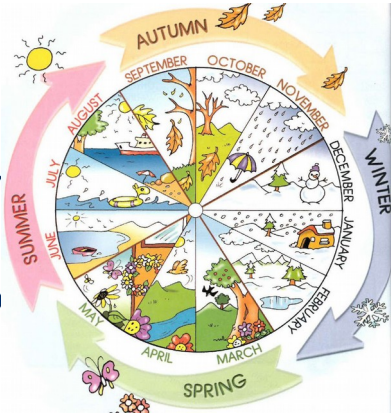
ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ



Machine learning modeling of phenological phases in Poland

(lightning talk)

Bartosz Czernecki [1]. Jakub Nowosad [1,2], Katarzyna Jabłońska [3]

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

[2] Space Informatics Lab, Department of Geography
University of Cincinnati, USA

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

What is phenology?

Phenology *is the study of periodic plant and animal life cycle events and how these are influenced by seasonal and interannual variations in climate*

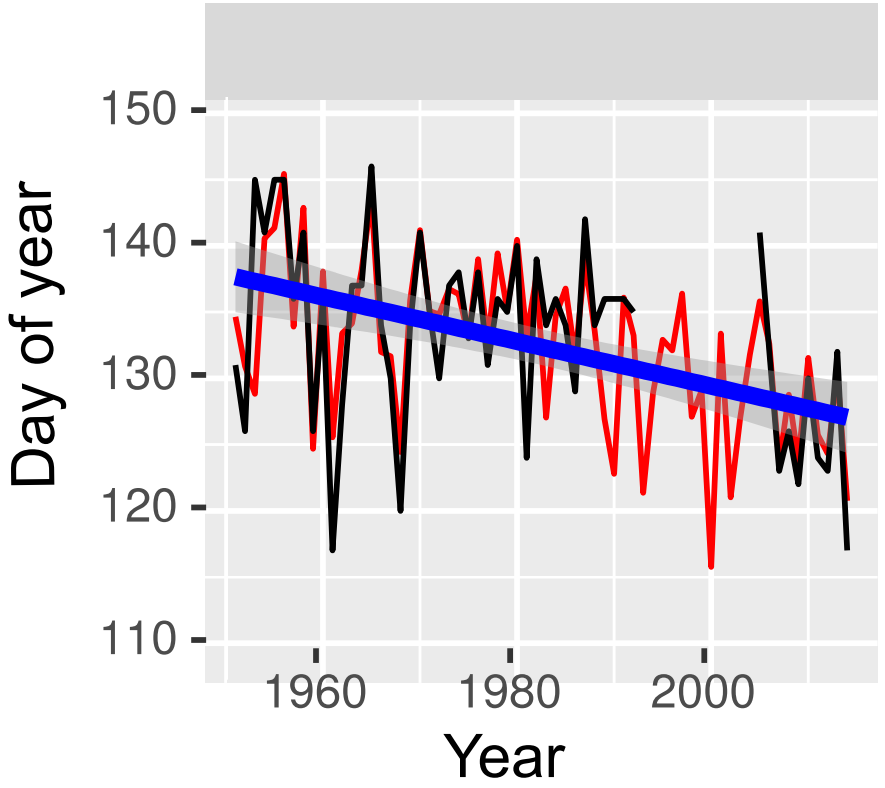
Early spring → Spring → Late spring → Summer → Autumn → Early Winter → Winter

Past records (human observations) → several main plant phenophases are consistently recorded over 150 years in Europe; development stages described usually by onset dates

Relevance of phenological data

Plant phenomena are very sensitive to small variations in climate. Therefore, changes in timing of phenological phases are important proxies in contemporary climate research:

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

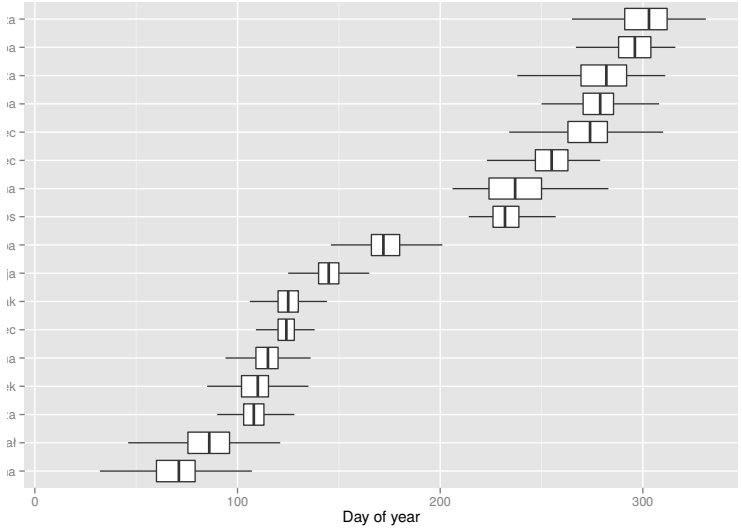


Changes in timing of Early spring in Poland

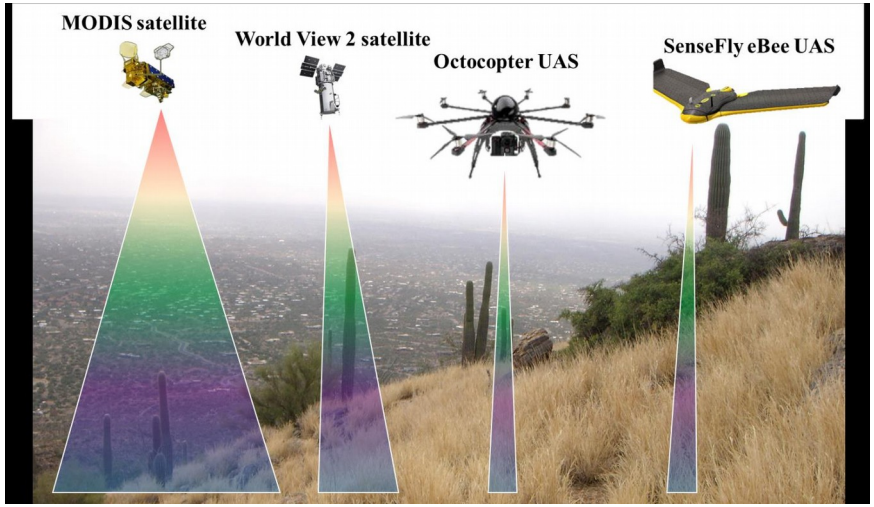
Czernecki and Jabłońska (2016)

Modern phenology

Human observations *– based on very specific (BBCH) regulations for monitoring network (since 2007 in Poland; over 60 locations)*



Remote sensing observations - observing the phenology of whole ecosystems using the vegatation's typical low reflection in red and strong in NIR → creating Vegetation indices like NDVI



Satellite data → complement traditional phenological methods → continuous in space and time information, non-user dependent; often noisy, only an approximation of the true biological growth stages

Aims

- The main aim → create and evaluate ML algorithms for **reconstructing** and **predicting** occurrence of selected phenological phases

- only free of charge satellite and meteorological data as predictors:
  - (1) distinguish the amount of information provided by both sources of data
  - (2) overlapping or not-overlapping ?
  - (3) may (or may not) robustly contribute to phenological modeling (predicting)

- **Tools → in order to automatize entire procedure everything written in R and its packages → (34!!! uniques)**

  github code: ~/github/polphen/R$ cat *.R |wc → 1393 4379 55196

Predictors and data preparation packages

1. **Satellite derived products → MODIS level-3 vegetation products:** ***mostly HDF and GIS files: (raster, sp, rgdal, modiscloud, maps, mapdata)***
   - Vegetation Indices: (NDVI, EVI, LAI, fraction of photosynthetically active radiation) → Modis HDF files not readably by GDAL-based packages!
   - Interactive Multisensor Snow and Ice Mapping System (IMS) products
   - Highly noisy data → pixel reliability information taken into account

