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

14 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 futures données d'extraction. Les courbes des valeurs des données d'extraction donnent une classification de la qualité des puits. L'étiquetage de la qualité de chaque puits à été effectué par des experts. 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 régressions vont être envisagées afin d'avoir la meilleure prédiction/classification possible.

Table des matières

TODOLIST

— EXPLICATIONS

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.

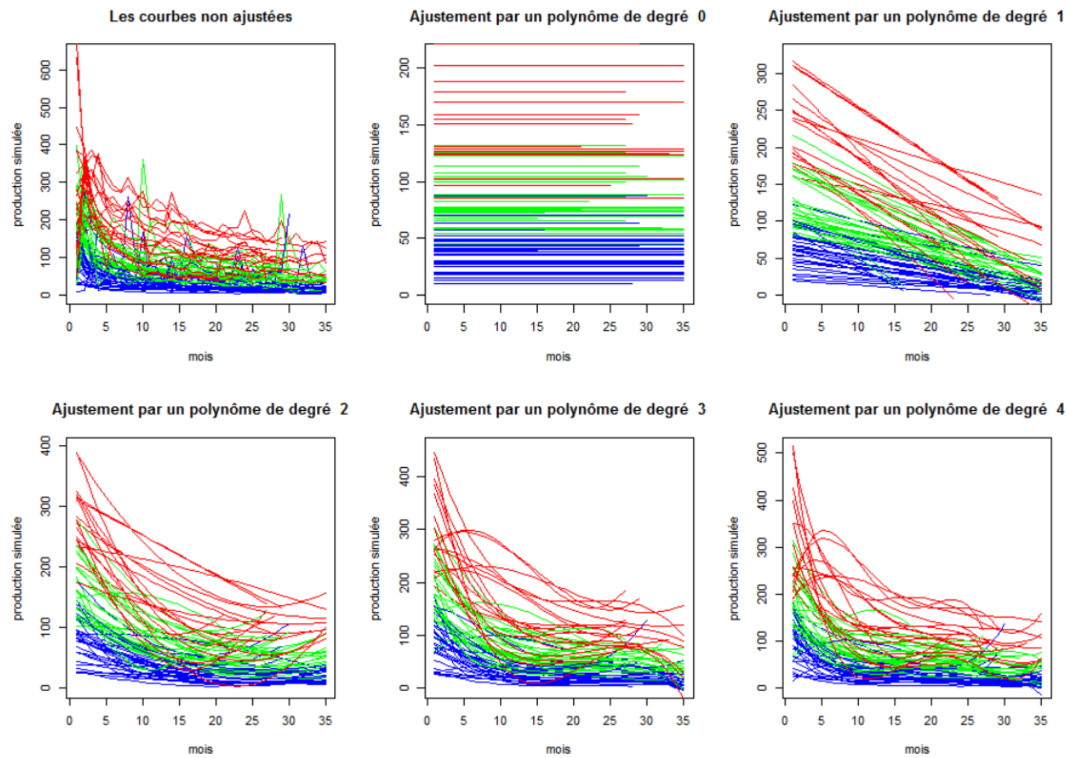

FIGURE 1 – Courbes non régressées, et courbes de régressions polynomiales de degrés 0,1,2,3 et 4

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 critère 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 à déterminer.

