RAPPORT PROJET PREVISON METEO MLG

Samir Aidoudi

09/01/2022

NB : Le Rapport est un complément des commentaires disponible dans le code R.

1) Analyse transverse et transformation des données

Objectif : Le but de ce chapitre et de se familiariser avec les données et essayer de trouver des premières relations sans pour autant se lancer dans une analyse statistique, pour schématiser passer les variables à l'entonnoir afin de trouver celles qui semblent être significatif pour prédire les précipitations.

Le plan est à changer de la partie 1

1.1 Recherche et Définition

Avant de se lancer tête baissée dans une analyse statistique qui pourrait être longue étant donné le nombre de variables, on va dans un premier temps, chercher à définir la pluie et son fonctionnement pour comprendre quelles sont les variables d'intérêts les plus susceptibles d'influencer notre modèle de précipitation.

D'après météo France :

« Le phénomène de précipitation (qui a pour conséquences de déposer au sol de l'eau liquide ou solide) est essentiellement dû au grossissement des particules constitutives du nuage. L'accroissement de la taille des éléments, et donc de leur poids, permet à un certain stade de vaincre les forces d'agitation. Les particules sont emportées par la gravité, une précipitation est en cours. »

De ce fait on comprend qu'un des phénomènes expliquant l'apparition des précipitations est la coalescence dont nous allons voir une définition appliquée dans le cadre de la météorologie :

« En microphysique des nuages, les gouttelettes croissent à des vitesses différentes par condensation et par effet Bergeron, selon la concentration de vapeur d'eau. Elles auront donc une variété de diamètres et se déplaceront à une vitesse différente reliée à leur diamètre dans le courant ascendant. La coalescence est l'amalgamation subséquente de deux ou plusieurs gouttelettes par collision pour en former de plus grosses. (Cf. Wikipédia) »

On peut donc penser que **l'humidité** et la **pression** (ou vitesse) sont des variables significatives dans notre modèle de pluviométrie, ainsi qu'une couche nuageuse de **moyenne** et **haute attitude** ce qui est confirmé dans un article de science et vie.

« Si les nuages ne s'approchent pas du niveau de la mer, c'est parce que les minuscules gouttelettes qui les constituent se forment par condensation de la vapeur d'eau quand l'air humide se refroidit. Or,

ce refroidissement de l'air se produit lorsqu'une masse d'air s'élève dans l'atmosphère et subit une décompression. " Un refroidissement suffisant pour provoquer une condensation nécessite que la masse d'air monte à quelques centaines de mètres ", précise François Jobard, prévisionniste à Météo-France. »

On s'aide de ces liens :

http://education.meteofrance.fr/dossiers-thematiques/observer-et-mesurer/les-precipitations/comment-se-forment-les-precipitations

https://www.science-et-vie.com/questions-reponses/pourquoi-les-nuages-restent-ils-en-altitude-au-lieu-de-tomber-comme-la-pluie-5751

https://fr.wikipedia.org/wiki/Coalescence_(physique)

1.2 Data Quality.

En observant le summary et en parcourant le jeu de données nous ne voyons pas de problèmes de data quality. On peut estimer que les données sont cohérentes.

1.3 Organisation du Dataset

Je vais introduire une nouvelle variable construite à partir des dates afin de retranscrire l'impact du calendrier sur la météo. Nous verrons par la suite à quelle variable nous associons cette nouvelle variable.

Nous avons fait tout d'abord une première modification, de la première colonne (clef sur les lignes, à minutes [1 :5] on a regroupé ces colonnes en 1 clef sur les dates (à convertir) il n'y a pas de pertes d'informations cela fait data set 3.

A partir de la colonne 24 du data set (excepté pluie) les colonnes sont des relevés min et max des variables présents. Ainsi par exemple on à la température colonne « » qui est la moyenne arithmétique du relevé max en colonne « » et relevé min en colonne « ».

Je vais garder uniquement dans un premier temps ces valeurs moyennes afin de simplifier mes modèles et mon étude et puis je n'ai pas l'impression que je vais beaucoup d'informations par rapport au bénéfice de « simplicité » d'étude.

Cette impression devra être confirmé par la suite notamment au travers d'une **ACP** et de l'étude des liens entre ces variables min, max et moyenne.

1.4 Graphes

Graphes de ces variables

1 Graphe Pluie vs Couverture nuageuse.

On voit que plus la couverture nuageuse est élevée plus il y'a des « chances » de pleuvoir.

2 Graphe Pluie vs Sunshine Duration

On voit que plus la Sunshine duration est élevée moins il y'a des « chances » de pleuvoir. Il sera utile cependant de voir le lien entre cette variable et la précédente.

