INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

INSA ROUEN

Projet de Fin d'Études

Pluie en Australie

Théophile THIERRY

M. PORTIER

2021-2022

Table des matières

Première partie

Description des données

1 Présentation des données

Les données utilisées pour ce projet peuvent être trouvées sur le site Kaggle. Elles ont été récupérées des données du gouvernement australien, dans la partie Daily Weather Observations, et ont été complétées avec les données de la partie : Climate Data Online.

Ces données contient 10 années d'observations de la météo australienne sur 49 lieux différents entre 2007 et 2017. Une observation est constituée de (presque) toutes ces variables :

- Date : date de la mesure.
- Location : localisation de la mesure.
- MinTemp : température minimale dans les 24h jusqu'à 9h du matin (en °C).
- MaxTemp : température maximale dans les 24h jusqu'à 9h du matin (en °C).
- Rainfall : précipitation dans les 24h jusqu'à 9h du matin (en mm).
- Evaporation : bac d'évaporation de classe A dans les 24h jusqu'à 9h du matin (en mm).
- Sunshine : ensoleillement en heure dans les 24h jusqu'à minuit.
- WindGustDir : direction de la plus forte rafale dans les 24 heures jusqu'à minuit (16 points cardinaux/intercardinaux).
- WindGustSpeed : vitesse de la plus forte rafale dans les 24 heures jusqu'à minuit (16 points cardinaux/intercardinaux).
- WindDir9am : direction du vent à 9h du matin.
- WindDir3pm : idem à 15h.
- WindSpeed9am : vitesse du vent à 9h.
- WindSpeed3pm : idem à 15h.
- Humidity9am : taux d'humidité relative à 9h.
- Humidity3pm : idem à 15h.
- Pressure9am : pression atmosphérique réduite au niveau moyen de la mer à 9h.
- Pressure3pm : idem à 15h.
- Cloud9am : fraction du ciel couverte par un nuage à 9h (en huitième).
- Cloud3pm : idem à 15h.
- Temp9am : température à 9h (en °C).
- Temp3pm : idem à 15h.
- RainToday : s'il a plu le jour même.
- RainTomorrow : s'il a plu le lendemain.

Jetons tout d'abord un coup d'oeil aux variables numériques de nos données.

1.1 Variables numériques

Comme nous pouvons le constater dans la Table 1, pour certaines d'entre elles, il manque beaucoup d'observations (voir la ligne *NA's*). Nous allons devoir remédier à cela dans les futures parties, et principalement sur les variables *Sunshine*, *Evaporation*, *Cloud9am* et *Cloud3pm*, dont nous avons moins de 60% des observations.

Nous pouvons de plus noter certaines choses : comme on pouvait s'y attendre, il fait en général moins froid à 15h qu'à 9h, et il fait aussi moins humide.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	Std.
MinTemp	-8.50	7.60	12.00	12.19	16.90	33.90	1485.00	6.40
MaxTemp	-4.80	17.90	22.60	23.22	28.20	48.10	1261.00	7.12
Rainfall	0.00	0.00	0.00	2.36	0.80	371.00	3261.00	8.48
Evaporation	0.00	2.60	4.80	5.47	7.40	145.00	62790.00	4.19
Sunshine	0.00	4.80	8.40	7.61	10.60	14.50	69835.00	3.79
WindGustSpeed	6.00	31.00	39.00	40.04	48.00	135.00	10263.00	13.61
WindSpeed9am	0.00	7.00	13.00	14.04	19.00	130.00	1767.00	8.92
WindSpeed3pm	0.00	13.00	19.00	18.66	24.00	87.00	3062.00	8.81
Humidity9am	0.00	57.00	70.00	68.88	83.00	100.00	2654.00	19.03
Humidity3pm	0.00	37.00	52.00	51.54	66.00	100.00	4507.00	20.80
Pressure9am	980.50	1012.90	1017.60	1017.65	1022.40	1041.00	15065.00	7.11
Pressure3pm	977.10	1010.40	1015.20	1015.26	1020.00	1039.60	15028.00	7.04
Cloud9am	0.00	1.00	5.00	4.45	7.00	9.00	55888.00	2.89
Cloud3pm	0.00	2.00	5.00	4.51	7.00	9.00	59358.00	2.72
Temp9am	-7.20	12.30	16.70	16.99	21.60	40.20	1767.00	6.49
Temp3pm	-5.40	16.60	21.10	21.68	26.40	46.70	3609.00	6.94

