# Crowdsourcing EO datasets to improve cloud detection algorithms and land cover change

Matej Aleksandrov, Matej Batič, Grega Milčinski, Linda See, Christoph Perger, Inian Moorthy, Steffen Fritz

Involving citizens in science is gaining considerable traction of late. With positive examples (e.g. Geo-Wiki, FotoQuest Austria), a number of projects are exploring the options to engage the public in contributing to scientific research, often by asking participants to collect some data or validate some results. The International Institute for Applied Systems Analysis (IIASA), with extensive experience in crowdsourcing and gamification, has joined Sinergise, Copernicus Masters 2016 winners, to engage the public in an initiative involving ESA's Sentinel-2 satellite imagery.

Sentinel-2 imagery offers high revisit times and sufficient resolution for land change detection applications. Unfortunately, simple (but fast) algorithms often fail due to many false-positives: changes in clouds are perceived as land changes. The ability to discriminate of cloudy pixels is thus crucial for any automatic or semi-automatic solutions that detect land change.

A plethora of algorithms to distinguish clouds in Sentinel-2 data are available. However, there is a need for better data on where and when clouds occur to help improve these algorithms. To overcome this current gap in the data, we are engaging the public in this task. Using a number of tools, developed at IIASA, and Sentinel Hub services, which provide fast access to the entire global archive of Sentinel-2 data, the aim is to obtain a large data resource of curated cloud classifications. The resulting dataset will be published as open data and made available through Geopedia platform.

The gamified process will start by asking users if there are clouds on a small image (e.g. 8x8 pixels at the highest Sentinel-2 resolution of 10 m/px), which will provide us with a screening process to pinpoint cloudy areas, employing Picture Pile crowdsourcing game from IIASA. The next step will involve a more detailed workflow, as users will get a slightly larger image (e.g. 64x64 pixels) and will then be asked to delineate different types of clouds: opaque clouds (nothing is seen through the clouds), thick clouds (where the surface is still discernible through the clouds), and thin clouds (where the surface is unequivocally covered by a cloud); the rest of the image will be implicitly cloud-free. The resulting data will be made available through the Geopedia portal, both for exploring and downloading. This paper will demonstrate this process and show some results from a crowdsourcing campaign.

The approach will also allow us to collect other datasets in a rapid and efficient manner. For example, using a slightly modified configuration, a similar workflow could be used to obtain a manually curated land cover classification data set, which could be used as training data for machine learning algorithms.