ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

Statistical modeling of phenological phases in Poland based on coupling satellite derived products and gridded meteorological data

Bartosz Czernecki [1]. Jakub Nowosad [1], Katarzyna Jabłońska [2]

ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

[1] Faculty of Geosciences
Adam Mickiewicz University in Poznań, Poland
nwp@amu.edu.pl

[2] Institute of Meteorology and Water Management
- National Research Instiute, Warsaw, Poland

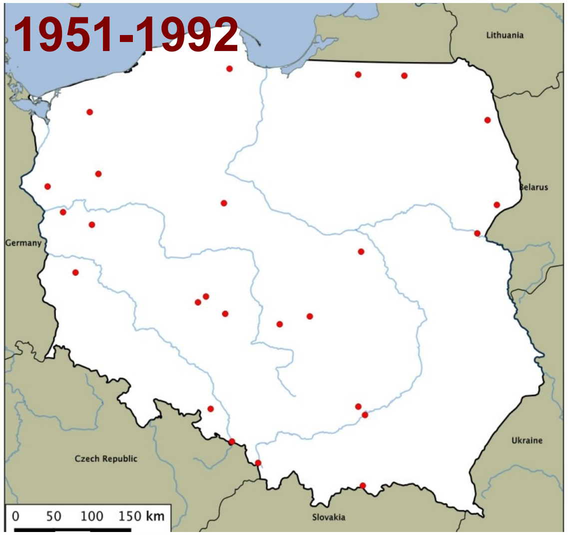
Phenological observation in Poland – after 1945

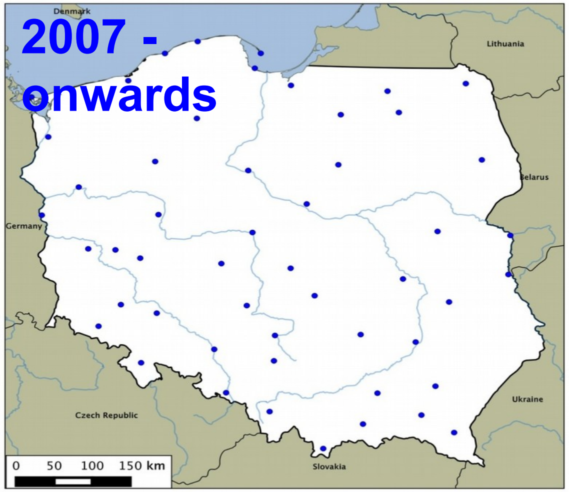
1951-1992 – network of phenological observations run by Polish Met Service (IMGW); strongly varying number of stations (over 700 in 1970s and below 100 in 1980s). Only ~30 stations with complete and reliable dataset.

1993-2004 – network of phenological observations completely abandoned

2005-2007 – reactivation of phenological observation, partly in old locations (~40 stations)

2007 – onwards – newly established network in location of meteorological stations (~60 stations). Measurements according to BBCH-scale

