Université Paris-Dauphine, Mémoire d'actuariat

ETUDE DE FAISABILITE D'UNE ASSURANCE RENDEMENT BASEE SUR INDICE CLIMATIQUE

présenté et soutenu en octobre 2011 par

ERWAN KOCH

Stage effectué au sein du Laboratoire de Météorologie Dynamique
C.N.R.S., Jussieu
et encadré par M.Filipe Aires
Mémoire rédigé sous la direction de M. Christian Robert.

Remerciements

Cette étude a été réalisée au Laboratoire de Météorologie Dynamique sur le site de Jussieu. Je remercie donc l'ensemble des personnes (chercheurs, ingénieurs, thésards...) qui m'ont particulièrement bien accueilli. L'ambiance s'est avérée très ouverte et chaleureuse. J'ai notamment été frappé par la grande diponibilité de tous, et ce qu'il y ait ou non un lien entre leur domaine et le sujet de mon étude.

Je remercie bien évidemment chaleureusement mon maître de stage Filipe Aires pour sa grande disponibilité, son attention à mon égard, ses compétences et enfin ses conseils avisés.

Un grand merci également à Frédérique Cheruy pour avoir effectué des simulations sur le Maroc (ce qui m'a permis d'utiliser des sorties du modèle LMDZ) ainsi que pour ses explications sur les composantes du modèle.

Enfin, un grand merci à Jacques Lefrère et Robert Fournisseur pour leur aide précieuse au niveau informatique.

Je remercie également chaleureusement mon directeur de thèse Christian Robert pour ses précieux conseils.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Abstracts

Mots-clés : assurance rendement, indice climatique, théorie de l'apprentissage statistique, techniques de régularisation, modèle climatique

Dans un contexte de changement climatique, il apparaît de plus en plus important de fournir aux agriculteurs des moyens de se protéger face aux risques naturels. Ceci est d'autant plus vrai dans un univers agricole mondial en pleine mutation (contraintes environnementales, mondialisation) les obligeant à investir dans de nouvelles technologies. C'est dans le cadre du plan Maroc Vert supposé répondre à ces nouveaux défis que la coopérative agricole COSUMAR a demandé une étude de faisabilité d'une assurance rendement dans le domaine de la production de betterave sucrière. L'idée que nous proposons ici est une assurance basée sur un indice climatique. Comme nous allons le voir, des solutions d'assurance traditionnelle (assurance basée sur les rendements effectifs) seraient très difficiles à mettre en place compte-tenu des risques d'anti-sélection et d'aléa moral ainsi que des coûts élevés de gestion des sinistres.

Afin de limiter le risque de base, l'indice climatique doit rendre compte au mieux du rendement réel. Nous cherchons donc la meilleure combinaison de variables prédictives et testons différents modèles reliant celles-ci aux rendements (modèle linéaire, non linéaire, hiérarchique). L'une des difficultés principales concerne la petite taille des échantillons de données. Nous mettons donc en place des techniques de régularisation permettant de rendre les modèles développés robustes, c'est-à-dire les plus efficaces possibles sur des données climatiques indépendantes de la base d'apprentissage. Nous nous attachons à évaluer de manière la plus honnête et objective possible la performance de notre produit d'assurance.

Nous étudions également brièvement la possibilité d'utiliser un tel indice comme sousjacent de produits financiers tels les dérivées climatiques ou les obligations catastrophes.

Enfin, l'étude est réalisée à partir de données climatiques issues de mesures de stations mais nous analysons la pertinence de l'utilisation d'autres sources de données telles les analyses de l'atmosphère et les sorties de modèles climatiques. Nous décrivons notamment les techniques de downscaling et d'upscaling.

Key words : yield insurance, climatic index, statistical learning theory, smoothing methods, climate models

In a context of climate change, it is particularly important to provide farmers with tools to cover meteorological risks. This is specifically true in a rapidly changing agricultural environnement (ecological constrains, globalisation) that obliges them to invest in new technologies. Morocco has elaborated a plan called "Maroc Vert" to face these new challenges. In this context, the cooperative COSUMAR has been looking for a study concerning the possibility of a yield

insurance in the field of sugar beet production. The solution we propose is an insurance based on a climatic index. As we will see, traditional insurance schemes (based on observed yields) are not valid due to the risks of adverse selection and moral hazard as well as to very high claim management costs. In order to limit the basis risk, the climatic index must be highly correlated with the real yield. We therefore look for the best combination of explanatory variables and test different kinds of models (linear model, non linear model, hierarchical model). One of the most important difficulties is the lack of data concerning both yield and climate. Thus we have to apply smoothing methods allowing our index to have a correct value when computed on climatic data independent of the training dataset (on which the model is calibrated). We assess the performance of our insurance product as objectively as possible.

The possible use of such an index as underlying of financial products such as climate derivatives or catastrophe bonds is also briefly studied.

Finally, we discuss the utility of using other sources of climatic data such as atmospheric analysis data or outputs of climate models. The latter could predict the future evolution of the productivity in a context of climate change. In particular techniques like downscaling and upscaling are discussed.

Chapitre 2

Note de synthèse

Mots-clés : assurance rendement, indice climatique, théorie de l'apprentissage statistique, techniques de régularisation, modèle climatique

2.1 Introduction

Le secteur agricole marocain apparaît pauvre et vétuste, dans un contexte mondial marqué par des évolutions rapides. Les échanges accrus ainsi que la prise en considération de la contrainte environnementale obligent les agriculteurs à améliorer leur techniques de production. Une telle amélioration requiert :

- La possibilité de se couvrir face aux risques principaux ;
- L'accès au micro-crédit afin de financer les investissements dans de nouvelles technologies.

Le risque climatique apparaît crucial dans un contexte de changement climatique qui tend à accroître la fréquence des périodes de sécheresse.

On constate néanmoins un manque de couverture très important provenant de l'offre insuffisante des compagnies privées. En effet, dans l'agriculture en général et a fortiori dans les pays en voie de développement, les phénomènes d'antisélection ainsi que d'aléa moral rendent les risques climatiques inassurables avec les solutions traditionnelles. A cela s'ajoutent les coûts très élevés de gestion des sinistres (envoi d'experts sur place pour comptabiliser les rendement effectifs).

Ce manque de couverture a trois conséquences principales :

1. Une diminution immédiate de la richesse du fait de la baisse de revenu mais aussi des dommages potentiels aux appareils de production.
2. Une stratégie agricole peu risquée parfois au détriment de la productivité.
3. Un accès très difficile au micro-crédit, ce qui a pour effet d'empêcher les agriculteurs de rénover leur technologie de production.

Il apparaît donc absolument crucial de trouver des solutions d'assurance viables. La protection des agriculteurs apparaît comme l'un des objectifs du Plan Maroc Vert mis en place en 2008 pour aider l'agriculture marocaine à répondre aux nouveaux défis. C'est dans ce cadre que la coopérative agricole COSUMAR a demandé une étude de faisabilité d'assurance rendement pour le secteur de la production de betteraves sucrières. Ce dernier est particulièrement important au Maroc car le pays souhaiterait parvenir à une auto-suffisance en sucre à hauteur de 50 % . Il est néanmoins très sensible aux aléas météorologiques.

L'idée proposée ici est une assurance rendement basée sur un indice climatique. Il s'agit d'une solution intéressante puisqu'elle élimine les problèmes d'antisélection, d'aléa moral et de gestion des sinistres évoqués précédemment.

2.2 Les données

L'étude porte sur la totalité de la production de betterave marocaine. Celle-ci se répartit en 16 zones. Pour chacune d'elle, nous disposons d'un historique de 30 ans (période 1979 à 2008) des variables climatiques ainsi que des rendements en sucre. L'historique n'est néanmoins complet que pour 8 zones.

Chacune d'entre elles comporte une station météorologique d'où proviennent les moyennes mensuelles de température, précipitation, température minimale, température maximale et degré-jour.

Un point crucial de toute étude d'impact concerne le pré-traitement des données. En effet, le rendement en sucre n'est pas seulement expliqué par les variables climatiques. Il possède une tendance croissante que l'on peut attribuer à d'autres facteurs, par exemple l'évolution des technologies de production, l'utilisation des engrais ou encore l'irrigation. Comme on peut le voir sur la figure 10.5, le rendement dans la zone de Ben Amir présente une tendance en escalier que l'on ne peut attribuer à des influences climatiques. L'origine est probablement une rupture technologique. Il convient de modéliser la tendance en introduisant toutes les connaissances a priori au sujet des facteurs autres que le climat. En la retranchant, on se ramène à un résidu probablement expliqué en grande partie par les facteurs climatiques. En normalisant par l'écart-type, on obtient l'anomalie. Ainsi, si le rendement s'écrit $S_t = z_t + r_t$ où z_t et r_t sont respectivement la tendance et le résidu, l'anomalie est :

$$a_t = \frac{r_t}{\sigma(S_t)} = \frac{S_t - z_t}{\sigma(S_t)}$$

Dans le cas des variables climatiques, il faut, outre la tendance, s'affranchir de la saisonnalité. On a donc :

$$a_t = \frac{r_t}{\sigma(S_t)} = \frac{S_t - (z_t + s_t)}{\sigma(S_t)}$$

où s_t désigne la saisonnalité.

2.3 Méthodologie

L'objectif est de déterminer la relation entre les anomalies de rendement en sucre et celles relatives aux variables climatiques pertinentes. Il s'agit donc de trouver : Il s'agit donc de trouver :

- **les bons prédicteurs (variables climatiques pertinentes) ;**
- **le modèle approprié (linéaire, non linéaire, hiérarchique) représentant la relation entre les anomalies de variables climatiques (entrées) et les anomalies de rendements (sorties).**
- **les paramètres du modèle.**

Les réseaux de neurones sont extrêmement efficaces mais ils nécessitent un nombre de données suffisant. Si ce n'est pas le cas, ils ont tendance à faire du sur-apprentissage : ils apprennent la base d'apprentissage par coeur (et en particulier le bruit des données) mais sont incapables de prédire correctement à partir de nouvelles données (très faible capacité de généralisation). Pour pallier cette difficulté, il faut mettre en place des techniques de régularisation :

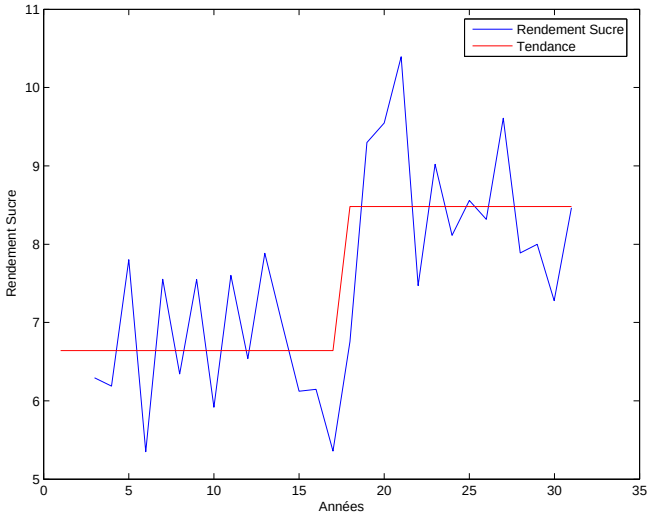


FIGURE 2.1 : Evolution du rendement en sucre dans la zone de Ben Amir et tendance associée

- l'input perturbation : on ajoute du bruit aux données pour éviter que le réseau apprenne les détails.
- le early stopping : on arrête le processus de minimisation de la fonction de coût avant d'avoir trouvé le minimum. Ceci permet également d'éviter d'apprendre les détails de la base.
- le weight decay : nous ajoutons à la fonction de coût un stabilisateur pénalisant les valeurs de poids trop importantes, souvent à l'origine de sur-apprentissage.
- les runs d'ensemble : il s'agit de déterminer différents jeux de coefficients par apprentissage sur différentes bases. La moyenne des prédictions issues des différents runs peut avoir un bon comportement en généralisation et l'écart-type fournit un indicateur de la précision.

Dans le cas de la régression linéaire, il est possible d'utiliser les techniques d'input perturbation et de runs d'ensemble. On ne parle pas à proprement parler de weight decay dans le cas linéaire mais il est équivalent de limiter la co-linéarité ou encore de pratiquer la "ridge regression".

Nous souhaitons évidemment obtenir l'indice ayant la meilleure capacité de généralisation. L'estimation du taux de généralisation requiert l'utilisation de bases d'exemples indépendantes de la base d'apprentissage : la base de validation ainsi que la base de test. Compte-tenu du manque de données, nous ne pouvons pas séparer la base totale en bases de tailles suffisantes. Nous avons donc recours au bootstrap et au leave-one-out pour mesurer la qualité de l'indice. L'idée du bootstrap est de constituer un grand nombre de nouvelles bases de données en considérant des sous-ensembles quelconques de notre base de données initiale. Le leave-one-out est un cas particulier du bootstrap où l'on laisse un exemple de côté. Ceci permet d'augmenter artificiellement la taille de la base de test ou de validation.

Nous voulons obtenir la meilleure combinaison de prédicteurs de manière statistique. Or cette procédure s'avère relativement coûteuse en temps de calcul car elle teste de nombreuses combinaisons de variables explicatives (sur des bases de validation et de test). Il est évidemment impossible de tester l'ensemble des modèles envisageables ainsi que des valeurs des paramètres de régularisation sur chaque combinaison de prédicteurs possible. Il convient donc de séparer le choix du modèle et des paramètres de régularisation d'une part et la recherche des meilleurs

prédicteurs d'autre part.

Ainsi, **nous fixons dans un premier temps** 4 **prédicteurs issus de l'expertise agronomique et cherchons le meilleur modèle ainsi que les paramètres de régularisation optimaux**. Considérer davantage de variables explicatives reviendrait à introduire des paramètres supplémentaires et donc à favoriser le sur-apprentissage compte-tenu du faible nombre de données. Les prédicteurs choisis sont :

- l'anomalie de précipitation en juillet : bénéfique si positive car juillet correspond aux dernières semaines de croissance de la plante à une époque souvent très sèche ;
- l'anomalie de précipitation en avril : bénéfique si positive car moment important de la croissance et les précipitations peuvent s'avérer faibles ;
- l'anomalie de température en février : bénéfique si positive car les températures sont souvent basses, notamment sur les hauts plateaux ;
- l'anomalie de température en septembre : bénéfique si positive car favorise la croissance juste après la période de germination ;

Nous souhaitons dans un premier temps élaborer un indice par zone. Néanmoins, les résultats en généralisation sont très mauvais, que ce soit en utilisant la régression linéaire ou le réseau de neurones, et ce même appliquant les paramètres de régularisation optimaux. Nous appliquons donc les modèles à l'ensemble des 16 zones puis seulement aux 8 zones pour lesquelles les données sont complètes. Les meilleurs résultats sont obtenus dans le dernier cas. On remarque alors que le réseau de neurones bien régularisé fournit un résultat similaire à celui de la régression linéaire mais pas meilleur. Ceci est logique puisque le cas linéaire est compris dans le cas non linéaire. Néanmoins, il ne peut faire mieux ici du fait du manque de données.

Finalement, le modèle le plus approprié est le modèle linéaire simple avec introduction de bruit de 10 % sur les entrées, appliqué aux 8 **zones pour lesquelles on possède des données complètes**.

Connaissant le modèle ainsi que les paramètres de régularisation, on peut désormais choisir la meilleure combinaison de prédicteurs. Notons que les variables explicatives testées correspondent aux anomalies des variables climatiques et non à des fonctions linéaires de ces dernières. Un tel cas est en effet inclus dans la classe des fonctions proposées par les réseaux de neurones et a donc été indirectement testé et rejeté.

Ainsi, la démarche est la suivante : Etant donné que l'on travaille sur 8 zones et que l'on dispose d'un historique de 30 ans pour chacun d'elles, on possède au total 240 observations pour chaque variable (à la fois pour les anomalies de rendement et des différentes variables météorologiques)

On considère ainsi 50 plages différentes comportant chacune 10 exemples test.

1. On choisit une première base de test et on considère le complémentaire dans les données noté *Comp*.
   (a) On veut choisir le premier prédicteur optimal.
      i. On teste un premier prédicteur.
         A. On effectue alors un premier leave-one out dans *Comp* ;
         B. On calcule les coefficients de la régression (du rendement sur ce prédicteur) sur les données de *Comp* privées de celles correspondant au leave-one-out (une donnée de rendement et une donnée correspondant au prédicteur). On calcule alors la prévision du modèle pour la donnée laissée de côté.
         C. On refait 300 leave-one-out, obtenant une série observée de taille 300 que l'on peut comparer à la série prédite grâce au modèle. On comprend alors

que l'intérêt du leave-one-out réside dans le fait qu'il permet de constituer artificiellement une base de validation de grande taille.

D. On calcule la corrélation linéaire ainsi que l'erreur quadratique entre les deux séries.

ii. On réitère ceci avec les 60 prédicteurs potentiels et on choisit celui qui maximise la corrélation ou minimise l'erreur quadratique entre les deux séries.

(b) On réitère alors la démarche précédente pour le deuxième prédicteur. Il s'agit alors de maximiser la corrélation (ou de minimiser l'erreur quadratique) sur la base de validation (constituée par les 300 leave-one-out) entre la série réelle et la série reconstruite à partir de l'indice comportant le premier prédicteur et celui que l'on est en train de choisir.

(c) Par récurrence, on réitère ce procédé pour les prédicteurs suivants.

2. On réédite la même démarche sur les autres bases de test.

On choisit finalement la combinaison de prédicteurs maximisant la corrélation (ou minimisant l'erreur quadratique) sur la base de test finale.

2.4 Résultats

L'indice optimisé complet est décrit dans le tableau 2.1.

Prédicteurs	Coefficients
Biais	-0,003
Anomalie de degré-jour en septembre	1,503
Anomalie de température minimale en novembre	-0,833
Anomalie de précipitation en août	-0,005
Anomalie de précipitation en novembre	-0,08

TABLE 2.1 : Indice optimisé final

Le tableau 11.10 résume les résultats en généralisation :

	Corrélation	Erreur
Moyenne	0,31	0,14
Ecart-type	0,29	0,06

TABLE 2.2 : Résultats en généralisation pour l'indice optimisé

On constate que l'on aboutit finalement à un indice d'impact présentant un score de 0,31 en corrélation avec une probabilité de 70 % d'être entre 0,2 et 0,6. Cette valeur assez faible est due au nombre limité d'années d'observations, à l'absence de données concernant certaines variables importantes (irrigation par exemple) ainsi que l'imprécicion concernant les rendements. La distribution des corrélations (entre la série de rendement réelle et la série reconstruite par l'indice) sur les différentes bases de test est visible sur la figure 2.2. Le risque

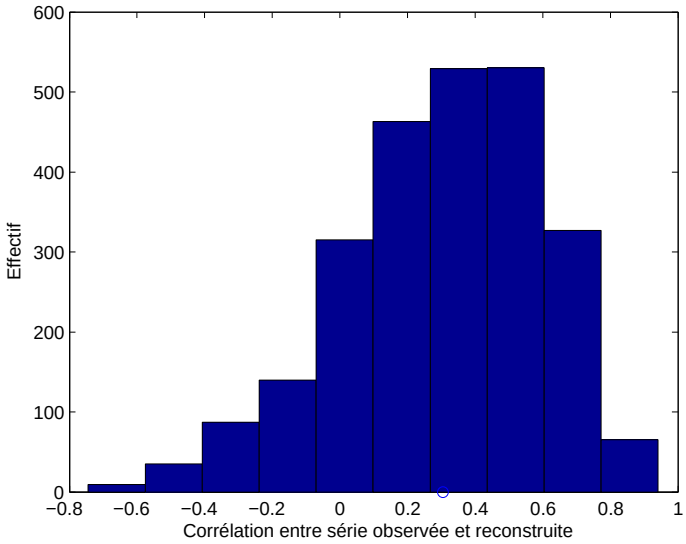


FIGURE 2.2 : Histogramme des corrélations pour l'indice optimisé

de base est donc particulièrement élevé et une solution d'assurance basée sur cette indice ne serait pas pertinente. Les agriculteurs ne seraient en effet pas toujours protégés en cas de besoin. Néanmoins, le risque lié à la valeur de l'indice est assurable.

2.5 Conclusion

L'agriculture marocaine a entrepris les réformes nécessaires en vue de s'adapter au nouveau contexte mondial. Ceci passe par des moyens permettant aux agriculteurs de se prémunir notamment face aux risques climatiques, dans un contexte où les épisodes de sécheresse sont de plus en plus fréquents. Néanmoins, les problèmes d'antisélection et d'alé moral ainsi que les coûts élevés de gestion des sinistres rendent les solutions traditionnelles d'assurances rendement difficilement applicables. Une solution intéressante permettant de pallier ces problèmes consiste à proposer une assurance basée sur un indice indirect, par exemple un indice climatique.

Nous avons testé la possibilité de mettre en place ce type de couverture dans le cas de la production sucrière marocaine. Nous en tirons les enseignements suivants :

- La phase de prétraitement des données est absolument cruciale.
- Il est difficile de mesurer la performance réelle de l'indice. En effet, il est crucial d'éviter le piège des scores en sur-apprentissage. Il faut évaluer la capacité de généralisation du modèle.
- Il est très délicat d'obtenir une bonne capacité de généralisation lorsque l'on dispose de peu de données, comme c'est le cas ici. Il convient alors de mettre en place des méthodes de régularisation. En particulier, le réseau de neurones ne donne pas de meilleurs résultats que la régression linéaire. Au mieux, bien régularisé, il est équivalent.
- La méthode de choix des variables prédictives apparait satisfaisante.

L'indice climatique optimal possède une corrélation de 0,31 avec le rendement en sucre. Le risque de base est donc très élevé. Outre ceci, le risque apparaît difficilement assurable du point de vue actuariel. Une assurance basée sur cet indice n'est donc pas envisageable. Néanmoins, nous avons la satisfaction **d'avoir mis en place une méthodologie permettant d'obtenir**

le résultat optimal et de tester la performance réelle de l'indice.

Il est également intéressant de mentionner le développement de produits financiers basés sur de tels indices et permettant de transférer le risque climatique aux marchés financiers. On utilise les dérivés climatiques pour couvrir les aléas courants et les cat bonds dans le cas des risques extrêmes.

Par ailleurs, comme on l'a vu, l'utilisation des analyses atmosphériques ainsi que des sorties de modèle climatiques peuvent être très intéressantes. Les dernières permettent notamment d'étudier l'évolution future dans un contexte de changement climatique.

Chapitre 3

Synthesis

Key words : yield insurance, climatic index, statistical learning theory, smoothing methods, climate models

3.1 Introduction

The moroccan agriculture appears poor and inadequate in a world context characterized by rapid changes. The increasing exchanges and the environmental constrains oblige the farmers to improve their production capacity. This requires :

- the possibility to be covered against the main risks ;
- the access to credit facilities in order to finance the investments in new technologies.

The climatic risk is crucial in a context of climate change tending to increase the frequency of drought periods. However there is an obvious lack of cover apparently due to a very low offer by insurance companies. In the agricultural sector in general and even more so in developing countries, adverse selection and moral hazard make it difficult to insure climatic risks by traditional means. Moreover, management claim costs are very high (need to send experts to measure the yields).

This lack of cover has 3 main consequences :

1. An immediate reduction of wealth because of the decrease of revenue and potential damages of the production devices.
2. The choice of a low risk sometimes unfavouring the productivity.
3. A difficult access to credit facilities that prevents them from renovating their production technologies.

Therefore it is crucial to find valid insurance solutions. The farmers' protection is one objective in the plan "Maroc Vert" of 2008 in order to help the moroccan agriculture to face the new challenges. In this context, the cooperative COSUMAR has been looking for a study concerning the possibility of a yield insurance in the field of sugar beet production. This is particularly important in Morocco since the country wants to be self-sufficient in sugar at a level of 50 %. However, this production depends highly on the meteorological conditions.

The solution proposed is a yield insurance based on a climatic index. It is very interesting because it eliminates the problems of adverse selection, moral hazard and claims management.

3.2 The dataset

The study concerns the entire production of sugar beet in Morocco. 16 production areas are considered. For each of them, we know the climatic variables and the sugar yields during 30 years (from 1979 to 2008). However, data are complete only for 8 zones.

Each of these areas has a meteorological center providing the monthly means of temperature, rain, minimal and maximal temperature and degree days.

A crucial point of all kind of impact studies is the pre-treatment of the dataset. Actually, the sugar yield does not only depend on the climatic conditions. It tends to increase because of other factors i.e. the improvement of production technologies and irrigation and the use of fertilizers. As to be seen on figure 3.1, the yield in the Ben Amir area shows a sudden increase which cannot be explained by meteorological conditions. It is apparently due to changes in technology. The tendency has to be modelled by introducing all the factors independant of climate. By eliminating this trend, we obtain data (residuals) mostly explained by climatic factors. The yield anomalies are computed by dividing by the standard deviation.

If the yield is expressed by $S_t = z_t + r_t$ where z_t and r_t are the trend and the residual respectively, the anomaly is :

$$a_t = \frac{r_t}{\sigma(S_t)} = \frac{S_t - z_t}{\sigma(S_t)}$$

In the case of climatic variables, seasonnal effects have also to be eliminated. Thus :

$$a_t = \frac{r_t}{\sigma(S_t)} = \frac{S_t - (z_t + s_t)}{\sigma(S_t)}$$

where s_t means seasonality.

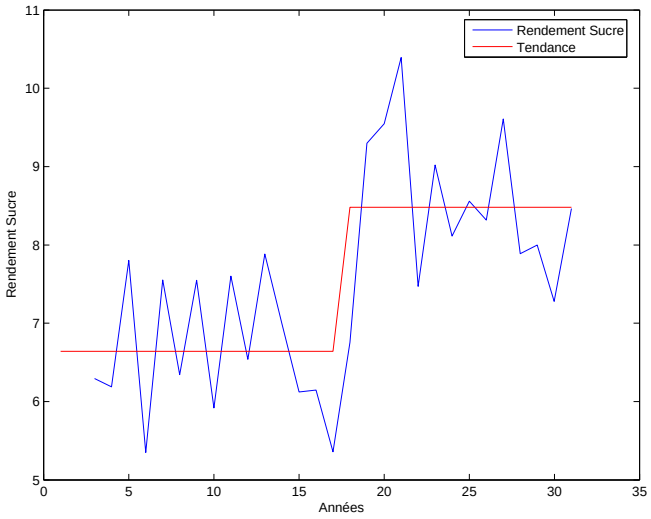


FIGURE 3.1 : Change in sugar yield in the Ben Amir zone and associated trend

3.3 Methods

The purpose of the study is to determine the relationship between anomalies of sugar yields and those of relevant climatic variables. It has to be defined :

- **relevant explanatory variables ;**
- **the appropriate model (linear, non linear, hierarchical) explaining the relationship between climatic and sugar yield anomalies ;**
- **the model parameters.**

The neural networks are very efficient but they need a sufficient number of data. If this is not the case, there is a tendency towards overfitting : the network memorizes all the details of the training dataset but has a very low generalisation capacity.

In order to avoid it, additional smoothing techniques have to be used :

- the input perturbation : noise is added to the data in order to eliminate the effect of details ;
- the early stopping : the minimisation process of the cost function is interrupted before the minimum is reached. This also permits not to memorize the details of the training dataset ;
- the weight decay : a penalty term is added to the cost function, eliminating the risk of high weight values that may generate overfitting ;
- the use of multiple runs : the model can be calibrated on different training datasets. Thus different sets of coefficients can be obtained. The mean prediction using all these sets can have a good generalisation capacity while the standard deviation indicates the uncertainty of the prediction.

In case of linear regression, techniques of input perturbation and multiple runs can be used.

The purpose is evidently to obtain the index with the highest capacity of generalisation. The generalisation rate estimation requires the use of data sets independent of the training data set : the validation and the test data sets. Due to the lack of data, the total data set cannot be divided in three sufficiently large data sets. Therefore bootstrapping and leave-one-out techniques are used to evaluate the quality of the index. Bootstrapping consists in creating a big number of new data sets by considering random subsets of the total data set. The leave-one-out is a particular example of bootstrapping. This allows to increase artificially the size of the validation or test data set.

The purpose is to obtain the best combination of predictors by use of statistical methods. This procedure is very time demanding since it is testing numerous combinations of explanatory variables (on the test and validation bases). Thus it is necessary to separate the choice of the model and of the smoothing parameters from the calculation of the best predictors.

Thus to begin with we define 4 predictors stemming from the agronomic expertise and look for the best model as well as the optimal smoothing parameters. Considering more predictors would introduce additional parameters and by this favour overfitting because of the small size of the data set. The following predictors have been chosen :

- rain anomaly in july : beneficial if positive since it concerns the terminal period of growth.
- rain anomaly in april ;
- temperature anomaly in february : important since temperatures are often low in particular at higher altitude ;
- temperature anomaly in september : high temperature favours the growth immediately after the period of germination.

To start with, an individual index for each zone was tested. However, the generalisation score was very bad, both using linear regression and neural networks, and even when optimizing the regularisation parameters. Thus the models were applied to all the 16 zones and then only to 8 zones having complete data. The best results were obtained in the latter case. It was shown that

the neural network well regularised gives similar results as those obtained by linear regression. However, it was not superior due to the low number of data.

Finally, the most appropriate model is the linear one applied to the data of the 8 zones to which 10 % noise was added (noise with normal distribution).

Knowing the appropriate model and the smoothing parameters, the best combination of predictors could be selected. Thus the procedure is the following :

Since 8 zones are considered and the data cover 30 years, there are 240 observations of each variable (both yield and climatic anomalies). 50 different random sets are considered, each one containing 10 test examples.

1. A first test data set is chosen among those 50 ones. The rest of the data is defined *Comp*.
   (a) The first optimal explanatory variable has to be chosen.
      i. A first explanatory variable is tested.
         A. A first leave-one out is applied to *Comp*;
         B. The regression coefficients (yield on explanatory variable) is computed using *Comp* data after eliminating one example randomly chosen (one yield data and one explanatory data). The model's prediction for the data left out is then computed.
         C. This technique is applied 300 times; thus a series of 300 observations is obtained and can be compared with the series predicted by the model. The leave-one-out technique allows to establish a validation data set of a big size.
         D. The linear correlation and the quadratic error between the 2 series are computed.
      ii. The same is done with all the 60 potential predictors and the one which maximizes the correlation or minimizes the quadratic error between the 2 series is chosen.
   (b) The same procedure is repeated to choose the second predictor. The real series is compared with the series predicted by the index containing the first predictor and the one which is to be tested.
   (c) The same method is applied to choose the following predictors.
2. The other test data set are treated in the same way.

Finally the combination of the predictors that maximizes the correlation (or minimizes the quadratic error) between the observed and the predicted series of the test data set is chosen.

3.4 Results

The optimised index is given in table 3.1.

The table 3.2 summarizes the generalisation capacity :

The optimised index has a mean correlation score of 0.31. There is a 70 % probability that the coefficient of correlation is situated between 0.2 and 0.6. The distribution of the correlation coefficients (between the observed and predicted series) is shown in figure 3.2. Thus the basis risk is particularly high. An insurance product based on this index would not be relevant.

3.5 Conclusion

The moroccan agriculture has started the necessary reforms in order to adapt to the new world context. This implies that the farmers have the possibilities to be covered against climatic

Predictors	Coefficients
Bias	-0.003
Degree day anomaly in september	1.503
Minimal temperature anomaly in november	-0.833
Rain anomaly in august	-0.005
Rain anomaly in november	-0.08

TABLE 3.1 : Optimised index

	Correlation	Quadratic error
Moyenne	0.31	0.14
Ecart-type	0.29	0.06

TABLE 3.2 : Performance of the optimised index

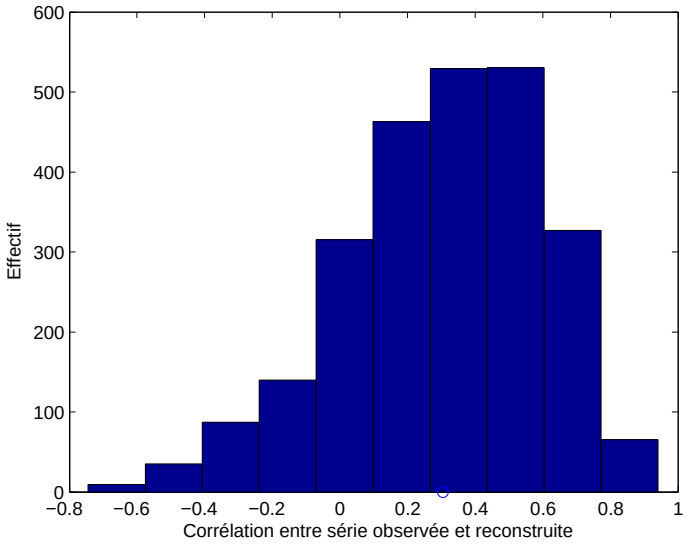


FIGURE 3.2 : Histogram of the correlations concerning the optimised index

risks in particular since drought episodes are increasingly frequent. However, problems of adverse selection and moral hazard as well as high claim management costs render traditional yield insurance solutions difficult. An interesting alternative consists in choosing an insurance based on an indirect index, for example a climatic index.

The possibility of using this type of cover was tested in the case of the moroccan sugar production. The following conclusions can be drawn :

- The pretreatment of data is crucial ;
- It is difficult to measure the true performance of the index. It is important to avoid giving scores based on overfitting. The generalisation capacity has to be evaluated.

– It is difficult to obtain a high generalisation capacity when the size of data is small. It appears that the neural network is not superior to linear regression. At the best, it is equivalent. In this case smoothing methods have to be applied.
– The method applied to choose the best combination of predictive variables appears satisfactory.

The yields computed with the index show a mean correlation of 0.31 with the real ones. Thus the basis risk is very high. Moreover the climatic risk is not insurable from an actuarial point of view. Thus an insurance based on this index would not be relevant. The rather low value of correlation is due to the limited number of years of observation, the lack of important variables (irrigation) and the uncertainty concerning the yields.

However, it is satisfactory to have been able to develop a methodology allowing to obtain the optimal result and to test the real performance of the index.

Financial products based on such indexes have been developed to transfer climatic risk to financial markets. Climatic derivatives and cat bonds respectively are used to cover moderate and extreme risks.

Finally, the utility of using other sources of climatic data such as atmospheric analysis data or outputs of climate models is shown. The latter could predict the future evolution of the productivity in a context of climate change.

Chapitre 4

Introduction

Les économistes ne s'intéressent à l'impact des variations météorologiques sur l'activité économique que depuis peu. En effet, pendant assez longtemps, les seuls travaux tentant de modéliser la relation entre économie et climat étaient l'oeuvre du néo-classique Stanley Jevons, datant du XIXe siècle. Ce dernier avait alors avancé la possible existence d'une corrélation négative entre le nombre de taches solaires et l'activité économique.

Néanmoins, certains événements tels la canicule de 2003 ont confirmé, dans un contexte humain hélas dramatique, la très grande sensibilité de notre société aux conditions climatiques. Les entreprises sont, elles, soumises à court terme à deux types de risque :

- ceux associés aux **risques extrêmes** de types inondations, ouragans, tempêtes ;
- ceux liés aux **variations courantes** du niveau de la température, des précipitations, du vent ou encore de la neige.

Dans certains secteurs, ces phénomènes sont à l'origine d'énormes variations de l'activité économique. Ainsi, les secteurs de l'énergie, du tourisme, de l'agroalimentaire, du textile, de la construction, des loisirs ou encore de la production agricole sont très affectés par la variabilité météorologique. Cette dernière a d'ailleurs, dans ces domaines d'activité, un impact au moins aussi important que les mouvements de taux d'intérêt, les variations du cours des devises ou encore celles relatives aux prix des matières premières. A titre indicatif, la variance de la consommation d'électricité, tout comme celle de la consommation de bière, est expliquée à plus de 90 % par la variance de la température. Celles-ci s'avèrent donc faiblement expliquées par la croissance économique, le niveau des taux d'intérêt ou encore le prix du pétrole.

La légitimité de la gestion du risque climatique dans l'entreprise repose en fait sur deux constats :

- d'une part, **la volatilité des indices climatiques est du même ordre de grandeur que celle des variables de marché** ;
- d'autre part, **la sensibilité des entreprises aux variations climatiques est, dans certains secteurs, plus importante que celle attachée aux variables de marché "traditionnelles"** .

Ainsi, dans un grand nombre de domaines, il s'avère absolument nécessaire de tenir compte de l'impact des chocs météorologiques réguliers ou extrêmes. Les entreprises peuvent couvrir les risques associés grâce à des **produits d'assurance ou de type bancaire**. Les produits dérivés conçus pour se prémunir face aux variations de taux d'intérêt, de cours de change, d'indices boursiers ou encore de prix des matières premières ont été étendus au climat et il est donc désormais possible pour les entreprises de couvrir leur exposition au risque météorologique par des options, des contrats à terme ou encore des swaps climatiques.

Néanmoins, le risque climatique ne se limite pas à ces perspectives à court terme. Il s'agit en effet d'appréhender les **grandes questions climatiques de long terme**. On entend par là-même le réchauffement climatique et ses conséquences sur les condtions météorologiques, de deux types :

- la modification du comportement moyen ;
- une volatilité accrue.

Il s'ensuivra par là-même un probable accroissement de la fréquence des événements extrêmes. Selon les hypothèses retenues, le réchauffement global de notre planète à horizon d'un siècle se situerait entre $+1.4^{\circ}$ et $+5.8^{\circ}$. Si la borne supérieure était atteinte, les températures exceptionnelles du mois d'août 2003 se situeraient alors dans la moyenne. Les conditions de la production agricole seraient alors bouleversées, tout comme la structure de consommation ou la productivité de certains secteurs économiques. Tout dirigeant d'entreprise doit avoir ces évolutions en perspective, même si son horizon de gestion est à plus courte échéance que celui de la prévision climatologique.

L'agriculture est probablement le secteur le plus sensible aux aléas climatiques, et plus particulièrement dans les pays en voie de développement et ne disposant que de faibles ressources en eau. Il y est absolument crucial d'aider les agriculteurs à sortir de la pauvreté et à s'adapter à un contexte mondial en pleine mutation. Ceci passe notamment par des solutions adéquates de gestion des risques.

C'est dans ce cadre que la coopérative agricole marocaine COSUMAR a demandé au courtier d'assurance AON de mener une étude de faisabilité au sujet de la conception de produits de couverture pour les producteurs de betterave sucrière. Le but de ce mémoire est de déterminer quels produits peuvent être utilisés et de mettre en place une méthodologie permettant de les élaborer et d'évaluer leur performance.