TABLE 1 – Résumé des variables numériques

1.2 Variables factorielles

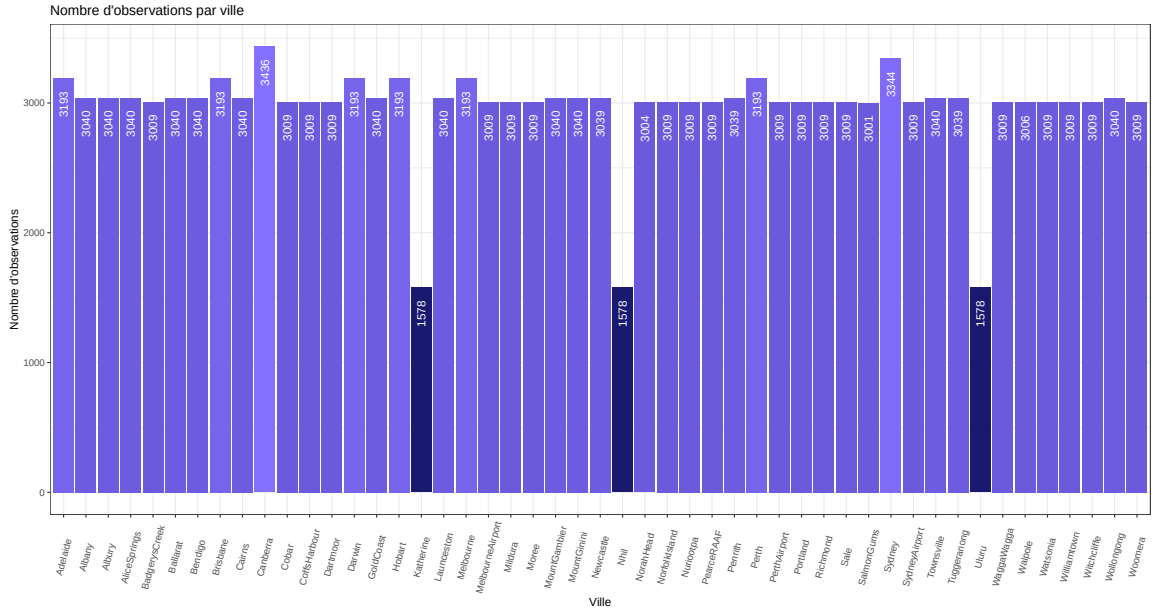

FIGURE 1 – Nombre d'observations pour chaque ville

Regardons désormais les variables avec des facteurs. Les dates, tout d'abord, vont du 2007-11-01 au 2017-06-25, ce qui représente 3524 jours. Si l'on regarde la distribution du nombre d'observations par ville, on

devrait donc voir que chacune d'entre elles en ont 3524 (voir Figure 1). On voit déjà qu'il manque certaines dates d'observations pour la plupart des villes, et pour 3 d'entre elles : Katherine, Nhil et Uluru, nous avons moins de la moitié. Ceci est dû au fait que les observations démarrent à ces endroits en 2013.

Penchons-nous désormais sur les variables de direction du vent en regardant la distribution de leur facteurs dans les Tables 2, 3 et 4.

Les 16 niveaux utilisés pour ces variables sont tous représentés avec à peu près la même distribution. Nous souhaitons changer ces facteurs en valeurs numériques, pour cela, nous allons remplacer chaque direction par sa valeur en degrés : utiliser l'angle pour la directions nous servira peut-être lorsque nous utiliserons des modèles de prédiction comme des arbres CART. En effet, baisser le nombre de facteurs que nous avons nous permettra sûrement de réduire le nombre de feuilles que nous aurions pu avoir. Nous utiliserons les valeurs indiquée dans la Table 5.

	E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW
Degré	0.0	22.5	45.0	67.5	90.0	112.5	135.0	157.5	180.0	202.5	225.0	247.5	270.0	292.5	315.0	337.5

TABLE 5 – Les 16 points cardinaux en degrés

On remarque une fois de plus qu'une partie de ses observations sons manquantes (plus de 7% pour *WindGustDir* et *WindDir9am* et un peu moins de 3% pour *WindDir3pm*)

Enfin, nous avons les deux dernières variables booléennes concernant la pluie (Tables 6 et 7). On voit que notre base de données est déséquilibrée : la variable que nous voulons prédire étant *RainTomorrow*, nous voulons avoir une équilibre entre les observations *Yes* et les observations *No*. Pour veiller à ceci, nous utiliserons des méthodes de resampling telle que SMOTE.

On remarque de plus que nous avons 2.25% des variables manquantes pour ces deux variables. Ceci s'explique par le fait qu'il manque 2.25% des observations de la variable *Rainfall*, sur laquelle sont basées ces deux variables.

Enfin, nous avons la variables *Location* qui contient tous les lieux d'observations de la base de données. Nous nous pencherons sur celle-ci dans la section Cartographie.

Avant de nous lancer plus loin, nous allons tout d'abord modifier la base de données pour la rendre utilisable. Pour cela, nous devons nous occuper des variables qui contiennent des facteurs et les changer en numériques, comme nous avons fait pour les variables de direciton du vent. Nous allons ensuite remplacer la variable lieu avec une variable de longitude et de latitude, pour des raisons que nous expliquerons dans la section suivante. Enfin, nous allons rajouter une variable correspondant aux climats de chaque lieu (on prendra pour cela une carte des climats et on pourra rentrer à la main chaque climat de chaque lieu). Pour ce qui est de la variable date, nous la remplacerons par une variable saison à seulement 4 niveaux. Ce choix sera expliqué dans une partie sur le climat et sur les périodicités. Au final, nous pourrons nous occuper du traitement des observations manquantes.

2 Cartographie

Notre base de donnée comprend donc une variable *Location*, qui est une variable qualitative avec le nom du lieu de mesure. Nous en avons 49 différentes, et afin de visualiser un peu mieux ces différents points d'observation, nous voulons les afficher sur une carte.

