Data Mining: Analyse de données océaniques et météorologiques

Ambrosino-Ielpo Gwenaël Fort Tz'ayik Gomès Antoine Jourdan Luca
Touati Gaïs

19 juin 2019

1 Introduction

1.1 Dataset

Dataset : NOAA ICOADS, A global marine meteorological and surface ocean dataset (`https://www.kaggle.com/noaa/noaa-icoads`)

Le "International Comprehensive Ocean-Atmosphere Data Set" contient de nombreuses données océaniques et météorologiques (plus de 75 variables différentes sont prises en compte : température de la surface de l'eau, pression, météo, direction du vent, force du vent...) qui ont été relevées sur une centaine d'années. Les données proviennent de nombreuses mesures nationales et internationales : des bateaux marchands, des bouées, des stations côtières, des plates-formes marines, etc.

Nous soulignons qu'un gel du budget de la NOAA rend actuellement leur site inaccessible pour une période indéterminée, qui était malheureusement la principale source d'information et de renseignement disponibles sur les données océaniques, ce qui a considérablement compliqué nos recherches ...
`https://www.noaa.gov/`

1.2 Objectifs

Notre dataset est très complet, et énormément de pistes auraient pu être suivies. Nous nous sommes donc fixés 3 objectifs principaux :

Tout d'abord, nous verrons s'il est possible de corréler des phénomènes océaniques, des anomalies, avec des phénomènes météorologiques. Y a t-il un lien entre les différentes mesures et le climat (pluie, brouillard...) ?

Nous voudrions détecter des courants océaniques tel que le Gulf Stream ou El niño, ces courants devraient théoriquement être visible, notamment El niño qui est marqué par un net changement de température. Mais nos données étant des relevés sporadiques réalisés par des bateaux et des bouées, nous n'en avons pas la certitude.

Nous allons également tenter de détecter les ouragans, notamment l'ouragan Katrina, qui a dévasté la Nouvelle-Orléans au mois d'août 2005.

2 Préparation et prise en main des données

Comme mentionné précédemment, la manipulation des données a été une première grande difficulté. L'ampleur du dataset étant très importante (170 Go) la plupart de nos analyses ont été faites sur l'année 2005 uniquement, que nous avons sélectionnée en sachant que c'était une année intéressante, puisque c'était à la fois l'année de l'ouragan Katrina, et d'un événement "El niño". De plus nous avons identifié 18 variables qui nous semblaient les plus intéressantes : la latitude et la longitude, la direction et la vitesse du vent, la météo, la pression de l'air, la température de l'air et de la surface de l'eau, la température au wetbulb, la quantité et la taille des nuages, la direction, la période et la hauteur des vagues ainsi que celles de la houle et le timestamp. Nous nous contenterons donc de traiter uniquement ces variables.

Nous avons utilisé la librairie Python BigQuery de Google pour récupérer et manipuler les données. BigQuery permet de faire des requêtes SQL efficace sur des grands jeux de données. BigQuery permet aussi de prévoir la quantité de données scannées par nos requêtes et de prévoir leur temps d'exécution. En utilisant BigQuery depuis Kaggle, nous avons le droit de scanner jusque 5To de données par mois, ce qui a été suffisant pour ce projet. En revanche, en dehors de Kaggle, nous n'avons pu récupérer les données que 1go par 1go ce qui fut très contraignant. Au final, nous avons téléchargé en dehors de Kaggle l'entièreté des données de l'année 2005 pour pouvoir les manipuler de partout, et nous avons utilisé Kaggle lorsque nous avions besoin de données supplémentaires.

L'année 2005 à elle seule représente encore 1.2Go et est toujours difficilement manipulable. Nous avons donc discrétisé l'année pour chaque mois, ce qui nous permet de visualiser l'évolution des résultats au cours de l'année. Cela nous a permis d'accélérer considérablement la vitesse de calcul de nos différents algorithmes pour certains problèmes spécifiques, en utilisant parfois seulement des échantillons quand les données étaient encore trop conséquentes.

Notre première approche pour comprendre les données a été de construire une matrice de corrélation sur les paramètres que nous avons décidé de garder (Figure 1). Ainsi nous pourrions identifier tout de suite les variables qui seraient intéressantes à traiter ensemble.

