Abreschviller 2007

Overview

Our data was provided by Cyrille Rathberger. In this report, we are studying data from 2007 Abreschviller, Donon, Vosges Mountains, France. We are extracting information from two datasets:

- Cell count data for 15 trees sampled weekly during growth season.
- Meteorological data sampled every day of the year.

Our goal is to predict cell count data given meteorological data.

Our first step is to perform topological data analysis (TDA). This method enables to determine the *shape* of the data, hence helping to deduce its structure. It is particularly useful in exploratory data analysis on complex datasets were many of the variables are correlated with each other. It helps to determine essential, large-scale features in the data. An important attribute of this technique is robustness to noise, which is advantageous when dealing with biological data.

The data

Cell count data

```
#Cell count data for Abreshviller 2007
head(ABR_07)
```

```
##   Site Year    Species Tree Sample  DY RF CZ EZ WZ MZ   PR
## 1  ABR 2007 Abies alba   48      1  93  1  5  0  0  0 0.77
## 2  ABR 2007 Abies alba   48      1  93  2  6  0  0  0 0.77
## 3  ABR 2007 Abies alba   48      1  93  3  6  0  0  0 0.76
## 4  ABR 2007 Abies alba   48      2 100  1  4  0  0  0 0.44
## 5  ABR 2007 Abies alba   48      2 100  2  5  0  0  0 0.44
## 6  ABR 2007 Abies alba   48      2 100  3  6  0  0  0 0.43
```

```
#DY: Day of the year
#RF: Radial file
#CZ: n° of Cambial cells
#EZ: n° of Enlarging cells
#WZ: n° of Wall-thickening cells
#MZ: n° of Mature cells
#PR: Precision of sampling
```

Each tree was sampled 31 times during the year 2007 (if less we have added the missing samples with a cell count of NA). For each sample there are three measurements which are referred to as RF, corresponding to the radial files. In each measurement there is the count of cambial cells, enlarging cells, wall-thickening cells and mature cells (CZ, EZ, WZ, MZ resp.) in the corresponding radial file.

Meteorological data

```
#Meteorological data for Abreshviller 2007
head(meteo_ABR_07)
```

```
##   annee jour     vent pluie tsec hum      rgl tmin tmax
## 1  2007    1 4.368530  14.0  7.3  89 180.9282  4.6 10.8
## 2  2007    2 3.727789  10.6  2.8  97 146.6629  1.1  4.2
## 3  2007    3 2.902885   0.0  4.4  92 266.9465  2.5  6.9
## 4  2007    4 4.154965   5.8  5.8  88 142.6212  3.6  7.0
## 5  2007    5 3.584772   4.6  6.4  91 127.0658  5.3  8.2
## 6  2007    6 4.045815   4.2  8.6  93 175.3882  6.8 10.3
```

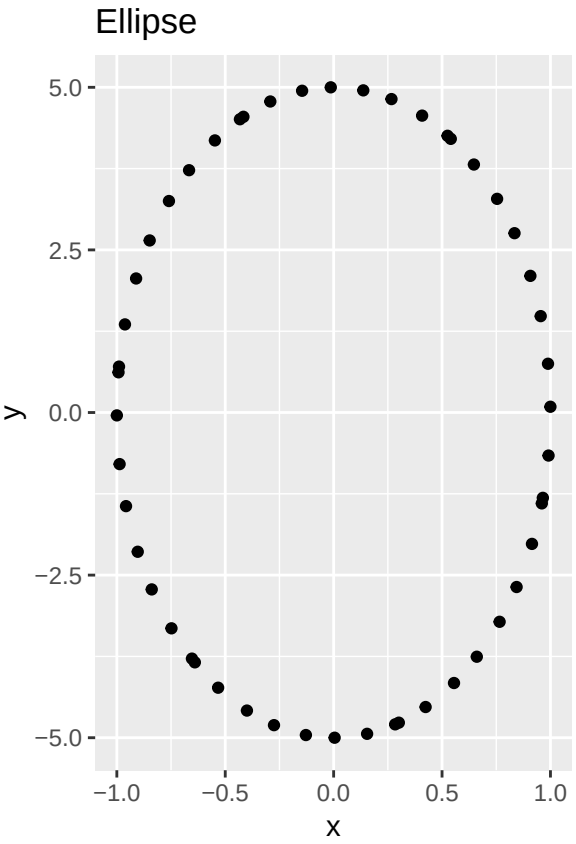
```
#annee: Year
#jour: Day of the year
#vent: Wind
#pluie: Precipitation
#tsec: Mean temperature
#hum: Humidity
#rgl: Solar radiation
#tmin: Minimum temperature
#tmax: Maximum temperature
```

