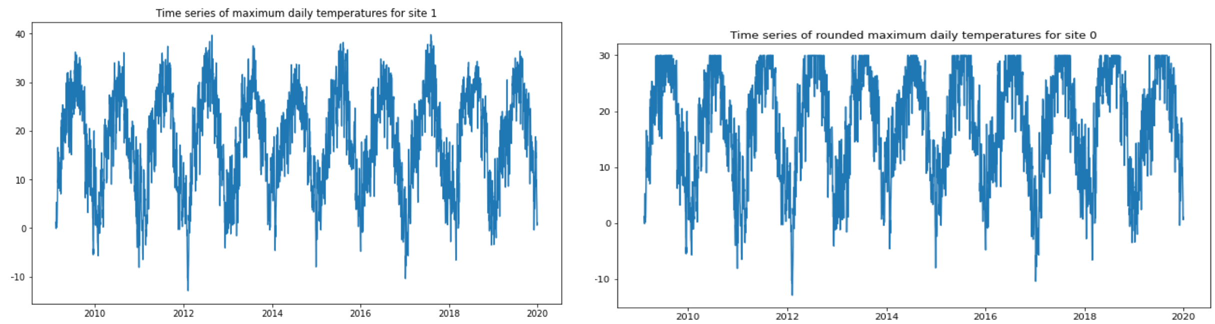

Figure 12: Time series of maximum daily temperatures for site 0 (a) before and (b) after rounding

Time series becomes stationary after only one differentiation, and therefore parameter ***d*** is 1.

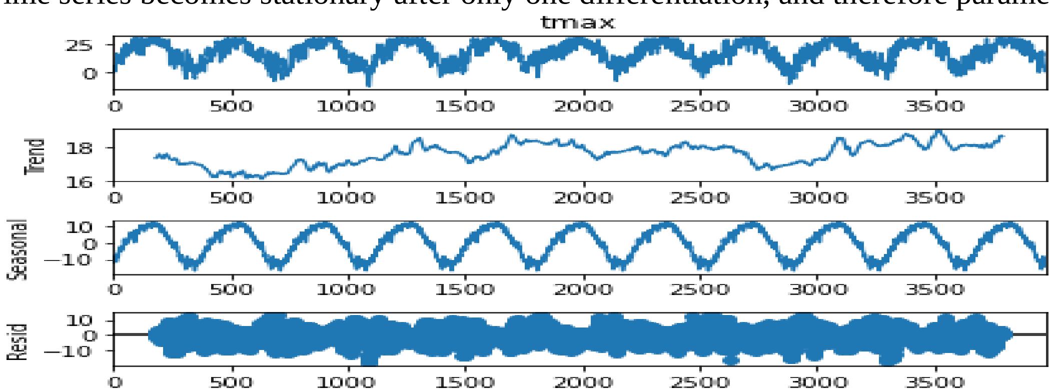

Figure 13: Decomposition of time series of maximum daily temperatures for site 0

Parameter ***q*** is maximum 5, since the number of outliers of ACF is 5.

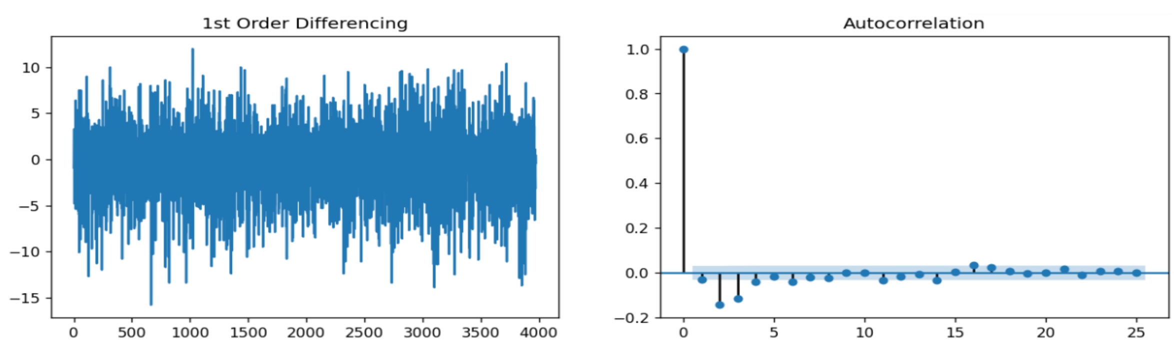

Figure 14: Autocorrelation function of maximum daily temperatures time series for site 0

The lowest AIC gives the MA(3), and predictions made with this model are used as input for daily maximum temperature predictions.

After finding both minimum and maximum temperature predictions, the GDUs are calculated following the formula for calculation stated in subsection 4.3.2. The calculated GDUs provided the input for optimization model. In the end we will see how different temperatures prediction affected the output.

The same analysis was conducted for site 1. In the figures below show the time series of minimum and maximum daily temperatures, before and after we applied the constraints stated in GDU formula.

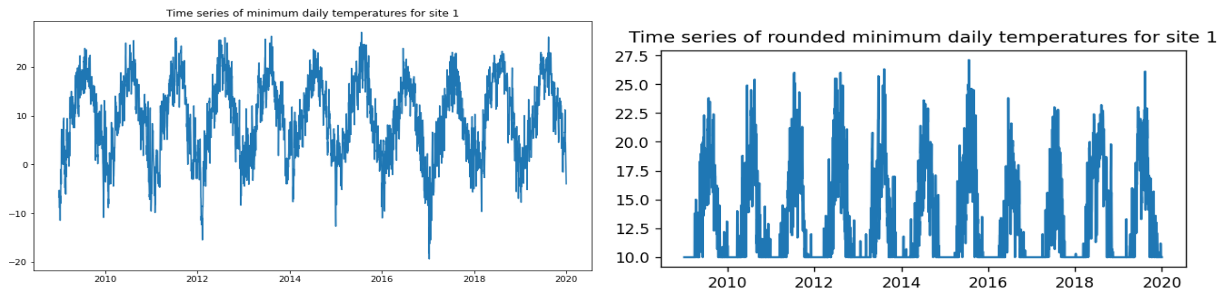

Figure 15: Time series of minimum daily temperatures for site 1 (a) before and (b) after rounding

Both time series have seasonality component, and both become stationary after first order differentiation, therefore ***d*** is 1.

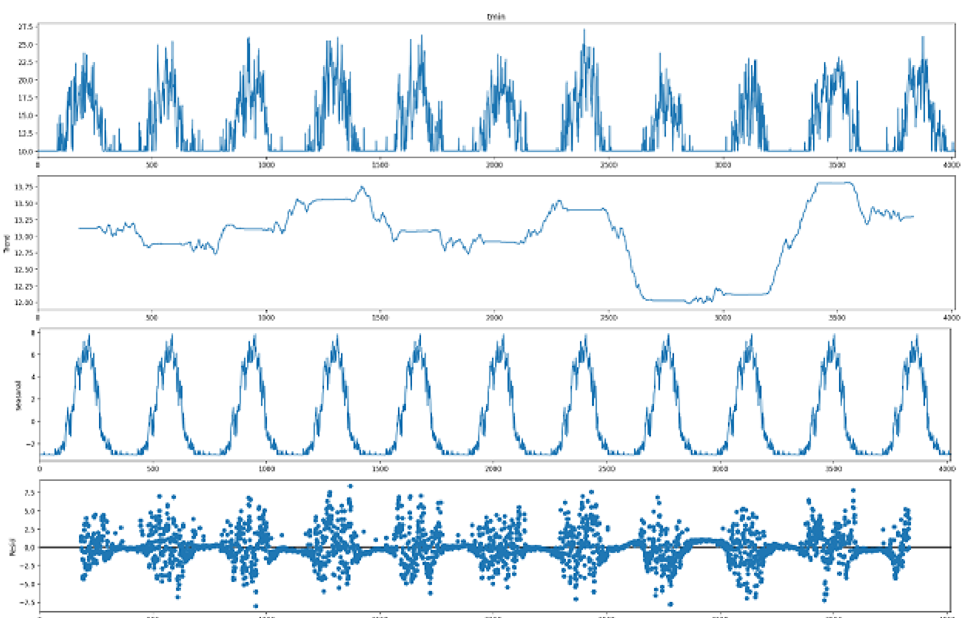

Figure 16: Decomposition of time series of minimum daily temperatures for site 1

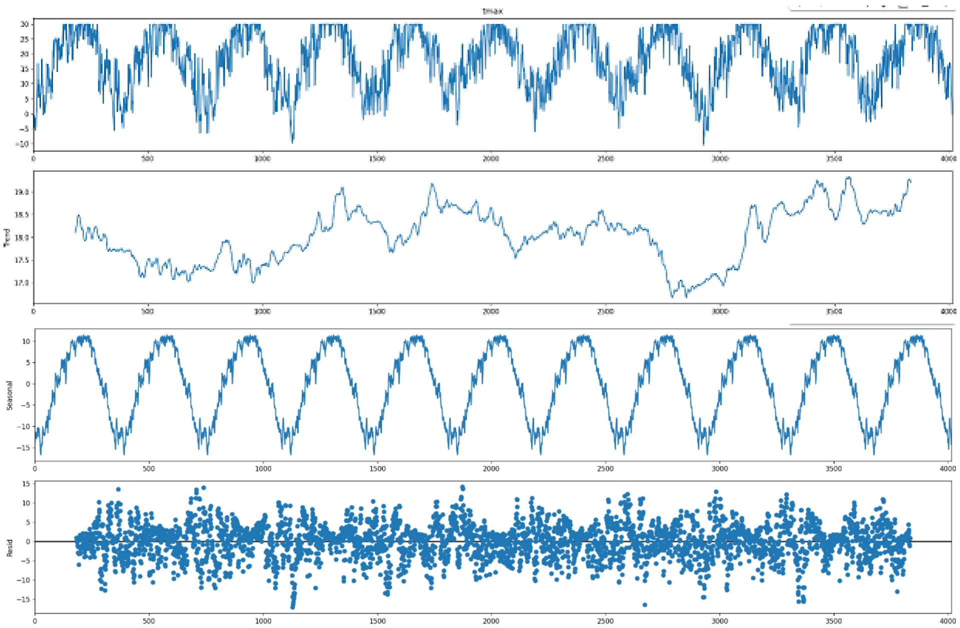

Figure 17: Decomposition of time series of maximum daily temperatures for site 1

Parameters ***q*** for minimum daily temperatures prediction is 5, since that is the number of outliers of ACF of differentiated series.

