Modèle Linéaire Généralisé

Nathalie Ung

2023-07-08

Contents

Introduction

La météo représente un aspect essentiel dans la prise de décision dans différents domaines tels que la planification des évènements urbains, l'agriculture ou encore au niveau des individus. La capacité à prédire avec précision l'occurence des précipitations peut également avoir des impacts économiques, les individus modifiant leur organisation pour s'y adapter.

Dans ce contexte, nous cherchons à comprendre, étudier et in fine prédire la probabilité de précipitations le lendemain, dans la ville de Bâle. A à fin, nous procéderons dans une première partie à une analyse exploratoire afin de d'identifier et comprendre les facteurs météorologiques qui influencent le phénomène de précipitations ainsi que leur potentielles interactions. Ainsi, à l'aide ces éléments, nous développerons dans une deuxième partie un modèle de prédiction des précipitations estimation d'un modèle sur un échantillon d'apprentissage. La validation du modèle sera ensuite réalisée dans une quatrième partie. Enfin, dans la dernière partie, nous réaliserons une prédiction quant aux chances de précipitation pour le lendemain pour la ville de Bâle.

1. Préparation des données

Dans le cadre de notre étude, nous disposons de deux ensembles de données *meteo_train* et *meteo_test* pour étudier et tenter de prédire s'il va pleuvoir le lendemain dans la ville de Bâle.

Le premier ensemble, constituant l'ensemble d'entraînement, sera utilisé pour explorer et analyser les observations ainsi qu'ajuster un modèle de prédiction puis en évaluer la qualité, en le confrontant aux observations, et enfin le valider.

La prédiction finale sera réalisée en mobilisant le second ensemble *meteo_test*.

Nous cherchons dans un premier temps à identifier des valeurs manquantes dans les ensembles de données.

	Train set	Test set
Nb de N/A	0	0

Aucune données manquantes n'ayant été identifiées, nous pouvons analyser la typologie des données constituant l'ensemble d'entraînement. Ce dernier est constitué de 1 180 observations réalisées pour 47 variables.

```
## [1] 1180   47
```

La premiere variable que nous avons nommé X semble correspondre à un identifiant de ligne. N'en ayant pas l'utilité dans notre analyse, nous la retirerons dans la suite de notre étude.

Afin de faciliter la lecture de nos observations, nous avons également choisi de renommer les variables explicatives.

```
## 'data.frame':    1180 obs. of  46 variables:
##  $ Year                  : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ Month                 : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ Day                   : int  2 4 6 8 10 12 14 16 18 20 ...
##  $ Hour                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Minute                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Temperature           : num  15 17.3 21.6 20.2 22.6 ...
##  $ Humidity              : num  76.5 77.6 69.5 75.1 73.5 ...
##  $ Sea.Level             : num  1015 1017 1015 1008 1004 ...
##  $ Precipitation         : num  1 0 3.7 0.2 0 2.2 1.8 1.8 17.5 1.2 ...
##  $ Snowfall              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cloud.Cover.Total     : num  79.8 4.7 42.1 67.5 56.3 ...
##  $ Cloud.Cover.High      : num  3 0.67 21.21 54.71 50.25 ...
##  $ Cloud.Cover.Medium    : num  31.6 0 25.9 65.8 55.3 ...
##  $ Cloud.Cover.Low       : num  79.2 4.5 35.3 18.9 34.2 ...
##  $ Sunshine.Duration     : num  287.2 821.4 441.3 41.9 473.2 ...
##  $ Shortwave.Radiation   : num  6710 7974 4834 5390 7216 ...
##  $ Wind.Speed.10m        : num  11.64 6.34 8.4 5.4 9.16 ...
##  $ Wind.Direction.10m    : num  275 230 215 205 179 ...
##  $ Wind.Speed.80m        : num  14.99 8.92 10.38 6.53 11.91 ...
##  $ Wind.Direction.80m    : num  268 199 208 206 186 ...
##  $ Wind.Speed.900mb      : num  20.6 27.9 18.9 10.4 21.9 ...
##  $ Wind.Direction.900mb  : num  180.4 93.7 250.1 238.6 153 ...
##  $ Wind.Gust.            : num  14.88 9.48 13.5 5.31 12.21 ...
##  $ Temperature.max       : num  18.5 25 26.2 24.2 30.7 ...
##  $ Temperature.min       : num  11.1 10.4 17.7 14.7 16.9 ...
```

```
##  $ Humidity.max          : int  94 92 91 89 97 92 96 96 97 95 ...
##  $ Humidity.min          : int  59 54 57 62 39 65 69 64 74 61 ...
##  $ Sea.Level.max         : num  1017 1019 1016 1010 1006 ...
##  $ Sea.Level.min         : num  1014 1016 1013 1006 1001 ...
##  $ Cloud.Cover.Total.max : num  100 28 100 100 100 100 100 100 100 100 ...
##  $ Cloud.Cover.Total.min : num  0 0 0 0 0 0 0 100 0 0 ...
##  $ Cloud.Cover.High.max  : int  16 11 100 100 100 28 100 100 100 24 ...
##  $ Cloud.Cover.High.min  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cloud.Cover.Medium.max: int  100 0 100 100 100 100 100 100 100 41 ...
##  $ Cloud.Cover.Medium.min: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cloud.Cover.Low.max   : int  100 28 100 100 100 100 100 100 100 100 ...
##  $ Cloud.Cover.Low.min   : int  0 0 0 0 0 0 0 29 0 0 ...
##  $ Wind.Speed.max.10m    : num  22 15.5 22.7 10.7 20.5 ...
##  $ Wind.Speed.min.10m    : num  5.62 1.08 2.41 0 2.52 2.28 1.3 4.32 7.2 8.05 ...
##  $ Wind.Speed.max.80m    : num  23.8 18.7 32 10.2 23.4 ...
##  $ Wind.Speed.min.80m    : num  8.65 0 0.51 1.44 2.97 ...
##  $ Wind.Speed.max.900mb  : num  32.1 48.1 44 22.2 40.8 ...
##  $ Wind.Speed.min.900mb  : num  12.25 6.62 5.48 4.69 4.68 ...
##  $ Wind.Gust.max         : num  25.2 20.2 41.8 11.2 24.1 ...
##  $ Wind.Gust.min         : num  6.48 2.16 1.08 0.36 1.44 ...
##  $ pluie.demain          : logi  FALSE FALSE TRUE TRUE TRUE TRUE ...
```

La variable pluie étant binaire, nous la transformons sous les conditions suivantes:

- La valeur *FALSE* est remplacée par 0
- La valeur *TRUE* est remplacée par 1

```
meteo_train$pluie.demain=as.integer(as.logical(meteo_train$pluie.demain))
```

2. Analyse exploratoire

Nous réalisons en premier lieu une première analyse statistique (minimum, moyenne, ...) pour l'ensemble des variables, dont *pluie.demain* qui correspond à la variable d'intérêt.

```
##       Year          Month             Day             Hour       Minute
##  Min.   :2010   Min.   : 1.000   Min.   : 1.0   Min.   :0   Min.   :0
##  1st Qu.:2012   1st Qu.: 3.000   1st Qu.: 8.0   1st Qu.:0   1st Qu.:0
##  Median :2014   Median : 6.000   Median :16.0   Median :0   Median :0
##  Mean   :2014   Mean   : 6.436   Mean   :15.8   Mean   :0   Mean   :0
##  3rd Qu.:2016   3rd Qu.: 9.000   3rd Qu.:23.0   3rd Qu.:0   3rd Qu.:0
##  Max.   :2018   Max.   :12.000   Max.   :31.0   Max.   :0   Max.   :0
##   Temperature       Humidity       Sea.Level      Precipitation
##  Min.   :-7.63   Min.   :38.33   Min.   : 978.9   Min.   : 0.000
##  1st Qu.: 6.71   1st Qu.:64.82   1st Qu.:1012.4   1st Qu.: 0.000
##  Median :12.08   Median :72.21   Median :1017.0   Median : 0.100
##  Mean   :12.23   Mean   :71.40   Mean   :1017.0   Mean   : 2.085
##  3rd Qu.:17.54   3rd Qu.:78.63   3rd Qu.:1022.0   3rd Qu.: 2.300
##  Max.   :29.45   Max.   :95.54   Max.   :1042.4   Max.   :31.500
##     Snowfall       Cloud.Cover.Total Cloud.Cover.High  Cloud.Cover.Medium
##  Min.   :0.00000   Min.   :  0.00    Min.   :  0.000   Min.   :  0.00
##  1st Qu.:0.00000   1st Qu.: 23.80    1st Qu.:  1.657   1st Qu.:  1.83
```

```
##  Median :0.00000   Median : 51.67    Median : 11.880   Median : 24.98
##  Mean   :0.04965   Mean   : 50.76    Mean   : 20.284   Mean   : 31.50
##  3rd Qu.:0.00000   3rd Qu.: 78.53    3rd Qu.: 33.260   3rd Qu.: 54.21
##  Max.   :8.61000   Max.   :100.00    Max.   :100.000   Max.   :100.00
##  Cloud.Cover.Low  Sunshine.Duration Shortwave.Radiation Wind.Speed.10m
##  Min.   :  0.00   Min.   :   0.0    Min.   : 265.2      Min.   : 1.260
##  1st Qu.:  9.42   1st Qu.: 114.3    1st Qu.:2096.2      1st Qu.: 6.428
##  Median : 36.35   Median : 366.8    Median :3675.3      Median : 9.195
##  Mean   : 39.34   Mean   : 373.1    Mean   :3984.6      Mean   :10.707
##  3rd Qu.: 65.76   3rd Qu.: 587.7    3rd Qu.:5723.6      3rd Qu.:12.977
##  Max.   :100.00   Max.   :1015.8    Max.   :8363.3      Max.   :42.210
##  Wind.Direction.10m Wind.Speed.80m  Wind.Direction.80m Wind.Speed.900mb
##  Min.   : 11.19     Min.   : 1.34   Min.   : 12.18     Min.   : 2.25
##  1st Qu.:152.40     1st Qu.: 8.68   1st Qu.:157.42     1st Qu.:13.02
##  Median :206.36     Median :12.41   Median :213.78     Median :19.57
##  Mean   :201.82     Mean   :14.28   Mean   :206.23     Mean   :24.57
##  3rd Qu.:254.19     3rd Qu.:17.61   3rd Qu.:259.06     3rd Qu.:32.10
##  Max.   :331.67     Max.   :54.03   Max.   :333.43     Max.   :97.06
##  Wind.Direction.900mb   Wind.Gust.    Temperature.max Temperature.min
##  Min.   : 17.37       Min.   : 2.25   Min.   :-3.84   Min.   :-12.520
##  1st Qu.:144.02       1st Qu.: 9.48   1st Qu.:10.58   1st Qu.:  3.350
##  Median :233.47       Median :14.06   Median :16.54   Median :  8.005
##  Mean   :206.22       Mean   :16.69   Mean   :16.54   Mean   :  8.062
##  3rd Qu.:265.93       3rd Qu.:21.15   3rd Qu.:22.36   3rd Qu.: 13.092
##  Max.   :344.82       Max.   :79.38   Max.   :35.77   Max.   : 23.940
##   Humidity.max     Humidity.min   Sea.Level.max    Sea.Level.min
##  Min.   : 59.00   Min.   :19.00   Min.   : 981.9   Min.   : 977
##  1st Qu.: 83.00   1st Qu.:45.00   1st Qu.:1015.4   1st Qu.:1009
##  Median : 89.00   Median :54.00   Median :1019.5   Median :1015
##  Mean   : 87.69   Mean   :54.04   Mean   :1019.9   Mean   :1014
##  3rd Qu.: 94.00   3rd Qu.:63.00   3rd Qu.:1024.7   3rd Qu.:1019
##  Max.   :100.00   Max.   :92.00   Max.   :1045.4   Max.   :1039
##  Cloud.Cover.Total.max Cloud.Cover.Total.min Cloud.Cover.High.max
##  Min.   :  0.00        Min.   :  0.000       Min.   :  0.00
##  1st Qu.:100.00        1st Qu.:  0.000       1st Qu.: 15.00
##  Median :100.00        Median :  0.000       Median : 97.00
##  Mean   : 88.23        Mean   :  8.692       Mean   : 60.17
##  3rd Qu.:100.00        3rd Qu.:  2.400       3rd Qu.:100.00
##  Max.   :100.00        Max.   :100.000       Max.   :100.00
##  Cloud.Cover.High.min Cloud.Cover.Medium.max Cloud.Cover.Medium.min
##  Min.   :  0.0000     Min.   :  0.00         Min.   :  0.000
##  1st Qu.:  0.0000     1st Qu.: 22.75         1st Qu.:  0.000
##  Median :  0.0000     Median :100.00         Median :  0.000
##  Mean   :  0.9432     Mean   : 70.94         Mean   :  2.097
##  3rd Qu.:  0.0000     3rd Qu.:100.00         3rd Qu.:  0.000
##  Max.   :100.0000     Max.   :100.00         Max.   :100.000
##  Cloud.Cover.Low.max Cloud.Cover.Low.min Wind.Speed.max.10m Wind.Speed.min.10m
##  Min.   :  0         Min.   :  0.000     Min.   : 2.52      Min.   : 0.00
##  1st Qu.:100         1st Qu.:  0.000     1st Qu.:12.32      1st Qu.: 1.14
##  Median :100         Median :  0.000     Median :17.36      Median : 2.41
##  Mean   : 80         Mean   :  3.879     Mean   :19.06      Mean   : 3.57
##  3rd Qu.:100         3rd Qu.:  0.000     3rd Qu.:23.44      3rd Qu.: 4.45
##  Max.   :100         Max.   :100.000     Max.   :79.99      Max.   :27.73
##  Wind.Speed.max.80m Wind.Speed.min.80m Wind.Speed.max.900mb
```

```
##  Min.   : 3.98      Min.   : 0.000     Min.   :  4.02
##  1st Qu.:18.27      1st Qu.: 1.140     1st Qu.: 24.54
##  Median :23.85      Median : 2.600     Median : 37.12
##  Mean   :25.35      Mean   : 4.727     Mean   : 41.82
##  3rd Qu.:29.92      3rd Qu.: 5.830     3rd Qu.: 54.37
##  Max.   :93.84      Max.   :37.700     Max.   :136.25
##  Wind.Speed.min.900mb Wind.Gust.max   Wind.Gust.min     pluie.demain
##  Min.   : 0.00        Min.   : 4.32   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 3.05        1st Qu.:19.08   1st Qu.: 2.160   1st Qu.:0.0000
##  Median : 6.73        Median :26.10   Median : 3.960   Median :1.0000
##  Mean   :11.09        Mean   :29.31   Mean   : 6.502   Mean   :0.5093
##  3rd Qu.:15.31        3rd Qu.:37.08   3rd Qu.: 8.280   3rd Qu.:1.0000
##  Max.   :76.13        Max.   :95.04   Max.   :57.960   Max.   :1.0000
```

