

ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

I Warsztaty Geofizyczne – Programowanie w środowisku R



Borucino, 24-29.06.2018

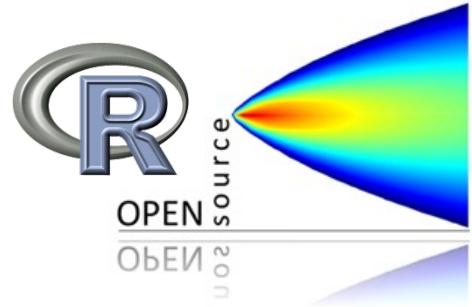
# Język programowania w naukach atmosferycznych

Bartosz Czernecki

Department of Climatology  
Faculty of Geographical and Geological Sciences  
Adam Mickiewicz University in Poznań, Poland  
[nwp@amu.edu.pl](mailto:nwp@amu.edu.pl) | [enwo.pl/przetwarzanie](http://enwo.pl/przetwarzanie)



[www.amu.edu.pl](http://www.amu.edu.pl)



prof. UWr dr hab.  
**Maciej Kryza**

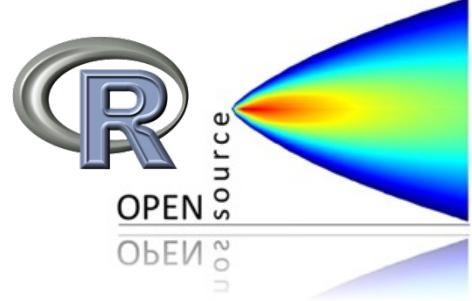
# Prowadzący warsztaty



dr **Michał Marosz**



dr **Bartosz Czernecki**



# Plan działania – prowadzenie zajęć



Poniedziałek



Wtorek

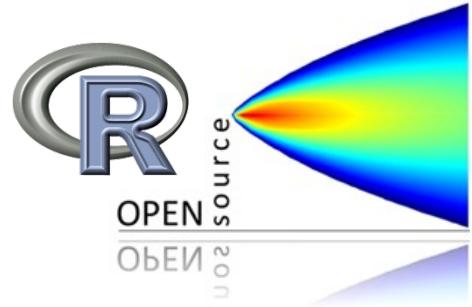


Środa

Czwartek

Piątek

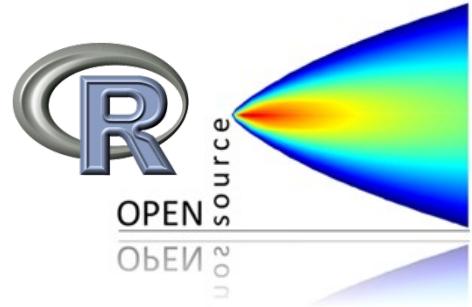




## Plan działania (2)

---

- Nacisk na **praktykę** (= mało teorii)
- brak „sztywnego” harmonogramu:  
poszczególne działy realizowane w zależności od postępów (umiejętności + problemy techniczne)  
→ założenie: każdy zdażył przerobić tutorial
- Nie boimy się dyskusji i pytań (także w sesji „wieczornej”)



# Plan prezentacji

---

## Kilka słów o

- Czym jest R? - o projekcie R kilka (na tle innych języków programowania)
- Dlaczego warto (na)uczyć się R?
- O czym warto pamiętać przy nauce programowania w R?

## R w naukach atmosferycznych:

- R jako *lingua franca* (*R* czy *Python*?)
- Przykładowe zastosowania:
  - Analiza danych
  - Statystyka
  - Wizualizacja



# Czym jest ?

Dynamicznie rozwijający się język programowania używany przede wszystkim jako narzędzie do:

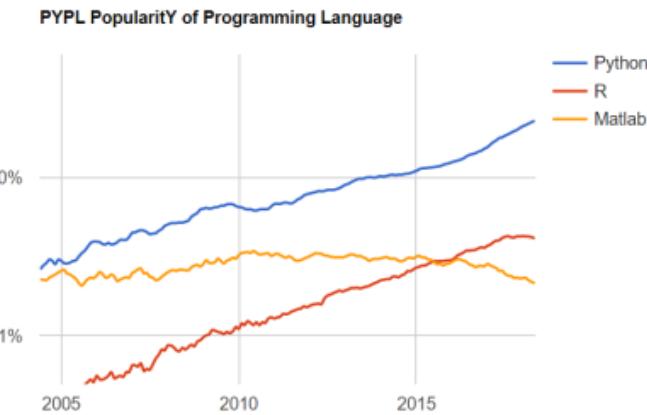
- analiz statystycznych
- modelowania
- przetwarzania
- wizualizacji danych

Początki – połowa lat 90. (R&R)  
→ środowisko akademickie  
(statystyczne)



Popularność R w ostatnich latach wyraźnie rośnie →

Worldwide, Jun 2018 compared to a year ago:				
Rank	Change	Language	Share	Trend
1	↑	Python	23.04 %	+5.2 %
2	↓	Java	22.45 %	-0.6 %
3	↑↑	Javascript	8.6 %	+0.3 %
4	↓	PHP	8.21 %	-1.6 %
5	↓	C#	8.01 %	-0.4 %
6		C/C++	6.15 %	-1.1 %
7	↑	R	4.14 %	+0.1 %
8	↓	Objective-C	3.46 %	-1.0 %
9		Swift	2.75 %	-0.8 %
10		Matlab	2.15 %	-0.4 %

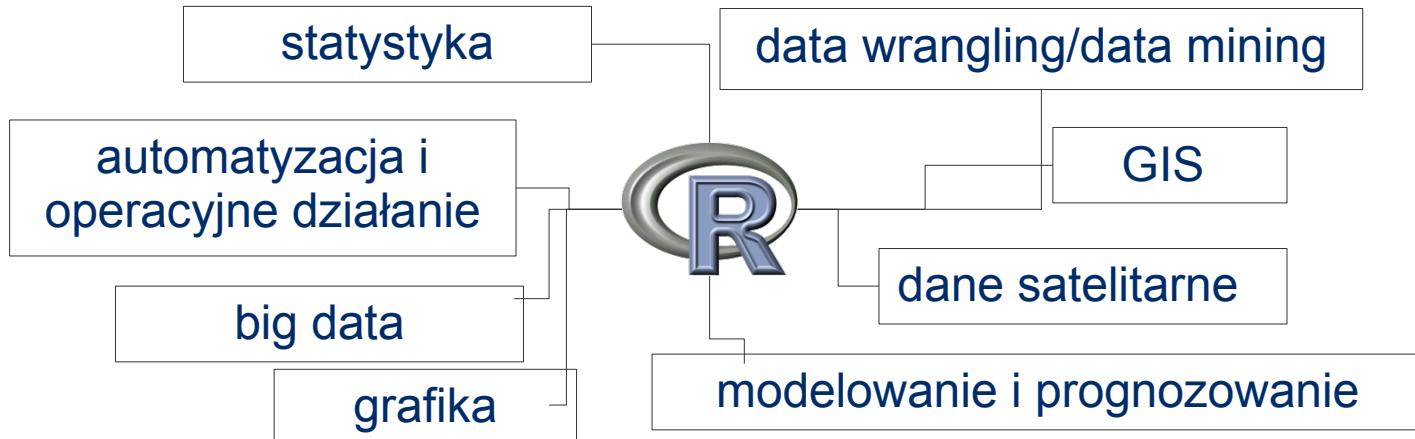


# Dlaczego popularność R wzrosnie?

- **Łatwy w nauce** (w porównaniu do innych języków programowania)
- Prosta składnia
- Brak kompilacji
- Zintegrowane i darmowe IDE (Rstudio)
- Multiplatformowość, skalowalność i mnogość obsługiwanych standardów
- **Bezplatny i otwartoźródłowy** → mali tna koszty, duzi (Google, Yahoo, New York Times, Facebook, Linde, NYSE...) dostosowują kod do swoich potrzeb
- **Ogrom gotowych rozwiązań** → każdy może napisać swoją paczkę... (CRAN >10k, + bioconductor, github ...)
- Liczba danych na których pracujemy jest ograniczona (w teorii) jedynie RAMem (limit MS Excel to... ?)
- **Rynek pracy**
- Społeczność R
- + ...

# Dlaczego używam ?

- Początki z R: od **2009**, jako **główne narzędzie** od ~**2010/2011**
- Background: Linux + Fortran + GIS
- Zastosowanie: nauki atmosferyczne + GIS (data mining + modelowanie + statystyka)
- Inny program do każdego typu problemu != **automatyzacja**



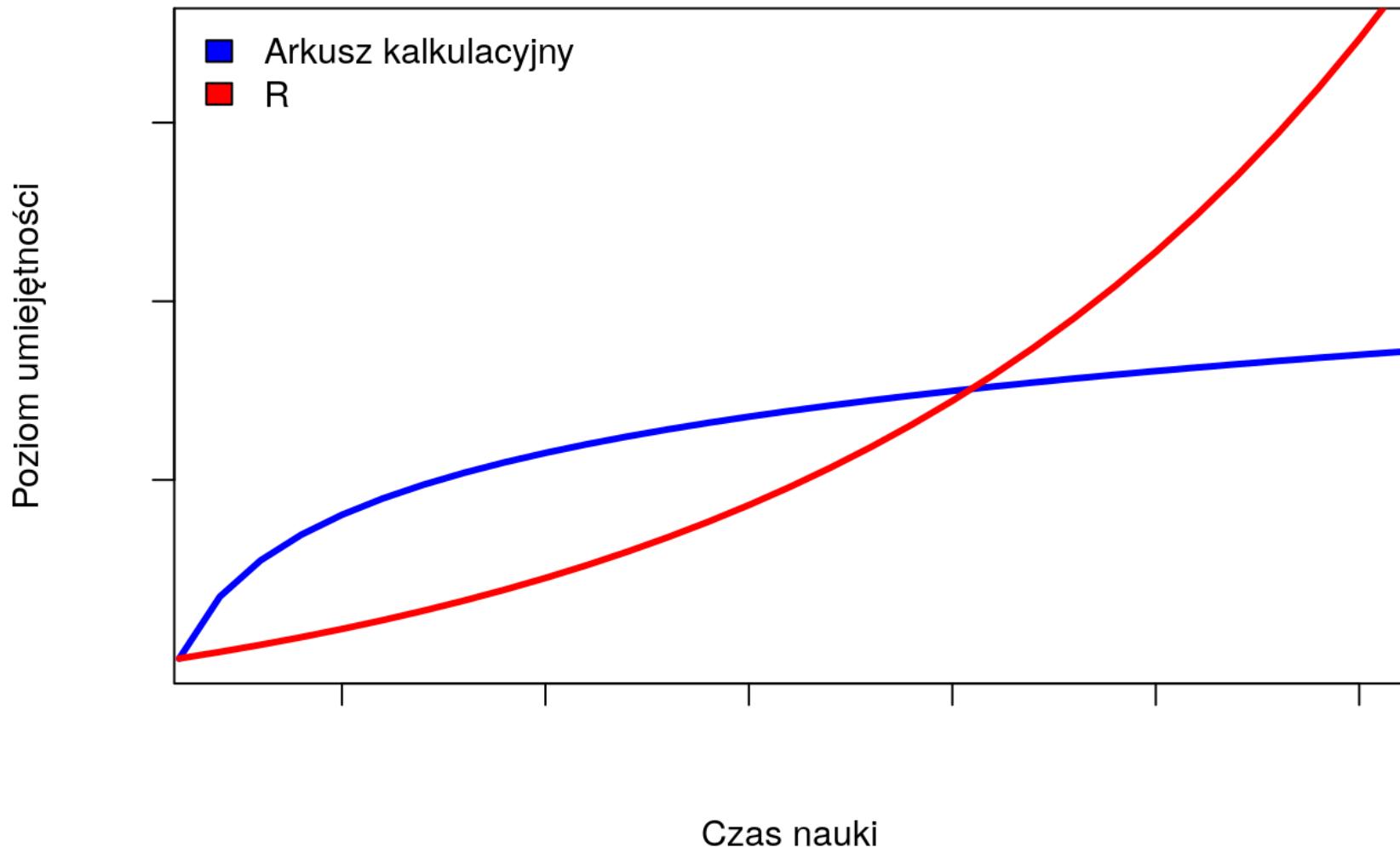
# Filozofia pracy w R

	<b>Arkusz kalkulacyjny</b>	<b>R</b>
<i>Interfejs użytkownika</i>	Klikane GUI	Linia komend
<i>Podgląd danych</i>	Do „ogarnięcia” wzrokiem	Ukryty w obiekcie
<i>Zadanie: Policzenie średniej</i>	Zaznaczamy komórki i klikamy funkcję	Piszemy komendę
<i>Zadanie: stworzenie wykresu</i>	Zaznaczamy komórki i klikamy 5x w opcje kreatora	Piszemy komendę
<i>Zadanie: stworzenie 12 identycznych wykresów</i>	Jak wyżej 12x	... dodajemy 1 słowo do wcześniejszej komendy

## Wniosek:

- nauka R wymusza nauczenia się (i stosowania) wielu komend i funkcji
  - wymaga **porzucenia** własnych **przyzwyczajeń** i nawyków (trudne)

## Krzywa uczenia R i arkuszy kalkulacyjnych



# X przykazań nauki



## 1) Nie bój się stromej krzywej uczenia =

uczenie tylko przez praktykę →  
*ćwicz ile się da, eksperymentuj ze składnią kodu  
i sprawdzaj różnice w wynikach*



## 2) Interpretuj błędy po każdej nieudanej komendzie =

```
> meanNLR<-mean(NLR) #sample mean of the normalized test statistic
Error: cannot allocate vector of size 304.5 Mb
```

## 3) Pracuj na własnych, dobrze znanych zbiorach danych =

Łatwiej zrozumieć działanie poszczególnych funkcji → wyłapać błędy

## 4) Staraj się unikać (początkowo) dużych zbiorów danych

(jeśli nie jest to wymagane)



## 5) Używaj wbudowanego systemu dokumentacji oraz rozwiązań online:



# X przykazań nauki



## 6) Zrozumienie podstaw jest kluczowe do analizy

bardziej zaawansowanych przypadków → dających dużo większą satysfakcję

## 7) Stosuj możliwie wiele komentarzy

- Łatwiej będzie zrozumieć po czasie filozofię własnego postępowania

## 8) Porządkuj kod oraz planuj działania, strukturę katalogów i nazwy plików

- docelowo twórz funkcje i paczki

## 9) R nie jest środowiskiem idealnym do każdych zastosowań

Teoretycznie można zrobić wszystko (ale nie zawsze ma to sens)

## 10) Rób przerwy. Czasem najlepsze pomysły przychodzą

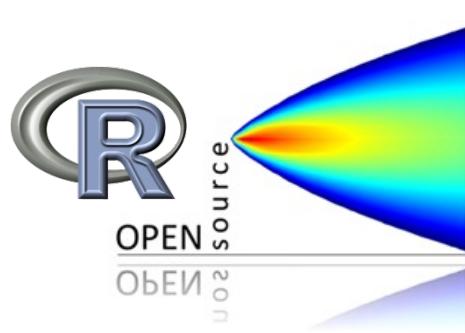
w najmniej oczekiwanych momentach. Niekoniecznie przy komputerze.