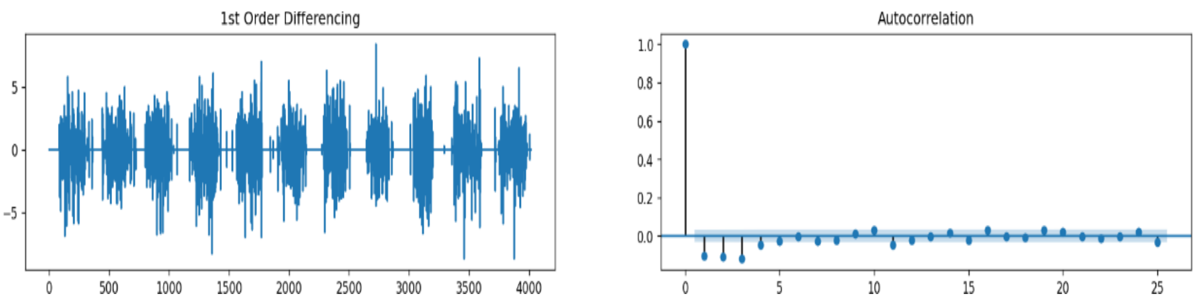

Figure 18: Autocorrelation function of minimum daily temperatures for site 1

The same conclusion holds for parameter ***q*** of maximum daily temperatures, which is 4.

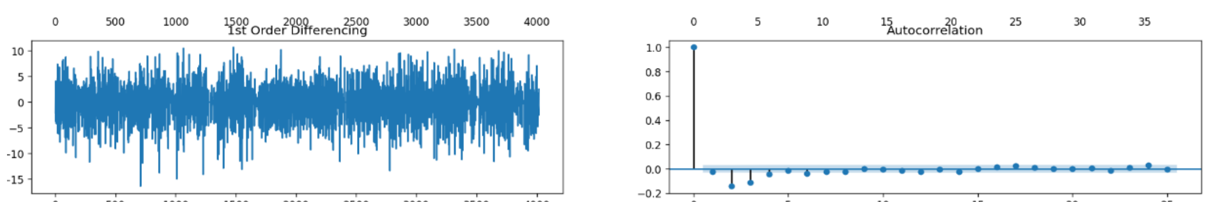

Figure 19: Autocorrelation function for maximum daily temperatures for site 1

The lowest AIC have the models MA(5) and MA(3) for minimum and maximum daily temperature predictions, respectively. After that, the GDUs are calculated using the formula for GDU calculation stated in Section 3.

4.3.2.2. ARIMA Model

The ARIMA model has three different parameters that need to be determined. In subsection 4.2.1 the parameter ***d*** and ***q*** have already been determined, for minimum and maximum temperatures, and for both sites. Here, the order of parameter ***p*** will be found, which is the number of outliers of PACF of differentiated time series.

Time series of minimum daily temperatures for site 0 are all shown in subsection 4.2.1. Figure below shows the PACF of time series, and we can read that maximum ***p*** is 14.

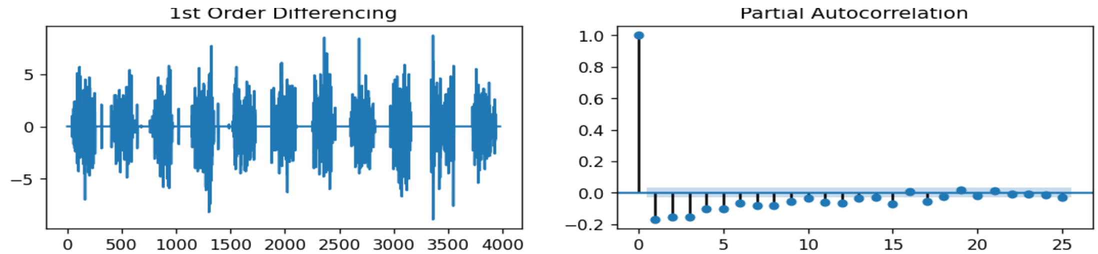

Figure 20: Partial Autocorrelation function of minimum daily temperatures for site 0

The ARIMA(10,1,4) has the lowest AIC, therefore we used it for predictions of minimum daily temperatures for site 0.

PACF of maximum daily temperatures is shown in Figure 21. We can read that maximum ***p*** is 12.

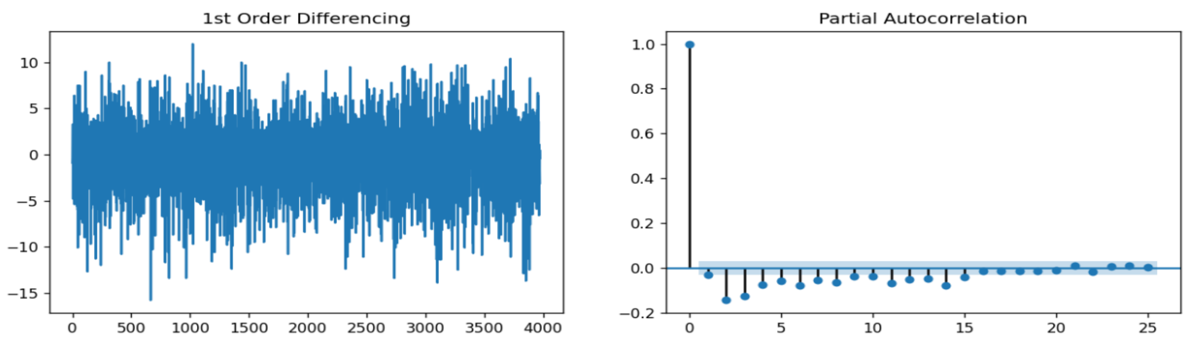

Figure 21: Partial Autocorrelation function of maximum daily temperatures for site 0

The ARIMA(9,1,3) has the lowest AIC, therefore it is used for predictions of maximum daily temperatures for site 0. The predictions made with those two models are further used for calculation of the GDUs for site 0.

The PACF graphs of differentiated time series of minimum and maximum daily temperatures for site 1 are shown in Figure 22 and 23, respectively. We can read that the number of outliers is 6 and 11, therefore maximum ***q***=6 and ***q***=11, for minimum and maximum daily temperatures, respectively.

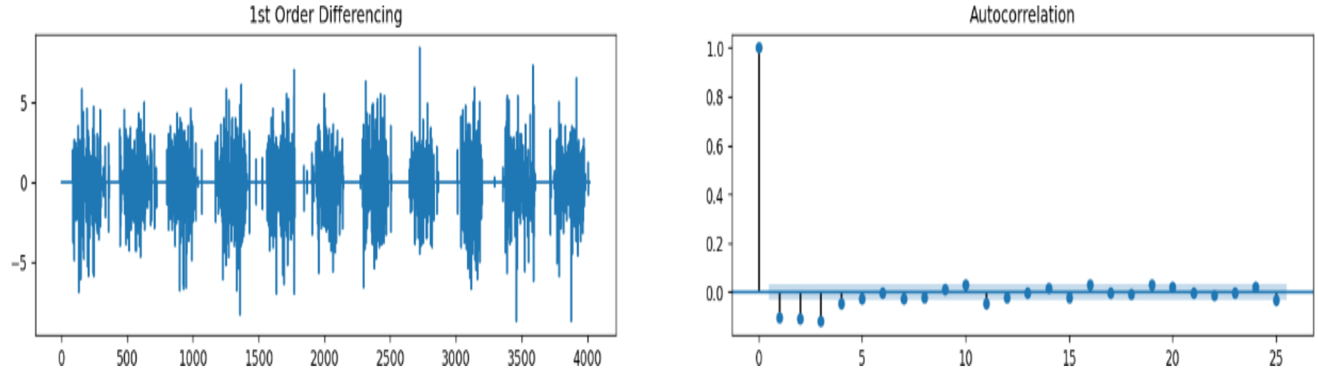

Figure 22: Partial Autocorrelation function of minimum daily temperatures for site 1

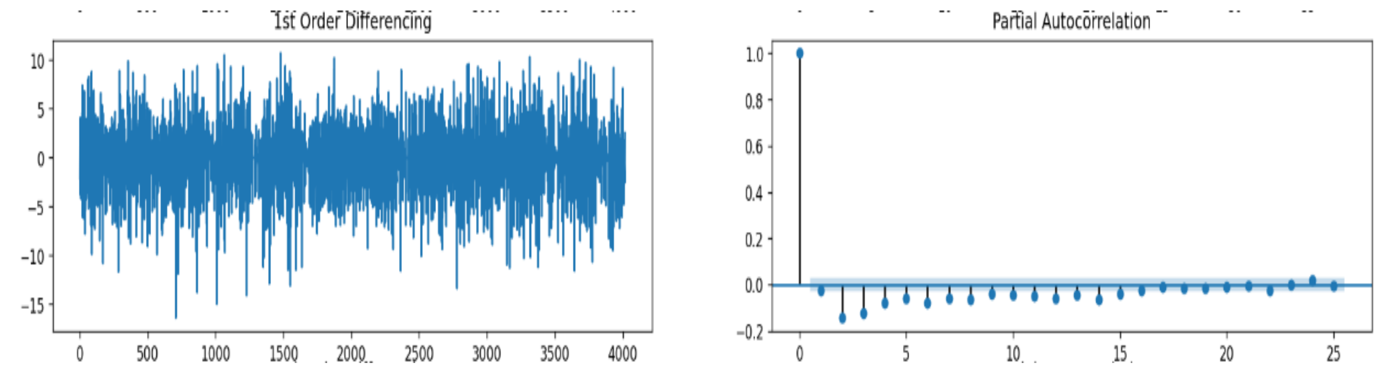

Figure 23: Partial Autocorrelation function for maximum daily temperatures for site 1

The best predictions for minimum daily temperatures provided ARIMA(6,1,3), with the lowest AIC. The ARIMA(6,1,3) is also selected as the best model for maximum daily temperature predictions for site 1. The GDUs predictions for site 1 are again calculated using the formula stated in the third Section.

ARIMA models have a lower AIC than MA models, for both sites and temperatures, so we expect to get better results, I.e., shorter harvesting period, smaller surplus, and greater tendency toward equalizing weekly harvest quantities.

