Unsupervised clustering individual growth

Marko Smiljanic

10/8/2021

Background

Short summary: we have two large sites selected to study pollution signal archived within tree growth signal. Overarching question is whether there is the significant difference in the pattern of the tree growth at two sites which could be potentially traced back to lowering pollution signal in the late 1980 to early 1990.

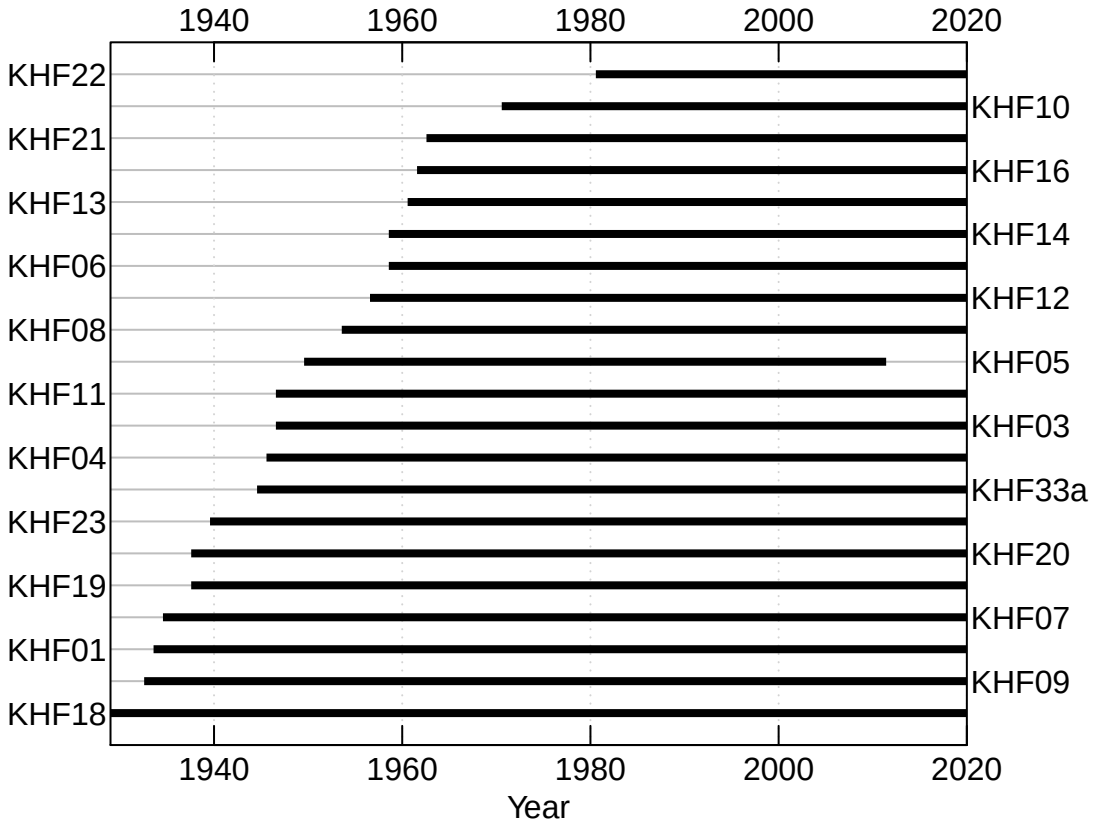
Introduction

To estimate the significance of the potential differences we decided to conduct hierarchical agglomerative clustering, HAC, of the time series. The analysis should help with the objective evaluation of the different growth patterns and indicate groups of the similar trees growth series. If those groupings are more or less clustered within two big areas, or smaller subsites, they might indicate dissimilar growth patterns within the sample.

Data import and transformation

Hierarchical clustering algorithm expects dissimilarity matrix between all tree series mostly generated by dist function. For this we first need to load all data, do individual transformations (i.e. detrending) of the tree series, and finally collect the data into one large data frame. First step is to import all the data into the list of the rwl data frames.

```
library(dplR)
library(tidyverse)
files <- list.files("../Crossdated complete/", full.names = TRUE) #lists all files within the folder
raw_data <- lapply(files, function(x) {  # returns the list of all rwl files
  dplR::read.rwl(x)  })
plot(raw_data[[1]]) # plot series from the first site using dplR/just for check
```



Second step is to detrend the series from our data frames. I have opted for the simple negative exponential here, but we can adjust later.

```
detrended_data <- lapply(raw_data, function(x) { # detrends all of the series individually with negativ
  x %>%
  as.data.frame() %>%
  dplR::detrend(method="ModNegExp") })
```

Finally, third step is to collect all of the detrended data frames into single data frame. We will use years for this. As the start and end years are not from the start I have to use the joining operation.

