DR 544046

CD - 1H 5996

Académie de Montpellier

Université Montpellier I

Université Montpellier II
Sciences et Techniques du Languedoc

Ecole Nationale Supérieure Agronomique
Montpellier

CIRAD

Interpolation de la pluie sur l'île de la Réunion

Rapport de stage de DEA de Biostatistique
Option : Méthodes et Outils pour la Biostatistique

par :

Mohamed CHAOUCH

sous la direction de :

Eric GOZÉ

soutenu le 20 juin 2005 devant le jury composé de

C. ABRAHAM	Maître de conférences, ENSA.M
G. DUCHARME	Professeur, Université, Montpellier II
J.F. DURAND	Professeur, Université, Montpellier II
N. MOLINARI	Maître de conférences, Montpellier I

CIRAD-DIST
Unité bibliothèque
Lavalette

000101444

Avant-propos

Ce stage a été effectué au Cirad (Centre de coopération internationale en recherche agronomique pour le développement) sous la direction de M. Eric Gozé.

Ce stage, outre son caractère pédagogique, se propose d'étudier la faisabilité et la comparaison de deux méthodes, la régression sur composantes principales et la PLS, pour interpoler la moyenne mensuelle des pluies journalières en tout point sur l'île de la Réunion.

Remerciements

J'exprime mes sincères remerciements

à mon maître de stage Monsieur Eric Gozé, pour son encadrement, ses nombreux conseils et sa grande disponibilité dont il a fait montre tout au long de ce stage

à Monsieur Philippe Letourmy, responsable de l'Unité de recherche aide à la décision et Biostatistique, de m'avoir accepté dans son unité et pour ses nombreux conseils durant ce stage

à Monsieur Gilles Ducharme, responsable du DEA Biostatistique, de m'avoir accepté dans ce DEA et pour les efforts qu'il a fait durant cette année

à Monsieur Ibnou Dieng, doctorant, pour sa collaboration et tous ses conseils tout au long de ce stage

à Mademoiselle Sabine Laurent, doctorante, pour ses conseils

à toute l'équipe de l'unité de recherche: aide à la décision et Biostatistique pour la convivialité

Table des figures

2.1 Cartographie en courbe de niveau de la pluviomètrie en mètres/an 5

2.2 Cartographie de la moyenne mensuelle des pluies journalières des postes pour les 12 mois de l'année 1996 : les mois sont arrangés par colonnes par exemple la première colonne contient Janvier, Février et Mars. 7

3.1 Démarche de construction de la matrice des variables explicatives 9

3.2 Projection du nuage des individus sur le plan factoriel 1-2, ACP centrée sur les altitudes des points voisins des stations 11

3.3 Premier vecteur propre, de l'ACP centrée sur les altitudes des points voisins des stations, cartographié en courbes de niveaux 13

3.4 Premier vecteur propre, de l'ACP bi-centrée sur les altitudes des points voisins des stations, cartographié en courbes de niveaux 14

3.5 nuage des points des 1000 points du MNT (a) et premier vecteur propre cartographié en courbes de niveaux (b) 15

3.6 Organigramme de la démarche à suivre dans la modélisation de la pluie avec la RCP 18

3.7 Distribution de pluie pour les mois Janvier-Décembre-Mai-Juillet 19

3.8 QQ-plot des résidus de la régression de la moyenne mensuelle des pluies journalières sur les composantes principales et la moyenne des altitudes des points au tour des stations- mois de Janvier 23

3.9 Evolution des résidus de la régression en fonction des valeurs prédites- mois de Janvier 24

3.10 Pluie en fonction des pluie ajustées- mois de Janvier 26

3.11 Variation de la pente de la droite de régression des valeurs absolues des résidus sur les valeurs prédites de la pluie-mois de Janvier 28

3.12 La distribution de la pluie après transformation-mois de Janvier 29

3.13 Les résidus de la régression en fonction des valeurs ajustées de pluie-mois de Janvier 30

3.14 Les résidus de la régression en fonction des valeurs ajustées de pluie-mois de mai 34

4.1 Critères d'aide à la décision pour le choix du nombre de composantes - Mois de Janvier 38

4.2 Validation croisée : observations vs prédictions par une régression PLS à 11 composantes 41

4.3 Comparaison entre les représentations pseudo-factorielles et factorielles pour le plan (1,2) des individus. Visualisation des individus dans les régressions de t* sur t pour les axes 1 et 2- Mois de Janvier 43

4.4 Comparaisons du PRESS de la régression PLS à celui de la Régression sur composantes principales- Mois de Janvier 44

4.5 Comparaisons du critère d'ajustement des données de la régression PLS à celui de la Régression sur composantes principales- Mois de Janvier 45

Liste des tableaux

3.1 pourcentage de variance résumé par chaque composante principale 20

3.2 Modèle de régression de la moyenne mensuelle des pluies journalière du mois de Janvier sur les composantes principales + la moyenne des altitudes des points du masque autour des stations 22

3.3 Test de Glejser : test de significativité du coefficient de régression de la valeur absolue des résidus sur les valeurs ajustées de la pluie 25

3.4 Modèle de régression de la transformée de la moyenne mensuelle des pluies journalières du mois de Janvier sur les composantes principales + la moyenne des altitudes des points du masque autour des stations 29

3.5 Test de Glejser sur les résidus après transformation de la variable pluie 30

3.6 Modèle de régression de la transformée de la moyenne mensuelle des pluies journalières du mois de Janvier sur les composantes principales + moyenne des altitudes des points du masque autour des stations 31

3.7 Modèle final de la régression descendante de la transformée de la moyenne mensuelle des pluies journalières du mois de Janvier sur les composantes principales + moyenne des altitudes des points du masque autour des stations 32

Glossaire

ACP	Analyse en composantes principales
RCP	Régression sur Composantes Principales
PLS	Partial Least Squares
CIRAD	Centre de coopération internationale en recherche agronomique pour le développement
CA	Cultures Annuelles
SIG	Système d'Information Géographique
MNT	Modèle Numérique de Terrain

Table des matières

Avant-propos i

Remerciements ii

Glossaire vi

1 Introduction 1

1.1 Problématique 2

1.2 But du stage 2

2 Présentation de la pluviométrie à la Réunion 3

2.1 Pluviométrie moyenne annuelle 4

2.2 Pluviométrie mensuelle et saisonnière 4

2.3 Données disponibles 5

2.3.1 Le Modèle Numérique de Terrain (MNT) 5

2.3.2 Les données pluviométriques 6

2.3.3 Sélection de l'année 6

3 Modélisation de la pluie en fonction du relief : Régression sur composantes principales 8

3.1 Caractérisation du relief local autour des postes pluviométriques par l'altitude des points d'un voisinage carré 8

3.2 ACP centrée sur les altitudes des points voisins des stations météo 10

3.3 ACP bi-centrée sur les altitudes des points voisins des stations météo 13

3.4 ACP bi-centrée sur l'échantillon aléatoire de 1000 points du MNT 14

3.5 Comparaison des deux ACP bi-centrées et conclusion 15

3.6 Régression de la pluie sur les composantes principales : Elaboration du modèle sur la base des données de mois de Janvier 17

3.6.1 Etude préliminaire de la variable pluie 19

3.6.2 Sélection du nombre des composantes à retenir de l'ACP bi-centrée faite sur l'échantillon de 1000 points du MNT . 20

3.6.3 Modèle de régression multiple de la pluie sur composantes principales 20

3.6.4 Analyse des résidus du modèle de régression 22

3.6.4.1 Test de normalité des résidus 22

3.6.4.2 Test d'homogénéité de la variance 24

3.6.4.3 Vérification visuelle de l'hypothèse de linéarité du modèle 25

3.7 Transformation de la variable pluie et régression sur les composantes principales : élaboration du modèle sur la base du mois de Janvier 26

3.7.1 Les avantages de la transformation Box-Cox apportés à la régression 27

3.7.2 Résultats du modèle après transformation Log 29

3.7.3 Sélection des composantes principales explicatives par régression descendante 31

3.7.4 Les limites de la transformation Box-Cox pour les mois les moins arrosés 33

4 Régression PLS **35**

4.1 Rappel sur la PLS1 35

4.2 Régression PLS sur les données du mois de Janvier 38

4.2.1 Choix du nombre de composantes PLS 38

4.2.2 Le modèle PLS 41

4.2.3 Représentation des individus 42

4.3 Conclusion 44

5 Conclusion et Perspectives **46**

Bibliographie **48**

Chapitre 1

Introduction

Le laboratoire d'accueil

Le Centre de coopération internationale en recherche agronomique pour le développement (CIRAD) est un organisme français de recherche agronomique au service du développement durable des pays tropicaux et subtropicaux. Ses activités de recherche s'appliquent aux domaines des sciences du vivant et des sciences sociales appliquées à l'agriculture, à la forêt, à l'élevage, à la gestion des ressources naturelles, à l'agroalimentaire, aux écosystèmes et aux sociétés du Sud.
L'Unité de Recherche "Aide à la décision et biostatistique" appartient au département Cultures annuelles (Ca) du Cirad. Ses activités de travail concernent l'informatique et les mathématiques. L'activité de Biométrie a pour objectif d'améliorer l'utilisation des techniques statistiques. L'activité d'informatique touche la gestion de projets, la programmation d'applications spécifiques sous windows, les Systèmes d'Information Géographique (SIG), les bases de données et l'analyse d'image.

1.1 Problématique

L'île de la Réunion présente un relief marqué et des paysages variés. La canne à sucre en est la principale source de revenus. Elle fait vivre de nombreux petits planteurs et entretient une industrie de transformation en sucre. La prévision de la richesse des cannes récoltées est un enjeu important pour cette industrie de façon à en organiser au mieux la collecte.
La production étant en bonne partie tributaire des conditions climatiques, l'estimation hebdomadaire de la pluie, du vent et de la température en tout point cultivé est nécessaire à la prévision de la récolte.
Or ces variables ne sont mesurées au mieux qu'en 90 points. Alors que la température et la vitesse du vent s'interpolent facilement, l'estimation de la pluie en tout point de l'île est délicate, notamment parce que les méthodes classiques d'interpolation telles que le krigeage ne prennent pas en compte le relief de manière satisfaisante.
Des mesures d'altitude sont disponibles sur une grille au pas de 25 m x 25 m, formant un MNT (Modèle Numérique de Terrain). Elles rendent possible la prise en compte du relief.
Une tentative de régression sur des variables calculées à l'aide du MNT et caractérisant divers aspects de ce relief (pente, exposition, ...) n'a pas permis la prédiction de la pluie moyenne mensuelle avec une précision suffisante.
Dans des conditions similaires de relief volcanique, sur l'île de Tahiti, une approche plus empirique de régression des pluies extrêmes sur des composantes principales issues d'un MNT donne des résultats meilleurs que le krigeage, ce qui montre l'intérêt de la prise en compte du relief (Wotling et al., 2000).

1.2 But du stage

Le but de ce stage est à la fois l'étude de faisabilité et la comparaison des performances de deux méthodes, pour estimer en tout point de l'île les pluies hebdomadaires en tenant compte du relief. Ces deux méthodes sont la Régression sur Composantes Principales (RCP) et la Partial Least Squares (PLS). Mais pour commencer, on fait des essais sur quelque chose de plus facile : pluies mensuelles moyennes car elles sont plus régulières et correctement relevées.

Chapitre 2

Présentation de la pluviométrie à la Réunion

Le caractère tropical et montagneux de l'île de la Réunion détermine un climat contrasté, dont les précipitations sont le phénomène météorologique le plus remarquable. L'île possède tous les records mondiaux de pluies pour les périodes comprises entre 12 heures et quinze jours. Le fait essentiel est la grande variabilité spatio-temporelle de ces précipitations. Sur un mois, plusieurs mètres d'eau peuvent tomber sur la côte Est alors que pas une goutte ne tombera sur la côte Ouest. Par ailleurs, pour un lieu donné, la pluie peut être absente pendant plusieurs mois et tomber ensuite en abondance sur une courte période lors du passage d'une dépression ou d'un cyclone. Les précipitations que connaît La Réunion peuvent être classées en deux grandes catégories :

- celles d'origine advective, provoquées par le passage de perturbations nuageuses venues du large : depuis les modestes ondes circulant dans le courant d'alizé jusqu'aux cyclones tropicaux les plus violents, en passant par tout l'éventail des perturbations tropicales et des remontées d'origine polaire.
- celles liées aux phénomènes convectifs d'évolution diurne qui, tout en obéissant à un cycle relativement immuable, seront d'autant plus marquées que la masse d'air sera plus humide et instable.

