ABSOLUT v1.2 – Directions for use

Some information on the five R programs and their application

Tobias Conradt

Contents

1 What is ABSOLUT v1.2 ?

The five R programs which make up ABSOLUT contain a staged modelling algorithm for crop yield prediction. It is exclusively applicable for agricultural landscapes with distributed weather and yield observations, e.g. the administrative subdivisions of a country for which respective statistical and meteorological data are available.

At the core of ABSOLUT are multiple linear regressions, one per district, estimating the annual yields of a given crop from a linear time trend and up to four multi-month aggregates of meteorological variables from before the individual harvests. In contrast to other regression approaches, the combinations of weather aggregates are not predefined but exhaustively searched and selected by their out-of-sample prediciton performance. Since version 1.1 the out-of-sample condition is also observed for the regression composition itself, i. e. for each target year whose crop yields should be predicted, only data from other years are used to both determine the most suitable predictor combinations and estimating the regression parameters before they are applied for the prediction.

1.1 The name of the game

The acronym ABSOLUT stands for “Assessing Best-predictive Sets fOr multiple Linear regressions throUgh exhaustive Testing” – the idea behind has been described above. It also hints to the *absolute* value yield predictions, in contrast to setups using relative yield differences the author had worked with before (Conradt et al. 2016).

1.2 Version history and publications

This document is about version 1.2, released in December 2021. Improvements over the preceding v1.1 are the exclusion of overlapping weather features of the same meteorological variable and the inclusion of regressions with less than four weather features, down to the pure time trend.

Version 1.1, released in October 2021, removed a major inconsistency regarding the out-of-sample approach (see the paragraph on v1.0 below).

Version 1.0 was available since January 2021. It is extensively described in a discussion paper (Conradt 2021), but the impressive hindcast performances presented there for an application to the districts of Germany dramatically overestimate the expectable prediction accuracy due to partially violated out-of-sample conditions: While regression testing was correctly made with separate training and testing data (the latter being single years, “leave-one-out”), the input feature combinations of the district regressions were selected based on the entire data set. (Actual yield predictions in operational applications or scenario modelling results are hardly affected by this error, though.)

As one of the reviewers of Conradt (2021) pointed out the technical flaw in v1.0, a revision of the manuscript was made based on v1.1, but this was not accepted for final publication. The editor’s justification was that ABSOLUT could not be considered as substantial advancement for geoscientific model development, and there would be limited scientific significance for the community. Dear reader, thank you for dissenting from this opinion by your actual interest, otherwise you would not read this! This release of version 1.2 is made in conjunction with a new manuscript submission to Biogeosciences entitled “Choosing multiple linear regressions for weather-based crop yield prediction with ABSOLUT v1.2 – Initial tests for the districts of Germany and an over-confidence trap in statistical modelling”. If it is not outright rejected it should become another discussion paper in the first instance, so keep your eyes open.

1.3 Code and data files

The program code consists of the five R scripts (programs 1–5) deposited along this document. Their three-digit numbers consist of one digit for the program number and two digits indicating the version:

```
112_absolut.R
212_absolut.R
312_absolut.R
412_absolut.R
512_absolut.R
```

The required input files (some of which are specific to the Germany test case) can be found at `https://zenodo.org/record/5625774`:

`absolutcontrol.dat` Text file (UTF-8) with case-specific settings for program execution, to be edited by the user before the R scripts are run. The settings include directory paths, the principal weather variables to be considered (usually temperature, precipitation, and sunshine), time limits for optionally excluding recent years from the input, the last month of weather data to be considered before harvest, the minimum requirement of years of historic yield data in a district, and a restrictiveness parameter for the regression selection process of program 3.

`yield-indat.csv` Table of crop yields in dt ha^{-1} for different crops in German administrative areas, annual values for the years 1999–2020. Each line in this file represents a certain district–year combination and the associated yield observations for different crops. City districts without agriculture must not be included, and the program can also deal with individual data gaps indicated by NA strings.

`crop-areas.csv` Table of crop areas in hectares for different crops in German administrative areas in the year 2016. The structure is similar to the yield data except that a subdiffentiation by years is (still) missing.