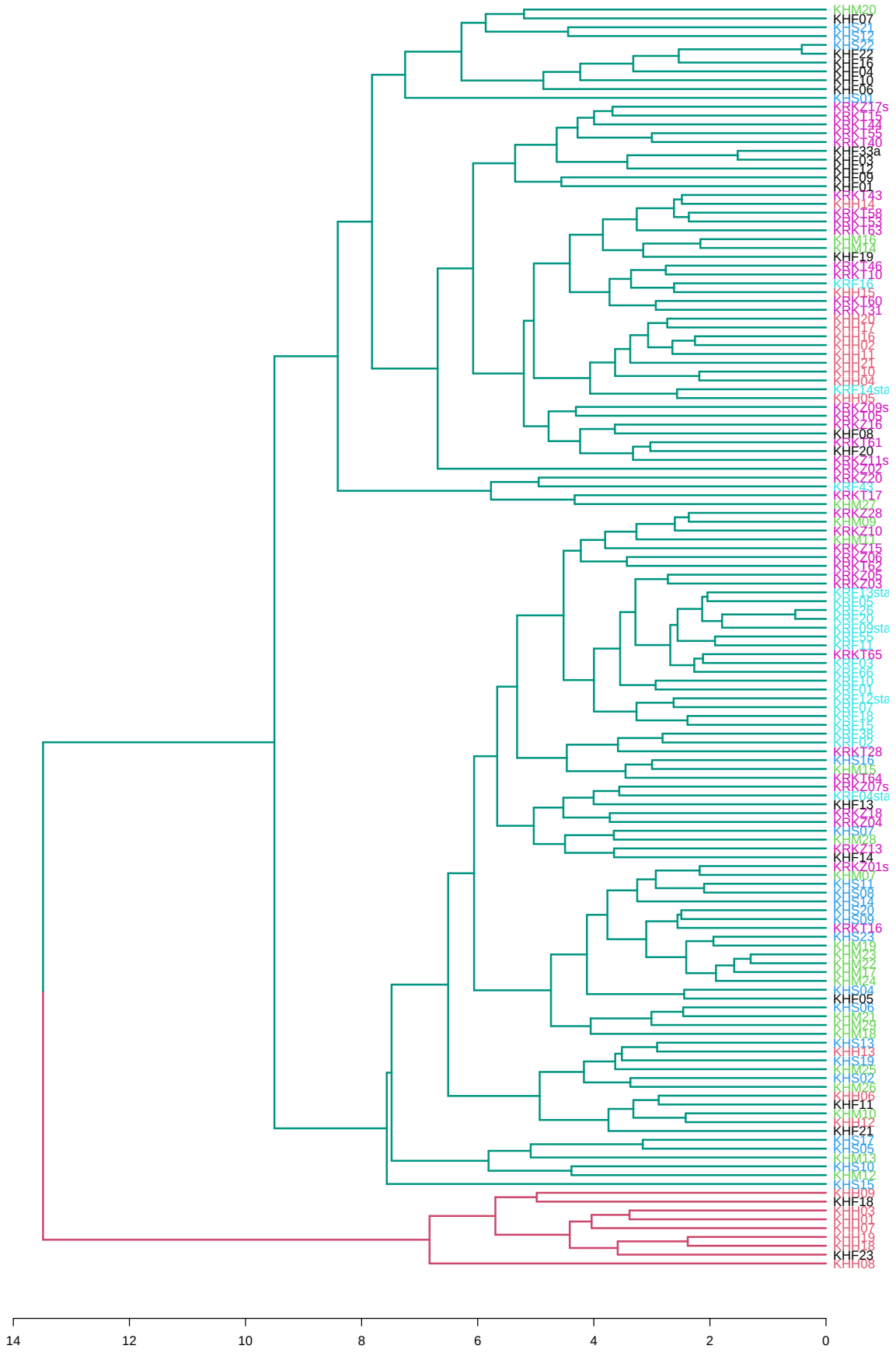
```
detrended_data_collection <- lapply(detrended_data, function(x) {
  x %>%
    rownames_to_column("Years") %>%  # First make a column from rownames
    mutate(Years = as.numeric(Years)) }) %>% # Second transform it to numeric
    plyr::join_all() %>% # Use new column to stick all of the individual data frames together
    column_to_rownames("Years") # finally reverse the manipulation of the data frame in order to have r
```

With that we are done with data manipulation and have a data in the form we need and HAC algorithm expects.

