Journal of Environmental Geography 13 (1–2), 53–61.

DOI: 10.2478/jengeo-2020-0006
ISSN 2060-467X

IDENTIFICATION OF RIPARIAN VEGETATION TYPES WITH MACHINE LEARNING BASED ON LIDAR POINT-CLOUD MADE ALONG THE LOWER TISZA'S FLOODPLAIN

István Fehérváry[1]*, Tímea Kiss[2]

[1]Lower Tisza Hydrological Directorate, Stefánia 4, 6720 Szeged, Hungary
[2]University of Szeged, Department of Physical Geography and Geoinformatics, Egyetem str. 2-6, 6722 Szeged, Hungary
*Corresponding author, e-mail: FehervaryI@ativizig.hu

Research article, received 7 April 2020, accepted 5 May 2020

Abstract

The very dense floodplain vegetation on the artificially confined floodplains results in decreased flood conveyance, thus increase in flood levels and flood hazard. Therefore, proper floodplain management is needed, which must be supported by vegetation assessment studies. The aims of the paper are to introduce the method and the results of riparian vegetation classification of a floodplain area along the Lower Tisza (Hungary) based on automatized acquisition of airborne LiDAR survey. In the study area 15x15 m large training plots (voxels) were selected, and the statistical parameters of their LiDAR point clouds were determined. Applying an automatized parameter selection and 10-fold cross-validation he most suitable decision tree was selected, and following a series of classification steps the training plots were classified. Based on the decision tree all the pixels of the entire study area were analysed and their vegetation types were determined. The classification was validated by field survey. On the studied floodplain area the accuracy of the classification was 83%.

Keywords: airborne LiDAR, scikit-learn, Gini impurity, decision tree, riparian forest,

INTRODUCTION

During the last one and half century several environmental effects (e.g. climate change, land cover alterations) affected the floodplains and river channels altering their characteristics. However, these semi-natural effects were exceeded by the consequences of various river engineering works: the channel and floodplain regulations works altered the hydrological processes, and as one of the consequences the riparian vegetation changed too.

The riparian vegetation highly influences the channel-floodplain connections. For example, the vegetation along a river stabilises the banks (Abernethy and Rutherfurd, 1998), or decreases the overbank flow velocities (Kiss et al., 2019a). Along the river-banks the density of vegetation primarily influences the development of natural levees (Nagy et al., 2018), whilst in the distal floodplain areas the vegetation influences the flood flow directions and velocities (Rátky and Farkas, 2003; Zellei and Sziebert, 2003; Brooks, 2005; Corenblit et al., 2007; Geerling et al., 2008), thus the vertical aggradation pattern (Steiger et al., 2001; Kiss and Sándor, 2009).

These processes are related to the roughness controlling function of the vegetation, which influences the flood stages too (Jalonen et al., 2015; Kiss et al., 2019a). The highest peak flood level on record was set in 1970 along the Tisza River, Hungary, however along the Lower Tisza this record was overprinted in 2000 and also in 2006 by higher stages of 80 cm (Kiss et al., 2019b), while the discharge of these record floods were lower than in 1970 (Kovács and Váriné Szöllősi, 2003). These hydrological changes draw attention to flood conductivity decline of the floodplain: nowadays only 13% of the flood discharge is drained on the floodplain while it was 23% in 1970 (Kovács and Váriné Szöllősi, 2003). It could be partly related to the very dense floodplain vegetation, therefore proper floodplain management is needed, which must be supported by vegetation assessment studies. Their first step should be the identification of riparian vegetation types; therefore, our aim is to apply the latest methods for this identification using automatized acquisition of LiDAR survey data.

Lately several researches relied on the statistical analysis of point-clouds of LiDAR surveys to identify vegetation characteristics and types. Most of these researches were made in the field of forestry or ecology. For example, Hudak et al. (2008) identified various tree species in forest patches based on various statistical parameters of the LiDAR points representing the canopy. Heurich and Thoma (2008) did similar research, but they measured some dendrological parameters too (e.g. tree height, canopy width) based on LiDAR data. Naesset et al. (2004) combined airborne and terrestrial LiDAR data to calculate the main parameters (i.e. number of stems, volume of harvestable wood) of forest units. Jung et al. (2011) calculated not only the parameters of trees (i.e. tree height, lower canopy height, canopy volume, stem diameter), but also the temporal changes between two survey campaigns including both LiDAR technologies. Though these researches slightly differ in the identification of vegetation types and in defining their parameters, it could be concluded, that some high-

resolution parameters could be applied just on small areas or on individual level, while the measurements on larger areas have limited resolution and less accuracy.

During the last years the LiDAR based vegetation analysis has been applied also in hydrological studies. For example, Vetter et al. (2011) determined the vegetation roughness based on the spatial connection of voxels (3D pixels) and the rate of reflections, using an airborne LiDAR survey with very high point density (>25 point/m^2). The resulted vegetation density values were applied in a 2D hydraulic model, and the derived hydrological data were compared to modelled data based on classical land-use category maps. Vetter et al. (2011) concluded, that the LiDAR based modelling gave much more reliable results, as the hydrological data were closer to the actually measured ones. Similarly, Manners et al. (2011) determined the role of Tamarix bushes in vegetation roughness using terrestrial LiDAR survey.

During the late 20th century forest became the dominant land cover type along the Tisza River and invasive species became widespread in the undergrowth, thus the vegetation roughness of the floodplain drastically increased, as it was indicated by point-based classical vegetation surveys (Kiss et al., 2019a). As the high vegetation roughness fundamentally decreases the overbank flow velocities and increases flood peak levels, high-resolution and up-to-date data would be needed for precise flood modelling. However, no such a dataset exists.

Therefore, our aims are to introduce the method and the results of riparian vegetation classification, based on automatized acquisition of airborne LiDAR survey. Within this article our goals are to describe the detailed methodology of the automatized classification, to classify the vegetation on a study area located along the Lower Tisza River, and to evaluate the feasibility of the method.