Pour cela, nous allons utiliser le paquet R "rnaturalearth", qui nous offre un moyen simple de dessiner nos propres cartes en utilisant le standard WGS84 (World Geodetic System).

	E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NA's
Compte	9181	8104	7372	9313	7133	6548	6620	8122	9168	9418	9216	8736	8967	9915	8252	9069	10326
%	6.31	5.57	5.07	6.40	4.90	4.50	4.55	5.58	6.30	6.47	6.34	6.01	6.16	6.82	5.67	6.23	7.10

Table 2 – Variable WindGustDir

	E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NA's
Compte	9176	7836	7630	11758	7671	8129	7980	8749	8659	9287	9112	7587	8423	8459	7414	7024	10566
%	6.31	5.39	5.25	8.08	5.27	5.59	5.49	6.01	5.95	6.38	6.26	5.22	5.79	5.82	5.10	4.83	7.26

Table 3 – Variable WindDir9am

	E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NA's
Compte	8472	7857	8505	8890	8263	6590	7870	8610	9926	10838	9399	8156	9354	10110	8874	9518	4228
%	5.82	5.40	5.85	6.11	5.68	4.53	5.41	5.92	6.82	7.45	6.46	5.61	6.43	6.95	6.10	6.54	2.91

Table 4 – Variable WindDir3pm

	No	Yes	NA's
Compte	110319	31880	3261
%	75.84	21.92	2.24

Table 6 – Variable RainToday

	No	Yes	NA's
Compte	110316	31877	3267
%	75.84	21.91	2.25

Table 7 – Variable RainTomorrow

2.1 Un standard de localisation et une projection

Afin de localiser avec précision un point sur Terre, nous avons besoin d'un standard de localisation. Un standard est basé sur un système de coordonnées géodésique. Il peut utiliser notamment un système de coordonnées en Longitude et Latitude.

2.1.1 Latitude et Longitude

Afin d'avoir une coordonnée pour n'importe quel point sur Terre, nous utilisons des coordonnées de Longitude et de Latitude. Ce sont des valeurs exprimées en degré à partir d'un degré 0 de référence.

La Terre ne peut être représentée comme une sphère car cela rendrait les coordonnées trop imprécises par rapport à la réalité. Elle est de plus arrondie aux pôles et c'est pour ces raisons que nous représentons la Terre par un éllipsoïde.

La Longitude est une coordonnée géographique représentée par une valeur angulaire, expression du positionnement est-ouest d'un point sur Terre [Wik21b]. Tous les points étant situés sur une courbure de l'élipsoïde reliant les pôle Nord et Sud et traversant l'équateur perpendiculairement ont la même longitude. Une courbure de référence, appelé "méridien" est choisi arbitrairement (le méridien de Greenwich) comme degré 0. Les valeurs de Longitude s'étendent de -180° vers l'ouest à 180° à l'est par rapport à ce méridien.

La Latitude est une coordonnée similaire mais qui à pour plan de référence l'équateur. Tous les points sur Terre ayant une même latitude forment un cercle dont le plan est parallèle à celui de l'équateur [Wik21a].

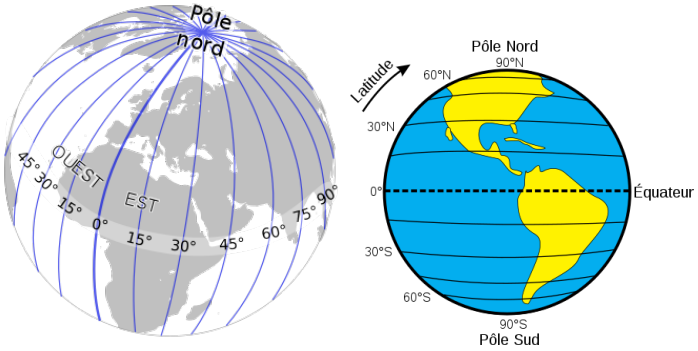

FIGURE 2 – Illustration du système de coordonnées de Longitude et Latitude

Lorsque l'on combine ce système de coordonnées et une représentation de la Terre en ellipsoïde (au travers de mesures précises des dimensions de la planète), on obtient un système géodésique.

2.1.2 Le WSG84

Le World Geodetic System 84 (WGS84) est un système géodésique, et nous pouvons l'utiliser pour nos cartes grâce au paquet "rnaturalearth". Il est notamment utilisé par le système GPS (Global Positioning System). Ce standard à été établi et est maintenu par le National Geospatial Intelligence Agency (NGA) des Etats-Unis [Wik22] depuis 1984. Il est basé sur un ellipsoïde de référence raffiné avec le temps pour représenter au mieux la Terre, ainsi que le système de coordonnées en Longitude et Latitude.

Nous avons maintenant un moyen de localiser précisément un point sur Terre grâce à deux valeurs numériques. Pour pouvoir les afficher sur une carte, il nous faut cependant une projection.

2.1.3 Projections

La projection cartographique est "un ensemble de techniques permettant de représenter la surface de la Terre dans son ensemble ou en partie sur la surface plane d'une carte" [Wik21c]. La Terre étant sphérique, afin de l'afficher sur une carte plane, il faut la projeter. Il existe différent types de projections, certaines permettent de conserver localement les surfaces, d'autres les angles ou encore les distances sur les méridiens.

