

# Projet de MRR

Abdoulaye SAKHO, Naïm SOUNI

Novembre 2021

## Introduction

Notre projet est d'être capable, à partir de données sur un ouragan, de prédire son intensité 24h après. L'intensité de l'ouragan 24h dans le futur est donc notre variable cible.

## 1 - Base De Données

### 1.1 Présentation

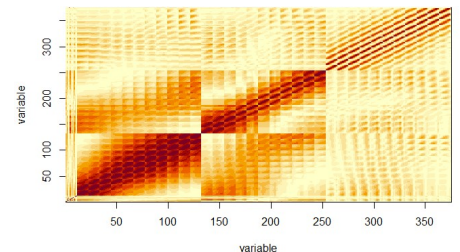
Notre BDD est composée à l'origine de 376 variables explicatives ( $p$ ) et de 10 308 observations ( $n$ ). Nous sommes donc dans un cas ici où  $n \gg p$ .

Les variables explicatives de notre problème sont: la vitesse longitudinale au sein de l'ouragan  $u$ , la vitesse latitudinale  $v$ , l'altitude correspondant à une pression de 700hPA  $z$ . Il y a aussi d'autres variables plus général concernant un ourgan: son identifiant, l'instant  $t$  auquel on l'observe, sa distance à la côte la plus proche entre autres. Ces variables supplémentaires sont au nombre de 13.  $u, v, z$  sont des images de  $11 \times 11$  pixels (soit donc  $3 \times 11 \times 11 = 363$  pixels, chacun représentant une variable). C'est pourquoi notre nombre de variables explicatives est aussi grand par rapport à leur description.

Notre variable cible correspond à la prédiction (24h dans le futur) de la vitesse maximale des vents au sein de cet ouragan en noeuds. On observe au sein de notre data set qu'elle varie entre 0 et 140. 140 noeuds correspond à un ouragan de catégorie 5 sur l'échelle de Saffir-Simpson.

### 1.2 Corrélations

Pour étudier les corrélations de nos variables entre elles et par rapport à notre variables cible, nous avons commencé par effectuer un tableau de corrélation (en valeurs absolues) avec les variables **quantitatives**. Dans un premier temps, nous avons observé ce tableau des corrélations des données d'origine sous forme d'image (Fig à droite). On observe beaucoup de cases sont très sombres donc, il existe une forte corrélation entre les deux variables concernées.



### 1.3 Restructuration des données

Au début et à la fin de chaque ouragan, on observe parfois dans notre dataset que la variable cible prend systématiquement la valeur 0. Cela vient probablement du fait que notre variable cible prédit 24h après un des paramètres de l'ouragan. Ces observations peuvent perturber notre construction de modèles, nous avons décidé de nous en passer car nous avons ici  $n \gg p$  et cela reste le cas même après la suppression. Afin de représenter au mieux l'aspect cyclique d'une journée, nous avons décidé "d'éclater en deux" la variable instant-t de notre dataset. Enfin, les colonnes des vitesses (respectivement longitudinale, latitudinale et l'altitude) sont très nombreuses et exprime la même dynamique. C'est pourquoi nous avons décidé de passer respectivement pour  $u, v$  et  $z$  de  $11 \times 11 = 121$  variables/colonnes à 5 colonnes/variables seulement (on a fait un sorte de compressions d'image). Le schéma de la (Fig.1) illustre notre méthode. Nous avons pris les moyennes de chaque  $Z_i$  pour construire les nouvelles variables.

Après toutes ces modifications, on a  $n = 9596$  et  $p = 29$ .

Nous avons à nouveau observé les corrélations de notre jeu de données modifié à l'aide de la fonction `corrplot()` (Fig.1) et `image()` de **R**.

Dans un second temps, nous avons reporté le tableau de corrélations dans un fichier .CSV. À l'aide d'une mise en forme conditionnelle sous Excel, nous avons observé le même phénomène (Fig.2). De plus nous avons

remarqué que notre cible était fortement corrélée à plusieurs variables dont notamment windspeed ce qui est cohérent.

## 2 - Cros Validation

Nous avons décidé pour la suite de notre projet de ne pas de Cross Validation dite "Classique". Nous préférons étudier les ouragans tout en entier.

Si nous avons  $m$  ouragans dans notre dataset, nous en utiliserons  $l$  pour l'apprentissage de nos modèles et  $m - l$  pour les tests de ces modèles. Cela nous semble plus cohérent.

## 3 - Difficultés anticipées

Nous pensons que les méthodes de construction de modèles vues en cours vont être inefficaces ici. Nous estimons qu'il nous faudra utiliser des modèles non vus en cours, plus particulièrement: le l'Elastic-Net et le Group-Lasso.

## 4 - Bibliographie

Échelle de Saffir-Simpson, Lasso Wikipédia et NOAA IBTrACS

## 5 - Annexes

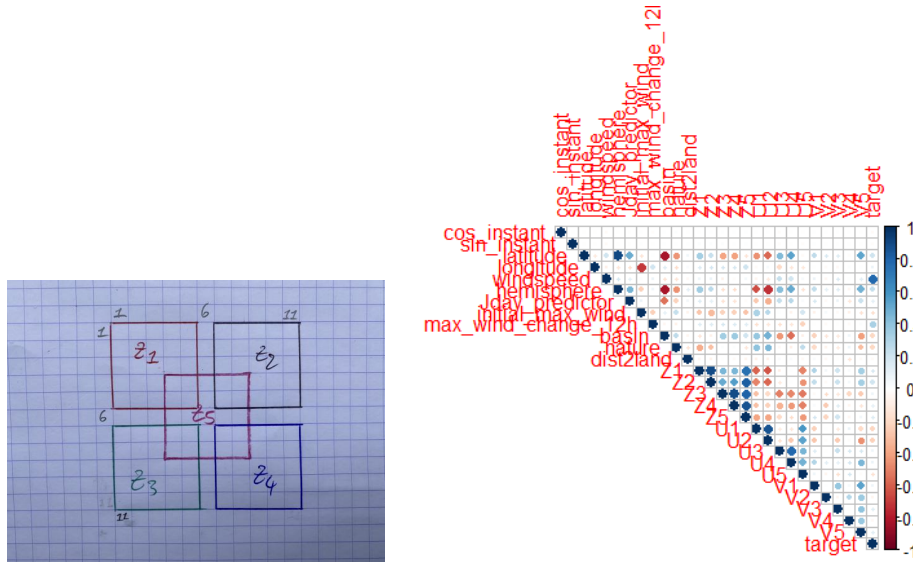


Figure 1: Schéma de la "compression" des images (on a pris la moyenne des  $Z_i$  à chaque fois). Sortie de la fonction *corrplot* après resturcturation de la BDD.

Figure 2: Excel: Mise en surbrillance des corrélations supérieur à 0.8

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	1.00																											
2	0.0324631	1.00																										
3	0.0324631	0.0412396	1.00																									
4	0.0412396	0.0412396	0.0412396	1.00																								
5	0.0412396	0.0412396	0.0412396	0.0412396	1.00																							
6	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00																						
7	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00																					
8	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00																				
9	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00																			
10	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00																		
11	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00																	
12	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00																
13	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00															
14	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00														
15	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00													
16	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00												
17	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00											
18	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00										
19	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00									
20	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00								
21	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00							
22	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00						
23	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00					
24	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00				
25	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00			
26	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00		
27	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00	
28	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00
29	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00
30	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	0.0412396	1.00