Evaluating Evapotranspiration By Using Data Mining Instead of The Physical Based Model In Remote Sensing

Lamya Neissi

Shahid Chamran University of Ahvaz

Mona Golabi

Shahid Chamran University of Ahvaz

Mohammad Albaji

Shahid Chamran University of Ahvaz <https://orcid.org/0000-0002-5483-5834>

Abd Ali Naseri (✉ abdalinaseri@scu.ac.ir)

Shahid Chamran University of Ahvaz

Research Article

Keywords: GIS, M5 decision tree, Remote Sensing, SEBAL.

Posted Date: August 27th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-707736/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Theoretical and Applied Climatology on November 16th, 2021. See the published version at <https://doi.org/10.1007/s00704-021-03822-7>.

Evaluating evapotranspiration by using data mining instead of the physical based model in remote sensing

Lamya Neissi¹, Mona Golabi², Mohammad Albaji^{*3}, Abd Ali Naseri⁴

¹Phd Student of Irrigation and Drainage, Department of Irrigation and Drainage, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran. E-mail: neissi.lamya@gmail.com

²Assistance Professor, Department of Irrigation and Drainage, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran
Email: mona_golabi@yahoo.com, m.golabi@scu.ac.ir

^{*3} Associate Professor, Department of Irrigation and Drainage, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
Corresponding author email: m.albaji@scu.ac.ir; m_albaji2000@yahoo.co.uk

⁴ Full Professor, Department of Irrigation and Drainage, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
Email: abdalinaseri@scu.ac.ir

Abstract

Precise calculations for plant water requirements and evapotranspiration is very crucial in determining the volume of water consumption for plant production. In order to estimate evapotranspiration in the extended area, different remote sensing algorithms required many climatological variables. Climatological variable measurements will cover small limited areas which can cause an error in extended areas. By using data mining and remote sensing, the evapotranspiration process can be modeled. In this research, the physical-based SEBAL evapotranspiration algorithm was modeled by M5 decision tree equations in GIS. Input variables of the M5 decision tree consisted of albedo, emissivity, and Normalized Difference Water Index (NDWI) which are represented as absorbed light, transformed light, and plant moisture, respectively. After extracting the best equations in the M5 decision tree model for 8 April 2019, these equations were modeled in GIS by using python scripts for 8 April 2019 and 3 April 2020. The calculated correlation coefficient (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for 8 April 2019 were 0.92, 0.54, and 0.42 and for 3 April 2020 were 0.95, 0.31, and 0.23, respectively. Also, sensitivity and uncertainty analysis were considered for more model evaluation. Those analysis revealed that evapotranspiration is sensitive to albedo more than the two other model inputs and the estimated evapotranspiration obtained by data mining is in acceptable range of certainty.

Keyword: GIS; M5 decision tree; Remote Sensing; SEBAL.

1. Introduction

These days precision agriculture has become more important and is applicable in most sections of agriculture, especially in the irrigation section. By using different methods (hardware and software) and high technologies, agricultural activities could be done with more precision. Different hardware methods including of pressurized irrigation systems, drones, and the internet of things for crop monitoring. Software methods, in this case, consist machine learning, deep learning, and data mining especially those regarding the decision trees.

Irrigation scheduling of crops can be done by using meteorological data for evapotranspiration calculation. By using satellite images and different algorithms, evapotranspiration can be estimated in an extended area and reach an accurate irrigation scheduling (Jaferian et al., 2019; Song et al., 2018; Goodarzi and Eslamian, 2018; Diarraa et al., 2017; Colaizzi et al., 2017; Anderson et al., 2012; Abdolhosseini et al., 2012).

Evapotranspiration estimation is a complicated process due to many independent parameters. Evapotranspiration is a dependent parameter which can be affected by many climatological parameters and crop conditions. For estimating evapotranspiration, different equations were obtained which can be used in different conditions such as FAO-Penman–Monteith, Blaney-Criddle, etc. Ground observations represent the results for one specific point in which high accuracy is needed to generalize them for extended region. Hence evaporation is different from one station to another. By using remote sensing technologies, one can reach acceptable and high accuracy for a specific extended region. By using satellite images as a remote sensing technique, ground observations (hard data) transformed to soft data. Among different methods of data

mining, the M5 decision tree was used to estimate the evapotranspiration in an extended area (Gibert et al., 2018).

This research is intended to establish an applicable different linear relation by using the M5 decision tree between independent remote sensing parameters (albedo, emissivity, and Normalized difference water index) with the dependent parameter (evapotranspiration) by using data mining which is the most important innovation of this research.

Landsat8 satellite images and SEBAL algorithm were used for evapotranspiration estimation which was used in many evapotranspiration estimations, and acceptable results were obtained by (Mhawej et al., 2020; Elnmer et al., 2019; Kong et al., 2019; Ochege et al., 2019; Gobbo et al., 2019; Kamali and Nazari, 2018).

One of input parameters for evapotranspiration estimation is Land surface temperature (LST). but the spatial resolution of this band is 100m. The estimated evapotranspiration image by using SEBAL algorithm through the use of this band has 100m spatial resolution, which is the innovation of this research to enhance spatial resolution by using the M5 decision tree. Input parameters have 30m spatial resolution and by applying the gained equations by the M5 decision tree, an evapotranspiration map with higher spatial resolution can be obtained.

According to sugarcane plantation in an extended area in the southwest of Iran (more than 94000 ha), an extremely high volume of water is consumed in this section, so spatially enhancing evapotranspiration estimation image, irrigation water scheduling can be calculated more precisely.

