# Rapport d’analyse des données

Table des matières

[Rapport d’analyse des données 1](#_Toc142071650)

[Valeurs nulles 1](#_Toc142071651)

[En globalité 1](#_Toc142071652)

[Répartition géographique 2](#_Toc142071653)

[Répartition temporelle 3](#_Toc142071654)

[Gestion des valeurs nulles 7](#_Toc142071655)

[Corrélations 8](#_Toc142071656)

[Variables quantitatives 8](#_Toc142071657)

[Variables qualitatives 10](#_Toc142071658)

[Géographie 11](#_Toc142071659)

[Répartition géographique 11](#_Toc142071660)

[Clusterisation 12](#_Toc142071661)

[Bilan général 15](#_Toc142071662)

[TODO / A réfléchir 15](#_Toc142071663)

### Gestion des valeurs nulles

#### Cas particulier de RainToday

Cette variable est supposée être à True lorsque Rainfall est supérieure ou égale à 1mm. Un contrôle permet de confirmer cette règle. De plus, lorsque RainToday est à NA, alors Rainfall l’est également et réciproquement. Il n’est donc pas possible d’utiliser Rainfall pour renseigner les valeurs manquantes de RainToday.

#### Cas particulier de RainTomorrow

La variable ‘RainTomorrow’ doit en théorie être égale à la valeur de ‘RainToday’ du lendemain pour la même Location. Après vérification, c’est en effet le cas. Une possibilité pour gérer les NA de cette variable est donc d’affecter la valeur de RainToday de la veille des dates pour lesquelles RainTomorrow vaut NA. Malheureusement, cette piste n’aboutit pas : lorsque RainTomorrow vaut NA, soit RainToday de la veille est à NA également, soit la date n’est pas présente pour la Location.

#### Autres variables

Après ces différents constats, plusieurs pistes s’offrent à nous pour gérer les valeurs nulles.

[à mettre au propre et à discuter/tester]

* Drop des 3 villes avec « seulement » la moitié des dates ?
* Drop des variables avec un taux de NA trop élevé ?
* Reprise des données météo pour les dates/variables manquantes à partir de la ville la plus proche (exemple de Melbourne pour laquelle il manque beaucoup de données, alors que la quasi-totalité des infos est disponible pour MelbourneAirport, situé à quelques kilomètres) ?
* Reprise de la moyenne des valeurs du même jour de l’année à partir des années renseignées ?
* Récupérer les données manquantes à partir d’autres sources de données (site de la météo australienne) ?

### Clusterisation

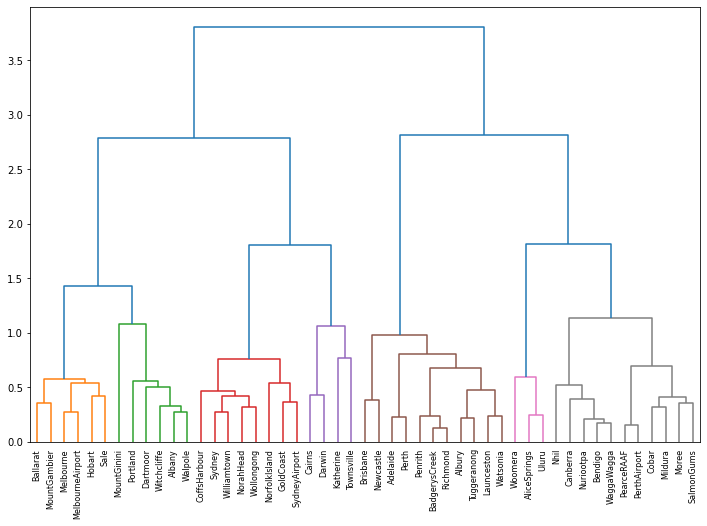
Nous pourrions ajouter une nouvelle variable précisant pour chaque Location le type de climat (désertique, subtropical, mousson, océanique,…) à partir de données complémentaires.

