

Mini Projet : Analyse de Données

Filière : Cycle d'Ingénieure en Génie Informatique s1

Présenté par :

- AOUIJIL NOURA
- BELKHIRI YOUSSEF
- EZ-ZINE HAMZA
- HDIDOU OUSSAMA
- HAMDAROU ALAROUÏ EL
HASSANE

Resp. : BOUARAFA SALEH

Faculté des sciences et techniques
d'Errachidia

Université Moulay Ismaïl

2022/2023

1. Description :

2.ACP

....

3. Analyse Discriminante :

Nous intéressons à la relation entre un groupe de variables indépendantes et une variable catégorielle. En utilisant cette relation, nous pouvons prédire une classification basée sur les variables indépendantes ou évaluer dans quelle mesure les variables indépendantes séparent les catégories dans la classification.

On prend la variable « plage » comme une variable catégorielle métrique qui contient 2 valeurs :

- 0 : La ville n'a pas une plage
- 1 : La ville a une plage

Et on prend les variables suivantes : Température maximale, température minimale, humidité et pluviométrie comme des variables indépendantes.

1^{er} Etape : On vérifie l'existence la différence entre les groupes

Nous utilisons trois indicateurs pour déterminer si les groupes sont statistiquement différents les uns des autres : la mesure de tendance centrale (moyenne ou variance), le test de F de Fisher et le coefficient de Wilks Lambda.

Cette première étape permet de déterminer les variables les plus discriminantes entre les groupes.

- La moyenne / l'écart type :

Statistiques de groupe					
plage		Moyenne	Ecart type	N valide (liste)	
				Non pondérées	Pondérées
non	TEMP_MAX	21.2917	11.13034	168	168.000
	TEMP_MIN	12.3393	9.10539	168	168.000
	HUMIDITE	73.3274	20.27727	168	168.000
	PLUVIOMETRIE	87.7976	79.32919	168	168.000
oui	TEMP_MAX	23.0985	7.17407	132	132.000
	TEMP_MIN	17.3598	7.37939	132	132.000
	HUMIDITE	83.5379	55.19174	132	132.000
	PLUVIOMETRIE	124.0833	98.09117	132	132.000
Total	TEMP_MAX	22.0867	9.62026	300	300.000
	TEMP_MIN	14.5483	8.74051	300	300.000
	HUMIDITE	77.8200	39.87498	300	300.000
	PLUVIOMETRIE	103.7633	89.75523	300	300.000

- Le test de F et lambda de Wilkes :

Tests d'égalité des moyennes de groupes					
	Lambda de Wilks	F	ddl1	ddl2	Sig.
TEMP_MAX	.991	2.622	1	298	.106
TEMP_MIN	.918	26.466	1	298	.000
HUMIDITE	.984	4.910	1	298	.027
PLUVIOMETRIE	.960	12.548	1	298	.000

L'application du test de F dans notre cas indique que la variable "TEMP_MIN" est la plus influence pour discriminer les groupes, étant donné que le F est élevé (26.46) et Sig est très proche de 0,00.

De plus lambda de Wilks des variables "TEMP_MIN" est la plus petite que les autres (0.918)

2^{eme} Etape : On vérifier la validité de l'étude

Cette etape contient aussi 3 indicateurs :

- Le test M de Box :

Résultats du test		
Test de Box		77.872
F	Approx.	12.834
	ddl1	6
	ddl2	554526.987
	Sig.	.000
Teste l'hypothèse nulle de matrices de covariance à égales populations.		

La significativité du test de F doit tendre vers 0. S'il est supérieur à 0,05, l'analyse n'est pas valide. Dans ce cas le modèle est valide car Sig de f est 0.00 (inférieur a 0,05)

- La Corrélation globale :

Valeurs propres				
Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	.679 ^a	100.0	100.0	.636
a. Les 1 premières fonctions discriminantes canoniques ont été utilisées pour l'analyse.				

Plus la corrélation globale est proche de 1, plus le modèle est meilleur, Et dans notre cas on a dans la corrélation canonique est proche de 1 d'un montant de 0,636.

- Le Lambda de Wilks :

Lambda de Wilks				
Test de la ou des fonctions	Lambda de Wilks	Khi-carré	ddl	Sig.
1	.596	153.575	3	.000

Plus la valeur de Wilks est faible et sa significative tend vers 0 , plus le modèle est bon. On a Sig. = 0.00 donc il tend vers 0.

- Le Lambda de Wilks :

3^{ème} Etape : On Vérifie la pouvoir discriminant des axes

Coefficients des fonctions discriminantes canoniques standardisées	
	Fonction 1
TEMP_MAX	-3.858
TEMP_MIN	4.120
PLUVIOMETRIE	-.206

Ce tableau permet d'obtenir la fonction discriminante, dans notre cas la fonction est :

$$Y = -3.858 \times (\text{TEMP_MAX}) + 4.12 \times (\text{TEMP_MIN}) + 0 \times (\text{HUMIDITE}) - 0.206 \times (\text{PLUVIOMETRIE})$$

Matrice de structure	
	Fonction 1
TEMP_MIN	.362
PLUVIOMETRIE	.249
HUMIDITE ^a	.167
TEMP_MAX	.114
Les corrélations intragroupes combinés entre les variables discriminantes et les variables des fonctions canoniques standardisées sont ordonnées par la taille absolue des corrélations à l'intérieur de la fonction.	
a. Cette variable n'est pas utilisée dans l'analyse.	

La table «Matrice de structure» vérifie aussi que «TEMP_MIN» se sont les variables qui separent le mieux entre les groupes.

4^{eme} Etape : La qualité de la représentation du modèle

Résultats du classement ^{a,c}					
		Appartenance au groupe prévu			Total
		plage	non	oui	
Original	Effectif	non	130	38	168
		oui	32	100	132
	%	non	77.4	22.6	100.0
		oui	24.2	75.8	100.0
Validé-croisé ^b	Effectif	non	130	38	168
		oui	32	100	132
	%	non	77.4	22.6	100.0
		oui	24.2	75.8	100.0

a. 76.7% des observations originales sont classées correctement.

b. La validation croisée n'est effectuée que pour les observations de l'analyse. Dans la validation croisée, chaque observation est classée par les fonctions dérivées de toutes les autres observations.

c. 76.7% des observations validées-croisées sont classées correctement.

Ce tableau est utilisé pour évaluer dans quelle mesure la fonction discriminante fonctionne et si elle fonctionne aussi bien pour chaque groupe de la variable dépendante. Ici, il classe pour le groupe 1 (les villes qui n'ont pas la plage) 130 ville ont été bien reclassées (77,4%) grâce a la fonction discriminante et 38 villes on été mal reclassées (22.6%) .

Et pour le groupe 2 (les villes qui ont la plage) on observe 32 villes ont été mal reclassées (24.2%) et 100 villes ont été bien reclassées (75.8).