# Extraction de connaissances sur les défaillances de compteurs d'essieux

Iwo Doboszewski*,**,*** Simon Fossier**, Christophe Marsala***

*AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow
idobosz@agh.edu.pl,
**Thales Research & Technology France, 1 av. Augustin Fresnel, 91767 Palaiseau, France
simon.fossier@thalesgroup.com
***Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606,
4 place Jussieu 75005 Paris
christophe.marsala@lip6.fr

**Résumé.** Cet article propose une méthode d'analyse pour des enregistrements opérationnels d'un ensemble de compteurs d'essieux, qui constituent un élément central à l'infrastructure ferroviaire. Notre objectif est de fournir une façon efficace d'extraire automatiquement des éléments de connaissance concernant les défaillances de ces systèmes.

Puisque les données fournies ne contiennent pas de vérité de terrain sur les causes de défaillances, les informations et leurs causes doivent être extraites des relations sous-tendant les événements enregistrés. Après une phase de prétraitement, les événements sont groupés en fonction des relations qui ont été mises en lumière entre eux. Ces regroupements peuvent ensuite être utilisés pour créer des classes d'événements en utilisant un système de classification adapté.

Au delà de cette application spécifique, cette approche est une façon nouvelle d'aborder les problèmes d'analyse de fiabilité.

## 1 Introduction

Pour la sûreté et l'opérabilité des trains, l'infrastructure des réseaux ferroviaires est massivement surveillée, en temps réel, par des opérateurs de contrôle de trafic. Les données sont enregistrées et analysées pour améliorer leur fiabilité (Rosenberger et Pointner, 2015). Les sorties principales du système de surveillance sont les détections de pannes et leur diagnostic. En outre, les informations sur les contrôles, réparations et révisions sont parfois enregistrées, ce qui facilite l'extraction d'une information utile à partir des données.

Cet article propose une méthode d'extraction de connaissances pour une telle situation. Nous travaillons avec des données composées d'événements enregistrés automatiquement dans le système : passage de trains, mouvements d'aiguillages, défaillances d'équipement, etc. Il n'y a pas de vérité terrain sur les causes de défaillance, ni d'information sur les réparations. Nous savons aussi qu'une partie des défaillances enregistrées ne résultent pas d'une défaillance

physique des dispositifs, mais plutôt d'une absence de réponse, possiblement provoquée par des causes externes indépendantes.

L'étude se concentre sur un sous-ensemble des dispositifs : les compteurs d'essieux. Pour extraire une connaissance sur les défaillances à partir des données, nous proposons une procédure en deux étapes : d'abord, une détection et un filtrage des rapports de défaillance qui semblent provenir de conditions externes ; puis, à partir des relations entre événements consécutifs, une classification des défaillances par effets et fréquence. Cette procédure est une étape préliminaire pour une possible analyse de fiabilité (Schroeder et Gibson, 2007).

Cet article est structuré de la façon suivante. Dans la section 2, les compteurs d'essieux et le jeu de données utilisé sont décrits, suivis par les étapes de prétraitement des données. Dans la section 3, le modèle de traitement est présenté. Enfin, nous présentons nos résultats et conclusions. Notons que, dans cet article, la terminologie utilisée pour le domaine de la maintenance est cohérente avec celle du European Committee for Standardization (2010).

# 2 Compteurs d'essieux et jeu de données

## 2.1 Compteurs d'essieux

Un compteur d'essieux est composé de bobines inductives placées le long des voies et d'un estimateur, connectés électriquement. Quand l'essieu d'un train passe, il perturbe le champ électromagnétique entre les bobines, conduisant à un changement de tension dans le circuit, mesuré par l'estimateur et comparé avec un seuil prédéfini. Le passage du seuil est signalé au système et enregistré comme un passage de train (Rosenberger (2011), Wei et al. (2010)). Dans notre cas, les valeurs sont uniquement accessibles dans la couche basse du système et ne sont pas enregistrées par le système de surveillance. L'estimateur compte par ailleurs le nombre d'essieux entrant et sortant de la section, et se signale comme libre (non-occupé) au module d'enclenchement quand ces nombres sont égaux. Ces comptes ne sont pas accessibles ici.

