MODÉLISATION DES RETARDS DE TRAINS POUR LA ROBUSTESSE DES OPÉRATIONS EN GARE

+ SÉMINAIRE MSDMA 19 JANVIER 2018

Marie Milliet de Faverges
Marie.milliet-de-faverges@reseau.sncf.fr

SOMMAIRE

01.
GÉNÉRALITÉS SUR LA PONCTUALITÉ

02.
LA PLANIFICATION EN GARE

03.
PRÉSENTATION DES DONNÉES

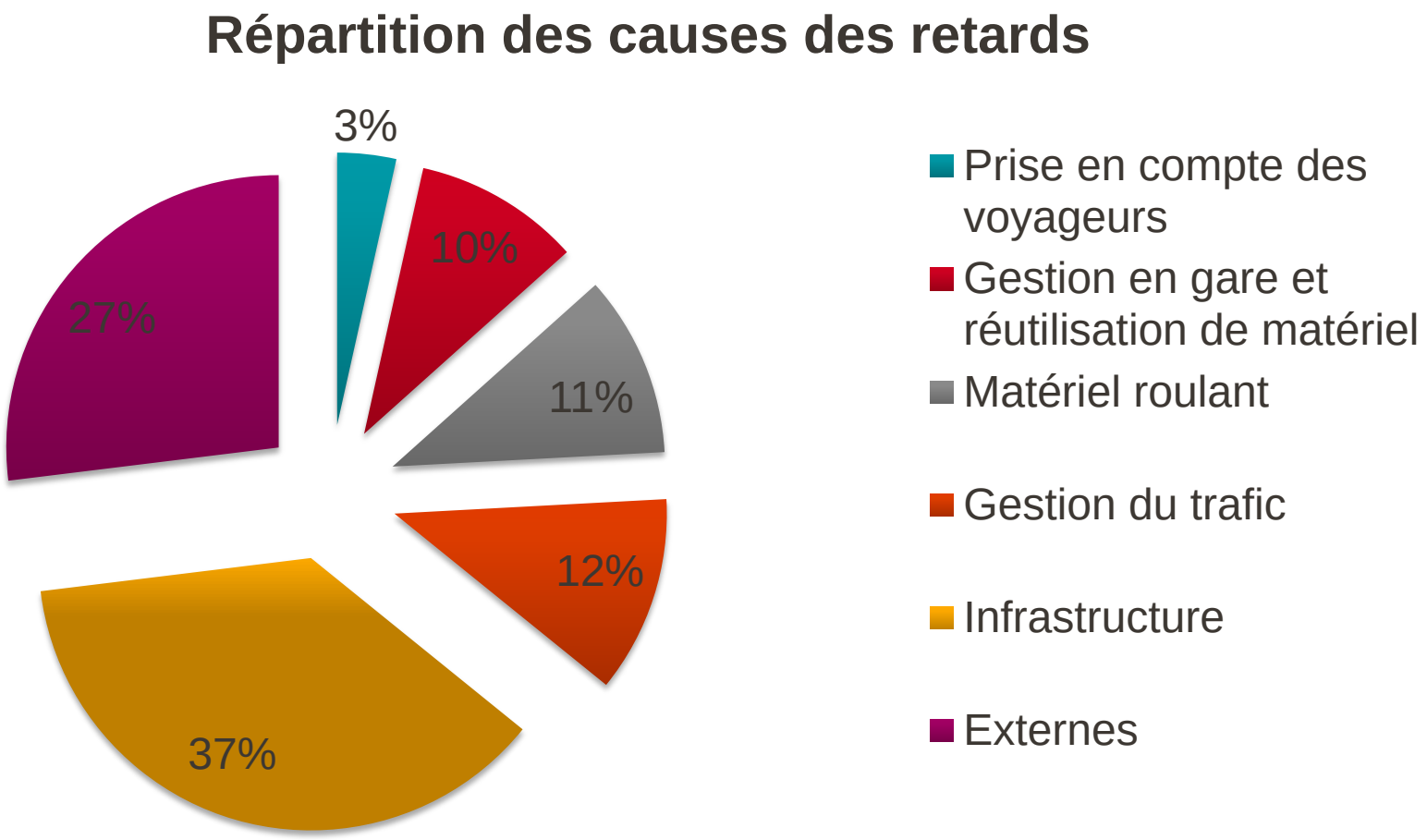
04.
MODÉLISATION DES RETARDS

01.
PLANIFICATION FERROVIAIRE: CONTEXTES ET ENJEUX

PONCTUALITÉ DES TRAINS

CAUSES DES RETARDS

DONNÉES 2016 POUR LA LGV ATLANTIQUE



Source : Rapport Annuel 2016, Autorité de la qualité de service dans les transports (AQST), Juillet 2017

CAUSES DES RETARDS

Famille	Exemples
Gestion du trafic	Gestion des interactions entre les différentes circulations sur le réseau, planification, replanification
Gestion en gare et réutilisation du matériel	Personnel de bord, roulement de matériel
Matériel roulant	Pannes
Prise en compte des voyageurs	Gestion des affluences, des correspondances, prise en charge des passagers à mobilité réduite
Infrastructure	Maintenance et incidents (voies, caténaires, signalisation,…)
Causes externes	Conditions météorologiques, obstacles sur les voies, malveillance, mouvements sociaux, colis suspects, …

BILAN

Certaines causes imprévisibles

+ Dysfonctionnement (matériel ou infrastructure)
+ Causes externes (météo, malveillances, obstacles sur les voies,…)

Propagation

+ Gestion du trafic, synchronisation des circulations

Création et accumulation de retards en gare

+ Croisement de ligne = risques de retards liés à la gestion du trafic
+ Gestion des affluences : flux voyageurs, temps de descente et montée dans une rame,…
+ Réaffectation du matériel, horaires du personnel
+ Gestion des correspondances
+ Itinéraires en conflit, voie occupée, etc

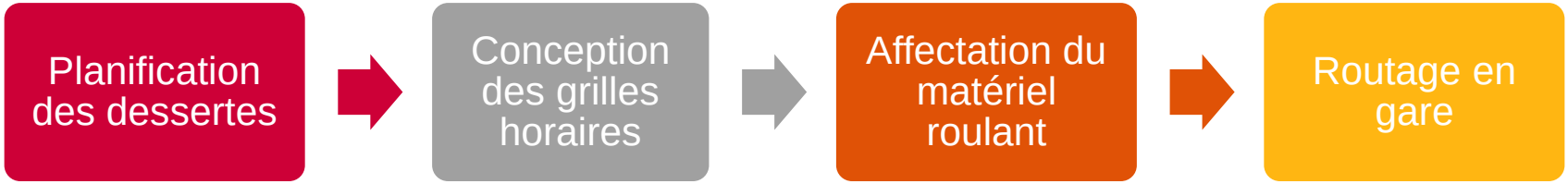
Un train en retard risque fortement d'accumuler plus de retard ou de propager son retard à d'autres trains, en particulier au niveau des gares