Régression exponentielle $log(y) = ax + b$

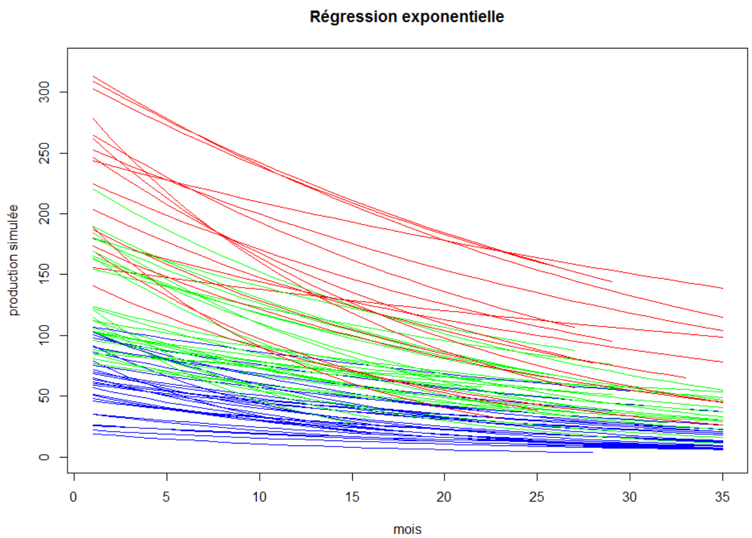

FIGURE 2 – Courbes ajustées avec une fonction exponentielle et l'outil lm de R

Régression exponentielle $y = k_0 e^{-k_1 x}$

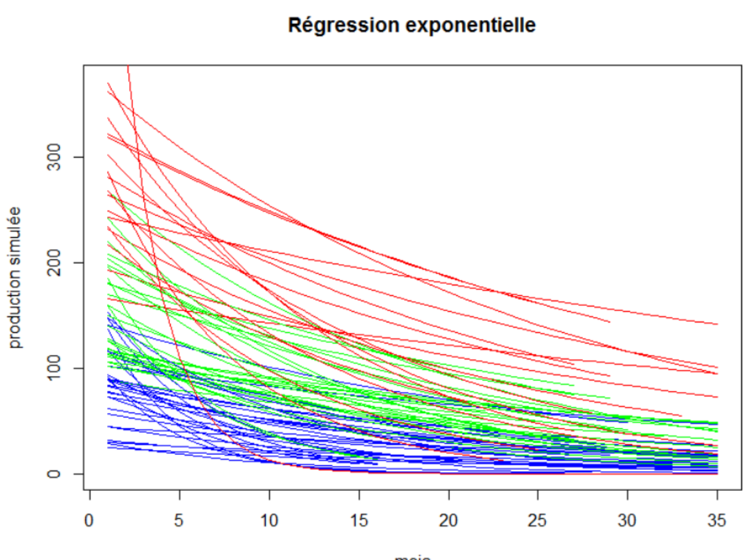

FIGURE 3 – Courbes ajustées avec une fonction exponentielle et l'outil nls de R

La régression $log(y) = ax + b$ faite avec lm est mieux, on s'intéressera donc à 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 à priori que les régressions polynomiales de degré supérieures ou égale à 3.
Néanmoins, quand on s'intéresse au détail des R-Squared, on trouve de très 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 à la fonction exponentielle de s'adapter aux courbes.
La partie 5 et le lissage des courbes permettra de pallier à ce problème et de probablement montrer que la régression exponentielle est adaptée si un travail d'atténuation des "pics" est fait au préalable.

10 premiers R-Squared :
0.8011383, 0.4182548, 0.7403922, 0.7585257, 0.4186529, 0.6888943, 0.3297291, 0.818889, 0.7745934 et 0.7527809.

Avez-vous d'autres idées ?
Régression avec fonction inverse, ou gaussienne ?
Ou bien une composée de polynôme dans un logarithme.

3 Courbes hautes et basses à 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 (à 95%) et la courbe basse (toujours à 95%)