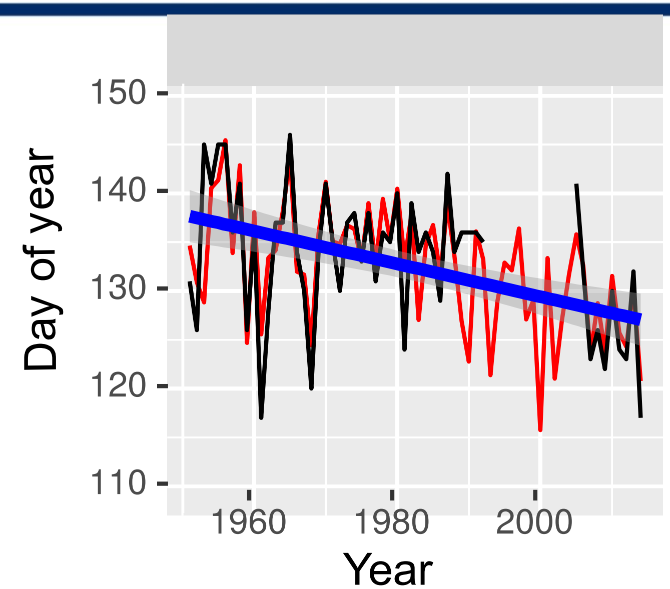


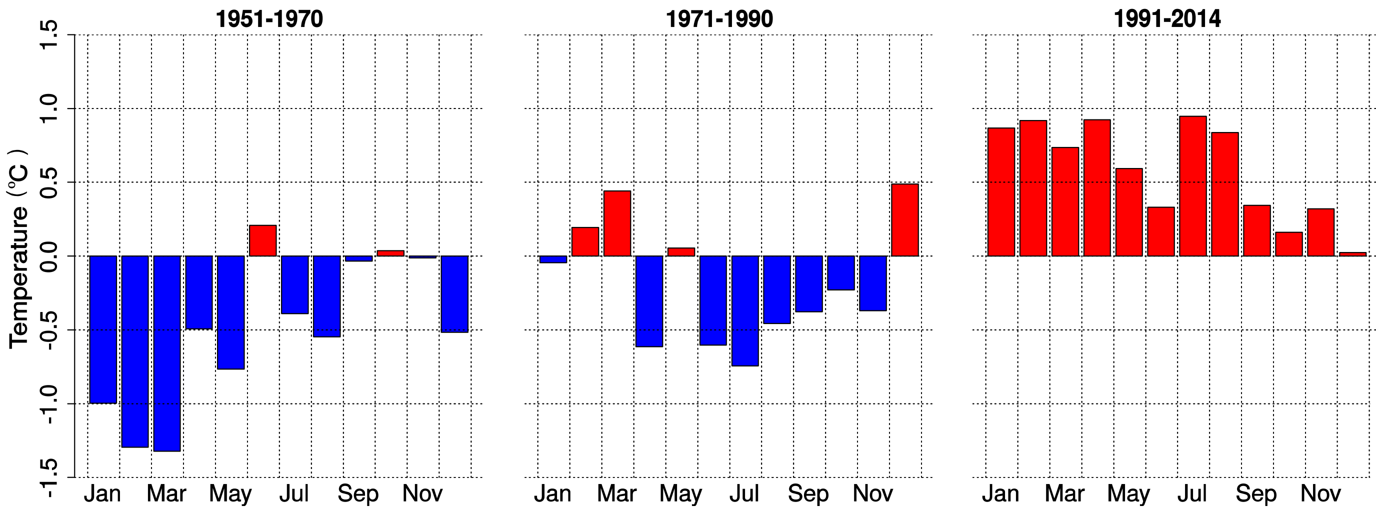


Why to use phenological data?

Global warming determine the advance of phenological events. Therefore, changes in timing of phenological phases are important proxies in contemporary climate research:

- climate change
- climate proxy
- dynamics of climate seasonality
- food production
- aerobiology (pollen)





