Prévision de la récolte de canne à sucre à partir d'un modèle de croissance Exemple de La Réunion

Pierre Todoroff[1], Jean-François Martiné[2], Eric Gozé[3]

[1] *CIRAD UR SCA Station de Ligne Paradis 97410 Saint-Pierre La Réunion*
pierre.todoroff@cirad.fr
[2] *CIRAD UR SCA Station de La Bretagne, BP 20 97408 Saint-Denis Messagerie Cedex 9 La Réunion*
[3] *CIRAD UR SCA Avenue Agropolis 34398 Montpellier Cedex 5 - France*

Résumé

La prévision de récolte de canne à sucre est une étape cruciale dans l'organisation logistique et la rentabilité de l'ensemble de la chaine de production de sucre. Nous présentons une méthode d'estimation de la récolte simple, robuste automatisable et de faible coût. Son principe repose sur l'ajustement d'un modèle de régression linéaire multivariée entre des variables de croissance simulées par un modèle de croissance de la canne et un historique de production.

La performance de cette méthode est évaluée en l'appliquant à la production de canne à sucre de l'île de La Réunion. Elle est basée sur la modélisation de la croissance des parcelles de canne par le modèle de culture MOSICAS qui fournit des variables explicatives à des régressions linéaires de plusieurs variables ajustées sur un historique de production de 11 années des cinq bassins de production de La Réunion. Pour cela, la performance des méthodes de régression Stepwise et Least angle sont comparées. La précision des prévisions issues des deux modèles de régression a été calculée par validation croisée. Des prévisions de rendement sont simulées jusqu'à quatre mois avant le début de la récolte. Sont comparées, la précision des prévisions fournies par les deux modèles de régression, les prévisions à dire d'expert, et la moyenne des productions réelles, aux échelles du bassin de production, de l'usine, et de l'ile entière. Les résultats ont montré que les deux modèles de régression ont les meilleures performances à toutes les échelles. A l'échelle des bassins de production, la méthode de Least angle, avec une erreur quadratique moyenne de 7%, est légèrement meilleure que la méthode Stepwise (7.4%), et plus précise que la variation interannuelle moyenne du rendement (9.2%). Cette erreur descend à 3.6% à l'échelle de l'ensemble de l'ile. La précision à l'échelle du bassin est très satisfaisante même quatre mois avant le début de la récolte (erreur maximale de 10%). Cette méthodologie de prévision de récolte a l'avantage d'être semi-automatisable. Les prévisions de rendement peuvent ainsi être actualisées en temps quasi-réel à condition de disposer d'un réseau de stations météorologiques automatiques.

Mots clés : *Prévision de rendement, modèle de croissance, canne à sucre, régression linéaire, précision, variables explicatives, climat.*

Introduction

La productivité de la filière canne à sucre est en partie conditionnée par le bon déroulement de la phase de récolte. Il s'agit d'ajuster au mieux les outils industriels (définition de la période de fonctionnement des sucreries et approvisionnement des intrants) et les moyens de récolte (ajustement de la logistique de récolte, calcul et répartition géographique des quotas de livraison), de façon à pouvoir broyer dans les meilleures conditions la totalité des cannes récoltables.

Pour cela, la connaissance du volume de canne à récolter est une variable clé.

La prévision de cette production reste une tâche épineuse dans les zones où la canne est cultivée par une multitude de producteurs dont on connaît rarement avec précision les surfaces en culture, l'état végétatif et sanitaire des cultures, l'itinéraire technique, etc…

La plupart du temps les unités sucrières utilisent leur propre système de prévision de récolte basé sur des échantillonnages dans des champs de référence.

Des modèles de cultures ont été développés depuis une trentaine d'années (Bouman, van Keulen et al. 1996) et sont largement utilisés dans les systèmes de prévision de rendement (Singhi and Pariyar 1995).

La plupart d'entre eux reposent sur le calcul du rendement simulé à la date de la récolte en complétant les données météorologiques mesurées par des prévisions climatiques (Bezuidenhout and Singels 2007) ou des données calculées par un générateur de climat (Soltani and Hoogenboom 2007).

Parallèlement des méthodes d'estimation basées sur des données de télédétection ont vu le jour (Rasmussen 1997).

Des systèmes de prévision plus sophistiqués utilisent les données de télédétection pour initialiser et ajuster des modèles de cultures (Moran, Maas *et al.* 1995).

Tous les pays producteurs ne disposent pas des données ni des moyens financiers et techniques nécessaires à la mise en œuvre de ces techniques.

Nous proposons dans cet article une méthode de prévision de récolte simple, robuste, de faible coût, et adaptée aux cas où les données statistiques disponibles sont géographiquement très agrégées (bassin de production, région, …).

Son principe consiste à ajuster un modèle de régression linéaire de plusieurs variables explicatives, fournies par le modèle de croissance de la canne à sucre MOSICAS (Martiné and Lebret 2001; Martiné and Todoroff 2002), sur les données historiques de production agrégées à l'échelle d'un bassin de production.