Intervalles de confiance avec nls

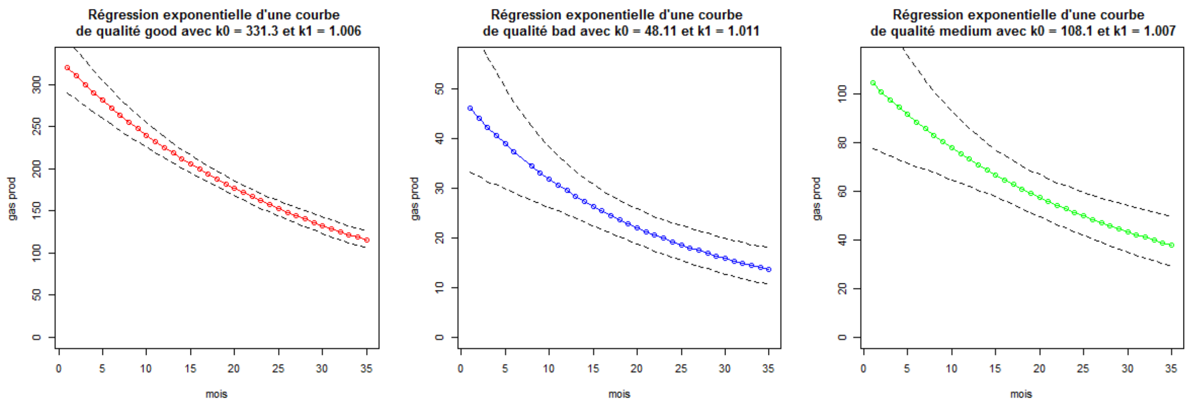

FIGURE 4 – Courbes hautes et basses à 95% pour un exemple de chaque classe de courbes

Intervalles de confiance avec lm

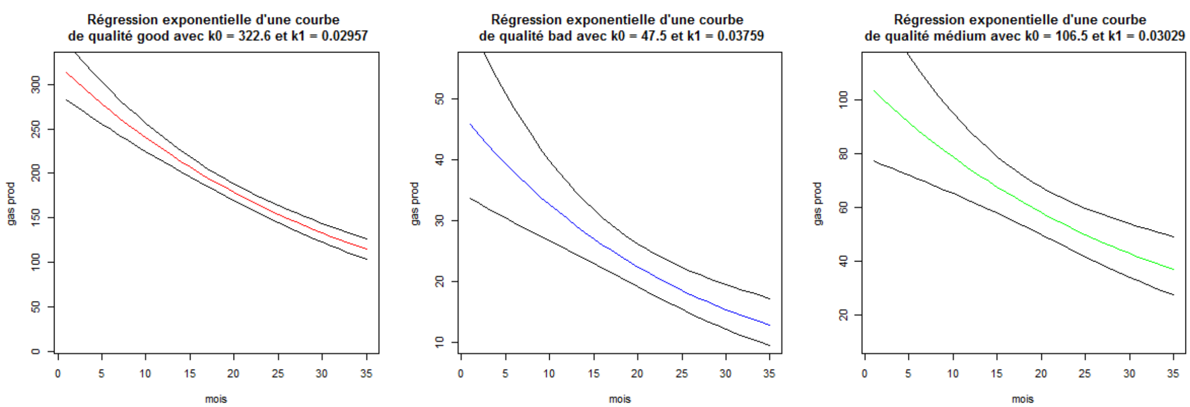

FIGURE 5 – Courbes hautes et basses à 95% pour un exemple de chaque classe de courbes

4 Reclassement avec régression logistique

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).

Clustering des courbes en fonction de $k0$ et $k1$ avant reclassement

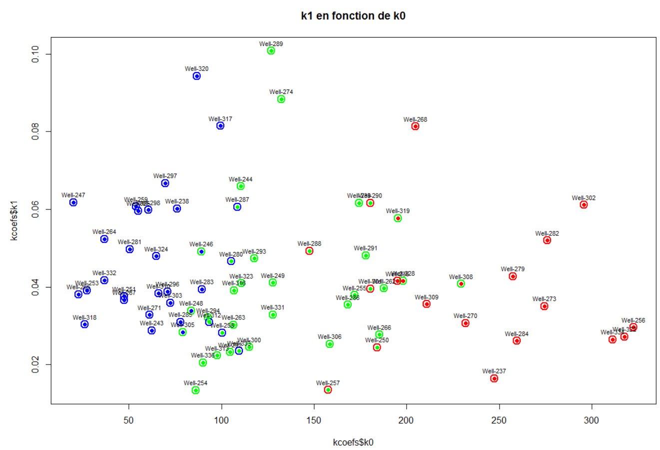

FIGURE 6 – Coefficients $k1$ en fonction de $k0$ et classes des courbes

La couleur du cercle extérieur est la classe de départ (observée) et le cercle intérieur est sa classe prédite.
16 courbes sont mal classées d'après les prédictions.

L'AIC de la régression est de 74.63465.

Après plusieurs tests de changements de courbes et de comparaison d'AIC, on change la classe de 5 puits :
Les courbes des puits '$Well - 288$' de good à médium, '$Well - 333$' de bad à médium, '$Well - 246$' de médium à bad, '$Well - 257$' de good à médium, et '$Well - 258$' de bad à médium.