`DistrictWeather/` Directory containing district-wise files `dw_01001.dat`, `dw_01002.dat,...,dw_16077.dat` with monthly weather variables. Each line represents a month in chronological order, and the columns indicate years, months, and the weather variables like temperature, precipitation, etc.

`DistrictFeatures/` Directory initially empty.

`ClimateScenarios/` Directory containing subdirectories with example climate scenario realisations in the same format as `DistrictWeather/`.

`ClimateScenarioFeatures/` Directory initially empty.

2 Hard- and software requirements

The R software (`https://www.r-project.org/`) is required in version 3.5.1 or newer, and the following packages must be installed to run the code as provided (version numbers indicate the versions used by the author):

leaps Exhaustive search for regression subset selection (3.0) and

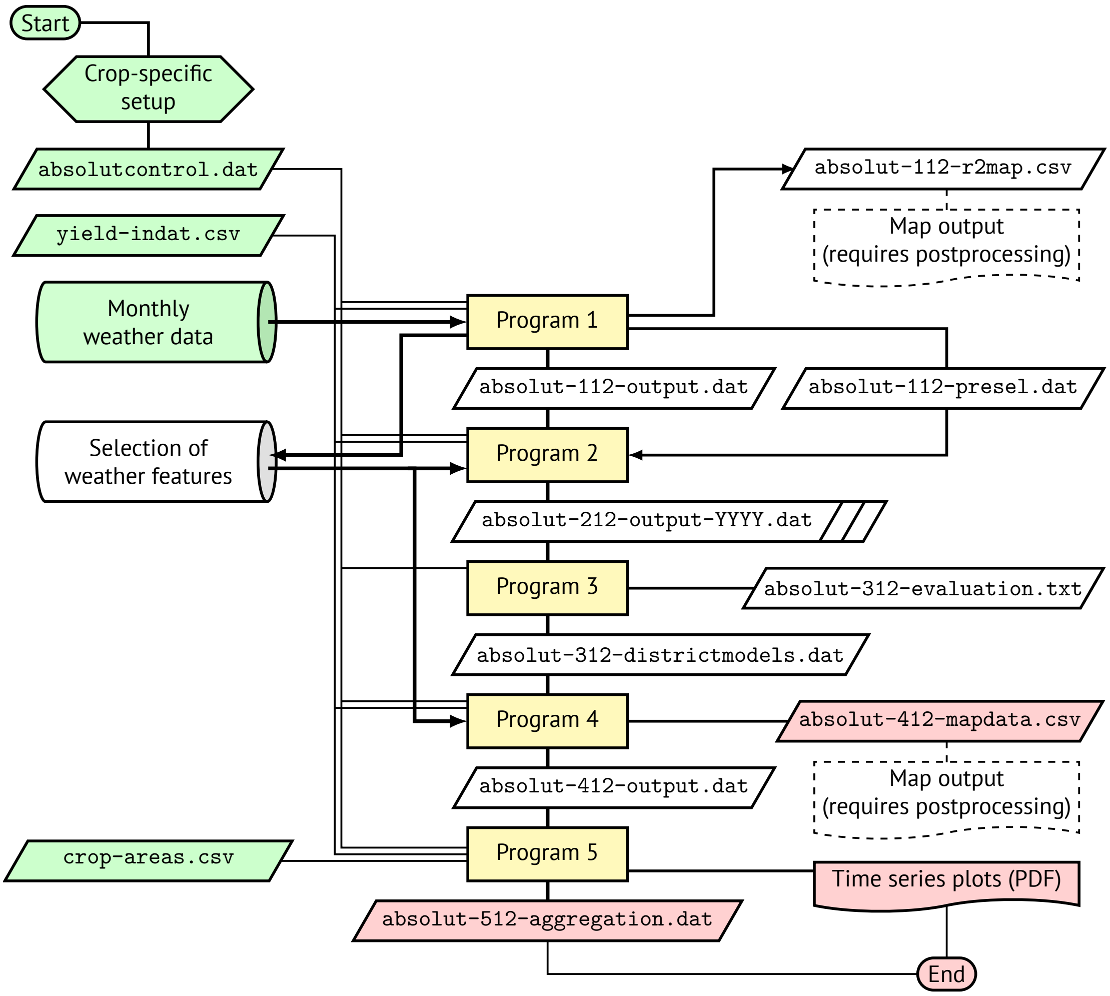
either *doMPI* Parallel computing via MPI (0.2.2) which also loads *foreach* (1.4.4)

or *foreach* on its own to run the `doParallel` command in case the hardware does not support MPI.