4.3.2.3. Holt-Winters Model

Again, as in the first approach, the goal was to find the model with better results, the one that has AIC lower than MA and ARIMA. The Holt Winters model provided desired results. Due to the fact that we have daily predictions, seasonal frequency is always set to 365. In the previous subsections it is stated that all time series have seasonality component.

Firstly, let us describe how we got the model for predicting minimum temperatures for site 0. The Holt Winters model has two parameters that refer to the nature of trend and seasonal component, already stated. The definition of multiplicative trend and seasonality implies that values of time series must be positive to calculate predictions. In the previous case, it could only be stated that time series has additive components. However, this time all combinations of additive and multiplicative components can be tested since all negative temperatures are set to be equal to 10. Again, additive trend and additive seasonal component gave the lowest AIC.

The best model, after manual parameter tuning was performed, for site 0 minimum daily temperature predictions, is the model with following parameters: $\alpha \approx 0.91$, $\beta \approx 0.0011$, $\gamma \approx 0.0034$.

Regarding the maximum daily temperature predictions for site 0, we can say that only one constraint was present in parameter tuning. After rounding temperatures, we still have some negative values, since we did not round lower maximum temperature values only those greater than 30. This fact implies that trend and seasonal component could only be additive. Again, after manually tuning the parameters, the best Holt Winters model is obtained and it has the following parameters: $\alpha \approx 0.97$, $\beta \approx 0.00003$, $\gamma \approx 0.000012$.

The same analysis was conducted while searching for the best models for temperature predictions of site 1. As in the case of site 0 minimum temperatures, here we can also test all the combinations of additive and multiplicative components. Additive components provided the best results for minimum daily temperature predictions. Manual parameter tuning was again shown as a viable choice since it provided the model with the lowest AIC, which are $\alpha \approx 0.84$, $\beta \approx 0.0064$, $\gamma \approx 0.015$.

And finally, the model for maximum daily temperature predictions for site 1 was selected. Temperature values again implied that trend and seasonality could only be additive. Lowest AIC is obtained with: $\alpha \approx 0.95$, $\beta \approx 0.000019$, $\gamma \approx 0.00021$.
It can be noticed that in both cases, for both temperatures, recent predictions have the greater impact on the future values, since parameter α is very close to 1.

In the end, the daily accumulated GDUs are calculated, using the given temperature predictions. The results will be validated through the output of the optimization model.

4.3.2.4. Conclusion

In the second approach, the AIC scores of time series prediction models are very high, and that result in imprecise GDUs predictions. That is the reason why it is decided not to validate them with optimization model. In the Chapter 5, only models from first approach provided the input for optimization model. Tables 1 and 2 provides the AIC values for all prediction models, in both approaches, for site 0 and site 1.

	AIC (site 0)	AIC (site 1)
MA	4764	8689
ARIMA	4760	8629
Holt Winters	-6304	925

Table 1: AIC values of GDU prediction models for site 0 and site 1

	AIC (Tmin, site 0)	AIC(Tmax, site 0)	AIC(Tmin, site 1)	AIC(Tmax, site 1)
MA	18.643	21.131	14.541	20.613
ARIMA	18.588	21.088	14.528	20.591
Holt Winters	2.982	9.637	3.822	9.867

Table 2: AIC values of minimum and maximum daily temperature prediction models for site 0 and site 1

4.4. Crop Planting Scheduling Problem

The definition of the CPSP:

In order to produce the proper amount of corn, the planting schedule for the corn crop must be optimized. However, this quantity must be distributed throughout the time horizon as evenly as possible, and going over the available storage capacity must be avoided. In the first scenario, the predetermined capacity for the site is given. In the second scenario, there is no predefined capacity and the goal is to find which capacity will give the best results.

For solving this optimization problem, a mathematical model was constructed in [4]. As described in section above, the model was solved by using ALNS meta-heuristic.

Mathematical model

The model corresponding to scenario 1 is presented in Section 4.4.1. Then, some modifications of this model are made to solve the problem for scenario 2 and that is presented in Section 4.4.2. The sets and indices are shared by the two models; therefore, they are presented here after.

Indices

p- the population id {1, …, P}

w the week id {1, …, W}

d the day of the week {1, …, D}

scen the scenario id {1, 2}

s the site id {0,1}

Sets

P the population {1, …, 2569}

W the weeks {1, …, 60}

D the days in a week {1, …, 7}

FS (p, w): set of feasible combinations of population ***p*** and harvest week ***w***

4.4.1. Scenario 1

In this scenario, both sites have pre-defined capacity values for each week.

Parameters for scenario 1

$plSite_p$-planting site of population p

$reqGDU_p$ -required GDU units for population p

$obtGDU_{d;w;s}$-obtained GDU units on day d of week w on site s

$hq_{p;scen}$ -harvesting quantity of population p in scenario scen

$c1_{s;w}$ -harvesting capacity on site s in week w for scenario 1

Parameters h_q and c1 are used in the mathematical model, the rest of the parameters are used for pre-processing and post-processing.

Decision variables

pl_p^w=1 if population p is harvested in week w/ 0 otherwise

$used_w$=1 if week w is used for harvesting/ 0 otherwise

$m^+_{s,w}$=the positive deviation from the capacity of harvesting week w on site s

$m^-_{s,w}$=the negative deviation from the capacity of harvesting week w on site s

$n^+_{w1,w2}$=the increase in harvesting quantity from week w1 to week w2

$n^-_{w1,w2}$=the decrease in harvesting quantity from week w1 to week w2

Objectives

The problem definition implies a multi-objective approach which is presented below.

$$\min \quad f_1 = \sum_{w \in W} used_w \tag{27}$$

$$\min \quad f_2 = \sum_{s,\, w \in W} m^+_{s,\, w} + m^-_{s,\, w} \tag{28}$$

$$\min \quad f_3 = \sum_{w \in W} n^+_{w_1,\, w_2} + n^-_{w_1,\, w_2} \tag{29}$$

Objective (27) minimizes the number of weeks used for harvesting. Whit this objective, harvesting time horizon is shortest possible. Objective (28) minimizes the total deviation from the capacity.

The positive and negative deviations are equally important. Finally, objective (29) minimizes the total deviation of harvest quantity between two consecutive weeks.

Constraints

$$\sum_{w:(p,w)\in FS} pl_p^w = 1 \;\; \forall p \in P \tag{30}$$

$$\sum_{p:s=plSite_p, p:(p,w)\in FS} hq_{p,1}\, pl_p^w = C1_{s,w} + m_{s,w}^{+} + m_{s,w}^{-} \qquad \forall w \in W,\; s \in \{0,1\} \tag{31}$$

$$\sum_{p,\, w:(p,w)\in FS} pl_p^w \le used_w \le M_1 \tag{32}$$

$$\sum_{p:\, s=plSite_p, p:(p,w)\in FS} hq_1\, pl_p^w - \sum_{p:\, s=plSite_p, p:(p,w)\in FS} hq_1\, pl_p^{w+1} = n_{w,w+1}^{\pm} n_{w,\,w+1}^{-} \quad \forall\, w \in W : w \ne |W| \tag{33}$$

$$pl_p^w \in \{0,1\} \tag{34}$$

$$used_w \in \{0,1\} \tag{35}$$

$$m_{s,\,w}^{+},\, m_{s,\,w}^{-} \ge 0 \tag{36}$$

$$n_{w_1,w_2}^{+},\, n_{w_1,w_2}^{-} \ge 0 \tag{37}$$

With constraint (30) we make sure that all populations are planted. The weekly deviation from the harvesting capacity is determined with (31) and with (32) we get all weeks that are used for harvesting. M1 is a big-M value that is at most $\sum_{p,\,w:\,(p,w)\in FS} pl_p^w$. The difference between two consecutive days is defined with (33). The constraints, (34)-(37), are domain related constraints. The variables *pl* and *used* are binary variables whilst the deviation variables, i.e., *m* and *n*, are non-negative variables.

4.4.2 Scenario 2

In scenario 2, there is no pre-defined capacity. That implies that we need to find for which capacity the objectives are minimized, i.e. that $c1_{s;\,w,}$ becomes a decision variable, i.e., $c2_{s;\,w.}$ There is additional value to be minimized, the maximum value of this variable, i.e., Cs, to determine the capacity value for scenario 2. The rest of the model is the same and the model remains linear.

Additional decision variables

$c2_{s;\, w}$ -harvesting capacity on site s in week w for scenario 2

Cs minimum possible capacity for site s

Additional objective

$$\min f_4 = Cs \tag{38}$$

Objective (37) minimizes the maximum capacity level over all weeks for site s.

Additional constraints

$$c2_{s,w} \leq C_s \qquad \forall w \in W,\ s \in \{0,1\} \tag{39}$$

$$C_s \geq 0 \ \ \forall s \in S \tag{40}$$

$$c2_{s,\,w} \geq 0 \ \ \forall s \in S,\ w \in W \tag{41}$$

The constraint set (39) implies that the minimum possible capacity for site is bigger than the maximum capacity among all weeks for each site s. Additional objective actually manage to equalize those two capacities. Constraints (40) and (41) are clear [4].

4.5. Heuristic Algorithm

The optimization problem described in Section 4.4 has four objectives (27), (28), (29) and (38) and eight problem-specific operators. These operators improve the model solution and they stop the model from being stuck in local optima. There are 5 five operator groups. Rebalancing operators 1, 2, and 3 switch some populations' harvest weeks from high-quantity harvest weeks to low-quantity harvest weeks. Operator 4 is a stability operator, and it balances the amount of harvest collected over successive weeks. Operators 5 and 6 are empty operators, and they make the situation that all populations are harvested in one week, impossible. The harvest quantity for a given week is set as near to capacity as possible, by Operator 7. Only in scenario 2 operator 8 is used, and it updates the suggested lowest site capacity. The eight neighborhood operators' pseudocodes will be presented in the paragraphs below [4].