Nous étudions tout d'abord le contexte actuariel et économique et montrons en quoi il est crucial d'aider les agriculteurs à se prémunir face aux risques. Nous regardons ensuite la question de l'assurabilité et dressons un panel des solutions de couverture traditionnelles. Nous montrons que celles-ci sont difficilement applicables dans les pays en développement. Nous introduisons alors l'assurance basée sur indice (et plus particulièrement celle basée sur indice climatique) et montrons qu'elle apparaît comme une solution prometteuse. Nous présentons ensuite la méthodologie et exposons les résultats.

Chapitre 5

Contexte actuariel de l'étude

L'agriculture marocaine a toujours été un secteur stratégique pour le développement socio-économique du pays. Depuis l'indépendance, elle a connu de nombreux programmes de développement ainsi que de multiples réformes structurelles afin de permettre à la nation d'assurer sa sécurité alimentaire. Elle a par ailleurs une contribution décisive aux grands équilibres économiques et à la balance commerciale du pays. Néanmoins, malgré le soutien dont elle a bénéficié, l'agriculture reste un secteur sous-développé. Elle souffre en effet d'un déficit de croissance chronique, dans un contexte économique mondial en pleine mutation. Avant de s'intéresser plus en détail au cas de l'agriculture marocaine, nous présentons les caractéristiques principales du contexte agricole au niveau mondial.

5.1 Le nouveau contexte mondial de l'agriculture

L'agriculture revient au premier plan des préoccupations sur la scène internationale. Le dernier rapport de la Banque Mondiale au sujet du développement dans le Monde a mis en exergue le rôle crucial de l'agriculture dans la lutte contre la pauvreté ainsi que les mesures à prendre pour la rendre plus efficace et au service du développement. La satisfaction des besoins agricoles de la planète redevient en effet un enjeu stratégique comme le suggère l'augmentation des prix des produits agricoles sur les marchés internationaux. Celle-ci a été extrêmement rapide durant le premier trimestre 2008 et a même provoqué des émeutes de la faim dans certains pays.

Ainsi, les débats au sujet des réformes des politiques agricoles prennent de plus en plus d'importance. On peut notamment citer la politique agricole commune, les négociations au niveau du commerce international ainsi que les partenariats économiques. Paralèllement, les questions environnementales occupent une place de plus en plus notable dans la conception que l'on se fait du développement économique. Or l'agriculture apparaît comme le principal utilisateur de ressources naturelles telles les terres, l'eau et la biosphère.

5.1.1 L'augmentation des prix

Comme nous l'avons vu, les prix ont connu une forte hausse depuis 2008 et les prévisions annoncent le maintien de cette tendance dans la prochaine décennie. Les raisons sont les suivantes :

- La demande croissante en biocarburants induite par les politiques incitatives américaines, européennes et brésiliennes. Celle-ci est également étroitement liée au prix du pétrole ;
- Les besoins alimentaires issus de la croissance démographique ;

– Les nouvelles demandes, souvent engendrées par la hausse des revenus dans les pays émergents tels que la Chine et l'Inde. Celles-ci concernent principalement les produits animaux, mais aussi les oléagineux.

5.1.2 Les négociations internationales et les réformes des politiques agricoles

Le contexte est également marqué par les négociations au sein de l'Organisation Mondiale du Commerce (OMC), les réformes de la Politique Agricole Commune (PAC), de la politique américaine (FARM BILL), ainsi que les accords bilatéraux, notamment les Accords de Partenariat Economique (APE) entre l'Union Européenne, les pays d'Afrique, les Caraïbes et le Pacifique.

Selon B.Bachelier (2008), la période s'étendant de 2008 à 2013 devait sans doute être décisive puisque les négociations devaient déterminer les nouvelles règles des échanges ainsi que les principes de définition des politiques agricoles. Aussi, 2013 est la date de révision de la Politique Agricole Commune et la date prévue de mise en oeuvre de nombreuses dispositions de l'Organisation Mondiale du Commerce.

Les objectifs sous-jacents à ces différentes dynamiques concernent la libéralisation des échanges, la suppression des entraves au commerce ainsi que la réduction des aides publiques entraînant des distorsions de marché. L'hypothèse qui sous-tend ces évolutions est que l'accroissement des échanges entraîne la création de richesse, une répartition plus équitable des revenus en faveur des pays en développement et améliore donc le bien-être des plus pauvres.

Cependant, ces négociations connaissent actuellement un certain nombre de difficultés. En effet, les pays en développement ne cessent de réclamer aux pays développés (notamment les Etats-Unis et l'Europe) une diminution des subventions agricoles, ce qui leur permettrait un accès plus facile aux marchés correspondant. La question de la suppression des subventions aux exportations est également désormais soulevée.

5.1.3 La contrainte environnementale

La prise de conscience de l'absolue nécessité de préserver les ressources naturelles s'est récemment élargie à l'ensemble de l'opinion publique et des décideurs. Auparavant, on ne se préoccupait que des effets locaux de l'agriculture sur l'environnement (pollution). Désormais, on s'intéresse beaucoup plus à son impact sur les ressources naturelles ainsi que la biodiversité, en lien avec le changement climatique. En effet, ce dernier pourrait s'imposer comme une contrainte majeure voire insurmontable dans certaines régions. Certaines parties du monde telles les Tropiques ou le bassin Méditerrannéen risquent de connaître des épisodes de sécheresse dramatique. Dans un tel contexte, on comprend que l'eau sera extrêmement précieuse et qu'il sera absolument crucial de l'utiliser avec autant de parcimonie que possible. En ce qui concerne la préservation de la biodiversité, celle-ci passe par une augmentation de la productivité des surfaces déjà cultivées afin d'empiéter le moins possible sur les habitats naturels (dans le cas de l'Amérique latine et de l'Afrique notamment). Ceci en ayant en tête les problèmes d'érosion et de perte de fertilité des sols.

Il apparaît dès lors que seule une modernisation des techniques de production pourrait permettre de faire face à ces nouvelles contraines :

– Amélioration du système d'irrigation pour gaspiller le moins d'eau possible ;
– Utilisation d'intrants de meilleure qualité et de manière plus systématique ;

Il convient de noter que les aides dispensés par les bailleurs de fonds et les institutions internationales sont de plus en plus conditionnées au respect de l'environnement, et ce par l'intermédiaire

d'institutions spécialisées comme le Fonds Mondial de l'Environnement. Ainsi, les pays développés sont tentés d'imposer des priorités écologiques aux pratiques de toutes les agricultures du monde.

5.1.4 La crise financière internationale

Les indices boursiers des principales places du monde ont connu de très fortes baisses à partir de septembre 2008. Ceci a engendré une perte de confiance chez les principaux moteurs de la croissance, à savoir les investisseurs et les épargnants, d'où une propagation à l'économie réelle. Les principales puissances sont alors entrées en récession et il en est allé de même pour la plupart des pays en voie de développement du fait des étroites relations internationales. L'agriculture étant l'un des secteurs stratégiques de ces pays, elle a également subi les conséquences néfastes de la crise.

5.2 Le contexte de l'agriculture marocaine

Après avoir exposé les principaux défis de l'agriculture à l'échelon mondial, il convient désormais de présenter les grandes lignes de l'agriculture marocaine. Ceci nous permettra de mettre en exergue l'étendue des réformes et modernisations à entreprendre en vue d'une adaptation aux nouvelles exigences du contexte mondial.

Le secteur agricole et rural occupe une place économique et sociale importante :
- La population rurale constitue actuellement environ la moitié de la population totale du pays ;
- L'emploi direct dans l'agriculture représente à lui seul 80 % de l'emploi rural, soit 3 à 4 millions de personnes ;
- 60 mille emplois dans l'agroalimentaire.
- En terme de valeur ajoutée, l'agriculture contribue au PIB à hauteur de 14 à 25 % selon les conjonctures climatiques.

Cette importance du secteur agricole découle des nombreux efforts du gouvernement marocain depuis l'indépendance, ayant pour finalité la lutte contre la pauvreté rurale :
- L'opération labour ;
- La construction de barrages ;
- L'aménagement hydro-agricole ;
- La promotion de la production végétale et animale ;
- L'encadrement des agricultures ;
- L'élaboration de programmes de développement rural intégré ;

Néanmoins, des contraintes diverses empêchent l'agriculture marocaine d'améliorer sa compétitivité et par la-même de relever les défis relatifs à la sécurité alimentaire, la globalisation des marchés, la concurrence ainsi que la protection environnementale . On peut notamment citer :
- Le déficit de gouvernance : les modèles d'intervention de l'Etat ont été standardisés, ce qui a largement inhibé le potentiel d'innovation qu'offre le pays. Par ailleurs, les rapports entre l'Etat et les acteurs professionnels locaux semblent inadaptés, ce qui ne favorise pas une dynamique de rattrapage économique rapide.
- Le foncier : le manque de sécurité et de clarté dans ce domaine a limité les investissements et les incitations à une bonne gestion des terres.
- Une population non réellement préparée à la modernisation : l'âge moyen de la population d'exploitants est de l'ordre de 52 ans et le taux d'analphabétisme est élevé. En ce

qui concerne les technologies, la proportion des exploitations agricoles ayant recours à la mécanisation ne s'élève qu'à 47 et 31 % respectivement dans le cas des travaux du sol et de la moisson. En ce qui concerne les engrais et semences sélectionnées, seuls respectivement 51 et 33% des exploitants les utilisent.
- Une gestion et économie de l'eau peu maîtrisées : le taux d'irrigation figure parmi les plus bas de la région et les pertes sont élevées.
- Un manque d'organisation de la profession de manière générale.

C'est en vue de pallier ces difficultés qu'a été mis en place le Plan Maroc Vert que l'on étudie à la section suivante.

5.3 Le Plan Maroc Vert

Cette stratégie de relance de l'agriculture marocaine a été élaborée en 2008 et a pour objectifs principaux :
- Imprimer au secteur agricole une dynamique d'évolution harmonieuse, équilibrée et évolutive et tenant compte de ses spécificités ;
- Exploiter les marges de progrès et valoriser au mieux les potentialités ;
- Faire face aux nouveaux enjeux tout en préservant les équilibres sociaux et économiques ;
- Accompagner la profonde mutation que connaît le système agro-alimentaire mondial .

Le plan prône notamment le regroupement des producteurs autour d'un agrégateur, opérateur performant et structuré et s'articule autour de deux piliers majeurs :
- Développer une agriculture performante et adaptée aux règles de marché, en s'appuyant sur des investissements privés ;
- Elaborer une approche orientée vers la lutte contre la pauvreté, en augmentant par exemple significativement le revenu des exploitants agricoles les plus fragiles **tout en les aidant à faire face aux différents risques, et en particulier celui dû aux aléas climatiques.**

5.4 Risques climatiques et appauvrissement

Les événements climatiques peuvent avoir des conséquences très étendues au sein d'une économie régionale. En effet, les pertes agricoles affectent non seulement les revenus des exploitants mais aussi les salaires de leurs employés ainsi que les réserves de nourriture. Il peut ensuite y avoir contamination progressive à l'économie rurale non agricole. En outre, dans le cas de pertes d'actifs productifs, les foyers risquent de sombrer dans une pauvreté dont ils auront du mal à se sortir. Des études de plus en plus nombreuses démontrent que des pertes d'actifs ainsi que des bouleversements récurrents du revenu peuvent concourir à piéger les foyers dans la pauvreté (Barnett, Barrett et Skees 2008). On peut citer la sécheresse (risque covariant au sens où elle affecte un grand nombre de personnes au même moment) qui met en évidence de façon dramatique les insuffisances des méthodes traditionnelles de gestion des risques. Des études détaillées sur l'impact des sécheresses sévères en Éthiopie (Webb et von Braun 1994), dans l'est de l'Inde (Pandey, Bhandari et Hardy 2007) ainsi que dans le sud de l'Inde (Hazell et Ramasamy 1991) démontrent toutes que les pertes de revenus à terme peuvent de loin excéder les pertes de production initiales.

Des défauts de paiement généralisés des prêts sont alors observés, ce qui dissuade progressivement les établissements financiers de contracter de nouveaux prêts.

On comprend donc l'absolue nécessité de la mise en place de techniques de couverture efficaces.

5.5 La demande d'assurance par la COSUMAR

En conclusion, le secteur agricole marocain est pauvre et assez vetuste dans un contexte de pleine mutation à l'échelle mondiale. Le Plan Vert apparaît comme une stratégie de rénovation et d'adaptation de l'ensemble du secteur. Afin de sortir les agriculteurs de la pauvreté et de leur permettre d'investir afin de satisfaire les exigences de modernisation nécessaires dans le contexte de mutation mondiale (augmentation de la demande, accroissement de la concurrence, contraintes environnementales), il apparaît crucial de donner aux exploitants la possibilité de se prémunir face aux risques. Par ailleurs, il convient de souligner que l'un des objectifs du Maroc est de parvenir à **une auto-suffisance en sucre à hauteur de 50 %** et il convient donc de favoriser le choix de la betterave pour les agriculteurs. Or il s'agit d'une **culture à risque puisque très sensible à la variabilité climatique**.

C'est dans ce contexte que la Coopérative agricole COSUMAR, probablement incitée politiquement, a cherché à se procurer une assurance climatique pour les producteurs de sucre. Le commercial AON en charge du Maghreb et de l'Afrique était alors en étroite relation avec le risk manager du groupe ONA (Omnium Nord Africain) auquel appartient la COSUMAR et c'est donc tout naturellement qu'AON a été choisi pour entreprendre cette étude. AON a proposé la construction d'un indice climatique plutôt qu'une assurance dommage classique qui aurait exposé la compagnie à différents risques tels l'antisélection et l'aléa moral et aurait réclamé une organisation locale d'expertise des sinistres trop lourde à mettre en place. Il convient de noter que ce type d'assurance basée sur un indice climatique est peu répandu au Maroc. La Mutuelle Agricole Marocaine d'Assurances (MAMDA) offre en effet le plus souvent des solutions classiques, en ce qui concerne l'assurance grêle ou la mortalité des animaux par exemple.

En raison d'un lien entre AON et le Laboratoire de Météorologie Dynamique, j'ai eu la possibilité de travailler sur cette étude.

Chapitre 6

Assurabilité

On expose ici les principaux concepts relatifs à l'assurabilité. L'étude l'assurabilité du risque que l'on souhaite couvrir ici se fera dans la section 12

6.1 La notion d'assurabilité

En se référant à la classification proposée par Arthur Charpentier, on peut retenir sept critères d'assurabilité :

1. la survenance du risque doit être aléatoire pour que le contrat soit valide juridiquement ;
2. le cadre juridique doit être stable dans le temps ;
3. la perte maximale possible ne doit pas être catastrophique au regard de la solvabilité de l'assureur ;
4. le montant moyen des pertes doit être identifiable et quantifiable ;
5. les risques doivent être mutualisables ;
6. absence d'alé moral et d'anti-sélection ;
7. il doit y avoir un marché, c'est-à-dire une demande et une offre qui se rencontrent ;

Les points 1 et 2 relèvent de l'assurabilité juridique, les points 3, 4 et 5 de l'assurabilité actuarielle et les points 6 et 7 de l'assurabilité économique.

6.2 L'assurabilité juridique

Du point de vue juridique, un risque est assurable s'il y a un aléa et que le droit ne s'oppose pas à sa couverture. L'assurance faisant intervenir des contrats, elle est soumise au droit des contrats. Michel Denuit et Arthur Charpentier rappellent l'évolution historique de l'inassurabilité juridique : autrefois les bonnes meurs s'opposaient à l'assurabilité sur la vie et à l'assurance des fautes. Ils remarquent également que les évolutions jurisprudentielles peuvent se présenter comme un facteur d'inassurabilité. Un contexte juridique mouvant peut en effet engendrer une crise assurancielle. Ils citent l'exemple de la responsabilité civile médicale en France en 2002. On comprend alors que des difficultés puissent se produire dans des pays tels le Maroc, où le droit des contrats n'est pas encore parfaitement établi.

6.3 L'inassurabilité actuarielle

Comme on l'a vu, le montant des pertes moyenne doit être identifiable et quantifiable. La tarification des contrats se base alors sur :

- La loi des grands nombres : la moyenne empirique des pertes tend vers l'espérance.
- Le théorème central limite : la dispersion relative diminue avec la taille du portefeuille ;

L'utilisation de ces deux théorèmes requiert les hypothèses suivantes :

- risques indépendants ;
- risques homogènes (coûts de même loi)
- portefeuille de grande taille (la loi des grands nombres est un résultat asymptotique)

6.4 L'inassurabilité économique

6.4.1 Le problème d'antisélection

L'antisélection est un phénomène qui apparaît lorsque l'entreprise d'assurance ne peut pas distinguer les bons et les mauvais risques. Elle est néanmoins au courant de l'existence de ces deux catégories de risque. En conséquence, le montant de la prime d'assurance est compris entre celui correspondant aux bons risques et celui correspondant aux mauvais. Il y a asymétrie d'information dans le sens où les assurés, eux, savent à quelle catégorie ils appartiennent. En conséquence, les bons risques savent que la prime qui leur est demandée est trop élevée et se dirigent donc vers les assureurs concurrents.

Plus généralement, considérons le cas de la coassurance. Une augmentation de la sinistralité (remplacement d'une sinistralité X par une sinistralité X' dominant stochastiquement la première à l'ordre 1, c'est-à-dire jugée plus élévée par tout agent dont les préférences sont représentables par une fonction d'utilité espérée et respectant l'hypothèse de non satiété) entraîne une augmentation du taux de couverture (sous l'hypothèse d'aversion partielle inférieure à 1). Cela signifie que, à aversions pour le risque identiques, à richesses initiales identiques, les mauvais risques se couvrent davantage que les bons. L'assureur se doit alors de demander une prime supérieure à celle qu'implique la sinistralité moyenne de la population. Cette hausse de la prime entraîne une diminution du taux de couverture des bons risques, ce qui entraîne une nouvelle hausse de la prime. Ce phénomène d'attirance des mauvais risques et de rejet des bons est appelé antisélection (ou sélection adverse) car il est contraire à ce que souhaite l'assureur.

De manière générale, l'antisélection désigne le phénomène suivant : en cas d'asymétrie d'information, la personne mal informée (ici, l'assureur) est conduite à proposer un contrat qui repousse les personnes informées qu'elle voudrait attirer et n'intéresse que celles avec lesquelles il souhaite le moins contracter.

Ce problème est très fréquent dans les pays en développement car il est très difficile pour les assureurs de bénéficier d'informations fiables sur les différentes catégories de risques.

6.4.2 Le problème d'aléa moral

Le second problème posé par l'asymétrie d'information, de nature différente de l'antisélection, est appelé aléa moral. Une fois assuré, un agent peut être, plus ou moins consciemment, moins vigilant. Il y a donc modification de sa sinistralité (la sinistralité X prévalant avant l'action de s'assurer est remplacée par une sinistralité X' qui la domine stochastiquement à l'ordre 1). Le nom donné à ce changement de comportement vient du fait qu'un individu d'une grande moralité ne changerait pas de comportement. Contrairement à l'antisélection, il s'agit d'un phénomène qui se produit après la signature du contrat d'assurance.

Dans le cas particulier de l'agriculture, on peut citer deux types d'aléa moral :

- Un aléa moral ex ante : l'agriculteur peut par exemple préserver ses réserves en eau, diminuer la quantité d'intrants utilisée ou encore mal effectuer sa récolte. Cet aléa se manifeste donc par une modification de comportement avant la comptabilisation de la récolte.
- Un aléa moral ex post : il peut également effectuer de fausses déclarations sur ses récoltes sans que l'assureur puisse le vérifier à moindre coût.

Dans la littérature, deux thèses s'opposent sur la présence d'aléa moral ex ante dans le cas de l'assurance rendement. Selon Horowitz, la décision de s'assurer a lieu après l'achat des intrants et il n'y a donc pas d'aléa moral ex ante.

Cet argument est balayée par Goodwin et Smith ainsi que par Babcock et Hennessy qui considèrent que toutes les décisions se prennent simultanément. Les agriculteurs engagent ainsi moins de dépenses en intrants étant donné que le rendement est garanti.

On peut également objecter à Horowitz que l'aléa moral ex ante ne concerne pas seulement les intrants mais également par exemple les ressources en eau.

Chapitre 7

Les différents outils de gestion du risque agricole

Les exploitations agricoles encourent deux types de risques principaux : les risques liés aux événements climatiques menaçant les récoltes (risque de quantité) et les risques économiques qui menacent la valeur qu'ils peuvent tirer de leur activité (risque prix).

7.1 La gestion interne du risque

Il s'agit des stratégies mises en place au sein de l'activité agricole, à savoir l'auto-assurance et la diversification.

7.1.1 L'auto-assurance

Il s'agit des techniques visant à transférer les revenus des bonnes périodes vers les moins bonnes. Ceci permet aux agiculteurs d'amortir leurs pertes et par là-même de pérenniser leur activité. Celle-ci passe par des stratégies d'épargne ou des lissages professionnels via les fonds de lissage. Les agriculteurs peuvent donc s'auto-assurer en constituant des épargnes de précaution dans lesquels ils peuvent puiser. Ces fonds propres permettent de surmonter des difficultés ponctuelles d'ordre opérationnel. Ceux-ci sont néanmoins le plus souvent insuffisants et doivent être complétés par un transfert de risque ou un aménagement des techniques de production.

Les agriculteurs peuvent également souscrire à des fonds de lissage. Il s'agit d'organismes le plus souvent subventionnés par le public qui prélèvent une partie du revenu des agriculteurs lors des bonnes périodes et leur viennent en aide lors des mauvaises. Néanmoins, ceux-ci ne permettent de couvrir que le risque prix. Ils analysent les tendances de prix et fixent un intervalle autour de celui-ci dont la borne inférieure (ou plancher) définit le seuil d'intervention et la borne supérieure plafonne l'intervention. Ce système serait efficace si la fluctuation des prix représentait l'unique source de risque (absence du risque de quantité) et si le mode de fonctionnement des fonds était stable. Or ce dernier varie significativement d'une année à l'autre, d'où la difficulté de mesurer leur efficacité.

7.1.2 La diversification

Celle-ci peut être pratiquée sous différentes formes :

- La diversification des cultures (ou assolement) permet à l'agriculteur de diversifier le risque prix de ses cultures. Cette activité est très réglementée au niveau européen du fait de la contrainte environnementale. Pratiquer plusieurs cultures permet à l'agriculteur de se positionner sur plusieurs marchés et donc de s'affranchir du caractère systémique du risque prix d'une agriculture monoculturale. L'optimisation des quantités de chaque type s'effectue donc au regard des prix de ventes, de la main d'oeuvre nécessaire, de la nature des sols ainsi que des coûts fixes et variables induits. Néanmoins, le revenu moyen est souvent plus bas du fait de coûts de production plus élevés. En effet, des équipements supplémentaires sont nécessaires, les effets d'économie d'échelle sont réduits et l'adaptation est souvent délicate en raison d'un manque d'expertise managériale.
- La diversification des moments auxquels il vend ces produits. Celle-ci peut se faire au niveau collectif ou individuel. Au niveau individuel, ce mode de gestion du risque prix n'est pas très développé car il nécessite de la part de l'agriculteur des aptitudes au maniement des contrats à termes ou de livraison différée afin de pouvoir arbitrer judicieusement sur les marchés. Au niveau collectif, les coopératives fournissent aux agriculteurs un prix correspondant au prix de vente moyen pour tous les apporteurs de la coopérative. L'appartenance à une coopérative permet donc aux agriculteurs de mutualiser leurs prix en fonction des périodes de vente. En raison de besoins de financement et de stratégies de gestion différents, ils ne vendent pas tous au même moment. Pour chaque campagne, la coopérative collecte les produits de ses différents apporteurs et verse un acompte au moment de la livraison. Celle-ci effectue ensuite des études de marché afin de mettre en place une stratégie optimale. Les bonus sont ensuite redistribués aux différents apporteurs en complément de l'acompte initial. L'effet de lissage est avéré car on constate des écarts entre le prix final payé à l'agriculteur et le prix moyen de marché.

7.1.3 L'intégration verticale

Il s'agit pour la firme de gérer plusieurs niveaux d'activité. Cela permet de réduire les risques associés à une variation de la quantité et de la qualité des inputs (intégration vers le bas) ou des outputs (intégration vers le haut). L'intégration verticale est principalement répandue dans le secteur du bétail (intégration vers le bas en ce qui concerne la production de fourrage) ainsi que dans celui des légumes frais (intégration vers le haut au sujet de l'assemblage et du packaging).

7.2 La gestion par des organismes externes : l'Etat, les compagnies d'assurance et les marchés financiers

Une partie importante des risques agricoles doit impérativement être traitée sur des marchés indépendants de l'activité elle-même. Certains risques tels le risque météorologique affectent systématiquement l'ensemble des agriculteurs d'une même zone et doivent donc être transférés à d'autres acteurs économiques.

7.2.1 Fonds publics ou Fonds Calamités

Ceux-ci sont régulés par les gouvernements et alimentés selon une base annuelle. Ils reçoivent également des contributions du secteur privé, généralement sous la forme de taxes obligatoires sur la production ou les primes. L'aide est fournie si la déclaration de catastrophe est prononcée. L'avantage principal de ces fonds par rapport à l'aide ad hoc réside dans le fait qu'ils permettent d'éviter une importante distorsion du budget gouvernemental.

7.2.2 Fonds mutuels

Il s'agit d'un moyen de partage du risque pour un ou plusieurs groupes de producteurs souhaitant conserver l'entière responsabilité du management du risque. Organisés selon une initiative privée, ils sont principalement mis en place à un niveau spécifique du secteur où les producteurs partagent des risques de même nature, ou alors à une échelle régionale. Ils peuvent être vus comme un système de compensation spécifique, mais dont la capacité de financement est néanmoins limitée. Lorsqu'un membre enregistre une perte, celle-ci est remboursée partiellement ou en totalité grâce à l'argent disponible dans le Fonds, selon des règles pré-définies (souvent avec une collecte additionnelle des participants). Néanmoins, les ressources disponibles sont souvent limitées, et ce en particulier durant les premières années d'existence. A cela s'ajoute le risque que plusieurs et même tous les agriculteurs subissent des pertes en même temps. Dans un tel cas, chaque agriculteur se doit outre sa perte personnelle de financer le Fonds afin de rembourser les pertes des autres membres. Les solutions à ce problème sont la réassurance ou la collaboration avec des Fonds mutuels d'autres régions.

L'un des avantages d'une organisation régionale est que les agriculteurs se connaissent très bien, ce qui réduit les problèmes liés à l'aléa moral et à l'anti-sélection.

Néanmoins, comme le dit le CEA, le statut légal des ces institutions n'est pas clair. Ils peuvent s'apparenter à des fonds de garantie, des fonds de solidarité ou encore à des mutuelles. Un grand nombre de mutuelles ont en effet des caractéristiques similaires. Ce sont des compagnies d'assurance appartenant partiellement ou en totalité à leurs adhérents. Comme les fonds mutuels, elles présentent comme principe l'absence de profit. Elles ne donnent donc pas lieu à l'émission d'actions et ne sont pas liés à des actionnaires : elles ne sont donc pas sur le marché. Par ailleurs, à la différence des compagnies d'assurance, elles possèdent un comité délégué représentant les agriculteurs.

La différence principale entre un fonds mutuel et une mutuelle d'assurance réside dans le fait que le fonds forme une organisation privée qui ne prévoit pas de remboursements de manière légale. Une mutuelle se doit en revanche de se conformer au règlement. Enfin, la participation dans le cadre d'un fonds correspond à un montant fixé indépendamment du risque, à la différence d'une mutuelle dans laquelle la cotisation est calculée de manière actuarielle.

7.2.3 L'assurance agricole

Celle-ci fonctionne généralement selon un principe indemnitaire. L'indemnité correspond à peu près au manque à gagner, contrairement au principe forfaitaire où un montant pré-établi est versé à chaque sinistre.

Les caractéristiques des assurances agricoles diffèrent selon que l'on considère le secteur du bétail ou des récoltes. L'assurance bétail couvre principalement les maladies non épidémiques ainsi que les accidents. L'assurance récolte la plus répandue concerne l'assurance contre la grêle, qui inclut souvent d'autres risques comme le risque incendie. D'autres polices d'assurance couvrent également le risque de gel ou un nombre limité d'événements météorologiques. Il s'agit des assurances "multi-risques".

On appelle assurance rendement le type de contrat qui couvre une perte de rendement correspondant à une culture en particulier et due à un certain événement météorologique. L'origine météorologique du dommage doit nécessairement être identifiée afin d'éviter les phénomènes d'aléa moral et d'antisélection. En général, tous les champs d'une ferme comportant la même culture doivent être assurés. L'assurance complète de rendement s'applique à l'ensemble des cultures produites par la ferme. Un perte concernant une culture ne sera pas remboursée si la réduction de production globale n'atteint pas le seuil fixé.

L'assurance revenu combine l'assurance rendement et l'assurance prix. L'agriculteur reçoit une indemnité si la valeur totale de sa production descend sous un certain seuil. L'assurance gain tient en plus compte des coûts de production. Elle n'est appliquée qu'aux Etats-Unis. Néanmoins, comme nous allons le voir, le risque prix est le plus souvent couvert grâce à des produits financiers.

Les types d'assurance précédents sont basés sur les résultats des agriculteurs pris individuellement et les pertes sont mesurées sur le terrain. Il existe également des systèmes d'assurance indicielle basés sur un indice commun à une certaine zone. Dans le cas de l'assurance basée sur un indice de rendement, le remboursement payé à l'agriculteur dépend du rendement statistique de l'année dans une région prédéterminée, généralement une unité administrative. L'assurance indicielle sur le revenu est basé sur le rendement sur une certaine zone multiplié par le prix dans cette même zone. Si le rendement ou le revenu moyens sont respectivement inférieurs à un certain seuil, l'ensemble des agriculteurs de la région assurés pour cette récolte sont indemnisés. Enfin, il existe des systèmes d'assurance basés sur des indices indirects. Ceux-ci ne font pas intervenir le rendement moyen dans une zone mais un indicateur météorologique ou des images satellite. Comme nous le verrons, ces derniers semblent les plus adaptés dans le cadre de notre étude.

7.2.4 Les produits financiers

Le risque prix peut-être transféré aux marchés financiers grâce à des produits dérivés.

Définition 7.1: *Un produit dérivé (ou plus simplement « dérivé » est un actif dont la valeur dépend d'autres variables plus fondamentales comme les prix d'autres actifs négociés sur les marchés tels les taux d'intérêt, les taux de change ou encore, comme nous allons le voir, les indices climatiques.*

Définition 7.2: *Le sous-jacent d'un produit dérivé est l'actif fondamental dont le prix dépend. Dans le cas présent, il s'agit de la récolte dont on veut couvrir le prix.*

Il existe principalement deux types de dérivés permettant de se couvrir contre le risque prix :
- Les contrats futures ;
- Les options .

Définition 7.3: *Un contrat futures, ou à terme, est, comme un contrat forward, un accord entre deux parties pour acheter ou vendre un actif donné à une date future pour un prix convenu. Contrairement aux contrats forward, les contrats futures sont négociés sur des marchés organisés. Les autorités de marché définissent des contrats standardisés pour assurer la liquidité. Comme, dans ce cas, les deux parties prenantes d'un contrat ne se connaissent pas, il existe un mécanisme qui permet d'assurer à l'acheteur et au vendeur la bonne fin des opérations.*

Ainsi, un agriculteur peut utiliser un tel contrat en vue de couvrir le risque de baisse du prix de sa production mais également de couvrir la hausse du prix des intrants. Ainsi, le prix de vente et les coûts deviennent certains. Etant donné qu'il y obligation de vendre ou d'acheter au prix fixé dans le contrat, l'agriculteur ne peut bénéficier des évolutions du marché. Ceci est rendu possible par le mécanisme des options.

Définition 7.4: *Il existe deux types d'options. Une option d'achat (call) donne le droit à son détenteur d'acheter une certaine quantité d'un actif sous-jacent à une date future donnée et à un prix convenu. Une option de vente(put) donne le droit à son détenteur de vendre une certaine quantité d'un actif sous-jacent à une date future et à un prix convenu. Ce prix est appelé prix d'exercice (strike price).*

Le fait de pouvoir bénéficier des évolutions favorables de marché a un coût : la prime d'émission.

Chapitre 8

Le cas des pays émergents

Les agents économiques des pays en voie de développement font le plus souvent appel aux solutions de micro-assurance. La micro-assurance agricole a pour objectif d'assurer les petits exploitants agricoles dans les pays en voie de développement. Ceci place les assureurs face à plusieurs difficultés, que sont notamment les risques d'antisélection, d'aléa moral et de fraude.

8.1 Le paysage de la micro-assurance et du micro-crédit agricoles

Comme nous l'avons vu dans la section 5, en l'absence de gestion adéquate, les risques agricoles peuvent nettement ralentir le développement économique, entraver la lutte contre la pauvreté et contribuer à l'apparition de crises humanitaires. Certains risques peu covariants (touchant des individus isolés) sont gérables au niveau interne (en diversifiant les cultures ou en mutualisant les risques au sein d'une communauté). En revanche, les risques covariants (affectant un grand nombre de personnes au même moment) sont particulièrement délicats à gérer et nécessitent une aide extérieure. Néanmoins, celle-ci est très souvent difficile à obtenir car les prestataires de services financiers ont réduit leurs activités dans les zones rurales.

8.1.1 La micro-assurance, le manque de couverture

Les assureurs privés sont généralement peu enclins à assurer les rendements des cultures et de l'élevage. Le manque d'informations permettant d'évaluer le risque complique considérablement la conception du produit.

Par ailleurs, les problèmes d'anti-sélection ainsi que d'aléa moral rendent généralement ces risques inassurables. Ces phénomènes sont particulièrement répandus dans le cas de l'agriculture et a fortiori dans les pays émergents et disposant de ressources en eau très restreintes. En effet, la quantité d'information sur les différentes catégories de risques est faible et la fiabilité pas forcément avérée, d'où un risque d'antisélection très important. Par ailleurs, les conditions souvent très délicates dans lesquelles évoluent les agriculteurs peuvent conduire à des comportements contre-productifs et donc à un aléa moral particulièrement important. Typiquement, considérons un agriculteur cultivant deux types de plantes et bénéficiant d'une assurance rendement sur l'une de ses cultures. Il serait alors tenté de consacrer l'ensemble de ses ressources en eau (irrigation) à sa culture non assurée, sachant que le rendement de l'autre est de toute façon garanti par l'assurance.

En outre, le coût de gestion des sinistres est extrêmement élevé dans le cas de l'assurance traditionnelle étant donné que des experts doivent se rendre chez chaque agriculteur afin de déterminer sur le terrain le montant de la perte subie. Enfin, du fait de la fréquence élevée ainsi que la nature covariante de certains risques, les indemnités à verser sont souvent élevées. Il s'ensuit que les primes sont souvent trop chères pour les agriculteurs en l'absence de subvention. Ainsi, les assureurs privés vendent le plus souvent des produits de micro-assurance contre les risques indépendants, à savoir des assurances vie, incendie et accidents. Lorsqu'ils proposent une assurance récolte, celle-ci ne concerne que des risques spécifiques comme les dommages causés par la grêle ou le gel.

En 2007, le Micro Insurance Centre a lancé une étude destinée à mieux comprendre la portée de la micro-assurance agricole dans le monde, en étudiant des aspects spécifiques de la micro-assurance tels que la distribution régionale, les types de couverture, les porteurs de risques, les régimes et programmes pilotes et, enfin, l'état de la réglementation. Les principales conclusions de ce travail sont les suivantes :

- Dans les pays en voie de développement, il existe très peu de régimes d'assurance agricole accessibles aux agriculteurs pauvres. L'étude en a répertorié un total de 122 au niveau mondial, dont certains ne sont pas opérationnels.
- La micro-assurance agricole se concentre en Amérique latine.
- Il y a très peu de régimes spécifiques et dédiés en matière de micro-assurance agricole. Ils fonctionnent en effet sur la base d'infrastructures et d'agents préexistants, qui ne font que perpétuer des modèles commerciaux non-viables.
- La micro-assurance agricole est fortement subventionnée et fonctionne sur la base de modèles commerciaux non durables

En ce qui concerne l'aide fournie par les gouvernements et organisations humanitaires, les problèmes principaux sont la difficulté de cibler les personnes qui en ont réellement besoin ainsi que le délai parfois important. Les gouvernements ne parviennent généralement pas à distinguer clairement les personnes pouvant payer une assurance de celles qui ne le peuvent pas. Cette confusion débouche le plus souvent sur des initiatives publiques très fortement subventionnées à destination de tous. Celles-ci sont donc très coûteuses pour le gouvernement et inefficaces du point de vue économique puisqu'elles n'encouragent pas les populations à prévenir les risques.

8.1.2 Le micro-crédit agricole et la difficulté d'obtenir des prêts

Cette exposition au risque ainsi que le manque de couverture engendrent une certaine réticence chez les banques et autres prestataires de services financiers à accorder des prêts. Ceci s'ajoute aux autres difficultés posées par la prestation de services financiers aux communautés rurales. En effet, la population se trouve dispersée sur une superficie plus importante et les infrastructures sont parfois défectueuses. Par ailleurs, les clients ont généralement besoin de produits différents de ceux des clients urbains et plus personnalisés. Il convient de surcroît de sensibiliser les populations aux différentes solutions. Tout ceci nécessite notamment des ressources en personnel importantes ainsi que des campagnes marketing.

8.1.3 Les conséquences économiques

Ainsi, en la quasi absence de protection, la seule solution viable passe l'utilisation de terres dans des zones peu risquées, ce qui peut générer des coûts d'opportunité élevés. En effet, il se peut que les terres les plus rentables ne soient pas exploitées en raison d'un risque trop important. Certaines études estiment que les revenus agricoles moyens pourraient être de 10 à 20 % plus

élevés en l'absence d'aversion au risque (Gautam, Hazell et Alderman 1994; Sakurai et Reardon 1997).

Par ailleurs, le manque de couverture des risques ainsi que la difficulté à obtenir des prêts n'incitent pas les agriculteurs à moderniser leurs outils de production. Ceci entraîne donc une sous-productivité qui contribue à les maintenir dans une situation précaire et les empêche de répondre aux défis posés par la mutation mondiale du secteur.

8.2 Conclusion

Comme nous l'avons vu, les mécanismes de micro-assurance inspirés des produits traditionnels des pays développés (tels l'assurance revenu ou l'assurance rendement) ne sont pas viables dans les pays en voie de développement. De plus, le support de l'Etat ainsi que les aides humanitaires sont le plus souvent insuffisants. Il s'avère donc absolument nécessaire d'élaborer de nouveaux outils de gestion des risques. L'assurance basée sur un indice (et notamment un indice climatique) ouvre des pistes très prometteuses. Nous étudions ce type d'assurance dans le chapitre suivant.