01.
PLANIFICATION FERROVIAIRE: CONTEXTES ET ENJEUX

OPEN GOV : MODÈLES DE ROUTAGE DES TRAINS EN GARE

LA PLANIFICATION FERROVIAIRE

Planification séquentielle

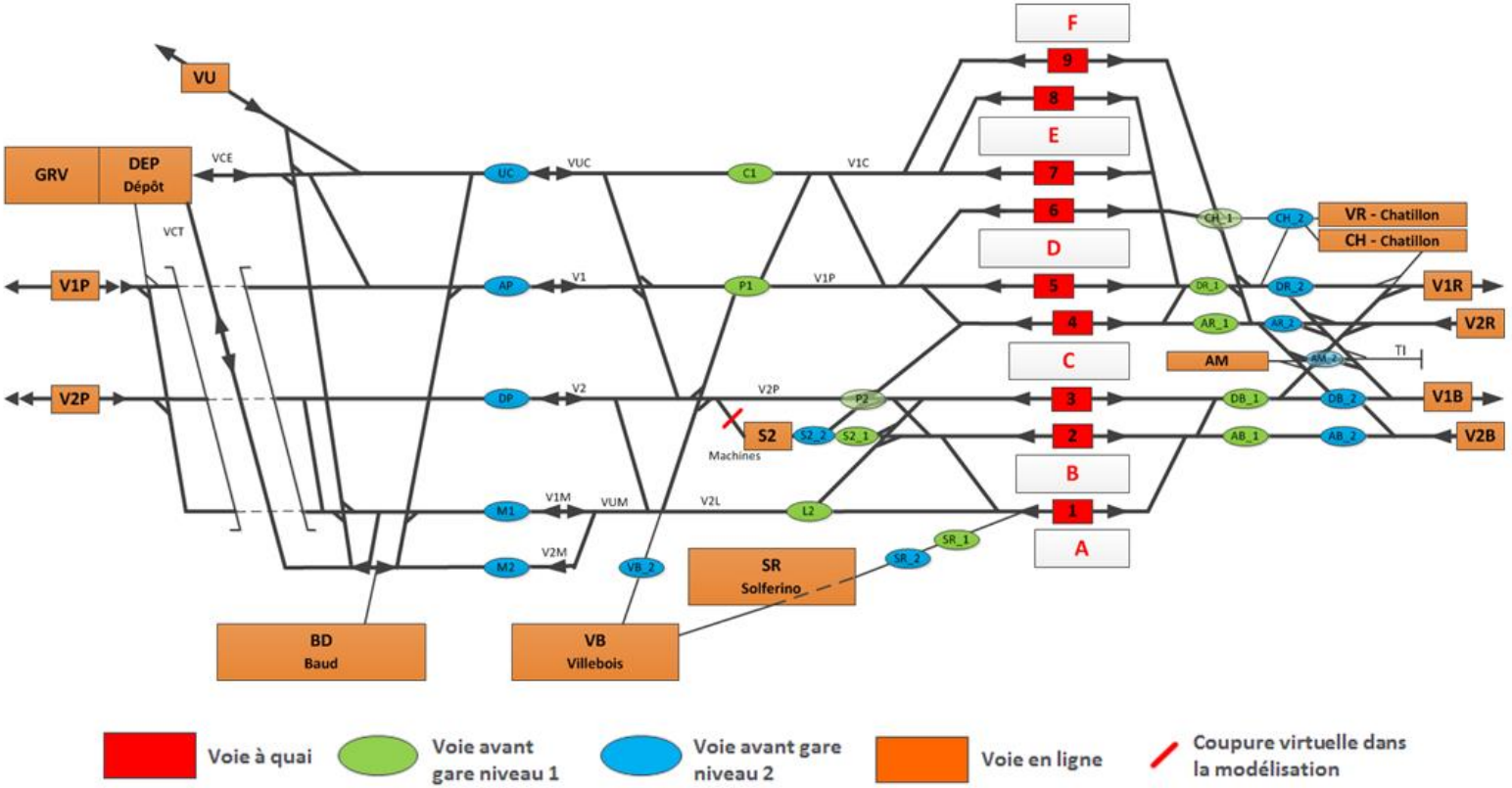


Stratégie de robustesse

+ Marges: temps additionnels sur les temps de trajets pour rattraper du retard
+ Temps de séparation: temps additionnels entre deux utilisation d'un élément d'infrastructure pour éviter la propagation

PLANIFICATION EN GARE

DIFFICULTÉ DU ROUTAGE



PLANIFICATION EN GARE

OBJECTIF D'OPEN GOV

Entrée :

- Liste des trains, leurs caractéristiques, leurs horaires
- Infrastructure de la gare et règles de fonctionnement

Sortie :

- Graphique d'occupation des voies faisables et robuste

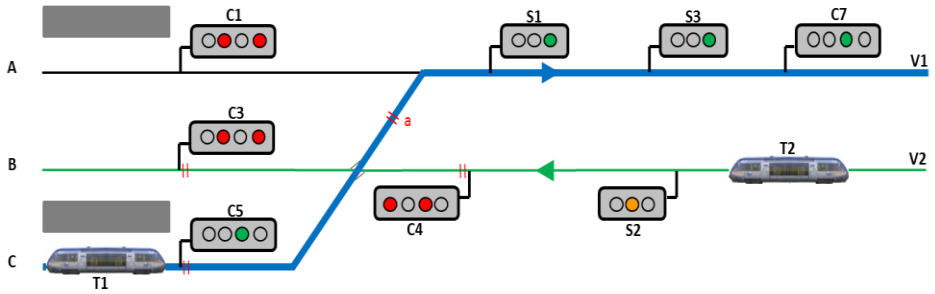
CONTRAINTES

Affectation des voies en gare

- Contraintes de réoccupation
- Voies interdites ou voies dédiées
- Contraintes de flux voyageurs, correspondances
- Longueur des trains

Routage des trains en gare : affectation d'itinéraire

- Contraintes de cisaillements



MODÈLE LINÉAIRE EN NOMBRES ENTIERS

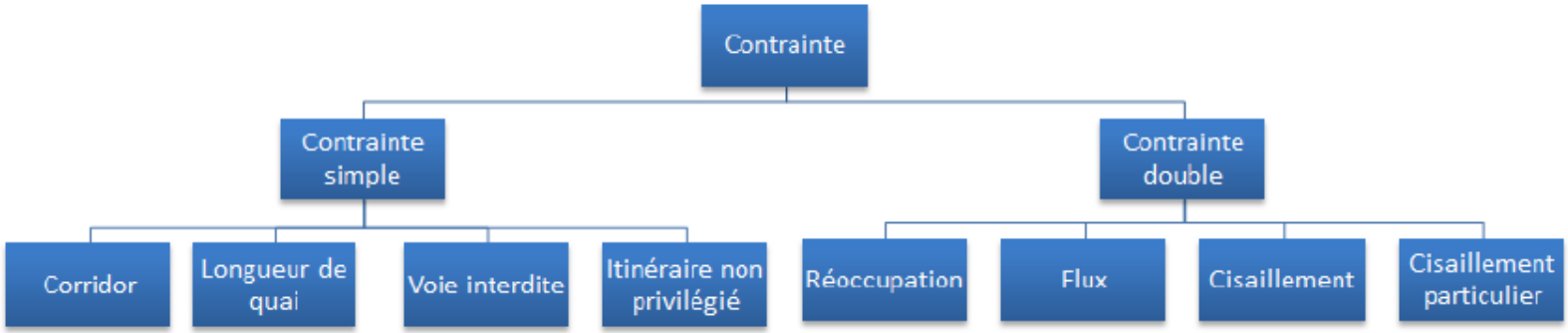
Variables

+ variable booléenne qui associe un chemin à un train (arrivée et départ)
+ *Variable booléenne pour les contraintes non respectées*

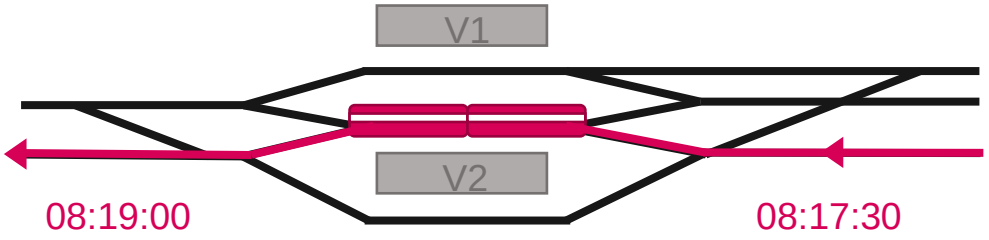
Contraintes

Deux types :

