Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-59-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "Using SHAP to interpret XGBoost predictions of grassland degradation in Xilingol, China" *by* Batunacun et al.

Anonymous Referee #1

Received and published: 14 July 2020

This manuscript tests whether XGBoost can provide alternative insights that conventional land-use models are unable to generate. The overall methodology is interesting. I have a number of major comments before I can suggest the paper for publication.

-Line 54: "Some such models are spatial (e.g. CLUE-S, GeoSOS-FLUS, LTM, Fu et al., 2018; Liang et 55 al., 2018; Pijanowski et al., 2002, 2005; Verburg & Veldkamp, 2004; Zhang et al., 2013); others are not (e.g. Markov models; Iacono et al., 2015; Yuan et al., 2015)." Authors should be aware that all land use change models are spatial models. Markov models are used to estimate the quantity of change from one land use state to another but are not land use change simulators.

-Line 57: "Hybrid models, which combine different approaches to make the best use of the advantages of each model, are another important variety. This type of model

is used to characterise the multiple aspects of LUCC patterns and processes (Li and Yeh, 2002; Sun and Müller, 2013)." Authors did not discuss important other land use modeling approaches such as Cellular Automata (CA), Agent-Based (AB) and a hybrid CA-AB (e.g., Mustafa et al., 2018, 2017; Vermeiren et al., 2016).

>Mustafa, A., Cools, M., Saadi, I., Teller, J., 2017. Coupling agent-based, cellular automata and logistic regression into a hybrid urban expansion model (HUEM). Land Use Policy 69, 529–540.

>Mustafa, A., Heppenstall, A., Omrani, H., Saadi, I., Cools, M., Teller, J., 2018. Modelling built-up expansion and densification with multinomial logistic regression, cellular automata and genetic algorithm. Computers, Environment and Urban Systems 67, 147–156.

>Vermeiren, K., Vanmaercke, M., Beckers, J., Van Rompaey, A., 2016. ASSURE: a model for the simulation of urban expansion and intra-urban social segregation. International Journal of Geographical Information Science 30, 2377–2400.

-Line 143: "The DEM data were extracted from the SRTM 90m resolution and, after resampling, all data were processed into 1×1 km2 raster files." Why do you resample the data to such low resolution? and what is the resample method do you employ?

-Line 146: "All distance measures were extracted from LUCC datasets from the years 2000 and 2015 using ArcGIS Euclidean distance". Euclidean distance is a basic GIS process that can be performed by many tools. No need to mention specific software for such a basic GIS analysis.

-Table 1 presents data with inconsistent dates (2000, 2015, or 2000, 2010). Please justify as this will bias the results.

-Line 207: "In our case study, 18,190 pixels (about 10% of the total) were selected by different sampling methods (Fig. S 3) to train (66% of the sample size) and test (34% of the sample size) the model." Please provide more details about your sample. Is it a

binary (0 no changes, 1 changes) excluding grassland with no change between 1975 and 2015?

-Figures 3 and 4: this evaluation of model performance was done for which period 1975-2000 or 2000-2015? AND do you consider all cells in the study are or the observed changes between two dates? Also, there is a sharp difference in performance between the Logit model and XGB, why? According to many studies that compared Logit with machine learning (ML) methods, ML outperformed logit but not such huge differences as presented in this study.

-Figure 6: can you present the variables' importance (Odds ratio) of the logit model as well? This will help readers to understand the differences between the two methods.

-Figure 9: I am confused about this probability map. I see that almost all pixels have a probability of either 100% (1) or 0% (0). So, is it really a gradient probability map? Another fundamental question, if we need to simulate future scenarios that assume a change of 100 pixels out of 1000 pixels (as an example) then this map is not useful as many pixels have a probability value of 100%. Should the model make a random selection from pixels with a 100% probability??

-English needs improvements.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-59, 2020.