Le jeu de données est construit à partir des données de production de 75 puits dans un champ pétrolier du Canada. Les données sont publiques.

Tous les puits ne fournissent pas des informations sur la même période, mais les données récupérées ont toutes été ramenées à une durée de production de 36 mois.

Si certaines valeurs sont manquantes, elles sont remplacées par 0. C'est le cas quand le puits n'a pas communiqué son activité pendant un ou quelques mois, ou quand il l'a interrompu, ou que le puits a cessé de produire avant le terme des 36 mois.

Les productions ont été classées en 3 classes, 'bonne' 'moyenne' ou 'mauvaise', dessinées en rouge vert et bleu respectivement sur la figure 1.

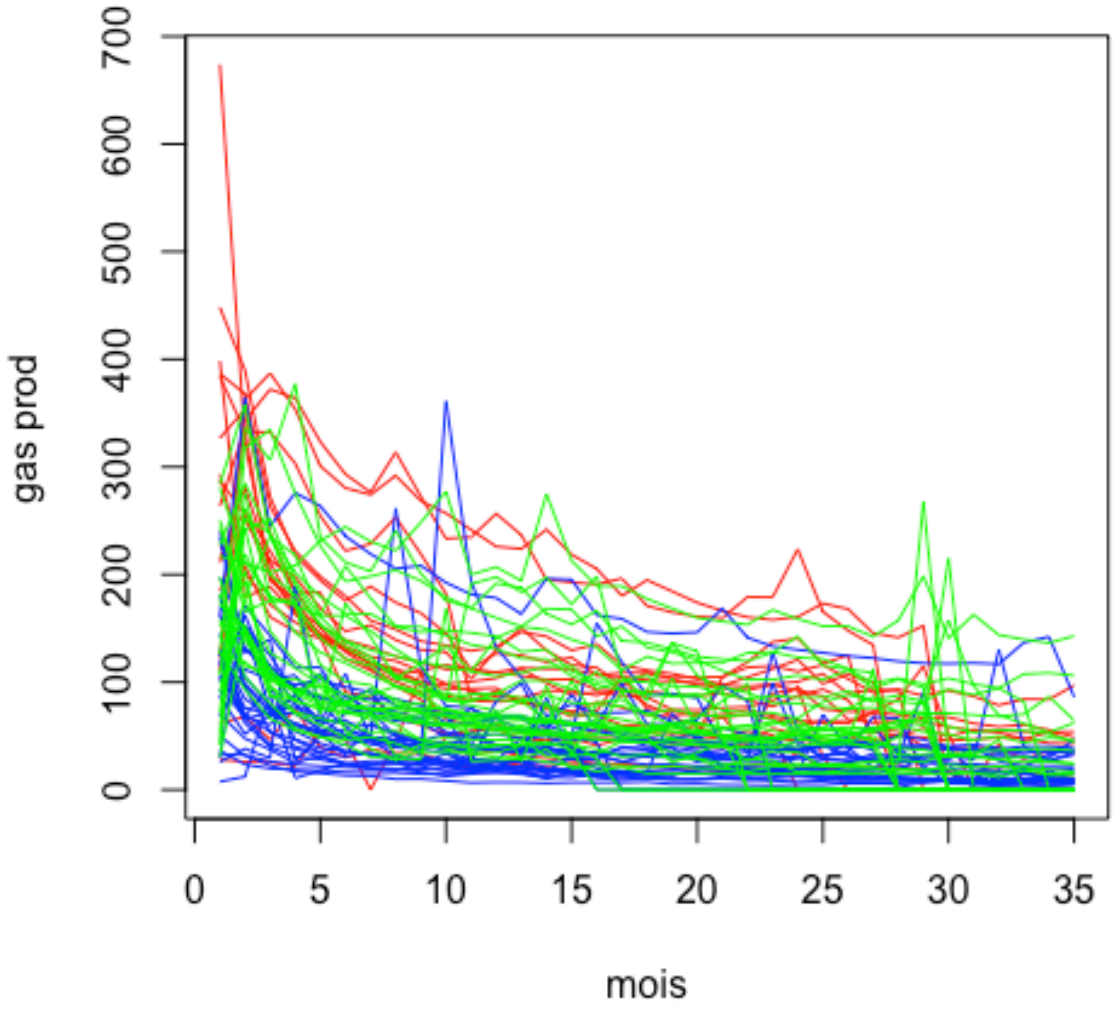

figure 1 : les données de production

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.
1) 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.

