

# Avancement Stage TDF

26 février 2014

## 1 Mercredi 22/01

Première séance, il y a eu une prise de contact entre Emile, Baptiste et Hoël afin de jeter les bases du stage, présentations, présentation du projet, nous avons également légèrement survolé les procédés que nous allons mettre en oeuvre durant cette période.

Puis nous avons travaillé sur l'outil Matlab afin de déterminer certaines caractéristiques des données cf stage1.m, dans le but de voir ce qui était le plus pertinent, et de réfléchir si nous aurions à demander d'autres informations à Tdf, et le cas échéant lesquelles.

## 2 Vendredi 24/01

Prise de contact de Laure Quivy avec les étudiants, les premières estimation de dates ont été jetées, un code grossier devrait être fourni pour le mois de Mai (TDF a une réunion concernant les tarifs, ils aimeraient présenter ça à ce moment là), puis si on peut l'améliorer, continuer pour ne fournir une version plus performante en Juin ou Juillet.

Mise en place du Git hub en commun, vision ou révision de Matlab, méthodes d'interpolation, et de régression pour les deux étudiants.

## 3 Mercredi 29/01

### 3.1 Réunion à TDF

Aujourd'hui nous avons eu la réunion avec Olivier Marzouk de TDF, pour lui présenter les premières analyses de données.

Apparemment la donnée qui sensiblement pourrait nous manquer porte sur la date de construction des différents sites, il faut donc voir par la suite si elle explique les points étranges. Nous n'aurons pas accès au code du logiciel de l'ARCEP, mais une partie du calcul est disponible en ligne (à voir sur le lien suivant que nous a fourni Olivier : <https://docs.google.com/open?id=0B4dfKcelvACpNFJIdDN2SEFXN00>).

Le premier nombre de panneau serait à priori inutile, c'est plutôt le nombre maximum de panneaux (donc utiliser  $X(:,2)$  plutôt que  $X(:,1)$ ).

### 3.2 Hors réunion

Il reste à vérifier si l'apparence de droites des graphes  $\text{scatter}(X(:,3), Y(:,1))$  exhibées par Baptiste reste valable en faisant varier les valeurs de  $X(:,2)$ ,  $X(:,4)$  et  $X(:,5)$ . Si c'est le cas, il ne reste plus qu'à trouver ce qui fait varier le coefficient directeur de la droite, et on obtient une relation affine entre le prix et les critères... ce qui paraît trop simple.

Les valeurs étranges ont été envoyées à Olivier afin qu'il vérifie qu'elles ne sont pas fausses (l'erreur est humaine).

Hoël a exhibé les configs qui apparaissent étranges sur les graphes (du moins celles qui sautent aux yeux), c'est-à-dire ceux qui ont des caractéristiques proches d'autres sites mais qui présentent avec eux un écart de prix d'au moins 25

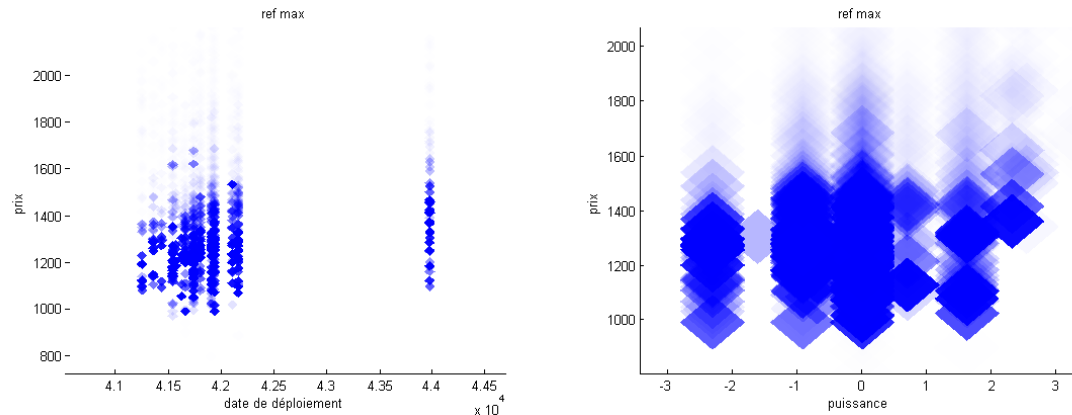
Baptiste a écrit un script permettant d'exhiber toutes les configs qui ont les mêmes caractéristiques mais pas le même prix, il y a 391 groupes comprenant chacun au moins 2 configs qui correspondent à cette description. Cela confirme qu'il nous faut d'autres critères.

## 4 Vendredi 14/02

Aujourd'hui Hoël a fait de la biblio sur "The Elements of statistical learning" sur la régression et la méthode des moindres carrés. Baptiste de son côté a étudié la régression par noyau (Kernel regression).