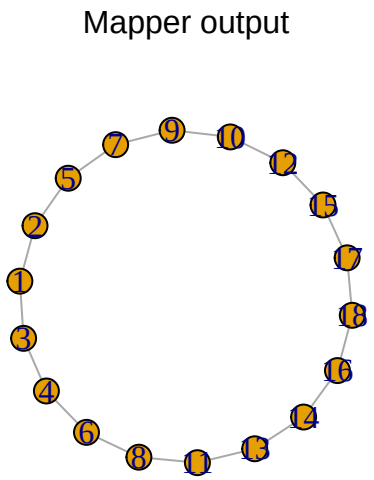
The Mapper algorithm

The TDA algorithm we use is called Mapper. This algorithm among other things inputs a data frame of numerical values and outputs a graph. The vertices in the graph correspond to clusters and edges link vertices when the corresponding clusters share points.

As an example, if we sample a dataset from an ellipse, on the left we plot the data, on the right we plot the output of the Mapper algorithm. Of course, if our data frame contains two variables, this is no use, but when it contains more this can be very informative.

```
ellipse = data.frame(x = cos(1:50), y= 5*sin(1:50))
```





It is difficult to explain the algorithm without using pictures nor math. An excellent explanation of it is 0:00-15:00 of https://www.youtube.com/watch?v=zDe72aINF2s. A more mathematical and in-depth explanation is 0:00-20:00 of https://www.youtube.com/watch?v=3Z73Wd2T1xE&t=936s.

To run Mapper we need to specify a way to visualise the data as input, which we call a *lens* or *filter function*. In practice a lens is a one or two number summary of each observation in the data. It is most useful to use a lens that summarises important features in the data. However, in exploratory data analysis we usually do not know what these may be. This is why it is important to run the algorithm many times with different lenses. These will enable to build incremental knowledge about the data even with poorly chosen lenses.

Methodology

Mapper gives information about the "shape" of data in a data frame. Since our meteorological data and cell count data are in two separate data frames, we must merge them into one to be able to apply Mapper.

We observe the following:

- Cell count data is sampled weekly throughout the growing season while meteorological data is sampled daily throughout the year.
- To be able to merge the two data frames, we need both data sets to be sampled at the same times.
- Ultimately, we would like to account for all the variability in the meteorological data in its influence on cell count data. Therefore, it would be counterproductive to reduce meteorological data to weekly samples during the growing season. Rather, it would be ideal to know the daily cell count data.

Our attempt is to approximate daily cell count data by linear interpolation on the available data.

We bear in mind the following:

- Linear interpolation of cell count data is a naive approximation that does not take meteorology into account. We intend to incorporate more sophisticated methods of interpolation, taking meteorology and biological factors into account, later in the project.
- We think that Mapper, which gives insights about large scale features in the data while being robust to noise, will be largely insensitive to the type of interpolation that we use.
- We approximate cell count data outside of the growing season by the values of the first or last sample. That is, we assume that there is no growth in trees outside of the sampling range.

In the process of implementation we have encountered missing values in the cell count data. Mapper cannot deal with NAs so it is important to replace them by numerical values.