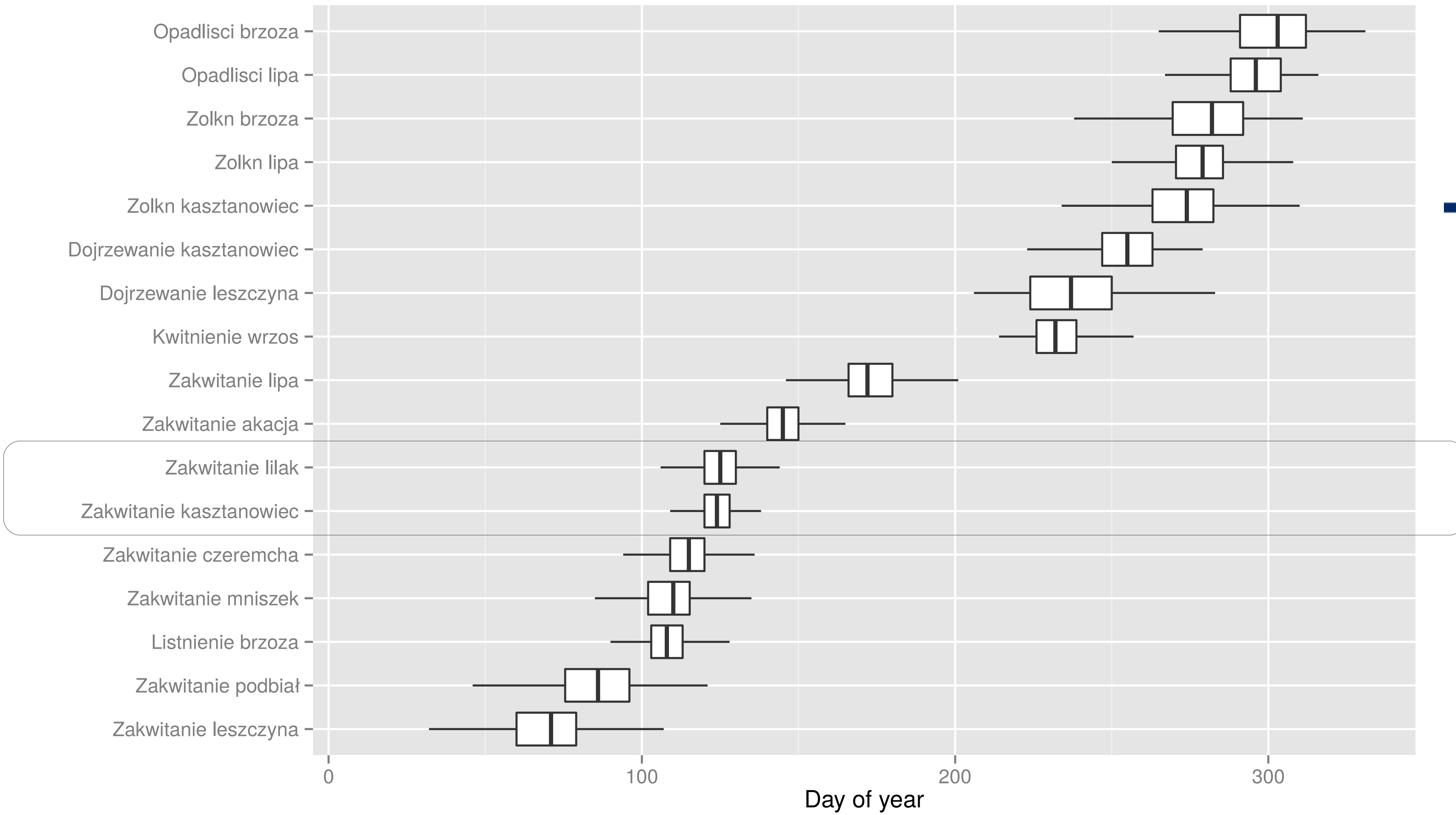
Czernecki and Jabłońska (2016) (in print)

Aim

- The main aim → create and evaluate different statistical models for **reconstructing** and **predicting** day of year of selected phenological phases occurrence using the most recent data → finding a robust predictors

- Evaluate possibilities of using only free of charge data remote sensing and meteorological data as predictors
  - (1) distinguish the amount of information provided by both sources of data
  - (2) define whether they are unrelated and contain possible sources of not overlapping information,
  - (3) and thus may (or may not) robustly contribute in phenological research, especially in terms of phenological modeling

- Tools → everything written in R programming languages and its packages to automatize entire procedure

