Soil Organic Carbon Predictor based on temperature, precipitation, soil tilling, cover crop, and crop

Mekaela Stevenson
28525

Introduction

Farm management techniques have an impact on soil organic carbon, including tilling, cover crops, and fertiliser use (Krause et. al, 2023), presenting an opportunity to find practices to improve this soil. Climate change has exemplified these issues and made management changes increasingly important, and so landowners are increasingly under pressure to change their management practices. Increasing soil carbon in all types of agriculture, forestry, and urban land holds potential for sequestering large amounts of carbon dioxide. This tool will allow landowners and policymakers to estimate the carbon levels of the land they manage by inputting the management techniques and site details. Users can then change the management choices to find the best choices for their site, which will help them choose how to manage their land.

Data

The SoilHealthDB created by Jian et. al (2020) was used to train this tool, which provides data collated from studies across 354 sites across 42 countries. This dataset includes several different types of soil organic carbon measurements. There are many inconsistencies and missing fields in this dataset, and so the data cleaning step of this project required several simplifications of the data. The fields were chosen for their impact on soil carbon and the ease of someone knowing the details for the property they want to predict.

One area of missing data that severely restricts its use is that the vast majority of records do not include annual temperature and precipitation. As these have a large impact on soil organic carbon levels, it means that many records could not be used without significant extra work of calling a weather API with the provided latitude and longitude. It was not possible to find an API that provided mean annual temperatures by latitude and longitude, and so weekly records over several years would have had to be fetched for each location and an average calculation done for each. As there were so many records needing this, it was decided to train the model with only the records which contained annual temperature and precipitation data. This provides a clear area to improve the model, with data that is already available. After removing these records, there were 414 records, which becomes 828 as each record contains two data points - the control and the treatment data.

The use of fertiliser was initially proposed to be used with the model, however the data from this field was very difficult to clean and inconsistent. Some fields only had the nitrogen fertiliser used, whereas others had nitrogen and phosphorus, and others listed NPK, ‘varied’, ‘Fertilized’ or other comments. Many do not have data recorded for fertiliser, and it is not clear whether these had fertiliser used or not, as each row has a conventional record, so one would assume they are fertilised. For these reasons it was decided not to include fertiliser use on the tool. Organic carbon also had inconsistencies, and despite having a field for organic carbon concentration, it is mostly missing, so it was decided to use the organic carbon (OC_C and OC_T) field, however it may have inconsistent units, which will have had an impact on training the model and the analysis of the accuracy of predictions.

Other inconsistencies arrived in grain crop and cover crop, where it was sometimes not clear if a field was rotated between crops, or the property had several crops growing. If several crops were growing, the table could have had a record for each crop, however as this was not clear, the grain crop group and cover crop group were used instead. These still needed to be cleaned as they had fields of different levels, for example ‘wheat’ and ‘vegetable’. All fields were changed to the higher level category. Tillage also had a similar problem, and so was also cleaned and changed to the higher level category, e.g. ‘spring tillage’ became ‘Conventional’.

Cleaning this dataset was done in the Jupyter notebook ‘clean_data.ipynb’, with the final output of the notebook being a new csv called ‘soc_dataset.csv’

Data types and descriptions of each variable used are:
‘OC’: float: Soil organic carbon,
‘Tannual’: float: annual temperature,
‘Pannual’, float, annual precipitation,
‘Tillage’, string, the type of tillage used (“No”, “Conventional”, “Reduced”, “Other”),
‘CoverCropGroup’, string: the group of cover crop used (“No”, “Legume”, “Rye”, “Grass”, “Brassica”, “Mixed”, “Other),
‘GrainCropGroup’, string: the group of crop most often planted (“Grain”, “Grain-rotation”, “Vegetable”, Other”)

Data Organization

Both kfolds/test and train/test/validation datasets were tried when defining the model. KFold was slightly better, with an R^2 of 0.76 as compared to 0.758. Using KFolds ended up putting more importance on the annual temperature, as compared to using training/validation/test sets. This is not as useful for helping farmers to improve their soil (they cannot do a lot about the temperature on their farms), however it is more important for the model to be accurate, and so using KFolds was chosen. Shuffle was set to True for KFolds, as it was beneficial for the data to be randomly selected. Data was related to one another, both through many being from the same studies,

and from each original record being split into a control and a treatment record. Neither of these fields (Type - control or treatment, or ExperimentID) were used for training, and so they were not used to ensure equal splits for KFolds, however this could introduce some error.