## 11) To samo rozwiążanie można osiągnąć na wiele sposobów

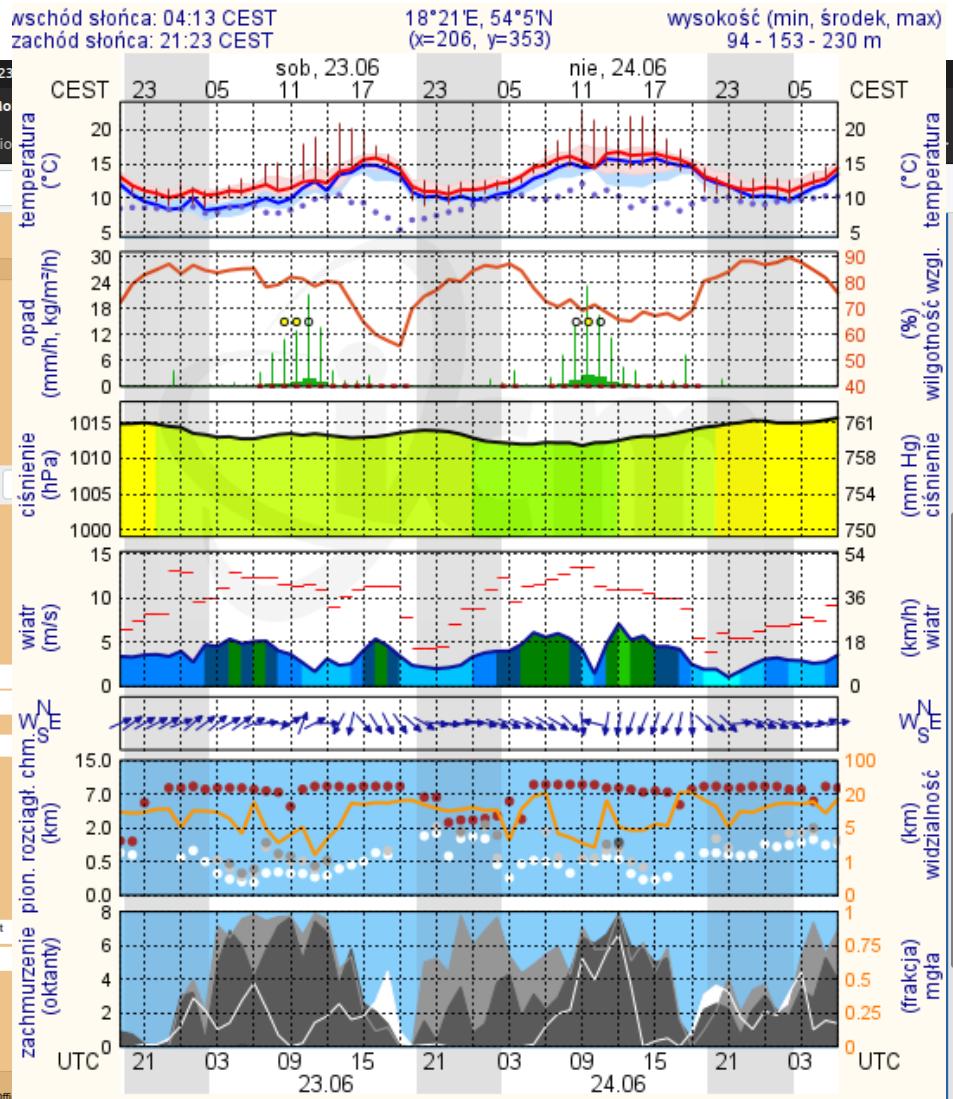
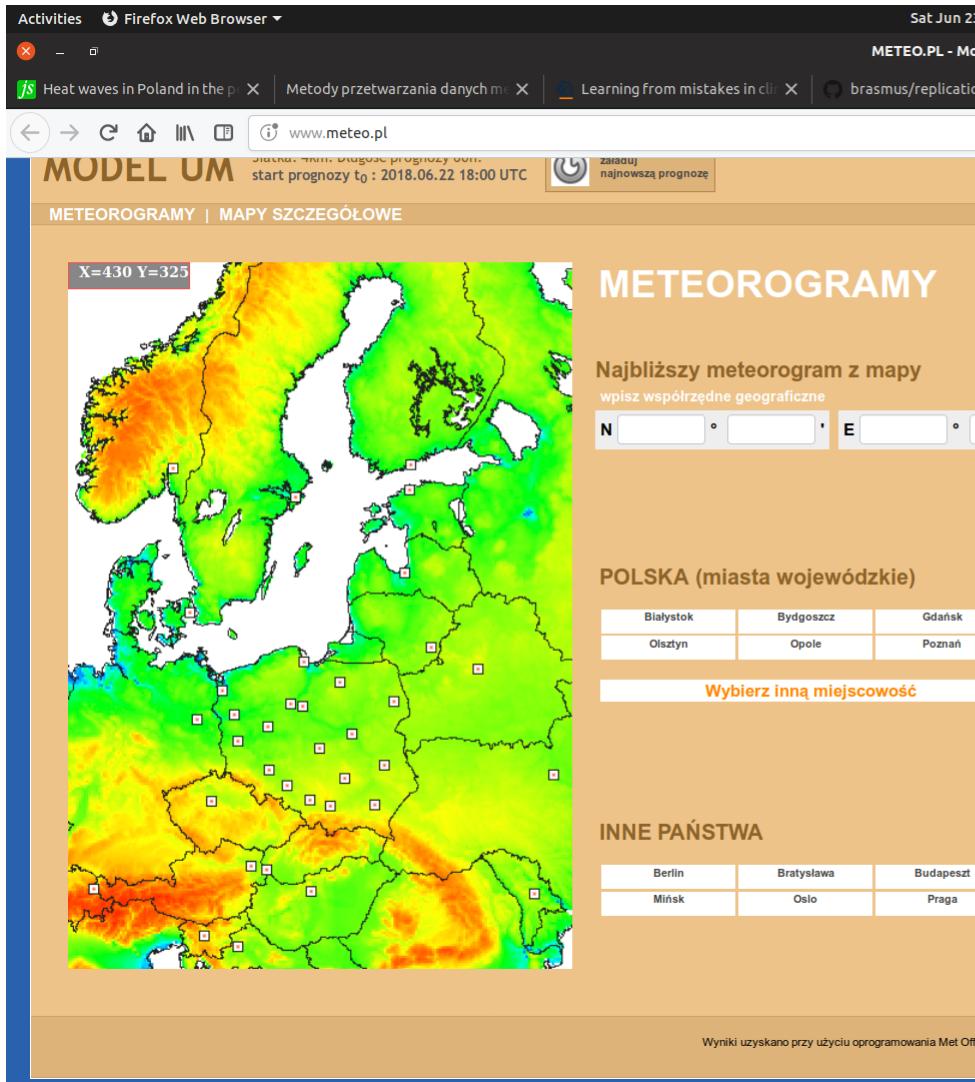
- Nie warto przejmować się tym, że gdzieś ktoś zrobił coś inaczej, o ile naszym nadziednym celem nie jest wydajność i profilowanie kodu



# w naukach atmosferycznych



# Przykładowe produkty oparte o R (o których prawdopodobnie nie wiedzieliśmy)



# Przykładowe produkty oparte o R

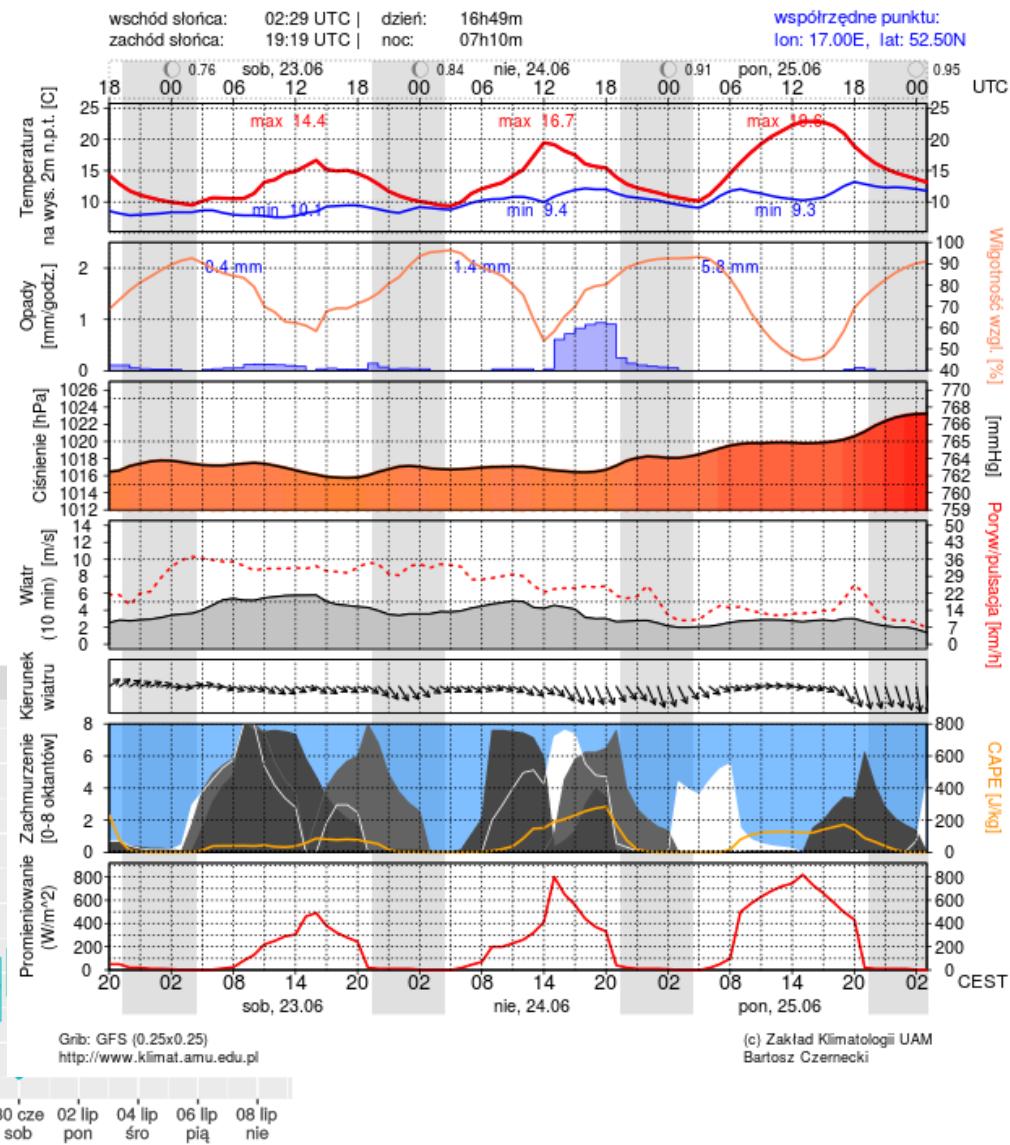
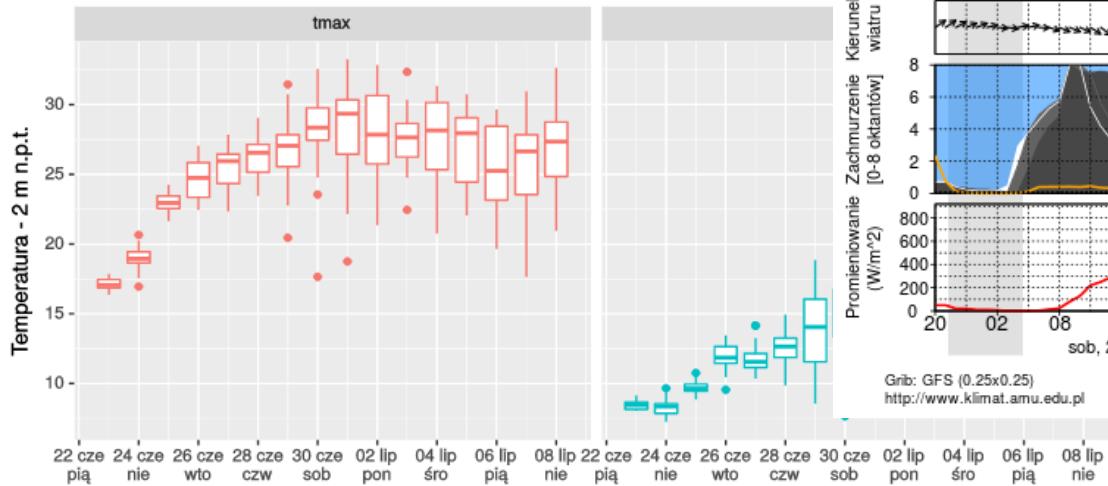
European  
Numerical  
Weather  
Outlooks

Meteogramy (PL)

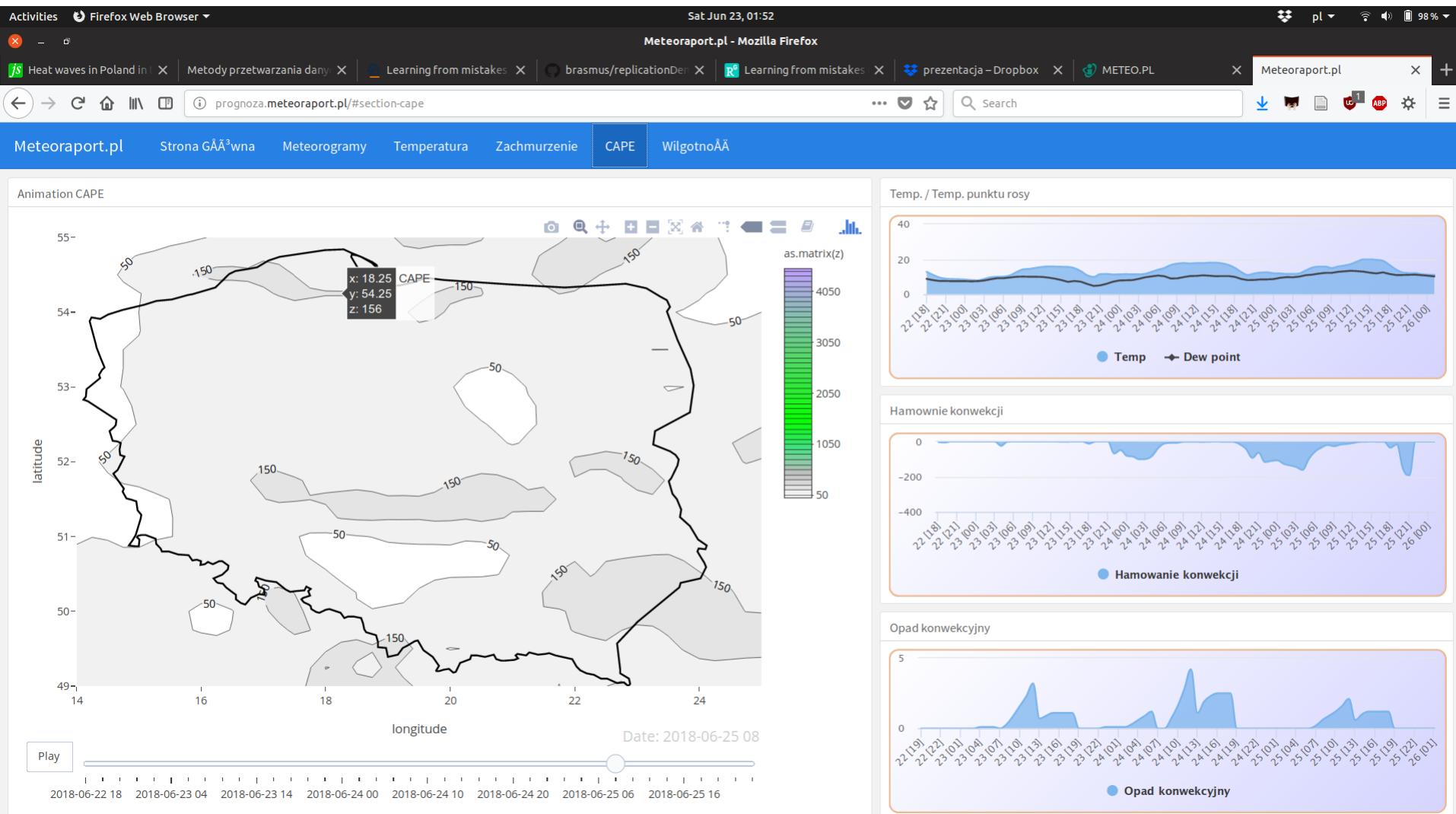


enwo.pl

Prognoza długoterminowa

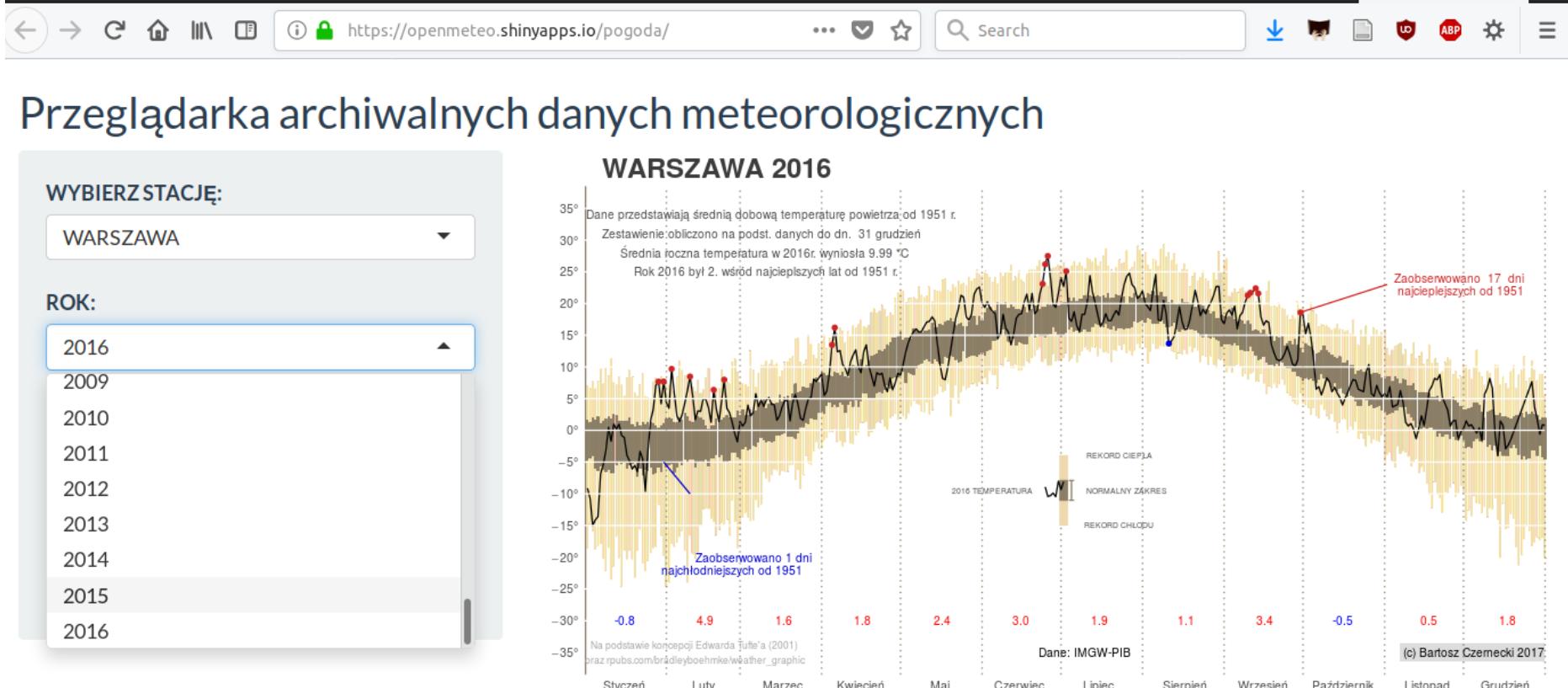


# Przykładowe produkty oparte o R



prognoza.meteoraport.pl

# Przykładowe produkty oparte o R



<https://openmeteo.shinyapps.io/pogoda/>

# Przykładowe paczki

→ ‘climatol’

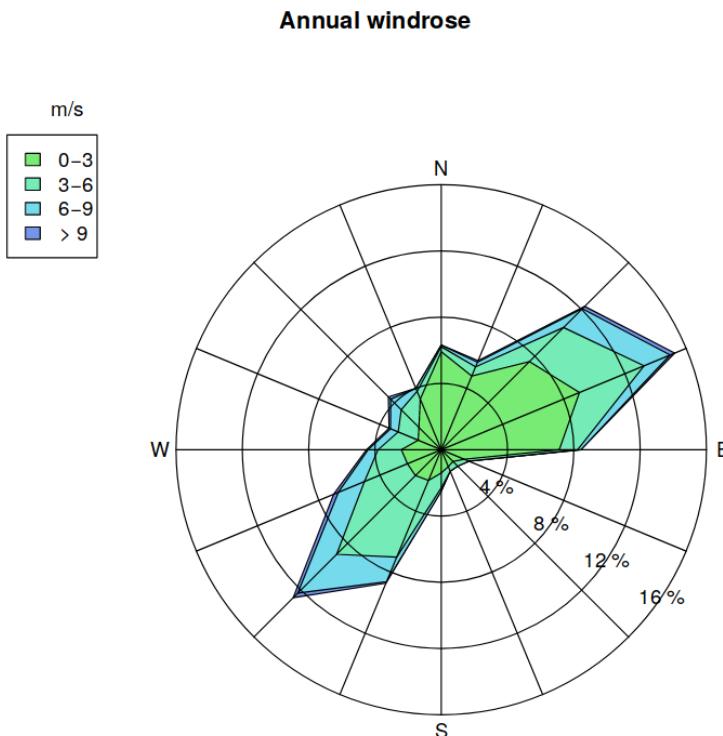


Figure 17: Example of a wind rose obtained with the `rosavent` function.

