UNIVERSITÉ CLAUDE BERNARD LYON I

MASTER DATA SCIENCE

MODÈLES DE RÉGRESSION

Régression des données d'extraction de puits de pétrole au Canada

Auteurs :
Bruno DUMAS
Jules SAUVINET

Professeur :
François WAHL

12 janvier 2017

Résumé

L'objectif est d'effectuer des régressions des valeurs d'extractions des puits de pétrole du Canada afin de pouvoir prédire les future données d'extraction. Les courbes des valeurs des données d'extraction donne une classification de la qualité des puits. L'objectif de ce travail est de mettre en place une classification automatique de ces puits : La démarche proposée est décrite ci-dessous. L'idée est de remplacer ces courbes par des fonctions paramétriques. Pour cela, plusieurs r égressions vont être envisagées afin d'avoir la meilleure prédiction/classification possible.

Table des matières

1 Régressions polynomiales

Une façon simple est d'ajuster un polynôme de degré faible sur chacune des courbes et de voir si les coefficients présentent des clusters, c'est à dire des groupes de points distincts quand on les regarde dans l'espace. On essaiera des polynômes de degré 0, 1, 2, 3, et 4. On présentera les courbes de production simulées obtenues, comme dans la figure ci-dessous.

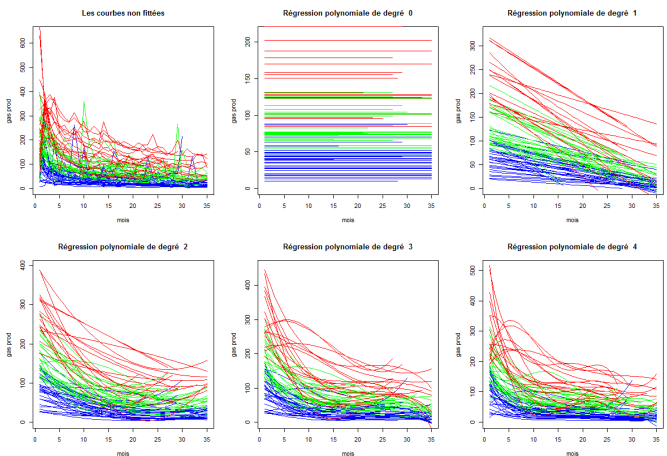

FIGURE 1 – Courbes non régressées, et courbes de régressions polynomiales de degrés 0,1,2,3 et 4

La moyenne des R-Squared ajusted pour les 75 courbes pour les 5 types de régression sont respectivement :
0.0000000, 0.4853609, 0.6622269, 0.7202006 et 0.7557633.
On voit que plus le degré du polynome augmente, plus la régression est de qualité au crière du R-Squared (les valeurs prédites se rapprochent des valeurs des observations.

Si on se réfère à la figure ci-dessus, on observe que la plupart des simulations présente une remontée au bout de quelques mois, que certaines simulations sont concaves au lieu d'être convexes, voire que certaines d'entre elles pourraient avoir des valeurs négatives (autour des 35 mois d'exploitation).

2 Régressions exponentielles

Une idée simple pour corriger ces défauts est d'utiliser une autre forme paramétrique pour les simulations. Une suggestion immédiate pour qui a un peu l'habitude de ces courbes est une forme exponentielle du style : oú y est la production, t le mois, et $k0$ et $k1$ deux paramètres á déterminer.

Régression exponentielle avec lm

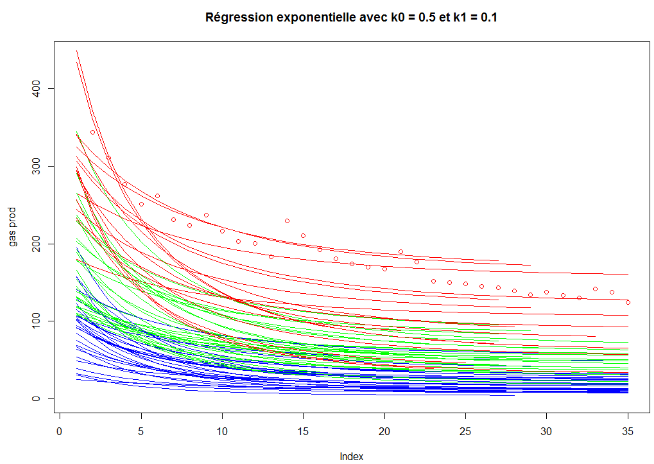

FIGURE 2 – Courbes régressées avec une fonction exponentielle et et la fonction lm de R

