

PROCEEDINGS OF SPIE

SPIDigitalLibrary.org/conference-proceedings-of-spie

Building rooftop classification using random forests for large-scale PV deployment

Dan Assouline, Nahid Mohajeri, Jean-Louis Scartezzini

SPIE.

Building rooftop classification using Random Forests for large-scale PV deployment

Dan Assouline^a, Nahid Mohajeri^a, Jean-Louis Scartezzini^a

^aSolar Energy and Building Physics Laboratory
Ecole Polytechnique Federale de Lausanne
CH-1015 Lausanne, Switzerland

ABSTRACT

Large scale solar Photovoltaic (PV) deployment on existing building rooftops has proven to be one of the most efficient and viable sources of renewable energy in urban areas. As it usually requires a potential analysis over the area of interest, a crucial step is to estimate the geometric characteristics of the building rooftops. In this paper, we introduce a multi-layer machine learning methodology to classify 6 roof types, 9 aspect (azimuth) classes and 5 slope (tilt) classes for all building rooftops in Switzerland, using GIS processing. We train Random Forests (RF), an ensemble learning algorithm, to build the classifiers. We use (2×2) [m^2] LiDAR data (considering buildings and vegetation) to extract several rooftop features, and a generalised footprint polygon data to localize buildings. The roof classifier is trained and tested with 1252 labeled roofs from three different urban areas, namely Baden, Luzern, and Winterthur. The results for roof type classification show an average accuracy of 67%. The aspect and slope classifiers are trained and tested with 11449 labeled roofs in the Zurich periphery area. The results for aspect and slope classification show different accuracies depending on the classes: while some classes are well identified, other under-represented classes remain challenging to detect.

Keywords: Geographic Information Systems, LiDAR, Roof classification, Random Forests, Roof mounted Photovoltaics

1. INTRODUCTION

Solar energy is arguably one of the most promising sources of renewable energy, and has been vastly studied over the last few years. While various solar energy technologies exist on the market, photovoltaic (PV) panels are getting more and more attention as their electricity output make them very volatile and usable for all kinds of tasks. Most specifically, the price of PV panels are continuously decreasing and their efficiency is significantly increasing over time. Consequently, PV panels mounted over existing building rooftops have proven to be a viable large scale resource of clean energy in urban areas.^{1,2} However, a potential study is necessary, before local governments and municipalities can plan a large scale PV deployment. Studies on solar PV potential typically consider 3 main estimations: the area available for PV installation over rooftops, the geometric properties of the roofs including roof shape, slope (tilt) and aspect (azimuth), and finally the solar radiation over the roofs.

Several studies suggest methodologies to estimate the available roof area for PV installation.³ However, very limited studies suggest an accurate methodology to estimate the geometric properties of the roofs. Depending on the available source of data, the suggested methods to estimate the geometric characteristics of rooftops can be different. A first general method consists in GIS processing, using LiDAR data and good quality footprint vector polygons. Some studies use LiDAR data integrated with footprint polygons using Triangulated Irregular Network (TIN) to classify roof types as either flat or pitched.⁴ Alternatively, rooftops segments can be created using slope and aspect classes, along with different building types, with LiDAR data and footprint polygons.⁵ Other studies use LiDAR data to estimate roof characteristics by using a catalogue of common roof shapes fitted with the studied roofs.⁶ A second common method is to use satellite images in order to detect different roof

Further author information: (Send correspondence to D.A.)

D.A.: E-mail: dan.assouline@epfl.ch, Telephone: +41 21 69 34557

N.M.: E-mail: nahid.mohajeri@epfl.ch, Telephone: +41 21 69 34547

J-L.S.: E-mail: jean-louis.scartezzini@epfl.ch, Telephone: +41 21 69 35549

shapes. An example of such a method uses 2D satellite images to classify roof shapes into 4 common types, and other combinations of these types.⁷ Other various methodologies include the direct use of LiDAR data and satellite images to detect buildings^{8,9} and extract some geometric characteristics. An original learning-based roof classification was recently presented using bags of words features extracted from a point cloud.¹⁰ However, very few studies attempted to use similar methodologies at a large scale, given the precision needed for the data to obtain good results. Large scale potential studies often use adapted methods that do not focus on geometric characteristics of the building roofs. A good example of these kind of study is a recent one for Spain.¹¹

The present paper uses Random Forests with low resolution building LiDAR data to classify building rooftops and estimate their aspects and slopes using generalised footprint polygons at the national scale in Switzerland. The machine learning approach aims at counterbalancing the lack of precision of the footprint polygons and the LiDAR data, which are widely available datasets, but not designed for precise geometric analysis. In particular, the location and the shape of the footprint polygons is rather approximative. The study aims at using Random Forests with elevation raster features and geometric features extracted from the GIS data in order to: (1) Classify roof types into 6 predefined classes, (2) Estimate the main slope and aspect class for each roof class, with predefined 5 classes for slope, and 9 predefined classes for aspect, (3) Make these classifications embeddable in a solar rooftop PV potential study.

