Pierre-André Cornillon
Département Miashs

R et data-table

1 Premiers pas avec data-table

1. Importer `LTER.csv` dans le data-table `lter`
2. Afficher l'objet `lter`
3. Afficher le résumé de l'objet `lter` [`summary`]
4. Afficher les 10 premières lignes de l'objet `lter` pour la colonne `Year` (sous forme de data-table puis de vecteur)
5. Supprimer les colonnes `N.Obs`, `Order`, `Species`, `Feeding`, `Habitat` et `Pollinator`.
6. Creer le facteur codeLocale×`Species`×`Species.code` en concatenant les chaines de caractères `LTER.site`, `Locale`, `Species.code`, `Species.code` [`paste`, `factor`]. Ce facteur sera nommé `LLS`.
7. Trouver le type de chaque colonne [`lapply`, `.SD`, `class`]
8. Sélectionner les 3 premières lignes de l'objet `lter` pour chaque niveau de `LLS` et mettre le résultat dans `tmp`
9. Trouver le pourcentage d'abondance nulle sur tout le jeu de données [`mean`]
10. Trouver la moyenne et l'écart-type des pourcentages d'abondance nulle par niveau de LLS. On nommera ces variables `moy` et `ec` [`mean`, `sd`]
11. Trouver le pourcentage des series d'abondances par niveau de LLS qui ont au moins un zéro.
12. sélectionner les séries de LLS qui comporte
    — 8 annees de mesures ou plus ET
    — au moins 4 valeurs non nulles
    Ne garder que les colonnes `Abundance LLS` et `Year` ; appeler le data-table résultat `dtm`.
13. Créer pour le fun une clef avec les variables `LLS` et `Year` pour `dtm`.
14. Toutes les abondances sont elles positives ou nulles ? Si non compter les lignes dont les données sont erronées et éliminer les.
15. Pour chaque série i de LLS passer au log l'Abondance (A_{ij} dénote l'Abondance de l'observation j du niveau i de LLS ou, dit autrement de la série i)

$$A^*_{ij} = \begin{cases} \log(A_{ij}) & \text{si } A_{ij} > 0 \\ \log(c_i + A_{ij}) & \text{sinon} \end{cases}$$

    où $c_i = 0.5 \times \min_{j|A_{ij}>0} A_{ij}$
16. Pour chaque série i de LLS ramener les années (notées t_{ij}) entre 0 et 1

$$T_{ij} = \frac{t_{ij} - \min_j t_{ij}}{\max_j t_{ij} - \min_j t_{ij}}$$

17. Pour chaque série i de LLS centrer et réduire les données (donnant $\tilde{A}^*_{ij}$)
18. Pour chaque série i de LLS, faire un modèle de régression simple

$$\tilde{A}^*_{ij} = \beta_{i1} + \beta_{i2} T_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$$

    et garder les estimations de β_{i1}, β_{i2}, σ_i et son AIC.
19. Pour ceux qui s'ennuient refaire la même chose avec

$$\tilde{A}^*_{ij} = \beta_{i1} + \beta_{i2} T_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} = \rho \varepsilon_{ij-1} + \eta_{ij}, \quad \eta_{ij} \sim \mathcal{N}(0, \sigma_i^2)$$

    et garder les estimations de β_{i1}, β_{i2}, σ_i, ρ et son AIC.

2 La figure

1. Télécharger les données `https://trends.google.com/trends/explore?date=2012-01-01%202020-09-30&q=dplyr,data.table`
2. En utilisant les package suivants

```
library(ggplot2)
library(lubridate)
library(data.table)
```

3. obtenir ce tableau

```
don
     dplyr DT       date
  1:     0  2 2012-01-01
  2:     0  1 2012-02-01
  3:     0  1 2012-03-01
 ---
104:    79 17 2020-08-01
105:    83 16 2020-09-01
106:    76 13 2020-10-01
```

avec les types suivants

```
     dplyr      DT    date
1: integer integer POSIXct
2: integer integer  POSIXt
```

4. Puis celui ci

```
b
           date package search
  1: 2012-01-01   dplyr      0
  2: 2012-02-01   dplyr      0
 ---
211: 2020-09-01      DT     16
212: 2020-10-01      DT     13
```

5. Puis finir par un `ggplot2` pour le graphique (revision de la séance précédente/ sortez une feuille ;)