Cet article montre qu'il est possible d'obtenir un système de prévision de récolte fiable et automatisable. Les résultats présentés mettent en œuvre 2 méthodes de calcul de modèles de régression qui permettent de sélectionner les variables explicatives les plus pertinentes et d'ajuster les coefficients de régression de façon semi-automatique : la régression stepwise, très répandue et la régression « least angle » (Efron, Hastie et *al.* 2004) développée plus récemment.

Les principes des méthodes sont illustrés et les résultats obtenus comparés sur l'exemple de l'île de La Réunion.

Matériel et méthode

Contexte agricole et climatique

La canne à sucre est la principale production agricole de l'île de La Réunion. Elle est cultivée sur 25 000 ha et occupe plus de la moitié des surfaces agricoles utilisées (Agreste Réunion, 2008).

Ces terres sont exploitées par près de 3 400 exploitants individuels, répartis dans 5 bassins de production (Figure 1). La majorité d'entre eux possède une exploitation de taille moyenne (de 5 à 20 ha).

Il est très difficile dans ces conditions d'anticiper à coût raisonnable et avec une précision satisfaisante les volumes à récolter de chaque exploitant.

De plus les rendements, de 70 tonnes/ha en moyenne, sont très hétérogènes sur l'île de par la diversité des conditions pédoclimatiques des zones de production : littoral Est chaud et pluvieux, littoral Ouest sec et irrigué, zones d'altitude fraîches et pluvieuses, zones Sud semi-humides et irriguées (Figure 2).
A cette variabilité géographique s'ajoute une forte variabilité interannuelle des précipitations.

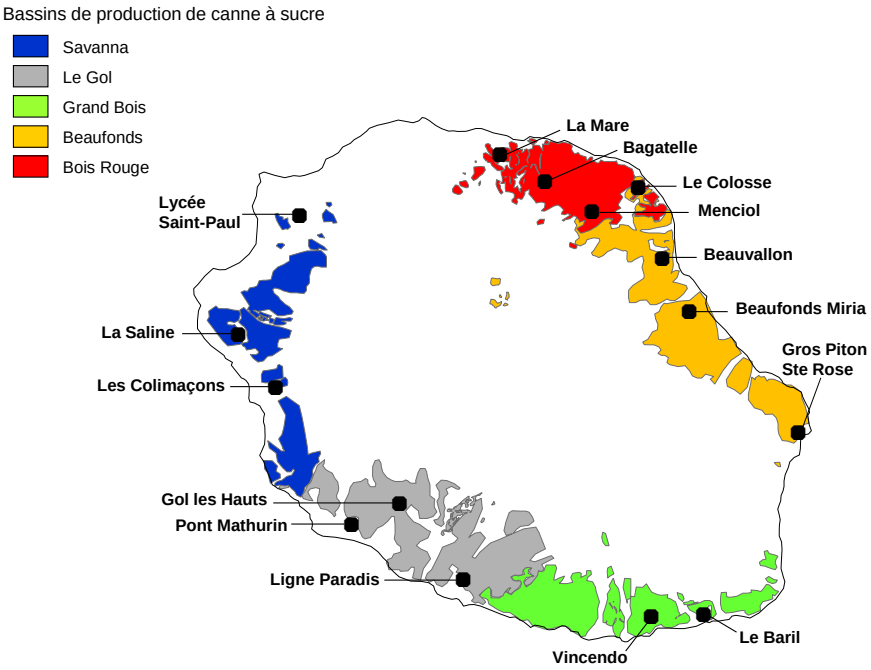

Figure 1: Stations climatiques de référence et bassins de production

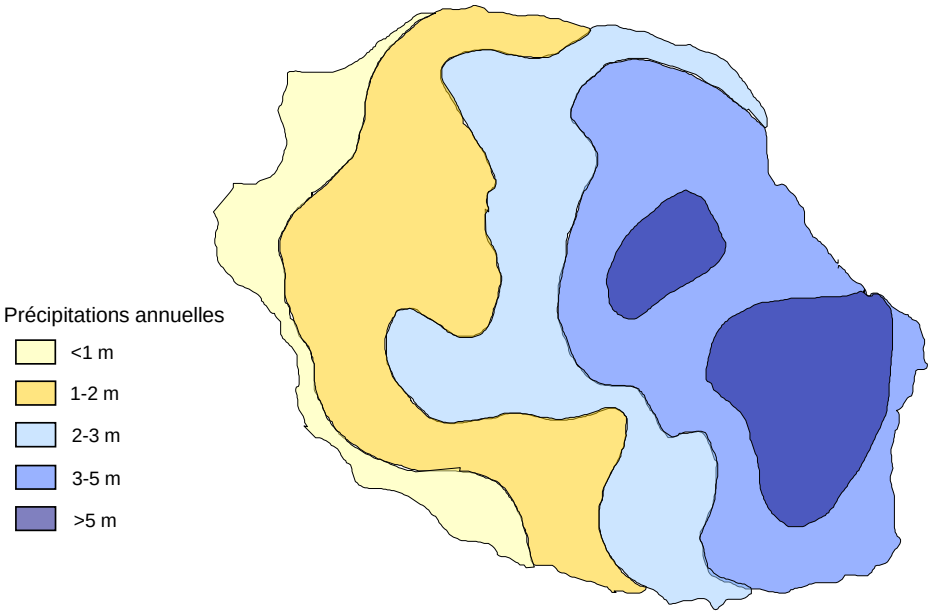

Figure 2 : Précipitations annuelles moyennes

Données de production

La filière canne à sucre conserve les données de rendement à l'échelle des planteurs depuis 2003 grâce au registre parcellaire graphique mis en place par la Communauté Européenne chez ses États Membres dans le cadre de la Politique Agricole Commune (Léo and Lemoine 2001).