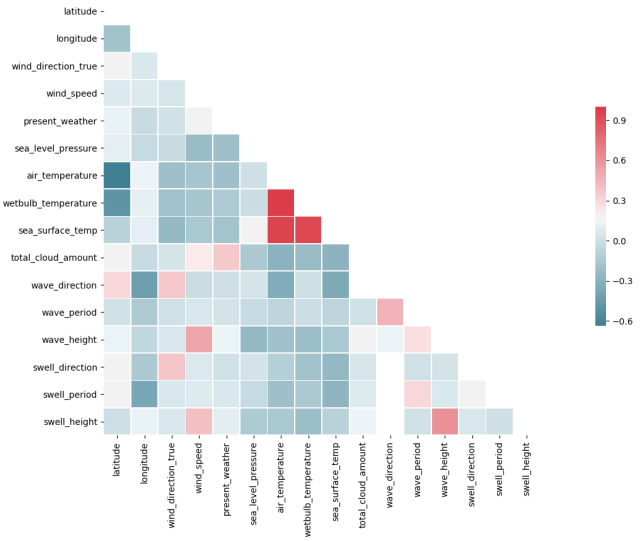

FIGURE 1 – Matrice de corrélation des variables

Nous observons quelques corrélations évidentes. Nous pouvons voir que la température de l'eau, la température de l'air et la température de surface de la mer sont fortement corrélées. Nous pouvons également observer une forte corrélation négative entre la latitude et la température de l'air tandis qu'il ne semble pas y avoir de corrélation entre température et longitude. Ces résultats sont plutôt décevants et nous n'en tirons aucune information utile.

3 Analyse du climat

3.1 Détection des vagues de froids

L'objectif ici est de détecter et de visualiser les variations de températures de notre dataset sur une carte, et d'essayer de trouver des liens entre cette variation et d'autres paramètres. Pour ce problème nous avons donc travaillé sur l'évolution jour par jour de la température pour chaque mois. Nous n'avions donc besoin ici que de la latitude, la longitude (pour l'affichage), la température, la pression (pour trouver un lien) et le timestamp.

Nous avons ensuite utilisé une méthode de clustering, l'algorithme k-means, afin de repérer les différentes zones de température. Le premier problème que nous avons rencontré est le choix du nombre de clusters, car il a fallu trouver un compromis entre la précision et la robustesse des résultats. En effet si le nombre de cluster est trop faible nous n'obtenons pas beaucoup d'informations, mais s'il est trop élevé,

des faibles changements dans la température peuvent modifier complètement la structure des clusters. Après plusieurs essais, nous avons finalement choisi de garder 6 clusters. (Figure 2)

