UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

Centre de Coopération Internationale en Recherche Agronomique pour le Développement

DESS : METHODES STATISTIQUES APPLIQUEES AUX INDUSTRIES AGRONOMIQUES AGROALIMENTAIRES ET PHARMACEUTIQUES

Université Montpellier II

RAPPORT DE STAGE

Effectué au C.I.R.A.D-C.A.
Du 1er Mars 2005 au 30 Août 2005

Par

Anissa TLILI

Interaction génotype environnement du Sorgho & interpolation des données climatiques sur l'île de la Réunion

Soutenu le 07 Septembre 2005 devant le jury composé de

Mr Ali Gannoun Mr Robert Sabatier et Mme Catherine Aliaume

Sous la direction de

Monsieur Philippe LETOURMY

Monsieur Jacques CHANTEREAU

Remerciements

Je remercie mes maîtres de stage M. Philippe LETOURMY (chef de l'unité Aide à la décision et biostatistique) et M. Jacques CHANTEREAU (chef de l'unité Agrobiodiversité des plantes de savanes) pour m'avoir permis d'effectuer ce stage dans d'excellentes conditions.

M. Eric Gozé (biométricien à l'unité) pour ses précieux conseils.

Mlle Sabine Laurent (Thésarde à l'unité) pour son indispensable aide et sa gentillesse.

MM. Ibnou Dieng et Mohamed Chaouch pour leurs soutiens permanents.

Toute l'équipe de l'unité (Jean, Michel, Jean-Baptiste, Sandrine, et Vincent) pour la bonne ambiance qui régnait.

A ma famille de l'autre côté de la méditerranée et spécialement à ma mère que je ne remercierai jamais assez.

Enfin à mon compagnon qui a cru en moi.

Liste des tableaux

Table des figures

A - Présentation de l'entreprise :

Le CIRAD : Centre de coopération Internationale en Recherche Agronomique pour le Développement est un organisme scientifique français au service du développement durable des pays tropicaux et subtropicaux. Ses services de recherche s'appliquent aux domaines des sciences du vivant et des sciences sociales appliquées à l'agriculture, à la foret, 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 (UPR13) appartenant au département Cultures annuelles (Ca), ses activités de travail concernent l'informatique et les mathématiques. Son activité repose d'une part sur l'amélioration de l'utilisation des techniques statistiques et d'autre part en informatique sur la gestion des projets, la programmation sous WINDOWS, les SIG, les bases de données et l'analyse d'image.

A- 1 - Objectif du Stage :

Ce stage se compose en deux grandes parties : une première partie qui évalue une « core collection » de Sorgho sous différents régimes hydriques ; et une seconde partie concernent l'interpolation des données climatiques sur l'île de la Réunion.

La première partie du stage est une étude qui entre dans le cadre d'une évaluation d'une « core collection » de Sorgho sous différents régimes hydriques : en conditions d'alimentation hydrique non limitante, en cas de déficit hydrique en phase pré florale et en cas de déficit hydrique en phase post florale : elle vise globalement à identifier des variétés adaptées à un stress hydrique .

La seconde partie du stage s'intègre au travaux d'une thèse sur la modélisation de données longitudinales appliquée à la richesse en sucre de la canne à sucre sur l'île de la réunion ; l'objectif étant d'estimer des variables climatiques susceptible d'influencer la richesse de la canne et cela à une échelle temporelle hebdomadaire en tenant compte du relief de l'île.

B -Première partie : Interaction Génotype*Environnement :

I – Introduction :

Le sorgho est la cinquième céréale (après le blé, le riz, le maïs et l'orge) la plus cultivée au monde , il est utilisé pour l'alimentation humaine en Afrique, en Asie du Sud et en Amérique centrale. C'est une culture vivrière importante dans les régions semi-arides tropicales, où la production est largement auto consommée.(J. Chantereau et R. Nicou, 1991).

Aux Etats-Unis et dans les pays développés en général, le grain de sorgho est réservé à l'alimentation animale. Récemment, de nouveaux débouchés industriels sont apparus : fibres de sorgho pour la papeterie et sorghos sucrés pour la production de biocarburants.

Le Sorgho présente l'énorme avantage de supporter la chaleur et la sécheresse mais aussi les sols salins, calcaires ou même gorgés d'eau. Cette plante robuste pousse bien dans les régions à climat chaud comme l'Afrique, cependant dans ces régions son développement peut être limité par le déficit pluviométrique .

La core collection (ou collection noyau) étudiée est une population comprenant 210 accessions de sorgho représentatives de l'ensemble de la diversité génétique de la collection mondiale. C'est un outil de travail qui essaye de rassembler dans une collection limitée de variétés le maximum de diversité existant dans l'espèce étudiée.

La core collection est ensuite utilisée pour explorer la gamme de réaction de l'espèce à un problème donné, ici la tolérance au stress hydrique.

L'analyse de la réponse de la core collection permet d'identifier des types de matériels et des origines de variétés qui se distinguent par le caractère étudié et reconnaître à l'intérieur de ces groupes des génotypes intéressants pour des programmes de sélection.

Cette core collection a été testée en contre saison en 2002 / 2003 au *CERAAS* à Bambey (Sénégal), en condition non limitante d'alimentation hydrique, avec un dispositif en blocs augmentés (10 blocs de 6 témoins communs plus 21 accessions différentes dans chaque bloc et présentes qu'une seule fois) ; 28 variables quantitatives et qualitatives ont été suivies. Les analyses de la variance de cet essai ont déjà été faites (B.Sine, 2003) .

En 2003 / 2004 la même core collection, toujours en contre saison et avec les mêmes 6 témoins a été semée à Bambey pour être soumise à un stress hydrique post floral, avec un dispositif constitué en 3 alpha-plans et un essai en blocs incomplets à 2 répétitions, les accessions de la core collection sont ainsi mis en 4 groupes de précocité ce qui assure que le stress hydrique intervient à la même phase physiologique, d'un groupe à l'autre un nombre limité de variétés communes servent de ponts pour comparer l'ensemble des accessions, ainsi les mêmes caractères que ceux de l'essai non stressé ont été suivis.

II - Matériel et méthodes :

1 - Présentation des données :

Le matériel végétal utilisé est constitué des 210 accessions de la core collection de sorgho provenant de 37 pays différents, et des 6 témoins : 211 : BF 201, 212 :B3042, 213 : IRAT204, 214 : Vacares, 215 : hybride medium DK18, 216 : tardif semence Provence (ARGENCE).

Les variables étudiées sont majoritairement des caractères agromorphologiques de la plante, on distingue :

Les paramètres liés à la longueur du cycle : La durée du semis à la floraison « dsflo » en jours, Le temps thermique du semis à la floraison « ttsflo » en degré jour, La durée de la floraison à la maturité « dflomat » en jours, Le temps thermique de la floraison à la maturité « ttflomat » en degré jour,

Les paramètres liés à la croissance : La hauteur du sol au sommet de la panicule « hspan » en cm, la quantité de matière sèche aérienne par pied « ms » en gramme, le taux de production de matière sèche aérienne par pied et par jour « txms » en gramme / jour, le taux de production de matière sèche aérienne par pieds et par jour « txmsd » en gramme/ degré / jour.

Les composantes du rendement : Le nombre des grains par pieds « ngr », le nombre de panicules par pieds « npan »,le poids des grains par pieds « pgr » en gramme, le poids de mille grains « pmgr » en gramme, le poids sec tiges et feuilles « pstf » en gramme.

2 - Dispositifs expérimentaux utilisés :

Pour cette étude les deux essais qui vont être considérés sont : l'essai en condition d'évapotranspiration maximale « ETM », ainsi que l'essai en condition de stress hydrique post floral « STR ».

L'essai ETM : a été mis en place selon un dispositif en blocs incomplets randomisés ou blocs augmentés : il se compose de 10 blocs, chaque bloc contient 21 accessions tirées au hasard dans les 210 accessions et présentes une seule fois dans l'essai plus les 6 témoins répartis au hasard et présents une seule fois dans chaque bloc (voir annexe).

L'essai STR : a été réalisé suivant un dispositif constitué de 3 petits essais en alpha plans à deux répétitions et un essai en bloc complet à deux répétitions, représentant 4 essais de précocités (ce qui a permis de décaler les dates de semis afin d'homogénéiser les périodes de floraison et d'appliquer le stress hydrique à la même date), le premier alpha plan contient 56 variétés réparties en 7 blocs de 8 parcelles à 2 répétitions, le second alpha plan contient 104 variétés réparties en 13 blocs de 8 parcelles à 2 répétitions, le troisième alpha plan contient 56 variétés réparties en 7 blocs de 8 parcelles à 2 répétitions le quatrième essai contient 16 variétés réparties en 2 répétitions.
D'un essai à l'autre un nombre limité de variétés communes servait de ponts pour comparer l'ensemble des accessions (voir annexe).

La méthode utilisée dans toute la suite est celle de l'analyse de la variance du regroupement de ces essais.(Philipe Letourmy, Eric Gozé, 1999).

3 - Analyse de la variance de chaque essai :

L'analyse de variance (ANOVA pour ANalysis Of VAriance) recouvre un ensemble de techniques de tests et d'estimations destinés à apprécier l'effet de variables qualitatives (facteurs) sur une variable numérique et revient dans le cas simple à comparer plusieurs moyennes d'échantillons gaussiens.(G.Saporta, 1990).

