Fiche TD avec le logiciel R : MastingYearSelection

Sélection des années de masting

Pr Jean R. Lobry

Analyse du critère de LaMontagne. Proposition d'une version étendue.

Contents

1 Introduction

On s'intéresse ici à la méthode proposée par Jalene M. LaMontagne [3, chap. 3] pour décréter qu'une année donnée est de *masting* ou non. Cette méthode se comporte mieux que 5 autres compétitrices [4]. Notons $\boldsymbol{x}$ une série temporelle de n éléments génériques x_i. Dans les études de *masting* les éléments sont de $\mathbb{R}_+$, autrement dit, nous avons à faire à des séries temporelles non négatives :

$$\forall i \in \{1, \ldots, n\} : x_i \geq 0 \tag{1}$$

La première étape va consister à transformer notre série, $\boldsymbol{x}$, en son pendant en valeurs centrées-réduites, $\boldsymbol{x}^*$, en retranchant la moyenne et en divisant par l'écart-type :

$$\forall i \in \{1, \ldots, n\} : x_i^* = \frac{x_i - \bar{x}}{s_x} \tag{2}$$

On note classiquement ici $\bar{x}$ la moyenne de l'échantillon et s_x l'écart-type de l'échantillon, c'est à dire la racine carrée de la variance de l'échantillon s_x^2 :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

Soit M_i la variable aléatoire de Bernoulli indicatrice de l'évènement : « la ième observation x_i correspond à une année de *masting* ». La règle de décision proposée par Jalene M. LaMontagne est :

$$M_i = 1 \iff x_i^* > |\min \boldsymbol{x}^*| \tag{4}$$

Dans les séries de *masting* la distribution des x_i suit typiquement une distribution log-normale enrichie en valeurs nulles. Il est donc vraisemblable qu'une telle série comporte au moins un zéro. Notons ${}^0\boldsymbol{x}$ une série comportant au moins un zéro, dans ce cas particulier on a alors :

$$M_i = 1 \iff {}^0x_i^* > |\min {}^0\boldsymbol{x}^*| = |\frac{0 - {}^0\bar{x}}{{}^0s_x}| = \frac{1}{\mathrm{PCV}({}^0\boldsymbol{x})} \tag{5}$$

Si la série comporte au moins un zéro, le seuil de décision n'est rien d'autre que l'inverse du coefficient de variation de Pearson [6], ${}^{\mathrm{P}}\mathrm{CV}({}^0\boldsymbol{x})$. Ceci explique la relation « curvilinéaire » rapportée par LaMontagne et Boutin [4, Fig. 2], qui n'est rien d'autre qu'une branche d'hyperbole équilatère. Ceci permet de comprendre le caractère « autoadaptatif » du seuil de décision : pour des séries ayant un fort coefficient de variation, donc avec une asymétrie à droite très prononcée, on va pouvoir se permettre de couper plus bas qu'avec une série de moindre variabilité (on illustrera ce point dans la section suivante).

Dans le cas particulier d'une série ${}^0\boldsymbol{x}$ comportant au moins une valeur nulle, le critère de LaMontagne revient à décréter comme étant une année de *masting* toutes celles dont la valeur dépasse deux fois celle de la moyenne de la série, ${}^0\bar{\boldsymbol{x}}$:

$$M_i = 1 \iff {}^0x_i^* > \frac{{}^0\bar{x}}{{}^0s_x} \iff \frac{{}^0x_i - {}^0\bar{x}}{{}^0s_x} > \frac{{}^0\bar{x}}{{}^0s_x} \iff {}^0x_i > 2\ {}^0\bar{x} \tag{6}$$

Pour des espèces suffisamment longévives, observer à long terme une année avec une valeur nulle est un évènement quasi-certain. On peut donc définir un critère LaMontagne « extrapolé » en ajoutant par la pensée une valeur nulle à notre série $\boldsymbol{x}$. La règle de décision est alors simplement :

$${}^0M_i = 1 \iff x_i > 2\bar{x} \tag{7}$$

2 Le jeu de données

```
load("../CVisDead/batch/myws.Rda")
load("../CVisDead/batch/KCV.Rda")
load("../CVisDead/batch/PCV.Rda")
```

On utilise un jeu de données[1] de séries de *masting* déjà exploité par ailleurs [5], c'est un sous-ensemble de MASTREE+ [1] contenant 1433 séries quantitatives avec au moins 12 années documentées (max. 69, moy. 22.5). Les valeurs du coefficient de variation de Pearson [6], ^{P}CV, et de Kvålseth [2], ^{K}CV, ont déjà été calculées pour toutes ces séries dans les tables `eisPCV` et `eisKCV`, respectivement. Chaque série temporelle est univoquement identifiée [1] par son `ID` obtenu en concaténant `Alpha_Number`, `Site_number`, `Variable_number` et `Species_code`. Les données sont rangées ici dans la table `ms`, pour plus de détails voir la fiche de TD[2] « MASTREE ».

3 Implémentation de la méthode

La fonction `MastingYearSelection()` définie ci-après retourne le vecteur de booléens `Mi` indicateur de l'évènement : « la ième observation x_i correspond à une année de *masting* selon de le critère de LaMontagne (eq. 4) ». Elle renvoie également sa version « extrapolée » (eq. 7) dans le vecteur `Mi0`. Elle commence par tester si la série est constante : dans ce cas on ne peut pas calculer les données centrées-réduites puisque l'écart-type vaut zéro, par convention on décide qu'aucune valeur ne satisfait au critère. Dans les cas non pathologiques elle calcule également un booléen `zeroflag` indicateur de la présence d'une valeur nulle dans la série. Elle renvoie également la valeur du seuil dans les unités originelles avec le critère de LaMontagne (`seuilorig`) et avec sa version « extrapolée » (`seuilorig0`).

```
MastingYearSelection <- function(x, ...){
  n <- length(x)
  if(isTRUE(all.equal(x, rep(x[1], n)))){ # série constante
    tF <- rep(FALSE, n)
    return(list(seuil = NA, Mi = tF, zeroflag = NA, seuilorig = NA, seuilorig0 = NA, Mi0 = tF))
  } else { # série "normale"
    xs <- (x - mean(x, ...))/sdn(x, ...)
    seuil <- abs(min(xs))
    seuilorig <- sdn(x, ...)*seuil + mean(x, ...)
    seuilorig0 <- 2*mean(x, ...)
    return(list(seuil = seuil, Mi = ifelse(xs > seuil, TRUE, FALSE),
      zeroflag = any(as.logical(lapply(x, all.equal, 0.0)), na.rm = TRUE),
      seuilorig = seuilorig, seuilorig0 = seuilorig0,
      Mi0 = ifelse(x > seuilorig0, TRUE, FALSE)))
  }
}
```

Il suffit maintenant de faire rouler notre fonction `MastingYearSelection()` sur toutes les séries disponibles. On complète la table `ms` avec une colonne `Mi` et une colonne `Mi0` indiquant si la valeur de la ligne est classée en *masting* ou non. On complète également la table `eisPCV` avec les valeurs du seuil et l'indicatrice de la présence d'une valeur nulle dans la série.