Plutôt que de le faire manuellement (ce qui reste une option, vu le nombre raisonnable de localités), nous pourrions mettre en place un modèle de clusterisation qui se chargerait lui-même de regrouper les Location présentant des similitudes.

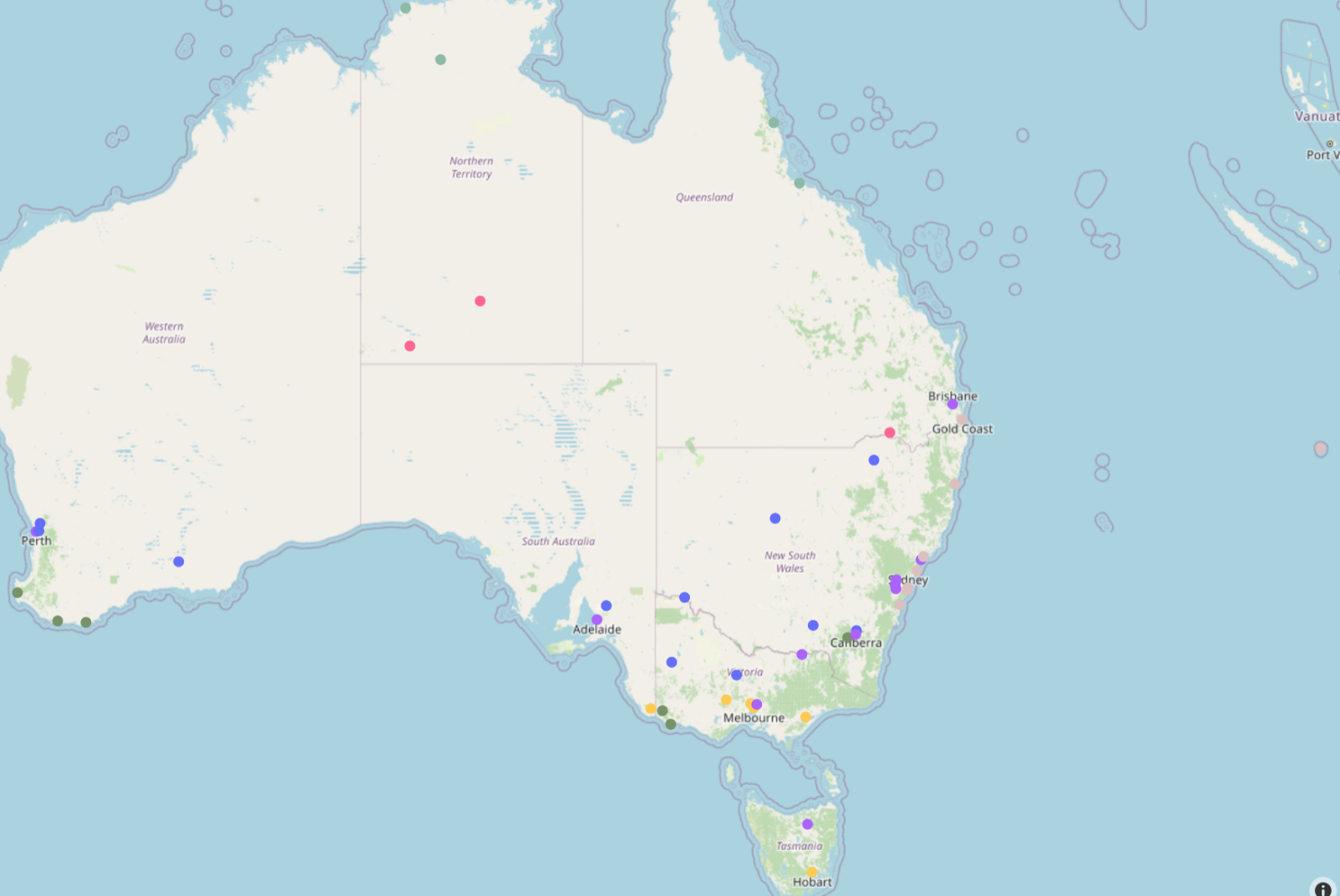
Une approche simple est de calculer la moyenne des valeurs pour chaque ville, puis de tracer le dendrogramme et effectuer une clusterisation par CAH. Nous avons choisi ici de faire 7 clusters (méthode *clusterisation\_groupee()* )