+ les contraintes simples : concernent un seul mouvement
+ les contraintes doubles : concernent un couple de mouvements



MODÈLE LINÉAIRE EN NOMBRES ENTIERS



Données

+ T l'ensemble des trains de la journée
+ I l'ensemble des itinéraires, I_t ensemble des itinéraires éligibles pour $t \in \mathrm{T}$

Variables de décision

+ variable booléenne qui associe un chemin à un train (arrivée et départ)

$$x_{t,i} \in \{0,1\}, \qquad t \in T, i \in I_t$$

Objectif

+ On prend en compte les préférences dans le choix des quais, soit $p_{t,i}$ la pénalité associée à l'affectation de i à t :

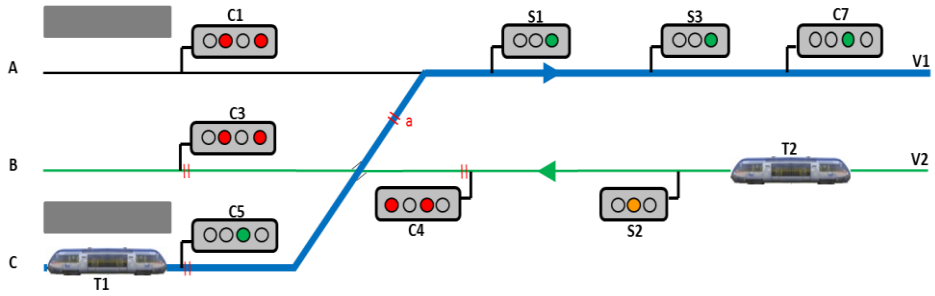
$$\min \sum_{t \in T} \sum_{i \in I_t} p_{i,t} \,.\, x_{t,i}$$

Contraintes simples

+ Chaque train doit utiliser un itinéraire

$$\sum_{i \in I_t} x_{t,i} = 1\,, \qquad \forall\, t \in T$$

MODÈLE LINÉAIRE EN NOMBRES ENTIERS



Données

+ T l'ensemble des trains de la journée
+ I l'ensemble des itinéraires,
  I_t ensemble des itinéraires éligibles pour $t \in \mathrm{T}$
+ U ensemble des doubles affectations conflictuelles :

$$U = \{(t_1, i_1, t_2, i_2)\ tels\ qu'on\ ne\ peut\ pas\ affecter\ simultanément\ i_1 \text{à}\ t_1 et\ i_2 \text{à}\ t_2\}$$

Variables de décision

+ variable booléenne qui associe un chemin à un train (arrivée et départ)

$$x_{t,i} \in \{0,1\}, \qquad t \in T, i \in I_t$$

Objectif

+ On prend en compte les préférences dans le choix des quais, soit $p_{t,i}$ la pénalité associée à l'affectation de i à t :

$$\min \sum_{t \in T} \sum_{i \in I_t} p_{i,t} \,.\, x_{t,i}$$

Contraintes simples

+ Chaque train doit utiliser un itinéraire

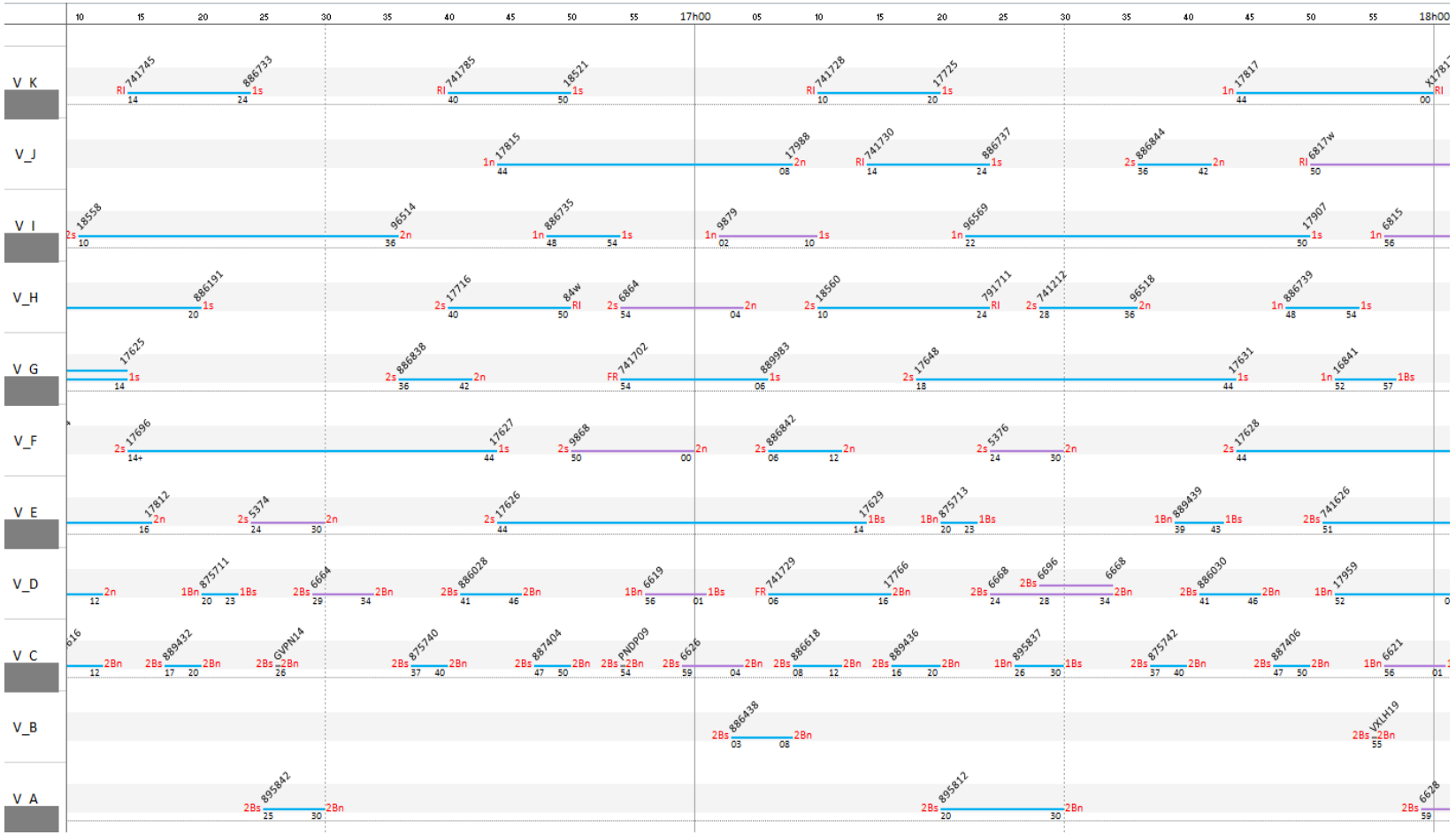
$$\sum_{i \in I_t} x_{t,i} = 1\,, \qquad \forall\, t \in T$$

Contraintes doubles

+ Pas d'affectations incompatibles

$$x_{t_1 i_1} + x_{t_2 i_2} \leq 1, \qquad (t_1, i_1, t_2, i_2) \in U$$

ROBUSTESSE DES GOV



ROBUSTESSE DES GOV

But

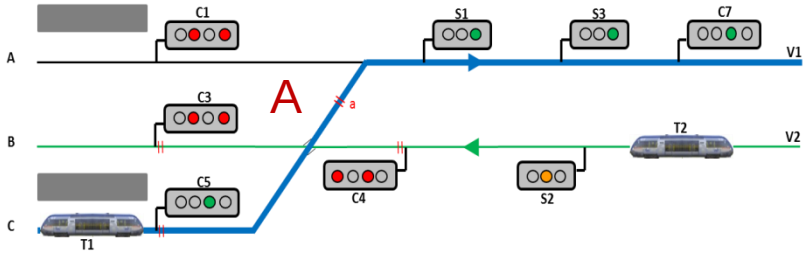
+ Fournir une solution restant réalisable malgré de petites perturbations