2.1 Pluviométrie moyenne annuelle

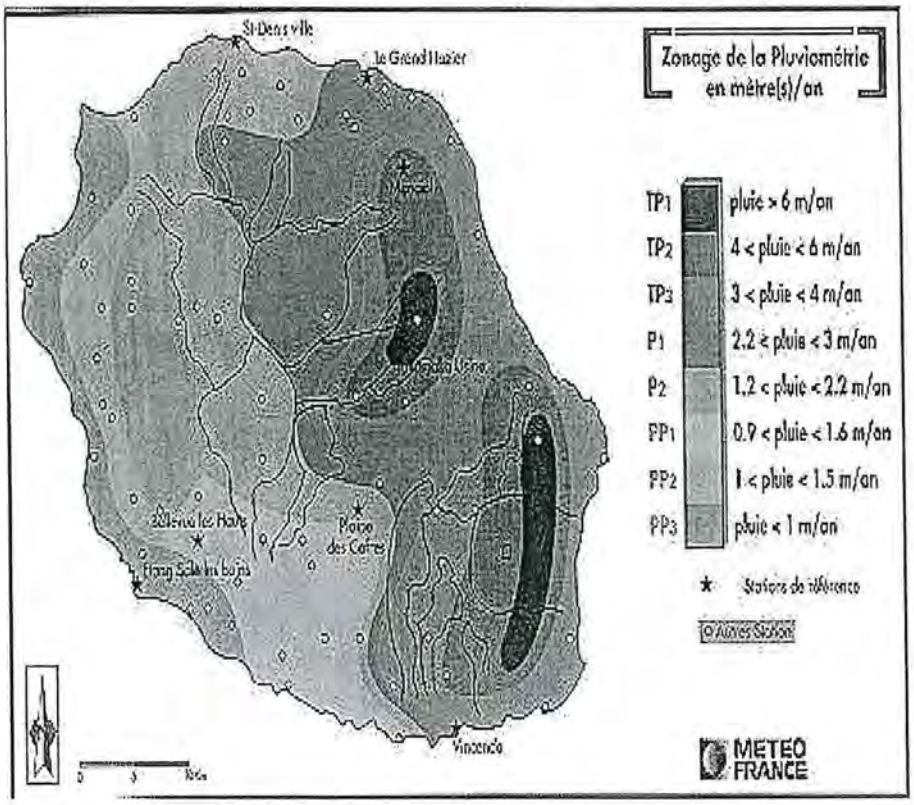
La carte des isohyètes annuels montre une grande dissymétrie entre l'est et l'ouest de La Réunion. A l'ouest les précipitations sont peu abondantes. En revanche, plus on se décale vers l'est, plus la moyenne des pluies augmente, jusqu'à atteindre des valeurs dépassant 10 mètres par an, ce qui est tout à fait exceptionnel.
A l'ouest, c'est sur le littoral qu'il pleut le moins, en particulier autour de St-Gilles-les-Bains (525 mm/an). Par contre la pluviosité augmente au fur et à mesure que l'on s'élève.
On note ainsi, par exemple, un petit noyau pluvieux au niveau de la plaine des Makes (2 m/an). A l'est, deux zones sont particulièrement arrosées : la région de Takamaka (7m/an) et tout l'est du volcan où l'on enregistre les valeurs les plus élevées (plus de 11 m/an sur les Hauts de Ste-Rose).
Les hauts reliefs de l'île, massifs du Piton des Neiges et du Piton de la Fournaise, sont la cause de cette dissymétrie est-ouest. Qu'elles soient d'origine adventive ou d'évolution diurne, les précipitations tendent à se concentrer sur les régions directement exposées à l'humidité océane apportée par la circulation générale de secteur est.
On parle de région "au vent". A l'inverse, les régions "sous le vent", qui bénéficient de l'abri du relief sont beaucoup plus sèches. Pour un secteur donné, quand on s'élève du littoral au sommet de l'île, les foyers de précipitations les plus intenses sur l'île, se rencontrent aux altitudes intermédiaires, entre 1000 et 2000 mètres : Plaines des Makes, Petite-France, Takamaka, Hauts de Sainte-Rose et Forêt de Saint-Philippe...

2.2 Pluviométrie mensuelle et saisonnière

Saison des pluies : la saison des pluies peut être définie entre janvier et mars. Sur l'ensemble de l'île, le mois de février est de loin le plus arrosé. Il pleut en un mois plus de 1,5 m à la station Hauts de Ste-Rose.
Sur l'ouest, il pleut davantage au mois de janvier qu'au mois de mars. C'est l'inverse sur l'est. Les mois de décembre et d'avril sont eux aussi bien arrosés. Ils représentent la transition entre les deux périodes "humide" et "sèche". Saison sèche : elle débute au mois de mai pour s'achever au mois de novembre.

A l'ouest la pluviomètrie est quasiment nulle en particulier sur la côte (4,2 mm en juillet à St-Gilles-les-bains). Dans l'est, même si l'on parle de saison sèche, il pleut tout de même plus de 700 mm sur le volcan au cours du mois le moins arrosé, c'est à dire plus qu'à Paris en une année.

FIG. 2.1 – Cartographie en courbe de niveau de la pluviomètrie en mètres/an



2.3 Données disponibles

2.3.1 Le Modèle Numérique de Terrain (MNT)

Le Modèle Numérique de Terrain (MNT) est une série de relevés d'altitudes effectués à chaque noeud d'une grille de 25m de côté. Le modèle numérique de terrain permet d'établir facilement des courbes de niveaux et de connaître l'altitude de chaque point du territoire.

2.3.2 Les données pluviométriques

Notre but est d'interpoler les pluies hebdomadaires, mais on va commencer par les pluies mensuelles. Elles devraient être plus faciles à interpoler car généralement la régularité spatiale d'un cumul de pluie (ou d'une moyenne) s'améliore quand la moyenne porte sur des intervalles de temps plus grands.

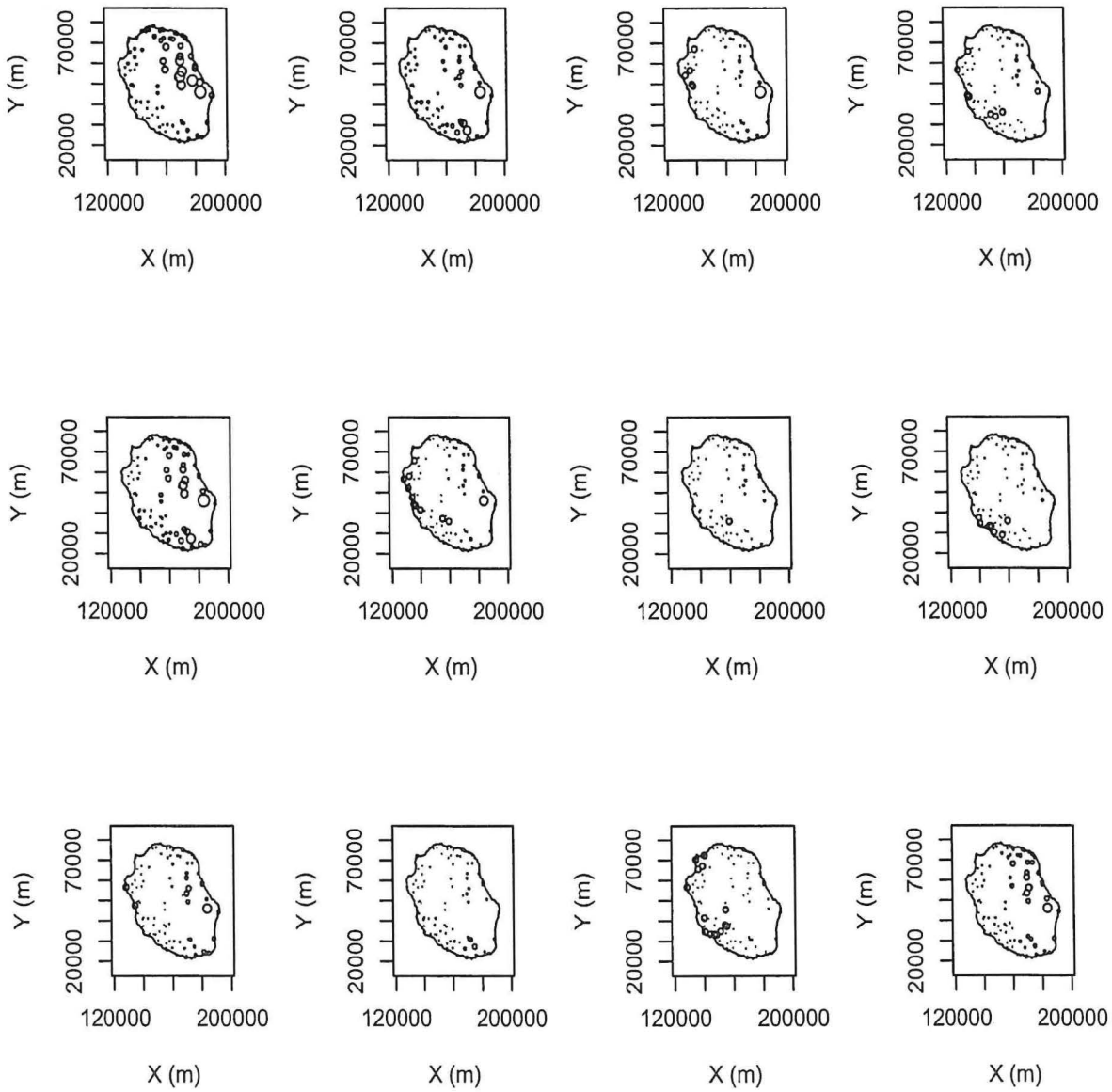
Nous disposons des mesures journalières de pluie pour chaque poste météo, de l'année 1990 jusqu'à 2001. Nous avons remarqué que le fichier des pluies journalières contenait de nombreuses données manquantes qui peuvent être dues à divers causes tel que : le pluviomètre est en panne, la mesure n'a pas été enregistrée,...
Donc se baser sur un tel fichier pour interpoler la pluie peut engendrer des résultats biaisés, pour cela nous proposons, pour une année et un mois fixé, de ne conserver que les postes dont les relevés sont complets pour le mois considéré. Ceci permet d'avoir des fichiers sans données manquantes au prix de la perte d'un certain nombre de postes.
Par la suite nous viendrons à l'étape de sélection de l'année sur laquelle nous allons nous baser pour mettre au point la méthode, c'est à dire l'année sur laquelle nous allons nous baser pour interpoler la pluie en tout point de l'île.

2.3.3 Sélection de l'année

Après examen des années (de 1990 à 2001) nous retiendrons l'année 1996 car elle contient plus de stations par mois que les autres années (et donc moins de données manquantes), en plus les quantités de pluie mesurées sont proches de la situation ordinaire et ne représentent aucun phénomène spécial tel qu'un cyclone ou autre.

La figure (2.2) représente la cartographie de la moyenne mensuelle des pluies journalières pour les 12 mois de 1996 (une colonne représente un trimestre). La quantité de pluie mesurée sur chaque poste est représentée par un cercle dont le diamètre varie selon la quantité. Un poste apparaît sur la carte s'il n'y a aucune donnée manquante dans le mois. On remarque que cette observation est bien conforme au climat présenté ci-dessus (voir présentation de la pluviométrie sur la Réunion). Par la suite nous modélisons la moyenne mensuelle des pluies journalières, en fonction du relief, pour chaque mois de l'année 1996, pour pouvoir interpoler la moyenne mensuelle des pluies journalières en tout point de l'île sur les 12 mois de l'année.

FIG. 2.2 – Cartographie de la moyenne mensuelle des pluies journalières des postes pour les 12 mois de l'année 1996 : les mois sont arrangés par colonnes par exemple la première colonne contient Janvier, Février et Mars.



Chapitre 3

Modélisation de la pluie en fonction du relief : Régression sur composantes principales

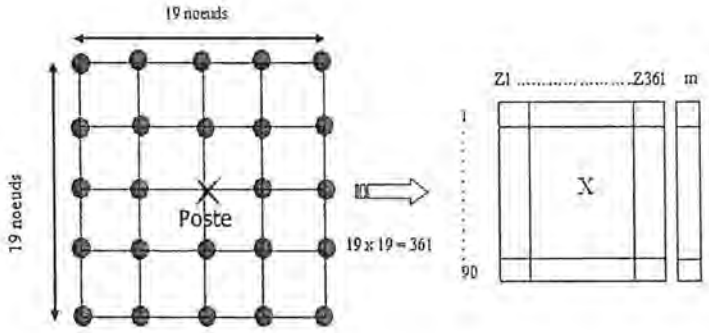
Dans le paragraphe suivant on explique la méthode utilisée pour caractériser le relief local autour des postes pluviométriques par l'altitude des points d'un voisinage carré ainsi que les étapes de construction de la matrice des variables explicatives de la moyenne mensuelle des pluies journalières.

