Report descriptions

Figures

Residual plots

Generated by 4.2_Residuals+VIF.R. Two different types of plots are generated for each crop. First the studentized residuals of the model are plotted against the fitted values on the link scale (*crop*_residuals_link.png).

The second plot (*crop*_residuals.png) is comprised of four individual plots. Here, the calculations are based on the standardized deviance residuals and the fitted values on the response scale. Just like the first plot, the top left quarter shows the residuals plotted against the fitted values. In the top right quarter, a quantile-quantile-plot is displayed. It plots the quantiles of the sample data against the quantiles to be expected in the assumed underlying distribution (in this case the gamma distribution). To confirm the assumption that the distribution is fitting, the plot should show something close to a straight line. The bottom left quarter plots the square root of the absolute value of the standardized residuals on the y-axis and the predicted values on the x-axis. It is closely related to the standard fitted vs. residual plot and makes it easier to spot heteroscedastic tendencies. The residuals should be evenly distributed across the whole plot to support the assumption of homoskedasticity. The last plot, a residual-leverage-plot can help detect outliers and influential datapoints. For each data point, the standardized deviance residual is plotted against the leverage. The leverage is a measure to quantify the influence of a data point by determining how much the fitted model would change if the y value of the data point was changed while all other points remain constant. To assess if points with much leverage are influential, the plot also has a line which delimits the threshold for an outlier based on Cook's distance (0.5). In the plots for this analysis, the cook's distance contours are not visible so there are no extremely influential outliers in the data set.

Loss of industry plots

Generated by Spatial Distribution Plots.ipynb. The loss of industry plots show the spatial distribution of the projected yield loss for each crop in phase 1 and 2. They are based on the *crop*_phase_*x*_rc_mean.asc files in the processed folder.

Plots LoI1.png and LoI2.png show yield reductions for all crops for phase 1 and 2 respectively. The LoI*crop*.png plots each depict the yield reduction of one crop in both phases.

Yield reduction by region plot

Generated by Reduction_by_continent.ipynb. The yield_reduction_by_region.png plot portrays the projected yield loss for each crop and phase broken down by continent. It is based on the zonal continent statistics saved in the Continent_statistics sheet of the Prediction_statistics.xlsx file.

Statistics

Raw column area

Intermediary file generated by 2_data_preprocessing.py. It contains the total crop area of the n_fertilizer and pesticides data sets for each crop before any of the preprocessing operations are carried out. The values are used in 3_LoI_scenario_data.py to calculate the distribution of nitrogen and pesticides in year 1 after the catastrophe.

Descriptive statistics

Generated by 2_data_preprocessing.py. This Excel file contains five sheets, one for overview statistics and then one for each crop. These statistics characterize the input data for each crop along the process of data cleaning by using several descriptive metrics.

Each statistic is given for four different steps in the data cleaning process as well as for the outliers. The steps are named raw, step1, step2, clean and outliers.

Raw: Data set for each crop is made up of the following columns: latitude, longitude, area, yield, n_fertilizer, n_manure, n_total, pesticides, irrigation_total, irrigation_reliant, mechanized, thz_class, mst_class, soil_class and continents. It contains all rows where the area for the respective crop is above 0 ha.

Step1: Data set for each crop containing all rows where the area for the respective crop is above 100 ha.

Step2: Data set for each crop after the following operations are carried out: levels 7, 8, 9 and 10 of the thz_class are combined into one level; levels 1 and 2 and 6 and 7 of the mst_class are combined into one level each; levels 0, 7 and 8 of the soil_class are replaced with values between 1 and 6; missing values in the fertilizer column are forward filled; missing values in the pesticides and mechanized column are dropped from the data set.

Clean: Data set for each crop after outliers in the columns Yield, n_fertilizer, n_manure, n_total and pesticides are eliminated and the no data values in the continent column are filled.

Outliers: Data points in the columns Yield, n_fertilzer, n_manure, n_total and pesticides which are larger than the 99.9th quantile of their respective column. Data points in the n_manure column which are larger than the 99th quantile of n_manure.

Overview sheet

The *Overview sheet* provides two descriptive metrics for each crop and step: total area in ha and the number of rows (aka data points) in the data set. These values do not vary depending on the column therefore they are only presented once for each crop data set.

Crop-specific sheets

The *crop-specific sheets* provide 5 metrics for all steps and 2 additional metrics unique to the outlier data set. As the metrics presented below are unique to every variable, they are not reported for the whole data set (like the overview statistics) but rather for each model relevant factor individually (each represented by one column in the data set), namely Yield, n_fertilizer, n_manure, n_total, pesticides, irrigation_total, irrigation_reliant, mechanized, thz_class, mst_class, soil_class and continents. Consequently, each metric is reported for each step and each variable except for the outlier specific statistics which are only presented for the variables listed in Outliers (see above).

