Statistical machine learning for adaptive ground truth sampling design

Invited submission for FAO-UN Expert Group Meeting (EGM) on Ground Truth Data (13-15 September 2022)

Dr Jacinta Holloway-Brown (University of Adelaide, Australia) and Distinguished Professor Kerrie Mengersen (Queensland University of Technology, Australia)

Abstract
Earth observation (EO) data is well acknowledged and widely used as a highly valuable data source for producing selected SDG indicators and official statistics by National Statistical Offices (NSOs), United Nations groups and the research community. Many satellite images and EO datasets are freely available at regular intervals globally and provide long term historical time series. The promise of beneficial insights from these data is unlocked by a suite of statistical machine learning methods, including random forests, support vector machines and neural networks. However, many of these methods require collection of appropriate ground truth data to train the models and validate analyses. Sampling strategies for ground truth data need to be carefully designed, balancing often limited financial and personnel resources with sufficient rigour to produce reliable statistics. We propose that selected statistical machine learning methods can effectively inform the sampling design for ground truth collection at the beginning of the pipeline, in addition to being used to analyse satellite images. We present our method, Stochastic spatial random forest (SS-RF) as an example of how statistical machine learning methods can be used to highlight areas of uncertainty of land cover based on EO analysis and prioritise areas for ground truthing. Our method produces predictions of land cover classes based on current modelled data and past observed images in addition to spatial maps of probabilities. These maps highlight areas of high and low uncertainty of the estimates. These maps of probabilities can provide additional information for an existing sampling design, or become the basis of a new adaptive design.

Introduction

The United Nations has recognised that prohibitive costs of field data collection is a key factor preventing some countries from monitoring the SDGs (Sachs et al., 2019) (Espey et al., 2015). Satellite image analysis based on free imagery is a proven solution to this challenge, allowing monitoring of important official statistics and SDGs over large areas over time. However, some amount of ground truth is typically needed to train and/or validate the models. The trade off between statistically sound sampling design and resource constraints is a known challenge (Delince et al., 2017). Given the resource constraints that have led to the use of satellite image analysis, it is essential that any ground truth sampling plan deliver as much useful information as possible for analysis and validation, while keeping costs low.

Adaptive sampling design

Adaptive sampling design, which is commonly known for its use in medical trials, involves changing experimental designs in response to collected data and outcomes to reduce costs and improve efficiency while maintaining the validity of the overall study (Chang, 2016; Mahajan & Gupta, 2010). This approach is useful in a SDG monitoring context by informing adaptive prioritisation of data collection in response to statistical modelling results to ensure optimal allocation of limited resources over large areas. It is also important for satellite image analysis from an accuracy perspective as land cover and land use change over time. Sampling strategies that are able to reflect change in a more timely way can enhance both the validation of statistical models and accuracy of the model results. We propose that selected statistical machine learning methods can effectively inform the adaptive sampling design for ground truth collection at the beginning of the satellite image analysis pipeline, in addition to its customary use in the training and validation stages (see Fig 1).