We have removed the missing values using the following two-step algorithm:

1) If not all XZ of a certain sample for a certain tree are NA, replace XZ NAs by the mean of the non-NA XZ values.

2) If all XZ of a certain sample for a certain tree are NA, replace XZ NAs by linear interpolation on the previous and next available sample.

Finally, we obtained a 16425 observation data frame with daily meteorology and daily cell count data for all 15 trees.

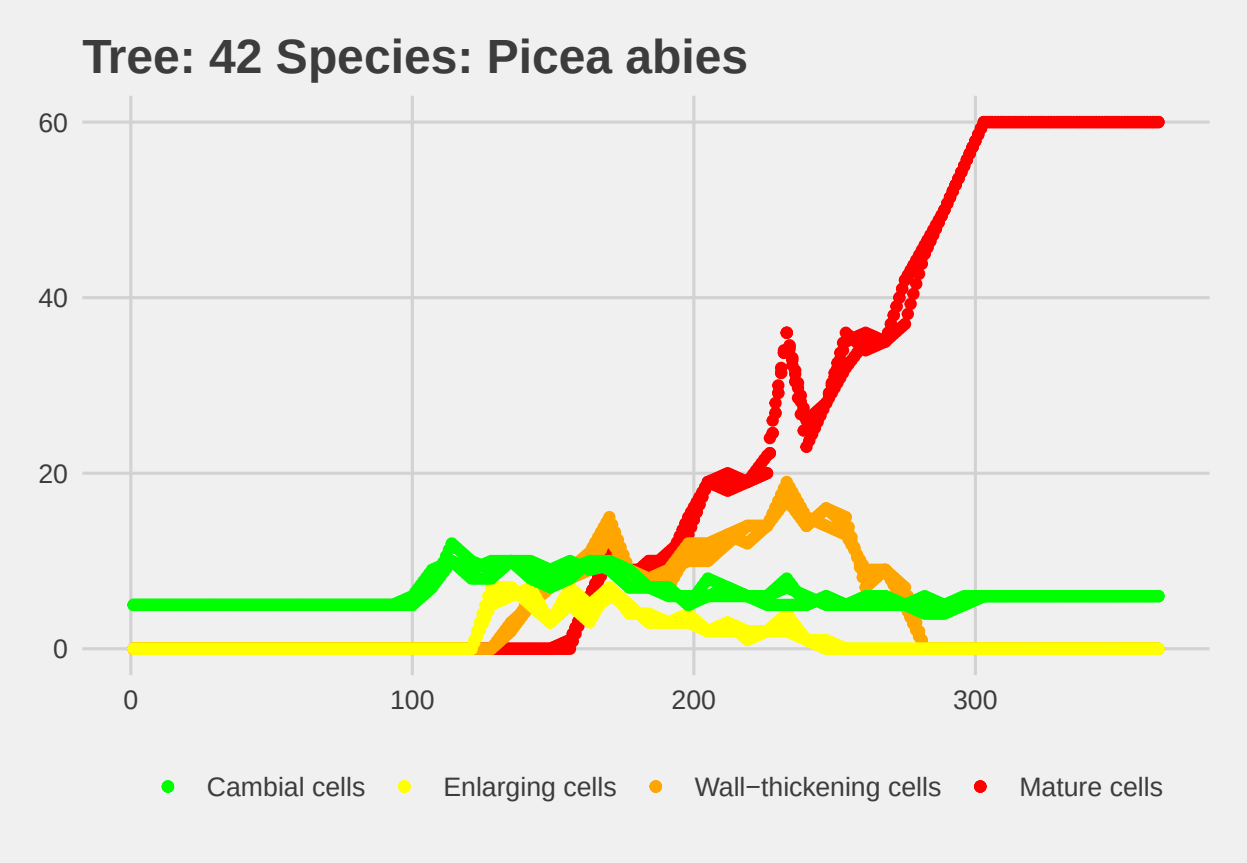
```
##      annee jour     vent pluie tsec hum      rgl tmin tmax Site
## 9760  2007  217 1.911441   0.2 19.4  74 2720.942 10.4 28.7  ABR
## 9761  2007  217 1.911441   0.2 19.4  74 2720.942 10.4 28.7  ABR
## 9762  2007  217 1.911441   0.2 19.4  74 2720.942 10.4 28.7  ABR
## 9763  2007  217 1.911441   0.2 19.4  74 2720.942 10.4 28.7  ABR
## 9764  2007  217 1.911441   0.2 19.4  74 2720.942 10.4 28.7  ABR
## 9765  2007  217 1.911441   0.2 19.4  74 2720.942 10.4 28.7  ABR
## 9766  2007  218 1.846738  14.0 19.9  84 2267.409 13.3 29.4  ABR
## 9767  2007  218 1.846738  14.0 19.9  84 2267.409 13.3 29.4  ABR
## 9768  2007  218 1.846738  14.0 19.9  84 2267.409 13.3 29.4  ABR
## 9769  2007  218 1.846738  14.0 19.9  84 2267.409 13.3 29.4  ABR
## 9770  2007  218 1.846738  14.0 19.9  84 2267.409 13.3 29.4  ABR
## 9771  2007  218 1.846738  14.0 19.9  84 2267.409 13.3 29.4  ABR
##               Species Tree Sample RF        CZ        EZ        WZ
## 9760       Abies alba   54      0  1  7.142857 2.8571429 18.428571
## 9761       Abies alba   54      0  2  7.142857 2.8571429 20.571429
## 9762       Abies alba   54      0  3  7.142857 2.2857143 18.714286
## 9763 Pinus sylvestris   55      0  1  5.000000 1.0000000  6.142857
## 9764 Pinus sylvestris   55      0  2  5.285714 1.0000000  6.142857
## 9765 Pinus sylvestris   55      0  3  5.142857 0.8571429  5.571429
## 9766      Picea abies   41      0  1 18.000000 5.7142857 42.714286
## 9767      Picea abies   41      0  2 17.142857 6.2857143 41.714286
## 9768      Picea abies   41      0  3 17.571429 6.7142857 41.285714
## 9769      Picea abies   42      0  1  6.142857 1.2857143 13.857143
## 9770      Picea abies   42      0  2  6.142857 2.1428571 12.142857
## 9771      Picea abies   42      0  3  6.000000 2.0000000 13.714286
##             MZ PR
## 9760 24.142857  0
## 9761 22.428571  0
## 9762 24.714286  0
## 9763  7.857143  0
## 9764  7.285714  0
## 9765  8.000000  0
## 9766 59.000000  0
## 9767 59.000000  0
```

```
## 9768 59.000000  0
## 9769 19.142857  0
## 9770 18.857143  0
## 9771 19.000000  0
```

