

INSA ROUEN

PROJET DE FIN D'ÉTUDES

Pluie en Australie

Théophile THIERRY

M. PORTIER
2021-2022

Table des matières

I	Description des données	2
1	Présentation des données	2
2	Cartographie	5
2.1	Un standard de localisation et une projection	5
2.1.1	Latitude et Longitude	5
2.1.2	Le WSG84	6
2.1.3	Projections	6
2.2	Affichage sur une carte	7
3	Etude des climats	7
3.1	Particularités des climats et périodicité	8
4	Complétion des données	10

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 contiennent 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 mais 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.

et pour 3 d'entre elles : Katherine, Nhil et Uluru, nous avons moins de la moitié. En regardant de plus près, ceci est dû au fait que les observations démarrent à ces endroits en 2013.

Penchons-nous désormais sur les variables de direction du vent :

E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NA's
9181	8104	7372	9313	7133	6548	6620	8122	9168	9418	9216	8736	8967	9915	8252	9069	10326

TABLE 1 – Variable WindGustDir

E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NA's
9176	7836	7630	11758	7671	8129	7980	8749	8659	9287	9112	7587	8423	8459	7414	7024	10566

TABLE 2 – Variable WindDir9am

E	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NA's
8472	7857	8505	8890	8263	6590	7870	8610	9926	10838	9399	8156	9354	10110	8874	9518	4228

TABLE 3 – Variable WindDir3pm

Les 16 niveaux utilisés pour ces variables sont tous représentés, on peut donc considérer cette variable comme un entier allant de 1 à 16. Avec 1 la direction Est et 16 la direction Ouest-Sud-Ouest (dans l'ordre alphabétique). Ici aussi, on remarque qu'il manque plus de 10 000 observations pour les deux premières variables et 4 000 pour la dernière.

Enfin, nous avons les deux dernières variables :

No	Yes	NA's
110319	31880	3261

TABLE 4 – Variable RainToday

No	Yes	NA's
110316	31877	3267

TABLE 5 – Variable RainTomorrow

Maintenant que nous connaissons nos variables, nous allons pouvoir les "simplifier" afin de les rendre plus facilement utilisables.

Dans un premier temps, nous allons transformer les valeurs des points cardinaux des directions de vent avec une valeur de 1 à 16 pour chaque niveau. Au vu de la distribution de ces 16 variables différentes, nous devons toutes les garder.

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. Ainsi, nous utiliserons la localisation géographique des lieux à la place de leur nom.

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.

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 "rnatuarearth", qui nous offre un moyen simple de dessiner nos propres cartes en utilisant le standard WGS84 (World Geodetic System).

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 ellipsoïde.

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'ellipsoïde 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 a 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].

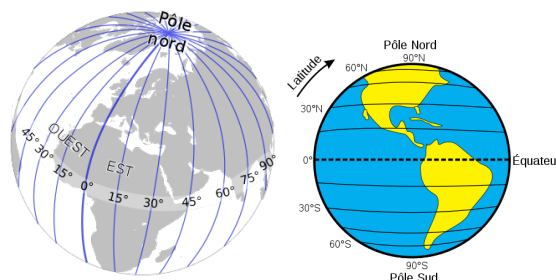


FIGURE 2 – Illustration du système de coordonnées de Longitude et Latitude

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 WGS84

Le World Geodetic System 84 (WGS84) est un système géodésique, et nous pouvons l'utiliser pour nos cartes grâce au paquet "rnatrualearth". Il est notamment utilisé par le système GPS (Global Positioning System). Ce standard a 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 différents 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 "rnatualearth" 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 désormais 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'autres 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.