Notre paquet utilise de base une projection dite géographique : elle consiste simplement à prendre les valeurs de latitude et de longitude et des les utiliser comme si elles étaient les coordonnées X et Y (respectivement) d'un repère en deux dimensions. Cette "projection" peut avoir des résultats différents en fonction du système géodésique utilisé.

Le plus gros inconvénient de cette pratique est la distorsion des surfaces lorsque l'on s'éloigne de l'équateur. Cependant, cela est suffisant dans notre cas, où nous voulons avoir seulement une idée globale de la position des lieux observés des uns par rapport aux autres. De plus, comme nous ne prévoyons pas de mesurer précisément la distance entre deux points, ce système de "projection" géographique est le plus pratique.

2.2 Affichage sur une carte

Nous avons désormais tous les éléments pour placer les lieux sur une carte. Le paquet "rnaturalearth" nous permet donc d'avoir une liste de polygone de pays. Le paquet "ozmaps" nous permet d'avoir les polygones des états australiens. Pour les lieux, nous récupérons les latitudes et longitudes manuellement grâce à n'importe quelle base que nous pouvons trouver sur internet et nous les rajoutons à chaque observation en ajoutant deux colonnes. Au final nous pouvons afficher notre carte grâce à ggplot2 :

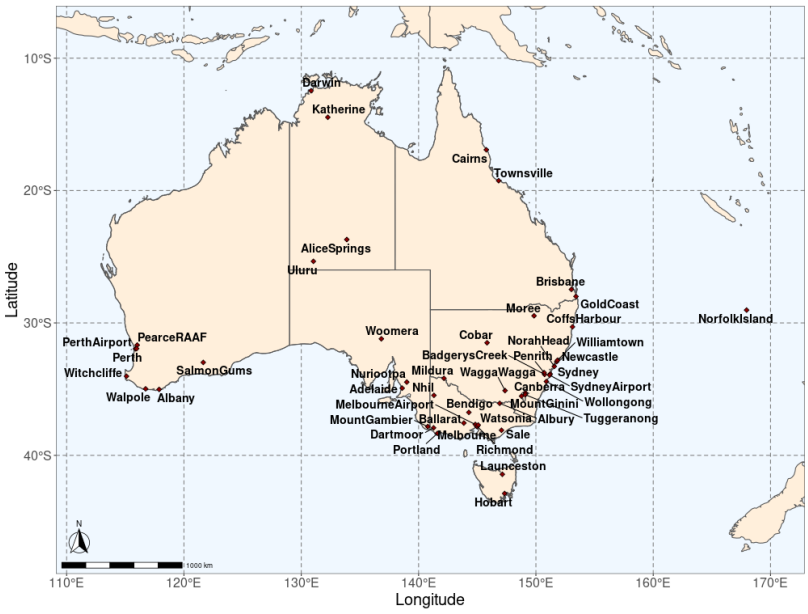

FIGURE 3 – Carte de l'Australie et villes dont la météo est observée

Nous pouvons desormais utiliser les colonnes de longitude et latitude à la place de la colonne localisation.

3 Etude des climats

Maintenant que nous pouvons afficher les lieux sur une carte, nous pouvons déterminer à quelle zone climatique appartient chaque point.

Comme on peut le voir sur la carte précédente, la plupart des observations ont lieu dans le sud-est du pays, où la concentration d'habitants et de ville est la plus grande. Cette zone correspond à un climat tempéré pour les villes les plus au sud et subtropical pour les villes plus au nord comme Brisbane. Au nord du pays nous avons les villes sur les littoraux dans une zone plus tropicale, et enfin au sud-ouest nous avons d'autre villes subtropicales. Plus à l'intérieur des terres, où il le climat est désertique, nous avons les observations de Uluru, Alice Springs et Woomera. Enfin, nous avons aussi les données de villes sur l'île de Tasmanie et l'île Norfolk.

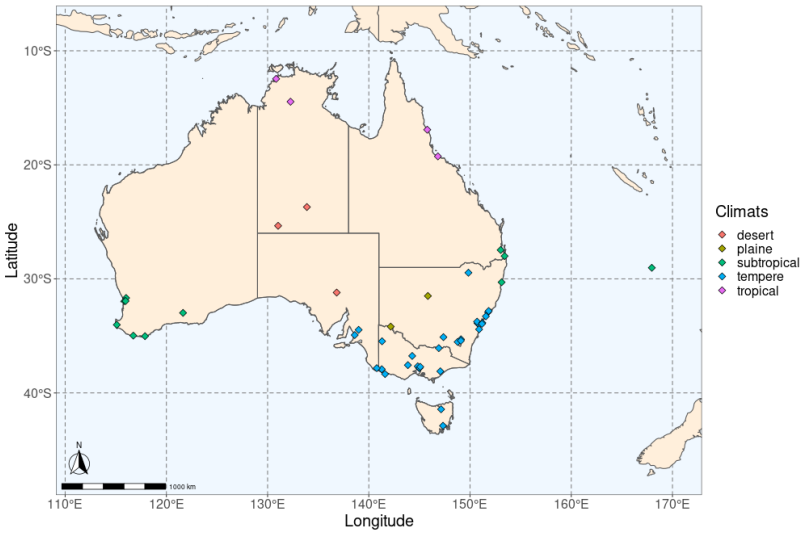

FIGURE 4 – Carte de l'Australie et climat des villes

Nous avons donc les observations de 4 lieux tropicaux, 3 lieux désertiques, 2 lieux dans les plaines (le climat de transition entre désertique et tempéré, en quelques sortes), 11 lieux subtropicaux et enfin 29 lieux tempérés. Certaines données sont liées à des villes mais d'autres à des aéroport ou encore des lieux touristiques.