En première lecture nous observons que la période d'étude couvre une période de 8 ans, allant de 2010 à 2018. La présence des variables Day et Month semblent indiquer que les données sont collectées de manière quotidienne. Par ailleurs, l'heure (Hour) et les minutes du relevé des donnnées sont toutes égales à 0. Par conséquent, ces variables ne semblent pas fournir d'information pour la prédiction à réaliser. La date de relevé des données ne semble pas de prime abord avoir d'impact sur la prévision météorologique. Les variables associées seront de ce fait retirées de notre analyse. Néanmoins, nous conserverons la variable Month qui peut donner une indication sur la saison, laquelle peut être utile (il y a une plus grande propension à pleuvoir en automne par exemple).

Les autres variables sont analysées plus en détail dans la section suivante.

2.1. Analyse descriptive

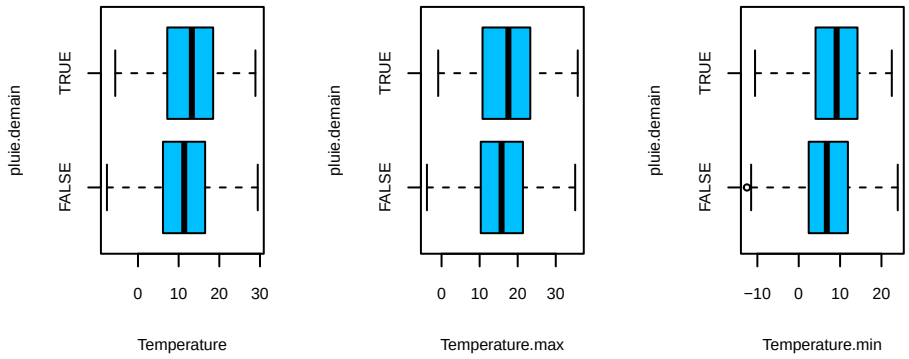
Dans l'ensemble d'entraînement, nous constatons un léger déséquilibre entre la proportion du nombre de jours où il a plu et celle où il n'a pas plu.

	Fréquence	Pourcentage
FALSE	579	49
TRUE	601	51

Nous réalisons ci-dessous une analyse descriptive par nature de variables.

Température moyenne, minimale et maximale

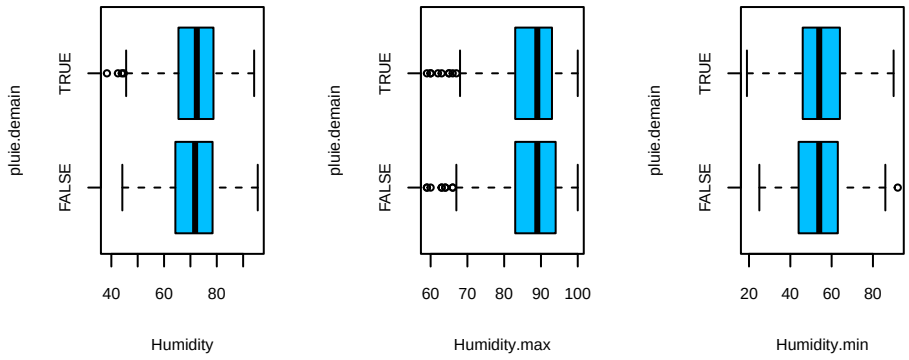
Nous analysons l'impact des trois variables liées à la "Température" sur la variable d'intéret.



Nous remarquons que la variable *Temperature.min* est celle qui présente le plus fort lien avec la variable *pluie.demain*.

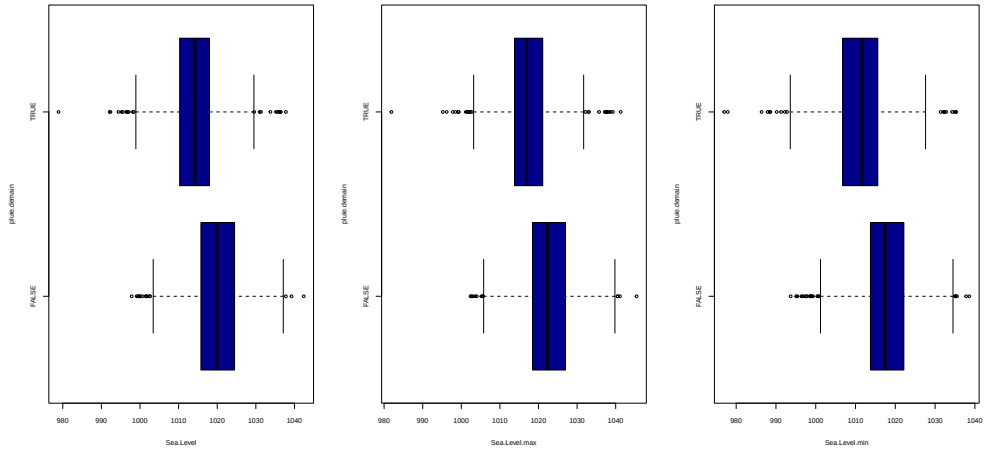
Le pourcentage d'Humidité relative moyenne, minimale et maximale

Les trois variables liées au taux d'humidité relative, présentées ci-dessous, ne semblent pas liés sur les précipitations.



Pression au niveau de la mer moyenne, minimum et maximum

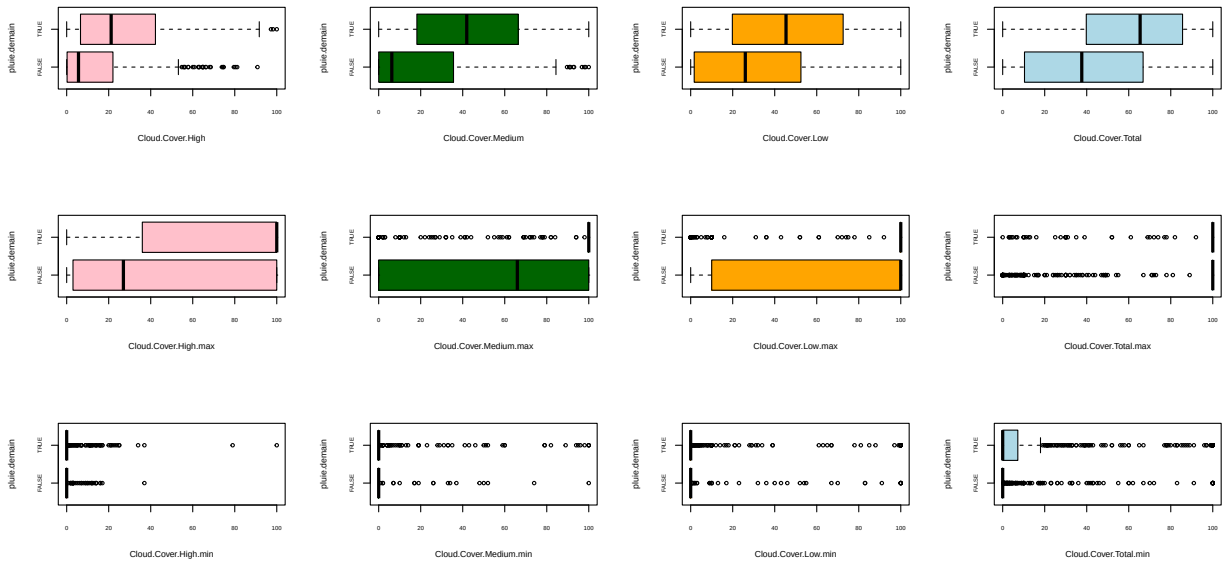
Trois variables sont liées à la pression au niveau de la mer.



Les trois variables semblent présenter un lien avec les précipitations. Le niveau de pression est faible en présence de précipitation. A contrario, il est élevé lorsqu'il ne pleut pas.

Nébulosité moyenne, minimale et maximale

Nous analysons dans un premier temps le potentiel lien entre la nébulosité à haute, moyenne et faible altitude ainsi que la couverture nuageuse totale.

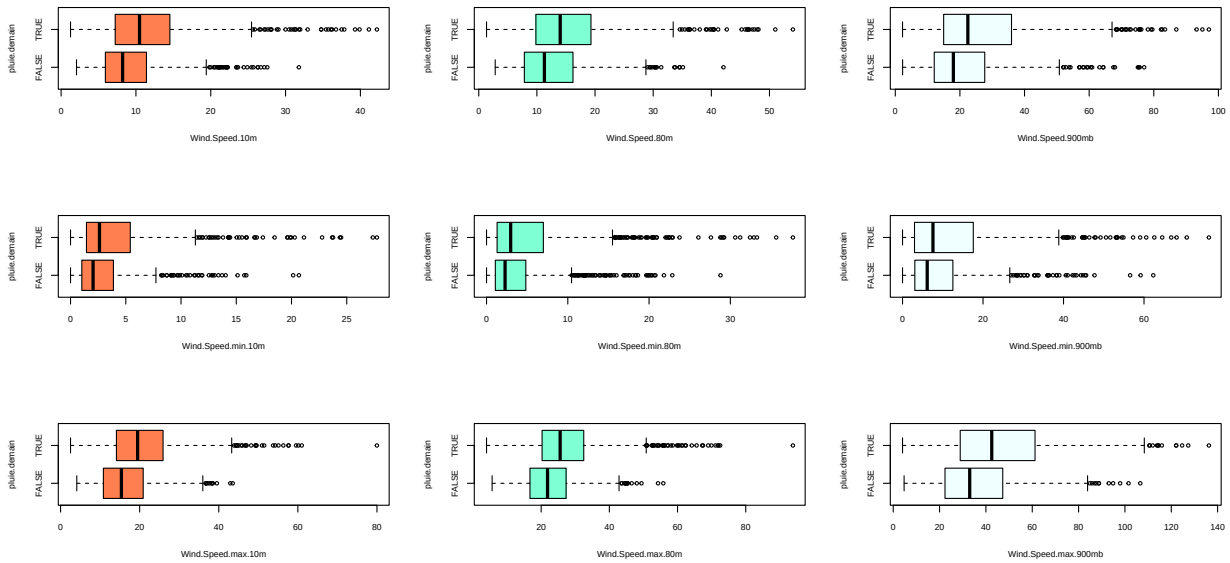


Le taux moyen de nébulosité (faible, moyenne, haute altitude et totale) est relié à la présence de précipitation. Il est plus élevé en cas d'occurence de la pluie. Dans ce cas, il est d'autant plus élevé que l'altitude est faible et, est le plus élevé pour la nébulosité totale.

Le lien est néanmoins moins affirmé entre le taux de nébulosité maximal à haute, moyenne et basse altitude et la variable d'intérêt. A ce stade, le taux de nébulosité minimal des trois altitudes et totale ne semblent pas être reliés à la présence de la pluie. Il sera nécessaire de réaliser des analyses plus détaillées dans la suite de notre étude.

Vitesse du vent moyenne, minimale et maximale

La vitesse du vent est mesurée à différentes altitudes: 10 mètres, 80 mètres et 900 milibares.