L'analyse de la variance appliquée à un plan d'expérience permet d'opérer une discrimination entre les facteurs selon qu'ils ont ou non une influence significative. (Aide –mémoire statistique. Edition CISIA)

L'analyse de la variance a été réalisée individuellement pour chaque essai, et cela dans le but de comparer les differentes variétés de Sorghos entre elles.Pour cela nous avons utilisé la procédure **Glm** du logiciel SAS.

a - L'essai ETM : Le programme SAS est le suivant :

```
     proc glm data=senegal;
class codvar Bloc;
model npan =codvar bloc;
lsmeans codvar;
run;
quit;
```

b - L'essai STR : Le programme SAS est le suivant.

```
     proc glm data=ceraas;
class codvar Essai Rep Bloc;
model ms = codvar Essai Rep(Essai) Bloc(Essai Rep);
lsmeans codvar;
run;
quit;
```

4 - Test d'Homoscédascticité des variances :

4 – 1 - Egalité des variances entre les 4 essais STR :

Afin de savoir si nous pouvons considérer les 4 petits essais de l'essai STR comme un seul essai, nous avons vérifié l'égalité des variances en utilisant le test de Fisher-Snédécor bilatéral. Après avoir récupéré les différentes variances de chacun des 4 essais, on fait le rapport de la plus grande sur la plus petite (s^2 max / s^2 min), cette valeur est comparée, dans une table de Fischer à une valeur théorique et doit lui être inférieure pour un seuil de risque choisi ; démarche dite de Bonferonni pour les variances au lieu des moyennes (à $1-\alpha / 6 = 0,995$) pour conserver l'hypothèse d'homogénéité des variances.

Le programme SAS est le suivant :

```
proc sort data=cerasas;
by Essai;
run;
proc glm data=cerasas;
by Essai;
class codvar Rep Bloc;
model pgr=codvar Rep Bloc(Rep);
run;
quit;
```

4 – 2 - Egalité des variances entre les essais ETM et STR :

De même pour tester l'égalité des variances entre les essais ETM et STR nous avons utilisé le test de Fisher-Snédécor bilatéral. Selon les résultats obtenues nous avons classé nos variables en trois groupes :

Le groupe A : où les variances sont homogènes entres les quatre essais STR et entre les essais ETM et STR .

Le groupe B : où les variances sont homogènes entres les 4 essais STR mais pas entre les 2 essais ETM et STR .

Le groupe C : où les variances sont hétérogènes.

Tableau 1: Classement des différentes variables selon la variance.

Homogénéité des variances	Variables	Procédure
Variances homogènes	1- pgr , ppan , ngr	A :Proc glm
Variances homogènes entres les 4 essais STR mais pas entre les 2 essais ETM et STR	2- pmgr, hspan, ttflom, dfloma, ttsflo, dsflo	B: Proc mixed (group=eau)
Variances hétérogènes	3- ms, tms, tmsd, pstf, npan	C: Proc mixed (group = essai*eau)

5 - Regroupement des essais ETM et STR :

Dans le but de chercher à obtenir une réponse des écarts entre les différentes variétés de Sorgho entre les deux environnements non stressé ETM et stressé STR , nous avons regroupé ces deux essais .

Pour cela nous avons procédé de la manière suivante :

- Création d'une nouvelle variable appelée **eau** avec 2 modalités représentant les 2 situations : **0** pour l'essai stressé STR et **1** pour l'essai non stressé ETM.
- Pour l'essai ETM, on considère la variable **essai** = 1 partout , la variable **rep** pour répétition à 10 modalités identiques aux 10 modalités de la variable **bloc**
- Selon les résultats des sections 4-1 et 4-2 nous utilisons une procédure SAS adaptée.

6 - Analyse de la variance du regroupement :

L'analyse différera selon les résultats de l'analyse de la variance de chaque essai individuellement .Les programmes SAS sont les suivants :

A : utilisation de la procédure **Glm** dans le cas d'égalité des variances entre les essais :

```
proc glm data= regrou;
class codvar eau essai bloc rep;
model  pgr=eau codvar eau*codvar essai(eau) rep(essai eau);
lsmeans  codvar*eau;
```

B : utilisation de la procédure **Mixed** qui tient compte de l'hétérogénéité des variances entre les essais:

```
proc mixed data= regrou;
class codvar eau essai bloc rep;
model   ms=eau   codvar   eau*codvar   essai(eau)      rep(essai   eau);
random  intercept / group = eau;
lsmeans  codvar*eau;
```

C : utilisation de la procédure **Mixed** qui tient compte de l'hétérogénéité des variances entre les essais:

```
proc mixed data= regrou;
class codvar eau essai bloc rep;
model    pgr=eau    codvar    eau*codvar    essai(eau)rep(essai    eau);
random  intercept  / group =   essai*eau;
lsmeans  codvar*eau;
```

7 - Intervalle de confiance :

Afin de détecter les variétés de sorgho où l'effet du stress hydrique est le plus significatif au niveau de la différence entre les essais stressé et non stressé nous traçons un intervalle de confiance à l'intérieur duquel le stress a le plus touché les variétés et cela grâce à l'option ***estimate*** de la procédure GLM ou Mixed ; cet intervalle est égale à 2 fois l'erreur standard de l'estimation.

III - Résultats :

L'analyse de la variance de l'essai non stressé indique qu'il y a une influence hautement significative du facteur variété (*pvalue* < 0.0001) pour chaque variable étudiée, par contre ce n'est pas le cas pour le facteur bloc, qui est significatif pour les variables de la longueur du cycle mais pas pour celles du rendement.(tableau 2).

Tableau 2: Résultats de l'analyse de la variance de l'essai non stressé.

	codvar		bloc	
	F	P	F	P
Pgr	10.69	<.0001	1.94	0.0706
Ngr	9.90	<.0001	1.18	0.3330
ppan	10.42	<.0001	1.93	0.0714
hspan	214.06	<.0001	2.57	0.0176
Pmgr	15.85	<.0001	1.05	0.4180
dfloma	3.45	<.0001	6.31	0.0001
ttfloma	3.14	<.0001	6.10	<.0001
dsflo	123.63	<.0001	2.37	0.0273
npan	401.94	<.0001	1.00	0.4540
Ms	34.80	<.0001	1.96	0.0672
Tms	21.24	<.0001	1.37	0.2283
tmsd	21.87	<.0001	1.48	0.1849
ntige	194.12	<.0001	1.00	0.4540
Pstf	158.05	<.0001	1.33	0.2476

Le model peut s'écrire : $Y_{ij} = m + \alpha_i + \beta_j + \xi_{ij}$

Où Y_{ij} est la réponse de la variété i du bloc j , m représente la moyenne générale, α_i l'effet variété (génotype) , β_j l'effet bloc et ξ_{ij} représente l'erreur aléatoire qui est la somme de l'erreur technique + l'erreur unitaire.

L'analyse de la variance de l'essai stressé a montré qu'il y a une influence hautement significative du facteur variété, pour chacune des variables étudiées, qu'il n'y avait pas d'effet du facteur bloc sauf pour la variable ttflom, pour les effets des facteurs essai et rep ils sont non significatifs pour les variables pmgr, dflomat, hspan et npan.(tableau 3).

Tableau 3: Analyse de la variance de l'essai stressé.

	codvar		essai		Rep (essai)		Bloc(essai*rep)	
	F	P	F	P	F	P	F	P
pgr	2.92	<.0001	1.28	0.2825	14.15	<.0001	0.77	0.8663
ngr	2.90	<.0001	1.28	0.2825	14.15	<.0001	0.77	0.8663
ppan	2.95	<.0001	1.59	0.1933	14.30	<.0001	0.87	0.7492
pmgr	1.22	0.0809	0.37	0.7783	0.44	0.7807	0.72	0.9176
hspan	7.72	<.0001	2.65	0.0504	23.46	<.0001	1.17	0.2275
ttfloma	5.67	<.0001	2.12	0.0987	9.61	<.0001	1.78	0.0031
dfloma	5.70	<.0001	1.12	0.3426	1.79	0.1317	1.02	0.4504
dsflo	11.77	<.0001	5.45	0.0013	5.79	0.0002	1.41	0.0535
ms	2.85	<.0001	2.48	0.0624	12.74	<.0001	0.93	0.6123
tms	2.86	<.0001	2.92	0.0355	14.02	<.0001	0.96	0.5524
tmsd	2.86	<.0001	2.92	0.0355	14.02	<.0001	0.96	0.5524
pstf	3.72	<.0001	3.27	0.0224	4.49	0.0017	1.05	0.3995
npan	1.54	0.0012	0.72	0.5400	3.10	0.0167	0.95	0.5731

Le modèle peut s'écrire : $Y_{ijkl} = m + \alpha_i + \beta_j + \delta_{jk} + \lambda_{jkl} + \xi_{ijkl}$

Où Y_{ijkl} est la réponse de la variété i de l'essai j, la répétition k et du bloc l , m représente la moyenne générale, α_i l'effet variété (génotype), β_j l'effet essai, δ_{jk} l'effet répétition à l'intérieur d'un essai, λ_{jkl} l'effet bloc à l'intérieur de chaque répétition de chaque essai et ξ_{ijkl} représente l'erreur aléatoire .

Comme le montre le tableau 4 l'analyse de la variance du regroupement a révélé un effet interaction génotype * environnement (codvar*eau) hautement significatif (p<.0001) et cela pour toutes les variables.

Tableau 4: Analyse de la variance du regroupement d'essais.

	eau		codvar		Codvar*eau		Essai (eau)		Rep(eau*essai)	
	F	P	F	P	F	P	F	P	F	P
pgr	539.76	<.0001	8.11	<.0001	2.93	<.0001	3.38	0.0188	5.77	<.0001
ngr	490.58	<.0001	6.93	<.0001	2.85	<.0001	3.49	0.0163	5.50	<.0001
ppan	504.08	<.0001	7.51	<.0001	2.66	<.0001	3.64	0.0132	5.72	<.0001
hspan	286.66	.	18.71	<.0001	2.28	<.0001	3.24	0.0225	8.34	<.0001
pmgr	7.38	.	1.58	0.0002	1.75	<.0001	0.78	0.5073	0.22	0.9984
dsflo	144.08	.	15.34	<.0001	2.97	<.0001	4.63	0.0035	2.06	0.0163
ttfloma	5.29	.	4.64	<.0001	2.61	<.0001	2.30	0.0779	4.72	<.0001
dfloma	16.87	.	5.23	<.0001	2.78	<.0001	0.85	0.4678	0.4678	0.0001
npan	18.46	.	5.68	<.0001	4.44	<.0001	1.02	.	1.18	0.2930
ms	1008.94	.	9.52	<.0001	3.48	<.0001	4.20	.	4.43	<.0001
tms	1255.45	.	8.67	<.0001	3.85	<.0001	4.20	.	4.89	<.0001
tmsd	1012.55	.	8.13	<.0001	3.43	<.0001	4.72	.	5.28	<.0001
ntige	0.00	.	4.10	<.0001	1.66	<.0001	3.09	.	3.08	0.0003
pstf	888.65	.	14.88	<.0001	4.86	<.0001	4.25	.	1.81	0.0414

Comme l'ANOVA que nous avons utilisé est non équilibrée (les effectifs des essais étant inégaux) , il a été nécessaire de réajuster les moyennes des traitements (codvar * eau) en fonction des blocs ou ils se trouvaient. Pour chaque variables; nous avons obtenu les moyennes ajustées grâce à la commande de SAS : ***lsmeans*** codvar * eau .