2. DATA AND METHODS

2.1 Data pre-processing

The vector and raster data sources are as follows:

- **VECTOR25:** a generalised footprint polygon data for clusters of buildings in Switzerland. The data is available from Swisstopo (<https://shop.swisstopo.admin.ch>). It includes 1,825,678 polygons in total. These polygons will be used to capture the approximate geometry and locations of buildings all across Switzerland and will be called VECpolygons throughout the study.
- **DOM:** the swiss Digital Surface Model available from Swisstopo (also called DSM). DOM is a LiDAR data offering real elevation values for the urban and rural surface while taking into account vegetation and buildings. Its resolution is (2×2) [m²] and the data is available in the form of (3000×4400) [m²] rectangles.
- **Sonnendach data:** a precise roof surface polygon data for around 800 communes in Switzerland (11449 buildings), available from (<http://www.bfe-gis.admin.ch/sonnendach/?lang=de>). The data is originally extracted from the swissBUILDINGS3D 2.0 data. It offers the projected polygon geometry of each roof surface in the area of interest as well as aspect and slope values for each surface. It, however, does not show superstructures and other more detailed structures over the roofs. It will be used to provide examples for the aspect and slope classification models.

ModelBuilder (an ArcGIS tool) and Python codes were used to automate the process of extraction of useful statistics from the DOM elevation data. We first split the entire DOM into medium-sized parts (about 25 parts for the entire remaining Switzerland territory once the communes covered by Sonnendach data are discarded), to allow for reasonable processing time on each of them. Then, we built a model using Model Builder which extract aspect and slope data over VECpolygons. These data is used as features for the classifications. The convention for aspect values is shown in Figure 1. A python script was written to run the model steps autonomously outside of ArcGIS to speed up the computational time.

The model, for each portion, performs the following tasks: (i) Upsample it to $(0,5 \times 0,5)$ [m²] to gain in precision; (ii) Compute aspect and slope raster from the DOM raster using the Spatial Analyst toolbox; (iii) Perform a Re-classification of raster values (from Spatial Analyst toolbox) by bins:

- For slope, with 9 bins: $[0^\circ, 10^\circ]$, $[10^\circ, 20^\circ]$, $[20^\circ, 30^\circ]$, $[30^\circ, 40^\circ]$, $[40^\circ, 50^\circ]$, $[50^\circ, 60^\circ]$, $[60^\circ, 70^\circ]$, $[70^\circ, 80^\circ]$, $[80^\circ, 90^\circ]$.

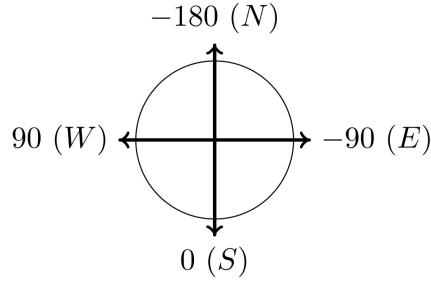


Figure 1. Convention used for aspect calculation.

- For aspect, with 2 different bins configuration: (a) 5 bins, including flat, $[-135^\circ, 135^\circ]$ (North), $[-135^\circ, -45^\circ]$ (East), $[-45^\circ, 45^\circ]$ (South), $[45^\circ, 135^\circ]$ (West); (b) 19 more precise 20° bins, including flat, $[-170^\circ, 170^\circ]$, $[-170^\circ, -150^\circ]$, $[-150^\circ, -130^\circ]$, $[-130^\circ, -110^\circ]$, ..., $[130^\circ, 150^\circ]$, $[150^\circ, 170^\circ]$. These two different set of bins are used separately for two tasks. The first is used to build features for roof classification, as it expresses the main changes in aspect across the roof and to avoid dilution of the feature information. The second more complete configuration is used to build labels for the aspect estimation.

(iv) Compute frequencies of raster cells for each slope and aspect bin over each VECpolygon to obtain the frequencies of cells with an aspect in each of the 5 aspect bins, the frequencies of cells with an aspect in each of the 19 aspect bins, and the frequencies of cells with a slope in each of the 9 slope bins; (v) Compute statistics for each of the three histograms frequency data extracted in (iv) (mean bin, mode bin etc.); (vi) Export these frequencies and statistics in csv format, ready to be used as features for further classifications. Illustrations of reclassified slope and aspect rasters can be seen in Figure 2 for two different roof types.

2.2 Classification in Machine Learning

Machine Learning (ML) methods are algorithms that learn patterns from examples in order to perform predictions. In a classical supervised learning framework, the examples are gathered in a dataset $(\mathbf{x}_i, y_i)_{i=1, \dots, N}$, where N is the number of points in the data set. Each data point (\mathbf{x}_i, y_i) includes a p -dimensional input vector \mathbf{x}_i , and an output value y_i . The input vector is a realization of the input variables of interest X_1, X_2, \dots, X_p (e.g. number of sides of the roof, roof area etc.), and the output value is the realization of the corresponding output variable Y (e.g. roof type). Note that the input variables are also called *features* or *predictors*, the input values are *samples* or *instances*, the output variable is the *target*, and the output values are *targets* or *labels*. The dataset of observed data points (examples) is called the *labeled set*.

Given the data, the aim of a machine learning task is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, where \mathcal{X} and \mathcal{Y} are respectively the input and output spaces, so that predictions $f(\mathbf{x})$ are close as possible as the corresponding targets y . Note that in case of a classification, \mathcal{Y} is a finite set of classes (y can only take discrete values), and f is called a *classifier*. In order to maximize the performance of the classifier and allow it to generalize well outside of the labeled set, we use the following classical strategy: (i) separate the labeled set into a *training set* (75% of the data) and a *test set* (25% of the data), (ii) train a model using solely the training set, (iii) use the trained model to predict output values for points in the test set and measure the discrepancy with the known labels to compute the test error. Most models include parameters, usually called *hyperparameters*, that have to be tuned in order to obtain the best model possible for a given data. The hyperparameters are often tuned while training the model, using a procedure called *k-fold cross validation*.¹²