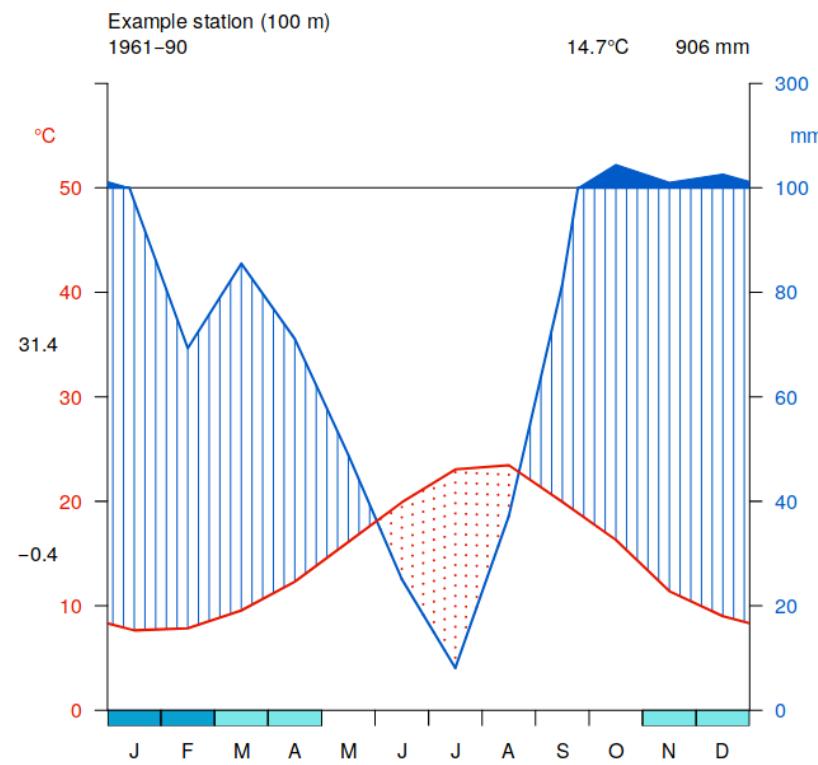


Figure 18: Example of a Walter&Lieth diagram obtained with the `diagwl` function.

# Przykładowe paczki

→ ‘climatol’



Tmin at S03(1), La Perla

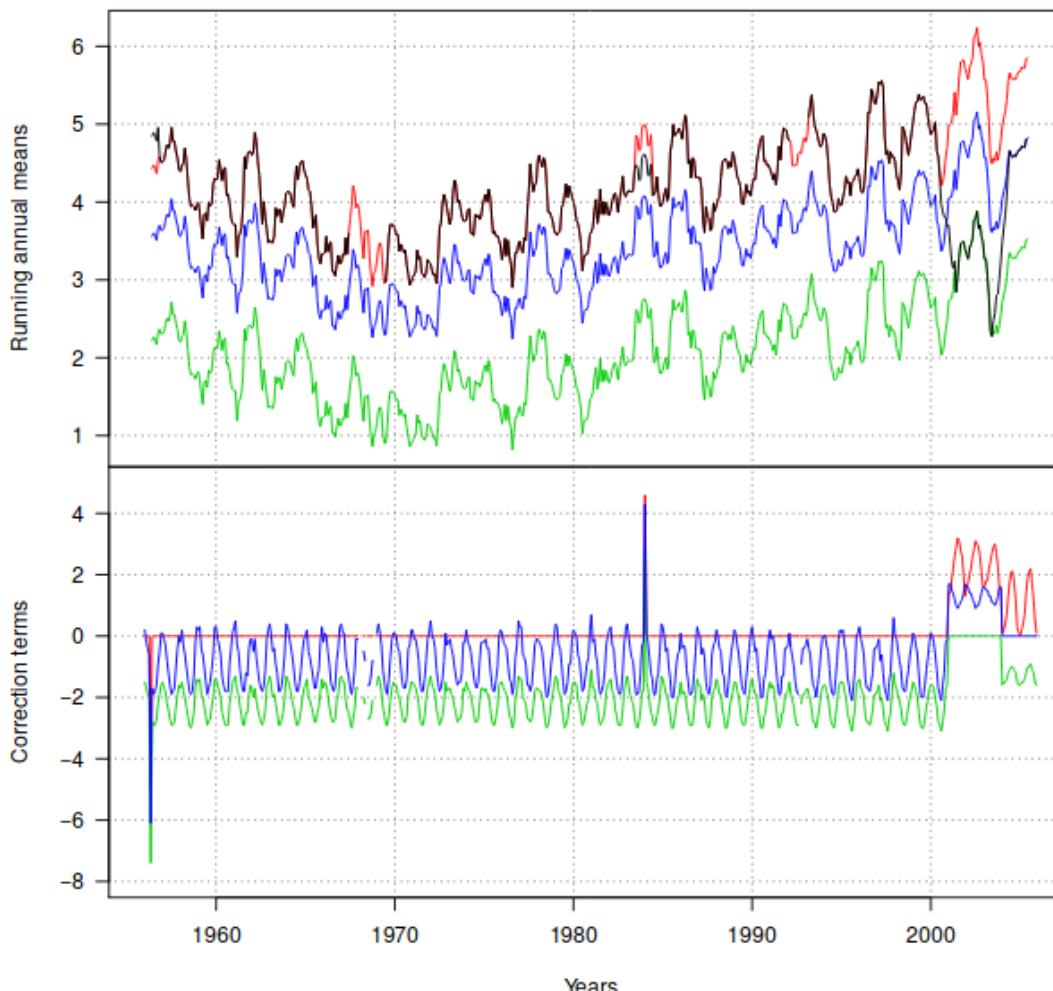
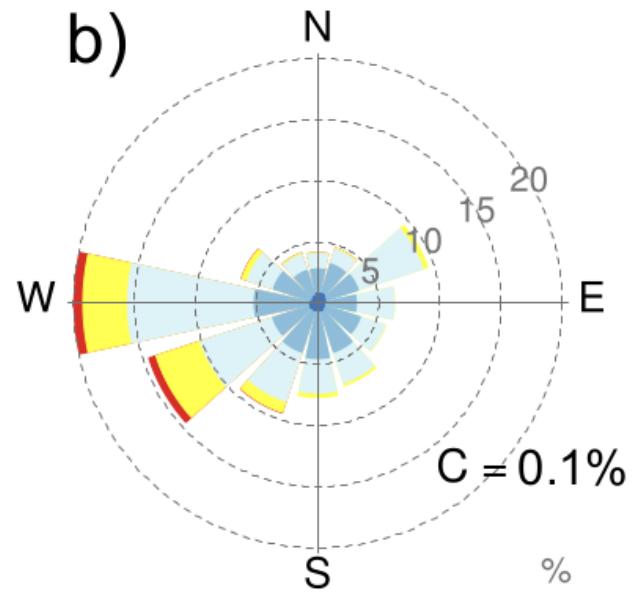
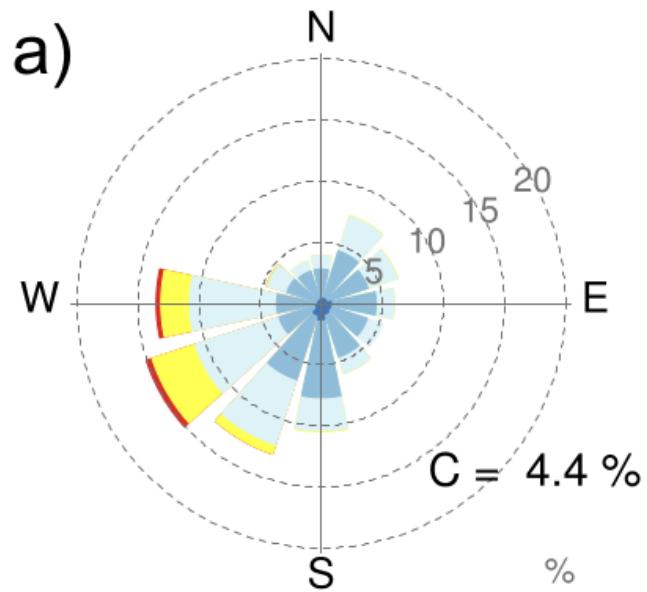


Figure 15: Original (in black) and reconstructed running annual series (top), and corrections applied to each fragment (bottom).

# Przykładowe paczki → ‘bReeze’



m/s  
0.1-2.0  
2.1-5.0  
5.1-10.0  
10.1-15.0  
>15  
procentowy  
C - udział cisz wiatrowych

# Przykładowe paczki → ‘solaR’

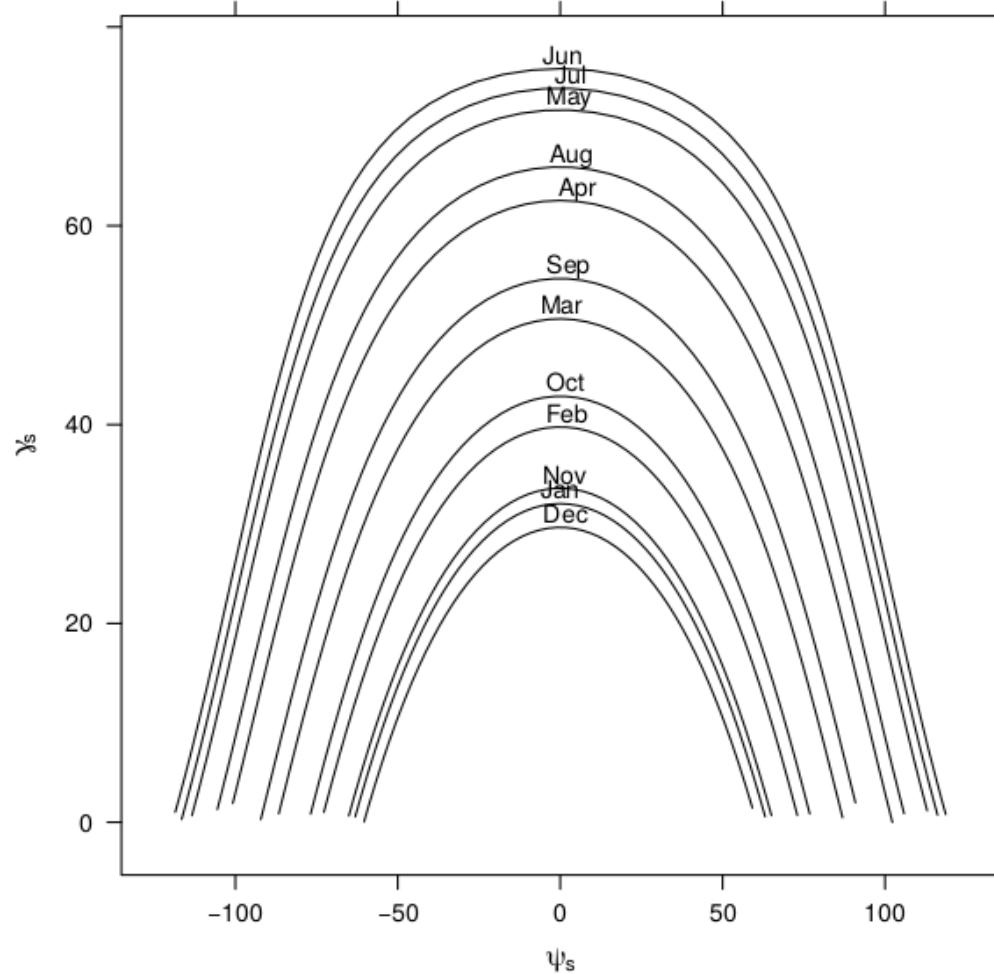
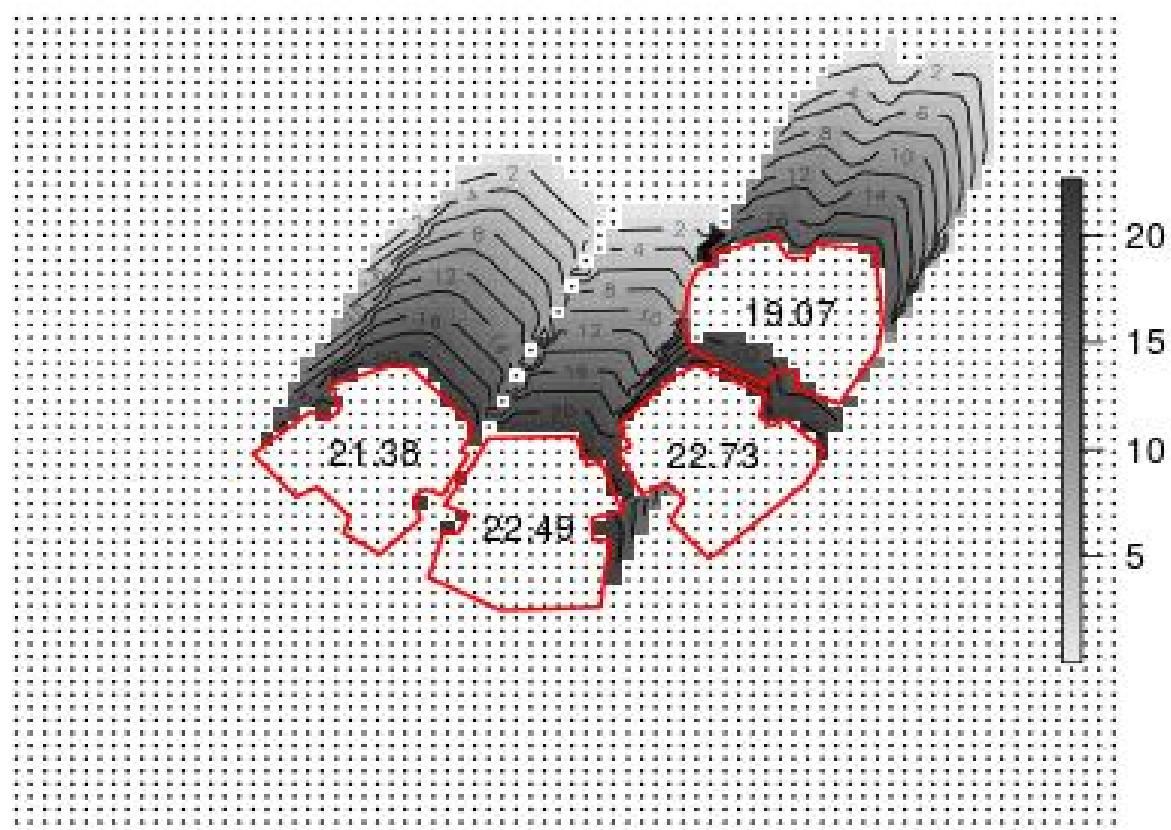
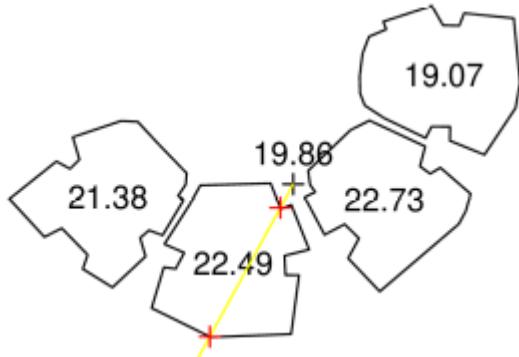


