ახალგაზრდა მეცნიერთა საერთაშორისო კონფერენცია „დედამიწის შემსწავლელი მეცნიერების თანამედროვე პრობლემები“
შრომები, ISBN 978-9941-36-044-2, თბილისი, საქართველო, 21-22 ნოემბერი, 2022 წ.
International Conference of Young Scientists "Modern Problems of Earth Sciences"
Proceedings, ISBN 978-9941-36-044-2, Tbilisi, Georgia, November 21-22, 2022

REVIEW OF MACHINE LEARNING METHODS IN THE ESTIMATION OF GREENHOUSE GAS EMISSIONS

Kerimov B., Chernyshev R.

Grozny State Oil Technical University, Chechen Republic

Abstract. *An ongoing climate change poses increasing challenges to the public interest. At the same time, digital transformation fosters the development and application of a multitude of different machine learning (ML) models. This work contains a scientific review of recent applications of ML models in the estimation and modelling of greenhouse gas emissions. We provide an overview of the main challenges and the performances of implemented methods and possibilities for future directions.*

Keywords: Greenhouse Gas Emissions, Machine Learning

Introduction

Anthropogenic influence in global warming cannot be underestimated. Increasing greenhouse gas (GHG) emissions caused by humans are a major force of global warming. These emissions are caused by a wide range of human activities from heavy industry to agriculture and daily routine activity. According to IPCC reports significant anthropogenic GHGs are CO_2, N_2O, CH_4 and CFCs. Agricultural GHGs emission has a major role in several countries without developed heavy industry (or with regularised heavy industry) in example Canada and Brazil [1, 2]. Agricultural soils both are a source and a sink for GHGs, however, small changes in the soil respiration process may cause significant changes in GHGs emission balance from sink to emission. According to the IPCC report, 60% of total N_2O and 50% of total CH_4 emissions come from agriculture. And emission of agriculture GHGs increasing year by year and emissions may escalate due to global population growth. CO_2 emits from microbiological decay or burning plants, and CH_4 is produced from organic decomposition and oxygen-deprived fermentation. N_2O produces from microbiological transformation, especially in wet conditions with high humidity or large precipitations with a predicted increase from 30% to 2030s. Modelling and researching of GHGs agriculture emissions is reasonably justified.

For this purpose, in Chechen republic in 2021 year was constructed system of carbon polygons. These polygons are located in former agriculture fields, former oil-development polygons, natural forests and former anthropogenic landscapes. Measurements on polygons presented by GHGs in-situ chambers measurements, flux towers and regular meteorological observations may be used in GHGs modelling. Also, NDVI and EVI indexes calculated with remote sensing are useful data for describing conditions on agricultural fields and should be used in modelling. Various biophysical models for GHGs cycle simulation have been developed, such as DNDC, DAYCENT, DSSAT. These models have proven their effectiveness, but are very sensitive to data conditions and physical parametrizations which are not always can be well identified and also required qualified users to operate with them. Machine learning (ML) algorithms are good for overcoming classic model problems. Machine learning algorithms are rapidly developing lately, with the increasing of their generalizing ability. A wide range of ML algorithms has been developed based on different ML techniques with special advantages and disadvantages of specific methods. This article presents

the results of some of the research works for ML modelling of GHGs. This experience is very useful for future modelling based on carbon polygons data.

This review focuses on the applications of machine learning methods on a local scale of the researched site. Some of the related studies centre around the prediction of global GHG emissions based on socioeconomic and geographical factors, which is outside of the scope of the paper. To our knowledge, there have not yet been systematic reviews of recent works in estimating GHG emissions with machine learning models. Toby et. al. [3] provided a comparison of the application of 4 families of machine learning in the estimation of CO2 fluxes in cropping fields.

Research Methodology

In our analysis, we highlight the following aspects of the application. Firstly, we investigate the context, scale, and purpose of the researched site. Naturally, the investigated site defines the mechanism, the scale, and the nature of emitted greenhouse gases. This includes cropping and various agriculture. The context is crucial in describing the problem and the domain knowledge largely defines the structure of the model. Lastly, we study the models applied in the research, as well as the nature and the scale of the measurements used in the modelling, and the machine learning method.