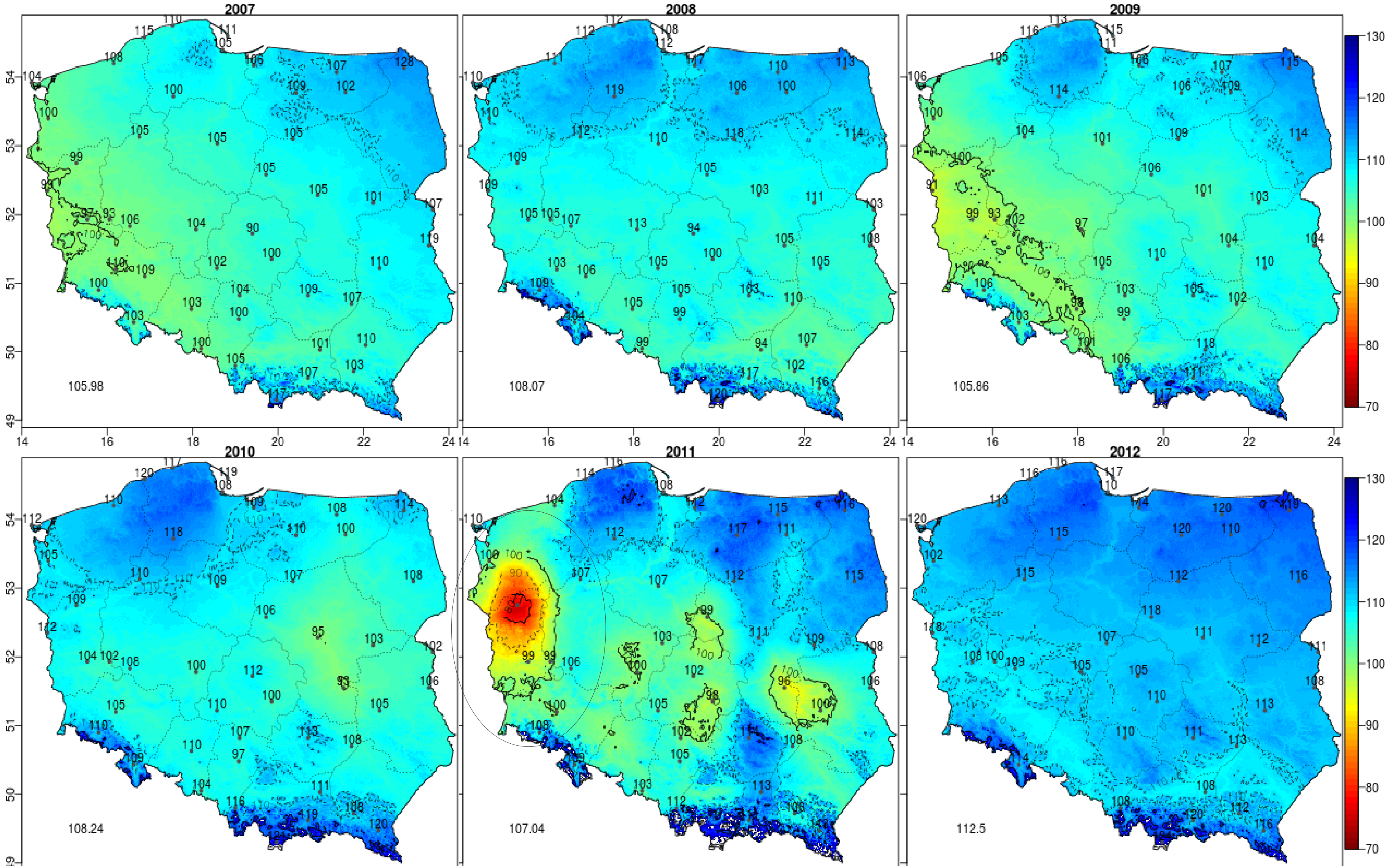


The study period covers years 2007-2014 and contains only quality-controlled dataset of *Syringa vulgaris* and *Aesculus hippocastanum* flowering dates (i.e. late spring phenophase) on 52 stations in Poland

Syringa vulgaris
By Ulf Eliasson - Own work, CC BY 2.5,
https://commons.wikimedia.org/w/index.php?curid=1387269

Aesculus hippocastanum
By H. Zell - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=8951128

Error detection
(data visualisation → expert decision)



Predictor variables and data preparation

Three types of data sources were used as predictors:

1. **Satellite derived products → MODIS level-3 vegetation products:**
   - NDVI, EVI, LAI, fPAR (fraction of photosynthetically active radiation)
   - Interactive Multisensor Snow and Ice Mapping System (IMS) products
   - Highly noisy data → pixel reliability information taken into account

2. **Preprocessed gridded meteorological data → ECA&D**
   - cumulative growing degree days (GDD), cumulative growing precipitation days (GPD), average monthly temperatures, monthly temperatures over the previous year

3. **Spatial features** (longitude, latitude, altitude, distance to Baltic Sea, etc.)