Algorithm 2: Operator 1

Given solution S, define the vector h of the harvest quantities h_w for each week $w \in W$
Find $w_0 \in W$ such that $w_0 = \text{argmax}_{w \in W}\ h_w$
Select a population $p \in P$ such that $pl_p^{w'} = 1$
Find week w'' in the set W_p of feasible harvest weeks of p such that $w'' = \text{argmin}_{w \in W_p}\ h_w \wedge h_w > 0$
Generate new solution S^* by assigning $pl_p^{w'} = 0$ and $pl_p^{w''} = 1$

Algorithm 3: Operator 2

Draw integer r from set {1,...,10}. Define set R = {1,.., r}
for $n \in R$ do
    Given solution S, define the vector h of the harvest quantities h_w for each week $w \in W$
    Find $w_0 \in W$ such that $w_0 = \text{argmax}_{w \in W} h_w$
    Select a population $p \in P$ such that $pl_p^{w'} = 1$
    Find week w'' in the set W_p of feasible harvest weeks of p such that $w'' = \text{argmin}_{w \in W_p} h_w \wedge h_w > 0$
    Update new solution S^* by assigning $pl_p^{w'}=0$ and $pl_p^{w''}=1$

Algorithm 4: Operator 3

```
Draw maximum number of iteration maxIter from set {1, 2,..,10}
Define a value Q > capacity for the maximum desirable harvest
i←1 while  ∧ max\sum_{p∈P} hq_p pl_w^p > Q i ≤ maxIter do
   Given solution S, define the vector h of the harvest
   quantities h_w for each week w ∈ W
   Find w' ∈ W such that w'=argmax_{w∈W} h_w
     Select a population p ∈ P such that plw 0 p = 1
    Find week w in the set Wp of feasible harvest weeks of p
        such that w''=argmin_{w∈Wp} h_w ∧ h_w>0
   Update new solution S* by assigning pl_p^w'=0 and pl_p^w''=1
   i ← i+1
```

Algorithm 5: Operator 4

```
Given solution S, define the vector h of the harvest
quantities h_w            for each week w ∈ W
Find (w, w + 1) ∈ W such that h_w > 0, h_{w+1} > 0 and |h_{w+1} − h_w| is
         maximized
Define w^- = argmin(h_w, h_{w+1}),  w^+ = argmax(h_w, h_{w+1})
Select a population p ∈ P such that pl_p^{w+} = 1
Update new solution S* by assigning pl_p^{w+} = 0 and pl_p^{w-}= 1
```

Algorithm 6: Operator 5

```
Given solution S, define the vector h of the harvest quantities
hw   for each week w ∈ W
Find (w, w + 1) ∈ W such that h_w > 0, h_{w+1} > 0 and h_w + h_{w+1} is
         minimized
Define w^- = argmin(h_w, h_{w+1}), w+ = argmax(h_w, h_{w+1})
for p ∈ P : pl_p^{w-}= 1 do
   if w^+ is a feasible harvest week for p then
      Update new solution S* by assigning pl_p^{w-} = 0 and pl_p^{w+}= 1
   else
     Find other feasible harvest week w_new for p such that

                      hw_new > 0
    Update new solution S* by assigning pl_p^{w-}=0 and pl_p^{wnew} = 1
```

Algorithm 7: Operator 6

Given solution S, define the vector h of the harvest quantities h_w for each week $w \in W$
if $\exists w \in W : h_w = 0 \wedge h_{w+1} > 0$ **then**
  for $p \in P : pl_p^{w+1} = 1$ **do**
   if $\exists\ w_{new}\ \ W \notin w, w+1 : w_{new}$ is a feasible harvest week for p **then**
    Update new solution S* by assigning $pl_p^{w+1} = 0$ and $pl_p^{wnew} = 1$

Algorithm 8: Operator 7

Draw integer r from set {1, 2,.., 10}. Define set R = {1,.., r}
for $n \in R$ **do**
  Given solution S, define the vector h of the harvest quantities h_w for each week $w \in W$
Select random week w' such that $h_{w'} > 0$
while $h_{w'} > \mathit{Capacity}$ do
  Select a population $p \in P$ such that $pl_p^{w'} = 1$
  Find week w_{new} in the set Wp of feasible harvest weeks of p such that
   $w_{new} = \mathrm{argmin}_{w \in Wp} \mid Capacity - (h_w + hq_p) \mid$
  Update new solution S* by assigning $pl_p^{w'} = 0$ and $pl_p^{wnew} = 1$

Algorithm 9: Operator 8

Calculate $M^+ = \sum_{w \in W} m_w^+$ and $M^- = \sum_{w \in W} m_w^-$
if $M^+ > M^-$ then
  Capacity ← Capacity + 100
else
  Capacity ← Capacity − 100

5. Results

The assumptions made in previous subsections about the performance of each model, for first approach, will be validated hereafter. Each model provided the input for the optimization model, in the form of GDUs predictions. We want to discuss the result of optimization model, I.e., which predictions gave the shortest harvest period and lowest deviation from the capacity, for each site and each scenario.

Results for Scenario 1, Site 0

The results of the optimization problem for Scenario 1 and site 0, using the MA, ARIMA and Holt Winters model, respectively, will be presented here after.

In Figure 24 with yellow and green colors, respectively, the initial and final harvest quantities distributions obtained with predictions made with MA models, are presented. Final solution shows that the algorithm has succeeded to create four periods with fairly constant weekly harvest quantities: 1) from week 19 to week 38, 2) from week 40 to week 44, 3) from week 57 to week 63 and 4) from week 66 to week 70. The algorithm managed to create three periods with weekly harvest quantities remarkably close to the capacity of 7000.

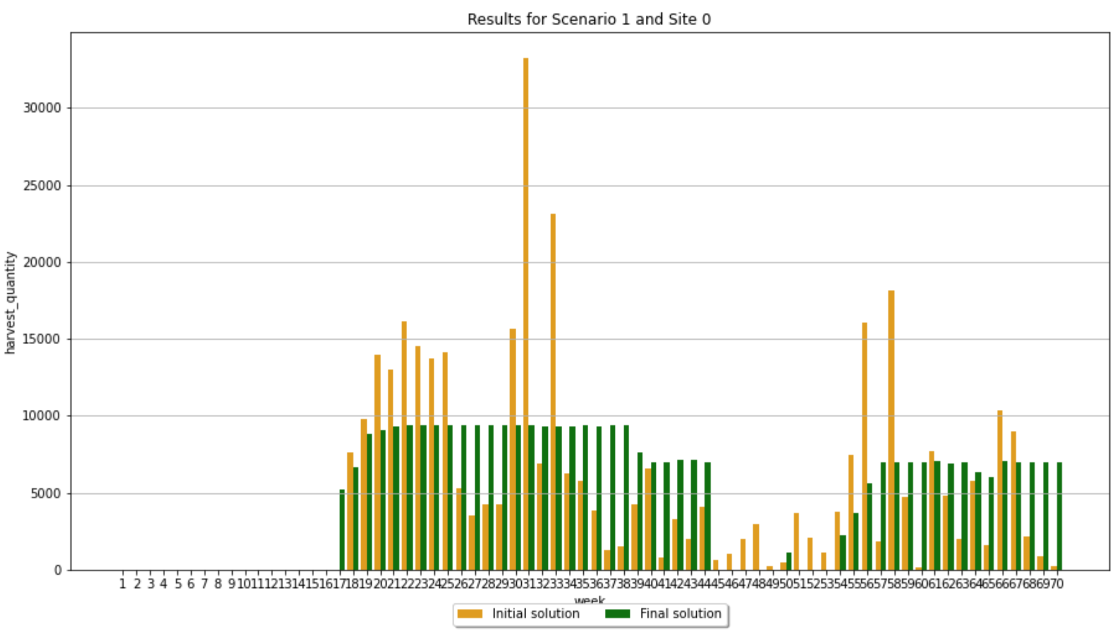

Figure 24: Results for Scenario 1 and Site 0-MA model

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	53	46
Deviation from the capacity	9,524,361	9,370,033
Deviation between consecutive weeks	261,044	38,474

Table 3: Objective functions results obtained with MA model

The results obtained with an ARIMA model as GDU predictor are shown in the Figure 25. Again, yellow, and green colors, respectively, represent the initial and final harvest quantities distributions.

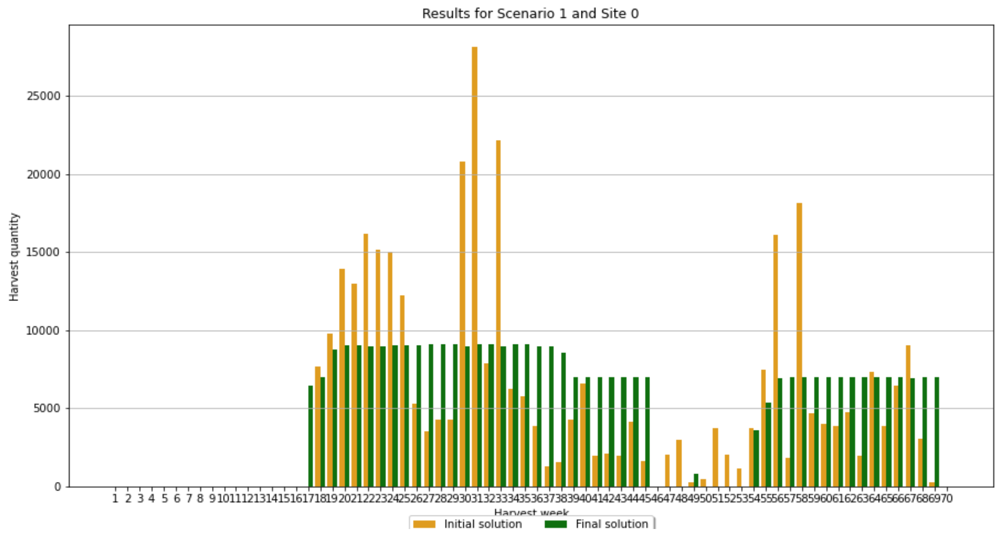

Figure 25: Harvest quantity results for Scenario 1 Site 0- ARIMA model

From the final solution, we can see that the algorithm has succeeded to create three periods with constant weekly harvest quantities: 1) from week 19 to week 38, 2) from week 39 to week 45, and 3) from week 56 to week 69. Optimization algorithm, with ARIMA GDUs predictions as input, managed to improve the results obtained with MA model. It extended the second period obtained by two weeks, and connected third and fourth period in one. The latter two periods have weekly harvest quantities very close to the capacity of 7000. In Table 1, objective function values for the initial and final solutions, are presented.