STUDY AREA

The research was conducted on a 3 km^2-large floodplain area of the Lower Tisza (197-194 fkm) between the settlements of Algyő and Szeged (Fig. 1). The floodplain is artificially confined by artificial embankments to 800 m; and the river channel is 130 m wide in average. At Algyő gauging station the greatest flood-stage was at 84.65 m asl, while the lowest stage was measured at 71.55 m asl, thus the absolute change in water level is ca. 13 m. The flooding of the floodplain starts when the water is at 80 m asl, and up to 5-6 m deep water column develops over the floodplain at the time of record floods. The slope of the water is very low (2.9 cm/km), therefore the overbank flow velocity is also low (max. 0.1-0.2 m/s).
In the mid-19th century the river regulation works (e.g. artificial cut-offs, building artificial levees) resulted in considerable land-use changes of the floodplain (Kiss et al. 2019a). At the beginning of the 20th century the former wetlands were replaced by meadows, pastures and plough fields, while the proportion of forests remained low. However, as the result of intensive afforestation in the 1970-80s the proportion of forests increased above 70% in the 1980s, and nowadays it is above 80%. These land-cover changes resulted in fourfold increase in vegetation roughness (from 0.02 to 0.08), which is even higher (0.13) if the dense stands of invasive species in forest and on fallow lands are considered. Among the invasive species the *Amorpha fruticosa* is the most abundant (11%), which creates impenetrable shrubbery. According to our latest

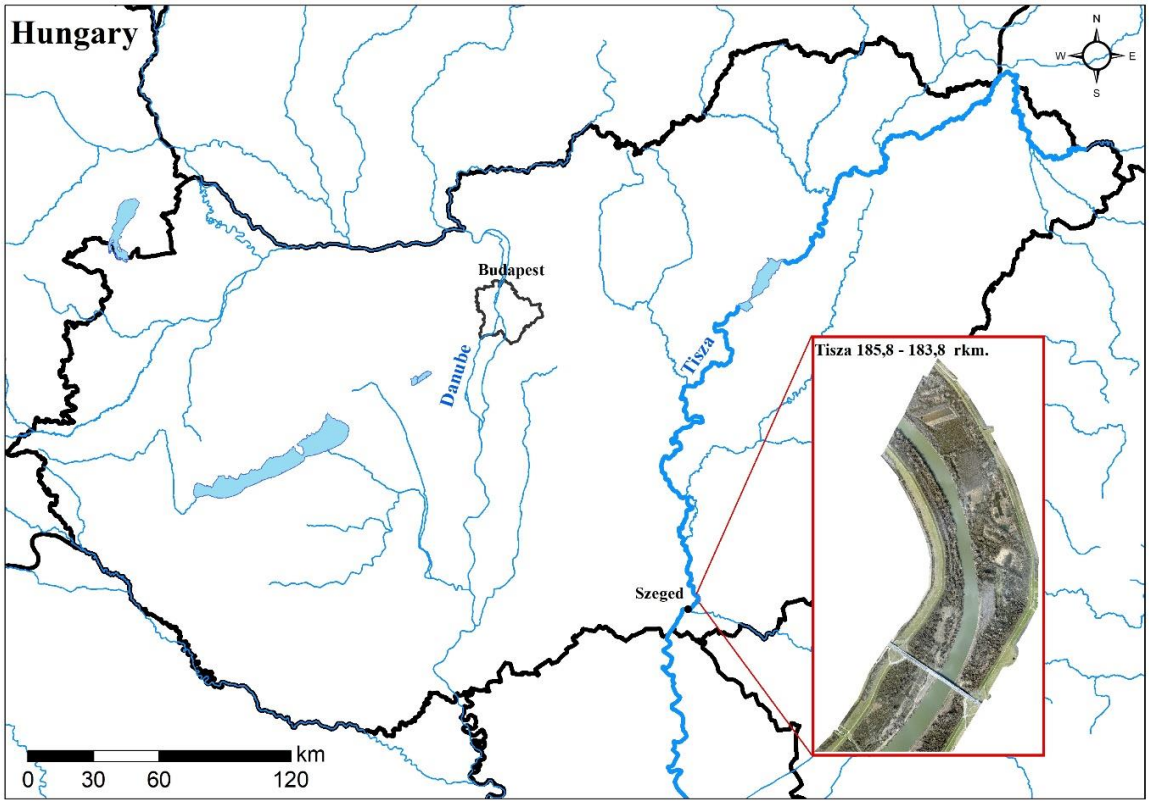

Fig. 1 The study area is located along the Lower Tisza, north of the city of Szeged

results its thickets decrease the flood flow velocity to one third, creating 20-30 cm increase in flood levels at our study area (Kiss et al., 2019a).

DATA SOURCE

The analysed point cloud is based on a LiDAR (full-waveform) survey was made at the early spring of 2015 during low stage of the Tisza River. There were no leaves on the trees, thus the canopy structure of the individual trees is nicely visible on the survey. Simultaneously to the LiDAR survey an ortho-photo was also made with 10 cm resolution. The study area is represented by 22.5 million reflected points on the airborne LiDAR survey, which are stored in eight .las files. The lowest points represent the bank of the river at the actual water-surface (at 75.1 m asl), while the highest indicate the topmost points of the 30-35 m high trees. The water surfaces (e.g. open water, wetlands) do not reflect the LiDAR beams, therefore, these areas were excluded from the analysis. After this exclusion the point-density of the study area is 9 point/m^2, thus it could be considered as suitable for the analysis, as according to Laes et al. (2008) the analysis of forests requires at least 4 point/m^2 point-density. Only 3% of the study area does not fulfil this requirement, thus the quality of the survey was good for further analysis. The digital elevation model (DEM) made of the LiDAR data has 0.5 m resolution.

METHODS

For the analysis the Fusion 3.8 and ArcMap10.6.1 software were used, while the algorithm of the vegetation classification (decision tree) was written in Python using the scikit-learn (0.22.1) library (see Pedregosa et al., 2011). The decision tree were generated with *DecesionTreeClassifier* class-based on the Gini impurity (Grabmeier and Lambe 2007) which is in the *sklearn.tree* module. To find the most ideal and relevant parameters for the decision tree algorithm, the *GridsearchCV* class was applied using K-fold cross validation method. The detailed description of the decision tree can be found in the Results 5.1. chapter. Based on the results of the decision tree the vegetation types of all pixels in the study area were determined automatically, and finally the results were checked and validated based on a field survey.

Data preparation

As the DEM was in a *.flt* format, it had to be converted to *.dtm*, thus it could be used in the Fusion software for further calculations. As a first step the DEM was exported in ArcMap software to .ascii format using the *RastertoAscii* tool, then in Fusion it was converted to *.dtm* applying the *ASCII2DTM* tool.

In the next step the quality assessment of the LiDAR point-clouds stored in *.las* files had to be done, analysing the extreme values and the number of reflections. The quality check was made in Fusion software applying the *Catalog* tool. As a result, the software made a quality report of each .las file, including the number of points, their minimum and maximum heights, and the point-density (point/voxel). Using the *FilterData* tool the extreme values were deleted from the file.

