SKOGSSTYRELSEN

The Swedish Forest Agency is the government body for forests and forestry issues. We are interested in knowing where there are forests with high conservation value. These are forests that are important for the conservation of biodiversity. They have been able to develop naturally without the impacts of modern-day forestry such as logging and thinning, so display in general a greater variation than managed forests (for example, variation in species composition, structural diversity). Since 1993, the Forest Agency has mapped high conservation value forests though field surveys. We have approximately 67 000 areas (=nyckelbiotoper in Swedish) delimited in our database, ranging from a single ancient tree to hundreds of hectares of native forest. It is however costly and time consuming to conduct field surveys, so we're interested in seeing how far we can come with national continuous cover datasets, such as the laser scanning that has been done over the country, satellite images and aerial photos, to identify high conservation value forests.

The natural word is complex, and high conservation value forests vary in their properties. They can be coniferous forests, deciduous or a mixture; they can be on steep ground or flat ground; they can be near water or on very dry soil; tree height and dimensions vary based on local conditions, to name a few factors. Our database over high conservation value forests consists of delimited polygons. For each polygon, data has been gathered in the field on a number of qualities, such as tree-species composition, soil moisture, type of habitat etc. The properties of the forest within the polygon can vary. Different types of habitat can be present within the same polygon, though all have high conservation value. For example, a sparse pine forest on a hill descends onto a rocky slope dominated by tall spruce and this is delimited as one polygon.

We would like to know if our database over forests with high conservation value could be used in machine learning, to train a model to recognise similar forests.

1. Is this dataset appropriate to use as training data?
2. If yes, how should the dataset be prepared to be optimal training data? For example:
    - Should the data set be divided into subsets of sites that exhibit similar characteristics, based on the information that has been registered?
    - In this case, how should the division be done?
    - Should the dataset be divided into subsets of sites that exhibit similar characteristics in some other way, for example by extracting information from the national continuous cover datasets?
    - Should the polygons be edited in some way to improve accuracy?
    - Is a dataset of 'negative' data necessary?