Development of statistical models

A few methods were tested and evaluated against the onset dates of phenophases:

- multiple linear regression with (**lmAIC**) and without stepwise selection (**lm**)
- generalized linear model with (**glmAIC**) and without stepwise selection (**glm**)
- random forest (**RF**)

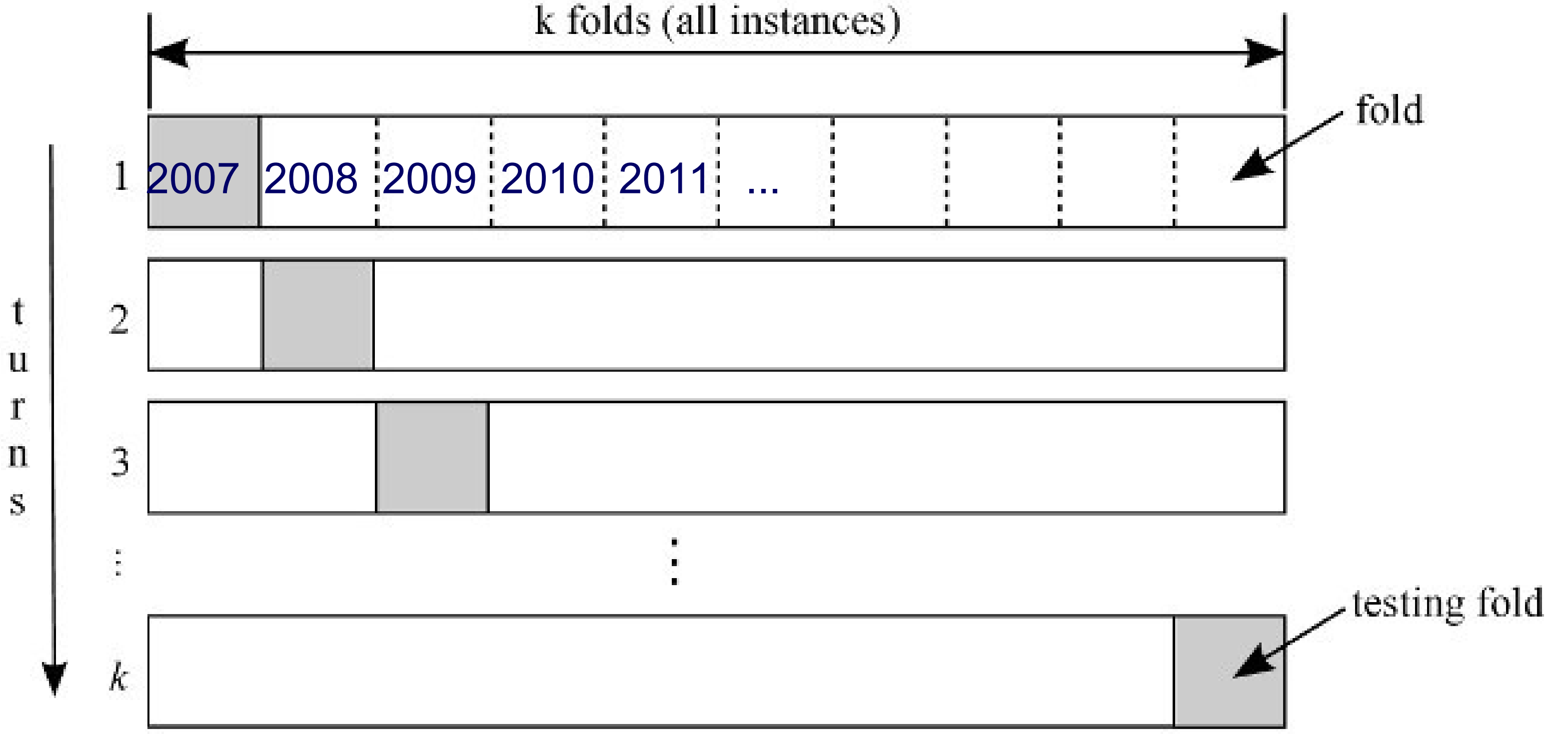
Potential predictors splited into four sub-groups (to estimate their importance):

1. Only meteorological-derived variables and locations' features
2. MODIS-derived predictors
3. All available variables pre-processed with the use of Boruta algorithm (that finds all-relevant features) (Kursa 2010)
4. All available variables without any pre-selection

