RÉGRESSION MULTIPLE ET PRÉVISION DE RENDEMENTS AGRICOLES EN FONCTION DE DONNÉES MÉTÉOROLOGIQUES[1]

P. Dagnelie et R. Palm

Faculté des Sciences agronomiques
B-5030 Gembloux (Belgique)

pierre@dagnelie.be

SUMMARY

Under contracts with the Statistical Office (Luxembourg) and the Joint Research Centre (Ispra, Italy) of the European Community, we had, from 1982 to 1991, a set of research programs about the prediction of the main crops yields in the twelve countries of the Community. The aim of this paper is to give a summary of these works, a detailed review being published elsewhere [Palm and Dagnelie, 1993].

The data were related, on the one side, to about twenty annual and perennial crops, and on the other side, to the minimum, mean and maximum temperatures, and to the rain, for periods of ten days.

Our approach was of the empirical-statistical type, the yields being considered, by regression, as functions of time and meteorological variables.

The time factor models were classical least squares linear and quadratic regressions, robust regressions, exponential smoothing techniques, and Box-Jenkins autoregressive and moving average models.

Considering the meteorological data, the main problem was the selection of a few influential variables, within a very large set of possible explanatory variables. Several solutions were considered, including aggregation of meteorological stations, calculation of sums of temperatures and sums of rains, recursive selection process, selection according to the values of correlation coefficients, and the stepwise algorithm.

Altogether, the classical least squares linear and quadratic regression models seem to be the best time series models, and the contribution of meteorological variables to explain the residuals coming from these regression models seems to be very limited, whatever the models considered are.

1. INTRODUCTION

Dans le cadre de contrats conclus avec l'Office Statistique (Luxembourg) et le Centre commun de Recherche (Ispra, Italie) de la Communauté européenne, nous avons effectué, de 1982 à 1991, une série de travaux relatifs à la prévision des rendements des principales cultures, dans les douze pays de la Communauté. Le but de cet article est de présenter une synthèse de ces travaux, qui sont exposés par ailleurs de façon plus détaillée [Palm et Dagnelie, 1993].

Nous envisagerons tout d'abord les principes méthodologiques (paragraphe 2), puis nous donnerons quelques résultats (paragraphe 3) et quelques conclusions (paragraphe 4).

---

[1] *In: Biométrie et analyse des données spatio-temporelles*. Vannes, Société française de Biométrie, 50-57, 1993.

2. MÉTHODOLOGIE

2.1. Données disponibles

Les données agricoles concernent les superficies, les rendements et les productions d'une vingtaine de cultures annuelles et pérennes. Ces données ont trait aux différents pays de la Communauté, toutes les cultures considérées n'étant cependant pas présentes dans tous les pays.

Les données météorologiques disponibles sont relatives aux températures minimums, moyennes et maximums et aux précipitations, pour des périodes décadaires, ainsi qu'aux sommes de températures supérieures à certains seuils et aux sommes de précipitations, pour des périodes plus longues. Ces données proviennent d'une centaine de stations météorologiques.

Dans la plupart des cas, les données existent pour des périodes de 20 à 30 ans. Toutefois, les séries d'observations comportaient de nombreuses imperfections. Nous y avons remédié dans toute la mesure du possible, par la réalisation de tests de cohérence, par l'utilisation de fourchettes de vraisemblance, et par le calcul de distances et la recherche de valeurs anormales, à une et à plusieurs dimensions.

2.2. Principes généraux

Le but poursuivi était d'obtenir, aussi tôt que possible dans l'année, des prévisions de rendements par hectare, pour les différentes cultures et les différents pays, ainsi que pour l'ensemble de la Communauté.

Les modèles de prévision, du type statistique-empirique, ont été construits indépendamment pour les différentes combinaisons culture-pays prises en considération et pour chacun des mois de la période de croissance des plantes. À cette fin, nous avons tout d'abord calculé la tendance générale du rendement, pour chaque combinaison culture-pays considérée, et nous avons ensuite mis en relation, pour chaque mois, les résidus par rapport à la tendance avec les données météorologiques, par régression multiple.