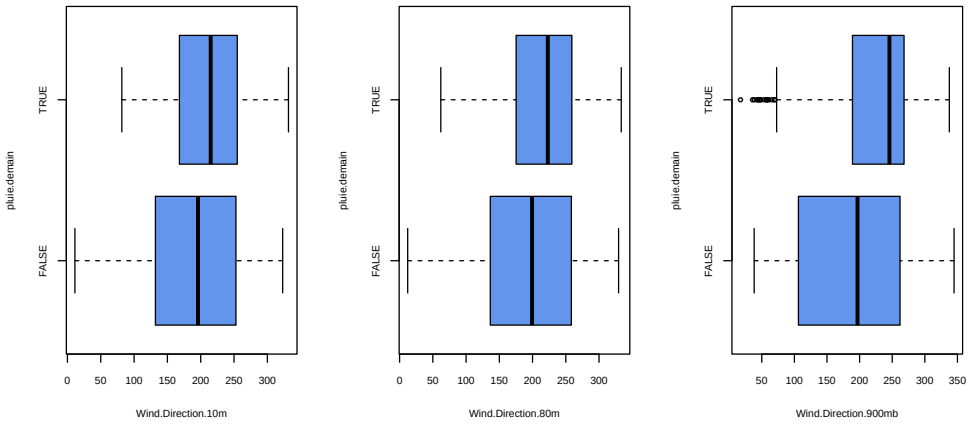


La vitesse moyenne et maximale du vent, pour les trois altitudes données, est plus élevée en cas de précipitation. Nous remarquons qu'à 900mb elle est la plus élevée.

La vitesse du vent minimale semble a contrario non reliée à la présence ou non de la pluie.

Direction du vent

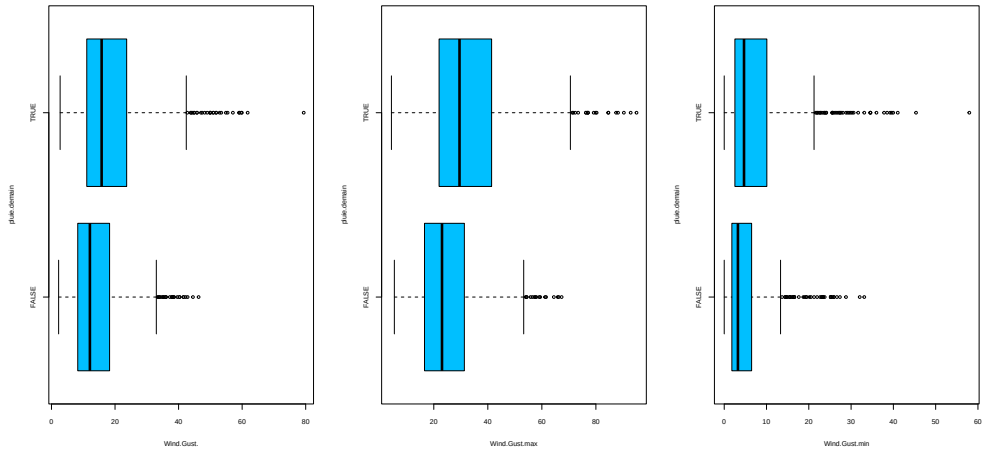
La direction du vent est ci-dessous représentée à différentes altitudes, 10m, 80m et 900mb.



Nous constatons que la direction du vent a un lien avec la variable pluie demain. Néanmoins, seul celle qui est mesurée à une altitude de 900mb a un lien fort avec la variable d'intérêt.

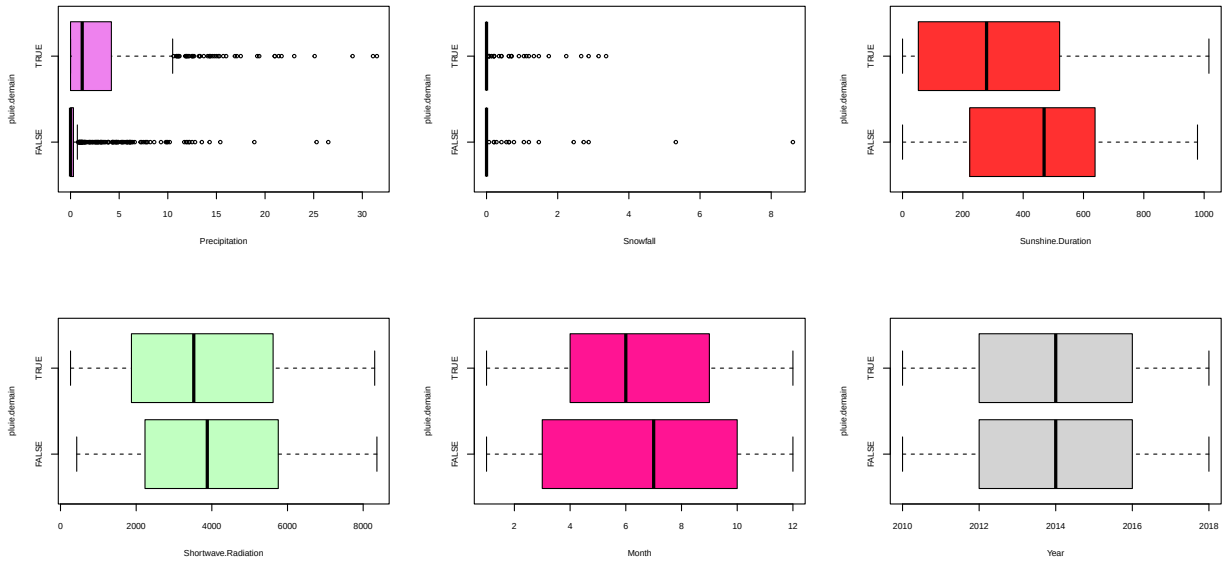
Rafales de vent moyenne, maximales et minimales

Parmi les trois variables relatives à la vitesse des rafales de vent, celle représentant la vitesse maximale des rafales de vent semble être celle la plus en lien avec la variable d'intérêt *pluie.demain*.



Précipitation, chutes de neige, minutes d'ensoleillement, rayonnement solaire et mois

Nous observons que les variables durée d'ensoleillement et rayonnement solaire sont négativement liées à la variable d'intérêt. Nous en déduisons ainsi qu'un niveau élevé d'ensoleillement et de rayonnement solaire limitent les chances de précipitation.



La présence de neige et de précipitation au jour J ne semblent pas être liées à la variable d'intérêt.

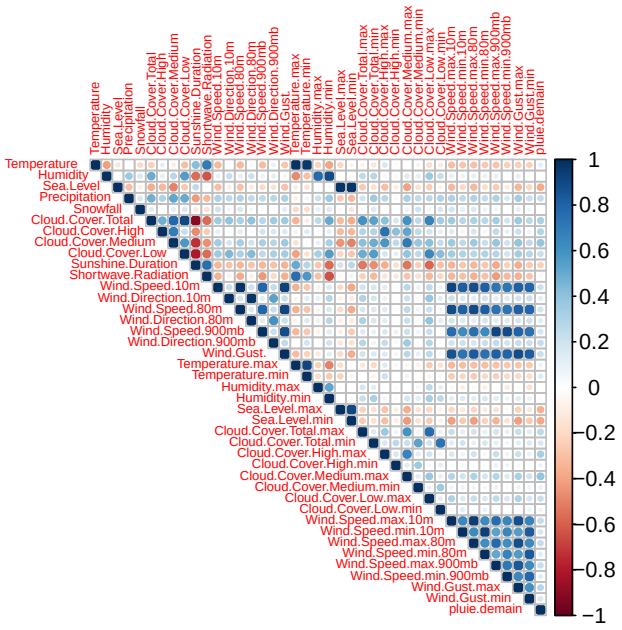
La variable *Month*, traduisant les saisons semble, quant à elle être liée à la variable pluie.demain.

Afin de poursuivre notre analyse descriptive, nous étudions dans la section suivante la corrélation entre chacune des variables explicatives.

2.2. Corrélation

Dans cette section nous analysons la corrélation entre les différents variables. La synthèse des corrélations entre les paires de variables est synthétisée dans la matrice de corrélation ci-dessous.

Nous observons qu'il y a une propension de forte corrélation positive par nature de variables, par exemple pour la température ou le vent. A l'inverse, nous constatons une forte corrélation négative pour certaines variables telles que Sunshine.Duration et Cloud.Cover.Total, ce qui nous semble météorologiquement cohérent.



Par ailleurs, nous remarquons que toutes les variables explicatives ont une corrélation plus ou moins faible avec pluie.demain, allant approximativement de -0,4 à +0,4. Aucune d'entre elles n'est fortement corrélée à la variable d'intérêt. Cela nous laisse penser que la prévision de la pluie est basée sur une combinaison plus ou moins complexe des différentes variables. Nous vérifierons cette hypothèse dans la troisième partie, lors de la définition de notre modèle de prédiction.

Nous tentons d'affiner notre analyse des corrélation en identifiant ci-dessous les variables dont la corrélation est supérieure ou égale à 90%.

Variable	Variable	Corrélation
Cloud.Cover.Total	Cloud.Cover.Low	0.9
Cloud.Cover.Total	Sunshine.Duration	-0.91
Wind.Speed.10m	Wind.Speed.80m	0.98
Wind.Direction.10m	Wind.Direction.80m	0.97
Wind.Speed.10m	Wind.Gust.	0.92
Wind.Speed.80m	Wind.Gust.	0.92
Temperature	Temperature.max	0.98
Temperature	Temperature.min	0.97
Temperature.max	Temperature.min	0.91
Sea.Level	Sea.Level.max	0.97
Sea.Level	Sea.Level.min	0.97
Sea.Level.max	Sea.Level.min	0.9
Wind.Speed.10m	Wind.Speed.max.10m	0.92
Wind.Speed.max.10m	Wind.Speed.max.80m	0.95

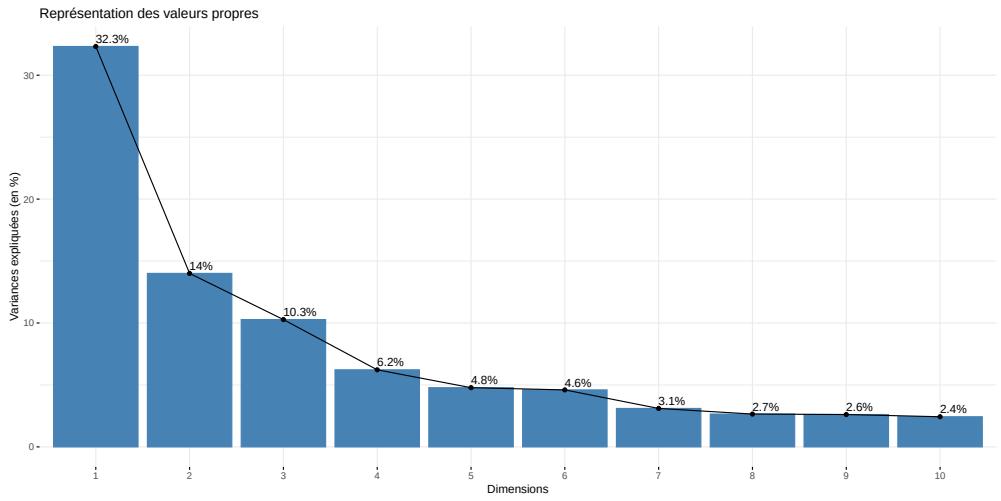
Variable	Variable	Corrélation
Wind.Speed.min.10m	Wind.Speed.min.80m	0.93
Wind.Speed.900mb	Wind.Speed.max.900mb	0.92

Nous observons que les plus fortes corrélations positives sont portées par les variables de même nature. Par exemple, pour Temperature et Sea.Level, la valeur moyenne est extrêmement corrélée à la valeur maximale, qui est elle-même très corrélée à la valeur minimale.

Les très fortes corrélations affichées peuvent prendre potentiellement impliquer une colinéarité entre les variables. Nous vérifierons ce point dans la troisième partie.