3.1 Caractérisation du relief local autour des postes pluviométriques par l'altitude des points d'un voisinage carré

Le relief local autour d'un poste météo est caractérisé par l'altitude des points d'un voisinage carré autour de la station.

Comment est-il-choisi ce voisinage carré ?

On définit autour de la station météo une grille de 19x19 points (dans ce qui suit on appellera cette grille masque), de telle sorte que la station soit au centre, et si elle ne se coïncide pas à un noeud on l'amène au noeud le plus proche.



avec :

$Z_1^i, \ldots, Z_{361}^i$: Les altitudes des 361 points voisins de la station i $\quad i \in [1, 90]$

$m^i = \frac{1}{361} \sum_{j=1}^{361} Z_j^i$: La moyenne des altitudes des points autour de la station i.

Ces variables permettent de décrire le relief autour de n'importe quelle station météo et elles seront des variables explicatives de la pluie.

Pour modéliser la moyenne mensuelle des pluies journalières on envisage un modèle linéaire de type :

$$Y = X\beta + \varepsilon \tag{3.1}$$

mais on ne peut pas utiliser directement une régression linéaire de ce type car le nombre de variables (361) est très supérieur au nombre d'individus (90) et que la matrice X dans notre cas n'est pas de plein rang et par conséquent elle n'est pas inversible(présence de multicolinéaritée entre les variables). Le calcul de la matrice de variance-covariance montre que certaines variables sont fortement corrélées entre elles, donc dans une telle situation on a le choix entre la Régression sur

Composantes Principales (RCP) ou la régression PLS.

Commençons d'abord par la RCP. Une étude préliminaire (Gigliobianco, 2003)a montré que les vecteurs propres sont proches pour une ACP faite sur la totalité des points du MNT et sur un échantillon aléatoire de 275 points.

La question qui se pose c'est de savoir si on fait l'ACP sur les seuls points où il y a des stations, l'échantillon d'apprentissage, ou bien sur tous les points du MNT, la population où nous devons fournir des prédictions.

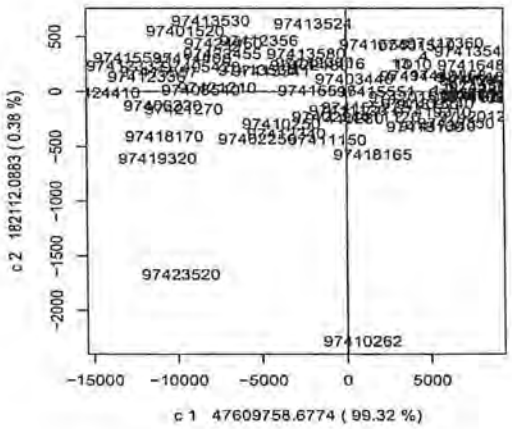
La première option offre l'avantage d'orthogonalité des composantes sur l'échantillon d'apprentissage, ce qui fait le principal intérêt de l'ACP avant la régression, la deuxième option présente un autre avantage, pour le cas où les stations ne seraient pas représentatives de l'ensemble de la population. En effet, elle permet de ne pas restreindre le sous espace résumé par l'ACP à celui décrit par les stations.

Cela laisse la possibilité de mettre en évidence sur les composantes principales des types de relief locaux que l'on n'aurait peu ou pas rencontrés autour des stations. Ces situations d'extrapolation doivent alors se caractériser par une estimation de l'erreur-type de prédiction plus élevée que pour les situations d'interpolation.
Pour comparer les deux ACP nous mettons à l'épreuve cette hypothèse :***"les stations météo constituent un échantillon représentatif du reste des points du MNT et par conséquent elles peuvent être utilisées pour interpoler la pluie en tout point de l'île."***
Pour aller plus loin dans notre étude nous allons vérifier cette hypothèse mise sur les stations météo, pour cela on commence par faire une ACP centrée sur le tableau de données X composé de 90 lignes representant les mesures des altitudes des 361 points voisins de chaque station.

3.2 ACP centrée sur les altitudes des points voisins des stations météo

On choisit de faire une ACP centrée, non réduite car les variables sont de dimension homogène c'est à dire elles ont la même unité de mesure (le mètre).

FIG. 3.2 – Projection du nuage des individus sur le plan factoriel 1-2, ACP centrée sur les altitudes des points voisins des stations



- La figure (3.2) montre que la première composante principale résume 99.32% de la variance totale et le reste est réparti entre les composantes suivantes : 0.38 % pour la deuxième, 0.17 % pour la troisième et 0.03 % pour la quatrième.
- A cause de l'importance particulière de la première composante principale, cette ACP peut poser des problèmes pour le calcul des composantes suivantes.
- La première composante est presque égale à la moyenne des altitudes de tous les points du masque.

En effet :

$$c_1 = XQa_1 = \sum_{i=1}^{361} a^i X^i \tag{3.2}$$

Une composante principale est une combinaison linéaire des mailles du voisinage de chaque station et les coefficients de cette combinaison linéaire caractérisent chaque voisin, et ce sont les éléments du vecteur propre. les représenter sous forme de courbes de niveaux facilite leur interprétation.

Or les composantes du premier vecteur propre sont presque égales entre elles,elles varient entre -361 et -366(voir figure 3.3).

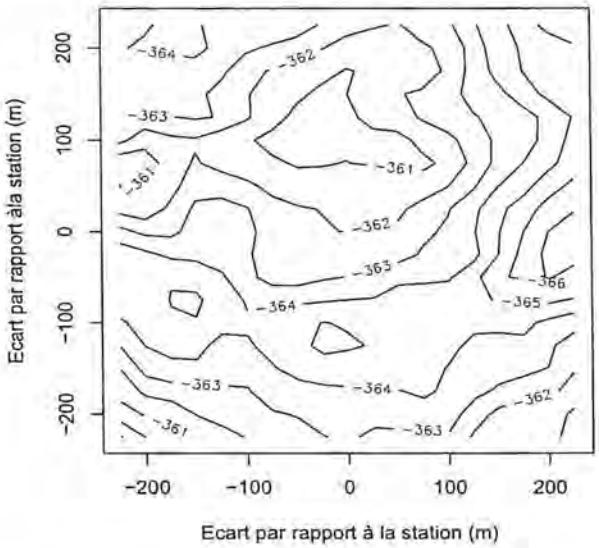
Donc $a^i \simeq a^j = \alpha, \qquad \forall i \neq j\,.$

Avec α : coefficient de proportionnalité

par la suite,

$c_1 = \alpha \sum_{i=1}^{361} X^i$= moyenne des altitudes de tous les points du masque.

FIG. 3.3 – Premier vecteur propre, de l'ACP centrée sur les altitudes des points voisins des stations, cartographié en courbes de niveaux



On a montré que la première composante résume plus que 99% de la variance et qu'elle est quasiment proportionnelle à la moyenne des altitudes de tous les points du masque. Pour cette raison, nous allons éliminer cette moyenne en centrant en ligne la matrice des variables explicatives. Cela nous permettra d'obtenir une ACP où les valeurs propres seront moins contrastées.

3.3 ACP bi-centrée sur les altitudes des points voisins des stations météo

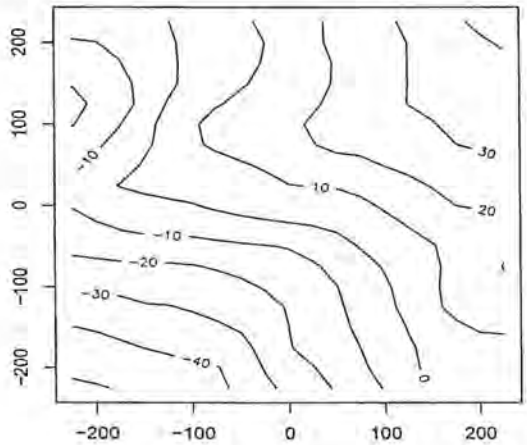
La représentation en courbes de niveaux du premier vecteur propre de l'ACP bi-centrée [1] faite sur les altitudes des points voisins des stations météo reste encore difficile à interpréter (Fig. 3.4). On ne retrouve pas les résultats de Wotling (2000) ou de Gigliobianco (2003), où l'ACP était fait sur les points du MNT. Donc le bicentrage élimine l'effet taille.

On peut aller plus loin dans la vérification de l'hypothèse imposée sur les stations météo, pour cela nous allons refaire une ACP centrée mais cette fois ci sur

[1] La matrice des variables explicatives X est centrée à la fois en ligne et en colonne

CIRAD-DIST
Unité bibliothèque
Lavalette

FIG. 3.4 – Premier vecteur propre, de l'ACP bi-centrée sur les altitudes des points voisins des stations, cartographié en courbes de niveaux



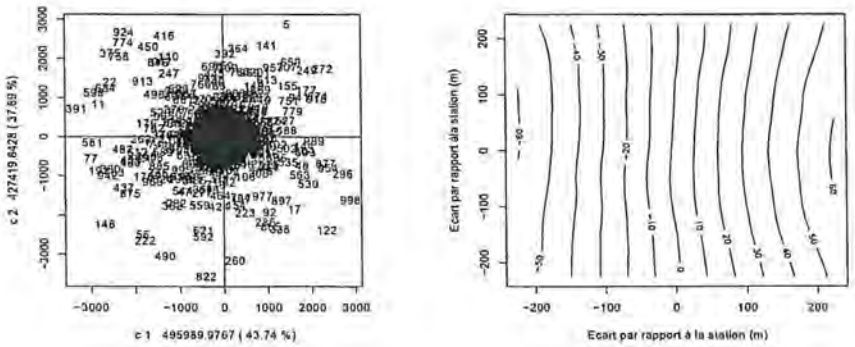
un échantillon aléatoire de 1000 points du MNT et puis comparer les courbes de niveau du premier vecteur propre issu respectivement de l'ACP bi-centrée faite sur l'échantillon de 1000 points du MNT et celle de l'ACP bi-centrée faite sur les stations météo.

3.4 ACP bi-centrée sur l'échantillon aléatoire de 1000 points du MNT

Contrairement à l'ACP sur les stations météo, la figure du nuage des points sur le plan factoriel (1-2) (Fig 3.5(a)) ne montre rien d'organisé , la première composante principale résume 43.74 % de la variance totale et la deuxième composante 37.69 %.

La représentation en courbes de niveaux du premier vecteur propre (Fig 3.5(b)) montre que la première composante principale est proportionnelle à la pente moyenne dans le sens Ouest-Est. Elle est conforme à l'étude de Gigliobianco (2003) et Wotling (2000).

FIG. 3.5 – nuage des points des 1000 points du MNT (a) et premier vecteur propre cartographié en courbes de niveaux (b)



(a) (b)

3.5 Comparaison des deux ACP bi-centrées et conclusion

- Le centrage des variables par rapport aux lignes est important car il permet d'obtenir des différences par rapport à l'altitude moyenne du voisinage (on parle d'altitude relative) ; cela permet de prendre en compte uniquement le relief local dans l'explication de la pluie.
  D'autre part, le tableau obtenu a la propriété d'être centré en ligne comme en colonne , ce qui permet de remplacer l'ACP sur ce tableau par une ACP sur le tableau transposé, ce qui peut alléger les calculs s'il y a plus de variables que d'individus et ce qui peut être fort utile par la suite si on veut effectuer une ACP sur un voisinage plus grand, donc plus de variables.
- Le relief local autour des stations n'est pas représentatif du relief local autour de n'importe quel point de l'île. Ceci car la courbe de niveau du premier axe principal de l'ACP bi-centrée faite sur les stations météo est différente de celle faite sur un échantillon aléatoire de 1000 points du MNT.

En conclusion :

- Dans tout ce qui suit nous allons travailler avec des tableaux de données bi-centrés.

- Nous proposons de faire une ACP bi-centrée sur l'échantillon aléatoire de 1000 points du MNT car ceci permet d'avoir des composantes principales qui résument le relief, autour de n'importe quel point du MNT, à partir des altitudes de ses 361 points voisins. Puis on introduit les postes météo dans l'ACP comme individus supplémentaires et on calcule les composantes supplémentaires via la formule :

$$(C_{su})_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X_{su} Q a_\alpha \tag{3.3}$$

  avec :
  - $(C_{su})_\alpha$: la α^e composante supplémentaire
  - λ_α : la α^e valeur propre de l'opérateur Escoufier VQ.
  - Q : une métrique sur l'espace des variables (dans notre cas Q = I_{361})
  - a_α : le α^e vecteur propre de la matrice de variance-covariance V.

