
Le Vendredi 1 Décembre 2017

–

MTH2302D

# Travail de session – Partie 2

## Analyse des données

Emir K. Belhaddad (1825569)

Anthony Dentinger (1718526)

# Contexte général

**Tetris** fut le jeu le plus vendu pour la console rétro *GameBoy* [1]. Il est encore joué ajourd'hui. C'est pour cette raison que l'on a choisit de faire une analyse statistique sur ce jeu. Nous nous somme basés sur les données d'un site internet, **Tetris Friends** [2], qui permet à des joueurs, ayant ou pas un compte, d'avoir accès à différentes version du jeu. Le site donne aussi accès à des statistiques sur les parties des utilisateurs possédant un compte. Pour notre analyse statistique, nous avons choisit la version **Tetris 1989** [3] du jeu. Dû aux différences majeurs entre les versions, nous avons estimé qu'il serait difficile d'obtenir des résultats possédant une réelle signification si l'on prenait nos données de toutes les versions de Tétris confondues.

# Provenance, obtention et description des données

La section statistique de Tetris Friends contient les données d'une partie d'un joueur ayant un compte:

- Pseudonyme du joueur
- Score obtenu (**entier** entre 0 et 999 999 – valeur discrète)
- Niveau final (**entier** entre 0 et 20 – valeur discrète)
- Nombre de lignes complétées (**entier naturel** – valeur discrète)
- Temps de jeu (mesuré à la **centiseconde** près – valeur continue)
- Nombre de lignes complétées par minute (**réel** – valeur continue)
- Nombre de tétriminos (blocs Tetris) déposés (**entier naturel** – valeur discrète)
- Nombre de tétriminos par minute (**réel** – valeur continue)
- Nombre de combinaisons de lignes complétées (pour simples, doubles, triples et tetris) (**entiers naturels** – valeur discrète).

Nous avons également déterminé la proportion de la contribution de chaque type de combinaison de lignes aux nombre de lignes complétées total pour chaque type de combinaison.

Puisque le site web ne fournit pas de moyen de récupérer et enregistrer les statistiques, nous avons écrit un **script JavaScript** (disponible sous 1825569_1718526.js) injectable qui récupère automatiquement les données des parties jouées pendant une journée et produit un texte au format CSV qui peut facilement être copié dans un fichier. La récupération des données s'est faite **tous les jours entre 16h55 et environ 18h30 du 24 Septembre au 30 Septembre, sauf le vendredi 29**. Nous avons donc obtenu **579 statistiques de parties jouées**.

## Questions ouvertes

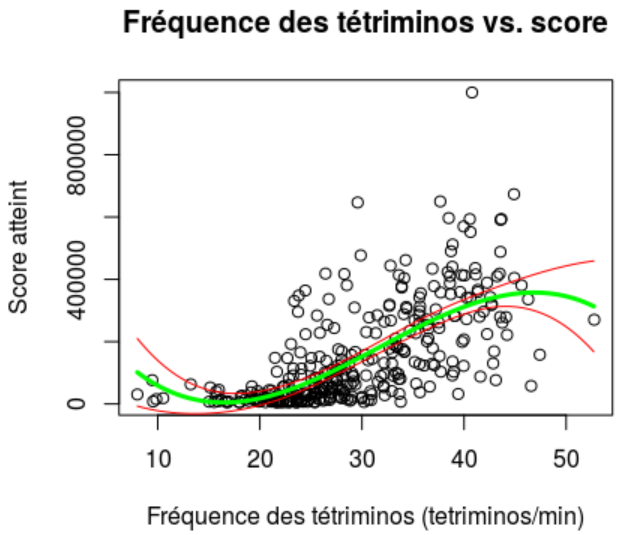
Notre analyse statistique nous a permis de répondre à des questions concernant des **stratégies maximisant le score**. Voici les questions auxquelles nous répondront dans ce rapport:

- Les joueurs rapides font-ils plus de point?
- Quelle est la vitesse optimale à laquelle il faut placer les tétriminos?
- Quelle stratégie de complétion de lignes (simple, double, triple ou tetris) faut-il avantager?
- Est-ce qu'une partie plus longue signifit que l'on va atteindre un niveau plus élevé?