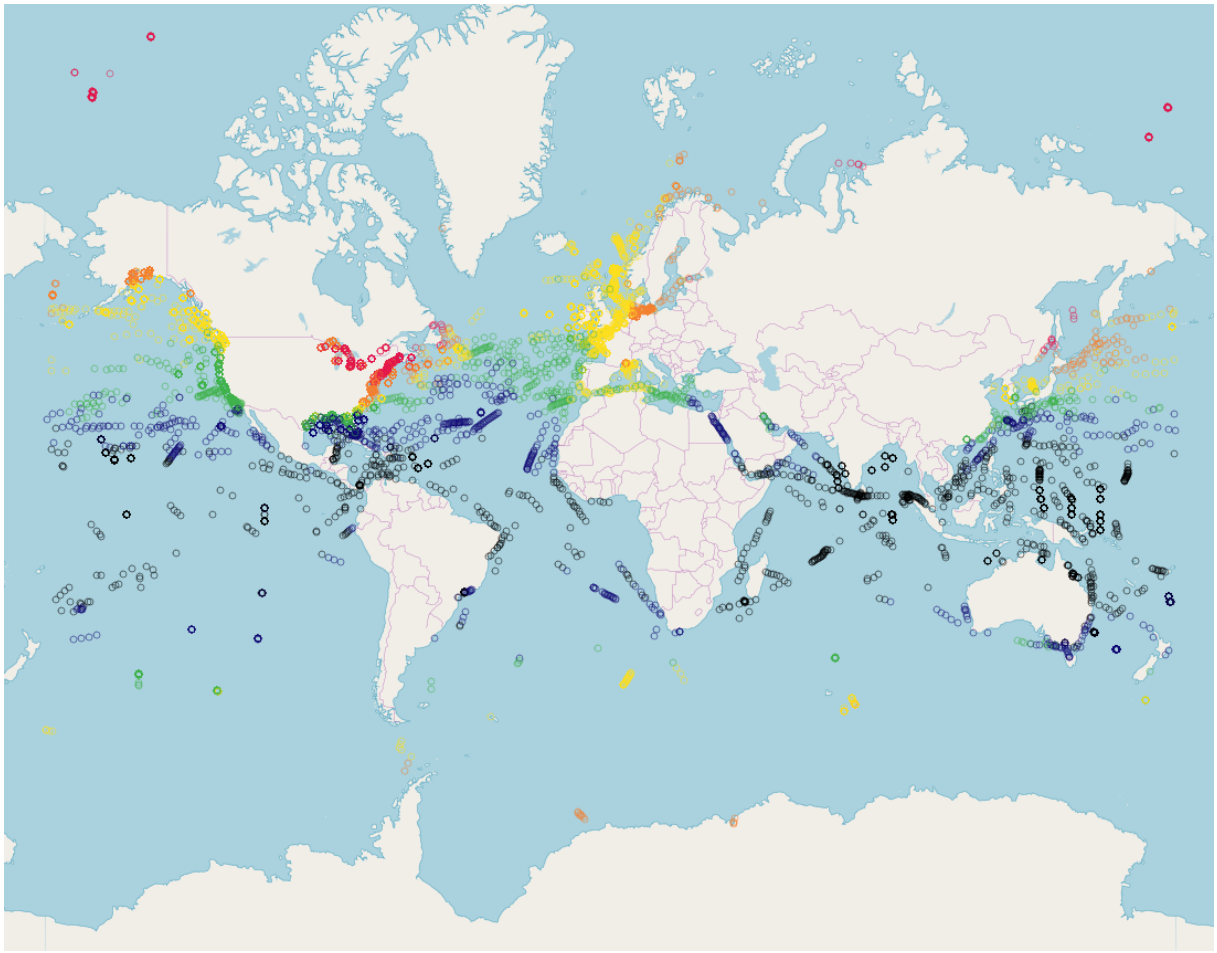

FIGURE 2 – K-means sur les températures d'un jour en janvier 2005
Du moins au plus chaud : rouge, orange, jaune, vert, bleu, noir

En observant ensuite les résultats jour par jour, nous avons pu déterminer les déplacements des vagues de froid sur notre carte. Nous avons également pu détecter certaines "anomalies", lorsque nous constatons un changement brutal de couleur (et donc de cluster), on passe par exemple une fois du vert à l'orange.

Nous avons utilisé ensuite la même méthode sur la pression au niveau de la mer pour essayer de trouver un lien entre la pression et la température. Nous pensions que selon la loi des gaz parfait (PV = NRT), la pression augmenterait avec la température de manière presque proportionnelle. Nous nous sommes rendus compte que ce n'est pas le cas du tout (Figure 3 et 4) Une pression élevée est en fait signe de temps calme, et une pression faible de temps agité, ce qui explique que les pressions faibles se trouvent à des températures autour de 5 degrés, qui correspondent à un temps venteux et pluvieux.

Nous avons également essayé d'autres méthodes de clustering tel que DBSCAN ou Meanshift mais nous les avons abandonnés car les résultats étaient peu concluant.

Afin de mieux voir ces différents changements chaque jour, nous avons construit deux visualisation retrouvables sur github :

- température : `https://bobsloth.github.io/icoads_dm/temp_janvier_2005/visu`
- pression : `https://bobsloth.github.io/icoads_dm/pressure_janvier_2005/visu`

(`https://github.com/Bobsloth/icoads_dm`)