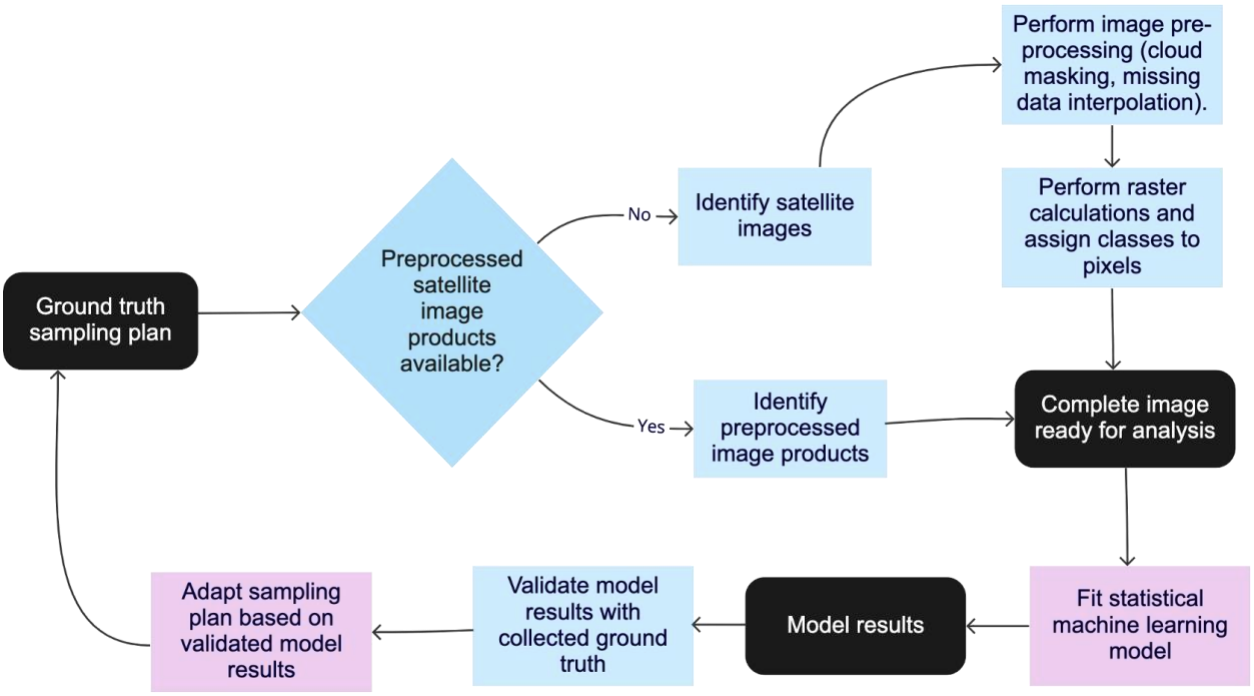

Fig 1: Satellite image analysis pipeline. Black indicates an output, blue and purple are actions in the pipeline, with purple indicating actions focused on in this discussion paper.

Stratified sampling has been recommended as best practice for land cover estimation to ensure that classes of interest and rarer classes are appropriately identified (Food and Agriculture Organization of the United Nations, 2016). Under an adaptive sampling approach, a baseline stratified sample design can be augmented by more targeted sampling, depending on the analytic aim. See table 1.

Analytic aim	Sampling strategy

Improve predictions of areas of interest, under-represented areas and/or small classes	Targeted sampling (TS) of these areas
Reduce overall prediction uncertainty	Increased sample size for current stratified design; TS of specific areas
Model validation	TS of areas of interest and/or smaller classes
Gain knowledge of a previously unsampled area	Stratified or cluster sampling

Table 1: Analytic aims that can benefit from adaptive designs and potential sampling strategies

Statistical machine learning methods to inform adaptive sampling design: Stochastic spatial random forest

Under an adaptive sampling design approach, we propose that the outputs of selected statistical machine learning methods can inform future ground truth data collection for each of the aims listed in Table 1. We present our method, Stochastic spatial random forest (SS-RF) as an example of how statistical machine learning methods can be used to highlight areas of uncertainty of land cover based on EO analysis and prioritise areas for ground truthing. Our method uses a Bayesian stochastic framework to produce predictions of land cover classes based on current modelled data and past observed images (if available). The outputs from the model are posterior probabilities of belonging to a pixel class and a pixel classification. From these outputs we can produce spatial maps of probabilities, which highlight areas of high and low uncertainty of the estimates (Holloway-Brown et al., 2021). These probabilities can provide additional information to adapt an existing sampling design or form the basis of a new design.

As an example, consider only the left panel of Fig 2. The class of the dark blue pixels in the centre of this panel (highlighted by an orange oval) is relatively uncertain based on the model results because they have neither a high (close to 100) nor low (close to 0) probability of belonging to the forest class. These results would be categorised as uncertain and a subset of these could be proposed as additional ground truth sample points under an adaptive design. Due to the fact they are georeferenced and analysed at a per pixel level, these probabilities from the model which are presented in Fig 2 can be plotted in various ways, including being layered on top of a map of the original ground truth samples and a satellite image for the area to highlight areas of uncertainty in the original landscape of interest.