Determine the spatial resolution

To select the most suitable spatial resolution is crucial point of the research. If it is too high, the point cloud will be over dissected, and the typical parameters of a given vegetation type could not be distinguished. However, if the spatial resolution is too low, the spatial differences could disappear and there will be a greater chance to have mixed classes. Laes et al. (2008) suggested, that the spatial resolution should be fitted to the mean canopy width. Therefore, in the study area 15x15 m spatial resolution was selected, thus the point-cloud was split into voxels with 15x15 m cell-size, and the height of the voxels was determined by the highest point of the vegetation.

Calculation of statistical parameters (metrics)

In the following step the statistical parameters of the LiDAR point-clouds representing the vegetation were calculated applying 15 m resolution for the voxels. The calculation was made by Fusion software using its *GridMetrics* tool. The input data included the filtered and quality-checked point-cloud and the DEM in *.dtm* format. In the programme a *heightbreak* could be set, thus the program could recalculate some statistical parameters of the voxels split into two parts by a given elevation (for example the proportion of reflected points above a given height). In our case we had selected 6.0 m as a height brake, and the voxel parameters for the forests were calculated above this value. This value was selected because (i) the bushes (especially invasives) are never taller than 6 m, and we wanted to classify the vegetation regardless of the rate of invasives in the underwood; and (ii) this is the limit of the overbank flood height. The calculated statistical parameters of each voxel were stored in .csv files. The programme provided 74 parameters for each voxel. Not all these parameters are introduced in this paper (following McGaughey 2018), only those, which were used to classify the vegetation types of the study area applying the decision tree.

The *canopy relief ratio* (CRR) was calculated as the ratio of the difference between the mean and minimum heights and the maximum and minimum height of the points of a given voxel [(mean-min)/(max-min)]. Thus, the greater the difference is between the mean and maximum values, the CRR is lower. This parameter refers to the spreading of the canopy: in case of large and wide canopy the mean and maximum values have relatively small differences. At the study area the old white poplars have huge canopy with 0.2-0.3 CRR values, whereas the young planted black poplars have slender canopy with 0.03-0.04 CRR values. The open surfaces (e.g. short grasslands) have the highest CRR values (0.4-0.5), as their mean and the maximum height values are almost similar.

The standard deviation of the height values of a voxel (*Elev_std*) refers to the diversity of the points, thus to the vertical dissection and density of the canopy: the flat and at a given elevation dense canopy is reflected by almost homogenous point-cloud at a given elevation of

the voxel, thus it is characterised by low standard deviation. For example, in the study area the open surfaces, where almost every point represents the land or the grass, the standard deviation is 0.03 m, while in case of the riparian willow stands the standard deviation is higher by two orders (3-4 m). The greatest standard deviation (8-10 m) was measured at the *Populus alba* stands which has variegated canopy.

The 99th percentile value for a voxel (Elev_P99) refers to that elevation, where the height values reach the 99% from the ground, thus it refers almost to the maximum height of the voxel. For its calculation the height values of the point-cloud of a voxel are ordered, and that value is selected which represents the 99% of the dataset. For example, on the study area this value is much lower for the willow stands (~15 m) than for the white poplar trees (~25 m).

The 95th percentile value for a voxel (Elev_P95) refers to that elevation, where the height values reach the 95% from the ground. It is calculated similarly as the previous parameter (Elev_P99). This parameter is useful during the calculations, because it is a good indication of the maximum height of the vegetation, however it does not contain points with survey errors.

Skewness of the heights for the points in the voxel (Elev_skewness) provides some indication of how asymmetric the distribution of the values is. In case of symmetric (standard normal) distribution the skewness is zero. The skewness is influenced by abundant and extreme values: if they are at low values, then the skewness will be negative, while in case of higher values it will be positive. In the study area high (4-6) and positive skewness characterises the young poplar plantations, and the lonely and slim trees: in their case the canopy is not perfectly closed, therefore high proportions of the reflected points are from the ground or from the top of the canopy. In case of grasslands the skewness is low (0.1-0.4), as no extreme values are present, the reflected points originate from almost a flat surface.

Selection of training-plots and definitions of vegetation types

Before the training plots were selected, the main vegetation types had to be determined at the study area. The following categories were identified based on our preliminary field survey and the Forestry WebMap of Hungary (https://erdoterkep.nebih.gov.hu) (1): open surface, *Amorpha* thicket, young poplar plantation, poplar plantation, riparian willow forest, and riparian poplar forest with *Populus alba*.

Based on our field-survey and the available ortho-photo we had selected 15x15 cell-sized pixels with homogenous vegetation as training plots. They were selected for each vegetation types, at least 40-50 cells per type. During the selection of training plots we aimed to select homogenous pixels, thus the pixel should not be affected by side effects of other vegetation types. The selection of the training plots was supported by the ortho-photo providing idea on the character of the cell, while the point-cloud of the voxel gave idea about the height conditions of the vegetation and the shape of the canopy (Fig. 2).