Les données fiables de production et de surfaces remontent à 1998 mais sont agrégées à l'échelle du bassin de production.

Pour constituer une série historique suffisamment longue, c'est cette échelle géographique modélisation qui est retenue. Les données de rendement sont calculées de 1998 à 2008.

Le modèle de croissance de la canne MOSICAS

Il s'agit d'un modèle dynamique de type semi-mécaniste au pas de temps journalier qui simule la croissance en biomasse des différents organes de la canne (tiges, feuilles, racines).

Ce modèle simule la croissance d'une parcelle de canne homogène conduite en conditions « idéales », c'est-à-dire à l'optimum thermo-radiatif limité par la seule contrainte hydrique. On considère donc que les autres facteurs limitant potentiels de la croissance (alimentation minérale, adventices, maladies, mauvaises pratiques culturales,…) sont maitrisés et négligeables.

La réalité est parfois très éloignée de ces hypothèses. Les résultats ne peuvent donc pas être appliqués directement pour décrire la croissance des parcelles commerciales. C'est pourquoi le recours à un ajustement statistique est nécessaire pour tenir compte des écarts entre le potentiel calculé et les conditions réelles de croissance.

Les données d'entrée du modèle sont constituées des variables décrivant la parcelle considérée (sol, écartement des rangs), l'itinéraire technique (variété, stade, date de plantation) et les données climatiques journalières (pluie, température, rayonnement global et ETP).

Les sorties sont constituées de variables décrivant la croissance en biomasse de la parcelle (matière sèche des tiges usinables, hauteur des tiges, indice foliaire, teneur en sucre,…), les variations du bilan hydrique (taux de remplissage de la réserve utile, …) et les conditions agroclimatiques de croissance (rayonnement photosynthétique actif « PAR » intercepté, évapotranspiration maximale, âge thermique, …).

Les paramètres de configuration et d'initialisation des parcelles sont fixés dans cette étude en fonction des données culturales et pédologiques moyennes observées (voir Tableau 1).

Tableau 1 : Paramètres de configuration des parcelles de simulation

Ecartement des rangs	1,5 m
Variété	R570
Stade de repousse	1ère repousse
Coefficient cultural initial	0,2
Réserve utile (RU)	100 mm/m
Taux de remplissage initial de la RU	25%
Profondeur d'enracinement	1 m

Principe de la méthode

Une série de 11 années de données de production (traduites en rendement), de 1998 à 2008, est disponible à l'échelle de chaque bassin.

Il s'agit d'ajuster un modèle de régression du rendement (variable expliquée) constitué de plusieurs variables explicatives. Le produit de ce rendement avec les surfaces à récolter donne la masse totale de canne à récolter.

Les variables explicatives potentielles sont constituées des sorties calculées par le modèle MOSICAS : des variables caractérisant la croissance (biomasse) de la canne, des variables décrivant le bilan hydrique et des variables climatiques (Figure 3).

Pour chaque bassin nous simulons la croissance d'une parcelle fictive représentative des conditions de croissance de ce bassin. Les données climatiques retenues pour simuler la croissance de chacune des 5 parcelles fictives sont les moyennes des mesures des stations présentées dans la Figure 1.