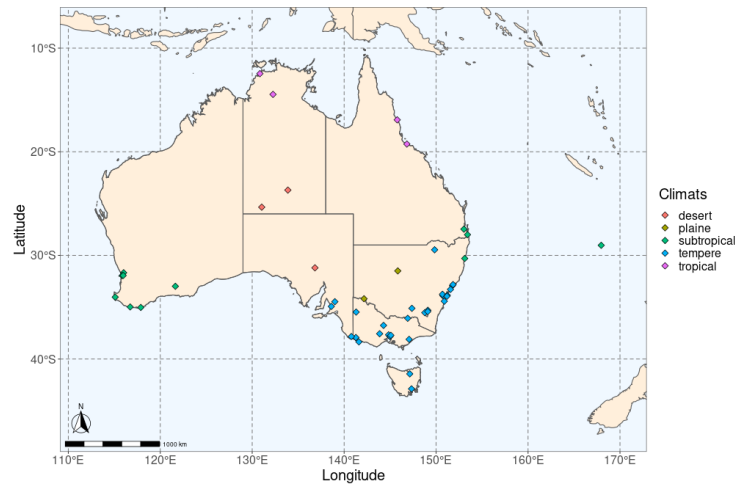


FIGURE 4 – Carte de l'Australie et climat des villes

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 pourrions 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 des graphiques comme suit :

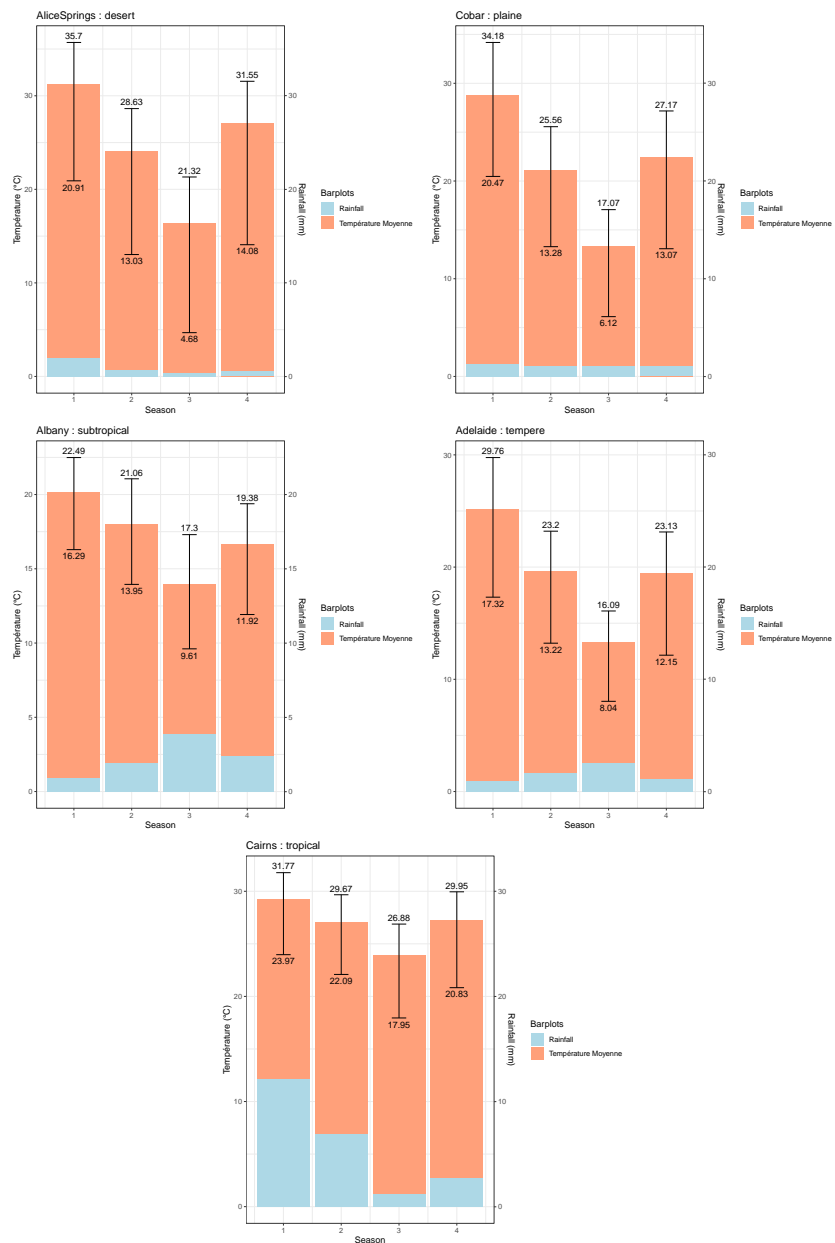


FIGURE 5 – Température Minimale et Maximale ainsi que pluviométrie au cours d’une année pour certaines des villes des données (une par climat)

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 précisément 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 débarrassons 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
- Rainfall
- Evaporation
- Sunshine
- WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm
- Humidity9am, Humidity3pm
- Pressure9am, Pressure3pm
- Cloud9am, Cloud3pm
- Temp9am, Temp3pm
- RainToday
- RainTomorrow
- Season
- Latitude, Longitude
- Climate

Qui sont toutes des variables quantitative ou des variables qualitatives avec moins de 16 niveaux (numérotés en entier).

