

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

<b>1</b>	<b>Régressions polynomiales</b>	<b>2</b>
<b>2</b>	<b>Régressions exponentielles</b>	<b>3</b>
<b>3</b>	<b>Courbes hautes et basses à 95%</b>	<b>5</b>
<b>4</b>	<b>Reclassement avec régression logistique</b>	<b>6</b>
<b>5</b>	<b>Gestion des spikes et lissage des courbes</b>	<b>8</b>

# 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.

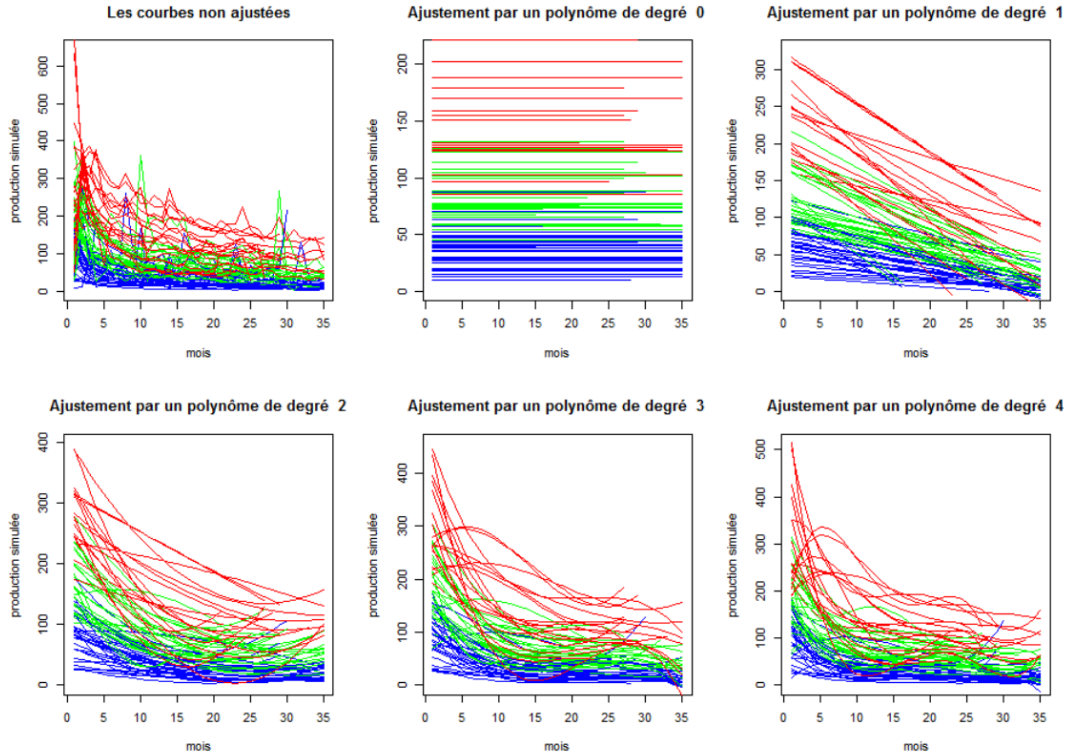


FIGURE 1 – Courbes non régressées, et courbes de régressions polynomiales de degrés 0,1,2,3 et 4

La moyenne des R-Squared ajustés 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 polynôme 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 :  $y = k_0.e^{-k_1.t}$  où  $y$  est la production,  $t$  le mois, et  $k_0$  et  $k_1$  deux paramètres à déterminer.