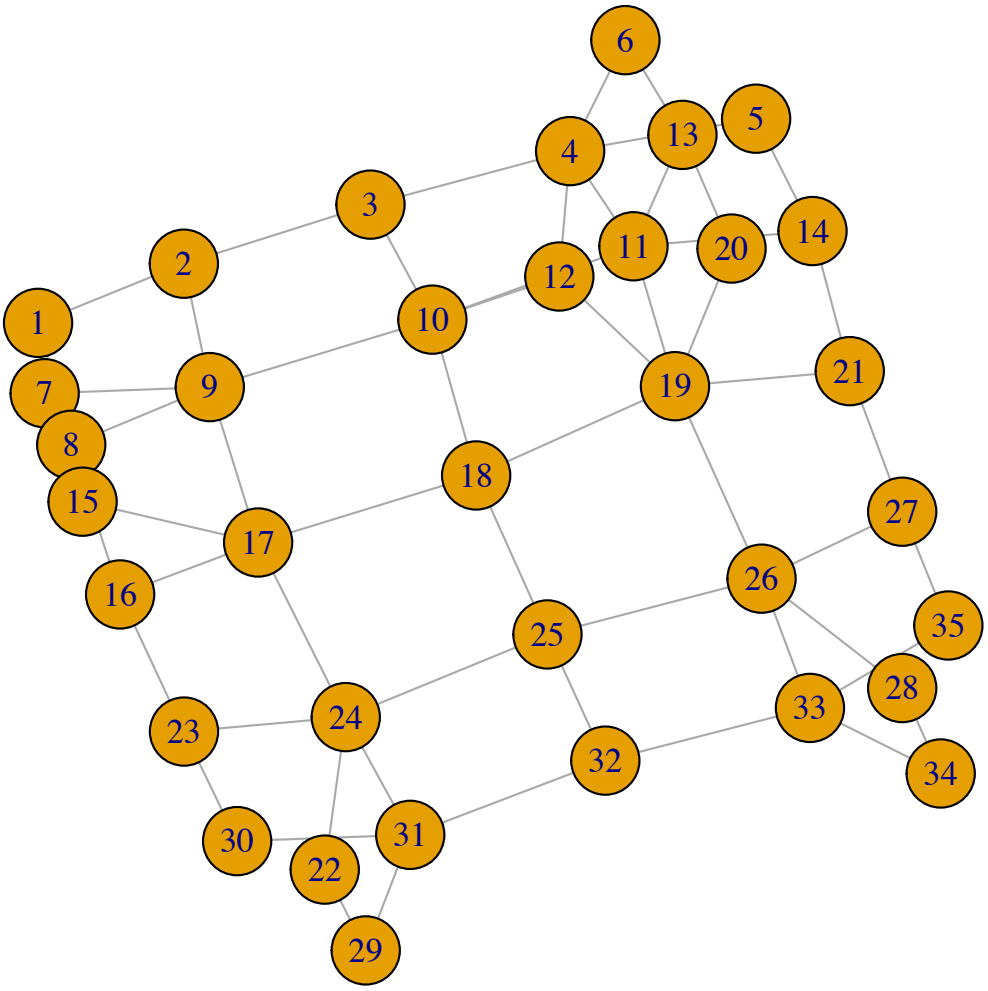
As you can see, the precision and sample variables are zero at these observations. This is because the cell count data at day 217 and 218 has been linearly interpolated from the information in the original data set. We did not interpolate the sample number (Sample) nor the sample precision (PR) because it does not make sense. We remove these two variables and the non-numerical variables (Site, Species) before applying Mapper.