Nous effectuons la clusterisation à partir des données climatiques, et non de la latitude et de la longitude. La proximité géographique des résultats n’est donc pas le fruit de l’exploittaion de ces deux paramètres, mais bien de la cohérence de clusterisation, deux villes proches ayant généralement le même climat.

Le dendrogramme obtenu montre en particulier 2 petits clusters : l’un est composé de Woomera, Alicesprings et Uluru. Il s’agit de 3 villes très arides, deux dont sont au cœur du désert. La second petit cluster est constitué de Cairns, Darwin, Katherine et Townsville. Il s’agit des 4 villes situées le plus au nord.

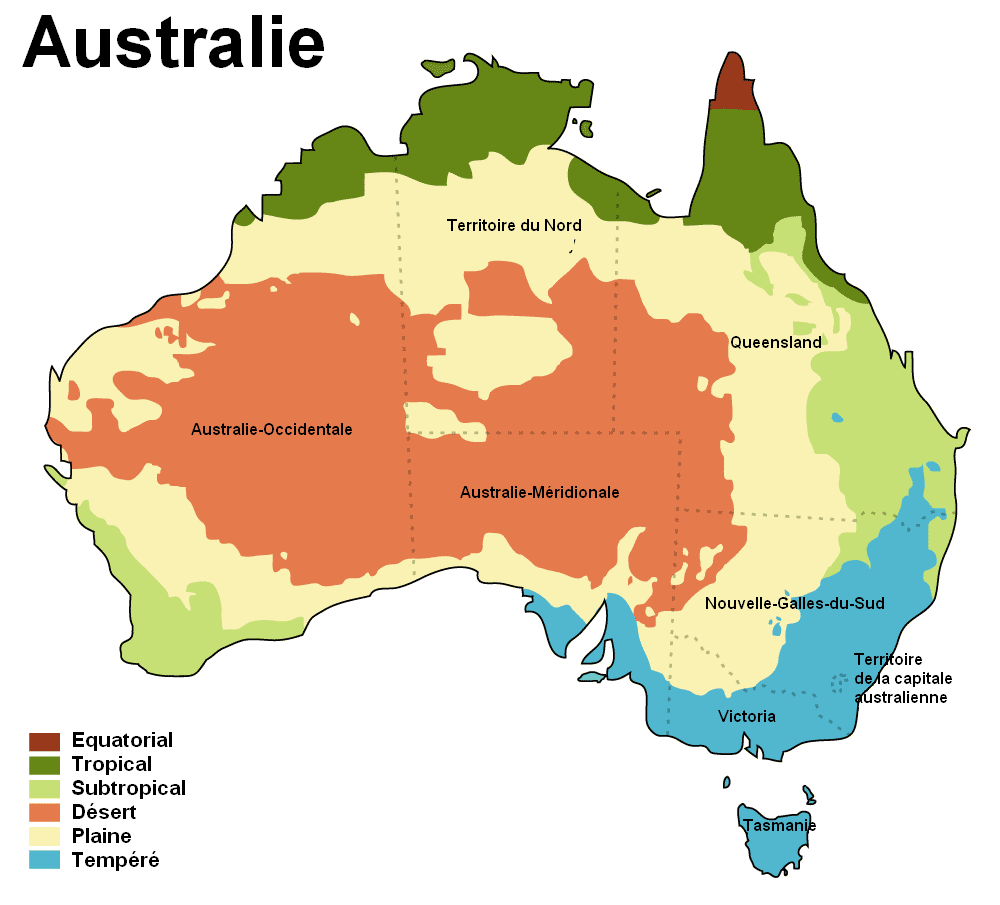


Regardons plus en détail sur la carte comment sont répartis les 7 clusters. Outre les deux exemples cités plus haut, nous retrouvons un groupe sur la côte est, un sur des villes côtières du sud, un autre qui est intermédiaire entre les villes côtières et le désert.