Chapitre 9

Le produit d'assurance à mettre en place : Assurance basée sur un indice climatique

9.1 Présentation du concept d'assurance sur indice

L'assurance indicielle est un produit financier lié à un indice présentant une forte corrélation avec les rendements locaux. Les contrats sont rédigés de façon à protéger le contractant contre des risques ou des événements spécifiques (par exemple perte de rendement, sécheresse, ouragan, inondation) définis et consignés à l'échelle régionale (par exemple dans une station météorologique locale). Les indemnisations sont déclenchées lorsque l'indice atteint une certaine valeur ou possède une certaine tendance et ne sont pas basées sur les rendements effectifs. Tous les acheteurs d'une même région se voient proposer les mêmes conditions contractuelles par dollar de couverture d'assurance. En d'autres termes, tous paient le même taux de prime et, lorsqu'un événement déclenche une indemnisation, tous reçoivent le même taux d'indemnisation. L'indemnisation totale est alors proportionnelle à la valeur de la couverture d'assurance souscrite. Le montant des indemnités peut être structuré de différentes façons. On peut notamment trouver des contrats :

- de type "tout ou rien" : lorsqu'un seuil est atteint, le taux de paiement s'élève à 100 % ;
- utilisant un barème de paiement échelonné (par exemple un taux de paiement d'un tiers lorsqu'un premier seuil est atteint, deux tiers quand un deuxième est dépassé, et 100 % quand le dernier est atteint) ;
- avec remboursement proportionnel à l'écart avec le seuil.

L'assurance basée sur un indice permet notamment de pallier les difficultés évoquées dans le chapitre 8 :

- Le risque d'antisélection : Le risque auquel est soumis la société d'assurance dépend uniquement de la valeur de l'indice. Tous les souscripteurs d'un même contrat payent la même prime et reçoivent la même indemnité par unité assurée. Ainsi, le risque est totalement indépendant des différents profils des agriculteurs : il n'y a pas de bons ou de mauvais risques. Par exemple, l'assureur ne supporte que le risque climatique et non pas le risque lié au retard technologique.
- L'aléa moral : De la même façon, celui-ci est totalement éliminé étant donné que le comportement de l'agriculteur ne peut influencer un indice basé sur les conditions météorologiques (sous réserve que l'organisme de collecte des données météorologiques soit indépendant).

– Les coûts de gestion : ceux-ci sont fortement réduits étant donné que seules les mesures météorologiques sont nécessaires pour connaître la valeur de l'indice. Ainsi, les contrats pourraient quasiment prendre la forme de chèques voyages ou de billets de loterie à disposition de tout acheteur intéressé. Néanmoins, la législation relative aux assurances (assez stricte dans de nombreux pays) n'autoriserait pas cette option.

En outre, le produit d'assurance étant basé sur un indice vérifiable de façon indépendante, il peut être réassuré. Une partie du risque peut même être transférée aux marchés internationaux via des produits de type dérivés climatiques ou cat bonds.

Néanmoins, l'assurance indicielle possède également des inconvénients :
– Son lancement s'avère coûteux car des ressources et une expertise technique importantes sont nécessaire pour mener à bien les travaux initiaux de recherche et de développement.
– Le risque de base, c'est-à-dire le risque de corrélation entre le rendement réel et l'indice, peut être important.
– Elle nécessite de surcroît de renforcer les capacités des assureurs locaux ainsi que des autres intervenants des réseaux de distribution. Il convient également de sensibiliser de manière efficace les clients potentiels.
– L'accès aux données climatiques peut parfois être délicat.

9.2 Utilisation d'un tel indice

L'assurance indicielle est potentiellement utile à différents niveaux. Au niveau micro, les agriculteurs peuvent bénéficier d'une stratégie de gestion des risques supplémentaire. Au niveau méso, les prestataires de services financiers ainsi que les fournisseurs d'intrants peuvent se couvrir contre les défauts de paiement. Enfin, au niveau macro, l'assurance indicielle peut épauler les gouvernements et les organisations humanitaires dans leurs démarches de développement et de gestion des catastrophes naturelles.

Des études menées en matière de protection sociale (Grosh et al. 2008) ainsi qu'une publication récente de l'Institut international de recherche pour le climat et la société (Hellmuth et al. 2009) montrent que l'assurance indicielle peut être envisagée soit comme outil d'assistance en cas de catastrophe, soit comme outil de développement.

9.2.1 L'assurance indicielle comme outil d'assistance en cas de catastrophe

Les organisations humanitaires publiques ou encore les ONG pourraient tout à fait recourir à l'assurance indicielle dans le cas d'événements extrêmes tels les ouragans, les inondations ou encore les sécheresses sévères. L'indice se doit alors d'être sensible à des événements covariants et extrêmement peu fréquents (prise en compte des valeurs extrêmes de températures et précipitations).

Une première approche pour l'organisation est de contracter une assurance basée sur un indice climatique et de financer ses propres efforts humanitaires grâce aux indemnités versées par la compagnie d'assurance.

L'avantage d'un indice objectif est qu'il permet à l'assureur d'effectuer des versements rapides aux organisations humanitaires et aux foyers. Ceci permet ainsi d'éviter les retards habituels subis lorsque les organisations humanitaires se doivent dans un premier temps de démontrer qu'il existe une situation d'urgence, puis de faire appel à la générosité des gouvernements ainsi que des donateurs. Des études ont démontré que l'aide humanitaire protège d'autant mieux les populations des impacts sociaux négatifs qu'elle arrive tôt après un bouleversement. Elle leur

évite ainsi de devoir vendre leurs actifs dans l'urgence et accélère par là-même le processus de rétablissement (Dercon, Hoddinott et Woldehanna 2005).

Une autre approche serait de distribuer directement des bons d'assurance aux foyers ciblés. Ceux-ci pourraient alors les faire valoir en cas de situation dangereuse. L'avantage de cette approche est qu'elle permet de mieux cibler les foyers les plus vulnérables et donc d'éviter les attributions hasardeuses susceptibles de se produire lorsque des aides d'urgence sont dispensées dans la précipitation.

9.2.2 L'assurance indicielle comme vecteur de développement

On parle ici d'une assurance visant à aider les foyers ainsi que les organismes financiers à se protéger des risques covariants peu à moyennement fréquents.

Ceci aide les agriculteurs à adopter des stratégies agricoles plus risquées mais également plus rentables. Elle leur permet par ailleurs d'accéder aux marchés de produits à valeur élevée, aux intrants ainsi qu'aux technologies modernes car les prestataires financiers ainsi que les fournisseurs d'intrants accordent beaucoup plus facilement des prêts à des agriculteurs assurés. Dans cette même perspective, on pourrait tout à fait imaginer un lien formel entre l'assureur et le prestataire de façon à garantir au prêteur un accès direct à une partie de l'indemnité versée par l'assurance.

Un autre approche possible pour les organismes financiers serait d'utiliser une assurance indicielle en vue de couvrir leur propre portefeuille de crédits contre les événements peu fréquents à covariance élevée. Ceci les protégerait d'une vague de défauts de paiement généralisés.

Des études ont montré le caractère bénéfique de l'assurance indicielle dans la croissance du niveau de vie des agriculteurs. Néanmoins, certains problèmes subsistent :

- L'assurance indicielle devrait s'accompagner des subventions au moins dans les premières phases de son développement, afin d'aider les agriculteurs à payer les primes.
- La vente de tels produits directement aux individus apparaît très délicate. Le recours à des instances fédératrices (par exemple les entreprises de transformation des produits agricoles, les fournisseurs d'intrants, les prestataires de services financiers, les associations d'agriculteurs) semble cruciale pour réduire les coûts de transaction et élargir la portée des produits à une échelle significative.

Dans notre étude, le lien entre l'assureur et les agriculteurs n'est pas direct mais s'effectue par l'intermédiaire de la coopérative agricole, la COSUMAR.

9.3 Les différentes catégories d'assurances indicielles

Les assurances sur indices sont divisées en deux groupes :

- celles basées sur indices régionaux : l'indice correspond directement au rendement/revenu moyen dans une certaine région ;
- celles basées sur indices indirects : par exemple les indices climatiques ou encore les indices de végétation obtenus à partir d'images satellites.

La deuxième catégorie est la plus complexe.

9.4 L'historique des solutions d'assurance basées sur indices

Le concept d'assurance indicielle n'est pas récent. Halcrow (1948) et Dandekar (1977) ont été les premiers à proposer ce type d'assurance.

9.4.1 Au niveau mondial

La première catégorie d'indices (régionaux) a été expérimentée depuis quelques années ou décennies dans certains pays tels le Brésil, le Canada, l'Inde ou encore les Etats-Unis (Miranda 1991 ; Mishra 1996 ; Skees, Black et Barnett 1997). Ce type d'assurance est le plus souvent basé sur les rendements d'une région homogène. Un exemple est le GRP (Group Risk Plan) aux Etats-Unis. Un autre type d'assurance disponible aux Etats-Unis concerne le GRIP(Group Risk Income Protection) dans lequel l'indice représente le revenu régional, c'est-à-dire le rendement régional multiplié par le prix du produit. En 2004, ces deux types d'assurances représentaient 7,4 % de la superficie assurée totale mais moins de 3 % des primes.

Un exemple particulier concerne la Mongolie. Une assurance basée sur indice régional pourrait voir le jour sous peu dans le secteur du bétail, et serait basée sur les taux de mortalité régionaux. Ceci est possible car la Mongolie effectue un recensement annuel de chaque espèce (Skees et al. 2005).

En ce qui concerne les solutions d'assurance basées sur des indices indirects, celles-ci sont très récentes et seulement en phase d'étude dans la plupart des pays. Le gouvernement australien avait demandé une étude de faisabilité portant sur l'assurance précipitations au milieu des années 1980 mais a finalement décidé de ne pas donner suite (IAC 1986).

La Banque Mondiale vise clairement à promouvoir le développement de telles solutions. Ainsi, une assurance basée sur indice climatique a été proposée en 2005 au gouvernement du Nicaragua mais celui-ci l'a considérée comme inutile.

9.4.2 Au niveau européen

Les assurances basées sur indices régionaux n'existent quasiment pas en Europe. Seuls la Suède et le Royaume-Uni ont connu des programmes subventionnés. Au Royaume-Unis, une telle assurance a été lancée en 1998. L'indice reposait sur les statistiques de rendement du HGCA (Home Grown Cereals Authority) ainsi que sur les prix des matières premières du LIFFE (London International Financial Futures). La couverture fournissait une indemnité pour une baisse de 10 % de rendement et de 5 % de prix. Les taux de primes variaient de 1,1 à 3,5 %. L'offre a été annulée la saison suivante.

En ce qui concerne les indices indirects, nous pouvons essentiellement citer deux exemples.

Le premier pays où une telle assurance a été commercialisée de manière durable est probablement l'Espagne. Un produit d'assurance pour pâturage est disponible depuis 2001. Il est basé sur un indice de végétation calculé à partir d'images satellites de résolution assez grossière.

En Autriche, une assurance sécheresse basée sur indice climatique a été commercialisée pour la première fois en 2007.

9.5 Le produit à mettre en place dans la cas présent

Ici nous souhaitons mettre en place une telle assurance basée sur un indice climatique. Les autres types d'assurances sur indices apparaissent peu appropriés. En effet, un indice basé sur rendement régional pose le problème des coûts et de la fiabilité des mesures. Par ailleurs, l'utilisation d'un indice indirect basé sur images satellites apparaît impossible car il est très délicat de mesurer le rendement de manière précise à l'aide de telles images. Seuls des enregistrements satellites fournissant des informations au sujet de la sécheresse des sols ou de certaines variables climatiques auraient pu s'avérer utiles mais nous n'en possédions pas.

Etant donné que nous disposions des relevés météorologiques dans différentes stations, il est apparu naturel de s'intéresser à un indice climatique. Nous possédons également des données de rendement en sucre. Notre but est donc de fabriquer un indice basé sur les variables météorologiques présentant une forte corrélation avec le rendement.

Une indemnité sera versée dès que la valeur de l'indice sera inférieure à un certain seuil (par exemple 50 % du revenu moyen) et ce de manière proportionnelle à l'écart entre ce seuil et la valeur de l'indice.

Même un indice ne représentant que les variations autour de la moyenne s'avère très difficile à constituer. En effet, quand on demande aux agronomes d'expliquer l'évolution du rendement, on s'aperçoit que chaque année correspond à un cas particulier. Ceci est très visible sur le graphique 9.1. On constate qu'il n'y a pas réellement d'argument général pouvant expliquer la baisse ou la hausse des rendements. Ainsi, on se doute dès à présent de la difficulté d'élaboration de cet indice.

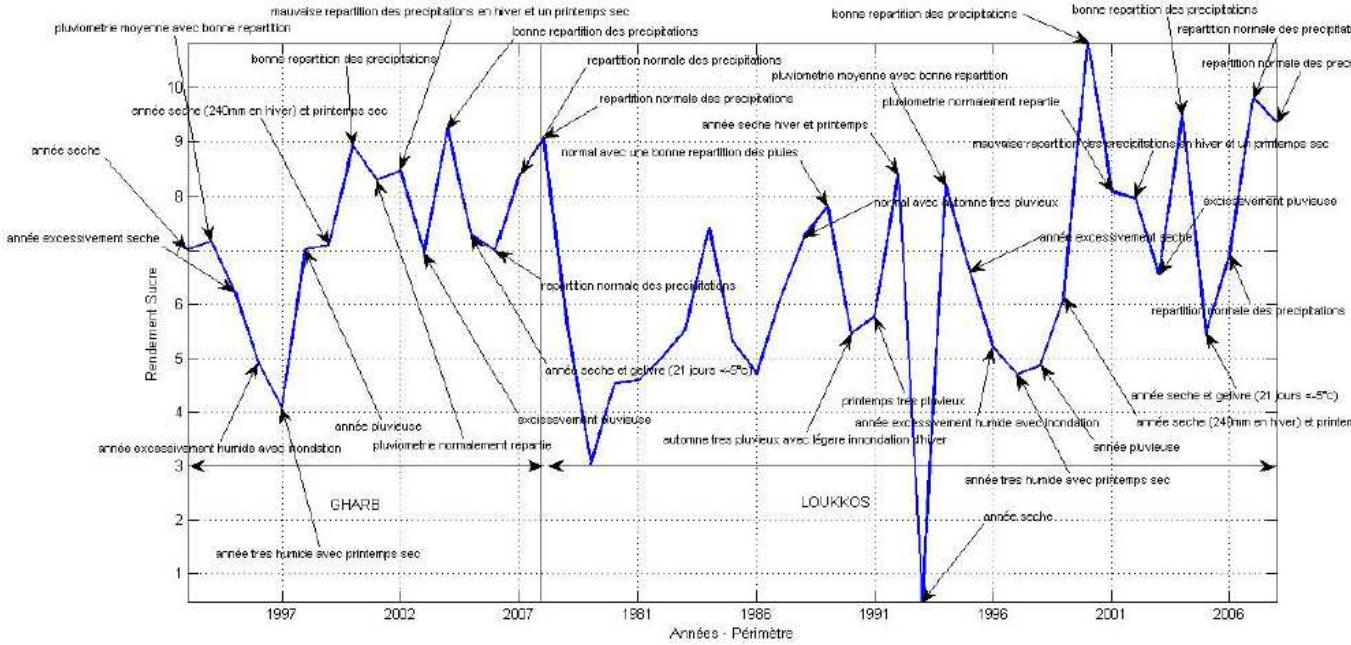


FIGURE 9.1 : Explication de l'évolution du rendement pour la zone Gharb

Chapitre 10

Présentation et prétraitement des données

10.1 Présentation des données

Nous présentons ici les deux grandes catégories de données que nous tentons de mettre en correspondance, à savoir les données agronomiques (rendements en sucre) et les données climatiques. Auparavant, il convient de décrire rapidement le contexte géographique de l'étude. Celle-ci porte sur la globalité de la production sucrière du Maroc. Comme nous pouvons le constater sur la figure 10.1, l'exploitation s'effectue dans plusieurs périmètres de production. Chacun d'eux est découpé en différentes zones comme on peut le voir dans le tableau 10.1.

Périmètre	**Zone**
Doukkala	Doukkala
	Sidibennour
	Zemamra
Gharbloukkos	Gharb
	Kek
	Loukkos
	Mbk
	Sidi Allal Tazi
	Sidi Sliman
Moulouya	Berkane
	Moulouya
	Nador
Tadla	B.Amir
	B.Moussa
	Tadla

TABLE 10.1 : Périmètres et zones de production

Nous disposons d'un hitorique de 31 ans (de 1978 à 2008). Néanmoins, celui-ci n'est complet que pour 8 zones. **Le nombre de données disponibles est donc faible**, ce qui rend très délicate la conception d'un indice fiable. Ce problème est commun à la plupart des études agricoles. En effet, les cycles sont en général annuels et ne donnent lieu qu'à une valeur de rendement par

an.

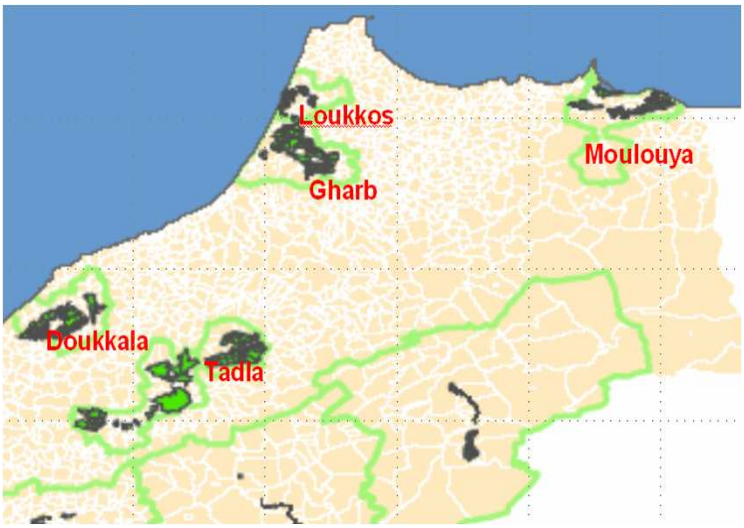


FIGURE 10.1 : Périmètres de production

Les données ont été fournies par la COSUMAR et peuvent être considérées comme fiables. Même si les méthodes de mesures évoluent avec le temps et que par conséquent certains retraitements ont été nécessaires, le problème principal ne concerne pas l'âge des données mais leur nombre.

10.1.1 Données agronomiques

Compte-tenu de la capacité limitée des sucreries, le traitement de la betterave sucrière au niveau des usines doit se faire progressivement et de ce fait, il convient d'étaler l'arrachage et donc également les semis. Ainsi, deux paramètres sont à prendre en considération lors de la mise en place de la culture, à savoir la date de semis et la date de récolte, ce qui fixe la durée du cycle. L'effet du cette dernière sur le comportement de la betterave a fait l'objet de multiples investigations de la part d'auteurs dont on peut citer Hull(1970) et Jaggar(1983). Généralement, la durée optimale se situe entre 220 et 250 jours pour les semis précoces de septembre-octobre et entre 170 et 220 jours pour les semis tardifs de fin décembre ou début janvier, et ce quel que soit le type de variété.

En effet, on distingue essentiellement trois types de vaiété de betterave sucrière :

- Recoltapoly, type Z. Ce dernier possède la particularité d'avoir un cycle court, un rendement en racines faible par rapport aux autres types de variété mais en revanche une plus grande richesse en sucre.
- Sultan, type N. Le cycle est moyen, le rendement en racines et la richesse en sucre présentent également des valeurs moyennes.
- Desprez poly E, type E. Le cycle est long, le rendement en racines élevé mais la richesse en sucre faible.

Les données fournies par la Cosumar contiennent pour chaque zone les chiffres de rendement total en racines (en $t.ha^{-1}$) ainsi que la richesse moyenne en saccharine (en %) et le rendement total en sucre obtenu en faisant le produit des deux grandeurs précédentes.

Afin d'obtenir la donnée de richesse en saccharine, un échantillon des racines récoltées est prélevé, lavé et râpé suivant les procédures décrites par le comité technique de normalisation des sucres (SNIMA, 2005) à l'aide d'un saccharimètre électronique à quartz. La polarisation est exprimée en pourcentage de sucre brut contenu dans la râpure. Néanmoins, un certain nombre d'éléments diminuent la quantité de sucre que l'on peut extraire par les méthodes classiques : il s'agit des éléments mélassigènes que sont Na, K et N α-aminé. Les concentrations de ces éléments dans le jus de sucre fournissent des indications au sujet de la quantité de sucre mélasse, c'est-à-dire le sucre qui ne peut pas être extrait par le processus classique des sucreries.

Comme on a pu le constater, **les données de rendement sur lesquelles nous effectuons notre étude prennent en fait en compte à la fois des cycles mais également des variétés distinctes**. Or, comme nous le verrons par la suite, les conséquences d'un même facteur météorologique peuvent être bénéfiques ou néfastes selon le degré de maturité de la plante. Il apparaît donc d'autant plus délicat d'exhiber l'impact de telle ou telle variable. On peut donc parler **d'imprécision dans les données de rendements**.

10.1.2 Données climatiques in situ

Nous disposons, pour chaque périmètre et zone décrits précédemment, d'un historique de relevés mensuels pour les données suivantes :

- température moyenne ;
- précipitations moyennes ;
- température minimale moyenne ;
- température maximale moyenne ;
- degrés-jours moyens (degrés de bénéfice pour la plante) ;
- les données de barrage.

Les données de barrage étant très incomplètes, elles s'avèrent véritablement inexploitables. Nous les laissons par conséquent de côté dans le reste de l'étude. En outre, il est très important de souligner que les données d'irrigation ne sont pas connues. **Nous ne disposons donc pas des quantités d'apport en eau artificiel**. Ceci rend évidemment notre étude très compliquée, étant donné que notre indice ne pourra pas expliquer la variance de rendement due à la variabilité de l'irrigation. Il y a véritablement **manque d'informations**.

Chronologiquement, les premières données correspondent à août 1978. Le cycle de production se terminant en juillet, on ne peut donc pas expliquer le rendement 1978 à l'aide des variables climatiques. De même, les données climatiques postérieures à juillet 2008 sont inutiles car nous ne connaissons pas le rendement 2009. Ainsi, l'étude se base sur un historique de 30 ans.

10.2 Prétraitement des données

Il faut bien comprendre que la première étape de toute étude d'impact concerne le traitement des séries temporelles que nous avons en notre possession. Toute série temporelle peut être décomposée en trois termes que sont la tendance globale, la saisonnalité, et le résidu. Ainsi, pour une série quelconque S_t, on a :

$$S_t = z_t + s_t + r_t$$

où z_t représente la tendance, s_t la saisonnalité et r_t le résidu. Afin d'obtenir le modèle, il est nécessaire, comme nous le verrons par la suite, de mettre en relation les anomalies (résidus normalisés). Dans le cas contraire, il est impossible de se limiter à l'influence climatique. Par ailleurs, subsistent alors des auto-corrélations entre les variables.

Le prétraitement diffère selon que l'on s'occupe des variables météorologiques ou agronomiques. Les premières évoluent de manière naturelle suivant le cycle saisonnier mais aussi suivant une tendance globale (qui correspond au changement climatique). En revanche, étant donné que l'on observe une seule récolte par an, les données de rendements sont annuelles. Elles ne présentent que tendance globale et résidu.

10.2.1 Variables climatiques

Nous utilisons une méthode de régression de la série qui nous permet d'exhiber simultanément z_t et s_t. On effectue cette régression sur une matrice obtenue par une fonction matlab (dummyvar) à laquelle on rajoute une colonne. Les 12 premières colonnes (issues de dummyvar) correspondent à la saisonnalité et la dernière à la tendance. Notre matrice de régression contient ainsi dans ses 12 premières colonnes la valeur 1 dans celle correspondant au mois de la valeur condidérée de la série et 0 dans les autres. On obtient ainsi le coefficient correspondant à chaque mois, ce qui donne la saisonnalité. La tendance est obtenue grâce à la dernière colonne contenant les dates. On obtient ainsi directement la série $z_t + s_t$. Nous avons tracé celle-ci pour la température et la précipitation dans le cas de Ben Amir. Ainsi, il est très facile d'en déduire l'anomalie a_t :

$$a_t = \frac{r_t}{\sigma(S_t)} = \frac{S_t - (z_t + s_t)}{\sigma(S_t)}$$

Remarque 10.1: *On observe clairement sur la figure 10.2 la tendance moyenne à la hausse à laquelle s'ajoute le cycle saisonnier. On met ainsi en évidence le réchauffement climatique. En ce qui concerne les précipitations, on constate sur la figure 10.3 qu'elles sont en baisse, d'où une sécheresse accrue. Il s'agit là de la mise en évidence d'une conséquence prévue du changement climatique. Enfin, les anomalies de précipitation ont une plus grande amplitude que celles des températures (valeurs de -0,2 à 0,3 contre -4 à 6). Les précipitations sont plus aléatoires et difficiles à prévoir.*

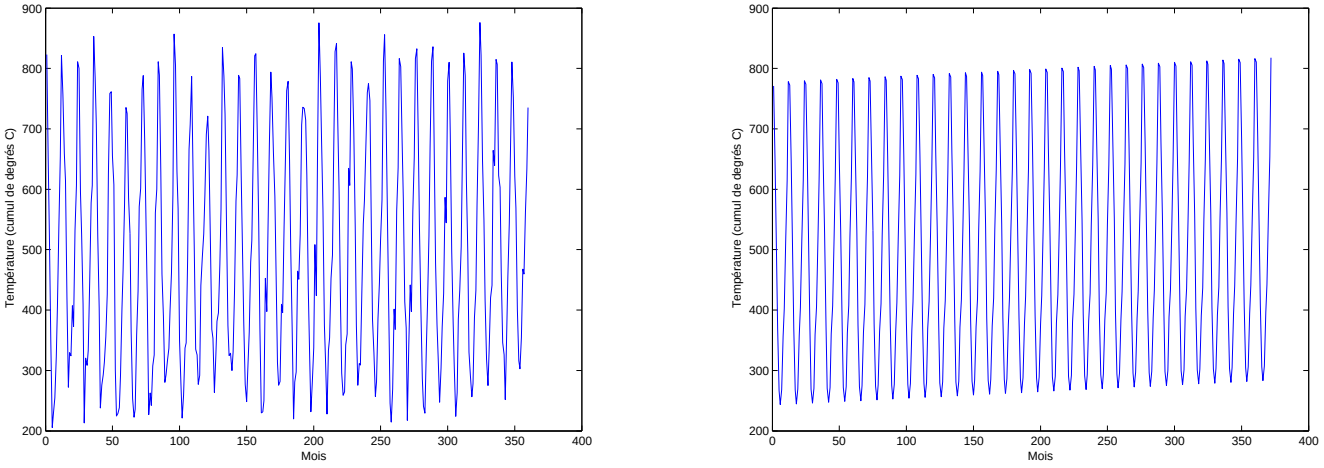


FIGURE 10.2 : Série de températures mensuelles (g) et tendance associée (avec saisonnalité) (d) dans la zone de Ben Amir

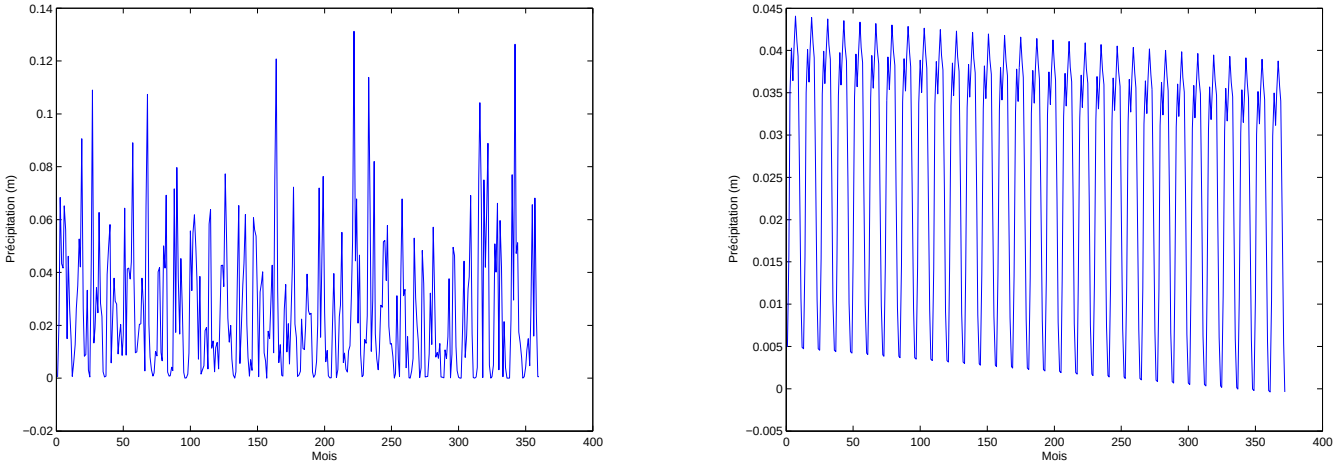


FIGURE 10.3 : Série de précipitations mensuelles (d) et tendance (avec saisonnalité) associée (d) dans la zone de Ben Amir

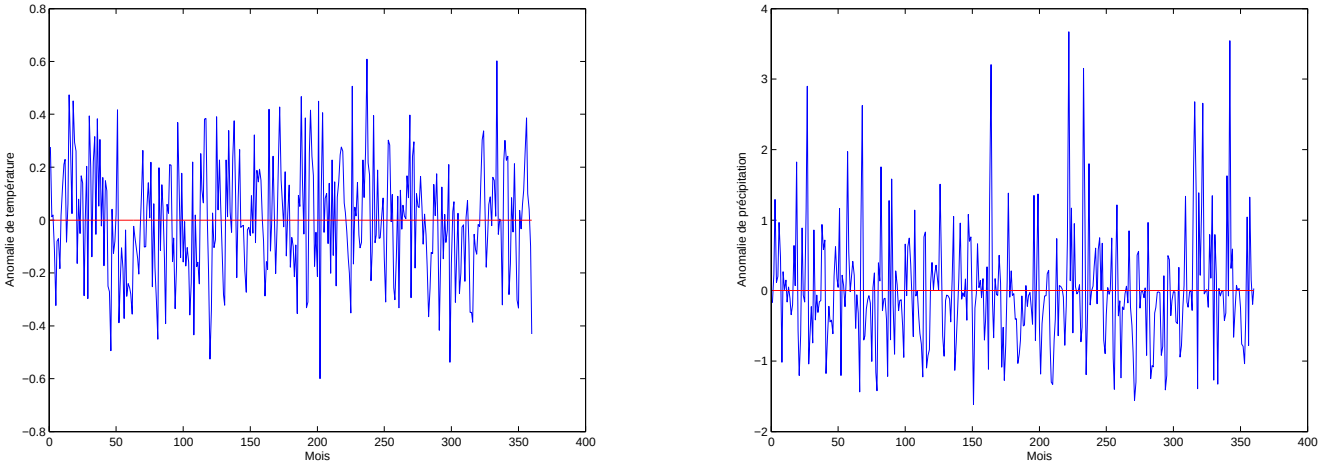


FIGURE 10.4 : Séries des anomalies de températures et précipitations mensuelles dans la zone de Ben Amir

Les graphes des autres variables climatiques (température minimale, maximale et degrés-jour) et de leur tendance associée ainsi que de leurs anomalies pour cette même zone sont renvoyés en annexes A.1 page 111.

10.2.2 Rendements

En ce qui concerne les données de rendements, compte-tenu du profil chaotique de leur évolution, il faut bien comprendre que la modélisation de tendance revêt un certain caractère subjectif. Il convient **d'introduire dans ce modèle de tendance l'ensemble des a priori que nous pouvons avoir au sujet des phénomènes autres que les conditions climatiques**. Dans le cas présent, il peut s'agir de nouvelles espèces de plantes ou encore de ruptures technologiques

modifiant considérablement le rendement. Certains de ces a priori peuvent être validés mais d'autres ne sont pas vérifiables en raison d'un manque d'information. Ainsi, on essaie de prendre en compte dans la tendance tous les facteurs indépendants des causes climatiques. De ce fait, on peut légitimement espérer que les anomalies calculées sont expliquées en majeure partie par les anomalies associées à certaines variables climatiques.

Il faut bien comprendre que l'obtention des anomalies grâce à ce prétraitement est absolument cruciale. La plupart des études de sensibilité (en climatologie ou en finance) ne détaillent en général quasiment pas cette étape alors qu'il s'agit en fait de **la pierre de base de l'ensemble de l'étude**. Il faut bien avoir à l'esprit que les hypothèses que l'on fait pour modéliser la tendance de rendements conditionnent la suite de la démarche. Idéalement, un processus de rétroaction sur ces hypothèses serait nécessaire. Néanmoins, ceci demande des calculs assez coûteux.

Dans le cas de Ben Amir, comme on peut le voir sur la figure 10.5, on voit assez clairement émerger **une structure en escalier que l'on peut interpréter comme une rupture technologique**. En tous cas, il apparaît clair qu'une telle discontinuité ne peut être imputable aux conditions climatiques. Il convient donc de l'intégrer à la tendance afin de l'éliminer dans le reste de l'étude.

Dans d'autres cas, comme dans la zone de Berkane, il convient de choisir une tendance quadratique. Les graphes des rendements en sucre et tendances associées pour les 8 zones pour lesquelles nous possédons des données complètes sont présentés en annexe A.2 page 111.

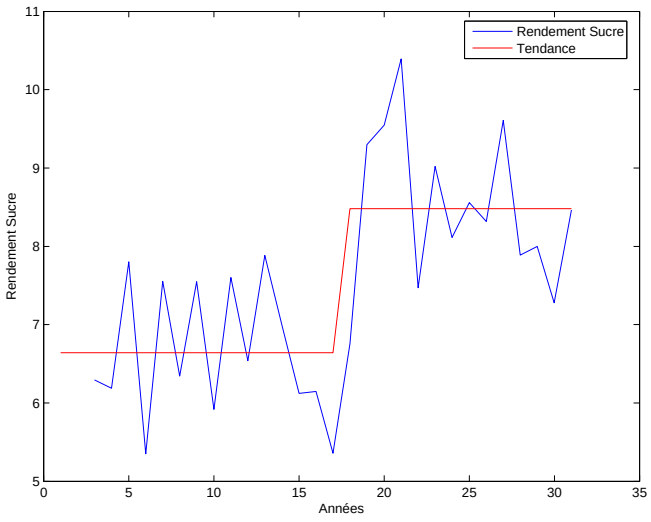


FIGURE 10.5 : Evolution du rendement en sucre dans la zone de Ben Amir et tendance associée

10.2.3 Tests sur données traitées

Un processus X_t est qualifié de bruit blanc si les X_t sont indépendants et identiquement distribués. On parle de bruit blanc gaussien si les X_t suivent $N(\mu, \sigma)$.

Différents tests de bruit blanc existent. Pour ce qui est du bruit blanc gaussien :

Un test basé sur la valeur des R_k

On dispose d'une réalisations du processus X_1 , X_2 , ..., X_T. Sous l'hypothèse H_0 de bruit blanc gaussien, pour T grand,

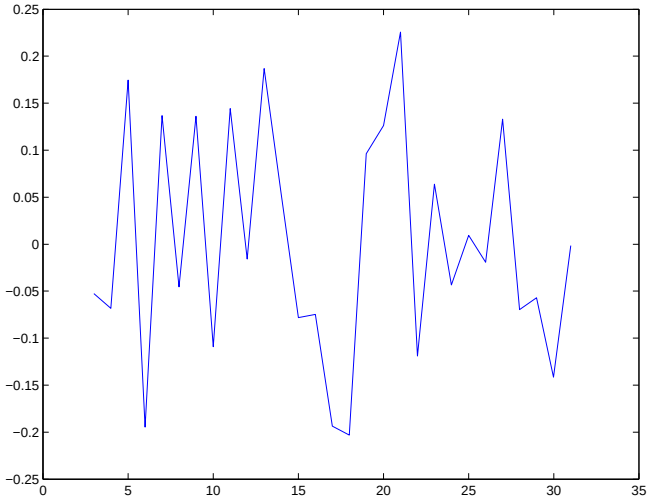


FIGURE 10.6 : Anomalies de rendements en sucre dans la zone de Ben Amir

$R_k \longrightarrow N(\frac{-1}{T}, \frac{1}{\sqrt{T}})$ Or, pour T grand, $\frac{1}{T} \approx 0$ donc on rejette l'hypothèse de normalité si $R_k > \frac{1.96}{\sqrt{T}}$.

En pratique, on trace sur le corrélogramme les droites d'équation $\frac{2}{\sqrt{T}}$ et $\frac{-2}{\sqrt{T}}$.

Le test de Box et Pierce

On choisit un entier K positif. Sous l'hypothèse H_0 de bruit blanc gaussien, $T\sum_{k=1}^{K} r_k^2 -> X_K^2$ en distribution. On effectue le calcul avec différentes valeurs de K (notamment 12 ici étant donné que l'on considère des anomalies mensuelles)

Le test de Ljung et Box

Pour un entier positif K choisi et sous l'hypothèse de bruit blanc H_0, $T(T+2)\sum_{k=1}^{K} \frac{r_k^2}{T-k} \longrightarrow X_K^2$

Ce test est souvent plus robuste que celui de Box et Pierce.

L'hypothèse de stationnarité est généralement accepté de justesse au risque de 5 %. Cela montre que nos séries d'anomalies sont à peu près stationnaires. Le test de Kolmogorov-Smirnov montre par ailleurs qu'elles ont une distribution à peu près normale.

Néanmoins, ces tests ne sont pas très importants (c'est pour cela que nous ne détaillons pas les résulats) car nos séries d'anomalies ne doivent pas nécessairement s'apparenter à des bruits blancs. En effet, on espère qu'elles contiennent le plus d'informations possible. Si les rendements en sucre présentent une certaine structure d'auto-corrélation, on peut espérer qu'il en soit de même pour les variables climatiques. Le but du prétraitement effectué n'est pas d'obtenir un bruit blanc.

Chapitre 11

Explication de la méthodologie

11.1 Les différents constituants de l'indice

L'idée est d'évaluer la sensibilité de l'activité, ici en l'occurence la production agricole, à la variabilité climatique. Il convient donc de trouver une relation entre les anomalies de rendements en sucre et les anomalies de variables météorologiques à déterminer.
Il s'agit donc de trouver :

- **les bons prédicteurs (variables climatiques pertinentes) ;**
- **le modèle approprié (linéaire, non linéaire, hiérarchique) représentant la relation entre les anomalies de variables climatiques (entrées) et les anomalies de rendements (sorties).**
- **les paramètres du modèle.**

11.1.1 Choix des prédicteurs

Nous faisons face dans cette étude à un problème délicat dans la mesure où nous possédons un grand nombre de variables explicatives potentielles, néanmoins sans savoir a priori lesquelles peuvent s'avérer pertinentes. Commme nous l'avons vu dans la partie 10.2, les 5 grandeurs que sont les anomalies de température, précipitation, température minimale, température maximale et degré-jour sont susceptibles d'avoir un impact. Nous disposons donc au total de 60 prédicteurs potentiels. Afin de sélectionner la combinaison de ceux-ci la mieux reliée aux données de rendement, **deux méthodes sont envisageables :**

- La première utilise l'expertise agronomique.
- La seconde s'avère être **purement statistique** : elle construit de manière itérative **la combinaison optimale**.