[1] Disponible à `https://esb.univ-lyon1.fr/donnees/CVisDead.zip`
[2] `https://esb.univ-lyon1.fr/pdf/MASTREE.pdf`

```
IDs <- unique(ms$ID)
for(the_ID in IDs){
  sel <- which(ms$ID == the_ID)
  res <- MastingYearSelection(ms[sel , "Value"])
  ms[sel , "Mi"] <- res$Mi
  ms[sel , "Mi0"] <- res$Mi0
  eisPCV[eisPCV$ID == the_ID, "seuil"] <- res$seuil
  eisPCV[eisPCV$ID == the_ID, "seuilorig"] <- res$seuilorig
  eisPCV[eisPCV$ID == the_ID, "seuilorig0"] <- res$seuilorig0
  eisPCV[eisPCV$ID == the_ID, "zeroflag"] <- res$zeroflag
}
```

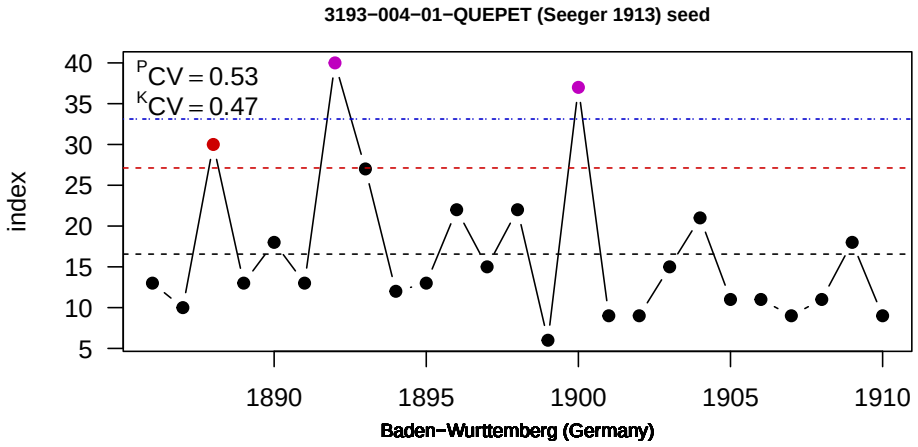
4 Visualisation des séries

On définit la fonction `plotID()` pour représenter une série temporelle donnée. Les années classées en *masting* sont représentées par des points en couleur : rouge avec le critère de LaMontagne et violet avec sa version « extrapolée ». La valeur moyenne est représentée par un trait en pointillé en noir, le seuil de décision par un trait pointillé en rouge et le seuil de décision « extrapolée » par un trait pointillé en bleu. On porte en information supplémentaire la valeur du coefficient de variation de Pearson et Kvålseth.

```
plotID <- function(the_ID, las = 1, pch = 19, cex.main = 0.75,
    cex.sub = 0.75, ...){
  par(mar = c(3, 4, 2, 0) + 0.1)
  df <- ms[ms$ID == the_ID, ]
  x <- df$Year ; y <- df$Value
  col <- ifelse(df$Mi, "red3", "black")
  col[df$Mi0] <- rgb(0.75, 0, 0.75)
  main <- paste(the_ID, " (", unique(df$Reference), ") ",
                unique(df$Variable), sep = "")
  plot(x, y, main = "", xlab = "", ylab = unique(df$Unit),
       las = las, type = "p", pch = pch, col = col, cex.main = cex.main, ...)
  points(x, y, pch = pch, type = "b", cex = 0)
  title(main = main, line = 1, cex.main = cex.main)
  title(sub = paste(df$Site, " (", df$Country, ")", sep = ""), line = 2, cex.sub = cex.sub)
  abline(h = mean(y), lty = 2)
  thres <- eisPCV[eisPCV$ID == the_ID, "seuilorig"]
  abline(h = thres, lty = 2, col = "red3")
  thres0 <- eisPCV[eisPCV$ID == the_ID, "seuilorig0"]
  abline(h = thres0, lty = "1331", col = "blue3")
  PCV <- round(eisPCV[eisPCV$ID == the_ID, "est"], 2)
  KCV <- round(eisKCV[eisKCV$ID == the_ID, "est"], 2)
  legend("topleft", inset = 0.01, legend = c(bquote(phantom(0)^P*CV == .(PCV)),
    bquote(phantom(0)^K*CV == .(KCV))), bty = "n", adj = c(0.4, 0))
}
```

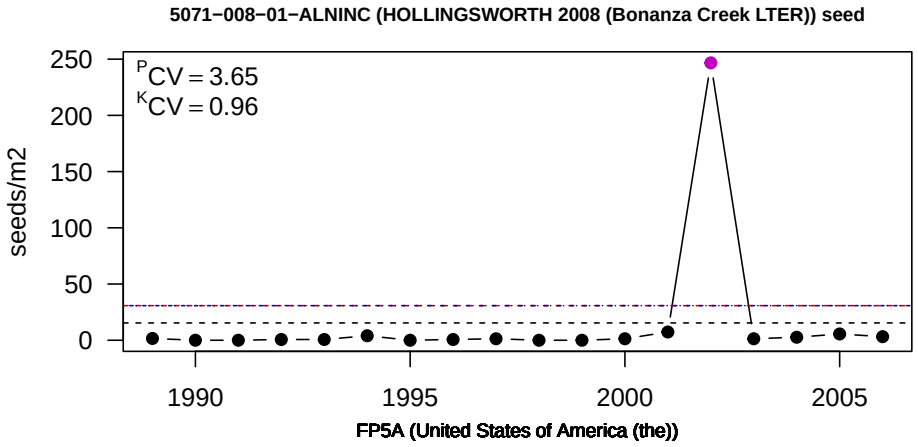
Voici un exemple de représentation pour une série ne comportant pas de valeurs nulles. Le critère « extrapolé » est plus strict et sélectionne moins d'années de *masting*.

```
plotID("3193-004-01-QUEPET")
```



Voici un exemple de représentation pour une série comportant une valeur nulle. Les deux lignes de seuil sont fusionnées, d'où l'aspect violet de la ligne. Il n'y a plus que des points violets puisque dans ce cas les deux critères sont identiques.

```
plotID("5071-008-01-ALNINC")
```



5 Fréquence des années de *masting*

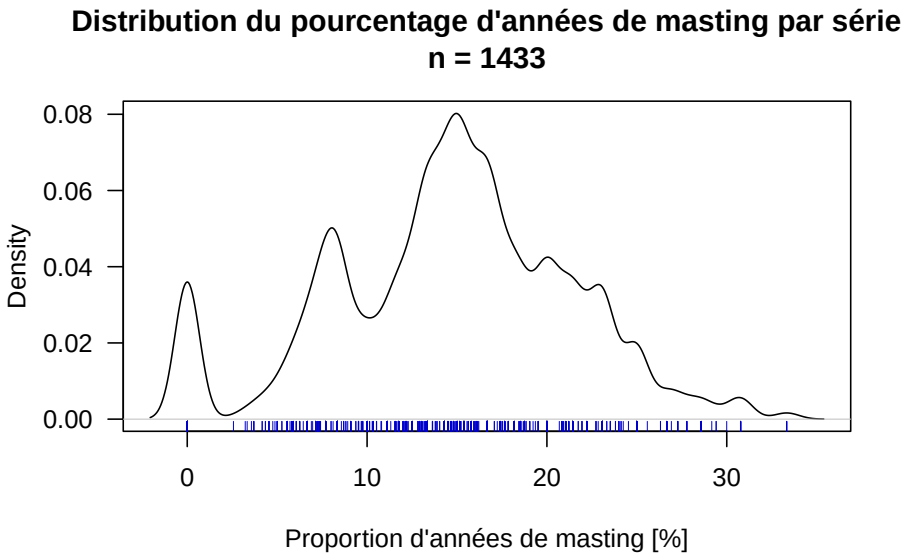
5.1 Critère de LaMontagne

Globalement, sur les 32304 années documentées, 4655 sont déclarées comme étant de *masting*, soit 14.4 %, une valeur assez proche de celle de 17 % trouvée par LaMontagne et Boutin [4]. Au vu de la distribution du pourcentage d'années de *masting* on serait tenté de définir quatre grandes classes : les séries à 0 %, 7 %, 15 % et plus de 20 %.

```
sum(ms$Mi)/nrow(ms) # pourcentage global d'années de masting
```

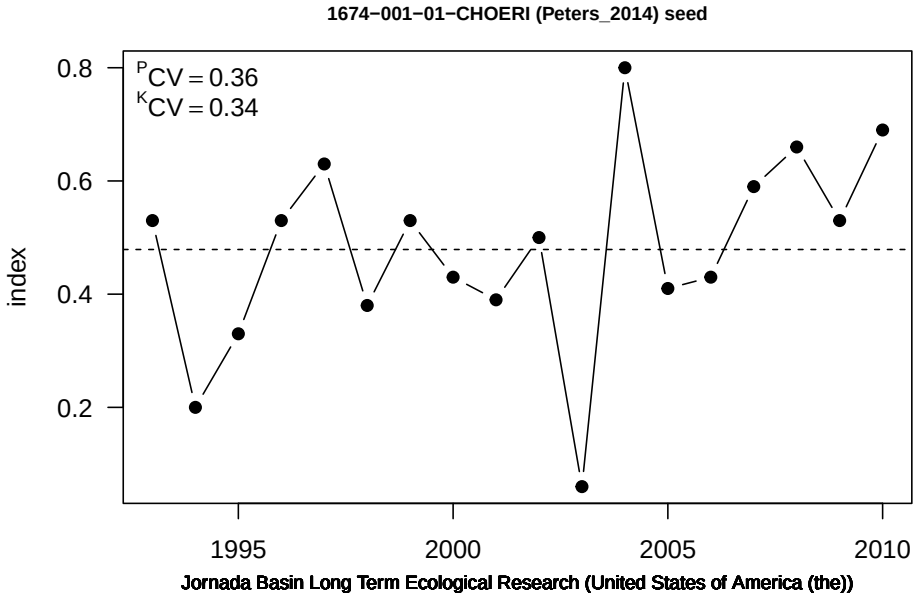
```
[1] 0.1440998
```

```
pcMastYear <- with(ms, tapply(Mi, ID, \(x) 100*sum(x)/length(x)))
main <- paste0("Distribution du pourcentage d'années de masting par série\nn = ",
  length(pcMastYear))
plot(density(pcMastYear, adjust = 0.5), main = main,
  las = 1, xlab = "Proportion d'années de masting [%]")
rug(jitter(pcMastYear), col = "blue3")
```



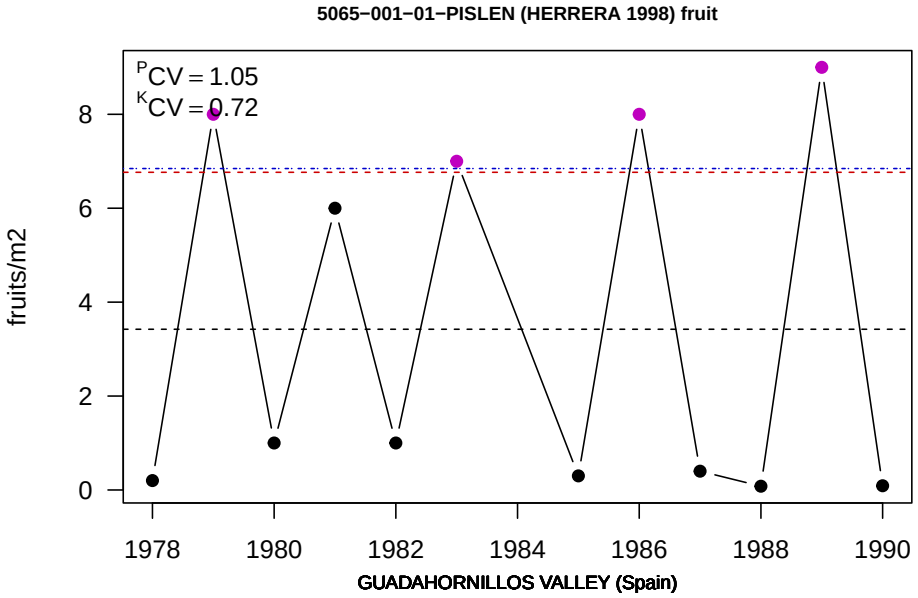
Voici un exemple de série sans année de *masting*. C'est une série avec un faible niveau de variabilité, et donc aucune année ne se distingue fortement de la moyenne.

```
plotID("1674-001-01-CHOERI")
```



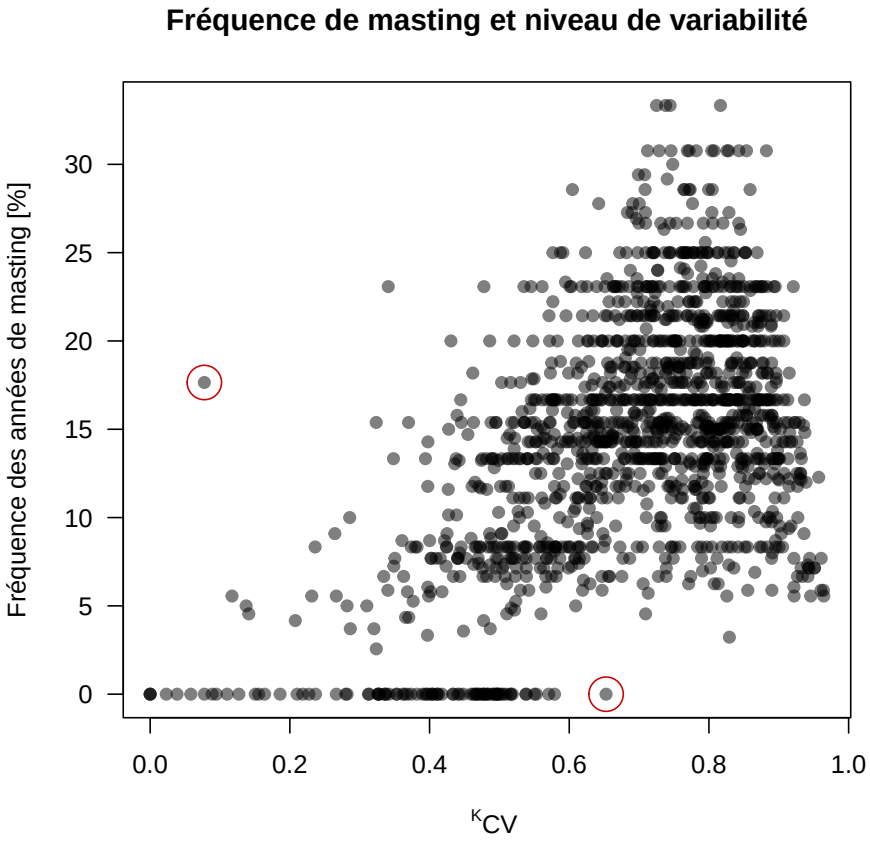
À l'autre extrémité du spectre, voici un exemple de série dont le tiers des années (4 sur 12) sont déclarées en *masting*. Le niveau de variabilité est bien plus fort qu'avec l'exemple précédent.

```
plotID("5065-001-01-PISLEN")
```

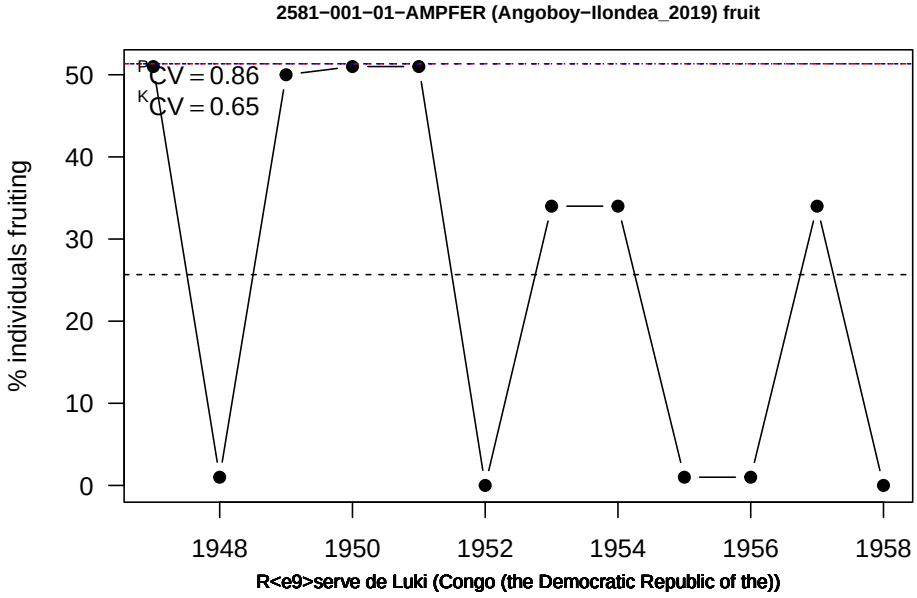


En généralisant, avec le graphique ci-après, on voit qu'il y a une sorte de relation triangulaire entre la fréquence de *masting* et le niveau de variabilité : une forte variabilité est une condition nécessaire mais non suffisante pour avoir beaucoup d'années de *masting*.

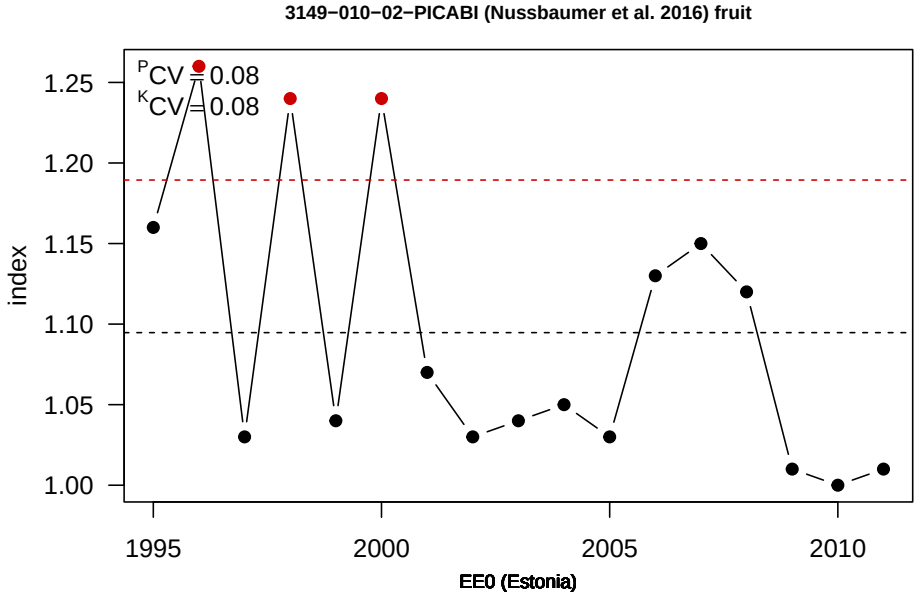
```
eisKCV <- merge(eisKCV, data.frame(ID = names(pcMastYear), pcM = pcMastYear))
with(eisKCV, {
  plot(est, pcM, pch = 19, col = rgb(0,0,0,0.5), las = 1,
    main = "Fréquence de masting et niveau de variabilité",
    xlab = bquote(phantom(0)^K*CV),
    ylab = "Fréquence des années de masting [%]")
  out1ID <- "2581-001-01-AMPFER" ; ii1 <- which(out1ID == ID)
  points(est[ii1], pcM[ii1], col = "red3", cex = 3)
  out2ID <- "3149-010-02-PICABI" ; ii2 <- which(out2ID == ID)
  points(est[ii2], pcM[ii2], col = "red3", cex = 3)
})
```



On regarde par curiosité les deux séries mises en évidence par les cercles rouges dans le graphique précédent. La première n'a pas d'années de *masting* malgré un niveau de variabilité relativement élevé.



La seconde a un nombre d'années de *masting* non négligeable malgré un niveau de variabilité très faible.



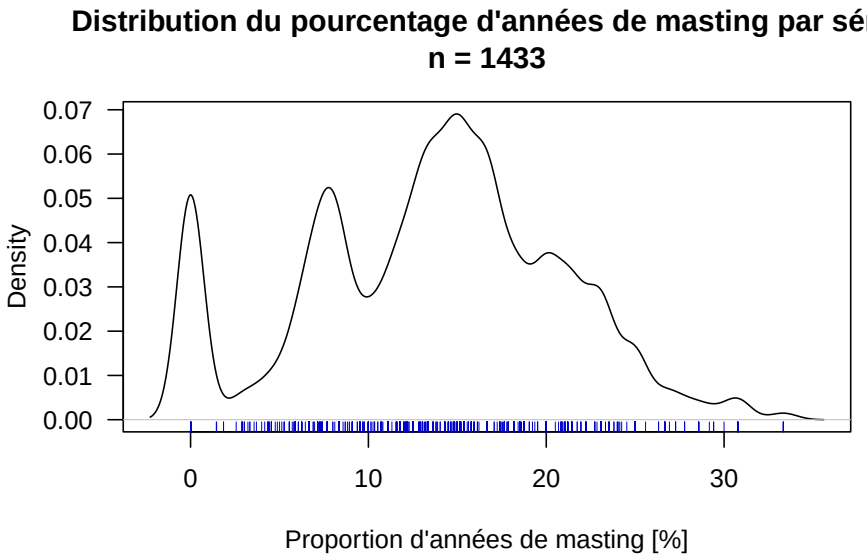
5.2 Critère de LaMontagne « extrapolé »

Globalement, sur les 32304 années documentées, 4237 sont déclarées comme étant de *masting*, soit 13.1.4 %, une valeur assez proche des 14.4 % avec le critère de LaMontagne. La distribution du pourcentage d'années de *masting* ressemble à ce que l'on avait précédemment avec un gonflement de la classe des séries sans aucune année de *masting*.

```
sum(ms$Mi0)/nrow(ms) # pourcentage global d'années de masting
```

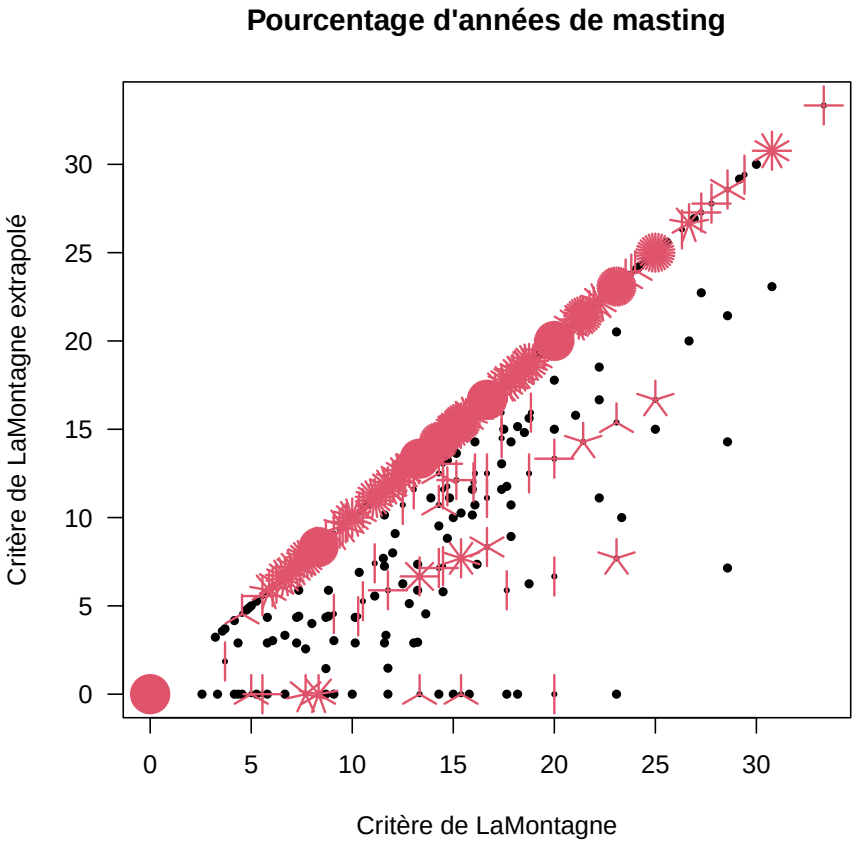
```
[1] 0.1311602
```

```
pcMastYear0 <- with(ms, tapply(Mi0, ID, \(x) 100*sum(x)/length(x)))
main <- paste0("Distribution du pourcentage d'années de masting par série\nn = ",
  length(pcMastYear0))
plot(density(pcMastYear0, adjust = 0.5), main = main,
  las = 1, xlab = "Proportion d'années de masting [%]")
rug(jitter(pcMastYear0), col = "blue3")
```



Le graphiques suivant montre que les deux critères sont cohérents. Le critère extrapolé étant plus exigeant, le pourcentage est toujours inférieur ou égal avec ce dernier qu'avec le critère de LaMontagne.

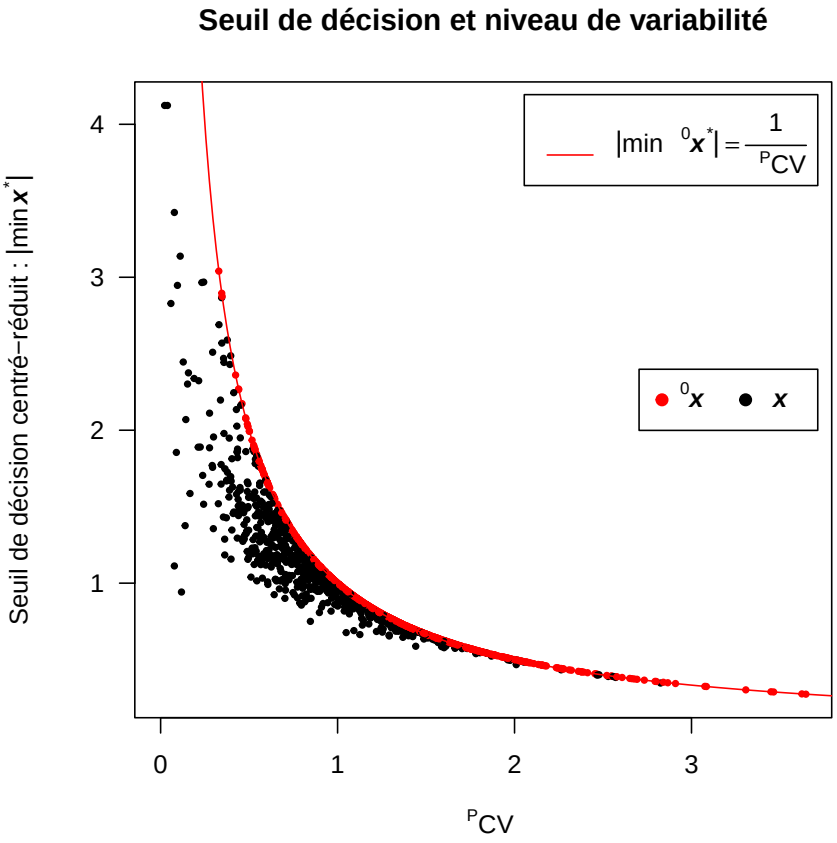
```
sunflowerplot(pcMastYear, pcMastYear0, main = "Pourcentage d'années de masting",
  xlab = "Critère de LaMontagne", ylab = "Critère de LaMontagne extrapolé", las = 1)
```



6 Impact du niveau de variabilité sur le seuil

On reproduit ci-dessous une figure du type [4, fig. 2], mais avec la valeur absolue du seuil de décision. En coloriant en rouge les séries contenant au moins un zéro on vérifie empiriquement la relation donnée par l'équation 5. On voit clairement ici que plus le niveau de variabilité est élevé, plus on peut se permettre de couper bas. Avec le critère extrapolé, tous les points seraient sur la ligne rouge.

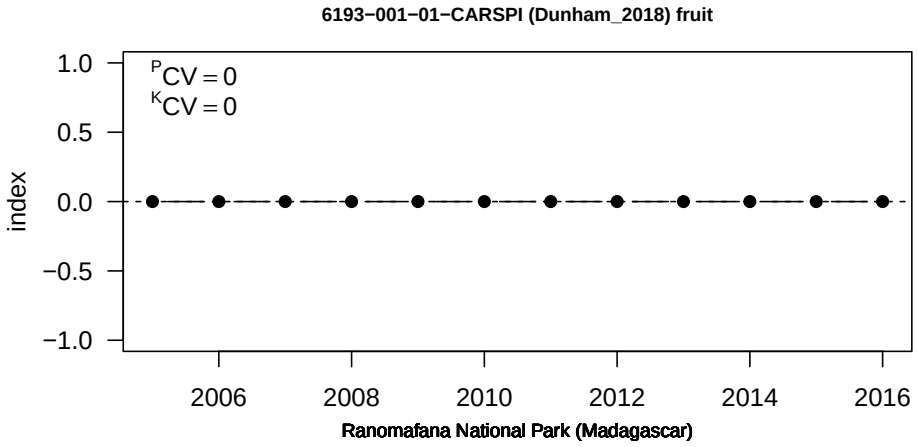
```
par(mar = c(5, 5, 4, 2) + 0.1)
plot(eisPCV$est, eisPCV$seuil, col = ifelse(eisPCV$zeroflag, "red", "black"),
     pch = 19, cex = 0.5, las = 1, xlab = bquote(phantom(0)^P*CV),
     ylab = bquote(paste("Seuil de décision centré-réduit : ", group("|",
        min(bolditalic(x)^plain("*")), "|"))),
     main = "Seuil de décision et niveau de variabilité")
xx <- seq(0, 5, le = 255)
lines(xx, 1/xx, col = "red")
legend("topright", inset = 0.02, lty = 1, adj = c(0, 0.3),
  legend = bquote(group("|",
    min(phantom(0)^0*bolditalic(x)^plain("*")),
    "|") == frac(1, phantom(0)^P*CV)), col = "red")
legend("right", inset = 0.02, pch = 19, col = c("red", "black"),
  legend = c(bquote(phantom(0)^0*bolditalic(x)), bquote(phantom(0)^phantom(0)*bolditalic(x))),
  adj = c(0.5, 0.3), ncol = 2)
```

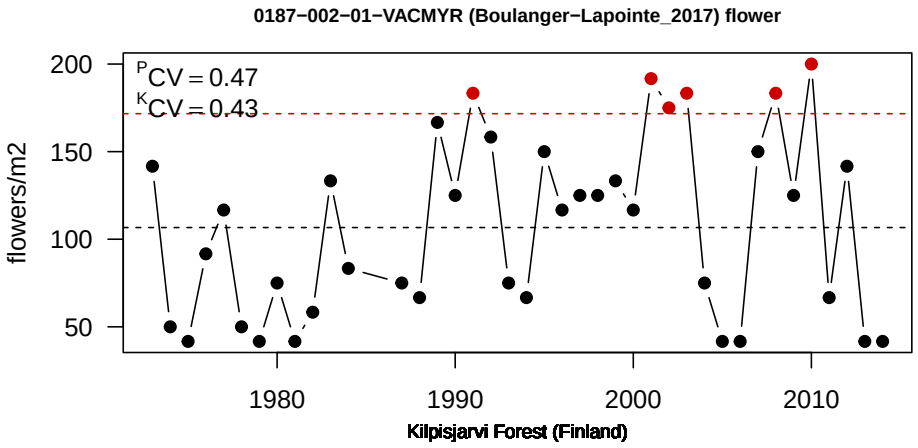


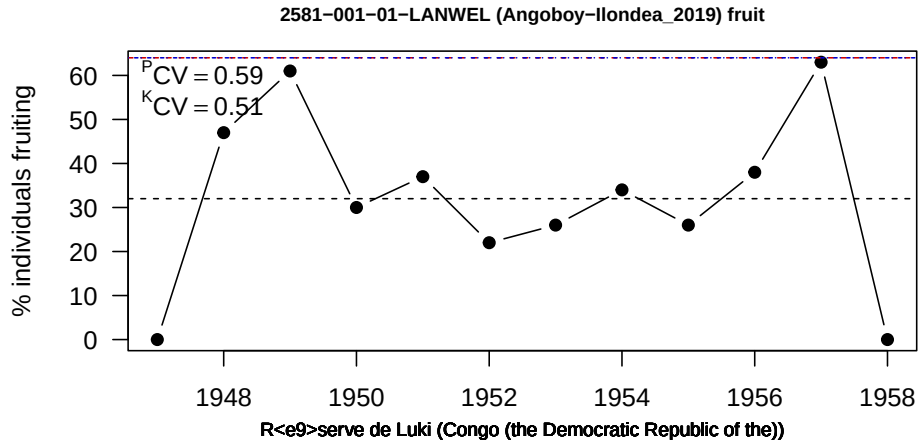
7 Quelques illustrations

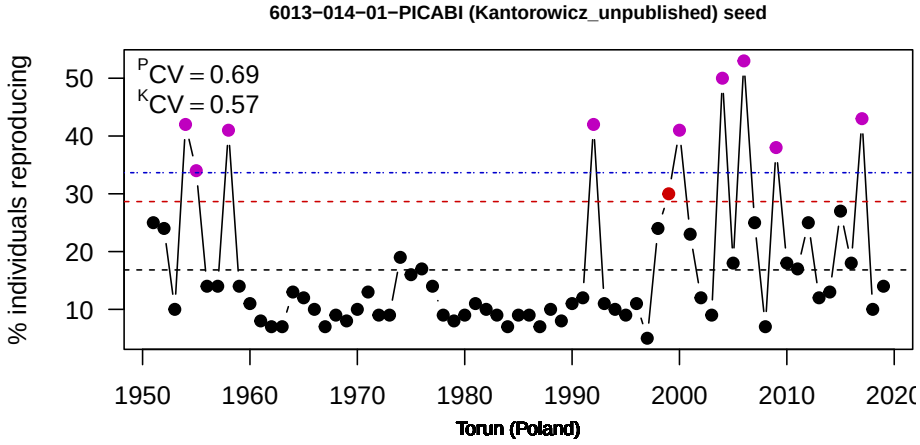
On produit ici quelques illustrations en sélectionnant quelques séries dans l'ordre croissant de variabilité. Pour les produire toutes il suffit d'ajuster la valeur du paramètre `nillustr`.

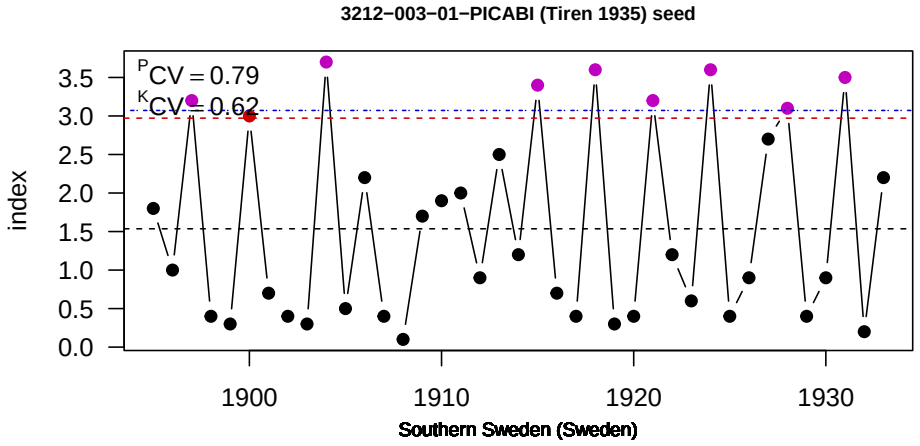
```
#nillustr <- nrow(eisKCV)
nillustr <- 15
eisKCV <- eisKCV[order(eisKCV$est), ]
iseq <- as.integer(seq(1, nrow(eisKCV), le = nillustr))
for(i in iseq){
  the_ID <- eisKCV[i, "ID"]
  fname <- paste0("figs/", the_ID, ".pdf")
  pdf(fname, ,width = 6, height = 3)
  plotID(the_ID)
  dev.off()
  cat(paste0("\n\\includegraphics[width=\\textwidth]{", fname, "}\n"))
}
```

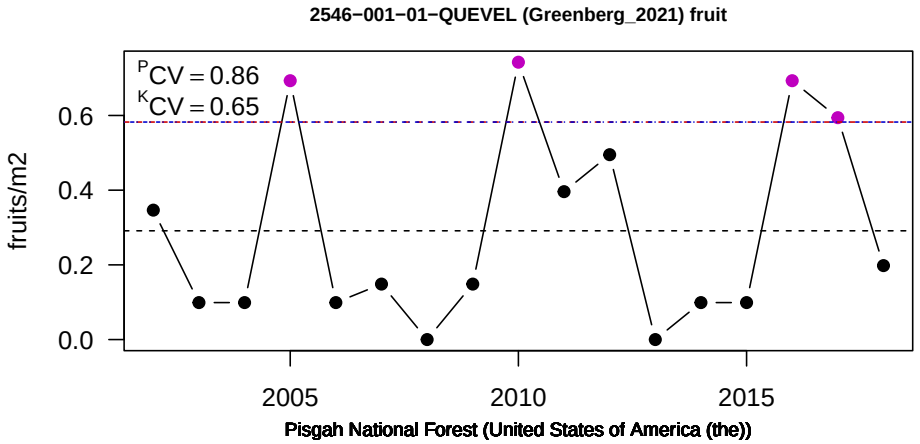

6193−001−01−CARSPI (Dunham_2018) fruit
PCV = 0
KCV = 0
index
1.0
0.5
0.0
−0.5
−1.0
2006
2008
2010
2012
2014
2016
Ranomafana National Park (Madagascar)

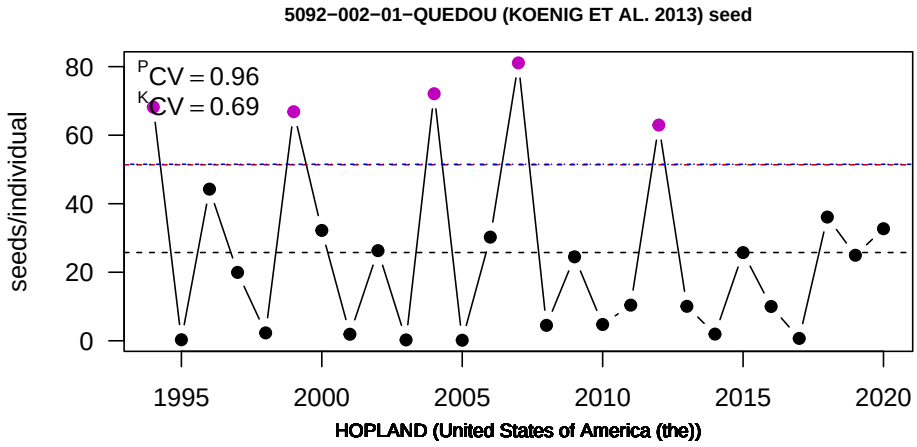

0187−002−01−VACMYR (Boulanger−Lapointe_2017) flower
PCV = 0.47
KCV = 0.43
flowers/m2
200
150
100
50
1980
1990
2000
2010
Kilpisjarvi Forest (Finland)

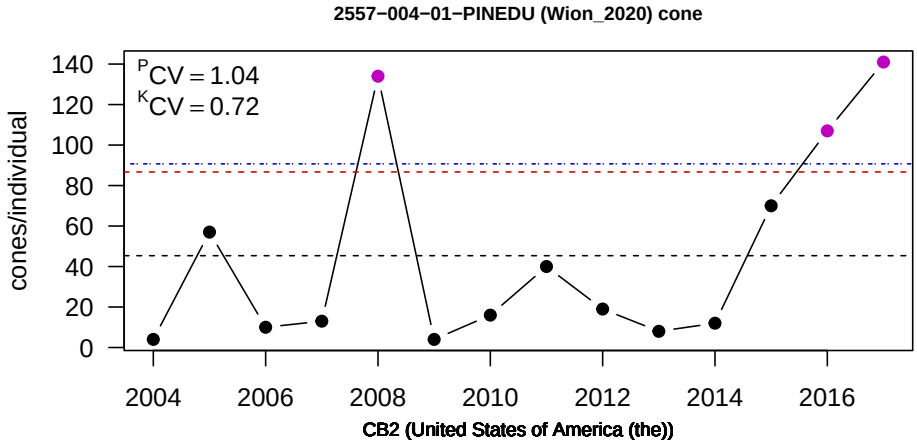

2581−001−01−LANWEL (Angoboy−Ilondea_2019) fruit
PCV = 0.59
KCV = 0.51
% individuals fruiting
60
50
40
30
20
10
0
1948
1950
1952
1954
1956
1958
R<e9>serve de Luki (Congo (the Democratic Republic of the))

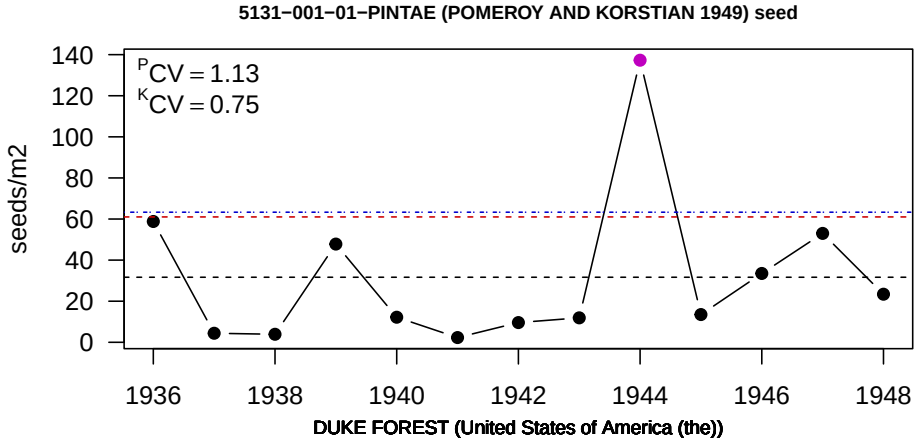

6013−014−01−PICABI (Kantorowicz_unpublished) seed
PCV = 0.69
KCV = 0.57
% individuals reproducing
50
40
30
20
10
1950
1960
1970
1980
1990
2000
2010
Torun (Poland)

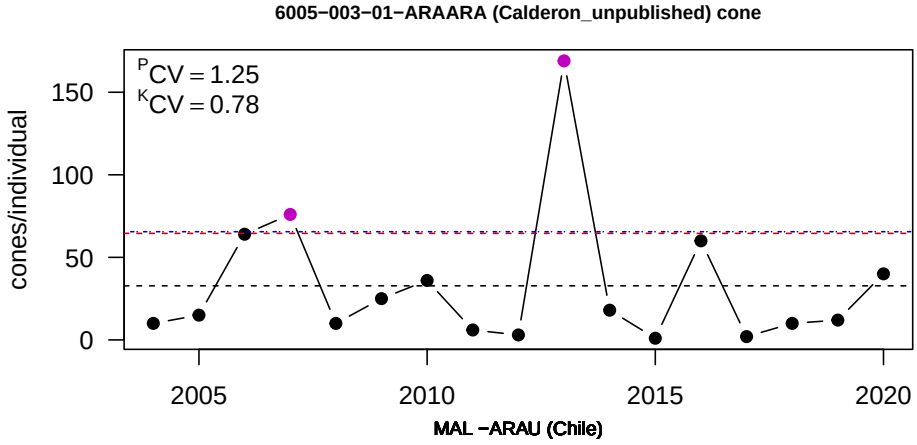

3212−003−01−PICABI (Tiren 1935) seed
PCV = 0.79
KCV = 0.62
index
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1900
1910
1920
1930
Southern Sweden (Sweden)

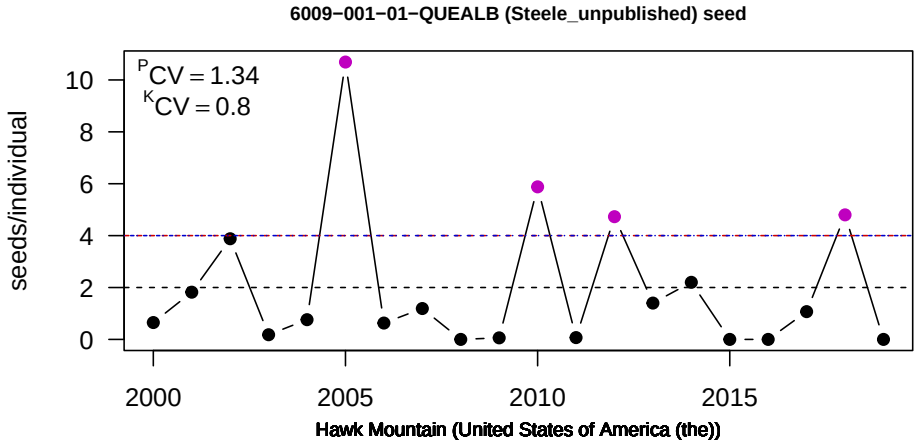

2546−001−01−QUEVEL (Greenberg_2021) fruit
PCV = 0.86
KCV = 0.65
fruits/m2
0.6
0.4
0.2
0.0
2005
2010
2015
Pisgah National Forest (United States of America (the))

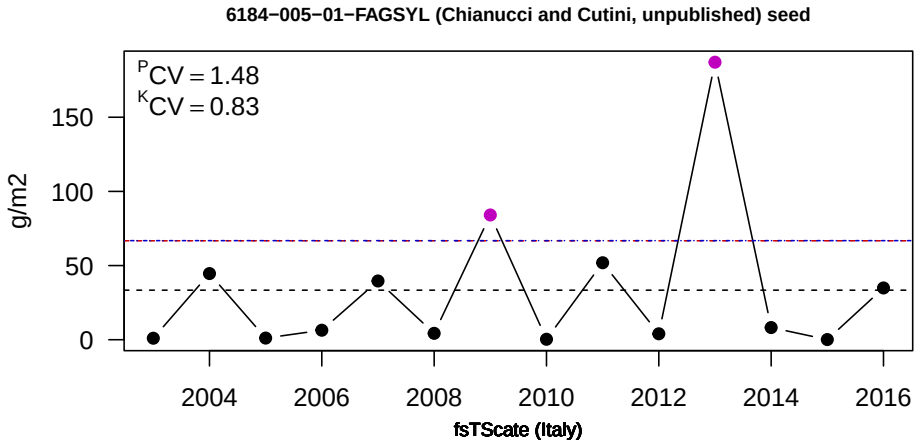

5092−002−01−QUEDOU (KOENIG ET AL. 2013) seed
PCV = 0.96
KCV = 0.69
seeds/individual
0
20
40
60
80
1995
2000
2005
2010
2015
2020
HOPLAND (United States of America (the))

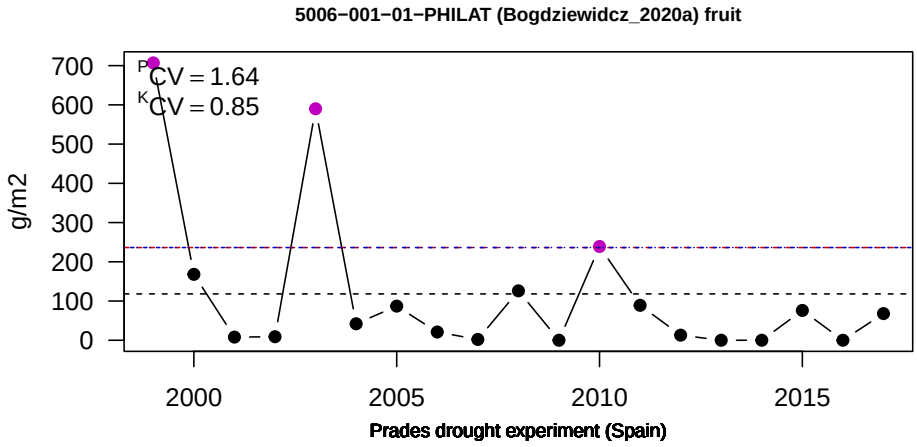

2557−004−01−PINEDU (Wion_2020) cone
PCV = 1.04
KCV = 0.72
cones/individual
0
20
40
60
80
100
120
140
2004
2006
2008
2010
2012
2014
2016
CB2 (United States of America (the))

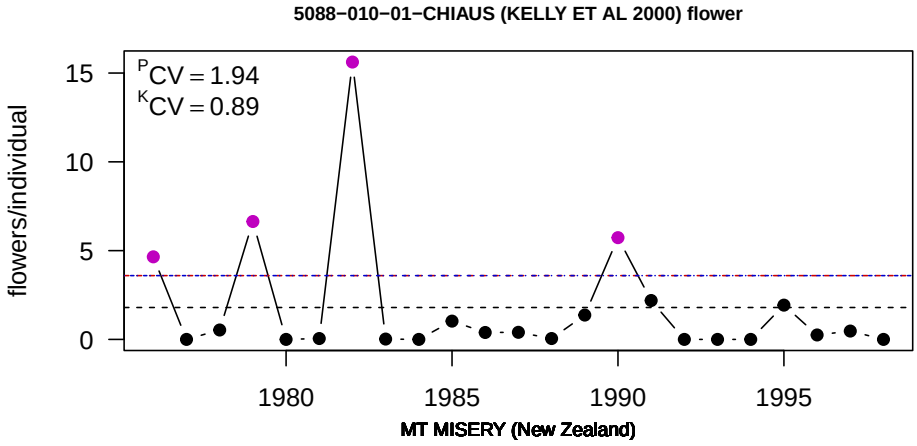

5131−001−01−PINTAE (POMEROY AND KORSTIAN 1949) seed
PCV = 1.13
KCV = 0.75
seeds/m2
0
20
40
60
80
100
120
140
1936
1938
1940
1942
1944
1946
1948
DUKE FOREST (United States of America (the))

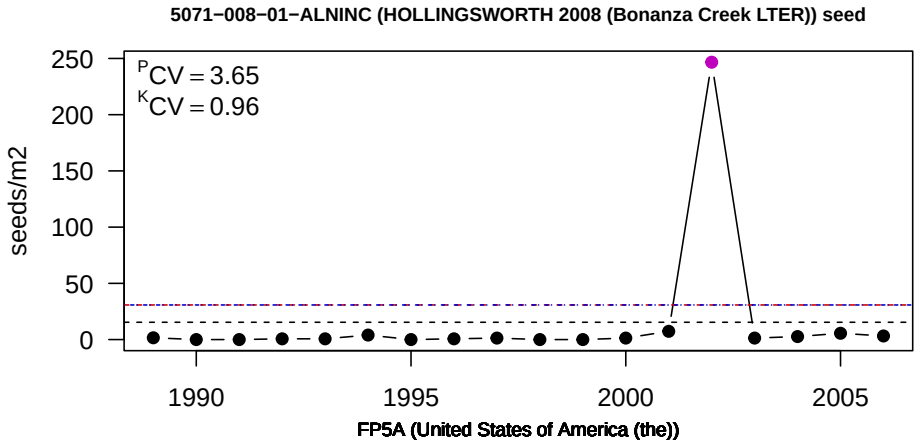

6005−003−01−ARAARA (Calderon_unpublished) cone
PCV = 1.25
KCV = 0.78
cones/individual
0
50
100
150
2005
2010
2015
2020
MAL −ARAU (Chile)


6009−001−01−QUEALB (Steele_unpublished) seed
PCV = 1.34
KCV = 0.8
seeds/individual
0
2
4
6
8
10
2000
2005
2010
2015
Hawk Mountain (United States of America (the))


6184−005−01−FAGSYL (Chianucci and Cutini, unpublished) seed
PCV = 1.48
KCV = 0.83
g/m2
0
50
100
150
2004
2006
2008
2010
2012
2014
2016
fsTScate (Italy)


5006−001−01−PHILAT (Bogdziewidcz_2020a) fruit
PCV = 1.64
KCV = 0.85
g/m2
700
600
500
400
300
200
100
0
2000
2005
2010
2015
Prades drought experiment (Spain)


5088−010−01−CHIAUS (KELLY ET AL 2000) flower
PCV = 1.94
KCV = 0.89
flowers/individual
15
10
5
0
1980
1985
1990
1995
MT MISERY (New Zealand)


5071−008−01−ALNINC (HOLLINGSWORTH 2008 (Bonanza Creek LTER)) seed
PCV = 3.65
KCV = 0.96
seeds/m2
250
200
150
100
50
0
1990
1995
2000
2005
FP5A (United States of America (the))

References

[1] A. Hacket-Pain, J.J. Foest, I.S. Pearse, J.M. LaMontagne, W.D. Koenig, G. Vacchiano, M. Bogdziewicz, T. Caignard, P. Celebias, J. van Dormolen, M. Fernández-Martínez, J.V. Moris, C. Palaghianu, M. Pesendorfer, A. Satake, E. Schermer, A.J. Tanentzap, P.A. Thomas, D. Vecchio, A.P. Wion, T. Wohlgemuth, T. Xue, K. Abernethy, M.-C. Aravena A., M.D. Barrera, J.H. Barton, S. Boutin, E.R. Bush, S.D. Calderón, F.S. Carevic, C.V. de Castilho, J.M. Cellini, C.A. Chapman, H. Chapman, F. Chianucci, P. da Costa, L. Croisé, A. Cutini, B. Dantzer, R.J. DeRose, J.-T. Dikangadissi, E. Dimoto, F.L. da Fonseca, L. Gallo, G. Gratzer, D.F. Greene, M.A. Hadad, A.H. Herrera, K.J. Jeffery, J.F. Johnstone, U. Kalbitzer, W. Kantorowicz, C.A. Klimas, J.G.A. Lageard, J. Lane, K. Lapin, M. Ledwoń, A.C. Leeper, M.V. Lencinas, A.C. Lira-Guedes, M.C. Lordon, P. Marchelli, S. Marino, H. Schmidt Van Marle, A.G. McAdam, L.R. . Momont, M. Nicolas, L.H. de Oliveira Wadt, P. Panahi, G. Martínez Pastur, T. Patterson, P. Luis Peri, Ł. Piechnik, M. Pourhashemi, C. Espinoza Quezada, F.A. Roig, K. Peña Rojas, Y. Micaela Rosas, S. Schueler, B. Seget, R. Soler, M.A. Steele, M. Toro-Manríquez, C.E.G. Tutin, T. Ukizintambara, L. White, B. Yadok, J.L. Willis, A. Zolles, M. Żywiec, and D. Ascoli. MASTREE+: time-series of plant reproductive effort from six continents. *Global Change Biology*, 00:1–17, 2022.

[2] T.O. Kvålseth. Coefficient of variation: the second-order alternative. *Journal of Applied Statistics*, 44:402–415, 2017.

[3] J.M. LaMontagne. *Spatial and temporal variability in white spruce (Picea glauca) cone production: individual and population responses of North American red squirrels (Tamiasciurus hudsonicus)*. PhD thesis, University of Alberta Edmonton, Alberta, Canada, 2007.

[4] J.M. Lamontagne and S. Boutin. Local-scale synchrony and variability in mast seed production patterns of *Picea glauca*. *Journal of Ecology*, 95(5):991–1000, 2007.

[5] J.R. Lobry, M.-C. Bel-Venner, M. Bogdziewicz, A. Hacket-Pain, and S. Venner. The CV is dead, long live the CV! *Methods in Ecology and Evolution*, 14:2780–2786, 2023.

[6] K. Pearson. VII. mathematical contributions to the theory of evolution.—III. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, (187):253–318, 1896.