Use of Spatial Visualization for Pattern Discovery in Evapotranspiration Estimation

Fernando Xavier[1,2], Maria Luíza Correa Brochado[3]

[1]Centro Universitário do Distrito Federal (UDF)
SHCS Q704/904 – Asa Sul - 70390-045 – Brasília – DF – Brazil

[2]Polytechnic School – University of São Paulo (USP)
Av. Prof. Luciano Gualberto, 380 - Butantã – 05508-010 – São Paulo – SP – Brazil

[3]Geography Department – University of Brasília (UnB)
Campus Universitário Darcy Ribeiro, Asa norte, Distrito Federal, Brazil

fxavier@usp.br, luizacorreaenf@gmail.com

Abstract. *In Water Resources area, data are obtained from various sources, such as measuring instruments and satellites. Often such data may contain patterns that are not easily identified, either because of the large volume of data sets or because the analysis requires the use of several data dimensions. In this way, this study proposes the application of machine learning resources and spatial visualization to identify patterns in the estimation of an important component of the hydrological cycle: the evapotranspiration. This work is expected to contribute to an approach to estimate evapotranspiration, using spatial resources for pattern identification and model generation.*

1. Introduction

In Big Data scenario, a big challenge for researchers is how to extract useful information in large volumes of data which are obtained from various sources and at increasing rates. In areas such as Water Resources, data are generated by mathematical models, sensors, conventional measurement instruments. Processing this data and extracting information in this context requires the use of new approaches or even the adaptation of existing approaches, with application of many techniques and concepts in an integrated way.

Among these approaches, stands out the machine learning, by use of techniques in which data is processed in an automated way using algorithms with a variety of objectives, such as patterns discovery in large volumes of data. Another widely used technique is spatial visualization, in which georeferenced data are analyzed combining spatial layers. Spatial analysis provides many visualization benefits in areas of high data density but, in other hand, occurs frequently overlapping that makes it difficult to distinguish between the points. One possibility that solves this problem is the mapping by heat maps, technique which uses a color gradient to represent the geographic density of elements on a map.

The use of these techniques, applied alone or together, may reveal patterns that are not clear from the preliminary analyzes and should be used to confirm or refute hypotheses about the data, adding new information to the experiments.

Based on this, this article is an application report of an experiment using two of these techniques, data mining and spatial visualization, for the discovery of regional patterns in an important activity in the Water Resources area: the evapotranspiration estimation.

Using data available from the Brazilian National Institute of Meteorology (INMET), it was applied the data mining technique to identify a model for estimating evapotranspiration in a meteorological station, named reference station. This model, in turn, was applied to other meteorological stations in Brazil, aiming to identify the validity of this model in data for these stations. Finally, the data generated were visualized on maps in order to verify if there was a regional pattern, be it related to latitude or vegetation cover.

It is expected, with this study, to demonstrate how the integrated use of two techniques for data analysis can aggregate information that would not be clear if they were applied in an isolated way. In addition, the experiment may reveal information to researchers in the Water Resources area about possible patterns in the evapotranspiration estimation.

In Section 2, it is described the theoretical reference used in this work, detailing concepts related to evapotranspiration and its estimation methods, such as use of machine learning. The following section describes the approach that was used to solve the research problem as well as the methods for evaluating the solution. In Section 4, it was detailed the steps of the experiment, with information about its execution. The application of spatial visualization is illustrated in Section 5, followed by the analysis about results obtained in Section 6. Finally, in Section 7, final considerations of this work are made, including suggestions of future research works.

2. Background

2.1. Evapotranspiration

Evapotranspiration is a component of the water cycle, defined as the loss of surface water through the combination of soil evaporation processes and vegetation transpiration [Di Bello 2005], which returns returns to atmosphere as vapor, as shown in Figure 1.

There are many local and meteorological factors that affect the amount of the water lost by the surface in the evapotranspiration process [FAO 2017]. Among the local factors, are included the type of vegetation and the soil, which influence the surface capacity to absorb the water received from rain. In addition, weather conditions such as temperature, wind speed, humidity and cloudiness also contribute to the process, directly or indirectly, affecting the energy amount used in the transformation of water molecules from the liquid to the gaseous state.