Objective function	Initial solution value	Final solution value
Number of harvest weeks	51	46
Deviation from the capacity	9 516 785	9 355 171
Deviation between consecutive weeks	225 228	34 738

Table 4: Objective functions values for initial and final solution-ARIMA

If we compare Table 3 and Table 4, we can see that ARIMA also decreased the deviation from the capacity as well as deviation between consecutive weeks.

Holt Winters algorithm also managed to make three periods with almost equal harvest quantity, those are: 1) from 19th to 36th week, 2) from 43rd to 45th week and 3) from 56th to 69th.

Optimization model can give longer periods of almost equal weekly harvest quantities if ARIMA predictions are used as input.

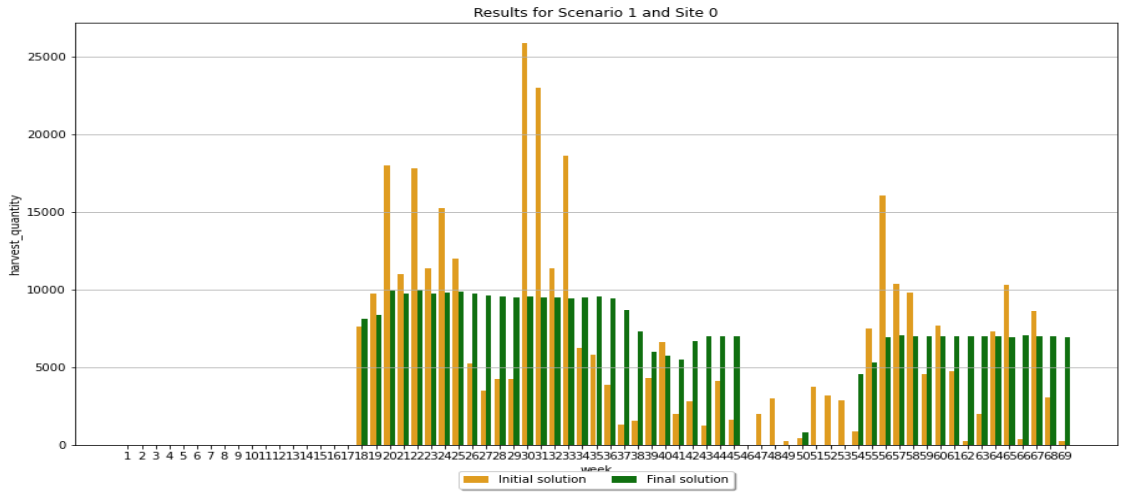

Figure 26: Harvest quantity results for Scenario 1 and Site 0, obtained with Holt-Winters model

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	51	45
Deviation from the capacity	9,514,073	9,374,465
Deviation between consecutive weeks	221,000	39,894

Table 5: Objective functions values for initial and final solution-Holt Winters model

Table 5 shows that optimization model based on the predictions made with Holt Winters model, manages to give the shortest harvesting period of 45 weeks. On the other hand, the surplus is bigger than the surplus that we got with alternative models, MA and ARIMA. To summarize, the Holt Winters model has to lower AIC than ARIMA model, but we can see that a less accurate predictions can lead to a falsely good solution.

Results for Scenario 1, Site 1

The MA, ARIMA and Holt Winters are inputs for the optimization model. Here, the output of optimization model for Scenario 1 and Site 1 is presented and compared, after each of stated algorithm provided the input for the model.

Figure 27, presents the initial (represented by the yellow bars) and final (green bars) harvest quantities distributions, corresponding to the initial and selected solutions, respectively. These results are obtained as output of optimization model, which has as an input, prediction made with MA model. We can notice that for this site, the algorithm was not as successful as for Site 0. Nevertheless, the results demonstrated the algorithm tendency towards equalizing the weekly harvest amounts, which are still far from 7000.

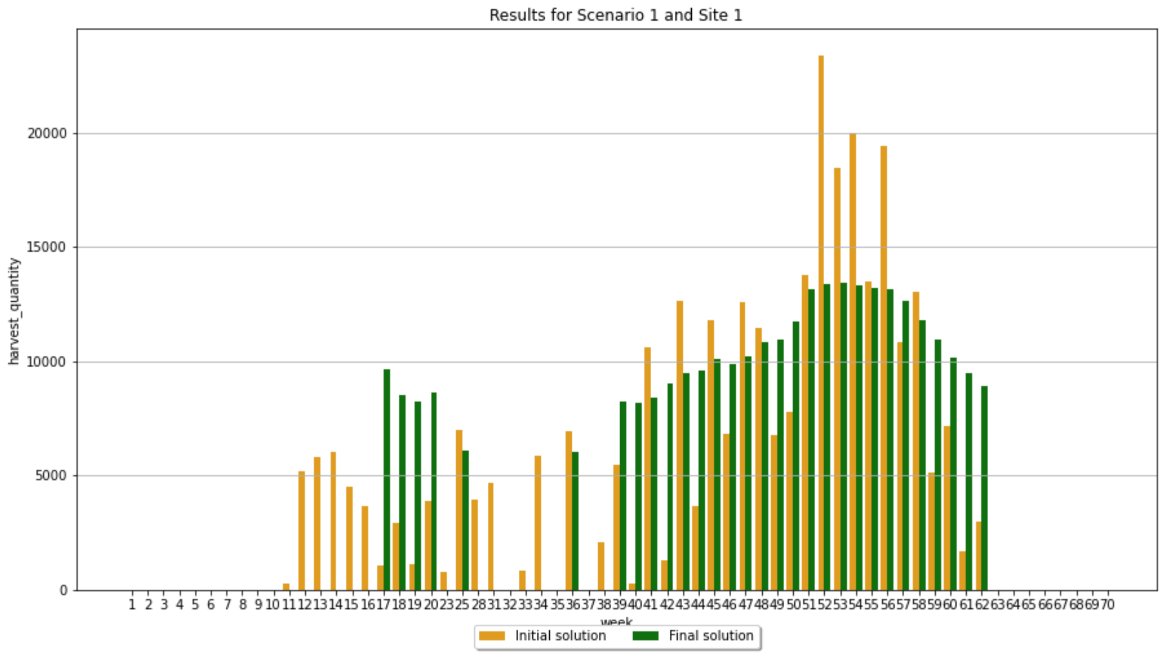

Figure 27: Harvest Quantity output for Scenario 1 and Site 1 –MA model

Table 6 denotes objective function values for the initial and final solutions.

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	42	30
Deviation from the capacity	7,096,757	7,111,339
Deviation between consecutive weeks	221,212	71,724

Table 6: Objective function values for initial and final solution- MA model

GDUs predictions made with ARIMA model, provided the results presented in Figure 28 and Table 7.

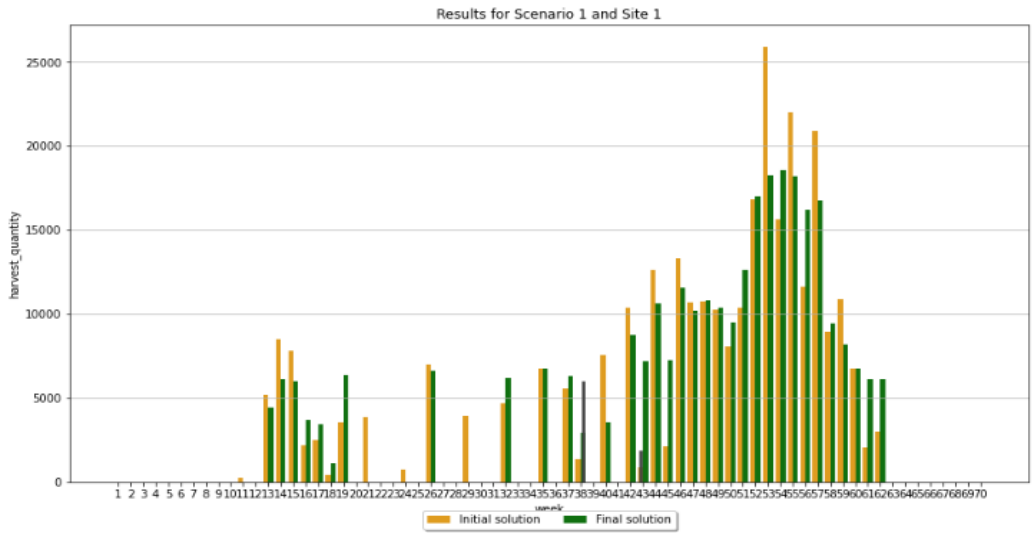

Figure 28: Harvest quantity results for Scenario 1 Site 1-ARIMA

The results demonstrated the improvement in algorithms tendency towards equalizing the weekly harvest amounts, and amounts are closer to 7000 units. On the other hand, the harvest period is longer and we have slightly bigger deviation from the capacity than those obtained with MA model predictions. It can also be noticed that the initial solution is better, but the algorithm did not manage to preserve that improvement till the end.

