Guide for mini-projects

Here, you'll find links to data that can be used to work during the mini-projects, as well as some example research questions that you can explore using this data. Feel free to choose a dataset and explore it! All suggested research questions should be adapted to low and high coding skills (some might require a little help to shape the data).

We highly recommend that people interested in a research question work together in groups.

Automatic data and corresponding manual timeseries

Data

- Switzerland (Poleno and Hirst):
    - Timeseries:
        - Manual: https://www.meteoswiss.admin.ch/services-and-publications/applications/ext/download-data-without-coding-skills.html
        - API: https://opendatadocs.meteoswiss.ch/a-data-groundbased/a7-pollen-stations
    - /!\ hourly data since 2023: automatic. Check ogd-pollen_meta_datainventory.csv file.

Research questions

- Is a shift in season observed over the long term? Can it be identified over the short history of automatic data?
    - Choose one taxa over several years and explore seasonal shifts across time / countries (mixed Manual / automatic data).
- Long-term exploration of correlation between pollen time series and weather.

Literature

- https://pubmed.ncbi.nlm.nih.gov/35387022/
- https://link.springer.com/article/10.1007/s40629-026-00374-y
- https://doi.org/10.1007/s00484-024-02796-1
- https://doi.org/10.1038/s41467-022-28764-0

Diurnal cycles in automatic hourly pollen data

Data

Hund BAA500 and Poleno

- EBAS (see lectures to use the API or download manually from EBAS)

Research questions

- (Low to high coding skills) Use hourly pollen concentrations for one or several taxa and explore day / night cycles evolution. See if it correlates with some weather variables. Does it depend on the biogeographical region?

Imprecision in a classification algorithm: where can they come from?

Poleno

Data

- Automatic, with MeteoSwiss operational model 2025: https://drive.google.com/file/d/1_lCwhxJTlaAuhGj2pnOv0twaKMRf-1fU/view?usp=drive_link
- Manual, MeteoSwiss: https://drive.google.com/file/d/145Zm8vzr3mMPvr0Vsj1VAAq8SUsF5FpG/view?usp=drive_link

- Meteorological data from Sonnblick (automatic and manual): https://data.hub.geosphere.at/en/
- Other measurement data from Sonnblick: https://data.sonnblick.net/index

Research questions

- Using the predictions of a model and corresponding manual timeseries, explore the model scaling factors, confidence thresholds etc. for one or two taxa. Can you identify the best Scaling factors and Confidence Thresholds?
- Meteorological condition at Sonnblick impacts the classification. What condition? How and why? Can you propose some ways to improve the bioaerosol classification?

Literature

- https://amt.copernicus.org/articles/19/3427/2026/

Munich 2021 Intercomparison (March-Jully 2021)

Data

https://docs.google.com/spreadsheets/d/1XMt7A7kMRi3S3k0CnLXmLsE49XNcddcF/edit?gid=800918148#gid=800918148

- 9 taxa (Alnus, Betula, Corylus, Fraxinus, Pinus, Poaceae, Quercus, Taxus, Urtica) + total counts
- 19 combinations (device + algo) + Averaged Hirst manual timeseries

Research questions

- Choose one / two taxa, one / two combinations, and explore day/night cycle patterns
- Choose one / two taxa, one / two combinations, and explore correlation with meteorological variables. Data from https://meteostat.net/en/place/de/munich?s=10865&t=2021-03-01/2021-07-31. The csv can be found here: https://drive.google.com/file/d/1PsaxWHiIMAyEWChsV838bjzpbGMVdOna/view?usp=drive_link
- Anything you might find interesting (comparing models / instruments / combination of model and device concentrations for on eor several taxa)

Literature

- https://www.sciencedirect.com/science/article/pii/S0048969722083243?via%3Dihub

Explore size and distribution of the training datasets and develop an automatic cleaning tool

Bioaerosol forms and shapes can change over season / years / geographical region. Depending on when training datasets are collected and used to create a sample, the distribution of size, shape etc can vary.

Poleno

Data Training

- *Corylus* – Switzerland – 2026 (not cleaned): https://drive.google.com/file/d/1yrnY5RXrcUlrUJ-9BxHSEV2Xsk1T7naN/view?usp=drive_link
- *Corylus* – Switzerland – 2023 (cleaned): https://drive.google.com/file/d/1Npy9XSyWExFJupHqGrVIOZI1mqs3P_Pl/view?usp=sharing
- *Alnus* – Switzerland – 2023 (cleaned): https://drive.google.com/file/d/1VWWyx2iJDOOM_mXX4sR1IO4lZZrfooIw/view?usp=sharing
- *Alnus* – Switzerland – 2026 (not cleaned): https://drive.google.com/file/d/1T4qfYCDFLOhmv1jyDxU_Ei2lkzuka8Fn/view?usp=drive_link
- Metadata training datasets cleaned : https://docs.google.com/spreadsheets/d/1E2kmUQg04iBXu5h7odLfxGP8K6zlvgl-/edit?usp=drive_link&ouid=110945135499934629859&rtpof=true&sd=true
- SYLVA datasets : raw data can be downloaded, metadata must be extracted https://data.sylva.bioaerosol.eu/workspace/referencedatasets/poleno/

Operational data

- *Alnus* – Switzerland – 2026: https://drive.google.com/file/d/1jbMfuvcZbRTaoh1QfjPjRZ-YbO6JO10F/view?usp=drive_link
- *Corylus* – Switzerland – 2026: https://drive.google.com/file/d/1Kul-Nlf28xKImmnSwnXYhoOqWYqsuBOZ/view?usp=drive_link

Can you use this to conceptualize / develop an automatic cleaning tool?