Il y a différentes sources de défaillance pour les compteurs d'essieux :

— une dégradation physique. Les bobines sont placées le long de la voie, en environnement hostile (vibrations, pluie, humidité, poussière et large gamme de températures), ce qui dégrade les circuits, les sortant de leur zone de fonctionnement nominale.

— une erreur de comptage. La perturbation du champ électromagnétique entre bobines peut provoquer une erreur de comptage d'un essieu. Dans ce cas, indépendemment de son état réel, la section de voie est traitée comme perturbée, et une défaillance est enregistrée par le système. L'opérateur doit alors réinitialiser le compteur. Ces situations sont, a priori, les causes de défaillance les plus fréquentes, bien que ces erreurs puissent aussi provenir d'une dégradation des rails.

— des causes externes, conduisant à un problème opérationnel du compteur. Parmi celles-ci : rupture d'alimentation, fermeture de section de voie en maintenance, ou redémarrages système et mises à jour.

## 2.2 Description des logs

Les données prennent la forme de logs journaliers enregistrés par le système de supervision ferroviaire. Ils proviennent d'une station polonaise, sur 11 mois, du 1er janvier au 17 novembre

2015. Nous en avons extrait des enregistrements provenant de 79 compteurs d'essieux, qui ont signalé 692 défaillances au total.

Nous ne disposons pas d'un historique des données et ne pouvons donc pas observer la vie complète des équipements. Les systèmes de surveillance enregistrent les événements dans des disques locaux de taille limitée. Quand de nouvelles données arrivent et que le disque est plein, les plus anciennes sont effacées. La période d'enregistrement ne correspond donc pas au cycle de vie des équipements ou au plan de maintenance, rendant une estimation directe de la distribution du temps de vie résiduel impossible (par exemple via l'estimateur Kaplan-Meier (Hosmer et al., 2008)).

Les maintenances, informations constructeur et autres données techniques ne sont pas présentes. Nous ne connaissons donc pas les causes et solutions au problèmes signalés. Seuls sont enregistrés les défaillances détectées et le passage des trains, sous la forme de messages `occupé` ou `libre`, couplés à l'identifiant d'équipement et l'horodatage de l'événement.

### 2.3 Prétraitement des données

Parfois, plusieurs équipements signalent des défaillances dans un temps court (typiquement 30 minutes). Selon les experts, de tels rapports quasi-simultanés sur différents équipements ne sont pas des défaillances physiques, et doivent provenir d'événements externes : redémarrages systèmes, phénomènes naturels, coupures de courant, etc. nous parlerons ici d'*interruptions de service*. Des groupes de défaillances sont ici considérés comme interruptions de service si au moins 4 équipements sont affectés dans un intervalle de 30 minutes, et ils sont alors retirés des données avant analyse.

## 3 Introduction au modèle de traitement

Comme expliqué précédemment, les données analysées sont non-étiquetées, et il est difficile de construire explicitement un modèle du problème, ce qui nous a dirigé vers des techniques d'apprentissage non-supervisé. De nombreuses techniques de traitement peuvent être utilisées pour ce type de problème : recherche de règles d'association, partitionnement de données, cartes auto-adaptatives, réduction de dimensionnalité (par ex. analyse en composantes principales) (James et al., 2009).

Nous avons choisi une analyse de partitionnement des données qui offre une très bonne méthode de représentation de plusieurs modes de défaillance. En outre, notre base de défaillances est de taille modeste et sa représentation vectorielle est de faible dimension, ce qui rend la réduction de dimensionnalité peu utile.

La méthode de partitionnement choisie est le clustering hiérarchique, dans lequel une heuristique permet de décider du nombre de groupes à retenir à partir de l'analyse du dendrogramme généré (James et al., 2014, chapter 4). Les regroupements en clusters sont effectués à l'aide de la distance euclidienne avec une approche *complete linkage* pour les calculs de distances entre groupes [1].