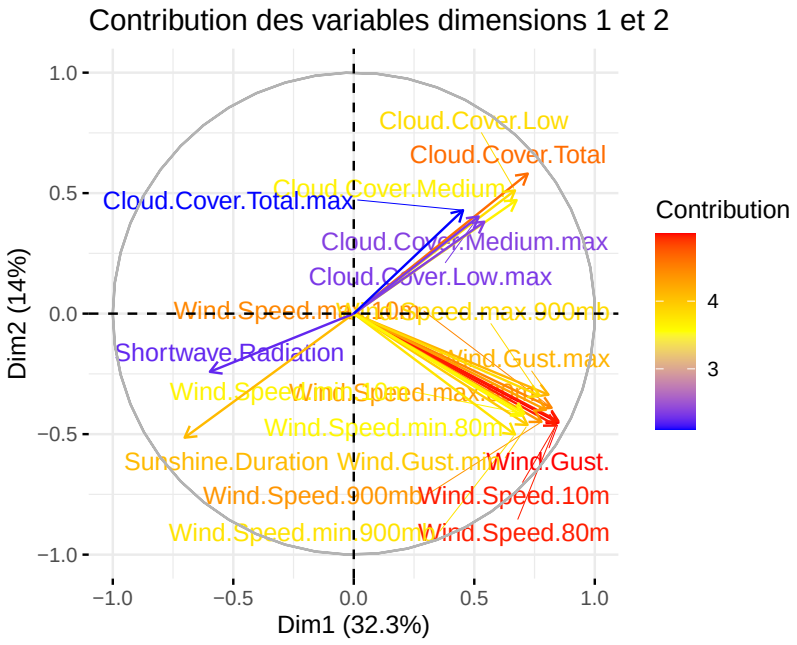
2.3. Analyse factorielle des composants

Dans cette section, nous réalisons une analyse factorielle des composants afin d'une part d'affiner notre vision sur la relation entre les différentes variables et d'autre part tenter d'identifier celles qui contribuent le plus sur les dimensions, qui portent le principal de l'information.



Nous constatons que 62,8% de la variance totale est portée par les dimensions 1 à 4. Le graphique indique que la plus grande part d'inertie est portée par le premier axe, qui résume 32,3% de l'information. Le deuxième et le troisième axes résument respectivement 14% et 10.3% de l'information. Enfin, le quatrième axe porte quant à lui 6.2% de la variance totale.

Nous analysons ci-dessous l'information portée par les dimensions 1 et 2, soit 46,3% de l'inertie totale.



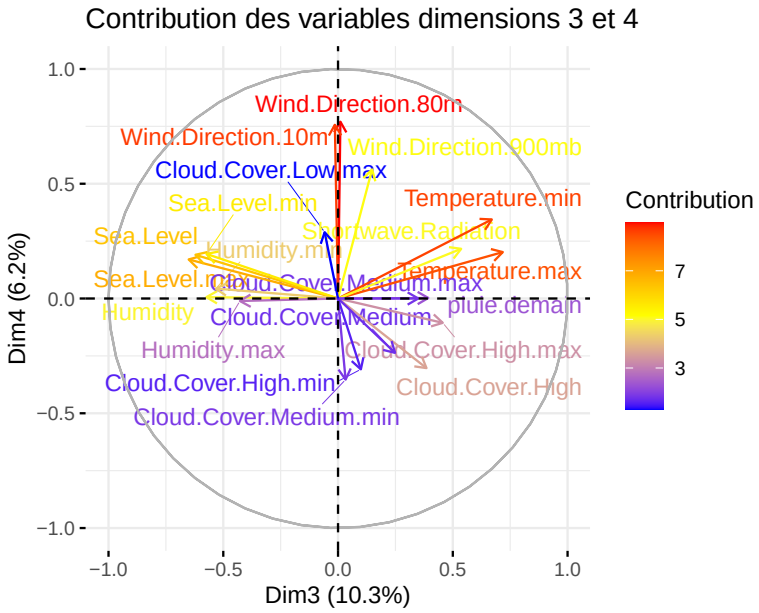
Nous constatons que l'axe 1 est fortement corrélé (positivement) aux variables liées au vent (rafales, vitesse et direction). Il l'est également, avec une corrélation un peu plus faible, aux variables liées à la nébulosité (aux différentes altitudes). L'axe 1 est en revenche très fortement et négativement corrélé aux variables durée d'ensoleillement et radiation solaire.

L'axe 2 indique que les variables liées à la nébulosité sont fortement et négativement corrélées à la durée d'ensoleillement et dans une moindre mesure aux radiations solaires. Cela rejoint nos premières observations, dans lesquelles nous avons constaté que plus il y a de nuages moins le ciel est ensoleillé et inversement.

De manière générale, les variables liées au vent semblent être en direction opposées à celles liées à la nébulosité. Néanmoins, il convient de nuancer ces propos en considérant les caractéristiques particulières du vent telles que sa vitesse ou encore le niveau d'altitude.

De la même manière, nous représentons sur les 20 variables les plus contributrices des axes 3 et 4.

Néanmoins, l'axe 3 est le plus fortement et positivement corrélé aux variables de température (notamment minimale et maximale). A l'inverse, il est négativement corrélé aux variables liées au niveau de la mer (moyenne, minimale et maximale). Les taux d'humidité relative (moyenne, minimale et maximale) sont, dans une moindre mesure, fortement et négativement correlés à l'axe 3. L'axe 4 semble être le plus fortement et négativement corrélé aux variables de nébulosité. Et est à l'inverse très positivement corrélée à la direction du vent.



Les variables liées à la température et à la pression au niveau de la mer sont les plus représentatives de l'axe 3. Elles sont par ailleurs à l'opposé les unes des autres.

En complément de nos observations sur les dimensions 1 et 2, la nébulosité à une tendance opposée à la direction du vent.

Dans cette partie nous retenons que la variable d'intérêt n'est pas explicable par certaines variables en particulier. Les "groupes" de variables de même nature (ou sous-groupes dans le cas de la nébulosité) ont des interactions positives ou négatives entre elles. Par ailleurs, l'existence de très fortes corrélations au sein de chaque goupe ou sous-groupe de variables de même nature indique la potentielle présence de colinéarité. Dans ce cadre, il sera nécessaire de ne retenir que certaines d'entre elles lors de l'ajustement de notre modèle de prédiction.

3. Modélisation

Afin de pouvoir réaliser analyser la qualité de la modélisation, nous commençons par découper l'ensemble d'entraînement en deux parties en utilisant le *holdout aléatoire*. La première partie, nommée *trainset*, sera constituée de 80% des observations et la seconde partie (*valset*) sera constituée des 20% d'observations restantes. Cela nous permettra d'entraîner notre modèle avec le *trainset* et de vérifier la qualité de notre ajustement avec le *valset*.

```
# holdout aléatoire
trainset = sample(c(T, F), nrow(meteo_train), replace = TRUE, prob = c(.8, .2))
```

Nous obtenons ainsi une répartition des deux parties de taille suivante:

trainset	valset
952	228

3.1. Identification des colinéarités

Nous étudions ci-dessous les potentielles colinérarités entre les variables. Dans le cas d'existence avérée de colinéarité, il conviendra de n'en retire qu'une dans l'ajustement de notre modèle de prédiction. La sélection sera effectuée lors de la définition et de l'analyse de nos différents modèles, dans la section suivante.

Analyse de la colinéarité des variables liées à la Température

```
modele3=glm(pluie.demain~ Temperature+Temperature.max+Temperature.min,
            data=meteo_train[trainset, ],
            family = "binomial")
```

	VIF	Correlation avec pluie.demain
Temperature	168.54	0.12
Temperature.max	63.05	0.09
Temperature.min	38.50	0.15

Nous constatons bien une colinéarité entre ces trois variables. Par ailleurs, la corrélation entre les variables *Température.min* et *pluie.demain* est la plus élevée.

Analyse de la colinéarité des variables liées à la pression au niveau de la mer

```
modele4=glm(pluie.demain~ Sea.Level+Sea.Level.max+Sea.Level.min,
            data=meteo_train,family = "binomial")
```

	VIF	Correlation avec pluie.demain
Sea.Level	154.93756	-0.37
Sea.Level.max	40.25781	-0.35
Sea.Level.min	51.06136	-0.39

Ces trois variables sont colinéaires. Par ailleurs, les variables *Sea.Level.min* et *pluie.demain* sont les plus fortement corrélées (négativement).

Analyse de la colinéarité des variables liées à la couverture nuageuse

```
modele5=glm(pluie.demain~ Cloud.Cover.Total+Cloud.Cover.Low+
                          Cloud.Cover.Medium+Cloud.Cover.Medium.max+
                          Cloud.Cover.High+
                          Cloud.Cover.High.max,
            data=meteo_train[trainset,],family = "binomial")
```

	VIF	Correlation avec pluie.demain
Cloud.Cover.Total	13.620922	0.32
Cloud.Cover.Low	8.164013	0.23
Cloud.Cover.Medium	4.466721	0.39
Cloud.Cover.Medium.max	2.339938	0.40
Cloud.Cover.High	3.218258	0.30
Cloud.Cover.High.max	2.801753	0.33

Nous constatons l'existence d'une colinéarité entre la nébulosité totale et la nébulosité à basse altitude. La variable *Cloud.Cover.Total* est celle qui est la plus corrélée avec la variable *pluie.demain*.

Nous analysons ci-dessous la colinéarité de la nébulosité maximale à haute et moyenne altitude.

```
#Modele 6: avec les variables liées à la couverture nuageuse 1
modele6=glm(pluie.demain~ Cloud.Cover.Medium.max+Cloud.Cover.High.max,
            data=meteo_train,family = "binomial")
```

```
##                          vif.modele6.
## Cloud.Cover.Medium.max      1.535452
## Cloud.Cover.High.max        1.535452
```

Malgré une forte corrélation entre les variables, nous ne constatons pas de colinéarité.

Nous analysons ci-dessous la colinéarité entre les variables de nébulosité maximale à basse altitude et totale.

```
modele7=glm(pluie.demain~ Cloud.Cover.Low.max+Cloud.Cover.Total.max,
            data=meteo_train[trainset,],family = "binomial")
```

```
##                         vif.modele7.
## Cloud.Cover.Low.max         1.82913
## Cloud.Cover.Total.max       1.82913
```

Malgré la forte corrélation entre ces deux varaiables, nous ne constatons pas de colinéarité.

Analyse de la colinéarité des variables liées à la direction du vent

L'ajustement réalisé ci-dessous permet d'identifier une potentielle colinéarité entre les variables liées à la direction du vent à différentes altitudes.

```
modele8=glm(pluie.demain~ Wind.Direction.10m+Wind.Direction.80m+
              Wind.Direction.900mb,
            data=meteo_train[trainset, ],family = "binomial")
```

	VIF	Correlation avec pluie.demain
Wind.Direction.10m	21.466719	0.12
Wind.Direction.80m	24.677557	0.13
Wind.Direction.900mb	1.829503	0.24

Nous observons une colinéarite entre les variables *Wind.Direction.10m* et *Wind.Direction.80m*. Leur corrélation avec la variable *pluie.demain* est sensiblement identique.

Analyse de la colinéarité entre les variables liées à la direction du vent et les rafales de vent

```
modele9.1= glm(pluie.demain~Wind.Direction.80m+Wind.Gust.,
               data=meteo_train[trainset, ],family = "binomial")
```

```
## Wind.Direction.80m         Wind.Gust.
##           1.025702           1.025702
```

```
modele9.2= glm(pluie.demain~Wind.Direction.10m+Wind.Gust.,
               data=meteo_train[trainset, ],family = "binomial")
vif(modele9.2)
```

```
## Wind.Direction.10m         Wind.Gust.
##            1.02892            1.02892
```

Malgré une forte corrélation, nous ne constatons pas de colinéarité entre la direction du vent à 10m et à 80m d'altitude et les rafales de vent.

Analyse de la colinéarité entre les variables liées à la vitesse du vent

```
modele9.3= glm(pluie.demain~Wind.Speed.10m+Wind.Speed.max.10m+Wind.Speed.max.80m,
               data=meteo_train[trainset, ],family = "binomial")
```

	VIF	Correlation avec pluie.demain
Wind.Speed.10m	6.014853	0.21
Wind.Speed.max.10m	11.260214	0.25
Wind.Speed.max.80m	8.075257	0.24

```
modele9.4= glm(pluie.demain~Wind.Speed.min.10m+Wind.Speed.min.80m,
               data=meteo_train[trainset, ],family = "binomial")
```

	VIF	Correlation avec pluie.demain
Wind.Speed.min.10m	6.640704	0.17
Wind.Speed.min.80m	6.640704	0.13

Nous constatons dans les deux tableaux ci-dessus une colinéarité entre les variables.

```
modele9.6=glm(pluie.demain~ Wind.Speed.min.900mb+Wind.Speed.900mb,
              data=meteo_train[trainset,],family = "binomial")
```

	VIF	Correlation avec pluie.demain
Wind.Speed.min.900mb	5.218588	0.12
Wind.Speed.900mb	5.218588	0.19

Une faible colinéarité est constatée. La variable Wind.Speed.900mb est celle qui est la plus corrélée avec *pluie.demain*.

Analyse de la colinéarité entre la vitesse du vent et les rafales de vent

```
modele9.5=glm(pluie.demain~ Wind.Speed.max.10m+Wind.Gust.max,
            data=meteo_train,family = "binomial")
```

```
## Wind.Speed.max.10m      Wind.Gust.max
##           3.510419           3.510419
```

Nous ne constatons pas de colinéarité entre la vitessse maximale du vent à 10m d'altitude et les rafales de vent maximales.