3 Graphe Pluie vs pression

On voit que plus la pression est élevée moins il y'a des « chances » de pleuvoir. Ce qui est normal en vue des études préliminaires sur la formation de la pluie (cf partie...)

4 Graphe Pluie vs Direction Vent

Le vent est également une variable significative.

5 Graphe Température vs Saison

Les résultats sont intéressants … Je m'attendais à des variations plus significatives de températures entre les saisons

6 Graphe Précipitation vs Saison

Il semble que la saison à un impact (pas déterminant) mais non négligeable sur la précipitation.

Ce qui en somme toute semble être logique.

Ainsi cette étude préliminaire nous permet de voir les premières variables significatives et pertinentes pour notre construction de modèle, et de réorganiser notre dataset.

Nous allons poursuivre cette analyse en étudiant les liens entre variables et individus via une analyse en composante principale.

2) ACP

2.1 Matrice de Corrélation et Analyse.

D'abord on visualise les matrices de corrélations. Effectivement dans la matrice des données complètes on s'aperçoit que des variables moyennes sont fortement corrélés à celle min et max par exemple :

	Temperature
Temperature.daily.max..2.m.above.gnd.	0.98066559040582
Temperature.daily.min..2.m.above.gnd.	0.968386279796393
	Pressure
Mean.Sea.Level.Pressure.daily.max..MSL.	0.972200760824231
Mean.Sea.Level.Pressure.daily.min..MSL.	0.973547753679361
	Humidity
Relative.Humidity.daily.max..2.m.above.gnd.	0.773905615608494
Relative.Humidity.daily.min..2.m.above.gnd.	0.893969518525286
	Wind.Gust
Wind.Gust.daily.max..sfc.	0.885366971826856
Wind.Gust.daily.min..sfc.	0.822333664244249

On voit aussi que les vitesses moyennes de vent sont fortement corrélées et que la précipitation est corrélée positivement à la couverture nuageuse, humidité et négativement à la pression et Sunshine duration

2.2 Choix des axes.

Cf commentaire code R

2.3 Analyses des individus

Cf commentaire code R

2.4 Analyse Automatiques des variables

Cf commentaire code R

2.5 Conclusion ACP

L'ACP nous donne de nombreuses informations notamment sur les variables, leur corrélations (positives ou négatives), le lien entre elles, ce qui nous sera fortement utile par la suite pour sélectionner nos variables et construire des modèles pertinents.

De plus le fait de réduire le data set à des données moyennes ne constitue pas une perte significative d'informations. Mais par acquis de conscience nous allons comparer 2 recherches de variables entre les 2 datasets.

3)Recherche Variables et Modèles.

3.1 Méthode Automatique de choix de variables

Cf commentaire code R

3.2 Choix de Modèles

Cf commentaire code R

3.3 Modèles Logistique vs Probit.

Cf commentaire code R

3.4 Analyse des résidus du modèle selectioné.

Cf commentaire code R

4)Validation Croisée et Prédiction

4.1 Validation Croisée

Cf commentaire code R

4 .2 Erreur de Prédiction

Cf commentaire code R

4.3 Prédiction sur le dataset de test et export des résultats.

Cf commentaire code R

5) Conclusion

Le projet fut très intéressant et m'a permis de mettre en application ce que nous avons étudier en cours

J'ai essayé de pas me laisser "emporté" par les méthodes et algorithmes j'espère notamment que vous comprendrez mes choix (notamment la réduction du data set), j'ai voulu faire un effet entonnoir dans mon approche.

J'ai tenté de justifier qualitativement et quantitativement mon raisonnement, il n'empêche que le degré de précision de ma fonction de prédiction tend à être améliorer

Cependant j'ai essayé de chercher d'autres covariables que la saison, et fait des recherches de modèles à partir du dataset complet mais comme les résultats n'était pas satisfaisant je ne les ai pas affichés (afin notamment de pas encombrer ce rapport avec des données inutiles).

Ce projet m'amène à me poser quand des questions, dont j'espère que vous aurez les réponses :

1° La limite ou opposition entre la recherche de modèles et la compréhension des variables exemple : modèle physique de météorologie VS analyse statistique.
Ainsi pour paraphraser une citation (dont j'ai oublié l'auteur), « Avec le big data et les milliers de variables qu'une machine sait traitée on a plus besoin de construire des modèles »

2°Equilibre entre réduction ou complexité du modèle et son gain en précision.

3°Au moment où je finis ce projet nous débutons les cours de régression non paramétrique, ainsi utiliser une régression non param est ce que on aurait pu avoir de meilleurs résultats ?

Je vous remercie pour le sujet, son étude fu très intéressant, ainsi que pour vos critiques et commentaires de mon rapport qui je suis sur me permettrons d'améliorer mon analyse.