Nous les étudions de manière très détaillée dans les sections 11.5 et 11.6.

11.1.2 Modèles

Nous étudions ici principalement deux types de modèle : le modèle linéaire (dans lequel les sorties sont des fonctions linéaires des entrées) et le modèle non linéaire (réseaux de neurones).

Modèle linéaire

Nous cherchons à régresser une variable aléatoire Y (dans notre cas le rendement en sucre) sur un nombre **d de variables aléatoires** (nos prédicteurs). Le modèle $Y = {}^t\beta X + \epsilon$, avec

Y variable aléatoire sur $\Re$, $X \in \Re^d$, $\beta \in \Re^d$ permet ainsi de prédire la variable Y à partir des d variables contenues dans le vecteur aléatoire X. Pour une réalisation de l'ensemble de ces variables aléatoires (c'est-à-dire une année dans notre problème), on a donc :

$$y^1 = {}^t\beta x^1 + \epsilon$$

Dans le cas où l'on dispose de **P observations** i.i.d. $(x^1, y^1), ..., (x^P, y^P)$ de même loi que (x^1, y^1), il s'agit du modèle d'échantillonnage décrit par le système d'équations :

$$y^e = {}^t\beta x^e + \epsilon^e, \quad e \in 1, ..., P$$

où les $\epsilon_1, \cdots, \epsilon_P$ sont des variables aléatoires i.i.d. et indépendantes de $X^1, \cdots, X^P$. Il est pratique d'adopter une écriture matricielle de ce système. On pose désormais les notations suivantes :

$$X = (x^1, ..., x^P)^T$$

matrice P*d

$$Y = (y^1, ..., y^P)^T \in \Re^P$$

$$\epsilon = (\epsilon_1, ..., \epsilon_P) \in \Re^P$$

Le système décrit précédemment s'écrit alors :

$$Y = X\beta + \epsilon$$

On note que dans ce modèle, le vecteur aléatoire ϵ suit une loi gaussienne multivariée $N(0, \sigma^2 I_n)$. Il y a un donc un paramètre d'intérêt qui est β et un paramètre de nuisance qui est la variance σ^2 du bruit. L'estimation de β permet d'utiliser le modèle en prédiction mais celle de σ s'avère également utile pour l'analyse.

L'estimateur classique dans le contexte de la régression linéaire est l'estimateur des moindres carrés. Il est obtenu en résolvant le problème de minimisation de l'erreur quadratique

$$\min_{\beta \in \Re^d} \sum_{e=1}^{n} (y^e - {}^t\beta x^e)^2 = \| Y - X\beta \|^2$$

On le résout en notant que la solution est la projection orthogonale notée $X\hat{\beta}$ de Y sur l'espace $\{X\beta : \beta \in \Re^d\}$. On a donc :

$${}^t(X\beta)(Y - X\hat{\beta}) = 0, \forall \beta \in \Re^d$$

, ce qui est équivalent à :

$${}^tX(Y - X\hat{\beta}) = 0$$

ou encore :

$${}^tXY = {}^tXX\hat{\beta}$$

On en déduit alors l'estimateur des moindres carrés de β et celui de σ^2 par substitution :

$$\hat{\beta} = ({}^tXX)^{-1} \quad {}^tXY$$

Nous ne prenons pas ce dernier en compte étant donné qu'il ne sert pas à la prévision.

Réseaux de neurones

Nous utilisons dans cette étude un réseau de neurones particulier, le Perceptron Multi-Couches (PMC), en tant qu'approximateur de fonction pour faire de la régression des rendements sur les variables climatiques. Comme nous le verrons, **ce type de méthode demande un nombre important d'exemples** du processus à étudier, ce qui n'est pas le cas ici. Afin d'effectuer la modélisation statistique, on recherche des dépendances implicites sur des données empiriques de la base d'exemples échantillonnés. Ce type de modélisation ne requiert presque aucune information autre que celle induite par la population des données. Toutefois, nous verrons dans la suite qu'il est néanmoins possible, et même **nécessaire dans le cadre de notre étude, d'ajouter toute l'information** *a priori* **disponible sur le problème à résoudre.** Ce dernier point est particulièrement important ici compte-tenu du manque de données.

Le PMC est de loin le réseau de neurones le plus utilisé de nos jours. Ce type de réseau de neurones est plutôt utilisé pour faire de l'interpolation car ses facultés à extrapoler ne sont pas évidentes. Nous détaillons ce point dans la section 11.2.1. Il possède plusieurs avantages : une fois éduqué (i.e. paramétré), le PMC est capable de généraliser son comportement à des données qui ne lui sont pas connues. Il est très rapide en mode opérationnel (une fois l'apprentissage effectué) et il est de plus robuste quant à la présence d'un bruit sur les données d'entrée lorsque l'apprentissage a été régularisé car il existe une redondance de l'information dans les processeurs parallèles que sont les neurones. Il possède également l'avantage en régression sur d'autres méthodes de travailler globalement sur l'espace des données d'entrée et non pas localement, ce qui est important pour le traitement de données de dimension élevée. De plus, la souplesse de cette technique permet l'introduction de connaissance *a priori* sur le problème sous des formes diverses, ce qui est, comme on l'a dit, capital dans un certain nombre d'applications.

Lors de l'utilisation des réseaux de neurones, les étapes sont les suivantes : sélection de données pour la construction de la base d'apprentissage, « pre-processing » des données (de haut et de bas niveau[1]), détermination de l'architecture adoptée (i.e. sélection du modèle), identification des paramètres du RN par apprentissage (i.e. estimation des paramètres) et validation du modèle obtenu.

Définition du PMC

Le perceptron multi-couches [Hertz et al., 1991] est un modèle paramétrique non linéaire qui calcule une sortie (multi-variée) lorsqu'on lui présente une entrée (multi-variée).

Comme tous les réseaux de neurones, il est constitué de plusieurs « processeurs » interconnectés qui effectuent des calculs locaux : les neurones. Ces derniers sont organisés selon une certaine « architecture » : une structure de graphe avec des arêtes orientées. Ce graphe est appelé architecture neuronale. Dans le PMC, les neurones sont organisés en couches successives de neurones indépendants. Traditionnellement, on ne compte pas la couche d'entrée dans le nombre de couches du réseau. Toutes les couches, sauf les couches d'entrée et de sortie, sont appelées couches « cachées ». Sur une couche, chaque neurone est relié à tous les neurones de la couche précédente (figure 11.1) par une « liaison synaptique » à laquelle est associé un poids synaptique w_{ji}. À architecture fixée (nombre de couches cachées, nombre de neurones par couches et fonction d'activation choisis), toute l'information du réseau est incluse dans ces poids synaptiques.

Le neurone du perceptron (modèle de McCulloch et Pitts [1943]) effectue successivement deux opérations (figure 11.2) :

1. La distinction entre un traitement de haut et de bas niveau se fait par le niveau de spécificité de celui-ci quant à l'application traitée.

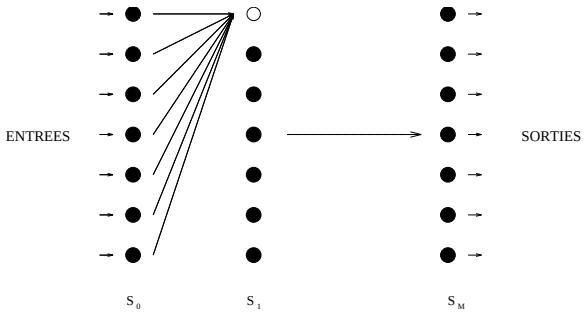


FIGURE 11.1 : Architecture du perceptron multi-couches

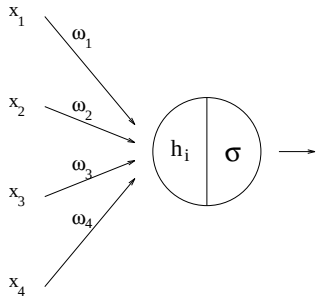


FIGURE 11.2 : Neurone i

– le neurone i calcule la somme pondérée « h_i » (appelée entrée totale) de ses entrées x_j :

$$h_i = \sum_{j=1}^{p} w_{ji} x_j,$$

avec p le nombre d'entrées du neurone i,
– et applique à cette entrée totale h_i, une fonction d'activation (ou de seuil, ou de transfert) sigmoïde σ (définition 11.2) plus un biais b_i :

$$y_i = \sigma(h_i) + b_i.$$

Remarque 11.1: *Le biais b_i de sortie du neurone i est un paramètre du RN qui doit être déterminé. Dans la pratique, on ajoute toujours un neurone supplémentaire à la couche d'entrée du RN, il est appelé neurone de biais. Sa valeur est fixée à 1 et reste constante durant l'apprentissage. Tous les neurones du réseau auront comme entrée ce neurone de biais et la liaison synaptique w_{0i} entre le neurone de biais et un neurone i est égale au biais b_i du neurone i. Dans la suite du texte, on omettra ce neurone de biais par souci de concision dans les notations.*

Définition 11.2: *Une fonction sigmoïde est une fonction σ telle que $lim_{x\to-\infty}\sigma(x) = -1$ et $lim_{x\to+\infty}\sigma(x) = +1$.*

On peut par exemple prendre comme fonction sigmoïde la fonction seuil, la fonction arctangente ou la fonction logistique (exemple 11.3). La fonction identité est parfois, elle aussi,

utilisée, par exemple pour faire de la discrimination, bien que ne satisfaisant pas la définition 11.2.

Cette fonction sigmoïde doit être impérativement continue si on veut faire de l'apprentissage en minimisant la fonction de coût, il faut, en effet, que cette fonction de coût soit différentiable par rapport aux paramètres du réseau. On pourra alors utiliser une technique d'optimisation pour la minimisation, par exemple une descente de gradient. Nous nous placerons toujours, par la suite, dans le cas d'une fonction sigmoïde continue en utilisant systématiquement la fonction logistique suivante.

Exemple 11.3: *La fonction logistique (figure 11.3) est une fonction sigmoïde continue, définie par :*

$$\sigma(x) = \frac{1 - e^{-2\beta \cdot x}}{1 + e^{-2\beta \cdot x}},$$

où β est la dérivée de la fonction à l'origine.

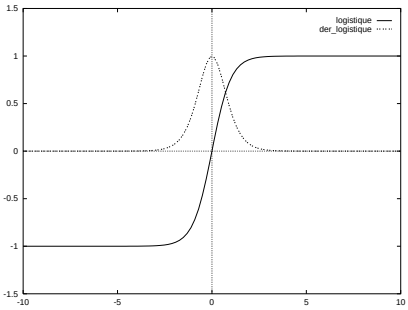


FIGURE 11.3 : Fonction logistique (trait plein) et sa dérivée (pointillés)

Remarque 11.4: *Les neurones auront donc une sortie entre -1 et +1 mais on peut aussi choisir des sorties entre 0 et 1 en introduisant un simple biais.*

Remarque 11.5: *Il existe des modèles plus complexes de neurones, faisant, par exemple, intervenir des réponses impulsionnelles (dans le temps). Mais ce type de considérations dépasse le cadre de ce travail.*

L'intégration globale des neurones conduit à l'architecture en réseau du PMC. Par exemple, pour un perceptron à une couche cachée, la $k^{\text{ème}}$ sortie y_k du RN sera définie par :

$$y_k = \sigma(\sum_{j \in S_1} w_{jk} \cdot \sigma(\sum_{i \in S_0} w_{ij} \cdot x_i)).$$

Les PMC sont souvent appelés réseaux à propagation avant (« feedforward network ») car la mise-à-jour des neurones se fait couche par couche, de la couche d'entrée à la couche de sortie. La mise-à-jour des neurones dans une couche est synchrone (i.e. en même temps).

Les inter-relations entre neurones spécifient des compétitions et des coopérations entre neurones : c'est la base du terme de *connexionisme*. Des architectures plus complexes (boucles, récurrences, connexions latérales, éliminations de certains poids, ...), spécifiant des inter-relations plus complexes, peuvent aussi être utilisées dans certains cas.

Le fait de rechercher une architecture de façon automatique (« weight pruning », « weight growing »,...) rend le modèle du PMC non paramétrique.

Apprentissage supervisé des PMC

Le modèle général de l'apprentissage supervisé à partir d'exemples possède 3 composantes :
- un générateur d'exemples x^e ; $e = 1, \ldots, P$ mutuellement indépendants dans l'espace des entrées de la fonction à estimer, suivant une distribution de probabilité $F(x)$ inconnue. Ici il s'agit de nos anomalies de variables climatiques.
- un superviseur qui associe à chacune des entrées x^e une sortie y^e suivant une distribution de probabilité conditionnelle $F(y/x)$ (par exemple $F(y/x) = y - g(x) + \varepsilon$, avec g une fonction déterministe inconnue et ε, un bruit aléatoire). Il s'agit des observations de rendements en sucre.
- un modèle paramétrique capable d'implémenter un ensemble de fonctions $\{g_W \ ; \ W \in \Theta\}$ qui puissent se rapprocher de la fonction désirée g.

On dispose, à architecture fixée, d'un estimateur non linéaire g_W paramétrique. L'apprentissage consistera à fixer les paramètres W de l'estimateur g_W pour inférer « au mieux » (nous allons voir dans quel sens) la fonction désirée g.

Inférence statistique des paramètres du PMC

Dans l'apprentissage supervisé, on se place dans le cas où une probabilité jointe inconnue $\mathcal{P}(y/x)$ relie des couples $z = (x, y)$. Et on dispose d'un ensemble d'apprentissage $\mathcal{B} = \{(x^1, y^1), (x^2, y^2), \ldots, (x^P, y^P)\}$ issu de cette loi de probabilité.

On considèrera ici que :

$$y = g(x) + \varepsilon \tag{11.1}$$

où ε, variable inobservable, est un bruit aléatoire et g une fonction déterministe inconnue.

L'apprentissage consistera à estimer la régression, c'est-à-dire à associer à une certaine observation x une estimation $\hat{y} = g_W(x)$ de $y = g(x)$. La qualité de la base de données $\mathcal{B}$ est essentielle pour la qualité des résultats finaux. Il existe essentiellement deux approches pour faire de l'apprentissage supervisé :
- l'inférence Bayésienne ;
- la maximisation d'un critère de qualité (on utilisera indifféremment, dans la suite du texte, fonction de coût ou critère de qualité).

La première cherche à déterminer la fonction de distribution (ou densité) de probabilité (f.d.p.) *a posteriori* $\mathcal{P}(W/\mathcal{B})$ et à estimer ensuite $W^{opt} = \int W \cdot \mathcal{P}(W/\mathcal{B}) dW$.

La deuxième approche spécifie une fonction de coût $C(\mathcal{B}, W)$ que l'on cherche à minimiser : $W^{opt} = argmin_W C(\mathcal{B}, W)$.

Ces deux techniques ne sont pas équivalentes, mais nous allons voir que pour un nombre important P de données dans la base d'exemple, le critère des moindres carrés et l'inférence Bayésienne sont équivalents dans l'hypothèse Gaussienne.

Nous implémentons dans le cadre de notre étude le PMC grâce à **Netlab** qui est un module de Matlab. L'inférence se fait par minimisation d'une fonction de coût. Ainsi nous n'utilisons pas le modèle bayésien dans le cadre des seuls réseaux de neurones. En revanche, nous l'utilisons dans les modèles hiérarchiques, en linéaire comme en non linéaire.

Minimisation d'une fonction de coût

On définit une fonction de coût $C(\mathcal{B}, W)$ qui mesure la dissimilarité des réponses de l'estimateur g_W avec les vraies réponses de la base d'apprentissage $\mathcal{B}$. On prend souvent l'erreur quadratique moyenne sur la base d'apprentissage :

$$C(W) = \frac{1}{2P} \sum_{e=1}^{P} \| \, y^e - g_W(x^e) \, \|^2,$$

où $\|\cdot\|$ est une norme issue d'une certaine distance. On peut prendre aussi le maximum de vraisemblance :

$$\mathcal{P}(\mathcal{B}/W) = \prod_{e=1}^{P} \mathcal{P}(z^e/W) \stackrel{def}{=} \mathcal{L}(W),$$

On définit alors la nouvelle fonction de coût (analogue d'une énergie) par :

$$E(W) = -ln\left(\mathcal{L}(W)\right) = -\sum_{e=1}^{P} ln(\mathcal{P}(x^e/W)).$$

Proposition 11.6: *Moindres Carrés et maximum de vraisemblance (i.e. estimateur Bayésien) sont deux critères équivalents dans l'hypothèse Gaussienne.*

Démonstration : En effet, si ε dans l'équation (11.1) suit une loi Gaussienne $\mathcal{N}(0,\sigma)$, la log-vraisemblance

$$ln(\mathcal{L}(W))$$

est égale à :

$$\frac{-1}{2\sigma^2}\sum_{e=1}^{P}\left(y^e - g_W(x^e)\right)^2 - Pln(\sqrt{2\pi}\sigma).$$

Puisque $ln(\sqrt{2\pi}\sigma)$ ne dépend pas de W, maximiser $\mathcal{P}(\mathcal{B}/W)$ revient à minimiser :

$$\sum_{e=1}^{P}\left(y^e - g_W(x^e)\right)^2,$$

le critère des MC. ■

Vapnik [Vapnik, 1997] introduit la fonction de risque :

$$R(W) = \int L(y, g_W(x))dF(x,y),$$

où $F(x,y)$ est la fonction de distribution de probabilité de la variable aléatoire (x,y) et L une mesure d'erreur sur l'espace des y. Mais $F(x,y)$ n'est pas connue, seul un échantillon $\mathcal{B} = \{(x,y)^e \ ; \ \ e = 1,\ldots,P\}$ est disponible, on minimise alors le risque empirique :

$$R_{emp}(W) = \frac{1}{P}\sum_{e=1}^{P} Q(x^e, y^e, W)$$

C'est souvent l'approche consistant à minimiser une fonction de coût $C(W)$ (souvent les Moindres Carrés) qui est prise pour l'utilisation des RN. Mais, comme on va le voir dans la partie 11.1.2, l'intérêt de l'inférence Bayésienne est grand. En effet, toute information *a priori* sur les f.d.p. des variables aléatoires peut être utilisée.

Modèle hiérarchique Bayésien

Comme on l'a vu, les méthodes d'apprentissage conventionnelles (ou fréquentistes) pour les perceptrons multicouches peuvent être interpétées comme des variantes de l'estimation par maximum de vraisemblance (équivalent dans l'hypothèse gaussienne à la méthode des moindres carrés). L'idée est alors de trouver un jeu de poids unique permettant de se rapprocher au mieux des observations correspondant à la base d'apprentissage.

L'Ecole bayésienne des statistiques est basée sur une conception très différente de l'apprentissage à partir de données. Celle-ci prône en effet la notion de probabilité pour représenter l'incertitude au sujet de la relation à trouver. En effet, considérons un modèle statistique M muni d'un paramétrage Θ $\{g_W \ ; \ \ W \in \Theta\}$. Dans le cadre bayésien, on fait l'hypothèse que le paramètre W est lui-même une variable aléatoire. Ainsi, on doit compléter la description du modèle statistique par la donnée d'une loi a priori sur Θ. Ainsi, dans le cas des réseaux de neurone, avant-même d'avoir pris connaissance du jeu de données, notre opinion a priori au sujet de la vraie relation (pouvant provenir d'avis d'experts par exemple), peut s'exprimer en termes de probabilité de distribution des valeurs des poids du réseau définissant cette relation.

Cette approche utilise le point de vue des subjectivistes dans la statistique plutôt que celui des fréquentistes (ou objectivistes). Les fréquentistes manipulent les probabilités dans le cas d'expériences aléatoires qui peuvent être reproductibles un grand nombre de fois ; c'est le cadre de la loi des grands nombres. Les subjectivistes, dans le cadre de l'inférence Bayésienne, associent au concept de probabilité une notion de plausibilité (le coefficient de vraisemblance ou de plausibilité étant situé entre 0 et 1) : le calcul de probabilité pourra ainsi être utilisé sur des évènements non reproduisibles. On en arrive presque ici à la théorie de la logique floue où une proposition possède elle aussi un degré de vraisemblance. Le théorème de Bayes (voir ci-dessous) donne l'expression de la probabilité *a posteriori* $\mathcal{P}(y/x)$ de y sachant x que l'on peut interpréter comme la « vraisemblance » de l'événement y connaissant la réalisation de x.

Théorème 11.7: *(de Bayes)*

$$\mathcal{P}(y/x) = \frac{\mathcal{P}(x,y)}{\mathcal{P}(x)}$$

On se sert de cette notion pour définir la régression en prédiction (trouver un y pour une observation x).

Définition 11.8: *La régression de y sur x est $E[y/x] = \int y \cdot \mathcal{P}(y/x)dx$, ce qui est une fonction déterministe.*

Proposition 11.9: *La régression est la meilleure prédiction de y sachant x au sens des moindre carrés.*

Démonstration :

$$\forall \ g_W, \forall \ x \quad :$$

$$E\left[(y - g_W(x))^2/x\right] = E\left[(y - E[y/x] + E[y/x] - g_W(x))^2/x\right]$$

$$\Rightarrow \quad E\left[(y - g_W(x))^2/x\right] = E\left[(y - E[y/x])^2/x\right] + (E[y/x] - g_W(x))^2$$

$$\Rightarrow \quad E\left[(y - g_W(x))^2/x\right] \geq E\left[(y - E[y/x])^2/x\right]$$

■

Pour faire de la prédiction complète (détermination d'une estimation g_W de la fonction g une fois pour toutes) on tire parti de tous les P exemples de la base d'apprentissage $\mathcal{B}$ en utilisant :

$$\mathcal{P}(W/\mathcal{B}) = \mathcal{P}(\mathcal{B}/W) \cdot \frac{\mathcal{P}(W)}{\mathcal{P}(\mathcal{B})} \tag{11.2}$$

avec :

- $\mathcal{P}(\mathcal{B}/W) = \prod_{e=1}^{P} \mathcal{P}(z^e/W)$, la vraisemblance ;
- $\mathcal{P}(\mathcal{B}) = \int \mathcal{P}(W') \prod_{e=1}^{P} \mathcal{P}(z^e/W')dW'$, un coefficient de normalisation.

On détermine W^{opt} :

$$W^{\text{opt}} = \int W \cdot \mathcal{P}(W/\mathcal{B})dW.$$

L'intérêt essentiel de l'inférence Bayésienne est de pouvoir utiliser l'information *a priori* éventuellement disponible sur les f.d.p. des variables aléatoires du problème.

Dans le cadre de notre étude, nous utilisons cette notion d'inférence bayésienne dans le cas des **modèles hiérarchiques**. Nous utilisons les modules implémentés en R que sont **lme (linear mixed effect) et nlme (non linear mixed effect)**.

11.2 Problèmes

11.2.1 Généralisation

Dans un modèle de régression (linéaire ou non linéaire), on vise durant l'apprentissage à obtenir l'écart le plus faible entre les sorties du modèle de régression et les observations réelles. Cette différence faible sur la base d'apprentissage peut être intéressante en tant que telle, notamment pour des questions de mémorisation. Néanmoins, dans la plupart des cas, et notamment ici, l'intérêt principal d'un tel modèle est sa capacité de généralisation, c'est-à-dire de fournir des résultats satisfaisants sur des données indépendantes de celles de la base d'apprentissage. Dans le cadre de notre étude, l'indice climatique doit permettre de rendre compte du rendement avec précision pour les années futures, indépendantes de la base d'apprentissage. Néanmmoins, cette capacité de généralisation n'est pas toujours possible, même dans le cas non-linéaire, et ce contrairement aux affirmations de certains auteurs. Par exemple, Caudill et Bluter, en 1990, prétendent : "Un réseau de neurone est capable de généraliser "sans jutifier cette affirmation et en négligeant l'ensemble des aspects complexes intervenant dans la capacité à généraliser.

Ainsi, il est possible de distinguer environ trois conditions nécessaires à celle-ci. Ces dernières ne sont néanmoins pas suffisantes.

1. L'existence d'une relation réelle entre les entrées et les sorties. En effet, il doit exister une relation mathématique entre les entrées et les sorties témoignant d'un phénomène réel. Une erreur trop souvent commise est de tenter d'obtenir un modèle prédictif à l'aide d'entrées choisies de manière arbitraire sans étude préalable. Dans le cas des réseaux de neurones, même si ceux-ci sont en mesure de capter des phénomènes d'interaction très complexes, ils ne peuvent en aucune façon trouver une relation qui n'existe pas.
2. Une certaine régularité de la fonction (relation) que l'on tente d'établir. En d'autres termes, à de faibles variations des entrées doivent correspondre de faibles variations des sorties. Ceci implique la continuité de la fonction à trouver. Néanmoins, certains réseaux peuvent apprendre des discontinuités sous réserve que la fonction soit continue par morceaux.
3. Une base d'apprentissage de taille suffisamment importante et représentative de l'éventail des possibilités (c'est-à-dire la population globale). Ceci est lié à l'existence de deux types de généralisation : l'interpolation et l'extrapolation. L'interpolation s'applique à des données relativement proches de celles de la base d'apprentissage. Le reste est appelé extrapolation. Sous réserve que les deux conditions précédentes soient réalisées, l'interpolation aboutit en

général à un résultat assez fiable. L'extrapolation, en revanche, conduit la plupart du temps à une prévision nettement erronée. Le réseau a en effet, durant la phase d'apprentissage, tenté de mettre en correspondance observations et données reconstruites mais sur des données très éloignées. Il n'a donc pas appris les phénomènes particuliers correspondant aux années pour lesquelles on lui demande de prévoir. Ainsi, il est très important de posséder une base d'apprentissage comportant suffisamment de données pour pouvoir éviter d'avoir à extrapoler. Les réseaux de neurones sont extrêmement efficaces car ils peuvent capter énormément de phénomènes du fait de leur non-linéarité. Par ailleurs, on peut choisir un nombre de degrés de liberté aussi grand qu'on le souhaite. Mais ceci signifie alors que le réseau doit être contraint par un échantillon d'apprentissage de très grande taille.

11.2.2 Dilemme biais/variance

Comme nous l'avons vu précédemment, l'approche la plus courante pour l'inférence des réseaux de neurone est la minimisation de l'erreur quadratique moyenne.

L'erreur quadratique moyenne, sur l'ensemble des exemples de $\mathcal{B}$, de g_W comme estimateur de la régression $E[y/x]$ est :

$$E_{\mathcal{B}}\left[(g_W(x,\mathcal{B}) - E[y/x])^2\right]$$

L'influence de $\mathcal{B}$ est primordiale, c'est pourquoi nous allons décomposer cette erreur quadratique en deux termes : le biais et la variance.

$$\forall x \ : \quad E_{\mathcal{B}}\left[(g_W(x,\mathcal{B}) - E[y/x])^2\right]$$

$$= E_{\mathcal{B}}\left[\ (g_W(x,\mathcal{B}) - E_{\mathcal{B}}[g_W(x,\mathcal{B})] + E_{\mathcal{B}}[g_W(x,\mathcal{B})] - E[y/x])^2 \ \right]$$

$$= E_{\mathcal{B}}\left[(g_W(x,\mathcal{B}) - E_{\mathcal{B}}[g_W(x,\mathcal{B})])^2\right] + E_{\mathcal{B}}\left[(E_{\mathcal{B}}[g_W(x,\mathcal{B})] - E[y/x])^2\right] +$$

$$2 \cdot \underbrace{E_{\mathcal{B}}\left[g_W(x,\mathcal{B}) - E_{\mathcal{B}}[g_W(x,\mathcal{B})]\right]}_{=E_{\mathcal{B}}[g_W(x,\mathcal{B})]-E_{\mathcal{B}}[g_W(x,\mathcal{B})]} \cdot (E_{\mathcal{B}}[g_W(x,\mathcal{B})] - E[y/x])$$

$$= \underbrace{(E_{\mathcal{B}}[g_W(x,\mathcal{B})] - E[y/x])^2}_{biais} + \underbrace{E_{\mathcal{B}}\left[(g_W(x,\mathcal{B}) - E_{\mathcal{B}}[g_W(x,\mathcal{B})])^2\right]}_{variance}$$

Ainsi, le fait que l'estimateur $g_W(x,\mathcal{B})$ ne coïncide pas forcément avec E[y/x] s'explique par le fait que l'estimateur ne soit pas nécessairement égal en moyenne à la vraie valeur et par la variabilité propre de l'estimateur.

Le biais caractérise la sous-paramétrisation (i.e. pas assez de paramètres dans l'interpolateur) : la classe de fonctions que peut simuler le modèle paramétrique est insuffisante pour la tâche que l'on a à effectuer. Ce biais est mesuré sur la base d'apprentissage et c'est lui que l'on minimise pour estimer g_W.

Définition 11.10: *Si la moyenne de $g_W(x,\mathcal{B}) \neq E[y/x]$ alors l'estimateur est dit biaisé.*

La variance caractérise la sur-paramétrisation (i.e. trop de paramètres dans l'interpolateur). La complexité du modèle est alors trop importante et cela pour deux raisons possibles :

– la fonction g à approcher est moins complexe que g_W. Le réseau de neurones (ou le modèle de régression linéaire RL) comporte alors trop de paramètres quelle que soit la taille de la base d'apprentissage. Dans le cas du RN comme de la RL, il se peut alors que le nombre de variables explicatives (les entrées du modèle) soit trop important.
– la base d'apprentissage n'est pas assez vaste compte-tenu de la complexité du problème : un nombre trop important de paramètres libres W nécessite une base d'apprentissage de grande taille, sinon le RN apprendra les données de la base d'apprentissage « par cœur » et aura un mauvais comportement pour des données hors de la base d'apprentissage. On parle alors de sur-apprentissage.

Remarquons que dans le cas du réseau de neurones ou de la régression linéaire, l'augmentation du nombre de variables explicatives (et donc d'entrées) conduit à une augmentation du nombre de paramètres. Compte-tenu du faible nombre de données dans notre étude, il faudra donc veiller à ne pas prendre un nombre trop important de prédicteurs (4 semble être optimal).

Ces problèmes de manque de données et de sur-apprentissage sont traités dans les sections suivantes.

Ainsi, pour minimiser le biais, on augmente le nombre de paramètres libres, mais alors la variance de l'erreur augmente ! Cette compétition entre la variance et le biais est appelée « principe d'incertitude ». La résolution du dilemme « biais/variance » [Geman et al., 1992] consiste à trouver la bonne complexité de l'estimateur neuronal. Il faudra faire un compromis entre les erreurs sur la base d'apprentissage et la complexité du RN. Ce dilemme « biais/variance » est donc un problème délicat et c'est le taux de généralisation qui nous permettra de choisir le meilleur compromis possible.

11.2.3 Manque de données

Comme on vient de le voir, posséder un nombre de données suffisant est une condition nécessaire pour une bonne capacité de généralisation. Ceci a été théorisé par Vapnik et Chervonenkis. Un résumé de leurs travaux est présenté en annexe C.1 page 125.

11.2.4 Sur-apprentissage

Comme nous l'avons dit, l'un des nombreux intérêts des perceptrons multi-couches pour l'approximation de fonctions est leur capacité à généraliser, c'est-à-dire leur capacité à donner une bonne réponse à une donnée qui n'appartient pas à la base d'apprentissage qui a servi à l'éduquer. Cette capacité à traiter des données non connues de la base d'apprentissage, la faculté de généralisation, permet de faire de l'interpolation ou de l'extrapolation (on pourra voir [Friedman, 1994] pour un article de vulgarisation sur ce sujet).

Néanmoins, en vue d'obtenir une bonne capacité de généralisation, il convient de se méfier ce problème de sur-apprentissage.

Dans un tel cas, l'interpolateur passe exactement par tous les points de la base d'apprentissage mais entre ces points, les erreurs commises peuvent être très importantes. La dérivée de l'interpolateur aux points de la base d'apprentissage $\mathcal{B}$ n'a rien à voir avec la dérivée de la fonction désirée. En effet, le réseau a appris la base d'apprentissage par coeur et donc en particulier ses détails (ou bruit) au détriment des grandes tendances caractérisant le comportement global. Cette information sur les dérivées de l'interpolateur peut être utilisée pour régulariser le problème.

Pour estimer la qualité d'un réseau de neurones (comme de tout estimateur fonctionnel), on fait généralement appel à la racine carré de l'erreur quadratique moyenne (ou « Root Mean Square ») :

$$\mathrm{RMS}_i = \sqrt{\frac{1}{P}\sum_{e=1}^{P}({y_i}^e - g_W(x^e)_i)^2}$$

Comme précédemment, on sépare la RMS en deux termes :

$${\mathrm{RMS}_i}^2 = {\mathrm{biais}_i}^2 + {\text{écart-type}_i}^2$$

avec :

- $\mathrm{biais}_i = \frac{1}{P}\sum_{e=1}^{P}({y_i}^e - g_W(x^e)_i)$
- $\text{écart-type}_i = \sqrt{\frac{1}{P}\sum_{e=1}^{P}({y_i}^e - g_W(x^e)_i - \mathrm{biais}_i)^2}$

L'indice i désigne la i-ème composante de la sortie.

Pour analyser le comportement du RN, on observe donc indépendamment ces deux quantités. Mais pour pallier le problème de sur-apprentissage, le bon indicateur est le taux de généralisation, c'est-à-dire la RMS calculée sur une base autre que celle utilisée durant l'apprentissage (la base de généralisation).

La figure 11.4 illustre le problème de sur-apprentissage en calculant la RMS sur la base d'apprentissage et la base de généralisation. Au-delà d'un certain nombre d'itérations dans le processus de la minimisation de la RMS sur la base d'apprentissage, l'erreur sur la base de test augmente. A partir de ce moment, on oblige en fait le réseau à apprendre des détails et ce au détriment des caractéristiques globales de la fonction à estimer.

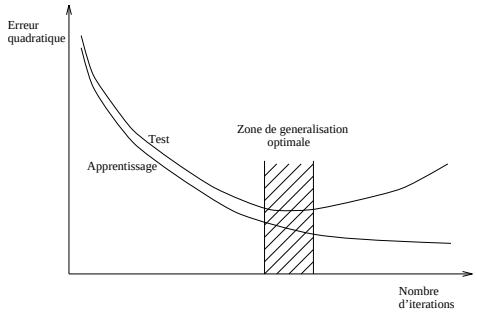


FIGURE 11.4 : Courbes d'apprentissage et de généralisation

Nous étudions en détail les mesures de qualité de l'indice dans la partie 11.3

11.3 Mesure de la qualité de l'indice

11.3.1 Bases d'apprentissage, de test et de validation

Cette séparation des données en trois bases distinctes intervient principalement en théorie des réseaux de neurones. Mais comme on le verra dans la suite de l'exposé, on peut également avoir recours à celle-ci dans le cas linéaire.

Nous cherchons évidemment à obtenir le réseau de neurones (ou plus généralement l'indice dans le cadre de notre étude) ayant les meilleures performances sur un jeu de données inconnu.

Il convient donc d'optimiser le taux de généralisation. Celui-ci n'est malheureusement pas accessible directement et il faut donc l'estimer. Les différents réseaux sont entraînés sur la base d'apprentissage par minimisation d'une certaine fonction d'erreur, comme on l'a vu précédemment. Néanmoins, le taux d'erreur sur la base d'apprentissage (que l'on appelle taux d'apprentissage) est un estimateur biaisé du taux de généralisation car il dépend fortement de la base d'apprentissage. C'est pourquoi le taux d'apprentissage n'intervient pas dans notre choix pour le meilleur réseau (ou opur l'architecture de l'indice dans le cas général)

Les performances des différents réseaux (différant par leur architecture : nombre de couches cachées, nombre d'époques de minimisation de la fonction de coûts), sont alors comparées sur une base de validation, ou de généralisation. Pour chacun d'eux, on calcule une estimation du taux de généralisation réel, estimation qui cette fois sera non biaisée. En effet, on teste l'aptitude du réseau à bien traiter des données hors de la base d'apprentissage. Le réseau possédant l'erreur la plus faible sur celle-ci est alors choisi. Néanmoins, cette procédure peut conduire à un surapprentissage de la base de validation. En effet, il se peut que le réseau choisi ait en effet par hasard appris par coeur les données de cette base. Nous comprenons alors la nécessité d'introduire une troisième base, la base de test, sur laquelle la performance du réseau choisi doit être testée. Ceci permet de confirmer ou d'infirmer le choix du réseau (ou de manière plus générale de l'indice).

Nous appliquons une méthodologie similaire lorsque nous cherchons la combinaison optimale des variables climatologiques explicatives. Les différentes combinaisons possibles seront testées sur la base de validation et celle minimisant l'erreur (ou maximisant la corrélation) sera choisie. La performance de la combinaison retenue sera enfin évaluée sur une base de test.

Notons que d'autres estimateurs du taux de généralisation existent (« leave-one-out », « bootstrapping », *etc*) [Bishop, 1996].

Ceux-ci sont étudiés dans les sections suivantes.

11.3.2 Bootstrap

Les techniques de bootstrap sont très utiles lorsque l'on dispose de petites bases de données. En effet, il est alors délicat d'obtenir des grandeurs statistiques de manière fiable. L'idée est alors de constituer un grand nombre de nouvelles bases de données en considérant des sous-ensembles quelconques de notre base de données initiale.

Typiquement, dans notre étude, supposons que notre indice soit construit. Nous aimerions évaluer la dispersion de ses performances. Ainsi, nous choisissons plusieurs plages d'années tests. Nous réalisons à chaque fois l'apprentissage des paramètres et nous calculons le résultat en généralisation. Ceci nous fournit alors l'écart-type et un intervalle de confiance au sujet de la performance de l'indice.

11.3.3 Leave-one-out

Le leave-one-out, tout comme le bootstrap, permet de tester les capacités de généralisation de notre modèle. Il s'agit d'ailleurs d'un cas particulier de bootstrap. Remarquons que nous utilisons dans cette étude le bootstrap sur les bases de test et le leave-one-out sur les bases de validation.