88

89   **2.   Materials and methods**

90   **2.1. Study area**

91   This study was conducted in Amir-Kabir Agro-Industry Sugarcane fields located in southwest of
92   Iran (Figure 1). The soil texture is clay-loam and annual average evapotranspiration for 20 years
93   was 3331.812 mm. The total area of Amir-Kabir agro-industry is over 17000 ha with a 14000 ha
94   cultivated area. Each farm has a 25ha area with a low-pressure hydro flume irrigation system and
95   a subsurface drainage system with a 40m distance and 1.8 m depth for each drain tile. The total
96   irrigation water consumption is 3000 mm and the peak of irrigation water was applied in July.

97

98   **2.2. Landsat8 satellite**

99   Landsat8 OLI satellite images were the main data for remote sensing processes
100   (<http://glovis.usgs.gov>). Thermal bands have lower resolution compared to other optic bands. As
101   for Landsat8, band10 image represent a thermal band that provides less spatial resolution
102   (100m). Thermal bands are critical for evapotranspiration estimation and Landsat8 has the most
103   appropriate thermal band for agricultural evapotranspiration estimation in a great variety
104   extended regions. Amir-Kabir agro-industry plantation with area of 14000 ha is extended enough
105   for evapotranspiration calculations by using remote sensing images.

106

107   **2.3. Ground measurements**

ET estimation requires meteorological data. Therefore, these data were obtained from Amir-Kabir agro-industry plantation local weather station. Weather data including Max and Min temperature, the relative percentage of humidity, wind speed, and sunshine hours were used for evapotranspiration calculations and Ref-et software was used for reference evapotranspiration calculations.

2.4. SEBAL algorithm

SEBAL algorithm is known as one source algorithm for evapotranspiration calculation and in this study SEBAL algorithm is used for calculating evapotranspiration of the sugarcane fields of Amir-Kabir agro-industry plantation. Energy balance equation is the main equation of this algorithm. Equation1 represents the main energy balance equation:

$$\lambda ET = R_n - G - H \quad (1)$$

Where λET latent heat flux (W/m^2), R_n is the net radiation at the surface (W/m^2), G is the soil heat flux (W/m^2) and H is the sensible heat flux to the air (W/m^2) (Bastiaanssen et al., 1998).

The net sun radiation (R_n) is a balance equation between the incoming and outgoing short and long wave components as shown in equation (2) (Bastiaanssen et al., 1998):

$$R_n = (1 - \alpha)R_{s\downarrow} + R_{L\downarrow} - R_{L\uparrow} - (1 - \epsilon_0)R_{L\downarrow} \quad (2)$$

Where $R_{s\downarrow}$ is incoming shortwave radiation (W/m^2), $R_{L\downarrow}$ is incoming longwave radiations emitted by the atmosphere (W/m^2), $R_{L\uparrow}$ is out coming longwave radiations (W/m^2), α is surface albedo also known as reflection coefficient and ϵ_0 is the broadband surface emissivity.

126 The incoming shortwave radiation $R_{s\downarrow}$ is estimated from the radiation received at the top of the
 127 atmosphere as (Allen et al., 2002):

$$R_{s\downarrow} = G_{sc} \times \cos\theta \times dr \times \tau_{sw} \quad (3)$$

128 Where G_{sc} is the solar constant values $1367 \text{ (W/m}^2\text{)}$, and $\cos\theta$ is the cosine of solar Zenith
 129 angle, dr is the inverse relative squared distance of Earth to Sun and τ_{sw} is atmospheric
 130 transparency factor.

131 Stephan-Boltzmann equation is used to calculate the incoming long wave radiation, emitted by
 132 the atmosphere (equation 4) (Allen et al., 2002):

$$R_{L\downarrow} = \varepsilon_a \times \sigma \times T_a^4 \quad (4)$$

133 Where ε_a is the atmospheric emissivity (dimensionless), σ is the Stefan-Boltzmann constant
 134 ($5.67 \times 10^{-8} \text{ W/m}^2/\text{K}^4$), and T_a is the air temperature in K.

135 The Stephan-Boltzmann equation is used to estimate the long wave radiation emitted by the
 136 surface (Allen et al., 2002):

$$R_{L\uparrow} = \varepsilon_0 \times \sigma \times T_s^4 \quad (5)$$

137 Where ε_0 is the broadband surface emissivity (dimensionless), σ is the Stefan-Boltzmann
 138 constant ($5.67 \times 10^{-8} \text{ W/m}^2/\text{K}^4$), and T_s is the surface temperature in K. Calculation, the amount
 139 of the surface temperature of the Earth has used equation 6:

$$T_s = \frac{K_2}{\ln\left(\frac{\varepsilon_{NB} \times K_1}{\rho_{10}} + 1\right)} \quad (6)$$

Where K_1 and K_2 are the constant coefficients of Landsat 8 image equal to 774.88 and 1321.0789, $e_{NB}=1$ (Allen et al., 2011) and ρ_{10} is thermal correction band (band 10) of Landsat 8 image.

The amount of albedo coefficient is given by equation 7.

$$\alpha = \frac{\alpha_{toa} - \alpha_{path-radiance}}{\tau_{sw}} \quad (7)$$

Where; $\alpha_{path-radiance}$ is the average portion of the incoming solar radiation by considering all bands that is back-scattered to the satellite before it reaches the earth's surface, and τ_{sw} is the atmospheric transmissivity (Allen et al., 2002).