Clustering analysis

First clustering analysis. To start, we transpose the data collection so that the rows are tree series and columns years, then build distance matrix and pass it through an hclust algorithm. Finally, we save the plot as dendrogram for further colouring and annotation. Distance matrix evaluate common Euclaedian distance for now and Ward process. Colour of the labels indicate the site of origin, while colour of the branches indicate two most distinct sites, i.e. two most pronounced groups...

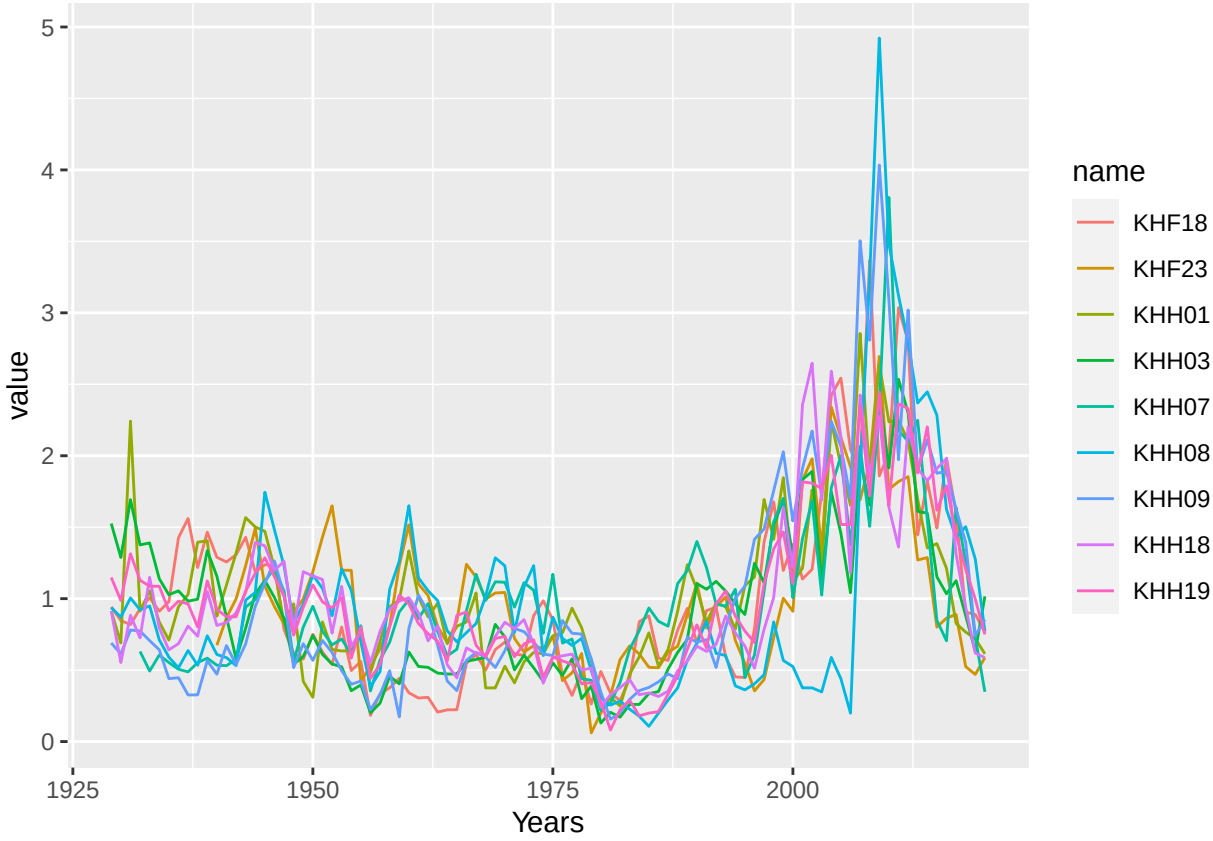
```
library(dendextend) # for fancy dendrogram colouring

dd <- detrended_data_collection %>%
  t() %>% #transpose
  dist() %>% #create distance matrix
  hclust() %>% #algorithm tool
  as.dendrogram() #save as dendrogram

#
cols <- as.factor(substr(colnames(detrended_data_collection),1,3))
labels_colors(dd) <- as.numeric(cols)[order.dendrogram(dd)]
dd <- dd %>% colour_branches(k=2) %>% set("branches_lwd", 2)

par(cex=0.8)
plot(dd, horiz=TRUE)
```