# Analyse des données

### 1 et 2) Influence de la fréquence des tétriminos sur le score

Étant donné que nous souhaitions obtenir la fréquence optimale de tétriminos qui permettrait d'augmenter le score, nous avons cherché une corrélation entre les deux variables aléatoires considérées : la fréquence des tétriminos et le score. Nous avons choisi une **régression cubique** puisque ni la régression linéaire ni la quadratique ne permetaient d'obtenir une fréquence optimale. La figure suivante illustre la régression déterminée par notre script R, ainsi que les régressions minimales et maximale avec un seuil de confiance de 95%.




Ce graphique X-Y est approprié puisqu'il **permet de représenter visuellement à la fois les données et les régressions**. Nous pouvons par exemple voir que la valeur maximale de la fréquence

de tétriminos est guidée **en assez grande partie par les 4 ou 5 observations les plus à droite de notre graphique**, ce qui nous indique que la fréquence obtenue ne sera pas statistiquement significative.

Nous avons décidé de ne considérer, dans cette étude, que les parties durant lesquelles au moins 100 tétriminos ont été déposés sur la pile. En effet, **les parties perdues quasi-immédiatement ne nous intéressent pas** dans notre étude puisque nous cherchons à déterminer une stratégie de jeu pour battre ses records, or il est possible qu'avant de jouer une bonne partie, un joueur fasse quelques parties perdues très rapidement.

Le **modèle théorique de régression vu en cours demande que la variance des données soit uniforme.** Nous constatons ici que **ce n'est pas le cas** puisqu'une grande partie des données est répartie dans la zone [15, 30] x [0, 100000], alors que les données ailleurs sont beaucoup plus dispersées. Cela est bien-sûr normal puisque le score et la fréquence des tétriminos n'ont pas une relation triviale, et cela ne nous empêche pas de néamoins tenter d'estimer la fréquence optimale. De plus, la régression est assez peu significative ($R^2 = 0.4581$). Nous devons tout de même garder cela à l'esprit lors de l'évaluation de la régression.

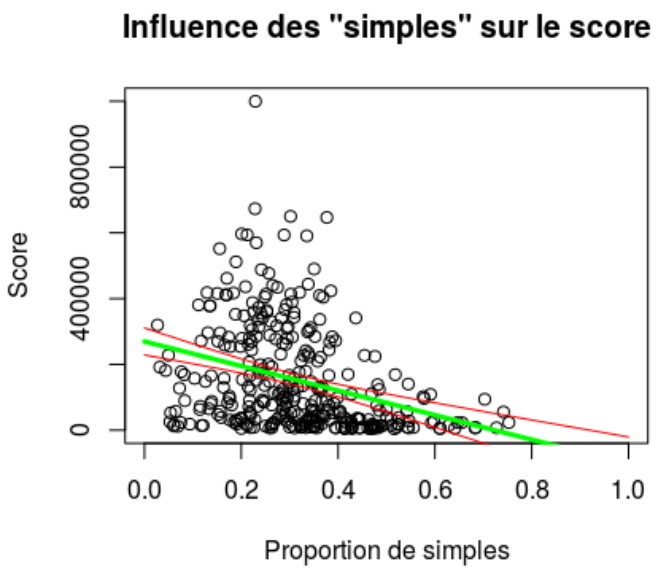
Nous avons finalement obtenu la fréquence optimale prédite par chacune des trois courbes de régression présentées sur la figure précédente. Les résultats sont présentés au tableau ci-dessous.