Regardons maintenant une carte représentant les zones climatiques de l’Australie. L’exemple ci-après est issu de la page Wikipédia ‘Climat de l’Australie’.

Zones climatiques australienne (source : Wikipédia : <https://fr.wikipedia.org/wiki/Climat_de_l'Australie> )



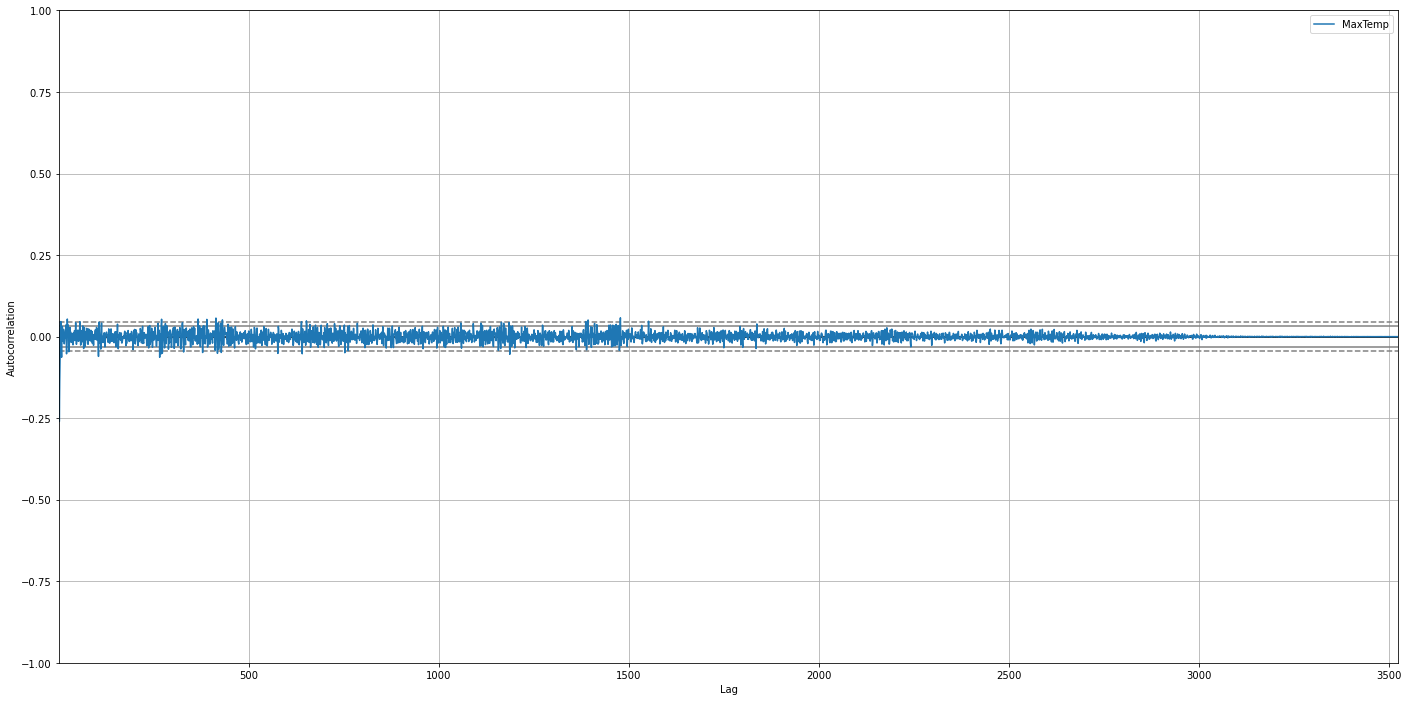
Nous retrouvons bien la cohérence des 4 villes nordiques, correspondant à la zone climatique tropicale. Les villes des plaines forment également un cluster à part entière. Les zones subtropicales et tempérées correspondent aux 4 derniers clusters, mais avec des frontières assez différentes. Le fait qu’un de notre cluster soit spécifique aux villes côtières orientales semble être un signe d’une cohérence intéressante à approfondir.