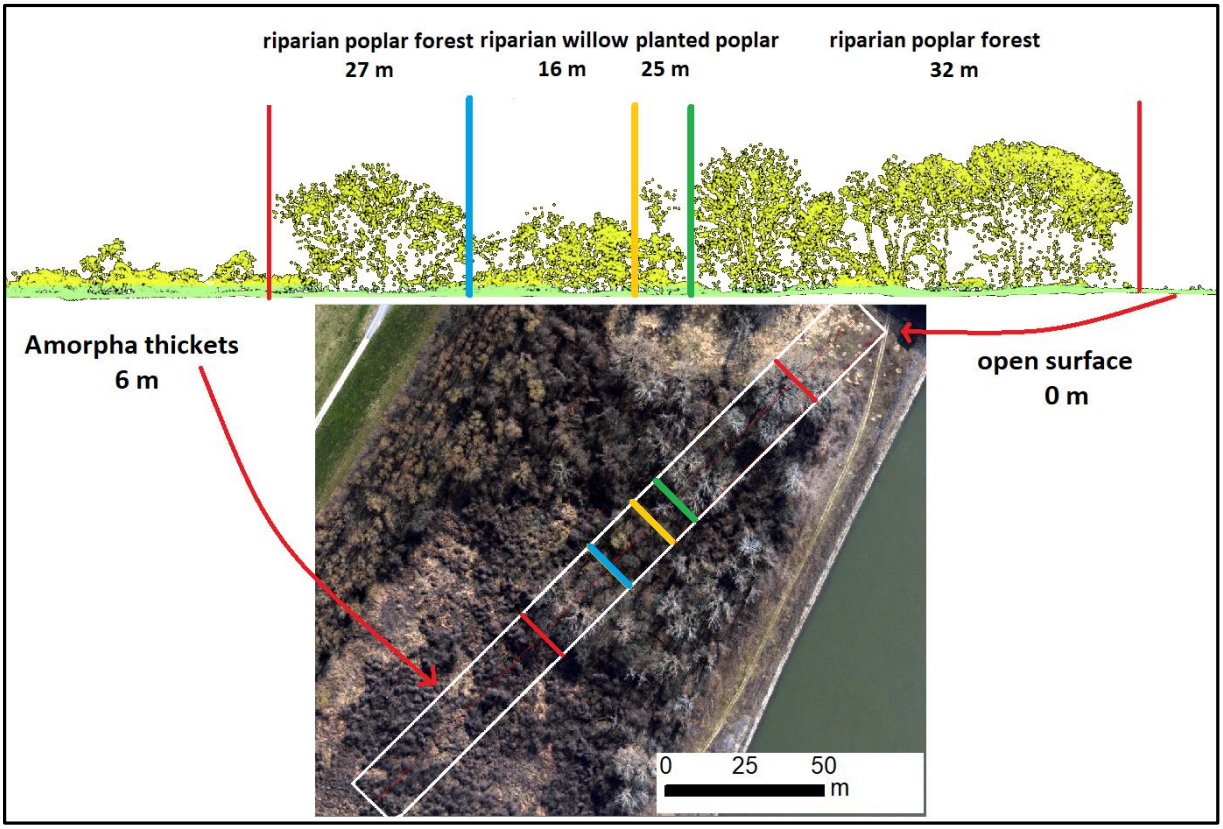

Fig. 2 Various vegetation types represented by the vertical view of the LiDAR point-cloud, and its appearance on the ortho-photo

Grasslands and open surfaces are mostly located on the artificial levees. The LiDAR survey was made in early spring when the grass was very short and dead, therefore the reflected points actually represent the land surface, therefore here the height of the vegetation is almost 0 m. The dense *Amorpha thickets* have height points at 1-6 m, the top of the canopy is almost flat, and great proportion of the LiDAR point-cloud is at the upper part of the canopy. The *riparian willow forests* are high (16-20 m), and on the ortho-photo the almost bunding willow-branches have brownish-orange colour. The *poplar plantations* are planted in rows, thus the skewness of their point-cloud is high, as lots of points are reflected from the ground and also from the top of the canopy. On the LiDAR point-clouds their canopy structure is quite specific, as their branches are thin, most of the points are around the stem, and on the ortho-photo the top of these plantations is quite homogenous. Within the poplar plantations we distinguished the group of *young poplar plantations*. Their trees are shorter (>5.6 m), and more points are reflected from the ground due to their undeveloped canopy. This forest type also includes lonely trees with sparse bushes. The *riparian poplar forest* patches are characterised by tall and easily distinguishable *Populus alba* trees. The white poplar has greyish-white branches on the ortho-photo, and it has very special canopy structure both on the ortho-photo and the LiDAR point-cloud (Fig. 2).

Creating the decision tree

The statistical parameters of the training plots were saved in a *.csv* file. In this file the names of vegetation types and all of the statistical parameters were given. The next step was the parametrisation of the decision tree algorithm. The decision tree classifies the elements – in our case the attributes of the voxels – based on the series of classification steps, aiming to have the most homogenous classes. The selection of attributes and the thresholds between the classes are based on a calculation algorithm. In our case the threshold values were based on the Gini impurity, because it could be run faster than the entropy based calculations, besides, there is no qualitative difference between the accuracy of these classifications. The Gini impurity refers to the probability of the classification of an element to a wrong class (Grabmeier and Lambe, 2007). If its value is zero, it means that the given selection criteria perfectly divided a class from the main population, while 1.0 refers to a totally diverse class.

The setting of the parameters of the decision tree was made automatically applying the *GridsearchCV* module, considering the (i) maximum depth of the decision tree referring to the number of decision levels; (ii) the minimum element number of the leaves; and (iii) the minimum element number for split. To precisely determine the above mentioned settings of the decision tree, each setting was set to an interval (e.g. decision tree depth: 1-10; minimum element number of leaves: 2-10; minimum split: 2-20), and finally best setting combination was selected, which resulted in the most precise decision tree.

To check the accuracy of the decision tree algorithm, a cross-validation method was applied, which is very common at automatized learning technologies. We had selected the method of *10-fold cross-validation*. As a first step the dataset (training plot voxels) were divided to 10 groups, and one of them was selected for validation by the algorithm. The cross-validation lasts for 10 iterations, until every group will be used exactly once as a training set (Bengio and Grandvalet 2004). To estimate the accuracy, the average of 10 results were needed. The advantage of this kind of cross-validation is that each point (voxel) of the dataset will be used for automatic learning and for validation too, however its disadvantage is that it is a quite long process, as in our case the automatic learning was repeated 10 times.

The accuracy of the final classification based on the decision tree is expressed in percentage, referring to the rate of well-classified elements, though this value does not refer to the efficiency of the classification. During the application of the decision tree two kind of methodological mistakes could be made. In the first case the classification process and the applied thresholds do not determine the classes homogenously, thus the decision tree will have low accuracy and it is underfit, however if there are lots of data and several parameters, this case is quite rare. Much more often the decision tree will be overfit, thus the selection criteria will be too specific and valid just for some elements of the dataset. In this case the accuracy is very high (>95%), however the decision tree could not be applied on other datasets (Schaffer 1993).

The aim of the automatized parameter selection and of the *10-fold cross-validation* was to find the most suitable decision tree, which eliminates the errors of the overfitting. During the cross-validation runs it became obvious, that the decision tree is greatly influenced by its depth. Our results suggest, that if the depth of the decision tree is greater than 4, the accuracy won't be considerable better, however the risk of overfitting increases (Fig. 3). Therefore, the depth of decision tree was determined to be 4. The minimum element number and minimum split number was determined to be 2. In this way the accuracy of the created decision tree based on the *10-fold cross-validation* was 92%.