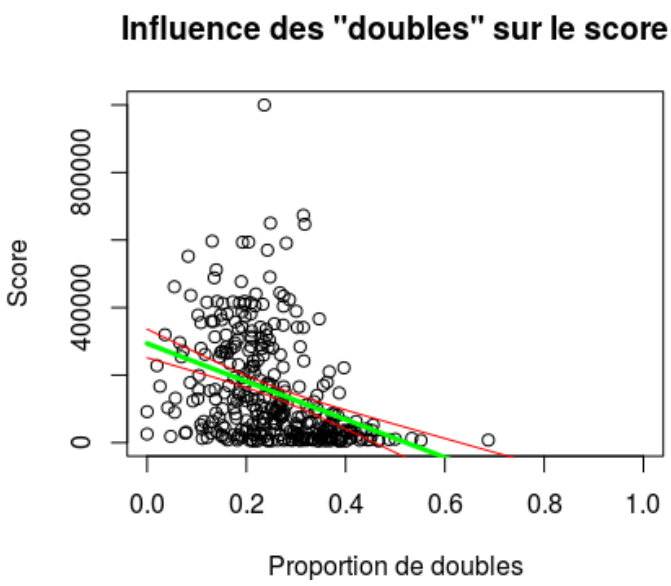
| | Fréquence de tétriminos optimale | Score optimal prédit |
|---|---|---|
| Courbe à 2.5% | 44.1 | 313092.1 |
| Courbe prédite | 46.9 | 357447 |
| Courbe à 97.5% | 52.7 | 459088.2 |

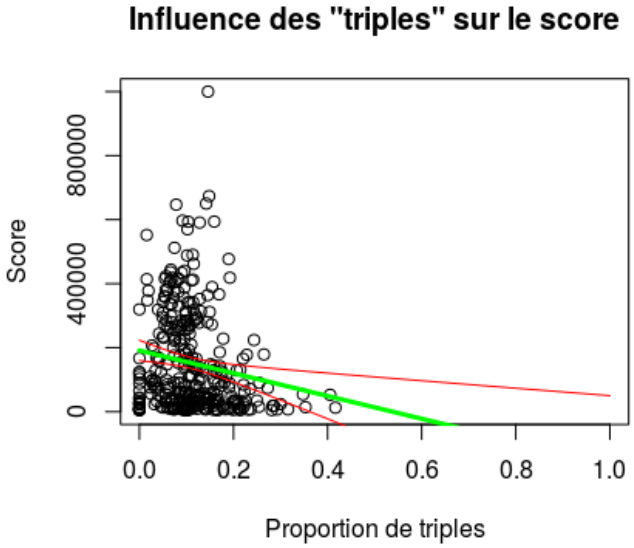
Comme nous l'avons mentionné plus tôt, ces fréquences ne sont pas très significatives statistiquement puisque nous avons relativement peu d'observations au-dessus de 45 tétriminos/min.

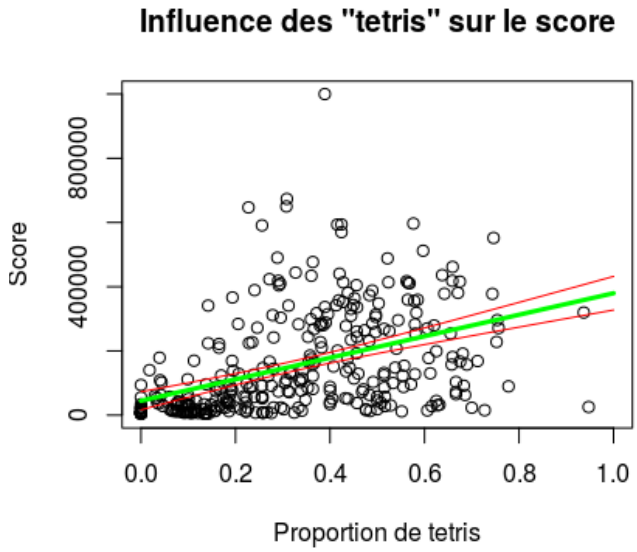
## 3) Combinaison à privilégier pour optimiser le score

Nous avons suivi une **méthode presque identique** que celle utilisée pour les questions 1 et 2, sauf que nous avons effectué une régression linéaire. Comme précédemment, nous avons **filtré les parties** dont le nombre de tétriminos est inférieur à 100. Les quatre graphiques 2D suivants illustrent les résultats obtenus. Comme précédemment, les régressions extrêmes sont à 2.5% et 97.5%.