Dans le cadre de notre étude, le leave-one-out nous permet principalement :

- d'augmenter artificiellement la base de validation
- de robustifier le modèle

En effet, comme on vient de le voir, on a besoin de trois bases. On possède $N_{total} = 240$ données de rendements. On choisit une base de test comportant $N_{test} = 10$ années de test. Il nous reste donc 230 données pour l'apprentissage et la validation. Afin de bénéficier de l'apprentissage

sur le plus grand nombre de données possible (afin de limiter le sur-apprentissage), nous voudrions limiter le nombre d'années de la base de validation. L'idée est alors de n'en prendre qu'une et de répéter ce procéder un grand nombre de fois (300 par exemple). Ainsi, nous créons une série de rendements estimés (les leave-one-out successifs) que l'on peut comparer avec la série réelle. Ainsi, tout se passe comme si nous possédions une base de validation de grande taille.

De surcroît, ce procédé engendre des jeux de coefficients différents pour chaque nouveau leave-one-out différent (étant donné qu'un élément de la base d'apprentissage a changé). Ceux-ci engendrent ce que l'on appelle un modèle d'ensemble. Ceci sera étudié en détail dans la partie 11.4.4.

Après avoir étudié les obstacles à la capacité de généralisation

11.4 Régularisation de l'indice

La régularisation d'un problème de prédiction mal posé est l'introduction d'une information *a priori* supplémentaire dans la formulation du problème, afin de le rendre bien posé. L'information *a priori* additionnelle peut concerner :

- La solution du problème, c'est-à-dire la relation entre les entrées et les sorties ;
- Les données d'entrée ;

11.4.1 Input pertubation

Cette technique est très fréquente dans l'utilisation des réseaux de neurones mais peut également être appliquée dans le cas de la régression linéaire.

Comme nous l'avons vu, le critère de qualité le plus répandu en statistique est indéniablement l'erreur quadratique moyenne :

$$C(W) = \frac{1}{2} \sum_{k=1}^{M} \int \int (y_k - g_k(x;W))^2 \cdot P(y_k/x) \cdot P(x) \cdot dy_k \cdot dx \tag{11.3}$$

où y_k est la $k^{\text{ème}}$ composante de la sortie désirée, g_k la $k^{\text{ème}}$ composante de sortie du RN, $P(.)$ la densité de probabilité des données d'entrée x et M le nombre de sorties du RN g_W. Dans la pratique, $C(W)$ est approché par :

$$\overline{C}(W) = \frac{1}{2P} \sum_{e=1}^{P} (g_k(x;W) - y_k)^2 \tag{11.4}$$

Pour rendre l'estimation moins sensible au bruit sur les entrées, la technique de régularisation par « Input Perturbation » peut être utilisée. C'est une technique heuristique pour controler la complexité effective de l'application neuronale (ou linéaire) g_W. Fondamentalement, elle limite le nombre de paramètres libres dans l'approximateur g_W pour rapprocher les complexités de la fonction désirée g et de son estimation neuronale.

Dans la pratique, l'« Input Perturbation » consiste a ajouter un vecteur aléatoire dans les entrées durant la phase d'apprentissage. Ce vecteur aléatoire peut être le bruit de mesure envisagé sur les entrées, mais il peut être aussi complètement artificiel dans un but de stabilisation du modèle, comme nous allons le voir.

En effet, il a été démontré [Bishop, 1996] que sous certaines hypothèses de bruit faible, l'apprentissage avec bruit est étroitement lié aux techniques de régularisation classiques par

lissage. En effet, la fonction de coût d'erreur quadratique moyenne prend, lorsque l'on introduit du bruit en entrée du RN, la forme suivante :

$$\tilde{C}(W) = \frac{1}{2}\sum_{k=1}^{M}\int\int\int \left(y_k(x+\eta;W) - t_k\right)^2 \cdot P(t_k/x) \cdot P(x) \cdot P(\eta) \cdot dt_k \cdot dx \cdot d\eta$$

Si le bruit η est suffisamment faible, la fonction neuronale $y_k(x+\eta;W)$ peut être développée au premier ordre. On obtient alors :

$$\tilde{C}(W) \simeq C(W) + \nu \cdot \Omega(W)$$

où ν est la variance du bruit et où $\Omega(W) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\int \left(\frac{\partial y_k}{\partial x_i}\right)^2 \cdot P(x) \cdot dx$.

On reconnaît dans le terme $\Omega(\cdot)$ le stabilisateur de Tikhonov, pénalisant les solutions g_W avec de grands gradients. Ainsi, la minimisation du nouveau critère $\tilde{C}(W)$ contraint la solution g_W à être lisse.

11.4.2 Early stoping

Cette technique est propre à l'apprentissage par réseau neuronal.

Durant l'apprentissage d'un RN, nous avons déjà montré qu'il était nécessaire de tester l'erreur de généralisation estimée sur une base de validation indépendante de la base d'apprentissage (section 11.3.1). Nous avons vu que durant la minimisation de la fonction de coût, à partir d'un certain nombre d'itérations, le phénomène de sur-apprentissage apparaissait (Fig. 11.4). Le « early stopping » consiste à arrêter l'apprentissage dès que l'erreur de généralisation estimée commence à croître, c'est-à-dire juste avant le phénomène de sur-apprentissage (section 11.2.4). Cela est dû au phénomène suivant : au début de l'apprentissage, le RN apprend les caractéristiques générales de la base d'exemples (on estime les grands vecteurs propres de la matrice de corrélation). Puis, l'apprentissage se focalisera sur les détails de la base d'exemples (petites valeurs propres). Ainsi le « early stopping » permet de « voir » les détails et d'éviter de ce fait le sur-apprentissage [Vallet, 1990].

Ainsi, l'architecture du RN peut être trop riche au regard de la complexité induite par la base d'apprentissage ; l'arrêt de l'apprentissage, en limitant l'adaptation des poids synaptiques aux exemples de la base d'exemples, limitera le nombre de paramètres libres dans le RN.

11.4.3 Weight decay

Cette technique est également propre aux réseaux de neurones.

Ici, nous ne prenons pas comme critère de qualité la seule somme des erreurs quadratiques (c'est-à-dire utilisation de la distance Euclidienne) pour l'estimation des paramètres W du RN g_W durant l'apprentissage. Nous lui adjoignons un stabilisateur $\Omega(\cdot)$:

$$C(W) = C(g_W, \mathcal{B}) + \gamma \cdot \Omega(g_W)$$

avec

$$\Omega(g_W) = \sum_{w \in \tilde{W}} {w_i}^2$$

où :

– $\tilde{W} \in W$ sont les poids du RN non connectés au neurone de biais ;

– γ est un méta-paramètre indiquant l'importance relative de l'un ou l'autre des deux types d'information utilisés durant l'apprentissage : information statistique des données $C(g_W, \mathcal{B})$ et information *a priori* du stabilisateur $\Omega(g_W)$.

Ce processus a pour but de réduire l'amplitude des poids synaptiques w_i du RN afin d'éviter de trop grandes variations dans le réseau. Il ne faut pas tenir compte, dans la somme $\sum_{w \in \tilde{W}} {w_i}^2$, des poids du neurone de biais ; ces paramètres ayant une fonction toute particulière dans le réseau.

L'utilisation de ce stabilisateur est équivalent à la technique classique en approximation de fonction de « ridge regression ». Ce nouveau critère pénalise les poids synaptiques w_{ji} du RN trop importants. Ainsi, l'activité $h_i = \sum_{j=1}^{p} w_{ji} x_i$ des neurones est, en moyenne, plus proche de zéro et donc se situe plus fréquemment dans leur partie linéaire. Le RN se rapproche ainsi des modèles linéaires. Empiriquement, on peut justifier l'utilisation de cette contrainte en se référant au sur-apprentissage. En effet, le sur-apprentissage apparaît lorsque de grandes variations sont possibles dans le RN. Ces dernières sont engendrées notamment par des poids synaptiques importants, d'où la pénalisation de tels poids [Bishop, 1996].

Remarquons que dans le cas de la régression linéaire, les coefficients prennent des valeurs très élevées dans le cas de variables explicatives co-linéaires. Ainsi, l'équivalent du weight decay est en quelque sorte le fait d'éviter les cas de co-linéarité.

11.4.4 Runs d'ensemble

Les méthodes précédentes, comme on l'a vu, visent à améliorer la capacité de généralisation du modèle, pour un apprentissage donné. L'idée des runs d'ensemble, elle, permet également d'améliorer la généralisation, mais en exploitant en fait l'ensemble des résultats d'une série d'apprentissages effectués sur des données différentes. Par exemple, considérons que nous possédons en tout N_{total} données de rendement. Laissons alors de côté N_{test} données qui constitueront notre base de test. Il faut alors séparer les $N_{total} - N_{test}$ données restantes en une base d'apprentissage et une base de validation. L'idée est alors de pratiquer un grand nombre de fois le leave-one-out. Notons N_{lo} ce nombre. Comme on l'a vu précédemment, ceci permet non seulement d'augmenter artificiellement la taille de la base de validation mais aussi d'obtenir plusieurs jeux de coefficients. Plusieurs manières de procéder sont alors envisageables. Afin de simplifier le formalisme, nous envisageons par la suite le cas d'un modèle linéaire à une variable.

Coefficients moyennés

On effectue N_{lo} leave-one-out. Ceci est donc équivalent à considérer Nlo bases d'apprentissage. On a donc :

$$Y_1 = a_1 X_1 + b_1 \tag{11.5}$$
$$Y_2 = a_2 X_2 + b_2 \tag{11.6}$$
$$Y_{Nlo} = a_{N_{lo}} X_{N_{lo}} + b_{N_{lo}} \tag{11.7}$$

On choisit alors comme jeu de coefficient :

$$a_{moy} = \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} a_i$$

et

$$b_{moy} = \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} b_i$$

La prévision sur la base de test est alors :

$$Y = a_{moy}X + b_{moy}$$

On comprend alors que l'on régularise l'indice puisque l'on considère la moyenne et donc le barycentre affecté de poids unité, des coefficients. Les coefficients finaux sont donc nécessairement compris entre le minimum et le maximum. Il s'agit en quelque-sorte d'une contrainte similaire à la régularisation de Tychonov. Supposons, par exemple, que le jeu de coefficient (a_1, b_1) apprenne une caractéristique d'une année particulière correspondant à du bruit (et donc en aucun cas à quelque-chose de généralisable). Les jeux de coefficient issus d'apprentissages ne prenant pas cette année particulière en compte ne présenteront pas ce défaut. Moyenné avec ceux-ci, l'impact du sur-apprentissage dû à (a_1, b_1) sera alors limité.

Runs d'ensemble

Le principe est de considérer les N_{lo} jeux de coefficients obtenus par leave-one-out et de les appliquer un-à-un sur la base de test. On obtient ainsi un nombre N_{lo} de prédictions sur cette base de test. La prévision est alors la prévision moyenne :

$$Y = \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} Y_i$$

Or on sait que :

$$Y_i = a_i X + b_i$$

D'où :

$$Y = \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} a_i X + b_i$$

$$= (\frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} a_i)X + (\frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} b_i)$$

$$= a_{moy}X + b_{moy}$$

On aboutit donc au fait que, dans le cadre d'un modèle linéaire, prévision moyenne et prévision obtenue en utilisant les coefficients moyennés sont identiques. L'approche par runs d'ensemble s'avère tout de même plus riche puisqu'elle fournit la dispersion des résultats.

Coefficients obtenus par inversion

Comme précédemment, nous effectuons un nombre N_{lo} de leave-one-out, ce qui nous fournit N_{lo} coefficients. L'idée est alors de calculer la prévision moyenne sur les $N_{total} - N_{test}$ années correspondant aux bases de test et de validation, puis de retrouver le coefficient correspondant par inversion de cette prévision moyenne. En théorie, comme on vient de le voir étant donné la linéarité du problème, il va de soi que nous devrions retrouver les coefficients moyennés. Néanmoins, ceci n'est absolument pas le cas du fait du mauvais conditionnement de la matrice sur laquelle nous effectuons la régression (car il s'agit d'anomalie). Celle-ci contient en effet des valeurs extrêmement faibles. Il s'agit d'un très bon exemple de problème inverse mal posé.

11.5 Indice avec prédicteurs obtenus grâce à l'expertise agronomique

Comme nous le verrons dans le chapitre 11.6 page 85, nous souhaitons mettre au point un indice construit de manière purement statistique (c'est-à-dire que le choix des prédicteurs s'effectue de manière statistique). Or cette procédure s'avère relativement coûteuse en temps de calcul car elle teste de nombreuses combinaisons de prédicteurs (sur des bases de validation et de test). Par ailleurs, comme nous avons pu le constater dans la section 11.1.2 page 56, il existe différents modèles d'apprentissage de la relation reliant les entrées aux sorties et ceux-ci doivent être accompagnés de **méthodes de régularisation nécessitant l'optimisation de paramètres**. Il est évidemment impossible de tester l'ensemble des modèles envisageables ainsi que des valeurs des paramètres de régularisation sur chaque combinaison de prédicteurs possible. Il convient donc de séparer le choix du modèle et des paramètres de régularisation d'une part et la recherche des meilleurs prédicteurs d'autre part.

L'idée que nous avons proposée est la constitution d'un **indice utilisant en partie l'expertise agronomique**. Nous ne considérons que 4 variables prédictives, ce qui fixe l'architecture de notre indice. Considérer davantage de prédicteurs revient à rajouter des entrées dans le modèle et donc des paramètres, ce qui diminue la performance en généralisation du fait du sur-apprentissage.

Nous utilisons alors ces prédicteurs pour déterminer la nature du modèle à utiliser (linéaire ou non-linéaire) et optimiser les paramètres de régularisation évoqués précédemment.

Ainsi, dans la section suivante, nous cherchons les meilleurs prédicteurs de manière statistique en utilisant le modèle ainsi que les paramètres de régularisation trouvés ici.

Cette démarche occasionne donc un gain de temps très important en vue de la constitution de notre indice purement statistique. Elle permet par ailleurs de voir in fine si les prédicteurs optimaux du point de vue agronomique le sont aussi du point de vue statistique.

11.5.1 Le choix des variables prédictives

Nous souhaitons trouver ici des prédicteurs pertinents sans avoir recours à un processus d'optimisation. Il s'agit donc d'isoler des variables prédictives ayant un sens physique, c'est-à-dire présentant un lien de cause à effet certain avec le rendement. Nous faisons donc appel à **l'expertise agronomique**. Le seul souci, comme nous l'avons annoncé précédemment, provient du fait que nous disposons des données de rendement globales. Ainsi, pour une zone donnée, le chiffre fourni par la Cosumar mélange des cycles de production et variétés différents. Or, comme nous allons le voir, un prédicteur donné peut avoir un impact positif dans le cas d'un certain cycle et un impact négatif dans le cas d'un autre.

Prenons un exemple simple afin de comprendre la difficulté induite par ce phénomène. Une anomalie positive de précipitations en décembre entraîne une anomalie de même signe au niveau du rendement dans le cas d'un cycle associé à un semis en septembre alors que la même anomalie engendre une anomalie négative dans le cas d'un cycle associé à un semis tardif. En effet, dans le premier cas, la plante a atteint la maturité suffisante pour profiter de l'accroissement de précipitations alors que dans le second, il s'agit d'une jeune plantule risquant de se retrouver noyée par des quantités d'eau trop importantes. Ainsi, si les anomalies engendrées sont du même ordre de grandeur en valeur absolue et que les deux cycles sont représentés de manière équitable, l'anomalie finale est proche de zéro et l'impact du prédicteur en question n'est donc pas visible.

Par ailleurs, du fait des diverses localisations géographiques et donc des conditions météorologiques différentes, il est possible que les anomalies aient des effets distincts selon les périmètres considérés. Une anomalie positive de température en hiver aura un effet plus important dans

des zones de montagne que dans des zones côtières où il fait déjà assez doux. Or, comme nous le verrons par la suite, nous **sommes contraints d'élaborer un seul modèle d'impact pour l'ensemble du Maroc compte-tenu du manque de données.** On comprend donc la difficulté de l'obtention de prédicteurs ayant un réel impact.

Maintenant que nous sommes conscients des problèmes et limites, tentons tout de même d'isoler quelques facteurs clés grâce aux études agronomiques. Le développement de la culture sucrière est tributaire de contraintes majeures. En ce qui concerne la qualité de la betterave, la forte chaleur de fin de cycle associée à des températures supérieures à 30° à partir de la deuxième décade favorise la respiration et entraîne une réduction de la matière sèche et de la teneur en sucre. En revanche, une anomalie positive de précipitations a un impact assez fort étant donné qu'il s'agit des dernières semaines de croissance. Cette dépendance physique se traduit, comme on peut le constater sur la figure 11.5, par des corrélations non nulles. En effet, on observe une anticorrélation entre l'anomalie de rendement et l'anomalie de température en juillet (de -0,19) alors que pour le même mois, la corrélation entre l'anomalie de rendement et l'anomalie de précipitations s'avère positive (0,21).

Par ailleurs, il est reconnu que les précipitations au mois d'avril ont un impact positif. En effet, dans certaines régions sèches, les précipitations peuvent s'avérer assez faibles au printemps et un apport supplémentaire (anomalie positive) se retrouvera de manière significative dans les rendements. Ceci peut une nouvelle fois se vérifier sur la figure 11.5 (corrélation de 0,27). En revanche, celles-ci sont en général suffisamment importantes en hiver. Des valeurs supérieures à la normale n'engendrent qu'un faible bénéfice en termes de rendement voire même causent des dégâts, comme on l'a dit dans le cas de plantule. Au contraire, les températures hivernales étant souvent assez basses (notamment sur les plateaux de l'intérieur du pays), les anomalies positives s'avèrent souvent très bénéfiques. On choisit alors de considérer la température en février (corrélation de 0,17).

Enfin, une température élevée en tout début de cycle favorise largement la croissance dès la fin de la période de germination. C'est pourquoi la température en septembre semble être une variable ayant un impact réel. Encore une fois, ceci se vérifie sur la figure 11.5 (corrélation de 0,26).

Les variables prédictives choisies sont résumées dans le tableau 11.1

Anomalie précipitation juillet
Anomalie précipitation avril
Anomalie température février
Anomalie température septembre

TABLE 11.1 : Tableau récapitulatif des prédicteurs provenant de l'expertise agronomique

Remarque 11.11: *Il convient de préciser que l'on aurait pu effectuer d'autres choix. Notamment, on aurait pu remplacer la température en février par une autre température hivernale ou même automnale.*
Par ailleurs, il faut bien être conscient que les corrélations présentées dans le tableau ont été calculées sur de courtes séries (30 données). Ainsi, il est tout à fait envisageable d'avoir des valeurs non nulles bien qu'il y ait indépendance. Ceci est d'autant plus vrai compte-tenu de l'imprécision sur les données. C'est justement pour trouver des relations de cause à effet réelles qu'il a fallu interroger les agronomes.

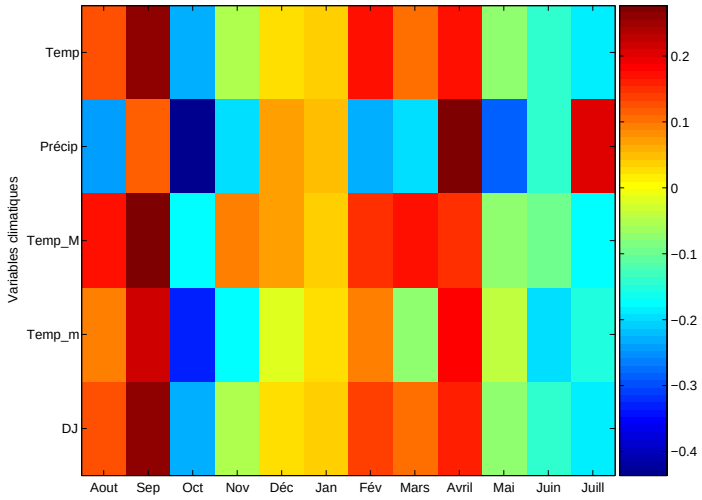


FIGURE 11.5 : Corrélations entre le rendement en sucre et les différents prédicteurs potentiels pour la zone Tadla

11.5.2 Choix entre un indice par zone ou un indice global

Dans un premier temps, nous envisagions de construire un indice par zone. En effet, comme on peut le voir sur la carte 10.1, les situations géographiques de celles-ci sont extrêmement diversifiées. Ainsi, l'intensité des variables météorologiques peut varier significativement de l'une à l'autre (les précipitations sont notamment plus importantes à proximité de l'océan). De ce fait, l'impact d'une même variable peut alors différer d'une zone à l'autre. A titre indicatif, l'anomalie de précipitation en avril aura un impact plus important dans les zones centrales très sèches que dans les zones côtières où les précipitations moyennes sont plus importantes. Ceci s'illustre parfaitement si on compare les zones Tadla (à l'intérieur) et Nador (à proximité de la côte). Dans un cas, la corrélation vaut 0,27 alors que dans l'autre, elle est seulement de 0,02.

Nous avons donc tout d'abord utilisé nos prédicteurs zone par zone, étudiant en premier lieu le modèle linéaire sans régularisation. La figure 11.6 obtenue en appliquant le modèle à 50 bases de test différentes (comportant 5 ou 10 valeurs), montre des résultats en généralisation extrêmement mauvais : la corrélation moyenne est de 0,04 et la dispersion très importante. Il advenait donc que, soit la relation entre les variables climatiques et les rendements comportait des effets non linéaires qu'on ne pouvait capter, soit le véritable problème était le manque de données. Nous nous sommes alors tournés vers le non linéaire et avons appliqué les méthodes de régularisation décrites précédemment. Nous avons tenté d'optimiser les paramètres de weight decay, de early stopping et de bruit (input perturbation) obtenant les valeurs dans le tableau 11.2. Néanmoins, comme on peut le voir sur la figure 11.6, nous avions systématiquement des corrélations négatives.

Bruit	Nombre d'époch s	Weight Decay
20 %	10	1

TABLE 11.2 : Paramètres de régularisation dans le cas d'une zone :Tadla

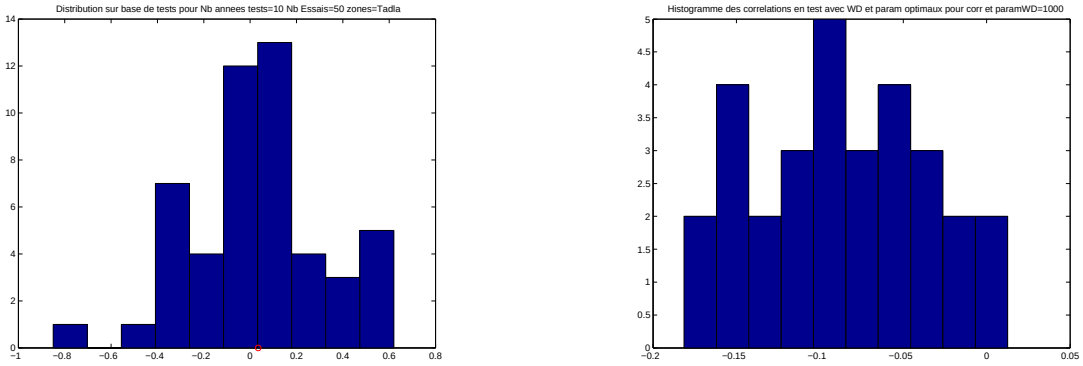


FIGURE 11.6 : Histogramme des corrélations pour Tadla dans les cas linéaire(g) et non linéaire(d)

La conclusion est alors apparue clairement : compte-tenu du manque de données, il est impossible de construire un indice spécifique à chaque zone ayant une bonne faculté de généralisation. Comme on l'a vu dans la section 11.5.3 page 78 , les réseaux de neurones comportent un nombre de degrés de liberté aussi grand qu'on le souhaite et s'avèrent donc très efficaces, néanmoins à conditions d'être contraints par suffisamment de données. Ici, les 30 données de rendement par zone dont on dispose s'avèrent être beaucoup trop insuffisantes. **Il est donc quasiment impossible d'utiliser un réseau de neurone dans un tel cas.**

Il s'est par conséquent avéré nécessaire de considérer l'ensemble des zones afin de **pouvoir effectuer l'apprentissage sur un nombre de données suffisant.** Un premier test a été effectué en non linéaire avec les paramètres de régularisation optimaux que l'on avait obtenus dans le cas d'une zone. Celui-ci a donné de meilleurs résultats (corrélation de 0,1), ce qui nous a encouragés à poursuivre sur cette voie.

Il restait alors dans un premier temps à choisir les zones sur lesquelles nous allions effectuer l'apprentissage. Il s'agissait de déterminer si l'on considérait la totalité de celles-ci ou alors simplement celles pour lesquelles on possédait suffisamment de données. Comme on l'a précisé dans la section 11.3, on applique désormais systématiquement bootstrap et leave-one-out pour réaliser ce genre d'étude. Comme on peut l'observer sur la figure 11.7, en linéaire, les résultats obtenus sont meilleurs lorsque l'on se contente des 8 zones pour lesquelles les données sont complètes. La distribution semble légèrement translatée. Ceci s'interprète par le fait que pour les 8 zones restantes, nous ne possédons que 6 données de rendement en sucre. Ainsi, **notre modélisation de tendance ne peut qu'être très imprécise dans une telle situation.** Les anomalies de rendement calculées sur ces zones sont donc probablement encore expliquées par d'autres facteurs que les facteurs climatiques, d'où la difficulté d'établir un lien avec ceux-ci et par conséquent de mauvais résultats en généralisation. Les résultats en apprentissage et en test sont donnés dans le tableau 11.3

Remarque 11.12: *Ce dernier point est particulièrement important puisqu'il illustre l'importance de la qualité de modélisation des tendances.*

Nous décidons donc par la suite de ne considérer que les 8 zones pour lesquelles nous possédons suffisamment de données. Ceci est légitime dans le cas linéaire et nous l'extrapolons au cas non linéaire compte-tenu de l'interprétation qu'on a émise précédemment. D'ailleurs, comme on le répétera par la suite, pour des raisons de temps de calcul, il n'est pas envisageable de tester l'ensemble des possibilités.

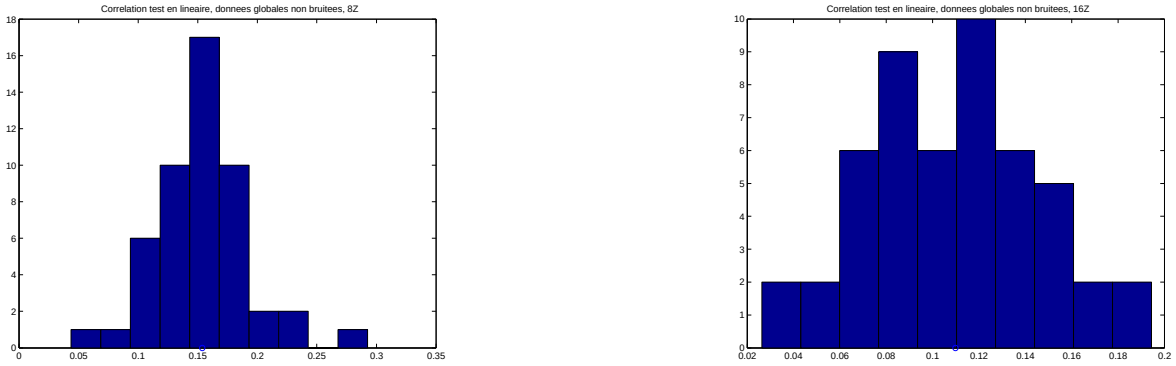


FIGURE 11.7 : Histogramme des corrélations en généralisation sur les données de 8(g)et 16(d) zones non perturbées

	8 zones	16 zones
Moyenne Corrélation en apprentissage	0,22	0,18
Moyenne Corrélation en test	0,16	0,11

TABLE 11.3 : Comparaison prise en compte de 8 ou 16 zones

Remarque 11.13: *Il est délicat de construire un test de significativité des différences de corrélation moyenne obtenues. En effet, les corrélations (entre série réelle et reconstruite) calculées sur les différentes bases de test, ne sont pas indépendantes (on pratique le bootstrap et donc les différents apprentissages s'effectuent toujours en partie sur des données communes). Néanmoins, les moyennes de corrélations sont assez stables même lorsque le nombre bases de test pris en compte est assez faible, et donc quand l'indépendance est à peu près respectée. Les différences observées semblent donc significatives.*

11.5.3 Choix du modèle

Maintenant que l'on sait précisément que l'indice doit être conçu à partir des 8 zones pour lesquelles nous possédons suffisamment de données, **il convient d'arbitrer entre les différentes catégories de modèles étudiées dans la section 11.1.2, à savoir :**

- **Le modèle linéaire ;**
- **Les réseaux de neurones ;**
- **Les modèles hiérarchiques bayésiens .**

Régression linéaire

Optimisation des paramètres

La première étape est de trouver les paramètres de régularisation optimaux. Dans le cas linéaire, le seul paramètre est le niveau de bruit introduit dans les données. On le détermine par leave-one-out. Comme on peut le voir sur les figures 11.8 et 11.9, **le niveau de bruit optimal (qui maximise la corrélation et minimise la RMS en généralisation) semble être de 10 %.** Néanmoins, la variabilité des corrélations obtenues pour ce niveau de 10 % (écart-type

pour plusieurs séries obtenues par leave-one-out), est du même ordre de grandeurs que celle des corrélations sur l'ensemble des niveaux de bruit. Ainsi, le choix de 10 % s'avère un peu arbitraire car il n'est pas forcément plus légitime que des niveaux de 15 ou 20 %. On peut même s'interroger sur la nécessité d'en introduire. Ce phénomène provient de la mauvaise qualité des données. **Dans le cas général, l'apport de l'introduction de bruit est véritablement quantifiable.**

Remarque 11.14: *On observe clairement sur les figures 11.8 et 11.9 que l'erreur décroît et la corrélation croît en apprentissage lorsque le bruit diminue. En généralisation, ce n'est pas le cas : il y a un niveau optimal.*

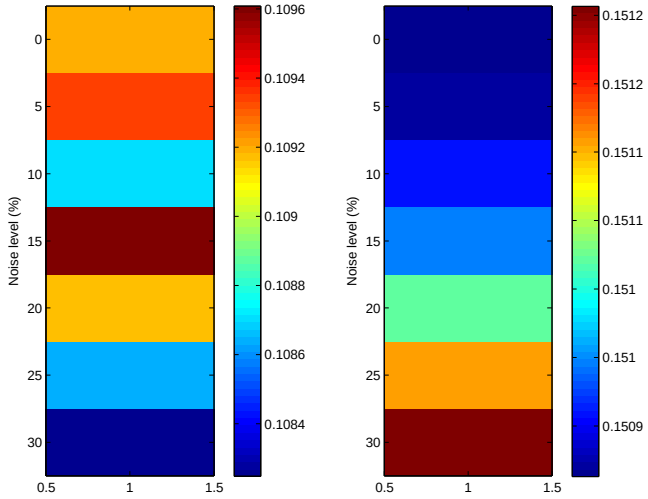


FIGURE 11.8 : Tableau des erreurs quadratiques en généralisation (g) et en apprentissage (d) en fonction du niveau de bruit dans le cas de la régression linéaire

Résultats en généralisation

Ceux-ci sont résumés dans le tableau 11.4

	Corrélation apprentissage	Corrélation test	Erreur apprentissage	Erreur test
Moyenne	0,22	0,18	0,15	0,15
Ecart-type	0,004	0,06	0,001	0,01

TABLE 11.4 : Résultats en linéaire régularisé

Réseaux de neurones

Optimisation des paramètres

Nous avons fait le choix de netlab pour l'implémentation des réseaux de neurones. Le temps de calcul étant relativememt important, il s'est avéré impossible d'effectuer toutes les optimisations en parallèle. Remarquons tout d'abord qu'étant donné que nous considérons 4 entrées, il nous

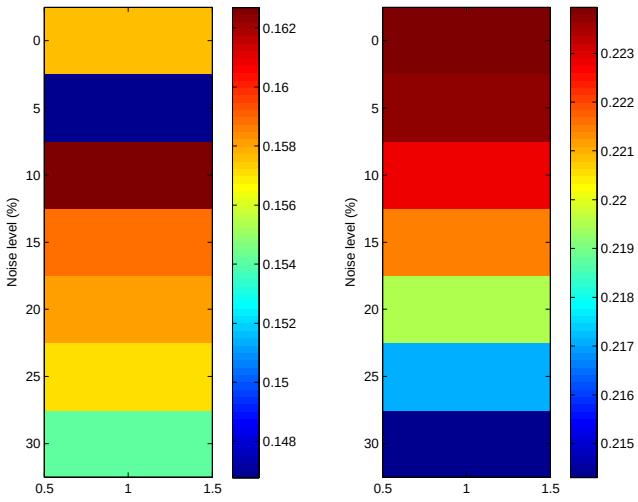


FIGURE 11.9 : Tableau des corrélations en généralisation (g) et en apprentissage (d) en fonction du niveau de bruit dans le cas de la régression linéaire

est apparu pertinent de ne prendre que 4 unités dans la couche cachée. En effet, compte-tenu du problème de sur-apprentissage provenant du manque de données, il est inutile et même néfaste en vue de la généralisation de considérer un nombre d'unités trop important. Les paramètres de régularisation à optimiser dans le cas du réseau de neurone sont : le niveau de bruit, le nombre d'époch (early stopping) et enfin le paramètre de weight decay.

Compte-tenu des contraintes de temps mentionnées plus haut, il était impossible de tester l'ensemble des triplets correspondant à ces trois paramètres. Nous avons alors choisi de prendre comme niveau de bruit celui que l'on avait obtenu en linéaire. De toute façon, comme on l'a vu, l'introduction de bruit est nécessaire car elle limite le sur-apprentissage mais néanmoins, on ne peut pas précisément définir de valeur optimale. En ce qui concerne le paramètre de weight decay, nous choisissons 0.1. Dans le cas d'une zone, le paramètre optimal, si tant est que l'on puisse en définir un dans ce cas, était de 1. Dans le cas présent, le sur-apprentissage étant plus faible compte-tenu du nombre de données plus élevé, nous pouvons choisir un paramètre plus bas. Enfin, le nombre d'époch est obtenu selon le même principe que le niveau de bruit dans le cas linéaire.

Remarque 11.15: *On voit clairement sur les figures 11.10 et 11.11 que les erreurs décroissent (et que simultanément les corrélations croissent) avec le nombre d'époques. Ceci se comprend aisément puisque cela correspond à des étapes supplémentaires de la descente de gradient visant à minimiser la fonction de coût. En revanche, ceci n'est pas le cas en généralisation étant donné qu'à partir d'un moment, les étapes supplémentaires de cette minimisation ne servent qu'à apprendre le bruit des données.*

Nous résumons les valeurs des paramètres de régularisation dans le tableau 11.5

Résultats en généralisation

Ayant obtenu les valeurs des paramètres optimaux, il convient désormais de tester la capacité de généralisation de notre modèle grâce, comme précédemment, à une série de leave-one-outs. On obtient l'histogramme en généralisation figure 11.12.

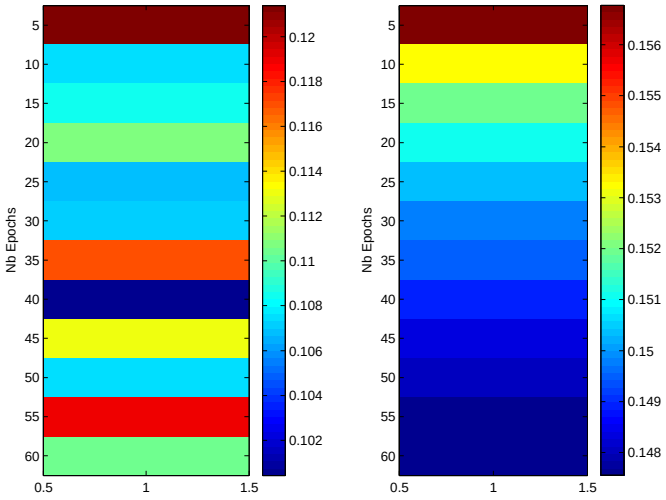


FIGURE 11.10 : Tableau des erreurs quadratiques en généralisation (g) et en apprentissage (d) en fonction du nombre d'époques dans le cas des réseaux de neurones

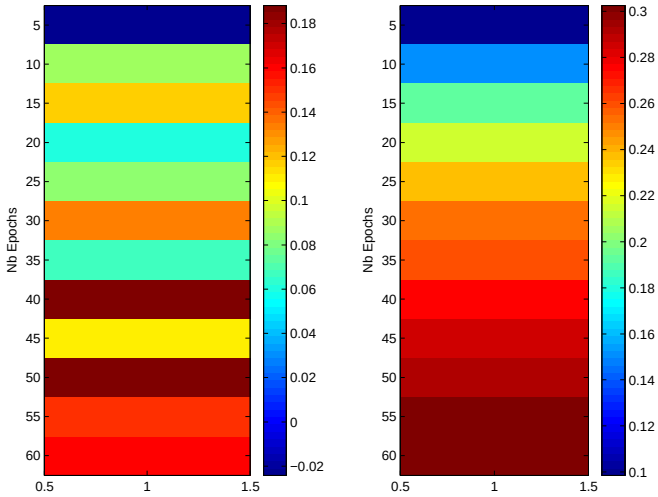


FIGURE 11.11 : Tableau des corrélations en généralisation (g) et en apprentissage (d) en fonction du nombre d'époques dans le cas des réseaux de neurones

Bruit	Nombre d'époch s	Weight Decay
10 %	40	0,1

TABLE 11.5 : Paramètres de régularisation quand on utilise l'ensemble des données

Il semble enfin intéressant de comparer les résultats précédents avec ceux obtenus dans le cas

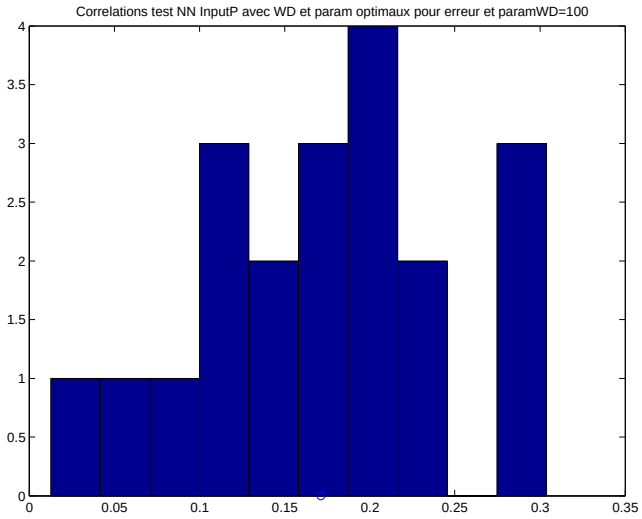


FIGURE 11.12 : Histogramme des corrélations en généralisation dans le cas du réseau de neurones régularisé utilisant les données bruitées

d'un réseau de neurones sans aucune régularisation, figure 11.13

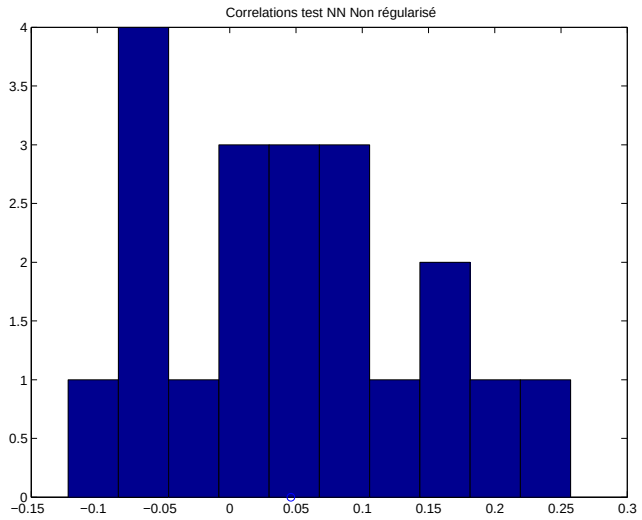


FIGURE 11.13 : Histogramme des corrélations en généralisation dans le cas du réseau de neurones non régularisé sans introduction de bruit

On observe sur la figure 11.12 que les différentes méthodes de régularisation permettent de rendre les corrélations en test quasiment toutes positives. La distribution se resserre autour d'une valeur moyenne plus élevée (0.17 contre 0.05).