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

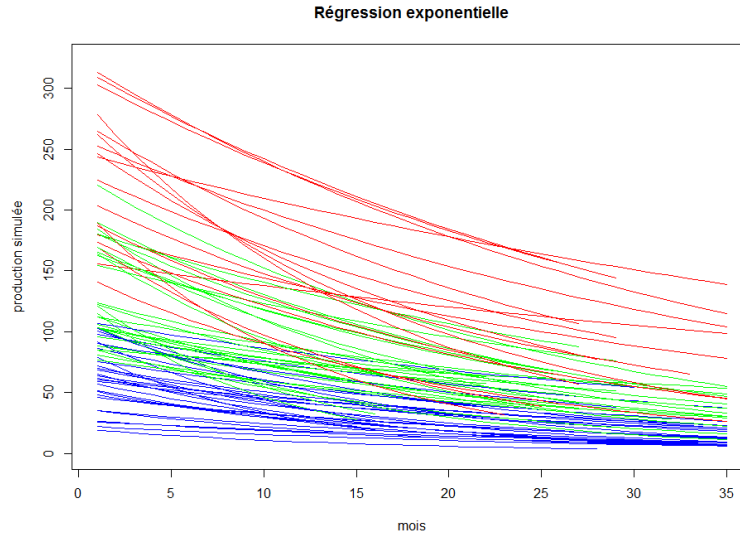


FIGURE 2 – Courbes ajustées avec une fonction exponentielle et l'outil lm de R

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

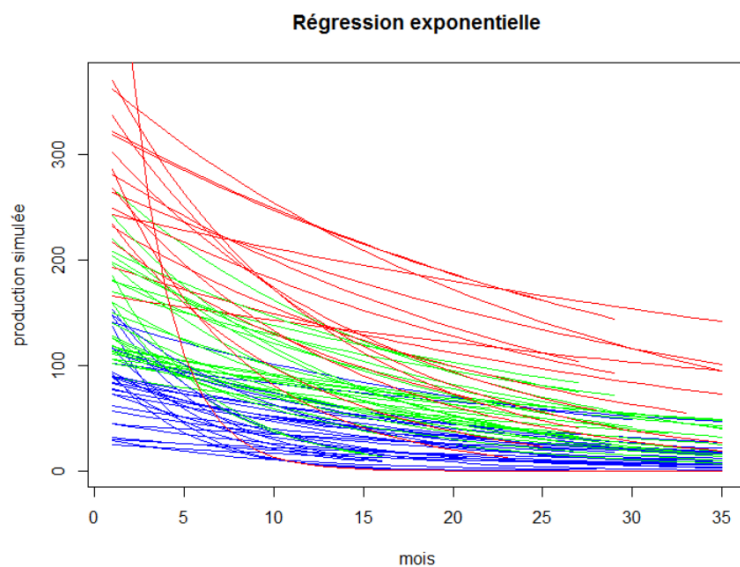


FIGURE 3 – Courbes ajustées avec une fonction exponentielle et l'outil nls de R

La régression  $\log(y) = ax+b$  faite avec `lm` est plus fidèle aux données, on s'intéressera donc à celle-ci.

La moyenne du "adjusted R-Squared" pour les 75 régressions exponentielles est cette fois-ci de 0.6754964, donc de moins bonne qualité 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 ce problème et probablement de 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.

On peut d'ores et déjà constater que les trois classes de puits définies par les experts sont visuellement distinguables, bien qu'un peu entremêlées. Ceci indique qu'il sera possible de mettre en place un modèle de classification des puits.

### **Autres possibilités**

Nous avons pensé à un ajustement par fonction inverse du type  $y = \frac{a}{x} + b$  avec  $y$  la production,  $x$  le mois et  $a$  et  $b$  des constantes à déterminer. On peut aussi penser à une gaussienne, dont l'allure s'approche de celles des courbes.

### 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

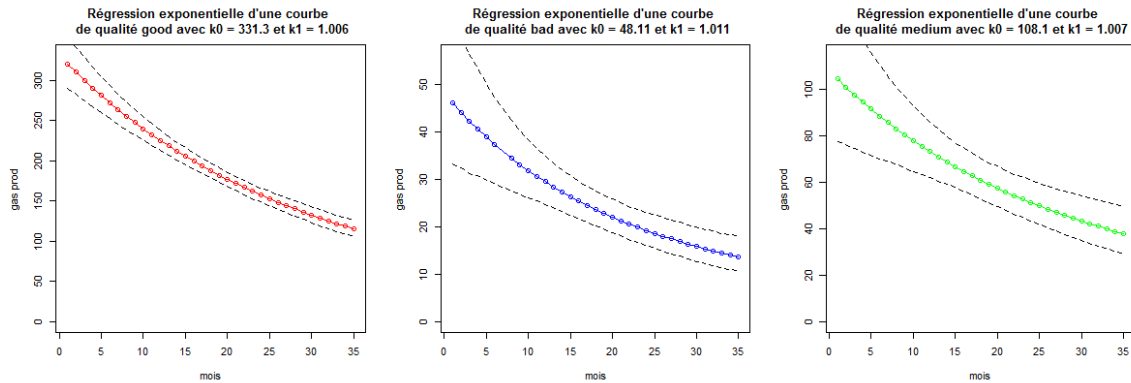


FIGURE 4 – Courbes hautes et basses à 95% pour un exemple de chaque classe de courbes

