



Apprenant : Attaf . Lapasset . Lepicaut  
Objet : projet entreprise  
Période : Centre de formation  
Jury Académique : Serge Gratton - Directeur MS Valdom  
Intervenant : Serge Gratton - Professeur  
Date : 25/03/19



**Mots clefs :** vitesse moyenne . route. mesures TomTom© . anova . R2

#### CADRAGE

date	20 février 2019
version	<b>28 mars 2019</b>
source des données	Tom Tom
nombre de classes	Ø
nombre d'enregistrements	92
périmètre	Apprentissage statistique > machine learning + deep learning
problème	[GLM > Régression Multiple > Anova ] + AFCM
type d'analyse	Méthode d'apprentissage machine supervisée
langage de programmation	Langage Python + Scikit-learn
<b>Note destinée au lecteur</b>	résultats clefs



- DISCLAIMER -

Pour la réalisation de ce livrable, L'équipe Ms-Valdom a mobilisé les compétences :

- Marie-Jo HUGUET, enseignant-chercheur en Informatique à l'INSA de Toulouse au Département Génie Electrique et Informatique, et chercheur au LAAS-CNRS,
- L'équipe Génie Modélisation Mathématique Insa de Toulouse.



## **- SOMMAIRE -**

(1) Modèle GLM : Etude du problème	p. 3
(2) Modèle GLM : Résultats	p. 4
(3) Modèle d'Analyse Factorielle	p. 5
(4) Annexes	p. 6

Problème: vérifier si 3  $X_i$  exogènes (Temp /J, Precipit /J, Day-off) sont des descripteurs utiles (?), importants (?) pour prédire  $Y$  = vitesse moyenne / J /  $\Sigma$  les routes)

Données:

Nombre  $X_i = 3$

(training = 0.77 = 61) + (test = 0.33 = 31)

- . Temp /J : Indicatrice, 0 =  $d^o \leq 25$ , 1 =  $d^o > 25$
- . Precipit /J : Indicatrice, 0 = pas pluie, 1 = pluie
- . Day-off : Indicatrice, 0 = pas jour férié et:ou pas WE, 1 = jour férié et:ou WE

Audit du problème:

- . Utilisation du Modèle Linéaire Généralisé (régression multiple) afin de produire un  $R^2$  et vérifier si les  $X_i$  sont des descripteurs utiles, et/ou des descripteurs importants pour expliquer les variations de  $Y$  (analyse de la variance ~ Reg. ANOVA).

Ecriture du Modèle ANOVA:  $Y(\text{Quanti}) = f(X_i[\text{Quali}])$

- .  $Y = b_1.\text{Temp} + b_2.\text{Precipit} + b_3.\text{Day-off} + \text{erreur}$

Méthode de résolution:

- . Test statistique = réaliser un test de  $\chi^2$  afin de tester les hypothèses

**H0 : variation de Y est indépendante de la variation des  $X_i$**

**H1 : variation de Y est dépendante de la variation des  $X_i$**

Si H0 acceptée et H1 rejetée, pas de relation entre  $Y$  et  $X_i \rightarrow$  pas utile de faire une régression ANOVA

Warning:

- . Présence d'une mesure supplémentaire, 1<sup>er</sup> jour du mois octobre dans le dossier du mois septembre ?

On appelle **pourcentage de variance expliquée** ou bien encore **coefficient de détermination multiple**, la quantité définie par :

$$R^2 = \frac{\text{SCR}(M_0) - \text{SCR}(M_p)}{\text{SCR}(M_0)} = 1 - \frac{\text{SCR}(M_p)}{\text{SCR}(M_0)}$$

**Remarques :**

- Intuitivement, ce coefficient quantifie la capacité du modèle à expliquer les variations de  $Y$ .

1. Test du Khi-2:

- . Pour faire le test, construction de la table de contingence entre Y et matrice (Xi)  
=> À partir des données fournies, pas possible.

2. Régression ANOVA:

- . Même en absence de résultat de test de khi-2, réalisation de la régression ANOVA pour avoir une « base-line » empirique.

- . Résultat : R2 empirique >

```
Training set score: 0.13
Test set score: 0.03
```

- . Analyse :

- En train / test, R2 empirique non significatif,

Warning:

- . Dans le cas où R2 empirique est significatif, confirmer le résultat avec les tests statistiques Two-way ANOVA, CHI-square test -> Goodness of fit [Ø] ...