A potentially useful package for preprocessing weather data is *ncdf4* (1.16.1) which can handle NetCDF grids. It however requires a NetCDF system library.

Further program recommendations for preprocessing and output visualization are the open source programs GDAL (3.0.4), GRASS (7.8.3), and GMT (6.0.0).

3 The workflow



The above graphic illustrates the workflow and shows also the intermediate and precheck files (white) and the final outputs (red-tinted). The input data are tinted green.

The programs 1–5 (code files `112-absolut.R` to `512-absolut.R`) have to be run in sequential order; programs 1 and 2 are only in themselves prepared for parallel processing on a multi-CPU platform. Program 5 is optional, it aggregates the district-wise crop yield predictions and produces time series plots. All code is written under and for Linux systems – Windows users will have to edit file paths (backslashes), system calls, and probably more.

Program 1 should be run on as many CPU cores as there are (target) years covered by the input data, because for each of these targets all possible input feature combinations are tested in parallel for their regression fits over the remaining years. (Remember: the final goal is to predict the target year from information of the remaining years only.) Based on the frequency of weather aggregate features used in the best-performing combinations, features are pre-selected for each year and stored in `absolut-112-presel.dat`. To enable regression testing, the weather features have to be aggregated beforehand, and they are stored in district files filling the `DistrictFeatures/` directory. At the very end of program 1, the columns in these district files are reduced to the preselected features present in `absolut-112-presel.dat`. The run time of program 1 for the Germany example with two decades of data in 326 districts was about 15 minutes using 24 CPU cores.

Program 2 should be provided a number of CPU cores which either slightly exceeds the number of districts with sufficient crop yield data – the minimum requirement per district can be set in `absolutcontrol.dat` – or a basic fraction (half, third, etc.) of that number. If, for instance, there are computing nodes of 16 cores each, using seven of them would be a sound decision for the Germany example, because $7 \cdot 16 = 112$ is a bit more than a third of 326. This was actually the setup used by the author, and it took several hours for program 2 to complete. The computational demand is closely related to the number of weather features selected by target year. For winter wheat, their numbers ranged between 29 and 36 which theoretically means up to 66 712 different combinations. These are however stripped from overlaps of the same meteorological variable (e.g. January–March and February–May precipitation) which leaves only 4050–15567 allowed combinations for the example years. For each target year, program 2 produces a big table (e. g. `absolut-212-output-2003.dat`) with Pearson correlations of the out-of-sample regression predictions compared to observed yields, for each district (columns) and features (lines) combination.

Program 3 selects, based on these tables and separately for each target year, the regressions finally to be used for district yield prediction. This information is collected in `absolut-312-districtmodels.dat`. The other file produced, `absolut-312-evaluation.txt`, is a human-readable ASCII document, originally it had been a diagnostic screen output. The program run time in the Germany example was about 15 minutes on a single CPU.

Program 4 calculates crop yield predictions for each target year using the input features selected. These usually change in most districts between target years. The test run took about 45 seconds to complete. Program 4 includes the capability for scenario modelling: to activate, you need to edit the code file `412_absolut.R`. Line 32 must be changed to `scenariorun <- TRUE`, and in line 33 a base trend selection can be made.

Program 5 is not as generally applicable to other modelling domains as the other programs. It is strongly adapted to the German case using a two-stage aggregation from districts via federal states to the national results, with hard-coded two-digit keys of federal states. There are other customizations galore, and users are advised to modify and carefully check this program according to their individual needs.

References

Conradt, T. (2021): The multiple linear regression modelling algorithm ABSOLUT v1.0 for weather-based crop yield prediction and its application to Germany at district level. *Geoscientific Model Development Discussions* 2021: 21 (34 pp). doi:10.5194/gmd-2021-21.

Conradt, T., C. Gornott & F. Wechsung (2016): Extending and improving regionalized winter wheat and silage maize yield regression models for Germany: Enhancing the predictive skill by panel definition through cluster analysis. *Agricultural and Forest Meteorology* 216: 68–81. doi:10.1016/j.agrformet.2015.10.003.