According to the FAO, the main energy source used in this process comes from solar radiation, which varies according to factors such as latitude, altitude, cloudiness among others. In addition, according to Figure 2, the radiation also depends on the period of the year and the time of day.

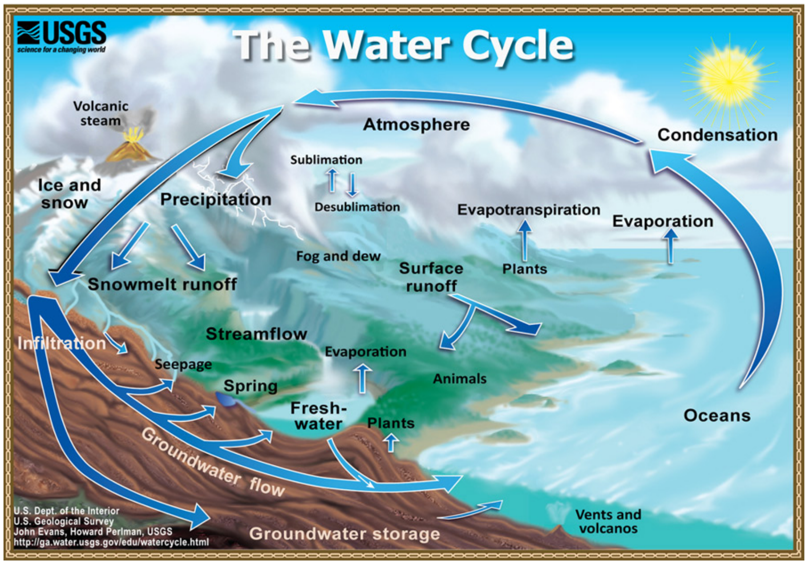

Figure 1. Evapotranspiration in the hydrological cycle [Evans & Perlman 2015]

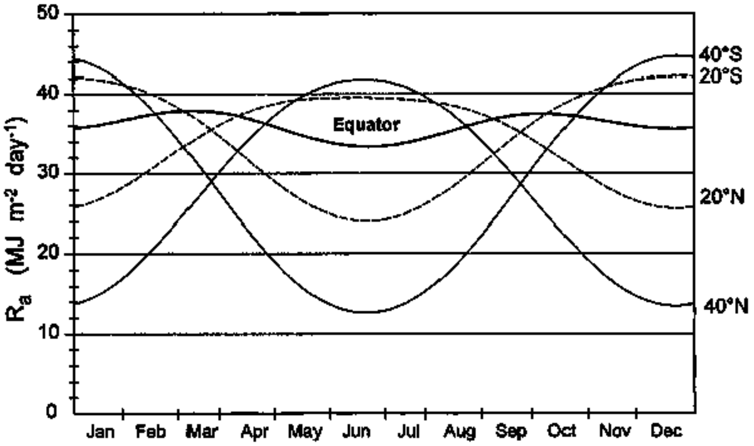

Figure 2. The solar radiation variation according to latitude and month [FAO 2017]

The estimation of evapotranspiration can be done through instruments such as lysimeters, remote sensing methods and mathematical models, such as the Penman-Monteith reference equation, FAO reference method. Due to the difficulty of obtaining the values for all parameters of the equation [Majidi et al 2015], other alternative models have been proposed, such as the Thornwaite and Hargreaves equations [Camargo et al 1999].

According to the FAO Guide, one of the alternatives for evapotranspiration estimation in the missing data scenario would be using data from nearby meteorological stations, since the conditions that affect evapotranspiration may be similar in geographically close regions.

2.2. Machine Learning application in Evapotranspiration Estimation

Machine learning was used to estimating evapotranspiration, based on historical data from the National Institute of Meteorology (INMET), obtaining models for evapotranspiration estimation by locality [Xavier 2016]. In this work, it was applied data mining to discover a model in each station using data from 2010 to 2014, aiming to generate a model with less attributes than the Penman-Monteith equation.