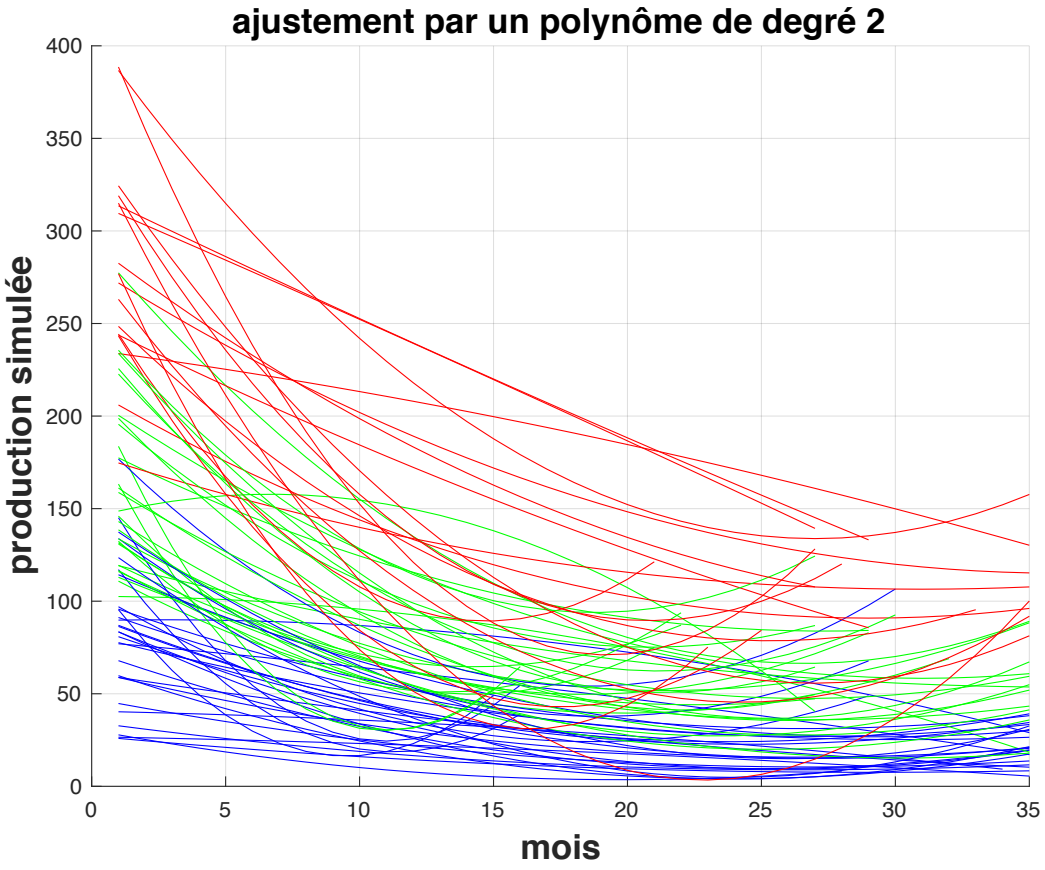

figure 2 : les courbes de production obtenues avec des polynômes de degré 2

2) Les courbes obtenues avec des polynômes ont des allures un peu choquantes pour l'œil d'un expert. 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).

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 : $y = k_0 \exp(-k_1 t)$, où y est la production, t le mois, et k_0 et k_1 deux paramètres à déterminer.

Est-ce que cela marche ? avez-vous d'autres idées ?

3) 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%)

4) En examinant le graphe k_1 fonction de k_0, 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 ; 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 à l'autre).

5) Les courbes de production présentent des 'spikes', c'est à dire des pointes soudaines qui peuvent correspondre à des problèmes de mesure. Dans ce genre de situation, il n'est pas anormal de lisser ('smoothing curves') les courbes de manière à enlever les spikes ou au moins à les atténuer. Plusieurs packages R font cela très bien. On vous demande d'en choisir un de le mettre en œuvre, et de réessayer les régressions avec un modèle polynômial de degré 3 et avec le modèle exponentiel du point 2.