Afin d'avoir les idées plus claires sur l'intérêt des différentes méthodes de régularisation employées, nous résumons les résultats obtenus dans le tableau 11.6.

L'introduction de bruit dégrade le résultat en apprentissage et l'améliore en gé-

		Corrélation en		Erreur en	
		appren-tissage	test	appren-tissage	test
RN régularisé, données bruitées	Moyenne	0,27	**0,17**	0,15	**0,14**
	Ecart-type	0,04	0,08	0,01*10 $^{-1}$	0,01
RN régularisé, données non bruitées	Moyenne	0,25	**0,11**	0,16	**0,15**
	Ecart-type	0,04	0,09	0,002	0,02
RN non régularisé, données non bruitées	Moyenne	0,09	**0,05**	0,16	**0,16**
	Ecart-type	0,08	0,10	0,005	0,02

TABLE 11.6 : Impact des méthodes de régularisation

néralisation. Ceci est tout à fait logique étant donné que celle-ci limite la possibilité de coller aux données et donc le sur-apprentissage. On remarque par ailleurs qu'elle a un effet similaire à la régularisation. Nous avons d'ailleurs montré dans la partie 11.4.1 que **l'introduction de bruit est équivalente aux méthodes de régularisation classiques de type "weight decay".** Ceci est d'ailleurs général à toutes les méthodes de régularisation.

Linear and non linear mixed effect

Linear mixed effect

Nous avons utilisé lme qui est un modèle hiérarchique bayésien utilisable en R. L'intérêt provient du fait que l'on peut utiliser l'ensemble des données. On effectue un apprentissage sur l'ensemble des zones (ce qui permet de limiter le sur-apprentissage et de dégager les informations) tout en permettant un comportement spécifique à chacune d'entre elles (les coefficients de la régression peuvent être dépendants de la zone) Dans le cadre de notre étude, aucune différence ne peut être dégagée entre les zones (la différence entre les coefficients étant de l'ordre de 10^{-12}) et les coefficients obtenus sont les mêmes que dans le cas du modèle linéaire.

Il n'est donc pas possible d'opérer de différenciation compte-tenu du manque d'informations.

Non linear mixed effect

Nous avons utilisé nlme, version non linéaire de lme qui permet de rentrer la fonction souhaitée. Nous avons implémenté la formule du réseau de neurone à 4 unités cachées, et, comme, **dans le cas linéaire, le modèle n'apporte rien du fait du manque de données, de leur imprécision, et du manque d'informations.**

11.5.4 Procédure adoptée

Comme on peut le constater par comparaison des résultats obtenus avec les différents types de modèle, il convient, dans le cas présent, de retenir le modèle linéaire avec introduction d'un bruit de 10 %. Le résulat en généralisation est proche de celui obtenu avec le modèle non linéaire régularisé mais le gain de temps est très important.

	Corrélation apprentissage	Corrélation test	Erreur apprentissage	Erreur test
Linéaire régularisé	0,22	0,18	0,15	0,15
Non linéaire régularisé	0,27	0,17	0,15	0,15

TABLE 11.7 : Comparaison résulats

Remarque 11.16: *Comme nous venons de le dire, on peut,* ***grâce aux méthodes de régularisation, obtenir une faculté de généralisation du réseau de neurones proche de celle du linéaire. Ceci est très rassurant car il s'agit d'une propriété générale : un réseau de neurones doit toujours être en mesure, s'il est bien régularisé, d'aboutir à des résulats au pire équivalents à ceux de la régression linéaire. Néanmoins, étant donné le manque de données, leur manque de précision et enfin le manque d'informations, le non-linéaire n'apporte rien par rapport au linéaire dans le cas présent.***

11.5.5 Résultats

La performance de l'indice est testée par bootstrap sur 2500 plages de 10 exemples test. L'histogramme correspondant est visible figure 11.14. On ne pratique pas de leave-one out, mais il y a équivalence des deux méthodes pour un grand nombre de tirages d'années test.

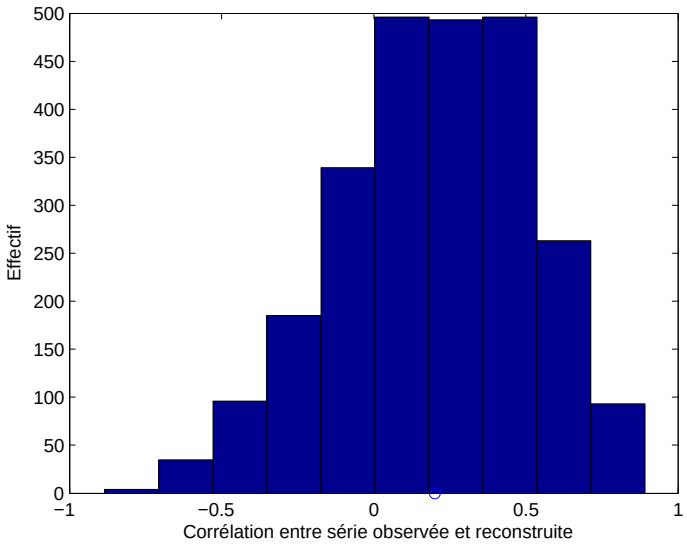


FIGURE 11.14 : Histogramme des corrélations en généralisation pour l'indice retenu

Il donne une idée de la dispersion et donc de la confiance que l'on peut avoir dans notre indice.

Les résulats sont résumés dans le tableau 11.8

La dernière étape consiste alors à fournir le jeu de coefficients du modèle de régression. Il convient de les déterminer sur la base d'apprentissage la plus large possible, et donc sur la totalité

	Corrélation	Erreur
Moyenne	0,20	0,15
Ecart-type	0,31	0,06

TABLE 11.8 : Résultats en généralisation

des données des 8 zones.

L'indice retenu (prédicteurs et coefficients correspondants) est donné dans le tableau 11.9

Prédicteurs	Coefficients
Biais	0,019
Anomalie de précipitation en juillet	0,033
Anomalie de précipitation en avril	0,044
Anomalie de température en février	0,442
Anomalie de température en septembre	2,913

TABLE 11.9 : Indice final

Remarque 11.17: *La procédure de runs d'ensemble fournit environ le même résultat. C'est pourquoi nous ne donnons pas les coefficients correspondants. Il s'agit d'un procédé utile dans le cas général. Néanmoins, ici, compte-tenu des raisons déjà évoquées (à savoir manque de données, imprécision, manque d'informations), le bénéfice n'est pas réellement quantifiable.*

Nous confrontons la prévision de l'indice aux données réelles dans le cas de la zone de Berkane, sur la figure 11.15.

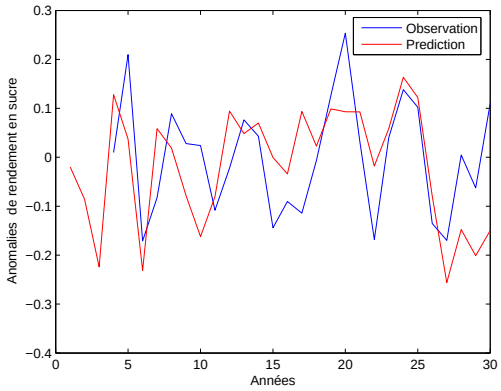


FIGURE 11.15 : Observation vs Prédiction de l'indice pour la zone de Berkane

Par ailleurs, afin de bénéficier d'une vision globale, nous traçons le scatter plot sur l'ensemble des données (prévisions en fonction des observations) sur la figure 11.16.

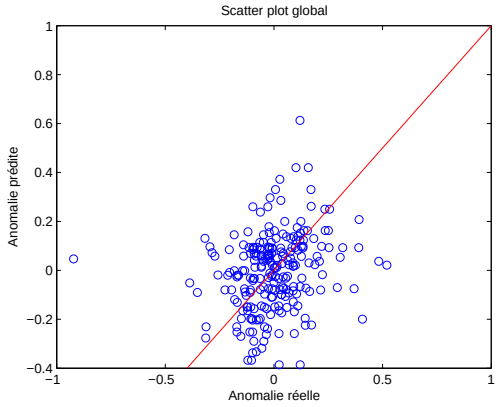


FIGURE 11.16 : Prédiction en fonction des observations sur l'ensemble des données

11.6 Indice avec prédicteurs obtenus par optimisation statistique

11.6.1 Introduction

Comme nous avons pu le voir dans le chapitre précédent, le modèle qui apparaît le plus pertinent dans le cas présent, compte-tenu du manque de données, de leur imprécision (du fait des dates de semis différentes) et du manque d'informations (notamment en ce qui concerne les données d'irrigation), est la régression linéaire simple avec comme outils de régularisation (ou de robustification) le bruitage des données avec comme valeur de bruit optimale pour chaque variable d'entrée 10 % de l'écart-type de cette variable.

11.6.2 Procédure adoptée

Comme nous l'avons dit, le choix des prédicteurs s'effectue ici de manière **purement statistique**. A la différence de la section précédente, il ne se base sur aucun a priori. L'idée est de construire l'indice par itération (méthode stepwise) en rajoutant à chaque étape le prédicteur qui maximise la capacité de généralisation de l'indice sur la base de validation. Cette dernière est construite par leave-one-out successifs. Ainsi, la démarche est la suivante :

On considère ainsi 50 plages différentes comportant chacune 10 exemples test.

1. On choisit une première base de test et on considère le complémentaire dans les données noté *Comp*.
   (a) On veut choisir le premier prédicteur optimal.
      i. On teste un premier prédicteur.
         A. On effectue alors un premier leave-one out dans *Comp* ;
         B. On calcule les coefficients de la régression (du rendement sur ce prédicteur) sur les données de *Comp* privées de celles correspondant au leave-one-out (une donnée de rendement et une donnée correspondant au prédicteur). On calcule alors la prévision du modèle pour la donnée laissée de côté.

C. On refait 300 leave-one-out, obtenant une série observée de taille 300 que l'on peut comparer à la série prédite grâce au modèle. On comprend alors que l'intérêt du leave-one-out réside dans le fait qu'il permet de constituer artificiellement une base de validation de grande taille.

D. On calcule la corrélation linéaire ainsi que l'erreur quadratique entre les deux séries.

ii. On réitère ceci avec les 60 prédicteurs potentiels et on choisit celui qui maximise la corrélation ou minimise l'erreur quadratique entre les deux séries.

(b) On réitère alors la démarche précédente pour le deuxième prédicteur. Il s'agit alors de maximiser la corrélation (ou de minimiser l'erreur quadratique) sur la base de validation (constituée par les 300 leave-one-out) entre la série réelle et la série reconstruite à partir de l'indice comportant le premier prédicteur et celui que l'on est en train de choisir.

(c) Par récurrence, on réitère ce procédé pour les prédicteurs suivants.

2. On réédite la même démarche sur les autres bases de test.

On choisit finalement la combinaison de prédicteurs maximisant la corrélation (ou minimisant l'erreur quadratique) sur la base de test finale.

Remarque 11.18: *Cette procédure revient en fait à trouver des prédicteurs « orthogonaux », c'est-à-dire apportant des informations indépendantes. On retrouve un peu l'idée de l'analyse en composantes principales.*

Remarque 11.19: *On ne teste que les variables climatiques elles-mêmes et non des fonctions non linéaires de celles-ci (on aurait pu par exemple effectuer la régression linéaire du rendement sur le carré de la température d'un mois donné). Néanmoins, ce cas est contenu dans la classe de fonctions proposées par le réseau de neurones et a donc été indirectement testé et rejeté dans la section 11.5.*

Comme on l'a vu dans la section 11.3.1, cette démarche de séparation en trois bases est générale. Elle est notamment souvent appliquée dans le cas des réseaux de neurones. La base de validation sert alors à choisir l'architecture (nombre d'unités cachées...) du réseau. Afin de s'affranchir du problème de sur-apprentissage sur la base de validation, on fait appel à une troisième base : la base de test. Il faut bien comprendre que la démarche adoptée ici est équivalente, étant donné qu'à l'architecture du réseau correspond l'architecture de l'indice, à savoir la combinaison des différents prédicteurs.

11.6.3 Résultats

Il advient finalement, quand on se limite à 4 prédicteurs, que l'indice optimal est constitué des prédicteurs suivants :

- L'anomalie de degrés-jours en septembre ;
- L'anomalie de température minimale en novembre ;
- L'anomalie de précipitation en août ;
- L'anomalie de précipitation en novembre ;

L'indice étant désormais fixé, on le teste par bootstrap sur 2500 bases d'exemples différentes. L'histogramme obtenu est présenté figure 11.17. On observe une distribution asymétrique, penchée vers la droite. Ceci est positif. Celle-ci est très similaire à celle obtenue avec l'indice issu de l'expertise agronomique, mais translatée d'environ 0.1 vers les valeurs supérieures.

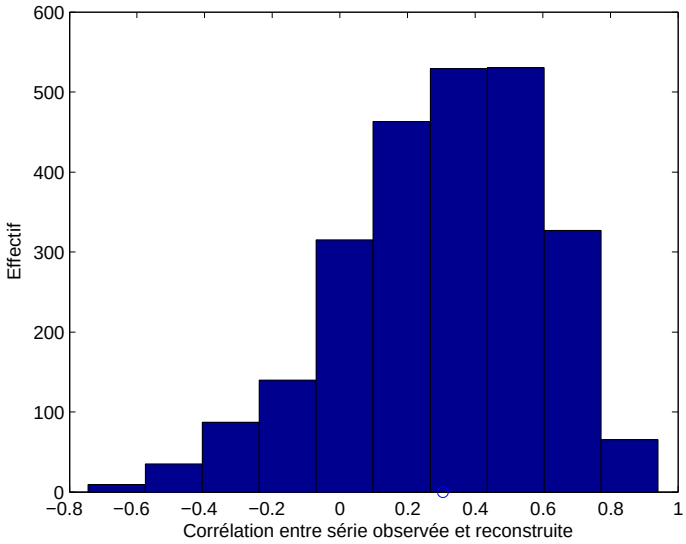


FIGURE 11.17 : Histogramme des corrélations pour l'indice optimisé

	Corrélation	Erreur
Moyenne	0,31	0,14
Ecart-type	0,29	0,06

TABLE 11.10 : Résultats en généralisation pour l'indice optimisé

Le tableau 11.10 résume les résultats :

Les coefficients de la régression sont calculés en utilisant la base de données complète. L'indice optimisé complet est décrit dans le tableau 11.11.

Prédicteurs	Coefficients
Biais	-0,003
Anomalie de degré-jour en septembre	1,503
Anomalie de température minimale en novembre	-0,833
Anomalie de précipitation en août	-0,005
Anomalie de précipitation en novembre	-0,08

TABLE 11.11 : Indice optimisé final

On constate donc que cet **indice obtenu de manière statistique permet de retrouver des variables ayant un réel impact d'après les expertises agronomiques**. Par exemple, l'anomalie de température en septembre, qui a une très grande importance pour les agronomes, se retrouve. A noter que l'on obtient ici les degrés-jours, qui sont les degrés ayant une réelle importance pour la plante. En effet, au-dessus d'une certaine valeur, l'augmentation de température

n'a plus d'impact réel. On observe également un impact négatif de la précipitation en novembre, ce qui rejoint ce que nous disions dans le chapitre 11.5 au sujet des fortes précipitations de la fin d'automne et du début d'hiver qui ont tendance à noyer les jeunes plantes.

Lorsque l'on construit un indice prenant en compte davantage de prédicteurs, des variables prédictives telles l'anomalie de température minimale en janvier apparaissent. Cette grandeur est une nouvelle fois en accord avec les constatations des agronomes.

Ainsi, on constate que la méthode statistique employée permet de retrouver des prédicteurs ayant une réelle signification physique. Ceci est tout à fait logique étant donné qu'elle est basée sur le principe de la maximisation des scores en généralisation. Même si les résultats ne sont pas exceptionnels compte-tenu des limites déjà évoquées, on est tout de même très satisfait de la capacité de la méthode statistique à trouver les prédicteurs les plus influents.

On constate que l'on aboutit finalement à un indice d'impact présentant un score de 0,31 en corrélation avec une probabilité de 70 % d'être entre 0,2 et 0,6 Il s'agit du résultat optimal dans le cadre de ce problème.

Ce résultat est extrêmement important car il signifie, comme on le verra dans le chapitre 12, que cet indice est difficilement exploitable. Il faut donc se méfier des résultats annoncés dans certaines études. **Parfois, en effet, seuls les résultats en sur-apprentissage sont présentés.**

A titre d'illustration, nous confrontons la prévision de l'indice optimisé aux données réelles pour la zone de Berkane sur la figure 11.18.

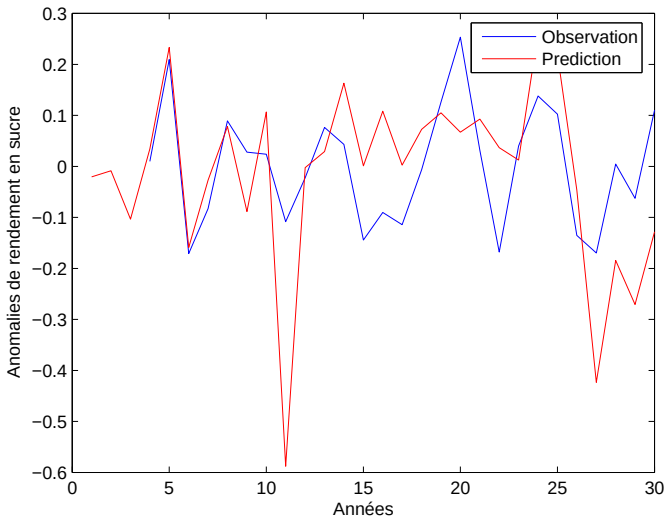


FIGURE 11.18 : Observation vs Prédiction pour la zone de Berkane

Par ailleurs, nous représentons l'ensemble des données prévues en fonction des observations sur la figure 11.19.

Remarque 11.20: *Nous ne présentons pas en détail les résultats issus des runs d'ensemble car ceux-ci sont sensiblement équivalents à ceux du modèle simple. On peut le voir dans le tableau 11.12. Il s'agit d'une méthode de régularisation très intéressante dans le cas général car il s'agit plutôt d'une régularisation de la fonction reliant les entrées et les sorties.*

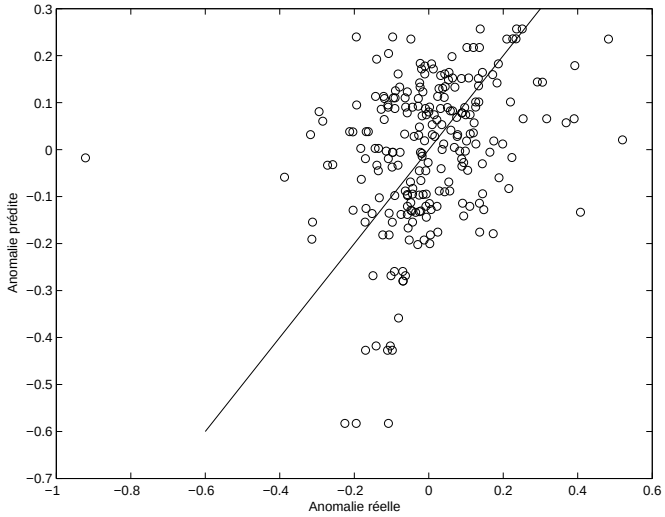


FIGURE 11.19 : Prédictions en fonction des observations sur l'ensemble des données

	Corrélation	Erreur
Moyenne	0,30	0,13
Ecart-type	0,35	0,05

TABLE 11.12 : Résultats en généralisation pour l'indice optimisé, version runs d'ensemble

Les points essentiels de la construction de nos deux indices sont donc :

- **Manque de données, données imprécises, manque d'informations (données d'irrigation) ;**
- **Nécessité de considérer les données globalement (aucune généralisation, en linéaire comme en non linéaire, dans le cas d'une zone), ce qui pose un problème compte-tenu de l'impact différent des anomalies selon que la région considérée est plutôt sèche (plateaux intérieurs) ou humide (à proximité de la mer) ;**
- **Nécessité de régulariser les données ;**
- **Difficulté d'utilisation du réseau de neurones quand on dispose d'une base de petite taille (degrés de liberté non contraints) et d'un manque d'information. On retrouve les mêmes résultats qu'en linéaire en régularisant mais pas mieux ;**
- **Méthode statistique de qualité car elle retrouve les prédicteurs pertinents ;**
- **Néanmoins, résulat assez médiocre en généralisation compte-tenu des limitations provenant des données .**
- **Très grande différence entre les chiffres que l'on peut donner en surapprentissage à propos d'un indice(arbitrairement élevés) et ses capacités réelles de généralisation.**

Chapitre 12

Conclusion assurancielle de l'étude

12.1 Contrats d'assurance

12.1.1 Méthode de tarification

Ici, le dommage correspond à une valeur I anormalement basse de notre indice, représentative de rendements très faibles. En-dessous d'une valeur seuil Is, l'indemnité est proportionnelle à l'écart $Is - I$. Si V est la valeur pécunière relative à la baisse d'un point de l'indice, l'indemnité versée vaut alors :

$$Indem = max(V(Is - I), 0)$$

Ici, nous choisissons V de manière à ce que l'indemnité corresponde exactement au dommage subi par rapport au seuil. V est donc égale à l'équivalent pécunier de la valeur de l'écart-type de rendement l'année considérée. En effet, l'ensemble de notre étude est basée sur des anomalies relatives à la valeur de l'écart-type de rendement. On se ramène donc à la perte de rendement estimée par l'indice en multipliant la valeur de l'anomalie par la valeur de l'écart-type. Nous prenons par ailleurs $Is = 0$, c'est-à-dire que tout rendement inférieur entraîne des indemnités.

La prime pure correspond à l'espérance mathématique de l'indemnité, soit $E(Indem)$ En effet, l'assurance est basée sur le principe de la mutualisation. L'assureur doit posséder un très grand nombre de contrats, et en vertu de la loi des grands nombres, la moyenne empirique de dépense par contrat tend alors vers son espérance mathématique (sous les hypothèse étudiées dans le chapitre 6 page 33).

Dans le cas présent, si l'assureur ne souscrit des contrats que dans une zone, il n'y a aucune mutualisation étant donné que l'indice possède une valeur commune à tous les contrats de cette zone. Le montant de l'indemnité varie alors énormément d'une année sur l'autre. Il y a davantage de mutualisation lorsque l'assureur possède des contrats dans l'ensemble des 16 zones. Néanmoins, ce nombre de 16 est faible et les fluctuations autour de la moyenne sont donc très importantes. Par ailleurs, les 16 risques considérés (correspondant aux différentes zones) ne sont pas indépendants compte-tenu de la corrélation spatiale de l'indice. Enfin, les risques ne sont pas réellement homogènes étant donné que l'indice n'a pas exactement la même loi dans toutes les zones. **Nous ne sommes donc pas dans le cadre d'application de la loi des grands nombres.**

Compte-tenu de ces limitations, l'assureur se doit donc d'ajouter à la prime pure un chargement de sécurité ou technique CT important. Ce chargement peut être proportionnel à l'espérance

ou mieux à l'écart-type (c'est-à-dire la volatilité). Il s'agit du choix que nous effectuons ici. Néanmoins, des chargements basés sur les moments d'ordre supérieur sont également envisageables. Un tel chargement est calculé de façon à minimiser le risque de ruine de l'assureur qui pourrait intervenir dans le cas d'un sinistre d'ampleur exceptionnelle. Ce risque de ruine sera d'autant plus faible que le chargement est élevé.

Cependant, il existe également des risques non mutualisables :

- Les erreurs de modèles (choix du modèle ou estimation des paramètres). Si le modèle fournit une mauvais espérance des sinistres, l'écart grandit avec le portefeuille. Les erreurs ne se compensent pas comme dans le cas des fluctuations autour de la moyenne.
- Les événements extrêmes, difficiles à capter dans le modèle.

Ici, les risques de modèles sont importants car le calcul de la prime pure n'est basé que sur un historique de 30 ans et s'avère donc imprécis. Pour davantage de précision, on peut calculer une prime unique pour l'ensemble des zones.

L'assureur peut alors diminuer les effets néfastes de ces risques non mutualisables en augmentant encore le chargement de sécurité et en disposant de fonds propres sous forme de capital supérieur à une exigence minimale fixée réglementairement. Il s'agit d'une source de capitaux supplémentaire, au niveau global.

Dans cette étude, on choisit de prendre $CT = \beta sigma(D)$ où D représente le dommage annuel, avec $\beta = 0.2$ ainsi qu'un seuil égal à 0. On considère donc qu'il y a dommage si la valeur de l'indice est inférieure à sa moyenne.

L'assureur se doit désormais de calculer le chargement commercial. Ce dernier vise à couvrir différents types de frais, comptabilisés de la manière suivante :

- Frais d'acquisitions des contrats : commissions des intermédiaires, campagnes de publicité ;
- Frais de règlement des sinistres (salaires des gestionnaires de sinistres) ;
- Frais de gestion des placements ;
- Frais d'administration des contrats (suivi des contrats...) ;
- Autres charges techniques ;
- Autres charges non techniques ;

Le plus souvent, les frais sont exprimés en pourcentage du chiffre d'affaires : On note F les frais totaux, PP la prime pure, PC la prime commerciale et N le nombre de contrats. On a alors :

$$F = N\Phi PC = N\Phi(PP + CT + \frac{F}{N}) = N\Phi(PP + CT) + \Phi F$$

d'où :

$$F = \frac{N\Phi(PP + CT)}{1 - \Phi}$$

donc :

$$PC = PP + CT + \frac{F}{N} = (PP + CT)(1 + \frac{\Phi}{1 - \Phi})$$

On choisit $\Phi = 10\%$.

La prime pure est évaluée de manière actuarielle. Dans le cas présent, nous utilisons la Burn Analysis qui conduit à calculer la prime d'assurance en se servant uniquement de l'historique des valeurs de l'indice (comme on vient de le voir, en équivalent pécunier) que l'on possède.

Une autre méthode serait de modéliser les données de coûts historiques évoquées précédemment et de modéliser leur fréquence à l'aide d'une distribution de probabilité de densité g(.) telle

que la loi Normale, Log-Normale, Gamma, Kernel, Valeurs extrêmes,...La prime pure est alors égale à la l'espérance mathématique de la fonction de coût pour l'assureur (prix des indemnités à verser) sous la mesure de probabilité choisie.

On a alors :

$$Prime\ pure = \int_R c(x)g(x)dx$$

Une telle intégrale est en général évaluée par des méthodes de type Monte-Carlo.

Cette méthode de calcul de la prime s'avère nettement plus robuste que la Burn Analysis (puisqu'on lisse la distribution empirique à l'aide d'une loi continue). Elle est néanmoins très sensible à la distribution choisie pour évaluer le risque. Dans le cas présent, l'objectif est d'établir un contrat par zone. Or, comme on l'a vu, on ne dispose que de 30 années. Parmi celles-ci, seules environ une dizaine engendrent des coûts pour l'assureur. Il est donc impossible d'effectuer un ajustement par une loi de probabilité.

Comme dans tout processus de tarification d'un assureur, celui-ci doit penser à préserver une marge pour les actionnaires ainsi qu'à se préoccuper des aspects commerciaux et concurrentiels.

12.1.2 Caractéristiques du produit

On applique la méthode de tarification décrite précédemment à notre produit. Nous menons dans un premier temps l'étude sur l'indice optimisé, avant d'analyser les cas de deux indices artificiels : l'un ne présentant aucun risque de base et l'un présentant un risque de base de 100 %.

Indice optimisé

La matrice de confusion est fournie dans le tableau 12.1. Comme on peut le voir, deux tiers des années où l'indice prévoit des dommages correspondent effectivement à des années où la récolte est inférieure à la moyenne. En revanche, on constate qu'il ne détecte que la moitié des années où se produisent des dommages. **Notre indice optimisé n'a donc pas trop tendance à prévoir des dommages qui ne se produisent pas, mais en revanche il ne rend compte que de la moitié des années où il y en a effectivement**.

	Années réelles	Années prévues
Années réelles	100 %	50 %
Années prévues	67 %	100 %

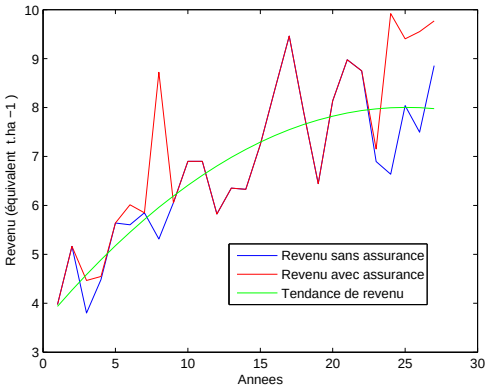
TABLE 12.1 : Matrice de confusion des dommages dans le cas de l'indice optimisé

La comparaison des revenus avec assurance et sans assurance est donnée figure 12.1.

Remarque 12.1: *Le revenu indiqué sur les graphiques ne prend pas en compte la prime. On voulait simplement illustrer le phénomène de compensation de pertes éventuelles. Par ailleurs, on ne tient pas compte de l'inflation. En effet, on suppose que le prix du sucre est constant. Ceci permet de se ramener aux graphes de revenus en rendements.*

Indice sans risque de base

Comme on peut le constater sur la matrice de confusion 12.2, l'indice prévoit exactement les années de dommage, ni plus ni moins. Il est donc parfait en termes de représentativité de la



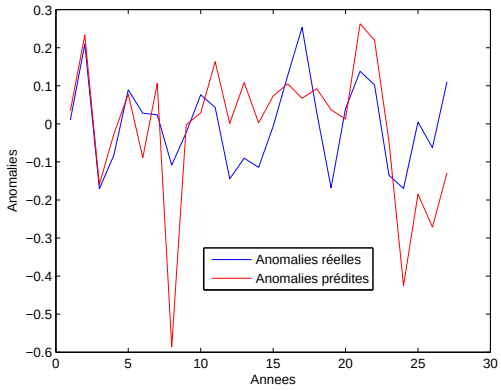
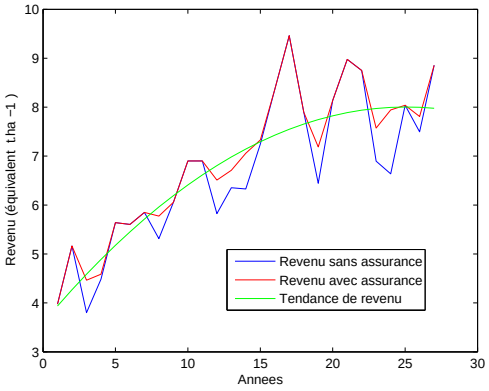


FIGURE 12.1 : Revenus de l'agriculteur(g) et comparaison des indices(d) (cas de l'indice optimisé)

réalité. On peut également observer ce phénomène sur la figure 12.2. **On voit en effet, que le revenu avec assurance est supérieur au revenu sans assurance si et seulement si il y a dommage**.

	Années réelles	Années prévues
Années réelles	100 %	100 %
Années prévues	100 %	100 %

TABLE 12.2 : Matrice de confusion des dommages dans le cas de l'indice de bonne qualité



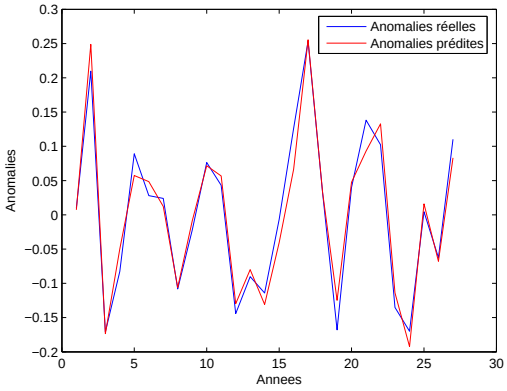


FIGURE 12.2 : Revenus de l'agriculteur(g) et comparaison des indices(d) (cas de l'indice en sur-apprentissage)

Indice de mauvaise qualité

La matrice de confusion est fournie dans le tableau 12.3. On remarque que parmi les années prévues par l'indice, seules un tiers correspondent à la réalité. Ainsi, il a fortement tendance à prévoir des dommages qui n'ont pas lieu en réalité. De manière générale, il a d'ailleurs tendance à surestimer le risque de dommages.

	Années réelles	Années prévues
Années réelles	100 %	50 %
Années prévues	43 %	100 %

TABLE 12.3 : Matrice de confusion des dommages dans le cas de l'indice de mauvaise qualité

La comparaison des revenus avec assurance et sans assurance est donnée figure 12.3 :

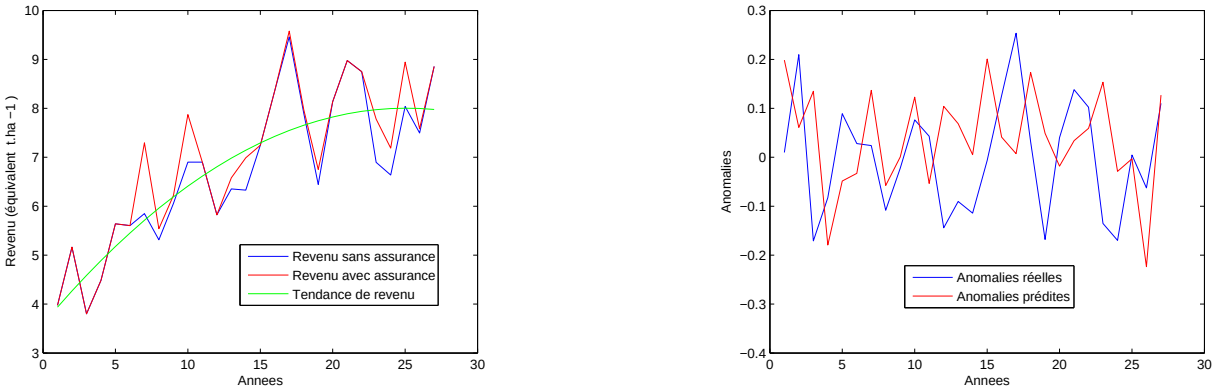


FIGURE 12.3 : Revenus de l'agriculteur(g) et comparaison des indices(d)(cas du mauvais indice)

Afin de clairement distinguer le très bon indice de celui de très mauvaise qualité, nous les comparons figure 12.4

Interprétations

Nous résumons les grandeurs attachées à chacun des contrats d'assurance dans le tableau 12.4 (les valeurs sont données en euros). **Finalement la prime d'assurance annuelle demandée à un agriculteur dans la zone de Berkane sera de 129 euros par hectare.** Cette prime est extrêmement élevée compte-tenu des chargements très importants nécessaires ici. Il est alors important de replacer ce chiffre dans le contexte réel. En effet, il est possible que l'agriculteur ne souhaite pas et ne puisse pas payer une prime aussi élevée. Par ailleurs, les concurrents de l'assureur proposant ce prix offriront peut-être des contrats plus avantageux. L'assureur doit alors pouvoir adapter la prime demandée aux réactions des autres acteurs de marché. Pour cela, il dispose, comme on l'a vu, de trois paramètres :

- le seuil Is ;
- la valeur pécunière d'un point d'indice V ;

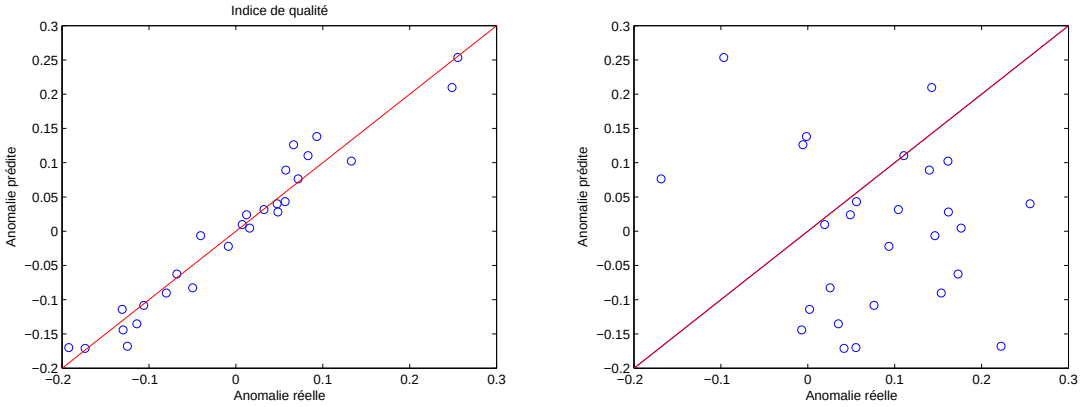


FIGURE 12.4 : Scatter plots pour le bon indice(g) et le mauvais(d)

– le coefficient de chargement technique β .

On observe que le mauvais indice engendre des indemnités proches du montant réel des dommages. Il faut bien comprendre que ceci est un hasard. En effet, il n'est absolument pas corrélé aux anomalies réelles de rendement mais descend en moyenne autant sous le seuil que les rendements réels. Néanmoins, comme on l'a vu, il ne se trouve pas en dessous du seuil les bonnes années. Ainsi, aucune indemnité ne sera versée certaines années de dommages alors que des versements pourront avoir lieu lors d'années ne présentant pas de dommage.

Notre indice optimisé engendre en moyenne des versements nettement plus importants que les pertes. Ceci n'est néanmoins pas satisfaisant. En effet, avec l'argent reçu du fait d'une mauvaise prévision de l'indice, l'agriculteur risque d'être tenté d'effectuer des dépenses plus importantes afin de satisfaire son confort et non d'épargner pour les années difficiles non prévues par l'indice. Il risque alors de se retrouver ruiner lors d'une année présentant un dommage non prévu.