Cross-validation

Repeated k-fold cross validation was used to avoid overfitting and to estimate the accuracy of the models

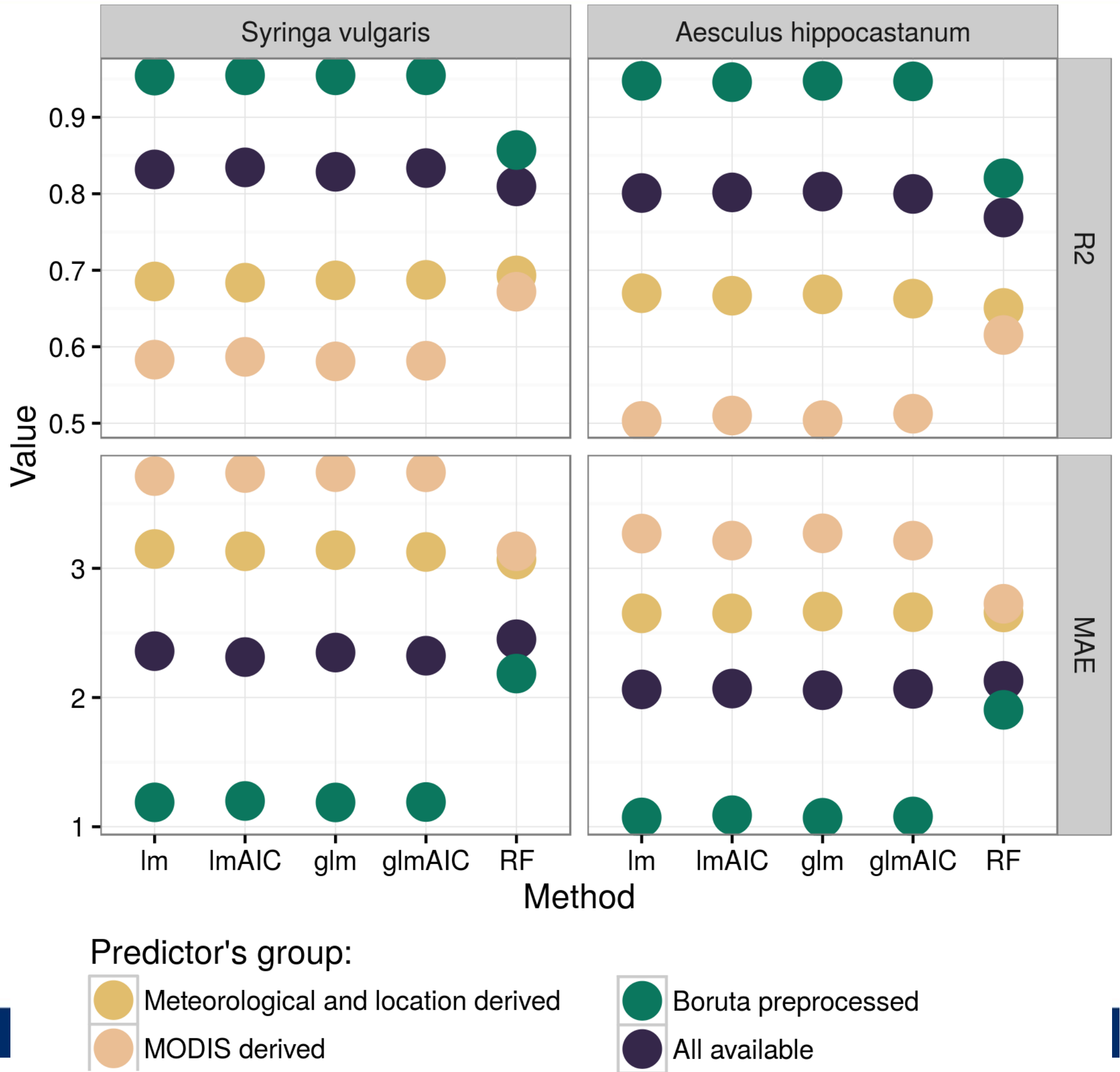
„Built-in" k-fold approach in caret package led to overfiting of models → spliting data on annual basis to avoid overfiting



Results

Substantial impact of predictors selection on final results

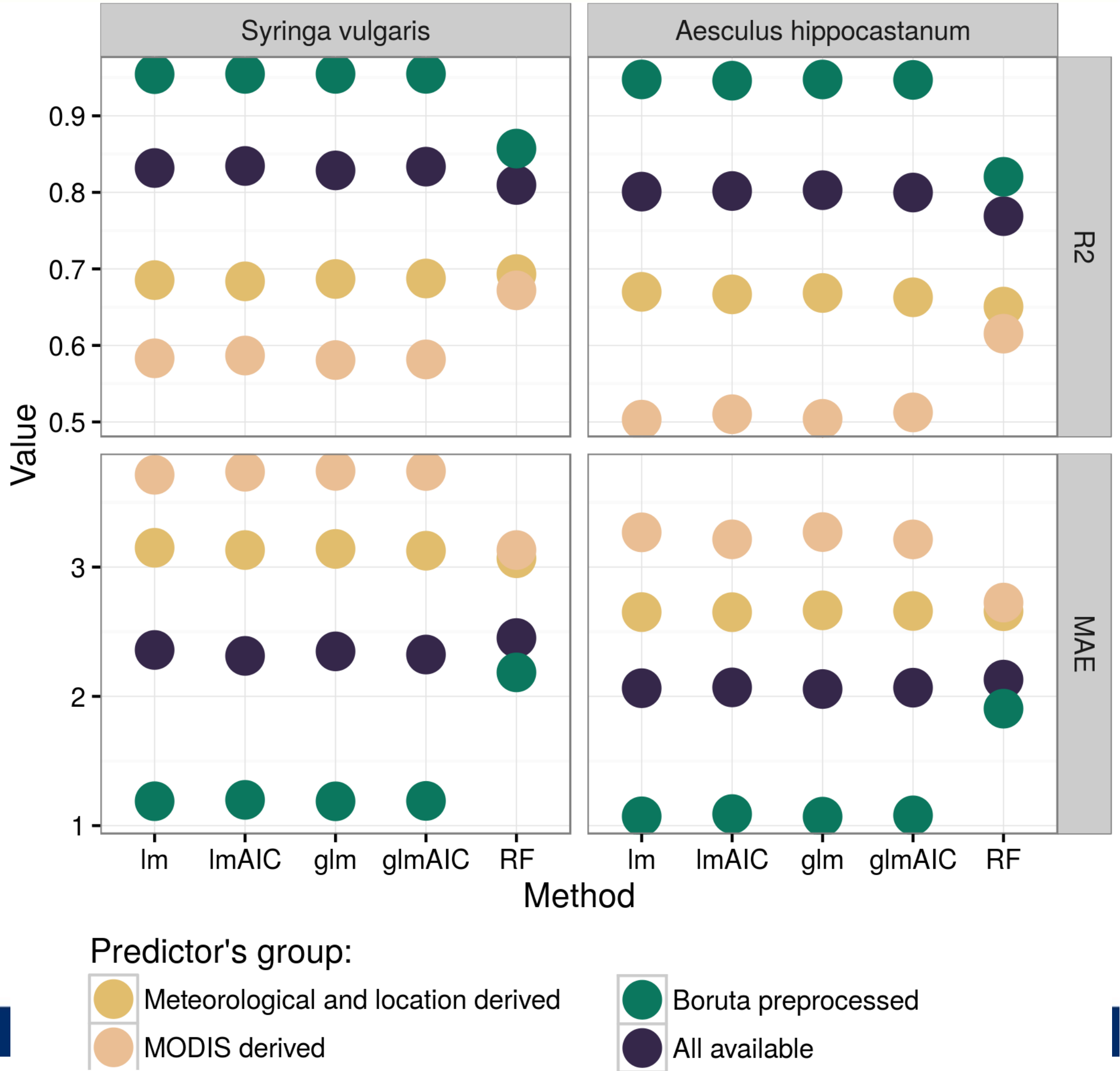
- The chosen set of predictors → very similar results in every of tested regression based algorithms
- The impact of selected predictors was **smaller on RF than on regression models**. However, the obtained pattern was similar in every of analyzed models, i.e. the best fitted models were preprocessed using Boruta algorithm.