Pourquoi une ACP sur un échantillon au hasard plutôt que sur les stations ?

On décide de faire une ACP sur un échantillon au hasard plutôt que sur les stations car :

- Les stations ne sont pas représentatives de l'îles, si on en juge par le relief local. En effet, les ACP sur le relief local ne sont pas identiques entre population des stations et échantillon pris au hasard dans le MNT.
- Il y a des configurations de relief local qui seraient sous-représentées ou pas représentées du tout sur la population de stations. Ces directions non explorées dans l'espace des individus seraient évitées par les premières composantes principales, qui s'orientent dans les directions à plus forte densité de points.
- Le filtrage par élimination des dernières composantes éliminerait donc toute information sur ces directions non explorées, ces configurations de relief local non trouvées autour des stations.

- La régression sur ces composantes principales ne permettrait donc pas d'être averti de ces situations d'extrapolation. En d'autres termes, le calcul d'incertitude sur la prédiction ne pourrait pas prendre en compte cette situation d'extrapolation. Pour éviter ces inconvénients, nous choisissons de faire l'ACP sur un échantillon du MNT, et non sur les stations qui ne sont gardées que comme individus supplémentaires.

3.6 Régression de la pluie sur les composantes principales : Elaboration du modèle sur la base des données de mois de Janvier

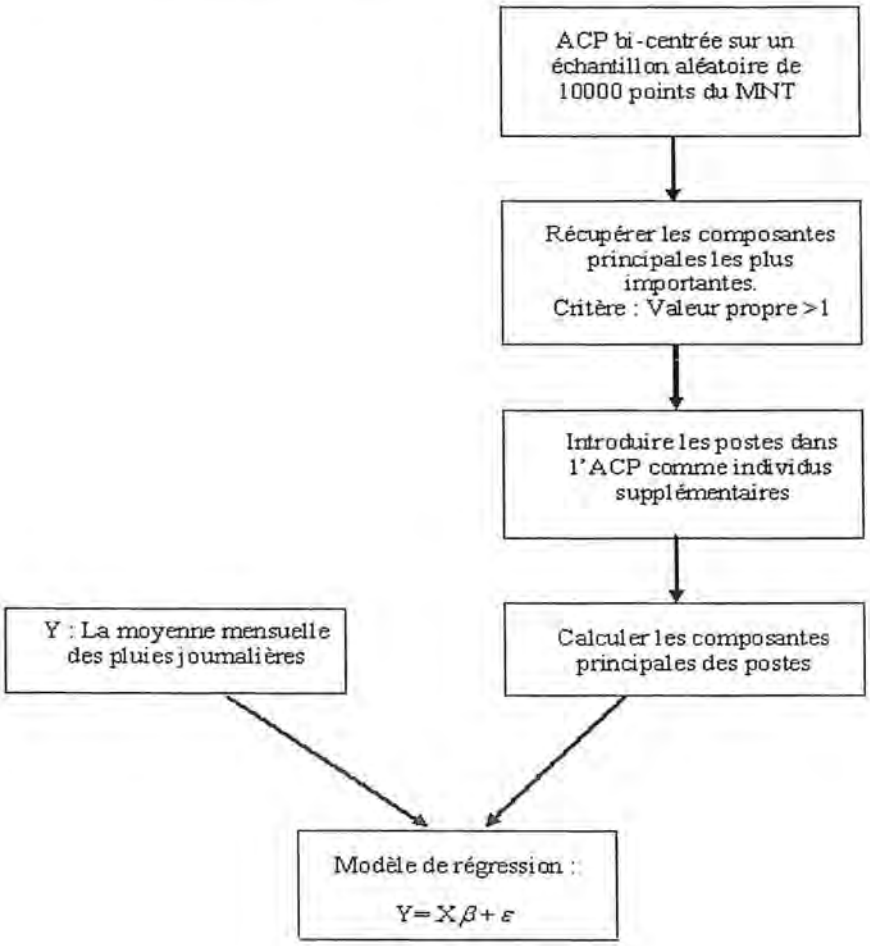
Dans cette partie nous mettons en oeuvre la méthode de régression de la moyenne mensuelle de la pluie journalière (Y) sur les composantes principales supplémentaires

.

On propose de construire le modèle sur la base des données du mois de Janvier et puis essayer d'appliquer ce modèle sur les autre mois et juger par la suite sa performance.

On a choisi le mois de Janvier car ce mois est caractérisé par des mesures de précipitation assez importantes ce qui rend plus facile l'élaboration du modèle de régression et l'interprétation des résidus.
Ci-dessous un organigramme qui explique la démarche suite pour construire le modèle de régression.

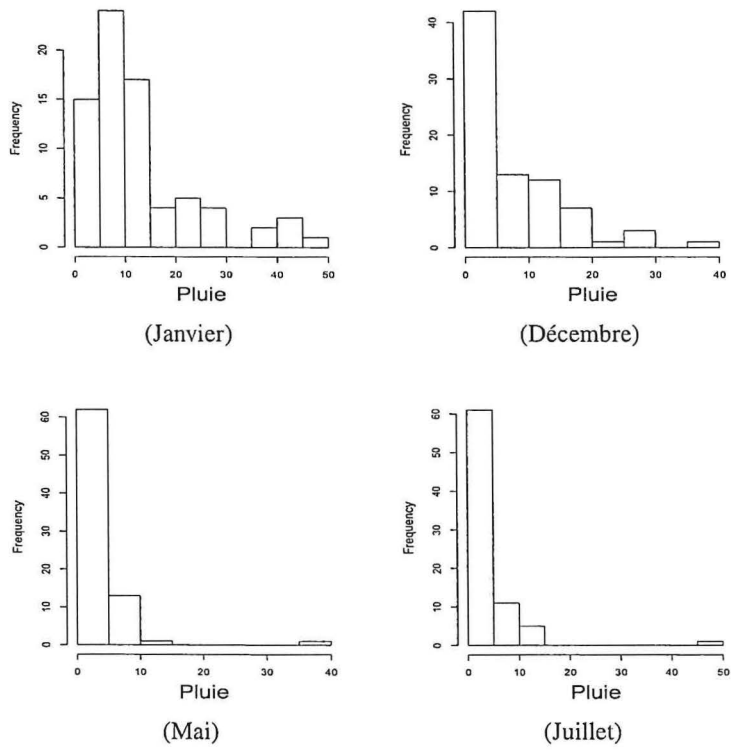
FIG. 3.6 – Organigramme de la démarche à suivre dans la modélisation de la pluie avec la RCP



3.6.1 Etude préliminaire de la variable pluie

L'examen des histogrammes de distribution de la moyenne mensuelle de la pluie journalière (Fig. 3.7) pour les mois Janvier, Décembre,Mai et Juillet montre une forte dissymétrie de cette variable.
Cette dissymétrie peut être due au fait que le vent dominant vient de l'océan Indien et traverse la côte Nord de l'île et s'arrête au niveau de la chaîne de montagne, dont les altitudes atteignent 3000 m, qui traverse le centre de l'île et donc la zone Nord est plus arrosée que la zone Sud : on pourrait donc avoir un mélange de distributions.

FIG. 3.7 – Distribution de pluie pour les mois Janvier-Décembre-Mai-Juillet



3.6.2 Sélection du nombre des composantes à retenir de l'ACP bi-centrée faite sur l'échantillon de 1000 points du MNT

Pour fixer le nombre de composantes principales à retenir nous avons deux critères : la règle de bâton brisé de l'histogramme des valeurs propres ou bien sélectionner les composantes relatives aux valeurs propres supérieures à 1.
En utilisant le deuxième critère et en se basant sur les résultats du tableau 4.10, nous nous arrêtons aux sept premières composantes principales.

indice	valeur propre	inertie	cumul d'inertie
1	**49.59**	**43.74**	**43.74**
2	**42.74**	**37.69**	**81.43**
3	5.38	4.75	86.18
4	4.62	4.08	90.26
5	3.01	2.66	92.92
6	1.41	1.25	94.17
7	**1.14**	**1.01**	**95.18**
8	0.75	0.67	95.85
9	0.68	0.60	96.45

TAB. 3.1 – pourcentage de variance résumé par chaque composante principale

3.6.3 Modèle de régression multiple de la pluie sur composantes principales

Dans cette partie nous allons chercher, dans l'espace des composantes principales, la combinaison linéaire qui explique au mieux la pluie.

Le modèle de régression à envisager est de la forme :

$$Y = C\beta + \varepsilon \tag{3.4}$$

avec :
- Y : La moyenne mensuelle des pluies journalières
- C : La matrice des composantes principales sur les stations.
- β : vecteur des coefficients de régression

Étant donné que la matrice C est fixée, la méthode consiste à trouver le vecteur β qui minimise la somme des carrés des résidus de la régression :

$$\widehat{\beta} = \arg\min(\sum_i (\widehat{u_i}) = \sum_i (y_i - \widehat{y_i})) \tag{3.5}$$

Cette méthode d'estimation des coefficients de régression est applée "Ordinary Least Squares" (OLS).Bien évidemment cette méthode d'estimation suppose certaines hypothèses qui permettent d'avoir un estimateur BLUE (Best Linear Unbiased Estimator).
Les hypothèses de OLS sont :

- h0 : Le modèle est correct
- h1 : Les erreurs ε_i sont centrées, (ie $E(\varepsilon_i)=0$, $\forall$ i)
- h2 : les ε_i ont la même variance, (ie $var(\varepsilon_i)=\sigma^2$)
- h3 : les $(\varepsilon_i)_i$ sont non corrélées, (ie $cov(\varepsilon_i,\varepsilon_j)=0$, $\forall i \neq j$)
- h4 : ε_i N(0, σ^2)

Quelques propriétés de l'estimateur $\widehat{\beta}$ de β par OLS :

Sous les hypothèses de la méthode OLS, les estimateurs des coefficients de régression ont les propriétés suivantes :

- sous h0 et h1 : les estimations de β sont sans biais c-à-d $E(\widehat{\beta}) = \beta$.
- sous h0,h1,h2 et h3 : les estimateurs sont optimaux ie efficaces parmi les estimateurs sans biais linéaires de β.
- sous les cinq hypothèses : les estimateurs $\widehat{\beta_i}$ suivent des lois normales et donc permettent d'utiliser le test de significativité de Student, ainsi que le test de Fisher pour la comparaison de modèles emboîtés.

Par la suite on estime les paramètres du modèle de régression tout en supposant que les hypothèses de OLS sont vérifiées et après nous vérifierons, à l'aide des tests sur les résidus, si ces hypothèses ont été respectées par les données ou non.

Pour expliquer la moyenne mensuelle des pluies journalières, on introduit comme variables explicatives les composantes principales et on propose d'introduire aussi la moyenne des altitudes des points voisins des stations météo. Cette moyenne été retranchée de toutes les altitudes des points du masque mais elle peut avoir un poids important dans l'explication de la pluie donc nous devons l'introduire dans le modèle.

Par la suite on teste la significativité de son coefficient de régression pour décider si on doit la conserver dans le modèle ou non.

Variable	$\widehat{\beta}$	Erreur-type	$t-Student$	$Pr(>\|t\|)$
intercept	10.0691	1.6810	5.990	9.24e-08
C1	-0.0118	0.0041	-2.840	0.0059
C2	0.0112	0.0040	2.760	0.007
C3	0.0238	0.0148	1.610	0.112
C4	0.0130	0.0107	1.214	0.228
C5	-0.0067	0.0176	-0.382	0.704
C6	-0.0085	0.0156	-0.549	0.584
C7	-0.0261	0.0242	-1.076	0.285
m	0.0067	0.0029	2.285	0.025

TAB. 3.2 – Modèle de régression de la moyenne mensuelle des pluies journalière du mois de Janvier sur les composantes principales + la moyenne des altitudes des points du masque autour des stations

3.6.4 Analyse des résidus du modèle de régression

L'analyse des résidus du modèle de régression permet de savoir à quel point les hypothèses de OLS ont été respectées par les données.