Ainsi, même dans l'hypothèse où les indemnités et les dommages s'équilibrent sur plusieurs années (voire où les indemnités sont supérieures aux dommages), une telle solution d'assurance n'est pas pertinente pour l'agriculteur. Il ne verrait pas l'argent reçu les mauvaises années comme une façon de couvrir ses risques futurs mais comme la possibilité d'augmenter momentanément son pouvoir d'achat. D'autant que cette hypothèse d'équilibre ne peut être validée comte-tenu du faible historique. En outre, point crucial, une année catastrophique non prévue par l'indice pourrait entraîner sa ruine. L'assureur n'est donc pas protégé. Un indice présentant un fort risque de base n'est donc pas pertinent en termes de risk management. **L'idée de base de l'assurance, à savoir le fait de remplacer le risque de perdre une somme très importante par la certitude de perdre une petite somme (la prime) n'est absolument pas respectée par ce genre d'indice**.

Conclusion

Nous pouvons retenir les points suivants :

– **Les risques sont difficilement mutualisables compte-tenu de leur caractère non indépendant et non homogène. Par ailleurs, le nombre de zones est insuffisant pour une bonne mutualisation. L'assureur devrait diversifier géographiquement ses risques. Par ailleurs, le risque de modèle est élevé dans l'évaluation de la**

	Dommages	Indemnités	Prime pure	Prime de risque	Prime totale
Indice optimisé	1679	2388	79	37	129
Bon indice	1679	1179	39	13	58
Mauvais indice	1679	1517	50	17	74

TABLE 12.4 : Tarification des différents contrats

prime compte-tenu de l'historique limité. Le risque climatique associé à cet indice n'est donc pas véritablement assurable du point de vue actuariel.
- **Les limitations précédentes obligent l'assureur à demander un chargement technique très important, ce qui rend la prime beaucoup trop chère.**
- **Enfin, le risque de base très important rend cette assurance inadaptée pour les agriculteurs.**

Une solution d'assurance rendement basée sur indice climatique apparaît donc peu envisageable compte-tenu des données dont on dispose.

12.2 Ouverture sur la réassurance

Comme on l'a dit, le risque associé à notre indice est difficilement mutualisable. L'assureur peut donc souhaiter se réassurer afin de couvrir le risque d'une perte trop élevée. La réassurance à partir d'un indice objectif est avantageuse. Le risque de base n'intervient plus dans ce cas puisque les pertes de l'assureur correspondent exactement à celles indiquées par l'indice. Néanmoins, les mêmes problèmes de non assurabilité actuarielle apparaissent. Ils sont d'ailleurs plus prononcés compte-tenu du nombre très faible de valeurs extrêmes de l'indice dans l'historique.

Une telle inassurabilité actuarielle est néanmoins moins problématique pour un réassureur compte-tenu de sa grande capacité ainsi que sa diversification géographique à l'échelle de plusieurs pays voire continents.

12.3 Mesures d'adaptation

Compte tenu des difficultés que pose l'assurance agricole, il apparaît nécessaire de limiter les risques par des interventions se situant en dehors du champ de l'assurance. Pour ce faire, la solution la plus évidente consiste à agir de manière préventive pour éviter les sinistres : par exemple améliorer les techniques d'irrigation ou encore semer des variétés plus résistantes à la sécheresse.

Chapitre 13

Autres utilisations potentielles de l'indice : les dérivés climatiques et les cat bonds

13.1 Introduction

Dans le cadre de cette étude, nous avons étudié une solution de couverture de type assuranciel. C'est en effet ce qui avait été demandé par la COSUMAR à AON. Il existe néanmoins un deuxième type de couverture fondamentalement différent et correspondant à une approche économique distincte : l'approche de marché. Il s'agit de se couvrir contre des événements climatiques à l'aide de produits dérivés du même type que ceux évoqués dans la section 7.2.4 page 39.

Comme nous allons le voir, la première semble nettement plus adaptée dans le cas qui nous intéresse ici. Néanmoins, nous analysons également la seconde car elle s'avère conceptuellement très intéressante et utilisable dans le cas général.

La première transaction portant sur un contrat climatique a été conclue entre deux entreprises énergétiques américaines en 1997. Les produits climatiques sont en effet apparus aux Etats-Unis dans le but de permettre aux compagnies énergétiques de se couvrir face au risque de baisse de la consommation lors des hivers doux (moindre utilisation du chauffage) et des étés frais (moindre usage de la climatisation).

Le cadre réglementaire, fiscal et comptable applicable aux contrats d'assurance est très différent de celui gouvernant l'usage des produits dérivés. Un contrat d'assurance permet de transférer un risque encouru par une personne physique ou morale vers un organisme d'assurance. Le paiement de la prime d'assurance est effectué en amont de la période assurée et le montant de l'éventuelle indemnité est en général estimé par un ou plusieurs experts, puis discuté par les deux parties lors de l'éventuel accident. Dans le cas présent, **l'accident correspond à une valeur anormalement basse de l'indice et correspond donc à une grandeur objectivement mesurable**. De ce fait, l'indemnité ne peut être contestée. De surcroît, **la prise en compte des pertes et dommages semble garantie si l'indice est suffisamment représentatif de la production**. Seules les compagnies d'assurance agréées ont le droit de vendre des contrats d'assurance. Un produit dérivé est un instrument financier négociable, qui peut être le support d'opérations de couverture et de "trading" (spéculation). Dans le cadre de cette approche, le paiement final, communément appelé le pay off, survenant à l'échéance du contrat, ne résulte pas des dommages subis mais d'un simple calcul paramétrique dont la formule a été fixée lors de

la signature de la transaction. Il se peut alors que celui-ci soit **assez éloigné de la valeur des pertes et dommages réellement subies**.

Outre cette différence dans le calcul de l'éventuel paiement à l'assuré, la réglementation des contrats est sensiblement différente. En effet, si les contrats d'assurance sont ouverts à tous, de nombreuses restrictions existent pour la vente ou l'achat des produits dérivés. A titre indicatif, certaines entreprises allemandes n'ont tout simplement pas le droit d'acheter ou de vendre des produits dérivés. Par ailleurs, il n'existe pas d'équivalent de la taxe sur les primes d'assurance pour les produits dérivés, du moins en France.

Dans la première approche, l'entreprise considère **la couverture du risque climatique comme un processus d'assurance**. Elle espère ainsi recevoir des indemnités en cas de valeur anormalement basse de l'indice mais **n'a pas le souhait de gérer une position financière au cours du temps**.

Dans la seconde, l'entreprise **ouvre une position de dérivés et gère de manière déconnectée cette position et l'exposition à couvrir**, sachant évidemment que les deux positions varient en sens opposé. Néanmoins, la gestion dynamique de la couverture exige des moyens humains et informatiques importants et est placée sous la responsabilité de la trésorerie qui dispose déjà des compétences techniques dans le domaine de la couverture. **On comprend alors pourquoi ce type de gestion n'est pas adapté dans le cas de la coopérative agricole COSUMAR qui n'a pas l'habitude de ce type de considérations.**

13.2 Organisation du marché

Le marché des produits climatiques est organisé comme tout autre marché financier.

On distingue en amont le **marché primaire**, sur lequel se retrouvent, d'un côté, les acheteurs, c'est-à-dire les entreprises s'assurant contre le risque climatique (end users) au travers de contrats hautement structurés et, de l'autre côté, les vendeurs tels les assureurs, les réassureurs, les banques et encore certaines compagnies énergétiques. Sur ce marché, les contrats sont en général conçus sur mesure et sont donc parfaitement adossés au besoin du client

En aval, on trouve le **marché secondaire** sur lequel les vendeurs de couverture négocient entre eux des contrats standardisés qui leur permettent de gérer dynamiquement leurs portefeuilles climatiques. La maturité des contrats traités sur le marché primaire peut aller de une heure à plusieurs années et porter sur différents sous-jacents tels que la température, la pluie, le vent, la hauteur des vagues ou la neige, ou sur une combinaison de certains de ces facteurs (dans le cas de notre indice par exemple). Comme on le verra par la suite, le marché secondaire se doit d'être liquide et les contrats y sont donc en général standardisés. Ainsi, les contrats ont le plus souvent des durées de cinq jours (du lundi au vendredi), un mois ou cinq mois (période de chauffage de novembre à mars et de climatisation de mai à septembre).

13.3 La gestion des risques courants : utilisation des dérivés climatiques

Les dérivés climatiques sont des produits dérivés financiers ayant comme sous-jacent un indice climatique (température, précipitations ou combinaison de différentes variables climatiques comme dans le cas de notre indice).

Comme nous l'avons dit, ils sont nés pour permettre aux compagnies d'énergie de se couvrir face aux risques d'hivers trop doux et d'étés trop frais. Les premiers étaient basés sur les deux indices suivants :

– Le HDD (Heating Degree Day). Si l'on note $Tmoy$ la température moyenne de la journée, $HDD = (65 - Tmoy)_+$. 65 °Fahrenheit correspond environ à 18 °C. Ainsi, le HDD indique de combien de degrés il est nécessaire de réchauffer un logement.
– Le CDD (Cooling Degree Day) vaut $CDD = (Tmoy - 65)_+$ et correspond au nombre de degrés dont il faut refroidir un logement.

La température est véritablement devenue une matière première négociable, au même titre que les taux d'intérêt, les cours de change, les actions ou les matières premières. S'il n'est évidemment pas possible de l'acheter ou de la vendre sur un marché au comptant, elle se traite à terme, le débouclement du contrat se faisant par cash settlement- règlement de la différence entre le prix d'achat et le prix de vente- et non par physical settlement-livraison physique de l'actif.

Nous étudions ci-dessous les différentes catégories de dérivés climatiques.

Les contrats futures et forward

Les deux contreparties s'engagent à acheter ou vendre un indice à un prix fixé et à une date prédéterminée. Considérons un agent économique, par exemple un producteur d'électricité, exposé au risque de température trop élevée au mois de décembre et ce sur la zone géographique de Paris. Il souhaite alors se couvrir face à la baisse du cumul de HDD. Il peut par exemple vendre à terme le cumul de HDD à un prix de $210°C$, cotation en vigueur au moment de la couverture. Si le cumul de HDD observé au mois de décembre est finalement de $180°C$, il peut alors acheter en quelque sorte le cumul de HDD à un prix de $180°C$, et le revendre à $210°C$, réalisant un gain égal à $(210°C - 180°C)$ multiplié par la valeur en euros de $1°C$. Ce résultat est supposé compenser la perte liée à une réduction de la consommation d'électricité.

Les options sur indice

Dans le cas des options, l'acheteur a le droit et non l'obligation "d'acheter" ou de vendre l'indice.

Ainsi, dans l'exemple précédent, l'entreprise peut acheter un put, c'est-à-dire une option de vente. Si le strike (valeur à laquelle l'entreprise peut vendre) vaut $210°C$ et la valeur observée est de $180°C$, l'entreprise exerce son option, achète à 180 et vend à 210. Si la valeur observée est en revanche de 220, l'entreprise n'est pas obligée d'exercer, c'est-à-dire d'acheter à 210 et de vendre à 220. Si l'option n'est pas exercée, l'acheteur perd le montant de la prime versée mais ceci est compensé par les conditions météorologiques favorables.

Lorsque l'on souhaite se protéger face à la hausse d'un indice, il convient d'acheter un call. A titre d'exemple, la pluie en période estivale peut affecter de manière très significative le chiffre d'affaire des parcs d'attraction. Ainsi, ce dernier peut acheter un call avec strike de 140 mm. Le parc peut alors recevoir un certain montant par millimètre au-dessus du strike. Ceci revient donc à acheter l'indice (donc les précipitations) au prix du strike et à la revendre au prix observé.

Les swaps

Ils sont échangés de gré à gré et ont l'avantage de n'avoir aucun coût.

Par exemple, une entreprise souhaitant se couvrir face au risque d'été trop frais peut vendre un CDD swap. Elle s'engage alors :
– à vendre le cumul de CDD à un prix fixé dans le contrat (strike) ;
– à acheter le cumul de HDD observé.

Ainsi, elle touche un montant pour chaque degré au-dessus de $65°F$ et paie pour chaque degré en-dessous.

Les caps et floors

Il s'agit respectivement de calls et de puts mais dont le gain est plafonné.

Les collars

Ce sont des combinaisons d'options put et call, ce qui permet de limiter à la fois le risque lié au climat ainsi que le risque de perte de la prime. Considérons l'exemple d'un Costless Collar : un industriel peut souhaiter se couvrir face au risque de températures trop élevées nécessitant l'utilisation de climatiseurs. Il peut ainsi acheter pour 750000 euros un call sur HDD avec un strike à 5200 HDD et vendre un put sur HDD avec un strike à 5050 HDD, créant un montage qui ne lui coûte rien initialement. Il recevra alors 10000 euros pour chaque degré au-dessus de 5200 HDD et devra payer 10000 euros pour chaque degré en-dessous de 5050 CDD. Entre ces deux niveaux, il n'y a pas de paiement.

Ces produits sont assez proches des swaps mais bénéficient de davantage de flexibilité sur les marchés.

Les actifs dérivés étudiés ici s'avèrent être des solutions alternatives intéressantes à l'assurance et à la réassurance. Néanmoins, ils ne permettent véritablement qu'une couverture contre les aléas, c'est-à-dire de petites variations et non contre des catastrophes. Il s'agit d'outils de lissage du bilan. Seules les options ayant pour trigger la survenance d'une catastrophe peuvent être utiles dans le cas d'événements extrêmes mais celles-ci ne sont pas très répandues.

L'outil le plus célèbre permettant de transférer le risque catastrophe aux marchés financiers est l'obligation catastrophe ou cat bond, que l'on étudie dans la section suivante.

13.4 La gestion des risques extrêmes : les cat bonds

Le cyclone Andrew qui a dévasté la Floride en 1992 a souligné la nécessité d'un grand changement dans la gestion des catastrophes naturelles. En effet, les marchés de l'assurance et de la réassurance n'ont pas le capital nécessaire pour couvrir de telles catastrophes. C'est dans ce contexte que sont nés les cat bonds ou obligations catastrophes, en vue de leur permettre de transférer ce risque aux marchés financiers.

Le principe est le suivant. L'entreprise cédante du risque (compagnie d'assurance ou de réassurance) émet (directement ou indirectement via un véhicule ad hoc : un SPV, ou Special Purpose Vehicle) une dette obligataire dont le remboursement est conditionné à la survenance d'un événement donné. Différents cas sont envisageables :

- nominal protégé/ coupons à risque ;
- nominal à risque/ coupons protégés ;
- nominal et coupons à risque.

13.5 Application au cas de notre indice

En théorie, il serait envisageable d'appliquer les principes énoncés précédemment au cas de notre indice.

La coopérative souhaite se couvrir face à une baisse de l'indice. Elle pourrait donc être tentée d'acheter un put ayant pour sous-jacent notre indice. Si l'indice descend sous le strike, les agriculteurs gagnent ainsi la différence, ce qui vient compenser en partie les pertes de rendement. Cela suppose néanmoins que l'indice soit coté. Supposons que l'indice 2011, en raison des tendances climatiques à long terme, soit coté à 0. Ainsi, l'agriculteur (plus exactement la coopérative) a la possibilité de le vendre à terme à ce niveau. Si la valeur réelle de l'indice est finalement de -0.3, il l'achète au cours au comptant à -0,3 et le revend en exerçant son contrat à terme au prix de 0. Il gagne donc au final 0.3 points d'indice, ce qui vient compenser sa perte.

Néanmoins, **seuls les contrats liquides et donc standardisés sont susceptibles d'être cotés**. En effet, il faut que deux parties sur le marché désirant couvrir des risques opposés se rencontrent et souscrivent au contrat. Le cas présent, comme on l'a vu fait intervenir un indice extrêmement complexe et nécessite donc un produit sur mesure. Il est très délicat de trouver une contrepartie désirant couvrir des risques opposés. La coopérative doit alors se tourner, si elle souhaite souscrire des contrats de type dérivés climatiques, vers des banques ou compagnies d'assurance (marché primaire). Contrairement aux transactions sur le marché secondaire, elle doit alors payer une prime. Par ailleurs, on a déjà dit que le **dédommagement engendré par un dérivé n'était pas forcément très fiable**. Ceci est particulièrement vrai dans le cas d'un indice non parfaitement corrélé avec les rendements réels. **Il y a alors un risque de corrélations très importants.** Le cas d'une anticorrélation est dramatique puisque toute couverture sur le principe précédent s'effondre.

Finalement, on peut dire que l'approche assurancielle semble la plus adaptée au cas de la coopérative COSUMAR.

Il est enfin possible que la compagnie d'assurance souscrivant le contrat avec la COSUMAR désire se couvrir face aux risques extrêmes en émettant un cat bond. Du point de vue conceptuel, on pourrait imaginer une telle obligation catastrophe indexés sur notre indice. Celle-ci pourrait être définie par l'absence de remboursement dès que l'indice est inférieur à un certain seuil extrême. Néanmois, comme on l'a dit, notre indice n'est pas très efficace pour représenter les risques extrêmes et ne permettrait donc pas de donner naissance à ce genre de produits.

Chapitre 14

Utilisation des données de l'ECMWF ainsi que des sorties du modèle LMDZ

14.1 Brève présentation

L'ECMWF (European Center for Medium-Range Weather Forecasts) est une organisation internationale indépendante supportée par 31 Etats, dont les principaux objectifs sont :

- Le développement de méthodes numériques pour les prévisions à moyen terme ;
- L'élaboration, sur une base régulière, de prévisions météorologiques à moyen terme, notamment à destination des services météorologiques nationaux des Etats membres ;
- Des recherches scientifiques et techniques ayant pour but d'améliorer ces prévisions ;
- Le stockage des observations météorologiques appropriées .

C'est le dernier point qui nous intéresse plus particulièrement dans le cadre de notre étude. L'ECMWF effectue en continu, grâce à son important réseau d'observations, des analyses et réanalyses de l'état de l'atmosphère. Les réanalyses sont en fait les analyses corrigées. Elles fournissent les valeurs des principales grandeurs climatiques (température, précipitation, vent) aux points d'une grille dont la résolution est de l'ordre de 100km.

Nous présentons un exemple de carte d'analyse sur la figure 14.1.

Le modèle LMDZ est le modèle climatique développé au sein de l'Institut Pierre Simon Laplace (IPSL). Il fournit les valeurs des grandeurs climatiques sur chaque cellule. Sa particularité est sa capacité à augmenter sa résolution sur une région donnée. La taille des cellules est alors de l'odre de 30 km. Des précisions au sujet de ce modèle ainsi que de l'IPSL sont données en annexes B.1 page 118.

14.2 Utilisation

Une première idée serait de tester la fiabilité des données in situ en les comparant aux analyses et réanalyses de l'ECMWF ainsi qu'aux sorties du modèle LMDZ. Ceci s'avère néanmoins impossible car ces données sont de nature très différente. En effet, les analyses de l'ECMWF répertorient les valeurs des grandeurs climatiques au points d'une grille. Ceux-ci ne correspondent généralement pas à la localisation des stations géographiques d'où proviennent les données in situ

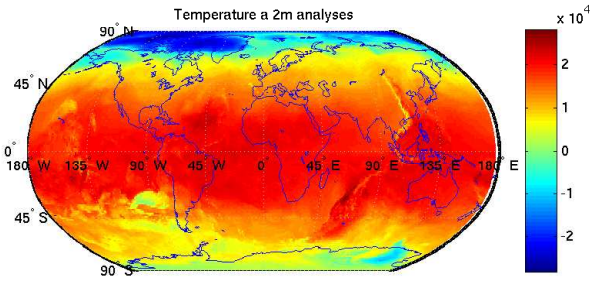


FIGURE 14.1 : Température à 2m moyenne en janvier 2003 issue des analyses de l'ECMWF

et les tendances des différentes variables peuvent donc différer assez largement. Ceci n'est pas trop problématique dans le cas des températures car celles-ci sont en général spatialement très corrélées. En revanche, les précipitations en deux sites très rapprochés peuvent avoir un comportement très différent (du fait de caractéristiques eurographiques spécifiques). Les précipitations ne varient pas spatialement de manière linéaire et donc même une interpolation bilinéaire entre les ponts de grille pour se ramener aux coordonnées effectives des stations n'est pas satisfaisante.

De même, les sorties de modèle fournissent une seule valeur des différentes variables par cellule, s'interprétant comme une moyenne. On illustre la comparaison entre les données in situ et celles issues du modèle dans la figure 14.2

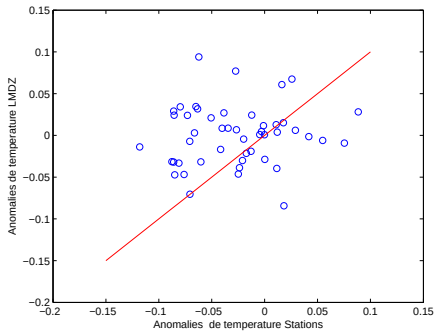


FIGURE 14.2 : Comparaison des anomalies de température à Berkane observées et prévues(LMDZ)

On constate donc que les données in situ et celles issues de LMDZ ne présentent une corrélation que d'environ 0,5. Les sorties du modèle présentent néanmoins une plus faible variabilité. Ceci est général car les modèles climatiques ont tendance à lisser les grandeurs.

Ces deux types de données peuvent néanmoins s'avérer très utiles :
- On peut souhaiter en déduire les valeurs des variables climatiques dans les stations. Ceci serait possible grâce aux techniques de downscaling si les stations météorologiques formaient une grille régulière.
- Il est également possible que les données de stations ne soient pas les plus pertinentes dans le cadre de cette étude. En effet, comme on l'a dit, les précipitations peuvent avoir un comportement très différent en des lieux très rapprochés. Ainsi, dans une même zone, les anomalies peuvent largement différer d'une exploitation agricole à l'autre et être très éloignées de celles mesurées à la station de référence. En effet, les stations ne sont pas toujours à proximité des exploitations. On comprend ainsi qu'une sortie de modèle qui représente une moyenne sur une cellule pourrait être plus pertinente car plus représentative de la véritable anomalie de précipitation à l'échelle de la zone. On parle alors d'upscaling puisque l'on considère des données à l'échelle du modèle plutôt qu'à l'échelle locale.

Les techniques de downscaling et de upscaling sont présentées dans la section suivante.

14.3 Techniques de downscaling et upscaling

14.3.1 Downscaling

Il est possible de distinguer deux catégories de techniques de downscaling : le downscaling physique et le downscaling statistique. Le modèle LMDZ zoomé/guidé appartient à la première catégorie.

Downscaling statistique

L'idée principale de cette technique repose sur le fait que le climat régional est majoritairement conditionné par deux facteurs :
- la circulation à grande échelle, bien résolue par les modèles ;
- les caractéristiques de petite échelle, telles l'eurographie, la répartition des étendues d'eau, etc...

Ainsi, on établit dans un premier temps **un lien empirique entre l'information de grande échelle (variables explicatives) et les variables locales ou régionales (variables prédites), valable pour le climat actuel**. Par application de cette relation empirique, il est alors possible de déduire des conditions de grande échelle simulées par un AOGCM (Atmosphere-Ocean General Circulation Model), les variables locales relatives au climat futur. Cette approche est donc basée sur une hypothèse très forte selon laquelle la relation empirique établie pour le climat actuel restera valable dans le futur sous des conditions climatiques différentes. Cette **hypothèse de stationnarité** constitue d'ailleurs la lacune théorique principale de ces techniques de downscaling statistique car elle ne s'avère pas vérifiable. Comme on peut le voir dans l'annexe B.1 page 118, la paramétrisation des modèles est également soumise à ce problème. L'un des objectifs de LMDZ était justement de rendre celles-ci robustes en introduisant le plus de physique possible.

Comme on vient de le voir pour les deux types de downscaling, on parle de conditions de grande échelle et de conditions locales. Ceci permet d'intuiter la contrainte de ces méthodes qui est la nécessité de posséder deux grilles de résolutions différentes :
- **Une grille basse résolution qui correspond à la grande échelle ;**
- **Une grille haute résolution qui correspond à l'échelle locale .**

Or, comme nous l'avons vu, les données in situ ne proviennent que de quelques stations ponctuelles, largement insuffisantes pour la constitution d'une grille haute résolution "locale". Il est donc impossible d'appliquer quelque méthode que ce soit pour retrouver les données de stations à partir des analyses ou des sorties de LMDZ. En revanche, on voit qu'il est possible de pratiquer un downscaling statistique de la grille des analyses vers celle du modèle LMDZ. Nous expliquerons l'intérêt d'une telle démarche par la suite.

Nous présentons deux techniques envisageables en vue d'effectuer ce downscaling :
- Une simple régression ;
- Une analyse par EOF (analogue à l'Analyse en Composantes Principales) .

L'analyse par EOF

Les données climatiques sont réparties en espace et en temps. Dans le cas qui nous intéresse, nous possédons une donnée par mois. Nous nous intéressons dans un premier temps à l'échelle haute résolution. L'information contenue dans un tel champ est très lourde à étudier et nous nous proposons donc de la réduire à quelques paramètres. D'ailleurs, il apparaît clair que seules les tendances principales pourront être déduites des conditions de grande échelle. Cela ne pose donc pas de problème dans le cadre du downscaling. L'idée est de déterminer des cartes (les EOFS) qui maximisent la variance du champ spatio-temporel $X(x,t)$. On a alors :

$$X(x,t) = \sum_k a_k(t) E_k(x)$$

Ceci permet de réduire la dimensionnalité des données et de séparer les mécanismes indépendants. En pratique, on ne garde que les K premières EOFS offrant un pourcentage de variance expliquée suffisant. Le calcul des EOFS est traditionnel.

L'idée du downscaling est alors de calculer les coefficients temporels $a_k(t)$ à partir des données de l'échelle basse résolution, par minimisation de la fonction de coût (le bon jeu de coefficients est celui qui minimise l'écart quadratique entre les observations et les estimations par EOF aux points de la grille basse résolution). Connaissant les conditions de grande échelle, il est alors possible d'en déduire les conditions locales.

Downscaling physique

Il s'agit d'une interpolation physique. Plutôt que de relier petite et grande échelle par une relation statistique (ce qui est fait dans la partie 14.3.1), on le fait de manière physique. Le zoom du modèle LMDZ correspond exactement à ce type de downscaling (voir annexe B.1 page 118) . Les valeurs aux bords du domaine sur lequel on effectue le zoom sont en fait contraintes par les observations de l'ECMWF. Il convient de noter que le zoom, c'est-à-dire l'augmentation de résolution, se fait de manière assez brutale. On passe en fait d'une résolution d'une centaine de km à une résolution de l'ordre d'une trentaine de km sur une faible distance (environ un millier de km). Ainsi, la résolution de la zone étudiée (sur laquelle on effectue le zoom) est à peu près homogène, ce qui facilite les calculs ainsi que l'utilisation.

14.3.2 Upscaling

L'idée de l'upscaling est de prendre en compte des données intégrées spatialement. Il convient alors d'utiliser une moyenne des valeurs des pixels de LMDZ zoomé/guidé chevauchant une zone donnée voire un périmètre donné. Comme on l'a dit, il est tout à fait envisageable que de telles données soient beaucoup plus représentatives que les données de stations et donc que le résulat soit meilleur, du fait d'entrées de qualité nettement meilleure. Néanmoins, étant donné

que l'on considère des données moyennées mensuellement, les valeurs correspondant à deux pixels consécutifs sont très proches. De ce fait, il revient quasiment au même de ne considérer qu'un seul pixel (ou cellule). Par ailleurs, ces pixels étant rectangulaires, ils ne recouvrent pas précisément les périmètres d'étude. On peut donc les choisir de manière un peu approximative.

Par ailleurs, le modèle LMDZ fournit les valeurs des principales variables climatiques dans le futur. Il s'agit d'un point crucial car cela signifie que de telles études d'impact peuvent être effectuées sur les prochaines décennies grâce aux sorties du modèle. **Ainsi, il est envisageable de travailler en changement climatique et donc de pouvoir étudier l'impact de ce dernier.**

14.4 Indice sur différents inputs

Nous avons fait tourner le modèle LMDZ sur le Maroc sur la période 2002/2006 (les calculs sont très longs) et avons appliqué les anomalies obtenues à notre indice optimisé de la section 11.6 page 85. Ceci donne des résultats très proches (corrélation d'environ 0.3) et le upscaling semble bénéfique. Néanmoins, il faut rester très prudent compte-tenu du très faible nombre d'années. Il serait également utile de chercher la meilleure combinaison de prédicteurs en utilisant ces anomalies provenant de LMDZ.

Même s'il reste encore du travail, on peut donc présupposer que cette technique de upscaling permis par le modèle LMDZ fournit des résulats intéressants. Elle permet par ailleurs une approche en changement climatique très utile.

Or les calculs de LMDZ sont très coûteux car il s'agit d'une méthode de downscaling physique et donc d'une interpolation physique des données. L'idée est alors de retrouver les sorties de LMDZ à partir des données d'analyses et de réanalyses grâce à un downscaling statistique. Les données obtenues sont alors utilisées pour élaborer le modèle d'impact. Cette idée est très intéressante mais nous n'avons pas encore eu le temps de la mettre en oeuvre.

Chapitre 15

Conclusion

15.1 Du point de vue scientifique

L'agriculture marocaine a entrepris les réformes nécessaires en vue de s'adapter au nouveau contexte mondial. Ceci passe par des moyens permettant aux agriculteurs de se prémunir notamment face aux risques climatiques, dans un contexte où les épisodes de sécheresse sont de plus en plus fréquents. Néanmoins, les problèmes d'antisélection et d'alé moral ainsi que les coûts élevés de gestion des sinistres rendent les solutions traditionnelles d'assurances rendement difficilement applicables. Une solution intéressante permettant de pallier ces problèmes consiste à proposer une assurance basée sur un indice indirect, par exemple un indice climatique.

Nous avons testé la possibilité de mettre en place ce type de couverture dans le cas de la production sucrière marocaine. Nous en tirons les enseignements suivants :

- La phase de prétraitement des données est absolument cruciale.
- Il est difficile de mesurer la performance réelle de l'indice. En effet, il est crucial d'éviter le piège des scores en sur-apprentissage. Il faut évaluer la capacité de généralisation du modèle.
- Il est très délicat d'obtenir une bonne capacité de généralisation lorsque l'on dispose de peu de données, comme c'est le cas ici. Il convient alors de mettre en place des méthodes de régularisation. En particulier, le réseau de neurones ne donne pas de meilleurs résultats que la régression linéaire. Au mieux, bien régularisé, il est équivalent.
- La méthode de choix des variables prédictives apparait satisfaisante.

L'indice climatique optimal possède une corrélation de 0,31 avec le rendement en sucre. Le risque de base est donc très élevé. Outre ceci, le risque apparaît difficilement assurable du point de vue actuariel. Une assurance basée sur cet indice n'est donc pas envisageable. Néanmoins, nous avons la satisfaction **d'avoir mis en place une méthodologie permettant d'obtenir le résultat optimal et de tester la performance réelle de l'indice.**

Il est également intéressant de mentionner le développement de produits financiers basés sur de tels indices et permettant de transférer le risque climatique aux marchés financiers. On utilise les dérivés climatiques pour couvrir les aléas courants et les cat bonds dans le cas des risques extrêmes.

Par ailleurs, comme on l'a vu, l'utilisation des analyses atmosphériques ainsi que des sorties de modèle climatiques peuvent être très intéressantes. Les dernières permettent notamment d'étudier l'évolution future dans un contexte de changement climatique.

15.2 Du point de vue personnel

Au-delà de l'aspect purement scientifique, ce stage m'a véritablement beaucoup appris en terme organisationnel. En effet, il s'agissait de ma première véritable expérience sur un sujet d'assez grande ampleur. J'ai pu tirer un grand nombre de leçons relatives à la gestion de projet. Cette étude nécessitait en théorie un nombre extrêmement important d'optimisations. Dans un premier temps, j'ai parfois eu tendance à vouloir explorer l'ensemble des possibilités afin d'obtenir **le** résultat optimal. Néanmoins, j'ai progressivement pris conscience qu'une telle démarche était délicate et qu'il fallait parfois accepter de faire des hypothèses sans chercher à les vérifier. La science est faite d'hypothèse et il n'y a donc pas toujours une solution déterministe meilleure que les autres.

Par ailleurs, il s'est avéré très enrichissant du point de vue humain : j'ai été accueilli par des chercheurs ouverts, disponibles et heureux de partager leurs connaissances. Les contacts se sont donc toujours révélés d'une grande sincérité.

Enfin, passionné par les questions climatiques depuis mon plus jeune âge, j'ai réellement été ravi de travailler sur un tel sujet.

Bibliographie

[1] Aires, F. (1999) *Problèmes inverses et réseaux de neurones : application à l'interféromètre haute résolution IASI et à l'analyse des séries temporelles, Rapport de thèse, Université Paris IX (Dauphine).*

[2] Bishop, C. (1996) *Neural networks for pattern recognition.* Clarendon Press - Oxford.

[3] Bielza Dia-Caneja, M. et Conte, C.G. et Gallego Pinilla, F.J. Stroblmair, J. et Catenaro, R. et Dittman, C. (2009) *Risk Management and Agricultural Insurance Schemes in Europe.* JRC Reference Reports.

[4] Botou, L. (1991) *Une approche théorique de l'apprentissage connexioniste ; application à la reconnaissance de la parole.* PhD thesis, Université d'Orsay.

[5] Brix, A. et Jewson, D. et Ziehmann, C. (2005) *Weather Derivative Valuation.* The Meteorological, Statistical, Financial and Mathematical Foundations. Cambridge.

[6] Carle, J. et Fourneaux, S et Holz, R. et Marteau, D. et Moreno, M. (2004) *La gestion du risque climatique.* Collection Gestion. Economica.

[7] Douira, A. et Mzibra, A. et Zehauf, M. (2007) *Effet du cycle de la culture sur le rendement qualitatif et quantitatif de la betterave sucrière dans la région du Gharb(Maroc).*

[8] Friedman, J.H. (1994) *An overview of predictive learning and function approximation.* From Statistics to Neural Networks. NATA ASI Seris.

[9] Geman, S. et Bienenstock, E. et Doursat, R. (1992) *Neural networks and the bias-variance dilemma.* Neural Computation, 1(4) :1-58.

[10] Hazell, P. et Anderson, J. et Balzer, N. et Hastrup Clemmensen, A. et Hess, U. et Rispoli, F. (2010) *L'assurance basée sur un indice climatique : potentiel d'expansion et de durabilité pour l'agriculture et les moyens de subsistance en milieu rural.* Fonds international de développement agricole et le Programme alimentaire mondial.

[11] Hertz, J. et Kroght, A. et Palmer, R. (1992) *Introduction to the theory of neural computation.* Cambridge Press.

[12] Hourdin, H. (2005) *Représentation du transport direct et inverse dans les modèles globaux de climat et étude des couplages entre composition et dynamique atmosphérique sur Titan.* Mémoire présenté pour obtenir une Habilitation à Diriger des Recherches.

[13] Hull, J. (2007) *Options, futures et autres actifs dérivés.* 6ème édition.

[14] Toumi, L. (2008) *La Nouvelle Stratégie Agricole au Maroc (Plan Vert) : les Clés de la Réussite*

[15] Mc Culloch, W. et Pitts, W. (1943) *A logical calculus of the ideas immanent in nervous activity.* Bulletin of mathematical biophysics, 5 :115-133.

[16] Pernot, E. (1992) *Choix d'un classifieur en discrimination.* PhD thesis, Université de Jussieu (Paris).

[17] Vallet, F. (1990) *Approche globale du problème de discrimination : aspects probabilistes.* Revue Technique Thomson, 22(1) :519-541.