Figure one shows the feature importance, where it can be seen the annual temperature has a much larger importance than any other feature. Not tilling is the next most important feature, and then precipitation. It would be interesting to also do this analysis with only tillage or no-tillage, and cover crop or no cover crop. This would restrict the options for the user to input, however may more accurately capture the importance of each option, as there are many options in this model with almost no importance.

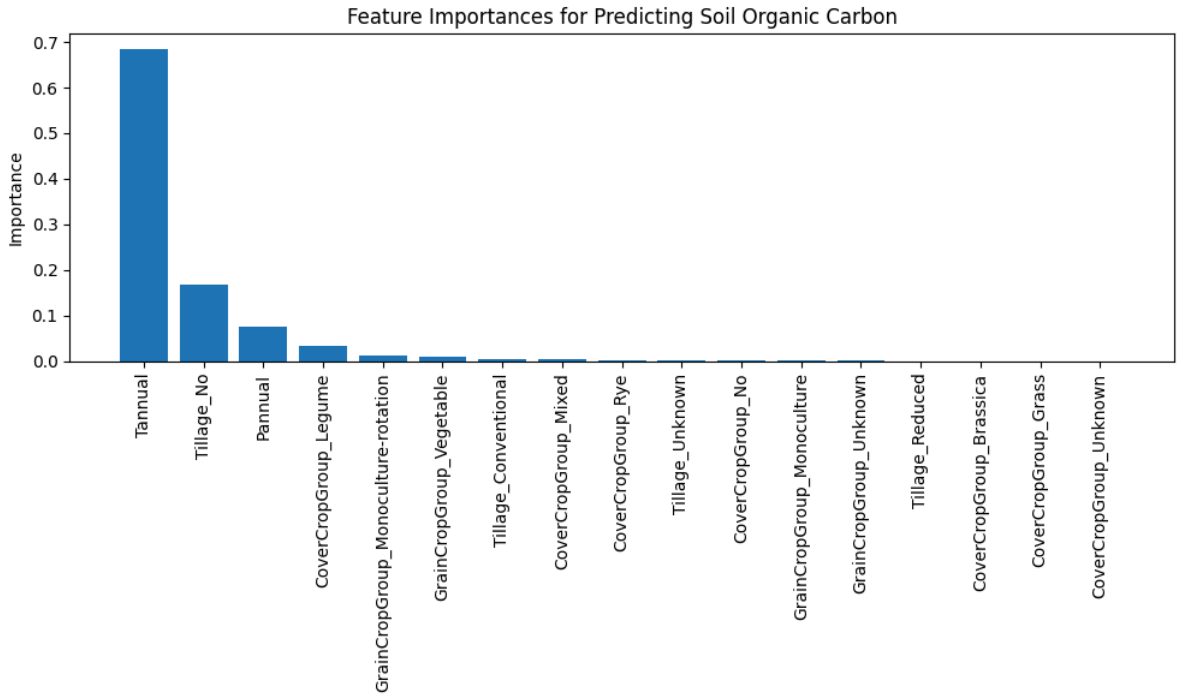

Figure 1: Feature importance for random forest regression model.

Methods

The scikit-learn RandomForestRegressor model was used after being compared with LinearRegression and Ridge scikit-learn models, as it was found to have the highest R^2. A pipeline was used, with a separate transformer for numeric and categorical features, as numeric features needed to be scaled and normalised, whereas categorical features only needed to be separated into binary variables with OneHotEncoder. As the dataset was clean, no cleaning transformations were made in the transformers. It would be useful to add an imputer to both transformers if used with non-cleaned data, however if there was an issue with this data, it was preferable to catch it during the cleaning process in the notebook, so that they could be manually categorised. This happened when the original dataset had the ‘No’ option named as “None”, e.g. for Tillage when there was no tillage. ‘None’ became nan,

which doesn't represent that the answer was 'None', so this was changed in the clean_data notebook and the soc_dataset was recreated.

As mentioned above, scikit-learn KFolds was used to split the data for training. The model and parameters were then chosen using scikit-learn GridSearchCV. The best model and parameters were then used to fit the model with the training data. This was then used for two things - in the notebook it was used to make predictions for the test dataset, which were then further analysed, and this trained model was used with the interface - in app.py - to predict soil carbon based on the user inputs.