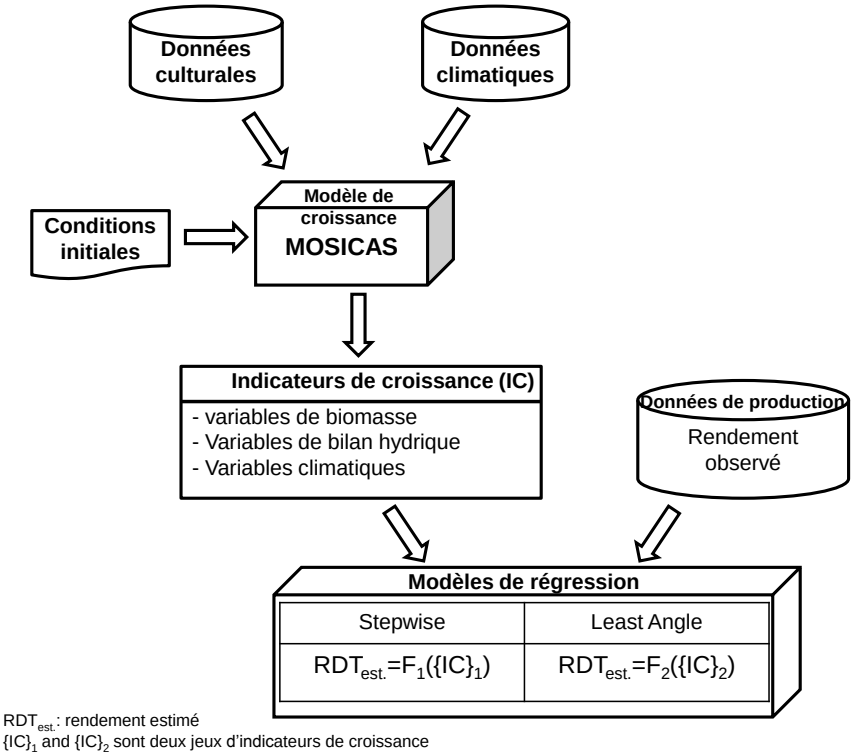

Figure 3: Diagramme de la méthodologie

Choix des variables explicatives

Les variables explicatives sont calculées à la date du jour des dernières données climatiques disponibles (en général la veille de la prévision de récolte). On l'appellera « jour j » dans la suite.

Exemple : pour des prévisions effectuées le 15 juin 2008 et des données climatiques disponibles jusqu'au 14 juin 2008, nous calculons les valeurs des variables explicatives au 14 juin de chaque année de l'historique.

La date de début de croissance des parcelles est estimée à la date de coupe médiane, celle où la moitié du tonnage de chaque bassin est livrée lors de la campagne précédente (on sous-estime la durée de croissance de la 1ère moitié du volume de canne du bassin autant qu'on surestime celle de la 2ème moitié).

La date de fin de simulation est le jour j de chaque année (Figure 4).

Le modèle produit 97 variables en sortie. Or il n'y a que 11 observations (rendement) disponibles dans chaque bassin. Toutes ces variables explicatives ne peuvent donc pas être retenues dans un même modèle de régression.

Des techniques de sélection bien connues (régression backward, forward ou stepwise) ont été élaborées pour retenir les variables les plus explicatives de la variable observée (Hocking 1976).

Ces méthodes utilisent des algorithmes d'inclusion ou de rejet des variables explicatives afin de minimiser les résidus. Ce qui conduit souvent à un sur-ajustement du modèle sur les données, donc de mauvaises capacités de prédiction (Allen 1974), et à une forte instabilité des variables retenues et des coefficients associés, de petites variations dans les données produisant de fortes variations de celles-ci (Prost, Makowski *et al.* 2008). Ceci est d'autant plus marqué que les variables explicatives sont susceptibles de présenter une colinéarité statistique significative malgré la sélection initiale (Foucart 2006).

Des méthodes minimisant ces inconvénients ont été mises au point plus récemment. Elles consistent à ajuster les valeurs des coefficients de régression par petits pas au lieu de calculer directement la valeur correspondant au critère des moindres carrés. Parmi celles-ci la méthode de régression « least

angle » (Efron, Hastie et al. 2004) met en œuvre des techniques mathématiques qui réduisent le nombre de calculs (Hesterberg 2008).