2.3. Étude de la tendance générale

Les différents modèles de tendance qui ont été progressivement étudiés sont:

- dix-sept modèles de régression au sens des moindres carrés, linéaires et quadratiques, sans et avec transformations logarithmiques de certaines variables;
- quatre types de régression robuste, à savoir la méthode de la ligne résistante de Tukey, la norme L_1 et deux méthodes basées sur les rangs des résidus (R-estimateurs);
- trois types de lissage exponentiel: lissage exponentiel simple, lissage de Brown et lissage de Holt;
- dix modèles autorégressifs et de moyennes mobiles de type ARIMA (modèles de Box et Jenkins).

2.4. Étude des données météorologiques

Le principal problème que nous avons rencontré, en vue d'expliquer les résidus par rapport à la tendance à l'aide des variables météorologiques, avait trait à la sélection d'un petit nombre de variables au sein d'un ensemble très vaste, comprenant parfois plusieurs centaines de variables explicatives potentielles.

Une première réduction du nombre de variables explicatives a été obtenue en définissant, pour chaque combinaison culture-pays, une station météorologique moyenne conventionnelle. Les caractéristiques de ces stations moyennes ont été déterminées, pour chaque combinaison, par le calcul de moyennes pondérées des températures minimums, moyennes et maximums et des précipitations observées dans les différentes stations météorologiques du pays considéré, le

coefficient de pondération, pour une station donnée, étant proportionnel à l'importance de la culture dans la région où est située cette station.

Les différentes solutions suivantes ont ensuite été adoptées, pour réaliser une réduction complémentaire du nombre de variables explicatives.

Dans un premier temps, nous avons utilisé un modèle récursif de régression, en considérant le mois comme période de base et en utilisant la prévision relative à chaque mois comme variable explicative pour le mois immédiatement suivant. Plus concrètement, les variables explicatives relatives à un mois donné (par exemple, le mois de mai) sont alors les variables météorologiques relatives à ce mois pour la station moyenne du pays considéré (par exemple, les températures minimum, moyenne et maximum et les précipitations de mai), ainsi que le rendement prévu le mois précédent (par exemple, le rendement prévu à la fin du mois d'avril, pour la prévision réalisée à la fin du mois de mai).

Dans un deuxième temps, nous avons introduit un processus de sélection basé sur le calcul de coefficients de corrélation simple, en écartant toutes les variables météorologiques qui n'étaient pas suffisamment corrélées avec les résidus par rapport à la tendance. En outre, à ce stade, les rendements prévus au cours du mois précédent ont été remplacés, comme variables explicatives, par les variables météorologiques qui intervenaient dans les prévisions du mois précédent.

Enfin, en un troisième temps, nous avons considéré les données décadaires, en maintenant le processus de sélection que nous venons d'évoquer, mais en abandonnant le caractère récursif des modèles, qui avait été considéré antérieurement. Nous permettions ainsi l'introduction, dans un modèle donné, de variables météorologiques qui n'avaient pas été introduites dans les modèles précédents.

En outre, en vue d'étudier le problème considéré à une échelle inférieure à celle des pays et avec un choix de variables météorologiques plus diversifié, nous avons également réalisé une étude particulière relative aux rendements du maïs dans une vingtaine de départements français. Nous disposions, dans ce cas particulier, des rendements observés dans chacun des départements considérés et d'une ou deux stations météorologiques par département, les températures et les précipitations, déjà utilisées antérieurement, étant complétées ici par des observations de rayonnement et d'évapotranspiration potentielle. Au cours de cette étude, nous avons aussi pris en considération une série de variables dérivées, dont l'utilisation nous avait été suggérée par des agronomes particulièrement qualifiés.

Enfin, nous avons également procédé à divers essais d'ajustement global de modèles, faisant intervenir simultanément la tendance générale et les données météorologiques.

2.5. Mesure de la qualité des prévisions

Dans les différents cas, nous avons tout d'abord considéré le coefficient de détermination multiple, R^2, comme mesure de la qualité des ajustements.

Nous avons ensuite utilisé un paramètre semblable, associé à une procédure de "*jackknife*". Pour un ensemble de n années, les erreurs de prévision sont alors obtenues à partir des prévisions réalisées pour chacune des années sur la base des $n-1$ autres années, les variables intervenant dans le modèle étant néanmoins choisies en fonction de l'ensemble des données (n années). Plus particulièrement, le paramètre considéré dans ce cas est :

$$R_j^2 = 1 - \mathrm{CM}_j / s_y^2 ,$$

R_j^2 étant le coefficient de détermination multiple du "*jackknife*", CM_j le carré moyen de l'erreur de prévision du "*jackknife*" et s_y^2 la variance marginale du rendement.