Results:
After fitting the model, predictions were made for the test dataset. Figure two shows the results of the actual versus the predicted Soil Organic Carbon(SOC) levels. It is visually simple to see that the model performed much better on lower values than higher values. This is likely due to there being many more lower values in the training dataset and dataset as a whole. This could also reflect on farmland in general having lower SOC values, and therefore there being fewer studies on higher SOC soils.

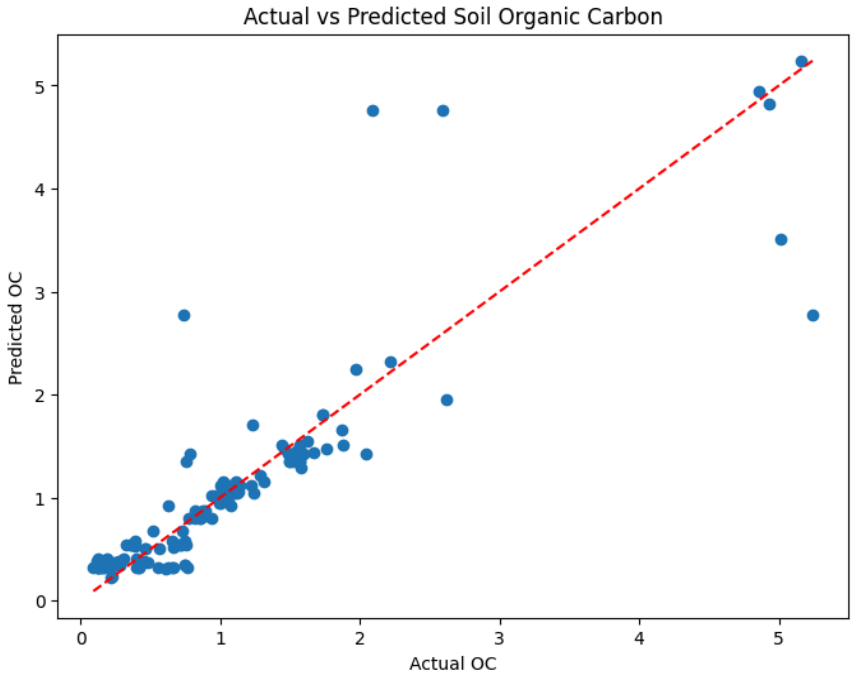

Figure 2: Actual vs predicted soil organic carbon for the test dataset

The user interface also allows a user to input values for annual temperature and precipitation, tillage, cover crop, and crop. This outputs a single value for soil organic carbon.

Analysis
Analysis of the predictions found the following results:
R^2: 0.784
Mean Absolute Error (MAE): 0.241
Root Mean Squared Error (RMSE): 0.486
Mean Absolute Percentage Error (MAPE): 41.59%

The R^2 is 0.764, which means the model explains 76.4% percent of the variance of SOC. As SOC is an ecological variable, it is expected to have a high variance, so this is a strong value, however having additional variable data, such as the fertiliser data which was intended to be used, could improve this value.
The mean SOC is 1.08, and so the RMSE of 0.486 has a 45% error, which is poor. This means that there are many outliers, which appears to be true visually from figure two for the higher SOC values. MAE is less sensitive to outliers, and calculates that predictions are off by 0.241 on average, which suggests that some large errors are inflating the RMSE, which does appear to be the case in figure two. MAPE calculates that predictions deviate from the observations by 41.59% on average, which is high, though considered acceptable for ecological studies which have high natural variability.

This model has difficulty predicting high and medium levels of SOC, however it does not have many data points in this range, so the most obvious solution is to 'unlock' the additional data that is available in the SoilHealthDB by finding the annual temperature and precipitation for this additional data. The other alternative is to only provide predictions for lower SOC levels, however this is counterintuitive to the intention of this tool, which is to inform users on a property's SOC with the aim of helping them to improve it.

Limitations and improvements
There were many limitations with the data used, one of which can be improved by finding the annual temperature and precipitation of records, as some crop groups could not be analysed, and even though there was sufficient data, the range of data across variables could be better. For example, the crop group only offers three options, and with the additional temperature and precipitation data, more crop types would be available, such as "Orchard". This dataset also had very inconsistent fertiliser data, and as fertiliser can increase soil organic carbon in some soil types (Yang et. al, 2024), it could have had an important impact on the analysis.