Soil heat flux (G) is the other main parameter for energy balance equation which represents the heat storage in the equilibrium (Bastiaanssen and Robeling, 1993). G and R_n ratio have a constant amount, and by dividing G on R_n the ratio is better to be between 0.3-0.6. For equation (8) is used for calculating G (Bastiaanssen et al., 1999):

$$\frac{G}{R_n} = \frac{T_s \times R_n}{\alpha} (0.0032\alpha + 0.0062\alpha^2) \times (1 - 0.978NDVI^4) \quad (8)$$

Where T_s is the surface temperature and a vegetation index NDVI (Normalized Difference Vegetation Index) is used for G calculations by considering equation 9 (Allen et al., 2005).

$$NDVI = \frac{\rho_{NIR} - \rho_R}{\rho_{NIR} + \rho_R} \quad (9)$$

Where near infrared band and red band are used for NDVI calculations and shown as ρ_{NIR} and ρ_R in equation 9, respectively. The NDVI of the study area is more than zero due to plant cultivation.

156 The last main parameter in the energy balance equation is heat flux (H). H calculation has high
 157 complexity and sensitivity comparing with other parameters. Determining anchor pixels for
 158 calculating H is very crucial. After determining anchor pixels, heat flux is calculated from
 159 equation (10) (Allen et al., 2002):

$$H = \frac{\rho_{\text{air}} \times C_{\text{air}} \times dT}{r_{\text{ah}}} \quad (10)$$

160 Where ρ_{air} represents the air density (Kg/m^3), C_{air} represents air specific heat (1004J/kg/K), dT
 161 (K) represents the temperature difference ($T_1 - T_2$) between two heights (Z_1 and Z_2) and r_{ah}
 162 represents the aerodynamic resistance to heat transport (m/s).

163 Aerodynamic resistance to heat transport is calculated with equation (11):

$$r_{\text{ah}} = \frac{\ln\left(\frac{z_2}{z_1}\right)}{u_* \times k} \quad (11)$$

164 Z_1 and Z_2 are the two heights; k is Von Karman constant (0.41) and u_* represents friction
 165 velocity which is calculated by using equation 12:

$$u_* = \frac{k \times u_{200}}{\ln\left(\frac{200}{z_{\text{om}}}\right)} \quad (12)$$

166 u_{200} (blended wind speed) is wind speed measured at height 200 m at the weather station. z_{om} is
 167 empirically estimated from the vegetation height.

168 And z_{om} is calculated by leaf area index (equation (13)):

169

$$z_{om} = 0.018LAI \quad (13)$$

170 The LAI (equation 14) is calculated by using SAVI (soil adjusted vegetation index) (equation
171 (15)):

$$LAI = \frac{(\ln(\frac{0.69 - SAVI}{0.59}))}{0.91} \quad (14)$$

$$SAVI = \frac{(1 + L)(\rho NIR - \rho R)}{(\rho NIR - \rho R + L)} \quad (15)$$

172 Where L is a correction factor for the background soil.

173 Surface temperature difference which is used in equation 10 is a linear relation between dT and
174 Ts presented in equation 16:

$$dT = aT_s + b \quad (16)$$

175 Where dT is the near-surface air temperature difference, Ts is the surface temperature, and a and
176 b are empirical coefficients.

177 In the main equation (equation 1) the term latent heat flux for each pixel is calculated by
178 equation 17:

$$ET_{ins} = 3600 \frac{\lambda_{ET}}{\lambda} \quad (17)$$

179 Where λ is latent heat of vaporization (J/kg) and can be computed as (equation 18):

$$\lambda = [0.2501 - 0.00236(T_s - 273)] \times 10^6 \quad (18)$$

180 The main aim of using SEBAL algorithm is to obtain ET₂₄ for each pixel by using equation 19
181 (Allen et al., 2002):

$$ET_{24} = ET_r \times ET_{r-24} \quad (19)$$

Where the ET_{r-24} is the total ET_r during a 24-hour period of the same day.

For this study, two cloud-free satellite images were obtained for 8 April 2019 and 3 April 2020.

Actual evapotranspiration (ET_a) maps in mm/day are generated by the SEBAL algorithm for each day.

2.5. Data mining

Data science analysis Data Mining (DM) algorithms are the most fundamental components.

Certain DM techniques such as artificial neural networks, clustering, and case-based reasoning or

Bayesian networks have been applied in environmental modeling (Gibert et al., 2018).

Decision Tree methods uses the explanatory variables with higher discriminant power by

considering the response variable, then iteratively subdivide the training sample by building a

tree where the internal nodes are associated with the variables and its corresponding branches are

the possible values of the variable (Gibert et al., 2018). M5 Model Tree (introduced by Quinlan

in 1992), has linear regression functions at the leaf nodes, which develops a relationship between

input and output variables. Data are split into subsets and a decision tree is created. The data in

child nodes of splitting criterion depends on treating the standard deviation of the class values

and calculating the expected reduction in this error as a result of testing each attribute at that

node. The standard deviation reduction (SDR) is calculated as Eq. (20) (Quinlan 1992):

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \quad (20)$$

201

202 Where T is a set of data that reaches the node, T_i is the subset of data that has the i th outcome of
203 the potential set and sd is the standard deviation (Rahimikhoob et al., 2013; Wang and Witten,
204 1997). The data in child nodes are purer due to a less standard deviation in comparison to parent
205 nodes. The M5 tree selects the one that maximizes the expected error reduction after scanning all
206 the possible splits.

207 Every inner node of a tree has multiple linear regression models by using the data associated
208 with that node and all the attributes that participate in tests in the sub-tree rooted at that node.
209 The linear regression models are simplified If the results have a lower expected error (Etemad-
210 Shahidi and Bonakdar, 2009). Figure 2 shows a M5 decision tree structure for two input
211 parameter domains of X_1 and X_2 with 4 linear models from Y_1 to Y_4 .