KHM20
KHF07
KHS21
KHS12
KHS22
KHF22
KHF16
KHF04
KHF10
KHF06
KHS01
KRKZ17s
KRKT15
KRKT47
KRKT44
KRKT55
KRKT40
KHF33a
KHF03
KHF17
KHF09
KHF01
KRKT43
KHH14
KRKT58
KRKT53
KRKT63
KHM16
KHM14
KHF19
KRKT46
KRKT10
KRKT16
KHH15
KRKT60
KRKT31
KHH20
KHH17
KHH16
KHH02
KHH11
KHH21
KHH10
KHH04
KRE14sta
KHH05
KRKZ09s
KRKT05
KRKZ16
KHF08
KRKT61
KHF20
KRKZ11s
KRKZ02
KRKZ20
KRKT45
KRKT17
KHM27
KRKZ28
KHM09
KRKZ10
KHM11
KRKZ15
KRKZ06
KRKT62
KRKZ05
KRKZ03
KRE13sta
KRE05
KRE26
KRE20
KRE09sta
KRE55
KRE11
KRKT65
KRE03
KRE28
KRE06
KRE10
KRE01
KRE12sta
KRE07
KRE19
KRE15
KRE38
KRE02
KRKT28
KHS16
KHM15
KRKT64
KRKZ07s
KRE04sta
KHF13
KRKZ18
KRKZ04
KHS07
KHM28
KRKZ13
KHF14
KRKZ01s
KHM07
KHS11
KHS08
KHS14
KHS20
KHS26
KHS09
KRKT16
KHS23
KHM19
KHM23
KHM22
KHM17
KHM24
KHS04
KHF05
KHS06
KHM21
KHM29
KHM18
KHS13
KHH13
KHS19
KHM25
KHS02
KHM26
KHH06
KHF11
KHM10
KHH12
KHF21
KHS17
KHS05
KHM13
KHS10
KHM12
KHS15
KHH09
KHF18
KHH03
KHH01
KHH07
KHH19
KHH18
KHF23
KHH08
14
12
10
8
6
4
2
0

Seems that there are two very distinct clusters, however one much smaller than the other. Now what could be the reason of this? Most of the sereis from the small cluster (marked red in the figure above) are from 'KHH' site (additionally KHF23 and KHF18). Let's first all of the tree series from the smaller cluster:

```
detrended_data_collection %>%
  select(KHH01, KHH03, KHH07, KHH08, KHH09, KHH18, KHH19, KHF18, KHF23) %>%
  rownames_to_column("Years") %>%  # First make a column from rownames
  mutate(Years = as.numeric(Years)) %>%
  pivot_longer(-Years) %>%
  ggplot(aes(x=Years, y=value, colour = name)) + geom_line()
```

```
## Warning: Removed 14 row(s) containing missing values (geom_path).
```



Right. They seem to follow the similar pattern, i.e. high release from late 1970 being more pronounced after late 1990s early 2000s. To test this hypothesis let's do a second stage detrending using spline.