- les 3 Xi exogènes (Temp /J, Precipit /J, Day-off) ne sont pas des descripteurs utiles pour expliquer les variations  $Y = \text{vitesse moyenne} / J / \Sigma \text{les routes}$ ,  
 → Les spécifications de notre modèle  $Y = b_1.\text{Temp} + b_2.\text{Precipit} + b_3.\text{Day-off} + \text{erreur}$ , **non valide (?)**  
 mais pas d'expérience de l'équipe MS-Valdom pour ce type d'étude,

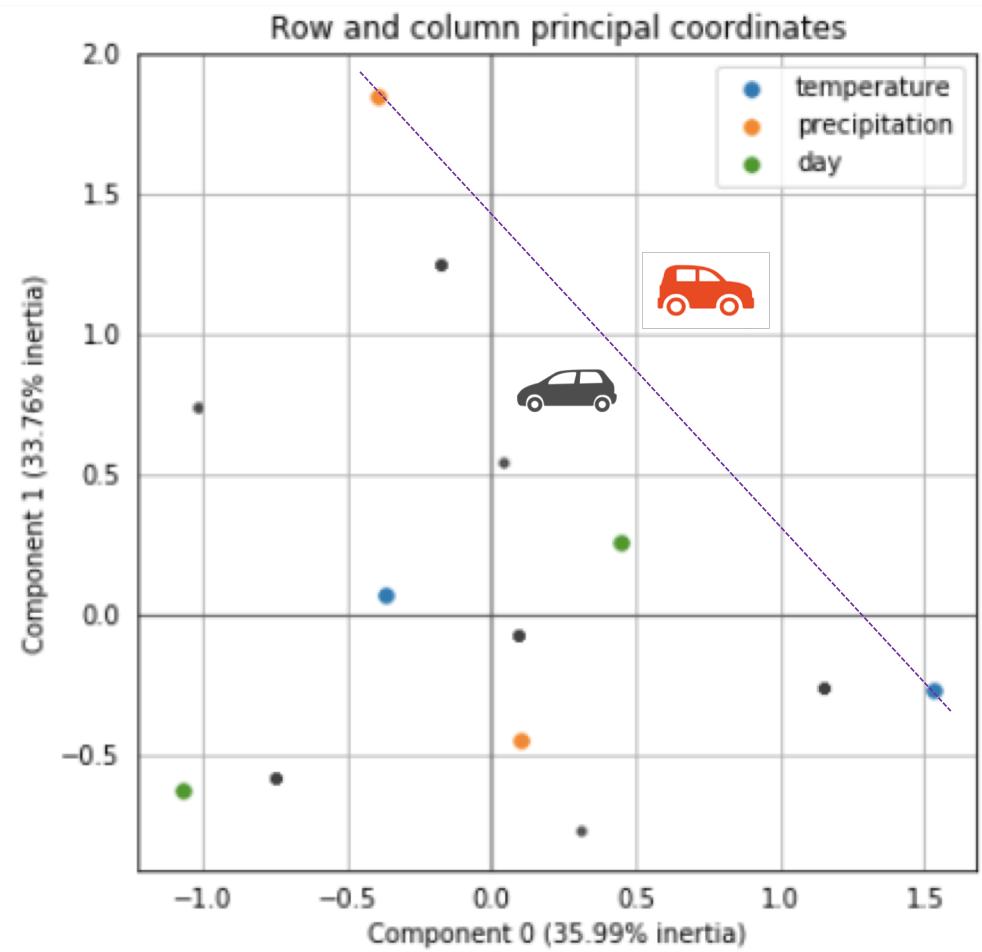
**→ Remarque sur les modèles d'analyse de la variance :**

- Temp /J, Precipit /J, Day-off → nécessité d'avoir des mesures / heure ou / minute pour ces variables explicatives,
- $Y = \text{vitesse moyenne} / J / \Sigma \text{les routes}$  → nécessité d'avoir des mesures / heure ou / minute / seconde pour la variable à expliquer,
- Travailler avec une mesure de la vitesse instantanée / seconde,

**→ SOLUTION :** pour ce type traffic routier (cf. Brendan Guillouet / Marie-Jo HUGUET – INSA)

- Utiliser une borne de comptage routier de la DDE pour mesurer le taux d'occupation d'une route



Analyse :