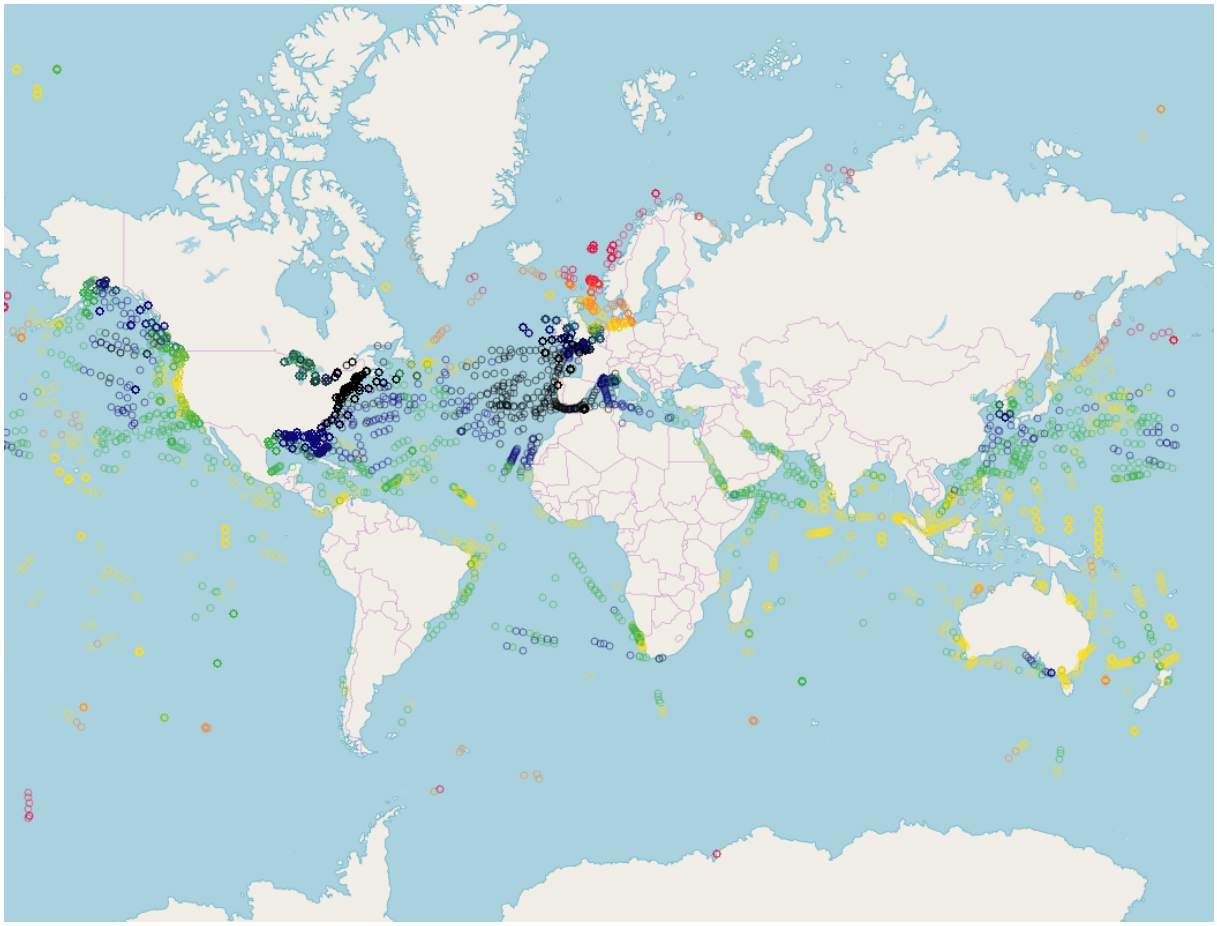

FIGURE 3 – Aperçu carte des pressions
Du moins au plus : rouge, orange, jaune, vert, bleu, noir

3.2 Anticyclones

Un anticyclone est une zone de haute pression atmosphérique par rapport à son voisinage. Ils sont caractérisés par un lent mouvement de l'air du ciel vers le sol, ce qui ralentit les vents et apporte un temps généralement calme et ensoleillé. Les mouvements des zones anticycloniques ont une utilité majeure dans la prévision de la météo.

Dans notre carte des pressions nous voyons que la zone entre l'Amérique du Nord et l'Europe de l'Ouest est constamment sous haute pression. Cela correspond à la ceinture anticyclonique subtropicale appelée "crête subtropicale" dans laquelle se forme beaucoup d'anticyclones. En particulier, on voit nettement se déplacer autour de la France et l'Espagne un anticyclone nommé "Anticyclone des Açores" qui a un impact majeur sur les prévisions météo en Europe de l'Ouest. En calculant des clusters sur la pression et la température on peut donc se faire une première idée de la météo. (Figure 4 et 5)