[18] Vapnik, V. (1997) *The nature of statistical learning theory.*

Annexe A

Données climatiques in situ et données de rendements

A.1 Données climatiques

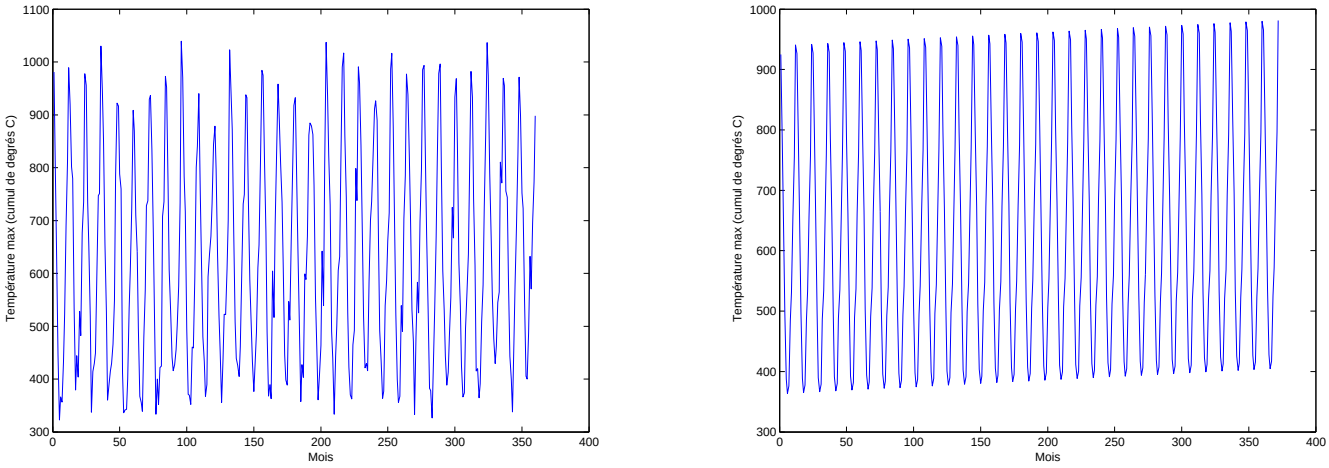


FIGURE A.1 : Série de températures maximales mensuelles (g) et tendance associée (avec saisonnalité) (d) dans la zone de Ben Amir

A.2 Données de rendements

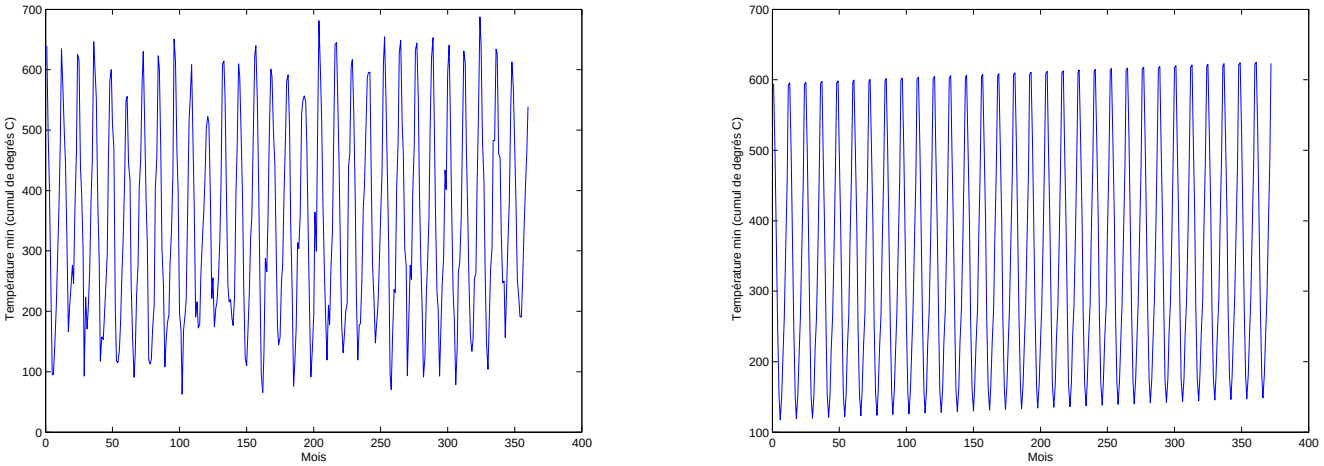


FIGURE A.2 : Série de températures minimales mensuelles (g) et tendance associée (avec saisonnalité) (d) dans la zone de Ben Amir

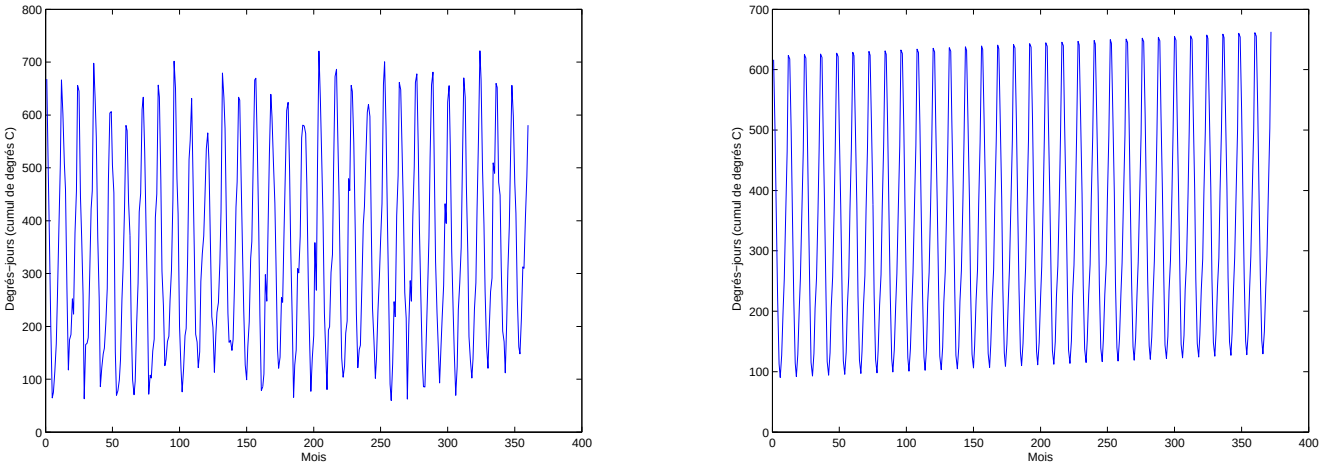
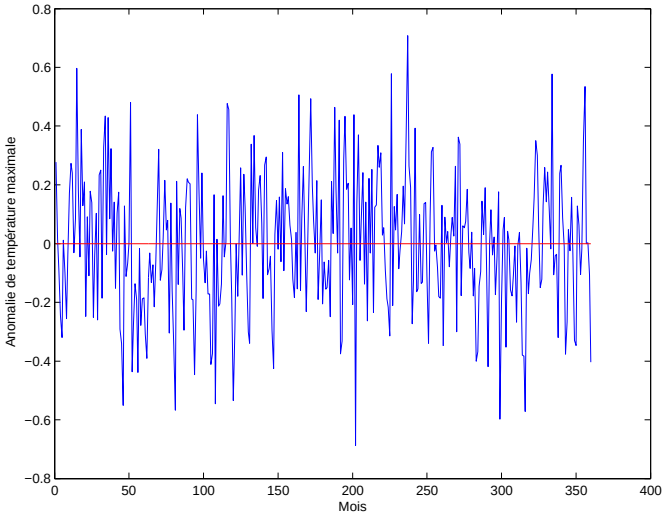


FIGURE A.3 : Série de degrés-jours mensuels (g) et tendance associée (avec saisonnalité)(d) dans la zone de Ben Amir



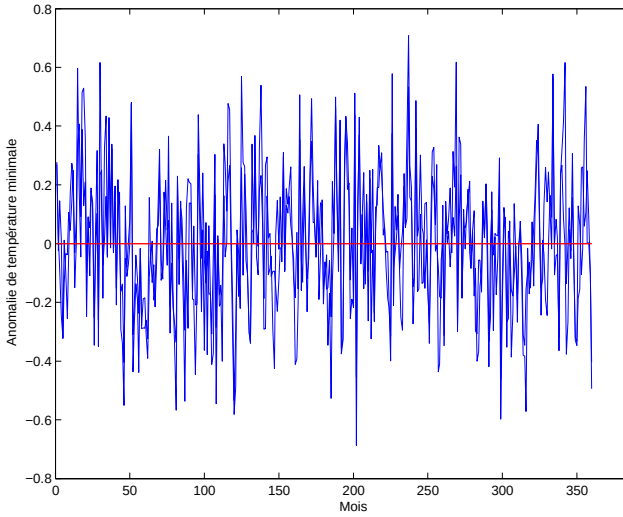


FIGURE A.4 : Séries des anomalies de températures maximales et minimales mensuelles dans la zone de Ben Amir

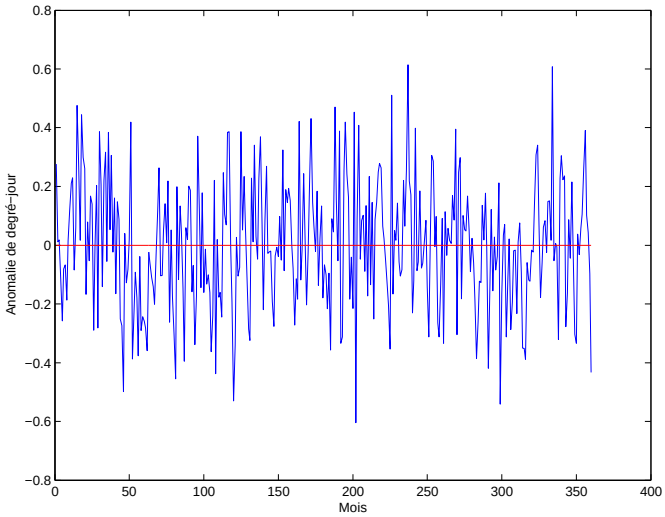


FIGURE A.5 : Série des anomalies de degrés-jours mensuels dans la zone de Ben Amir

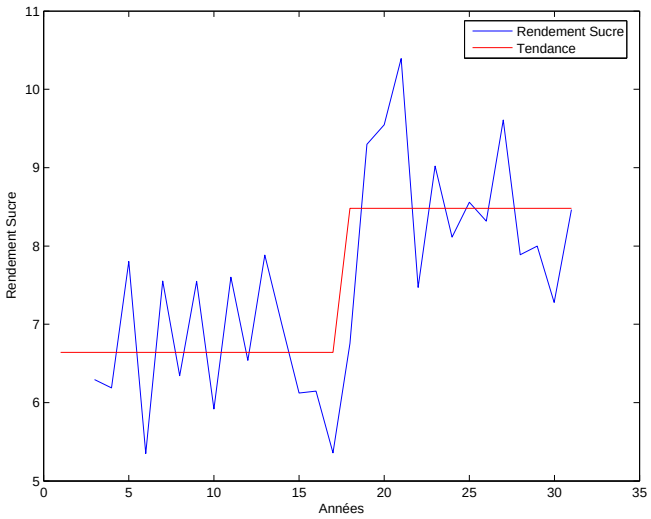


FIGURE A.6 : Rendement sucre et tendance associée pour la zone de Bamir

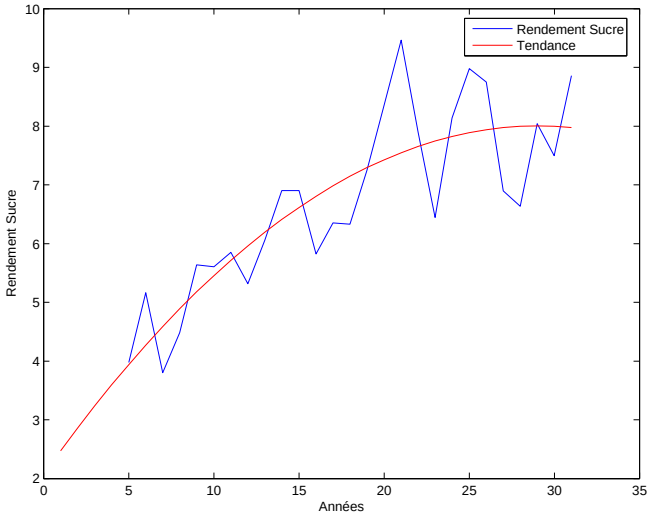


FIGURE A.7 : Rendement sucre et tendance associée pour la zone de Berkane

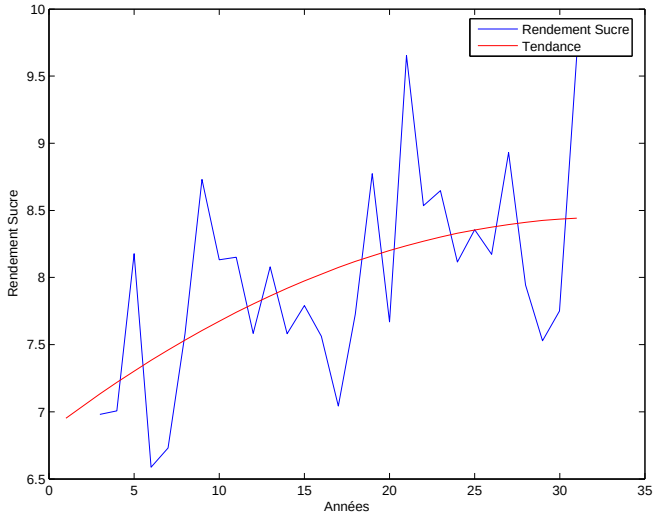


FIGURE A.8 : Rendement sucre et tendance associée pour la zone de Bmoussa

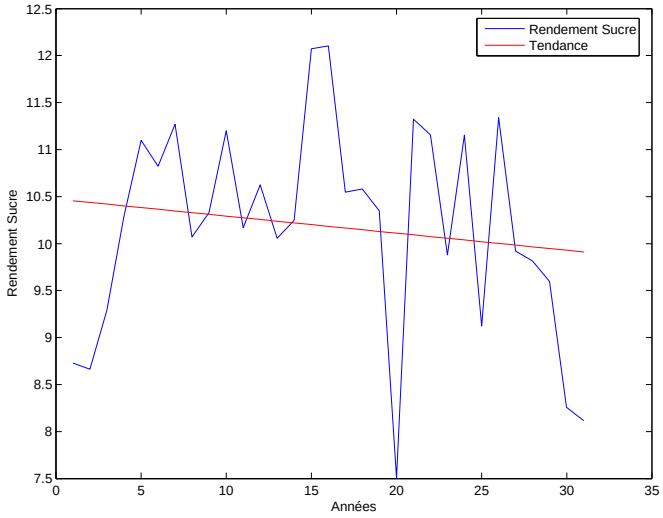


FIGURE A.9 : Rendement sucre et tendance associée pour la zone de Doukkala

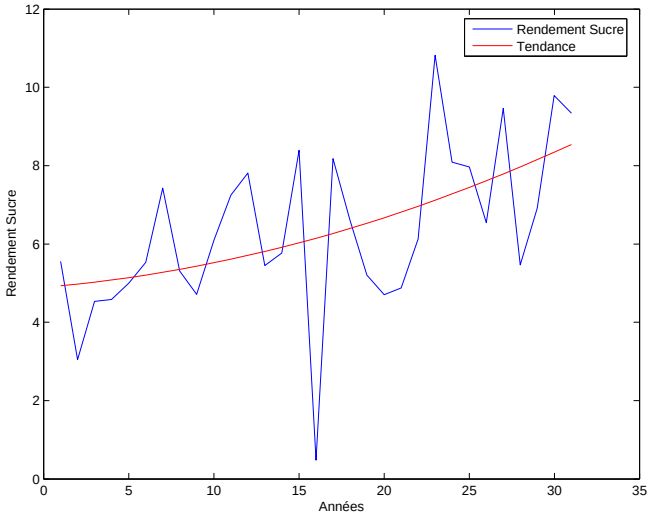


FIGURE A.10 : Rendement sucre et tendance associée pour la zone de Loukkos

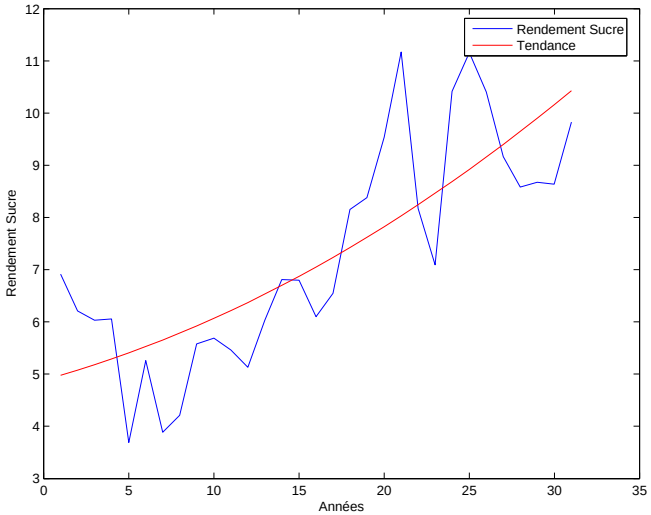


FIGURE A.11 : Rendement sucre et tendance associée pour la zone de Moulouya

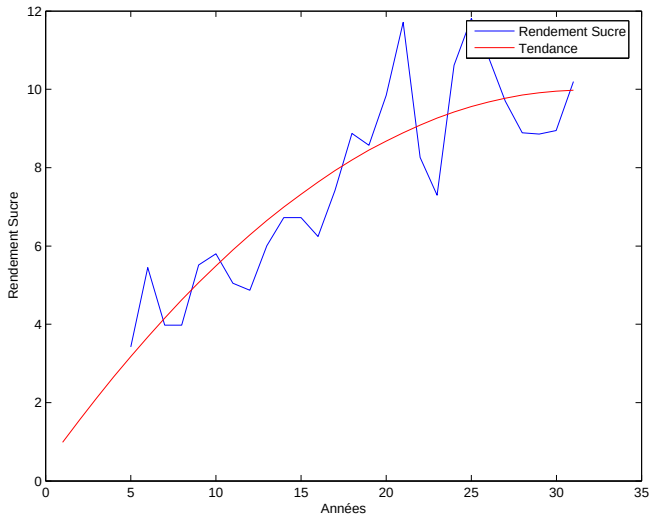


FIGURE A.12 : Rendement sucre et tendance associée pour la zone de Nador

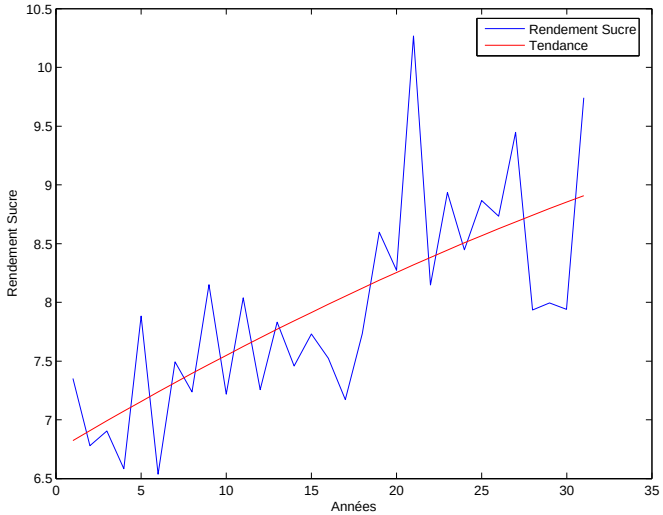


FIGURE A.13 : Rendement sucre et tendance associée pour la zone de Tadla

Annexe B

Données de modèles

B.1 Présentation du modèle LMDZ zoomé/guidé

Avant de présenter le modèle de climat en lui-même, il convient de situer le Laboratoire de Météorologie Dynamique dans son contexte. Nous débutons donc par une rapide description de l'Institut Pierre-Simon Laplace (IPSL).

B.1.1 Le Laboratoire de Météorologie Dynamique au sein de l'IPSL

L'IPSL est une fédération de recherche qui regroupe aujourd'hui 6 laboratoires et représente environ 750 personnes (280 chercheurs et enseignants-chercheurs, 240 ingénieurs, techniciens et agents administratifs et 230 non permanents - doctorants, post-doctorants et CDD), soit plus d'un tiers du dispositif national de recherche du CNRS et des universités dans le domaine des sciences de l'océan et de l'atmosphère.

L'IPSL est sous la tutelle conjointe de 4 établissements d'enseignement supérieur (université Pierre et Marie Curie, UVSQ, Ecole Polytechnique, Ecole Normale Supérieure), du Centre National de la Recherche Scientifique (CNRS), du Commissariat à l 'Energie Atomique (CEA) , du Centre National d'Etudes Spatiales (CNES) et de l'Institut de Recherche pour le Développement (IRD).

Il possède sous son égide cinq laboratoires que sont :
- le centre d'étude des Environnements Terrestres et Planétaires (CETP) ;
- le Laboratoire de Météorologie Dynamique (LMD) ;
- le Laboratoire d'Océanographie et de Climatologie : Expérimentations et Approche Numérique (LOCEAN) ;
- le Laboratoire des Sciences du Climat et de l'Environnement (LSCE) ;
- le Service d'Aéronomie (SA) .

Le Laboratoire de Météorologie Dynamique se décompose en trois groupements selon les problématiques traitées :
- modélisation du climat et impacts sur le site de Jussieu ;
- dynamique des fluides et turbulence à l'ENS Ulm ;
- instrumentation et télé-détection satellite à l'Ecole Polytechnique .

B.1.2 Le modèle de climat de l'IPSL

Le modèle de l'IPSL contient essentiellement trois composantes couplées :

– le modèle d'océan NEMO ;
– le modèle de surface Orchidée ;
– le modèle d'atmosphère LMDZ4 ;

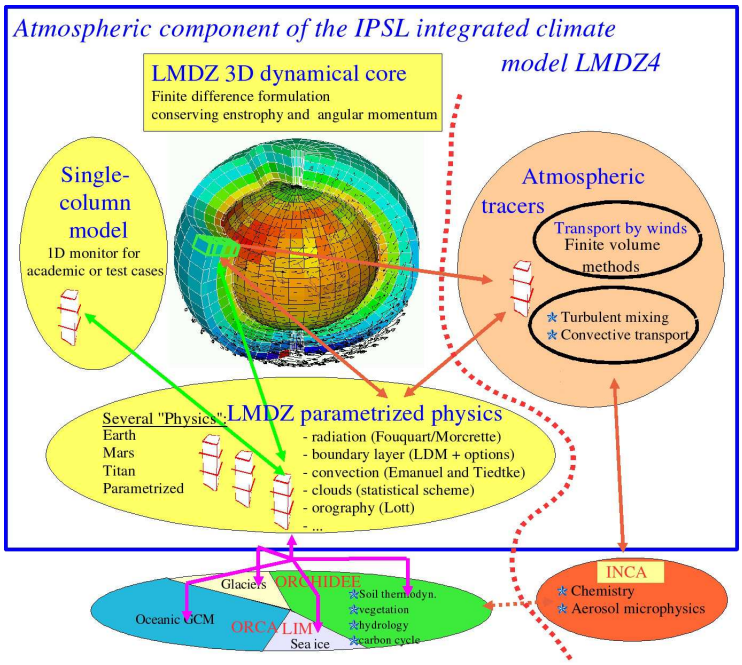


FIGURE B.1 : Modèle LMDZ

Le module **NEMO** modélise la dynamique océanique. Il est développé au LOCEAN.

Orchidée s'intéresse à la partie physique des **échanges avec le sol**. Il rend compte de la température de surface, des flux de surface (flux de chaleur latente et flux de chaleur sensible) ainsi que de l'évaporation. Il se décompose en deux sous-modules que sont **Sechiba** et **Orchidée**.

Sechiba fait intervenir un modèle de sol à 2 couches. Celui-ci est véritablement décrit dans sa profondeur mais la repésentation ne fait intervenir que deux niveaux. Il s'agit par exemple de déterminer si la pluie tombée reste en surface (et peut alors contribuer à l'évaporation) ou au contraire alimente des couches plus profondes (nappes phréatiques par exemple). Cette modélisation à 2 couches n'est néanmoins pas très satisfaisante. Des essais ont été réalisés sur un modèle à 11 couches et il s'avère que les améliorations sont très nettes.

Orchidée, le deuxième sous-module, modélise :
– le routage des pluies. Il analyse en effet l'identité des différents bassins dans lesquelles celles-ci se déversent (rivières, fleuves, mers) ainsi que les quantités mises en jeu. Ces grandeurs sont nécessaires lorsque l'on effectue des simulations à long terme puisqu'il convient de fermer le cycle de l'eau.
– le comportement du couvert végétal : l'eau alimente les racines des plantes, favorise par là-même leur respiration et donc l'évaporation. On constate de manière générale que l'évaporation est importante au-dessus des forêts.

Notons que celui-ci n'est pas activé lorsque l'on effectue des simulations à court terme. En effet, la dynamique de l'évolution du couvert végétal et des bassins possède un temps caractéristique assez élevé. On peut donc la négliger pour des simulations à quelques années. Remarquons que

la partie hydrologique (déversement dans les différents bassins, pénétration de l'eau dans le sol), est gérée par le LMD alors que les recherches au sujet du couvert végétal se font majoritairement au LSCE. Le couplage entre les différentes composantes intervenant dans un système climatique revêt une telle importance que l'on comprend l'utilité de partenariats très forts. Il s'agit d'ailleurs de l'un des avantages principaux de l'IPSL que de rassembler en son sein un très large panel de compétences issues de laboratoires en interaction étroite.

Le modèle **LMDZ4** comporte :

- Un **noyau dynamique**, qui résout les équations de la dynamique des fluides (Navier-Stokes) sur l'horizontale. On utilise en général des méthodes de type différences finies ou spectrales. Des filtres permettant de stabiliser les équations ont été mis au point. Ce noyau fournit les valeurs de vitesses horizontales et de température.
- Un **noyau physique**, dans lequel on trouve les paramétrisations. La ligne directrice du LMD est de mettre au point les paramétrisations les plus physiques possibles. Dans la plupart des modèles développés ailleurs, les paramétrisations effectuées sont statistiques. Or les statistiques peuvent évoluer alors que les lois physiques sont invariantes par translation dans le temps. L'approche "physique" appararaît donc beaucoup plus pertinente en vue d'une étude en changement climatique, objectif central du LMD. A titre indicatif, la représentation du rayonnement se doit d'être beaucoup plus physique dans un modèle de climat que dans un modèle de météorologie. En effet, le modèle de climat est libre une fois la condition initiale imposée alors que le modèle de météorologie est sans arrêt rappelé et guidé par les observations.

Les différentes paramétrisations concernent :

- **Les phénomènes de couches limites** : on appelle couche limite la zone au sein de laquelle les conditions de surface ont des conséquences. Elle est principalement définie en termes de température. Les flux de chaleurs étant nettement plus importants pendant la journée que la nuit, elle s'avère beaucoup plus épaisse en journée. Il ne faut pas la confondre avec la couche de surface (friction du vent, adhérence de l'air à la croûte terrestre). De toute façon, les 10 ou 20 premiers mètres de l'atmosphère ne sont pas pris en compte. Dans la couche limite sont principalement étudiés les phénomènes de turbulence et de diffusion verticale. Auparavant, seuls les phénomènes diffusifs étaient pris en compte au sein d'une même couche. On s'est aperçu qu'il existait des mouvements organisés dans le sens contraire au gradient. D'où un nouveau schéma de couche limite : le modèle en flux de masse. A la partie diffusive vient s'ajouter une partie convective organisée. On parle de convection peu profonde. A noter que la résolution verticale comporte 11 niveaux. On peut en trouver plusieurs dizaines dans certains modèles.
- **La convection profonde**, c'est-à-dire entre les différentes couches. On peut alors se poser la question du lien qu'elle peut entretenir avec la convection peu profonde. Il se peut en réalité que l'on modélise deux fois le même phénomène. Parfois, dans les modèles climatiques, certains processus sont représentés plusieurs fois alors que d'autres ne sont pas du tout pris en compte. La convection profonde agit sur la température et l'humidité de l'atmosphère, condense l'eau et donne de l'eau. En effet, l'air chaud et humide, plus léger, a tendance à monter dans les couches supérieures. C'est ce phénomène qui est à l'origine de la convection. Mais au fur et à mesure de sa montée, il se refroidit et la pression de vapeur saturante diminue. Quand celle-ci s'égalise avec la pression partielle de l'eau à l'état de vapeur, cette dernière finit donc par se condenser, ce qui entraîne la formation de masses nuageuses.
- **Les nuages** : La taille des gouttes permet en partie d'expliquer pourquoi des précipitations se déclenchent ou non. Celles-ci peuvent être de plus ou moins grande taille en fonction de la quantité d'aérosols. Typiquement, une goutte de taille importante peut se former

autour d'une impureté. Mais pour représenter très finement ce genre de phénomènes, il conviendrait de posséder un modèle de micro-physique plus précis que le modèle actuel. La taille des gouttes est donc importante car influe sur les précipitations et donc sur le cycle de l'eau. Elle impacte par ailleurs le rayonnement que nous étudions par la suite.

– L'interaction du **rayonnement** avec l'atmosphère. On distingue le rayonnement visible et infrarouge. En effet, la terre et l'atmosphère étant des corps à une certaine température, ils émettent un rayonnement infrarouge selon la loi de Planck et Stefan. La Terre étant en équilibre, Il y a égalité au sommet de l'atmosphère entre rayonnement incident et émergeant. La paramétrisation du rayonnement demande un temps de calcul très important car il faut représenter les interactions au niveau des raies d'absorption des gaz. On utilise des modèles à bandes étroites.

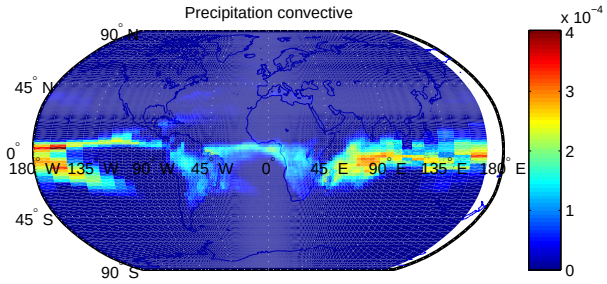
Enfin, un petit module, l'IOIPSL, gère les entrées et sorties.

Remarque B.1: *Au sujet du module de surface **Orchidée**. Le rayonnement incident s'égalise avec les flux de chaleur sensible et de chaleur latente au niveau du sol et les flux de conduction dans le sol. Le flux de chaleur intervient dans le cas de l'évaporation. En l'absence de disponibilité en eau, un flux d'énergie incident entraîne un réchauffement de l'atmosphère : il s'agit d'un flux de chaleur sensible. En revanche, lorsque de l'eau est disponible, l'énergie incidente est d'abord utilisée pour évaporer l'eau (flux de chaleur latente : elle est latente car imperceptible mais toutefois suscpetible de se libérer en cas de condensation) et ensuite pour chauffer l'air. Le réchauffement est donc moins rapide. Ainsi, on comprend pourquoi il fait si chaud pendant la journée dans les régions arides. Il n'y a pas d'eau donc toute l'énergie incidente est utilisée pour réchauffer l'air. Il est intéressant de remarquer que la canicule de 2003 peut également s'expliquer en partie par ce phénomène. Il avait très peu plu en particulier durant l'été précédent. De ce fait, la disponibilité en eau était très faible et le réchauffement s'est donc avéré massif.*

Maintenant que nous avons décrit les composantes principales du modèle LMDZ4, nous sommes en mesure de comprendre comment il peut rendre compte de l'évolution climatique. Comme nous l'avons vu, le noyau dynamique résout les équations d'évolution de types Navier Stokes. Ceci fournit les variables d'état telles la température et les vitesses du vent. Le pas de temps est de l'ordre de quelques minutes. Le noyau physique n'est pas appelé de manière aussi fréquente ; c'est-à-dire que certains paramètres des équations de Navier Stokes restent constants pendant quelques pas de temps. Il ne faut pas oublier de fournir des conditions aux limites (températures de surface,...) ainsi que des conditions initiales. Néanmoins, après un certain temps, ces dernières sont oubliées.

Nous présentons quelques résultats issus de LMDZ dans les figures B.2 et B.3

Enfin, un point crucial concernant ce modèle réside dans sa **capacité à zoomer**. Il s'agit véritablement de le l'une de ses spécificités par rapport aux modèles concurrents. Il faut bien comprendre que dans le cas présent, le modèle LMDZ est guidé par les observations de l'ECMWF. Il y a en effet un forçage externe par les données issues des analyses. Comme nous le verrons par la suite, il s'agit en fait en quelque sorte de downscaling physique. On relie les valeurs des grandeurs sur une grille assez fine (celle de LMDZ zoomé/guidé) à celles associées à une grille beaucoup plus grossière (celle des analyses) par une interpolation physique.



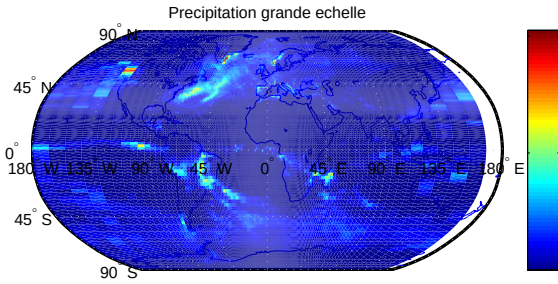
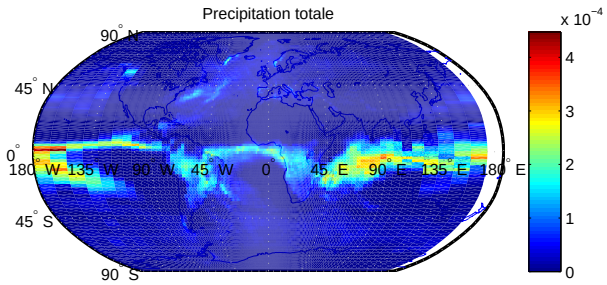


FIGURE B.2 : Cumul de précipitations convectives (g) et de grande échelle (d) en janvier 2003



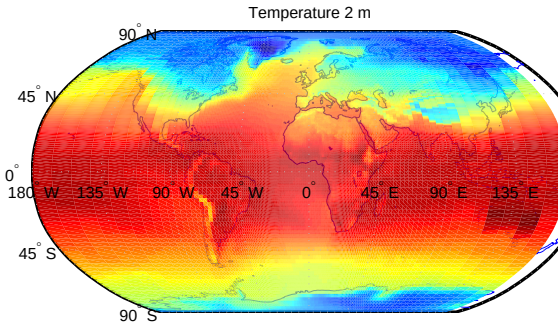


FIGURE B.3 : Cumul de précipitations totales (g) température (d) en janvier 2003

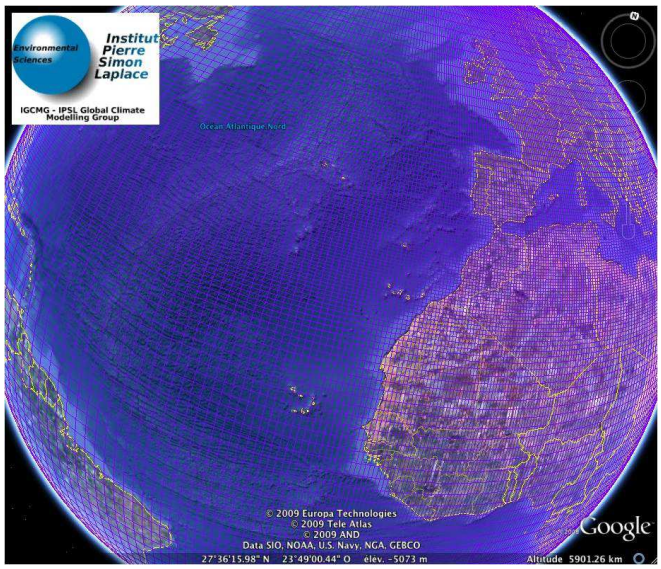


FIGURE B.4 : Grille zoomée globale

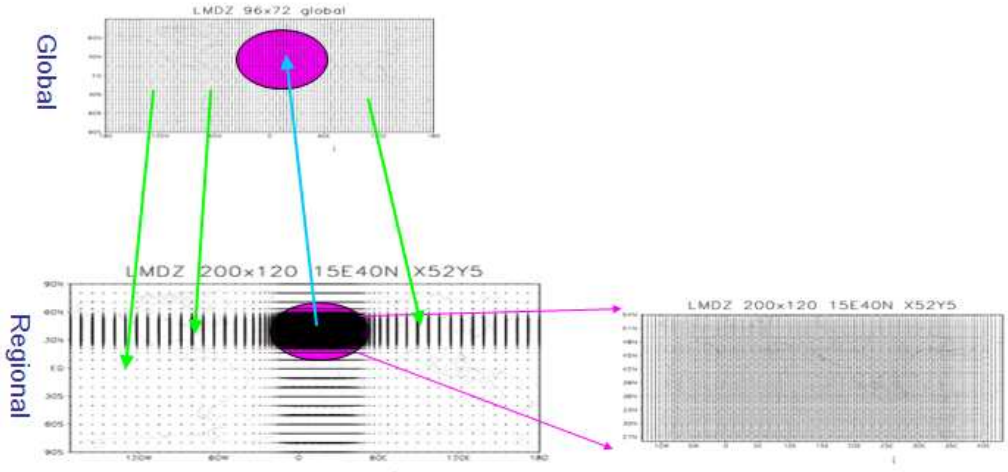


FIGURE B.5 : Principe du zoom sur une région

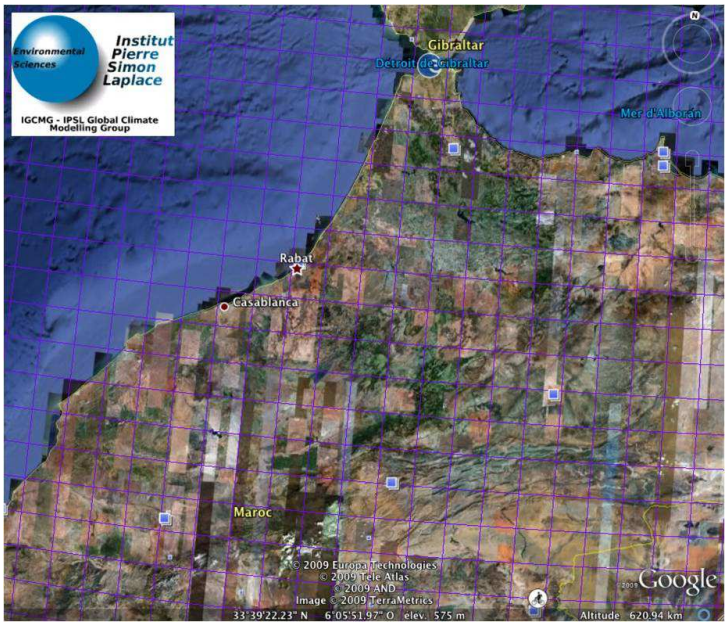


FIGURE B.6 : Résolution spatiale sur le Maroc

Annexe C

Aspects techniques de la modélisation

C.1 Manque de données

Dans leurs travaux [1971], Vapnik et Chervonenkis mesurent la capacité de généralisation par la fonction suivante : $h(g_W) = P[g_W(x) = g(x)]$, où P est la f.d.p. (i.e. fonction de distribution de probabilité) des données x. En fait, ne connaissant pas la f.d.p. P, on estime $h(g_W)$ sur la base d'apprentissage par $h_N(g_W)$, où g_W a été paramétré par les N exemples de la base d'apprentissage. La limite de $h_N(g_W)$ lorsque le nombre d'exemples N croît est $h(g_W)$, mais $h_N(g_W) > h(g_W)$ car N est toujours fini. Il y a sur-apprentissage lorsque $h_N(g_W) = 1$, l'estimation g_W sera alors dite biaisée. L'idée est alors d'étudier le cas le pire, i.e. le RN g_W pour lequel l'erreur est maximale $\|h_N(g_W) - h(g_W)\|$:

$$max_{\{g_W\}} \|h_N(g_W) - h(g_W)\|$$

où $\{g_W\}$ est la famille de fonctions pouvant être représentées par un RN g_W. Le résultat obtenu par Vapnik et Chervonenkis est que :

$$P[max_{\{g_W\}} \|h_N(g_W) - h(g_W)\| > \varepsilon] \leq 4\Delta(2N) \exp -\frac{\varepsilon^2 N}{8}$$

où $\Delta(2N)$ est une fonction de « croissance » indiquant le nombre de dichotomies qui peuvent être représentées par le RN pour avec les N exemples. Généralement, pour N faible, $\Delta(2N)$ croît est égal à 2^N. Ensuite, pour un nombre seuil d'exemples N_{seuil}, il croît plus faiblement. Ce nombre d'exemples seuil est appelé la dimension de Vapnik/Chervonenkis d_{VC}. On peut calculer des bornes pour d_{VC} mais elles ne sont pas exploitables dans la pratique.

Ainsi, les travaux de Vapnik et Chervonenkis font le lien entre l'erreur de généralisation et la capacité de représentation du RN pour spécifier le nombre d'exemples requis pour faire l'apprentissage.

La dimension de Vapnik-Chervonenkis ne donne qu'une borne supérieure pour le nombre d'exemples dans la base d'apprentissage. Cette borne est difficilement exploitable dans la pratique car elle se place dans le cas le pire, alors que l'on est souvent intéressé par le cas moyen. Bottou [1991] montre, sur un exemple, que l'on retrouve bien qualitativement le comportement attendu (c'est-à-dire qu'un problème plus complexe requiert plus d'exemples dans la base d'apprentissage) mais que quantitativement, il se trompe d'un ordre de grandeur sur la taille de la base d'exemples.

De plus, on peut faire la remarque que cette dimension ne tient pas compte des éventuelles régularisations intervenant dans le RN.