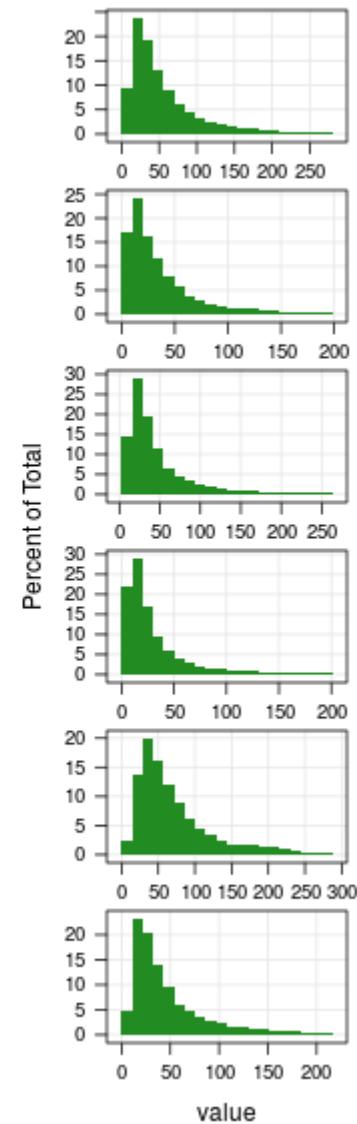
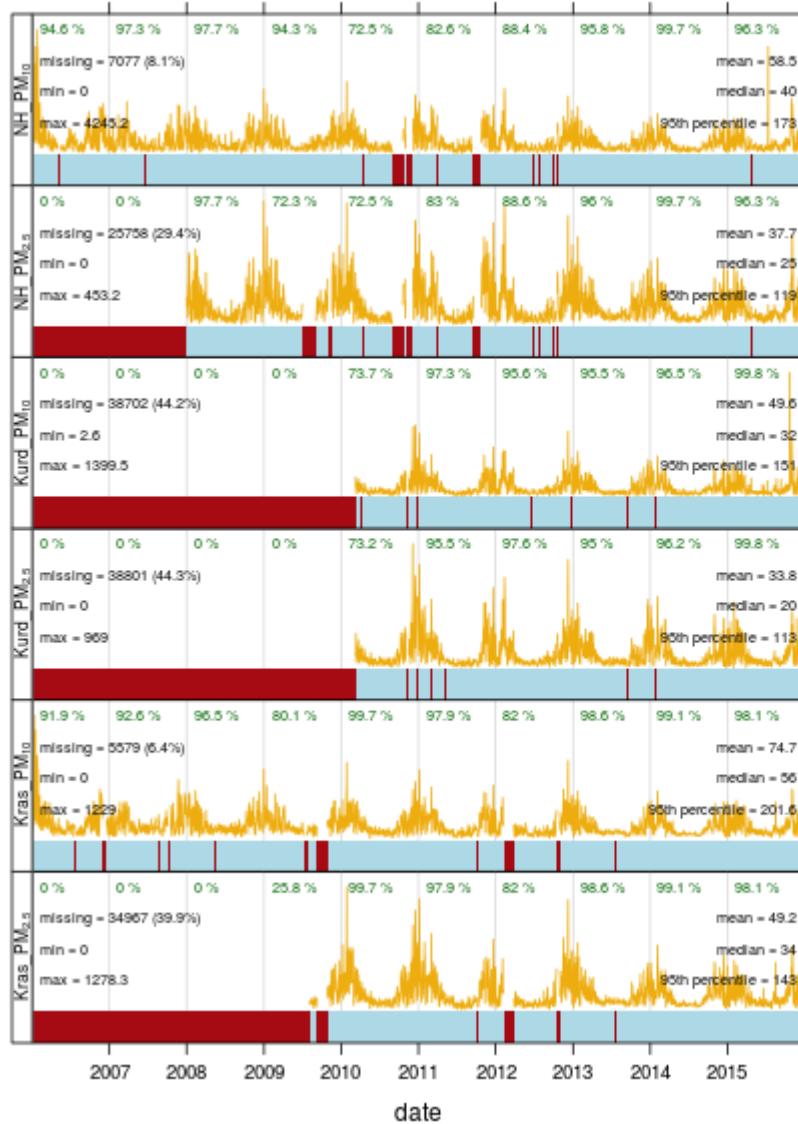
Figure 2: Azimuth and height solar angles during the “average days”.



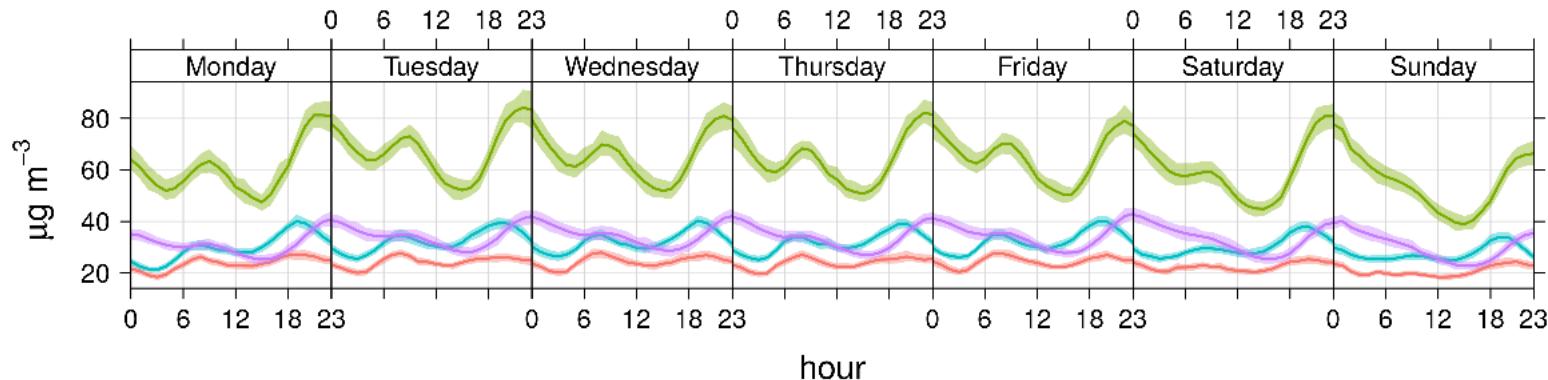
# Przykładowe paczki → ‘solaR’



# Przykładowe paczki → ‘openair’

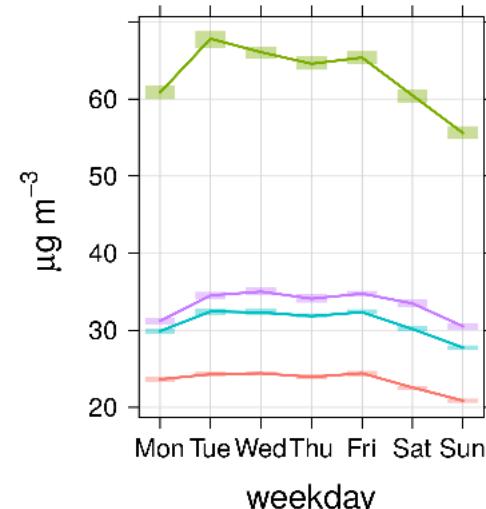
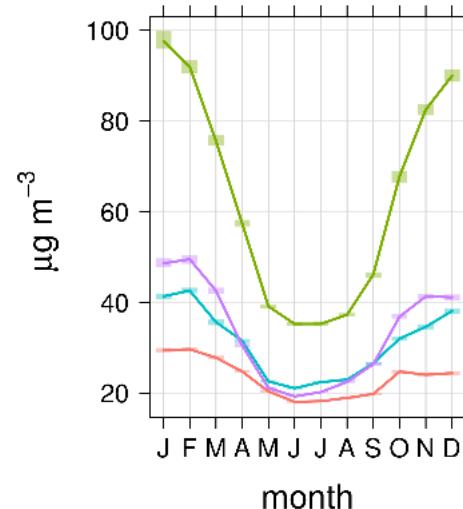
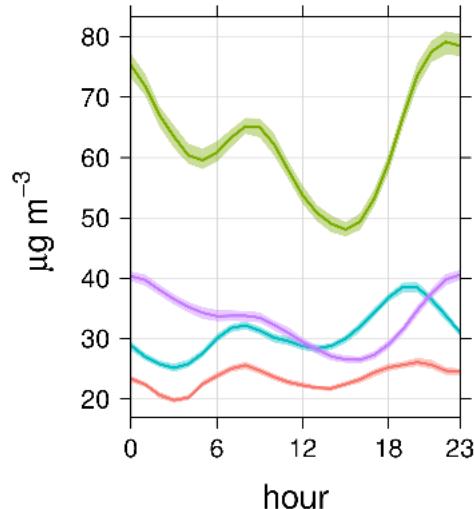


# Przykładowe paczki → ‘openair’



Tricity  
Kraków

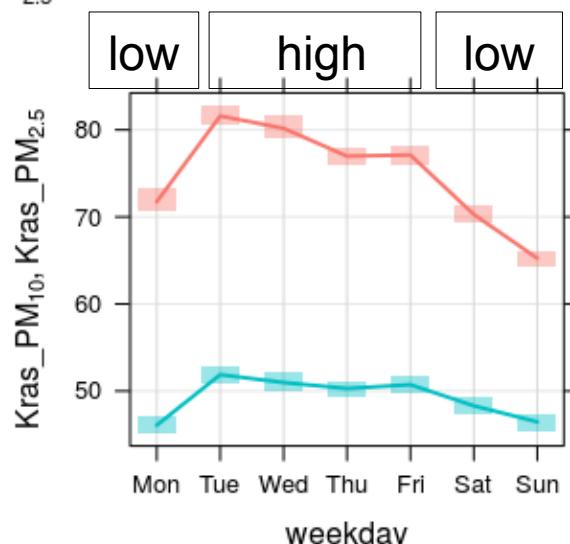
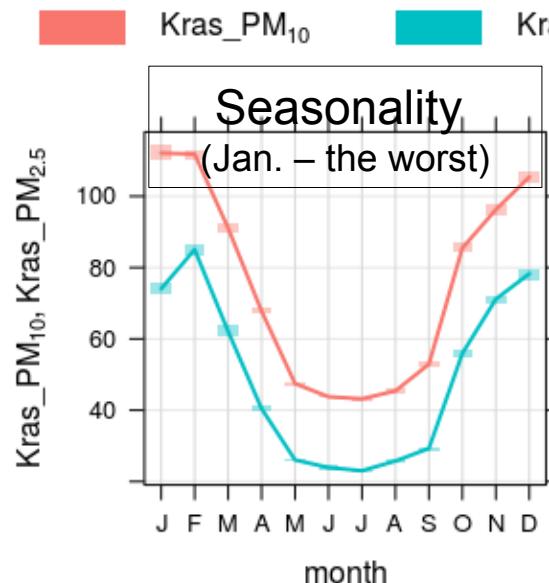
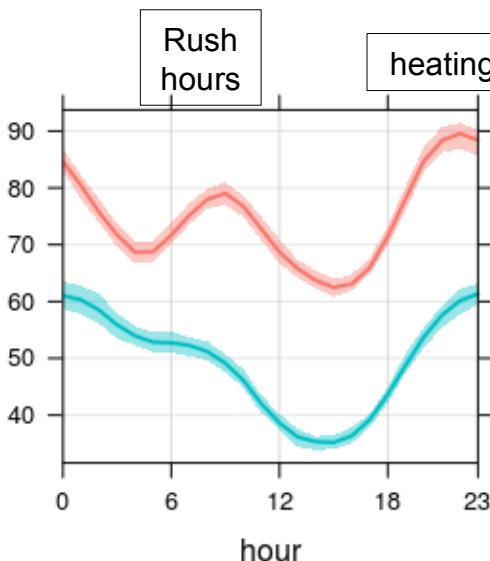
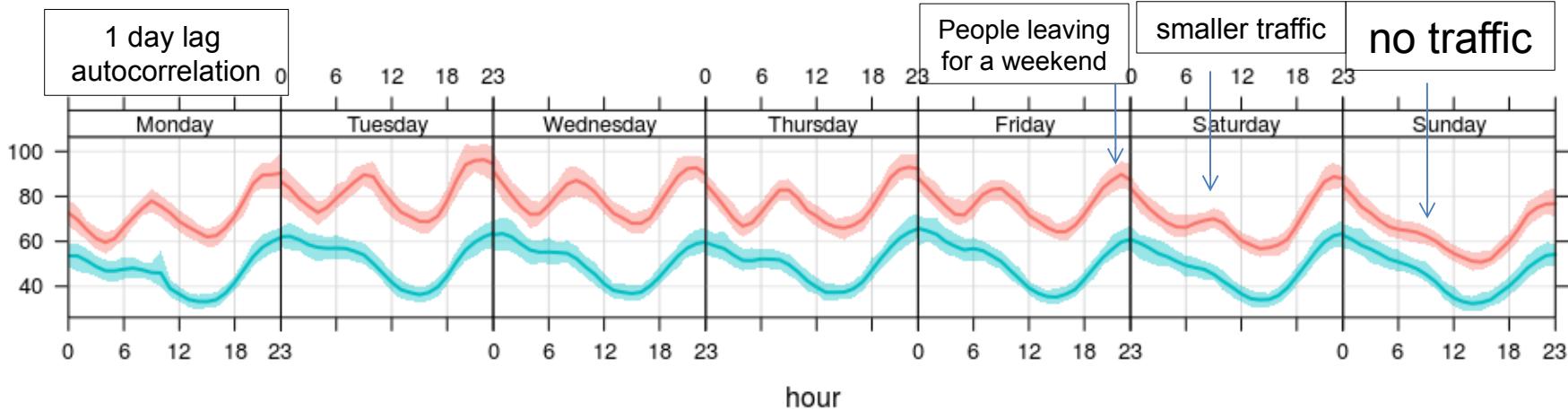
Łódź  
Poznań