Clustering des courbes en fonction de $k0$ et $k1$ après reclassement

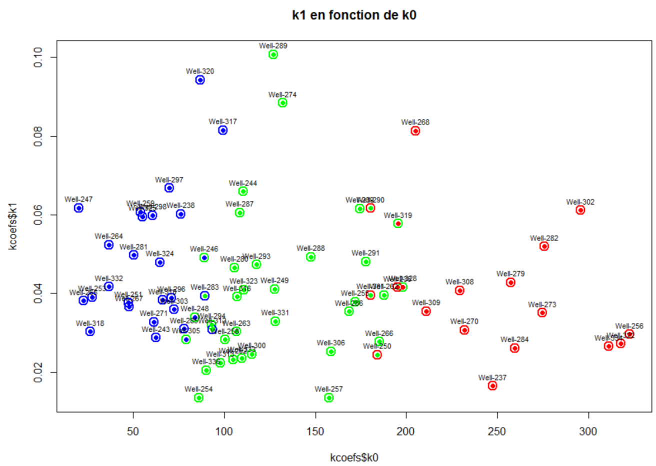

FIGURE 7 – Coefficients $k1$ en fonction de $k0$ et classes des courbes

L'AIC de la régression est de 52.52519.

5 Gestion des spikes et lissage des courbes

Régression polynomiale de degré 3 avec lissage loess

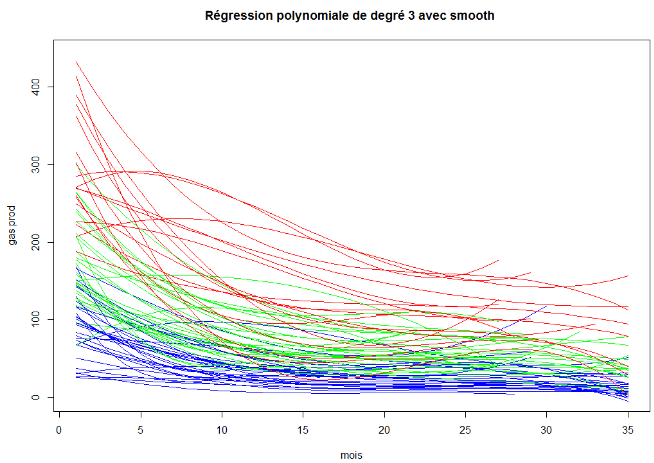

FIGURE 8 – Régression polynomiale de degré 3 avec courbes lissées au préalable avec loess

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 faite par le lissage des courbes et si les valeurs induites par les spikes avaient une pertinence intransigible.
Toutefois, la régression avec smooth ne règle pas les valeurs qui remontent, les simulations concaves au lieu d'être convexes, et les potentielles valeurs négatives autour des 35 mois.

Régression exponentielle avec lissage loess

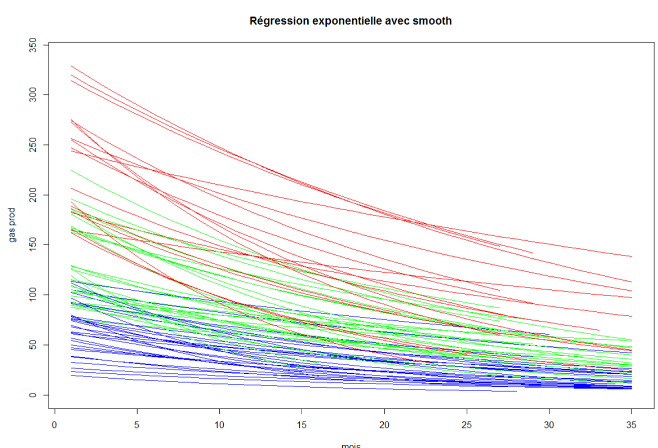

FIGURE 9 – Régression exponentielle lm avec courbes lissée au préalable avec loess

On obtient une moyenne de R-Squared de 0.7569813 cette fois-ci après lissage des courbes. La régression est sensiblement améliorée avec le lissage des courbes mais reste moins efficace que la régression polynomiale de degré 3.