3.1 Particularités des climats et périodicité

Les données étant étalées sur 10 ans, on peut trouver une périodicité dans les mesures à l'année. Nous pouvons alors, pour chaque ville, faire un graphique comprenant les moyennes des température maximales, minimales et moyennes de chaque jour sur 10 ans. Et faire de même pour les précipitations.

Nous pouvons afficher ces données en fonction des saisons. Nous pourrons ainsi remplacer la variable date par une variable avec uniquement 4 niveaux différents comme expliqué précédemment, à savoir les saisons. Nous prendrons comme saisons :

— L'été, de décembre à février
— L'automne, de mars à mai
— L'hiver, de juin à août
— Le printemps, de septembre à novembre

Nous obtenons les graphiques de la Figure 5

On remarque tout de suite que les températures les plus élevées sont aux alentours de décembre / janvier ; l'Australie étant dans l'hémisphère Sud, il s'agit de l'été.

On remarque ensuite quelques particularité dues aux climats. Dans les régions tempérée et subtropicale tout d'abord, nous avons des températures qui évoluent entre en dessous de 10 degrés et environ 30 degrés, avec en hiver (mai, juin, juillet) plus de pluie que sur le reste de l'année.

Du côté des régions dans les plaines, il pleut moins tout au long de l'année et nous n'observons pas de période de pluie comme pour les deux premières régions. Les températures sont en revanche à peu près les même, voire plus chaudes pendant l'été. Lorsque l'on se penche sur les régions désertiques, les températures sont encore plus hautes et les précipitations sont encore moins importantes, avec seulement quelques millimètres tout au long de l'année.

A l'opposé, dans les régions tropicales, la température tout au long de l'année évolue moins et reste plus proche de 30 degrés tout au long de l'année (avec une légère baisse en hiver). Dans ces régions, il pleut énormément pendant l'hiver et quasiment pas pendant l'été.

Le choix de ne garder que les saisons et pas les dates se justifie par le fait que si nous voulons prédire s'il pleut le lendemain, nous n'avons pas besoin de savoir précisemment quel jour nous sommes, voire quel mois. On remarque sur les graphiques des différences notables entre les saisons, et celle-ci suffiront sûrement pour nous aider à prédire ce que nous voulons. De plus, nous nous débarassons d'une variable avec beaucoup de facteurs différents, ce qui nous sera bénéfique lors de la mise en place de modèle de prédictions.

Nous nous retrouvons au final avec les variables suivantes :
— MinTemp, MaxTemp, Temp9am, Temp3pm
— Rainfall, RainToday, RainTomorrow, Evaporation
— Sunshine, Cloud9am, Cloud3pm
— WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm
— Humidity9am, Humidity3pm
— Pressure9am, Pressure3pm
— Season, Climate
— Latitude, Longitude

Qui sont toutes des variables numériques (0 ou 1 pour les variables booléennes RainToday et RainTomorrow).

4 Complétion des données

4.1 Pourquoi compléter ?

Afin de se rendre compte de la distribution des valeurs manquantes, on affiche ce qu'on appelle la missingness map de nos données (Figure 6).

Afin de faire des prédictions sur nos données, nous avons besoin de nous débarasser des observations avec des valeurs *NA*. Pour cela, on utilise la commande `na.omit`. On se retrouve avec 56420 observations sur les 145460 de base, ce qui est très peu. De plus, la plupart des lieux ne sont plus représentés comme nous l'indique la Figure 7.

