Production Agricole Mondiale

une analyse par l'apprentissage statistique

Orlando (Thomas) Da Costa, Jason Aubriot, Sarah Dupont

Janvier, 2024

Quelques liens

- Les données de ce projet correspondent à une remise en forme des données disponibles à cette adresse `https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset`, provenant initialement de la FAO et de `https://data.worldbank.org/`. Le code associé est disponible ici : `https://github.com/Dac-T/Agri_world_production`

Variables

- 10 types de cultures (*Item*)
  1. Maïs (*Maize*)
  2. Blé (*Wheat*)
  3. Pommes de terre (*Potatoes*)
  4. Sorgho (*Sorghum*)
  5. Riz (*Rice, Paddy*)
  6. Soja (*Soybeans*)
  7. Patates douces (*Sweet Potatoes*)
  8. Manioc (*Cassava*)
  9. Plantains et autres (*Plantains and others*)
  10. Igname (*Yams*)

- 101 pays (*Area*)
- 23 années d'observation (*Year*)
- Quantité de pesticide [T] (*pest*)
- Température moyenne annuelle [°C] (*temp* ou *avg_temp*)
- Pluie annuelle [mm/an] (*rain*)

→ Modifiée pour ajouter de la variabilité, à partir de WorldData.

Tableaux de données

```
1 kable(summary(data), "latex")
```

	Area	Year	Item	yield	rain	pest	temp	
	Cameroon: 230	2012 : 593	Potatoes :2091	Min. : 50	Min. : 51	Min. : 0.0	Min. : 1.30	
	Kenya : 230	2013 : 592	Maize :2028	1st Qu.: 18000	1st Qu.: 608	1st Qu.: 264.5	1st Qu.:16.23	
	Brazil : 207	2010 : 585	Wheat :1810	Median : 39544	Median :1083	Median : 2172.2	Median :20.86	
	Burundi : 207	2011 : 585	Rice, paddy:1502	Mean : 70969	Mean :1155	Mean : 14838.7	Mean :19.97	
	Colombia: 207	2008 : 584	Sorghum :1435	3rd Qu.: 97152	3rd Qu.:1622	3rd Qu.: 13335.2	3rd Qu.:25.87	
	Ecuador : 207	2006 : 583	Soybeans :1242	Max. :501412	Max. :3240	Max. :367778.0	Max. :30.42	
	(Other) :11842	(Other):9608	(Other) :3022					

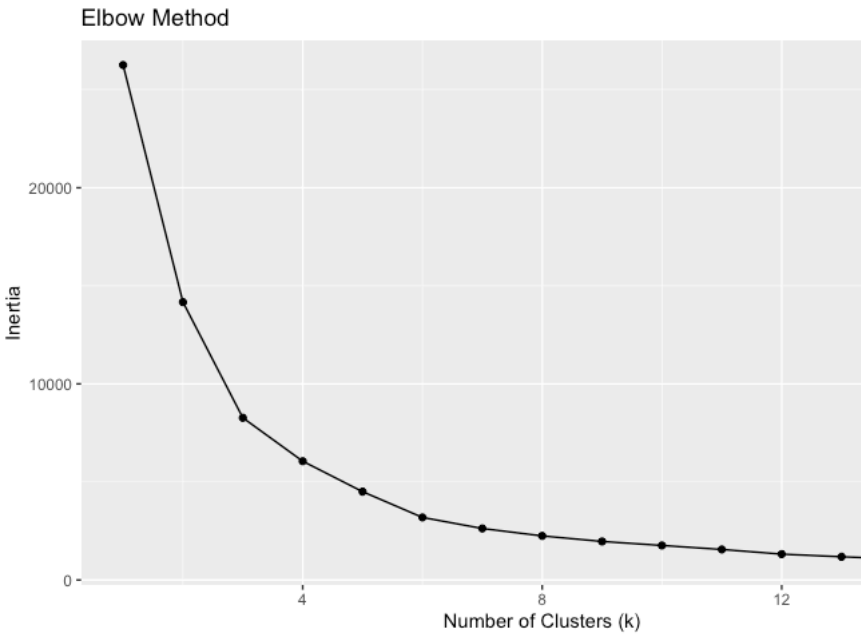
Tableaux de données

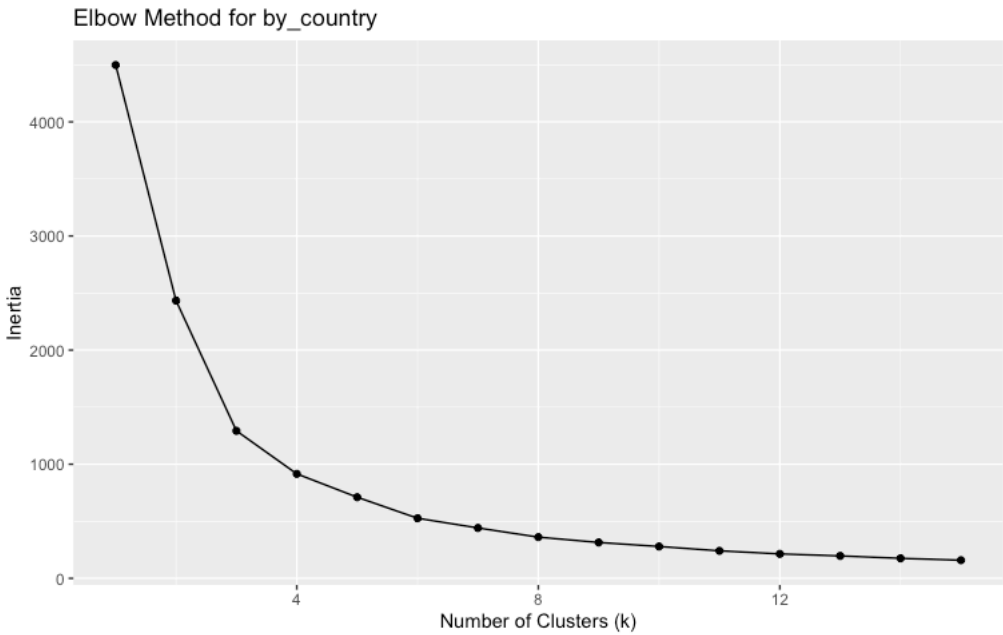
```
1 kable(summary(by_country), "latex")
```

	Area	Year	rain	pest	avg_temp	Maize	Potatoes
	Albania : 23	2012 : 101	Min. : 51	Min. : 0.0	Min. : 1.30	Min. : 849	Min. : 8406
	Algeria : 23	2013 : 101	1st Qu.: 589	1st Qu.: 278.7	1st Qu.:11.25	1st Qu.: 14068	1st Qu.:115023
	Angola : 23	2006 : 100	Median : 847	Median : 1841.2	Median :19.76	Median : 24489	Median :161890
	Argentina: 23	2007 : 100	Mean :1066	Mean : 12781.9	Mean :18.51	Mean : 36980	Mean :182602
	Australia: 23	2008 : 100	3rd Qu.:1513	3rd Qu.: 10960.2	3rd Qu.:25.46	3rd Qu.: 53186	3rd Qu.:236020
	Austria : 23	2009 : 100	Max. :3240	Max. :367778.0	Max. :30.42	Max. :207556	Max. :501412
	(Other) :2112	(Other):1648				NA's :222	NA's :159

	Rice..paddy	Sorghum	Soybeans	Wheat	Cassava	Sweet.potatoes	Plantains.and.others	Yams
	Min. : 2034	Min. : 578	Min. : 50	Min. : 1706	Min. : 11778	Min. : 8799	Min. : 21350	Min. : 11475
	1st Qu.: 22871	1st Qu.: 7192	1st Qu.:10000	1st Qu.:15890	1st Qu.: 58596	1st Qu.: 54090	1st Qu.: 58837	1st Qu.: 62844
	Median : 33554	Median : 12333	Median :16204	Median :24318	Median :100000	Median : 87500	Median : 97024	Median : 92404
	Mean : 37542	Mean : 17995	Mean :16564	Mean :30783	Mean :102705	Mean :104316	Mean :110838	Mean :103759
	3rd Qu.: 48933	3rd Qu.: 24126	3rd Qu.:22214	3rd Qu.:41287	3rd Qu.:132173	3rd Qu.:149969	3rd Qu.:125000	3rd Qu.:132997
	Max. :103895	Max. :206000	Max. :41609	Max. :99387	Max. :385818	Max. :400000	Max. :418505	Max. :250000
	NA's :748	NA's :815	NA's :1008	NA's :440	NA's :1309	NA's :1087	NA's :1786	NA's :1796

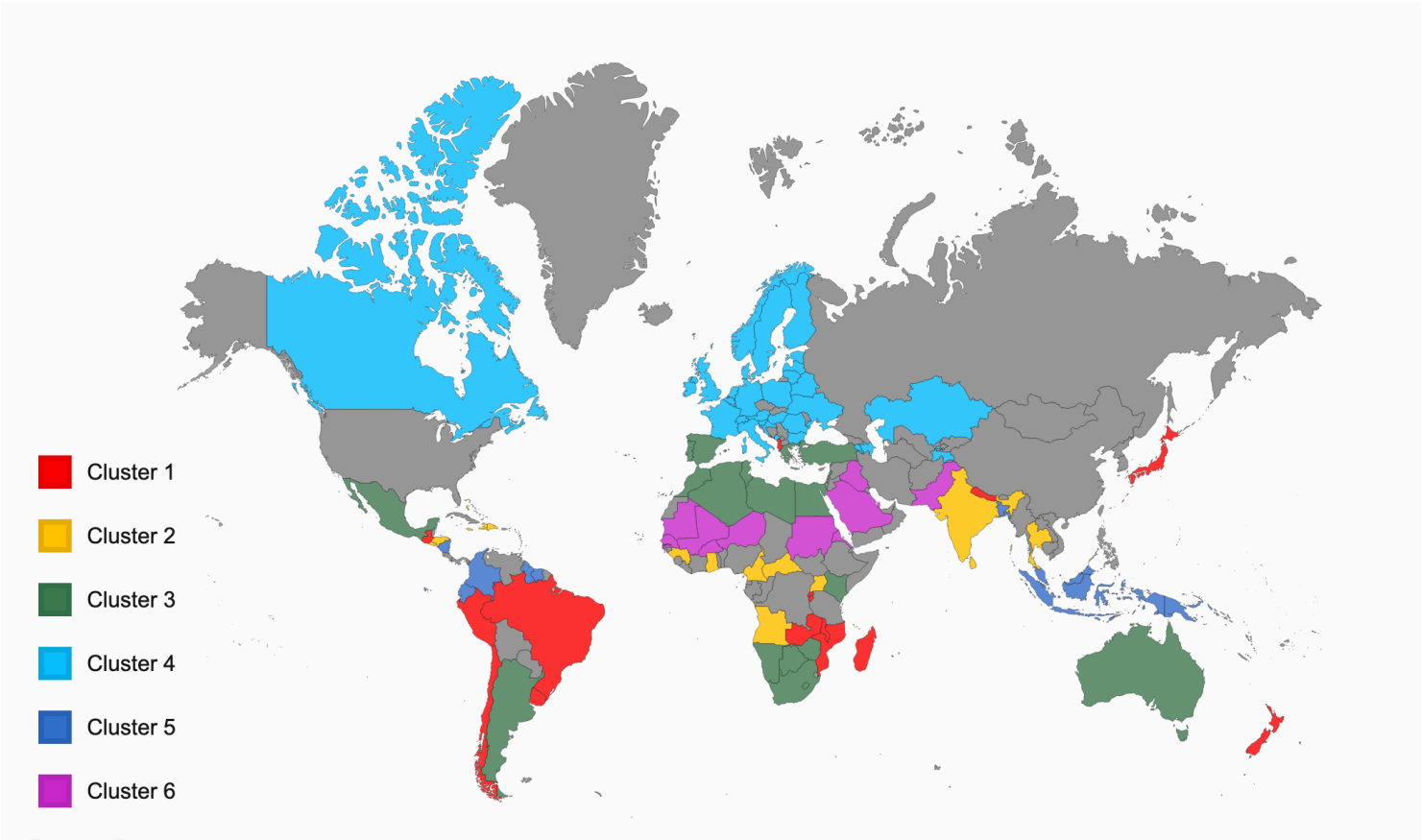
Détermination du nombre de clusters et de la méthode à choisir





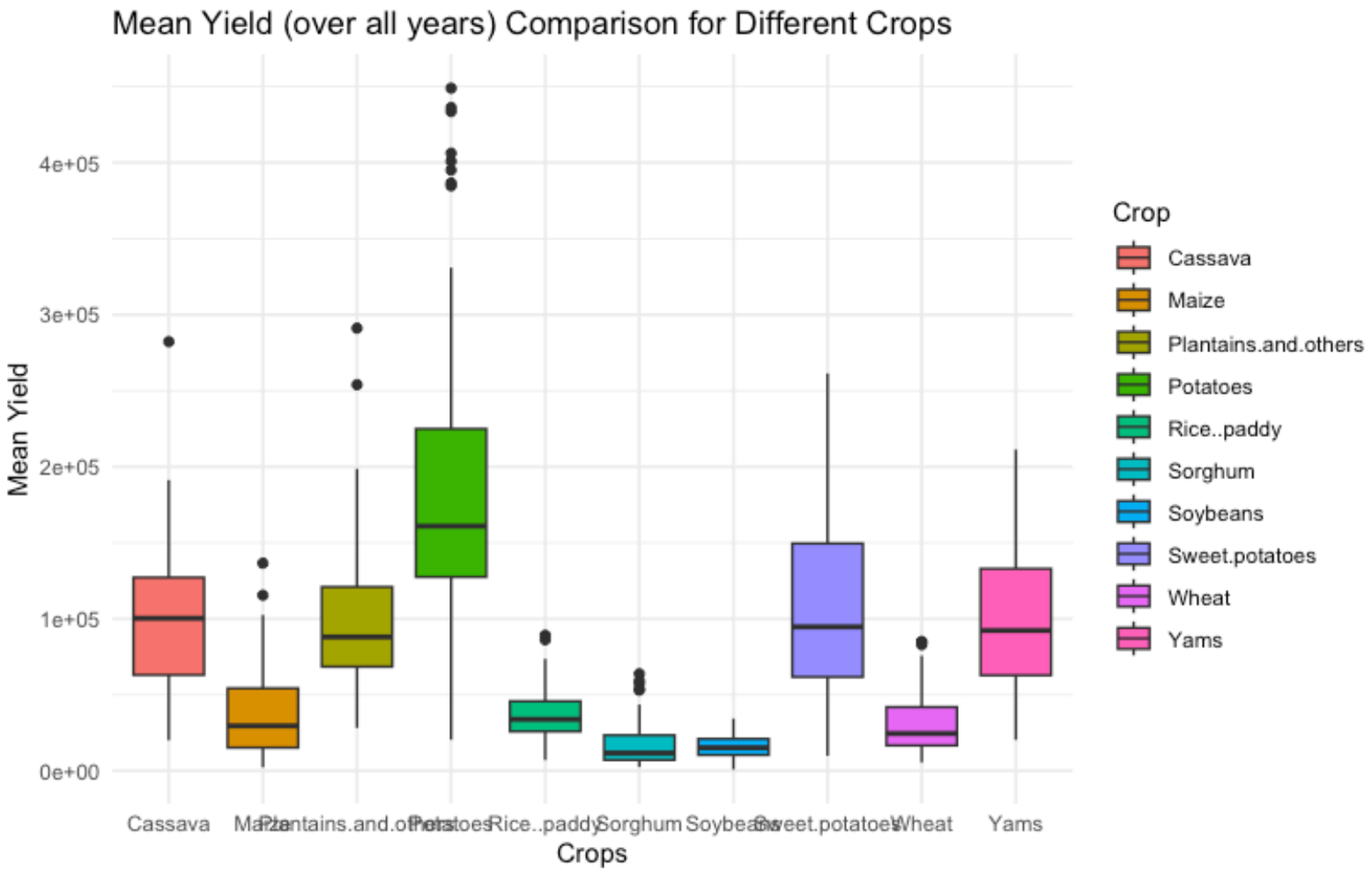
$$I_w(kmeans_{data}) = 3190.019 < I_w(HAC_{data}) = 3617.478$$
$$I_w(kmeans_{by\_country}) : 526.5913 < I_w(HAC_{by\_country}) = 531.2891$$

Résultat de la clusterisation



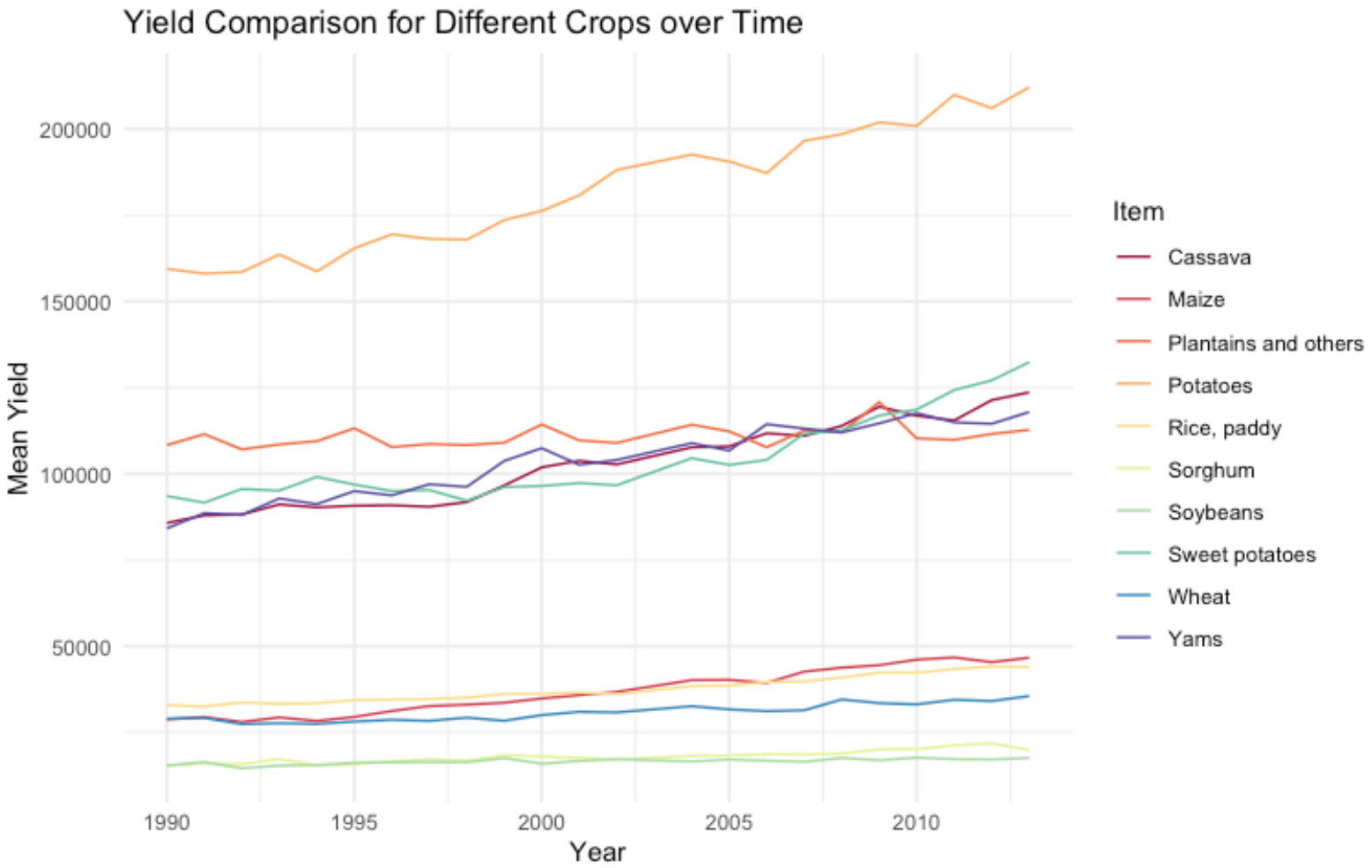
10 pays ("Azerbaijan" ; "Brazil" ; "Egypt" ; "Iraq" ; "Lesotho" ; "Malawi" ; "Montenegro" ; "Romania" ; "Rwanda" ; "Zambia") ont été attribué à deux clusters distincts selon l'année $\longrightarrow$ on les associe au cluster majoritairement attribué sur toutes les années d'observations

Cultures et rendements



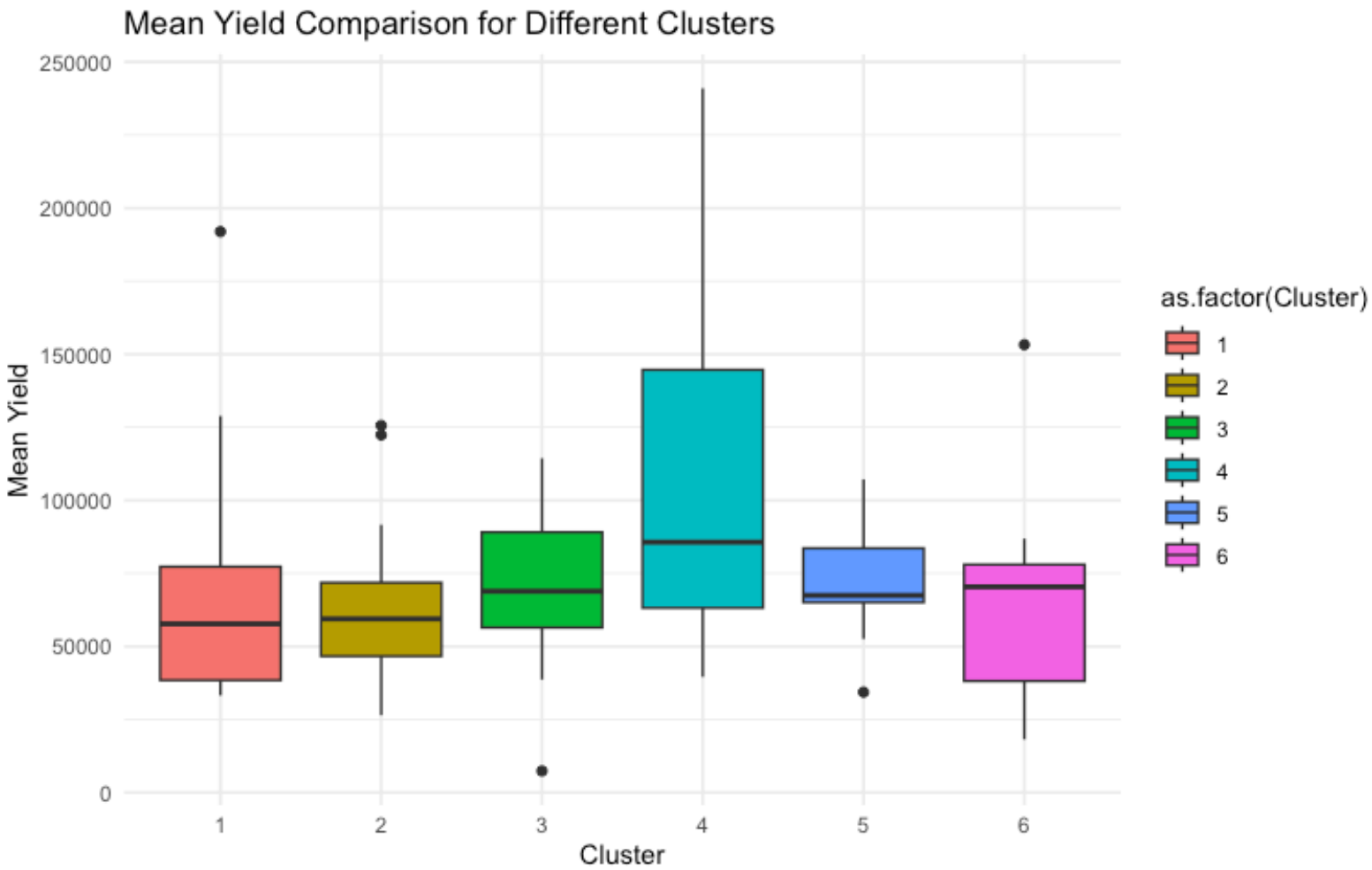
- variabilité des rendements inter-cultures et intra-cultures

Cultures et rendements - évolution temporelle



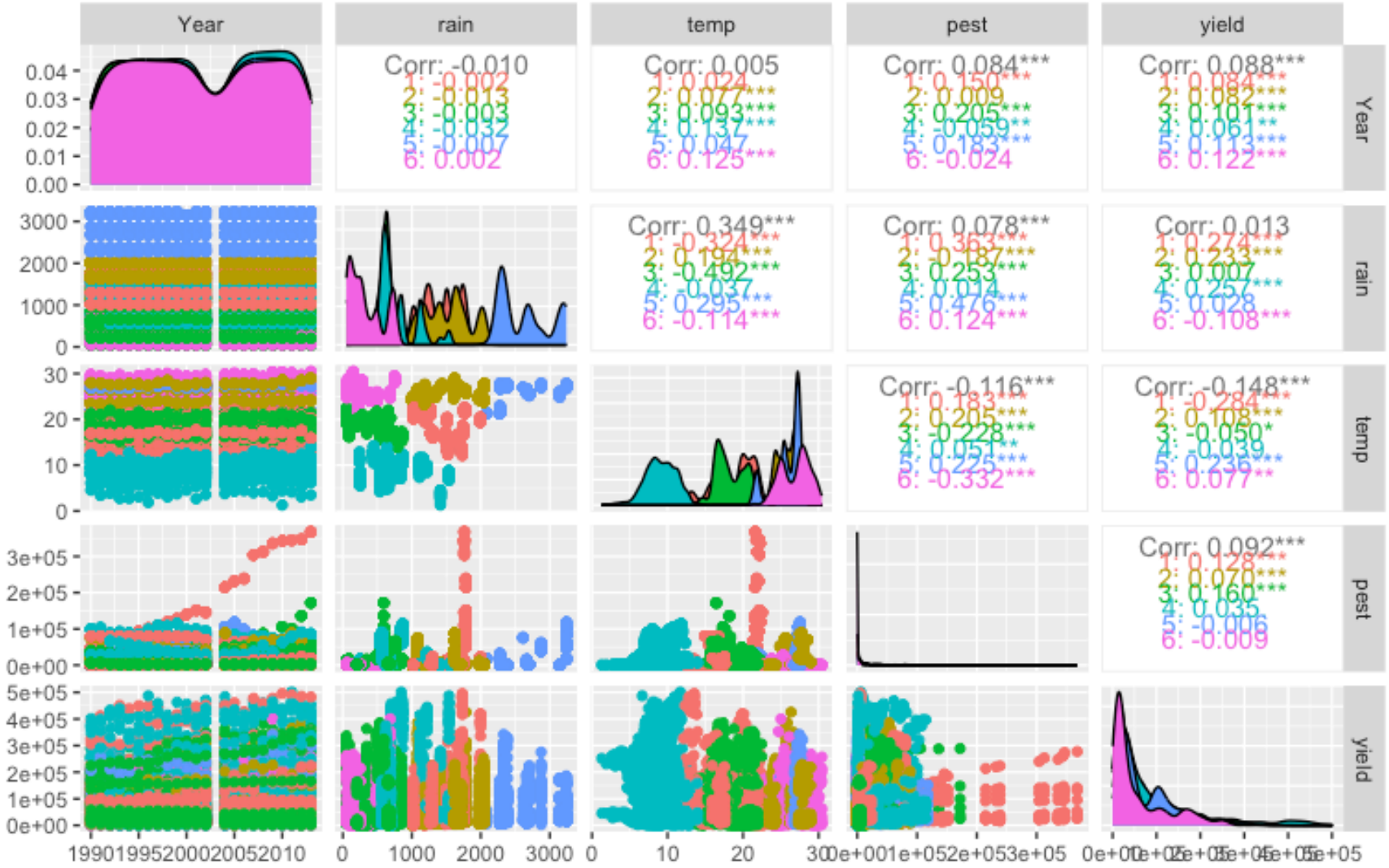
- +25% entre 1990 et 2013

Cluster et rendements

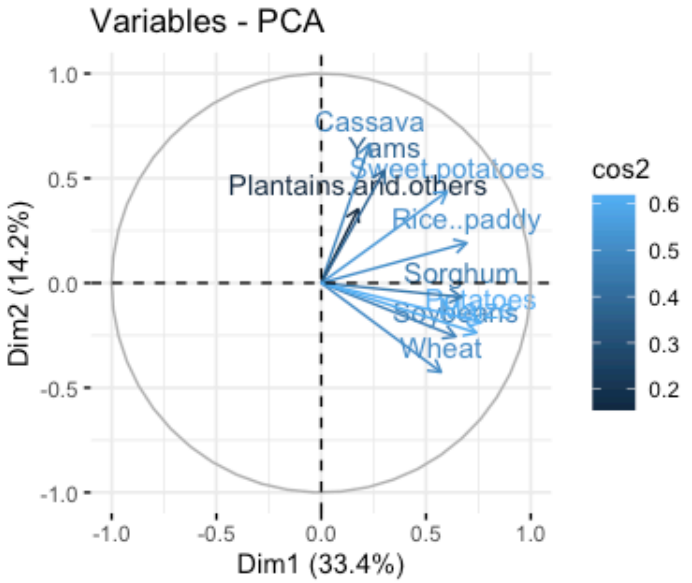


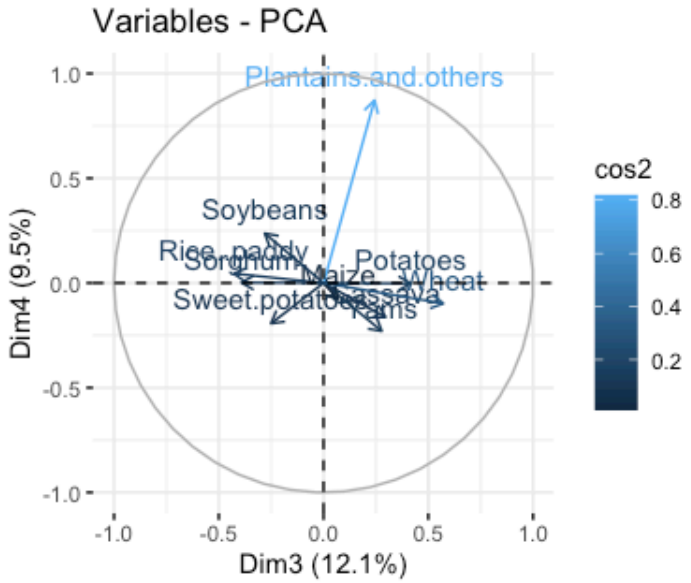
$\Rightarrow$ effet modéré des facteurs météorologique sur les rendements, toutes cultures confondues

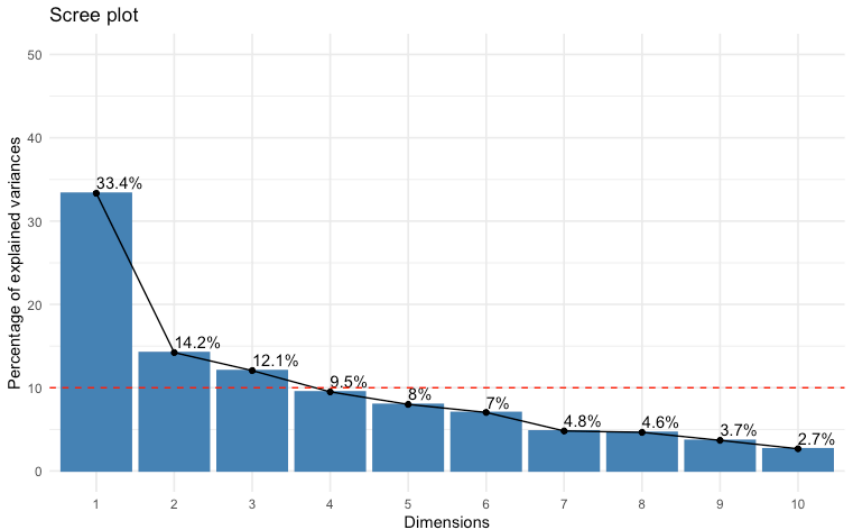
Pairwise plot



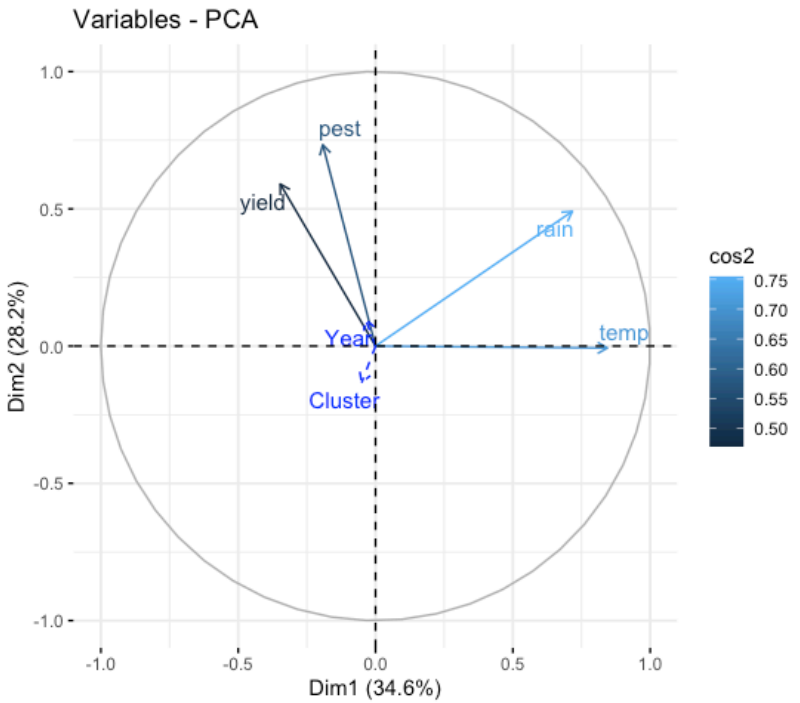
ACP sur les 10 variables de rendement

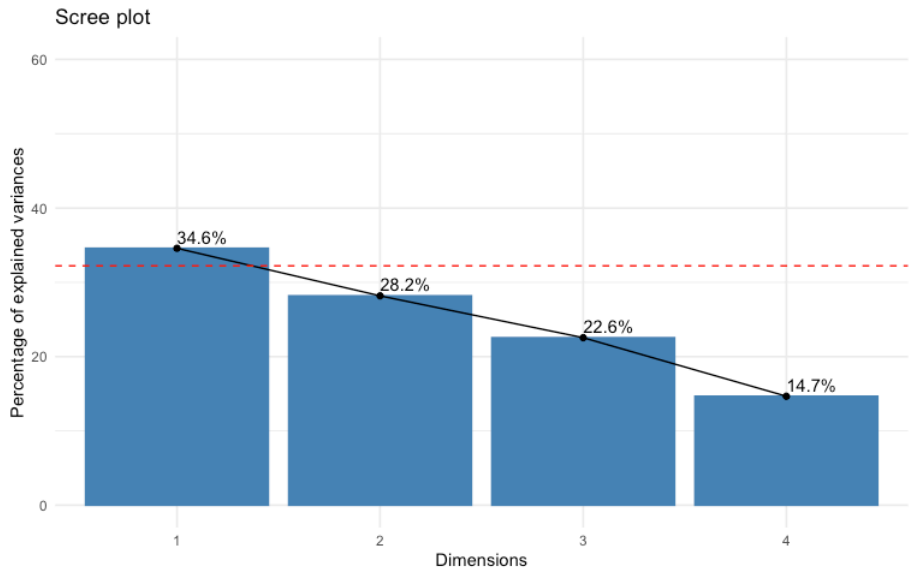






ACP sur les variables de *data*





- Pesticides et rendements relativement colinéaires, mais mal représentés.
- Pertinence d'une PCA sur 4 variables ? Faible.

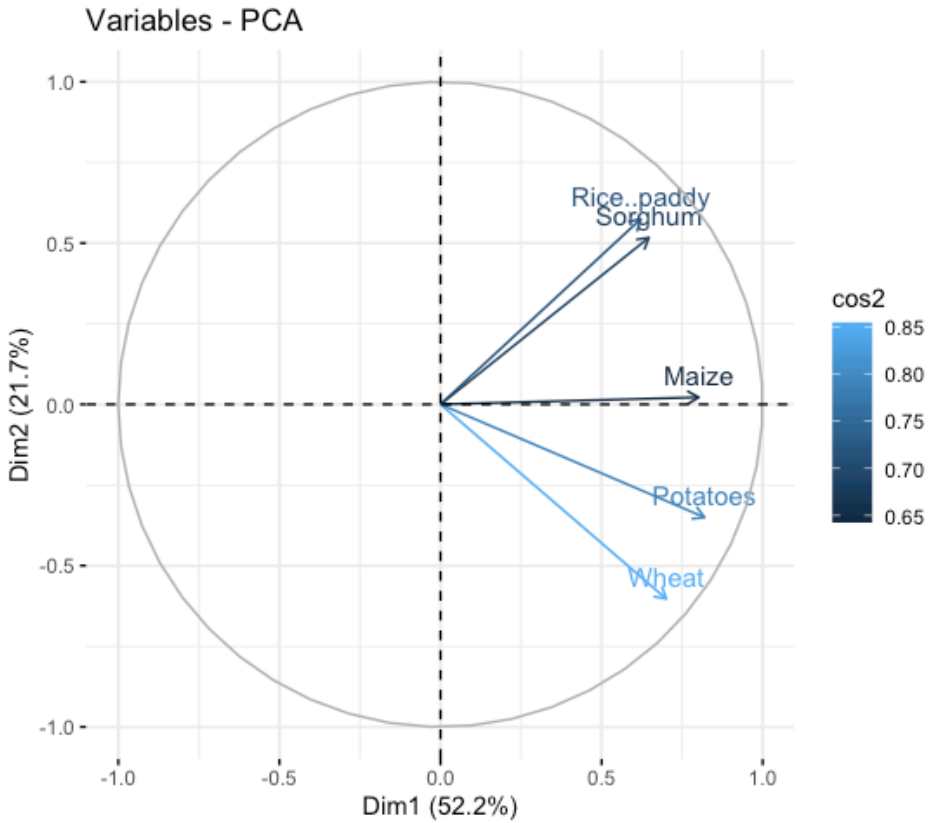
Réduction à 5 cultures

- On retire de nos données :
  1. Soja (*Soybeans*)
  2. Patates douces (*Sweet Potatoes*)
  3. Manioc (*Cassava*)
  4. Plantains et autres (*Plantains and others*)
  5. Igname (*Yams*)

→ Trop de valeurs manquantes

→ Meilleure ACP

⇒ Plan d'expérience complet



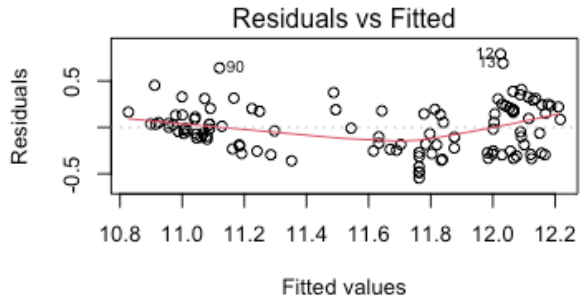
Questions

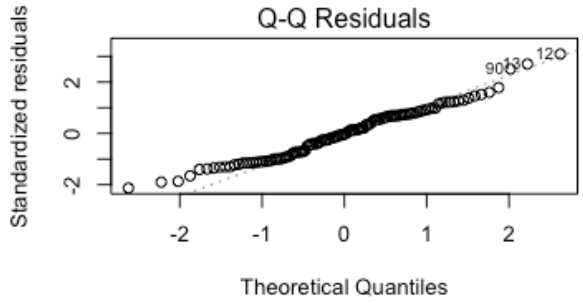
- Au sein de pays ayant les mêmes conditions climatiques, la quantité de pesticide utilisée influence-t-elle le rendement des cultures, pour une culture donnée ?
- Comment expliquer les variations de rendement selon les variables disponibles ?
- Quels sont les effets du cluster et de la culture, après contrôle du volume de pesticides ?

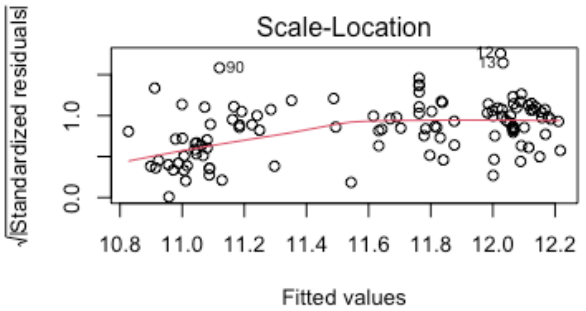
Régressions linéaires simples

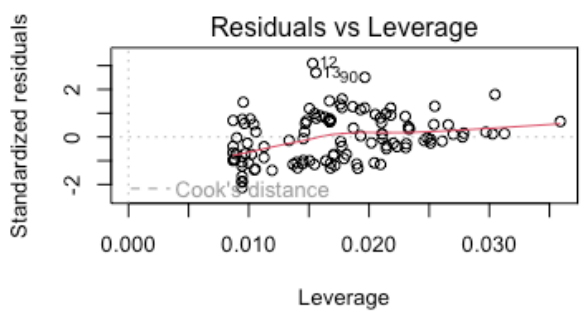
```r
crop_variables = c("Maize", "Wheat", "Rice..paddy", "Sorghum", "Potatoes")
models = list()

for (crop in crop_variables) {
  models[[crop]] <- list(models = list())

  for (N in 1:6) {
    regmod = lm(paste("log(",crop, ") ~ log(pest)"),
    data = fullscnona[fullscnona$Cluster == N, ])

    ad_stat = ad.test(residuals(regmod))$p.value        # Anderson-Darling test
    shapiro = shapiro.test(residuals(regmod))$p.value   # Shapiro test
    summary_info = summary(regmod) # Summary information

    # Store model
    models[[crop]]$models = c(models[[crop]]$models,
                              list(list(N = N, ad = ad_stat, shap = shapiro,
                              model = regmod, summary = summary_info)))
    }
}
# models[["X"]]$models[[N]]$model to get the reg_lin of the crop X and the cluster N
```

Graphes de diagnostic de la meilleure régression





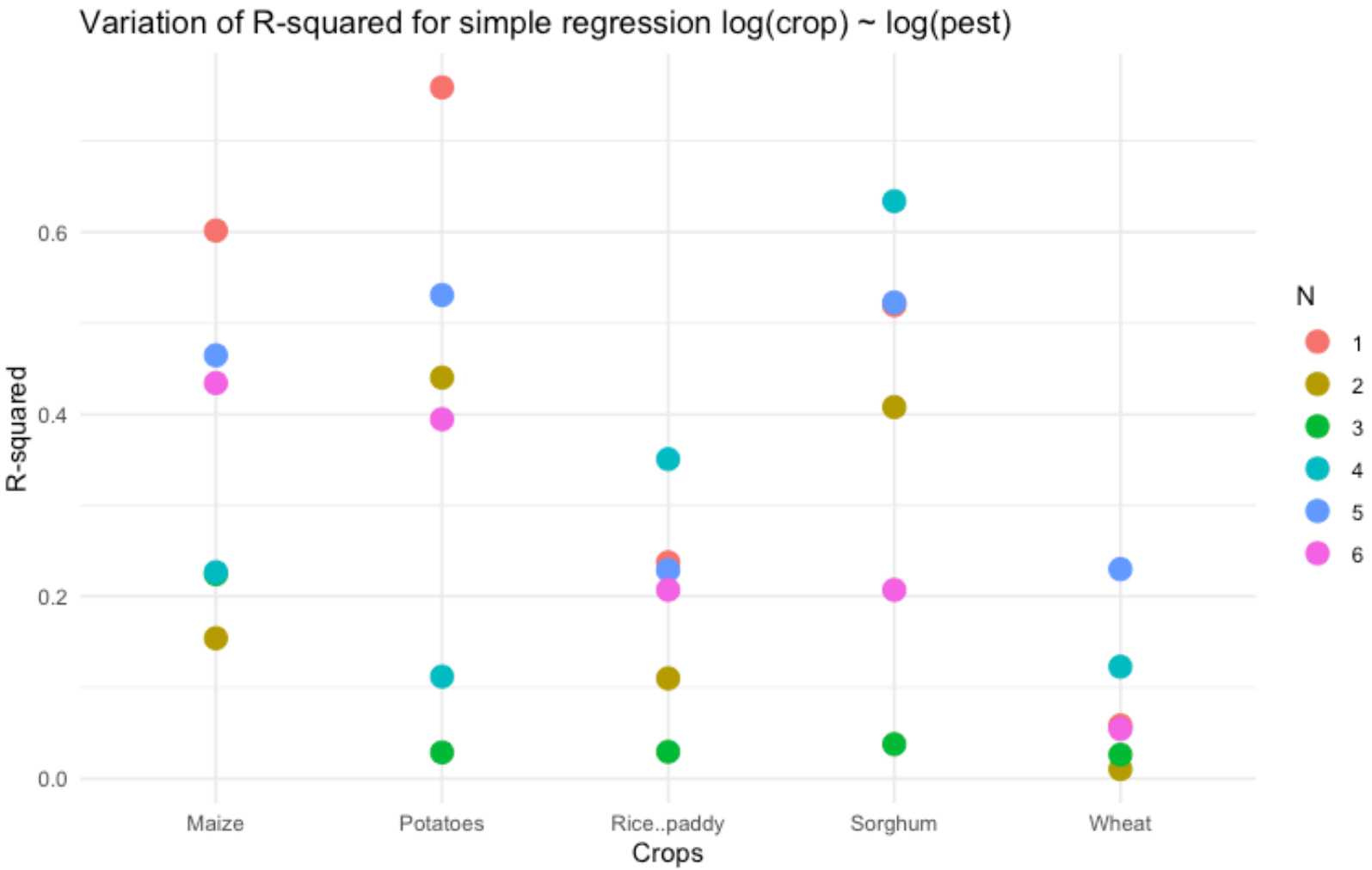




Pour le meilleur modèle :

- $R^2 = 0.76$
- Légère structure dans les résidus
- Hétérocédasticité
- Erreurs gaussiennes !

Coefficients de détermination linéaire R^2 pour tous les clusters et toutes les cultures



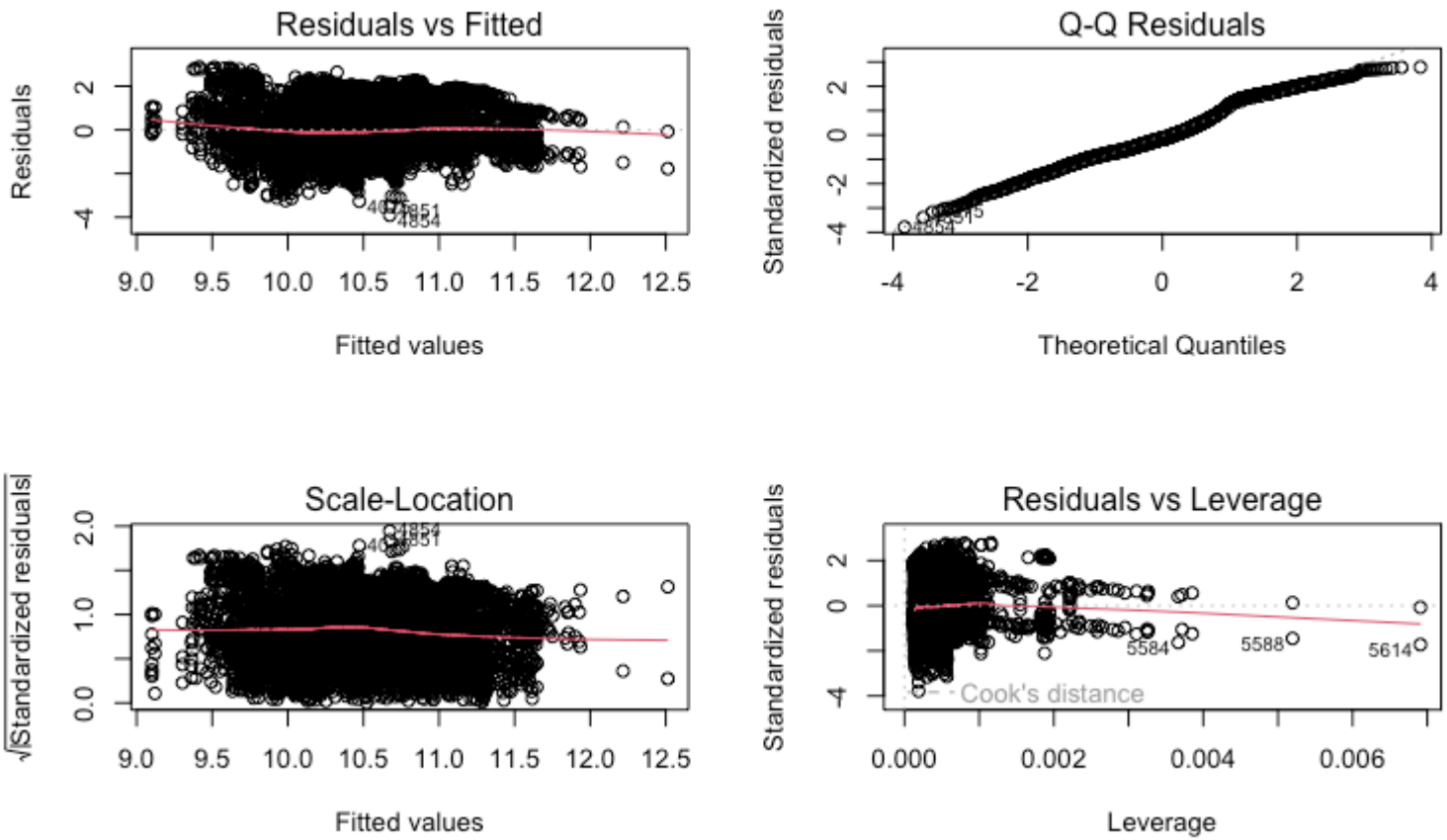
Conclusion 1

- Au sein de pays ayant les mêmes conditions climatiques, la quantité de pesticide utilisée influence-t-elle le rendement des cultures, pour une culture donnée ?

$\Rightarrow$ **La quantité de pesticide ne suffit pas à expliquer systématiquement le rendement d'une culture i dans une zone j**

Régression linéaire multiple

$$log(Yield_i) = \alpha + \beta_1 log(pest_i) + \beta_2 log(temp_i)$$



Régression linéaire multiple

```
> summary(mlr_select)

Call:
lm(formula = log(yield) ~ log(temp) + log(pest), data = fullscdata)

Residuals:
     Min       1Q   Median       3Q      Max
-3.9105 -0.7106 -0.1591  0.6495  2.8778

Coefficients:
             Estimate  Std.Error  t value  Pr(>|t|)
(Intercept) 12.03502    0.10853  110.89   <2e-16 ***
log(temp)   -0.80203    0.03127  -25.64   <2e-16 ***
log(pest)    0.10514    0.00458   22.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 1.033 on 7858 degrees of freedom
Multiple R-squared:  0.1815, Adjusted R-squared:  0.1813
F-statistic: 871.4 on 2 and 7858 DF,  p-value: < 2.2e-16
```

Tests multivariés issus d'une MANCOVA

```
mancova(fullscnona, deps = c("Maize", "Potatoes", "Rice..paddy", "Sorghum", "Wheat"),
        factors = c("Cluster", "Year"), covs = c("avg_temp", "pest", "rain"),
        boxM = T, shapiro = T, qqPlot = T)
```

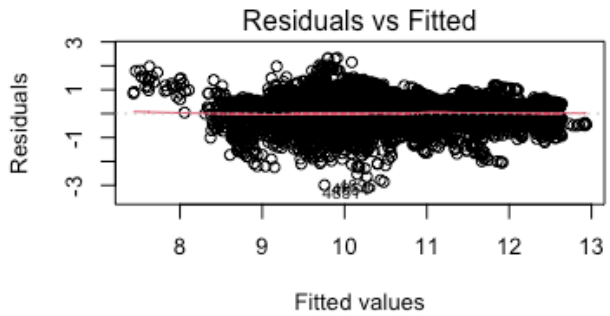
term[pillai]	stat[pillai]	f[pillai]	df1[pillai]	df2[pillai]	p[pillai]
Cluster	1.1245414	42.5969403	25	3670	0.0000000
Year	0.2107225	1.4679599	110	3670	0.0012438
Cluster:Year	0.1657757	0.2288219	550	3670	1.0000000
avg_temp	0.0458098	7.0093232	5	730	0.0000021
pest	0.3100519	65.6101264	5	730	0.0000000
rain	0.08543731	13.6204545	5	729	0.0000000

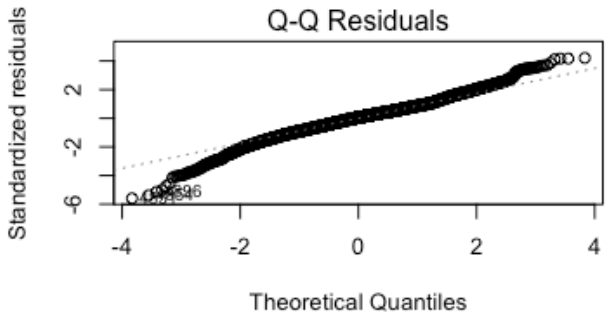
term[roy]	stat[roy]	f[roy]	df1[roy]	df2[roy]	p[roy]
Cluster	1.3973509	205.131111	5	734	0.0e+00
Year	0.2410270	8.041538	22	734	0.0e+00
Cluster:Year	0.0604368	0.403278	110	734	1.0e+00
avg_temp	0.0480091	7.009323	5	730	2.1e-06
pest	0.4493844	65.610126	5	730	0.0e+00
rain	0.09341876	13.6204545	5	729	0.0000000

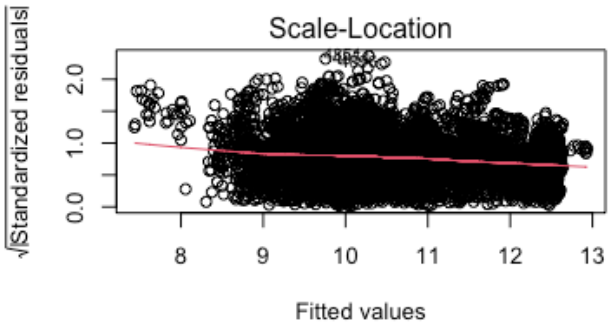
Régression linéaire générale

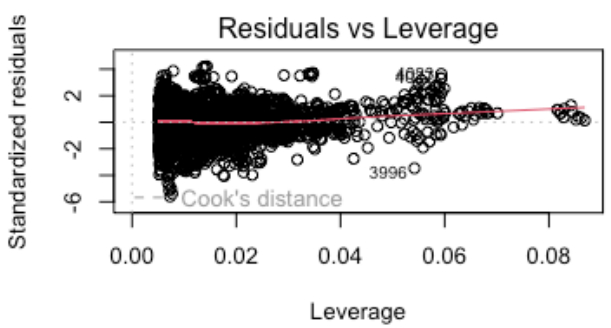
- Formula = log(yield) ~ Year*Item*Cluster*log(pest) ; AIC = -6960

⇒ Formula = log(yield) ~ Year + Item + Cluster + log(pest) + Item:Cluster + Year:log(pest) + Item:log(pest) + Cluster:log(pest) + Item:Cluster:log(pest) ; AIC = -9073







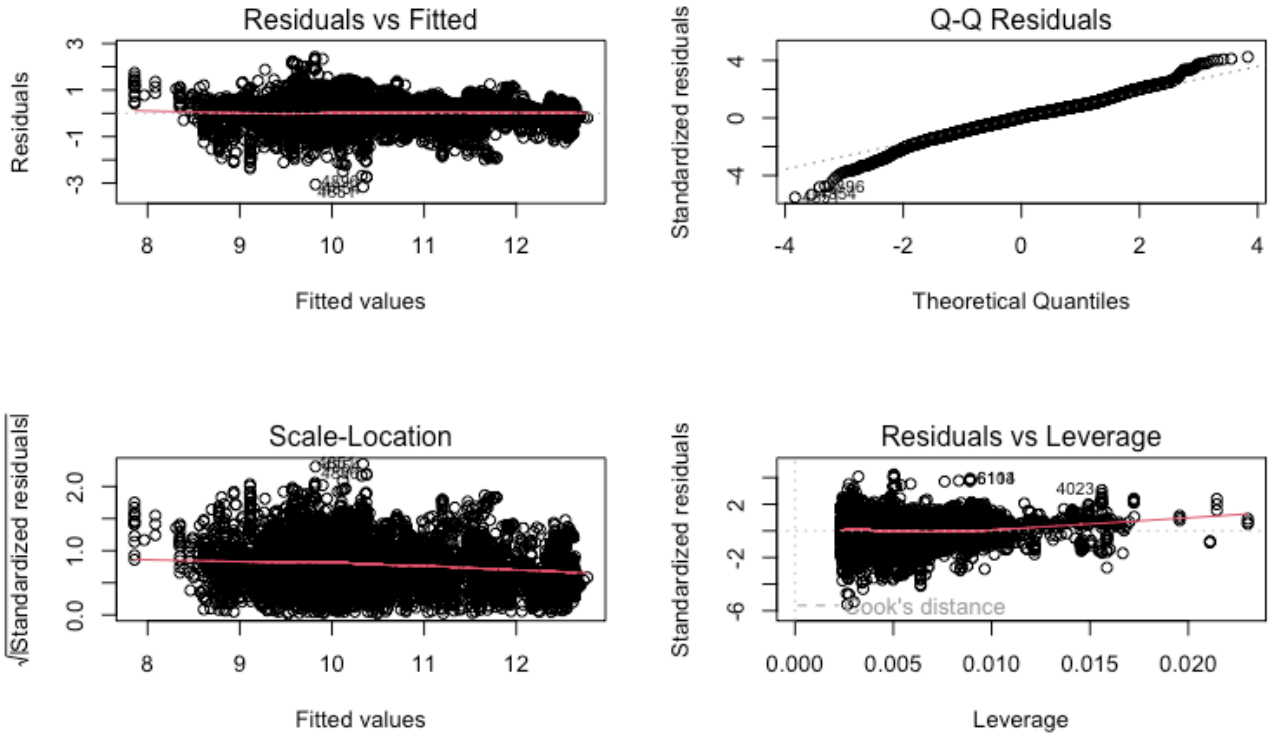


Régression linéaire générale

```
> summary(anc6)
[...]
Residual standard error: 0.5578 on 7757 degrees of freedom
Multiple R-squared:  0.7644,    Adjusted R-squared:  0.7613
F-statistic: 244.4 on 103 and 7757 DF,  p-value: < 2.2e-16

> car::Anova(anc6) # Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Anova Table (Type II tests)

Response: log(yield)
                      Sum Sq   Df   F value    Pr(>F)
Year                    37.9   22    5.5343 1.164e-15 ***
Item                  5377.6    4 4320.4661 < 2.2e-16 ***
Cluster                444.8    5  285.9029 < 2.2e-16 ***
log(pest)              560.4    1 1801.0247 < 2.2e-16 ***
Item:Cluster           173.4   20   27.8578 < 2.2e-16 ***
Year:log(pest)          16.0   22    2.3412 0.0003778 ***
Item:log(pest)          58.7    4   47.1685 < 2.2e-16 ***
Cluster:log(pest)       99.5    5   63.9734 < 2.2e-16 ***
Item:Cluster:log(pest) 102.8   20   16.5259 < 2.2e-16 ***
Residuals             2413.8 7757
```

Régression linéaire générale

⇒ Formula = log(yield) ~ Item + Cluster + log(pest) + Item:Cluster + Item:log(pest) + Cluster:log(pest) ;
AIC = -8707 ; $R^2 = 0.75$



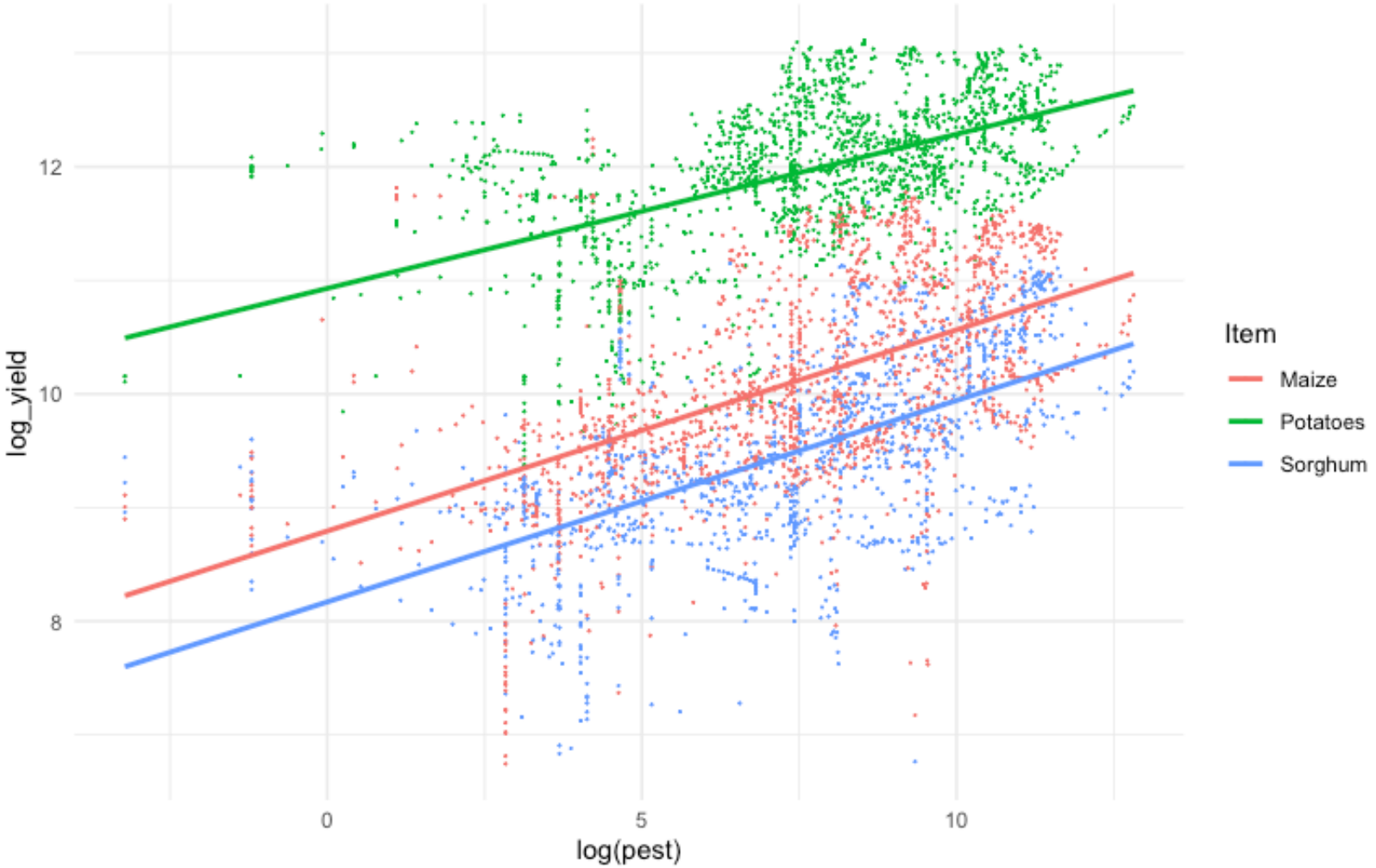
Conclusion 2

- Comment expliquer les variations de rendement selon les variables disponibles ?

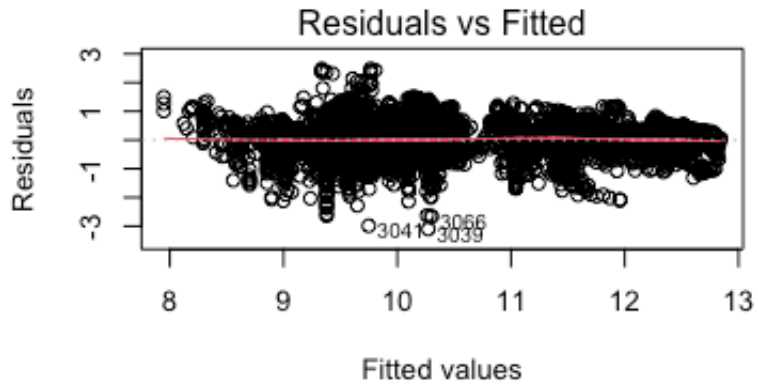
$\Rightarrow$ **Le logarithme du rendement est expliqué principalement par le logarithme du volume de pesticides, le type de culture, le cluster et leurs interactions respectives. L'année doit aussi, dans une moindre mesure, être prise en compte.**

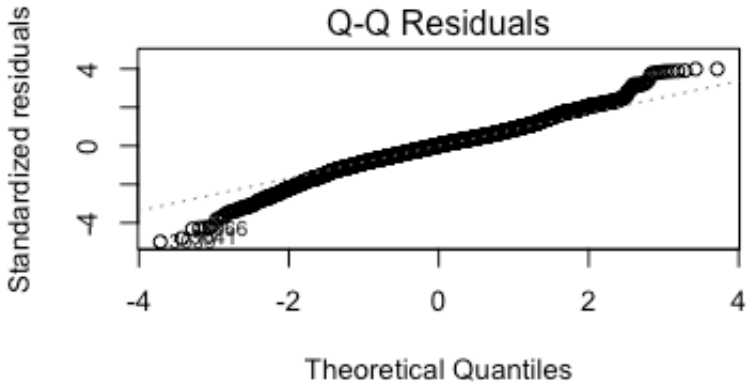
Modèle et hypothèses

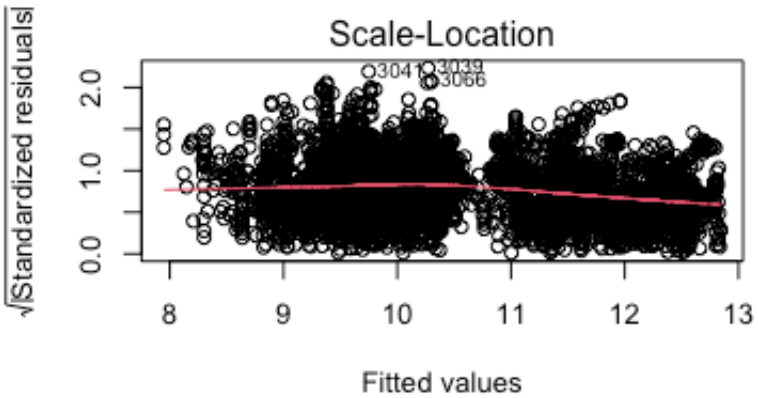
Formula = log(yield) ~ log(pest) + Item * Cluster

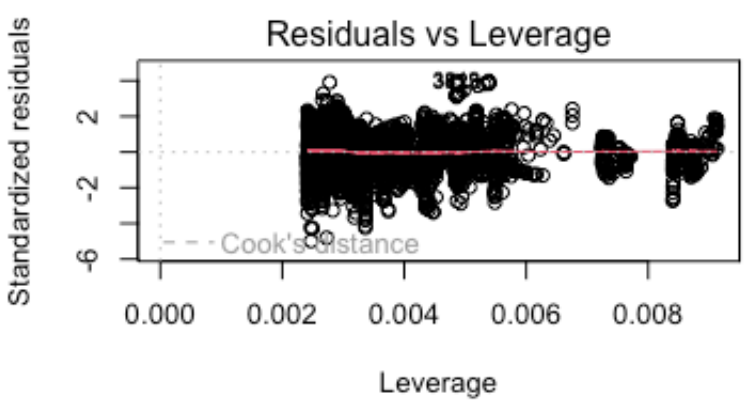


Diagnostic


Residuals vs Fitted
Residuals
Fitted values
3041
3039
3066


Q-Q Residuals
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Fitted values
3041
3039
3066


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Résultats

```
Res. std. error: 0.6223 on 4890 DF
Multiple R-squared:  0.7811
Adjusted R-squared:  0.7803
F-statistic: 969.2 on 18 and 4890 DF,
p-value: < 2.2e-16
```

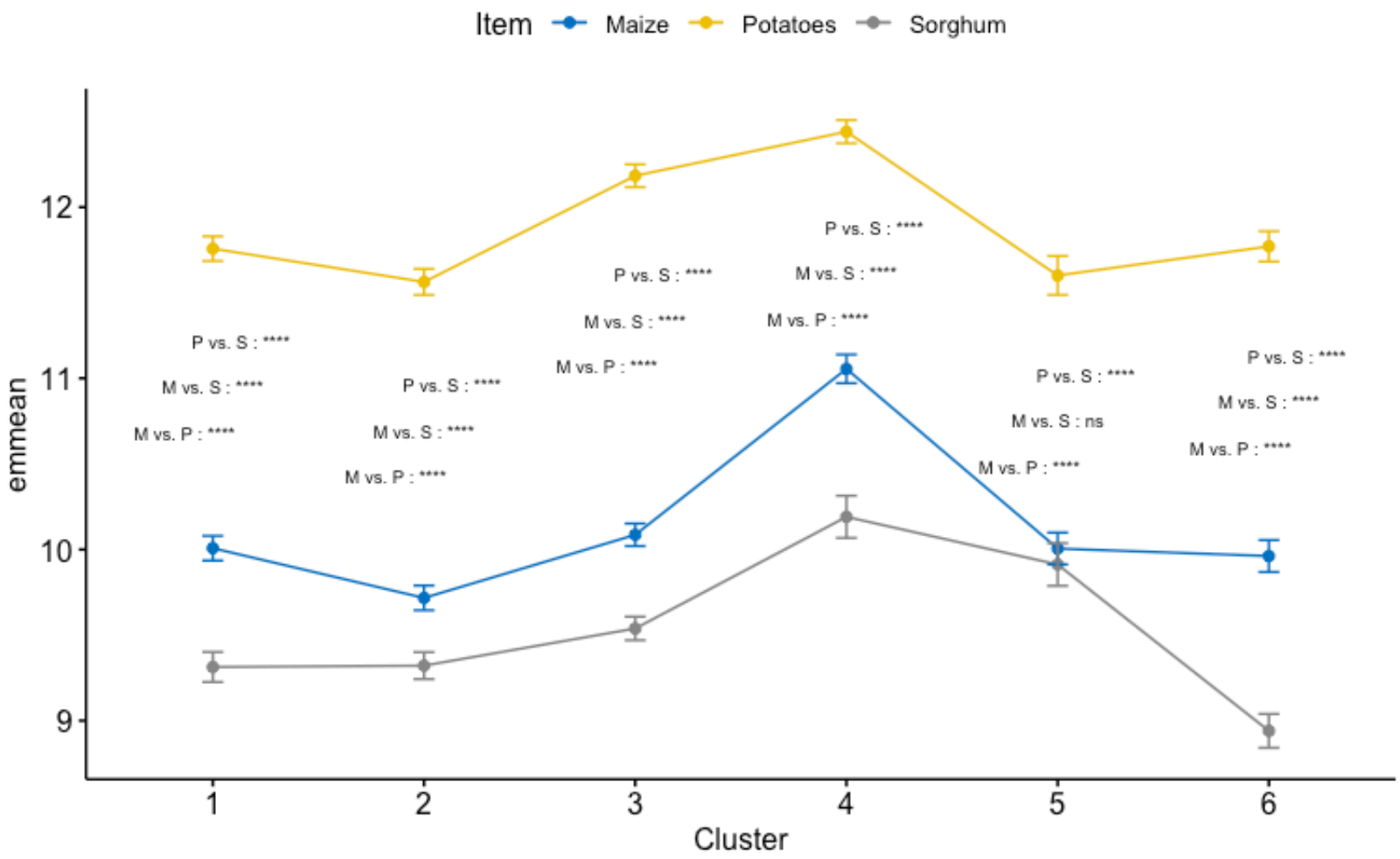
	Response	Sum Sq	Df	F value	Pr(>F)
log(pest)	542.5	1	1401.099	$< 2.2e-16$	***
Item	4919.1	2	6351.946	$< 2.2e-16$	***
Cluster	238.9	5	123.392	$< 2.2e-16$	***
Item:Cluster	106.3	10	27.447	$< 2.2e-16$	***
Residuals	1893.5	4890			

Résultats

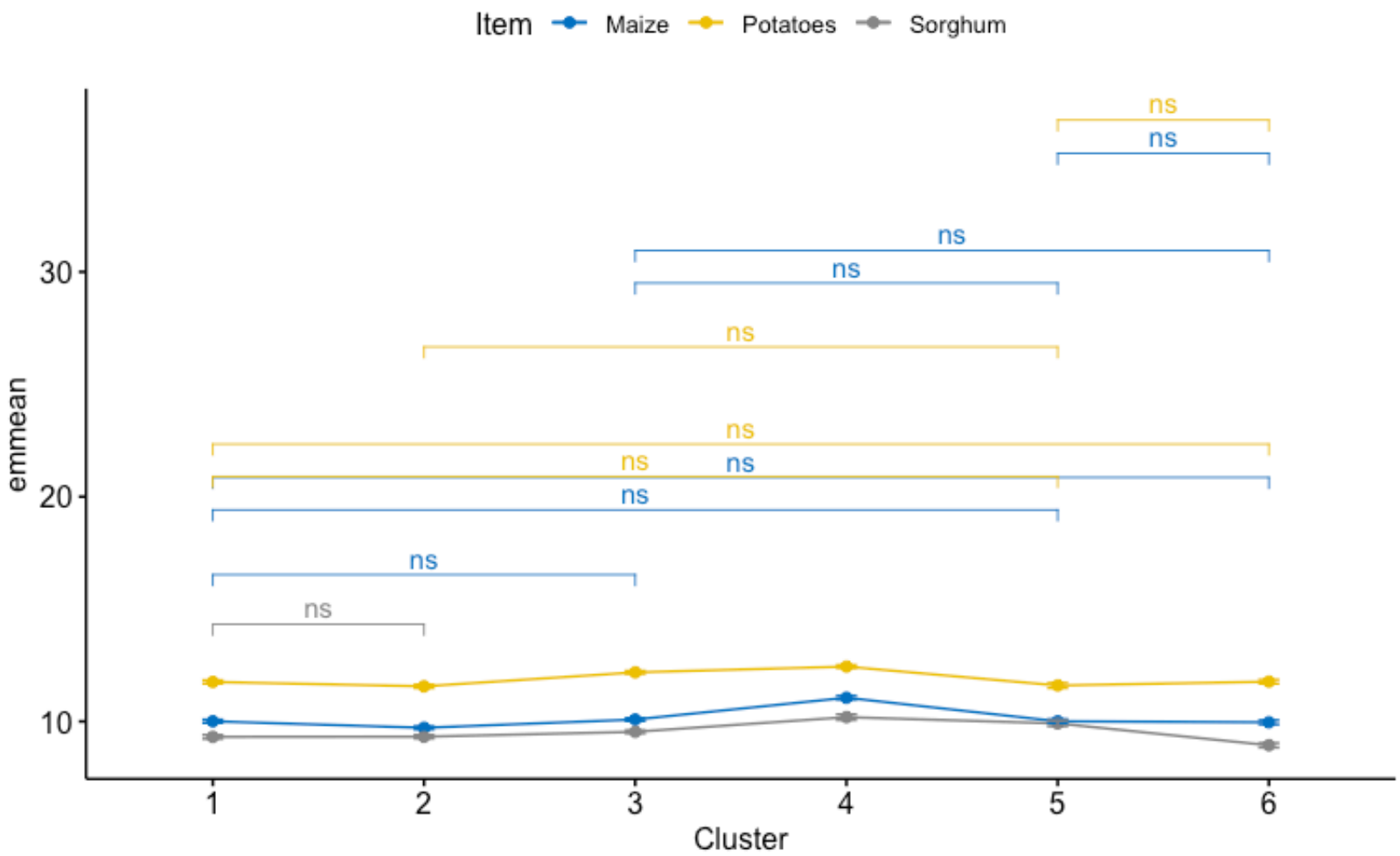
Cluster	Effect	DFn	DFd	F	p	$p_{\text{bonferroni}} < 8e-3$	ges
1	Item	2	919	1413	4.01e-281	*	0.755
2	Item	2	939	1516	8.66e-295	*	0.764
3	Item	2	1191	1347	1.64e-306	*	0.693
4	Item	2	763	1271	1.22e-243	*	0.769
5	Item	2	456	498	1.88e-115	*	0.686
6	Item	2	617	825	4.38e-175	*	0.728

Item	Effect	DFn	DFd	F	p	$p_{\text{bonferroni}} < 8e-3$	ges
Maize	Cluster	5	1768	68.6	$9.98e-66$	*	0.163
Potatoes	Cluster	5	1810	93.5	$1.01e-87$	*	0.205
Sorghum	Cluster	5	1310	24.7	$7.56e-24$	*	0.086

Différence entre les cultures selon les cluster



Différence entre les cluster selon les cultures



Conclusion 3

- Quels sont les effets du cluster et de la culture, après contrôle du volume de pesticides ?

⇒ **Au sein de chaque cluster on observe des différences significatives entre les rendements associés aux cultures. Pour une même culture, un changement de zone géographique affecte significativement le rendement.**

Take-Home Message

→ *k-means* : clusterisation à partir d'une paire de variables continues pour réduire les modalités d'une variable catégorielle

→ ACP : la réduction du jeu de données permet l'amélioration de la puissance statistique

⇒ Régressions linéaires : le rendement est intrinsèquement lié au type de culture et son environnement, en plus d'être expliquable par la quantité de pesticides utilisée.

Limites

- Pluie : une variable quasi-ordinale ?
- Manque de variables, manque de données
- Hypothèses de modèles linéaires non-satisfaites
- Performer un KNN aurait été vain (le code *ici*)

⇒ **L'étude de phénomènes non-linéaires demande des méthodes plus élaborées**

→ ... random forest ?

	Model	Accuracy	MSE	R2_score
0	Linear Regression	0.751364	1770624736.133630	0.751364
1	Decision Tree	0.978228	155044235.542397	0.978228
2	Random Forest	0.984811	108164948.657258	0.984811
3	Gradient Boost	0.865138	960402775.021678	0.865138
4	XGBoost	0.973514	188614498.872291	0.973514
5	Bagging Regressor	0.984792	108301368.373149	0.984792
6	KNN	0.332706	4752037374.447596	0.332706

[1] L. Rani, K. Thapa, N. Kanojia, N. Sharma, S. Singh, A. S. Grewal, A. L. Srivastav, and J. Kaushal, "An extensive review on the consequences of chemical pesticides on human health and environment," *Journal of Cleaner Production*, vol. 283, p. 124657, 2021.

[2] P. Nicolopoulou-Stamati, S. Maipas, C. Kotampasi, P. Stamatis, and L. Hens, "Chemical pesticides and human health: the urgent need for a new concept in agriculture," *Frontiers in public health*, vol. 4, p. 148, 2016.

[3] L. Beaumelle, L. Tison, N. Eisenhauer, J. Hines, S. Malladi, C. Pelosi, L. Thouvenot, and H. R. Phillips, "Pesticide effects on soil fauna communities—a meta-analysis," *Journal of Applied Ecology*, 2023.

[4] M. A. Beketov, B. J. Kefford, R. B. Schäfer, and M. Liess, "Pesticides reduce regional biodiversity of stream invertebrates," *Proceedings of the National Academy of Sciences*, vol. 110, no. 27, pp. 11039–11043, 2013.

[5] F. Geiger, J. Bengtsson, F. Berendse, W. W. Weisser, M. Emmerson, M. B. Morales, P. Ceryngier, J. Liira, T. Tscharntke, C. Winqvist, *et al.*, "Persistent negative effects of pesticides on biodiversity and biological control potential on european farmland," *Basic and Applied Ecology*, vol. 11, no. 2, pp. 97–105, 2010.

[6] C. Ateş, Ö. Kaymaz, H. E. Kale, and M. A. Tekindal, "Comparison of test statistics of nonnormal and unbalanced samples for multivariate analysis of variance in terms of type-i error rates," *Computational and mathematical methods in medicine*, vol. 2019, 2019.

Merci pour votre attention !