Analyse de la colinéarité des variables liées à l'humidité

```
modele11=glm(pluie.demain~ Humidity+Humidity.min+Humidity.max,
             data=meteo_train[trainset, ],family = "binomial")
```

	VIF	Correlation avec pluie.demain
Humidity	12.842272	0.03
Humidity.min	7.262620	0.03
Humidity.max	3.578251	0.00

Ces trois variables sont colinéaires. Les variables *Humidity* et *Humidity.max* sont corrélées avec *pluie.demain* de manière équivalente.

Suite à l'identification des colinéarités entre certaines variables, nous procédons dans la section suivante à l'ajustement de différents modèles.

3.2. Ajustement des modèles

Compte-tenu des différentes combinaisons de variables explicatives permettant de prévoir l'occurence de la pluie le lendemain, nous procédons dans cette section à l'ajustement de différents modèles.

Plusieurs critères permettant de comparer les modèles concurrents peuvent être mobilisés, tels que le critère AIC (Akaike Information Criterion) ou bien encore le BIC (Bayesian Information Criterion). Ces deux critères pénalisent les modèles plus complexes, avec une plus forte propension pour le BIC, ils permettent ainsi d'éviter un phénomène de sur-apprentissage. Du fait de notre hypothèse sur la complexité des combinaisons de variables explicatives, nous choisissons de retenir l'AIC comme critère de comparaison des modèles, celui-ci offrant le meilleur compromis entre ajustement et complexité.

Modèle complet

Dans un premier temps, nous ajustons un modèle avec l'ensemble des variables, à l'exception de Year, Day, Hour, Minute qui ne semblent pas avoir d'impact sur les chances de précipitation.

```
modele1=glm(pluie.demain~.-Year-Day-Hour-Minute,
            data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ . - Year - Day - Hour - Minute,
##     family = "binomial", data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5638  -0.8464   0.1215   0.8579   2.8179
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             6.653e+01  1.410e+01   4.719 2.37e-06 ***
## Month                  -2.693e-02  2.739e-02  -0.983 0.325499
## Temperature             1.754e-01  1.792e-01   0.978 0.327922
```

```
## Humidity                2.186e-02  3.539e-02   0.618 0.536826
## Sea.Level               5.372e-01  1.548e-01   3.469 0.000522 ***
## Precipitation           3.874e-02  3.112e-02   1.245 0.213189
## Snowfall               -6.201e-02  2.909e-01  -0.213 0.831201
## Cloud.Cover.Total       1.428e-02  1.300e-02   1.098 0.272011
## Cloud.Cover.High       -5.328e-03  7.446e-03  -0.716 0.474236
## Cloud.Cover.Medium      6.278e-03  7.536e-03   0.833 0.404804
## Cloud.Cover.Low        -8.419e-03  8.911e-03  -0.945 0.344783
## Sunshine.Duration       7.242e-04  9.719e-04   0.745 0.456162
## Shortwave.Radiation    -8.039e-05  1.107e-04  -0.726 0.467900
## Wind.Speed.10m         -9.177e-02  1.078e-01  -0.852 0.394407
## Wind.Direction.10m      2.696e-03  6.180e-03   0.436 0.662688
## Wind.Speed.80m         -6.593e-02  7.716e-02  -0.854 0.392847
## Wind.Direction.80m     -4.595e-03  6.439e-03  -0.714 0.475439
## Wind.Speed.900mb        1.993e-02  2.881e-02   0.692 0.489161
## Wind.Direction.900mb    4.753e-03  1.603e-03   2.965 0.003028 **
## Wind.Gust.              2.316e-02  4.122e-02   0.562 0.574274
## Temperature.max        -1.132e-02  1.049e-01  -0.108 0.914047
## Temperature.min        -1.215e-01  9.547e-02  -1.273 0.203042
## Humidity.max           -1.315e-02  2.237e-02  -0.588 0.556693
## Humidity.min           -1.407e-02  2.012e-02  -0.699 0.484253
## Sea.Level.max          -2.695e-01  8.246e-02  -3.269 0.001081 **
## Sea.Level.min          -3.364e-01  8.506e-02  -3.955 7.65e-05 ***
## Cloud.Cover.Total.max   5.391e-03  5.450e-03   0.989 0.322501
## Cloud.Cover.Total.min   6.438e-03  6.722e-03   0.958 0.338221
## Cloud.Cover.High.max    3.511e-03  3.221e-03   1.090 0.275698
## Cloud.Cover.High.min    2.607e-02  2.738e-02   0.952 0.341003
## Cloud.Cover.Medium.max  4.569e-03  3.493e-03   1.308 0.190904
## Cloud.Cover.Medium.min -8.676e-03  1.005e-02  -0.863 0.387968
## Cloud.Cover.Low.max     2.478e-03  3.792e-03   0.653 0.513457
## Cloud.Cover.Low.min     6.550e-05  7.696e-03   0.009 0.993210
## Wind.Speed.max.10m      7.839e-02  3.838e-02   2.042 0.041110 *
## Wind.Speed.min.10m      1.615e-01  7.232e-02   2.233 0.025548 *
## Wind.Speed.max.80m     -5.384e-03  3.208e-02  -0.168 0.866730
## Wind.Speed.min.80m     -4.123e-02  4.748e-02  -0.868 0.385256
## Wind.Speed.max.900mb   -1.238e-02  1.339e-02  -0.925 0.355019
## Wind.Speed.min.900mb   -7.900e-03  2.094e-02  -0.377 0.706015
## Wind.Gust.max           1.042e-02  1.890e-02   0.552 0.581233
## Wind.Gust.min           6.959e-04  3.048e-02   0.023 0.981782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1008.8  on 910  degrees of freedom
## AIC: 1092.8
##
## Number of Fisher Scoring iterations: 4
```

Nous constatons que relativement peu de variables sont significatives dans ce modèle complet. Seules les variables liées à Sea.Level (moyenne, minimum, maximum), *Wind.Speed.min.10m* et *Wind.Speed.max.10m* et *Speed.Direction.900mb* sont significatives.

Nous remarquons également que le modèle considère significatives des variables que nous avons identifiés

comme colinéaires, à l'instar des variables liées à Sea.Level.

Par ailleurs, nos précèdentes observations nous ont amené à supposer que l'occurence des précipitations implique une combinaison complexe de plusieurs facteurs météorologiques. Le nombre de variables significatives semble ainsi assez faible. De ce fait, nous tentons ci-dessous d'identifier si d'autres variables sont significatives dans le cadre d'un modèle qui sélectionne automatiquement les variables.

Modèle de sélection automatique des variables

Ci-dessous, nous mobilisons la méthode *backward* pour ajuster le modèle de sélection automatique des variables.

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature + Sea.Level + Cloud.Cover.Total +
##     Wind.Speed.80m + Wind.Direction.80m + Wind.Direction.900mb +
##     Wind.Gust. + Sea.Level.max + Sea.Level.min + Cloud.Cover.High.max +
##     Cloud.Cover.Medium.max + Wind.Speed.max.10m + Wind.Speed.min.10m,
##     family = "binomial", data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3856  -0.8518   0.1483   0.8716   2.7775
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            75.168872  12.849524   5.850 4.92e-09 ***
## Temperature             0.040318   0.014278   2.824 0.004747 **
## Sea.Level               0.480974   0.143080   3.362 0.000775 ***
## Cloud.Cover.Total       0.008331   0.003594   2.318 0.020449 *
## Wind.Speed.80m         -0.124438   0.036457  -3.413 0.000642 ***
## Wind.Direction.80m     -0.002859   0.001650  -1.733 0.083104 .
## Wind.Direction.900mb    0.004319   0.001388   3.112 0.001856 **
## Wind.Gust.              0.036803   0.020711   1.777 0.075580 .
## Sea.Level.max          -0.242009   0.075704  -3.197 0.001390 **
## Sea.Level.min          -0.315492   0.079581  -3.964 7.36e-05 ***
## Cloud.Cover.High.max    0.004457   0.002362   1.887 0.059126 .
## Cloud.Cover.Medium.max  0.007115   0.002741   2.596 0.009428 **
## Wind.Speed.max.10m      0.069608   0.021961   3.170 0.001526 **
## Wind.Speed.min.10m      0.084259   0.037710   2.234 0.025458 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1024.1  on 938  degrees of freedom
## AIC: 1052.1
##
## Number of Fisher Scoring iterations: 4
```

Le meilleur modèle, au sens de l'AIC, sélectionne automatiquement 12 variables explicatives. Nous constatons que la significativité des variables *Wind.Direction.80m*, *Wind.Gust.* et *Cloud.Cover.High.max* est moindre comparativement aux autres variables. Nous constatons que le modèle a sélectionné des variables colinéaires (*Sea.Level*, *Sea.Level.min* et *Sea.Level.max*).

Le critière AIC est quant à lui amélioré par rapport au modèle complet (1052.1 vs 1092.8).

Afin de tester l'existence potentielle d'autres variables significatives, nous ajustons un modèle sélectionnant automatiquement les variables, via la méthode *stepwise*.

```
##
## Call:
## glm(formula = pluie.demain ~ Sea.Level.min + Cloud.Cover.Medium.max +
##     Wind.Direction.900mb + Wind.Gust.max + Temperature.max +
##     Cloud.Cover.Medium + Month + Cloud.Cover.Low.max + Sea.Level +
##     Sea.Level.max, data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0098  -0.3503   0.0239   0.3396   1.1469
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            11.5049829  2.3090501   4.983 7.47e-07 ***
## Sea.Level.min          -0.0475923  0.0118277  -4.024 6.19e-05 ***
## Cloud.Cover.Medium.max  0.0016758  0.0004846   3.458 0.000568 ***
## Wind.Direction.900mb    0.0006615  0.0002058   3.214 0.001353 **
## Wind.Gust.max           0.0031645  0.0011815   2.678 0.007529 **
## Temperature.max         0.0083602  0.0021303   3.924 9.33e-05 ***
## Cloud.Cover.Medium      0.0023490  0.0007436   3.159 0.001635 **
## Month                  -0.0069657  0.0041670  -1.672 0.094929 .
## Cloud.Cover.Low.max     0.0006501  0.0004313   1.507 0.132062
## Sea.Level               0.0708872  0.0223790   3.168 0.001587 **
## Sea.Level.max          -0.0347081  0.0125769  -2.760 0.005899 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1862671)
##
##     Null deviance: 238.00  on 951  degrees of freedom
## Residual deviance: 175.28  on 941  degrees of freedom
## AIC: 1114.7
##
## Number of Fisher Scoring iterations: 2
```

Parmi les dix variables sélectionnées par le modèle, seul Cloud.Cover.Low.max n'est pas significative. Le modèle a retenu les variables Sea.Level.min et Sea.Level.max qui sont colinéaires.

Le critère AIC est pour sa part largement dégradé, en comparaison du modèle précédent (1114.7 vs 1052.1).

Le modèle complet ainsi que les modèles de sélection automatique de variables (*backward* et *stepwise*) n'ayant pas défini une combinaison de variables nous semblant satisfaisante au regard de l'analyse descriptive, nous définissons manuellement un nouveau modèle.

Modèle manuel

Nous ajustons un nouveau modèle avec les variables significatives identifiées dans le modèle 1, tout en retirant des variables colinéaires (Sea.Level, Sea.Level.max,...).

Sachant d'une part, que les nuages sont un facteur de causalité des précipitations, et que d'autre part, parmi les variables liée à la nébulosité, la variable *Cloud.Cover.Total* est celle qui est la plus liée à la variable *pluie.demain* (cf. partie descriptive). Nous choisissons donc de l'intégrer dans le modèle.

Météorologiquement, des températures plus basses au sein des nuages augmentant les risques de précipitations, nous choisissons d'intégrer la variable *Temperature.min* dans notre modèle. Par ailleurs, au sein du *groupe* de variables de nature Température, elle est celle qui est la plus liée à la variable d'intérêt (cf. partie descriptive).

De la même manière la variable *Sea.Level.min* est celle, parmi son groupe de même nature, qui présente le lien le plus élevé avec *pluie.demain*. C'est elle que nous conservons dans notre modèle.

Une vitesse du vent à 10m élevé impacte la dispersion des nuages, réduisant ainsi les chances de précipitations à très court terme. A l'inverse, la vitesse du vent à une altitude où la pression vaut 900hPa (i.e. plus de 1000m) transporte de l'humidité, ce qui favorise la création de nuages. Cela contribue à augmenter les risques de pluie sur les jours à venir.

Sur la base de ces postulats, nous incluons les variables *Wind.Speed.min.10m* et *Wind.Direction.900mb*.

```
modele2=glm(pluie.demain ~ Temperature.min + Sea.Level.min
                + Cloud.Cover.Total + Wind.Speed.min.10m
                +Wind.Speed.max.900mb,
            family = "binomial", data = meteo_train[trainset, ])
summary(modele2)
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.min + Sea.Level.min +
##     Cloud.Cover.Total + Wind.Speed.min.10m + Wind.Speed.max.900mb,
##     family = "binomial", data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3054  -0.9408   0.2356   0.9511   2.5553
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          83.319280  11.252074   7.405 1.31e-13 ***
## Temperature.min       0.060549   0.012733   4.755 1.98e-06 ***
## Sea.Level.min        -0.083960   0.011029  -7.612 2.69e-14 ***
## Cloud.Cover.Total     0.014563   0.002646   5.504 3.72e-08 ***
## Wind.Speed.min.10m    0.015481   0.024235   0.639  0.52296
## Wind.Speed.max.900mb  0.013465   0.004180   3.221  0.00128 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1084.4  on 946  degrees of freedom
## AIC: 1096.4
##
## Number of Fisher Scoring iterations: 4
```

Nous observons que seule la variable *Wind.Speed.min.10m* n'est pas significative, ce qui semble indiquer que à basse altitude, la vitesse du vent n'impacte pas ou peu la prédiction de précipitation dans le temps.

Au regard du critère AIC, celui-ci a augmenté par rapport au modèle de sélection automatique des variables *backward* (1096.4 vs 1052.1).

Modèle 12

Généralement outre la vitesse, la mesure du vent comprend un second paramètre, qui correspond à sa direction. Dans l'analyse exploratoire, nous avons identifié que *Wind.Direction.900mb* est la variable du groupe de nature "direction du vent", qui est la plus liée à la variable *pluie.demain*.

A cette altitude, la direction du vent combinée à sa vitesse peut favoriser la formation de dépressions et de fronts atmosphériques. Ces derniers étant des systèmes météorologiques qui sont souvent accompagnés de conditions propices à la précipitation. Par conséquent, nous intégrons la variable *Wind.Direction.900mb* dans un nouveau modèle.

```
modele12=glm(pluie.demain~ Temperature.min+Cloud.Cover.Total
             +Wind.Speed.max.900mb+Wind.Direction.900mb
             +Sea.Level.min,
             data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.min + Cloud.Cover.Total +
##     Wind.Speed.max.900mb + Wind.Direction.900mb + Sea.Level.min,
##     family = "binomial", data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2600  -0.9433   0.2450   0.9318   2.5976
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          85.058362  11.235972   7.570 3.73e-14 ***
## Temperature.min       0.049786   0.013157   3.784 0.000154 ***
## Cloud.Cover.Total     0.012664   0.002741   4.621 3.83e-06 ***
## Wind.Speed.max.900mb  0.011899   0.003971   2.996 0.002732 **
## Wind.Direction.900mb  0.003010   0.001105   2.724 0.006456 **
## Sea.Level.min        -0.085991   0.011023  -7.801 6.13e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1077.4  on 946  degrees of freedom
## AIC: 1089.4
##
## Number of Fisher Scoring iterations: 4
```

Nous constatons que l'ensemble des variables que nous avons sélectionné dans notre modèle sont significatives.

Le critère AIC est légèrement meilleur que le modèle précédent (1089.4 vs 1096.4).

Modèle 13

Nous avons constaté dans la partie exploratoire que la variable *Cloud.Cover.Medium.max* a un lien avec *pluie.demain*. Par ailleurs, une couverture nuageuse maximale à moyenne altitude peut renforcer la nébulosité totale. Nous l'ajoutons donc à notre précèdent ajustement pour en analyser l'impact.

```
modele13=glm(pluie.demain~ Temperature.min
             +Cloud.Cover.Total+Cloud.Cover.Medium.max
             +Wind.Speed.max.900mb
             +Wind.Direction.900mb
             +Sea.Level.min,
             data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.min + Cloud.Cover.Total +
##     Cloud.Cover.Medium.max + Wind.Speed.max.900mb + Wind.Direction.900mb +
##     Sea.Level.min, family = "binomial", data = meteo_train[trainset,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1889  -0.8795   0.2832   0.8831   2.6325
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            74.741634  11.388785   6.563 5.28e-11 ***
## Temperature.min         0.044009   0.013370   3.292 0.000996 ***
## Cloud.Cover.Total       0.005943   0.003151   1.886 0.059267 .
## Cloud.Cover.Medium.max  0.010211   0.002309   4.422 9.77e-06 ***
## Wind.Speed.max.900mb    0.010717   0.004010   2.672 0.007532 **
## Wind.Direction.900mb    0.002604   0.001132   2.301 0.021371 *
## Sea.Level.min          -0.076046   0.011162  -6.813 9.58e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1057.2  on 945  degrees of freedom
## AIC: 1071.2
##
## Number of Fisher Scoring iterations: 4
```

La considération de la variable *Cloud.Cover.Medium.max* semble diminuer la significativité des variables *Cloud.Cover.Total* et *Wind.Direction.900mb*. Ces dernières restent néanmoins significatives, ainsi que le reste des autres variables.

Nous constatons que le critère AIC est néanmoins plus faible que le modèle précèdent (1071.2 vs 1089.4).

Modèle 14

Compte-tenu de nos hypothèses concernant la vitesse et la direction du vent à une altitude où la pression est de 900hPa, nous nous questionnons sur l'impact de ces deux paramètres à une altitude de 80m. Le modèle suivant prend ainsi en compte les variables *Wind.Speed.max.80m* et *Wind.Direction.80m* dont nous tenterons d'analyser les effets sur le modèle.

```
modele14=glm(pluie.demain~ Temperature.min
             +Cloud.Cover.Total+Cloud.Cover.Medium.max
             +Wind.Speed.max.80m
```

```
            +Wind.Direction.80m+Wind.Direction.900mb
            +Sea.Level.max,
            data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.min + Cloud.Cover.Total +
##     Cloud.Cover.Medium.max + Wind.Speed.max.80m + Wind.Direction.80m +
##     Wind.Direction.900mb + Sea.Level.max, family = "binomial",
##     data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0321  -0.9074   0.3105   0.8812   2.5790
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            65.984920  12.144283   5.433 5.53e-08 ***
## Temperature.min         0.038061   0.013874   2.743 0.006082 **
## Cloud.Cover.Total       0.007461   0.003254   2.293 0.021862 *
## Cloud.Cover.Medium.max  0.010343   0.002316   4.466 7.97e-06 ***
## Wind.Speed.max.80m      0.030907   0.008064   3.833 0.000127 ***
## Wind.Direction.80m     -0.002154   0.001586  -1.358 0.174390
## Wind.Direction.900mb    0.004116   0.001333   3.088 0.002016 **
## Sea.Level.max          -0.067278   0.011846  -5.679 1.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1066.2  on 944  degrees of freedom
## AIC: 1082.2
##
## Number of Fisher Scoring iterations: 4
```

Nous constatons que la variable *Wind.Direction.80m* n'est pas significative. Par ailleurs, nous constatons que la considération des paramètres vitesse et direction du vent à une altitude de 80m dégrade le critère AIC comparativement à notre précédent modèle (1082.2 vs 1071.2). Par conséquent, nous ne la retenons pas pour notre modélisation.

Modèle 15

Nous avons précédement constaté que la variable *Month* peut apporter des informations quant à la saison. La ville de Bâle étant située sur une région montagneuse du continent européen, la saison autonmale est en général associée à une situation propice aux précipitations ainsi qu'à une baisse de la température. Ce postulat nous amène à considérer l'interaction entre les variables *Month* et *Temperature.min*.

```
modele15=glm(pluie.demain~ Temperature.min
            +Cloud.Cover.Total+Cloud.Cover.Medium.max
            +Wind.Speed.max.80m
            +Wind.Direction.900mb
            +Sea.Level.min
```

```
            +Month:Temperature.min,
            data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.min + Cloud.Cover.Total +
##     Cloud.Cover.Medium.max + Wind.Speed.max.80m + Wind.Direction.900mb +
##     Sea.Level.min + Month:Temperature.min, family = "binomial",
##     data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2064  -0.9121   0.2745   0.8927   2.7048
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            68.747776  11.841013   5.806 6.40e-09 ***
## Temperature.min         0.145295   0.030866   4.707 2.51e-06 ***
## Cloud.Cover.Total       0.006681   0.003162   2.113 0.034614 *
## Cloud.Cover.Medium.max  0.010000   0.002337   4.279 1.88e-05 ***
## Wind.Speed.max.80m      0.019774   0.008567   2.308 0.020986 *
## Wind.Direction.900mb    0.003118   0.001116   2.793 0.005221 **
## Sea.Level.min          -0.070352   0.011574  -6.078 1.22e-09 ***
## Temperature.min:Month  -0.013793   0.003738  -3.689 0.000225 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1044.8  on 944  degrees of freedom
## AIC: 1060.8
##
## Number of Fisher Scoring iterations: 4
```

Nous constatons que l'ensemble des variables, dont l'interaction entre la température et le mois sont significatifs dans le nouveau modèle que nous avons ajusté.

Comparativement au modèle précédent, le critère AIC en est par ailleurs amélioré (1060.8 vs 1082.2).

Modèle 16

Dans ce modèle nous considérons l'interaction entre les variables *Wind.Speed.max.80m* et *Wind.Direction.900mb*. En effet, l'interaction entre la vitesse et la direction du vent à différentes altitudes peut créer des zones de convergence de l'air au niveau de l'atmosphère. En météorologie, la convergence désigne une région de l'atmosphère où les flux d'air de différentes directions se rejoignent pour créer une accumulation de masse pouvant éventuellement mener à la formation de nuages et de précipitations.

Par ailleurs, une température élevée combinée à une convergence météorologique peut potentiellement avoir un impact sur les conditions favorisant les précipitations. Lorsque l'air monte, un refroidissement se produit, provoquant une condensation. Ce phénomène peut potentiellement être accentué lorsque la température est plus élevée et ainsi créer une situation propice à la formation de nuages et de précipitations. Ce postulat nous amène à considérer la variable *Temperature.max* au lieu de la *Temperature.min*. Nous conserverons néanmoins l'interaction *Month:Temperature.min*.

```r
modele16=glm(pluie.demain~ Temperature.max
             +Cloud.Cover.Total+Cloud.Cover.Medium.max
             +Wind.Speed.max.80m
             +Wind.Direction.900mb
             +Wind.Speed.max.80m:Wind.Direction.900mb
             +Sea.Level.min
             +Month:Temperature.min,
             data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.max + Cloud.Cover.Total +
##     Cloud.Cover.Medium.max + Wind.Speed.max.80m + Wind.Direction.900mb +
##     Wind.Speed.max.80m:Wind.Direction.900mb + Sea.Level.min +
##     Month:Temperature.min, family = "binomial", data = meteo_train[trainset,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2401  -0.8644   0.2369   0.8869   2.7959
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             64.6802141 12.0994160   5.346 9.01e-08
## Temperature.max                          0.1007648  0.0206683   4.875 1.09e-06
## Cloud.Cover.Total                        0.0130632  0.0036189   3.610 0.000307
## Cloud.Cover.Medium.max                   0.0098231  0.0023582   4.165 3.11e-05
## Wind.Speed.max.80m                      -0.0401666  0.0282221  -1.423 0.154668
## Wind.Direction.900mb                    -0.0020604  0.0030237  -0.681 0.495614
## Sea.Level.min                           -0.0662194  0.0117463  -5.637 1.73e-08
## Wind.Speed.max.80m:Wind.Direction.900mb  0.0002604  0.0001182   2.204 0.027554
## Month:Temperature.min                   -0.0094149  0.0028383  -3.317 0.000909
##
## (Intercept)                             ***
## Temperature.max                         ***
## Cloud.Cover.Total                       ***
## Cloud.Cover.Medium.max                  ***
## Wind.Speed.max.80m
## Wind.Direction.900mb
## Sea.Level.min                           ***
## Wind.Speed.max.80m:Wind.Direction.900mb *
## Month:Temperature.min                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1037.4  on 943  degrees of freedom
## AIC: 1055.4
##
## Number of Fisher Scoring iterations: 4
```

Nous constatons que l'ajout de l'interaction entre les variables *Wind.Speed.max.80m* et *Wind.Direction.900mb* induit la non significativité des variables en elles-mêmes. L'interaction est quant à elle significative. Nous constatons une forte significativité des autres variables explicatives que nous considérons dans notre modèle. Cela tend à confirmer nos suppositions.

Le critère AIC est de nouveau amélioré en comparaison du modèle précédent (1055.4 vs 1060.8).

Modèle 17

Le retrait des variables non significatives du modèle précédent (*Wind.Speed.max.80m* et *Wind.Direction.900mb*) nous amène à considérer le modèle suivant.