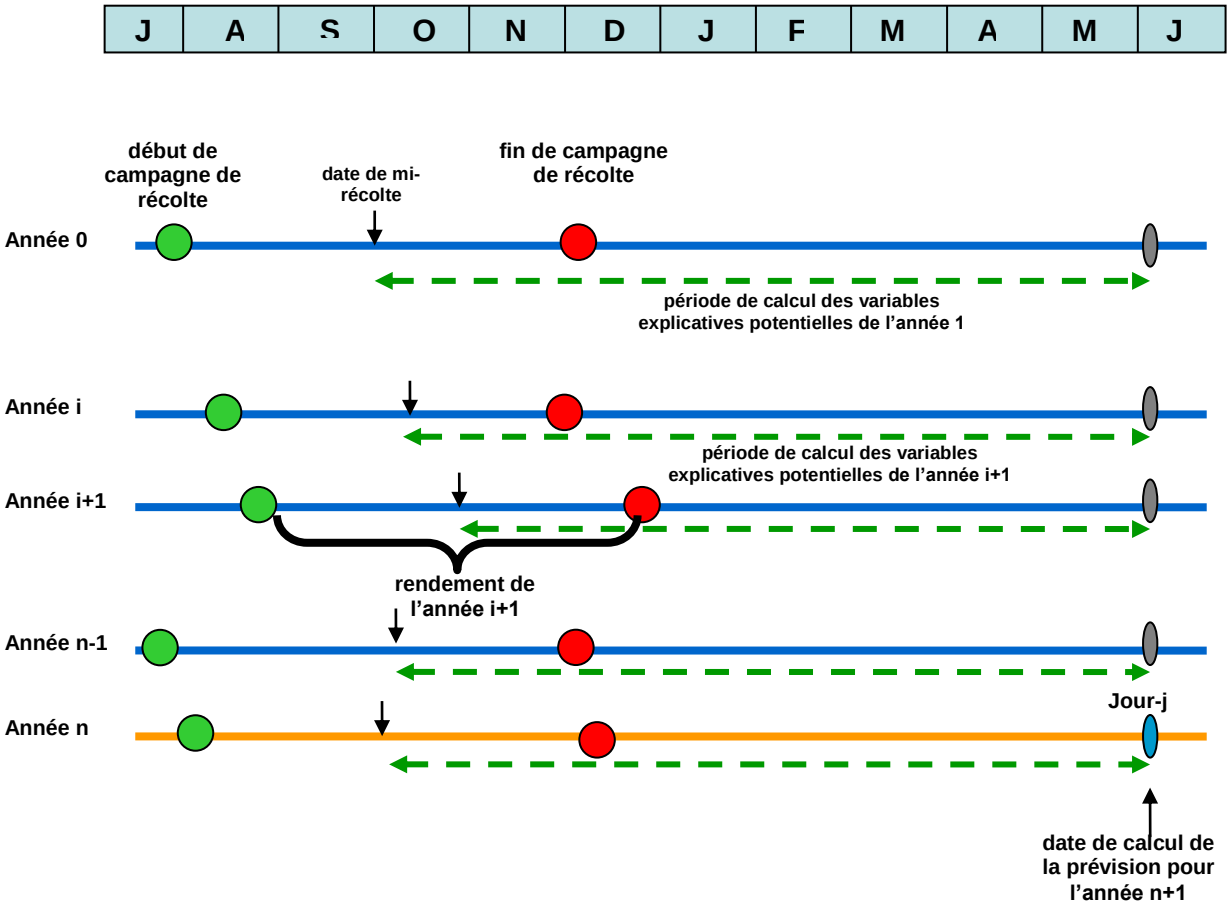

Figure 4 : Périodes de calcul des variables explicatives potentielles

Dans cet article nous choisissons de mettre en œuvre les 2 types de méthodes : la méthode stepwise très connue et la méthode plus récente de régression « least angle » et de comparer leur performances pour la prévision de rendement sur notre jeu de données.

Un premier filtre permet de ne retenir que les variables intégratives (moyennes ou cumuls tels que la somme des pluies, le taux de satisfaction hydrique moyen, etc…) calculées du 1er au dernier jour de la croissance, et ignorons les variables journalières.

Dans un 2ème temps, les variables trop fortement corrélées entre elles ainsi que celles qui, après raisonnement agronomique, sont les moins reliées au rendement sont éliminées. Pour éviter tout risque de sur-ajustement du modèle de régression, et au regard du nombre d'observations, le nombre maximal de variables explicatives utilisables dans les modèles est fixé à 3.

Les calculs ont été mis en œuvre grâce à la fonction lars (Hastie and Efron 2007) du logiciel statistique R (R Development Core Team 2009), qui intègre les 2 algorithmes.

Les modèles de régression associés à chacun des 5 bassins sont ainsi calculés. La production estimée pour chaque bassin (Figure 5) est déduite par produit avec la surface correspondante.

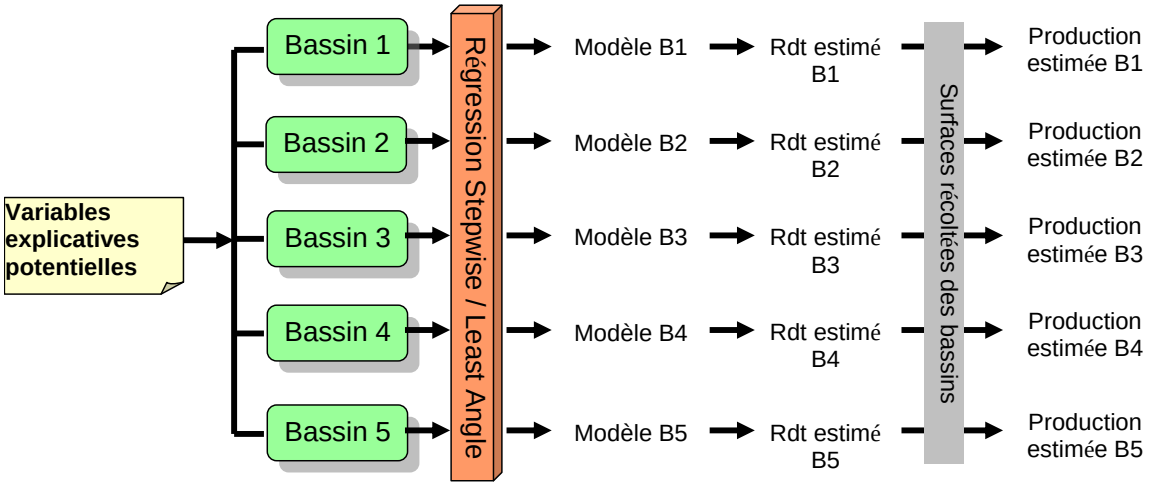

Figure 5: Calcul de la production estimée

Calcul de l'erreur

Le critère d'erreur retenu est l'erreur quadratique moyenne c'est-à-dire la moyenne de la somme des carrés des écarts (SCE) entre la valeur de rendement estimé et la valeur observée.

Cette erreur est calculée par validation croisée (de type Leave One Out) : les modèles de régression sont calculés à partir des données simulées et observées auxquelles sont retirées successivement une des 11 années de données de production.