In order to measure the performance of the classifier (by measuring the error between predicted outputs and labels), we use in this study the *accuracy*, which is a classical error measure for classification tasks. It computes the probability of being well classified in the test set, using the model built in the training set:

$$\text{Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}_{[f(\mathbf{x}_i)=y_i]} \quad (1)$$

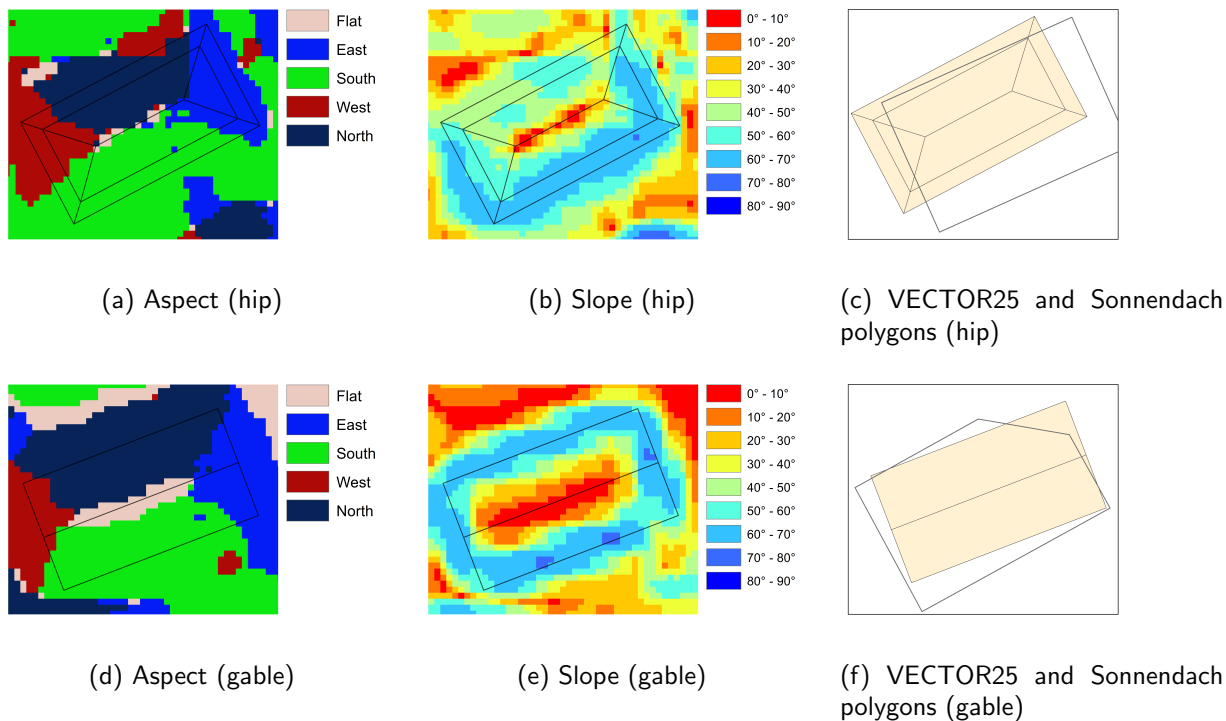


Figure 2. Aspect and Slope reclassified ($0,5 \times 0,5$) [m^2] rasters, along with the building polygons from VECTOR25 (grey thick line) and Sonnendach data (black thin line) for a building with a hipped (a,b,c) and a gabled roof (d,e,f). One can observe the significant delay of position between the two polygons. Also, the different aspect and slope patterns between the two types are clearly shown, specially regarding the amount of roof raster cells showing a flat surface, significantly larger in the gable case.

where N_{test} is the size of the test set, and $\mathbb{1}$ is the indicator function, defined by:

$$\mathbb{1}_{[f(\mathbf{x}_i)=y_i]} = \begin{cases} 1, & \text{if } f(\mathbf{x}_i) = y_i \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

2.3 Random Forests for classification

Random Forests (RF),¹³ is a machine learning algorithm that is part of the Ensemble Learning family of methods. Ensemble Learning aims at aggregating the results from multiple generated “weak” learners (simple and fast models with a poor performance) to obtain a better estimator. In case of RF, the weak learners are classification and regression Trees.¹⁴ Trees are decision models that have been widely used for many years, and in various applications. They are constructed with the training data by a series of binary splits, at each nodes of the tree, which virtually split the input space according to a query. The query is performed on one of the variable, for example “Is $X_2 < 3.5$ ”. At each node, if the answer is “Yes”, the left path is taken, otherwise, the right one. Terminal nodes are decision nodes, giving the predicted value for regression, or the predicted class for classification. The query (meaning the choice of the variable and the threshold) is optimized at each node in order to decrease the number of samples in the node as fast as possible (maximize the *impurity decrease*), which contributes to the performance of the model. The details of optimization, however, will not be presented here. New observations are predicted by passing their features through the tree down to a terminal node. As we use classification in this study, we will focus on RF for classification. However, the regression version of the algorithm is very similar to classification in principle.

The two ensemble algorithms attracting a lot of attention are originally Bagging¹⁵ and Boosting¹⁶ for classification. Bagging aims at reducing the variance of the trees by fitting a large number of trees on bootstrap-sampled

(sampled with replacement) versions of the training data. The predicted class is then the one with the majority of votes from the independent trees, a vote being the class predicted by one of the trees. Boosting also combines multiple trees. However, instead of building the trees on bootstrap samples of the training data it considers weighted versions of the training data. The final classifier is a weighted average of the tree classifiers. Boosting has proved to perform better than bagging in most problems, but it is delicate to tune its parameters.