The area weighted mean is calculated for all continuous variables and for the categorical variables the area weighted mode is provided (0_Weighted_Mean_Mode). The minimum and maximum value of each variable is presented (1_Minimum and 2_Maximum). Further, for each step the number of missing values and the number of zeros for each variable are given in the columns 5_NaN_count and 6_0_count.

The outlier specific metrics are the outlier threshold (3_Outlier_threshold) and the number of outliers (4_Number_Outliers). The outlier threshold is the absolute value of the 99.9th or 99th quantile for each variable.

Model results

Generated by 4.1_GLM_analysis.py. This Excel file contains five sheets portraying the results of the generalized linear model for each crop and additional metrics to assess goodness of fit, multicollinearity and the influence of each variable within each model.

Model_results sheet

The results of the model for each crop are reported in the *Model_results* sheet. The rows each represent one variable relevant to the model. Since the categorical climate variables are coded as dummies, each level of the three variables is represented by its own row. In a practical sense e.g., C(thz_class)[T.2] can only take the values 0 and 1: either the thz_class at the specific data point is T.2 or it is not. The corresponding coefficient therefore measures the difference between T.2 and the reference level. In dummy encoding, one level of each variable is always chosen as the reference level. In this case the levels T1=Tropics, lowland, M2=Length of Growing Period < 120 days and S1=Dominantly very steep terrain are the reference levels for thz_class, mst_class and soil_class respectively and therefore implicit in the intercept. In other words, if all other classes of a categorical variable are coded as zero, it represents a hypothetical 1 for the reference level. As a result, the intercept is the expected value for crop yield without any external nutrient, water or pesticides inputs, without the use of machinery in a tropical lowland climate with a growing period length below 120 days and in dominantly very steep terrain.

For each of these variables five different values are reported for each crop:

Coefficients: The raw generalized linear model coefficients on the link scale. Multiplying each coefficient with the value for the respective variable for a specific data point and summing up all of the results provides the yield predicted for this data point by the model.

Multiplicative_Coefficients: These coefficients are reported on the response scale. They are calculated by exponentiating the coefficients on the link scale. Unlike the coefficients on the link scale, they can not be substituted into the model formula to predict yield for a data point. The multiplicative coefficients represent the change in predicted yield when all variables are held constant except for a one unit increase in one variable. E.g. using the raw coefficients, the yield for n_total at level x is predicted. If this number is then multiplied with the multiplicative coefficient for n_total, it results in the predicted yield for n_total at level x+1 while all other variables are held constant:

$$Yield_{n\_total\ at\ x} \times\ Multiplicative\ Coefficient_{n\_total} = Yield_{n\_total\ at\ x+1}$$

lower_95%_Confidence_Interval/ upper_95%_Confidence_Interval: The upper and lower bounds of the 95% confidence interval (the real value of this parameter lies with a 95% certainty within this range) for the coefficients are reported for the multiplicative coefficient on the response scale as these values can be interpreted more intuitively.

p-value: The p-value for each coefficient. The p-value is a measure to assess if - given the model training data – it is reasonable to assume that the coefficient is unequal to zero, meaning that this specific independent variable has an impact on the outcome of the response variable beyond statistical noise. The significance level for this study is taken to be at 5%, so all p-values above 0.05 constitute a significant effect.

Normalized_results sheet

The normalized_results sheet is structured identical to the model_results sheet. Here the three continuous variables n_total, pesticides and irrigation_tot are normalized to achieve comparability between all coefficients. Normalization is achieved by scaling the values of the three continuous variables to fall between 0 and 1 (subtract the minimum value of the variable and divide by the maximum value). In consequence, all variables have the minimum value 0 and the maximum value 1. This allows to compare the coefficients across all variables: a large coefficient represents a large effect on the response variable and can not be attributed to the variables' units. The p-values are unaffected by this measure.

Model_statistics sheet

The Model_statistics sheet reports different metrics to assess goodness of fit for each crop.

McFaddens_roh: McFadden's ρ^2 is an alternative goodness of fit measure for non-normally distributed data. It is used as an analog to R^2. However, the interpretation differs: unlike for R^2, ρ^2 values ranging from 0.2 to 0.4 represent an excellent fit according to McFadden (1977).

RootMeanGammaDeviance: The root mean gamma deviance serves as a measure for model transportability. It is an alternative to the root mean square error for data following a gamma distribution.

AIC and BIC: The Akaike-Information-Cirterion and the Bayesian-Information-Criterion are two goodness of fit metrics based on the log likelihood which are primarily used to compare different models for one data set with each other. They are not comparable across different data sets and are mainly reported here for completeness.