Estimation of GHG Emissions

Tackling the problem of predicting greenhouse emissions from agricultural soils, Hamrani et. al. analysed a pool of several ML algorithms from classical regression models to deep learning neural networks. The models predicted CO_2 and N_2O emission based on measurements from agricultural fields in Quebec, Canada: air and soil temperature, soil volumetric water content, air humidity, precipitation, atmospheric pressure and crop N uptake as predictors; CO_2 and N_2O emission measured with chambers as predictands. Data was measured in 3 different plots size 75x15 meters during 6 years from 2012 to 2017 with crop rotation (corn, soybeans, oat). CO_2 data is cyclically emitted throughout every year, N_2O is seasonally emitted with sharp peaks. The authors identified the feature importance of each gas. Feature importance has been evaluated with neighbourhood component analysis (NCA) and minimum redundancy maximum relevance (MRMR) analysis. For CO_2 most important features are air and soil temperatures, soil volumetric water content and humidity, for N_2O soil water volumetric content and precipitation are the most important which corresponds with other authors [4, 5]. Models to choose from, presented by classic regression models: SCM, LASSO, Random Forest; shallow learning: FNN, RBFNN; deep learning: LSTM, CNN, DVN. The minimized quality metric is RMSE, also quality is evaluated with R^2. With selected models, the authors concluded that the LSTM model has the best performance for CO_2 and N_2O prediction with best score R^2 = 0.87, RMSE 30.3 mg $m^{-2}h^{-1}$.

On the example of a contrasting environment, Freitas et. al. accent on CO2 emissions over green cane fields in Brazil. Data was collected from three different fields after the harvest period with Li-Core 8100 flux gas analyzer system (2008, 2010 and 2012) year. The authors employed a multilayer perceptron (MLP) with three layers to estimate FCO2 emission. Training data soil samples were collected from 0 to 0.10 m in depth. The following routine tests were carried out: determining the content of organic matter (OM), the available P, K, Ca, Mg, and H + Al, and establishing the calculation of the sum of the bases and the exchange cation capacity. Custom feature engineering may help evaluate FCO2 with higher accuracy with lack of train data. The authors with their custom features reached good quality with mean absolute percentage error (MAPE) 18.4% and R2 = 0.92 for predicting 2012 years on 2008 and 2010 train data. The authors did not check the stability of the model with known validation methods, however, their model well agreed with FCO2 emission peaks in 2012 year, which may conclude custom feature engineering was performed thoroughly.

As a comparison, the work by Oertel et. al. introduces a random forest regression model in order to estimate N_2O emissions on the dataset collected from a set of experiments. The dataset contributed by Steh-

fest, E. et. al. [6] includes results of experiments from over 114 publications focused on the prediction of nitrous oxide emissions on various cropping fields. As a result of pre-processing, they extracted 19 input variables, categorized into soil contents, crop type, geographical, and fertilizing agent features. The authors show that RF systematically outperforms the regression model evaluated on the same data. The accuracy was calculated based on rooted mean squared error (RMSE) in kg N ha^{-1} $year^{-1}$. Additionally, the work shows the importance of fertilizer and crop type as identified by RF ranking of input variables. As the authors stated, the explainability of the method is limited due to the impossibility of explicitly including external factors, such as soil tillage and regional effects. However, they cause an indirect influence on some of the included variables.

Machine Learning models

Decision Trees and Random Forests

Decision trees are a family of tree-like algorithms that are based on binary decision rules. They have been particularly useful in various applications due to their explainability and robustness to missing data [5]. XGBoost and Random Forest extend the family by ensembling the prediction of individual trees and thus achieving higher robustness and generalisation.

Deep Neural Networks (DNNs)

Ongoing digitalization and advances in computation opened the way to deep learning models. These models usually are equipped with multiple layers of various neural networks, such as ANNs, convolutional operators, and recursive architectures. Convolutional neural networks (CNNs) are a standard in image processing applications. Analogously, they are widely used in hyper- and multi-spectral analysis, which was found particular use in the estimation of GHG emissions and carbon stock. One of the biggest drawbacks of neural networks is the “black box” nature, which hinders their explainability. Furthermore, these models usually require considerable amounts of data. This may be unattainable in many manual and in-situ measurements.