Another approach was used in a preliminary study [Xavier el at 2015], based on the hypothesis of evapotranspiration estimation by similarity criteria was used. In this study, the model discovered by the machine learning method for one locality was applied in six locations, three of them classified as similar and the other three classified as not similar , using the latitude as a factor of similarity. According the preliminary results, the model learned for a locality could be applied in places with similar characteristics.

By means of use of computational visualization resources, it is intended to evolve this preliminary study, verifying possible patterns in the estimation of evapotranspiration aggregating factors other than latitude and extend these approach to more stations than preliminary study mentioned previously.

3. Methodology

3.1. Solution Proposed

Using historical series of meteorological data obtained from INMET datasets, a model of evapotranspiration estimation for a locality will be obtained through the use of machine learning. This location will be called a reference location and will be used in comparison with other locations.

The model generated for the reference location will be used to estimate evapotranspiration using data from other locations, each called a test location. In the model application in each test location data, will be calculated the correlation between the evapotranspiration value obtained by the model learned and the historical values available in the INMET datasets.

In this way, the hypothesis to be verified in this work is: how much more a test location is similar, defined according to latitude or vegetation, to the reference location, better will be the correlation obtained by application of the reference equation in the test location data.

3.2. Evaluation of the proposed solution

By means of spatial visualization, the correlation values of each location will be compared using layers of latitude data, potential evapotranspiration, and state boundaries.

3.3. Experiment Planning

The visualization process consists of several steps [WARD et al 2010], from raw data collection to visualization by users. By means of this process, illustrated by the Visualization Pipeline in Figure 3, five stages were defined to execute the experiment proposed in this research work:

- Data Collection and Analysis: meteorological data will be collected from the

INMET database and will be applied exploratory evaluation of these datasets;

- Pre-Processing: application of transformations in data as well as preparation of datasets for processing, excluding non-relevant columns for this study;
- Processing: use of classification algorithms, to determine an estimation model of the evapotranspiration for the reference station and application of this model to data of the other stations;
- Spatial Visualization: visualization of correlation values in many maps combined to the layers that represent the similarity factors;
- Results Analysis: evaluation if the results confirm or refute the hypothesis defined for the experiment.

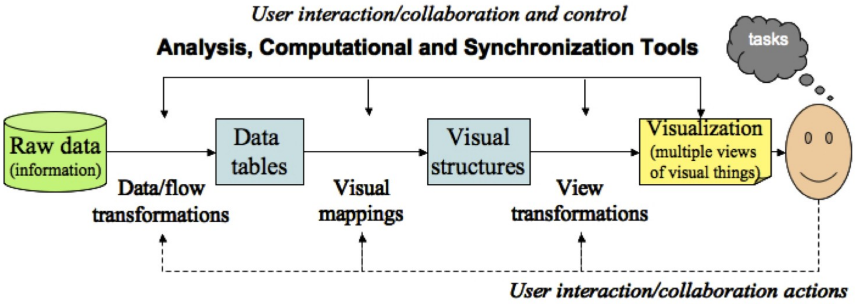
User interaction/collaboration and control
Analysis, Computational and Synchronization Tools
tasks
Raw data (information)
Data tables
Visual structures
Visualization (multiple views of visual things)
Data/flow transformations
Visual mappings
View transformations
User interaction/collaboration actions

Figure 3. Data visualization pipeline [WARD et al 2010]

In addition to the data visualization pipeline, the steps in this experiment were defined according to the steps of the Knowledge Discovery in Databases (KDD) process [Fayad et al 1996].

4. Experiment Description

4.1. Collection and Analysis Data Step

The data were collected from the Meteorological Database for Teaching and Research (BDMEP), a tool created by INMET to provide historical data to researchers. The following filters were used to collect the data:

- Period: from 01/01/1996 to 31/12/2016
- Measurements: All
- Stations: All

These datasets are available in CSV format (comma separated values and, in addition to the historical data series, contains information about the station, such as latitude, longitude and altitude.