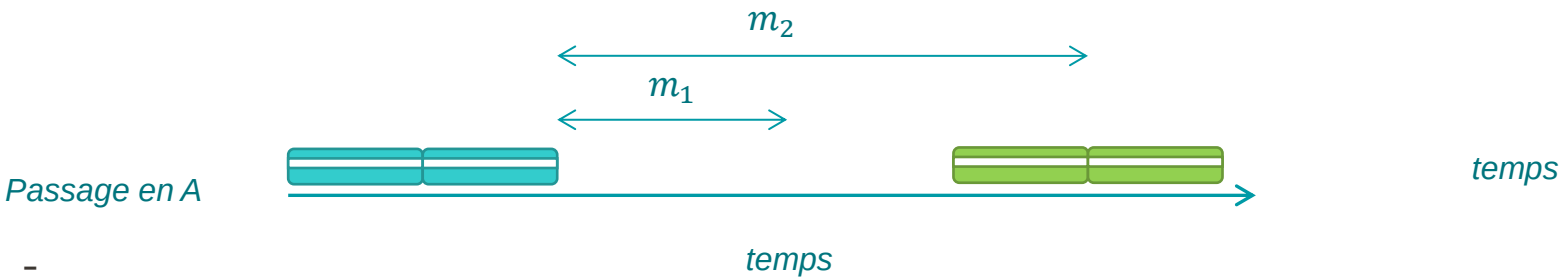
Approches classiques

1. Assurer la disponibilité des itinéraires pendant une période plus longue
2. Prévoir un itinéraire de repli si le premier n'est plus disponible

Open GOV

+ On va détecter les conflits potentiels qu'on va chercher à éviter pour avoir une planification qui reste réalisable pour des retards légers





MODÈLE ROBUSTE

Données

+ T l'ensemble des trains de la journée
+ I l'ensemble des itinéraires, I_t ensemble des itinéraires éligibles pour $t \in$ T
+ $U = \{(t_1, i_1, t_2, i_2)\ tels\ qu'on\ ne\ peut\ pas\ affecter\ simultanément\ i_1 \text{à}\ t_1 et\ i_2 \text{à}\ t_2\}$
+ $V = \{(t_1, i_1, t_2, i_2)\ tels\ que\ l'affectation\ i_1 \text{à}\ t_1 et\ i_2 \text{à}\ t_2\ est\ risquée\}$

Variables de décision

+ variable booléenne qui associe un chemin à un train : $x_{t,i} \in \{0,1\}, \quad t \in T, i \in I_t$
+ Variable booléenne pour affectations non robustes $\gamma_v \in \{0,1\}, v \in V$

Objectif

+ Préférence dans le choix de quai et on pénalise les affectations non robustes

$$\min \sum_{t \in T} \sum_{i \in I_t} p_{i,t} \,.\, x_{t,i} + \sum_{v \in V} p_v . \gamma_v$$

Contraintes cas idéal

$$\sum\nolimits_{i \in I_t} x_{t,i} = 1\,, \qquad \forall\, t \in T$$

$$x_{t_1 i_1} + x_{t_2 i_2} \leq 1, \qquad (t_1, i_1, t_2, i_2) \in U$$

Contraintes robustes

+ On détecte les affectations non robustes

$$x_{t_1 i_1} + x_{t_2 i_2} \leq 1 + \gamma_v, \qquad v = (t_1, i_1, t_2, i_2) \in V$$

BILAN ET PERSPECTIVES

Modèle

$$\min \sum_{t\in T} \sum_{i\in I_t} p_{i,t} \cdot x_{t,i} + \sum_{v\in V} p_v \cdot \gamma_v$$

$$\begin{cases} \sum_{i\in I_t} x_{t,i} = 1\ ,\ \forall\, t \in T \\ x_{t_1 i_1} + x_{t_2 i_2} \leq 1,\ (t_1, i_1, t_2, i_2) \in U \\ x_{t_1 i_1} + x_{t_2 i_2} \leq 1 + \gamma_v,\ v = (t_1, i_1, t_2, i_2) \in V \\ x_{t,i} \in \{0,1\},\ \ t \in T, i \in I_t \\ \gamma_v \in \{0,1\}, v \in V \end{cases}$$

Limites

+ Paramétrage complexe
+ Pas de prise en compte de la réalité

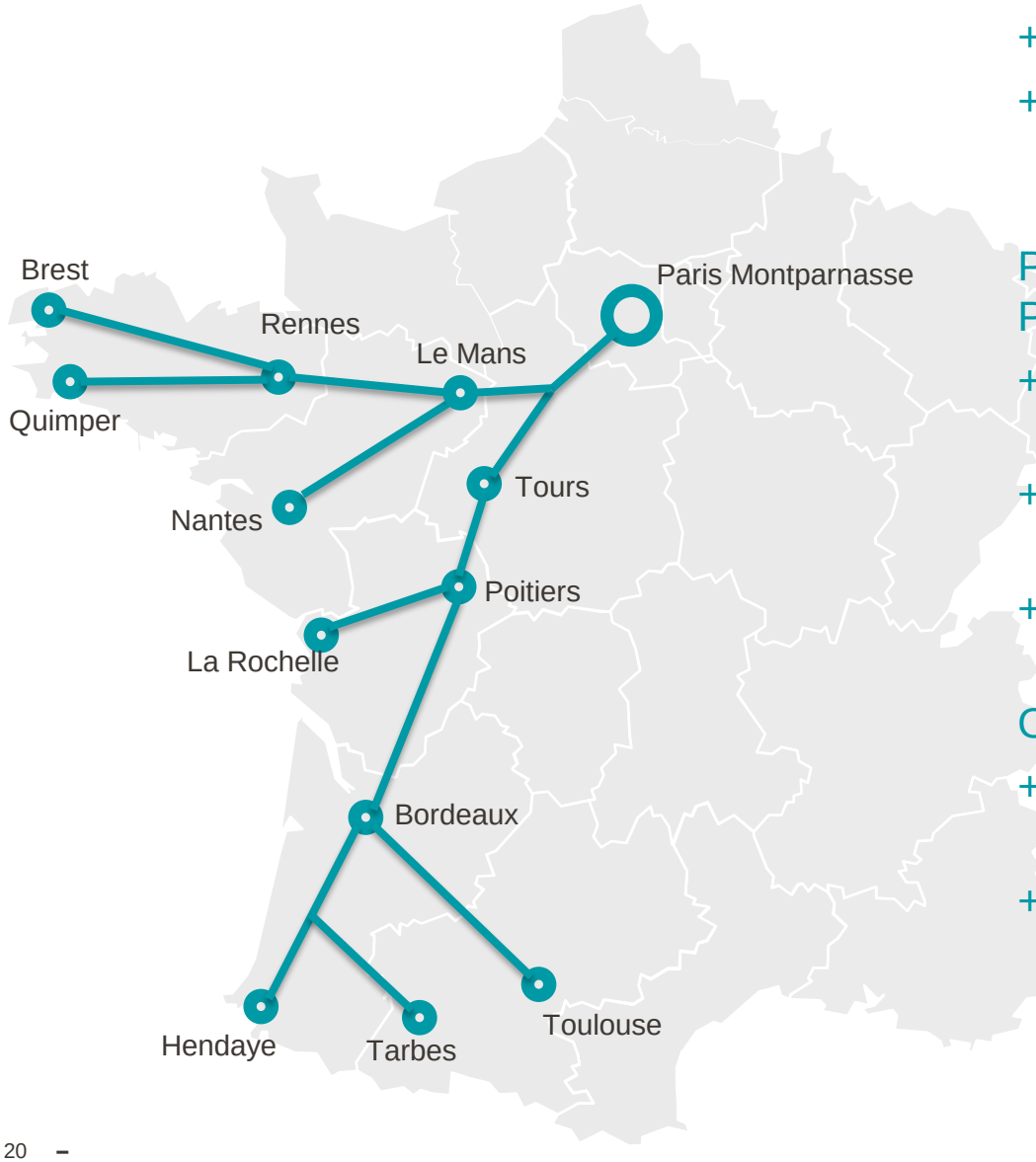
Idée

+ Utiliser des méthodes statistiques pour prendre en compte l'aléa
+ Algorithmes de replanification en temps réel
+ Quelques études pour la conception de grilles horaires (Vansteenwegen(2004), Sels(2013))