Enfin, nous avons considéré les erreurs de prévision des dernières années (par exemple, des 8 dernières années), en utilisant des modèles basés uniquement sur les années antérieures (par exemple, les 22 premières années). À partir de ces erreurs simulées, nous avons calculé un ensemble de valeurs R_s^2, semblables aux valeurs R_j^2, mais en ne prenant en considération que les dernières années, dans le calcul du carré moyen de l'erreur de prévision et de la variance marginale.

On notera que, bien qu'elles soient désignées pour la facilité par un symbole faisant intervenir une puissance 2, par analogie avec la définition du coefficient de détermination classique, les quantités R_j^2 et R_s^2 peuvent être négatives.

3. QUELQUES RÉSULTATS

3.1. Résultats relatifs à la tendance générale

En ce qui concerne la régression au sens des moindres carrés, aucun des quinze autres modèles pris en considération ne s'est avéré supérieur aux régressions linéaire et quadratique classiques. De même, aucune des méthodes de régression robuste ne s'est montrée globalement supérieure à la régression au sens des moindres carrés.

Par contre, mais dans le cas des cultures annuelles uniquement, deux des méthodes de lissage sont apparues très légèrement supérieures, en moyenne, à la régression linéaire ou quadratique au sens des moindres carrés, les valeurs moyennes du paramètre R_s^2 étant égales à 0,68 et 0,69, pour ces deux méthodes, par comparaison avec 0,67 pour les moindres carrés classiques. Toutefois, cette apparente supériorité, en moyenne, est ternie par une très grande hétérogénéité des résultats dans le cas des lissages, les prévisions étant très bonnes pour certaines séries et au contraire très mauvaises (valeurs R_s^2 négatives) pour d'autres séries.

De même, certains des modèles ARIMA se sont avérés légèrement supérieurs à la régression linéaire ou quadratique au sens des moindres carrés, les valeurs moyennes de R_s^2 étant ici égales à 0,70 pour les deux meilleurs modèles, toujours par comparaison avec 0,67, dans le cas des cultures annuelles, et à 0,32, par comparaison avec 0,26, dans le cas des cultures pérennes. On notera cependant que les modèles ARIMA ne peuvent être utilisés de manière systématique que dans le cas de séries complètes, c'est-à-dire en l'absence de toute donnée manquante ou anormale, ce qui en limite considérablement l'intérêt.

Tenant compte de ces diverses considérations, nous pensons que la régression classique, linéaire ou quadratique, au sens des moindres carrés reste la procédure la plus adéquate pour tenir compte de la tendance générale.

3.2. Résultats relatifs aux données météorologiques

En ce qui concerne les données météorologiques, une des principales conclusions a trait aux divergences d'interprétation, particulièrement importantes, auxquelles peuvent conduire les différentes mesures de la qualité des ajustements et des prévisions (R^2, R_j^2 et R_s^2).

Le premier paramètre, R^2, est bien une mesure de la qualité des ajustements, et non pas une mesure de la qualité des prévisions. Les valeurs observées pour ce paramètre sont souvent relativement élevées (de l'ordre de 0,7 par exemple), en particulier pour les cultures annuelles, mais même dans les cas les plus favorables, les prévisions obtenues sont de mauvaise qualité.

Le deuxième paramètre, R_j^2, est déjà mieux adapté pour juger de la qualité des prévisions, mais il reste néanmoins très trompeur. Les valeurs obtenues sont en général inférieures à celles

du premier paramètre, le plus souvent avec une différence de l'ordre de 0,1 à 0,2, mais elles restent généralement positives.

Le troisième paramètre, R_s^2 , est lui une réelle mesure de la qualité des prévisions, et non plus de la qualité des ajustements. Mais les valeurs observées sont ici souvent négatives, mettant en évidence le fait que les erreurs de prévision ont fréquemment une variance supérieure à celle des résidus de la tendance eux-mêmes(1).