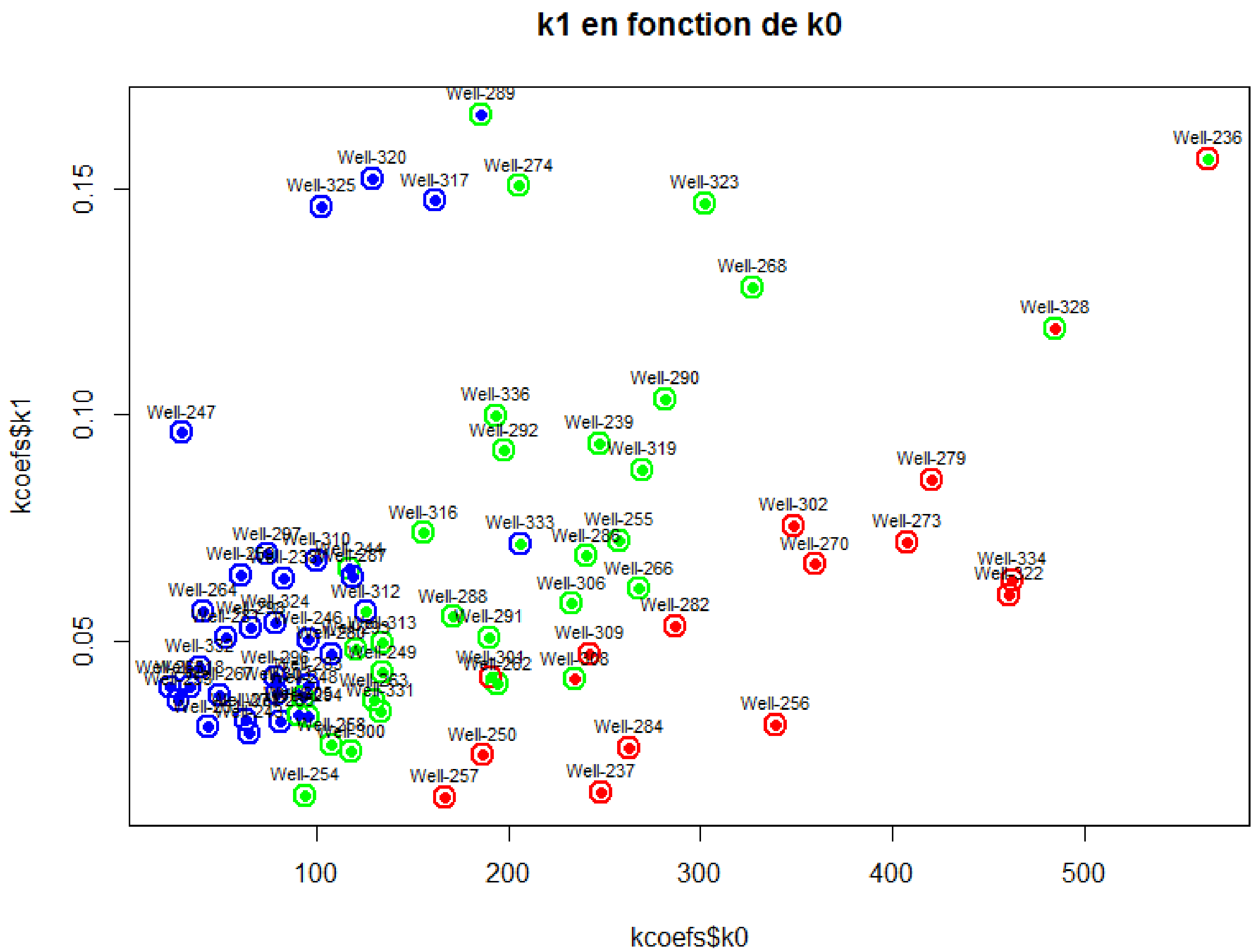

FIGURE 10 – Clustering sur modèle exponentiel avec courbes lissée

6 Conclusion

TODO TOO

Acknowledgments

Merci à François Wahl.