Results

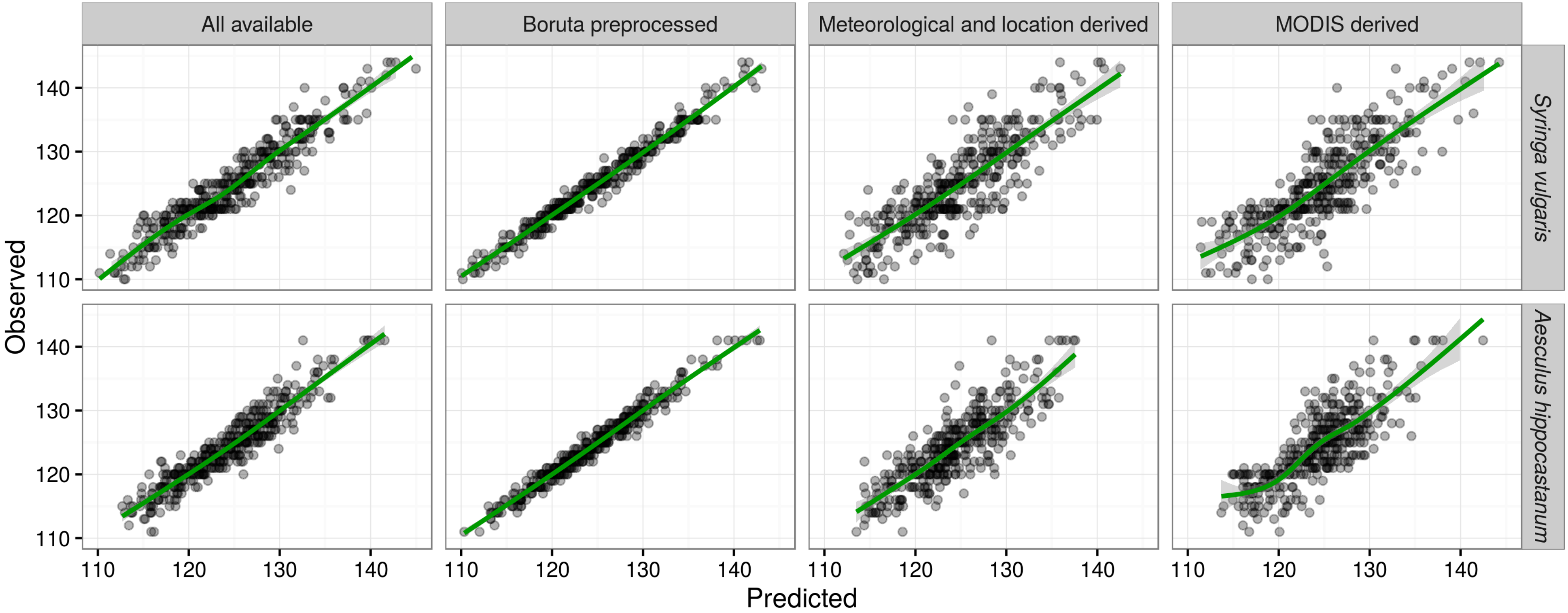
Substantial impact of predictors selection on final results:

- The AIC stepwise screening hardly influences the obtained results → do not redress computational time that is required while applying this procedure



Results

- Models based only on meteorological indices accounted for about 80% of variance in *Syringa vulgaris* and *Aesculus hippocastanum* flowering dates,
- applying remote sensing data and preprocessing by Boruta algorithm increased this value to over 95%



Conclusions

- The created models show high applicable potential
- The models based on meteorological characteristics were better fitted to observational time-series than remote sensing-based models
- Even though, conjunction of both data sources improve model's accuracy
- A strong improvement if preprocessing procedures (e.g. Boruta) were applied → numerous set of potential predictors

- Clear limitations of applying satellite observation in phenology modeling:
  - small contribution of satellite derived products to model's results
  - satellite data contain noisy information and thus were omitted while applying preprocessing procedures
- Therefore, most of the created phenology models are primarily based on climatological indices with only slight improvements of satellite products

Thank you for your attention