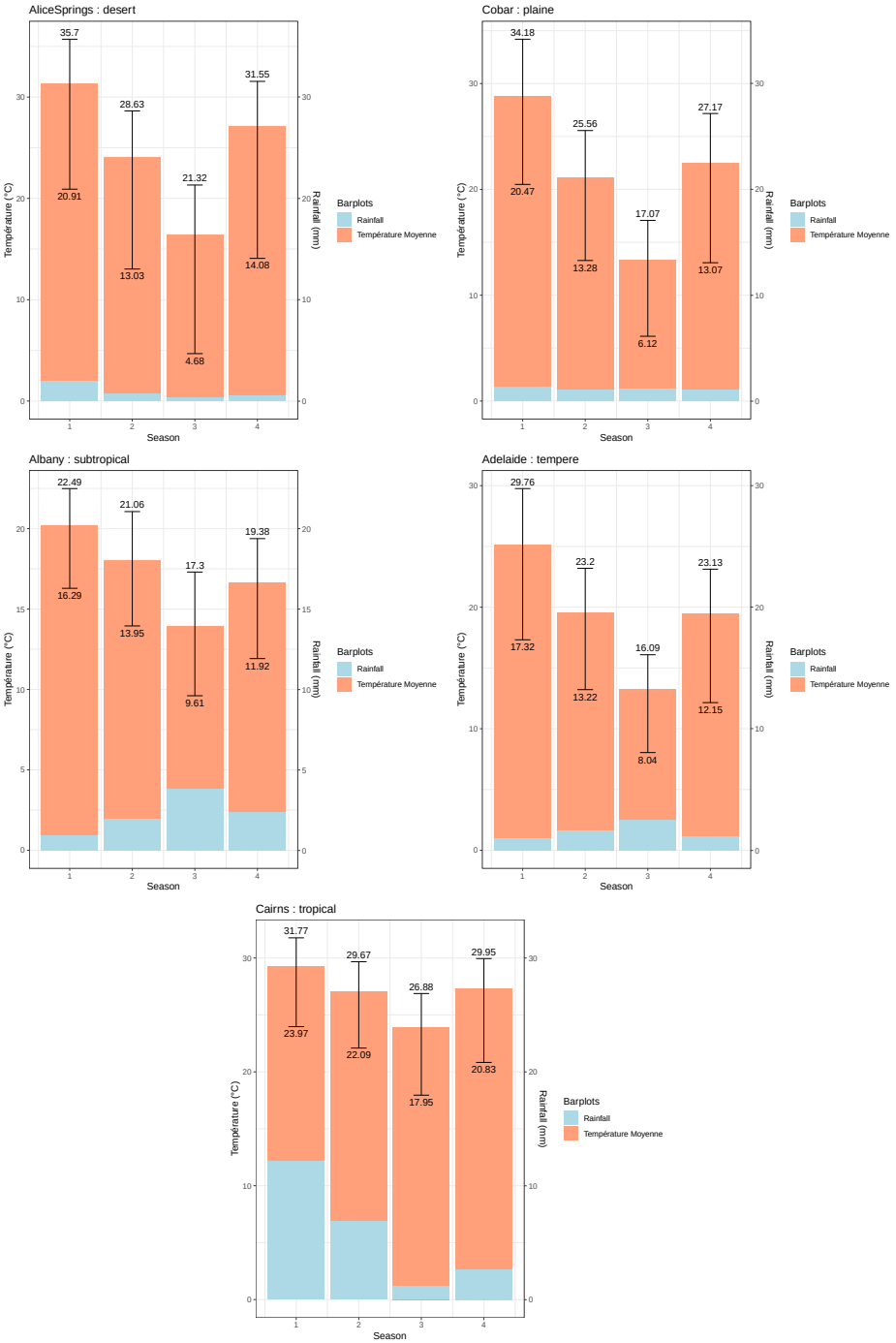

FIGURE 5 – Température Minimale et Maximale ainsi que pluviométrie au cours d'une année pour certaines des villes des données (une par climat)

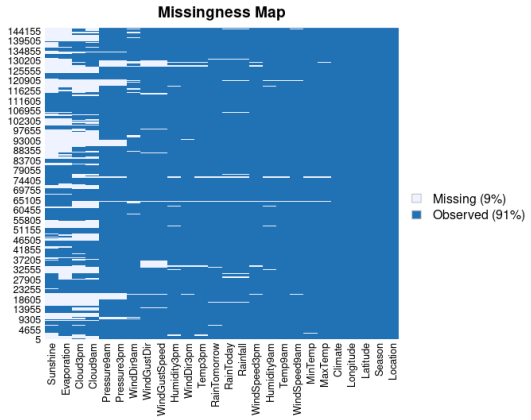

FIGURE 6 – Missingness Map des données avant la complétion.

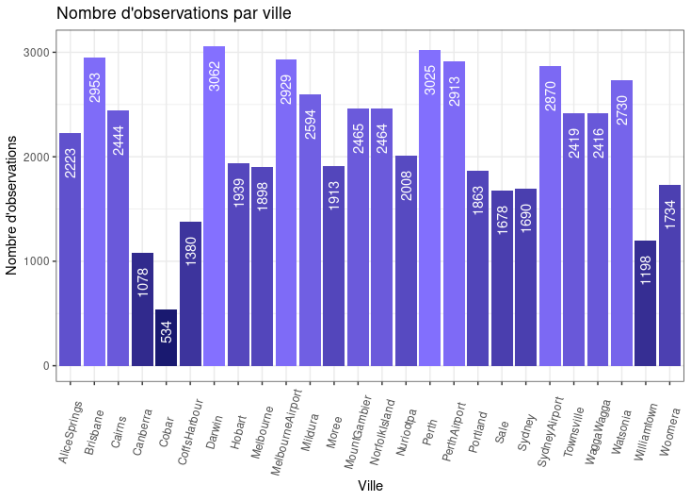

FIGURE 7 – Distribution du nombre d'observations après un *na.omit*.

Nous allons donc chercher un moyen de compléter les valeurs *NA*, et de le faire de façon à ne pas avoir trop de données redondantes et de garder une cohérence vis-à-vis des climats : nous allons copier les données d'un lieu à l'autre d'une certaine façon.