## Saisonnalité

Autocorrélation

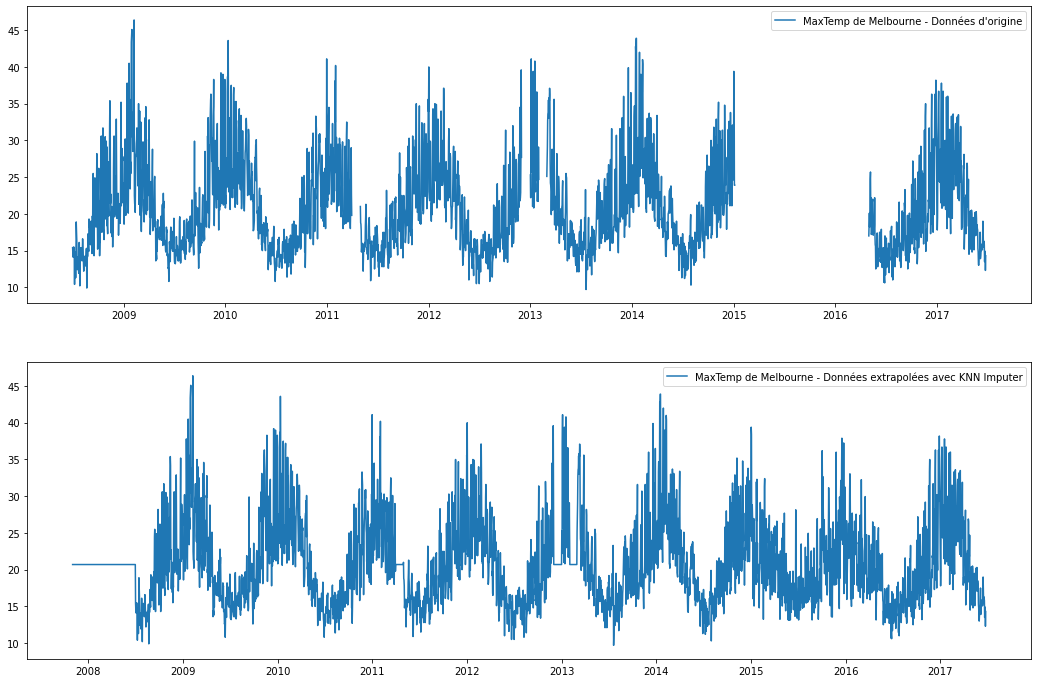
Le graphe d’autocorrélation des variables différenciées (ici MaxTemp) est proche de 0 en tout point.

* Est-ce à dire que la valeur pour une journée ne dépend jamais de celle de la veille ??
* Pourquoi ne voit-on ici aucune saisonnalité ?



## KNN Imputer

L’imputation par KNN nous permet de renseigner les NA d’une façon plus adaptée qu’en reprenant des moyennes. En particulier, cela nous permet de retrouver une saisonnalité et une dispersion des données semblables au jeu de données d’origine. Nous voyons sur le graphe ci-dessous la courbe des température maximales de Melbourne, avec en haut les données d’origines (comprenant plusieurs plages non renseignées) et en bas les données comprenant l’imputation knn. On note que les imputations faites entre 2015 et mi 2016 semblent globalement cohérentes et assez variées. En revanche, celles des petites plages (avril 2011, décembre 2012, février 2013) ont la même valeur. Il en va de même pour les données antérieures à la première valeur renseignée.



## Bilan général

[in progress]

* Enrichissement des données avec AmplitudeTemp, longitude, latitude, climat
* Drop des données redondantes
* Gestion des NA : KNN imputer (évaluer la valeur de knn ainsi que les colonnes pertinentes)