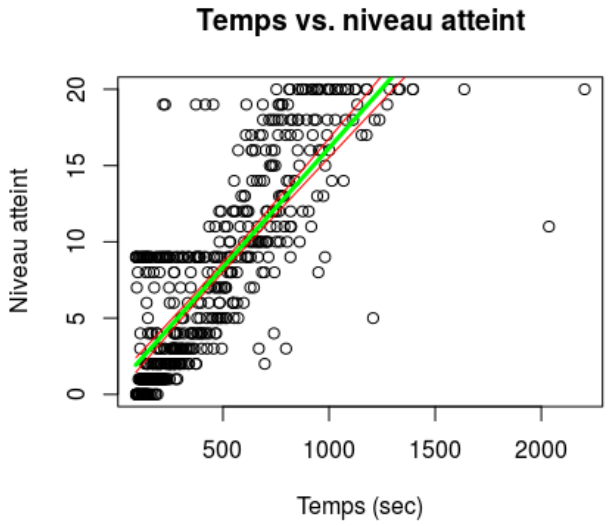

Nous avons choisi de déterminer également le coefficient de corrélation `r` pour les quatre régressions linéaires. Les résultats sont présentés dans le tableau ci-après.

| | Simples | Doubles | Triples | Tetris |
|---|---|---|---|---|
| Coef. de corrélation (`r`) | `-0.3229654` | `-0.3659157` | `-0.1582996` | `0.44389` |
| Coef. de détermination ($R^2 = (r)^2$) | `0.10430665` | `0.1338943` | `0.025058763` | `0.197038332` |

Visuellement et selon les coefficients de corrélation, nous constatons que **les tetris**, bien qu'il s'agisse de la combinaison la plus dangereuse à effectuer, donne un **gain de score suffisant pour le rendre plus avantageux que les trois autres combinaisons**. Les simples et doubles ont tendance à faire diminuer le score final ; il vaut mieux les éviter durant une partie. Les triples ont assez peu d'influence sur le score final. Nos **résultats sont très peu significatifs** étant donné la répartition de nos données. Il y a donc possiblement **plusieurs stratégies qui peuvent mener à un bon score**.

### 4) Niveau final par rapport au temps de jeu

Comme pour les deux études précédentes, nous avons effectué une régression, cette fois linéaire, entre les deux variables aléatoires considérées : le **temps de jeu et le niveau final atteint**. Le niveau final est une variable entière discrète. Les données et les résultats de la régression (régression prédite et régressions extrêmes avec un seuil de confiance de 95%) sont présentés dans la figure ci-après.




Visuellement, la répartition de nos données correspond assez bien à ce qui est requis par la théorie, à savoir que **la variance reste constante**. Nous pouvons donc avoir assez confiance dans les résultats de la régression.

La **régression est très significative** étant donné que les données viennent de parties jouées par des être humains ; nous obtenons un $R^2 = 0.8046185$.

## Conclusion

Nous avons déterminé à partir de données extraites par un script JavaScript sur Internet et à partir d'études statistiques, des **stratégies permettant de maximiser le score** obtenu par un joueur.

Nous en sommes arrivés aux conclusions suivantes:

1) La fréquence des tétriminos a, jusqu'à un point, tendance à augmenter le score du joueur

2) La **fréquence optimale des tétriminos à privilégier se situe entre 44.1 et 52.7 tétriminos par minute**, mais nos données ne sont pas suffisantes pour que ces données soient statistiquement significatives.

3) Les **tétris sont définitivement la combinaison à privilégier**, bien que nos données montrent qu'il y a **potentiellement plusieurs stratégies valides**.

4) Le niveau final atteint par les joueurs suit une relation linéaire avec le temps de jeu.

# Références

[1] "List of best-selling Game Boy video games", Wikipedia, 2017. [Online]. Available: https://en.wikipedia.org/wiki/List_of_best-selling_Game_Boy_video_games#Video_games

[2] "Free Tetris | Tetris Friends Online Games", TetrisFriends, 2017. [Online]. Available: http://www.tetrisfriends.com/

[3] "Tetris 1989 - Free online Tetris game at Tetris Friends", TetrisFriends, 2017. [Online]. Available: http://www.tetrisfriends.com/games/Mono/game.php