www.modelia.org

Approche de *machine learning* pour prédire les cultures

François Brun
(Acta, les Instituts Techniques Agricoles)

Séminaire. *L'utilisation des données de télédétection et des méthodes de deep learning pour traiter de problématiques autour des paysages agricoles.* 27-28 juin 2019 à Paris

Objectifs

- Illustrer de manière pratique comment on peut utiliser les images satellites pour prédire la culture en place, avec une approche d'apprentissage machine
  - Utilisation d'image Sentinel
  - Utilisation du RPG pour annoter
  - Random forest pour la procédure de machine learning
- Effort pour vous fournir tous les scripts, fonctionnel, mais cela reste une base à améliorer (optimiser) car temps de calcul très long.

Etape 1. Les données sentinelles

Acronym	Central (nm)	Width (nm)	Spatial resolution (m)	Potential Applications
B1	443	20	60	Atmosphere
B2	490	65	10	Atmosphere
B3	560	35	10	Vegetation
B4	665	30	10	Vegetation
B5	705	15	20	Vegetation
B6	740	15	20	Vegetation
B7	783	20	20	Vegetation
B8	842	115	10	Vegetation
B8a	865	20	20	Vegetation
B9	945	20	60	Atmosphere
B10	1375	30	60	Atmosphere
B11	1610	90	20	Vegetation
B12	2190	180	20	Vegetation

Etape 1. Les données sentinelles

Animation ndvi

Etape 1. Les données sentinelles

- **Accéder aux données sentinel**
- **Téléchargement en ligne**
- via R et (pas de level-2A pour 2016
  - 01_sat_sentinel_download.r
- Level-1C => level-2A (2016) – sen2cor
    - 01sup_sat_sentinel_1C_to_2A.r
- à la main : par exemple (cf autres presentations)
  - https://theia.cnes.fr/

Etape 1. Les données sentinelles

- Utilisation des données 2017
- Image avec nuages <5%
- 7 images disponibles que l'on va utiliser

date	tile	type	names1	names2		
20170225	**T31TCJ**	**L1C**	S2A_MSIL1C_20170225T105021_N0204_R051_T31TCJ_20170225T105020.SAFE	L1C_T31TCJ_A008771_20170225T105020		
20170413	**T31TCJ**	**L1C**	S2A_MSIL1C_20170413T104021_N0204_R008_T31TCJ_20170413T104021.SAFE	L1C_T31TCJ_A009443_20170413T104021		
20170602	**T31TCJ**	**L1C**	S2A_MSIL1C_20170602T104021_N0205_R008_T31TCJ_20170602T104212.SAFE	L1C_T31TCJ_A010158_20170602T104212		
20170622	**T31TCJ**	**L1C**	S2A_MSIL1C_20170622T104021_N0205_R008_T31TCJ_20170622T104021.SAFE	L1C_T31TCJ_A010444_20170622T104021		
20170705	**T31TCJ**	**L1C**	S2A_MSIL1C_20170705T105031_N0205_R051_T31TCJ_20170705T105605.SAFE	L1C_T31TCJ_A010630_20170705T105605		
20170814	**T31TCJ**	**L2A**	S2A_MSIL2A_20170814T105031_N0205_R051_T31TCJ_20170814T105517.SAFE	L2A_T31TCJ_A011202_20170814T105517		
20171013	**T31TCJ**	**L2A**	S2A_MSIL2A_20171013T105031_N0205_R051_T31TCJ_20171013T105315.SAFE	L2A_T31TCJ_A012060_20171013T105315		

Etape 2. Les données RPG

- Registre parcellaire graphique
  - http://professionnels.ign.fr/rpg#tab-3

Etape 2. Les données RPG

- Année 2017 disponible
- Simplification
  - 2017 et occitanie uniquement (question de temps d'extraction)
  - restriction à quelques cultures
    - CZH Colza d'hiver Colza
    - TRN Tournesol
    - MIS Mais
    - BTH Blé tendre d'hiver
  - 02_RPG_shapefile_extract.r
  - Output : rpg_poly_4c.rda

Etape 3. Constitutions du jeu de données annoté

- Extraire les données satellite sentinel correspondant aux parcelles RPG sélectionné
- Pour chaque image
  - 03_sat_sentinel_RPG_shapefile_extract.r
  - Output : dt3_L1C_T31TCJ_1.5_43.5_10000_20170225.rda
  - avec normalisation
  - 12 colonnes
    - id_parcelle, code_culture
    - 10 bandes spectrales
  - 618542 lignes correspondant aux pixels 10m
    - Regroupé en 1261 parcelles

BTH TRN MIS CZH

322 702 207 30

Etape 4. machine learning

- 04b_sentinel_RPG_machinelearning_severaldates.r
- 1) Agréger les données des différentes images
- 618542 lignes(pixels) X 52 colonnes (5 mois*10B+2)

id_parcelle code_culture B02_M0225 B03_M0225 B04_M0225 B08_M0225 B05_M0225
1 598430 BTH -0.1042163 -0.3229245 -0.07240643 -0.9301808 -0.4465191
B06_M0225 B07_M0225 B8A_M0225 B11_M0225 B12_M0225 B02_M0413 B03_M0413 B04_M0413
1 -0.8103806 -0.8875366 -0.8435823 -0.3829222 0.05112776 -0.6320252 -0.6097437 -0.5279185
B08_M0413 B05_M0413 B06_M0413 B07_M0413 B8A_M0413 B11_M0413 B12_M0413 B02_M0602
1 -0.630775 -0.5987597 -0.6407902 -0.6321541 -0.6323081 -0.599892 -0.5417055 -0.6444638
B03_M0602 B04_M0602 B08_M0602 B05_M0602 B06_M0602 B07_M0602 B8A_M0602 B11_M0602
1 -0.6275538 -0.5701629 -0.6469542 -0.6160388 -0.6474965 -0.6464624 -0.6463035 -0.6062286
B12_M0602 B02_M0622 B03_M0622 B04_M0622 B08_M0622 B05_M0622 B06_M0622 B07_M0622
1 -0.5535167 -0.6266638 -0.6179268 -0.5770553 -0.6207712 -0.6031526 -0.6204206 -0.6183736
B8A_M0622 B11_M0622 B12_M0622 B02_M0814 B03_M0814 B04_M0814 B08_M0814 B05_M0814
1 -0.6209396 -0.6079358 -0.5868632 -0.835708 -0.9388076 -0.8568184 0.200109 -0.7056911
B06_M0814 B07_M0814 B11_M0814 B12_M0814 B8A_M0814
1 0.7137623 0.7415566 -0.5630981 -0.7253052 0.8387944

Etape 4. machine learning

- 2) Sélectionner les mois

```
  month_max=6
> colsel = c(1,2, grep(pattern = paste0("M0[1-",month_max,"]"), names(dt3_all)))
> dt3_all = dt3_all[,colsel]
> dim(dt3_all)
[1] 618542     42
> names(dt3_all)
 [1] "id_parcelle"  "code_culture" "B02_M0225"    "B03_M0225"    "B04_M0225"
 [6] "B08_M0225"    "B05_M0225"    "B06_M0225"    "B07_M0225"    "B8A_M0225"
[11] "B11_M0225"    "B12_M0225"    "B02_M0413"    "B03_M0413"    "B04_M0413"
[16] "B08_M0413"    "B05_M0413"    "B06_M0413"    "B07_M0413"    "B8A_M0413"
[21] "B11_M0413"    "B12_M0413"    "B02_M0602"    "B03_M0602"    "B04_M0602"
[26] "B08_M0602"    "B05_M0602"    "B06_M0602"    "B07_M0602"    "B8A_M0602"
[31] "B11_M0602"    "B12_M0602"    "B02_M0622"    "B03_M0622"    "B04_M0622"
[36] "B08_M0622"    "B05_M0622"    "B06_M0622"    "B07_M0622"    "B8A_M0622"
[41] "B11_M0622"    "B12_M0622"
```

Etape 4. machine learning

- 52 variables
  - 1 variables à expliquer : code_culture
  - 50 variables explicative : BXX_MYYjj, variable combinée, 10 bandes XX pour le mois YY, jour jj de l'année 2017
  - 1 variable permettant d'identifier la parcelle
    - 1) diviser le jeu de données
      - un jeu de données de parcelle d'apprentissage
      - un pour l'évaluation des performances de prédiction
    - 2) permet de regrouper les résultats des pixels par parcelle
      - Intérêt par parcelle et la structure est « stable » d'une année à l'autre
      - Cela permet d'améliorer la qualité de prédiction, en prenant la prédiction par pixel majoritaire

Résultats – tous les mois - N=50000

- En ajustement

```
> system.time(rf.model <- randomForest(code_culture ~ ., data = dt3_sel))
utilisateur     système      écoulé
    200.83       0.53      202.09
> rf.model
Call:
 randomForest(formula = code_culture ~ ., data = dt3_sel)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 7
        OOB estimate of  error rate: 2.55%
Confusion matrix:
    BTH   TRN  MIS  CZH class.error
BTH 9915   169  147   15  0.03230529
TRN  157 29122  171    4  0.01127181
MIS  118   327 8031    6  0.05317142
CZH  135    25    2 1656  0.08910891
```

Résultats – tous les mois - N=50000

Et la qualité d'ajustement ?

1) Ajustement sur jeu d'apprentissae

```
rf.model <- randomForest(code_culture ~ ., data = dt3_sel)
```

2) prediction sur jeu de test

```
Ypred = predict(rf.model, newdata = dt3_test )
dt3_test2 = cbind(dt3_test[,c("id_parcelle","code_culture")],Ypred)
table(dt3_test2$code_culture,dt3_test2$Ypred)
```

```
      BTH    TRN    MIS    CZH
 BTH  48312    695   1339    632
 TRN    641 128889   1298     64
 MIS    757   5004  34011      7
 CZH   1233    259     10   2456
>
```

Résultats – tous les mois - N=50000

Statistique de performance de prédiction

```
caret::confusionMatrix(dt3_test2$code_culture, dt3_test2$Ypred)
```

```
Confusion Matrix and Statistics

          Reference
Prediction    BTH    TRN    MIS    CZH
       BTH  48312    695   1339    632
       TRN    641 128889   1298     64
       MIS    757   5004  34011      7
       CZH   1233    259     10   2456
```

```
Overall Statistics

               Accuracy : 0.9471
                 95% CI : (0.9461, 0.948)
    No Information Rate : 0.5977
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9077

 Mcnemar's Test P-Value : < 2.2e-16
```

```
Statistics by Class:
                     Class: BTH Class: TRN Class: MIS Class: CZH
Sensitivity              0.9484     0.9558     0.9278    0.77746
Specificity              0.9847     0.9779     0.9695    0.99325
Pos Pred Value           0.9477     0.9847     0.8550    0.62052
Neg Pred Value           0.9849     0.9371     0.9858    0.99683
Prevalence               0.2258     0.5977     0.1625    0.01400
Detection Rate           0.2141     0.5713     0.1508    0.01089
Detection Prevalence     0.2260     0.5802     0.1763    0.01754
Balanced Accuracy        0.9665     0.9669     0.9486    0.88535
```

Résultats – tous les mois - N=50000

Et si on rassembler l'info par parcelle ?

```
> res=data.frame()
> for(p in unique(dt3_test2$id_parcelle) ){
+ code_cultureP_dist = table(dt3_test2[dt3_test2$id_parcelle==p,"Ypred"])
+ code_cultureP_maj = names(code_cultureP_dist[code_cultureP_dist==max(code_cultureP_dist)])[1]
+ code_cultureRPG = paste(unique(dt3_test2[dt3_test2$id_parcelle==p,"code_culture"]))
+ res1 = data.frame(p=p,code_cultureRPG=code_cultureRPG,code_cultureP_maj=code_cultureP_maj)
+ res=rbind(res,res1)
+ }
caret::confusionMatrix(res$code_cultureRPG, res$code_cultureP_maj)
Confusion Matrix and Statistics
          Reference
Prediction BTH TRN MIS CZH
       BTH 117   0   5   0
       TRN   1 255   2   0
       MIS   0   8  66   0
       CZH   3   0   0   4
Overall Statistics
                Accuracy : 0.9588
                  95% CI : (0.9364, 0.975)
     No Information Rate : 0.5705
     P-Value [Acc > NIR] : < 2.2e-16
                   Kappa : 0.9296
Mcnemar's Test P-Value : NA
```

```
Statistics by Class:

                     Class: BTH Class: TRN Class: MIS Class: CZH
Sensitivity              0.9669     0.9696     0.9041   1.000000
Specificity              0.9853     0.9848     0.9794   0.993435
Pos Pred Value           0.9590     0.9884     0.8919   0.571429
Neg Pred Value           0.9882     0.9606     0.9819   1.000000
Prevalence               0.2625     0.5705     0.1584   0.008677
Detection Rate           0.2538     0.5531     0.1432   0.008677
Detection Prevalence     0.2646     0.5597     0.1605   0.015184
Balanced Accuracy        0.9761     0.9772     0.9417   0.996718
```

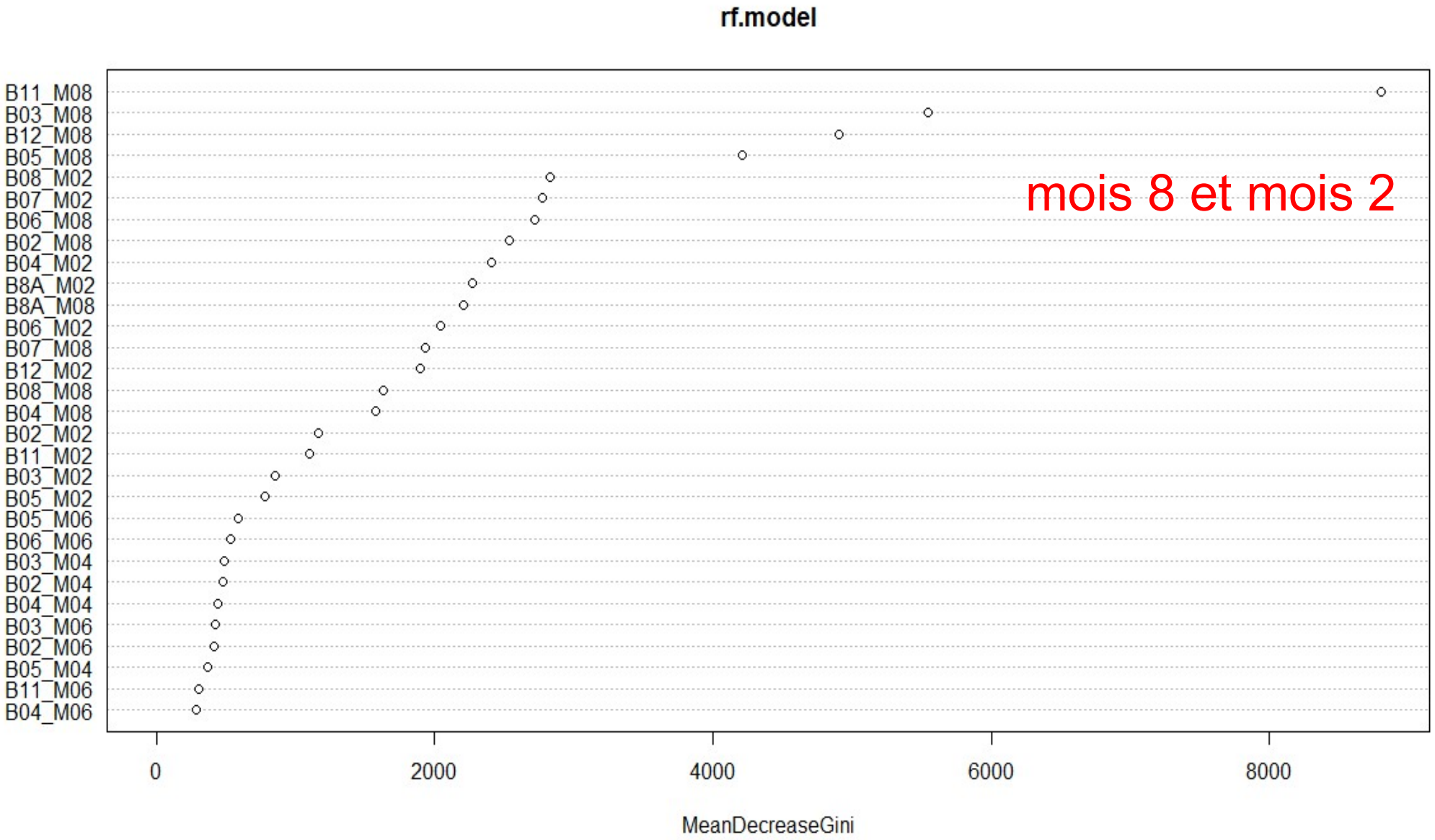
Résultats – tous les mois - N=50000

N parcelles	Time	Npix	pixel de parcelle	regroupe parcelle
800 / 461	900s	100000	Accuracy : 0.9456 95% CI : (0.9447, 0.9466) Kappa : 0.907	Accuracy : 0.9696 95% CI : (0.9496, 0.9833) Kappa : 0.9484

Résultats – comparaison d'échantillonnage

N parcelles	Time	Npix	pixel de parcelle	regroupe parcelle
800 / 461	900s	200000		
800 / 461		100000	**Accuracy : 0.9456** 95% CI : (0.9447, 0.9466) Kappa : 0.907	**Accuracy : 0.9696** 95% CI : (0.9496, 0.9833) Kappa : 0.9484
800 / 461	275s	50000	**Accuracy : 0.9471** **95% CI : (0.9461, 0.948)** Kappa : 0.9077	**Accuracy : 0.9588** 95% CI : (0.9364, 0.975) Kappa : 0.9296
800 / 461	23s	10000	Accuracy : 0.9479 95% CI : (0.9235, 0.9664) Kappa : 0.9132	Accuracy : 0.9309 95% CI : (0.9298, 0.9319) Kappa : 0.8873
800 / 461	1.4s	1000	Accuracy : 0.9143 95% CI : (0.9132, 0.9155) Kappa : 0.8526	Accuracy : 0.9523 95% CI : (0.9286, 0.9699) Kappa : 0.9185
800 / 461	0.5s	500	Accuracy : 0.8855 95% CI : (0.8842, 0.8867) Kappa : 0.7981	Accuracy : 0.9154 95% CI : (0.8862, 0.9392) Kappa : 0.8595

Résultats – variables explicatives



Résultats – et en cours de saison ?

- Avant fin avril ?

- Plus que 20 variables explicatives (fév et avril)

```
 [1] "id_parcelle"  "code_culture" "B02_M0225"   "B03_M0225"   "B04_M0225"
 [6] "B08_M0225"   "B05_M0225"   "B06_M0225"   "B07_M0225"   "B8A_M0225"
[11] "B11_M0225"   "B12_M0225"   "B02_M0413"   "B03_M0413"   "B04_M0413"
[16] "B08_M0413"   "B05_M0413"   "B06_M0413"   "B07_M0413"   "B8A_M0413"
[21] "B11_M0413"   "B12_M0413"
```

Résultats – et en cours de saison ?

Par pixel

Prediction	Reference BTH	TRN	MIS	CZH
BTH	37631	3954	8	272
TRN	2346	113838	3554	5
MIS	2528	40309	5558	522
CZH	602	471	2	3755

Overall Statistics

Accuracy : **0.7466**

95% CI : (0.7447, 0.7484)

No Information Rate : 0.7363

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5321

Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

	Class: BTH	Class: TRN	Class: MIS	Class: CZH
Sensitivity	0.8730	0.7179	0.60930	0.82455
Specificity	0.9754	0.8960	0.78976	0.99490
Pos Pred Value	0.8989	0.9507	0.11362	0.77743
Neg Pred Value	0.9684	0.5321	0.97859	0.99620
Prevalence	0.2002	0.7363	0.04236	0.02115
Detection Rate	0.1747	0.5286	0.02581	0.01744
Detection Prevalence	0.1944	0.5560	0.22715	0.02243
Balanced Accuracy	0.9242	0.8070	0.69953	0.90973

Résultats – et en cours de saison ?

Par parcelle

```
Prediction BTH TRN MIS CZH
       BTH 104   8   0   0
       TRN   1 253   3   0
       MIS   4  75   2   0
       CZH   3   1   0   7
Overall Statistics
Accuracy : 0.7939
95% CI : (0.7541, 0.8299)
No Information Rate : 0.731
P-Value [Acc > NIR] : 0.001079
Kappa : 0.612
Mcnemar's Test P-Value : NA
Statistics by Class:
Class: BTH Class: TRN Class: MIS Class: CZH
Sensitivity              0.9286     0.7507   0.400000    1.00000
Specificity              0.9771     0.9677   0.826754    0.99119
Pos Pred Value           0.9286     0.9844   0.024691    0.63636
Neg Pred Value           0.9771     0.5882   0.992105    1.00000
Prevalence               0.2430     0.7310   0.010846    0.01518
Detection Rate           0.2256     0.5488   0.004338    0.01518
Detection Prevalence     0.2430     0.5575   0.175705    0.02386
Balanced Accuracy        0.9528     0.8592   0.613377    0.99559
```

Conclusion et perspectives

- Tous les scripts R dispos, plus ou moins commenté
- Pour assurer la qualité de prédiction opérationnel
  - Plus de culture : question de regroupement ?
  - Il faudrait le faire sur plus de surface (France entière)
  - Il faut valoriser plus d'années
    - actuellement RPG X sentinel : uniquement 2016 et 2017
  - Tester différentes méthodes et procédure de validation croisée
  - Utiliser l'a priori de la succession (RPG)
- Mais François ≠ expert de la question
  - Maitrise du temps de téléchargement, traitement
  - Maitrise du métier données sat : pré-traitement atmo par exemple

=> collaboration nécessaire avec les personnes compétentes ! (CESBIO, TETIS,…)