1. Les calculs ont été réalisés à partir de la bibliothèque SciPy (Jones et al., 01 ).

### 3.1 Représentation vectorielle des défaillances

Nous souhaitons différencier les types de défaillances selon la fréquence à laquelle elles apparaissent et leur effet sur le fonctionnement, mesuré par la période d'inopérance qu'elles induisent. Les variables suivantes ont été choisies pour décrire les défaillances :

— $t_{last}$ : durée depuis la dernière défaillance (en secondes) ;
— $t_{next}$ : durée avant la prochaine défaillance (en secondes) ;
— $op_{last}$ : nombre de cycles d'opération depuis la dernière défaillance ;
— $op_{next}$ : nombre de cycles d'opération avant la prochaine défaillance ;
— $t_{off}$ : durée de fonctionnement incorrect depuis la dernière défaillance (en secondes).

La variable $t_{off}$ est associée à la perturbation introduite par la défaillance dans le système et la sévérité de celle-ci. En particulier, si la défaillance a été résolue par un redémarrage du compteur, le temps avant que l'équipement soit à nouveau opérationnel devrait rester court. Cette variable est critique pour déterminer les effets des défaillances sur l'infrastructure.

Rappelons que nous ne savons pas si une indisponibilité longue de l'équipement est due à une réelle défaillance ou si, pour une raison quelconque, l'équipe de maintenance a pris un temps inhabituel pour atteindre et réparer l'équipement, ou même si l'équipement a été réparé rapidement mais n'a pas pu être utilisé par la suite.

Les première et dernière défaillances d'un enregistrement, pour lesquelles $t_{last}$ ou $t_{next}$ n'est pas connu, ont été incluses dans l'analyse si une période suffisante (fixée à deux semaines) existe entre le début des enregistrements et la défaillance. C'est cette durée qui est alors utilisée dans les calculs.

La plupart des variables a une distribution dense sur les valeurs faibles, le reste étant plus épars, ce qui montre que la plupart des événements a lieu dans des intervalles de temps plutôt courts. Pour certains événements, ces intervalles peuvent aller jusqu'à plusieurs jours.

Il y a une corrélation nette entre les variables $t_{next}$ et $op_{next}$ (et entre $t_{last}$ et $op_{last}$). Intuitivement, si le trafic est réparti de façon homogène à l'échelle d'une année, la relation entre les deux variables doit être quasi-linéaire sur un compteur. Ceci correspond globalement à nos observations, mais un certain nombre d'événements s'éloignent de cette relation.

Pour le partitionnement des données, chaque variable a été normalisée dans $[0, 1]$ par division par son maximum global : les variables ont des unités différentes, et il n'est pas pertinent de les comparer directement. De plus, puisque l'écart entre les opérations est d'au plus quelques minutes, les variables temporelles à valeurs élevées auraient rendu insignifiant l'impact des autres variables.

Une fois le partitionnement effectué, nous caractérisons les clusters sur la base des distributions de variables présentées ci-dessus. Avec cinq variables, il est possible d'étudier les distributions manuellement, afin d'interpréter opérationnellement les résultats algorithmiques.

## 4 Résultats

À partir de l'analyse du dendrogramme, une coupure permettant d'obtenir 3 clusters a été choisie. Au total, 487 signalements ont été identifiés comme résultant d'un événement externe (section 2.2). Par ailleurs, 5 signalements ont eu lieu trop près du début ou de la fin des mesures pour que toutes les valeurs des variables puissent être fournies (section 3.1). Il y a 42 défaillances dans le cluster 1, 136 dans le cluster 2 et 22 dans le cluster 3.