Hund BAA500

Data Training

- SYLVA datasets : raw data can be downloaded, metadata must be extracted: https://data.sylva.bioaerosol.eu/referenceDatasets

Research question

- How do size / fluorescence / area etc distribution vary between datasets of the same taxa, across the season or across biogeographical regions?
- How do size / fluorescence / area etc distribution vary between training datasets and operational data?

Literature

- https://www.sciencedirect.com/science/article/abs/pii/S0034666724001994 (needs an institutional access)
- https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.70070
- https://www.mdpi.com/2037-0164/14/3/59 : Illustration of pollen features variability across various populations of the same species

Exploring a descriptive analysis of training datasets from reference fungal spores

Data

Poleno Jupiter datasets from Sciensano (https://drive.google.com/drive/folders/1MndRMJ4eiRS24iwBSIaIOIkTAXYxDTlx?usp=drive_link):

1) 17 reference species, including several *Alternaria* spp., *Cladosporium* spp., and *Aspergillus*/*Penicillium* spp.
2) Same batch of spores measured at different time points.

Research questions

1) Are the holography/fluorescence-derived features enough to differentiate several species of the same genus? Which ones matter the most depending on the genus?
2) How spore aggregates and chains affect the average features of uncleaned datasets? Is it possible to isolate *Alternaria* chains made of 2, 3 or more spores? Is there a linear correlation between size properties (chain length) and the fluorescence signal?
3) Are hyphae particles (=mycelium fragments) easy to clean from training datasets? What are the most useful holography/fluorescence-derived features in order to discriminate it from spores?
4) Does long-term natural aging of reference fungal spores, stored at room temperature in the dark, may affect their holography/fluorescence-derived features?

Literature

- Pogner C. et al., *Aerosol Science and Technology* 2024 (https://doi.org/10.1080/02786826.2024.2338544)
- Bruffaerts N. et al., *Aerobiologia* 2025 (https://doi.org/10.1007/s10453-025-09864-y)

Evaluation of the CAMS ensemble members pollen forecasts and long-term assessment of pollen trends for allergenic trees

Data *(the choice of dataset will depend on the research question selected)*

- EBAS pollen concentrations (see lectures to use the API or download manually)
- CAMS ensemble members pollen forecasts:
    - CAMS Regional (11 individual models and their ensemble): https://confluence.ecmwf.int/spaces/CKB/pages/202173092/CAMS+Regional+European+air+quality+analysis+and+forecast+data+documentation
    - Atmosphere Data Store (ADS): https://confluence.ecmwf.int/spaces/CKB/pages/202163324/Atmosphere+Data+Store+ADS+documentation
    - Download of pollen forecasts (CDS API requests) from ADS: https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-forecasts?tab=form (…check “Download" tab)
- SILAM pollen forecasts (http://silam.fmi.fi/pollen.html):
    - Europe (resolution ~10 km), from SILAM thredds server_--->> https://thredds.silam.fmi.fi/thredds/catalog/silam_europe_pollen_v6_1/catalog.html
    - Scandinavia, high resolution (~2.5 km), from SILAM thredds server_--->> https://thredds.silam.fmi.fi/thredds/catalog/silam_regional_pollen_v6_1/catalog.html

    - Finland, high resolution (~800 m), from SILAM thredds server_--->> https://thredds.silam.fmi.fi/thredds/catalog/silam_hires_pollen_v6_1/catalog.html
- European Pollen Reanalysis (EPR):
    - metadata description --- >> https://doi.org/10.57707/FMI-B2SHARE.85841086F9DB46B882D750EAA9E42515
    - gridded data --->> https://european-pollen-reanalysis.lake.fmi.fi/index.html

Research questions *(preliminary, to be agreed with students)*

- How accurate are CAMS ensemble members forecasts in terms of pollen season timing and season intensity? (Pollen species: alder, birch, grass, olive, ragweed, mugwort; use pollen observations from EBAS and model extracted timeseries: 11 models + ENSEMBLE; pollen season - 2025)
- Estimate the multidecadal trends of pollen season timing and intensity at several European locations with different bioclimatic characteristics (based on EPR hourly surface concentrations; species – alder, birch, olive; 1980-2022)
- (*) Estimate the deterioration over time (if any) of the predictive skills of the CAMS models.
- Student's suggestions (?)

Literature

Colette et al. Copernicus Atmosphere Monitoring Service – Regional Air Quality Production System v1.0, Geosci. Model Dev., 18, 6835–6883, https://doi.org/10.5194/gmd-18-6835-2025, 2025.

Sofiev, M., Palamarchuk, J., Kouznetsov, R., et al., 2024. European pollen reanalysis, 1980–2022, for alder, birch, and olive. Sci Data 11, 1082. https://doi.org/10.1038/s41597-024-03686-2

Sofiev, M., Palamarchuk, J., Kouznetsov, R., Gauss, M., 2025. Annual report on the evaluation of the CAMS Regional pollen production (daily forecasts) (https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/EQC-regional/Pollen/CAMS283bis_2024SC1_D83bis.1.4.1-2024_pollenEQC_v2.pdf)

NB! mini-project guidance and assistance will be provided remotely.

Other open datasets to go further

- https://www.kaggle.com/datasets/nataliakhanzhina/pollen20ldet : This Kaggle dataset contains 13-20 plant species, 2413 images containing 7745 single pollen grains from Perm Krai, Russia. Pictures are annotated, high-quality, and particles are treated with fuchsin. It should be possible to train a classifier on these data and develop a model that automatically recognizes pollen types from slides from manual pollen monitoring devices. [Higher coding skills needed]