212

213 **2.6. Model Inputs and Output**

214 For estimating evapotranspiration by SEBAL algorithm, meteorological data including
215 temperature, humidity, wind speed, etc are needed. Some inputs of SEBAL algorithm such as
216 albedo and emissivity are affected by the land surface temperature, so for using fewer variables
217 by data mining, albedo and emissivity were considered as inputs of M5 decision tree since these
218 two parameters are easy to get obtain? and better show the temperature variances. Also,
219 transpiration depends on the moisture of the plant. In data mining calculations, one vegetation
220 index must represent the plant moisture such as Normalized Difference Water Index (NDWI).
221 So, albedo and emissivity are represented as absorbed and transformed light to the atmosphere
222 and NDWI is represented as plant moisture. The main idea was to use the basic SEBAL equation

($ET = R_n - G - H$) as a simple equation: $ET = a(\text{Albedo}) - b(\text{emissivity}) - c(\text{NDWI})$ and calculate the constant values with M5 decision tree model. The three inputs of the M5 decision tree are explained hereafter in more details.

2.6.1. Evapotranspiration

One of the crucial parameters for evapotranspiration estimation is Land surface temperature (LST) the factor of which Radiation and the exchange of energy flux between the earth's surface and atmosphere depend on (Weng et al., 2019). A physical model such as the SEBAL algorithm has made it possible to estimate evapotranspiration for large areas. Surface biophysical characteristics such as albedo, greenness, and wetness are among the most important parameters affecting LST (Weng et al., 2019). The energy distribution is determined by albedo and emissivity of the surface and atmosphere. Previous studies show that surface emissivity strongly correlates to vegetation cover (Griend and Owe, 1993; Rechid et al., 2009); vegetation also strongly affects atmospheric properties through evapotranspiration (Gordon et al., 2005). Atmosphere emissivity is determined by atmospheric water vapor pressure (Staley and Jurica, 1972; Brutsaert, 1975).

2.6.2. Albedo

Albedo is a dimensionless diffuse reflectivity or reflecting power of a surface (Zhang et al., 2017) and is an important effective parameter on digital climate models and surface energy balance equations (Zhang et al., 2017). Surface albedo is computed by correcting the α_{toa} for

atmospheric transmissivity from equation 7. This parameter considered as one of decision tree input for calculating evapotranspiration.

2.6.3. Emissivity

The surface emissivity is the ratio of the actual radiation emitted by a surface which is emitted by a black body with the same surface temperature (Allen et al., 2002). Surface emissivity is an important variable for estimating land surface temperature and determining long-wave surface energy balance (Mira et al., 2010). Sobrino et al. (2004) proposed emissivity based NDVI in three different cases as Eq. (21):

$$\epsilon = \begin{cases} 0.973 & NDVI < 0.2 \\ \epsilon_v P_v + \epsilon_s (1 - P_v) + d\epsilon & 0.2 \leq NDVI \leq 0.5 \\ 0.986 & NDVI > 0.5 \end{cases} \quad (21)$$

Where ϵ_v is the vegetation canopy emissivity and ϵ_s is the bare soil emissivity; in this paper $\epsilon_v = 0.986$ and $\epsilon_s = 0.973$. The effect of the geometrical distribution of the natural surfaces is measured as $d\epsilon$ in Eq.6. P_v is the vegetation proportion obtained according to (Carlson and Ripley, 1997) as Eq. (22):

$$P_v = \left[\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s} \right]^2 \quad (22)$$

The minimum value of the NDVI for bare soil over the study region is presented as $NDVI_s$ and $NDVI_v$ is the highest NDVI for a fully vegetated pixel. The emissivity of land surfaces can differ

significantly by vegetation, surface moisture, and roughness (Nerry *et al.* 1988, Salisbury and D'Aria 1992).

2.6.4. Vegetation Index

The NDWI spectral index which represents the crop moisture is the normalized difference water index. NDWI has been used to estimate the equivalent water thickness of vegetation canopy (Yilmaz *et al.*, 2008). The NDWI considers two infrared bands with a central wavelength near about 0.86 μm (NIR), and a central wavelength of about 1.24 μm (SWIR). The equation is (Eq.23):

$$NDWI = \frac{\{\rho(0.86\mu m) - \rho(1.24\mu m)\}}{\{\rho(0.86\mu m) + \rho(1.24\mu m)\}} \quad (23)$$

The M5 decision tree model takes Albedo, emissivity, and a vegetation index as input and after the data mining process on these data, linear equations will be extracted. By inserting linear equations, the evapotranspiration map was obtained as an output with higher spatial resolution. Figure 3 shows the flowchart of the M5 decision tree and SEBAL algorithm.

2.7. Model Validation

2.7.1. Statistical Analysis

By using three inputs (albedo, emissivity, and NDWI) the accuracy of M5 decision tree was evaluated. The accuracy of the M5 decision tree model and the final evapotranspiration map which was combined with M5 decision tree was evaluated by correlation coefficient (R^2), root mean square errors (RMSE), and mean absolute errors (MAE) statistics (Eq. 24 to 26):

$$R^2 = 1 - \frac{\sum (ET_o - ET_p)^2}{\sum (ET_o - \overline{ET})^2} \quad (24)$$

$$RMSE = \sqrt{\frac{1}{n} \sum (ET_o - ET_p)^2} \quad (25)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |ET_o - ET_p| \quad (26)$$

Where N is the number of data, ET_o is the observed evaporation values calculated by the SEBAL algorithm and ET_p is the M5 decision tree model of estimated evapotranspiration.