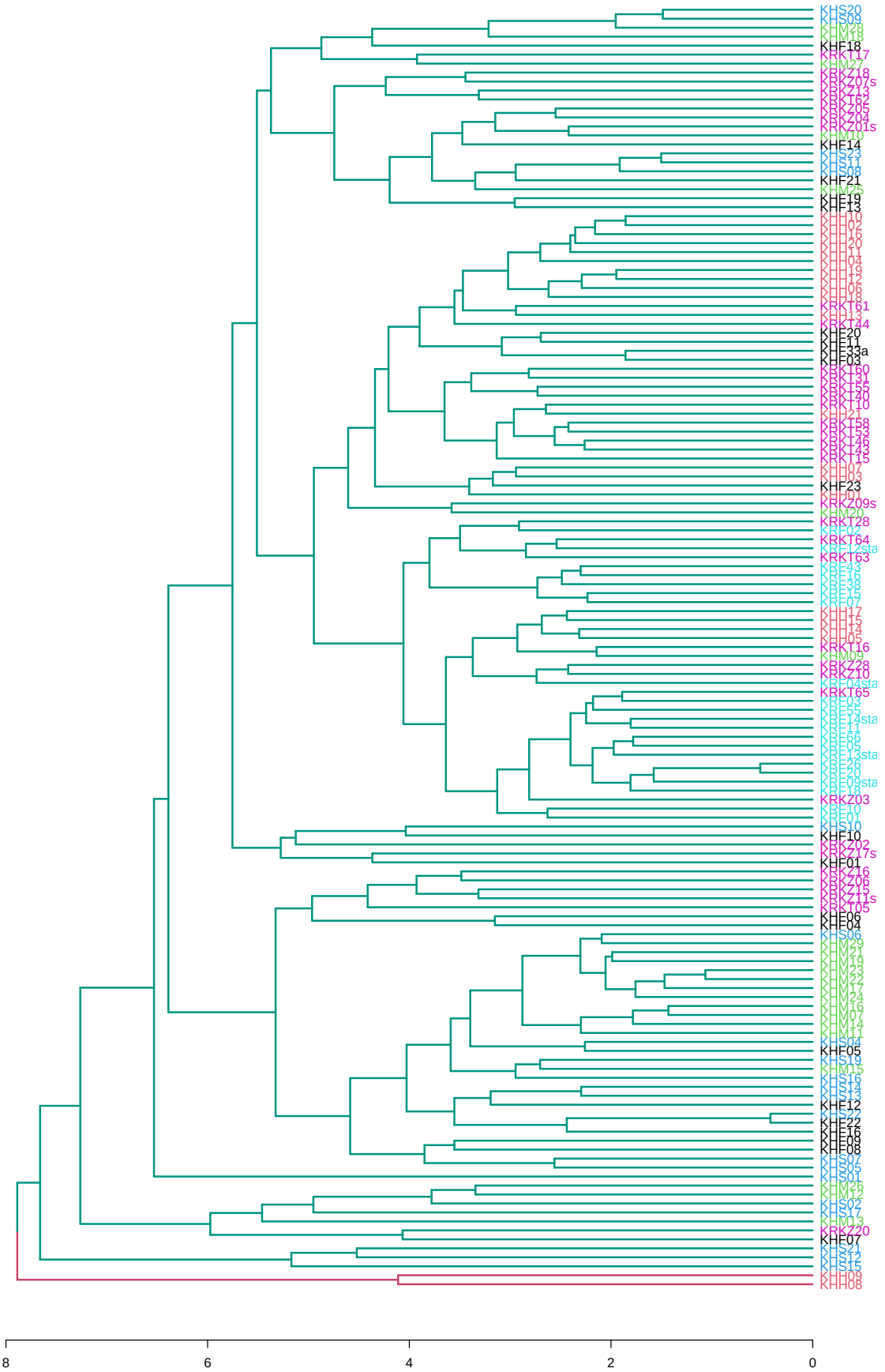
```
detrended_spline_data <- lapply(raw_data, function(x) { # detrends all of the series individually with
  x %>%
  as.data.frame() %>%
  dplR::detrend(method="Spline") })
```

Now repeat the manipulation of the two-stage detrending similar to before.

```
detrended_spline_data_collection <- lapply(detrended_spline_data, function(x) {
  x %>%
    rownames_to_column("Years") %>%
    mutate(Years = as.numeric(Years)) }) %>% plyr::join_all() %>% column_to_rownames("Years")
```

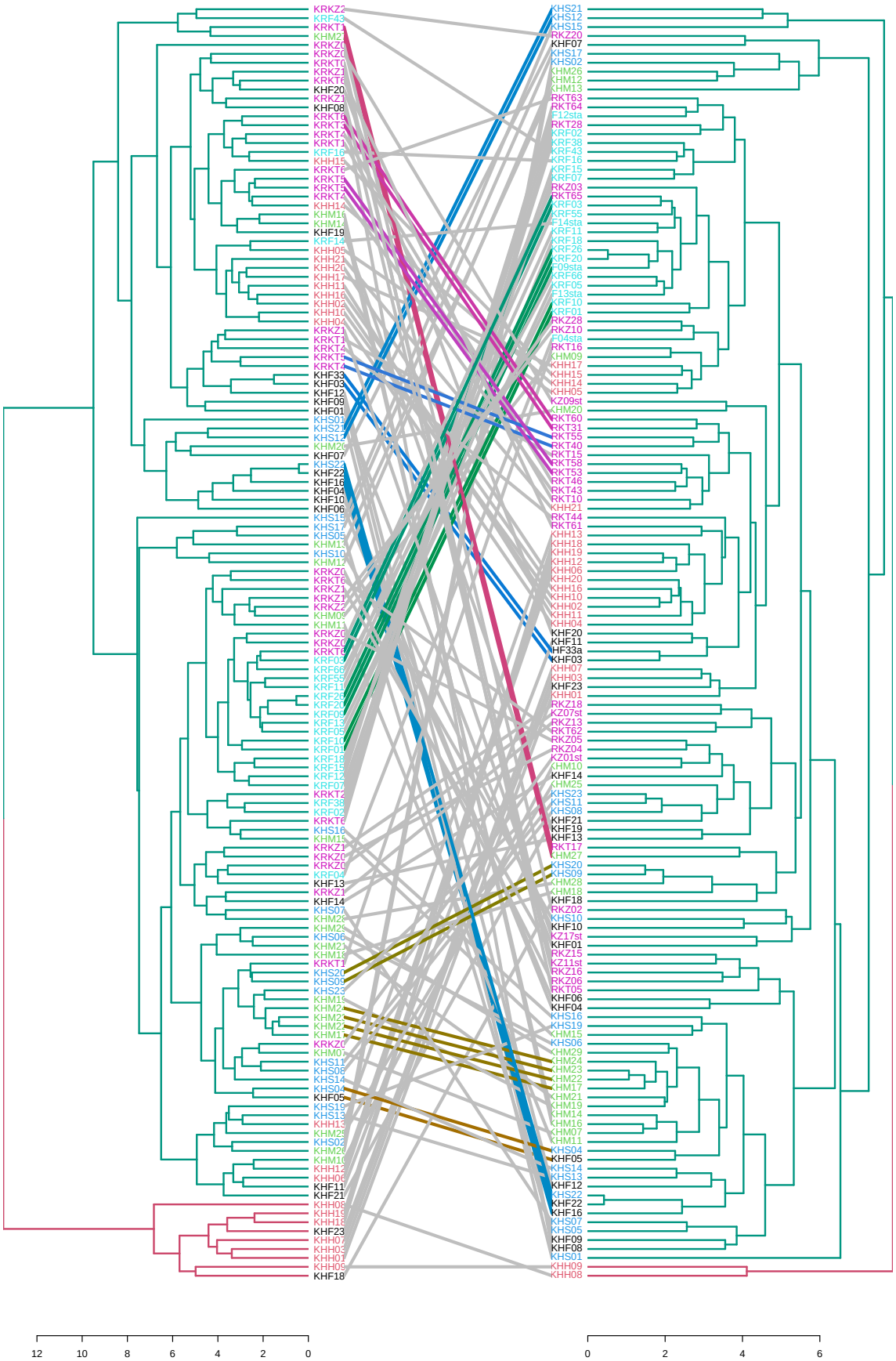
Clustering...

```r
library(dendextend)

dd_spline <- as.dendrogram(hclust(dist(t(detrended_spline_data_collection))))

cols <- as.factor(substr(colnames(detrended_spline_data_collection),1,3))
labels_colors(dd_spline) <- as.numeric(cols)[order.dendrogram(dd_spline)]

dd_spline <- dd_spline %>% colour_branches(k=2) %>% set("branches_lwd", 2)

par(cex=0.8)
plot(dd_spline, horiz=TRUE)
```


KHS20
KHS09
KHM28
KHM18
KHF18
KRKT17
KHM27
KRKZ18
KRKZ07s
KRKZ13
KRKT62
KRKZ05
KRKZ04
KRKZ01s
KHM10
KHF14
KHS23
KHS11
KHS08
KHF21
KHM25
KHF19
KHF13
KHH10
KHH02
KHH16
KHH20
KHH11
KHH04
KHH19
KHH12
KHH06
KHH18
KRKT61
KHH13
KRKT44
KHF20
KHF11
KHF33a
KHF03
KRKT60
KRKT31
KRKT55
KRKT40
KRKT10
KHH21
KRKT58
KRKT53
KRKT46
KRKT43
KRKT15
KHH07
KHH03
KHF23
KHH01
KRKZ09s
KHM20
KRKT28
KHE02
KRKT64
KHE12sta
KRKT63
KHE43
KHE16
KHE38
KHE15
KHE07
KHH17
KHH15
KHH14
KHH05
KRKT16
KHM09
KRKZ28
KRKZ10
KHE04sta
KRKT65
KHE03
KHE55
KHE14sta
KHE11
KHE66
KHE05
KHE13sta
KHE26
KHE20
KHE08sta
KHE18
KRKZ03
KHE10
KHE01
KHS10
KHF10
KRKZ02
KRKZ17s
KHF01
KRKZ16
KRKZ06
KRKZ15
KRKZ11s
KRKT05
KHF06
KHF04
KHS06
KHM29
KHM21
KHM19
KHM23
KHM22
KHM17
KHM24
KHM16
KHM07
KHM14
KHM11
KHS04
KHF05
KHS19
KHM15
KHS16
KHS14
KHS13
KHF12
KHS17
KHF22
KHF16
KHF09
KHF08
KHS07
KHS05
KHS01
KHM26
KHM12
KHS02
KHS18
KHM13
KRKZ20
KHF07
KHS21
KHS12
KHS15
KHH09
KHH08
8
6
4
2
0

Finally let's compare two dendrograms.

```
dl <- dendlist(dd, dd_spline)
tanglegram(dl, sort = TRUE, common_subtrees_color_lines = TRUE,
           highlight_distinct_edges  = FALSE, highlight_branches_lwd = FALSE)
```


12
10
8
6
4
2
0
0
2
4
6

So it seems that there is a joint release of some, but not all trees at KHH site. Otherwise, tree sereis seem to be quite randomly distributed.

Discussion & some first conclusions

Without going too deep into the matter for the first analysis, it seems that tree series cluster more or less randomly together, which in turn probably indicates that there are no obviously different clusters. KHH and release event is one possible outlier, which should probably be investigated as soon as possible. Specifically, what caused the sudden release in those 7-9 trees. First step would be to investigate if there were one of large management pressure/fire/dieback events at KHH site in the late 1970s or early 1990s?

Second point is to investigate whether the pollution acted as an modifying parameter, interacting with others. On the individual scale that can probably be achieved with (non)linear mixed effect models. However, for that our sample size might actually not be enough...

Phase 2 - moving window clustering comparison

In phase 2 of the clustering analysis we wanted to do a moving window comparison of the clustering dendrograms. Basically we compare clustering dendrogram from the reference period (1930-1960 in this case) with one from periods 1931-1961, 1932-1962, ... , 1989-2019, 1990-2020. We expect the difference to constantly increase with time, as pollution depositions are accumulated. Structural breaks (sudden changes) in the series might indicate a change of the growing conditions. As a measurement of the difference between two dendrograms I am using the entanglement, which is estimating the difference in the orders of the labels of two dendrograms after searching for optimal rotation.

Important dates to remember

1960 probable pollution start at KRK

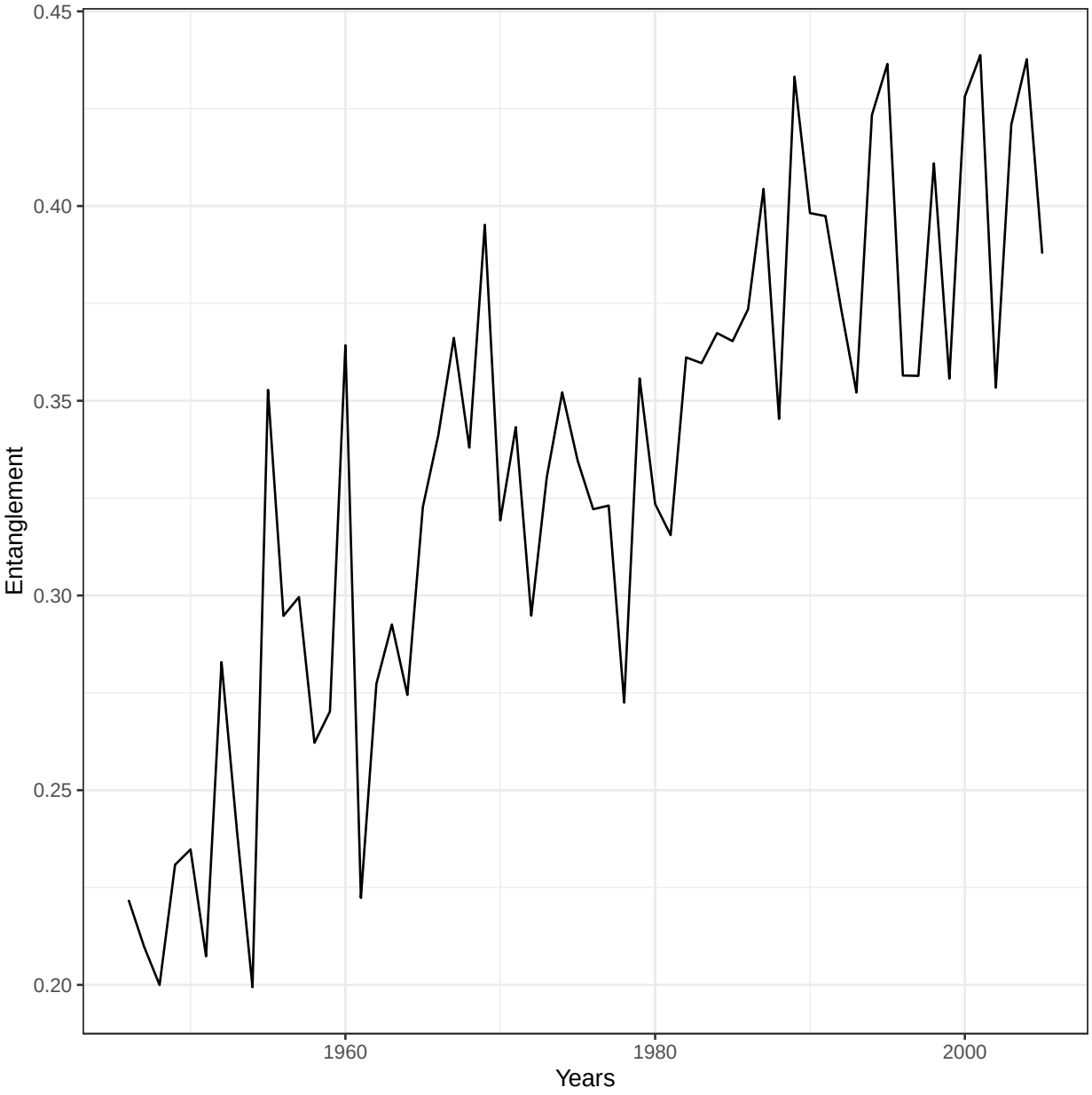
1968 environmental monitoring

1992 or 1993 sulfur filters

1995 or 1996 dust filters

```
two_period_compare <- function(period = 1931:1961, dat = detrended_spline_data_collection, ref_period =
  #extract reference period
  dat_ref <- dat %>%
    rownames_to_column(var="Years") %>%
    filter(Years %in% ref_period) %>%
    column_to_rownames("Years")

  #select the series which do not have near zero variance - removes trees that are too young for the re
  select_vec <- as.vector(caret::nzv(dat_ref))
  dat_ref <- dat_ref %>% select(-all_of(select_vec))

  #reference period clustering and dendrogram save
  dd_ref <- dat_ref %>%
    t() %>%
    dist() %>%
    hclust() %>%
    as.dendrogram()

  #extract comparison period and select the same trees as in ref
  dat_period <- dat %>%
    rownames_to_column(var="Years") %>%
    filter(Years %in% period) %>%
```

```
    column_to_rownames("Years") %>%
    select(-all_of(select_vec))

  # comparison period clustering and dendrogram
  dd_period <- dat_period %>%
    t() %>%
    dist() %>%
    hclust() %>%
    as.dendrogram

  # entanglement estimation
  dendlist(dd_ref, dd_period) %>% untangle(method="random", R=Replicates) %>% entanglement()
};


### generate periods
start <- 1931:1990
end <- start + 30
periods <- map2(start, end, seq)

### generage entanglements
set.seed(1234)
#entangles <- map_dbl(periods, two_period_compare, Replicates = 10000)
### Plot the figure
# pdf("Entanglement.pdf")
# data.frame(Years = floor((start+end)/2), Entanglement = entangles) %>%
#   ggplot(aes(x=Years, y=Entanglement)) + geom_line() +
#   theme_bw()
# dev.off()
knitr::include_graphics("Entanglement.pdf")
```



Discuasion Phase 2

The difference between cluster dendrograms has been monotonally increasing until 1990s (period 1975-2015), after which differences remained more or less constant. Therefore, we can not rule out the possibility that the increasing pollution was contributing to the increasing variability of the tree growth.

ToDo other dissimilarity measurements...