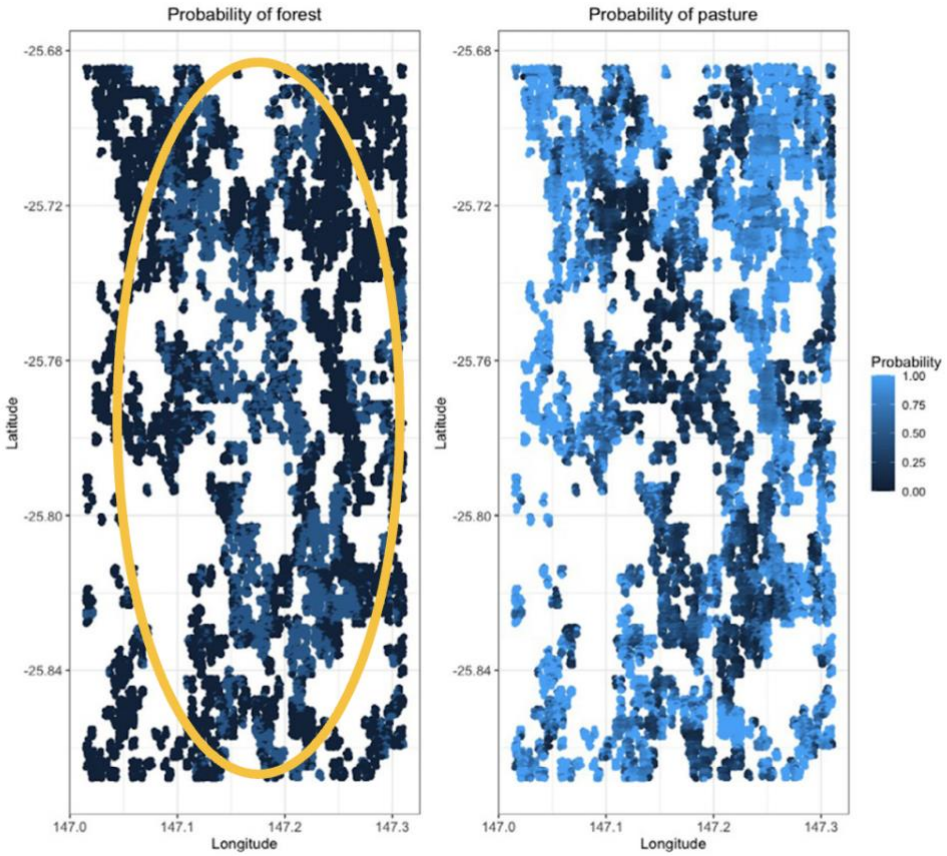

Figure 2: Probability of forest and pasture land cover for the same sample of pixels from a case study in Injune, Queensland, Australia. The probabilities were obtained from our method and plotted by spatial coordinates. The probability that each pixel belongs to a land cover class is indicated by shading: darker blue indicates low probability the pixel belongs to the class and light blue indicates high probability of belonging to that class. The pixels are in patches because they are part of a data sample taken at random. Where the plot is white no data were interpolated for that particular sample. Source: Holloway-Brown et al. 2021.

A key benefit of our method specific to SDG monitoring is accessibility. The spatial random forest model that underpins our SS-RF method relies on only absolute minimum data that can be extracted from free satellite images: latitude and longitude. Provided that there are some observed pixels to train the models at the edges of missing data due to clouds, or previous observations of the same pixels as identified by their geographic position, this is sufficient information to make fit for purpose predictions of land cover. Moreover, our method does not require high performance computing and can be run on millions of pixels on a laptop. Removing these barriers to access is an important contribution of the method, in addition to its accuracy and uncertainty quantification. Details of the method workflow are in the appendix (Fig 1).

As summarised in Table 1, we propose two main scenarios for using the results of the stochastic spatial random forest outputs to inform ground truth sampling. The first is where no current ground truth sampling strategy exists. For example, if the area has not been visited due to cost (Espey et al., 2015), remoteness or in some cases conflict in the region (Delince et al., 2017; Schneibel et al., 2017). By using models fit to free satellite images and producing probabilities of land cover, we are able to identify areas of uncertainty. These would be prioritised for ground truth collection, particularly where a more comprehensive sampling strategy is not possible due to resource or access limitations. The second, more ideal scenario, is that our method would be validated by an existing ground truth data set and the highlighted uncertain areas would be used to adapt and augment the existing ground truth collection strategy by reprioritising the focus of data collection to use resources more efficiently.

Conclusion

There is promise in an adaptive sampling design approach for planning ground truth collection that would enable increased and improved SDG monitoring. Under this fusion approach of ground truth and model estimates, we would use knowledge of where the model and ground truth agree and disagree to plan future ground truth collection. In situations where ground truthing resources are limited and have never been collected in the area of interest, our statistical machine learning model can act as a first parse to prioritise areas of uncertainty to focus on. Further research in the form of a case study to establish how our approach using land cover estimates with associated probabilities to inform adaptive ground truth sampling design performs in practice is of interest for future work.

References

Chang, M. (2016). *Adaptive design theory and implementation using SAS and R*. Chapman and Hall/CRC.