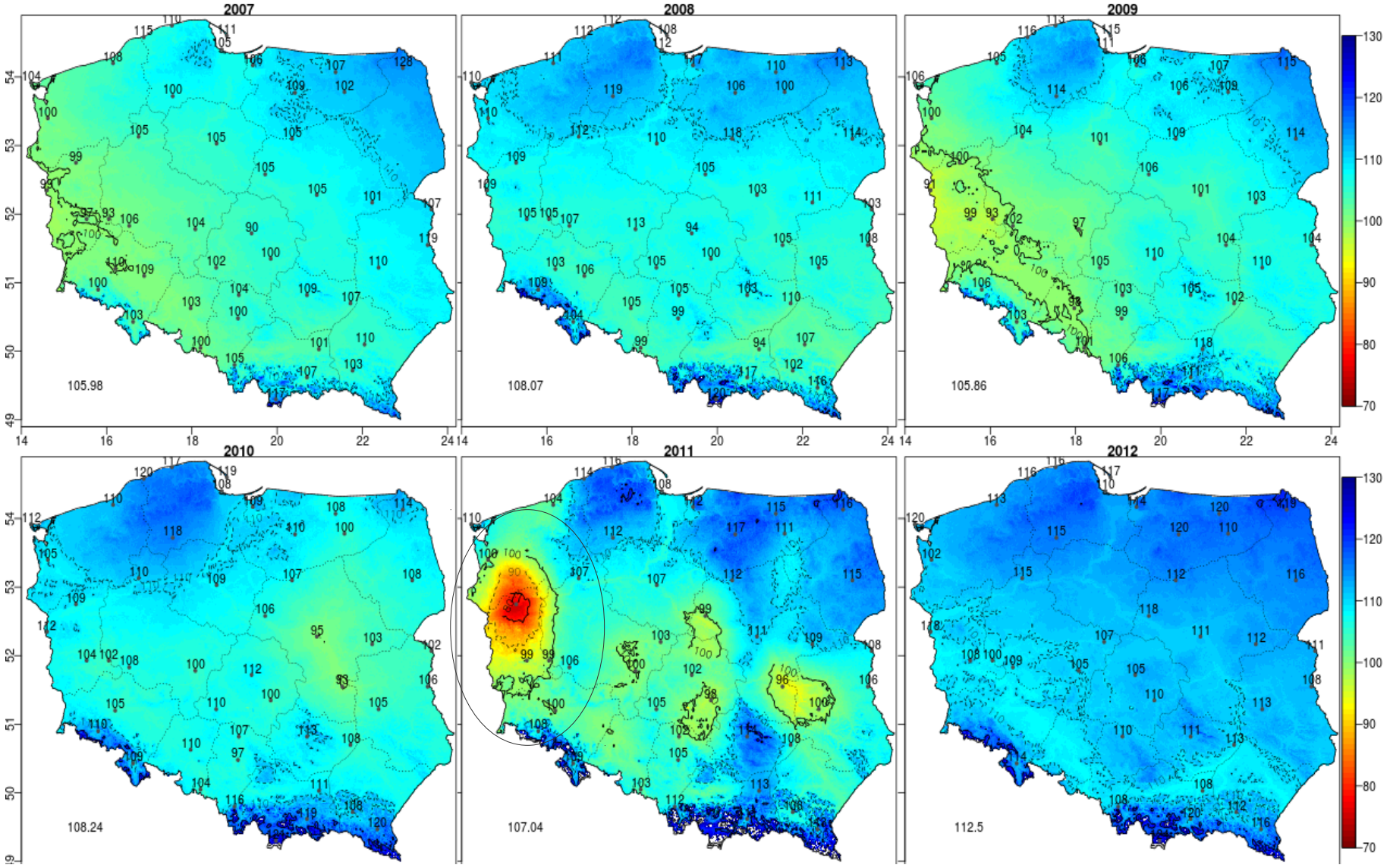
2. **Preprocessed gridded and in-situ meteorological data → ECA&D** ***mostly NetCDF4 files: (esd, ncdf4, polpred)***
   - cumulative growing degree days (GDD), cumulative growing precipitation days (GPD), average monthly temperatures, monthly temperatures over the previous year

3. **Spatial features** (longitude, latitude, altitude, distance to Baltic Sea, etc.) ***mostly GIS files: (gstat, rgdal, sp, maptools, raster, verification)***

Error detection

(data visualisation → expert decision)

gstat:: Regression kriging with external drift



Development of ML algorithms (caret package)

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**)
- Xgboost → !!! do not mix with parallelizing features included in library('doMC')) !!!