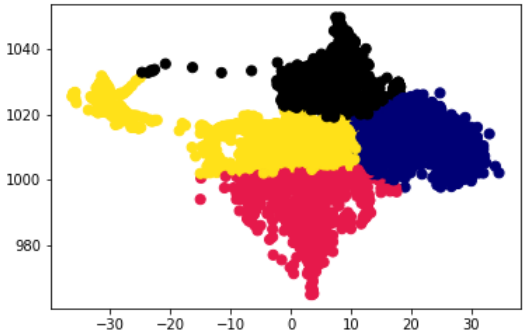

FIGURE 4 – Pression en fonction de la température le 25 Janvier 2005

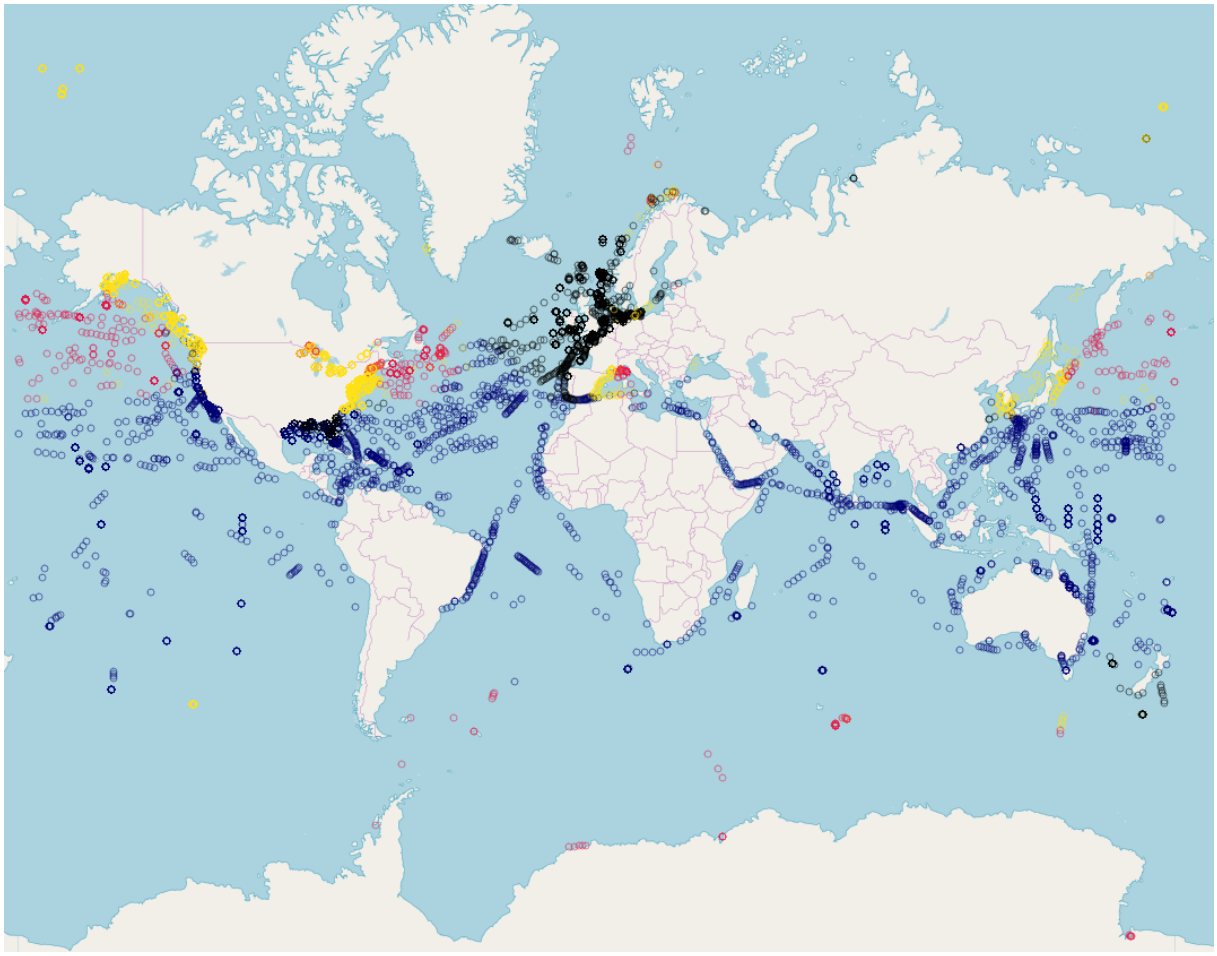

FIGURE 5 – Pression en fonction de la température le 25 Janvier 2005

Ici le cluster noir autour de l'Europe de l'Ouest indique un temps relativement froid et calme, tandis que le cluster rouge représente un temps agité. On s'attend donc à un temps calme sur le nord de la France et un temps très agité sur la méditerranée. Nous pouvons confirmer nos prédictions avec le site de météo France, qui nous indique que le 25 janvier, le temps était "éclairci" à Brest, tandis que Bastia et la côte-d'Azur était recouverts par 8 cm de neige.

3.3 Détection des ouragans

Nous nous sommes ensuite penchés sur les ouragans, qui sont des évènements extrême qui devraient être identifiables. Il faut savoir que les ouragans sont des phénomènes climatiques qui n'apparaissent qu'à une distance d'environ 10 degrés au-dessus ou au-dessous de l'équateur car c'est ici que la force de Coriolis agit sur le déplacement des masses d'air et permet à l'ouragan d'acquérir son mouvement rotatif. Les principales caractéristiques d'un ouragan sont : une très faible pression atmosphérique, une température de l'eau anormalement élevée et bien sur des vents très violents.

Plusieurs ouragans ont eu lieu en 2005, le plus marquant étant Katrina qui fit des ravages aux Etats-Unis. Nous avons donc essayé de repérer cet ouragan. Pour cela, nous avons observé l'évolution de la pression, de la température de l'eau et de la puissance du vent au cours du mois d'août. Cependant, les mesures de vent du dataset semblent être très incorrectes et inutilisables. Premièrement, l'unité donnée par la description du dataset est clairement fausse, et nous ne savons pas vraiment quelle unité est réellement utilisée. De plus, peu importe notre raisonnement, nous n'observons aucun vent violent, nulle part. Nous n'avons donc pas pu exploiter ce paramètre et nous nous sommes limités à l'analyse de la pression et la température de l'eau.

En utilisant à nouveau k-means, nous pouvons nettement voir l'ouragan avancer vers les côtes de la Louisianne à partir du 26 août tout en gagnant en puissance. Il atteint finalement la terre le 29 août et nous perdons sa trace. Le cluster violet représente une zone de haute température de l'eau et de basse pression, et est particulièrement présent au sud des États-Unis (Figure 6). On remarque tout de même que d'autres zones du globe sont sous les même conditions sans que ce soit des ouragans pour autant.

Nous nous étonnons que la pression dans l'ouragan ne soit pas bien plus basses, et nous soupçonnons des mesures imprécises dans des conditions extrêmes.