Objective function	Initial solution value	Final solution value
Number of harvest weeks	38	34
Deviation from the capacity	7 120 205	7 090 845
Deviation between consecutive weeks	242 630	135 486

Table 7: Objective function values for the initial and final solution-ARIMA

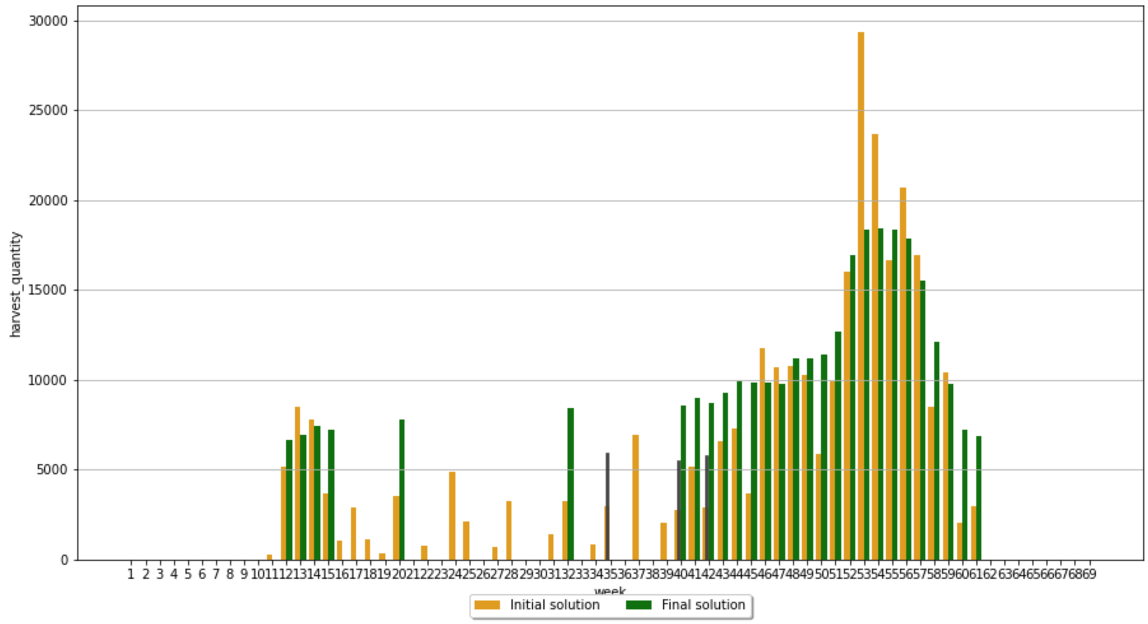

Figure 29: Harvest quantity for Scenario 1 and Site 1-Holt Winters

Figure 29 shows the initial (represented by the yellow bars) and final (green bars) harvest quantities distributions, that represent the output of optimization model, made with Holt Winters GDUs predictions. It can be noticed that the algorithm performed better in the case of Site 0. Nevertheless, comparing the results of optimization model obtained with predictions made with Holt Winters model with results obtained with alternative models, we can notice higher tendency towards equalizing the weekly harvest quantities. That can be noticed from some periods of almost equalized harvest quantities, which do not exist in other results.

Additionally, the smallest number of harvesting weeks is obtained-28. However, optimization model made a slightly greater surplus. The results are presented in Table 8.

Objective functions	Initial solution value	Final Solution value
Number of harvest weeks	43	28
Deviation from the capacity	7,106,417	7,135,339
Deviation between the consecutive weeks	169,638	85,086

Table 8: Objective functions values for initial and final solution-Holt Winters

Results for Scenario 2, Site 0

The results of optimization model for Scenario 2 and Site 0 using the MA, ARIMA and Holt Winters model are presented in this section.

In Figure 30, the initial and final harvest quantities distributions for Scenario 2 and Site 0 are presented, in yellow and green colors, respectively. With MA GDUs predictions, optimization model has not managed to equalize the weekly harvest quantities. Regarding the additional objective from Scenario 2, i.e., finding the minimal storage capacity, the result showed that the current capacity performs best. The found solutions with increased capacity only increased the harvest amount deviation from the capacity.

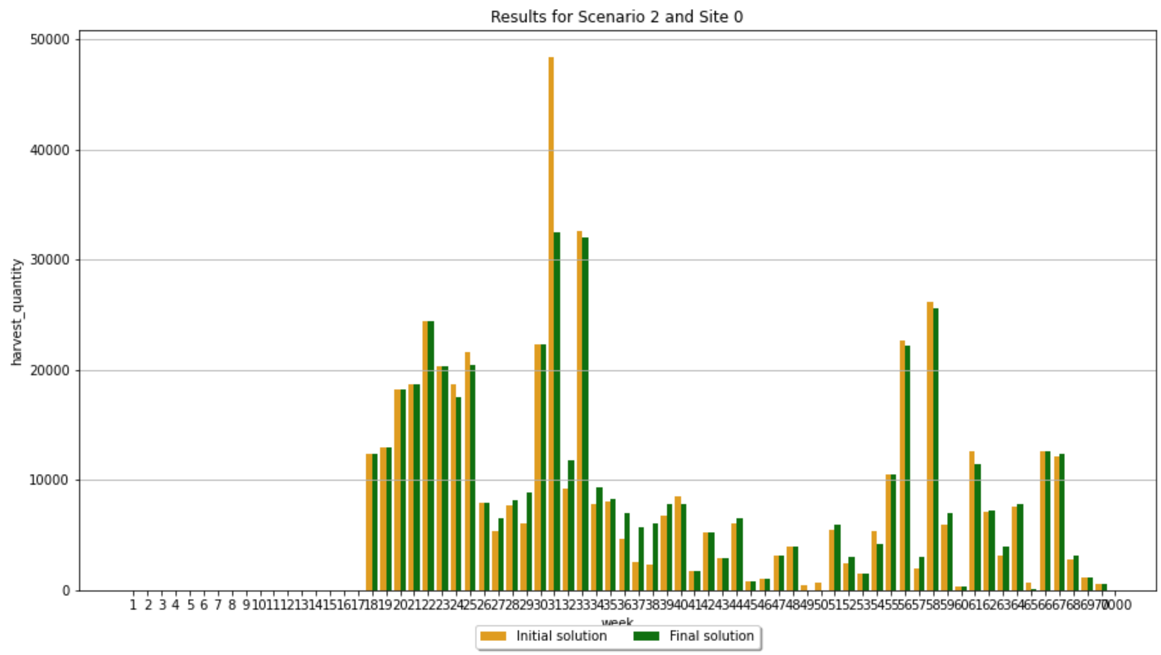

Figure 30: Harvest Quantity output for Scenario 2 and Site 1 –MA model

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	53	51
Deviation from the capacity	9,601,034	9,537,712
Deviation between consecutive weeks	381,094	325,068
Minimum storage capacity	7000	7000

Table9: Objective functions values for Initial and final solution- MA model

On the other hand, with the predictions made with the ARIMA model, results are improved. The optimization algorithm has managed to equalize the weekly harvest quantities, especially between 20th and 37th harvest week. Again, the result showed that the current capacity performs best. The found solutions with increased capacity only increased the harvest amount deviation from the capacity.

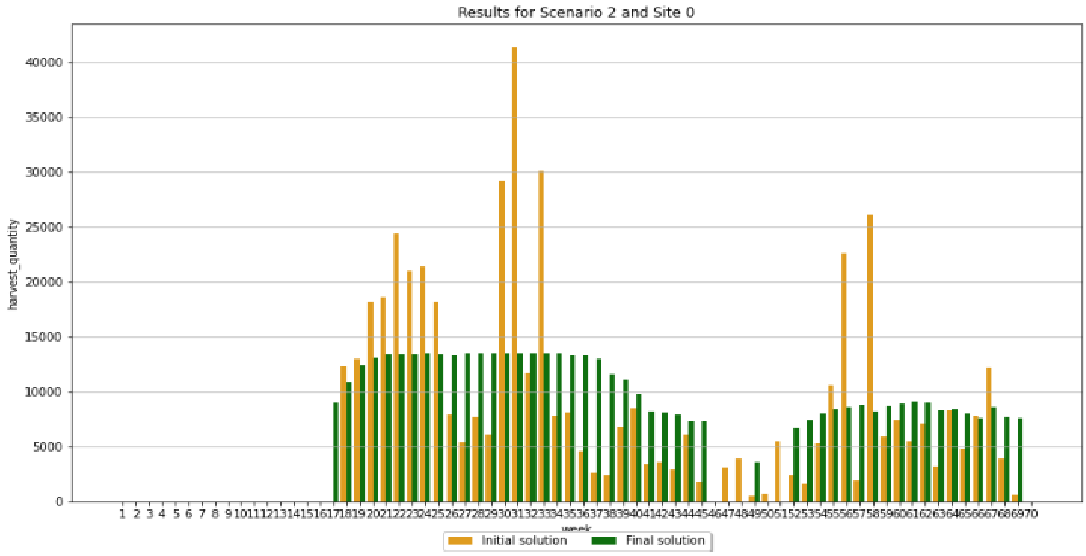

Figure 31: Harvest quantity output for Scenario 2 Site 0- ARIMA model

In Table 10 objective function values for the initial and final solutions, are presented. Comparing these results with the results of optimization model obtained with MA predictions, we can conclude that the algorithm managed to shorten the harvesting period, with noticeable decrement in deviation between consecutive weeks.

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	51	48
Deviation from the capacity	9,582,540	9,457,250
Deviation between consecutive weeks	321,472	56,136
Minimum storage capacity	7000	7000

Table 10: Objective function values for initial and final solution- ARIMA model

The predictions made with Holt Winters model, have not improved the result in case of equalizing weekly harvest quantities, there are no periods with equal harvest quantity. Algorithm minimized additional objective, and set storage capacity to 3000. However, from Figure 32, can be concluded that capacity is usually far from 3000.

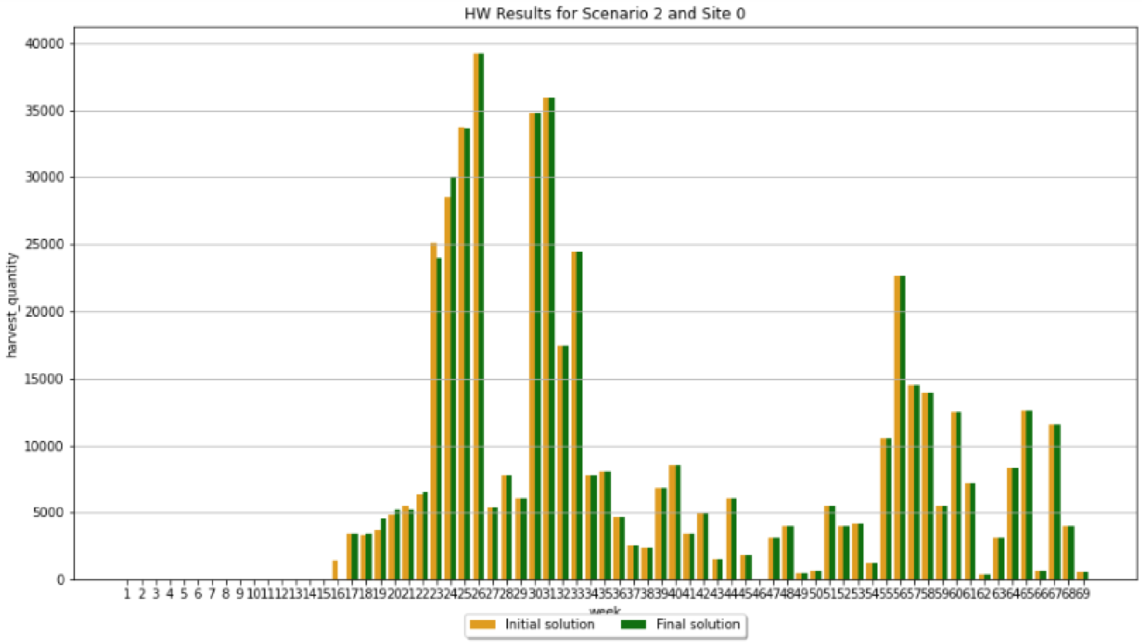

Figure 32: Harvest quantity output for Scenario 2 Site 0- Holt Winters model

Surprisingly, deviation from capacity is decreased by almost 50% comparing to other two models. However, deviation between consecutive weeks is still very high.