2.7.1. Uncertainty Analysis

Uncertainty analysis could be calculated from the percentage of the observed data by the 95 PPU (95 Percent Prediction Uncertainty), and the average distance $\bar{d}$ between the upper and the lower 95 PPU (or the degree of uncertainty) by using equation (27) (Abbaspour et al., 2007):

$$\overline{d_x} = \frac{1}{k} \sum_{i=1}^k (X_U - X_L) \quad (27)$$

295

296 Where k is the number of observed data, X_L is the 2.5th and X_U is the 97.5th percentiles of the
 297 cumulative distribution of every estimated data.

298 If 100% of the observed data are bracketed by the 95PPU and $\bar{d}$ is close to zero the results are in
 299 acceptable range of uncertainty analysis. And $d - factor$ calculated from equation (28):

300

$$d - factor = \frac{\bar{d}}{\sigma} \quad (28)$$

301

302 Where σ is the standard deviation of the measured variable. The desirable $d - factor$ is less
 303 than 1 (Abbaspour et al., 2007).

304

305 **2.7.2. Sensitivity analysis**

306 sensitivity coefficient is a dimensionless index (S) which is calculated by the ratio of the change
 307 in output to input on the condition when the other variables remain constant. Sensitivity of
 308 dependent variable (evapotranspiration) to a particular independent variable (Albedo, emissivity
 309 and NDWI) can be calculated from the derivative of evapotranspiration with independent
 310 variable, $\delta ETA / \delta X$ (Beven, 1979; McCuen, 1974). To evaluate the sensitivity of a variable, S

was divided into four classes from small to high sensitivity (Table 1). If the S was ranged in high level, the independent variable would have more impact on the dependent variable.

3. Results

3.7. M5 decision tree

Instead of a physical-based model evapotranspiration estimation, the M5 decision tree was used for using a data mining model which could increase the spatial resolution of an evapotranspiration map. By using data mining, several equations were obtained for estimating evapotranspiration of a single day on 8 April 2019. Then the extracted equations were used for estimating the next year evapotranspiration of 3 April 2020 and were compared with evapotranspiration calculated by SEBAL algorithm on this day. Hereafter, results are explained in more details.

3.7.1. Inputs

Input variables of the M5 decision tree consisted of four satellite images including Albedo, Emissivity, Normalized Difference Water Index (NDWI), and estimated evapotranspiration calculated by SEBAL algorithm .

Figure 4 shows the albedo maps for 8 April 2019 and 3 April 2020.

330 According to figure 4, calculated albedo for 8 April 2019 is higher than albedo of 3 April 2020
331 because in 2019, it was in drought conditions and albedo is influenced by atmospheric vapor and
332 temperature (Feng and Zou, 2019).

333 Figure 5 shows the obtained emissivity maps for 8 April 2019 and 3 April 2020. Emissivity is the
334 most important input variable for the M5 decision tree because this variable is required for land
335 surface temperature calculations and is affected by percentage vegetation.

336 According to figure 5, both maps almost have the same maximum and minimum range, but
337 emissivity has a different amount on the same farm by comparing both years. Emissivity changes
338 depending on the cultivation of farms. In April 2019 (figure 5a) the west part of the plantation
339 had more cultivated sugarcane based on agricultural activities and management decisions and in
340 the next year (figure 5b) there was less cultivated sugarcane which can affect the amount of
341 emissivity.

342 Vegetation moisture has an important role in evapotranspiration calculations. Normalized
343 Difference Water Index (NDWI) can be considered as vegetation moisture in sugarcane farms.

344 Figure 6 shows NDWI for 8 April 2019 and 3 April 2020. On 8 April 2019 vegetation moisture
345 is more prone to dry tension because of drought conditions, but on 3 April 2020 vegetation
346 moisture is in better condition compared to 2019.

347 Several climatological variables such as wind velocity, relative humidity, temperature, etc. are
348 considered in evapotranspiration calculations and these parameters could appear in the substance
349 of the SEBAL algorithm. One of the input variables of an M5 decision tree considered as the
350 target variable is the evapotranspiration raster obtained by the SEBAL algorithm for 2019 and
351 2020 which are presented in other sections.

3.7.2. Output

The values in parentheses under each label in the leaves indicate the number of segments resulting from the corresponding threshold. The second value indicates the number of times a misclassification occurred (Vieira et al., 2012).

Figure 7 shows the decision tree for the evapotranspiration of 8 April 2019. 20 different equations were extracted with a Correlation coefficient of 0.9429, an absolute error and root mean squared error of 0.4749 and 0.6479, respectively.

By using fewer input parameters including albedo, emissivity, and NDWI, many evapotranspiration equations were extracted. For the evapotranspiration of 8 April 2019, 20 conditional equations were extracted. According to figure 7, the albedo input variable was located at the top of a decision tree and the divisions were based on albedo amounts. Therefore, albedo has high importance in calculating evapotranspiration. In the first leaf (first equation), divisions are based on albedo and according to figure 7, most equations were extracted based on albedo. By considering the geographical location of the study area, it can be observed that there is a high amount of receiving light in this area and the albedo input variable was considered as absorbed light. Therefore, M5 decision tree divisions show that the amount of absorbed light in this area has a major role in the evapotranspiration process which most of the decision tree divisions were based on the albedo.

NDWI variable is considered as plant moisture which has an important role in evapotranspiration calculations after albedo variable. This shows that the plant moisture has an important role in extracting the decision tree equations besides the absorbed lights.

The emissivity variable was considered as the diffused light which has less importance in the decision tree divisions and by considering the geographical location of the study area

The extracted equations by using M5 decision tree and python scripts in the ArcMap environment are presented as Appendix1 and 2 respectively at the end of this article.