4 Complétion des données

Si l’on regarde les valeurs manquantes, on obtient la distribution suivante :

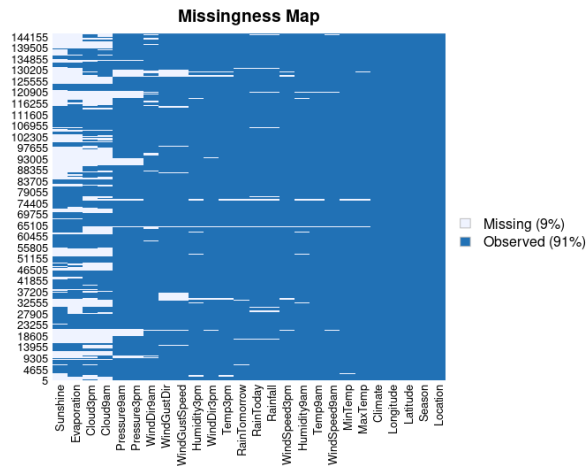


FIGURE 6 – Missingness Map des données

Afin de faire des prédictions sur nos données, nous avons besoin de nous débarrasser 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 :

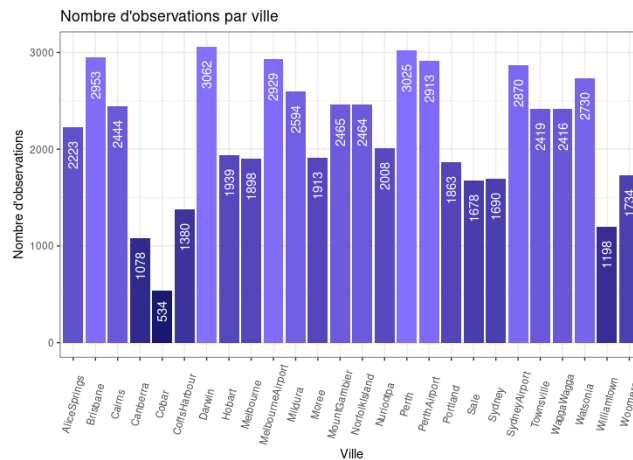


FIGURE 7 – Distribution des données après un `na.omit`

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'une ville à une autre.

Pour cela, nous allons regarder pour chaque variable quels sont les lieux qui ont besoin de complétion, disons les villes qui ont plus de 15% de *NA* pour cette variable, et quels sont ceux avec lesquels nous pouvons compléter : les autres qui ont plus de

2500 observations. En effet, comme nous l'avons vu, certaines villes n'ont d'observations que sur 1500 jours. Lorsque nous complétons nos données, nous voulons que les dates coïncident pour ne pas perdre en cohérence, ainsi nous voulons compléter les données avec les lieux pour lesquels nous avons des observations sur la période maximale d'observation de notre base de données, soit plus de 2500 jours.

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 appartiennent 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'une ville pour une autre ville, 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-ci :