From the analysis of the datasets, it was found that some attributes could be removed, due to the high degree of missing data or to contain values that would not influence in an equation, such as date and station code. In this way, it was defined that the attributes from the dataset used in the processing step would be:

- Average Wind Speed: Average wind speeds in the period;
- Average Max Wind Speed: Average maximum wind speeds in the period;
- Evapotranspiration Potential: Evapotranspiration estimated in the month;

- Total Insolation: The number of hours that sunlight has reached the surface of the Earth without cloud interference;
- Average cloudiness: The fraction of the celestial vault that is occupied by clouds;
- Total rainfall: amount of rainfall of the period;
- Maximum Average Temperature: average of the maximum temperatures of the period;
- Mean Compensated Temperature: average between maximum and minimum temperatures, in addition to three measures taken during the day (9, 15 and 21 hours);
- Average Minimum Temperature: average minimum period temperatures;
- Relative Humidity Average: percentage of water vapor in the air.

These attributes, defined by Branco (2014), are obtained by standards defined by the World Meteorological Organization (WMO) [WMO 2014].

4.2. Pre-processing Step

To use the data in the processing step, it was necessary a way to extract the metadata information from the historical data series. In this way, it was developed a simple Java program to extract data according to its use in the later steps:

- Stations information: latitude and longitude, to be used in the spatial visualization;
- Historical series data: for the application of evapotranspiration estimation models and including only the attributes defined in the previous step.

As a result of this step, files were generated in the CSV format (comma separated values) with the data of the historical series as well as a CSV file with the data of the stations.

4.3. Processing Step

In order to learn the baseline model of evapotranspiration estimation, it were data from the Resende-RJ station, due to its proximity to many stations in Brazil, which would guarantee a good number of stations considered similar, using latitude and vegetation criteria.

Data mining activity was executed using the Weka software [Hall et al 2009] [Frank et al 2016], a very-known tool to knowledge discovery in databases. By means of this tool, the data from the Resende-RJ station was processed using the M5P algorithm and a model for evapotranspiration estimation was generated.

This generated model have a correlation coefficient of 0.9715 for the Resende-RJ station data. After that, the generated model was applied to data of each station, obtaining correlation coefficient and storing it in a data file to be used in the next steps.

At the end of the processing stage, data were obtained for 252 stations from the model generated for the station of Resende-RJ. These data were recorded in CSV files for use in the later stage of this study.

5. Spatial Visualization

From the correlation coefficients obtained for each station in the previous step, three maps were generated to evaluate the results from different perspectives. For the three maps, the correlation coefficients obtained was classified according to the range of values defined in Table 1.

Table 1. Classification of the correlation coefficient by range

Correlation Coefficient	Classification
> 0.70	High
0.41 – 0.70	Medium
< 0.4	Low

This step, the spatial visualization, was developed using a tool to process data obtained from the previous steps and to combine them with a spatial layers related to the objectives of this study.

The tool choosen was QGis, free software to support spatial analysis and with support for multiple data formats [Qgis 2017], that was used to prepare the maps. In addition to the QGis, layers from the WorldClim climate database were used, which contains weather data for geographic information systems [Fick & Hijmans 2017], available in raster format.

The Figure 4 shows the correlation coefficients of each station used in the experiment, according to its geographic location, obtained from the metadata extracted in the pre-processing step. The use of this map aims to visualize the results to find some regional patterns, such as relations about the Brazilian biomes.