VI - Discussions et conclusion :

Sachant que l'ANOVA prévoit des hypothèses d'application dont la première est la normalité des populations considérées est vérifiée pour les deux essais, par contre l'homogénéité des variances entre les deux essais ne l'est pas pour toutes les variables étudiées ainsi, comme il a été précédemment dit dans le paragraphe 5, lors du regroupement des essais nous avons classé nos variables en 3 groupes :

Le groupe A : où les variances sont homogènes entres les quatre essais STR et entre les essais ETM et STR ; pour ce groupe l'utilisation de la *proc glm* est justifiée car il y a homogénéité des variances entre tous les groupes à comparer.

Le groupe B : où les variances sont homogènes entres les 4 essais STR mais pas entre les 2 essais ETM et STR ; comme il n'y a pas ici homogénéité des variances entre les deux essais ETM et STR nous considérons un effet aléatoire `group = eau`, et nous utilisons l'instruction `random / intercept group= eau,` de la procédure *mixed* qui est capable de prendre en compte des variances non homogènes.

Le groupe C : où les variances sont hétérogènes ; ici nous considérons un effet interaction essai*eau aléatoire et nous utilisons l'instruction `random / intercept group= essai*eau,` de la procédure *mixed* .

L'analyse de la variance du regroupement a révélé une interaction génotype * environnement hautement significative et cela pour toutes les variables (tableau 4), cela indique clairement que l'environnement joue un rôle primordial pour le bon développement de la plante.

Afin de pouvoir visualiser l'effet du stress sur les differentes variétés de Sorgho, nous avons tracé le graphique de la différence entre les moyennes ajustées de l'essai stressé et celles de l'essai non stressées en fonction de la moyenne des deux essais, et cela pour chaque variable étudiée.

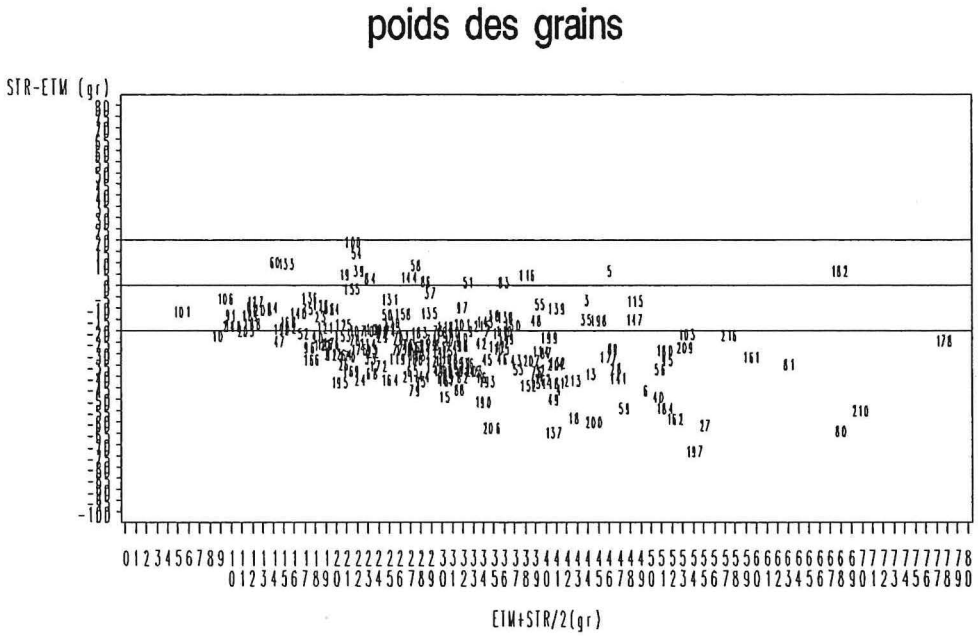


Figure 1:Répartition des variétés de Sorgho pour la variable poids des grains "pgr".

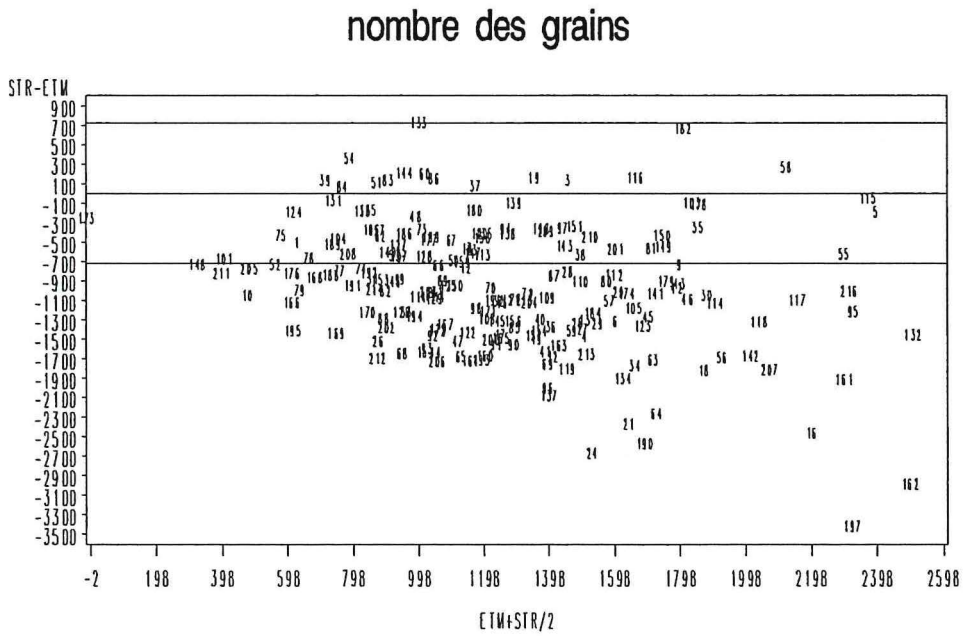


Figure 2: Répartition des variétés de Sorgho selon la variable nombre des grains « ngr »

Ainsi nous avons pu repérer quelques variétés de sorghos qui sont productives et statistiquement non interactives (il n'y a pas eu de baisse du rendement) après un stress hydrique post-floral, comme les variétés ayant le code : 3, 5, 7, 9, 30, 35, 38, 48,51, 55, 73, 83, 97, 110, 115, 116, 118, 138, 139, 143, 147, 150, 182, 198, 201.

Ce travail a permis de donner des indications quant à des variétés qui se comportent le mieux vis à vis d'un stress hydrique post-floral. En perspectives nous pourrions par la suite mieux modéliser l'interaction génotype*environnement et pouvoir ainsi prédire le rendement des variétés de sorghos à l'aide de covariables associées aux deux facteurs : pour le génotype des covariables qui permettent de caractériser les variétés de Sorgho selon leur sensibilité à la sécheresse comme l'indice de récolte, le coefficient de partition, ou le stay-green, ainsi que pour l'environnement des données climatiques comme la pluie et la température .

C – Deuxième partie : Interpolation des données climatiques sur l'île de la Réunion :

I - Introduction :

Cette deuxième partie du stage, s'intègre aux travaux de thèse en biostatistique de Melle Sabine LAURENT (CIRAD, UPR 13 : Aide à la décision et biostatistique) sur la modélisation de données longitudinales, appliquée à la richesse en sucre de la canne à sucre sur l'île de la Réunion.

La prévision de la richesse est un enjeu important pour l'industrie cannière Réunionnaise car la canne à sucre est sa première culture et l'une des principales ressources économiques de l'île.

Une bonne prévision de la richesse permettra une organisation de l'approvisionnement tenant compte des moments ou celle-ci est maximale, ceci provoquera un gain financier pour les planteurs car leur revenu dépend de la richesse des cannes livrées et pour les industries puisque la quantité de sucre extrait sera plus grande.

Le modèle de prédiction de la richesse établi dans le cadre de la thèse doit prendre en compte des facteurs climatiques. Les données climatiques disponibles sur l'île de la Réunion sont des données ponctuelles journalières par postes météos et pour le modèle on a besoin de données hebdomadaires.

L'objectif de cette partie est l'étude de la faisabilité d'une méthode régression la PLS (Partial Least Square) afin de pouvoir estimer des variables climatiques à l'échelle des ZECAS (Zones d'Etudes Canne A Sucre) et à l'échelle temporelle hebdomadaire, en tenant compte du relief de l'île : l'altitude, la latitude et la longitude.

Un premier stage a été mené au sein de l'unité par Mr Mohamed Chaouch (DEA bio-statistique), qui a modéliser la pluie mensuelle en fonction du relief sur l'île de la Réunion en comparant deux méthodes : la régression sur composantes principales (RCP) et La régression PLS. Ce stage a montré que la régression PLS donne de meilleurs résultats que la régression sur composantes principales.

Nous avons donc décidé de faire une régression PLS pour faire l'interpolation des données climatiques. Nous utiliserons le même principe c'est-à-dire faire la régression sur les altitudes des points voisins des postes météos.

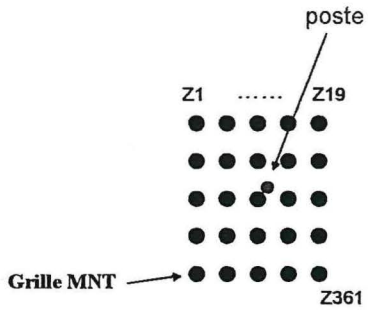
II – Matériel et Méthodes :

1 - Présentation des données disponibles :

1 - 1 - Données brutes :

a - Le modèle numérique de terrain (MNT) :

Un modèle numérique de terrain donne l'altitude des points d'une grille recouvrant un territoire ici toute l'île de la Réunion. Le MNT utilisé est celui de l'**IGN** (Institut Géographique National) à une échelle de 1 : 25000.
A partir du MNT nous avons pu extraire les altitudes des points voisins des postes. La grille était constituée de 361 points (19 x 19), de 475 mètre de coté et d'un pas de 25 mètre entre les points.



