Chapter 25
Precipitation Interception Modelling Using Machine Learning Methods – The Dragonja River Basin Case Study

L. Stravs, M. Brilly and M. Sraj

Abstract The machine learning methods M5 for generating regression and model tree models and J4.8 for generating classification tree models were selected as the methods for analysis of the results of experimental measurements in the Dragonja River basin. Many interesting and useful details about the process of precipitation interception by the forest in the Dragonja River basin were found. The resulting classification and regression tree models clearly show the degree of influence and interactions between different climatic factors, which importantly influence the process of precipitation interception.

Keywords Precipitation interception · forest hydrological cycle · the Dragonja River basin · machine learning · decision trees · M5 method · J4.8 method

25.1 Introduction

Hydrological science studies the circulation of water in nature, its phenomena, distribution on the earth, movement and physical-chemical characteristics (Chow, 1964). It mainly deals with circulation of water between the atmosphere, surface of the earth and its water systems (Brilly and Sraj, 2000). Forest hydrology studies the circulation of water in forested areas. It studies the course and ways of transition of water from the atmosphere through the forest ecosystem into the ground, groundwater and surface waters and its return back to the atmosphere (Smolej, 1988).

Precipitation is the main source of water in the hydrological cycle. Mostly, it is represented by rain and snow; however, in the coastline and in mountainous,

L. Stravs
Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova 2, SI-1000 Ljubljana, Slovenia, e-mail: lstravs@fgg.uni-lj.si

M. Brilly
Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova 2, SI-1000 Ljubljana, Slovenia, e-mail: mbrilly@fgg.uni-lj.si

M. Sraj
Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova 2, SI-1000 Ljubljana, Slovenia

R.J. Abrahart et al. (eds.), *Practical Hydroinformatics.* Water Science and Technology Library 68, © Springer-Verlag Berlin Heidelberg 2008

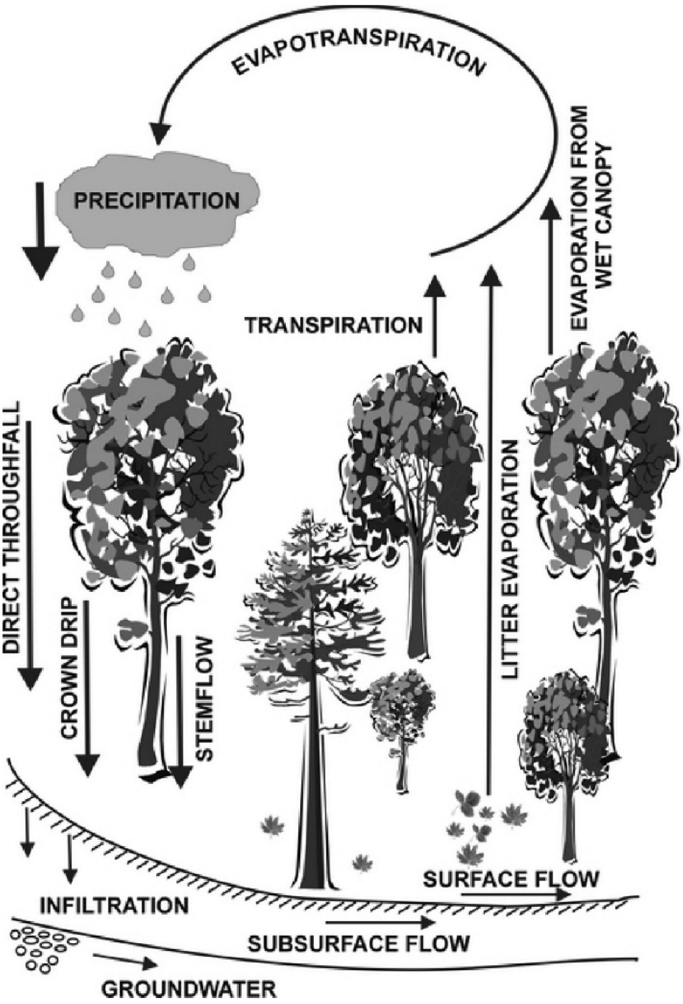


Fig. 25.1 Components of the forest hydrological cycle (Sraj, 2003a)

forested areas horizontal precipitation occurs, i.e. fog. In forested areas (Fig. 25.1) precipitation may be intercepted by the forest canopy and returned to the atmosphere via evaporation, or channeled downwards via throughfall, which is a portion of the rainfall that falls directly through gaps in the canopy or arrives on the ground as crown drip, or stemflow, which is the process that directs a portion of rainfall down tree branches and stems. Precipitation intercepted by the forest canopy (Sraj, 2003b) can be expressed as:

$$\mathrm{E_i} = \mathrm{P} - (\mathrm{T_f} + \mathrm{S_f}) \qquad (25.1)$$

where:

$\mathrm{E_i}$ … precipitation interception;

P … total precipitation amount above the forest canopy;

T_f … throughfall (sum of direct throughfall and crown drip);
S_f … stemflow.

