Early-Season Prediction of Wheat Yields in the North China Plain with Boosted Regression Trees

Stien Heremans¹, Qinghan Dong² and Jos Van Orshoven¹

¹: Department of Earth and Environmental Sciences, KU Leuven (Leuven University), Belgium
²: VITO, Mol, Belgium

Introduction

Wheat yield estimates should be both early and accurate to be useful in the context of food security.

Machine learning methods such as boosted regression trees (BoRTs) have proven to be accurate for predicting trends of biophysical parameters, but have not yet been applied for crop yield prediction from time series of remotely sensed images. Therefore, exploring the potential of BoRTs in this specific context is the purpose of this study.

Research Questions

• How does the accuracy of wheat yield prediction models evolve with growing time series?
• From which point in the growth season is district-level yield prediction from NDVI time series accurate enough?
• Which periods in the growth season are most relevant for wheat yield prediction in the North China Plain?

Methods

Study area: 17 districts situated in the Henan province in the North China Plain. Per district, the statistics for wheat yield and chemical fertilizer input (CFI) were collected for the last 15 years (2000-2014).

Boosted regression trees were trained for each district to predict wheat yield from NDVI time series progressively containing more dekads throughout the growth season. The inputs are time series of 10-day (dekadal) VGT NDVI values over the entire wheat growth season (September – June) + the annual CFI. Input selection was integrated into the modeling process: for each input series, up to ten features were iteratively selected. The course of the cross-validated $R^2$ ($R^2$ cv) throughout the growth season was evaluated.

Results

1. District typology based on $R^2$ evolution throughout the season

Districts can be assigned to five types based on the course of their $R^2$ cv (Fig. 2): high $R^2$ throughout season, rel. high start with small improvement, intern. start with large improvement, low $R^2$ throughout season, low start with large improvement.

2. Geographical distribution

Fig. 3 shows the geographical distribution of the five types in the Henan province.
- The districts with a low start-of-season accuracy (red and blue) are situated in the more remote parts of the study area.
- The central districts show the typical pattern of improving accuracy throughout the season.

3. Most relevant dekads

Fig. 4 displays the relevance of the dekads:
- Three periods stand out with higher relevance than the rest of the season: Sep-Oct, Jan and Apr-May.
- The last dekad has an exceptionally high relevance = unexpected!

Conclusions

• $R^2$ cv reaches (near-)maximum for most districts between Sep. and Jan. = early- to mid- season
• Good prediction is possible before January for almost every district (except Puyang)
• Relevant dekads occur throughout the entire season, even before sowing and after harvest and during winter when the area is often covered in snow (Fig. 4) = unexpected!
• Annual CFI may mask important dekads → analysis without CFI is subject for further research

Acknowledgements

We acknowledge the E-AGRI project, conducted in the EU’s Seventh Framework Programme for Research (FP7) for providing the input data.