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	53	52
Deviation from the capacity	4,342,266	4,345,200
Deviation between consecutive weeks	303,098	303,040
Minimum storage capacity	3000	3000

Table 11: Objective function values for initial and final solution- Holt Winters model

The predictions made with ARIMA model, gave the shortest harvest period with most evenly distributed harvest quantities. On the other hand, Holt Winters prediction provided the lowest deviation from total capacity, but harvest quantity is not uniformly distributed, there are a lot of fluctuations between consecutive week. Deviations between consecutive weeks are not significantly changed.

Results for Scenario 2, Site 1

The MA, ARIMA and Holt Winters provided the input for optimization model in Scenario 2 and Site 1, once more. The outputs are presented and discussed. As in previous case, the predictions made with MA did not provide a tendency of optimization model toward equalizing the weekly harvest amount. Also, from Figure 32 we can see that its weekly amounts are extremely far from 6000, which is the minimum storage capacity. As in Scenario 2 for Site 0 the best results are obtained if we keep the current capacity. All of this is presented in Figure 32 and Table 12.

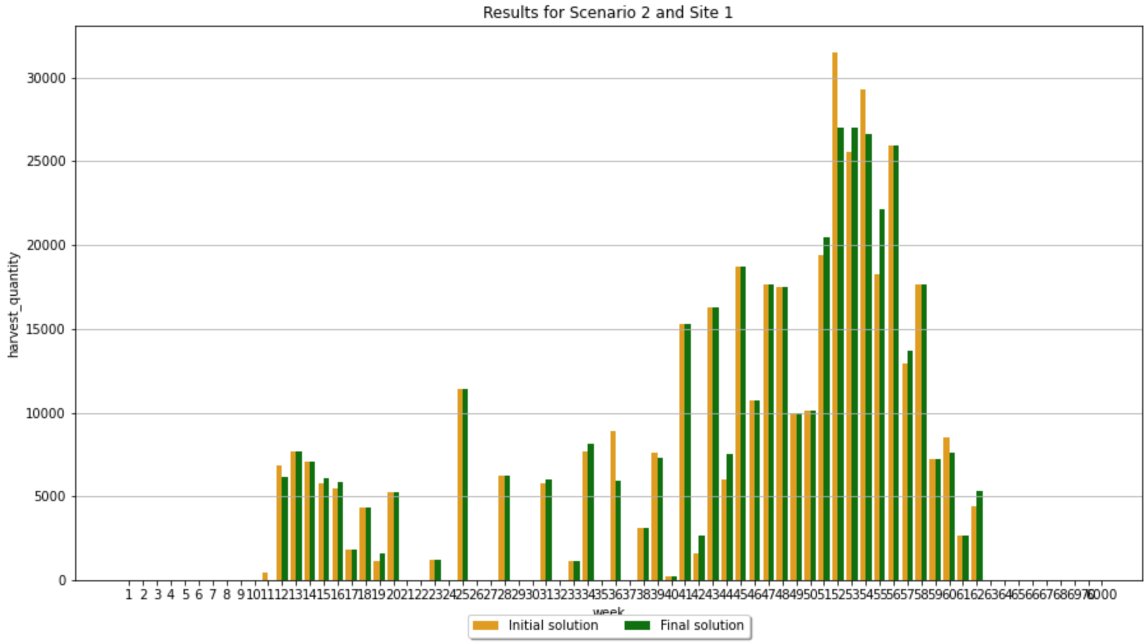

Figure 33: Results for Scenario 2 and Site 1 obtained with MA algorithm

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	42	41
Deviation from the capacity	7,177,404	7,172,690
Deviation between consecutive weeks	306,436	269,712
Minimum storage capacity	6000	6000

Table12: Objective function values for Initial and Final solution-MA model

The ARIMA predictions managed to improve the results. The oscillations in the weekly harvest amounts are reduced, compared to the initial planting schedule. Also, the algorithm managed to reduce the number of isolated harvest weeks, which are those weeks that are preceded and succeeded by weeks with no harvest, such as weeks 21 and 24 in the original solution. Additionally, algorithm shortened the harvesting period, and reduced both the capacity and weekly deviation. Finally, the weekly harvest amounts are closer to 6000 units, which is again chosen as minimum storage capacity. All of this is presented in Figure 34 and Table 13.

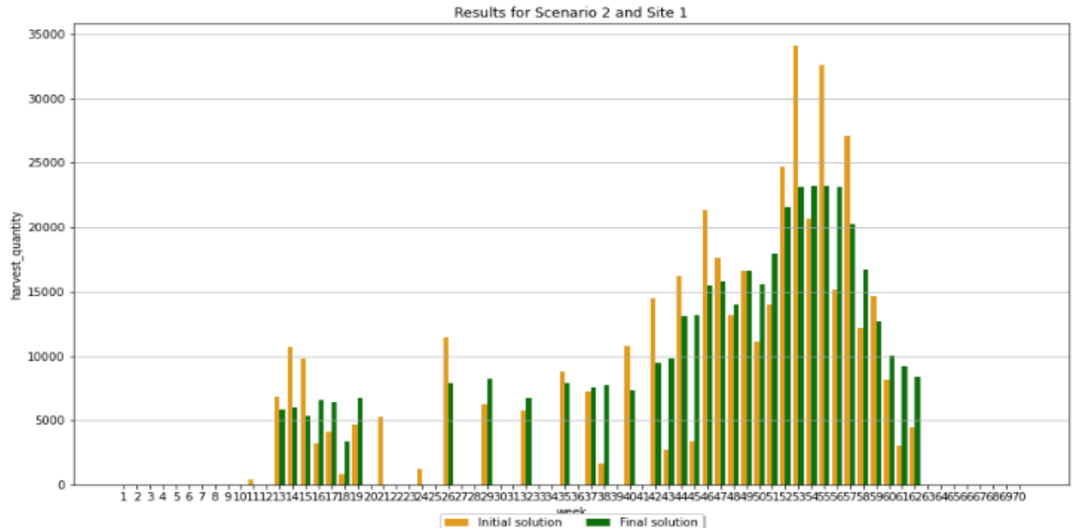

Figure 34: Harvest quantity results for Scenario 2 Site 1-ARIMA model

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	38	35
Deviation from the capacity	7,209,088	7,177,366
Deviation between consecutive weeks	345,422	165,344
Minimum storage capacity	6000	6000

Table 13: Objective function values for initial and final solutions-ARIMA model

Predictions made with Holt Winters model, have not helped the optimization model to make shorter harvest period. However, it can be noticed that there is higher tendency towards equalizing weekly harvest amount.

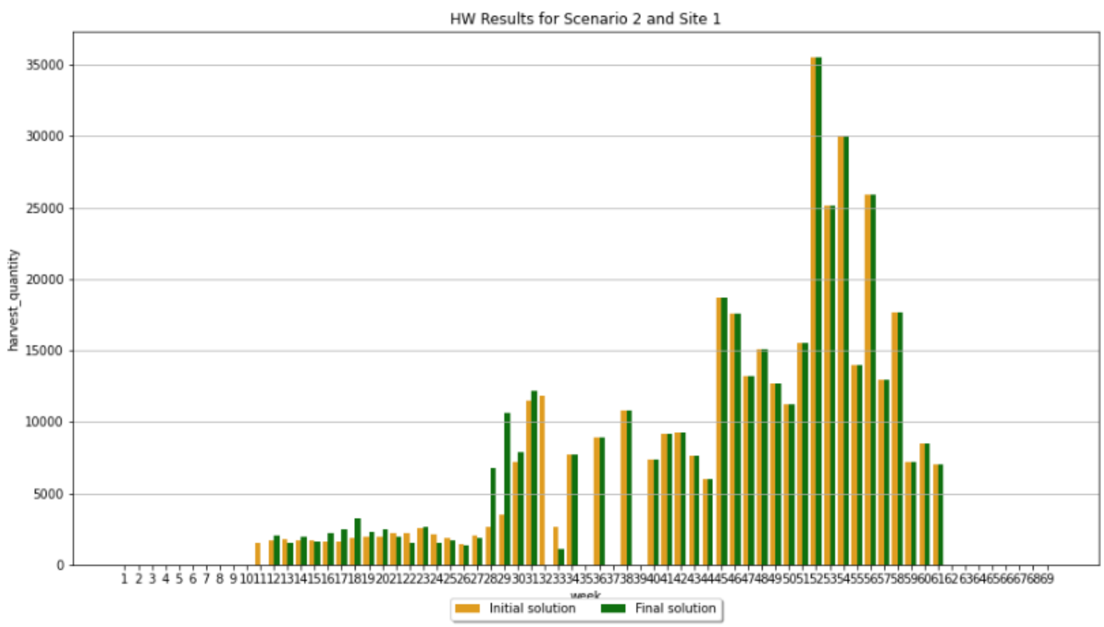

Figure 35: Harvest quantity results for Scenario 2 Site 1- Holt Winters model

Comparing the values of objective functions in Table 12, Table 13, and Table 14, it is obvious that, with Holt Winters predictions, optimization model managed to give 7 times smaller deviation from total capacity. On the other hand, algorithm have not managed to do the same in case of deviation between consecutive weeks.

Objective functions	Initial solution value	Final solution value
Number of harvest weeks	48	46
Deviation from the capacity	1,963,776	1,969,910
Deviation between consecutive weeks	217.926	235,262
Minimum storage capacity	1400	1400

Table 14: Objective function values for initial and final solutions-Holt Winters model

In case of this site and scenario, ARIMA predictions provided shortest harvesting period. However, in the case of minimizing the surplus, the predictions made with Holt Winters model provided the best results.