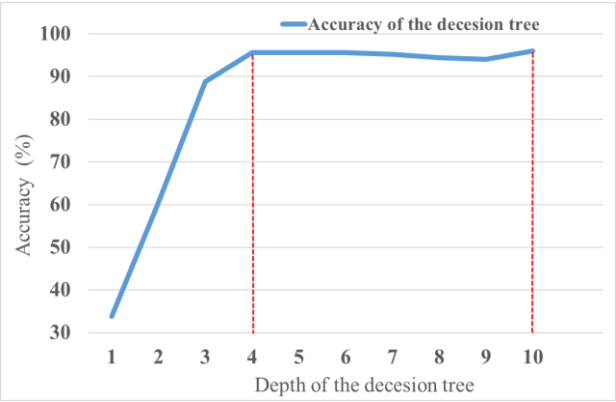

Fig. 3 Relationship between the depth of the decision tree and the accuracy of the classification

Fitting the decision tree to the entire study area and its validation

The decision tree was created based on training plots. In the next step, based on the decision tree all the 11656 pixels of the study area were analysed and their vegetation types were determined. The result and precision of the automatized classification were determined by field-survey. For the field validation aerial photos made by DJI Phantom III Pro drone were taken on 72 cells. The survey was made at 30-60 m height from orthogonal position (applying 90º camera axis). On these low-aerial photos the vegetation types of the cells were identified by expert judgement. The geo-coordinates of the photos were extracted; thus the identified vegetation types could be compared to the automatized classification of the same cell. The results of the comparison were evaluated applying a confusion matrix. The titles of columns and rows in the matrix refer to the vegetation categories. In the main diagonal line, the numbers refer to the proportion of precisely classified cells, while the other cells refer to the proportion of false vegetation classes.

RESULTS

The decision tree

The decision tree automatically selected the parameters of the voxels (see 4.5. chapter) for the identification of vegetation types (see 4.3. chapter). First, the *young poplar plantations* were selected based on *CRR≤0.039* criteria. This parameter well sunders the young and short poplar trees with undeveloped canopy from the higher trees with more complex canopy and from the grassland/open surfaces. The next step followed the false (no-)branch of the decision tree (*CRR≥0.039*). Here, based on the standard deviation of the voxel's point-cloud the *Amorpha* thickets and the grassland/open surfaces were identified. The two vegetation classes could be divided based on their height (*Elev_P99*). To identify the *grassland/open surfaces* the voxels had to be fulfil the following criteria: *CRR≥0.039* and *Elev_std≤1.783,* and *Elev_P99≤2.119*. The *Amorpha thickets* were identified by *CRR≥0.039* and *Elev_std≤1.783,* and *Elev_P99>2.119*. The Gini impurity (0.0) of these three categories reflects that they were identified with the greatest accuracy (Fig. 4.).

On the false (no-)branch of the standard deviation (*Elev_std≥1.783*) of the decision tree the older poplar plantations, the riparian poplar forests and riparian willow forests remained. On the floodplain the riparian poplar forests are characterised by tall *Populus alba* trees, thus they could be selected based on their height conditions (*Elev_P95>17.987*). However, this selection criterion is not totally clear, as some voxels with tall planted poplars also fall into this class. The natural and planted poplar forest could be divided based on the CRR parameter: the poplar plantations have less complex canopy than of the natural poplars, therefore the plantations have smaller CRR values, thus their selection criterion is *CRR≤0.103*. The *riparian poplar forests* were selected following *Elev_std>1.783* and *Elev_P95>17.987* and *CRR>0.103*. Some of the poplar plantation's cells (with older and higher trees) also fall into this class, only their CRR values differed (*Elev_std≥1.783* and *Elev_P95>17.987* and *0.039<CRR≤0.103*). Based on the tests the tall (≤18 m) and *old poplar plantations* were clearly separated (Gini impurity = 0.0) from the riparian poplar forests with *Populus alba* (Fig. 4.).

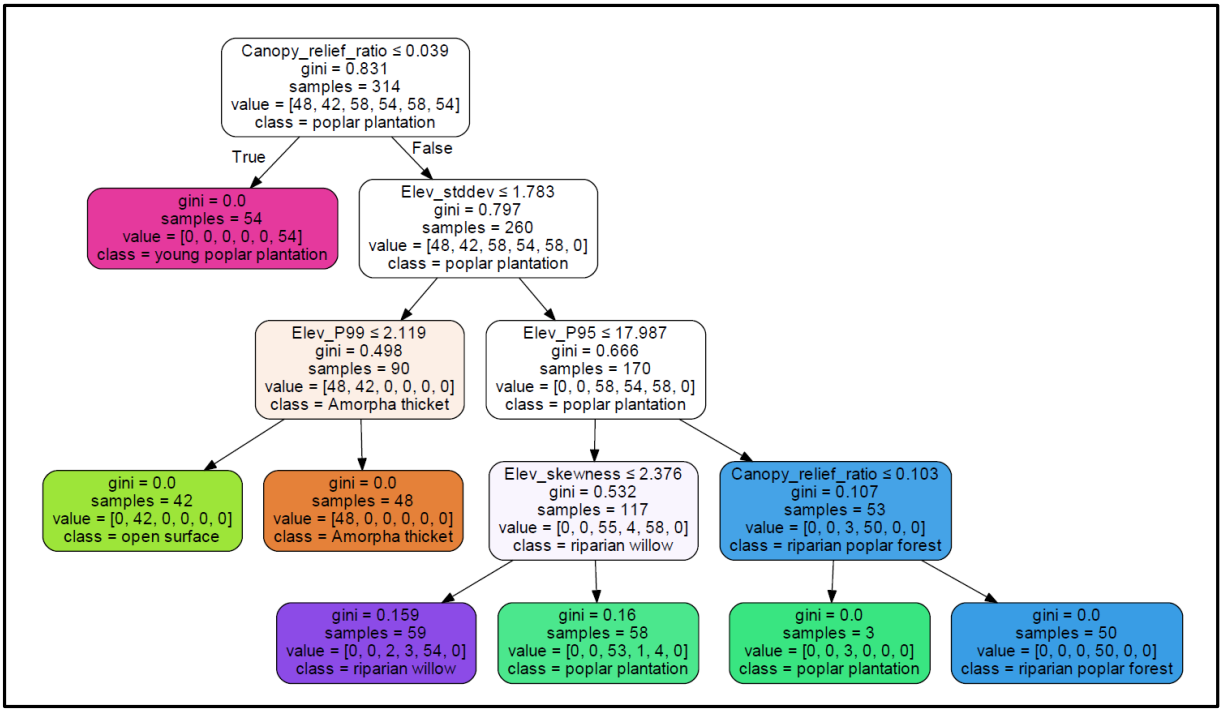

Fig. 4 Decision tree built up based on training plots of the study area. This decision tree was applied to classify the vegetation types on the entire study area