En considérant ce dernier paramètre, il apparaît, d'une manière tout à fait générale, que l'introduction de variables météorologiques dans les modèles, au lieu de réduire les erreurs de prévision, est une source d'accroissement de ces erreurs, en particulier quand le nombre de variables météorologiques figurant dans les modèles est élevé.

Ces divers phénomènes sont illustrés par les données du tableau suivant, qui concerne un des cas les plus favorables. Il s'agit en effet des valeurs moyennes de R^2 , R_j^2 et R_s^2 observées dans l'étude relative au rendement du maïs dans une vingtaine de départements français.

	R^2	R_j^2	R_s^2
Données brutes	0,71	0,58	- 2,96
Variables dérivées	0,31	0,22	- 0,50

Ce tableau a trait plus particulièrement aux prévisions basées, d'une part, sur les données météorologiques brutes et, d'autre part, sur les variables dérivées suggérées par des agronomes spécialisés. Dans le premier cas, le nombre de variables explicatives potentielles était de 144 et le nombre moyen de variables introduites dans l'équation était de 3,8. Pour le deuxième cas, ces nombres étaient respectivement de 13 et 0,7. On pourra ainsi constater que même cette dernière approche ne donne pas de résultats satisfaisants.

4. CONCLUSIONS

Dans l'ensemble de cette étude, nous avons pris en considération un très grand nombre d'équations de régression, dans des conditions très variées (différentes cultures et différents pays).

Globalement, le modèle classique de régression au sens des moindres carrés, linéaire ou quadratique, s'avère être le meilleur modèle de tendance générale, et la contribution des variables météorologiques pour expliquer les résidus par rapport à la tendance s'avère extrêmement limitée, quel que soit le modèle considéré.

Cette dernière conclusion, très décourageante, apparaît de façon très nette: les équations faisant intervenir des variables météorologiques s'ajustent très bien aux données, mais fournissent néanmoins de très mauvaises prévisions.

Cette conclusion semble due au processus de sélection des variables explicatives en fonction de leurs niveaux de signification. Comme l'a montré notamment Miller [1990], les processus classiques de sélection basé sur des niveaux de signification sont sans fondement et peuvent induire des erreurs importantes dans l'estimation des coefficients de régression. Les

---

(1) Pour la bonne compréhension de ce paragraphe, rappelons en effet que les données météorologiques interviennent ici comme variables explicatives des résidus, calculés par rapport à la tendance générale (paragraphe 2.2).

variables estimées de cette façon peuvent n'avoir aucune valeur prédictive et peuvent induire des erreurs de prévision importantes.

Cette étude fournit également une illustration du rôle fondamental de la validation croisée en régression multiple. Elle met notamment en évidence le fait que les sommes de carrés d'écarts résiduelles calculées par la procédure du "*jackknife*" peut conduire à une surestimation importante de la qualité apparente des modèles.

Une autre approche, complémentaire, pourrait être d'utiliser certaines des alternatives à la régression classique qui ont été proposées pour faire face aux problèmes de colinéarité: régression en fonction des composantes principales, méthode de Webster, Gunst et Mason, régression par les moindres carrés partiels, régression pseudo-orthogonale et estimateurs à "rétrécisseurs" de James et Stein [Palm et Iemma, 1993]. Le travail de Hébel *et al*. [1993a, 1993b] en fournit un exemple, dans un cas plus limité.

5. RÉFÉRENCES

Hébel P., Faivre R., Goffinet B., Wallach D. [1993a]. Estimateurs à rétrécisseurs appliqués à la prévision du rendement de blé d'hiver. *In: XXVes Journées de Statistique*. Vannes, Institut universitaire de Technologie, 1 p.

Hébel P., Faivre R., Goffinet B., Wallach D. [1993b]. Shrinkage estimators applied to prediction of French winter wheat yield. *Biometrics* **49** (1), 281-293.

Miller A. [1990]. *Subset selection in regression*. London, Chapman and Hall, 229 p.

Palm R., Dagnelie P. [1993]. *Tendance générale et effets du climat dans la prévision des rendements agricoles des différents pays de la Communauté Européenne*. Luxembourg, Office des Publications officielles des Communautés européennes, 132 p.

Palm R., Iemma A.F. [1993]. Quelques alternatives à la régression classique dans le cas de la colinéarité. *Notes Stat. Inform*. (Gembloux) 93/2, 27 p.