01.
DONNÉES

ETUDE DES RETARDS

CAS D'ÉTUDE



Gare Montparnasse

+ Gare terminus
+ infrastructure complexe
+ trafic hétérogène (40% LGV, 47% Transilien, 12% TER)

Prédiction des retards : TGV arrivant à Paris

+ Environ 40 000 trains sur la période du 1 juillet 2016 au 20 décembre 2017
+ 4 axes principaux : Bretagne, Pays de la Loire, Poitou-Charentes et Aquitaine
+ Plus de 60 gares TGV connectées

Objectif

+ Prédictions des retards pour une journée à venir (jusqu'à J-1)
+ Intégration des prédictions dans les modèles d'aide à la décision

LITTÉRATURE

Distribution des retards

+ Loi exponentielle, Weibull, Gamma, log-normale (Yuan, 2006)

Short-term delay prediction

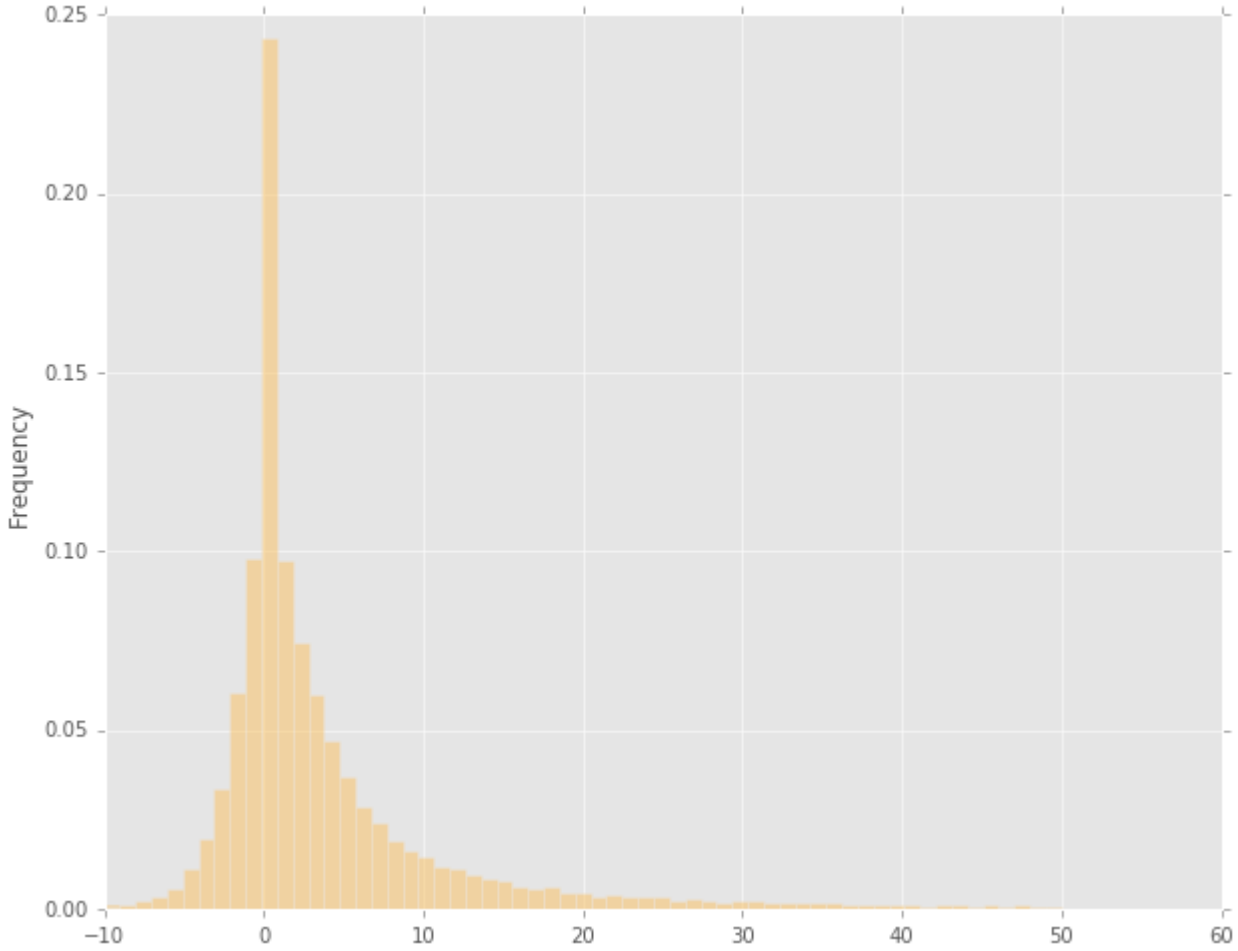
+ Réseaux de neurones (Peters(2005), Oneto et al. (2017))
+ Random forest (Kecman(2015))
+ Algorithme de propagation (Hansen(2010))

Long-term delay prediction

+ Peu de littérature ferroviaire (SVR et aNN de Markovic(2015))
+ Bus: random forest, SVR (J. Mendes Moreira et al, 2012, 2015)