Régression exponentielle avec nls

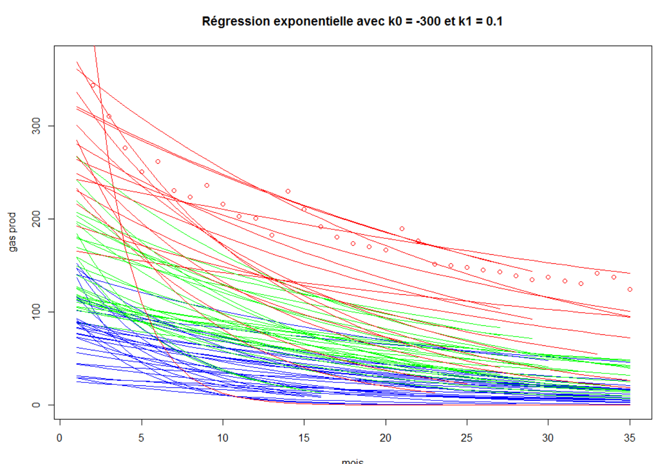

FIGURE 3 – Courbes régressées avec une fonction exponentielle et et la fonction nls de R

La régression avec lm paraît mieux, on s'intéressera donc á celle-ci.

Est-ce que cela marche ? La moyenne du "ajusted R-Squared" pour les 75 régressions exponentielles est cette fois-ci de 0.6754964, donc de moins bonne qualié á priori que les régressions polynomiales de degré supérieures ou égale á 3. Néanmoins, quand on s'intéresse au détail des R-Squared, on trouve de trés bonne régressions avec un R-Squared très bon comme des régressions avec un R-Squared très mauvais. Cela est probablement dû à certaines valeurs ”outliers” qui ne permettent pas á la fonction exponentielle de s’adapter aux courbes. La partie 5 et le lissage des courbes permettra de pallier á ce problème et de probablement monter que la régression exponentielle est adaptée si un travail d’atténuation des ”pics” est fait au préalable. 10 premiers R-Squared : 0.889547, 0.203835, 0.7613715, 0.936559, 0.2696669, 0.726263, 0.1067514, 0.8683789, 0.6881983, 0.8563925.

Avez-vous d’autres idées ? Régression avec fonction inverse, ou gaussienne, ou log ? ou combinaison.

3 Courbes hautes et basses á 95%

Quelles sont les incertitudes sur les régressions des points 2 ? Plus concrètement, on vous demande de tracer pour un exemple de chaque type de courbe, la courbe haute (á 95%) et la courbe basse (toujours á 95%)

Avec predict nls ....

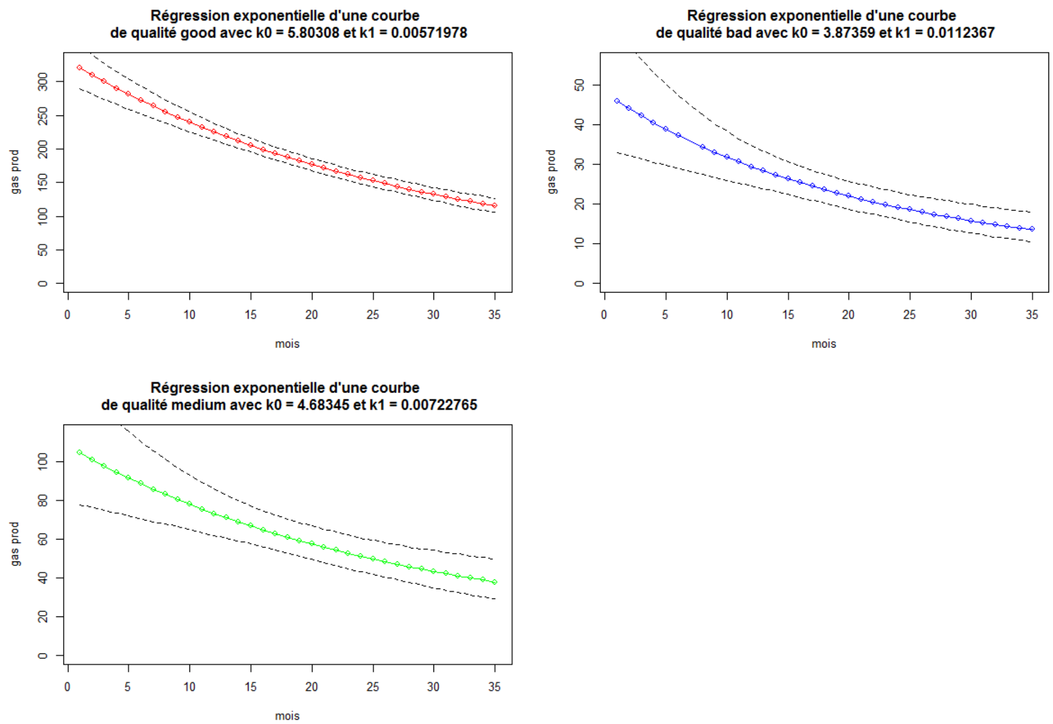

FIGURE 4 – Courbes hautes et basses à 95% pour un exemple de chaque classes de courbes