mean and 95% confidence interval in mean

Jędruszkiewicz et al. 2017

# Przykładowe paczki → ‘openair’

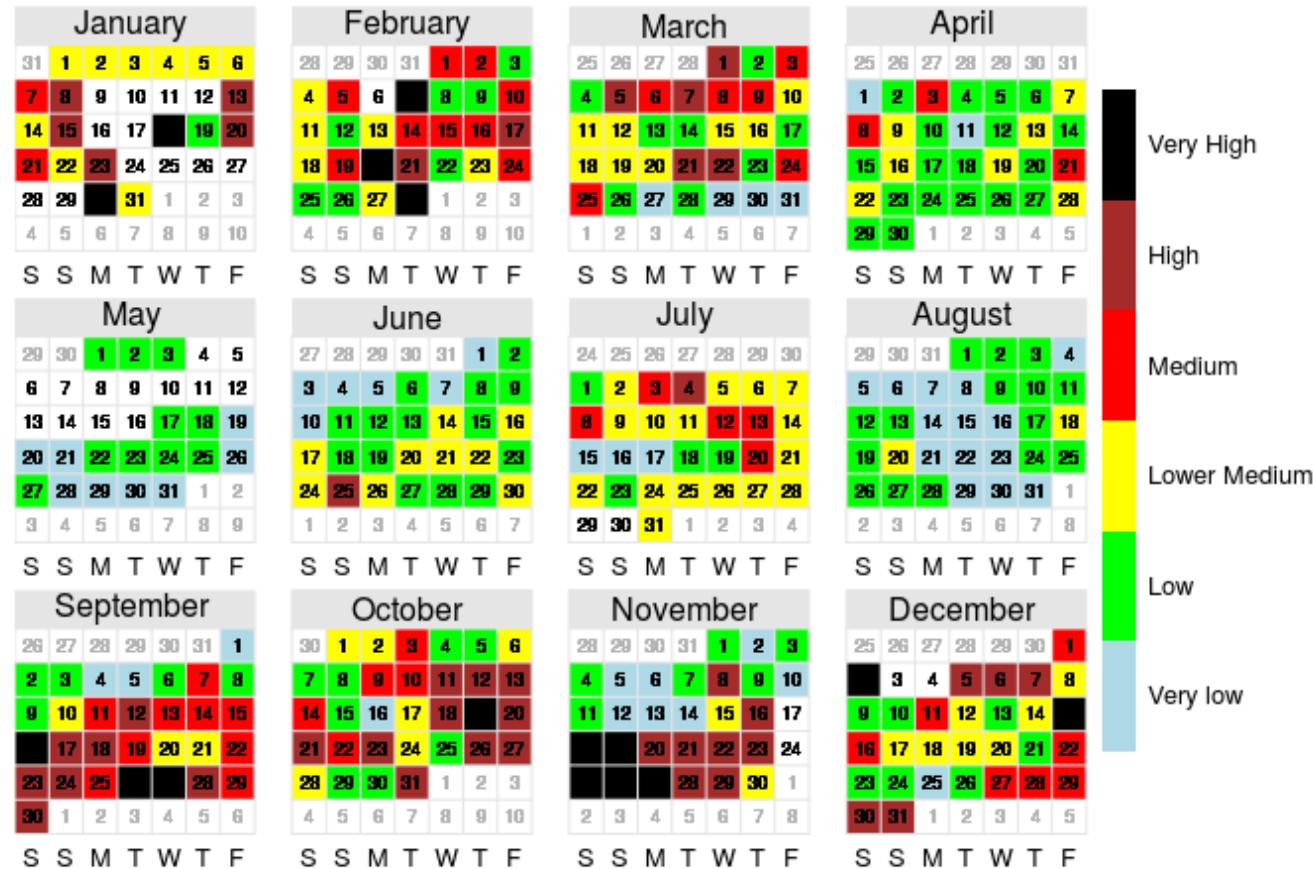


mean and 95% confidence interval in mean

# Przykładowe paczki → ‘openair’

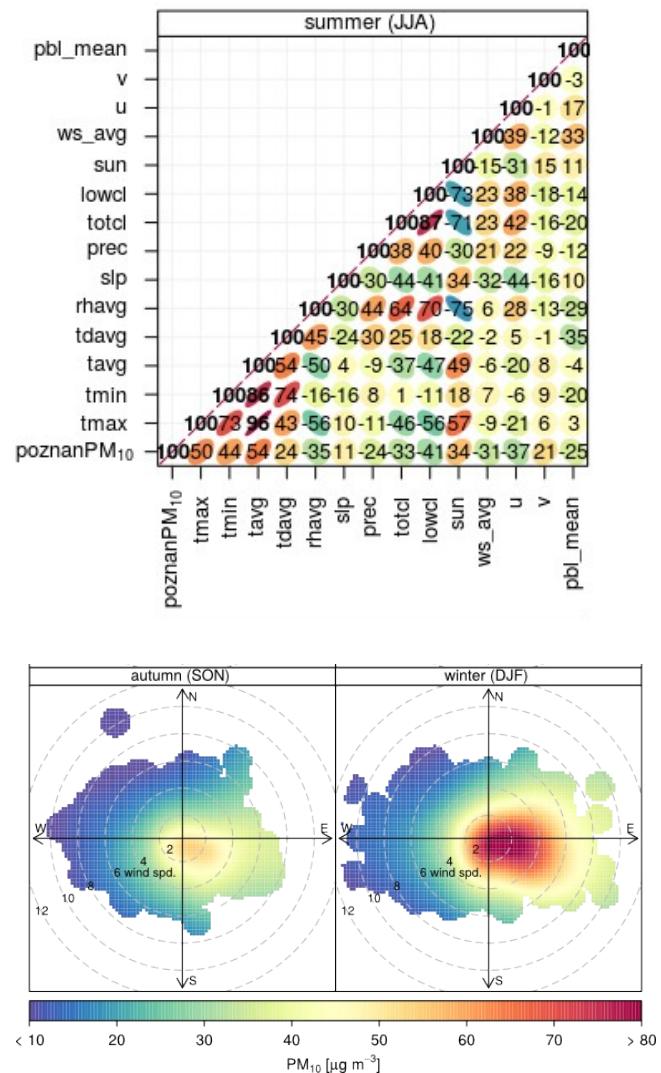
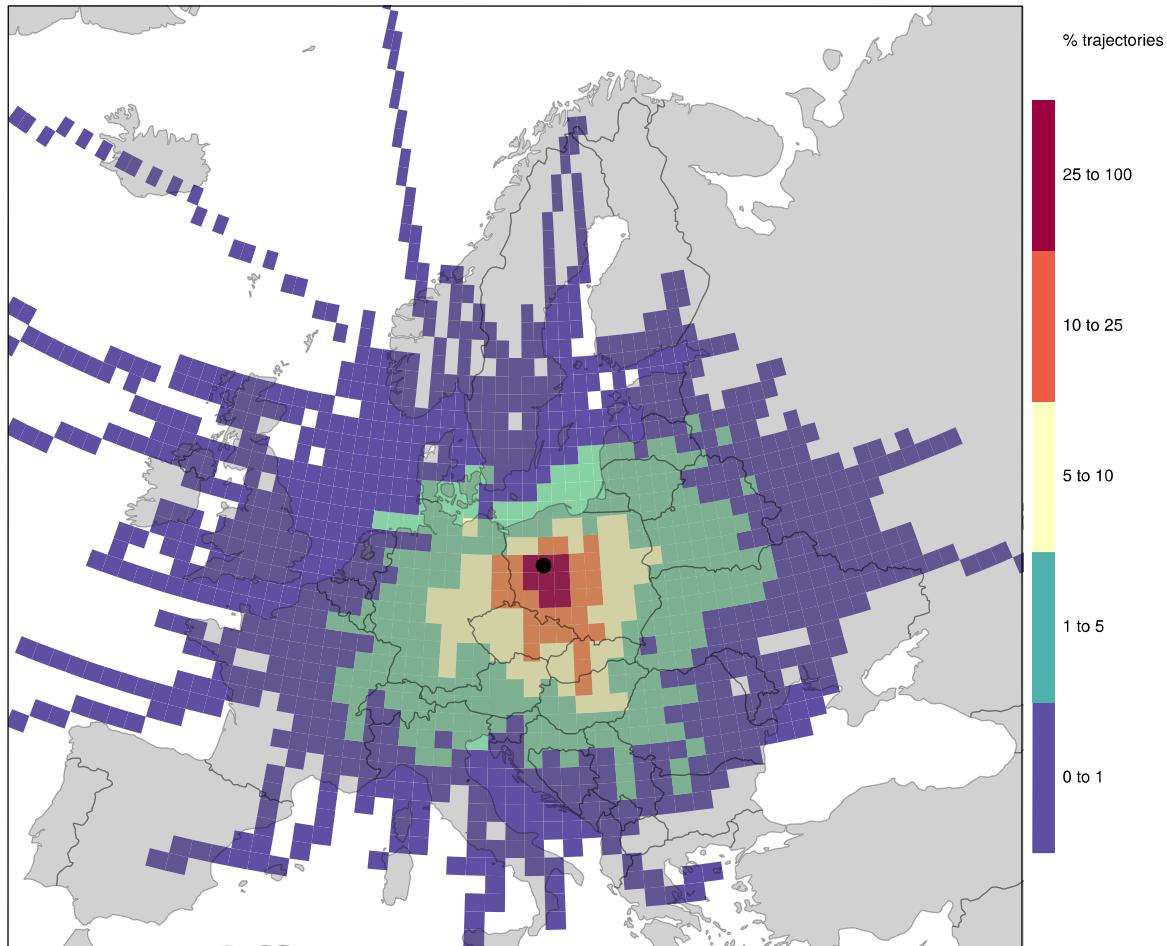


Nowa Huta (NH)



```
calendarPlot(tt, pollutant = "Kras_pm10", breaks = c(0, 25, 50, 75, 100, 150, 200), labels = c("Very low", "Low", "Lower Medium", "Medium", "High", "Very High"), cols = c("lightblue", "green", "yellow", "red", "brown", "black"), statistic = "mean", main='Krasinskiego', year=2006)
```

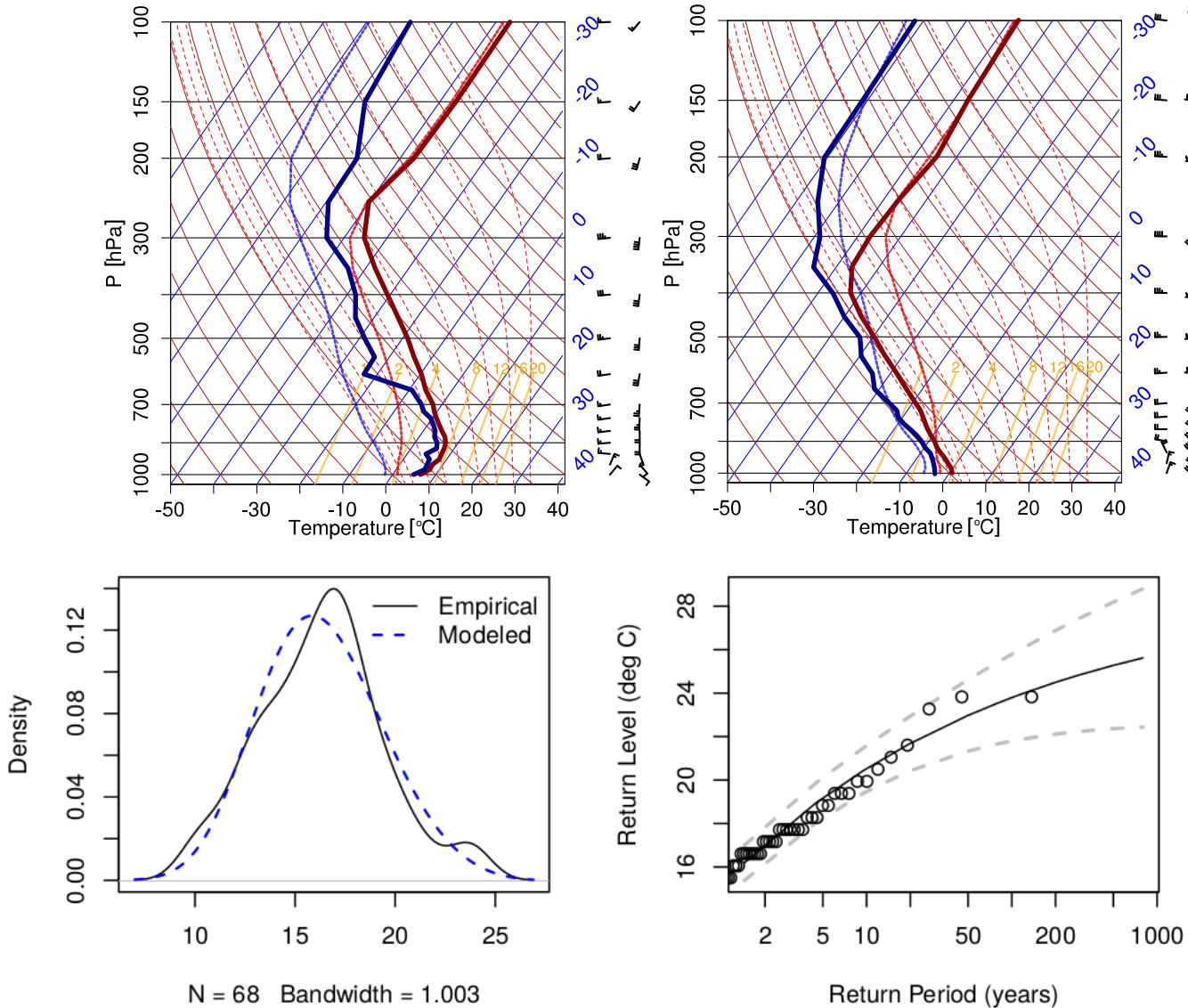
# Przykładowe paczki → ‘openair’



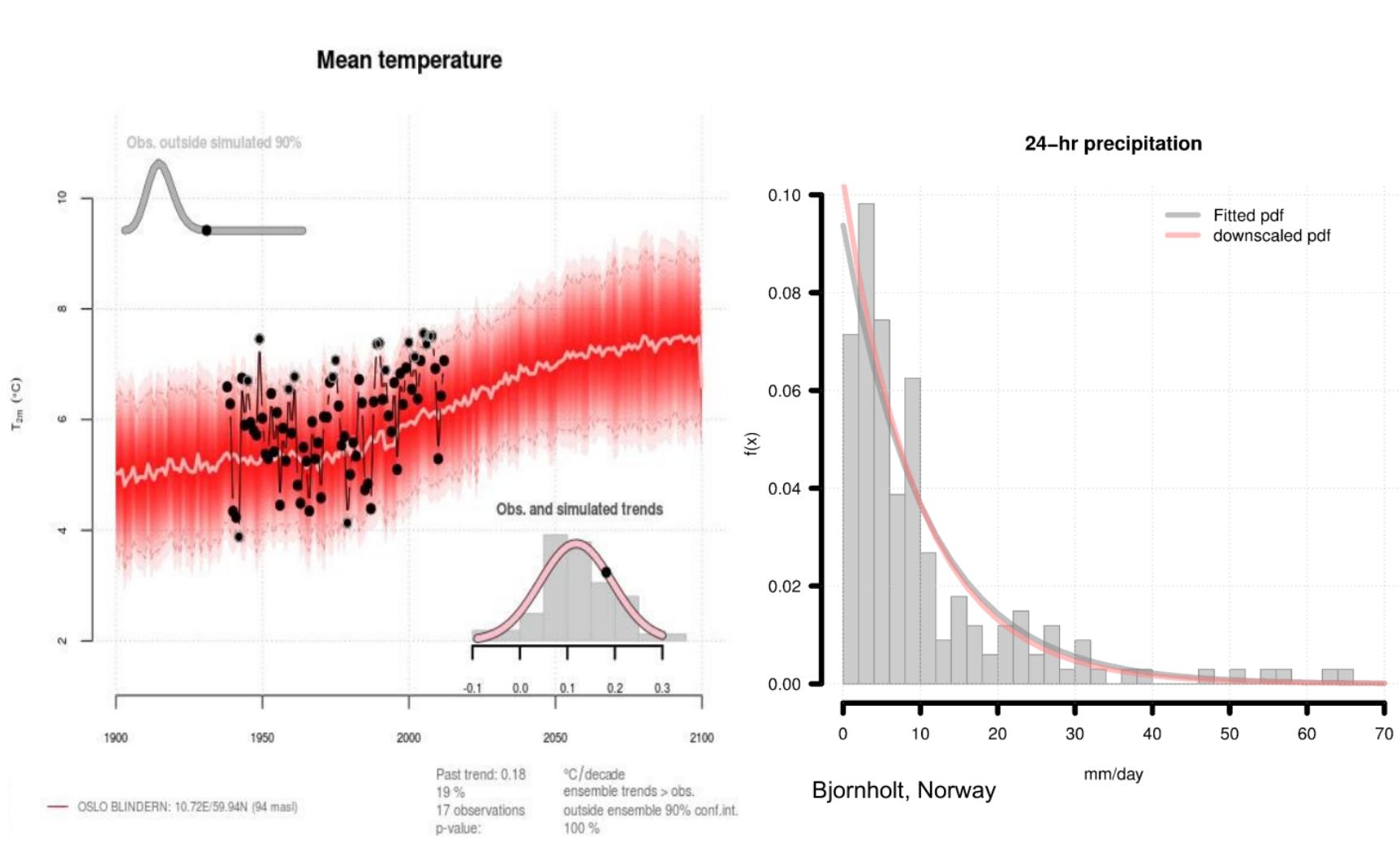
Czernecki et al. 2017

# Przykładowe paczki →

‘radiosonde’, ‘extRemes’



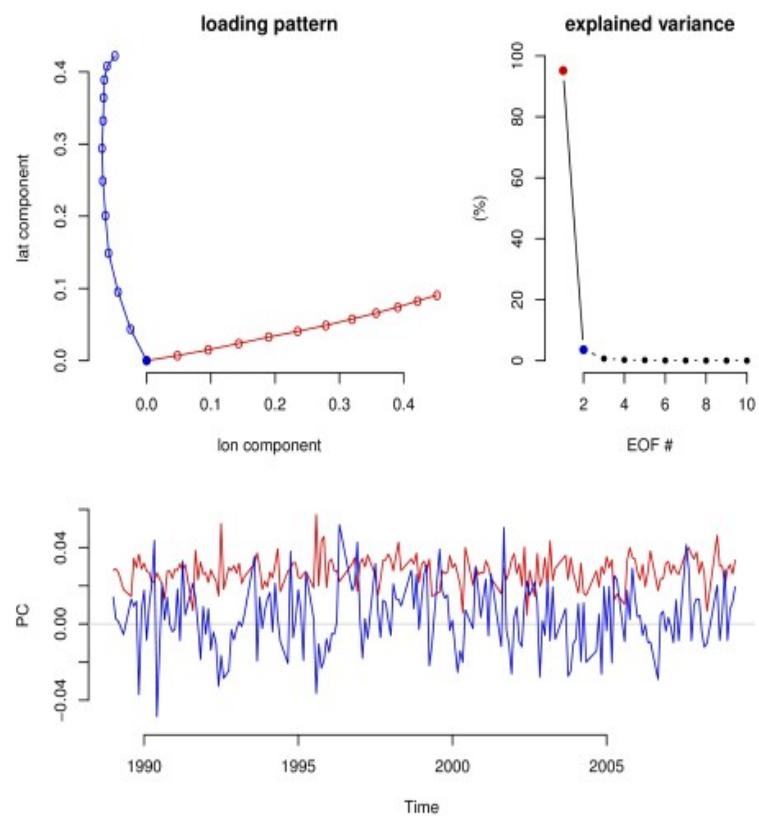
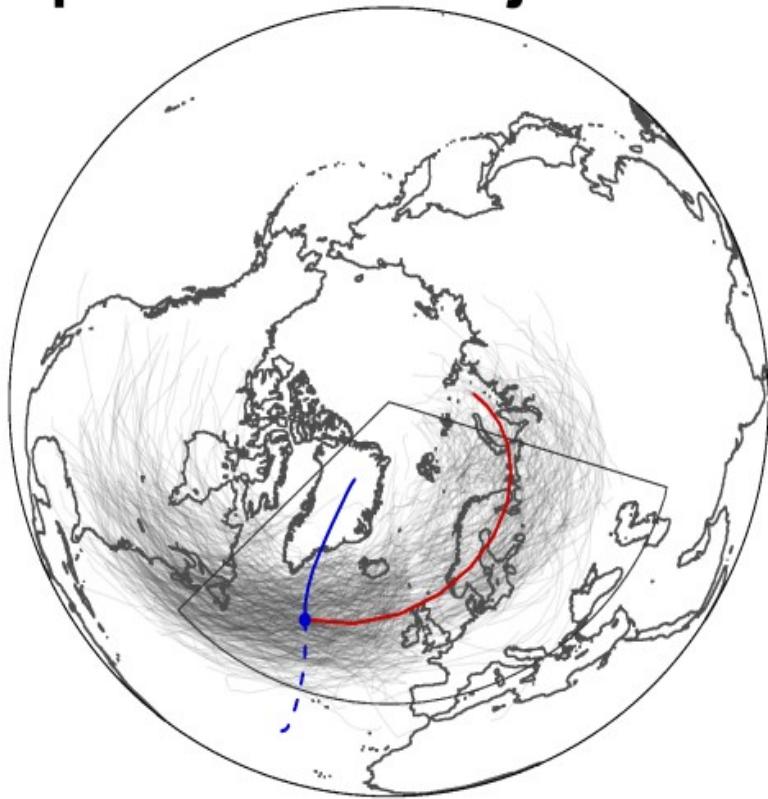
# Przykładowe paczki → ‘esd’



# Przykładowe paczki → ‘esd’



## 1. Explore storm trajectories



# Lingua fRanca

Popularność dzięki otwartości kodu źródłowego i reprodukowalności wyników badań:

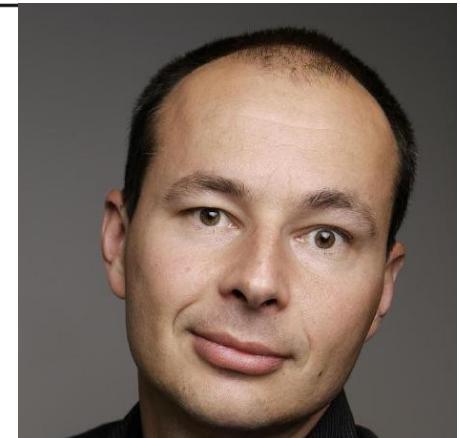
Theor Appl Climatol (2016) 126:699–703  
DOI 10.1007/s00704-015-1597-5



ORIGINAL PAPER

## Learning from mistakes in climate research

Rasmus E. Benestad<sup>1</sup> · Dana Nuccitelli<sup>2</sup> · Stephan Lewandowsky<sup>3,4</sup> ·  
Katharine Hayhoe<sup>5</sup> · Hans Olav Hygen<sup>1</sup> · Rob van Dorland<sup>6</sup> · John Cook<sup>2,7,8</sup>

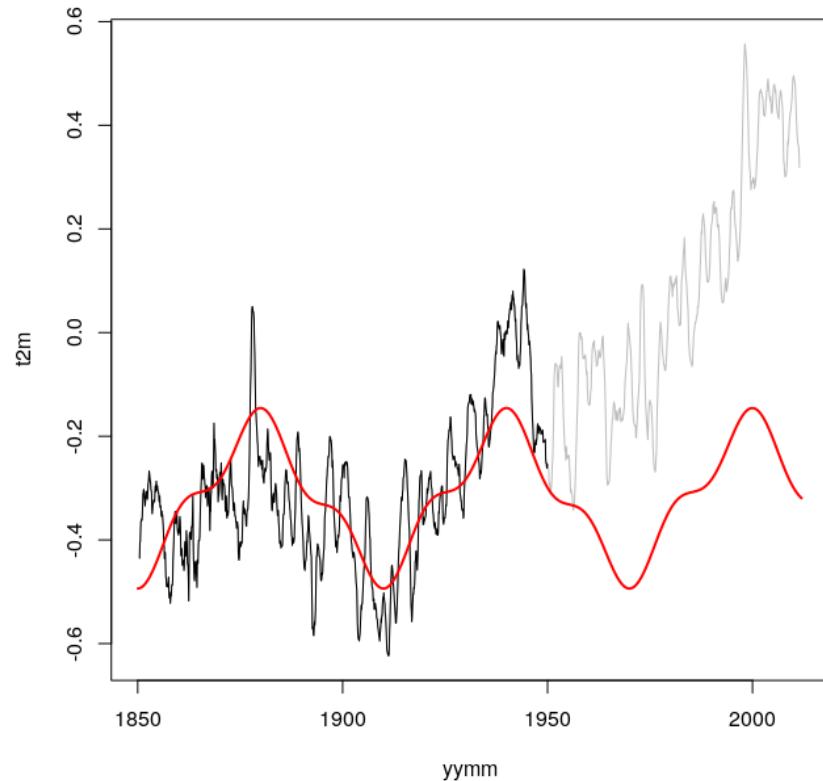
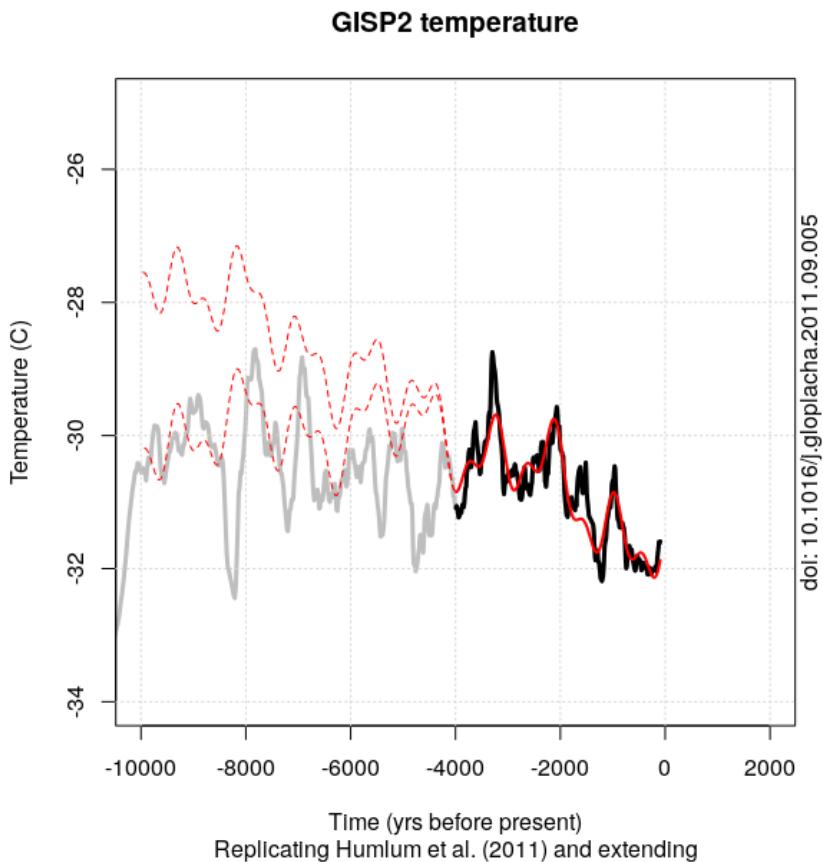


```
install.packages("replicationDemos_1.11.tar.gz", repos=NULL)
```

```
library(replicationDemos)
```

# ***Humlum et al. 2011 ; Loehle and Scafetta 2011***

„Odcięcie” danych niezbędnych do analizy oraz błędna analiza cykli:



# Czy zawsze są potrzebne paczki?

- do statystyki i podstawowej wizualizacji  
(zwykle) nie



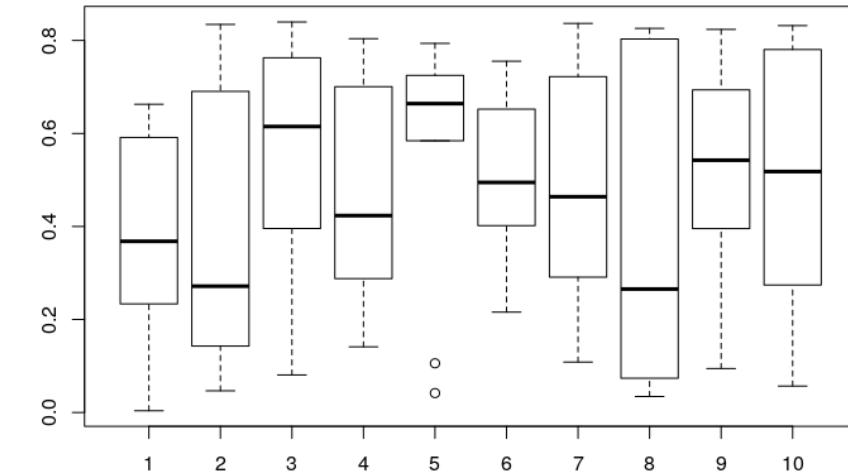
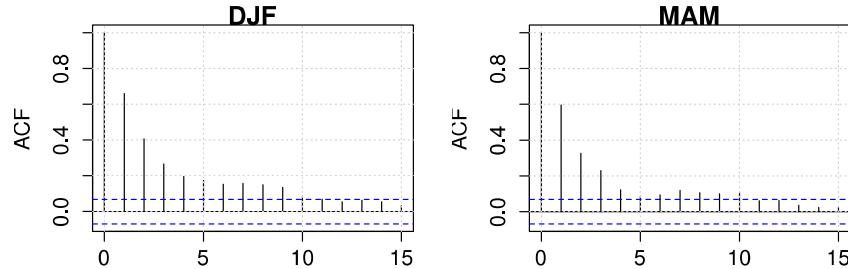
Information on package 'stats'

## Description:

```
Package:      stats
Version:     3.4.4
Priority:    base
Title:       The R Stats Package
Author:      R Core Team and contributors worldwide
Maintainer:  R Core Team <R-core@r-project.org>
Description: R statistical functions.
License:     Part of R 3.4.4
Imports:     utils, grDevices, graphics
Suggests:    MASS, Matrix, SuppDists, methods, stats4
NeedsCompilation: yes
Built:       R 3.4.4; x86_64-pc-linux-gnu; 2018-04-21 14:00:25 UTC; unix
```

## Index:

```
.checkMFClasses   Functions to Check the Type of Variables passed
                  to Model Frames
AIC               Akaike's An Information Criterion
ARMAacf           Compute Theoretical ACF for an ARMA Process
ARMAtoMA          Convert ARMA Process to Infinite MA Process
Beta              The Beta Distribution
Binomial          The Binomial Distribution
Box.test          Box-Pierce and Ljung-Box Tests
C                 Sets Contrasts for a Factor
Cauchy             The Cauchy Distribution
Chisquare          The (non-central) Chi-Squared Distribution
Distributions      Distributions in the stats package
Exponential        The Exponential Distribution
FDist              The F Distribution
GammaDist          The Gamma Distribution
Geometric          The Geometric Distribution
HoltWinters        Holt-Winters Filtering
Hypergeometric     The Hypergeometric Distribution
IQR               The Interquartile Range
KalmanLike         Kalman Filtering
Logistic            The Logistic Distribution
Lognormal           The Log Normal Distribution
Multinomial         The Multinomial Distribution
NLSstAsymptotic   Fit the Asymptotic Regression Model
NLSstClosestX     Inverse Interpolation
NLSstLfAsymptote  Horizontal Asymptote on the Left Side
NLSstRtAsymptote  Horizontal Asymptote on the Right Side
NegBinomial        The Negative Binomial Distribution
Normal             The Normal Distribution
PP.test            Phillips-Perron Test for Unit Roots
Poisson            The Poisson Distribution
```



```
library(help = "stats")
```

~300 wbudowanych funkcji statystycznych

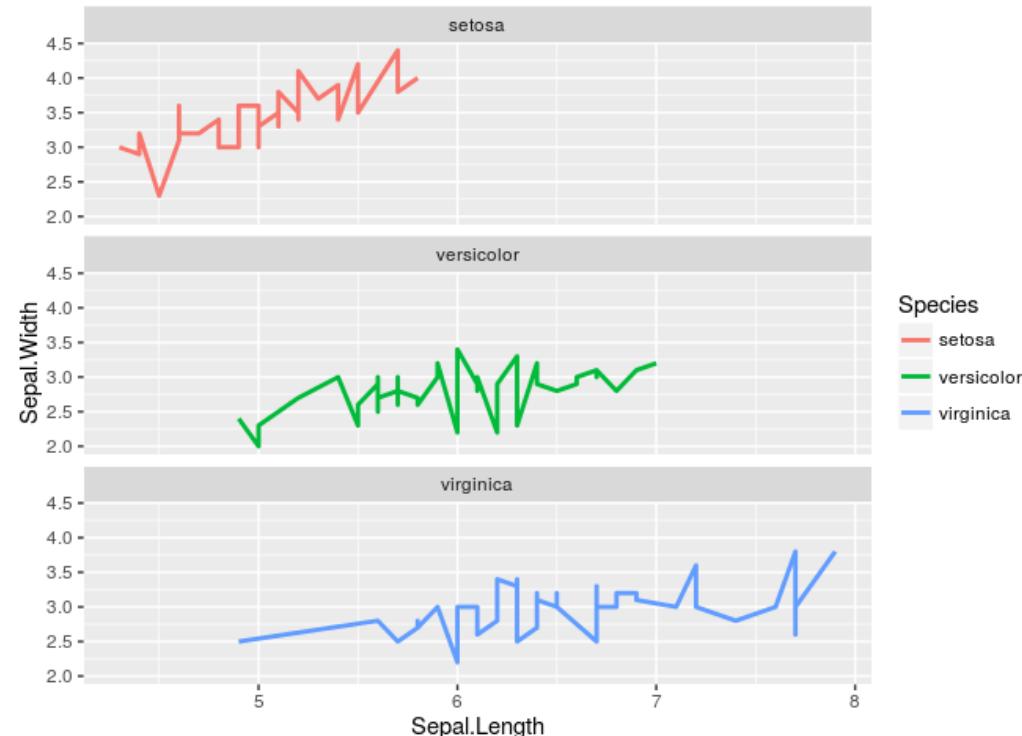
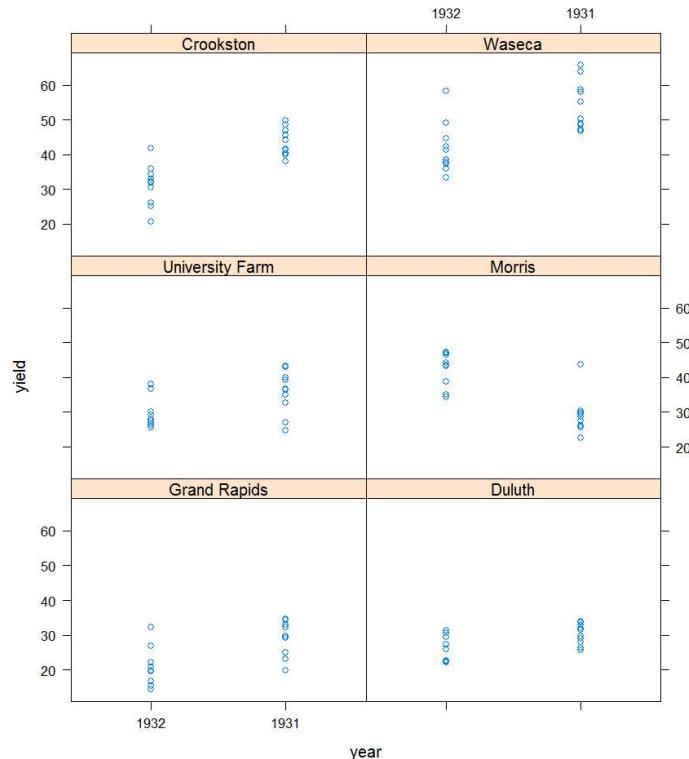
# Silniki graficzne R



R posiada natywnie wbudowany, wysoce konfigurowalny bazowy „silnik” graficzny (*'graphics'*) | > demo (graphics)

Najczęściej możliwości grafiki (statycznej) są rozszerzane za pomocą systemów:

*'lattice'* oraz *'ggplot2'*

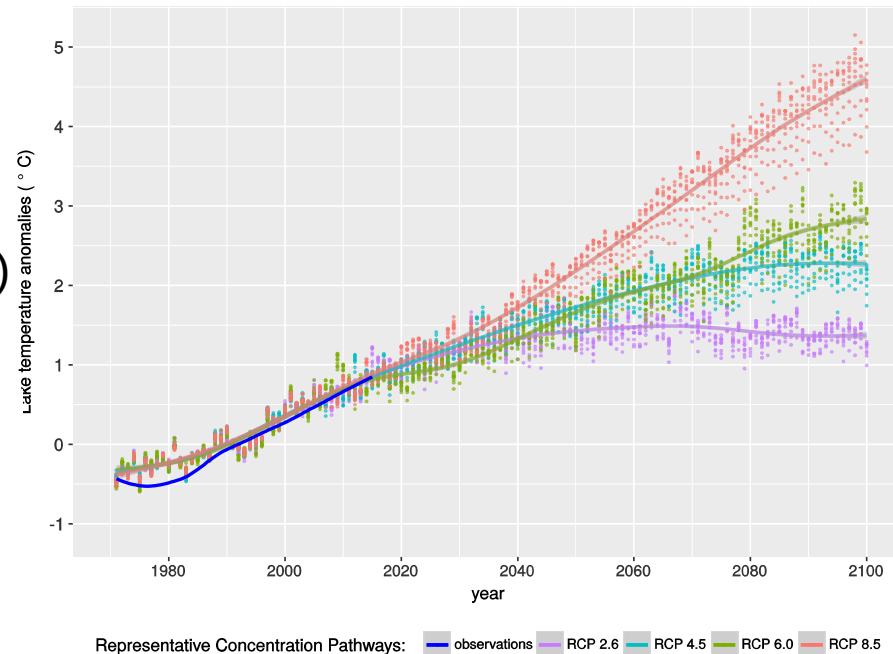
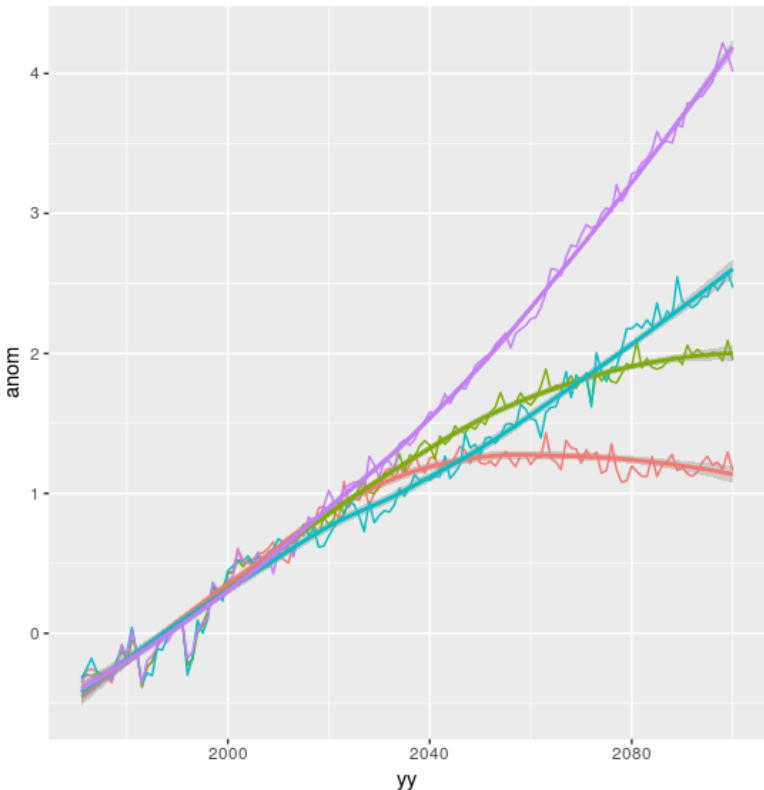