Nous avons 361 altitudes (Z1, Z2,...,Z361) autour des stations météos et elles vont constituer les variables explicatives.

Les travaux sur le MNT ont été effectués par Jean PARRIAUD informaticien spécialiste de SIG (Système d'Information Géo-référencé) du service.

b - Les variables climatiques à expliquer :

Ces données journalières sur 6 années sont fournie par Mrs : Mézino Michael et Pirot Roland de l'UPR5 (Système Cannier) de l'île de la Réunion qui gèrent les données provenant de 120 stations CIRAD et Météo France réparties sur toute l'île.

Les variables climatiques à expliquer disponibles sont des données ponctuelles journalières, pour les années de 1998 à 2003, et pour chaque poste météo (cela constituera les individus actifs) (figure 3)
La pluie mesurée en mm d'eau, par un pluviomètre.

La température : en degré Celsius, trois températures sont fournie la température maximale de la journée (Tmax), la température minimale de la journée (Tmin) et la température moyenne (Moyenne de tout les relevés de la journée donc différente de (Tmin+Tmax) / 2).

L'évapotranspiration potentielle (ETP) : en millimètre / jour cette variable est calculée par une fonction qui tient compte du jour de l'année, des températures, de l'altitude, de la latitude, de l'humidité, de la profondeur et de la capacité de chaleur sol.
Le rayonnement : en Joules/cm^2 cette variable est mesurée par un pyranomètre.

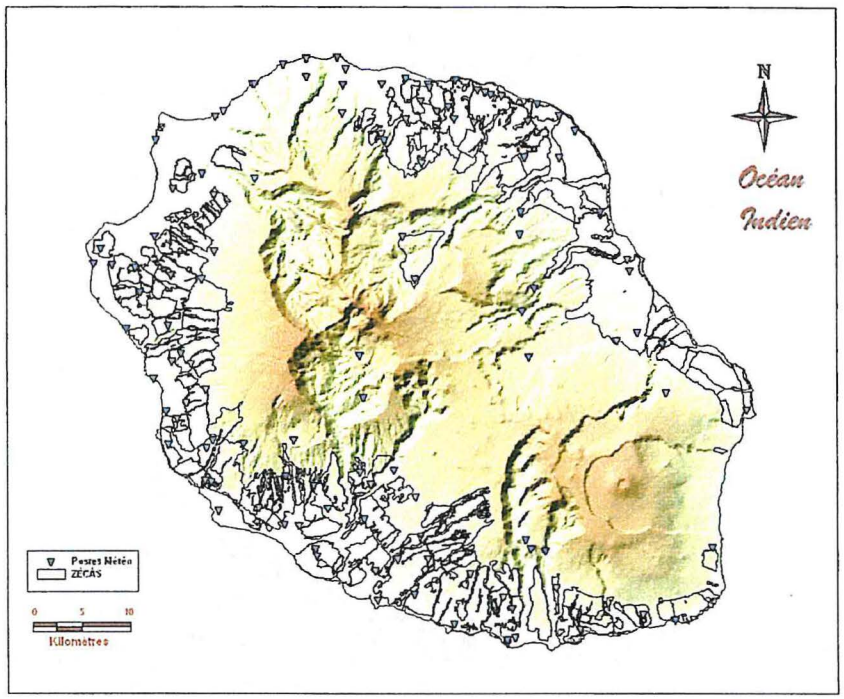


Figure 3: MNT, Postes météos et ZECAS

2 - Estimation demandée :

Notre objectif est d'interpoler les données météos sur les 82 ZECAS, par année et semaine, ce sont donc nos individus passifs. La ZECAS est l'échelle spatiale de travail de la thèse, elle représente une surface agricole homogène. (voir sur la carte de l'île de la Réunion figure 3 la délimitation des ZECAS et en triangle la localisation des postes météo).

3 - Obtention du jeu de données :

Pour la régression PLS nous avons besoins d'un jeu de données avec le N° de poste, les ZECAS, l'année, la semaine, la latitude, la longitude, les 361 altitudes du MNT autour des postes, des ZECAS, de la moyenne des altitudes par ligne, et enfin la variable climatique à expliquer.(tableau5).

Tableau 5: Forme du tableau du jeu de données final

Postes	ZECAS	Année	Sem	Lat	Lon	Z1	.....	Z361	mL	Var à exp
1015	-	1998	1	-22.5	55.3	235	....	325	251	10
....	-	...	...	...	...	...	...	...	...	...
97404	-	1998	53	-22.5	55.3	235	....	325	251	21
97405.	-	1999	1	-22.5	55.3	235	....	325	251	21
....	-	...	...	...	...	...	...	...	...	...
97501	-	1999	53	-22.5	55.3	235	....	325	251	52
....	-	...	...	...	...	...	...	...	...	...
97509.	-	2003	1	-22.5	55.3	235	....	325	251	52
....	-	...	...	...	...	...	...	...	...	...
97601	-	2003	53	-22.5	55.3	235	....	325	251	45
98401.	-	1998	1	-23.5	53.2	352	....	452	251	20
....	-	...	...	...	...	...	...	...	...	...
98409.	-	2003	53	-23.5	53.2	352	....	452	251	350
....	-	...	...	...	...	...	...	...	...	...
....	-	...	...	...	...	...	...	...	...	...
98601	-	2003	53	-23,8	53,8	325	...	485	254	345
-	1	1998	1	...	...	...	...	325	236	-
-	....	...	...	...	...	...	...	356	254	-
-	....	...	...	...	...	...	...	329	257	-
-	82	2003	53	...	...	...	...	453	261	-

Pour cela nous disposions dans une base de données ACCESS d'une table contenant pour chaque poste météo : Le N° de poste, la date, la semaine, l'année, La pluie (en mm), la température minimum, la température maximum et la température moyenne (en degré Celsius), l'ETP (en mm/jour) et du rayonnement (en joules/ cm^2), l'altitude la latitude et la longitude.

Nous avons extrait de cette table pour chaque variable à expliquer, une table contenant pour chaque semaine (53) de chaque année (6), la somme des pluies, ou la somme de l'ETP, ou la moyenne du rayonnement ou des températures journalières, la latitude et la longitude de chaque poste .

Grâce au MNT nous disposions d'un fichier EXEL contenant les 361 altitudes en lignes des points de la grille pour chaque poste météo que nous avons transposer pour avoir nos variables explicatives en colonnes (Z1....Z361), nous avons ensuite calculé les moyennes par lignes (mL) des altitudes des points voisins.

Nous disposons aussi d'un fichier ZECAS contenant pour les 82 ZECAS étudiées les parcelles qui y sont associées. Par parcelle nous disposons de la surface, ainsi que pour chaque géo-centre de parcelle les 361 altitudes des poins voisins. Là aussi nous avons du transposer ce fichier puis pour se ramener à l'échelle des ZECAS nous avons calculé les moyennes pondérées par la surface des Zi par ZECAS.

4 - Etude exploratoire :

Dans le but d'avoir une première représentation des données à étudier, nous avons tracé pour chaque poste et pour chaque année étudiée des courbes représentant les pluies, l'ETP, le rayonnement et les températures en fonction des 53 semaines de l'année pour chaque année étudiée.

Le programme SAS est le suivant :

```
proc gplot data=pluie;
     plot pluie*sem = annee;
     by poste;run;quit;
```

5 - La régression PLS :

La PLS (Word,1998; Tenenhaus,1998; Durand 2004) est une méthode de régression qui construit des modèles de prédictions , quand les variables explicatives sont très nombreuses ou fortement corrélées entres elles. Elle est spécialement utile lorsque le but étant la prédiction et qu'il n'y a pas de besoin pratique à limiter le nombre des facteurs à mesurer.

La régression PLS univariée (PLS1) est le cas ou il y a une seule variable **Y** à expliquer qui est un vecteur de dimension nx1, les variables explicatives sont constituées d'une matrice **Z** de dimension nxq.

La régression PLS1 est basée sur le calcul d'un jeu de α composantes orthogonales et non corrélées entre elles : $\mathbf{T} = [\ t_1,\ldots\ldots t_\alpha\ ]$ qui sont des combinaisons linéaires du tableau **Z** tel que les covariances entre $t_i = Z\ u_i$ et Y soit maximales.

Les composantes T sont alors utilisées comme de nouvelles variables explicatives. L'estimateur PLS étant obtenu par régression des moindre carrés de Y sur T.

Le programme SAS utilisé pour la régression PLS est le suivant :

```
proc pls data=donnee1 cv=one cvtest(stat=press seed=1982787);
     by annee sem;
     model sompluie=z1-z361 x y m;
     output out=lesychap predicted=ychap yresidual=yres xresidual=xres
xscore=xscr yscore=yscr ;
     run;
```

Choix du nombre de facteurs PLS :

Le choix du nombre des facteurs peut se faire en utilisant la validation croisée. Dans notre cas le choix du nombre de facteurs PLS est réalisé grâce à la statistique du PRESS (Predicted REsidual Somme of Squares) qui est basée sur les résidus générés par ce procédé de validation croisée. L'option (CVTEST STAT=PRESS) sélectionne le plus petit modèle ayant une statistique PRESS non significativement supérieur au PRESS minimum absolue.

Dans cette partie nous allons construire des modèles PLS (53 modèles par année d'étude) avec comme variables explicatives les 361 altitudes de la grille MNT (points voisins) qui entourent chaque poste météo, la moyenne ml des points voisins ainsi que les latitudes et les longitudes des postes, et comme variable réponse la somme des pluies et de l'ETP, le rayonnement moyen et les températures moyennes hebdomadaires et cela pour les années de 1998 à 2003, et avec comme individus actifs les postes météos et comme individus passifs les ZECAS (Zone d'Etude Canne A Sucres).