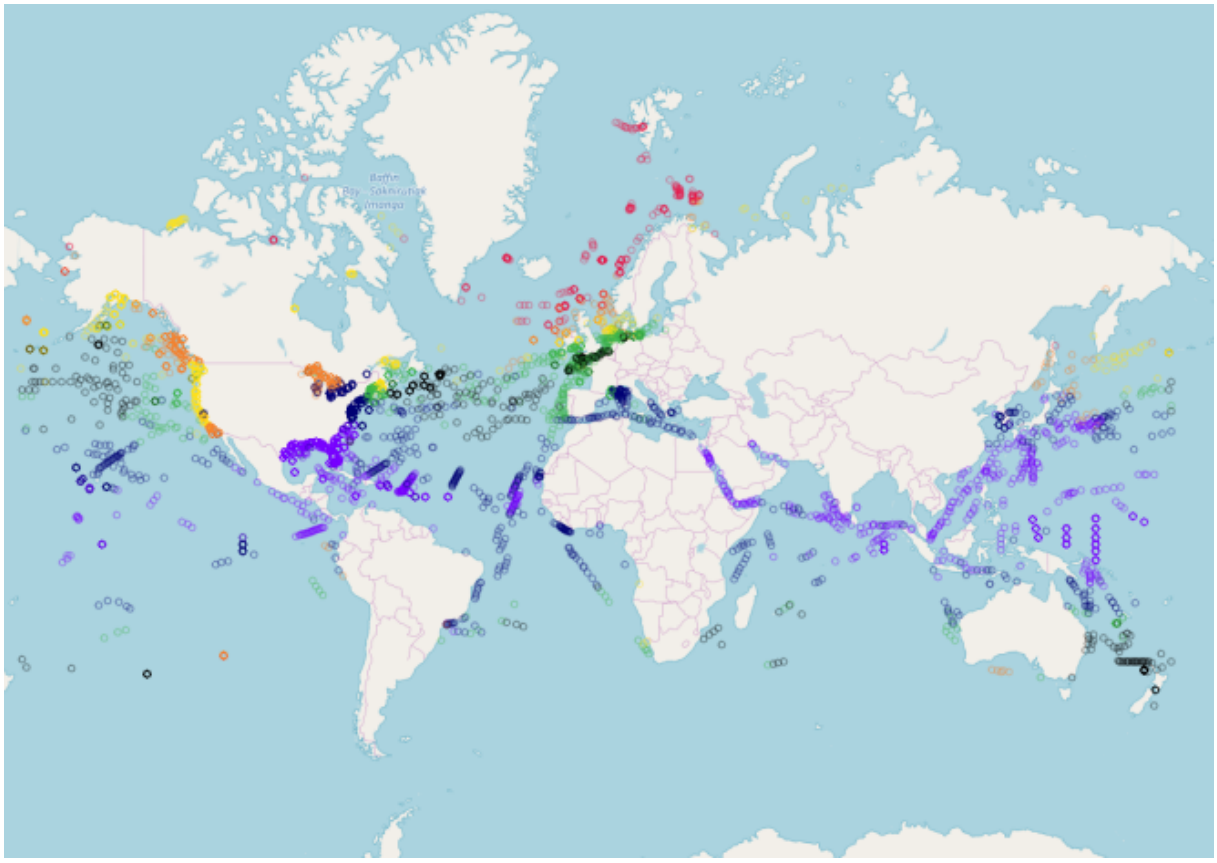

FIGURE 6 – Pression et température de l'eau pendant l'ouragan Katrina

4 Détection de El Niño

El Niño est le nom donné à l'évènement de l'apparition d'un courant marin particulièrement chaud à l'ouest de l'Amérique du Sud qui a un impact très important sur le climat local. Le courant a notamment un effet très récessif sur la faune maritime s'y trouvant, et il marque donc la fin de la saison de pêche en Amérique du Sud. Au niveau mondial, les effets sont moins reconnus mais de nombreuses recherches en cours semblent indiquer qu'il a un impact non négligeable. El Niño se produit de manière irrégulière, tous les 2 à 7 ans et dure entre 9 mois et 2 ans.

Pour déterminer si on est en période de El Niño, les chercheurs utilisent le South Oscillation Index (SOI), qui est une valeur calculée sans unité comprise entre -35 et 35. Le SOI est calculé pour chaque mois, si sa valeur est en dessous de -7 pendant plusieurs mois, cela indique qu'on est en présence d'El Niño. Si la valeur dépasse 7 pendant plusieurs mois, on est en événement "El Nina", qui est la formation d'un autre courant marin qui lui aussi a un impact conséquent sur le climat. Le SOI est calculé à partir de la différence de pression au niveau de la mer entre Tahiti et Darwin, Australie.

$$SOI = 10 \frac{[\ Pdiff - Pdiffav\ ]}{SD(Pdiff)}$$

where
Pdiff = (average Tahiti MSLP for the month) - (average Darwin MSLP for the month),
Pdiffav = long term average of Pdiff for the month in question, and
SD(Pdiff) = long term standard deviation of Pdiff for the month in question.

FIGURE 7 – Formule du calcul du SOI

Nous commençons donc par calculer le SOI pour chaque mois de chaque année de 2001 à 2016 afin de trouver quelques épisodes de El Niño. Pour cela nous utilisons BigQuery depuis Kaggle pour récupérer tous les données de pression au niveau de la mer, dans deux zones délimitées autour de Tahiti et de Darwin. Pour déterminer le SOI nous avons aussi besoin de Pdiff et SD(Pdiff) qui sont des valeurs définies sur une grande période. Avec notre dataset il serait trop fastidieux de les calculer, nous utilisons donc des valeurs utilisées dans un article de recherche de Kevin Trenberth en 1984. (Figure 7)

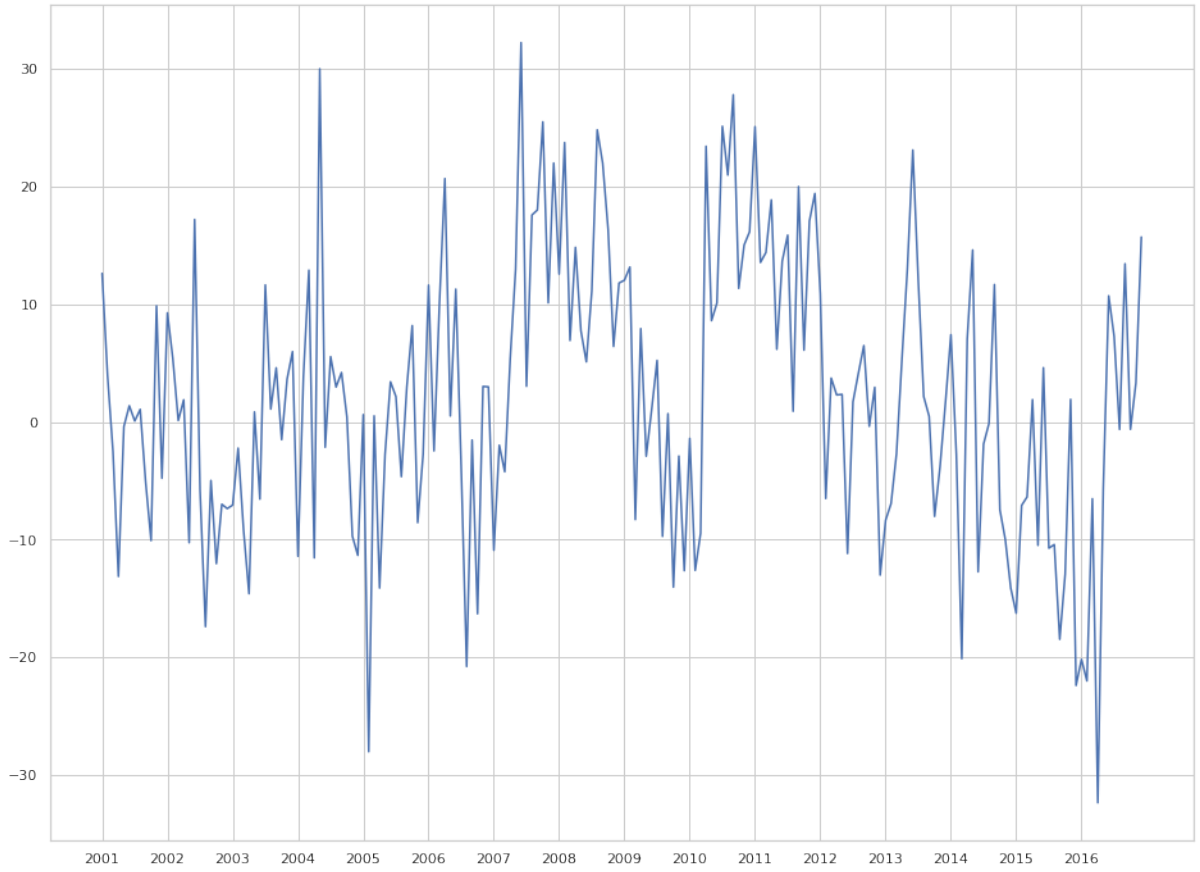

FIGURE 8 – Valeur du SOI entre 2001 et 2016

On remarque plusieurs épisodes El Niño : en 2002-2003, en 2005, en fin 2006, en fin 2009 et un important en 2015-2016. Ayant facilement accès aux données de 2005 nous avons essayé de visualiser El Niño sur une carte. (Figure 8) Mais nous n'avons malheureusement pas assez de données sur cette période à l'ouest de l'Amérique du sud pour que ce soit concluant. Cependant, le site de la NOAA nous confirme que 2004-2005 était bel et bien une période El Niño