GHG measuring technologies

Machine learning models are heavily reliant on the availability and quality of data. We highlight two main measuring methods

In-situ measurements

GHGs measures with flux methodology. Standard instruments are: chambers and optical gas analyzers. Chambers are especially useful and simple to construct and operate, also chambers are able to measure very low rates of fluxes in a short period of time. Unfortunately, close chambers have their disadvantages such as increasing gas concentration in the chamber may cause measurement errors. Chambers may cause their own greenhouse effect with recurrent problems. Optical gas analyzers are able to measure fluxes and concentrations of GHGs with more than sufficient quality, however, these analyzers are very expensive and affected by a territory's footprint.

Remote sensing

Concentration of gases and atmospheric composition can be measured with a network of land or aerial stations. These monitors are usually equipped with hyper- or multi-spectral imaging instruments and spectrometers and are practical for modeling CO_2 emissions via vegetation indices [7, 8]. For applications such as carbon stock inventory estimation, aerial imagery data provides a less costly and more standardised alternative source to manual labour. Furthermore, openly available measurements stimulate the development

of machine learning models. Although remote sensing with lidar measurements may evaluate a large territory's emission for a short time, however, it requires very careful preparation of the orthophotomap of the territory. Remote sensing doesn't have in-situ measurement quality. The scale of the errors of aerial imagery is radically higher than permissible deviations of in-situ measurements.

Conclusion

Availability of computing resources and ongoing digitalization of data fostered the development and adoption of ML methods. Although remote sensing and automated in-situ measurements of GHG emissions add up to the available data, manually labelled data is especially limited. This reinforces the drawbacks of some ML methods. One of the potential solutions to that is introducing inductive biases in the models or using advanced learning methods, such as semi- and self-supervised learning.

Additionally, we note that geographical and geoecological conditions impede the comparison of the performance of the models applied. At the same time, there is a positive outcome from compiling and sharing measurement results from different sources.

This work was carried out as a part of a scientific project:

Grozny State Oil Technical University FZNU-2021-0012

"Complex interpretation of geophysical and geoecological data in studies of greenhouse gas balance (on the example of Chechen Republic)".

References

1. Hamrani A., Akbarzade A., Madramootoo, C.A. Machine learning for predicting greenhouse gas emissions from agricultural soils//Science of the Total Environment, (2020), Volume 741 https://doi.org/10.1016/j.scitotenv.2020.140338
2. Freitas L.P.S., Lopes M.L.M., Carvalho L.B. et al. Forecasting the spatiotemporal variability of soil CO2 emissions in sugarcane areas in southeastern Brazil using artificial neural networks // Environ Monit Assess 190, 741 (2018). https://doi.org/10.1007/s10661-018-7118-0
3. Adjuik, Toby A., and Sarah C. Davis. Machine Learning Approach to Simulate Soil CO_2 Fluxes under Cropping Systems // Agronomy (2022), 12, no. 1: 197. https://doi.org/10.3390/agronomy12010197
4. Berglund Ö., Berglund,K., Klemedtsson L. A lysimeter study on the effect of temperature on CO2 emission from cultivated peat soils//Geoderma (2010) 154, 211–218. https://doi.org/10.1016/j.geoderma.2008.09.007
5. Oertel C., Matschullat J., Zurba K., Zimmermann F., Erasmi S. Greenhouse gas emissions from soils—A review // Geochemistry (2016), Volume 76, Issue 3, Pages 327-352, https://doi.org/10.1016/j.chemer.2016.04.002
6. Stehfest E., Bouwman L. N2O and NO emission from agricultural fields and soils under natural vegetation: summarizing available measurement data and modeling of global annual emissions // Nutrient Cycling in Agroecosystems (2006) Volume 74, pages 207–228 (2006). https://doi.org/10.1007/s10705-006-9000-7
7. Liu C., Xing C., Hu Q., Wang S., Zhao S., Gao M. (2022). Stereoscopic hyperspectral remote sensing of the atmospheric environment: Innovation and prospect//Earth-Science Reviews, 226, 103958. https://doi.org/10.1016/-J.EARSCIREV.2022.103958
8. Della-Silva J.L., da Silva Junior C.A., Lima M., Teodoro P.E., Nanni M.R., Shiratsuchi L.S., Teodoro L.P.R., Capristo-Silva G.F., Baio F.H.R., de Oliveira G., de Oliveira-Júnior J.F., Rossi F.S. CO2 Flux Model Assessment and Comparison between an Airborne Hyperspectral Sensor and Orbital Multispectral Imagery in Southern Amazonia // Sustainability 2022, 14, 5458. https://doi.org/10.3390/su14095458