```
modele17=glm(pluie.demain~ Temperature.max
             +Cloud.Cover.Total+Cloud.Cover.Medium.max
             +Wind.Speed.max.80m:Wind.Direction.900mb
             +Sea.Level.min
             +Month:Temperature.min,
             data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.max + Cloud.Cover.Total +
##     Cloud.Cover.Medium.max + Wind.Speed.max.80m:Wind.Direction.900mb +
##     Sea.Level.min + Month:Temperature.min, family = "binomial",
##     data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2538  -0.8819   0.2175   0.8824   2.7842
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             5.986e+01  1.177e+01   5.088 3.63e-07
## Temperature.max                         1.015e-01  2.062e-02   4.925 8.46e-07
## Cloud.Cover.Total                       1.351e-02  3.594e-03   3.760  0.00017
## Cloud.Cover.Medium.max                  9.491e-03  2.334e-03   4.066 4.79e-05
## Sea.Level.min                          -6.224e-02  1.148e-02  -5.424 5.83e-08
## Wind.Speed.max.80m:Wind.Direction.900mb  1.284e-04  2.746e-05   4.673 2.96e-06
## Month:Temperature.min                  -8.629e-03  2.777e-03  -3.108  0.00189
##
## (Intercept)                            ***
## Temperature.max                        ***
## Cloud.Cover.Total                      ***
## Cloud.Cover.Medium.max                 ***
## Sea.Level.min                          ***
## Wind.Speed.max.80m:Wind.Direction.900mb ***
## Month:Temperature.min                  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1040.7  on 945  degrees of freedom
## AIC: 1054.7
```

```
##
## Number of Fisher Scoring iterations: 4
```

L'ensemble des variables considérées sont très significatives dans notre modèle. Le retrait des variables non significatives du modèle précédent améliore sensiblement le critère AIC (1054.7 vs 1055.4).

Modèle 18

Dans ce nouveau modèle, nous ajoutons l'interaction entre les variables *Cloud.Cover.Medium.max* et *Humidity.max* afin d'étudier l'effet du taux d'humitité relatif (maximal) dans des conditions qui peuvent mener à une instabilité atmosphérique.

```
modele18=glm(pluie.demain~ Temperature.max
             +Cloud.Cover.Total+Cloud.Cover.Medium.max:Humidity.max
             +Wind.Speed.max.80m:Wind.Direction.900mb
             +Sea.Level.min
             +Month:Temperature.min,
             data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.max + Cloud.Cover.Total +
##     Cloud.Cover.Medium.max:Humidity.max + Wind.Speed.max.80m:Wind.Direction.900mb +
##     Sea.Level.min + Month:Temperature.min, family = "binomial",
##     data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2596  -0.8877   0.2155   0.8890   2.7817
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             6.035e+01  1.177e+01   5.127 2.95e-07
## Temperature.max                         1.038e-01  2.057e-02   5.045 4.53e-07
## Cloud.Cover.Total                       1.353e-02  3.652e-03   3.706 0.000211
## Sea.Level.min                          -6.272e-02  1.148e-02  -5.463 4.67e-08
## Cloud.Cover.Medium.max:Humidity.max     1.003e-04  2.665e-05   3.764 0.000167
## Wind.Speed.max.80m:Wind.Direction.900mb 1.320e-04  2.739e-05   4.818 1.45e-06
## Month:Temperature.min                  -8.739e-03  2.770e-03  -3.154 0.001610
##
## (Intercept)                             ***
## Temperature.max                         ***
## Cloud.Cover.Total                       ***
## Sea.Level.min                           ***
## Cloud.Cover.Medium.max:Humidity.max     ***
## Wind.Speed.max.80m:Wind.Direction.900mb ***
## Month:Temperature.min                   **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1043.2  on 945  degrees of freedom
```

```
## AIC: 1057.2
##
## Number of Fisher Scoring iterations: 4
```

Nous constatons une très bonne significativité de l'ensemble des variables considérées dans le modèle. Le taux d'humidité relatif (maximal) joue donc un rôle sur les chances de précipitations. Le critère AIC est sensiblement à la hausse (1057.2 vs 1054.7).

Modèle 19

Dans ce modèle, nous considérons une combinaison d'intéractions entre les différentes variables *Cloud.Cover.Total*, *Cloud.Cover.Medium.max* et *Humidity.max*.

```
modele19=glm(pluie.demain~ Temperature.max
             +Cloud.Cover.Total*Cloud.Cover.Medium.max:Humidity.max
             +Wind.Speed.max.80m:Wind.Direction.900mb
             +Sea.Level.min
             +Month:Temperature.min,
             data=meteo_train[trainset, ],family = "binomial")
```

```
##
## Call:
## glm(formula = pluie.demain ~ Temperature.max + Cloud.Cover.Total *
##     Cloud.Cover.Medium.max:Humidity.max + Wind.Speed.max.80m:Wind.Direction.900mb +
##     Sea.Level.min + Month:Temperature.min, family = "binomial",
##     data = meteo_train[trainset, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2499  -0.9053   0.2178   0.8875   2.9076
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                             6.172e+01  1.180e+01
## Temperature.max                                         1.042e-01  2.065e-02
## Cloud.Cover.Total                                       2.559e-02  6.625e-03
## Sea.Level.min                                          -6.440e-02  1.152e-02
## Cloud.Cover.Medium.max:Humidity.max                     1.643e-04  4.006e-05
## Wind.Speed.max.80m:Wind.Direction.900mb                 1.313e-04  2.731e-05
## Month:Temperature.min                                  -8.735e-03  2.778e-03
## Cloud.Cover.Total:Cloud.Cover.Medium.max:Humidity.max -1.777e-06  8.130e-07
##                                                        z value Pr(>|z|)
## (Intercept)                                              5.231 1.69e-07 ***
## Temperature.max                                          5.046 4.51e-07 ***
## Cloud.Cover.Total                                        3.863 0.000112 ***
## Sea.Level.min                                           -5.591 2.26e-08 ***
## Cloud.Cover.Medium.max:Humidity.max                      4.103 4.08e-05 ***
## Wind.Speed.max.80m:Wind.Direction.900mb                  4.807 1.53e-06 ***
## Month:Temperature.min                                   -3.145 0.001663 **
## Cloud.Cover.Total:Cloud.Cover.Medium.max:Humidity.max  -2.186 0.028841 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1319.7  on 951  degrees of freedom
## Residual deviance: 1038.4  on 944  degrees of freedom
## AIC: 1054.4
##
## Number of Fisher Scoring iterations: 4
```

Nous constatons que parmi l'ensemble des variables considérées, la significativité de l'interaction entre les variables *Cloud.Cover.Total*, *Cloud.Cover.Medium.max* et *Humidity.max* est la plus faible. Le critère AIC est sensiblement meilleur que celui du modèle précédent. Il est néanmoins le plus faible parmi l'ensemble des modèles que nous avons manuellement ajusté.

Nous procéderons dans la section suivante à la comparaison des différents modèles concurrents. ‘

3.3. Comparaison des modèles

Dans cette section, nous procédons à une comparaison des modèles concurrents afin de sélectioner le plus adéquat pour prédire l'occurence de précipitations le lendemain.

Suite aux résultats obtenus dans la section précèdente ainsi qu'à notre compréhension et postulats quant aux phénomènes météorologiques, nous retenons les modèles suivants à des fins de comparaison.

Modèle	Valeur AIC
Modèle auto backward	1052.06728606435
Modèle 17	1054.73031586504
Modèle 18	1057.23452615045
Modèle 19	1054.42027122384

4. Validation des modèles

Dans cette partie nous évaluons la qualité d'ajustement des modèles retenus dans la section précédente. Nous en analyserons également la qualité de prédiction avant de valider le modèle qui nous servira à prédire l'occurence de précipitations le lendemain, à l'aide de la partie *valset* de l'ensemble d'entraînement.

Afin de valider un modèle de prédiction, nous définissons ci-dessous les fonctions permettant d'évaluer le taux d'erreur de prédiction, la matrice de confusion pour un seuil de classification donné ainsi que le taux de prédictibilité.

```
#Fonction calculant les erreurs de prédictions
f.err=function(pred2){
          err.pred=mean(abs(pred2 - meteo_train[!trainset, "pluie.demain"]),
                        na.rm = T)
          return(err.pred)
}
```

```
#Fonction permettant de calculer la matrice de confusion
f.confusion=function(pred2,seuil){
                confusion=table(meteo_train[!trainset, "pluie.demain"], pred2>seuil)
                return(confusion)
             }
```

```r
#Fonction utilisée pour évaluer la qualité du modèle
f.qualite=function(pred2){
          qualite=mean(pred2 == (meteo_train[!trainset, "pluie.demain"]=="1"))
          return (qualite)
}
```

Choix de la fonction de coût

Une prédiction erronée quant à l'occurence des précipitations peut entraîner différents impacts aussi bien d'un point de vue des individus que d'un point de vue économique.

La pluviométrie est un facteur non négligeable dans la prise de décision d'un individu. En effet, pour toute activité de plein air de façon évidente mais également à but économique, la météo a un impact sur la programmation des activités.

Le fait de prédire qu'il ne va pas pleuvoir alors qu'il y aura des précipitations (faux négatif) peut avoir plusieurs type de conséquences. Par exemple, en cas d'estimation météorologique favorable, les activités en plein air pourront attirer les individus. Néanmoins, si la prédiction météorologique s'avère inexacte, la programmation des individus pourra en être modifiée.

De la même manière, prédire des précipitations alors qu'il n'y en a pas (faux positif) engendre des conséquences économiques. Par exemple, les restaurateurs disposant d'une terrasse pourront décider de ne pas l'ouvrir (non allocation des ressources pour la gestion de l'espace dédié à la terrasse), engendrant ainsi une perte de chiffre d'affaire.

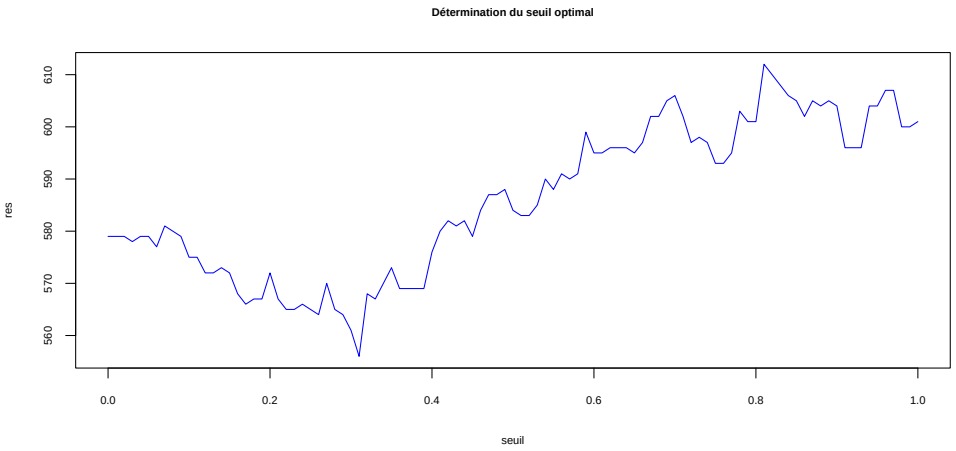
Pour ces raisons nous estimons, en termes de coûts, qu'un faux négatif a autant d'impact qu'un faux positif.

Qualité des modèles

Afin d'évaluer la qualité de chacun des quatre modèles retenus, nous procéderons à l'identification du seuil optimal, à partir duquel sera classé la prédiction d'occurence de pluie. Le taux d'erreur de prédiction ainsi que le taux de prédiction correcte seront ensuite mesurés.

Qualité du modèle de sélection automatique

Nous recherchons ci-dessous le seuil optimal du modèle de sélection automatique des variables.




```r
seuil[which.min(res)]
```

```
## [1] 0.31
```

Le seuil optimal est défini à 0.31.

Le modèle prédit l'occurence de pluie dès lors que ce seuil est atteint.

```
pred.auto2 = (pred.auto >= 0.31)
```

```
f.err(pred.auto2)
```

```
## [1] 0.2631579
```

Nous en déduisons l'erreur de prédiction du modèle de sélection automatique des variables, qui est de 26,3% au seuil 0.31.

La matrice de confusion nous renseigne sur les prédictions réussies ainsi que les prédictions erronées.

	FALSE	TRUE
0	51	53
1	7	117

Nous notons qu'au seuil optimal de 0.31, le modèle tend plus à prédire qu'il ne va pas pleuvoir demain alors que des précipitations sont observées (53 faux négatifs). Néanmoins, au regard des résultats obtenus de prédiction et des conséquences précédement évoquées, ce seuil nous semble trop faible.

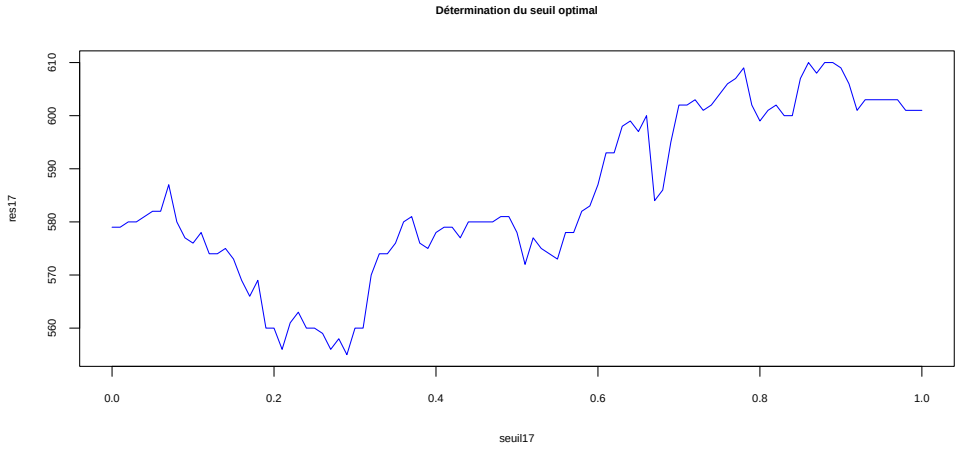
Nous procédons ainsi à une modification du seuil et analysons la qualité de prédiction avec différents autres seuils.

Seuil	Erreur de prédiction (%)	Qualité de prédiction (%)
0.31	26.32	73.68
0.41	25	75
0.42	25.44	74.56
0.43	23.68	76.32
0.44	24.12	75.88

Au seuil de 0.43, le taux d'erreur est le plus faible (23.68%) tandis que celui de la qualité de prédiction s'élève à 75.88%. Il s'agit également du taux le plus élevé. Classer une prédiction de pluie le lendemain dès lors que nous atteignons ce seuil semble raisonnable.

Qualité du modèle 17

Nous recherchons ci-dessous le seuil optimal du modèle considéré.



Le seuil optimal pour ce modèle est de 0.29.