Avec bootstraping

4 Reclassement avec régression logistique

$k1$ en fonction de $k0$

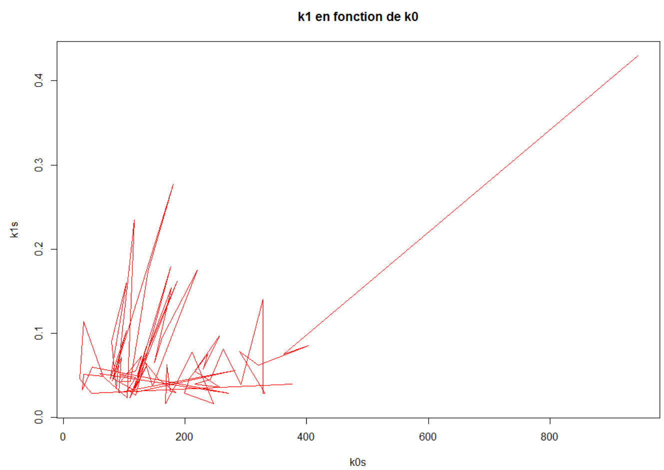

FIGURE 5 – Coefficients $k1$ en fonction de $k0$

En examinant le graphe k1 fonction de k0, on se rend compte que certaines courbes classées 'Good' par les experts donnent l'impression d'être plutôt 'medium', tandis que certaines 'bad' pourraient être aussi 'medium'. // Avez-vous des suggestions sur 5 courbes au plus qui pourraient être mal classées ? Justifiez vos choix (i.e. une façon de faire est d'effectuer une régression logistique dont le y est la classe prédite par l'expert et les x sont les coefficients k0 et k1, et d'examiner comment la régression est améliorée en changeant la classe d'un point ; une autre façon plus empirique est de déterminer deux droites x=k11 et x=k12 qui partitionnent au mieux les classes et d'examiner comme précédemment comment la classification est changée en basculant certains points d'une classe á l'autre).

TODO glm

5 Gestion des spikes et lissage des courbes

Régression polynomiale de degré 3 avec lissage loess

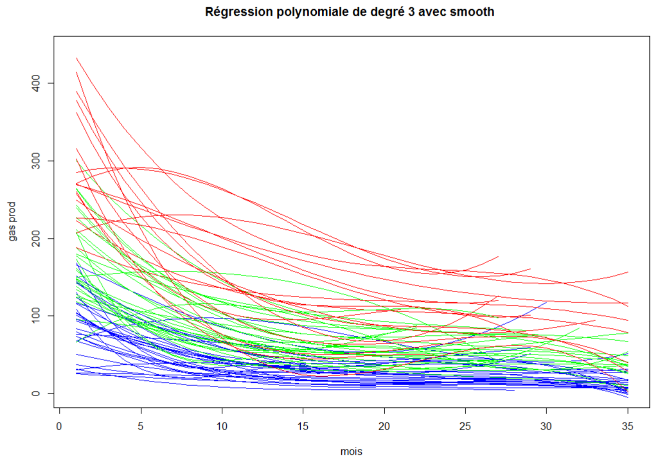

FIGURE 6 – Régression polynomiale de degré 3 avec courbes lissée au préalable avec loess

On obtient une moyenne de R-Squared de 0.9802724 cette fois-ci après lissage des courbes. La régression polynomiale devient ainsi très performante. Il reste à savoir si l'on peut se permettre l'approximation faîte par le lissage des courbes et si les valeurs induites par les spikes avait une pertinence intransigible.

Régression exponentielle avec lissage loess

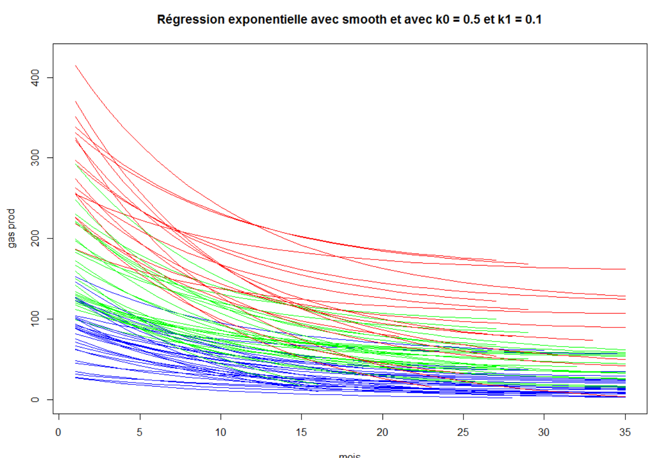

FIGURE 7 – Régression exponentielle lm avec courbes lissée au préalable avec loess

On obtient une moyenne de R-Squared de 0.845492 cette fois-ci après lissage des courbes. La régression n'est pas améliorée avec le lissage des courbes...

6 Conclusion

TODO TOO

Acknowledgments

Merci á François Wahl.