Random Forests are a refined version of the Bagging algorithm. The improvement is achieved by adding a layer of randomness to attempt to de-correlate the trees built from bootstrap samples of the data. In Bagging, each node is using the best possible split among all variables. Random forests, in contrast, randomly choose m variables out of all possible ones at each node and the best split among these m variables is used. This little addition considerably increases the performance of the algorithm, making it comparable to Boosting and other popular classifiers including Support Vector Machines or Neural Networks. What separates RF from other algorithms, however, is its easiness to use in practice, since it only has two parameters (m , and B , the number of trees). Furthermore, it is worth noting its various advantages in practice. These includes: (i) RF is not sensitive to outliers, (ii) it automatically performs feature selection while choosing the best split at each node, (iii) no need pre-processing such as scaling or normalizing the data is required before the training process, (iv) two extra measures are embedded by Breiman in the algorithm: the variable importance, estimating the impact of each variable in the model, and the Out-Of-Bag (OOB) error, a validation error provided during the training. The OOB error is the average prediction error computed for all training points using, for the prediction of each training point, only the trees that did not contain this training point in their bootstrap sample.

In addition, RF is not very sensitive to the choice parameters. The number of features considered for splitting (m) is usually chosen from a list of values that work well in practice, and $m = 1$ gives the optimal result for some data.¹⁷ m can nonetheless be fine tuned by k-fold cross validation. Also, the accuracy increases with the number of trees B . Thus, it is current practice to fix m and try increasing values of B , until an accuracy plateau is reached. Even though the number of trees required increases with the size of the training data, $B = 500$ trees appear to be enough to achieve optimal performance in most cases, from our experience.

3. ROOF GEOMETRIC FEATURES CLASSIFICATION

We describe here the methodology leading to the roof geometric features classification. It includes a roof type classification as well as an aspect and slope estimation for all buildings in Switzerland. As mentioned before, the building polygon data offers clusters of buildings (VECpolygons). Thus, for each cluster of buildings we estimate: one roof type, one main -most frequent- roof slope and one main -most frequent- roof aspect. The previously presented data will serve as input features for the machine learning algorithm used here, that is RF, for the classification tasks.

3.1 Roof classification

We use the presented random forests to classify roof types to further help future estimation of potential for PV deployment over roofs (specifically the area available for PV panels installation and the distribution of roof aspects across the building).

The first step is to choose the different classes that cover all possible types of roof. There are many possible roof types considered in the literature.^{6,10} For our purpose, we accounted for the differences both in roof shape and the general footprint geometry of the building. For example, one building can have a pyramidal roof, but with a rectangular or an L-shaped footprint, which will result in a very different aspect distribution. Roof shapes include mainly flat, gable, hipped, pyramidal, shed, mansard, and gambrel. Footprint geometries of gatherings of buildings can considerably vary. However usual forms include: rectangular, L-shaped, T-shaped, U-shaped, O-shaped, Triangle-shaped. Some of these shapes and geometries can be difficult to differentiate from one another due to their similarities and the relative lack of precision imposed by the large scale of our study. Thus, some choices were made to decrease the complexity of the task, thus increasing the performance of our classifier. It was decided to gather some of them in the same class (by similarity) in order to reduce the total number of classes. The classes are as follows: Gable and Shed, Hip and Pyramidal, L and T-shaped, O and U-shaped; and Complex. The complex class includes the Triangle-shaped buildings, and all roofs that do not fit in existing

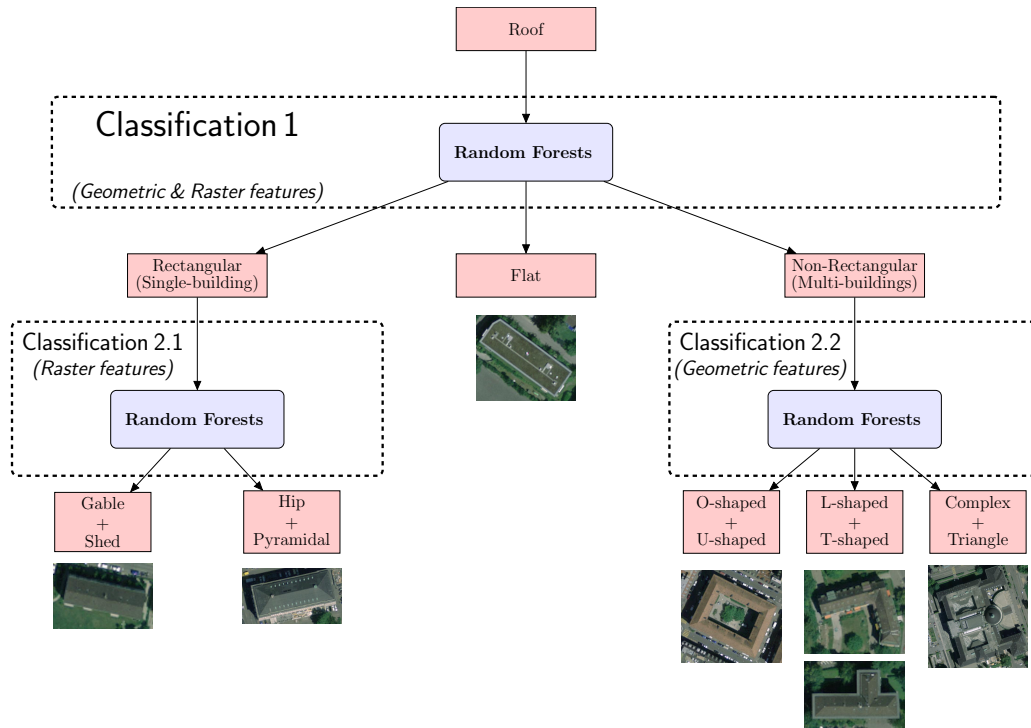


Figure 3. Roof classification scheme.

classes. Gambrel and Mansard were discarded because of their complex structure and the lack of examples found in the training process, as discussed later in the study.

