Basic information

1. Target group: Algeria, barley (23 provinces), soft wheat (20 provinces), hard wheat (24 provinces)

2. Timing: 2002 - 2018, sown in October - November, harvested in May - July

3. Source: ASAP, DSASI-MADR

4. Factors.

- Remote sensing data (including: NDVI (avg, max))

- Meteorological data (including: Rad(sum), Rain(sum), T(avg, min, max))

- Categorical variables: taking into account soil, management practices, other factors

Overall results

1. Overall yield: 0.5 - 2.5 t/ha

2. Overall climate: desert in the south, Mediterranean in the north, dry in the west, slightly better in the east

3. Overall performance of ML and comparison

- > NDVI: but need to choose best parameters

- SVR most used, followed by LASSO and MLP, with RF and GBR performing less well

- Similar results to null model in early months (small advantage), February/March start to improve, with errors in April at 28-34% rRMSEp (provincial) and 17-21% (national), reaching minimum in June (barley, hard wheat) and July (soft wheat); little improvement in accuracy after May

4. Comparable error for barley and soft wheat, lower error for durum wheat, little improvement for hard wheat

- Small planted area (smaller spatial scale)

- Low interannual variability (better accuracy of null models)

5. NDVI is usually more important than meteorological factors

- Similar performance of Met/Met- and RS/RS- in early season

- North: abundant precipitation, fertile soils -> meteorological conditions have little influence

- Only NVDI is usually retained for durum wheat (after April), both for barley and soft wheat

6. National level data is usually more accurately calculated than provincial level

- May barley, soft wheat and hard wheat, provincial R2p: 0.47, 0.50, 0.66, provincial time R2p: 0.46, 0.42, 0.34; national R2p: 0.81, 0.74, 0.61 (compared to NDVI: 0.52, 0.51, 0.58)

7. Low yielding years: ML's performance is clearly better than other models

- Average 6% increase in RMSEp for ML (compared to NDVI: 24%, null model: 33%)

- ML accuracy for barley and soft wheat: twice that of NDVI (with half the error), hard wheat: no significant optimisation or slightly reduced performance

8. OHEau use: median reduction in rRMSEp of 13% to 1%, but decreasing with time

9. Feature selection improves accuracy in 63.5% of cases: gains for LASSO, SVRlin, MLP are positively correlated with increasing season, gains for RF, GBR are negatively correlated with increasing season, gains for SVRrbf are unchanged; gains are greatest for RS&Met and least for RS and RS-

- tree can extract features more efficiently and/or have higher robustness than MRMR

10. RS&Met- works well for May-July -> weather affects flowering -> affects grain weight (widens or narrows grain filling period, mild -> lengthens, drought -> speeds up)

- Weather in April affects flower fertility -> affects grain weight

Model analysis

1. Use: Machine learning algorithms (ML)

- Model selection, feature engineering, hyperparameter selection and optimisation, testing

- Least Absolute Shrinkage and Selection Operator (LASSO); Random Forest (RF); Multilayer Perceptron (MLP); Support Vector Regression with linear/radial basis function kernels (SVRlin and SVRrbf); Gradient Boosting (GBR)

2. Cross-validation averaging - provincial data, averaging by province by year - temporal data, province weighted - national data

3. Difficulty: small data (but 97% rainfed, mostly useful); caution: data leakage

4. Data: retain sorted cumulative 90% of data (exclude marginal)

- Start of season (November, > 20% NDVI) - end of season (June, < 20% NDVI)

- Yield calculation: early December - early July, m-months using 12 - m-1 months data

5. Initialisation: no de-trending required (essentially no trend)

6. Data sets: training set, validation set (optimise parameters), test set

- Cross-validation: one external test year, corresponding to n calculations (n combinations of hyperparameters); for each calculation: 1 internal validation year

7. Feature selection: removal of irrelevant features, which in most cases will improve accuracy

- Set of manually defined features: RS&MET, RS, MET, RS&MET-, RS-, MET- (Table 3 - P4), the last three being the most important features of the first three (using the minimum redundancy and maximum relation method (MRMR))

- Percentage of input for each feature: 5, 10, 25, 50, 75, 100%

- Due to data dimensionality and crosstalk, the opposite situation may appear

8. Categorical variables (OHEau): are coded once and usually help to improve accuracy

9. Comparison with two benchmark models: comparison with null model (direct use of previous mean), NDVI peak model (yield = a \* max(NDVI) + b)

- The baseline model is calculated by cross-validation of a single layer

- Forecast accuracy statistics: root mean square error (RMSEp), relative root mean square error percentage (rRMSEp), mean error (MEp), coefficient of determination of model and observed yields (R2p)

- R2p for provincial data (spatially significant), temporal R2p by averaging data over time across provinces, and national R2p by weighted average across provinces

10. Bayesian hypothesis testing, comparing ML with other algorithms (in particular: Bayesian correlation t-test): better performance than significance tests for null hypotheses

- Solves the problem of small data (solves the problem of differences above the p-value)

- Presence of a region of practical equivalence (ROPE): range of equivalence parameters (for current purposes, subjective parameters)

- The null hypothesis δ (i.e. the region of practical equivalence) can be adjusted to find the best hypothesis

- Compare regions of P(algo1 <=> algo2) to draw conclusions: more, better/bad, uncertain (Fig. 3, P6)

11. Select models by comparing rRMSEp for different algorithms and different feature sets (Bayesian detection/direct comparison)

12. Focus on comparing low yield years (data from the first 1/4 yield years): critical for food security

13. Knowledge-driven feature selection slightly better than fully data-driven feature selection, but a combination would be better

Cautions

1. ML is usually better than other models under optimal parameters, but for LASSO, MLP and SVR (rbf, lin), the results are usually (I), i.e. > 5% of the region does not exceed the confidence level (90%), which does not ensure that it is necessarily better

2. ML with non-optimal parameters is often weaker than other models

3. ML works better when provincial information is more accurate; better at national level than at provincial level

4. Statistics in this paper do not take into account variation in cereal acreage -> excessive variation will lead to inaccuracy

Summary

1. The influence of different factors varies from one period to another and the main factors may be different (before-after effects, changing circumstances)

2. Same product may have different factors, or different proportions, in different contexts

3. Attention to cleaning of data

- Heads and tails may need to be removed

- Select the most appropriate feature, the most appropriate part of the feature to analyse

- Other factors may be condensed into a categorical variable

4. More complex models, more parameters, more accurate parameter ranges -> does not necessarily lead to optimisation, requires attempts

- Cumulative error of the model, extraction ability/robustness of the model itself

- Correlation of parameters, dimensionality of data, etc.

5. Don't perform operations that don't need to be performed (don't de-trend if there is no obvious trend)

6. Beware of data leakage (empty a column, cross-validate, or multi-layer cross-validate)

7. Use stronger methods such as Bayesian hypothesis testing for judging

8. Focus on some groups of data from a practical point of view (low yield -> food security -> need to be more accurate)

Chart, scatter chart

Description automatically generated

Table 3

Variables considered in the manually engineered feature sets.

Chart

Description automatically generated

Fig. 3.

Bayesian hypothesis test and diagnosis to compare two algorithms. The plain white lines delineate the regions of practical significance (at the 90% confidence level). Test outcomes falling outside of these regions are inconclusive.