Dans les tables et graphiques l'erreur de prévision est présentée sous la forme d'un écart relatif, c'est à dire la racine carrée de la SCE divisée par la moyenne de la variable observée.

Résultats et discussion

Les variables retenues par les algorithmes sont généralement les mêmes pour un même bassin, mais diffèrent d'un bassin à l'autre. Les modèles de régression sont donc distincts entre bassins. La différence entre les coefficients de régression calculés par les deux algorithmes pour un même bassin tient à la méthode de régression.

La précision des modèles obtenus est calculée dans chacun des bassins. La moyenne de l'erreur de prévision sur les 5 bassins est également présentée.

Des prévisions plus ou moins précoces ont été simulées, le premier de chaque mois entre mars et octobre. Les résultats sont présentés Figure 6.

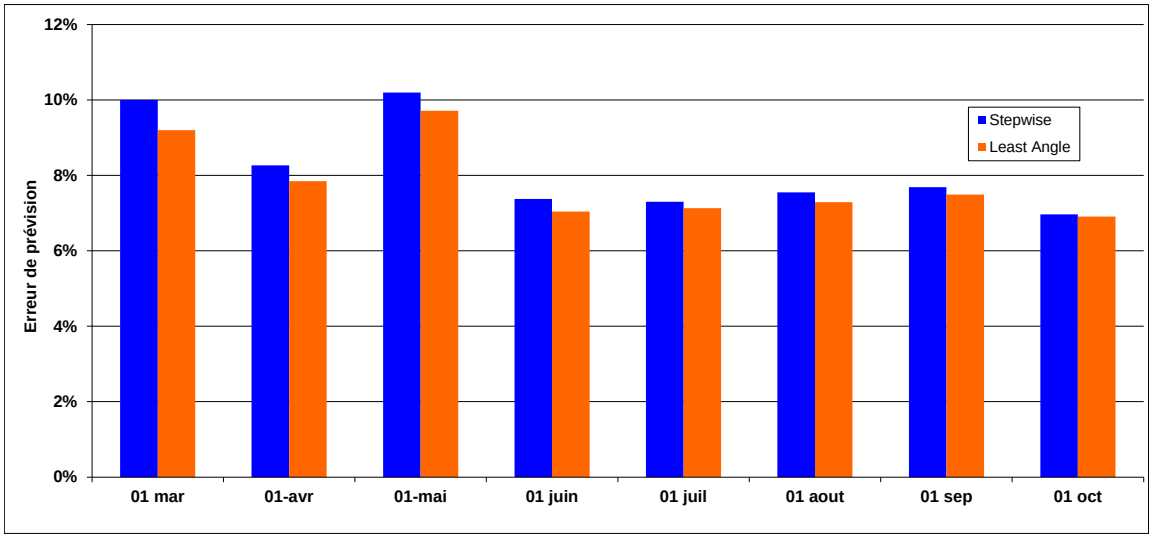

Figure 6: Erreur de prévision à différentes dates

La méthode se révèle relativement robuste à la date de calcul des prévisions puisque l'erreur reste inférieure à 10% quelque soit le mois de calcul, et proche de 7% à partir de juin.

On constate également que les erreurs obtenues avec la régression least angle sont plus faibles que celles obtenues avec la régression stepwise.

Ce qui signifie que cette technique de prévision peut être mise en œuvre tôt dans l'année, soit 4 à 5 mois avant le début de la récolte, avec une précision suffisante pour donner une bonne idée de la tendance de la récolte à venir.

Comparaison de la précision des méthodes de prévision à l'échelle des bassins de production

La comparaison des résultats simulés avec le modèle trivial constitué du rendement moyen observé au cours des 11 années d'historique (dénommé modèle « moyen » par la suite) permet d'évaluer le gain de précision apporté par les simulations du modèle de croissance et leur utilisation dans une régression linéaire. La Figure 7 compare les prévisions calculées par la moyenne des observations avec les résultats obtenus par les régressions (stepwise et least angle) pour chaque bassin au 1er juin.

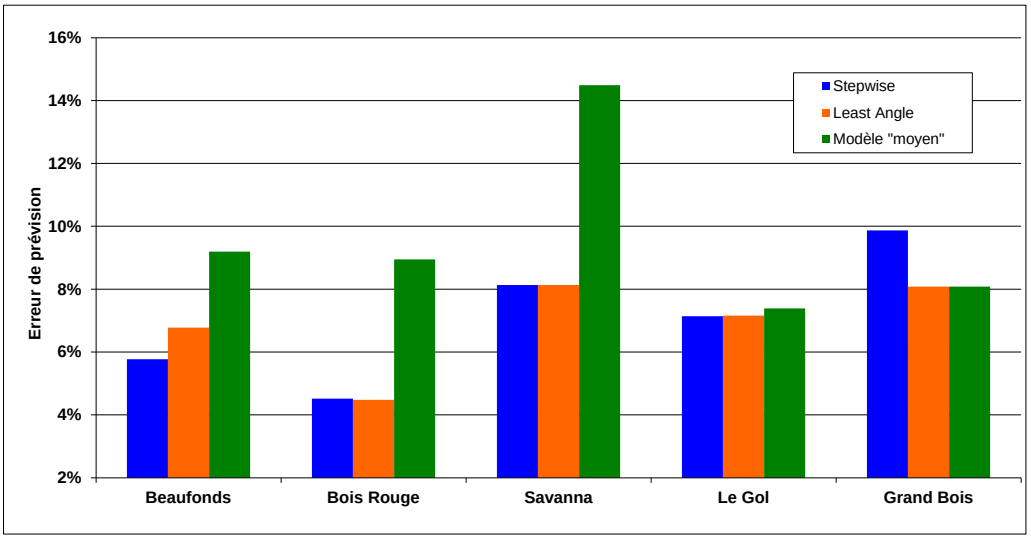

Figure 7 : Erreur de prévision pour les 5 bassins au 1er juin

Les modèles de régression présentent toujours une meilleure précision que la simple moyenne des observations, jusqu'à 6 points de mieux, sauf pour le bassin de Grand Bois pour lequel la précision du modèle « moyen » est du même ordre de grandeur que celle obtenue avec le modèle de régression least angle.

La méthode de régression least angle présente globalement une précision légèrement meilleure que la régression stepwise (respectivement 7% contre 7.4% de moyenne sur les 5 bassins dans ce cas). A titre de comparaison la variation interannuelle moyenne du rendement (écart-type) est de 6.1 t/ha (soit une variation relative de 9.2%).

Limites de la méthode

Cette méthode est très sensible aux données de production observées ; l'erreur associée à ces données se retrouve dans l'erreur du modèle de prévision. La principale source d'erreur dans notre cas est l'évaluation des surfaces associées aux livraisons qui sont utilisées pour calculer le rendement.

La méthode ne permet pas par ailleurs de prendre en compte des facteurs non corrélés au climat (changement de techniques culturales, accidents de cultures, problèmes logistiques,…).

Comparaison avec la méthode d'estimation à dire d'expert

A titre de comparaison, nous avons calculé les erreurs obtenues avec la méthode traditionnelle d'estimation par échantillonnage et expertise de terrain. Ces données ne sont disponibles à La Réunion qu'à l'échelle des 2 usines de l'île (Le Gol et Bois Rouge) et de l'île.

La Figure 8 compare l'erreur de prédiction calculée depuis 1998 par les 3 méthodes développées précédemment appliquées aux données ainsi agrégées, ainsi que l'erreur obtenue par le modèle « moyen ».

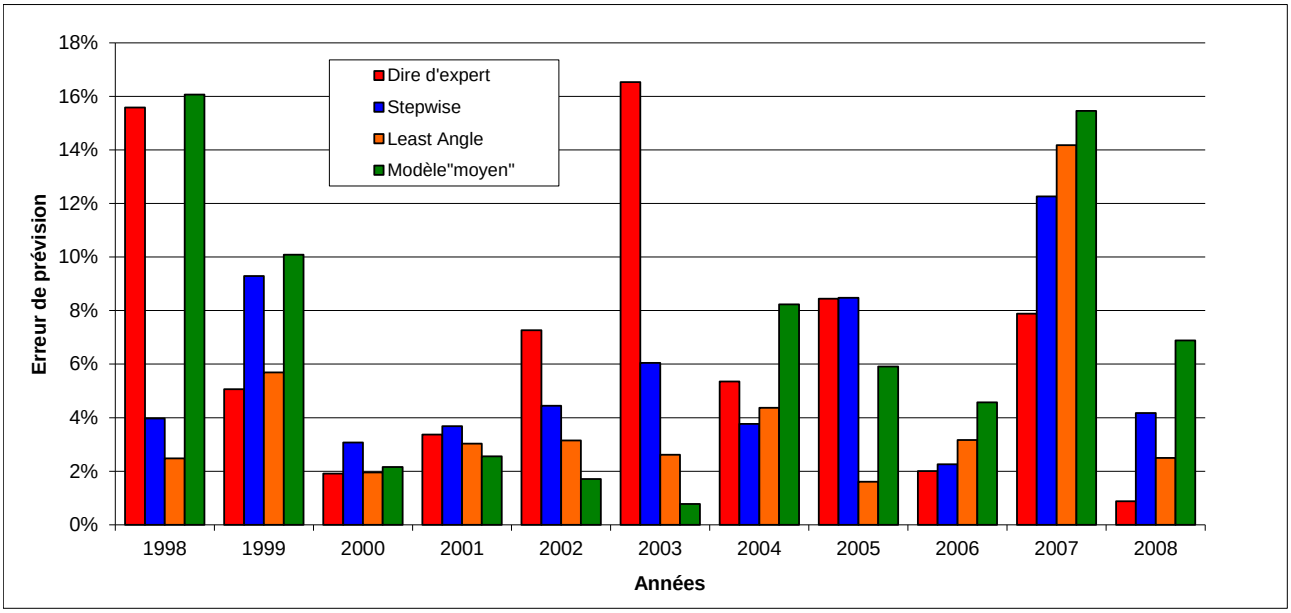

Figure 8 : Comparaison de l'erreur de prévision des différentes méthodes

Le Tableau 2 présente l'erreur de prévision moyenne sur les 11 années, pour chaque méthode et aux 2 échelles.