5.1. Results summarization and discussion

In the Scenario 1 Site 0, the best solutions of the optimization model are obtained with ARIMA predictions. The Holt Winters model gave the shorter harvesting period, by one week, but surpluses are bigger. Optimization model was not that successful in Scenario 1 Site 0, however models' tendency towards equalizing weekly harvest quantity is showed. The predictions of Holt Winters model provided the shortest harvesting period.

In Scenario 2, with no predefined capacity, predictions made with ARIMA model provided the shortest harvest period, for both sites. Additionally, optimization model managed to keep deviations between consecutive weeks lower than those obtained with other two models. With predictions made with Holt Winters model, optimization model significantly decreased the deviation from total capacity.

Table 13 shows for each prediction model the best results of the optimization model that it achieved, as well as the AIC values of that model. The best solution

	MA model	ARIMA model	Holt Winters model
AIC value	4766	4764	-6034
Number of weeks	46	46	45
Deviation from capacity	9.370.033	9.355.171	9,374.465
Deviation between consecutive weeks	38.474	34.738	39894
Capacity	7000	7000	7000

Table 15: Best Results

6. Conclusion

The Crop Planting Scheduling Problem is defined and solved using ALNS (Adaptive Large Neighborhood Search) meta-heuristic. It is discovered that output of the model, I.e., weekly harvest amount, deviation from total capacity and number of harvesting weeks, highly depends on GDUs predictions. The purpose of this thesis was to determine the most precise model for GDU prediction and investigate how sensitive the Optimization model is to different prediction models.

For GDU predictions we used three different models: *Moving Average* (MA), *Autoregressive Integrated Moving Average* (ARIMA) and *Holt Winters* model. For each model separately, hyperparameters are tuned and model with the lowest AIC score is considered as the best model and it is further used for an out-of-sample predictions.

All the models are tested in two different approaches. In the first approach accumulated daily GDUs for period 2009-2019 were given, and the task was to predict daily GDUs for the next two years. GDUs are calculated using minimum and maximum daily temperatures, therefore in the second approach the task was to predict daily temperatures, and then to calculate the GDUs.

In the case of second approach, we decided not to validate the results, since the AIC score were high. The imprecise predictions will not lead to significant results of optimization model.

Results of the first approach are validated through optimization model, on site 0 and site 1, in two different scenarios, with predefined capacity and without predefined capacity.

In the first approach, with MA model predictions, optimization model in majority of cases gives the greatest surplus and longest harvesting period. ARIMA predictions managed to improve the results of optimization model. Harvesting period is shorter, weekly harvest quantities are distributed more equally and they are closer to capacity, surpluses are decreased.

In most of the cases, result obtained with Holt Winters predictions are very close to those obtained with ARIMA predictions. Despite the fact that Holt Winters has lower AIC, the less precise ARIMA predictions, provided the better results of optimization problem. Hence, by selecting a less precise GDU prediction model, we might obtain falsely good solutions of optimization problem. This makes the choice of a model for predicting GDUs, extremely important.

Bibliography

[1] Saiara S, Guiping H, *Optimizing Crop Planting Schedule Considering Planting Window and Storage Capacity*. Front. Plant Sci., 2022

[2]Agweb: Who Produces What? Key Agriculture Stats from Around the Globe (2022). URL: https://www.agweb.com/markets/world-markets/who-produces-what-key-agriculture-stats-around-globe . Accessed: 2022-05-26

[3] Syngenta: Syngenta crop challenge in analytics 2021 (2021). URL https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php. Accessed: 2022-03-04

[4] Obrenović N., Ataç S., Bortolomiol S., *An ALNS-based heuristic for the Crop Planting Schedule Problem*. Zenodo, 2022, https://doi.org/10.5281/zenodo.7234591

[5] Obrenović N., Ataç S., Bortolomiol S., Brdar S., Marko O., Crnojević V., *The Crop Plant Scheduling Problem,* 2022.

[6] Ansafari J., Akhavizadegan F., Wang L., *Scheduling Planting Time Through Developing an Optimization Model and Analysis of Time Series Growing Degree Units,* 2022.

[7] Khalilzadeh Z., Wang L., A *MILP Model for Corn Planting and Harvest Scheduling Considering Storage Capacity and Growing Degree Units,* 2021

[8] Günder M., Piatkowski N., Von Rueden L., Sifa R., Bauckhage *C., Towards Intelligent Food Waste Prevention: An Approach Using Scalable and Flexible Harvest Schedule Optimization with Evolutionary Algorithms, 2021.*

[9]Machine Learning Mastery: *How to Decompose Time Series Data into Trend and Seasonality (2020)* URL: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/. Accessed: 2017-01-30

[10] Brockwell P., Davis R., *Introduction to Time Series and Forecasting,* https://doi.org/10.1007/978-3-319-29854-2 *,Springer Nature Switzerland AG, 2016*

[11]Brown, R.G., Meyer, R. F., *The fundamental theory of exponential smoothing*, Operations Research, 9, 1961, p. 673-685

[12] Bevans, R. (May 25, 2022). *Akaike Information Criterion | When & How to Use It (Example)*. Scribbr. Retrieved October 11, 2022, from https://www.scribbr.com/statistics/akaike-information-criterion/

[13] Bierlaire M., *Optimization: Principles and Algorithms,* EPFL Press, 2018

[14] Meteostat: *https://www.meteostat.net*.

[15] Google Colaboratory: https://colab.research.google.com/

[16] NDAWN: Corn Growing Degree Days Information URL: https://ndawn.ndsu.nodak.edu/help-corn-growing-degree-days.html

[17]otexts: Moving Average Models, URL: https://otexts.com/fpp2/MA.html

Biography

Dragana Šorak was born on 24th of February 1997
in Sremska Mitrovica, Serbia.
She attended elementary school *"Petar Kočić"*
and grammar school *"Gimnazija Inđija"* in Inđija.

In 2015 she started a B*achelor of Teaching in Mathematics at Faculty of Sciences, University of Novi Sad*
and finished in 2020 with a GPA of 8.44. In the same year she continued with master studies of *Data Science*
at the same faculty and passed all exams in 2022 with GPA of 9.23.

	UNIVERZITET U NOVOM SADU ● Prirodno-matematički fakultet 21000 Novi Sad, Trg Dositeja Obradovića 3
	KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj, RBR:	
Identifikacioni broj, IBR:	
Tip dokumentacije, TD:	Monografksa dokumentacija
Tip zapisa, TZ:	Tekstualni štampani materijal
Vrsta rada, VR:	Master rad
Autor, AU:	Dragana Šorak
Mentor, MN:	Dr Nikola Obrenović
Naslov rada, NR:	Optimizacija rasporeda setve kultura pod uticajem predikcije aktivnih temperatura
Jezik publikacije, JP:	Engleski
Jezik izvoda, JI:	Engleski
Zemlja publikovanja, ZP:	Republika Srbija
Uže geografsko područje, UGP:	Vojvodina
Godina, GO:	2022.
Izdavač, IZ:	Autorski reprint
Mesto i adresa, MA:	Novi Sad, Departman za matematiku i informatiku, PMF, Trg Dositeja Obradovića 2
Fizički opis rada, FO:	Poglavlja (6), strana (52), literaturnih citata (17), tabela (15), slika (35), grafikona (0), priloga (0)
Naučna oblast, NO:	Matematika
Naučna disciplina, ND:	Primenjena matematika
Ključne reči, KR:	Predviđanje vremenskih serija, Optimizacija rasporeda setve, Aktivne temperature, Moving Average model, ARIMA model, HoltWinters model

Univerzalna decimalna klasifikacija, UDK:	
Čuva se, ČU:	Biblioteka Departmana za matematiku i informatiku Prirodnomatematičkog fakulteta, u Novom Sadu
Važna napomena, VN:	
Izvod, IZ:	U ovom radu istražujemo kako različiti modeli za predviđanje aktivnih temperatura utiču na optimizaciju rasporeda setve, koja se ogleda u broju nedelja žetve i nedeljnoj količini žetve. Za to su koriščena dva pristupa i tri modela za predikciju vremenskih serija.
Datum prihvatanja teme od strane NN veća, DP:	07.06.2022.
Datum odbrane, DO:	28.10.2022.
Članovi komisije, KO: Predsednik: Mentor: Član:	 dr Oskar Marko, naučni saradnik Instituta BioSense, Novi Sad dr NikolaObrenović, naučni saradnik Instituta BioSense, Novi Sad Prof. Dr Dušan Jakovetić, vanredni profesor, PMF, Novi Sad

	UNIVERSITY OF NOVI SAD ● Faculty of Science 21000 Novi Sad, Trg Dositeja Obradovića 3
	KEY WORDS DOCUMENTATION

Accession number, ANO:	
Identification number, INO:	
Document type, DT:	Monograph type
Type of record, TR:	Textual material, printed
Contents code, CC:	Master thesis
Author, AU:	Dragana Šorak
Mentor, MN:	Dr Nikola Obrenović
Title, TL:	Optimization of Crop Planting Schedule Based on the Different Growing Degree Units Forecasting Models
Language of text, LT:	English
Language of abstract, LA:	English
Country of publication, CP:	Republic of Serbia
Locality of publication, LP:	Vojvodina
Publication year, PY:	2022.
Publisher, PU:	Autor's reprint
Publ. place, PP:	Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics, Trg Dositeja Obradovića 2
Physical description, PD:	Chapters (6), pages (52), references (16), tables (15), figures (35), graphs (0), additional lists (0)
Scientific field, SF:	Mathematics
Scientific discipline, SD:	Applied mathematics
Key words, KW:	Time series prediction, Crop Planting Scheduling Problem, Growing Degree Units, MA model, ARIMA model, Holt Winters model

Universal decimal classification, UDC:	
Holding data, HD:	Department of Mathematics and Informatics' Library, Faculty of Science, Novi Sad
Note, N:	
Abstract, AB:	In this thesis we wanted to investigate how different GDUs predictions impact the optimized planting schedule, expressed through the number of harvest weeks and quantities. For that we used two approaches and three models for time series prediction.
Accepted by the Scientific Board on, ASB:	07.06.2022.
Defended on, DE:	28.10.2022.
Thesis defend board, DB: Chairperson: Mentor: Member:	 dr Oskar Marko, scientific associate of BioSense Institute, Novi Sad dr Nikola Obrenović, scientific associate of BioSense Institute, Novi Sad Prof. Dr Dušan Jakovetić, professor at PMF, Novi Sad