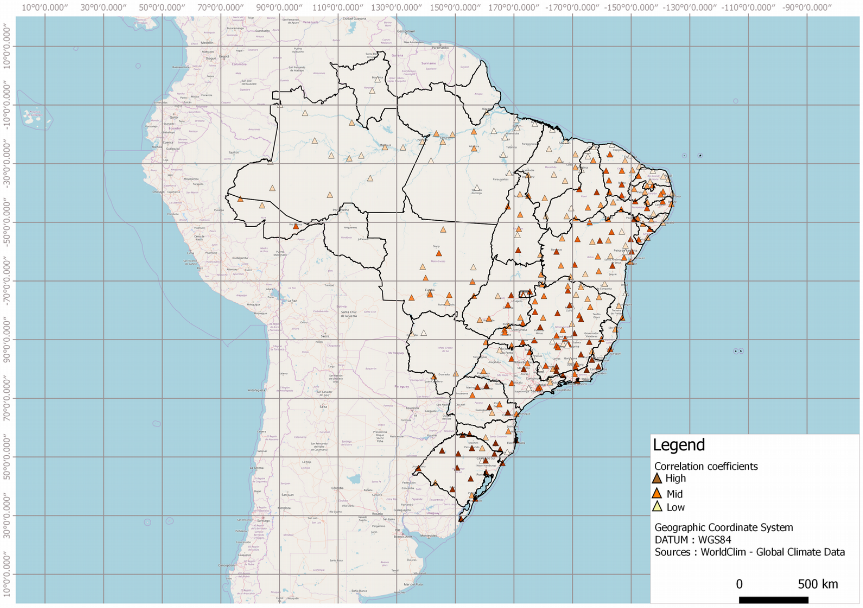

Figure 4. Correlation coefficients per station

In a later analysis, the correlation coefficients for each station were compared with the mean values of potential evapotranspiration in the month of September/2017 (Figure 5), obtained from Embrapa (2017).

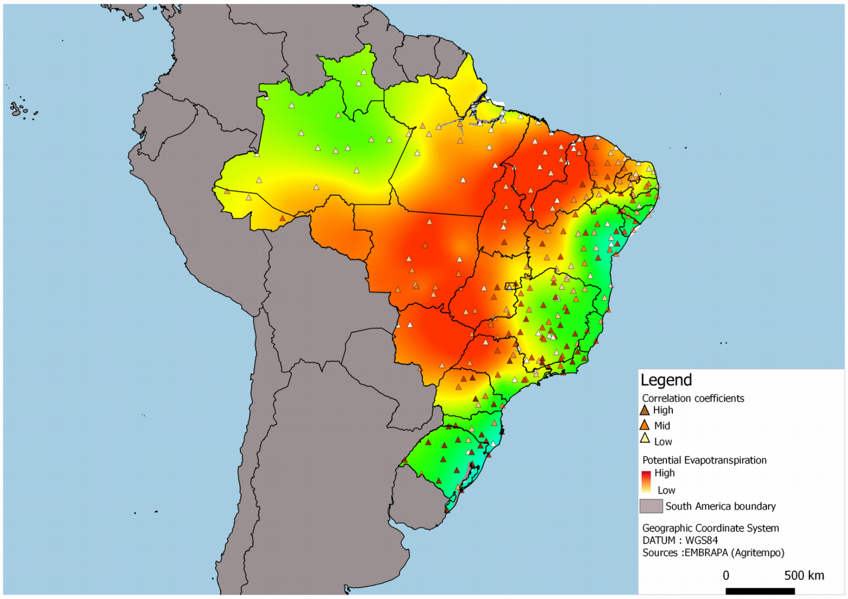

Figure 5. Visualization of correlation coefficients per station in relation to the potential evapotranspiration values