Amount of intercepted precipitation (Mikos et al., 2002) depends on vegetation and climatic factors:

– canopy capacity, which depends upon the class of species, size, shape and vegetational age, area and leaf orientation (coniferous trees intercept 20–40%, and deciduous trees 20–25% precipitation; the higher the vegetation age, the higher the intercepted precipitation (Geiger et al., 1995)).
– vegetational density (interception increases with tree density).
– intensity, duration and frequency of precipitation (smaller intensity or short duration results in a higher evaporation rate from the canopy, intensity of evaporation rate is highest at the beginning of storms, frequently occurring precipitation reduces interception).
– precipitation type (with coniferous species the water equivalent of intercepted snow exceeds the value of intercepted liquid precipitation).
– climate conditions (higher temperatures cause higher evaporation rate, the wind can have high influence on evaporation).
– periods in the course of the year (growing period, dormant period).

Based upon research, Ovington (1954) concluded that the quantity of intercepted precipitation may vary between 6 and 93%, i.e. in different conditions a very different interception rate may be achieved. The two most widely used modelling methods to estimate precipitation interception losses are process-based models of interception and evaporative loss, and empirical or semi-empirical regression models. In the field of precipitation interception modelling Rutter et al. (1971) were the first to move away from a site-specific empirical regression approach to estimate the interception loss. Rutter's model is a numerical model based on the water balance of the canopy and trunks and requires extensive climatic and canopy drainage data. The change in amount of water stored in the canopy is determined by the proportion of the rain that hits the canopy, the drainage from the canopy and evaporation of intercepted water (Schellekens et al., 1999). Gash (1979) proposed a simpler analytical model of precipitation interception based on Rutter's numerical model, in which he considered rainfall to occur as a series of discrete events and assumed for the canopy to have sufficient time to dry between events. Gash's model requires prior estimation of the canopy structure parameters.

In cooperation with the Vrije Universiteit, Amsterdam, extensive research of the hydrological processes in the Dragonja River basin was performed. The Dragonja River basin was chosen as an experimental river basin because intense natural reforestation has been identified in the last decades. This has caused a decrease in minimal and maximal flows of the Dragonja River, while at the same time no noticeable precipitation and temperature regime changes have been identified. The main intention of the research was to analyse the impact of reforestation on the water balance of the entire river basin and to determine the importance of individual climate factors influencing it.

Experimental equipment for measurements of the individual components of the forest hydrological cycle was set up. The machine learning methods M5 for generating regression and model tree models and J4.8 for generating classification tree models were selected as the methods for analysis of the results of experimental measurements in the Dragonja River basin. Successful applications of the machine learning techniques in modelling of hydrological processes like floods, debris flows and other water-related processes are well known (Stravs et al., 2004; Solomatine and Dulal, 2003). While usage of neural networks has already been widely researched and explored in the field of hydrological science (Govindaraju and Ramachandra Rao, 2000), new emerging methods from the framework of artificial intelligence like decision trees, instance-based learning, fuzzy based systems (Stuber and Gemmar, 1997), chaos theory and others have not gained much attention in the field of hydrology yet. Generally, machine learning methods are used for generating forecasting models or for generating descriptive models from which new knowledge about the modelled process can be learned.