# Dlaczego nie bazowy ‘graphics’?



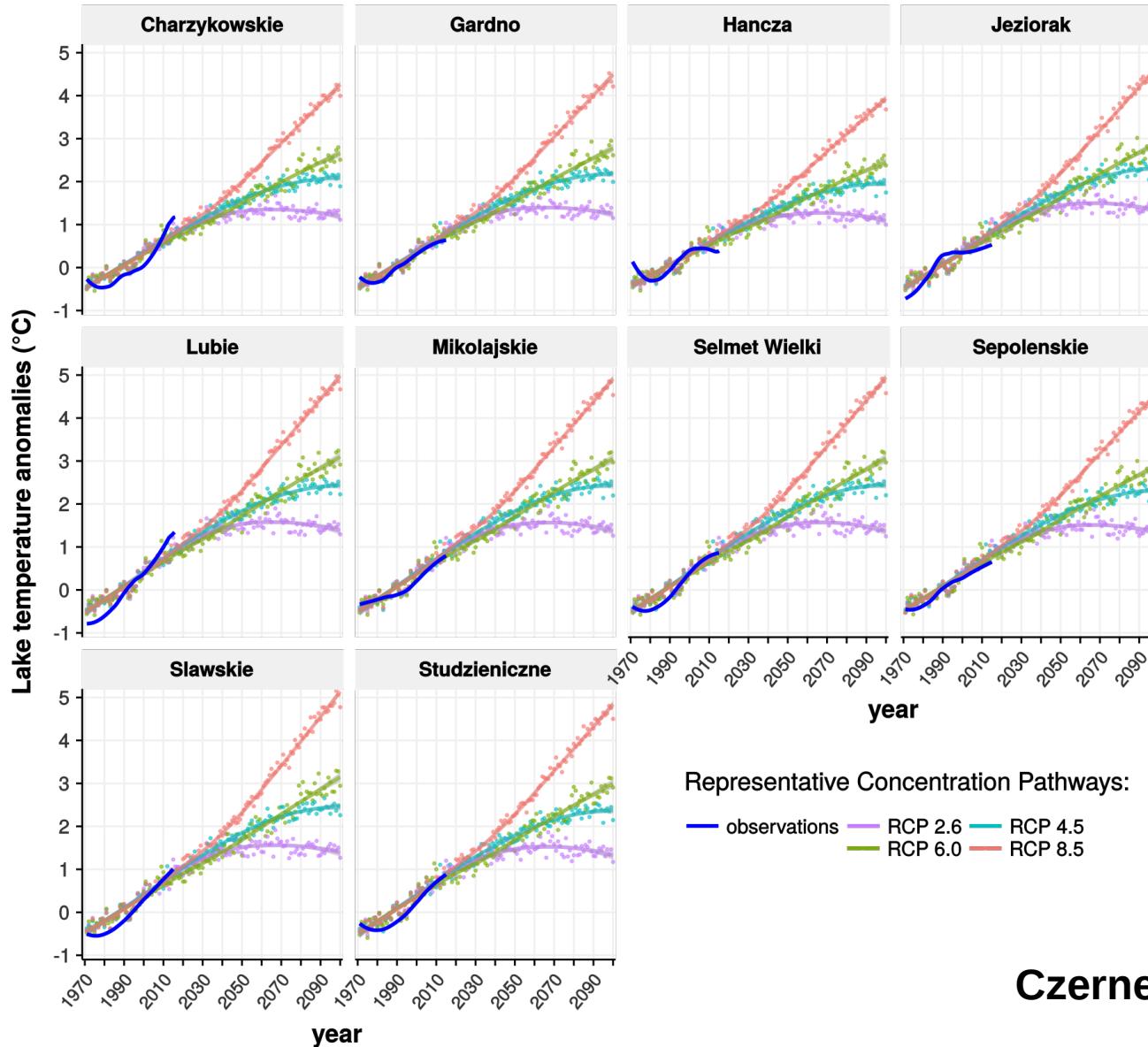
## GGPLOT2:

- styl (+biblioteki ze stylami)
- lepsze zarządzanie przestrzenią
- składnia kodu (grammar of graphics)
- prostota ‘dzielenia’ wykresów



- wiele wbudowanych elementów często używanych do wizualizacji  
(np. przedział ufności)

# Dlaczego nie bazowy ‘graphics’?



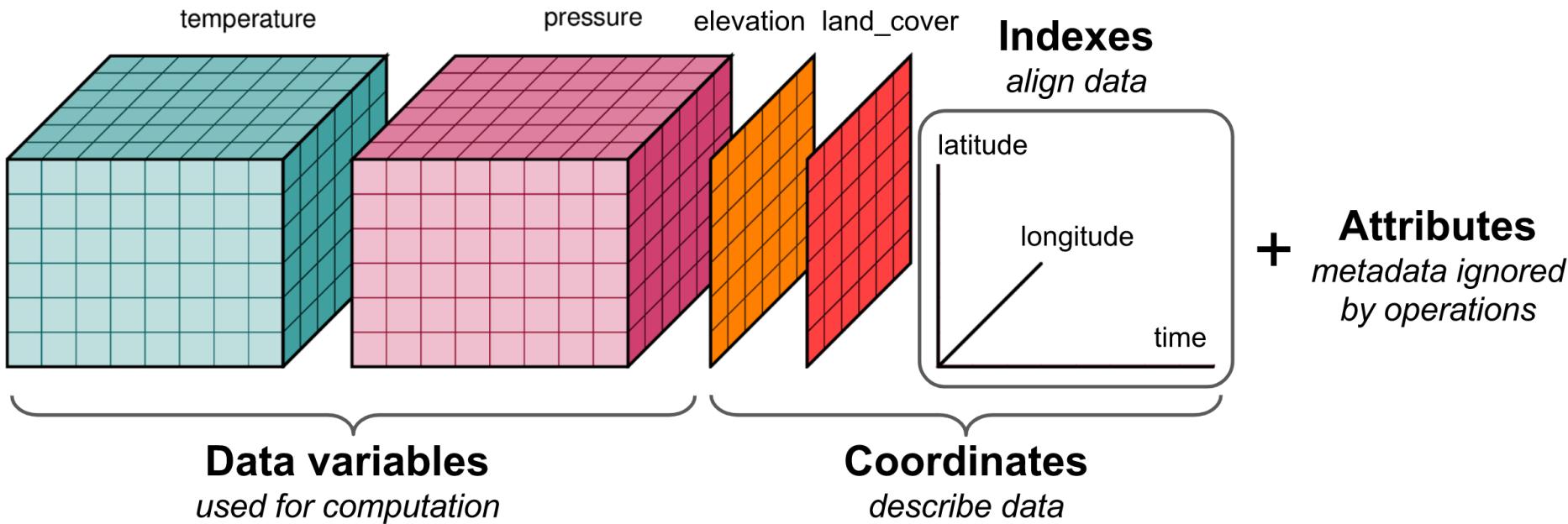
Podzielenie  
wykresów na  
10 paneli =  
2 „słowa” kodu



# R a dane przestrzenne

Obsługiwane praktycznie wszystkie standardy danych używane w naukach atmosferycznych (*bez radarów*):

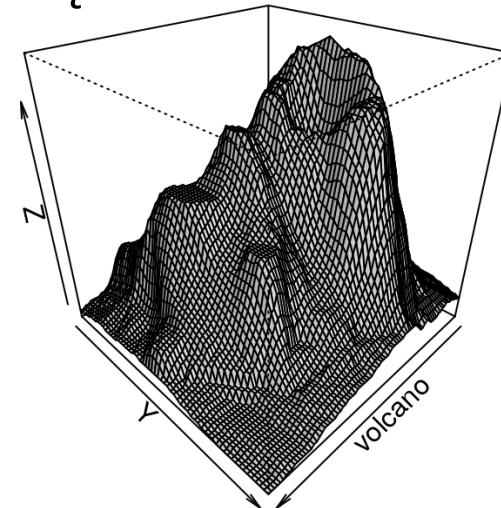
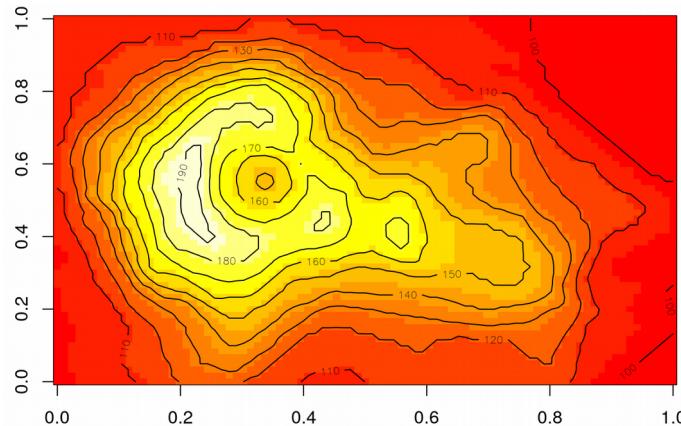
- NetCDF-3, NetCDF-4
- HDF (4-5)
- Grib (?)
- Popularne formaty danych rastrowych (GeoTIFF, ASCII grid ...)
- i wektorowych (ESRI Shapefile, GeoJSON,...)



# R a dane przestrzenne



Prosta mapa (w formie regularnej macierzy) możliwa do wizualizacji za pomocą wbudowanych narzędzi:



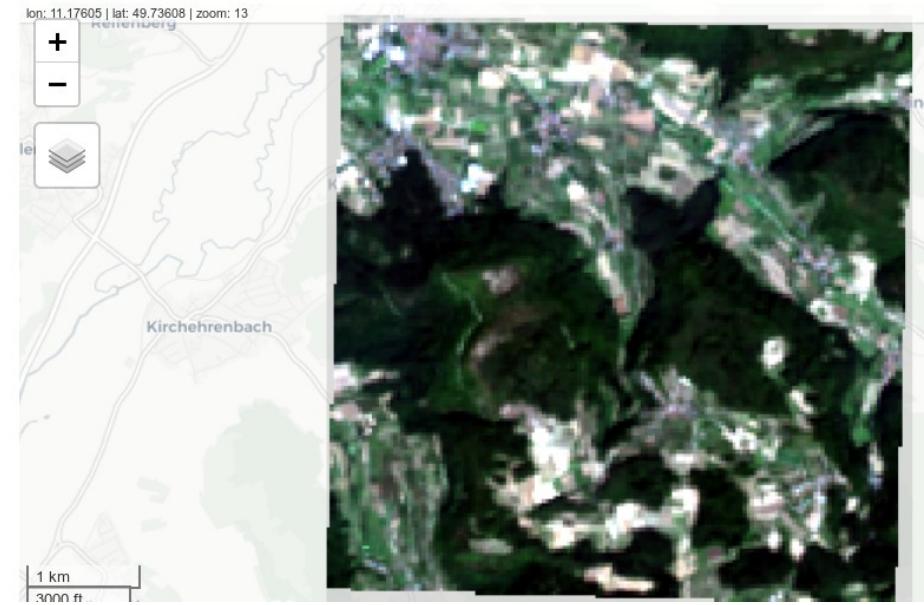
Nieco bardziej skomplikowane również nie powinny nastręczać trudności (np. granice krajów z pakietu 'mapdata')



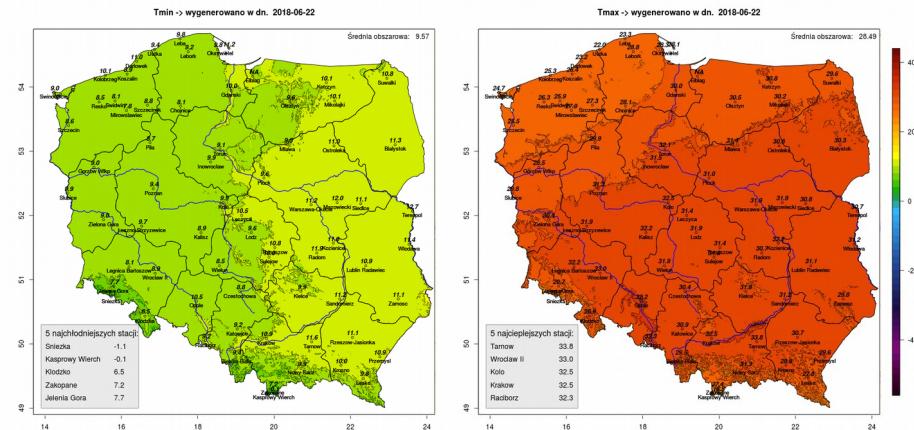
# R a dane przestrzenne



- Szerokie możliwości tworzenia map w panelach i map interaktywnych ('tmap', 'leaflet', 'mapview', 'sp')
- Im bardziej wchodzimy w świat standardów GIS (np. projekcje, rotowane siatki, itp.) tym bardziej rośnie poziom trudności (nieliniowo)
- ulubione narzędzie geostatystyczne wśród specjalistów rozwijających algorytmy analiz GIS
- Można natknąć się na niespójne lub niekonwertowalne pomiędzy sobą standardy → obecnie próba ujednolicenia

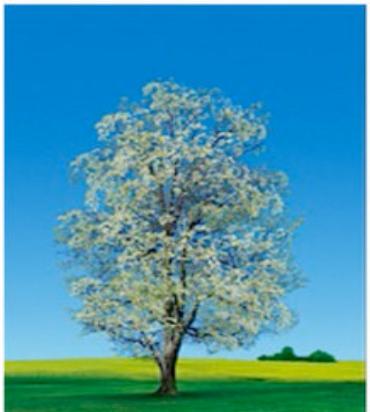


[https://r-spatial.github.io/mapview/articles/articles/mapview\\_05-extras.html#raster-only](https://r-spatial.github.io/mapview/articles/articles/mapview_05-extras.html#raster-only)



# Jak to działa w pRaktyce?

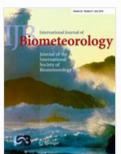
- przykład projektu interdyscyplinarnego



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

Springer Link

Search Home • Contact us • Login



[International Journal of Biometeorology](#)

[pp 1–13 | Cite as](#)

Machine learning modeling of plant phenology based on coupling satellite and gridded meteorological dataset

Authors

Authors and affiliations

Bartosz Czernecki , Jakub Nowosad, Katarzyna Jabłońska

Open Access | Original Paper

First Online: 11 April 2018

16  
Shares

500  
Downloads

[Download PDF](#)

[Cite article](#) ▾

[Share article](#)

Article

Abstract

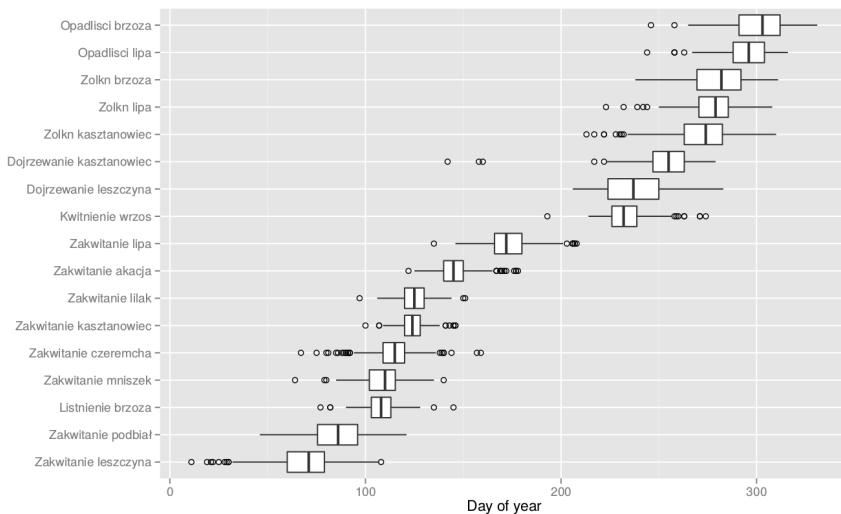
Introduction

Data and methods

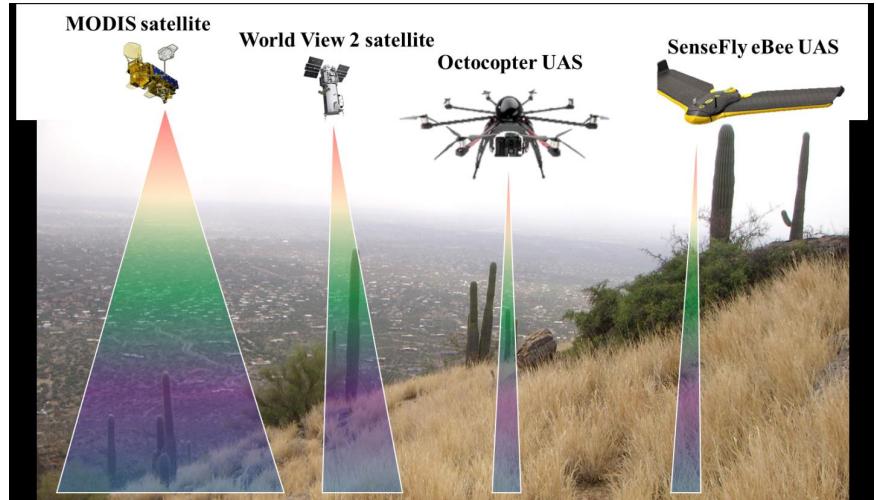
Results

# Współczesna fenologia

**Obserwacje naziemne – BBCH** (od 2007 roku w Polsce ponad 60 stacji)  
- wcześniejsze obarczone dużymi błędami obserwacyjnymi (obserwator, specyfika miejsca, rośliny, itp..)