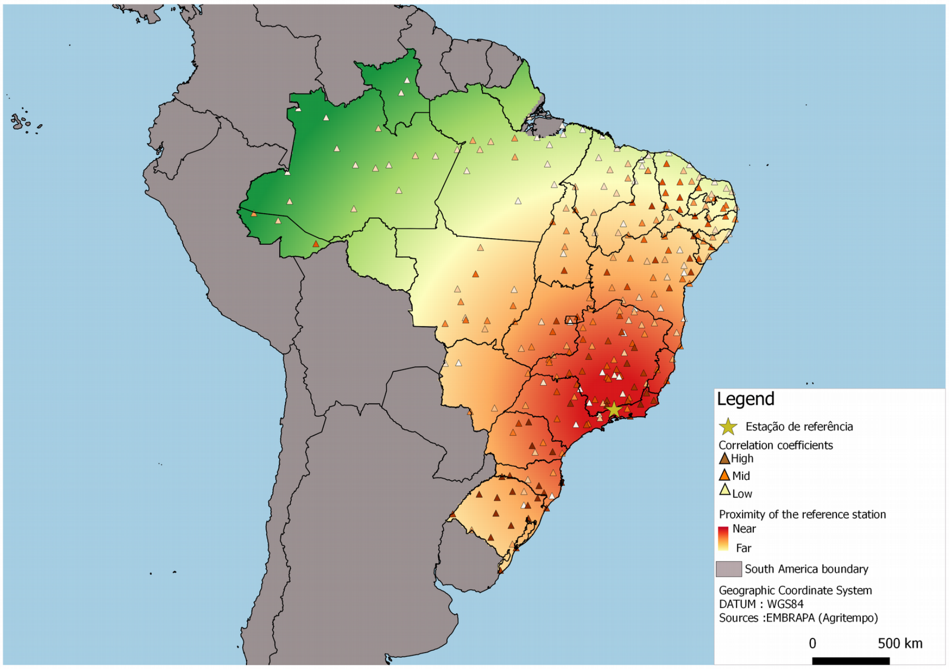

Figure 6. Visualization of the correlation coefficients of each station in relation to the reference station

In order to verify the quality of the correlation coefficients obtained in relation to the geographic distance of the reference station (Resende-RJ), a third heat map was created using the factor distance to the reference station (Figure 6). For this analysis, distance bands were established in relation to the reference station and its influence on the correlation coefficients was evaluated.

6. Results Analysis

The map in Figure 4 indicates that there was a higher concentration of high correlation coefficients in the South, Southeast and Northeast brazilians regions. This result is an evidence that the quality of the model generated from the data of the station of Resende-RJ is related to common characteristics of these regions, such as climate and vegetation.

It is also noticed that there was concentration of low correlation coefficients for the Amazon region, fact that can reinforce the relationship of the model with the vegetation. Another indication of the model's close relation to vegetation can be evidenced by the quality of the model for the Cerrado and Caatinga biomes (where there was a higher concentration of medium and high correlation coefficients) that are similar in relation to the type of vegetation (adapted to climates more dry), low levels of rainfall and less drained soils.

Regarding the mean values of potential evapotranspiration (Figure 5), no evident correlations were observed. In regions with low potential evapotranspiration value, the concentration of mean correlation coefficients in the South, Southeast and Northeast regions was observed, as well as the concentration of low correlation indices in part of the Amazon region. Regarding the mean correlation indexes, it was observed a higher concentration in the localities with high potential evapotranspiration value.

The latitude was one of the factors used in this work to classify similarity between stations. The hypothesis that the closer to the reference station, the better the quality of the model generated would be refuted (based on Figure 6), since in the experiment were obtained both low and high correlation coefficients for the same latitude of the reference station (similar stations by the latitude criteria). Also corroborating with the refutation of the hypothesis, different high-correlation coefficients were observed for locations of different latitudes, mainly in models applied to stations in the South and Northeast regions.

After analysis of the three maps, it was noticed that latitude would not be a good indicator of similarity between stations. On the other hand, vegetation could be a better indicator, given the concentration of high correlation indexes in the regions of the Atlantic Forest and Pampas, while in the Cerrado and Amazon, there were concentrations of medium and low correction index, respectively.

7. Conclusions

In the experiment described in this work, it was intend to study the use of spatial visualization as a way to identify patterns in the estimation of evapotranspiration. Using a machine learning approach, a model was generated for a reference station, which was used to estimate evapotranspiration in other INMET weather stations. The correlation coefficients generated between the evapotranspiration historical data and those calculated by the reference station model were placed on maps to identify possible patterns.

Using latitude and vegetation as factors to classify similarities between stations, it was defined a hypothesis that locations similar to the reference location would have better results with the model generated for this station. The latitude was defined as criteria to define similarity used because it is related to the radiation received by the localities, which is the main energy source in the evapotranspiration process. The vegetation, in turn, was used as criteria mainly due to the fact that part of the evapotranspiration value is originated from the transpiration of the vegetation.