#### Intervalles de confiance avec lm

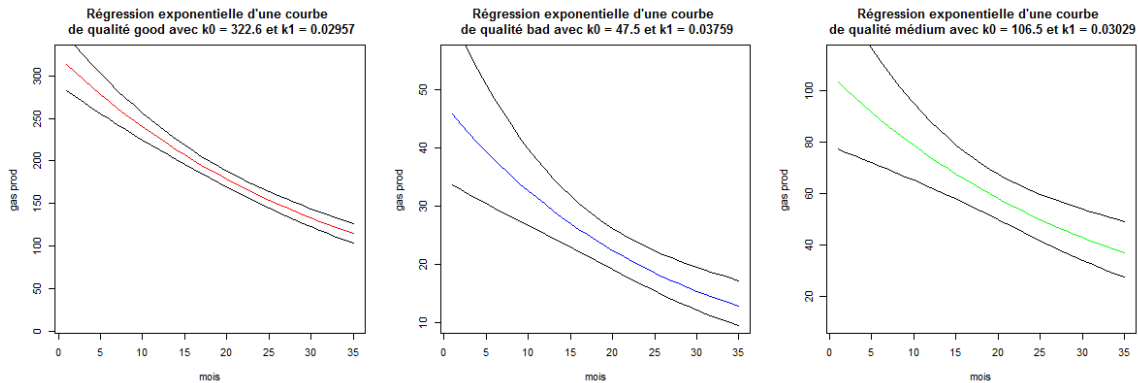


FIGURE 5 – Courbes hautes et basses à 95% pour un exemple de chaque classe de courbes

Comme le montre les figures ci-dessus, il y a davantage d'incertitude sur les valeurs des courbes de qualité 'bad' et 'médium'. Cela induit que les valeurs de ces courbes suivent des tendances plus compliquées à ajuster. En effet, les valeurs d'extraction pour ces puits sont probablement plus imprévisibles et suivent parfois quelques oscillations. En outre, les incertitudes aux valeurs d'extraction pour les premiers et derniers mois sont également plus fortes.

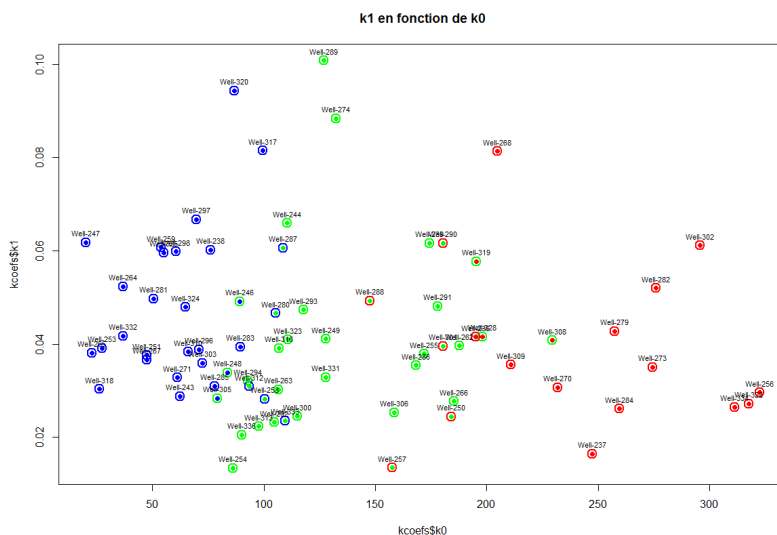
## 4 Reclassement avec régression logistique

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  $k_0$  et  $k_1$ , et d'examiner comment la régression est améliorée en changeant la classe d'un point).

### Clustering des courbes en fonction de $k_0$ et $k_1$ avant reclassement

FIGURE 6 – Coefficients  $k_1$  en fonction de  $k_0$  et classes des courbes

La couleur du cercle extérieur est la classe donnée par les experts et le cercle intérieur est sa classe prédite par le classifieur.

16 courbes sont mal classées d'après les prédictions et l'AIC de la régression est de 74.63465.

On essaye d'améliorer le taux du classifieur en modifiant la classe prédite par les experts sur certaines courbes.

On change ainsi la classe de 5 puits.

On se limite au reclassement de 5 puits car on ne veut pas trop modifier le modèle de départ et laisser de la souplesse au classifieur si celui-ci doit traiter la classification de nouveaux puits à l'avenir.

On procède de manière itérative. On choisit un puits qui semble très éloigné du reste de sa classe et on change sa classe donnée par les experts pour celle prédite par le classifieur. On relance le modèle du classifieur sur cette nouvelle base de données. On compte les points mal classés avec ce nouveau classifieur. Si le puits choisi ne diminue pas le nombre de points bien classés, on lui réattribue sa valeur originelle et on choisit un autre point. Quand on trouve un point qui convient, on répète ces

opérations jusqu'à atteindre 5 points.