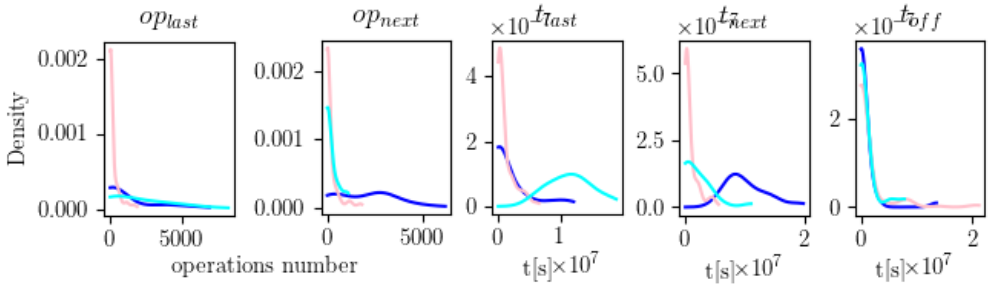


FIG. 1 – *Distribution des variables de partitionnement entre les trois clusters.*

La Fig. 1 présente les distributions lissées des variables introduites dans la section 3.1 pour chaque cluster. Dans le cluster 1, un grand nombre de défaillances ont de hautes valeurs de $t_{next}$ et $opnext$. Les valeurs de $t_{last}$ and $op_{last}$ sont aussi généralement plus hautes que dans le cluster 2. Nous pouvons interpréter ces défaillances comme aléatoires et ne résultant pas de la tendance de l'équipement à tomber en panne, mais plutôt comme des défaillances exceptionnelles. Dans le cluster 2, les variables $t_{last}$ et $t_{next}$ ont des valeurs plutôt basses. Ces défaillances sont séparées entre elles par des durées courtes. Elles peuvent être interprétées comme systématiques, le compteur d'essieux tombant en panne de façon répétée. Le cluster 3 contient des défaillances à haut $t_{last}$, et $t_{next}$ est légèrement plus haut que dans le cluster 2. Ces défaillances sont interprétées comme rares ou aléatoires, comme dans le cluster 1.

Les valeurs de $t_{off}$ ne sont pas spécifiques à un cluster. Bien que le cluster 2 présente des valeurs plutôt plus hautes, la tendance est peu claire. Ce résultat est légèrement décevant, dans la mesure où nous nous attendions à ce que $t_{off}$ permette de séparer les défaillances à conséquences sérieuses de celles qui n'ont eu que peu d'impact sur le trafic. Ainsi, le cluster dans lequel les valeurs de $t_{off}$ auraient été en général plus hautes et les autres variables plus basses aurait contenu les défaillances les plus sérieuses : fréquentes, et à conséquences fortes.

# 5 Conclusion

L'étude présentée dans cet article est la première étape d'une analyse de fiabilité d'équipements de trafic ferroviaire en présence de données limitées et incomplètes. Les étapes présentées fournissent une façon d'appréhender la validation des données et la construction d'un système de classification pour les défaillances, en fonction des spécificités du domaine. Bien que le problème soit loin d'être résolu, nous avons obtenus quelques résultats positifs.

L'approche par clustering hiérarchique a mis en évidence trois groupes dans les données, avec des caractéristiques différentes. Bien que n'expliquant pas complètement les modes de défaillances, les résultats sont prometteurs.

En perspective, on peut remarquer que certaines améliorations seraient intéressantes à étudier dans le but de proposer d'autres types de clusters, par exemple, l'ajout de nouvelles variables. Actuellement, l'espace décrit dans la section 3.1 est fortement hétérogène car la plupart des points ont des valeurs très basses pour toutes les variables, alors que quelques uns ont des

valeurs élevées et proches des maximums. Par ailleurs, l'utilisation d'autres techniques de clustering de données sera étudiée dans des travaux futurs car elle permettrait de mettre en avant une caractérisation des défaillances selon d'autres angles de représentation.

## Remerciements

Iwo Doboszewski est soutenu par une bourse d'étude du gouvernement français dans le cadre d'une thèse en co-tutelle internationale entre l'UPMC et l'université AGH.

## Références

European Committee for Standardization (2010). European standard 13306/2010.

Hosmer, D. W., Jr., S. Lemeshow, et S. May (2008). *Applied Survival Analysis : