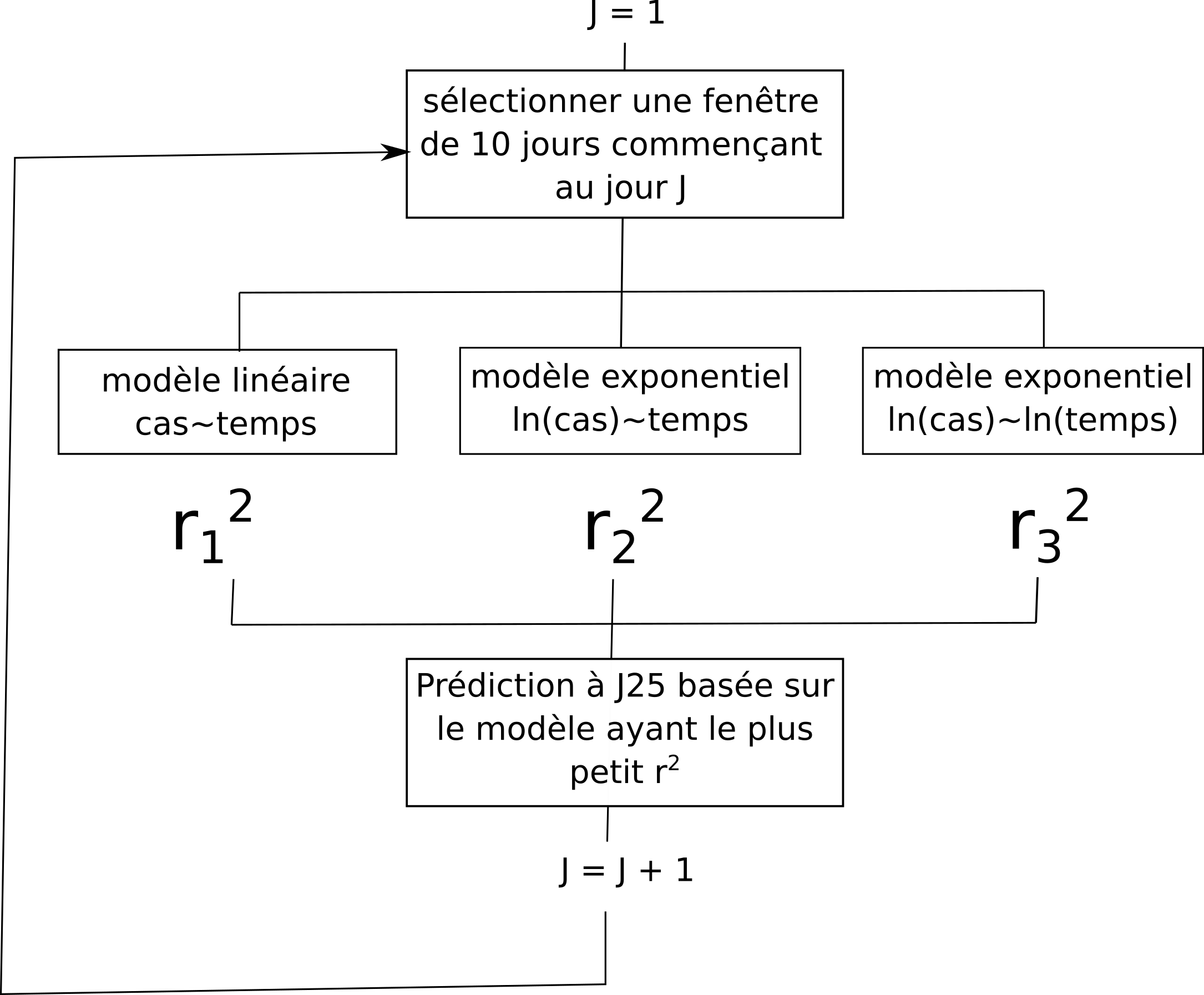
Fiabilité d’un modèle de prédiction statistique COVID-19  
basé sur 10 jours d’observation

PG

06 avril 2020 20:08

# Principe du modèle

On calcule les paramètres d’un modèle linéaire (), d’un modèle exponentiel () et d’un modèle log-log ( sur les 10 jours d’une série d’observation. On compare leurs r2 et on utilise le meilleur modèle (celui qui a le plus petit r2 pour prédire le nombre de cas au Jtième jour après la fin de la série.



# Principe de sa validation

En se basant sur des séries connues de 25 jours au moins, on calcule les paramètres du modèle sur les 10 premiers jours (à partir de j = 1) et on prédit la valeur à j+25 (donc 15 jours après), puis on décale la fenêtre de 10 jours à J+2, et on prédit la valeur à J+25, etc. jusqu’à j+16. Cette dernière fenêtre, à sa limite supérieure, inclut le jour pour lequel la prediction est calculée.

On oberve alors les écarts entre les valeurs prédites et les valeurs observées au fur et à mesure des itérations.

# Origine des données

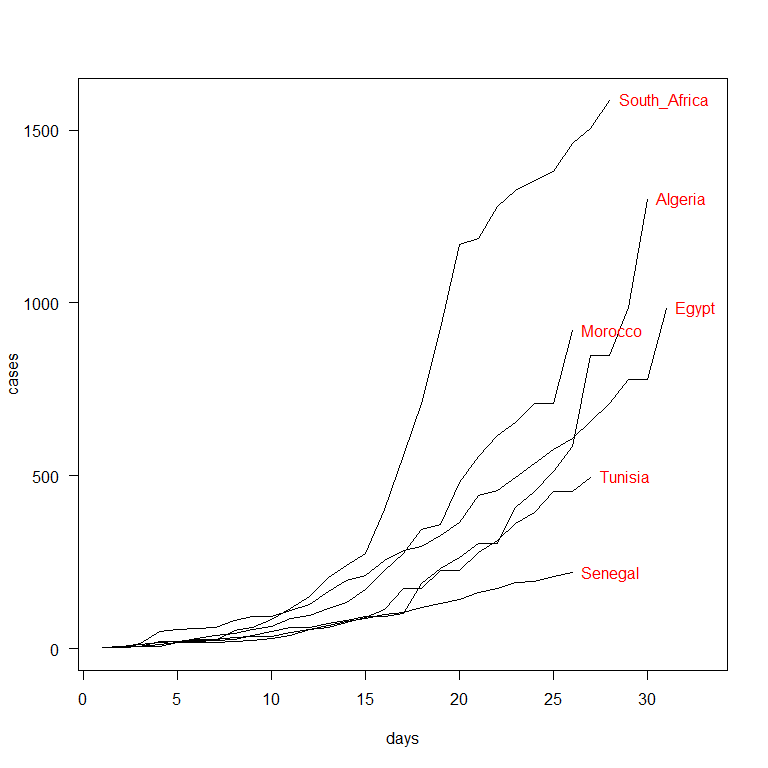
ECDC: <https://opendata.ecdc.europa.eu/covid19/casedistribution/csv>

# Sélection des séries

On bâti un tableau calculant la longueur des séries à ce jour, à partir du premier cas, et on identifie les séries ≥ 25 jours.

## country days  
## 2 Algeria 30  
## 15 Egypt 31  
## 33 Morocco 26  
## 39 Senegal 26  
## 43 South\_Africa 28  
## 47 Tunisia 27

6 pays s’avèrent disponibles pour l’exercice



Le moins qu’on puisse constater c’est la diversité des modes de croissance d’un pays à l’autre, et que donc l’utilisation d’un modèle unique *a priori* pour la prédiction précise des cas à plus de quelques jours est voué à l’échec.

# Calcul itératif

Le résultat des prédictions à J = 25 (J25, le 25ème jour) pour chaque fenêtre d’observation 1:10, 2:11, …. 16:25 est donné ci-dessous pour chaque pays, avec le r2 et le type de modèle sélectionné à l’itération correspondante.

Ci-dessous le résultat des observations à J25

## Algeria Egypt Morocco Senegal South\_Africa Tunisia   
## 511 576 708 207 1380 455

… et le résultat des itérations successives.

## $Algeria  
## pred r2 model  
## 1 2915 0.86 exp  
## 2 124 0.89 lin  
## 3 131 0.90 lin  
## 4 152 0.90 log  
## 5 191 0.94 log  
## 6 611 0.97 exp  
## 7 473 0.94 exp  
## 8 365 0.91 exp  
## 9 225 0.73 lin  
## 10 284 0.73 lin  
## 11 336 0.79 lin  
## 12 385 0.86 lin  
## 13 401 0.91 lin  
## 14 531 0.93 log  
## 15 605 0.93 exp  
## 16 576 0.92 exp  
##   
## $Egypt  
## pred r2 model  
## 1 88939 0.73 exp  
## 2 11571 0.64 exp  
## 3 1425 0.75 exp  
## 4 281 0.89 lin  
## 5 333 0.89 lin  
## 6 370 0.92 lin  
## 7 419 0.94 lin  
## 8 455 0.95 lin  
## 9 568 0.97 log  
## 10 856 0.97 exp  
## 11 760 0.96 exp  
## 12 721 0.97 exp  
## 13 542 0.97 lin  
## 14 554 0.97 lin  
## 15 565 0.98 lin  
## 16 570 0.98 lin  
##   
## $Morocco  
## pred r2 model  
## 1 238 0.94 log  
## 2 11483 0.94 exp  
## 3 6983 0.92 exp  
## 4 3562 0.90 exp  
## 5 263 0.97 lin  
## 6 299 0.95 lin  
## 7 359 0.90 lin  
## 8 427 0.90 lin  
## 9 508 0.90 lin  
## 10 551 0.94 lin  
## 11 631 0.93 lin  
## 12 707 0.95 lin  
## 13 759 0.97 lin  
## 14 1077 0.97 exp  
## 15 948 0.96 exp  
## 16 834 0.94 exp  
##   
## $Senegal  
## pred r2 model  
## 1 4118 0.82 exp  
## 2 1641 0.77 exp  
## 3 658 0.77 exp  
## 4 118 0.90 lin  
## 5 137 0.90 lin  
## 6 151 0.93 lin  
## 7 167 0.96 lin  
## 8 209 0.98 log  
## 9 345 0.97 exp  
## 10 320 0.96 exp  
## 11 273 0.98 exp  
## 12 256 0.99 exp  
## 13 205 0.99 lin  
## 14 210 0.99 lin  
## 15 227 0.99 exp  
## 16 218 0.98 exp  
##   
## $South\_Africa  
## pred r2 model  
## 1 197 0.85 lin  
## 2 254 0.86 lin  
## 3 319 0.87 lin  
## 4 406 0.87 lin  
## 5 485 0.91 lin  
## 6 547 0.95 lin  
## 7 672 0.92 lin  
## 8 843 0.87 lin  
## 9 1044 0.87 lin  
## 10 1282 0.87 lin  
## 11 1542 0.89 lin  
## 12 1663 0.93 lin  
## 13 1718 0.96 lin  
## 14 2624 0.92 exp  
## 15 2118 0.86 exp  
## 16 1718 0.81 exp  
##   
## $Tunisia  
## pred r2 model  
## 1 1567 0.82 exp  
## 2 600 0.90 exp  
## 3 97 0.82 lin  
## 4 113 0.84 lin  
## 5 137 0.87 lin  
## 6 162 0.92 lin  
## 7 192 0.93 lin  
## 8 251 0.85 lin  
## 9 281 0.90 lin  
## 10 325 0.91 lin  
## 11 341 0.94 lin  
## 12 366 0.96 lin  
## 13 611 0.96 exp  
## 14 555 0.96 exp  
## 15 505 0.95 exp  
## 16 471 0.96 exp

On obtient donc les écarts au prédiction suivants:

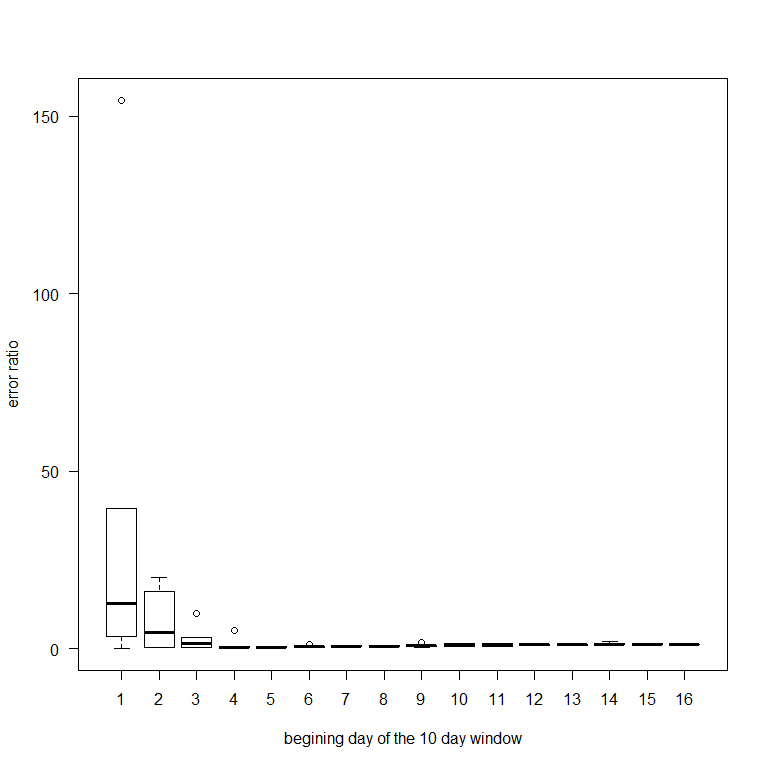
## Algeria Egypt Morocco Senegal South\_Africa Tunisia  
## [1,] 2404 88363 27208 3911 -1183 1112  
## [2,] -387 10995 10775 1434 -1126 145  
## [3,] -380 849 6275 451 -1061 -358  
## [4,] -366 -295 2854 -89 -974 -342  
## [5,] -351 -243 -445 -70 -895 -318  
## [6,] 100 -206 -409 -56 -833 -293  
## [7,] -38 -157 -349 -40 -708 -263  
## [8,] -146 -121 -281 -32 -537 -204  
## [9,] -286 -103 -200 138 -336 -174  
## [10,] -227 280 -157 113 -98 -130  
## [11,] -175 184 -77 66 162 -114  
## [12,] -126 145 -1 49 283 -89  
## [13,] -110 -34 51 -2 338 156  
## [14,] -67 -22 369 3 1244 100  
## [15,] 94 -11 240 20 738 50  
## [16,] 65 -6 126 11 338 16

Et si on raisonne en marge d’erreur (valeur absolue prédite/valeur observée)

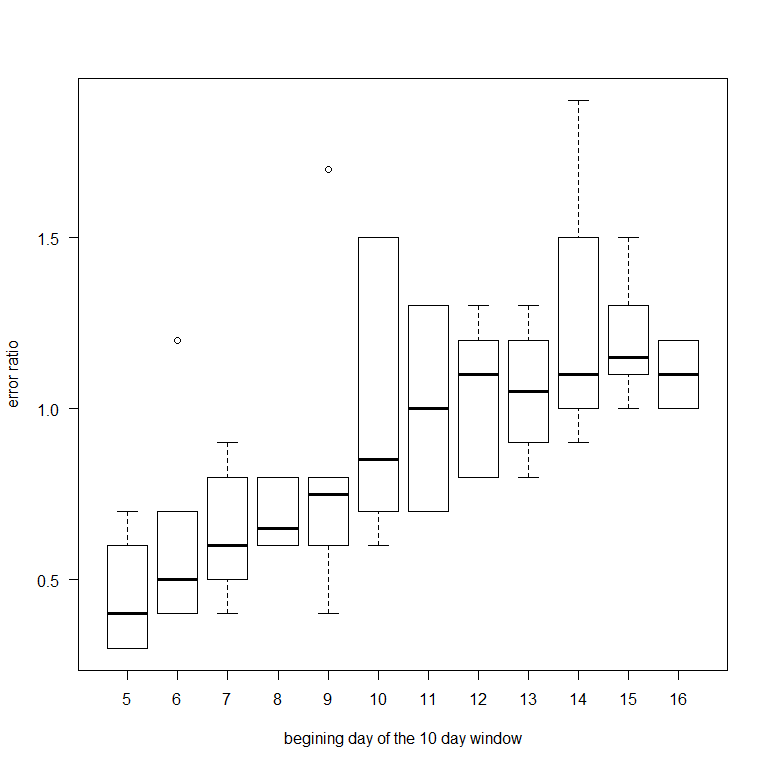
## Algeria Egypt Morocco Senegal South\_Africa Tunisia  
## [1,] 5.7 154.4 39.4 19.9 0.1 3.4  
## [2,] 0.2 20.1 16.2 7.9 0.2 1.3  
## [3,] 0.3 2.5 9.9 3.2 0.2 0.2  
## [4,] 0.3 0.5 5.0 0.6 0.3 0.2  
## [5,] 0.3 0.6 0.4 0.7 0.4 0.3  
## [6,] 1.2 0.6 0.4 0.7 0.4 0.4  
## [7,] 0.9 0.7 0.5 0.8 0.5 0.4  
## [8,] 0.7 0.8 0.6 0.8 0.6 0.6  
## [9,] 0.4 0.8 0.7 1.7 0.8 0.6  
## [10,] 0.6 1.5 0.8 1.5 0.9 0.7  
## [11,] 0.7 1.3 0.9 1.3 1.1 0.7  
## [12,] 0.8 1.3 1.0 1.2 1.2 0.8  
## [13,] 0.8 0.9 1.1 1.0 1.2 1.3  
## [14,] 0.9 1.0 1.5 1.0 1.9 1.2  
## [15,] 1.2 1.0 1.3 1.1 1.5 1.1  
## [16,] 1.1 1.0 1.2 1.1 1.2 1.0

On peut voir que les prédictions à 15 jours, sont complétement farfelues, que celles à 10 jours ont toutes produit des sous-estimations larges, et que même 3 jours avant, on peut avoir des erreurs grossières (voir l’Afrique du Sud, qui engage une surestimation de près de 2 (1.9), pour une fenêtre d’observation de 10 jours se terminant 3 jours avant le 25ème jour).

On peut matéraliser ça graphiquement:



… et si on s’arrange pour ne pas représenter les valeurs des 4 premières fenêtres, qui “compriment” la suite:



On pourrait donc en conclure qu’une estimation à J+5 serait “en moyenne” assez correcte (ratio d’erreur proche de 1), mais avec des possibilités de surestimation importante ( x 1.9 = ~ x 2). Ce qui peut arriver par exemple dans une transition vers un plateau de pic épidémique.

Ce que nous indique finalement ces tableaux et ce graphe c’est que, dans la phase de croissance épidémique:

* les estimations de J+15 à J+10 sont farfelues
* les estimations de J+11 à J+6 sous-estiment le nombre de cas (malgré la composante exponentielle de l’algorithme)
* les estimations de J+5 à J+1 sont plutôt correctes avec un risque de surestimation quand l’épidémie atteint un plateau

# Conclusion

* les dynamiques épidémiques sont très variables d’un pays à l’autre, et de ce fait, ne peuvent être décrites par un modèle unique; chaque tentative de prédiction est dépendante du type de dynamique observé et du stade dans lequel on est dans la dynamique
* les predictions > 10 jours sont impossibles
* les predictions les moins incorrectes sont obtenues à 1-5 jours, mais avec encore un important risque d’erreur, notamment lors d’une transition épidémique (passage de croissance à plateau)
* les prédictions à 10 jours sont possibles, mais, au moins dans la phase de croissance, sous-estiment les observations jusque du simple au double. Un troisième modèle log-log pourrait peut-être améliorer les choses (mais exposerait possiblement à des surestimations à d’autres période - à voir, je regarderai)

Pour le moment, les prédictions les plus fiables seront certainement faites en terme d’interprétation raisonnée, en combinant l’observation visuelle des courbes et cette approche de modélisation (ou une autre approche “modèle”), combinée à d’autres indicateurs locaux du tableau de bord épidémique. L’expérience terrain des épidémies est ici irremplacable, et aucun modèle, à l’heure actuelle ne peut *a priori* rendre compte de la diversité des situations. Bien sûr, de nombreux modèles s’ajustant au mieux aux données pourront être produits quand les séries seront complètes. Mais on sera alors là dans une position purement académique post-crise, qui n’est pas celle requise par la demande opérationnelle du présent.

# Linéarisation d’un exponentielle ascendante

(code caché)