25.2 River Basin Characteristics

The Dragonja River basin (Fig. 25.2) with a drainage area of 90.5 km^2 is situated in the southwest of Slovenia, on the Northern part of the Istria Peninsula. It flows from East to West to the North Adriatic Sea (the Piran Bay). On its mouth, there is a RAMSAR protected wetland (Secovlje salt pans), to which also the rivers Drnica and Jernejski potok flow. The Drnica used to be the tributary of the Dragonja, but after regulation of the Dragonja at its outflow to the sea, they became separated. The

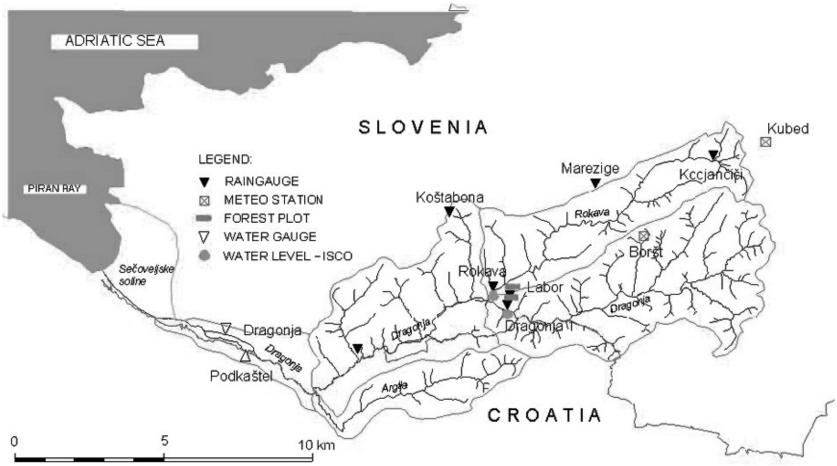


Fig. 25.2 The Dragonja River basin

surface runoff area of the Secovlje salt pans is 142 km^2, extending over Slovenia (116 km^2) and Croatia, the bordering state. Close to the Dragonja outflow to the sea, there are two karstic springs (Buzini and Gabrijeli) with a karstic river basin area. It mostly extends to Croatia (73 km^2 large area). Taking into consideration all river basin sub-units of the Secovlje salt pans, the area has 215 km^2 (Fig. 25.2).

The river basin consists of long flat ridges (up to 400 m a.s.l.) above the deep and narrow river valleys, where the majority of settlements have developed. There are 5860 inhabitants; depopulation has been noticed from 1960 on, but in the late 1990s it stopped (Globevnik, 2001). The area has rural character. The land, typically owned by one family, is traditionally very small; therefore there are hardly any large hillside farms. Larger farms can be found in the river valley at its outflow to the sea. Today, a few new plantation areas on the hills have developed (vineyards, olive groves).

The average annual temperature is 14°C on the coast and 10°C on the continental side. The average yearly precipitation on the sea coast is 900 mm, whereas on the eastern side of the river basin it is 1200 mm. The hydrological characteristics of the Podkastel water station (87 km^2) are (Globevnik, 2001):

– annual mean flow (1971–1995):	1.16 m^3/s;
– autumn high water peaks:	98 m^3/s.

The Slovenian coastal area is well known for its water supply shortages, especially in the summer season when the hydrological conditions are usually quite critical.

25.3 Methods

25.3.1 Measurements

Two forest plots (400 m apart and both at around 200 m a.s.l.) in the 30–35 year-old forest above the confluence of the Dragonja River and the Rokava River were selected as areas where thorough experimental measurements of individual components of the forest hydrological cycle (Fig. 25.1) would be performed; the first plot (1420 m^2) was on the north facing slope in the Rokava River basin and the second one (615 m^2) on the south facing slope in the Dragonja River basin.

Precipitation above the canopy, throughfall and stemflow were measured on both research plots. Rainfall above the canopy was measured with a tipping bucket rain gauge and with a totalizator (manual gauge) for control (Fig. 25.3). Throughfall was measured with two steel gutters in combination with ten manual gauges, which were emptied and moved randomly (Fig. 25.3). Stemflow was measured on two of the most typical species in each plot: on the north plot on oak and hornbeam trees and on the south plot on ash and oak trees.

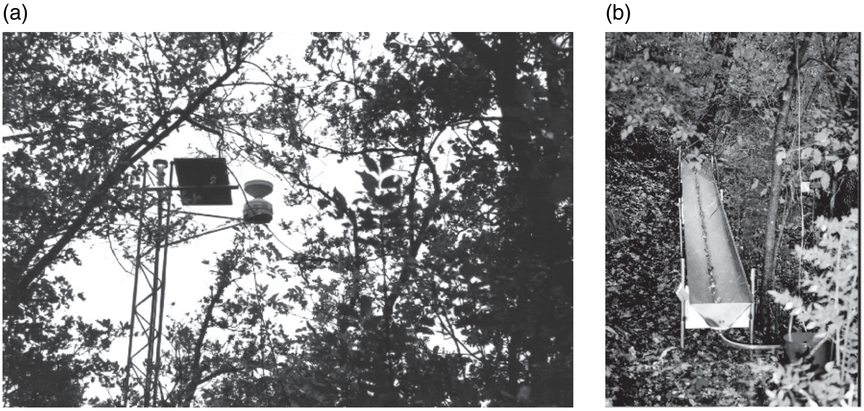


Fig. 25.3 Measurements of precipitation above the canopy (*left*) and throughfall (*right*)