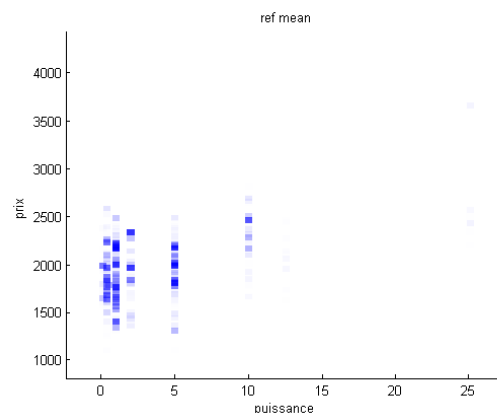
De plus des plots avec transparence (basée sur la distance à un point de référence après normation) ont été tracés en fonction de chacune des coordonnées, avec deux points de référence différents : la ref mean, où la référence est le point nul (sauf pour la coordonnées qui varie et pour celles qui paraissent moins utiles ie colonnes 1 5 et 6), et la ref max, où les coordonnées sont celles les plus représentées indépendamment.

Le prix ne semble pas influencé par les données puissance et date de déploiement (4 et 7). Le nombre de panneaux et la hauteur semblent par contre intéressants..



## 5 Mercredi 19/02

Aujourd'hui un script avec arguments optionnels a été écrit afin de faire des plots avec transparence bien plus facilement. On a ainsi exhibé une certaine corrélation entre la puissance et le prix (simple croissance globale), et que le réseau RC ne comprenait pas de config dont le prix est très élevé.



D'autre part Hoël a fait de la biblio sur les splines. Il faut s'intéresser aux smoothing splines pour obtenir une résistance au bruit sur les données. Il y a alors une fonction à minimiser, comportant des moindres carrés et un poids sur l'intégrale de la dérivée seconde. Le problème étant qu'il n'y a pas d'algorithme fonctionnant sur tous les datas sets donc ce sera du cas par cas. Les splines sont déjà codées sous matlab (cf fit pchip etc...) donc il faudra regarder ce que ça donne. (pour une base cf smoothing splines sur google)

Le code excel de l'ARCEP présente énormément de données que TDF ne nous a pas fournies, dont une bonne partie en libre accès sur le document en ligne. Il faut donc regarder si ces données peuvent avoir une influence (rapelons nous qu'il nous manque un élément déterminant, et que celà étant donné on se rapprocherait énormément d'un calcul multilinéaire...).

Baptiste a écrit plusieurs scripts permettant de faire des regressions kernels sur les datasets

## 6 Mercredi 26/02

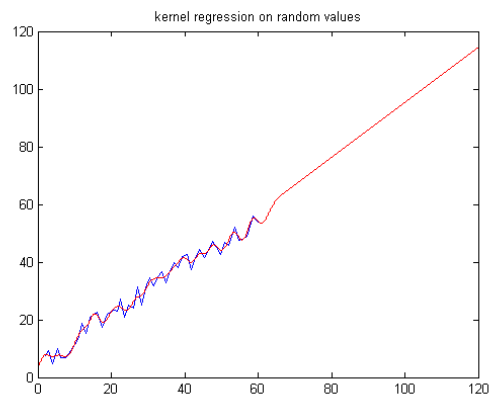
Aujourd'hui, une hypothèse sur le temps de calcul phénoménal du code de l'Arcep a été émise, après s'être rendu compte que le fichier mettait plusieurs minutes rien que pour être ouvert.. On suppose qu'il a beaucoup trop d'appels, et qu'il est beaucoup trop compliqué, et qu'on pourrait obtenir une précision suffisante en enlevant beaucoup de critères qu'il considère.

Dû à la complexité dudit fichier, il paraît complexe de travailler dessus (ne serait ce que d'extraire des pages à isoler) car libre office ne supporte pas de fichiers aussi lourds, il faudrait donc essayer avec excel.

Le fichier permettant de plotter avec transparence a été amélioré, en refaisant les légendes, une option pour choisir sa figure, et la taille des éléments plottés a été refaite. Il paraît donc plus ou moins complet pour l'instant.

Une page semblant intéressante du fichier Excel a été extraite (d'ailleurs il faut des licences Excel car on arrive aux limites de libre office), mais n'a pas été transférée sous matlab, parce qu'il y a plus de 200 critères à prendre en compte, donc c'est du travail lourd à faire à la main.

La priorité actuelle est de travailler sur de la regression Kernel, dans cette optique le script de Baptiste sur le premier script kernel linexp a été corrigé.



Maintenant il faut écrire différents scripts permettant de lancer un certain nombre (à entrer en tant que paramètre) de kernels en prenant différentes parties des datasets pour train et test. Puis calculant l'erreur sur le data test, faisant la moyenne de ces erreurs sur des valeurs lambda, sigma, et mu données. Puis grâce à des boucles déterminer les valeurs de lambda, sigma, et mu optimales. Pour l'instant on fait des boucles avec un nombre restreint de valeurs pour les paramètres, pour les faire tourner sur nos ordinateurs, puis on fera de plus grosses boucles pour les faire tourner sur le serveur pour améliorer les résultats.