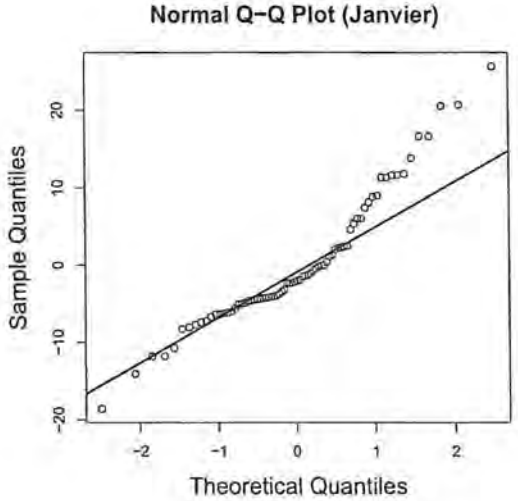
3.6.4.1 Test de normalité des résidus

L'hypothèse de normalité des résidus est importante car si elle est vérifiée on peut appliquer le test de Student [(2)] sur les estimateurs des coefficients de régression et le test de Fisher [(3)] sur la validité globale du modèle.

[2]test de significativité des coefficients de régression
[3]test de validité globale du modèle

QQ-plot

FIG. 3.8 – QQ-plot des résidus de la régression de la moyenne mensuelle des pluies journalières sur les composantes principales et la moyenne des altitudes des points au tour des stations- mois de Janvier



On remarque d'après les deux figures de (3.8(a)) que la partie gauche de la distribution des résidus est plus régulière que la partie droite qui s'éloigne progressivement de la droite de Henry[4], ceci ne permet pas de juger d'une manière rassurante la normalité des résidus. Pour prendre une décision nous faisons appel au test de Kolmogorov-Smirnov.

Test de Kolmogorov-Smirnov

En général, le test de normalité de Kolmogorov-Smirnov se comporte bien dans la majorité des situations et il est, en particulier, robuste en présence d'autocorrélation spatiale (Dutilleul et Legendre, 1992).
Dans notre cas, la statistique empirique du test de Kolmogorov-Smirnov, pour le mois de Janvier vaut **D= 0.1455** qui est inférieure à la statistique théorique $\mathbf{D}_{max}$**= 0.1534**, et donc on peut décider au risque habituel de 5% que la distribution des résidus soit normale.

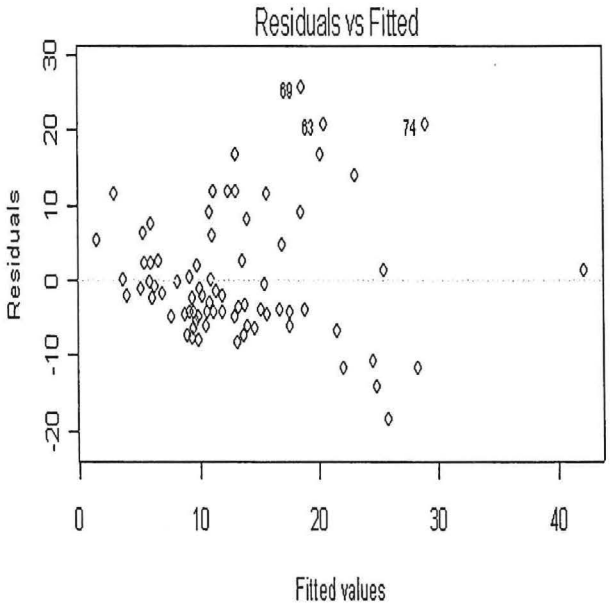
[4]la droite de Henry est une méthode pour visualiser qu'une distribution est Gaussienne ou non. Elle permet de lire rapidement la moyenne et l'écart type d'une telle distribution

CIRAD-DIST
Unité bibliothèque
Lavalette

3.6.4.2 Test d'homogénéité de la variance

L'homogénéité de variance des résidus de la régression peut être observés à partir du graphique des résidus en fonction des valeurs ajustées.

FIG. 3.9 – Evolution des résidus de la régression en fonction des valeurs prédites- mois de Janvier



Les résidus de la régression augmentent en fonction des valeurs de la moyenne mensuelle des pluies journalières ajustées par le modèle, pour le mois de janvier donc la variance des résidus n'est pas constante et l'hypothèse d'homogénéité de la variance des résidus est infirmée, ce qui met en doute l'efficacité de l'estimation des paramètres du modèle.

Test de Glejser

Le test de Glejser est un test d'hétéroscédasticité des résidus (Godfrey, 1996), il consiste à tester l'hypothèse H0 suivante :

| H0 : La variance des résidus est hétérogène.

Le principe du test de Glejser est :

1. Estimer le modèle de régression d'une variable réponse Y sur la matrice des régresseurs X.
2. Faire une régression de la valeur absolue des résidus sur les valeurs ajustées de Y.
3. Si la pente de la droite de régression de la valeur absolue des résidus sur les valeurs ajustées de Y est faible alors la variance des résidus est considérée comme homogène sinon elle est hétérogène.

Dans notre cas, le test de Glejser donne le résultat suivant :

Variable	Estimate	Erreur-type	$t - Student$	$Pr(>\|t\|)$
intercept	1.9021	1.1883	1.601	0.114
$\widehat{Pluie}$	**0.3378**	0.0798	**4.233**	6.54e-05

TAB. 3.3 – Test de Glejser : test de significativité du coefficient de régression de la valeur absolue des résidus sur les valeurs ajustées de la pluie

La statistique de Student du coefficient de régression de la valeur absolue des résidus sur les valeurs ajustées de la pluie est $t - student = 4.233 > 2$

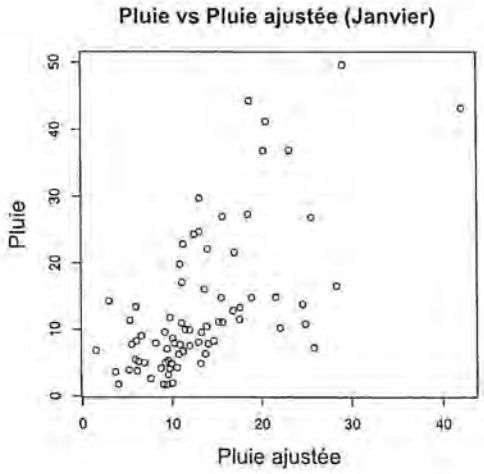
(Voir Tab. (3.3)), donc le coefficient de régression est significatif et par conséquent la variance des résidus est hétérogène.

Pour stabiliser la variance des résidus nous proposons une transformation de la variable pluie. Comme il y a plusieurs types de transformation : $\ln(y)$, $\sqrt{y}$, $\frac{1}{y}$, y^{α}, Box-Cox, ...
La transformation Box-Cox parait la plus judicieuse dans notre cas, car il est difficile de fixer à l'avance la transformation adéquate et en plus elle représente un cas assez général pour les transformations.

3.6.4.3 Vérification visuelle de l'hypothèse de linéarité du modèle

L'hypothèse de linéarité du modèle est vérifiée à partir du croisement des vraies valeurs de la pluie en fonction des valeurs ajustées par le modèle.

FIG. 3.10 – Pluie en fonction des pluie ajustées- mois de Janvier



Le nuage des points de la moyenne mensuelle des pluies journalières en fonction de ses valeurs ajustées par le modèle ne représente aucune courbure ou forme géométrique particulière donc il semble que l'hypothèse de linéarité du modèle soit respectée.

3.7 Transformation de la variable pluie et régression sur les composantes principales : élaboration du modèle sur la base du mois de Janvier

Les majeurs inconvénients du premier modèle de régression de la moyenne mensuelle des pluies journalières sur les composantes principales sont :

- La variance des résidus est hétérogène.
- Les résidus ne sont pas normaux pour les mois de Mai, Juillet, Septembre, Octobre et Novembre.

Donc pour résoudre la mauvaise adéquation du modèle aux données nous proposons une transformation de la variable réponse par la méthode de Box-Cox.

3.7.1 Les avantages de la transformation Box-Cox apportés à la régression

La transformation de Box-Cox consiste à construire à partir d'une variable Y, une autre variable Y^* qui s'écrit de la forme :

$$Y^* = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ \text{Log(Y)} & \text{si } \lambda = 0 \end{cases} \tag{3.6}$$

avec :
- Y : le vecteur à transformer
- Y^* : le transformé de Y
- λ : le paramètre de la transformation (en général $\lambda \in [-2, 2]$).

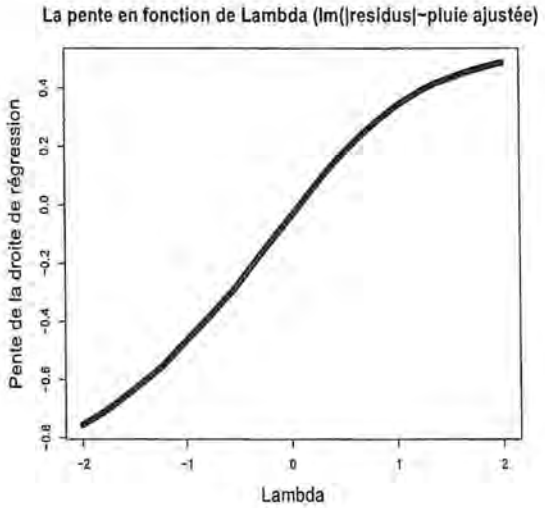
La transformation de Box-cox agit sur les trois propriétés suivantes ensemble, donc il faudrait choisir l'optimisation d'un des critères et se contenter de vérifier que les deux autres sont acceptables. Ce qui nous intéresse le plus c'est la stabilité de la variance des résidus.

- stabiliser la variance des résidus de la régression,
- normaliser les résidus,
- la linéarité de la relation.

Comment déterminer le paramètre λ de la transformation Box-Cox ?

Idée : Pour une série de valeurs de λ (en général $\lambda \in [-2, 2]$) on estime les paramètres du modèle puis on fait la régression de la valeur absolue des résidus sur les valeurs ajustées de la pluie et on retient le λ qui correspond à la pente la plus faible en valeur absolue de la droite de régression.

FIG. 3.11 – Variation de la pente de la droite de régression des valeurs absolues des résidus sur les valeurs prédites de la pluie-mois de Janvier



Dans ce cas ci le λ qui correspond à la pente minimale en valeur absolue vaut 0.08. Comme 0.08 est très proche de zéro ceci revient à une transformation Log. Comme la transformation Log ne sera pas applicable par la suite sur les données des mois les moins arrosés tel que Mai, Juillet et Octobre et qui sont caractérisés par des mesures nulles de pluie sur une proportion non négligeable de stations météo, nous avons retenu la transformation Log(Y+1) plutôt que Log(Y).

3.7.2 Résultats du modèle après transformation Log

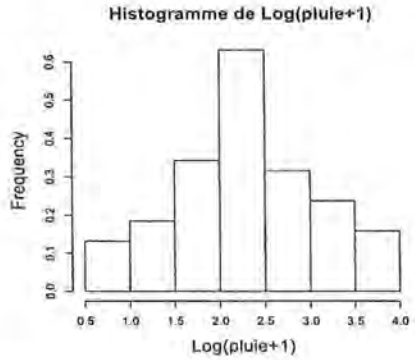
Le modèle de régression du logarithme de la pluie sur les composantes principales et la moyenne des altitudes est :

Variable	$\widehat{\beta}$	Erreur-type	$t - Student$	$Pr(>\|t\|)$
intercept	2.1819	0.1101	19.808	$< 2e-16$
C1	-0.00075	0.00027	-2.745	0.0077
C2	0.00072	0.00026	2.717	0.0083
C3	0.0016	0.0009	1.653	0.1029
C4	0.00089	0.0007	1.279	0.2052
C5	0.00039	0.0011	0.344	0.7321
C6	-0.0003	0.001	-0.298	0.766
C7	-0.0017	0.0015	-1.132	0.261
m	0.0004	0.0001	2.465	0.0162

TAB. 3.4 – Modèle de régression de la transformée de la moyenne mensuelle des pluies journalières du mois de Janvier sur les composantes principales + la moyenne des altitudes des points du masque autour des stations

Un premier résultat de la transformation c'est que la distribution de la moyenne mensuelle des pluies journalières suit une loi normale.

FIG. 3.12 – La distribution de la pluie après transformation-mois de Janvier



La deuxième conséquence de la transformation est l'homogénéité de la variance des résidus. En effet, par construction le paramètre λ de la transformation de

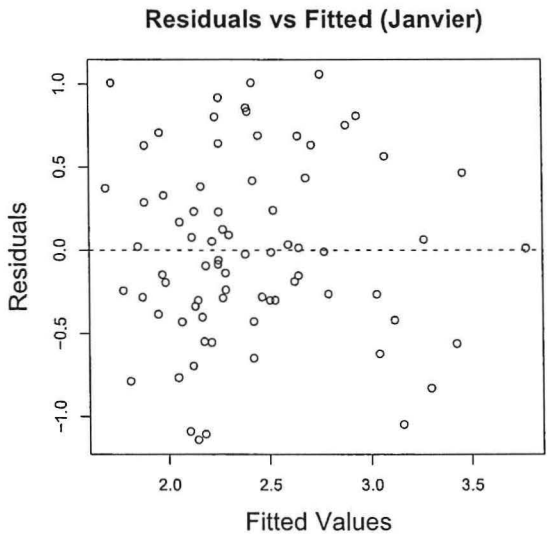
Box-Cox est choisie sous le critère de minimisation de la pente de la droite de régression des valeurs absolues des résidus sur les valeurs ajustées de la pluie.

Le test de Glejser après transformation donne le résultat suivant :

Variable	Estimate	Erreur-type	$t - Student$	$Pr(>\|t\|)$
intercept	0.4210	0.2073	2.030	0.0459
$\widehat{Pluie}$	**0.0072**	0.0848	**0.085**	0.9321

TAB. 3.5 – Test de Glejser sur les résidus après transformation de la variable pluie

FIG. 3.13 – Les résidus de la régression en fonction des valeurs ajustées de pluie-mois de Janvier



On peut confirmer, d'après les résultats ci-dessus, que la transformation Log de la pluie était nécessaire pour aboutir à un modèle plus performant.

3.7.3 Sélection des composantes principales explicatives par régression descendante

Après avoir testé l'adéquation des résidus aux hypothèses de la méthode d'estimation OLS, on doit sélectionner les composantes explicatives par une régression descendante.

Le choix de la régression descendante est basé sur le fait que les composantes principales des stations météo peuvent être corrélées. Donc si par exemple il y a deux composantes qui ont des coefficients de régression non significatives (ie $t - student < 2$) on ne doit pas les enlever les deux à la fois du modèle mais il faut retirer la composante la moins significative et après reconstruire le modèle et ainsi de suite jusqu'a ce que toutes les composantes retenues soient significatives.

Le premier modèle avec les 7 composantes principales et la moyenne des altitudes du voisinage des stations donne le résultat suivant :

Variable	$\hat{\beta}$	Erreur-type	$t - Student$	$Pr(>\|t\|)$
intercept	2.1819	0.1101	19.808	$< 2e - 16$
C1	-0.00075	0.00027	-2.745	0.0077
C2	0.00072	0.00026	2.717	0.0083
C3	0.0016	0.0009	1.653	0.1029
C4	0.00089	0.0007	1.279	0.2052
C5	0.00039	0.0011	0.344	0.7321
C6	-0.0003	0.001	-0.298	0.766
C7	-0.0017	0.0015	-1.132	0.261
m	0.0004	0.0001	2.465	0.0162

TAB. 3.6 – Modèle de régression de la transformée de la moyenne mensuelle des pluies journalières du mois de Janvier sur les composantes principales + moyenne des altitudes des points du masque autour des stations

La 6e composante est la moins significative ($p - value = 0.766 > 0.05$), donc nous devons la retirer du modèle.

En suivant cette démarche, le modèle final qui résulte de la régression descendante est :

Variable	$\hat{\beta}$	Erreur-type	$t - Student$	$Pr(>\|t\|)$
intercept	2.2012	0.1075	20.476	$< 2e-16$
C1	-0.0006	0.0002	-3.164	**0.0022**
C2	0.0008	0.0002	3.759	**0.0003**
C3	0.0018	0.0009	2.019	**0.047**
m	0.0004	0.0001	2.200	**0.031**

TAB. 3.7 – Modèle final de la régression descendante de la transformée de la moyenne mensuelle des pluies journalières du mois de Janvier sur les composantes principales + moyenne des altitudes des points du masque autour des stations

Finalement on conserve les trois premières composantes principales avec la moyenne des altitudes du voisinage du masque autour des stations, et le modèle s'écrit sous cette forme :

$$\widehat{Log(Y+1)} = \hat{\alpha}_0 + \hat{\alpha}_1 C_1 + \hat{\alpha}_2 C_2 + \hat{\alpha}_3 C_3 + \hat{\alpha}_4 m \tag{3.7}$$

avec :

- Y : la moyenne mensuelle des pluies journalières.
- $(C_i)_{i=1,2,3}$: les composantes principales retenues par la régression descendante.
- m : la moyenne des altitudes des points du masque autour des stations.
- $(\hat{\alpha}_i)_{i=0,\ldots,4}$: le i^e coefficient de régression.

L'équation (3.7) explique la moyenne mensuelle des pluies journalières, transformée par la fonction Log, en fonction de la moyenne des altitudes des points du masque autour de la station (m) et des trois premières composantes principales retenues après la régression descendante.

Par la suite on passe du modèle en les composantes principales à celui au variables initiales (les altitudes des points du masque autour de la station) :

$$\begin{aligned} \widehat{Log(Y+1)} &= \hat{\alpha}_0 + \hat{\alpha}_1 C_1 + \hat{\alpha}_2 C_2 + \hat{\alpha}_3 C_3 + \hat{\alpha}_4 m \\ &= \hat{\alpha}_0 + \hat{\alpha}_1 XA_1 + \hat{\alpha}_2 XA_2 + \hat{\alpha}_3 XA_3 + \hat{\alpha}_4 m \\ &= \hat{\alpha}_0 + X(\hat{\alpha}_1 A_1 + \hat{\alpha}_2 A_2 + \hat{\alpha}_3 A_3) + \hat{\alpha}_4 m \end{aligned}$$

avec :

- X : matrice des variables explicatives de dimension 90x361.
- $(A_i)_{i=1,2,3}$: les trois premiers vecteurs propres de l'opérateur d'Escoufier VQ [5].

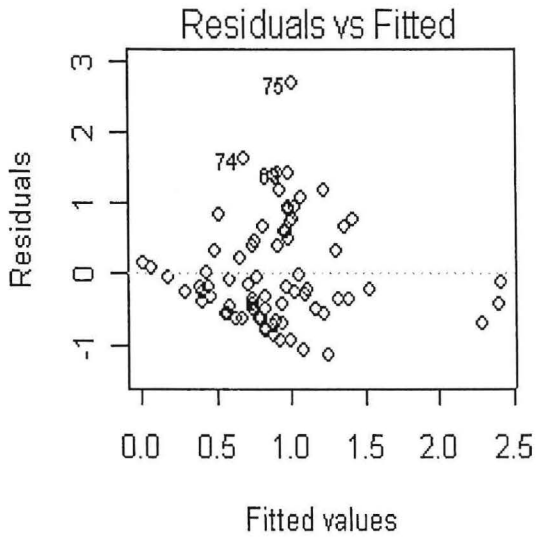
3.7.4 Les limites de la transformation Box-Cox pour les mois les moins arrosés

La transformation Box-Cox donne des résultats faciles à interpréter pour les mois les plus arrosés tel que Janvier, Février, Mars, Avril et Décembre. Par contre on ne distingue pas clairement son effet sur les mois les moins arrosés à savoir : Mai, Juillet, Septembre, Octobre et Novembre. Ces mois sont caractérisés par des mesures nulles de pluie sur plusieurs stations météo ce qui limite l'effet de la transformation sur la stabilisation de la variance des résidus.

[5] V_{pxp} est la matrice de variance covariance de X et Q_{pxp} matrice des poids sur les variables de X. Dans notre cas $Q = I_p$

CIRAD-DIST
Unité bibliothèque
Lavalette

FIG. 3.14 – Les r[illegible] pluie-mois de mai



Le croisement des résidus avec les valeurs ajustées de la pluie montre que pour les valeurs prédites les plus faibles, la variance n'est pas constante, mais qu'elle commence à se stabiliser à partir d'une valeurs prédite de 1. Cet effet d'hétérogénéité de la variance au debut est dû aux valeurs nulles de pluie. Les valeurs nulles de pluie forment un alignement de points sur la figure (3.14) d'équation $Y - \hat{Y} = -\hat{Y}$.

Conclusion :

La méthode de Régression sur Composantes Principales (RCP) a permis de réduire le nombre de variables explicatives, on a passer de 361 à 7 variables explicatives. Les variables latentes construites à partir de l'ACP bi-centrée résument 98 % de la variance de la matrice des variables originales, mais les 7 variables choisies ne sont pas forcément celles qui expliquent le mieux ou qui sont les plus corrélées à la variable réponse pluie.

La transformation de la variable pluie a permis de stabiliser la variance des résidus et de les normaliser dans le cas des mois les plus arrosés par contre on pouvait pas voir clairement son effet sur les données des mois les moins arrosés à cause des valeurs nulles de pluie. Cet effet sera constaté aussi pour les mois les plus arrosés, lorsqu'on passera à la modélisation de la pluie hebdomadaire vu qu'à cet échelle il y aura beaucoup plus de mesures nulles de pluie.

Chapitre 4

Régression PLS

4.1 Rappel sur la PLS1

La régression PLS (Wold, 1966 ; Tennenhaus, 1998 ; Durand, 2004) est une méthode de régression qui s'est beaucoup développée dans le cas où les variables explicatives sont fortement corrélées entre elles ou quand le nombre de variables explicatives est supérieur au nombre d'individus.

Dans ces situations, la méthode de régression multiple ordinaire est plus délicate à utiliser notamment dans sa partie test, à cause de l'obtention d'une matrice non inversible. Il est toujours possible alors d'obtenir des estimations des variables à expliquer, mais les coefficients de régression ne sont plus uniques. Un autre inconvénient de la régression multiple est sa difficulté d'interprétation étant donné que les coefficients de régression ne sont pas représentatifs de la corrélation entre la variable à expliquer et les variables explicatives (Elkettani, 2001). La régression PLS permet de remédier à ces inconvénients.

Partial Least Squares à une variable réponse PLS1 :

Les données sont constituées par une matrice $\mathbf{Z}$ de dimension nxq de variables explicatives, et un vecteur $\mathbf{Y}$ de dimension nx1 de variable à expliquer. On suppose de plus que $\mathbf{Z}$ et $\mathbf{Y}$ sont toutes deux centrées en colonnes par rapport à la métrique $n^{-1}I_n$ (matrice des poids sur les individus).

La régression PLS1 est basée sur le calcul d'un jeux de α composantes orthogonales entre elles $T = [t_1, ..., t_\alpha]$ qui sont des combinaisons linéaires du tableau $\mathbf{Z}$ telle que les covariances entre $t_i = \mathbf{Z}\, u_i$ et $\mathbf{Y}$ soient maximales. Les composantes $\mathbf{T}$ sont alors utilisées comme nouvelle variables explicatives, l'estimateur PLS étant obtenu par régression des moindres carrés de $\mathbf{Y}$ sur $\mathbf{T}$.
Plusieurs algorithmes ont été proposés pour calculer les composantes $\mathbf{T}$. Les plus répondus sont les algorithmes NIPALS (Wold, 1966) et SIMPLS (De Jong, 1993) le dernier étant plus rapide et donnant des résultats plus faciles à interpréter (Hubert & Vander Branden, 2003). Nous décrivons l'algorithme SIMPLS dans le cas de PLS1, sachant qu'il est équivalent à l'algorithme NIPALS dans ce cas là (Hubert & Vander Branden, 2003).

l'algorithme SIMPLS est :

On initialise tout d'abord $i = 0$, $\mathbf{Z}^{(0)} = \mathbf{Z} \qquad \mathbf{Y}^{(0)} = \mathbf{Y}$

A chaque étape $1 \leq i \leq \alpha$, la composante t_i est obtenu comme vecteur solution du problème suivant :

$$\begin{aligned} \max Cov(t, y) &= cov(\mathbf{Z}^{(i-1)}u, \mathbf{Y}^{(i-1)}) \\ u^t u &= 1 \end{aligned}$$

On peut montrer (Tenenhaus, 1998) que la solution u_α est le vecteur propre normé de la matrice symétrique $n^{-2}\mathbf{Z}^{(i-1)'}Y^{(\alpha-1)}Y^{(i-1)'}\mathbf{Z}(\alpha - 1)$ associé à la plus grande valeur propre.
La α^e composante s'obtient donc par $t_\alpha = \mathbf{Z}^{(i-1)}u_\alpha$ puis $\mathbf{Z}^{(i-1)}$ et $\mathbf{Y}^{(\alpha-1)}$ sont projetés sur la composante t_i pour obtenir les résidus $\mathbf{Z}^{(i)}$ et $\mathbf{Y}^{(i)}$.