All quantities were measured automatically with a 10-minute log time. Air temperature and relative humidity, wind direction and average wind speed were also measured at the nearby meteorological stations Kubed and Borst.

25.3.2 Modelling Methods

Machine learning generated models are mostly used for forecasting or prediction and for extracting new knowledge about the observed processes. In our case we used the machine learning methods M5 and J4.8 as they are implemented in the WEKA system (Witten and Frank, 2000), developed at the University of Waikato, New Zealand, to generate classification and regression tree models to analyse the impact of reforestation on the water balance of the entire river basin and learn more about the climate and other factors influencing it.

The basic idea of generating tree-like models is to develop simple, transparent models that are easy to use and interpret. The reason behind the choice of the decision trees for modeling the hydrological process of precipitation interception is obvious – we needed the result in the form and structure that can be easily interpreted and the resulting model that can uncover the empirically derived patterns of the underlying process.

By feeding the machine learning method with enough relevant input and output data of the modelled process it can automatically learn the patterns underlying the modelled process from the data only and it can divide the input data space (in machine learning theory called attributes) into subspaces where certain characteristic similarities or patterns exist.

Decision trees are generated through an iterative splitting of data into subspaces of the whole attribute space, where the goal is to maximize the distance between groups at each split (Breiman et al., 1984; Quinlan, 1986, 1992; Kompare, 1995;

Mitchell, 1997; Witten and Frank, 2000; Solomatine and Dulal, 2003). Basic components of a decision tree are the decision nodes, branches and leaves. The decision process starts with the top decision node (also called root node), which specifies a test to be carried out. The answer to this root node test causes the tree to split into branches, each representing one of the possible answers. Each branch will lead either to another subsequent decision node or to the bottom of the decision tree, called a leaf node.

Results of the modelling are decision tree models, which are a way of representing a series of rules that lead to a class value, numerical value or linear equation, and are therefore classified into:

- classification trees with class values as leaves of the model;
- regression trees with constant numerical values as leaves of the model;
- model trees with linear equations as leaves of the model.

25.3.3 Data

In the period of one year 369 events were recorded out of which 173 were recorded on the south plot and 196 on the north research plot. Events were separated by the rainless periods in which canopies could dry up. For each event the following attributes were available:

- plot orientation (North, South);
- rainfall quantity (expressed in mm);
- rainfall duration (hours);
- rainfall intensity (mm per hour);
- average air temperature (°C);
- relative humidity (%); and
- average wind speed (metres per second).

Precipitation above the canopy for single events varied from 0.2 to 100.2 mm, duration of rainfall varied from 5 minutes to almost 40 hours and rainfall intensity varied from 0.15 to 44 mm/h.

25.4 Results

Three decision tree models connecting some of the measured factors influencing the precipitation interception process in the Dragonja River basin and precipitation interception rate were developed.

In case #1 a classification tree (Fig. 25.4) was generated where attributes of each event were: plot orientation, rainfall quantity, duration and intensity, air temperature and relative humidity, and average wind speed. The output data or the modelled variable was precipitation interception percentage (relative to the precipitation amount above the canopy for each event) classified into 7 classes; R_0 meaning 0% interception loss, R_1_20, R_21_40, R_41_60, R_61_80, R_81_99 and R_100 meaning 100% interception loss.