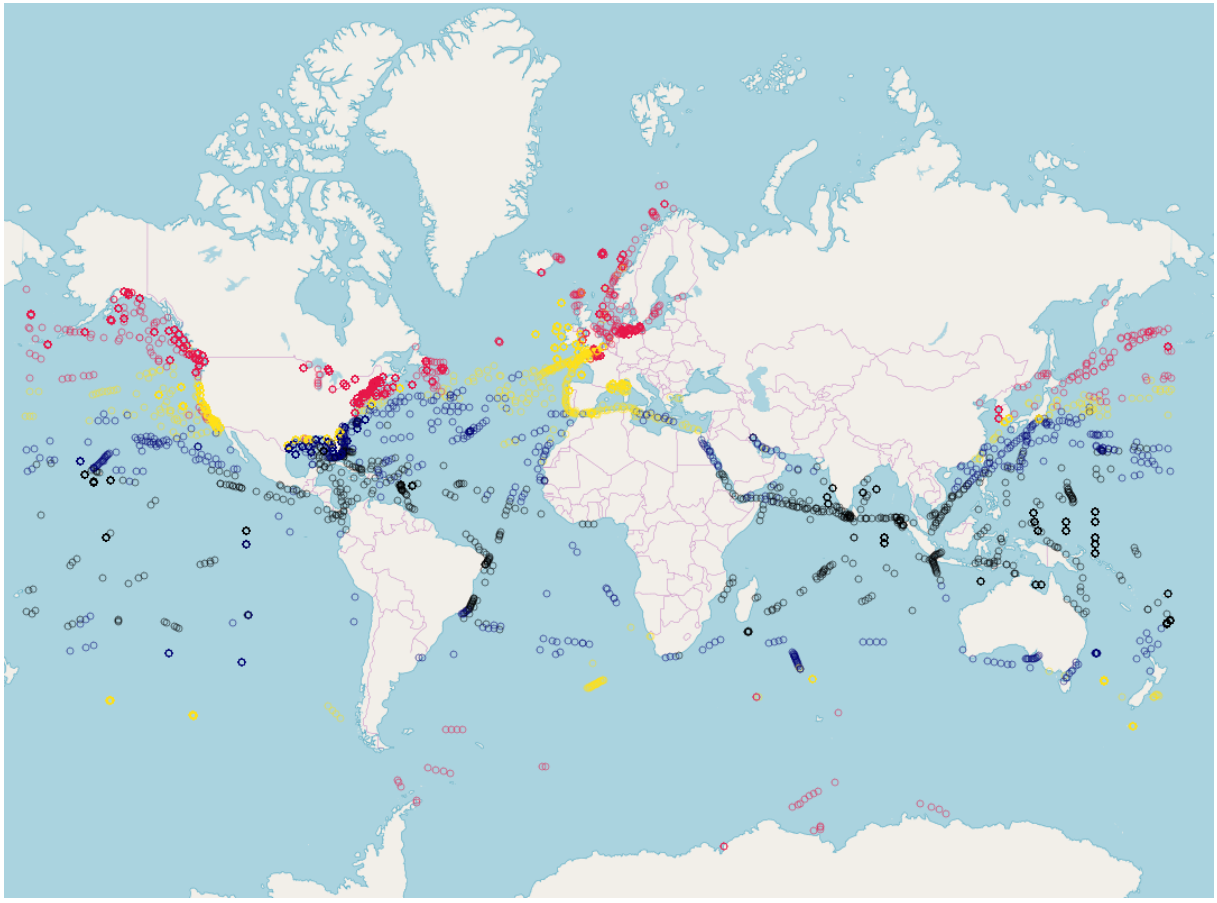

FIGURE 9 – Température de l'eau en janvier 2005

5 Conclusion

La taille importante du dataset, qui nous a semblée être un avantage au moment du choix, s'est finalement retournée contre nous. Nous pensions nous faciliter la recherche avec énormément de données, mais elles ont demandé beaucoup de pré-traitement pour n'en utiliser finalement qu'une petite partie. De plus nos tentatives d'interprétations des données globales au début du projet ont toutes été infructueuses et nous ont retardé. Nous tirons donc de cet exercice une première expérience de gestion de grand dataset et de filtrage de données. Cette difficuté aurait peut être pu être évité en utilisant des outils adaptés aux dataset de grandes tailles, tels que Spark, et c'est une piste que nous aimerions explorer dans le futur.

Une deuxième limite que nous avons rencontrée est la confiance que l'on peut avoir dans les données. Beaucoup d'attributs parmi les 75 disponibles dans le dataset étaient finalement inutilisables car très imprécis. L'exemple le plus gênant est celui de la vitesse du vent pour laquelle nous n'avons trouvé aucune interprétation des mesures convenable. Nous trouvons également les mesures de pression étranges, notamment lors des évènements extrêmes des ouragans, qui semblent anormalement élevées.

Nous arrivons donc à une troisième limite qui est notre manque certain de connaissances en météorologie, qui est pourtant nécessaire à l'interprétation des résultats. Nous avons donc du effectuer un grand nombre de recherches sur les événements que nous avons étudié pour pouvoir comprendre ce que l'on observait, ce qui a été une part importante du temps passée sur ce projet.

Finalement, nous avons pu observer un grand nombre de phénomènes météorologiques, et nous avons réussi à identifier des événements océaniques importants. A partir de toutes nos recherches, nous sommes maintenant capable d'avoir une bonne idée de la météo à partir des mesures de pression et de température. Il serait intéressant de croiser ces résultats avec ceux d'un dataset de météorologie uniquement, ou de les coupler avec des mesures sur Terre pour compléter les observations.