Comment valider le modèle PLS ?

Il y a trois critères pour valider le modèle PLS :

Critères MSE et MSEP

Un outil pour évaluer la performance d'une régression PLS est le calcul du critère MSE, ou Mean Squared Error. Il est défini pour α composantes par :

$$MSE(\alpha) = \sum_{i=1}^{n} \frac{1}{n} [y_i - \widehat{y}_{\alpha i}]^2 \tag{4.1}$$

Dans le cas de la régression PLS1, avec $\widehat{y}_{\alpha i}$ prédiction pour l'observation y_i calculée avec le modèle à α composantes.

Si l'on calcule les prédictions $\widehat{y}_{\alpha i'}$ associées à n' nouvelles observations y'_i (qui n'ont pas servi à calculer le modèle), alors on définit de manière similaire le critère MSEP, ou Mean Squared Error of Prediction :

$$MSEP = \sum_{i=1}^{n'} \frac{1}{n'} [y_{i'} - \hat{y}_{\alpha i'}]^2 \tag{4.2}$$

Ces critères ne sont cependant pas suffisants pour évaluer la qualité de la régression. En particulier, pour choisir le nombre α de composantes optimales, on réalise habituellement la validation croisée.

Validation Croisée (CV) :

La validation croisée permet d'obtenir le nombre optimal α de composantes PLS à retenir.

Dans le cas de la régression PLS1, pour chaque valeur de α on calcule les prédictions $\hat{y}_{\alpha[-i]}$ à l'aide du modèle à α composantes, calculé en utilisant la régression PLS utilisant toutes les observations sauf l'observation i.
La statistique utilisée est alors le PRESS ou PRediction Error Sum of Squares définit par (d'après Tenenhaus (1998)) :

$$PRESS_1(\alpha) = \sum_{i=1}^{n} \frac{1}{n} [y_i - \hat{y}_{\alpha i}]^2 \tag{4.3}$$

Le nombre optimal de composantes est choisi comme étant la valeur de α pour laquelle $PRESS_1$ est minimal ou suffisamment faible.

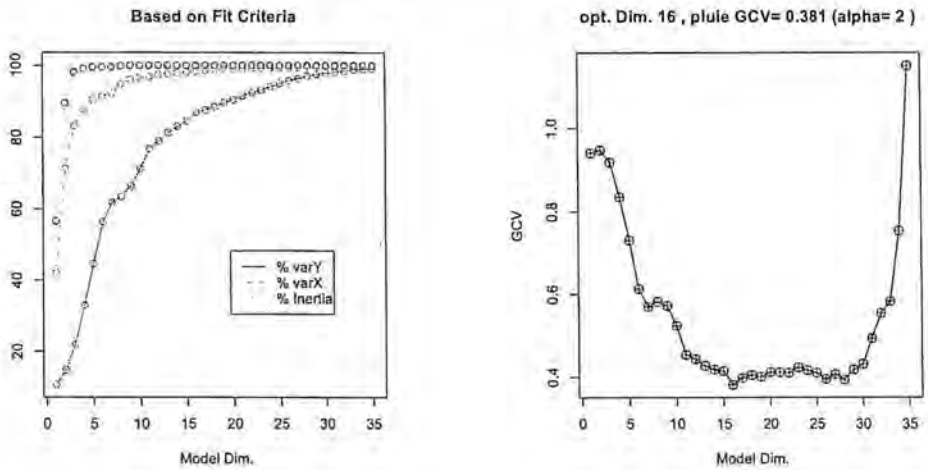
4.2 Régression PLS sur les données du mois de Janvier

Dans cette partie nous allons construire un modèle PLS avec comme variables explicatives les 361 variables, altitudes des points du masque autour de la station, la moyenne des altitudes des points voisins, m, et comme variable réponse la moyenne mensuelle des pluies journalières transformée par la fonction Log.
On a montré, dans le chapitre 3, que la transformation de la pluie est nécessaire dans le cadre d'une régression sur les composantes principales, et elle l'est aussi dans la PLS car se sont deux méthodes de projection qui différent essentiellement dans la manière de construction des composantes principales .

4.2.1 Choix du nombre de composantes PLS

Pour déterminer le nombre de composantes PLS, on ne dispose pas, ici, d'un jeu de données test. Seuls les critères basés sur l'ajustement aux données (Fig.4.1(a))et sur la validation croisée (Fig.4.1(a)) sont utilisables.

FIG. 4.1 – Critères d'aide à la décision pour le choix du nombre de composantes - Mois de Janvier



(a) (b)

La figure (4.1(a)) représente trois courbes :

- En rouge (courbe inférieure) : le pourcentage de la variance de **Y** reconstruit par les k composantes, $k \in [1, 35]$.
  La variance de **Y** reconstruite par les k composantes est donnée par la formule suivante :

$$I^y(k) = \sum_{i=1}^{k}(\sum_{l=1}^{q} r^2(y^l, t^i)var(y^l)) \tag{4.4}$$

  dans notre cas, q=1 (une seule variable réponse) donc l'équation(4.7)devient :

$$I^y(k) = \sum_{i=1}^{k}(r^2(y, t^i)var(y)) \tag{4.5}$$

  avec : $r^2(y, t^i)$: le carré du coefficient de corrélation entre y et la i^e composante PLS
  Le pourcentage de variance totale reconstruite par k axes est donnée par :

$$\%I^y(k) = 100 \frac{\sum_{j=1}^{k} I_j^y}{I^y} \tag{4.6}$$

  avec I^y : la variance totale de **Y** reconstruite par 361 composantes.
- En vert (courbe médiane) : le pourcentage de la variance de **X** reconstruit par k composantes, $k \in [1, 35]$.
  La variance de **X** reconstruite par les k composantes est donnée par la formule suivante :

$$I^x(k) = \sum_{i=1}^{k}(\sum_{l=1}^{p} r^2(x^l, t^i)var(y^l)) \tag{4.7}$$

  avec : $r^2(x^l, t^i)$: le carré du coefficient de corrélation entre la l^e variable explicative x^let la i^e composante PLS t^i
  La variance totale de **X** est donnée par l'expression suivante :

$$I^x(r) = tr(\mathbf{V}) = ||x||_D^2 = \sum_{l=1}^{p} var(x^l) \tag{4.8}$$

avec : **V** : matrice de variance-covariance de **X**, et r=rang(**X**)
Le pourcentage de variance totale reconstruite par k axes est donnée par :

$$\%I^x(k) = 100 \frac{\sum_{j=1}^{k} I_j^x}{I^x} \tag{4.9}$$

avec I^x : la variance totale de **X** reconstruite par 361 composantes.

- En bleu : le pourcentage de l'inertie des individus reconstruite par les k composantes PLS.

$$I^{ind}(k) = \sum_{i=1}^{k} var(t^{*i}) \tag{4.10}$$

avec : $t^{*i} = \mathbf{W}D \frac{t^i}{\|t^i\|_D}$ (Voir détails dans le livre de Tenenhaus).

L'inertie totale est :

$$I^{ind}(r) = \|\mathbf{V}\|_F^2 = \|\mathbf{X}\|_{V \otimes D}^2 = \sum_{h,l} cov^2(x^l, x^h) \tag{4.11}$$

Le pourcentage de l'inertie totale reconstruite par les k composantes est :

$$\%I^{ind}(r) = \sum_{i=1}^{k} 100 \frac{I_i^{ind}}{I^{ind}} \tag{4.12}$$

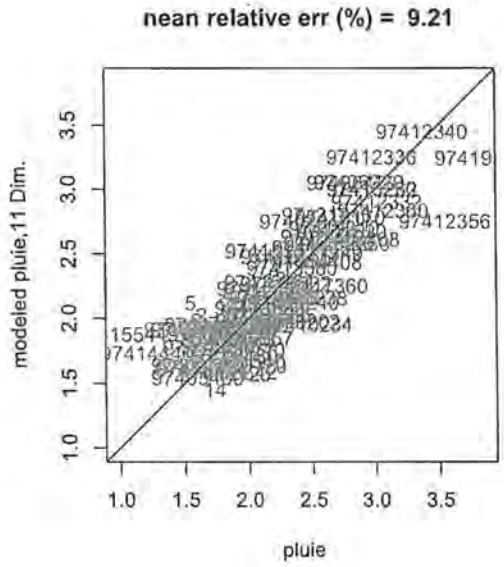
Interprétation des figures (4.1) :

Avec 16 composantes, PLS reconstitue 86.97 % de Y ainsi que 100% de l'inertie des individus et 98.36 % de la variance de X (Fig. 4.1(a)). La valeur optimale du PRESS est obtenue pour 16 composantes avec PRESS(16)=0.381 (Fig.4.1(b)). L'évolution des différents critères, Figure 4.1, indique que 16 est le meilleur choix pour la dimension du modèle.
On remarque que, entre la dimension 11 et la dimension 16, le PRESS ne change pas considérablement en effet PRESS(16)=0.38 et PRESS(11)=0.42. En plus le dernière décroissance remarquable du PRESS est celle de la dimension 10 à la

CIRAD-DIST
Unité bibliothèque
Lavalette

dimension 11. Donc on propose de s'arrêter jusqu'à la dimension 11.
La validation croisée permet aussi d'examiner quels sont les individus qui sont correctement prédits et quels sont ceux qui le sont moins. Pour cela la figure (4.2)présente le graphique des valeurs observés sur Y contre les valeurs prédites par validation croisée à 11 composantes. L'erreur relative moyenne de la prédiction est égale à 9.21 %.

FIG. 4.2 – Validation croisée : observations vs prédictions par une régression PLS à 11 composantes



Ce sont les individus {97412336, 97412356, 97415544, 5, 97414440} qui sont les moins bien prédits par le modèle à 11 dimensions.

4.2.2 Le modèle PLS

Finalement le modèle retenu, à 11 composantes, s'écrit sur les variables standardisées,

$$Log(\widehat{Pluie} + 1) \quad = \quad 0.04Z^1 + 0.043Z^2 + ... - 0.15Z^{360} - 0.17Z^{361} + 0.68m$$

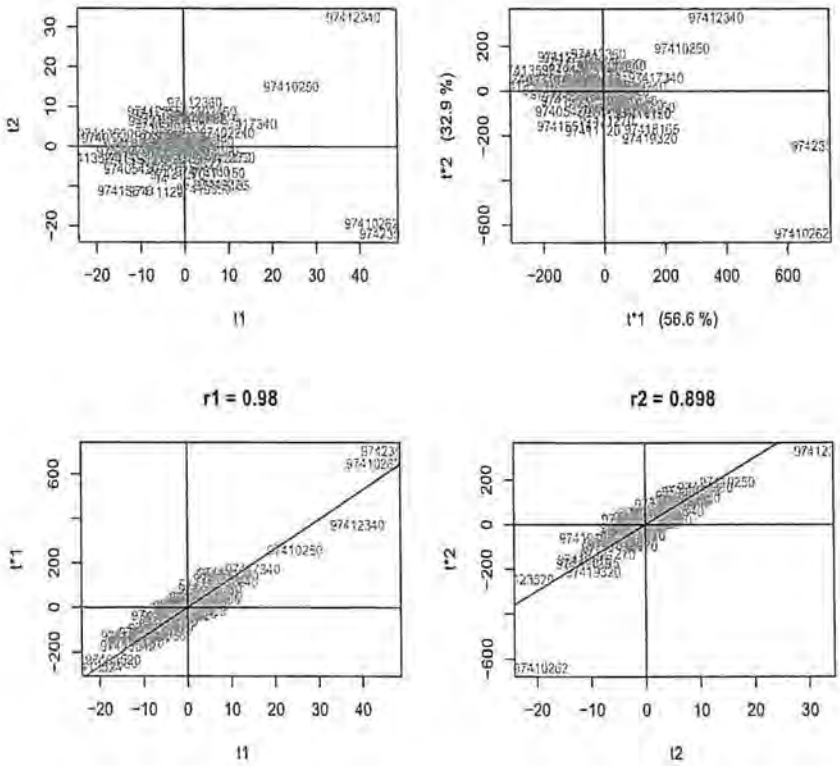
Sur les données d'origine, le modèle devient :

$$Log(\widehat{Pluie} + 1) \quad = \quad 1.912 + 0.0036Z^1 + 0.004Z^2 + ... - 0.0024Z^{361} + 0.001m$$

4.2.3 Représentation des individus

Rappelons que le plan factoriel (i,j) est la projection V-orthogonale des individus sur le plan (w^{*i}, w^{*j}). Les coordonnées du nuage des points projetés sont données par le couple (t^{*i}, t^{*j}). Les deux plans factoriels (i,j)individus-variables ne sont pas en dualité.
On ne peut interpréter un axe-coordonnée t^*, par rapport aux variables projetées sur l'axe-projection t, de même nom, que si le "saut de dualité" entre t^* et t est faible, c'est à dire si le coefficient de corrélation entre ces deux variables est voisin de un.
Dans notre cas, la figure (4.3) montre que les coefficients de corrélation sont voisins de un (r1=0.98 et r2=0.89), donc le saut de dualité est faible entre la projection des individus sur le plan factoriel (t^{*1},t^{*2}) et celle sur le plan (t^1,t^2).