3.8. Combining M5 and GIS

After extracting the most suitable equations from the M5 decision tree model, they were applied by using python scripts for faster and more accurate calculation (the python scripts could be provided by emailing the correspond author). The equations were obtained by using evapotranspiration of 8 April 2019 and were applied on input variables of the mentioned year to find if the extracted equations have acceptable performance. Figure 8 shows the evapotranspiration map calculated by the SEBAL algorithm and M5 decision tree for 8 April 2019 and 3 April 2020. According to figure 8 (a and c), The evapotranspiration maps for 2019 did not have many differences in the calculated evapotranspiration amounts. Figure 8b shows the calculated evapotranspiration by using the SEBAL algorithm for 2020 and figure 8d shows the calculated evapotranspiration by using the M5 decision tree equation extracted from 2019 SEBAL evapotranspiration map. Comparing figures 8b and 8d shows that the extracted equations from the M5 decision tree of 2019 can be applied on input variables of 3 April 2020 and obtain acceptable results due to less difference between the evapotranspiration map calculated by SEBAL algorithm and M5 decision tree for 2020.

Figure 9 shows the results of the comparison between the SEBAL algorithm and M5 decision tree for 8 April 2019 and 3 April 2020. Table 2 shows statistical coefficients for 8 April 2019

and 3 April 2020. According to figure 9 and table 2 by comparing the obtained results for the two different years, it could be possible to calculate evapotranspiration with fewer input variables and by using the most important variables due to high correlation coefficient and low errors. And according to Table 2, the calculated evapotranspiration for 3 April 2020 has obtained better results than 8 April 2019. This is due to the differences between input variables of the two years which makes the extracted equations with better accuracy on 3 April 2020.

3.9. Model evaluation

The evapotranspiration obtained by using data mining model was evaluated by using uncertainty analysis and sensitivity analysis. Uncertainty analysis of the obtained model was calculated by using two criteria, namely 95PPU and d-factor. By increasing in observed data in 95PPU level and the decrease in the average value of upper and lower bands (smaller than the standard deviation), also the bracketed value 95PPU must be in maximum range. In that case all observed data (calculated by SEBAL algorithm) are between the lower and the upper limit of the uncertainty defined as 2.5% and 97.5% (X_L and X_U) levels of the cumulative distribution of the output variables. When 80% of the calculated data (calculated by data mining model) are in 95PPU level, it makes clear that they are of high quality.

Table 3 shows the Uncertainty coefficient for 8 April 2019 and 3 April 2020. Due to this table the obtained d-factor is less than 1 for both images and 95PPU is more than 80%. The calculated uncertainty coefficients show that data mining model has high quality for estimating evapotranspiration by using less variables and could be used for evapotranspiration estimation with acceptable certainty.

Table 4 shows that the sensitivity coefficient of the variables is different. Albedo variable has the highest sensitivity for calculating evapotranspiration in both images of 8 April 2019 and 3 April 2020. The study area is located in a region with low sea level and receive lots of sun radiation. Net radiation (Rn) is one of the main variables for calculating evapotranspiration by using SEBAL algorithm and albedo could represent this variable as one of data mining model inputs. And the decision tree obtained for this research (figure7) shows that most of classifications occurred were the albedo variable considered in the node parts of the decision tree which makes clear that albedo has important role in extracting the equations and has high sensitivity. The other two parameters have less sensitivity in evapotranspiration calculations which in 2020 NDWI has small sensitivity due to late start of irrigating the cultivated sugarcane fields comparing with the last year (2019).

4. Discussion

Climatological variables can affect an evapotranspiration calculation which is a physical process. Measuring evapotranspiration in the extended area by using remote sensing is a possible process and can be done by physical-based algorithms like SEBAL. By transforming a physical-based model to a data mining model, the input variable can be decreased to the most important variable. By using an M5 decision tree, a physical-based evapotranspiration algorithm is transformed into a data mining model. The three inputs of the decision tree consisted of albedo, emissivity, and NDWI and the target function was evapotranspiration obtained by the SEBAL algorithm. Also, the possibility of calculating evapotranspiration by using linear regression equations obtained by M5 was evaluated. In this case, a complicated evapotranspiration process which was dependent on so many variables showed that even by using simple linear regression,

equations can be estimated. Another achievement of this research is to use the general equation: $ET = a(\text{albedo}) - b(\text{emissivity}) - c(\text{NDWI})$ which the constant values were obtained by using data mining. According to the basic SEBAL algorithm equation which is represented as energy balance: $\lambda ET = R_n - G - H$, this equation shows that evapotranspiration can be calculated as absorbed energy (net radiation at the surface) minus from transmitted energy (soil heat flux and sensible heat flux to the air) so instead of using energy term (heat fluxes), light terms were used as the main equation and evapotranspiration was calculated as the absorbed light (albedo) minus transmitted light (emissivity) due to the theory that light is another form of energy. Therefore, by using the M5 decision tree, the most acceptable result was obtained for calculating evapotranspiration in by using data mining method.

5. Conclusion

The main aim of this research was to transform a physical-based model by a data mining model since physical-based models need more input variables in comparison to data mining models which use fewer input variables and obtain acceptable results. By using three input variables including albedo, emissivity, and NDWI mathematical equations were extracted from M5 decision tree. These three variables were chosen according to the basic SEBAL algorithm equation: $ET = R_n - G - H$. the main idea is to calculate evapotranspiration as $ET = a(\text{albedo}) - b(\text{emissivity}) - c(\text{NDWI})$ which the constant values were calculated from M5 decision tree model. The extracted equations are applied on 8 April 2019 and the evapotranspiration calculations are almost as the same as SEBAL algorithm evapotranspiration values and for 3 April 2020 the extracted equations from 2019 are applied on this year which comparing to SEBAL evapotranspiration algorithm, is in the same range and make an acceptable result. In

conclusion, SEBAL algorithm, by a data mining model, M5 decision tree, by using fewer inputs
evapotranspiration can be calculated in an extended area. Also, this research makes it clear that
light term can be used instead of the energy term.