On the true branch of the *Elev≤17.987* criteria the riparian willow forests and the less-old and shorter poplar plantations remained. The planted poplars are slim with column-like canopy, and the reflected points from the ground between the tree-rows result in asymmetric distribution of the height points. As the skewness reflects well the asymmetric distribution of the values, the criteria of Elev_skewness> 2.376 were applied to identify the two vegetation classes. The identification of the ***poplar plantations*** followed the criteria of *CRR>0.039* and *Elev_std>1.783* and *Elev_P95≤17.987* and *Elev_skewness>2.376*. The parameters of the ***riparian willow forests*** are almost the same as of poplar plantations, only their skewness is different: *CRR>0.039* and *Elev_std>1.783* and *Elev_P95≤17.987* and *Elev_skewness≤2.376*. However, at the study site these vegetation types could be mixed even within a 15x15 m cell, which influences the accuracy of the classification, however, the identification of the classes was still effective (Gini impurity < 0.16).

Vegetation types of the study area based on the automatized classification

The decision tree created on training plots were applied on the entire, 3 km^2-large study area. The vegetation types of 11656 voxels were identified, and the land-cover map of the area was created (Fig. 5.).

In the study area the most abundant land-cover category (Fig. 6.) is riparian willow forest (30%, 80 ha). Willow patches appear mostly in deeper lying areas, like in front of the artificial levees and on the edges of clay-pits. The grasslands/open surfaces (24%, 63 ha) mainly cover the artificial levees, but in this category also some plough-fields are and clay-pits, where the dead herbaceous vegetation covers the surface like a mat. Planted poplar forests (15%, 40 ha) appear in great units. From the point-of view of flood conductivity the proportion of *Amorpha* thickets (10%, 25 ha) is crucial. They usually appear along the edges of other vegetation types and on the fallow lands. In the study area the smallest area is occupied by young poplar plantations (9%, 23 ha), but this category contains those patches as well, where young trees are mixed with bushes, but they do not create dense stands.

Validation of the results

The accuracy of the automatized classification was determined by comparing its results to field-surveys on 72 randomly selected cells. The results are summarised in a confusion matrix (Table 1.).

The accuracy of the automatized vegetation classification based on the decision tree algorithm was 83%. The open surfaces were classified with the lowest accuracy (75%). Some open surfaces were classified by

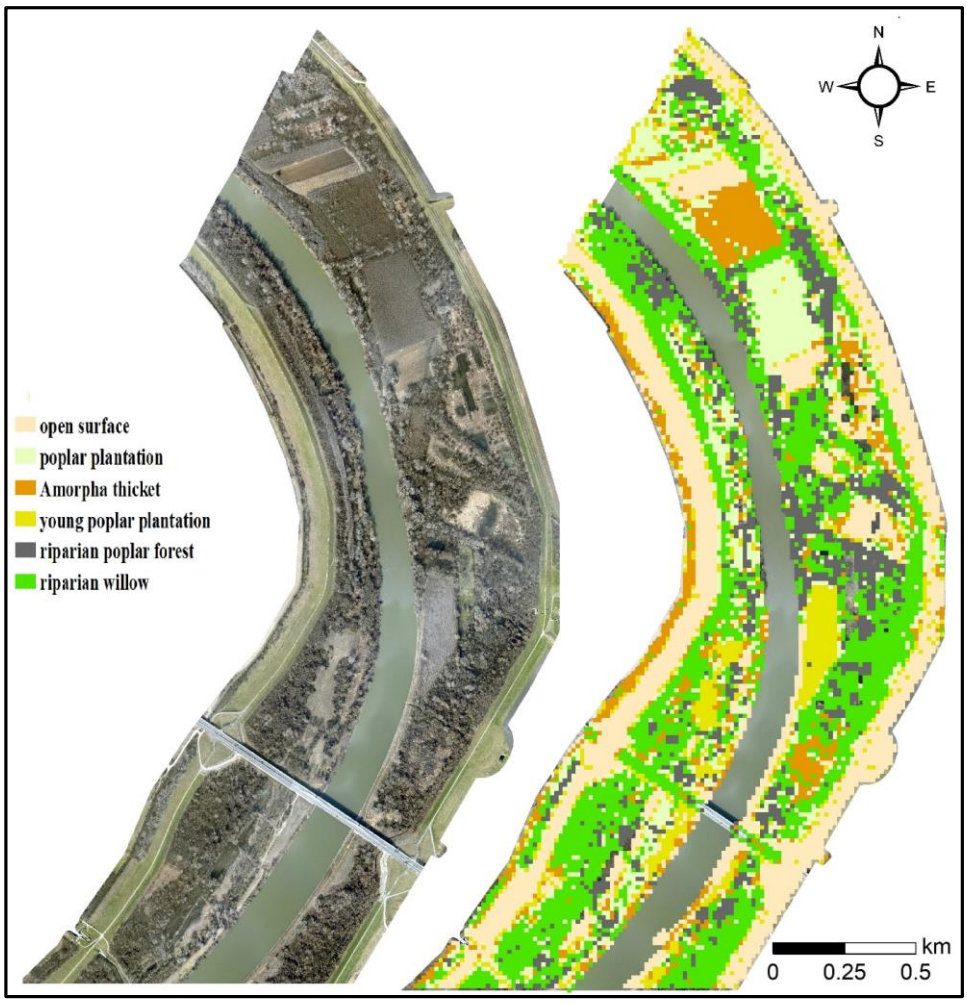

Fig. 5 Ortho-photo of the study area (A) and the automatized vegetation type map of the same area based on the classes of the decision tree (B)