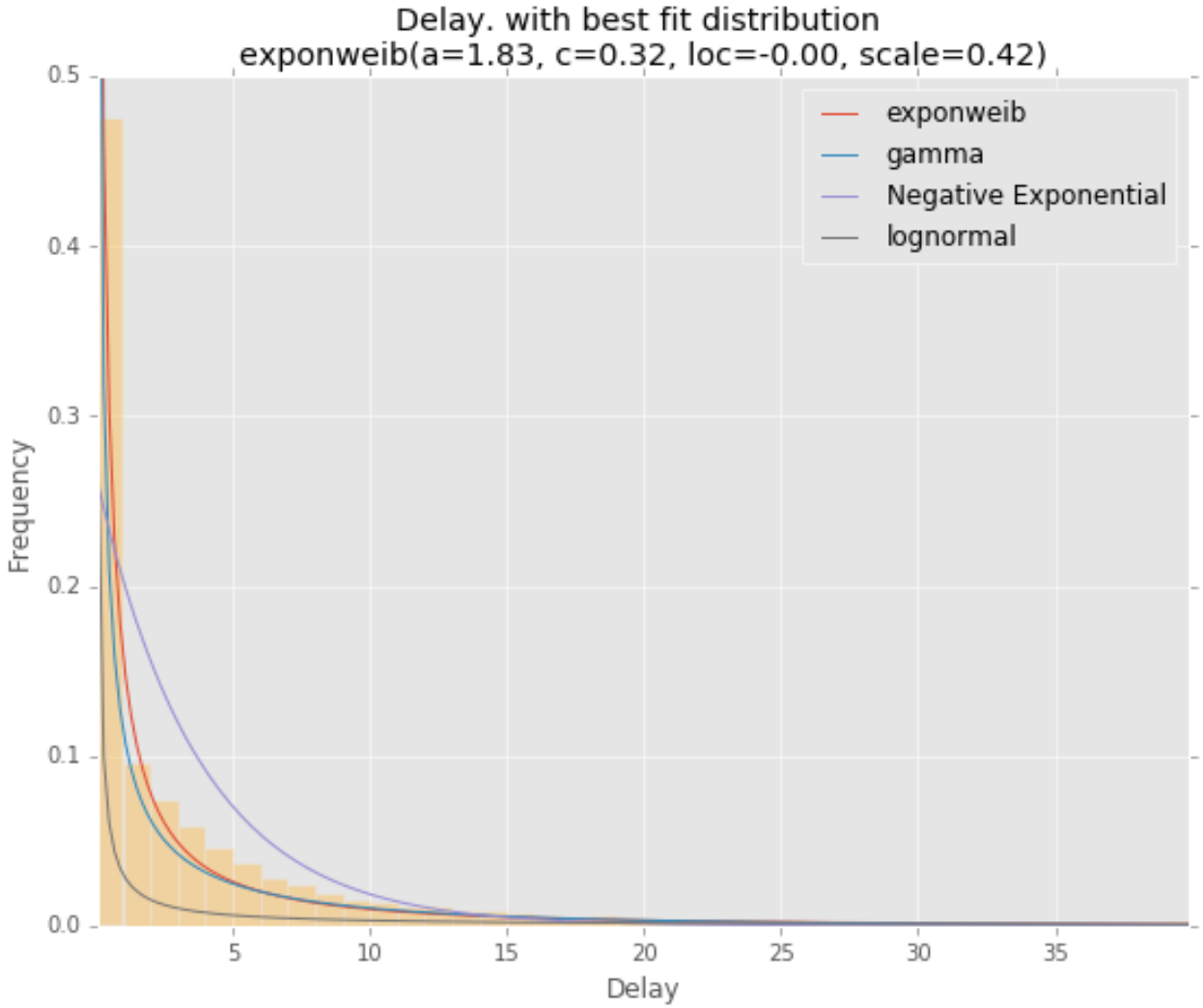
DISTRIBUTION DES RETARDS

RETARDS RÉELS



DISTRIBUTION DES RETARDS

MODÉLISATION SANS VALEURS NÉGATIVES



PRÉSENTATION DES DONNÉES

Données historiques

+ Numéro de train
+ Date et heure de l'observation
+ Lieu
+ Retard observé (positif ou négatif)
+ Sens du mouvement (arrivée/départ)
+ Type de train

Données travaux : marges sur le temps de trajet dues aux travaux

+ Nombre de minutes perdues à cause des travaux
+ Nombre de minutes de marge disponibles

Données météo

+ Données dans les grandes villes du réseau
+ Température, vent, visibilité

LE PROBLÈME

DÉFINITION DU CADRE D'INTERET

- On s'intéresse aux petits retards car :
  - Une planification ne peut a priori pas absorber les grosses perturbations, et dans le cas d'un GOV, pas beaucoup de différences entre un retard de 30 minutes et un retard de 2heures qu'il faut dans tous les cas replanifier.
  - Les grosses perturbations (entraînant les grands retards) sont très imprévisibles
  - Les petits retards sont très probables (propagation des retards et interdiction de départ en avance)
- Les avances ne sont pas acceptées

On filtre les retards entre -10 et 20 minutes et on passe les avances à 0

- A priori, impossible de prévoir précisément une valeur :
  - Les données disponibles donnent plutôt des infos sur les risques de retards légers et de propagation (météo, densité du trafic, etc)
  - Données très bruitées par les opérations ferroviaires
  - Le retard le plus probable est toujours l'heure prévue

On va plutôt chercher la loi de probabilité de retard de chaque train

01.
DONNÉES

MODÉLISATION DES RETARDS

RAPPEL GLM

Trois composantes dans le modèle :

+ La composante aléatoire : $\boldsymbol{y} = (y_1, \dots, y_n)$ suivant une loi de la famille exponentielle
+ Un prédicteur linéaire : $\eta_i = \sum x_i \beta_i$
+ Une fonction de lien g entre la composante aléatoire et le prédicteur linéaire : $\eta_i = g(\mu_i)$

Soit :

$$y \sim ExpFamily(\mu)$$
$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta}$$

Les paramètres $\boldsymbol{\beta}$ sont estimés par maximum de vraisemblance

Motivations :

+ Permet de connaître la loi du retard pour chaque train, ce qui peut être intéressant pour la partie RO
+ On respecte le domaine réduit de la variable de retard

MODÈLE CHOISI

Distribution

+ Loi Weibull : densité $f(y\,|\mu,\sigma) = \left(\frac{\sigma y^{\sigma-1}}{\mu^{\sigma}}\right).\exp\left\{-\left(\frac{y}{\mu}\right)^{\sigma}\right\}$,

  μ et σ paramètres d'échelle et de forme

+ Troncature à droite en 20 : on supprime les retards supérieurs à 20minutes (considérés comme des outliers)

Modèle

$$\boldsymbol{y} \sim Weibull(\boldsymbol{\mu}, \boldsymbol{\sigma})$$
$$log(\boldsymbol{\mu}) = \boldsymbol{X_1\beta_1}$$
$$log(\boldsymbol{\sigma}) = \boldsymbol{X_2\beta_2}$$

Données utilisées

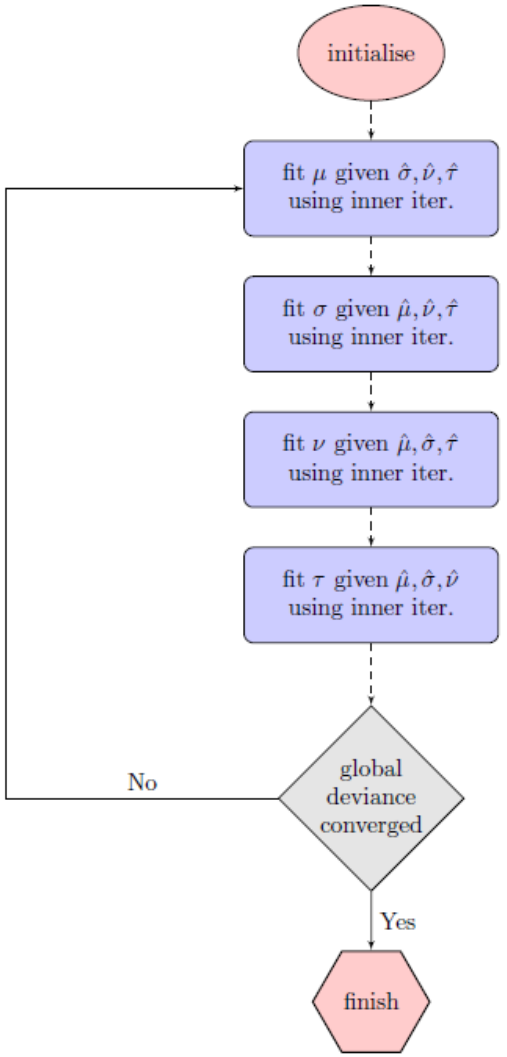
+ Desserte
+ Durée du trajet
+ Type de jour, Heure, vacances
+ Densité du trafic
+ Météo
+ Travaux

FONCTIONNEMENT GAMLSS

+ Optimisation deux paramètres simultanément

+ Modèles avec lois tronquées

+ Mesure de vraisemblance :
$D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}}) = -2[l(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}})]$



DIFFICULTÉ À L'ÉVALUATION

Grandeurs d'intérêt

+ Moyenne : pas vraiment de sens pour des données tronquées, et de manière générale pas représentative pour des distribution fortement asymétriques
+ Mode : dans notre cas, nulle tout le temps
+ Quantiles : plus robuste et utilisable par la suite