4.2 Comment compléter ?

Pour compléter nos données, nous allons regarder pour chaque variable quels sont les lieux qui ont besoin de complétion, disons ceux qui ont plus de 20% de *NA* pour cette variable, et quels sont ceux avec lesquels nous pouvons compléter : les autres qui ont plus de 2500 observations et qui ont au moins 80% de leurs observations complètes. Nous excluons ainsi les lieux dont nous n'avons pas les observations sur les dix années : lorsque nous complétons nos données, nous voulons que les dates coincident pour ne pas perdre en cohérence, ainsi nous voulons compléter les données avec les lieux pour lesquels nous avons des observation sur la période maximale d'observation de notre base de données, soit plus de 2500 jours. Le

seuil de 80% des données est choisi pour avoir un maximum d'observations sans *NA* et pour être sûr d'avoir une ville depuis laquelle copier les données.

Lorsque nous copions les données d'un lieu à un autre, nous nous soucions donc de la date d'observation. Cependant, cela ne suffira pas. Nous avons vu en effet que les différents lieux observés appartiennet à des zones climatiques très différentes. Pour chaque variable, nous allons donc associer nos deux listes de lieux (ceux à compléter et ceux avec lesquels compléter) en cherchant les lieux les plus proches de la même zone climatique.

Au final, nous nous retrouvons avec une liste de couples pour chaque variable.

Avec cette méthode, nous allons copier jusqu'à 5 fois maximum les données d'un lieu pour un autre lieu, et nous le ferons de manière "intelligente", sans perte de cohérence par rapport aux climats des lieux observés. Nous pouvons afficher quelles données de quelles villes vont compléter quelle autre villes sur des cartes comme celle de la Figure 8. On peut voir que pour les variables où il y avait le plus de *NA*, beaucoup de lieux sont complétés. Dans d'autre cas moins extrêmes, nous n'avons copié les données que d'un lieu à un autre.

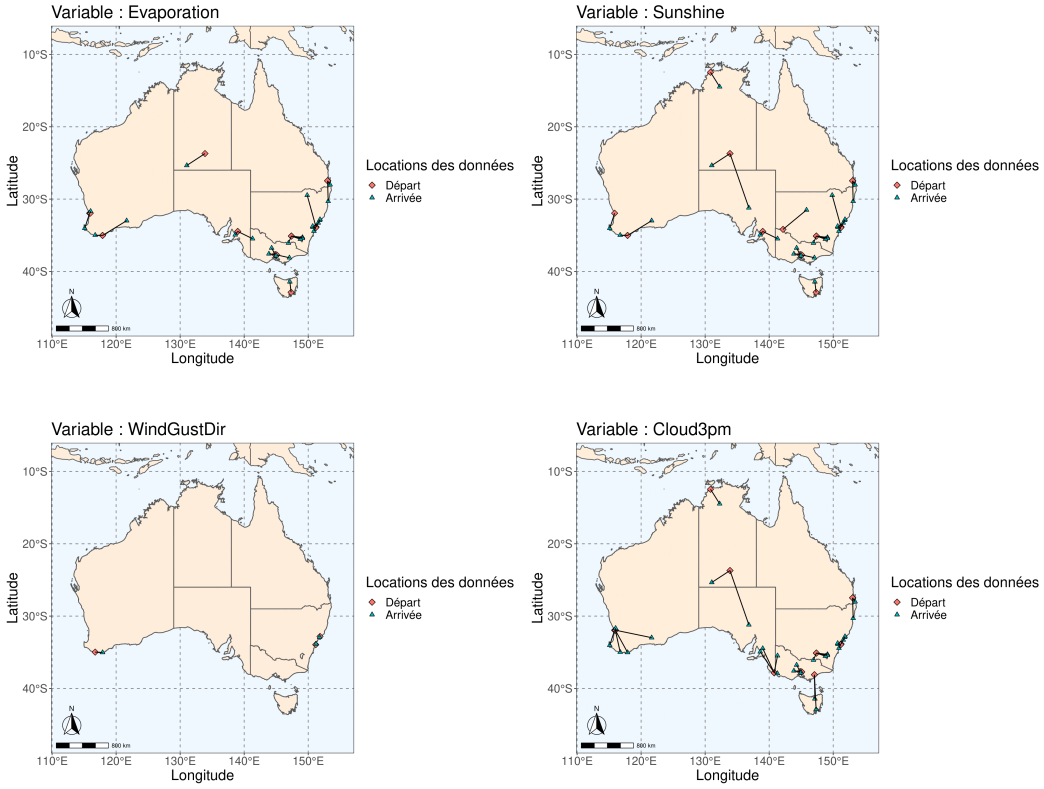

FIGURE 8 – Chemin des observations copiées (ville de départ et ville(s) d'arrivée(s)) pour certaines des variables complétées.