Since a split-sample approach is applied to calibrate and validate the model, McFadden's ρ^2 and the root mean gamma deviance are calculated for both, the calibration and the validation portions of the data set. This is done to verify that the model also performs well on data points that are not used for its calibration.

YieldReductionperFactor sheet

To assess the influence of each independent variable (excluding the climate variables as they are not subjected to change in the scenario of interest) on the yield of each crop (according to the model), the YieldReductionperFactor sheet presents the predicted reduction in yield for each crop when all variables are at their maximum value except for one, which is set to zero. The values are reported as relative change, so e.g. -0.45 stands for a 45% reduction in yield in comparison to the yield that can be achieved with the maximal input of that variable.

Model_VIF sheet

The variance inflation factor (VIF) (Rawlings et al., 1998) is a measure to check for multicollinearity within a set of predictors. The literature contains different threshold values for when the VIF indicates serious multicollinearity. The most prominent thresholds are specified as everything above 5 (Huang et al., 2010) or everything above 10 (Fox & Weisberg, 2011) constitutes the need for action. However, for dummy-coded variables, the VIF cannot be calculated for the entire predictor but is rather estimated for each level separately. The results of these separate VIF values strongly depend on the category which is taken as the reference level and are consequently highly unreliable. So instead, we report

Crop.Df: The degrees of freedom (Df) are determined by the number of levels of each variable. Consequently, a continuous variable has 1 Df while a categorical variable with 5 levels has 5 Dfs.

Crop.GVIF: The generalized variance inflation factor (GVIF) as introduced by Fox and Monette (1992).

Crop.GVIF..1..2.Df.: To make the GVIF comparable across predictors with a differing number of levels, Fox and Monette (1992) suggests using $GVIF^{\frac{1}{2\times Df}}$ (Df = *Crop*.Df).

Crop.GVIF2: Squaring $GVIF^{\frac{1}{2\times Df}}$ yields the regular VIF for predictors with one level (Df). Hence, *Crop*.GVIF2 can be seen as the standardized equivalent to the VIF. It allows to compare variables with different numbers of levels while still being able to apply the classical VIF rule of thumb (A VIF above 5 or 10 constitutes a problematic collinearity).

Prediction statistics

Generated by 4.1_GLM_analysis.py. This Excel file contains two sheets presenting the results of the yield prediction for each crop in phase 1 and phase 2.

Prediction_statistics sheet

Five descriptive metrics are presented for each crop:

Weighted_Mean: The area weighted mean for yield and the production weighted mean for predicted yield reduction.

Mean_1/2_ci: The upper and lower bounds of the 95% confidence interval (the real value of this parameter lies with a 95% certainty within this range) for the weighted mean can be constructed from this value: Subtracting the mean_1/2_ci value from the weighted mean results in the minimum value and adding it reveals the maximum value of the 95% confidence interval.

Minimum and Maximum: The minimum and maximum values of each data set.

Production: The total yearly crop production in kg for each data set.

Each metric is reported for the (predicted) crop yield (in kg) and the projected yield reduction at different points in time:

Yield_SPAM2010: Crop yield under current conditions based on the SPAM data set and representing conditions ca. 2010. The metrics for these data sets are presented as reference for the following data sets.

Yield_fitted_values: Crop yield under current conditions as predicted by the generalized linear model. In comparison to the original data the model was trained on (Yield_SPAM_2010), the range of the fitted values is notably less wide. The much higher minimum value in the fitted values data set shows that the model does not fit well for the low yields.

Yield_phase1 and Yield_phase2: Projected crop yield in phase 1 and 2 of a scenario with global catastrophic infrastructure loss.

RelativeChange_phase1 and Relative_Change_phase2: Projected yield reduction in phase 1 and 2 of a scenario with global catastrophic infrastructure loss in comparison to crop yield under current conditions. The reductions are reported as unitless relative change, e.g. -0.37 is equivalent to a 37% reduction. No confidence interval is presented as the reduction calculations lack information about their uncertainty.

Continent_statistics sheet

For each crop the Continent_statistics sheet reports the weighted mean of the (predicted) crop yield (in kg) and the projected yield reduction at different points in time (as described above) across the different continents.

References

Fox, J., & Weisberg, S. (2011). An R Companion to Applied Regression (2nd edition). SAGE Publications, Inc.

Huang, Y., Lan, Y., Thomson, S. J., Fang, A., Hoffmann, W. C., & Lacey, R. E. (2010). Development of soft computing and applications in agricultural and biological engineering. Computers and Electronics in Agriculture, 71(2), 107–127. https://doi.org/10.1016/j.compag.2010.01.001

McFadden, D. (1977). Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. Cowles Foundation Discussion Papers. (474). https://ideas.repec.org/p/cwl/cwldpp/474.html

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). Applied regression analysis: A research tool. New York : Springer. http://archive.org/details/appliedregressio00rawl_492