Since the footprint geometry and roof shape do not depend on the same features, we decided to perform two layers of classification, separating three main polygon classes in the first layer, and treating with shapes and geometry independently in the second layer. More Specifically, we perform: (i) Classification 1 to differentiate between Flat, Rectangle, and Non-Rectangle polygons (ii) Classification 2.1 on Rectangle polygons, to differentiate between Gable, and Hip; Classification 2.2 on Non-Rectangle polygons, to differentiate between O-shaped, L-T-shaped, and Complex. The complete classification process is illustrated in Figure 3. Note that in performing this classification, we did not differentiate between the different roof shapes when classifying the non-rectangular polygons. We show further in the study that the simple binary classification between Gable and Hip for rectangle polygons is a very hard task at a large scale, resulting in a quite poor accuracy. As it would lead to an even poorer accuracy for multi-buildings polygons, we focus on the geometry of the footprint and consider they are gable shaped.

The chosen roof classes are then considered to build the labeled set of examples. The labeled set was obtained by manually detecting different classes of buildings from examples using high resolution satellite images and the Sonnendach data (<http://www.bfe-gis.admin.ch/sonnendach/>). The VECTOR25 VECpolygons were layered over satellite images from Swisstopo (Swissimages 25cm) so that roof classes could be attributed to each polygon by visual observation. A total number of 1252 VECpolygons were manually labelled, which approximately corresponds to 1% of the total number of VECpolygons all around Switzerland. We considered 3 regions containing both rural/suburban parts and dense urban parts, including contemporary and old city center buildings: Baden region, Luzern region, and Winthertur region. A number of 268, 556, 232, 32, 88, and 76 VECpolygons were labeled respectively for Flat, Gable, Hip, O-shaped, L-shaped, and Complex classes. The training set was composed by 75% of the labels and the remaining 25% of the labels was used for the test set, leading to the accuracy computation for the classifiers.

Going in pairs with the labels, the features used for each polygon in the classification tasks are of two types: (i) the geometric features will serve as simple features to differentiate between different geometric footprint

shapes, (ii) the raster based features will be used to differentiate between different roof shapes (Flat, Gable, Hip) and will be extracted from the slope and aspect raster data.

The geometric features characterize the shape of the building footprints. They must be simple enough to be computable directly from the polygons, and aim at differentiating the different geometric footprint shapes. A natural feature is the number of vertices. Yet, it is clearly not sufficient to characterize the footprint shapes. To add information about the compactness of the polygon, the iso-perimetric quotient (*isoQ*) of the polygon is used as an extra geometric feature. This coefficient is defined as the ratio of the polygon area and the area of a circle with the same perimeter. A straight-forward calculation leads to the *isoQ* expression:

$$isoQ = \frac{4\pi A}{P^2} \quad (3)$$

where A and P are respectively the area and the perimeter of the polygon.

The raster features characterize the roof shape based on the elevation data. Since a roof shape is intuitively described by the arrangement of the roof different directions and tilts, the raster features used for training will be combination of various slope and aspect statistics extracted in the DSM raster processing. These raster features include:

- Statistics from the 5 bins aspect raster data: mean, standard deviation, variety, majority, minority, median.
- Statistics from the 9 bins slope raster data: mean, standard deviation, variety, majority, minority, median.
- Frequencies and percentages from the 5 bins aspect raster data: number of cells with aspect in each aspect bin, and proportion of cells with aspect in each aspect bin.
- Frequencies and percentages from the 9 bins slope raster data: number of cells with slope in each slope bin, and proportion of cells with slope in each slope bin.
- Ratios of flat cells frequencies and other directions frequencies: East/Flat ratio, South/Flat ratio, West/Flat ratio, North/Flat ratio.
- Ratios of flat cells frequencies and other slope bins frequencies: $[10^\circ, 20^\circ]$ /Flat ratio, $[20^\circ, 30^\circ]$ /Flat ratio ... etc.
- Ratios of slope bins with one another: $[10^\circ, 20^\circ]$ / $[20^\circ, 30^\circ]$ ratio, $[10^\circ, 20^\circ]$ / $[30^\circ, 40^\circ]$ ratio ... etc.
- Boolean variables to identify symmetry in roofs: EWsym, NSsym, BothSym respectively indicates an east-west symmetry, north-south symmetry, and a symmetry in both directions. They are simply computed: if the number of east cells is equal to the number of west cells, plus or minus 100, EWsym = 1, otherwise EWsym = 0. The computation is similar for NSsym. BothSym is given by EWsym \times NSsym.