Some data could be further grouped together, based on the lack of feature importance. For example, giving the user only the option of 'Tilled', or 'Untilled' would make the user experience simpler and make little impact on the results.

An extension to this project would be to take a property's latitude and longitude coordinates as input to provide the user with an analysis of the sentinel-2 imagery

and estimate the soil carbon of the user's property. This would give the user a more accurate estimate of their property's soil organic carbon, however would take more computational power and more time for the user to get an estimate.

Deployment:
The soil carbon predictor tool was made into a web application with two html pages - one with inputs, and one with the prediction. Flask was chosen to deploy this as it is simple to use html, css, and python with a flask app, it is quick to set up (as compared to alternatives like Django), and so is well suited to small projects. The inputs and outputs are shown in figure three.

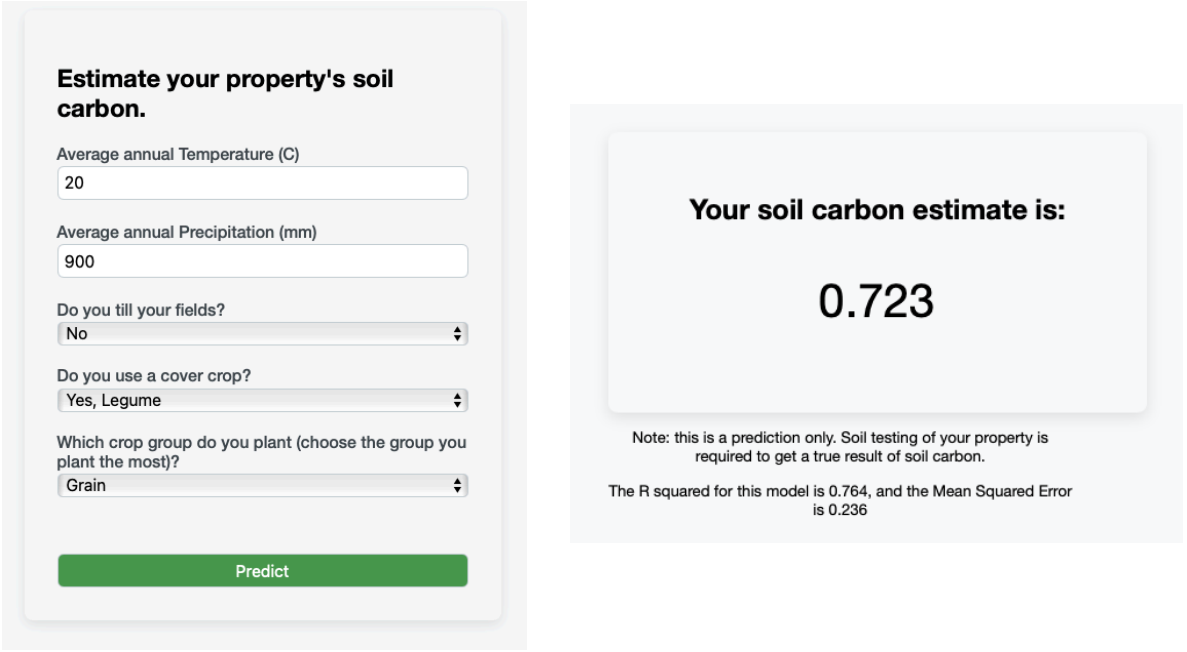

Figure 3: User interface input (left) and output (right)

References:

Jian J, Du X, Stewart RD (2020) A database for global soil health assessment. *Science Data* 7:16 https://doi.org/10.1038/s41597-020-0356-3
The related database SoilHealthDB for this article:
https://github.com/jinshijian/SoilHealthDB/tree/master/data

Krause HM, Stehle B, Mayer J, Mayer M, Steffens M, Mäder P, Fliessbach A (2022) Biological soil quality and soil organic carbon change in biodynamic, organic, and conventional farming systems after 42 years. *Agronomy for Sustainable Development*. 42:117 https://doi.org/10.1007/s13593-022-00843-y

Yang X, Bao Y, Li B, Wang R, Sun C, Ma D, Chen L, Zou H, Zhang J (2024) Effects of fertilization applications on soil aggregate organic carbon content and assessment of their influencing factors: A meta-analysis,
CATENA 242: 108135. https://doi.org/10.1016/j.catena.2024.108135.