4.3 Après complétion

Au final, le *na.omit* nous donne une base de données avec 105546 observations. On peut réafficher la missingness map et la distribution des observations par lieu (respectivement Figures 9 et 10)

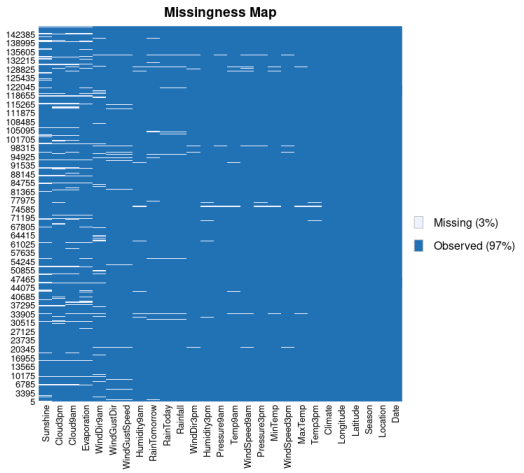

FIGURE 9 – Missingness Map des données complétées.

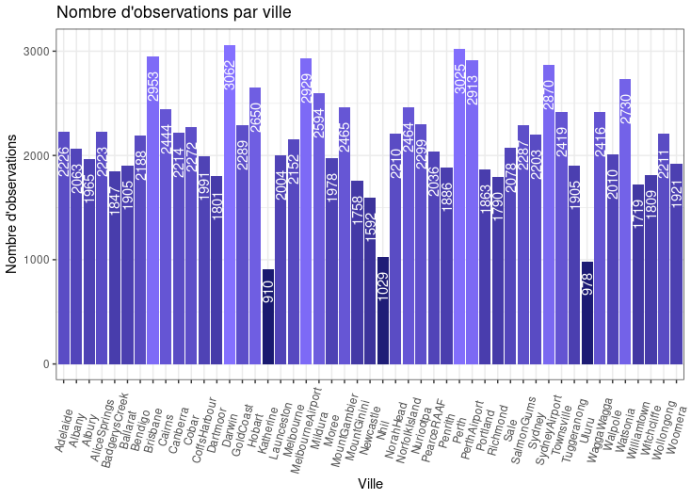

FIGURE 10 – Distribution des observations par villes dans notre base de données finale.

On remarque que certaines plages de dates n'ont pas été complétées pour certaines variables, il peut y avoir deux raisons à cela :

— Le lieu pour cette variable n'a pas été considéré comme à compléter, malgré quelques valeurs *NA* ;

— Le lieu qui a servi pour la complétion certaines dates en moins que celles du lieu à compléter.

Le nombre d'observations de notre base de données finale reste cependant satisfaisante et tous les lieux y sont représentés.

5 Relations entre les variables

Dans cette partie nous allons chercher les relations entre les variables.

5.1 Corrélations

Commençons tout d'abord par nous renseigner sur les corrélations entre les variables (Figure 11).

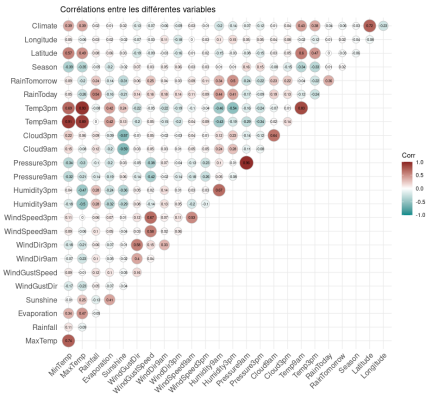

FIGURE 11 – Corrélations des variables deux-à-deux.

Deuxième partie
Prédiction

Maintenant que notre base de données est prête que nous la connaissons plus en détail, nous pouvons commencer à créer nos modèles de prédiction. Commençons tout d'abord par faire une analyse en composantes principales, pour avoir une meilleure idée de la distribution des observations.

6 ACP

Rappelons tout d'abord la distribution des observations où il pleut le lendemain et où il ne pleut pas :

RainTomorrow	Compte	%
0	82367	78.04
1	23179	21.96

TABLE 8 – Distribution des valeurs de la variable RainTomorrow dans nos données finales.

Et lorsque l'on affiche une analyse en composantes principales de ces observations, et en les colorant en fonction de leur valeur de RainTomorrow, on obtient le graphique de la Figure **??**.

Références

[Wik21a] Wikipédia. *Latitude — Wikipédia, l'encyclopédie libre.* [En ligne ; Page disponible le 29-décembre-2021]. 2021. url : %5Curl%7Bhttp://fr.wikipedia.org/w/index.php?title=Latitude&oldid=189341688%7D.

[Wik21b] Wikipédia. *Longitude — Wikipédia, l'encyclopédie libre.* [En ligne ; Page disponible le 6-décembre-2021]. 2021. url : %5Curl%7Bhttp://fr.wikipedia.org/w/index.php?title=Longitude&oldid=188614923%7D.

[Wik21c] Wikipédia. *Système de coordonnées (cartographie) — Wikipédia, l'encyclopédie libre.* [En ligne ; Page disponible le 9-avril-2021]. 2021. url : https://fr.wikipedia.org/w/index.php?title=Syst%C3%A8me_de_coordonn%C3%A9es_(cartographie).

[Wik22] Wikipedia contributors. *World Geodetic System — Wikipedia, The Free Encyclopedia.* https://en.wikipedia.org/w/index.php?title=World_Geodetic_System&oldid=1065796786. [Online ; accessed 15-January-2022]. 2022.