III- Résultats :

1 – Etude exploratoire des données :

D'après les graphiques ci-dessous nous remarquons que les variables ETP, températures et rayonnement ont une forme particulière d'une années à l'autre c'est à dire quelles ont une valeur maximale en début d'année, elle chute ensuite pour les semaines du milieu d'année et remonte enfin vers la fin de l'année. Par contre la variable pluies est quand à elle beaucoup plus elle est assez différente d'une année à l'autre.

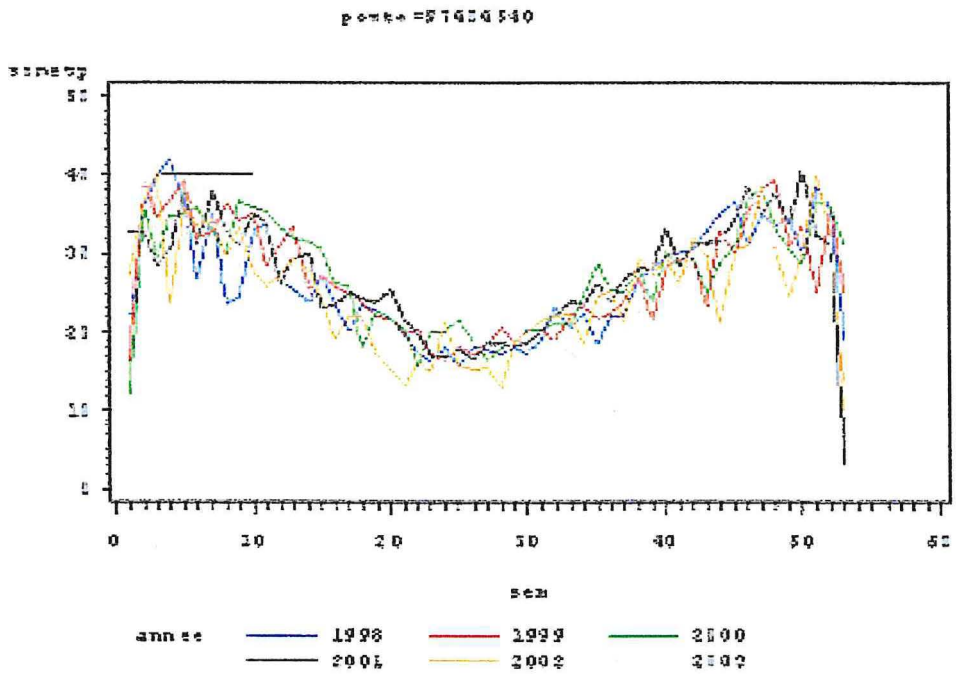


Figure 3: Etp hebdomadaire du poste 97404540

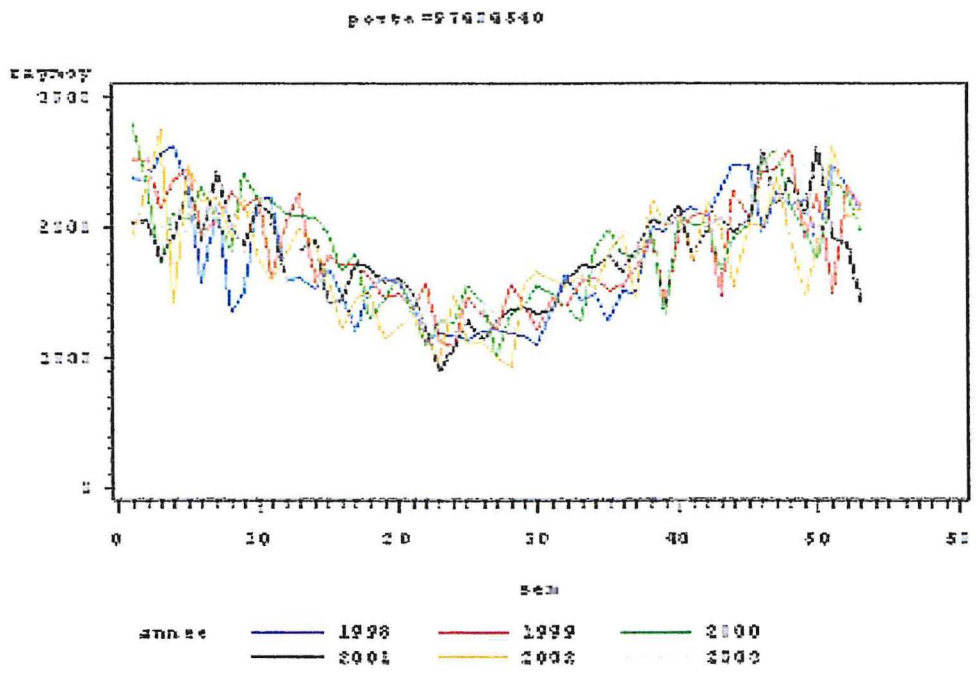


Figure 4: Rayonnement hebdomadaire du poste 97404540.

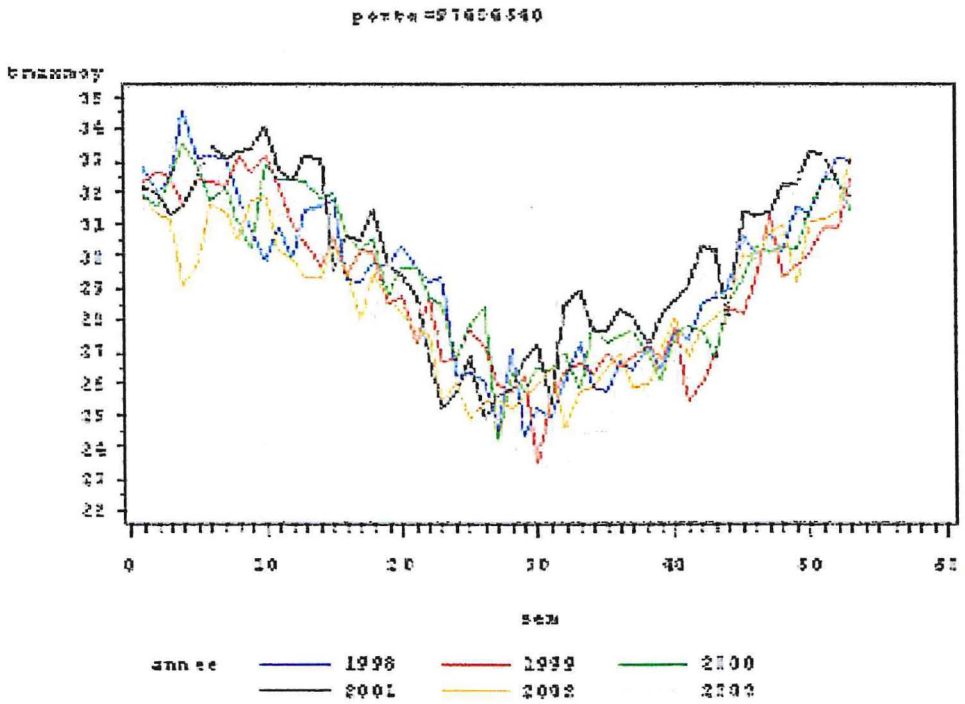


Figure 5: Températures maximales hebdomadaires du poste: 97404540.

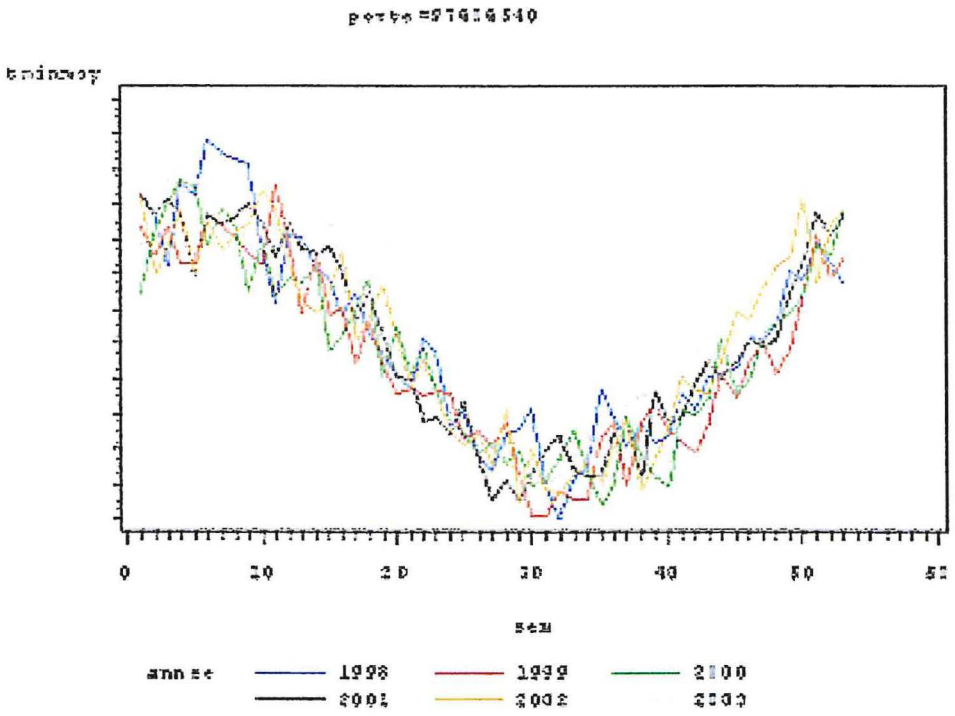


Figure 6: Températures minimales hebdomadaires du poste: 97404540.