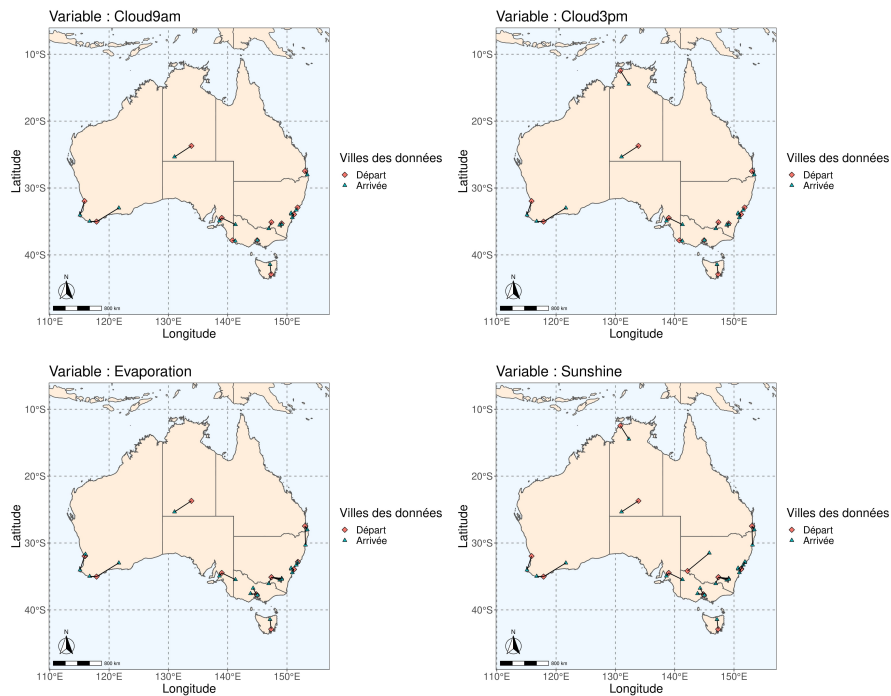


FIGURE 8 – Chemin des observations copiées (ville de départ et ville(s) d'arrivée(s)) pour certaines des variables complétées.

Au final, le *na.omit* nous donne une base de données avec 105546 observations. La missingness map et la distribution des observations par lieu sont les suivantes :

Missingness Map

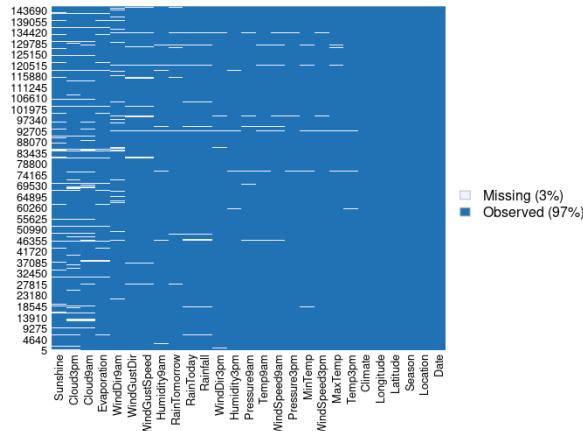


FIGURE 9 – Missingness Map des données complétées.

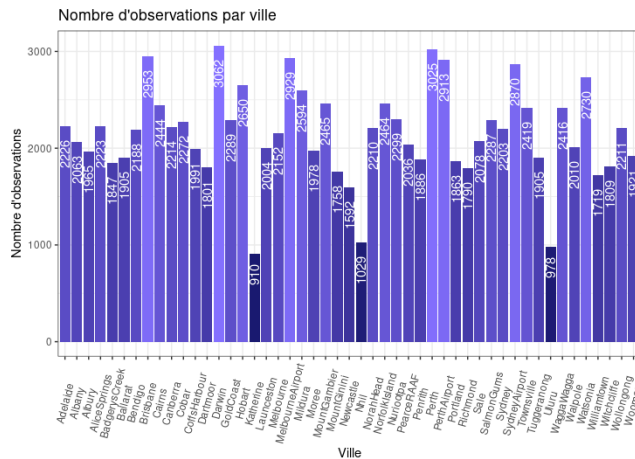


FIGURE 10 – Distribution des observations par villes dans notre base de données finale.

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.

Références

- [Wik21a] WIKIPÉDIA. *Latitude* — *Wikipédia, l'encyclopédie libre*. [En ligne ; Page disponible le 29-décembre-2021]. 2021. URL : <http://fr.wikipedia.org/w/index.php?title=Latitude&oldid=189341688>.
- [Wik21b] WIKIPÉDIA. *Longitude* — *Wikipédia, l'encyclopédie libre*. [En ligne ; Page disponible le 6-décembre-2021]. 2021. URL : <http://fr.wikipedia.org/w/index.php?title=Longitude&oldid=188614923>.
- [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_\(cartographie\)](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.