**Obserwacje teledetekcyjne** – obserwacje bardziej całego ekosystemu niż gatunku – wykorzystanie odbiciowości roślinności i absorpcji w paśmie czerwonym i podczerwonym – tworzenie indeksów roślinnych (np. NDVI)



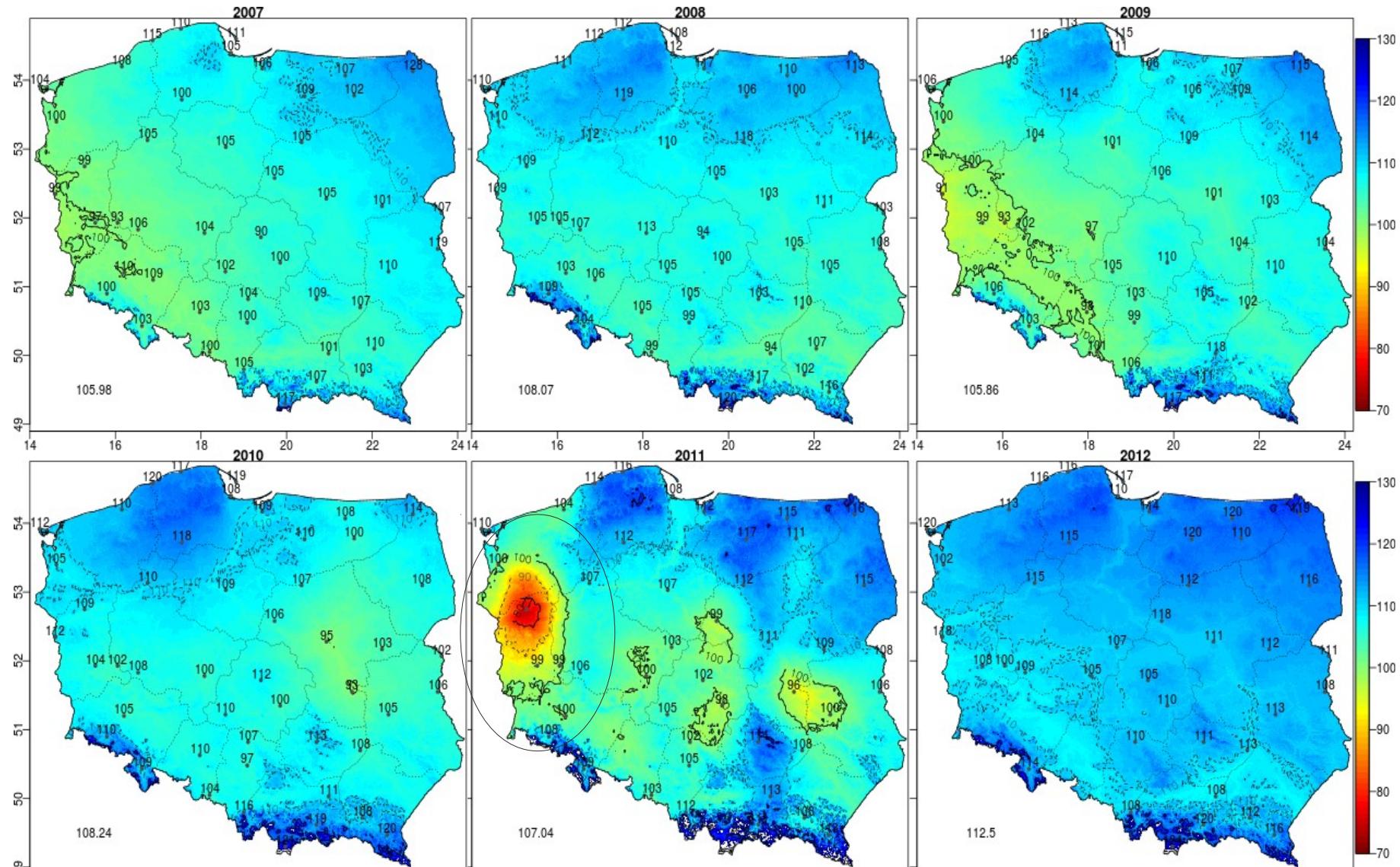
**Dane satelitarne** → komplementarne względem tradycyjnych obserwacji fenologicznych → informacje ciągłe w czasie i przestrzeni, niezależne od obserwatora; często „zaszumione”, są tylko aproksymacją prawdziwego stanu fizjologicznego rośliny

# Dane (2007-2014)

1. **Produkty satelitarne → MODIS level-3 vegetation products:**
  - pliki *Modis HDF* (*raster, sp, rgdal, modiscloud, maps, mapdata*)
    - Produkty „wegetacyjne”: (NDVI, EVI, LAI, fraction of photosynthetically active radiation) → automatyzacja pobierania + duże pliki!
    - Interactive Multisensor Snow and Ice Mapping System (IMS)
    - Operacje GIS: łączenie, przycinanie, reprojekcja + pętle
2. **Przetworzone zgridowane dane meteorologiczne (ECA&D)**
  - *NetCDF4* (*esd, ncdf4, polpred*)
    - cumulative growing degree days (GDD), cumulative growing precipitation days (GPD), średnie i sumy miesięczne, ...
3. **Dane przestrzenne** (longitude, latitude, altitude, distance to Baltic Sea, etc.) (*gstat, rgdal, sp, maptools, raster, verification*)
4. **Dane obserwacyjne (Excel)**

# Czyszczenie danych – błędne obserwacje

(automatyzacja wizualizacji (8 lat\*16 faz → decyzja eksperta)  
gstat:: Regression kriging with external drift



# Metody uczenia maszynowego (caret) → jedna z mocniej rozwijanych stron R (caret, H2O, keras...)

Kilka metod testowanych i ewaluowanych względem danych obserwacyjnych dla początku wystąpienia fenofaz:

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

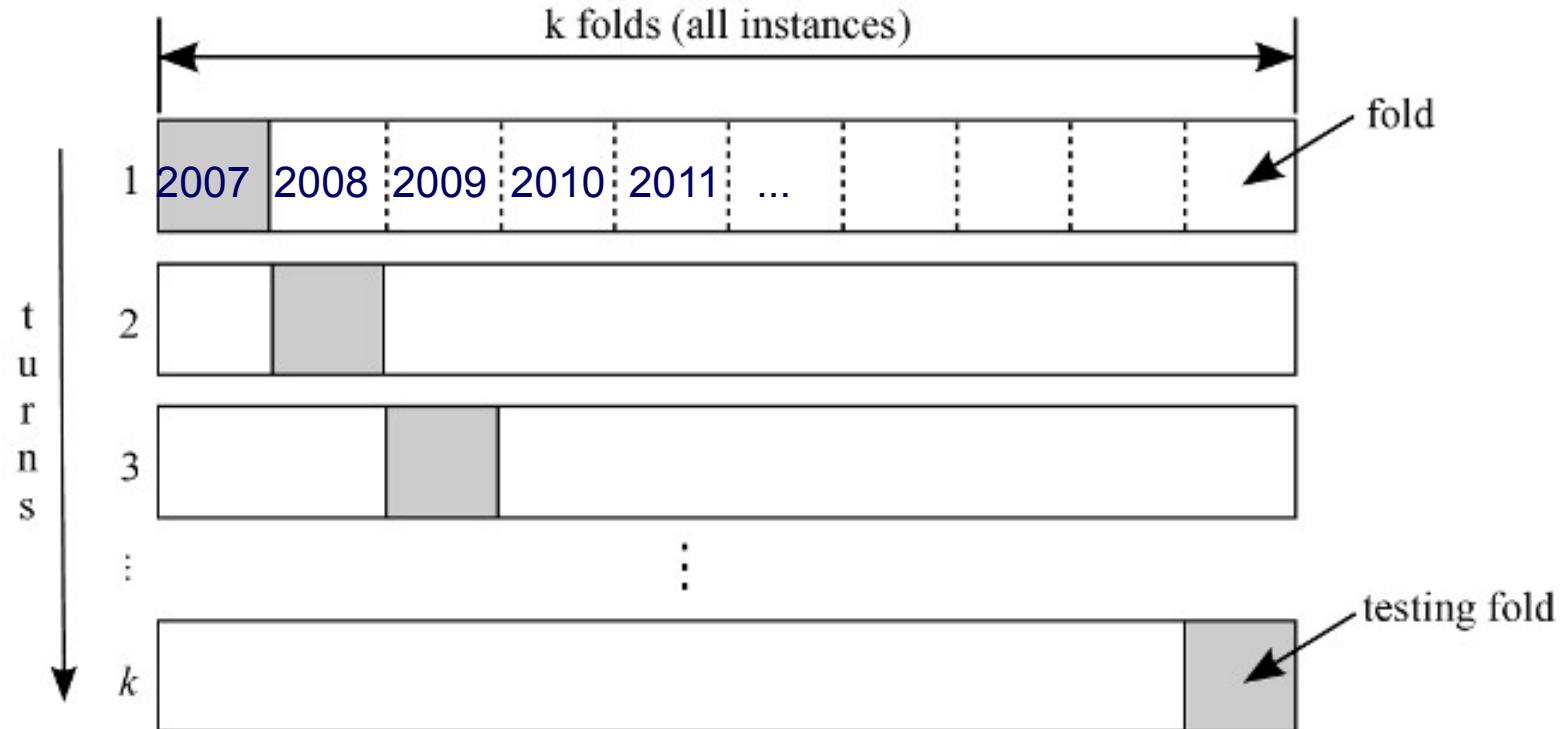
Rozbiecie potencjalnych predyktorów na kilka podgrup (w celu ustalenia które z nich są istotne):

1. Tylko dane meteo i dane lokalizacyjne
2. Dane satelitarne MODIS i dostępne indeksy roślinne
3. Meteo + satelita, ale przepuszczone wstępnie przez „algorytm szukający” czy jest w danych coś więcej niż szum (Boruta , Kursa 2010)
4. Wszystko co dostępne, bez żadnej obróbki

# Kroswalidacja

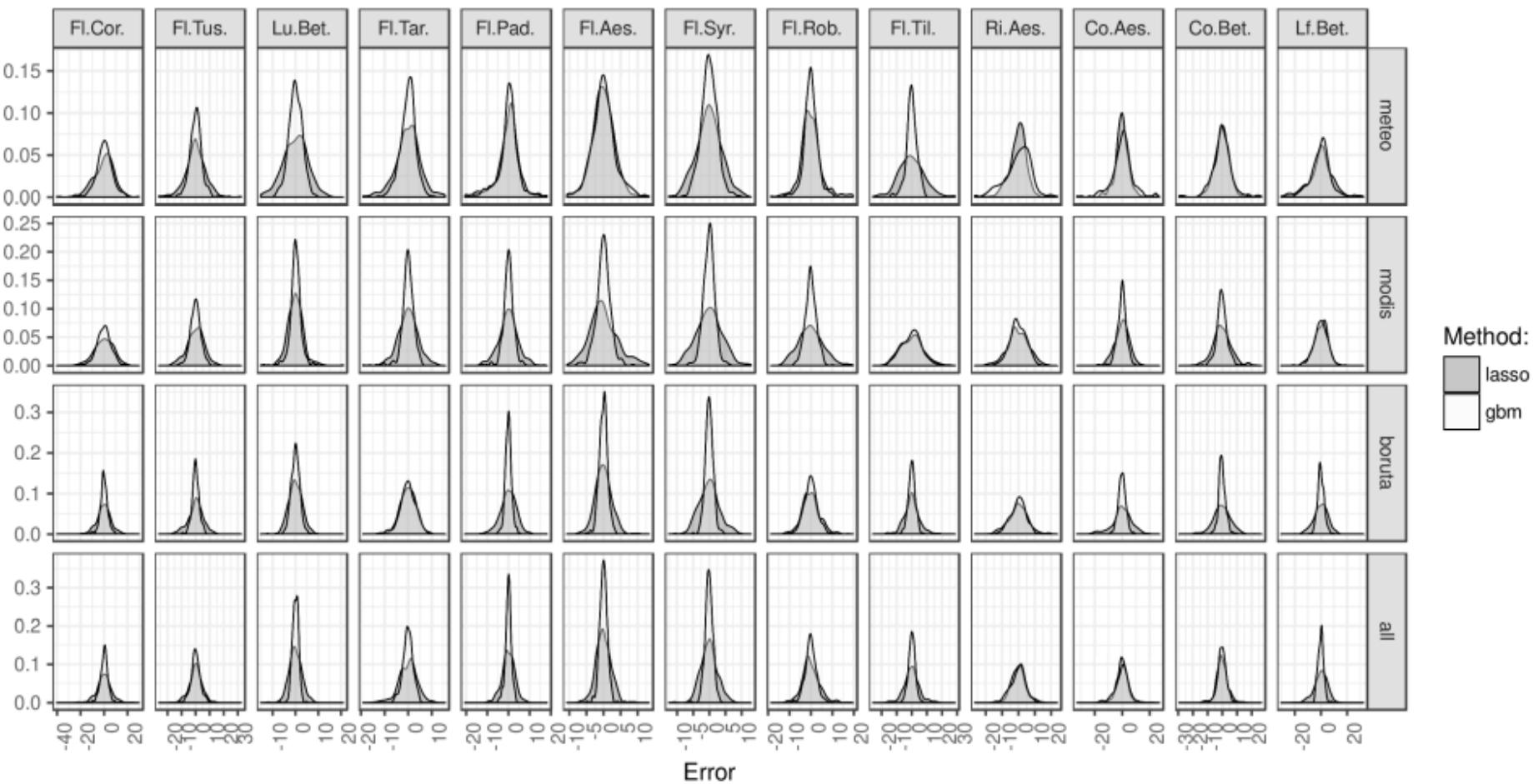
cross-walidacja *k-fold* – próba „nieprzeuczenia” modelu

Wbudowane opcja dostępne w oprogramowaniach statystycznych często prowadzą do przeuczenia → rozdzielenie ręczne na lata



# Przykładowe wyniki

- Średni błąd (RMSE) 6.3 dnia (min. 3.5 – max. 10)
- stopień zaawansowania modelu nie tak istotny jak dane wejściowe



# Github

## - Rozwijanie kodu i platforma współpracy

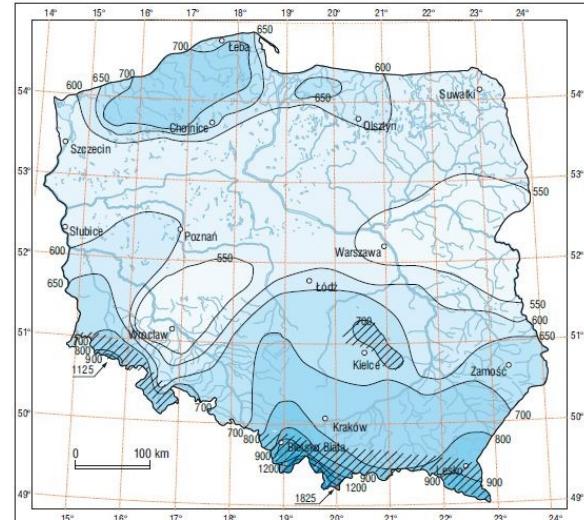


Activities Firefox Web Browser Sun Jun 24, 12:04 bczernecki/pm: Particulate Matter - Poland - Mozilla Firefox Machine learning modeling of pl Machine learning modeling of pl bczernecki/pm: Particulate M + GitHub, Inc. (US) https://github.com/bczernecki/pm ... Search ... + 1 Star 0 Fork 0 bczernecki / pm Private Unwatch 1 Star 0 Fork 0 Pull requests Issues Marketplace Explore Search or jump to... Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings Particulate Matter - Poland Edit Add topics 25 commits 1 branch 0 releases 1 contributor Branch: master New pull request Create new file Upload files Find file Clone or download bczernecki srednie roczne Latest commit 9d94454 on Aug 8, 2017 .Rproj.user srednie roczne 11 months ago R scatterploty 11 months ago data srednie roczne 11 months ago figs scatterploty 11 months ago man Initial commit a year ago .Rbuildignore Initial commit a year ago .Rhistory srednie roczne 11 months ago .gitignore zalegle 11 months ago DESCRIPTION Initial commit a year ago NAMESPACE Initial commit a year ago README.md baza danych dobowa - calosc a year ago

# Rzeczy których nie warto robić w



- wprowadzanie i edycja danych
- pojedyncze, mocno „customizowalne” mapy (np. do atlasów)
- edycja detali graficznych
- obliczenia numeryczne i inne w których ważna jest wydajność obliczeń (Fortran/C/C++/Java)





# R czy Python? w naukach atmosferycznych



## Python – język programistyczny ogólnego zastosowania

- łatwiejszy dla informatyków wchodzących w świat *data science* i *machine learning*
- bardziej stabilne rozwiązania
- lepszy do dużych projektów

## R:

- bardziej innowacyjne rozwiązania (i dostępne szybciej)
- bardziej popularny wśród *data scientistów*
- lepszy do obliczeń „na szybko”
- Shiny

Dziękuję za  
uwagę



Computations were carried out in the  
Poznań Centre for Networking and  
Supercomputing (Grant No. 331)