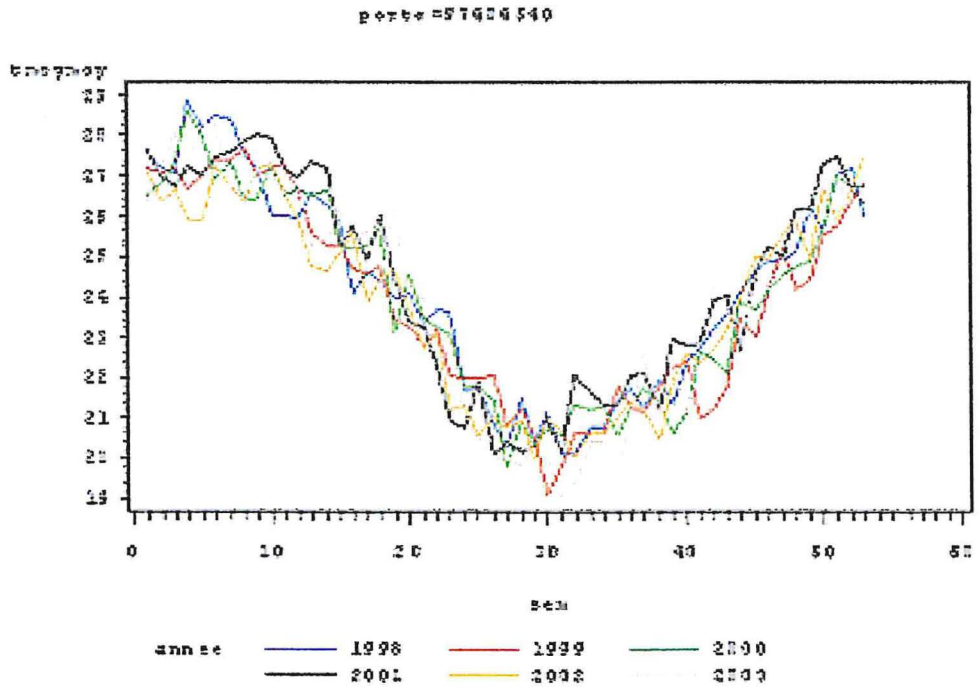


Figure 7: Températures moyenne hebdomadaires du poste: 97404540.

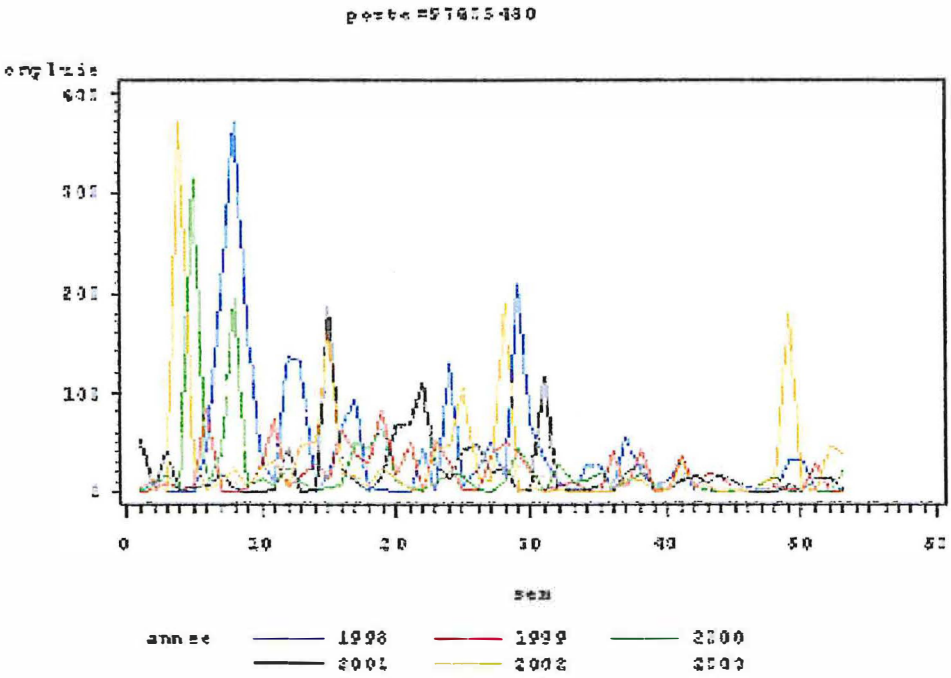


Figure 8: Pluies hebdomadaires du poste: 97405480.

2 – Régression PLS :

La régression PLS donnent pour chaque variable un modèle par semaine et par année. Certains modèles sont valides car ils ont un nombre de composantes PLS différent de zéro et d'autre non car ils n'ont pas de composantes dans ces cas, la régression PLS, considère que la moyenne des valeurs observées suffit à expliquer les variations de ces valeurs.

Le tableau 6, résume le nombre de modèles PLS avec un nombre de facteur à extraire différent de 0. C'est à dire le nombre de fois où la PLS considère qu'il y a au mois une composante PLS.

Année	Pluie	Log(pluie+1)	ray	Etp	tmax	tmin	tmoy
1998	36	35	6	1	35	53	48
1999	26	26	6	0	35	48	49
2000	30	30	3	1	39	50	52
2001	21	21	2	1	33	39	52
2002	21	21	7	1	49	50	50
2003	17	17	13	10	53	52	53
Total	125	124	13	14	214	292	304
%	39,3%	38,99%	4,02%	8,27%	67,29%	91,82%	95,59%

Tableau 6: Nombre de modèles avec un nombre de facteurs à extraire différent de 0.

a - Le rayonnement et ETP :

Pour les variables rayonnement et ETP nous obtenons que très peu de modèles qui expliquent la variation de ces variables avec au moins une composante PLS (4,02% et 8,28%).
L'exemple de la variable ETP pour l'année 1999 semaine 10 (tableau 7) montre qu'il n'y a pas de composantes PLS par conséquent pas de modèle PLS valide car il donne comme prédiction la valeur moyenne de l'ETP (figure 8).

Tableau 7 :Analyse PLS de l'ETP (année 1999 semaine 10) par validation croisée.

Cross Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS	Prob > PRESS
0	1.043478	0.3190
1	0.981988	1.0000
2	1.066617	0.3700
3	1.226078	0.3070
4	1.205153	0.3260
5	1.271178	0.4400
6	1.295838	0.3660
7	1.3839	0.1970
8	1.565536	0.1320
9	1.772109	0.0800
10	1.825661	0.0970
11	1.830648	0.0890
12	1.783332	0.0790
13	1.754068	0.0820
14	1.709026	0.0910
15	1.704575	0.0960

Minimum root mean PRESS	0.9820
Minimizing number of factors	1
Smallest number of factors with p > 0.1	0

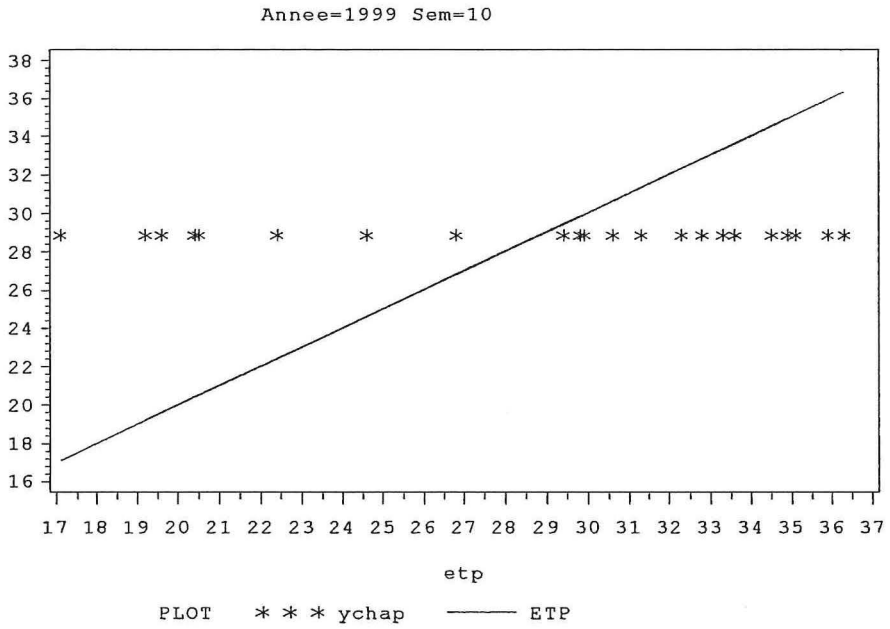


Figure 8 : Valeurs prédite de l'ETP en fonction des valeurs observées.

b- la pluie :

Pour la variable pluie nous remarquons qu'il y a 39,3% des modèles qui donnent une explication de la pluie avec au moins une composante PLS, certains modèles ont un bon pourcentage d'explication de la pluie d'autres un peu moins.

En prenant l'exemple de la prédiction de la pluie, de la semaine 10 de l'année 1998, les résultats de la validation croisée (Tableau 8) montrent que la procédure a sélectionnée un modèle avec 8 composantes PLS car c'est le plus simple modèle avec une statistique PRESS non significativement différente de la minimale du PRESS. Ce modèle PLS explique plus de 82,43 % de la variation de la pluie, ainsi que 92,28 % de la variation des prédicteurs.

Cross Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS	Prob > PRESS
0	1.010204	<.0001
1	0.897658	<.0001
2	0.894868	<.0001
3	0.86388	<.0001
4	0.887165	<.0001
5	0.857842	<.0001
6	0.799507	0.0090
7	0.746869	0.0650
8	0.61202	0.2470
9	0.600107	1.0000
10	0.648882	0.0380
11	0.697716	0.0190
12	0.726645	0.0030
13	0.69889	0.0130
14	0.722599	0.0040
15	0.715795	0.0130

Minimum root mean PRESS	0.6001
Minimizing number of factors	9
Smallest number of factors with p > 0.1	8

Percent Variation Accounted for by Partial Least Squares Factors

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	31.7922	31.7922	25.6021	25.6021
2	36.1838	67.9759	3.6122	29.2142
3	10.8700	78.8459	6.0896	35.3039
4	3.7545	82.6004	18.5397	53.8435
5	4.0784	86.6788	14.2185	68.0620
6	1.6716	88.3504	9.0860	77.1480
7	1.4294	89.7798	4.1802	81.3282
8	2.5080	92.2878	1.1068	82.4350

Tableau 8: Analyse PLS de la pluie (année 1998 semaine 13) avec validation croisée.

Afin de tester la qualité de ce modèle avec 8 composantes ; nous traçons le graphique des valeurs prédites par le modèle en fonction des valeurs observées. (Figure 9), nous remarquons que ce modèle prédit assez bien la pluie puisque il y a peu d'individus s'éloignant de la bissectrice.