From the resulting model (Fig. 25.4) we can learn that for events with less than 2.4 mm of rainfall and with duration shorter than 10 minutes 100% of the precipitation (class value R_100 on Fig. 25.4) is intercepted. This means that under such conditions no groundwater recharge or surface or subsurface runoff occurs. For events with less than 2.4 mm of rainfall, duration longer than 10 minutes and average temperature of the event less than 14°C, then approximately 50% of the precipitation above the canopy is intercepted. If the average temperature of an event with less than 2.4 mm of rainfall and duration longer than 10 minutes is higher than 14°C, then once again, almost all of the precipitation is intercepted (class value R_81_99 – from 81 to 99%). For events with rainfall amount ranging from 2.4 to 7.0 mm approximately half of the rainfall is intercepted (class value R_41_60 – from 41 to 60%). At events with more than 7 mm of rainfall, average wind power and rainfall intensity also influence the precipitation interception by the forest canopy. It is interesting to note that the generated model does not distinguish the differences in the process of precipitation interception between north and south research plots, which was expected. This could also be a result of different climatic conditions recorded

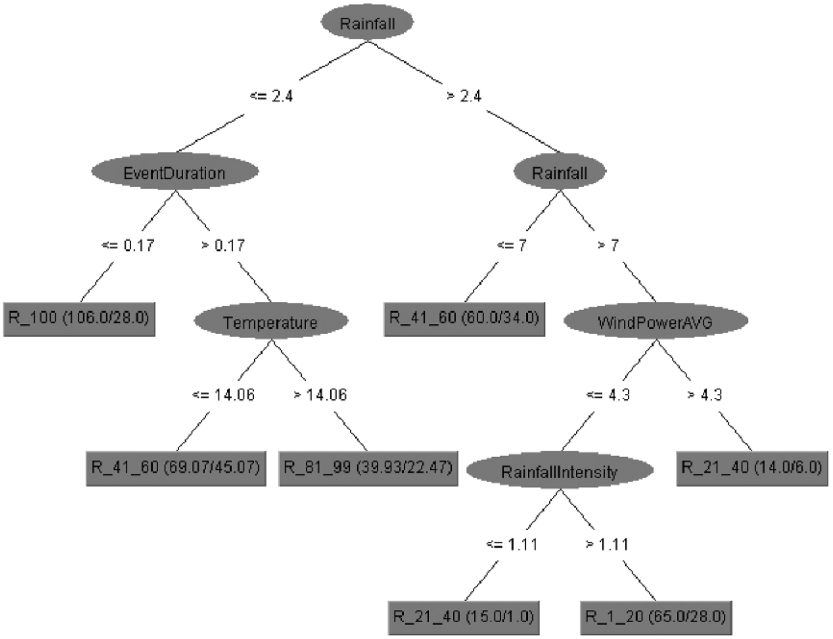


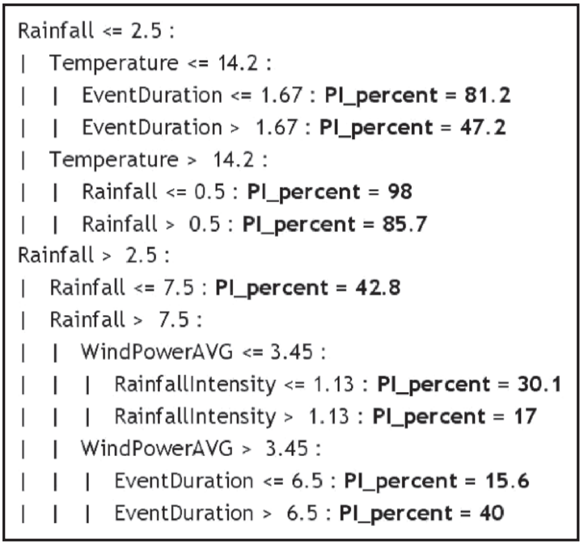
Fig. 25.4 Generated classification tree for case #1 (J4.8 method) – numbers in brackets of each leaf of the classification tree following the class value mean the number of instances that reached the leaf (*left*) and the number of incorrectly classified instances (*right*) in the model evaluation process

on each of the research plots resulting in different values of average air temperature, relative humidity and wind speed actually describing different climate conditions for each of the research plots.

The resulting classification tree (Fig. 25.4) correctly classifies 56% of the instances if it is evaluated on the training set and correctly classifies 48% of the instances in the process of 10-fold cross validation. More accurate classification trees were also generated, which correctly classify up to 65% of the instances if they are evaluated on the training set and correctly classify up to 55% of the instances in the process of 10-fold cross validation. However, they were pruned so that higher structural and explanatory transparency of the resulting models was achieved.