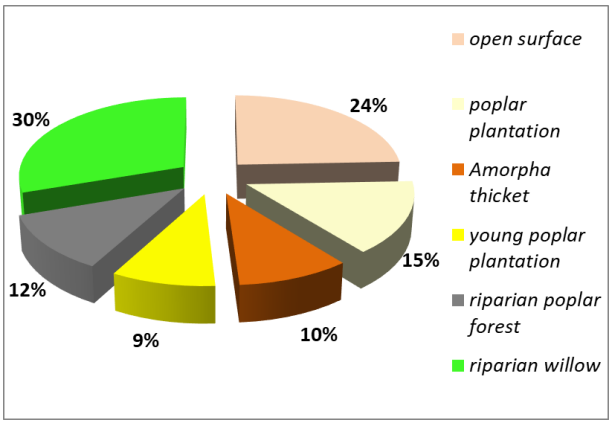

Fig. 6 Proportion of the different vegetation types in the study area

the algorithm as *Amorpha* ticket (8% of the cases) or as young poplar plantation (17%). Most of these mis-identifications occurred at the boundary between the grass-covered artificial levee and the arboreous vegetation, where many sprouts and low branches stretches over the grassland. The identification accuracy of the riparian willow forests is 85 %, in reality, the mis-identified patches belong to poplar plantations (5%), *Amorpha* thickets (5%) or riparian poplar forests (5%). This error has multiple sources: (i) between the LiDAR and the field surveys some forests were cleared and the clearances were colonised by *Amorpha*; (ii) these vegetation types could be mixed on a 15x15 m-sized cell; (iii) depending on the age of the forest patch the various vegetation types could have similar height and even similar canopy size. The *Amorpha* thickets were the most accurately (92%) classified by the algorithm. Only 8% of them were mis-classified, and got to the class of riparian willow forest. However, this error is not considerable, as it was detected on cells where the willow forest was highly invaded by *Amorpha*. The identification accuracy of riparian poplar forest was 83%, as some of their patches were identified as planted poplar (8%) and as young poplar plantation (8%). These mis-identifications were in

Table 1 Confusion matrix summarizing the validation results. Green colour indicates % of well -classified voxels (%), red colours refer to percent of unwell-classified voxels

		Based on decision tree					
		open surface	*riparian willow*	*Amorpha thicket*	*riparian poplar forest*	*young poplar plantation*	*poplar plantation*
Based on field work	*open surface*	0.75	0.00	0.08	0.00	0.17	0.00
	riparian willow	0.00	0.84	0.05	0.05	0.00	0.05
	Amorpha thicket	0.00	0.08	0.92	0.00	0.00	0.00
	riparian poplar forest	0.00	0.00	0.00	0.83	0.08	0.08
	young poplar plantation	0.00	0.00	0.17	0.00	0.83	0.00
	poplar plantation	0.00	0.00	0.00	0.18	0.00	0.82

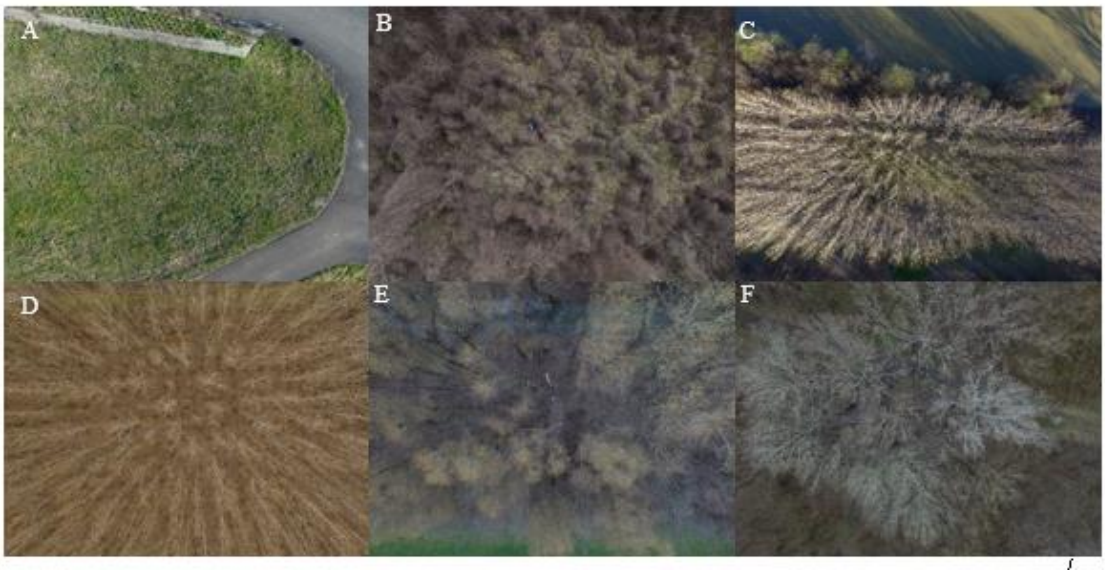

Fig. 7 Vegetation types based on drone photos. A: open surface, B: Amorpha thicket, C: young poplar plantation, D: poplar plantation, E: riparian willow forest, and F: riparian poplar forest with Populus alba

those cells, where the natural poplars were mixed to planted poplars. The young poplar plantations were identified by 83% accuracy too, as some of the cells were identified by the algorithm as *Amorpha* thickets. It could be explained by the fact, that very young plantations have similar height and density conditions as the thickets, besides, if the plantation is not managed properly, *Amorpha* could invade them very quickly. The accuracy of poplar plantations was 82%, as the greatest errors occurred when the algorithm classified them as riparian poplar forests. It could be explained by the similar height of old plantations and younger *Populus alba* trees.

CONCLUSION

The applied automatized machine learning-based classification is suitable to identify various vegetation types based on airborne LiDAR survey data. Not only land-cover types (e.g. forest), but various vegetation and forest types could be identified using the method, which has acceptable accuracy. For example, on the studied floodplain area the accuracy of the classification was 83% (based on 72 observation).

The data acquisition of LiDAR surveys combined to automatized machine learning enables us to precisely, effectively and quickly map the vegetation even on large, remote or impenetrable areas. In case of repeated surveys, the algorithm easily could be trained to the new dataset, thus the temporal changes in vegetation could be quickly and automatically detected, which is a great advantage for both researchers, stakeholders and decision makers.

The application of the resulted vegetation type map is quite wide. For example, it could be used by hydrologists, as up-to-date vegetation maps are needed during the planning and maintenance of flood-conductivity zones, or during the modelling of floods, when up-to-date data are needed on vegetation roughness to determine the Manning coefficient. In forestry these LiDAR-based vegetation maps could be used too, as the statistical parameters of forests could be calculated, and forest clearance plans could be supported.

Acknowledgement

The LiDAR data are owned and provided by the ATIVIZIG (Lower Tisza Hydrological Directorate). The research was supported by the Hungarian Research Fund (OTKA 119193).