```
seuil17[which.min(res17)]
```

```
## [1] 0.29
```

Le modèle prédit l'occurence de pluie dès lors que ce seuil est atteint.

```
pred17.2 = (pred17 >= 0.29)
```

```
#Erreur de prédiction du modèle
f.err(pred17.2)
```

```
## [1] 0.254386
```

A ce seuil, l'ajustement du modèle conduit à une prédiction, avec un taux d'erreur de 25,4%.

Pour ce seuil, nous obtenons la matrice de confusion suivante:

```
##
##      FALSE TRUE
##   0     52   52
##   1      6  118
```

Nous constatons qu'avec un seuil de 0.29, le modèle tend à prévoir plus de faux négatifs (52 occurences).

La proportion de faux négatifs étant très élevée par rapport à celle des faux positifs, nous choisissons d'analyser d'autres niveaux de seuils permettant de limiter le taux d'erreur tout en maximisant le taux de prédiction (qualité du modèle).

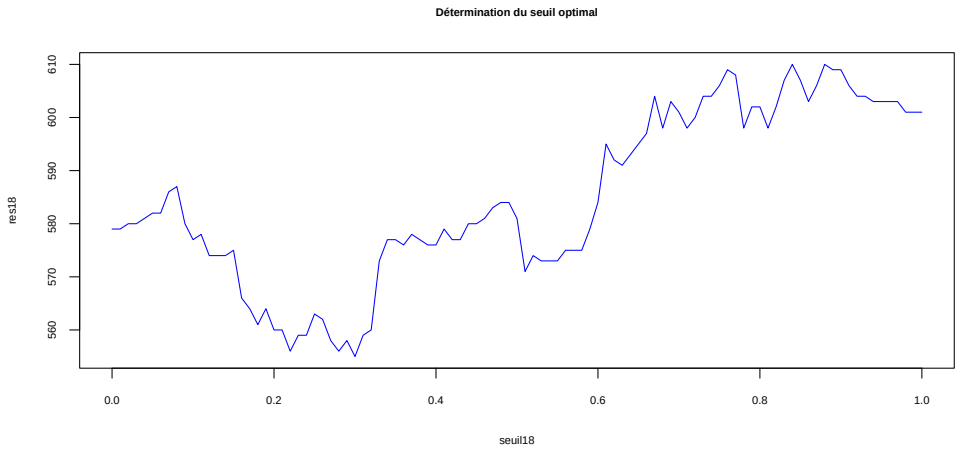
Seuil	Erreur de prédiction (%)	Qualité de prédiction (%)
0,29	25.44	74.56
0.32	24.56	75.44
0.38	23.25	76.75

Seuil	Erreur de prédiction (%)	Qualité de prédiction (%)
0.41	25.44	74.56
0.44	26.75	73.25

Nous constatons que le taux d'erreur le plus faible est de 23.25%. Le taux de bonnes prédictions le plus élevé est de 76,75%. Ces deux taux se situent au seuil de 0.38. Ces résultats nous paraissent être un bon compromis.

Prediction modèle 18

Nous identifions ci-dessous le seuil optimal du modèle considéré.



Nous constatons que le seuil optimal pour ce modèle est de :

```
seuil18[which.min(res18)]
```

```
## [1] 0.3
```

Le modèle classifie l'occurence de pluie à ce seuil.

```
pred18.2 = (pred18 >= 0.29)
```

Le taux d'erreur du modèle est présenté ci-dessous.

```
#Erreur de prédiction du modèle
f.err(pred18.2)
```

```
## [1] 0.254386
```

Au seuil optimal du modèle, la matrice de correlation est présentée ci-après.

```
##
##     FALSE TRUE
##   0    52   52
##   1     6  118
```

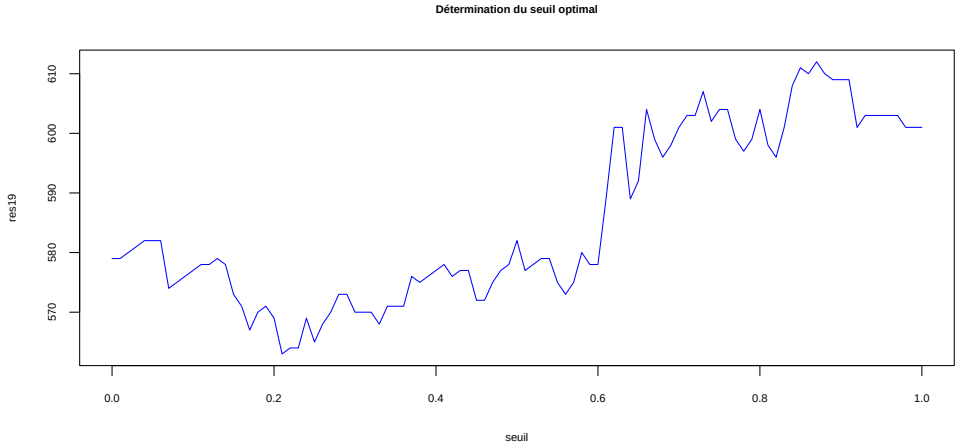
Nous analysons les résultats avec d'autres niveau de seuil

Seuil	Erreur de prédiction	Qualité de prédiction
0,29	25.44	74.56
0.3	25.44	74.56
0.32	24.56	75.44
0.38	23.68	76.32
0.45	26.75	73.25

Le taux d'erreur le plus faible est de 23.68% tandis que le taux de prédiction correct le plus élevé est de 76,32%. Ces deux taux sont constatés pour un seuil de 0.38.

Prediction mdoèle 19

Nous identifions ci-dessous le seuil optimal du modèle considéré.




```
seuil19[which.min(res19)]
```

```
## [1] 0.21
```

Nous constatons que le seuil optimal pour ce modèle est de 0.21.

Au seuil 0.21, le modèle classe l'occurence de précipitation.

```
#Classe en vrai si sup au seuil
pred19.2 = (pred19 >= 0.21)
```

Le taux d'erreur du modèle est présenté ci-dessous.

```
#Erreur de prédiction du modèle
f.err(pred19.2)
```

```
## [1] 0.3114035
```

A ce seuil la matrice de confusion est définie ci-dessous

```
##
##      FALSE TRUE
##   0     36   68
##   1      3  121
```

Nous analysons les résultats avec différents autres seuils.

```
#Classe en vrai si sup au seuil
pred19.2.2 = (pred19 >= 0.39)
pred19.2.3 = (pred19>= 0.4)
pred19.2.4 = (pred19 >= 0.41)
pred19.2.5 = (pred19 >= 0.42)
```

Seuil	Erreur de présdiction (%)	Qualité de prédiction (%)
0.21	31.14	68.86
0.39	24.56	75.44
0.4	23.68	76.32
0.41	25	75
0.42	24.56	75.44

Le taux d'erreur le plus faible est de 23,6% tandis que le taux de prédiction correcte le plus élevé est de 76,32%. Nous retenons le modèle à un seuil de 0.4.

Comparaison des modèles

Le tableau ci-dessous représente, pour chacun des quatres modèles, le meilleur taux de prédiction ainsi que le taux d'erreur de prédiction le plus faible.

Modèle	AIC	Erreur de Prédiction (%)	Seuil de prédiction	Qualité de prédiction (%)
Modèle auto	105206.73	26.32	0.42	73.68
Modèle 17	105473.03	23.25	0.38	76.75
Modèle 18	105723.45	23.68	0.38	76.32
Modèle 19	105442.03	24.56	0.4	75.44

Parmi les quatres modèles, nous retenons le modèle 17, au seuil 0.38. En effet, il présente le meilleur compromis erreur-qualité. D'une part, son taux d'erreur est le plus faible (23.25) et d'autre part sa qualité de prédiction est la plus élevée (76,75%).

5. Prediction

A l'instar de l'ensemble d'entraînement, nous avons également renommé le nom des variables explicatives dans l'ensemble de test *meteo.test* et procédons dans cette partie à la prédiction de l'occurence de précipitations pour le lendemain.

```
FALSE Les objets suivants sont masqués depuis meteo_train:
FALSE
FALSE     Cloud.Cover.High, Cloud.Cover.High.max, Cloud.Cover.High.min,
FALSE     Cloud.Cover.Low, Cloud.Cover.Low.max, Cloud.Cover.Low.min,
FALSE     Cloud.Cover.Medium, Cloud.Cover.Medium.max, Cloud.Cover.Medium.min,
```

```
FALSE     Cloud.Cover.Total, Cloud.Cover.Total.max, Cloud.Cover.Total.min,
FALSE     Day, Hour, Humidity, Humidity.max, Humidity.min, Minute, Month,
FALSE     Precipitation, Sea.Level, Sea.Level.max, Sea.Level.min,
FALSE     Shortwave.Radiation, Snowfall, Sunshine.Duration, Temperature,
FALSE     Temperature.max, Temperature.min, Wind.Direction.10m,
FALSE     Wind.Direction.80m, Wind.Direction.900mb, Wind.Gust.,
FALSE     Wind.Gust.max, Wind.Gust.min, Wind.Speed.10m, Wind.Speed.80m,
FALSE     Wind.Speed.900mb, Wind.Speed.max.10m, Wind.Speed.max.80m,
FALSE     Wind.Speed.max.900mb, Wind.Speed.min.10m, Wind.Speed.min.80m,
FALSE     Wind.Speed.min.900mb, Year
```

```
pred.finale = predict(modele17, newdata=meteo_test, type = "response")
```

```
pred.finale2 = (pred.finale >= 0.38)
```

```
##   pred.finale2
## 1         TRUE
## 2         TRUE
## 3         TRUE
## 4         TRUE
## 5         TRUE
## 6        FALSE
```

```
write.csv(prediction.meteo, file = "C:\\Users\\do\\Documents\\Cours\\EMSB\\MODULE 2\\1.Modele lineaire g
          row.names = TRUE)
```

Les prédictions réalisées sont disponibles dans le fichier nommé *Prediction pluie NU.csv*

Conclusion

Dans cette étude nous avons réalisé une analyse portant sur des différents facteurs météorologiques liés à la prévision de la pluie. Notre objectif était d'estimer la survenance de précipitations à J+1 pour la ville de Bâle.

A cette fin, après avoir observé les caractéristiques de ces facteurs, nous avons successivement scindé l'ensemble d'entrainement en deux sous-ensemble de données, identifié les colinéarités existantes entres certaines variables, défini et entraîné les modèles les plus adéquats, réalisé une comparaison à l'aide du critère AIC et enfin analysé la qualité des modèles de prévisions obtenus avec l'ensemble de validation. Cela nous a permis de valider un modèle et de réaliser la prédiction attendue.

Nous avons observé que les facteurs météorologiques considérés dans notre étude présentent tous une corrélation avec la variable d'intérêt. Cependant, aucune des variables n'influence à elle seule l'occurence de la pluie pour le lendemain. Comme on pouvait s'y attendre, la prédiction de la pluie repose sur une combinaison complexe de facteurs. Différentes interactions d'éléments crées des phénomènes météorologiques, tels que la convergence de l'air ou encore une convection ascendante, propice à la formation de nuages et par conséquent aux précipitations.

Afin d'affiner modèle de prédiction, une connaissance plus approfondie de la météorologie permettrait de mieux appréhender l'influence des interactions entres les facteurs sur la création et l'évolution des phénomènes météorologiques ainsi que leurs conséquences. La définition de paramètres liés à la localisation géographique permettrait également d'apporter plus de précision quant à la prédiction. En effet, les caractéristiques climatiques peuvent être spécifiques d'une localisation à une autre et peuvent impacter les prévisions.

Enfin, considérer l'évolution des paramètres météorologiques au fil du temps, par exemple en temps réel, ainsi que prendre en compte le facteur d'accélération du changement climatique et ainsi prédire les chances de précipitations avec plus de précision et de fiabilité.