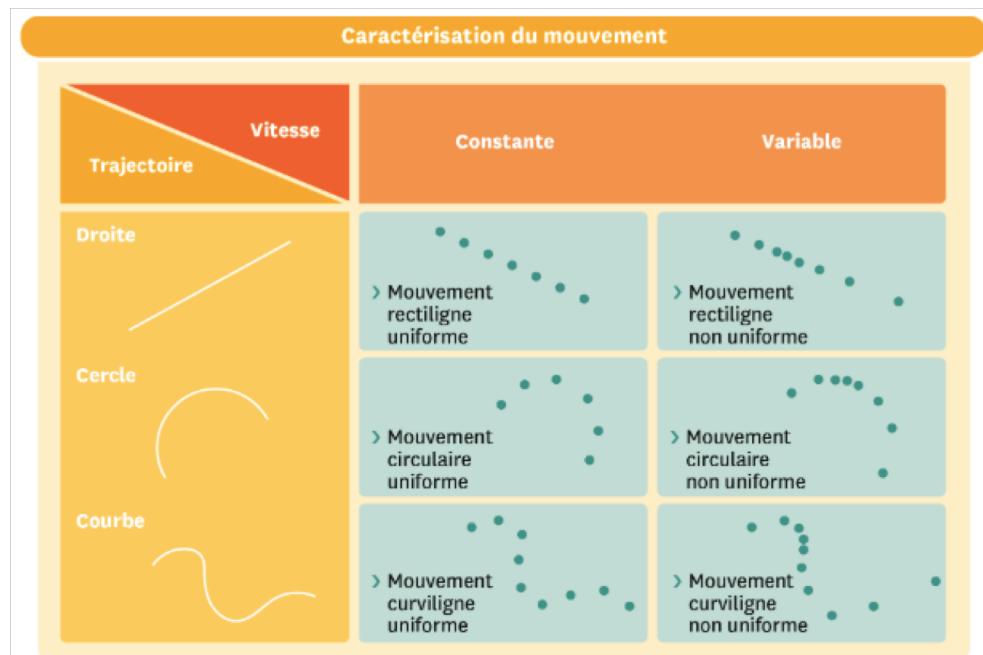
- . Inertie CP1 = 35.99%
- . Inertie CP2 = 33.76%
- . Inertie Plan Factoriel = **69.75%**

Sur l'axe 1 : il ne se passe rien entre les  $X_i \Rightarrow$  pas de relation entre  $X_i$

Sur l'axe 2 : il ne se passe rien entre les  $X_i \Rightarrow$  pas de relation entre  $X_i$

➔ SOLUTION : croiser les apports de l'ontologie avec l'expertise d'un ingénieur en modélisation du trafic routier afin de retravailler la phase de conception – modélisation du problème physique à observer.

Un mobile en mouvement... mesure de vitesse ?



## II] La vitesse ?

### 1°) Définition.

La vitesse instantanée est la vitesse à un instant  $t$  donné. C'est la vitesse moyenne sur un intervalle de temps très petit autour de l'instant considéré.

On représente la vitesse instantanée à l'instant  $t$  par un vecteur. Ce vecteur est tangent à la trajectoire de  $M$  à l'instant considéré, il est dirigé dans le sens du mouvement. Sa valeur est définie par :

$$V_i = d(M_{i-1}, M_{i+1}) / (t_{i+1} - t_{i-1})$$

La distance  $d_i$  entre le point  $M_{i-1}$  et le point  $M_{i+1}$  se calcule par la relation :

$$d_i = \sqrt{(x_{i+1}-x_{i-1})^2 + (y_{i+1}-y_{i-1})^2}.$$

La vitesse se calcule ensuite par :

$$V_i = d_i / (t_{i+1} - t_{i-1}) = \sqrt{(x_{i+1}-x_{i-1})^2 + (y_{i+1}-y_{i-1})^2} / (t_{i+1} - t_{i-1})$$

## 4.2 Annexes.

### Mesures en physique

Bonjour Monsieur,

Ce jour, après avoir travaillé sur l'étape de preprocessing des data traffic, nous nous sommes rendus compte de problèmes de cohérence des données qui fragilisent la production de nos futures analyses :

(1) lorsque l'on prend :

route1 et jour1 = on a 37568 points mesures de vitesses,  
route1 et jour2 = 37572 points mesures,  
route1 et jour3 = 37570 points mesures,

Problème 1 = on compare des points mesures différents, pour la même route et jour différents ?