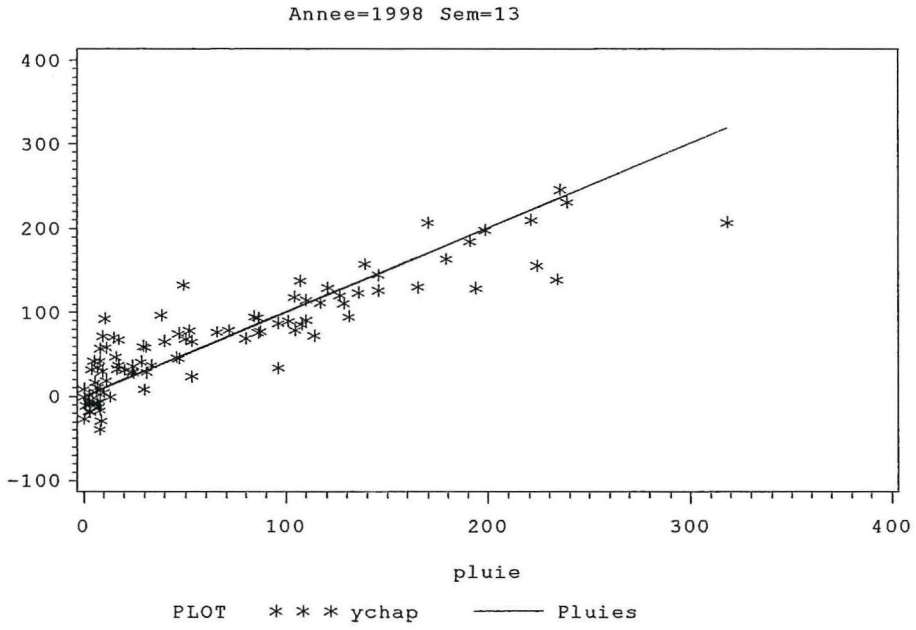


Figure 9: Valeurs prédites de la pluie en fonction des valeurs observées.

c- les températures :

Enfin pour les variables températures maximale, minimale et moyenne nous obtenons 67,29 % 91,82% et 95,59% de modèles qui expliquent la variation de ces variables avec au moins une composante PLS.
Prenons l'exemple des températures minimum de la semaine 38 de l'année 2002 les résultats de la validation croisée (Tableau 9) montrent que la procédure a sélectionnée un modèle avec 9 composantes PLS. Ce modèle PLS explique plus de 97.83 % de la variation de la température minimum, ainsi que 96.17% de la variations des prédicteurs.

Cross Validation for the Number of Extracted Factors		
Number of Extracted Factors	Root Mean PRESS	Prob > PRESS
0	1.029412	<.0001
1	1.020068	<.0001
2	1.058545	<.0001
3	1.287885	<.0001
4	1.022152	<.0001
5	0.747996	<.0001
6	0.568043	0.0020
7	0.521327	0.0070
8	0.495427	0.0130
9	0.416397	0.1120
10	0.402957	1.0000
11	0.440901	0.0070
12	0.421648	0.2840
13	0.472854	0.0190
14	0.539831	<.0001
15	0.563019	0.0020

Minimum root mean PRESS	0.4030
Minimizing number of factors	10
Smallest number of factors with p > 0.1	9

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	35.7639	35.7639	13.2805	13.2805
2	34.5425	70.3064	4.5219	17.8024
3	6.1889	76.4953	35.7036	53.5060
4	11.8908	88.3861	12.1321	65.6381
5	2.4487	90.8348	17.9532	83.5913
6	1.5425	92.3773	10.1880	93.7793
7	2.6533	95.0306	2.0886	95.8680
8	0.5310	95.5616	1.3603	97.2283
9	0.6158	96.1774	0.6033	97.8315

Tableau 9: Analyse PLS de la température minimum (annéé 2002 semaine 38) avec validation croisée.

Pour tester la qualité de ce modèle avec 9 composantes ; nous traçons le graphique des valeurs prédites par le modèle en fonction des valeurs observées. (Figure 10), nous remarquons que ce modèle prédit très bien la température m puisque il n'y a pas d'individus qui s éloignent de la bissectrice.

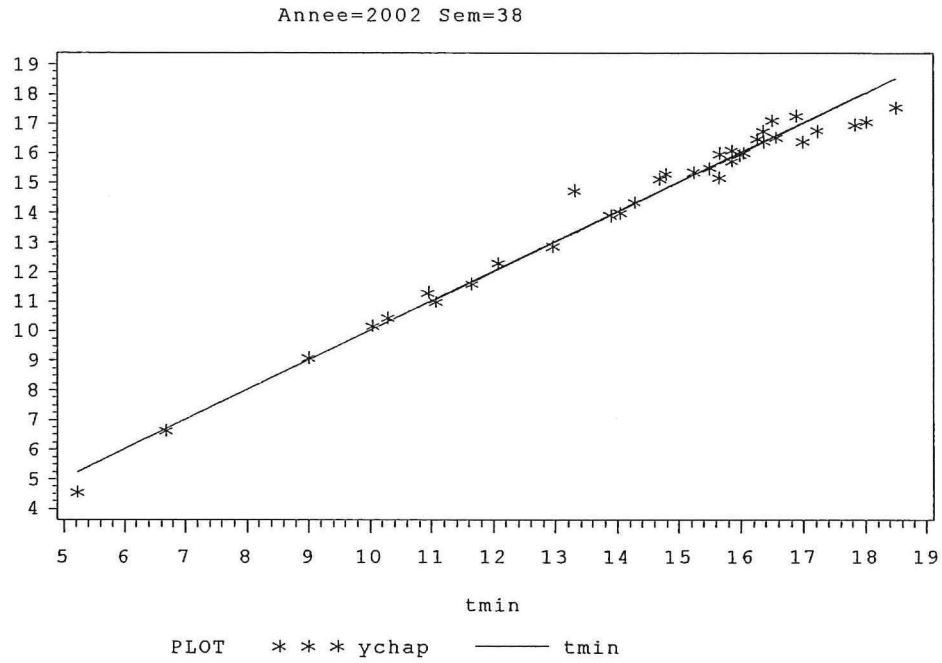


Figure 10: Valeurs prédites de la température minimum en fonction de valeurs observées.

IV- Discussion et conclusion :

L'objectif de cette partie du stage étant l'étude de faisabilité de la régression PLS est sans doute atteint, car avec les altitudes du MNT autours des stations météos et des ZECAS prises comme variables explicatives pour prédire certaines variables climatique hebdomadaire comme la pluie ou les températures, la PLS donne des prédictions à l'échelle hebdomadaire. Mais pour les variables ETP et rayonnement solaire la PLS donne très peu de prédictions, il faudra entre autres soit changer les variables explicatives ou en rajouter d'autres, ou changer de méthode de régression.

Nous avons mis en évidence les limites des méthodologies déjà existantes concernent le problème de l'interpolation des données climatiques. En perspective il sera important qu'une étude plus approfondie en statistique notamment en géostatistique soit menée afin de développer une méthodologie plus appropriée.

BIBLIOGRAPHIE

[1] J. Chantereau et R. Nicou (1991). Le Sorghos. Le technicien d'agriculture tropical. ACCT- Paris, CTA-Wageningen.

[2] B.Sine (2003). Evaluation d'une core collection de Sorgho en condition de déficit hydrique pré-floral. Rapport de DEA .Université Cheikh Anta Diop de Dakar.

[3] P.Letourmy, révision Eric Gozé (1999). Expérimentation Agronomique Planifiée. Support de cours, CIRAD, 50 p.

[4] G.Saporta (1991). Probabilités Analyse des données et statistique. Editions Technip, 493p.

[5] Aide –mémoire statistique. Edition CISIA*CERESTA,.285 p.

[6] G Philippeau (1989) Théorie des plans d'expériences application à l'agronomie. Service des études statistiques de l'I.T.C.F.

[7] M. Tenenhaus (1998) La Régression PLS .Edition TECHNIP.

[8] M. Chaouch (2005) Mémoire de DEA Interpolation mensuelle de la pluie sur l'ile de la Réunion. CIRAD de Montpellier.

[9] J.F.Durant (2004) Calcul matriciel et analyse factorielle des données. Université de Montpellier2.

ANNEXE

Randomisation de l'essai ETM :

bloc 1	bloc2	bloc3	bloc4	bloc5	bloc6	bloc7	bloc8	bloc9	bloc10
8	9	1	16	17	3	10	5	13	2
14	25	4	22	39	27	11	12	33	7
19	26	6	35	50	32	15	30	40	34
20	29	18	44	65	43	24	46	59	47
21	38	23	53	75	45	28	49	62	51
37	52	31	58	84	48	36	63	70	61
66	56	42	60	86	57	41	81	77	71
72	68	54	64	96	90	55	85	83	89
74	73	80	76	98	91	67	113	100	93
101	82	87	109	114	95	69	115	125	102
103	88	92	116	120	104	78	129	126	106
110	99	94	119	121	123	79	136	131	122
112	134	97	130	124	127	107	141	132	128
133	140	105	138	153	137	117	167	142	146
135	144	108	151	156	160	143	172	145	150
139	152	111	162	177	168	170	176	147	155
149	157	118	164	179	178	171	180	158	159
154	166	148	165	186	185	182	192	169	163
195	174	161	173	187	197	205	200	183	175
207	184	194	181	193	201	206	202	189	190
208	188	204	196	199	210	209	203	198	191
211	211	211	211	211	211	211	211	211	211
212	212	212	212	212	212	212	212	212	212
213	213	213	213	213	213	213	213	213	213
214	214	214	214	214	214	214	214	214	214
215	215	215	215	215	215	215	215	215	215
216	216	216	216	216	216	216	216	216	216

Essai STR :

Randomisation de l'essai 1 :

			Blocs						
Essai	Rép.	Parcelles	1	2	3	4	5	6	7
1	1	1	78	33	65	173	61	193	190
		2	195	164	162	93	34	13	202
		3	171	191	192	189	160	165	147
		4	215	170	15	90	17	89	98
		5	14	188	176	175	214	148	2
		6	168	203	47	208	172	166	111
		7	101	1	211	10	68	212	194
		8	64	26	205	94	161	169	44
	2	1	188	175	90	208	44	101	205
		2	93	195	26	190	33	47	61
		3	65	162	194	203	161	170	1
		4	193	211	172	168	13	68	111
		5	147	2	64	17	148	212	189
		6	171	164	173	176	215	98	214
		7	34	160	192	169	15	10	165
		8	191	89	166	202	94	78	14