6. Acknowledgments

We are grateful to the Research Council of Shahid Chamran University of Ahvaz for financial
support (GN: SCU.WI98.281).

Satellite images were supported through USGS earth explorer website. The authors would like to
thank the NASA agency for free access images to implementing this research.

Also Special thanks to Sugarcane Research and Training Institute for providing the equipment
and their expert advice.

Also Special thanks to Sugarcane Research and Training Institute for providing the equipment of
current research.

Conflict Of Interest- None.

7. Funding Statement

This study was funded by “Shahid Chamran University of Ahvaz”.

8. Author Contribution

All authors contributed to the study conception and design. Material preparation, data collection
and analysis were performed by [Lamya Neissi], [Mona Golabi], [Mohammad Albaji] and [Abd
Ali Naseri]. The first draft of the manuscript was written by [Lamya Neissi] and all authors

commented on previous versions of the manuscript. All authors read and approved the final manuscript. Also, [Mohammad Albaji] is the corresponding author of this manuscript.

9. Data availability statement

Data will be made available on request.

9. Reference

1. Abbaspour K.C, Yang J, Maximov I, Siber R, Bogner K, Mieleitner J, Zobrist J, Srinivasan R (2007) Modelling hydrology and water quality in the pre-alpine/ alpine Thur watershed using SWAT. J. Hydrol. 333(2-4): 413-430.
2. Abdolhosseini M, Eslamian S, Mousavi SF (2012) Effect of climate change on potential evapotranspiration: a case study on Gharehsoo sub-basin, Iran. 2(4): 362-372.
3. Allen R, Tasumi M, Trezza R (2002) SEBAL (Surface Energy Balance Algorithms for Land)-Advanced Training and User's Manual-Idaho Implementation, Version 1.0.
4. Anderson MC, Allen RG, Morse A , Kustas WP (2012) Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. Remote Sens. Environ. 122, 50–65.
5. Bastiaanssen W, Menenti M, Feddes R, Holtslag A (1998) A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. Jhyd. 212, 198–212.
6. Beven K (1979) A sensitivity analysis of the Penman-Monteith actual evapotranspiration estimates. Journal of Hydrology. 44: 3-4, 169-190.

7. Brutsaert W (1975)On a derivable formula for long-wave radiation from clear skies.
Water Resour. Res. 11, 742–744.
8. Carlson TN, Ripley DA (1997) On the relation between NDVI, fractional vegetation
cover, and leaf area index. Remote Sens. Environ. 62, 241–252.
9. Colaizzi PD, O’Shaughnessya SA, Evetta SR, Mounceb RB (2017) Crop
evapotranspiration calculation using infrared thermometers aboard center pivots. Agri.
Water Manag. 187, 173–189.
10. Diarraa A, Jarlan L, Er-Rakic S, Le Pageb M, Aouadec G, Tavernierb A, Bouletb G,
Ezzahard J, Merlinb O, Khabbaa S (2017)Performance of the two-source energy budget
(TSEB) model for the monitoring of evapotranspiration over irrigated annual crops in
North Africa. Agri. Water Manag. 193, 71–88.
11. Elnmer A, Khadr M, Kanae S, Tawfik A (2019) Mapping daily and seasonally
evapotranspiration using remote sensing techniques over the Nile delta. Agri. Water
Manag. 213, 682-692. ISSN 0378-3774, <https://doi.org/10.1016/j.agwat.2018.11.009>.
12. Etemad-Shahidi A, Bonakdar L (2009) Design of rubble-mound breakwaters using M5 ‘
machin learning method. Appl. Ocean Res. 31(3), 197-201.
13. Feng H, Zou B (2019) A greening world enhances the surface-air temperature difference.
Sci. Tot. Environ. 658, 385–394.
14. Gibert K, Izquierdo J, Sànchez-Marrè M, Hamilton SH, Rodríguez-Roda I, Holmes G
(2018) Which method to use? An assessment of data mining methods in Environmental
Data Science. Environ. Model. Software. 110, 3-27. doi:
<https://doi.org/10.1016/j.envsoft.2018.09.021>.

15. Gobbo S, Lo Presti S, Martello M, Panunzi L, Berti A , Morari F (2019) Integrating SEBAL with in-Field Crop Water Status Measurement for Precision Irrigation Applications—A Case Study. *Remote Sens.* 11, 2069.
16. Goodarzi M and Eslamian S (2018) Performance evaluation of linear and nonlinear models for the estimation of reference evapotranspiration. *International Journal of Hydrology Science and Technology*, 8(1): 1-15.
17. Gordon LJ, Steffen W, Jönsson BF, Folke C, Falkenmark M, Johannessen Å (2005) Human modification of global water vapor flows from the land surface. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7612–7617.
18. Griend AAVd, Owe M (1993) On the relationship between thermal emissivity and the normalized difference vegetation index for natural surfaces. *Int. J. Remote Sens.* 16, 1119–1131.
19. Jaferian V, Toghraie D, Pourfattah F, Ali Akbari O, Talebizadehsardari P (2019) Numerical investigation of the effect of water/Al O nanofluid on heat transfer in trapezoidal, sinusoidal and stepped microchannels. *Int. J. Num. Methods for Heat & Fluid Flow*. DOI: 10.1108/HFF-05-2019-0377.
20. Kamali MI, Nazari R (2018) Determination of maize water requirement using remote sensing data and SEBAL algorithm. *Agri. Water Manag.* 209, 197-205. ISSN 0378-3774, <https://doi.org/10.1016/j.agwat.2018.07.035>.
21. Kong J, Hu Y, Yang L, Shan Z, Wang Y (2019) Estimation of evapotranspiration for the blown-sand region in the Ordos basin based on the SEBAL model. *Int. J. Remote Sens.* 40, 5-6, 1945-1965, DOI: 10.1080/01431161.2018.1508919.

22. Lenhart T, Eckhardt K, Fohrer N, Frede H.G (2002) Comparison of two different approaches of sensitivity analysis. *Physics and Chemistry of the Earth, Parts A/B/C*. 27: 9-10, 645-654.
23. McCuen R.H (1974) A sensitivity and error analysis of procedures used for estimating evaporation. *Journal of the American Water Resources Association*. 10(3): 486-497.
24. Mhawej M, Elias G, Nasrallah A, Faour G (2020) Dynamic calibration for better SEBALI ET estimations: Validations and recommendations. *Agri. Water Manag.* 230, 105955. ISSN 0378-3774, <https://doi.org/10.1016/j.agwat.2019.105955>.
25. Mira M, EValor E, Caselles V, Rubio E, Coll C, Galve JM, Niclos R, Sanchez JM, Boluda R (2010) Soil Moisture Effect on Thermal Infrared (8-13- μ m) Emissivity. *Ieee Transactions on Geosci. Remote Sens.* 48(5), 2251-60. doi: Doi 10.1109/Tgrs.2009.2039143.
26. Nerry F, Labed J, Stoll MP (1988) Emissivity signatures in the thermal IR band for remote sensing: calibration procedure and method of measurements. *Appl. Optics*. 27, 758–764.
27. Ochege FU, Luo G, Obeta MC, Owusu G, Duulatov E, Cao L, Nsengiyumva JB (2019) Mapping evapotranspiration variability over a complex oasis-desert ecosystem based on automated calibration of Landsat 7 ETM+ data in SEBAL. *GISci. Remote Sens.* 56(8), 1305-1332. DOI: 10.1080/15481603.2019.1643531.
28. Quinlan JR (1992) Learning with continuous classes. In: *Proceedings of Australian Joint Conference on Artificial Intelligence* (Singapore: World Scientific Press). 343–348.

29. Rahimikhoob A, Asadi M, Mashal M (2013) A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. *Water Resour. Manag.* 27(14), 4815–4826.
30. Rechid D, Raddatz TJ, Jacob D (2009) Parameterization of snow-free land surface albedo as a function of vegetation phenology based on MODIS data and applied in climate modelling. *Theor. Appl. Climatol.* 95, 245–255.
31. Salisbury W, D’Arì DM (1992) Emissivity of terrestrial materials in the 8–14 mm atmospheric window. *Remote Sens. Environ.* 42, 83–106.
32. Singh RK, Irmak A, Irmak S, Martin DL (2008) Application of SEBAL model for mapping evapotranspiration and estimating surface energy fluxes in south-central Nebraska. *J. Irrig. Drain. Eng.* 134(3), 273–285.
33. Sobrino JA, Jiménez-Muñoz JC, Paolini L (2004) Land surface temperature retrieval from LANDSAT TM5. *Remote Sensing of Environment.* 90, 434–440.
34. Song L, Liu S, Kustas WP, Nieto H, Sun L, Xu Z, Skaggs TH, Yang Y, Ma M, Xu T, Tang X, Li Q (2018) Monitoring and validating spatially and temporally continuous daily evaporation and transpiration at river basin scale. *Remote Sens. Environ.* 219, 72–88.
35. Staley DO, Jurica GM (1972) Effective atmospheric emissivity under clear skies. *J. Appl. Meteorol.* 11, 349–356.
36. Vieira MA, Formaggio AR, Rennó CD, Atzberger C, Aguiar DA, Mello MP (2012) Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. *Remote Sens. Environ.* 123, 553–562.

37. Wang Y, Witten IH (1997) Induction of model trees for predicting continuous classes. In:
 Proceedings of the Poster Papers of the European Conference on Machine Learning.
 University of Economics, Faculty of Informatics and Statistics, Prague.

38. Weng Q, Karimi Firozjaei M, Kiavarz M, Alavipanah SK, Hamzeh S (2019)
 Normalizing land surface temperature for environmental parameters in mountainous and
 urban areas of a cold semi-arid climate. *Sci. Tot. Environ.* 650, 515–529.

39. Yilmaz MT, Hunt ER, Jackson TJ (2008) Remote sensing of vegetation water content
 from equivalent water thickness using satellite imagery. *Remote Sens. Environ.* 112 (5),
 2514–2522. <http://www.sciencedirect.com/science/article/pii/S0034425707004798>.

Figures

Figure 1

Amir-Kabir sugarcane Agro-Industry location area

Figure 2

Structure of M5 decision tree (Models Y1–Y4 are linear regression models)

Figure 3

Flowchart of the M5 decision tree and SEBAL algorithm

Figure 4

Albedo maps of (a) 8 April 2019 and (b) 3 April 2020

Figure 5

Emissivity maps of (a) 8 April 2019 and (b) 3 April 2020

Figure 6

NDWI maps of (a) 8 April 2019 and (b) 3 April 2020

Figure 7

Decision tree for the evapotranspiration of 8 April 2019

Figure 8

Evapotranspiration map calculated by SEBAL algorithm for (a) 8 April 2019 and (b) 3 April 2020 and evapotranspiration map calculated by M5 decision tree for (c) 8 April 2019 and (d) 3 April 2020

Figure 9

Comparing the results between the SEBAL algorithm and M5 decision tree