Results

Thanks to linear interpolation we have an estimate cell-count data for every tree throughout the year. Tree 42 is particularly neat:



Our Mapper analyses are giving us interesting shapes. Here is one of them:



The lens we used in this case was a two number summary for each observation given by *(vent + pluie + hum + tsec, MZ)*.

Current and future research

Currently are working on the following:

- Meteorological data does not influence tree growth on the day. Rather, tree growth is a function of say, the mean of the meteorological data over the 20 days before. Thus, we are also testing Mapper on modified datasets that at each day have the mean of the meteorological data over d days before, instead of having meteorological data of that day. We are doing this for different values of d.
- Running Mapper with many different lenses to get more intuition about the shape of our datasets. This will guide further analyses.
- Understanding which observations in the data correspond to which node in each output of Mapper.

In the near future we hope to work on:

- Doing correlation analyses to better and better approximate a function that would predict tree growth (in terms of cell count). Mapper analyses would hopefully give us some insight as to which functions to

test. In particular we would like to determine over how long do , and the impact in tree growth of frost events.
- Running Mapper with a lens that uses PCA.
- Using a better interpolation technique to estimate daily cell count data, ideally which takes meteorology into account.

Questions

- Mature cells being too low, which other interpolation for the end of the year should we use?
- Need to make EZ, WZ go down at the end of the growing season, which period of time should we use?
- Not taking into account the precision factor PR as do not know how to replace the missing values or interpolate it
- Ideas for lenses