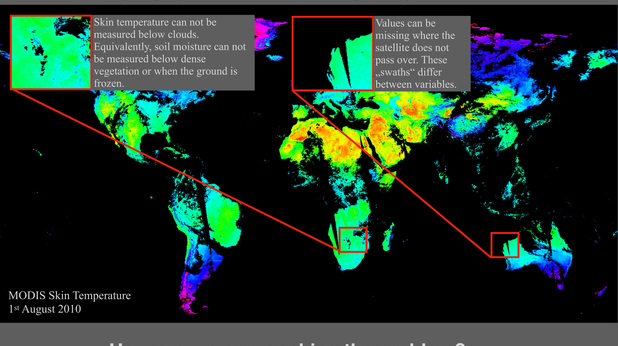
Data Science for Gapfilling Complex Earth Observations

authors: Verena Bessenbacher
Lukas Gudmundsson
Sonia I. Seneviratne

Verena Bessenbacher @climachine
7th May, EGU General Assembly
Land Climate Dynamics, ETH Zürich
verena.bessenbacher@env.ethz.ch

Why are we doing this?

Missing values in remote sensing datasets are ubiquitous, complex and unavoidable. Combining several remotely sensed variables is challenging, since missingness patterns are not the same and only taking the points where all variables are observed ignores a major fraction of the observed data.



How are we approaching the problem?

Gapfilling is common practice in the geosciences, but usually focuses on one variable only. This is often done with the help of other variables, as well as with spatial *or* temporal interpolation.

we attempt ***multivariate,*** i.e. using more than one variable
mutual, i.e. gapfilling each variable with the help of all others
multiple imputation, i.e. producing several estimates for each missing value

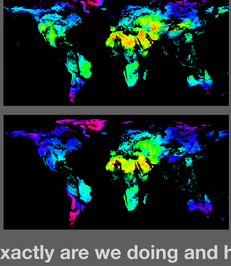
incorporating:
- covariance structure between variables
- spatial correlation among variables
- temporal autocorrelation among variables

How are we testing the gapfilling merit, since we cannot know what the „original“ values would have been?

Our algorithm „testbed“ — The perfect dataset approach

We use reanalysis data from the ERA5-Project, which provide gap-free estimates of essential climate variables. We employ a "perfect dataset approach", where we assume the reanalysis data to be the "true" state of the land-climate interactions and introduce artificial missing values that are subsequently imputed.

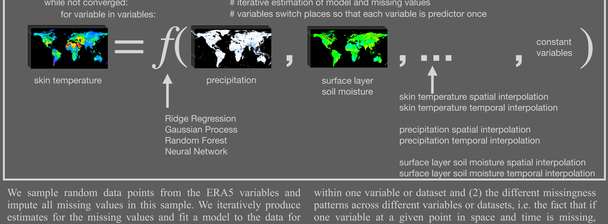
The analysis is confined to ***daily, global land-only ERA5 data from 2003 to 2012,*** at 0.25° resolution. Only ERA5 variables are considered that can be matched with available satellite remote sensing products: MODIS Aqua ***skin temperature***, GPM ***precipitation*** and ESA-CCI ***surface layer soil moisture*** of the uppermost soil layer. Additionally we assume constant maps of vegetation type, vegetation cover, topographic height and topographic complexity to be known and gap-free.



MODIS
Skin Temperature
1st August 2010

ERA5 Reanalysis
1st August 2010
with MODIS
missingness pattern

The gritty details: What exactly are we doing and how do we incorporate spatiotemporal context and covariance?



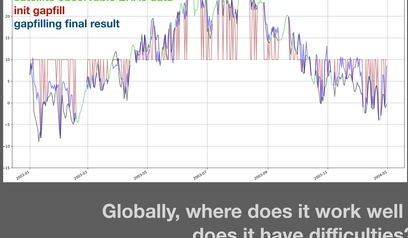
We sample random data points from the ERA5 variables and impute all missing values in this sample. We iteratively produce estimates for the missing values and fit a model to the data for each variable, in an expectation-maximisation alike fashion. This procedure is repeated until the estimates for the missing data points converge.

The method harnesses the highly-structured nature of gridded covarying observation datasets within the flexible function learning toolbox of data-driven approaches. The imputation utilises (1) the temporal autocorrelation and spatial neighborhood within one variable or dataset and (2) the different missingness patterns across different variables or datasets, i.e. the fact that if one variable at a given point in space and time is missing, another covarying variable might be observed and their local covariance could be learned.

A simple ridge regression is already able to outperform simple "ad-hoc" gapfilling procedures on high resolution daily satellite data, however, we are working on additionally testing a nonlinear method (Gaussian Process, Random Forest and Neural Networks).

How does this work exemplarily at one point?

ERA5 skin temperature in Basel year 2003



In **black**, the ERA5 skin temperature is plotted. In green, the same data is used, but only the values that would have been observed by a satellite are shown. Days where Basel was overcast with clouds cannot be seen by the satellite, for example much of December 2003.

In red, the initial gap filling procedure is shown. We use the temporal mean.

In blue, the final result is shown. The iterative procedure reduces the bias and increases the correlation of original data and gapfilled values by incorporating information
- from the other variables (soil moisture and precipitation)
- from the neighboring grid points
- from the day before and after

Globally, where does it work well and where does it have difficulties?

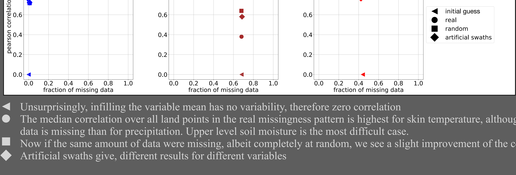
0 1

The Pearson correlation is high where much data can be observed, and low where data is missing a lot of times. However, correlation is never negative, showing that the gap filling procedure applied indeed improves the estimates for the missing values.

ESA-CCI soil moisture has an impressive 68% of missing values. Soil moisture measurements are therefore exposing a non-trivial missingness pattern with a comparatively high fraction of missingness among remote sensing products, making it especially challenging for imputation.

How good would the gapfilling work, if the satellite data would be missing completely at random? And what about if we only had swaths?

Because random missingness and missingness according to „artificial swaths“ patterns is an easier pattern to learn for the gapfilling, we want to observe how our gapfilling algorithm performs in this idealized experiments.



- Unsurprisingly, infilling the variable mean has no variability, therefore zero correlation
- The median correlation over all land points in the real missingness pattern is highest for skin temperature, although more data is missing than for precipitation. Upper level soil moisture is the most difficult case.
- Now if the same amount of data were missing, albeit completely at random, we see a slight improvement of the correlation
- Artificial swaths give, different results for different variables

What are our plans for the future?

- consider another initial gap fill, using climatology
- add non-linear method for gapfilling
- add net radiation as a variable
- check physical consistency of imputed values (e.g. soil gets wet when it rains)
- apply on real observations eventually!

Verena Bessenbacher @climachine
7th May, EGU General Assembly
Land Climate Dynamics, ETH Zürich
verena.bessenbacher@env.ethz.ch