In relation to vegetation, it was observed better rates in the Atlantic Forest biome, which is the same region of the reference station, and in the Pampas biome. In other regions, such as those delimited by the Amazon, the correlation coefficients were low, besides the Cerrado, where the average coefficients were concentrated.

In opposite to the observed in the preliminary study developed by Xavier et al. (2015), latitude was not noticed a good indicator of similarity for estimating evapotranspiration, since there were several points with low correlation coefficients at locations considered similar using this criteria, whereas several points with a high correlation coefficients in non-similar locations were identified by the latitude criteria.

As future works of this study, it is suggested application of the experiment using other reference locations. In addition, due to the variety of factors influencing evapotranspiration, it is also suggested application of the study using the combination of other layers of data, because is possible that exists other patterns not clearly visible.

This research work has demonstrated that the approach of spatial visualization to identify patterns in the estimation of evapotranspiration can be very useful, either to aggregate new information to studies carried out with other approaches or to discover new patterns that were not previously identified in other approaches of analysis of data. The use of spatial visualization, as demonstrated in this study, has brought new perspectives for analyzing data generated by mathematical models, which can be used in other areas of knowledge.

References

Branco, P. M. (2014). Elementos que caracterizam o clima. Available in http://www.cprm.gov.br/publique/Redes-Institucionais/Rede-de-Bibliotecas---Rede-Ametista/Canal-Escola/Elementos-Que-Caracterizam-o-Clima-1267.html [accessed in 29-August-2017].

Camargo, A. D., Marin, F. R., Sentelhas, P. C., & Picini, A. G. (1999). Ajuste da equação de Thornthwaite para estimar a evapotranspiração potencial em climas áridos e superúmidos, com base na amplitude térmica diária. Revista Brasileira de Agrometeorologia, 7(2), 251-257.

Di Bello, R. C. (2005). Análise do Comportamento da Umidade do Solo no Modelo Chuva-Vazão Smap II–Versão com Suavização Hiperbólica Estudo de Caso: Região de Barreiras na Bacia do Rio Grande-BA (Dissertação de Mestrado, Universidade Federal do Rio de Janeiro). Universidade Federal do Rio de Janeiro (2005).

Embrapa (2017). Agritempo – Sistema de Monitoramento Agrometeorológico. Available in: em https://www.agritempo.gov.br [Accessed in 20-september-2017].

Evans, J. & Perlman, H. (2015), “The Water Cycle”, U. S. GeologicalSurvey, Available in: http://water.usgs.gov/edu/watercycle.html. [accessed in 29-August-2017].

FAO (2017). Introduction to evapotranspiration. Available in http://www.fao.org/docrep/x0490e/x0490e04.htm [Accessed in 31-july-2017].

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3):37.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology.

Frank, E., Hall, M., and Witten, I. H (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. SIGKDD Explorations, 11(1).

Majidi, M., Alizadeh, A., Vazifedoust, M., Farid, A., & Ahmadi, T. (2015). Analysis of the effect of missing weather data on estimating daily reference evapotranspiration under different climatic conditions. *Water Resources Management*, *29*(7), 2107-2124. Management 29 (7) (2015) 2107{2124.

QGIS Development Team, 2009. QGIS Geographic Information System. Open Source Geospatial Foundation. URL http://qgis.osgeo.org

Xavier, F., Tanaka, A. K., and Revoredo, K. C. (2015). KDD application on Meteorological Data for Identification of Regional Patterns in Estimation of Evapotranspiration. In 30th Brazilian Symposium on Databases Posters Proceedings, pp. 27-32

Xavier, F. (2016). Application of Data Science Techniques in Evapotranspiration Estimation. Dissertation (Master in Informatics). Federal University of the State of Rio de Janeiro, p. 95. 2016.

Ward, Matthew. Grinstein, Georges G. Keim, Daniel. Interactive data visualization foundations, techniques, and applications. Natick, Mass., A K Peters, 2010

WMO (2014) Guide to Meteorological Instruments and Methods of Observation. World Meteorological Organization, WMO- No. 8, 2014.