On aurait pu automatiser cette opération qui est un problème d'optimisation simple. Ce problème peut se résumer par la phrase suivante :

*Trouver la séquence de 5 puits qui minimise le nombre de puits mal classés*

Après interprétations graphiques et des tests de prédiction de classe avec ou sans changement de classe des 16 puits mal classés, on détermine les 5 puits qui réajustent au mieux le modèle et améliorent le classifieur :

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

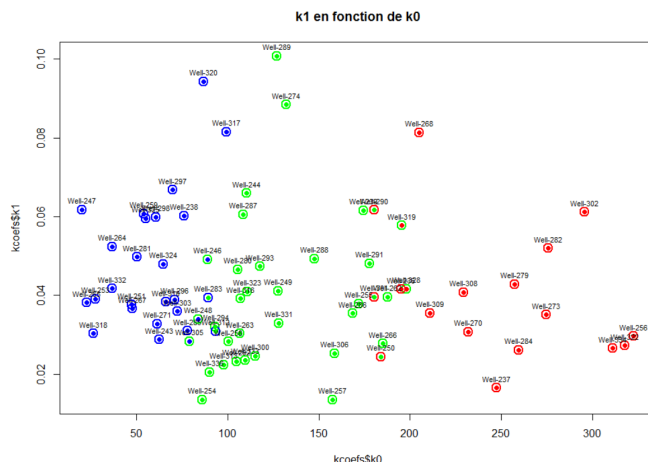


FIGURE 7 – Coefficients  $k1$  en fonction de  $k0$  et classes des courbes

L'AIC de la régression est de 52.52519 et il n'y a plus que 10 courbes mal classées pour 5 de changées sur les 16 de départ. Soit 1 courbe qui dont la prédiction de classe par le classifieur s'est accordée sur sa vraie classe.

A présent que nous avons établi un classifieur, nous pouvons nous demander s'il est possible d'en améliorer la précision. On peut par exemple tenter de produire un meilleur modèle sur les données en lissant les courbes avant d'effectuer la régression.



## 5 Gestion des spikes et lissage des courbes

### Régression polynomiale de degré 3 avec smooth

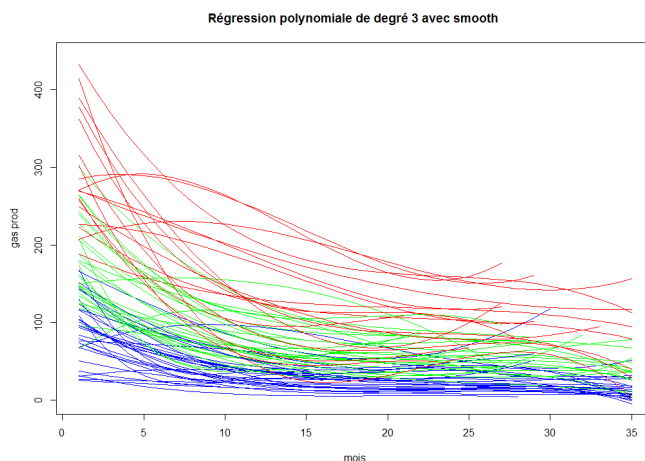


FIGURE 8 – Régression polynomiale de degré 3 avec courbes lissées au préalable avec loess

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 des spikes

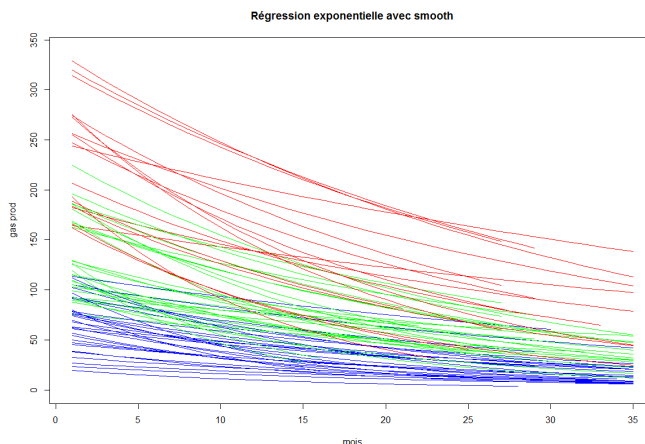


FIGURE 9 – Régression exponentielle lm avec courbes lissée au préalable avec loess

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.

On effectue à nouveau une régression logistique à partir des courbes lissées comme à la question 4. On obtient cette fois 16 courbes mal classées, un AIC de 85 et le graphique suivant :



9

## Acknowledgments

Nos remerciements vont à François Wahl pour l'enseignement de son cours "Modèles de régression" à l'Université Claude Bernard Lyon I et son accompagnement durant ce projet.