The first classification (Classification 1) uses both geometric and raster based features to differentiate between flat roofs, non-flat rectangular polygons and non-flat non-rectangular polygons. The features of Classification 1 include: number of vertices, *isoQ*, percentages from 5 bins aspect data, and ratios of flat cells $[20^\circ, 30^\circ]$ /Flat, $[30^\circ, 40^\circ]$ /Flat, $[40^\circ, 50^\circ]$ /Flat slope bins, for a total of 8 features. A choice of $B = 500$ trees is found to be sufficient to obtain optimal results, and m is chosen by 6-fold cross validation. Figure 4 shows the evolution of the OOB error with an increasing number of trees. The same number of trees and strategy for m tuning is used in the other roof classifications. The performance of the trained RF classifier is summarized in Table 1, in the form of a confusion matrix. This matrix exposes, for each class (each row), the number of polygons well classified, and the number of polygons wrongly classified in other classes. For example, the first row of the matrix shows that, out of the $45 + 13 + 9 = 67$ flat roofs considered in the validation set, 45 were well classified as flat roofs, 13 were wrongly classified as rectangular non-flat polygons, and 9 were wrongly classified as non-flat non-rectangular polygons. The last column gives the accuracy of the classifier for each class, meaning the percentage of well classified polygons. The Out-Of-Bag (OOB) estimate of the error is also provided in the table.

Table 1. Classification 1 confusion matrix.

OOB = 85%	Flat	Rect	Non-Rect	Acc.
Flat	45	13	9	70%
Rect	5	185	7	94%
Non-Rect	4	4	41	84%

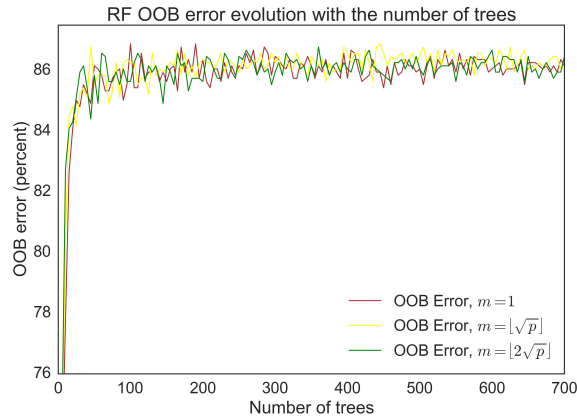


Figure 4. Evolution of OOB estimate of error rate with an increasing number of trees for classification 1. Note that $\lfloor \sqrt{p} \rfloor$ and $\lfloor 2\sqrt{p} \rfloor$ are values advised in practice.¹⁷ The plateau is reached very quickly, so that 500 trees seem more than sufficient. One can also observe that m does not have a very high impact on the OOB error.

The second classification (Classification 2.1) uses purely raster based features to differentiate between gable and hipped roofs. The features of Classification 2.1 include aspect and slope statistics, frequencies and percentages respectively for aspect and slope, and 13 different slope ratios. The performance of the trained RF classifier is summarized in Table 2.

Finally, the third classification (Classification 2.2) uses purely geometric features to differentiate between O-U-shaped, L-T-shaped and complex buildings. The features are simply the number of vertices and the *isoQ*. The performance of the trained RF classifier is summarized in Table 3. It is straight forward to obtain the final classification accuracy for each roof type, by multiplying the accuracies in each classification layer: $AE_{final} = AE_{class1} \times AE_{class2}$. It can be observed in Table 4. After the classifiers are built with the labeled data, they are used on the remaining polygons to determine their shape.

Table 2. Classification 2.1 confusion matrix.

OOB = 72%	Gable	Hip	Acc.
Gable	118	21	85%
Hipped	40	18	31%

Table 3. Classification 2.2 confusion matrix.

OOB = 65%	O-sh.	L-Sh.	Complex	Acc.
O-sh.	8	0	0	100%
L-Sh.	3	16	3	73%
Complex	2	0	17	90%

Table 4. Final accuracy of the overall classifier to detect each roof class.

Flat	Gable	Hip	O-shaped	L-Shaped	Complex	Mean
70%	80%	30%	84%	61%	76%	67%

The results of the classification vary greatly depending on the class. While the model identifies Gabled roofs quite well, it is very poor in classifying hipped roofs. This is mainly caused by the relatively low resolution of the LiDAR data and the lack of precision of the VECTOR25 polygons in shape and more particularly in the location. Thus, this prevents the model from detecting changes in aspect values in small areas, which is the key

to detect hipped roofs, characterized by the two lateral small “hips” (Figure 2). Besides the quality of the data at hand, the model performance is heavily depending on the size of the training data and the number of labels for each class. In case of the hipped roofs, a higher number of labels is desirable, and will be used in the future, to distinguish them from gabled roofs.

3.2 Aspect and slope estimation

The aspect and slope angle of the roofs are of course very significant when it comes to estimating the solar energy available over the roofs. As the current study is based on building VECpolygons available at a large scale (VECTOR25 polygons), aggregated values of aspect and slope are desirable for buildings. More specifically, we aim at one aspect value and one slope value for each polygon (cluster of buildings). Consequently, the modes (most frequent value) of the aspect and slope distributions were considered. These two quantities are real numbers, and naturally call for a regression estimation. Nevertheless, they revealed themselves to be quite delicate to predict, solely based on our raster features. As a consequence, we decided to relax the problem into a classification task by creating bins that act as classes, for both aspect and slope. The resulting classifications for each polygon consist of: (i) classification of the main aspect, meaning the center of the most frequent aspect bin represented, and (ii) classification of the main slope, meaning the center of the most frequent slope bin represented. In order to capture the different aspects of the various roof sides in each building, we consider as a prior information the predicted roof types from the previous part 3.1. We use the symmetry of each roof type to virtually distribute the different roof aspects from the main aspect estimation. Note that this symmetry allows us to gather aspect bins that are in the same direction, (meaning delayed by 180, as for example $[-50^\circ, -30^\circ] \cup [130^\circ, 150^\circ]$), which divides by two the number of aspect classes. Random Forests were used for both slope and aspect classifications. A table summarizing how the roof aspects are distributed from the main aspect for each roof type is shown in Table 7.

Classes were created as 20° bins for aspect estimation, and 10° bins for slope estimation. More specifically, the following bins were used:

- 5 bins for slope: $[10^\circ, 20^\circ]$, $[20^\circ, 30^\circ]$, $[30^\circ, 40^\circ]$, $[40^\circ, 50^\circ]$, $[50^\circ, 60^\circ]$, corresponding respectively to classes C_{s1} , C_{s2} , C_{s3} , C_{s4} , C_{s5} . Slope values beyond 60° are very rare and thus not considered.
- 9 bins for aspect:
 1. $[-10^\circ, 10^\circ] \cup [-170^\circ, 170^\circ]$
 2. $[-170^\circ, -150^\circ] \cup [10^\circ, 30^\circ]$
 3. $[-150^\circ, -130^\circ] \cup [30^\circ, 50^\circ]$
 4. $[-130^\circ, -110^\circ] \cup [50^\circ, 70^\circ]$
 5. $[-110^\circ, -90^\circ] \cup [70^\circ, 90^\circ]$
 6. $[-90^\circ, -70^\circ] \cup [90^\circ, 110^\circ]$
 7. $[-70^\circ, -50^\circ] \cup [110^\circ, 130^\circ]$
 8. $[-50^\circ, -30^\circ] \cup [130^\circ, 150^\circ]$
 9. $[-30^\circ, -10^\circ] \cup [150^\circ, 170^\circ]$

corresponding respectively to classes C_{a1} , C_{a2} , C_{a3} , C_{a4} , C_{a5} , C_{a6} , C_{a7} , C_{a8} , C_{a9} .

The labeled set was extracted from the Sonnendach data, containing aspect and slope values for each surface of all building rooftops in the covered area. The main aspect and slope were computed by extracting the most frequent aspect and slope value classes across the surfaces of each polygon, thus forming the label for each polygon. The entire Sonnendach data was considered, gathering 11449 polygons. The training and test set were built respectively with 75% and 25% of the labeled set. In both aspect and slope classifications, we use the respective frequencies and ratios to serve as features. For the aspect classification, the reclassified aspect values from the 20° bins were used to form the features of the input data. More specifically, the features include 9

aspect percentages and 20 ratios of aspect frequencies. For the slope classification, similarly, the reclassified slope values from the 10° bins were used to form the features of the input data. More specifically, the features include 7 slope percentages and 12 ratios of slope frequencies.

The strategy used for parameter tuning is similar than in part 3.1. A number of 500 trees is used and m is chosen by 6-fold cross validation in both classifications. A summary of the classifiers' performances is given in the form of the accuracy matrices depicted in Tables 5 and 6. The two classifiers can now be used on the unlabeled polygons to determine their main aspect and slope, and the distribution of aspect and slope for the remaining sides of each roof is assumed to be as shown in Table 7.

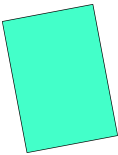
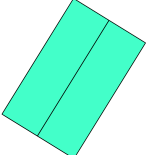
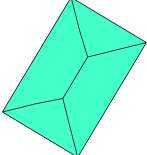
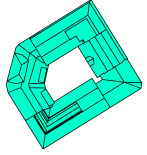
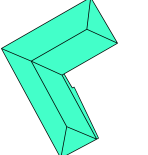
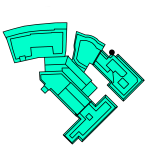
Table 5. Aspect estimation confusion matrix.

OOB = 63%	C_{a1}	C_{a2}	C_{a3}	C_{a4}	C_{a5}	C_{a6}	C_{a7}	C_{a8}	C_{a9}	Acc.
C_{a1}	128	19	1	1	18	5	1	0	12	69%
C_{a2}	14	267	22	3	4	35	37	1	1	70%
C_{a3}	1	21	186	11	2	0	24	18	3	70%
C_{a4}	1	2	5	127	7	0	5	18	14	71%
C_{a5}	17	6	4	5	124	6	1	0	12	70%
C_{a6}	28	37	2	1	11	83	7	1	2	48%
C_{a7}	1	37	31	1	0	6	174	5	0	68%
C_{a8}	2	2	15	20	4	0	5	81	11	58%
C_{a9}	10	4	1	17	21	0	2	1	97	63%

Table 6. Slope estimation confusion matrix.

OOB = 50%	[10,20]	[20,30]	[30,40]	[40,50]	[50,60]	Acc.
[10,20]	83	126	24	11	0	34%
[20,30]	55	414	123	24	0	67%
[30,40]	32	257	304	86	0	45%
[40,50]	30	113	177	115	5	26%
[50,60]	3	18	20	21	1	2%

Table 7. Roof characteristics considered for each roof type. β and γ are the center value of respectively the slope and aspect class predicted for the roof of interest.