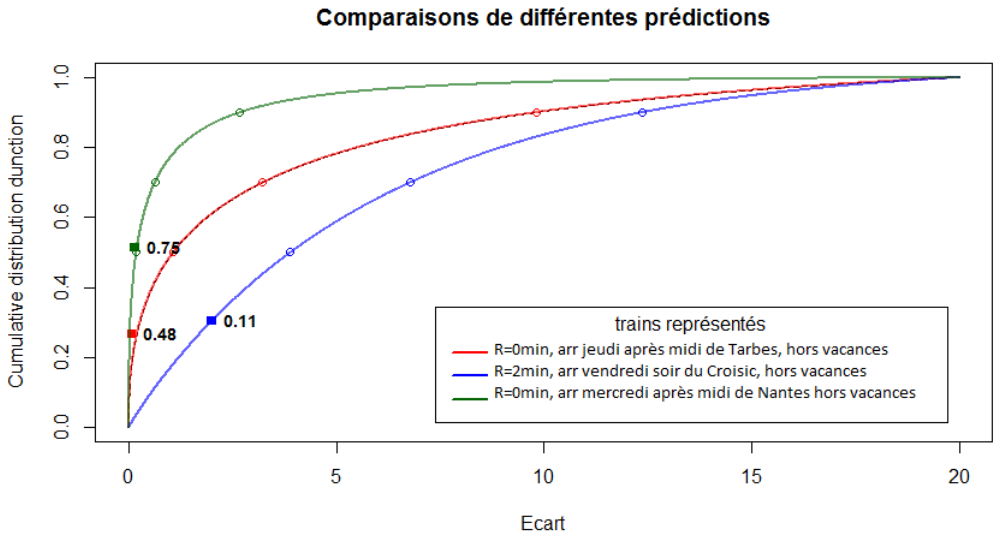
Deviance

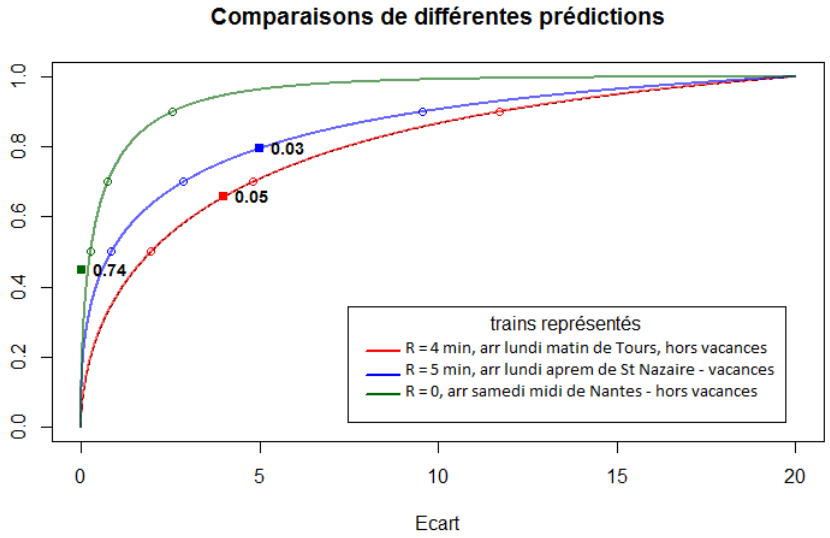
+ $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}) = -2[l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}})]$

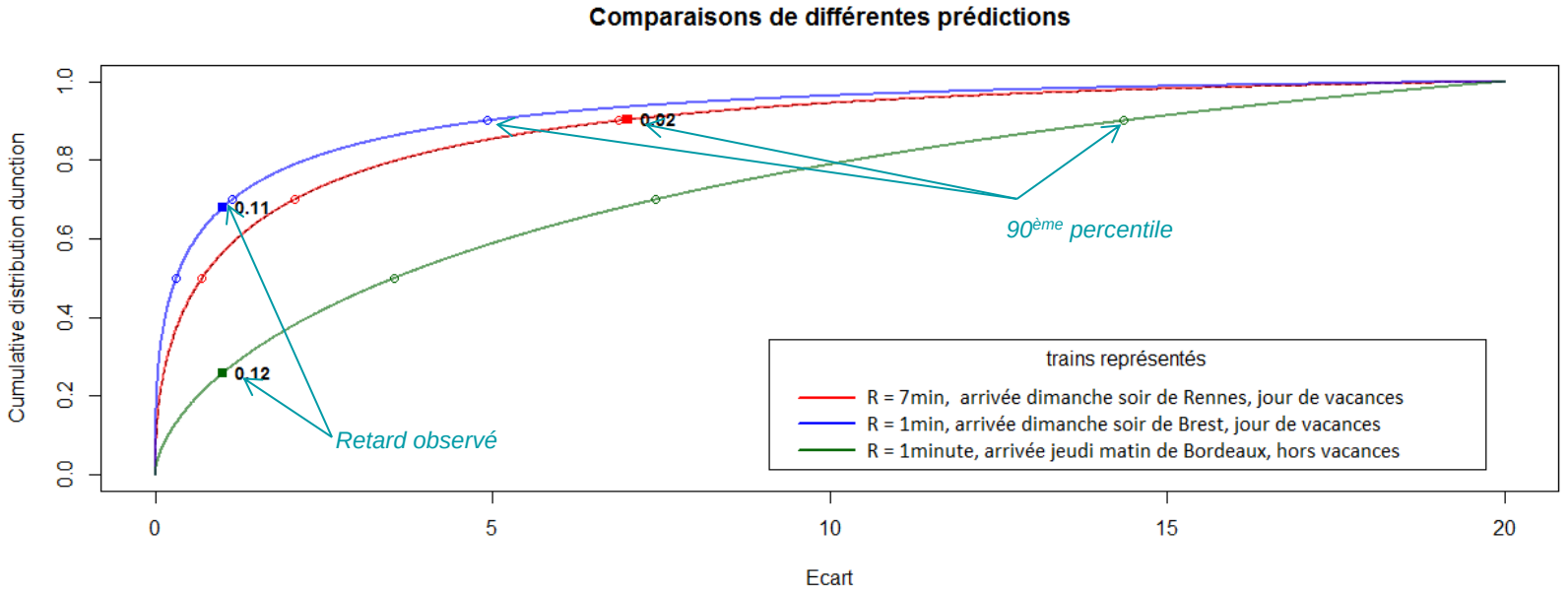
	Modèle nul	Modèle
Set1	73 000	70 310
Set2	27 130	26 085

+ Set 1 : TGV de juillet 2016 à juin 2017 (anciennes lignes) : train de 25000 lignes, test de 8300 lignes
+ Set 2 TGV de juillet 2017 à décembre 2017 (nouvelles lignes) : train de 7800 lignes, test de 2600 lignes

RÉSULTATS


Comparaisons de différentes prédictions
Cumulative distribution dunction
Ecart
0.75
0.48
0.11
trains représentés
R=0min, arr jeudi après midi de Tarbes, hors vacances
R=2min, arr vendredi soir du Croisic, hors vacances
R=0min, arr mercredi après midi de Nantes hors vacances


Comparaisons de différentes prédictions
Ecart
0.74
0.03
0.05
trains représentés
R = 4 min, arr lundi matin de Tours, hors vacances
R = 5 min, arr lundi aprem de St Nazaire - vacances
R = 0, arr samedi midi de Nantes - hors vacances


Comparaisons de différentes prédictions
Cumulative distribution dunction
Ecart
0.02
0.11
0.12
90ème percentile
Retard observé
trains représentés
R = 7min, arrivée dimanche soir de Rennes, jour de vacances
R = 1min, arrivée dimanche soir de Brest, jour de vacances
R = 1minute, arrivée jeudi matin de Bordeaux, hors vacances