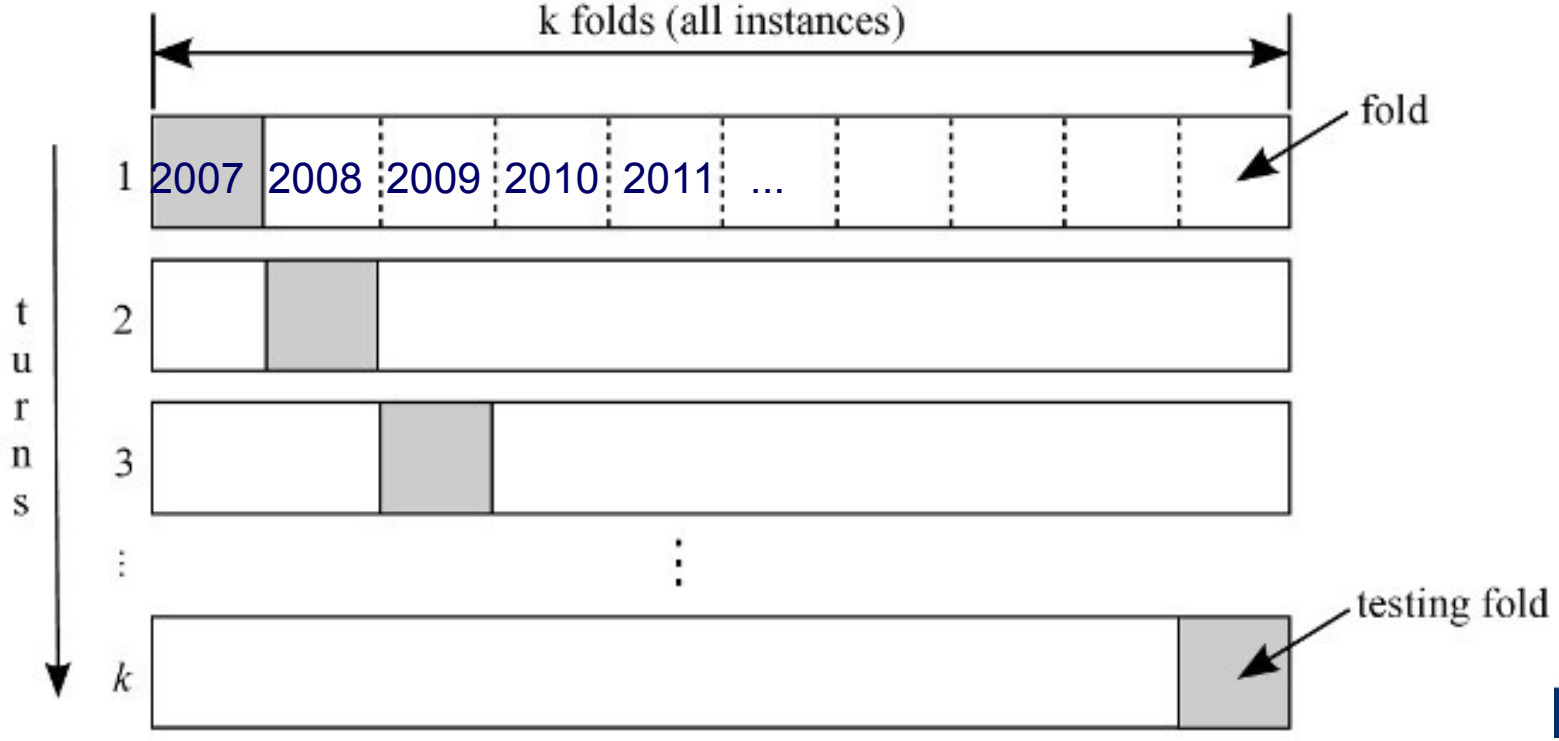
Potential predictors splited into sub-groups (to estimate their importance):