Dans tous les cas la précision s'améliore par effet de lissage lorsque l'on agrège les données (au mieux 3.7% d'erreur contre 7% à l'échelle des bassins).

Les prévisions à dire d'expert ne montrent pas de gain par rapport à une prévision basée sur la moyenne des observations. L'erreur est inférieure à 7%. Cette méthode, telle qu'utilisée à La Réunion, ne présente donc pas d'intérêt par rapport au calcul de la moyenne des rendements de l'historique.

Tableau 2 : Erreur de prévision des différentes méthodes à différentes échelles

Méthode de prévision	Erreur à l'échelle des usines	Erreur à l'échelle de l'île
Dire d'expert	6.8%	6.7%
Régression Stepwise	5.6%	3.8%
Régression Least Angle	4.1%	3.7%
Modèle "moyen"	6.8%	6.6%

En revanche les prévisions par régression linéaire permettent de réduire l'erreur d'un facteur 2 en descendant en dessous de 4% d'erreur. La précision obtenue avec la méthode least angle reste meilleure que celle obtenue avec la méthode stepwise.

Conclusion

La méthode de prévision de récolte basée sur le modèle de croissance de la canne à sucre MOSICAS et des modèles statistiques de régression avec les données de production observées est utilisée depuis plusieurs années à La Réunion.

Elle donne des résultats très satisfaisants à l'échelle de l'ile : 3.7%, et des usines : 4.1% sur la production totale estimée au mois de juin au cours des 11 dernières années. Elle est plus performante que les prévisions à dire d'expert en particulier lorsque les conditions climatiques sont inhabituelles ou contrastent avec celles de l'année précédente. Dans ces conditions en effet, les prévisions à dire d'expert, même partiellement basées sur des échantillonnages, ont du mal à quantifier les effets climatiques sur le rendement final.

Elle donne de bons résultats à l'échelle du bassin de production avec une précision de l'ordre de 4 à 8% selon le bassin.

Elle présente par ailleurs l'avantage de pouvoir être automatisée et fournir des prévisions à la demande par simple clic de souris via un site Web et actualisées en temps quasi réel grâce aux solutions de gestion de bases de données et réseaux de stations climatiques automatiques.

Bibliographie

Agreste Réunion, 2008. Données agricoles et rurales, Enquête sur la structure des exploitations en 2007, n° 36, Aout 2008, 4p.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. Technometrics **16**(1): 125-127.

Bezuidenhout, C. N. and A. Singels (2007). Operational forecasting of South African sugarcane production: Part 1 - System description. Agricultural Systems, **92**(1-3): 23-38.

Bouman, B. A. M., H. van Keulen, et al. (1996). The "School of de Wit" crop growth simulation models: A pedigree and historical overview. Agricultural Systems, **52**(2-3): 171-198.

Efron, B., T. Hastie, et al. (2004). Least angle regression. Annals of Statistics, **32**(2): 407-451.

Foucart (2006). Colinéarité et régression linéaire. Mathématics and Social Sciences, **44**(173): 5-25.

Hastie, T. and B. Efron (2007). lars: Least Angle Regression, Lasso and Forward Stagewise.

Hesterberg, T. C., N H Meier, L Fraley, C (2008). Least angle regression and l1 penalized regression: a review. Statistics Surveys, **2**: 61-93.

Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression. Biometrics**,** **32**(1): 1-49.

Léo, O. and G. Lemoine (2001). Land Parcel Identification System in the frame of Regulation (EC) 1593/2000 version 1.4. E. Commission. Ispra (Italy), Joint Research Centre, Space Application Institute. 1593/2000**:** 26.

Martiné, J.-F. and P. Lebret (2001). Modelling the water content of the sugarcane stalk. Proceedings of the South African Sugar Technologists Association, Durban.

Martiné, J.-F. and P. Todoroff (2002). The growth model "MOSICAS" and its simulation framework "SIMULEX". Congrès de la Société de Technologie Agricole et Sucrière de Maurice, Réduit - Mauritius -, STASM.

Moran, M. S., S. J. Maas, et al. (1995). Combining remote sensing and modeling for estimating surface evaporation and biomass production. Remote Sensing Reviews, **12**(3-4): 335-353.

Prost, L., D. Makowski, et al. (2008). Comparison of stepwise selection and Bayesian model averaging for yield gap analysis. Ecological Modelling, **219**(1-2): 66-76.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

Rasmussen, M. S. (1997). Operational yield forecast using AVHRR NDVI data: reduction of environmental and inter-annual variability. International Journal of Remote Sensing, **18**(5): 1059-1077.

Singhi, G. and M. P. Pariyar (1995). Crop yield forecasting, a review of methods used in developping countries of Asia. Crop Yield Forecasting Methods : Proceedings of the Seminar. FAO. Villefranche-sur-Mer - France**:** 167-179.

Soltani, A. and G. Hoogenboom (2007). "Assessing crop management options with crop simulation models based on generated weather data. Field Crops Research, **103**(3): 198-207.