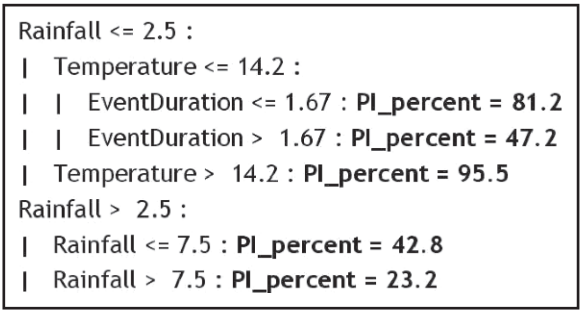
In cases #2 and #3 a regression tree (Fig. 25.5) was generated where attributes of each event were: plot orientation, rainfall quantity, duration and intensity, air temperature and humidity, and average wind speed. The class was the precipitation interception percentage (relative to the precipitation amount above the canopy for each event), this time in the form of a numerical value ranging from 0 to 100%. The difference between the resulting regression trees for cases #2 and #3 is in the complexity of the resulting regression tree (Figs. 25.5 and 25.6).

From the resulting models, especially from the pruned regression tree of case #3, we can learn that for events with less than 2.5 mm of rainfall and air temperature lower than 14.2°C, 81.2% of the rainfall is intercepted by the forest canopy if the event is shorter than 1.67 hours, and 47.2% of rainfall is intercepted if the event is longer than 1.67 hours. But if the temperature of the event with less than 2.5 mm of rainfall is higher than 14.2°C almost all precipitation is intercepted (95.5%). For events with more than 2.5 and less than 7.5 mm of rainfall, 42.8% of precipitation



```
Rainfall <= 2.5 :
|   Temperature <= 14.2 :
|   |   EventDuration <= 1.67 : PI_percent = 81.2
|   |   EventDuration >  1.67 : PI_percent = 47.2
|   Temperature >  14.2 :
|   |   Rainfall <= 0.5 : PI_percent = 98
|   |   Rainfall >  0.5 : PI_percent = 85.7
Rainfall >  2.5 :
|   Rainfall <= 7.5 : PI_percent = 42.8
|   Rainfall >  7.5 :
|   |   WindPowerAVG <= 3.45 :
|   |   |   RainfallIntensity <= 1.13 : PI_percent = 30.1
|   |   |   RainfallIntensity >  1.13 : PI_percent = 17
|   |   WindPowerAVG >  3.45 :
|   |   |   EventDuration <= 6.5 : PI_percent = 15.6
|   |   |   EventDuration >  6.5 : PI_percent = 40
```

Fig. 25.5 Generated regression tree (M5 method) – more complex (case #2) regression tree



```
Rainfall <= 2.5 :
|   Temperature <= 14.2 :
|   |   EventDuration <= 1.67 : PI_percent = 81.2
|   |   EventDuration >  1.67 : PI_percent = 47.2
|   Temperature >  14.2 : PI_percent = 95.5
Rainfall >  2.5 :
|   Rainfall <= 7.5 : PI_percent = 42.8
|   Rainfall >  7.5 : PI_percent = 23.2
```

Fig. 25.6 Generated regression tree (M5 method) – pruned regression tree (case #3)

is intercepted, and for events with rainfall amount higher than 7.5 mm 23.2% of rainfall is intercepted. The obtained regression tree of case #2 unveils some additional specific details for events with more than 7.5 mm of rainfall when wind speed becomes an important factor. If the average wind speed for such events is less than 3.5 m/s, approximately 10% less water is intercepted compared to events with average wind speed higher than 3.5 m/s.

25.5 Conclusions

Many interesting and useful details about the process of precipitation interception by the forest in the Dragonja River basin were found. In most cases, rainfall events with up to approximately 2.5 mm of rainfall contribute almost nothing to the recharge of groundwater at the Dragonja river basin and the Dragonja river discharge, with the exception of rainfall events longer than 1.67 hours and temperature lower than 14°C. The generated models also show that approximately 23 to 43% of the water at events with more than 2.5 mm of rainfall is intercepted by the forest as a direct consequence of natural reforestation in the last few decades.

The classification and regression tree models clearly show the degree of influence and interactions between different climatic factors, which importantly influence the process of precipitation interception. If the results obtained on both research plots are representative enough for the whole river basin where the process of reforestation occurred in the last decades, we can conclude that the impact of the land use change on the water balance of the Dragonja River basin is quite significant. In terms of water supply approximately one third of the water is lost in the river basin areas, which are covered by the forest.