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

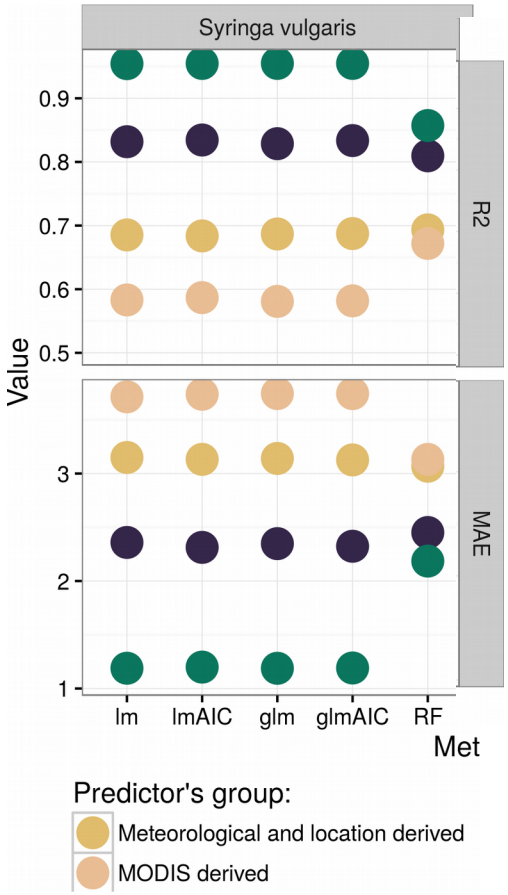
Cross-validation

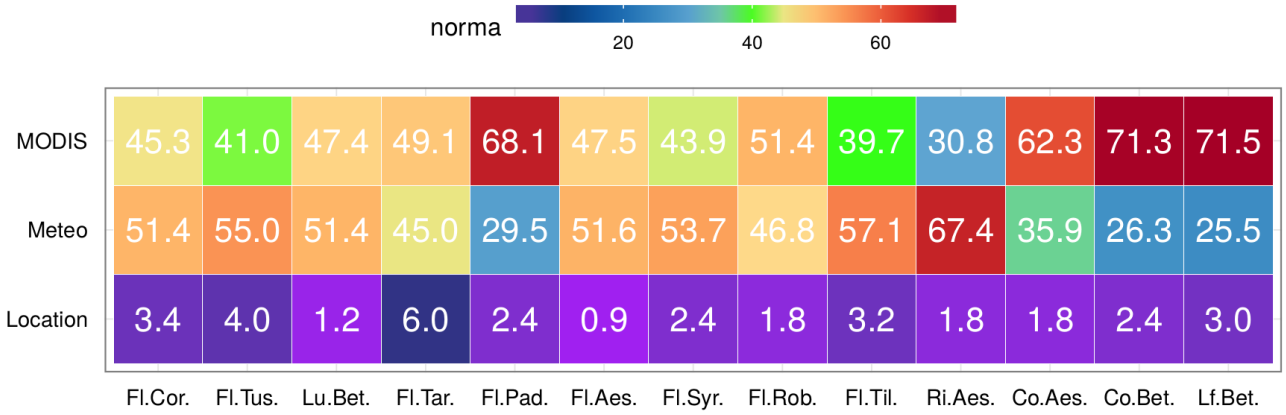
Repeated k-fold cross validation was used to avoid overfitting and to estimate the accuracy of the models

„Built-in” (default) k-fold or cv approaches in caret package led to overfiting of models → spliting data manually on annual basis to avoid overfiting



Results – variable importance (caret::varImp)





Early spring → Spring → Late Spring → Summer → Autumn → Early Winter

Conclusions

- ALMOST everything possible in R environment based on already existing packages (even if not efficient...)
- Careful CV procedures required to avoid overfitting when dealing with (climate) seasonality
- Many different strategies for parallelizing model building in caret package may lead to infinitive computations
  → keep in mind while mixing with own strategies

- Clear limitations of applying satellite observation in phenology modeling:
  - satellite data contain noisy information and thus are often omitted while applying preprocessing procedures
- Therefore, most of the created phenology models are primarily based on climatological indices with slight improvements of satellite products

Thank you for your attention

Computations were carried out in the Poznań Centre for Networking and Supercomputing (http://www.pcss.wroc.pl), Grant No. 295.