(2) en "crunchant" la donnée, nous nous sommes aperçus que la qualité de mesure de la vitesse est hétérogène d'un jour à l'autre, c'est à dire :  
jour1 et 2 = default  
jour3 à 7 = standard

Ce qui veut dire, lors de l'analyse de la route1 sur 31 jours du mois de juillet, on compare des jours du mois dont les conditions de mesures sont différentes. Donc le plan d'expérience des mesures n'est pas garanti, Que en pensez vous, quelles sont vos recommandations sur ces 2 points.

Très cordialement, équipe ms-valdom.

Bonjour,

Pour le premier point je ne suis pas sûr de comprendre votre question.

Faites-vous référence au jour 1 d'un mois particulier (ce qui me paraît improbable étant donné qu'il y a environ 20 000 mesures par jour tous segments confondus).

Pour chaque segment/route vous avez environ 8910 mesures pour le mois de juillet, 8870 pour le mois de août et 8370 pour le mois de septembre ; hormis pour 5 segments pour lesquels vous n'avez que des mesures pour le mois d'août et de septembre (vous pouvez ignorer ces 5 segments si vous le souhaitez pour plus de facilité).

De manière générale dans le cadre de ce TP nous vous avons fourni des données « brutes » c'est-à-dire des données issues directement de l'API de TOM-TOM, il est donc normal que celles-ci ne soient pas « lisses » (i.e. que le nombre de mesures de jour en jour soit variable).

Je vous conseille dans un premier temps d'analyser les fluctuations du trafic indépendamment des segments et de ne pas vous baser sur les noms des répertoires pour la date mais sur les timestamps des mesures afin d'avoir une granularité plus fine (analyse à l'heure ou à la minute plutôt qu'à la journée) ; grâce à cette connaissance vous pourrez ensuite effectuer des analyses de corrélations entre les variations du trafic et d'autres facteurs (la météo notamment). Puis si le temps vous le permet vous pourrez dans un second temps analyser les variations de trafic dans le temps entre segments.

Pour le deuxième point si vous faites référence au champ « measurement » ne le prenez pas en compte.

Comme vous pouvez le voir sur la nouvelle version de la documentation [  
<https://developer.tomtom.com/traffic-api/traffic-api-documentation-traffic-flow/flow-segment-data>] ce champ n'apparaît dorénavant plus.

Très cordialement,

Paul Pidou

## Apprentissage statistique : application au trafic routier à partir de données structurées et aux données massives par Brendan Guillouet



**Résumé :** Cette thèse s'intéresse à l'apprentissage pour données massives. On considère en premier lieu, des trajectoires définies par des séquences de géolocalisations. Une nouvelle mesure de distance entre trajectoires (Symmetrized Segment-Path Distance) permet d'identifier par classification hiérarchique des groupes de trajectoires, modélisés ensuite par des mélanges gaussiens décrivant les déplacements par zones. Cette modélisation est utilisée de façon générique pour résoudre plusieurs types de problèmes liés aux trafic routier : prévision de la destination finale d'une trajectoire, temps d'arrivée à destination, prochaine zone de localisation. Les exemples analysés montrent que le modèle proposé s'applique à des environnements routiers différents et, qu'une fois appris, il s'applique à des trajectoires aux propriétés spatiales et temporelles différentes. En deuxième lieu, les environnements technologiques d'apprentissage pour données massives sont comparés sur des cas d'usage industriels.



**Marie-Jo Huguet**  
Décision et optimisation

## Transport et mobilité



- Protection de la vie privée et mobilité

- Depuis 2014:
  - Objet: [Préservation de la vie privée et problèmes de covoiturage](#)
  - Participants: Marc-Olivier Killijian (LAAS-CNRS, TSF), Sébastien Gambs (IRISA, Rennes puis UQAM Montréal), Kevin Huguenin (LAAS-CNRS, TSF)
  - Stagiaire et doctorant: Ulrich Matchi Aïvodji

- Itinéraires multi-modaux et dépendants du temps

- 2013-2016:

- Objet: [Itinéraires multimodaux alternatifs](#)
- Participants: Sandra Ulrich Ngueveu (LAAS-CNRS, ROC)
- Doctorant: Grégoire Scano (MobiGIS)
- Stage: Céline Gimbertat (M1, 2015)

- 2008-2010:

- Objet: [Algorithmes dépendant du temps bi-objectif multimodaux](#)
- Participants: Christian Artigues (LAAS-CNRS, ROC)
- Doctorant: Fallou Gueye (MobiGIS)
- Stage: Arnaud Fradin (L3)