The generated models captured the important properties of the processes of the forest hydrological cycle at the Dragonja River basin. Results were in the context of what was expected and known about the precipitation interception process. Furthermore, many significant details about the process in this particular river basin were uncovered in a really short modelling time. We can conclude that the usage of machine learning methods for generating descriptive models like decision trees reduces the manpower and time spent in the process of extracting new knowledge about the processes that were measured and studied.

Usage of machine learning methods like decision trees for generation of structurally transparent and explanatory models from the data has offered great promise in helping scientists to uncover patterns hidden in their data. However, the development of models is only one of the steps in the acquisition of new knowledge; the selection, collection and preparation of data, the guidance of the model development and the interpretation of the generated models by the scientists who understand the modelled processes are equally important.

References

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadworth, Belmont

Brilly M, Sraj M (2000) Osnove hidrologije (Principles of Hydrology). University Textbook, University of Ljubljana, Faculty of Civil and Geodetic Engineering (in Slovene)

Chow VT (1964) Handbook of applied hydrology. McGraw-Hill, New York

Gash, JHC (1979) An analytical model of rainfall interception by forests. Quarterly Journal of the Royal Meteorological Society 105: 43–55.

Geiger R, Aron RH, Todhunter P (1995) The climate near the ground. Friedr. Vieweg & Sohn, Braunschweig/Wiesbaden

Globevnik L (2001) Celosten pristop k urejanju voda v povodjih (Intergrated approach to water resources management on the catchment level). Ph. D. Thesis, University of Ljubljana (in Slovene)

Govindaraju RS, Ramachandra Rao A (2000) Artificial neural networks in hydrology. Kluwer Academic Publishers, Dordrecht, Netherlands

Kompare B (1995) The use of artificial intelligence in ecological modelling. Ph. D. Thesis, Royal Danish School of Pharmacy, Copenhagen, Denmark

Mikos M, Kranjc A, Maticic B, Müller J, Rakovec J, Ros M, Brilly M (2002). Hidrolosko izrazje – Terminology in hydrology. Acta hydrotechnica 20/32: 3–324

Mitchell T (1997). Machine learning. MIT Press and The McGraw-Hill Companies, Inc

Ovington JD (1954) A comparison of rainfall in different woodlands. Forestry London 27, pp 41–53

Quinlan JR (1986) Induction of Decision Trees. Machine Learning 1: 81–106

Quinlan JR (1992) Learning with continuous classes. In: Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence, pp 343–348

Rutter AJ, Kershaw KA, Robins PC, Morton AJ (1971) A predictive model of rainfall interception in forests, Derivation of the model from observations in a plantation of Corsican pine. Agricultural Meteorology 9: 367–383

Schellekens J, Scatena FN, Bruijnzeel LA, Wickel AJ (1999) Modelling rainfall interception by a lowland tropical rain forest in northeastern Puerto Rico. Journal of Hydrology 225: 168–184

Smolej I (1988) Gozdna hidrologija (Forest hydrology). In: Rejic M, Smolej I (eds) Sladkovodni ekosistemi, varstvo voda in gozdna hidrologija (Freshwater ecosystems, water conservation, and forest hydrology). University of Ljubljana, Biotechnical Faculty, Forestry Departement, pp 187–225 (in Slovene)

Solomatine DP, Dulal KN (2003) Model trees as an alternative to neural networks in rainfall-runoff modelling. Hydrological Sciences Journal 48: 399–411

Sraj M (2003a) Estimating leaf area index of the deciduous forest in the Dragonja watershed – Part I: Methods and measuring. Acta Hydrotechnica 21/35: 105–127

Sraj M (2003b) Modeliranje in merjenje prestrezenih padavin (Modeling and measuring of rainfall interception). Ph. D. Thesis, University of Ljubljana (in Slovene)

Stravs L, Kobold M, Brilly M (2004) Modeli kratkorocnih napovedi pretokov visokih voda na Savinji (Short-term flood forecasting models for the Savinja River). Proceedings – Misicev vodarski dan, Maribor (in Slovene)

Stuber M, Gemmar P (1997) An approach for data analysis and forecasting with neuro fuzzy systems – demonstrated on flood events at river Mosel. In: Proc. International Conference on Computational Intelligence, 5th Fuzzy Days, Dortmund

Witten IH, Frank E (2000) Data mining: Practical machine learning tools and techniques with java implementations. Morgan Kaufmann Publishers, San Francisco, USA