Roof Type	Flat	Gable	Hip	O-shaped	L-Shaped	Complex
						
Number of directions	1	2	4	8	4	8
Roof sides aspect	γ	γ $\gamma + 180$	γ $\gamma + 180$ $\gamma + 90$ $\gamma - 90$	γ $\gamma + 180$ $\gamma + 90$ $\gamma - 90$	γ $\gamma + 180$ $\gamma + 90$ $\gamma - 90$	γ $\gamma + 180$ $\gamma + 90$ $\gamma - 90$
Roof sides Slope	10°	β	β	β	β	β

As in the roof classification step, the performance of the model changes significantly depending on the class of aspect or slope. While the aspect estimation offer a reasonable accuracy of around 67% for almost all aspect

classes, it seems still very challenging to estimate slope at a large scale without a very high resolution data. Indeed, if the most frequent slope class in Switzerland ($[20^\circ, 30^\circ]$) is relatively well identified, the other classes show a very low accuracy. Note that the use of Random Forests classification is not the first natural idea for aspect and slope estimation. One can simply compute the number of pixels in each aspect and slope bin, and assume that the bin with the highest frequency of cells is the main bin. The center of the bin is then the estimated aspect or slope value. Unfortunately, the lack of precision of the raster data resulted in poor results while using this simpler approach. Random forests offered significantly higher performance.

4. CONCLUSION

In this study, we use a machine learning methodology, namely Random Forests, for the large scale estimation of three geometric buildings characteristics: (i) the buildings roof type, (ii) the building roofs most frequent aspect value, and (iii) the building roofs most frequent slope value. We use Switzerland buildings as a case study. We use widely available low resolution data which includes (2×2) $[m^2]$ LiDAR data and generalized footprint vector polygons. The roof classifier learns from a training data of 1252 labeled buildings with two classification steps and is able to identify 6 roof types with an average accuracy of 70%. The slope and aspect classifiers learn from a training data of 11449 buildings around the Zurich region and are able to identify 9 classes of aspect and 5 classes of slope with varying accuracies depending on the classes. While highly represented classes are relatively well identified with an accuracy of 70%, under-represented classes suffer from a lack of labeled examples and remain challenging to identify. The building rooftop geometric estimation is designed to be embedded in a solar PV rooftop potential, which will be carried out in the future.

Future work will present improvements to increase the accuracy of the methodology, including: (1) the use of more labeled training data, particularly for roof shape classification, and (2) more complex feature engineering (trying different features that could have a better correlation with the outputs) to improve the performance of the RF.

ACKNOWLEDGMENTS

This research has been financially supported by the CTI (Commission for Technology and Innovation) within the SCCER Future Energy Efficient Buildings and Districts, FEEB&D (CTI.2014.0119).

REFERENCES

- [1] Gutschner, M., Nowak, S., and Toggweiler, P., "Potential for building integrated photovoltaics," *IEA-PVPS Task 7* (2002).
- [2] Wiginton, L., Nguyen, H., and Pearce, J. M., "Quantifying rooftop solar photovoltaic potential for regional renewable energy policy," *Computers, Environment and Urban Systems* **34**(4), 345–357 (2010).
- [3] Melius, J., Margolis, R., and Ong, S., "Estimating rooftop suitability for pv: a review of methods, patents, and validation techniques," *National Renewable Energy Laboratory (NREL), Golden, CO* (2013).
- [4] Alexander, C., Smith-Voysey, S., Jarvis, C., and Tansey, K., "Integrating building footprints and lidar elevation data to classify roof structures and visualise buildings," *Computers, Environment and Urban Systems* **33**(4), 285–292 (2009).
- [5] Boz, M. B., Calvert, K., and Brownson, J. R., "An automated model for rooftop pv systems assessment in arcgis using lidar," *AIMS Energy* **3**(3), 401–420 (2015).
- [6] Gooding, J., Crook, R., and Tomlin, A. S., "Modelling of roof geometries from low-resolution lidar data for city-scale solar energy applications using a neighbouring buildings method," *Applied Energy* **148**, 93–104 (2015).
- [7] Zang, A., Zhang, X., Chen, X., and Agam, G., "Learning-based roof style classification in 2d satellite images," in *[SPIE Defense+ Security]*, 94730K–94730K, International Society for Optics and Photonics (2015).
- [8] Wang, O., Lodha, S. K., and Helmbold, D. P., "A bayesian approach to building footprint extraction from aerial lidar data," in *[3D Data Processing, Visualization, and Transmission, Third International Symposium on]*, 192–199, IEEE (2006).

- [9] Rottensteiner, F., Trinder, J., Clode, S., and Kubik, K., “Building detection using lidar data and multi-spectral images,” in [*Digital Image Computing: Techniques and Applications*], **2**, 673–682, CSIRO (2003).
- [10] Zhang, X., Zang, A., Agam, G., and Chen, X., “Learning from synthetic models for roof style classification in point clouds,” in [*Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*], 263–270, ACM (2014).
- [11] Izquierdo, S., Rodrigues, M., and Fueyo, N., “A method for estimating the geographical distribution of the available roof surface area for large-scale photovoltaic energy-potential evaluations,” *Solar Energy* **82**(10), 929–939 (2008).
- [12] Kohavi, R. et al., “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in [*Ijcai*], **14**(2), 1137–1145 (1995).
- [13] Breiman, L., “Random forests,” *Machine learning* **45**(1), 5–32 (2001).
- [14] Breiman, L., Friedman, J., Stone, C., and Olshen, R., [*Classification and Regression Trees*], The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis (1984).
- [15] Breiman, L., “Bagging predictors,” *Machine learning* **24**(2), 123–140 (1996).
- [16] Freund, Y. and Schapire, R. E., “A decision-theoretic generalization of on-line learning and an application to boosting,” in [*European conference on computational learning theory*], 23–37, Springer (1995).
- [17] Liaw, A., Wiener, M., et al., “Classification and regression by randomforest,” *R news* **2**(3), 18–22 (2002).