References

Abernethy, B., Rutherfurd, I.D. 1998. Where along a river's length will vegetation most effectively stabilise stream banks? *Geomorphology* 23, 55–75. DOI: 10.1016/s0169-555x(97)00089-5

Bengio, Y., Grandvalet, Y. 2004. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research* 5, 1089–1105.

Brooks, G.R. 2005. Overbank deposition along the concave side of the Red River meanders, Manitoba, and its geomorphic significance. *Earth Surface Processes and Landforms* 30, 1617–1632. DOI: 10.1002/esp.1219

Corenblit, D., Tabacchi, E., Steiger, J., Gurnell, A.M. 2007. Reciprocal interactions and adjustments between fluvial landforms and vegetation dynamics in river corridors: A review of complementary approaches. *Earth-Science Reviews* 84, 56–86. DOI: 10.1016/j.earscirev.2007.05.004

Geerling, G.W., Kater, E., van den Brink, C., Baptist, M.J., Regas, A.M.J., Smits, A.J.M. 2008. Nature rehabilitation by floodplain excavation: The hydraulic effect of 16 years of sedimentation and vegetation succession along the Waal River, NL. *Geomorphology* 99, 317–328. DOI: 10.1016/j.geomorph.2007.11.011

Grabmeier, J. L., Lambe, L. A. 2007. Decision trees for binary classification variables grow equally with the Gini impurity measure and Pearson's chi-square test. *International Journal of Business Intelligence and Data Mining* 2(2), 213. DOI:10.1504/ijbidm.2007.013938

Heurich, M., Thoma, F. 2008. Estimation of forestry stand parameters using laser scanning data in temperate, structurally rich natural European beech (*Fagus sylvatica*) and Norway spruce (*Picea abies*) forests. *Forestry* 81, 645–661. DOI: 10.1093/forestry/cpn038

Hudak, A., Crookston, N., Evans, J., Hall, D., Falkowski, M. 2008. Nearest neighbour imputation of species-level, plot-scale forest structure attributes from lidar data. *Remote Sensing of Environment* 112, 2232–2245. DOI: 10.1016/j.rse.2007.10.009

Jalonen, J., Järvelä, J., Virtanen, J.P., Vaaja, M., Kurkela, M., Hyyppä, H. 2015. Determining Characteristic Vegetation Areas by Terrestrial Laser Scanning for Floodplain Flow Modelling. *Water* 7(2), 420–437. DOI: 10.3390/w7020420

Jung, S.E., Kwak, S.A., Park, T., Lee, W.K, Yoo, S. 2011. Estimating Crown Variables of Individual Trees Using Airborne and Terrestrial Laser Scanners. *Remote Sensing* 3, 2346–2363. DOI: 10.3390/rs3112346

Kiss, T., Fiala, K., Sipos, Gy., Szatmári, G. 2019b. Long-term hydrological changes after various river regulation measures: are we responsible for flow extremes? *Hydrology Research* 50(2), 417–430. DOI: 10.2166/nh.2019.095

Kiss, T., Nagy, J., Fehérváry, I., Vaszkó, Cs. 2019a. (Mis)management of floodplain vegetation: The effect of invasive species on vegetation roughness and flood levels. *Science of the Total Environment* 686, 931–945. DOI: 10.1016/j.scitotenv.2019.06.006

Kiss, T., Sándor, A. 2009. Land-use changes and their effect on floodplain aggradation along the Middle-Tisza River, Hungary. *AGD Landscape and Environment* 3(1), 1–10.

Kovács, S., Váriné Szöllősi, I. 2003. A Vásárhelyi Terv Továbbfejlesztését megalapozó hidrológiai és hullámtér hidraulikai vizsgálatok eredményei a Közép-Tiszán. MHT XXI. 2/12. 1–11.

Laes, D., Reutebuch, S., McGaughey, B., Maus, P., Mellin, T., Wilcox, C., Anhold, J., Finco, M., Brewer, K. 2008. Practical lidar acquisition considerations for forestry applications. RSAC-0111-BRIEF1. Salt Lake City, UT: U.S. Department of Agriculture, Forest Service, Remote Sensing Applications Center. 32 p.

Manners, R., Schmidt, J., Wheaton, M. J. 2013. Multiscalar model for the determination of spatially explicit riparian vegetation roughness. *Journal of Geophysical Research: Earth Surface* 118, 65–83. DOI: 10.1029/2011jf002188

McGaughey, R. 2018. Users Manual of Fusion/LDV: Software for LIDAR Data Analysis and Visualization. United States Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Naesset, E., Gobakken, T., Holmgren, J., Hyyppa, J., Maltamo, M., Nilsson, M., Olsson, H., Persson, A., Doderman, U. 2004. Laser scanning of forest resources: the Nordic experience. Scandinavian. *Journal of Forest Research* 19, 482–499. DOI: 10.1080/02827580410019553

Nagy, J., Kiss, T., Fiala, K. 2018. Hullámtér-feltöltődés vizsgálata az Alsó-Tisza mentén. II. Folyóhátak (parti hátak) feltöltődését befolyásoló tényezők. *Hidrológiai Közlöny* 98(1), 33–40.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(85), 2825–2830. DOI: 10.3389/fninf.2014.00014

Rátky, I., Farkas, P. 2003. A növényzet hatása a hullámtér vízszállító képességére. *Vízügyi Közl.* 85(2), 246–264.

Schaffer, C. 1993. Overfitting Avoidance as Bias. *Machine Learning* 10, 153–178. DOI: 10.1007/bf00993504

Steiger, J, Gurnell, A.M., Ergenzinger, P., Snelder, D.D. 2001. Sedimentation in the riparian zone of an incising river. *Earth Surf. Process. Landforms* 26, 91–108. DOI: 10.1002/1096-9837(200101)26:1<91::aid-esp164>3.0.co;2-u

Vetter, M., Höfle, B., Hollaus, M., Gschöpf, C., Mandlburger, G., Pfeifer, N. 2011. Vertical vegetation structure analysis and hydraulic roughness determination using dense ALS point cloud data–a voxel based approach. *Int. Arch. Photogr. Remote Sens. Spat. Inf. Sci.* 38(5), 200–206. DOI: 10.5194/isprsarchives-xxxviii-5-w12-265-2011

Zellei, L., Sziebert, J. 2003. Árvizi áramlásmérések tapasztalatai a Tiszán. In: Szlávik L. (szerk.): Elemző és módszertani tanulmányok az 1998-2001. évi ár- és belvizekről. Vízügyi Közlemények különszám 4, 133–144.