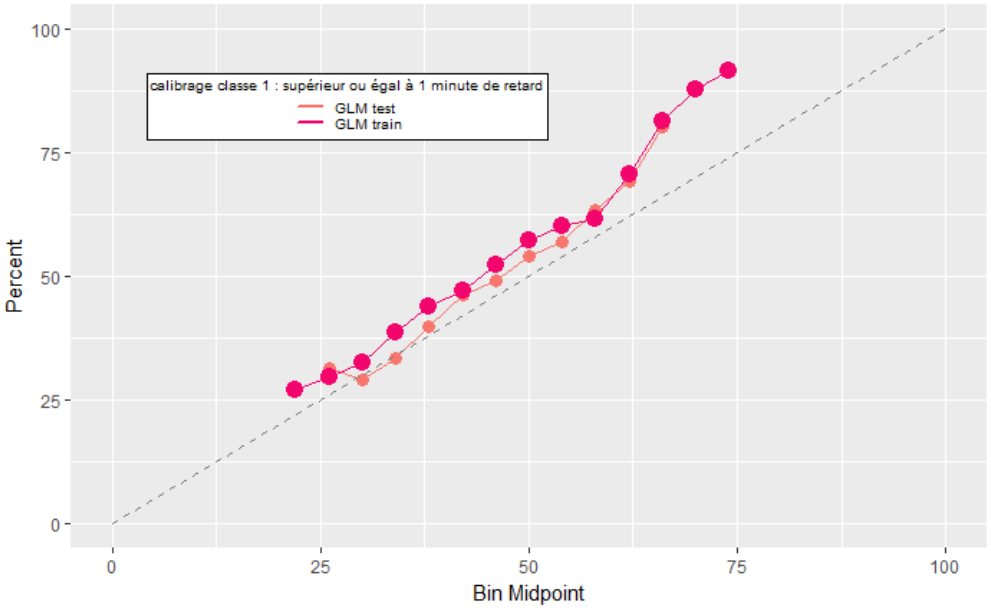
CALIBRATION PAR SEUIL

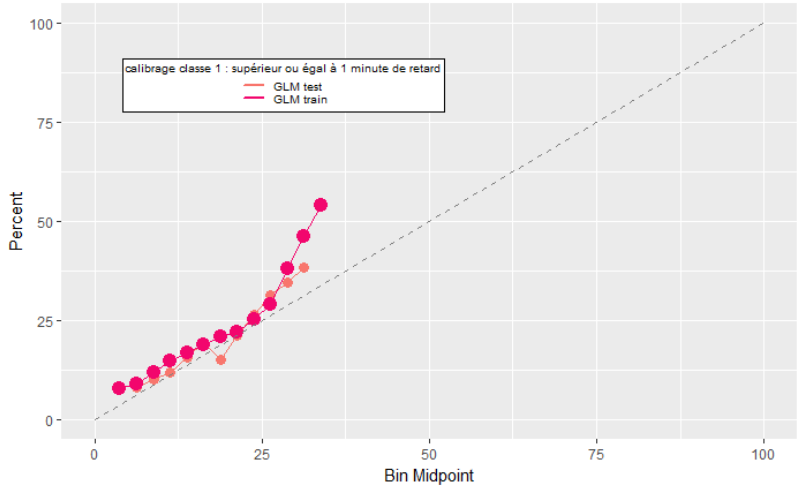
Proportions dans le dataset

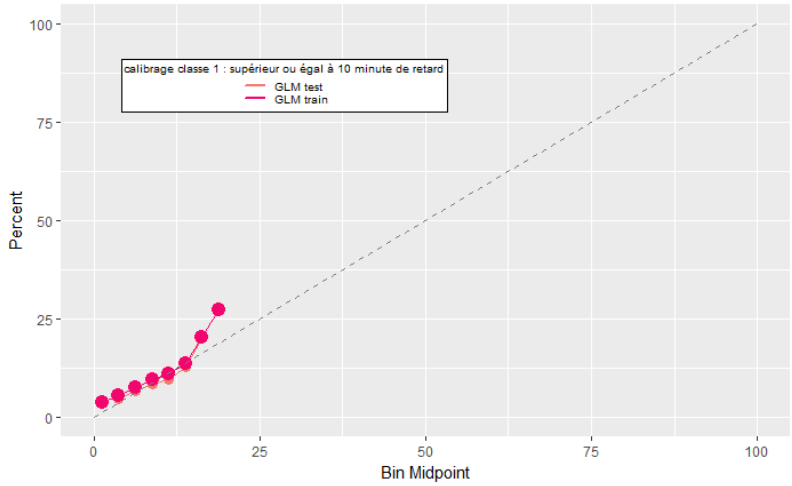
+Trains supérieurs à 1 minute : 51%

+Trains supérieurs à 6 minute : 18%

+Trains supérieurs à 10 minute : 9%







A FAIRE

Comparaison à d'autres modèles

+ GLM avec d'autres lois de probabilité
+ GLM en deux parties pour les cas nuls
+ Autres modèles permettant d'estimer une loi de probabilité

Mesures de performance

+ Calibration

Autres circulations

+ TGV au départ
+ Transiliens, TER

PERSPECTIVES

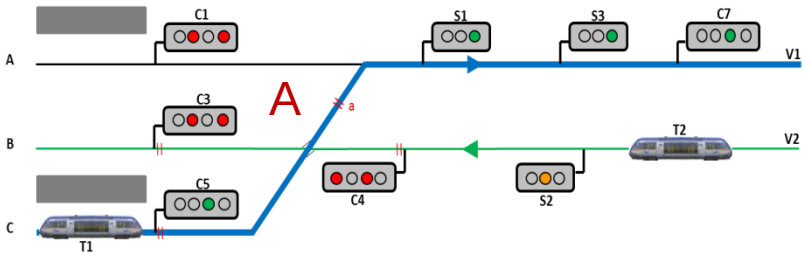
RETOUR SUR LE MODÈLE D'OPEN GOV

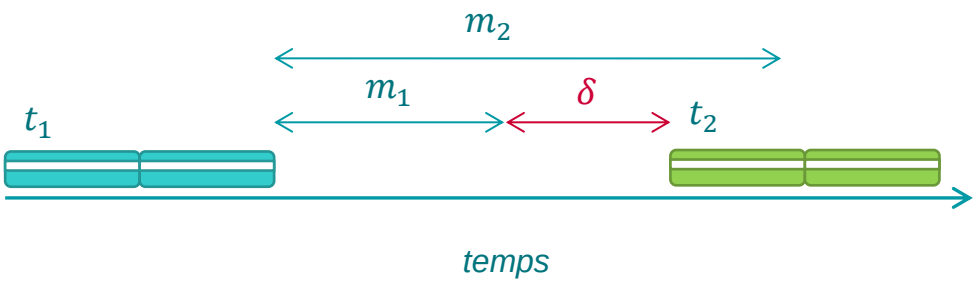
$$\min \sum_{t \in T} \sum_{i \in I_t} p_{i,t} \cdot x_{t,i} + \sum_{v \in V} p_v \cdot \gamma_v$$

$$\begin{cases} \sum_{i \in I_t} x_{t,i} = 1\ ,\ \forall\, t \in T \\ x_{t_1 i_1} + x_{t_2 i_2} \leq 1,\ (t_1, i_1, t_2, i_2) \in U \\ x_{t_1 i_1} + x_{t_2 i_2} \leq 1 + \gamma_v,\ v = (t_1, i_1, t_2, i_2) \in V \\ x_{t,i} \in \{0,1\},\ \ t \in T, i \in I_t \\ \gamma_v \in \{0,1\}, v \in V \end{cases}$$

PERSPECTIVES

DÉTECTION DE CONFLITS POTENTIELS





Options

+ Faire dépendre la pénalité $p_v, v \in V$ de la probabilité que le retard de t_1 soit supérieur à δ
+ Recalculer m_2 de façon à avoir une garantie en probabilité de robustesse, par exemple pour un seuil $s \in [0,1]$ fixé, $\mathbb{P}[r_1 \leq m_2 - m_1] \geq s$

MERCI DE VOTRE ATTENTION !
DES QUESTIONS ?