Randomisation de l'essai 2 :

			Blocs												
Essai	Rép.	Parc.	1	2	3	4	5	6	7	8	9	10	11	12	13
2	1	1	159	11	18	50	79	97	40	158	92	163	129	41	162
		2	36	110	118	186	30	142	13	66	136	45	122	167	109
		3	70	113	91	181	145	200	71	32	67	199	8	72	206
		4	77	9	213	134	119	149	96	75	184	201	112	155	82
		5	153	126	16	99	22	69	128	24	46	90	85	78	183
		6	135	216	156	204	74	31	143	106	62	157	196	140	187
		7	76	12	114	188	49	73	154	152	25	87	197	146	4
		8	121	207	52	102	151	137	105	125	120	88	6	21	108
	2	1	113	11	197	196	142	201	4	184	105	154	135	50	71
		2	67	24	206	187	36	128	151	149	32	157	143	146	85
		3	213	114	199	118	12	183	18	134	77	79	52	76	106
		4	97	159	136	13	40	110	200	108	45	140	75	87	163
		5	158	8	207	155	125	122	216	9	145	102	25	69	181
		6	22	73	156	126	74	72	112	90	62	82	137	49	167
		7	162	46	99	88	16	186	92	129	21	66	41	96	109
		8	6	119	78	204	120	121	188	91	31	153	30	152	70

Randomisation de l'essai 3 :

		Blocs							
Essai	**Rép.**	**Parcelles**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3	1	1	210	20	127	209	27	104	63
		2	177	7	124	28	150	23	86
		3	51	43	198	37	185	133	58
		4	57	178	29	131	84	117	35
		5	130	141	138	53	132	151	42
		6	118	55	144	107	59	56	95
		7	139	179	39	48	136	123	3
		8	80	159	158	174	83	38	149
	2	1	139	27	123	84	132	209	37
		2	159	53	177	133	48	210	39
		3	107	95	118	198	51	58	131
		4	178	29	124	130	56	23	141
		5	185	151	174	35	179	83	150
		6	86	20	59	28	63	149	57
		7	38	80	43	158	127	7	104
		8	138	117	42	55	136	144	3

Randomisation Essai 4 :

R1 : 115 100 182 5 198 54 60 19 131 103 81 116 180 131(2) 178 198(2)
R2 : 5 81 115 103 180 198 131 60 100 54 178 19 198(2) 182 131(2) 116

Programmes SAS

Graphiques figure 1 et 2 :

```
% include "regrou.sas";
ods rtf file="glm.rtf";
goptions reset=all;
ods output lsmeans=lsm;

proc glm data=regrou;
     class codvar eau essai bloc rep;
     model pgr=eau codvar eau*codvar essai(eau)   rep(essai eau); /*
bloc(essai rep eau); */
     lsmeans codvar*eau;

estimate 'effet eau sur var1'  eau -1 1 eau*codvar -1 1 0;
estimate 'effet eau sur var3'  eau -1 1 eau*codvar 0 0 0 0 -1 1 0;
estimate 'effet eau sur var37' eau -1 1 eau*codvar 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0 0
                                                   0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0 0
                                                   0 0 0 0 0 0 0 0 0 0
0 0 -1 1 0;
.....................................................................
run;
quit;

proc sort data=lsm;
     by  eau codvar;
run;

  data lsm0;
set lsm(keep=codvar eau pgrLSMEAN) ; esti0=pgrLSMEAN;
if eau=0;
run;

data lsm1;
set lsm (keep=codvar eau pgrLSMEAN);  esti1=pgrLSMEAN;
if eau=1;
run;

data differ (drop= pgrLSMEAN eau) ;
merge lsm1 lsm0;
by codvar;

data differ1;
set differ;
dif=esti0  -esti1;  moy=(esti1  +  esti0)/2  ;inverse=  dif  >0;  stress=
dif>=-20 and dif<0 ; productif =dif>-20 and  moy>=29;
run;
data differ2;
set differ1 ( keep= codvar inverse stress productif)   ;
run;

data annote ( keep=  x y xsys ysys text position size);
set differ1;
x=moy;  y=dif;  xsys='2';  ysys='2';  text=put(codvar,  10.);  size=0.7;
position='5';
```

```
proc gplot data=differ1;  title'pgr';
title 'poids des grains';

goptions symbol1 v=none;
     plot dif*moy/vref=0 vref=20 vref=-20   annotate=annote vaxis=axis1
haxis=axis2;symbol1 v=none;

axis1 label=('difference=str-etm') order=(-100 to 80 by 5) minor=none ;
axis2 label=('moyennes') order =(0 to 80 by 1) minor=none  ;
run;
quit;
ods rtf close;
```

Récupération du fichier individus supplémentaires:

```
/* Recuperation du fichier des individus supplémentaire zecasinsupmpas
*/

/* Transposition du fichier parcelle pour avoir toutes les alt en
variable*/

proc transpose data= parcelle out=parcelletrans prefix=z;
by parcelle;
id grille;
var z ;
run;

/* Calcul des moyennes par ligne et centrage en ligne*/

data parcellecent;
        set  parcelletrans (keep = parcelle z1-z361);
        m=mean(of z1-z361);
        array z(361) z1-z361;
        do i=1 to 361;
        z(i)=z(i)-m;
        end ;
        drop i;run;


/* classer le fichier zecas par parcelle*/
proc sort data=zecas out= zecasort;
by parcelle;
run;


/* fusion du fichier parcelle et zecas */
data zecasinsupint;
merge parcellecent zecasort;
by parcelle;
run;

/* fusion de zecasinsupint avec moylatlon*/
data zecasinsup;
merge zecasinsupint moylatlon;
by parcelle;
run;

/* Tri du fichier des zecas parcelles par zecas pour pouvoir fair la
moy pondere */
proc sort data=zecasinsup  out=zecasinsupsort;
```

```
by zecas;
run;

/* calcul des alt des point voisin par zecas par moy pond par la
surface*/
 proc means data= zecasinsupsort;
        var m Lat Lon z1-z361;
        weight surf;
        by zecas;
        output out=zecasinsupmp mean=m lat lon z1-z361 ;
        run;

/* Rajout des annee et sem dont on a besoin */
data zecasinsupmpas;
       set zecasinsupmp (keep = zecas m lat lon z1-z361);
       do annee= 1998 to 2003;
       do sem= 1 to 53;
       output;
       end ;
       end;
       run;
```

Régression PLS :

```
proc transpose data=vexpluie out=vexpluietrans ;
    by poste;
    id variable;
    var alt;
run;


data vexpluiecent;
   set vexpluietrans (keep = poste Z1-Z361);
   m=mean(of z1-z361);
   array z(361)  z1-z361;
   do i=1 to 361;
      z(i)=z(i)-m;
   end;
   drop i;
run;

data vexpluielatlon;
merge vexpluiecent latlonpluie;
run;

proc sort data=donpluie out=donpluiesort;
by poste;
run;

data  donplspluieint;
   merge donpluiesort vexpluielatlon;
   by poste;
run;

data donplspluie;
set donplspluieint zecasinsupmpas;
run;

proc sort data=donplspluie out=donplspluiesort;
 by annee sem;
run;
```

```
proc pls data= donplspluiesort cv=one cvtest(stat=press seed=1982787);
     by annee sem;
     model sompluie = z1-z361 Lat Lon m;
output out=lesychappluie predicted=ychap  yresidual=yres xresidual=xres
xscore=xscr  yscore=yscr  ;
run;

ods rtf close;
```

Graphiques figures 8 , 9 et 10 :

```
% include "resulttmoy.sas";
ods rtf file="graph.rtf";

proc sort data=resultpluie  out=result1;
by annee sem;
run;

symbol2 v=star  c=black ;
symbol3  i=join c=red;
axis3 label=none;
axis4 label=('pluie');

proc gplot data=result1;

plot (ychap pluies) * pluies/ overlay legend  haxis=axis4  vaxis=axis3;
by Annee Sem;
run;

quit;
```

Résumé

Ce stage du DESS MSIAAP (Méthodes statistiques appliquées aux industries agronomiques agroalimentaires et pharmaceutiques) était divisé en deux grandes parties :

La première partie : Interaction Génotype*Environnement du Sorgho entre dans le cadre d'une évaluation d'une « core-collection » de sorgho constituée de 210 variétés de sorgho qui représentent la diversité génétique de la collection mondiale, sous différents régimes hydriques : en condition non limitante d'alimentation hydrique, en condition de stress hydrique pré et poste floral. L'analyse a permis d'identifier les variétés de sorgho les plus résistantes au stress hydrique.

La seconde partie : Interpolation des données climatiques sur l'île de la Réunion, les variables climatiques étant : la température, le rayonnement solaire, l'évapotranspiration potentielle, et la pluie, une méthode de régression la PLS (Partial Least Square) nous a permis de proposer une estimation de ces variables climatiques à l'échelle de ZECAS (Zone d'Etude de la Canne A Sucre) et à une échelle temporelle hebdomadaire, en tenant compte de variables géographiques: l'altitude (modèle numérique de terrain : MNT), la latitude et la longitude.

Abstract

This training course of the DESS MSIAAP was divided into two parts :

The first part : Genotype* Environment Interaction of sorghum is in the frame of an evaluation of a core collection of sorghum composed of 210 sorghum varieties that represent the genetic diversity of the worldwide collection, under different soil moisture regimes : in the case of non limited hydric alimentation, in the case of pre and post floral hydric stress. Statistical analysis allowed us to identify varieties of Sorghum who are the best resistant to the hydric stress.

The second part: Interpolation of climatic data in Réunion island, the climatic variables are temperature, solar radiation, evapotranspiration and the rain, a PLS (Partial Least Square) regression method allowed us to propose an estimation of the climatic variables at the map scale of ZECAS (Zone d'Etude Canne A Sucre) and at a weekly time scale, and taking account of geographical variables : the altitude (Digital Terrain Model: DTM), the latitude and the longitude.