Delince, J., Lemoine, G., Defourny, P., Gallego Pinilla, F. J., Davidson, A., Ray, S., Rojas, O., Latham, J., & Frédéric, A. (2017). *Handbook on remote sensing for agricultural statistics*. https://doi.org/10.13140/RG.2.2.13259.69920

Espey, J., Swanson, E., Badiee, S., Christensen, Z., Fischer, A., Levy, M., Yetman, G., de Sherbinin, A., Chen, R., Qiu, Y., Greenwell, G., Klein, J., T., J., M., Jerven, Cameron, G., Aguilar Rivera, A. M., Arias, V. C., Lantei Mills, S., & Motivans, A. (2015). *Data for Development: A Needs Assessment for SDG Monitoring and Statistical Capacity Development*. https://sustainabledevelopment.un.org/content/documents/2017Data-for-

Development-Full-Report.pdf

Food and Agriculture Organization of the United Nations. (2016). *Map Accuracy Assessment and Area Estimation: A Practical Guide*. Food and Agriculture Organization of the United Nations. http://www.fao.org/3/a-i5601e.pdf

Holloway-Brown, J., Helmstedt, K. J., & Mengersen, K. L. (2021). Interpolating missing land cover data using stochastic spatial random forests for improved change detection. *Remote Sensing in Ecology and Conservation*, rse2.221. https://doi.org/10.1002/RSE2.221

Mahajan, R., & Gupta, K. (2010). Adaptive design clinical trials: Methodology, challenges and prospect. *Indian Journal of Pharmacology*, *42*(4), 201–207. https://doi.org/10.4103/0253-7613.68417

Sachs, J., Mccord, G., Maennling, N., Smith, T., Siamak, V. F.-T., & Loni, S. (2019). *SDG COSTING & FINANCING FOR LOW-INCOME DEVELOPING COUNTRIES 1 Acknowledgements Prepared by the SDSN Costing and Financing Team under the direction of Professor Jeffrey D. Sachs, Director of the UN Sustainable Development Solutions Network (SDSN) The SDSN*. Sustainable Development Solutions Network.

Schneibel, A., Frantz, D., Röder, A., Stellmes, M., Fischer, K., & Hill, J. (2017). Using Annual Landsat Time Series for the Detection of Dry Forest Degradation Processes in South-Central Angola. *Remote Sensing*, *9*(9), 905. https://doi.org/10.3390/rs9090905

Appendix

Fig 1:

Workflow of Stochastic Spatial Random Forest method (Holloway-Brown et al, 2021).

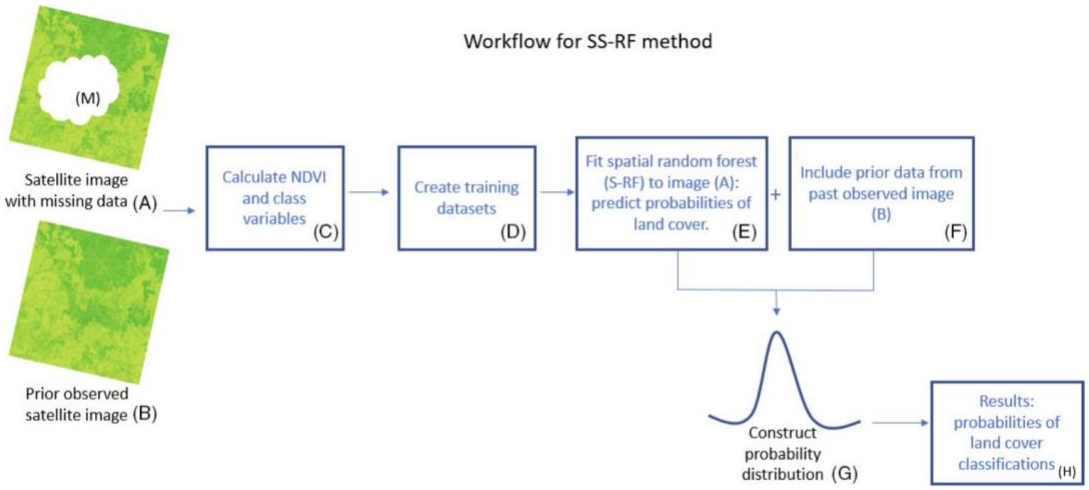

Figure 5. Overview of SS-RF method. SS-RF, stochastic spatial random forest. The components of the method are labelled A-H and M. The components are as follows: (A) Satellite image with missing data (M), (B) Prior observed satellite image, (C) Calculate NDVI and class variables, (D) Create training datasets, (E) Fit Spatial random forest (S-RF) to image (A), (F) Include prior data from image (B), (G) Construct probability distribution and (H) Produce probabilities of land cover classifications.