FIG. 4.3 – Comparaison entre les représentations pseudo-factorielles et factorielles pour le plan (1,2) des individus. Visualisation des individus dans les régressions de t* sur t pour les axes 1 et 2- Mois de Janvier



D'après le graphique de projection des individus sur le plan factoriel (t^{*1},t^{*2}), en haut à droite, il y a 4 individus qui sont éloignés de l'ensemble du nuage des points, ces stations sont respectivement {97410250, 97412340, 97410262, 97423520}, Takamaka PK12, Takamaka Usine, Trois-Bassins et Grand Galet. En revenant à la carte de la Réunion on remarque que ses stations météo se trouvent dans des ravines.

4.3 Conclusion

Théoriquement la régression PLS est meilleure que la régression sur composantes principales car elle permet de construire des composantes principales qui reconstituent suffisamment la variance de **X** et qui soient corrélées avec la variable réponse **Y**.

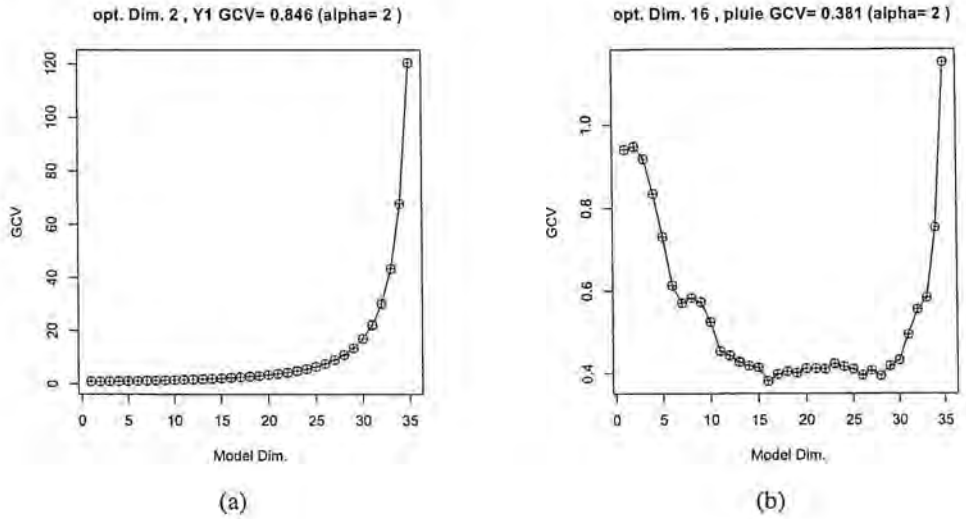
Pour mettre en oeuvre ce résultat théorique dans notre étude, nous allons comparer les résultats des deux méthodes, PLS et RCP, en se basant sur deux critères : le PRESS et la qualité de l'ajustement des données.

Pour comparer les résultats de la régression sur composantes principales avec ceux de la PLS nous allons se baser sur deux critères :

Le PRESS :

Pour la régression PLS le PRESS égale à 0.381 (Fig.4.4(b)), inférieur à celui fournit par la RCP, égale à 0.846 (Fig.4.4(a)), donc la PLS est meilleure que la Régression sur composantes principales.

FIG. 4.4 – Comparaisons du PRESS de la régression PLS à celui de la Régression sur composantes principales- Mois de Janvier

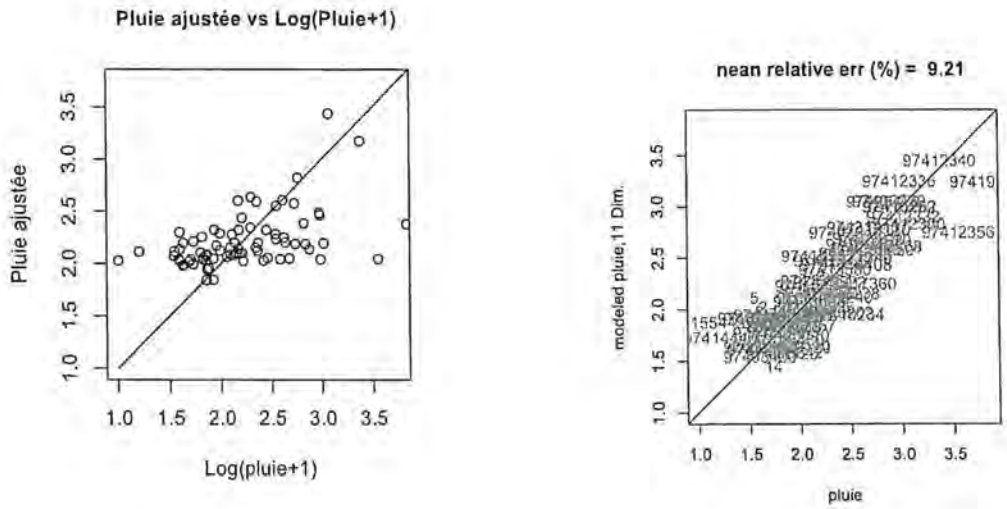


(a) (b)

Critère d'ajustement des données :

Le graphique des valeurs ajustés en fonction des valeurs initiales montre que la PLS donne un meilleur ajustement que la régression sur composantes principales.

FIG. 4.5 – Comparaisons du critère d'ajustement des données de la régression PLS à celui de la Régression sur composantes principales- Mois de Janvier



(a) (b)

Chapitre 5

Conclusion et Perspectives

Ce stage, interpolation de la pluie sur l'île de la Réunion, entre dans le cadre d'une étude faite pour une thèse : " Analyse de données longitudinales : Modélisation statistique de séries temporelles de richesse en sucre de la canne sur l'île de la Réunion".

La modélisation de la richesse devra être faite en tenant compte des covariables climatiques dont la pluviomètrie.

L'objet de ce stage c'est à la fois une étude de faisabilité et la comparaison des performances de deux méthodes, pour estimer en tout point de l'île les pluies hebdomadaires, en tenant compte du relief. Ces deux méthodes sont : la Régression sur Composantes Principales (RCP) et la Partial Least Squares (PLS).

Mais nous avons commencé par modéliser la moyenne mensuelle des pluies journalières, car elle est plus régulière et correctement relevée.

Pour la RCP, une ACP centrée faite sur les stations météo a montré une importance particulière de la première composante principale ce qui peut imposer des problèmes pour le calcul des composantes suivantes. En plus elle est presque égale à la moyenne des altitudes de tous les points du masque. D'où l'intérêt de centrer en ligne le tableau des données.

Dans une deuxième étape on a fait une ACP bi-centrée sur les altitudes des points voisins des stations météo, mais la cartographie en courbes de niveaux du premier vecteur propre est différente de celle pour une ACP sur un échantillon aléatoire

pris du MNT. Donc il y a des configurations de relief local qui sont sous représentées ou pas représentées du tout sur la population des stations.

En tenant compte de ce résultat, nous proposons de faire une ACP bi-centrée sur un échantillon aléatoire de 1000 points du MNT, représentatif de tous les points de l'île, et après introduire les stations météo comme individus supplémentaires.

Après avoir calculé les composantes principales des postes météo on a fait une régression multiple de la moyenne mensuelle des pluies journalières sur les composantes principales des stations. Les tests sur les résidus ont montré que les résidus ne sont pas normaux et que leurs variance n'est pas homogène. D'où la nécessite de transformer la variable pluie.

La transformation de Box-Cox a donné un paramètre $\lambda = 0.08 \simeq 0$ ce qui revient à une transformation Log.

Cette transformation a permis de stabiliser la variance des résidus, mais son effet est limité pour les mois les moins arrosées qui représentent de nombreuses valeurs nulles.

Pour la PLS, après transformation de la pluie, les résultats semblent plus satisfaisants que la RCP, en effet en se basant sur le critère de validation croisée, la PLS fournit un PRESS inférieur à celui de la RCP, en plus elle ajuste mieux les données que la RCP.

Cette étude a montrée que la modélisation de la moyenne mensuelle des pluies journalières avec la PLS est meilleure que celle avec la RCP. Ainsi, si nous voulons ultérieurement prédire les pluies hebdomadaires, il vaut mieux suivre une démarche de modélisation avec PLS.

Perspectives

Dans cette étude, la position des stations n'a pas été prise en compte, il s'agit d'interpolation par régression plutôt que spatiale. L'aspect interpolation spatiale pourrait être traité en complément, en introduisant une corrélation spatiale sur les erreurs.
La précision des prédicteurs n'est pas calculée que par validation croisée, nous n'avons pas calculé de prédiction individuelle par intervalle. Une telle prédiction serait sans doute utile pour cartographier la précision de l'interpolation plutôt que de calculer une MSEP ou un PRESS unique.

Bibliographie

[1] **Tenenhaus M.** 1998. La régression PLS, théorie et pratique. Paris : TECHNIP, 252p.

[2] **Saporta G.** 1990. Probabilités Analyse des données et Statistique. Paris : TECHNIP, 493p.

[3] **Bry X.** 1995. Analyses factorielles simples. Paris : Economica, 111p.

[4] **Lebart L.,Piron M.,Morineau A.** 1995. Statistique exploratoire multidimensionnelle. Paris : Dunod, 329p.

[5] **Tomassone R.,Audrain S.,Lesquoy de Turkheim E.,Millier C.** 1992. La régression, nouveaux regards sur une ancienne méthode statistique. Inra et Masson, Paris, ISBN 2-225-82696-X.

[6] **Wotling G.,Bouvier Ch.,Danloux J.,Fritsch J-M.** 2000. Regionalization of extreme precipitation distribution using the principal components of the topographical environment. Journal of Hydrology, 233 : 86-101.

[7] **Godfrey G.** 1996. Some results on the Glejser and Koenker tests for heteroskedasticity. Journal of Econometrics, 72 : 275-299.

[8] **Marazzi A.,Yohai V.** 2005. Robust Box-Cox transformations based on minimum residual autocorrelation. Computational Statistics & Data Analysis, : 1-18.

[9] **Dennis Cook R.,Sanford W.** 1994. Transforming a response variable for linearity. Biometrika, **81**(4) : 7-731.

[10] **Franses P.,de Bruin P.** 2002. On data transformations and evidence of nonlinearity. Computational Statistics & Data Analysis, 40 : 621-632.

[11] **Freeman J.,Modarres R.** 2005. Efficiency of test for independance after Box-Cox transformation. Journal of Multivariate Analysis, 95 : 107-118.

[12] **Atkinson C.,Riani M.** 2002. Tests in the plot for robust, diagnostic transformations in regression. Chemometrics and intelligent laboratory systems, 60 : 87-100.

[13] **Edwards J.,Hamilton A.** 1994. Errors in variables and the Box-Cox transformation. Computational Statistics & Data Analysis, 20 : 131-140.

[14] **Foucart T.**. Errors in variables and the Box-Cox transformation. Computational Statistics & Data Analysis, 20 : 131-140.

[15] **Besse P.** 2003. Pratique de la modélisation statistique. (On line). Toulouse : Université Paul Sabatier, 95p. (Publication du laboratoire de statistique et probabilités). <URL :http ://www.lsp.ups-tlse.fr/>.

[16] **Gigliobianco V.** 2003. Mise en oeuvre d'analyse en composantes principales en vue d'interpolation des pluies sur l'île de la Réunion. Montpellier : Université de Perpignan, 58p. Mémoire (IUT statistique et traitement informatique des données de Carcassonne).

[17] **Durand J.F.** 2004. Calcul matriciel et analyse factorielle des données. Montpellier : Université de Montpellier 2, 127p.

Résumé

L'objet de ce stage est à la fois une étude de faisabilité et la comparaison des performances de deux méthodes, pour estimer en tout point de l'île les moyennes mensuelles des pluies journalières, en tenant compte du relief. Ces deux méthodes sont : la Régression sur Composantes Principales (RCP) et la Partial Least Squares (PLS).

Mots clés : Régression sue Composantes Principales, PLS, ACP, Transformation Box-Cox.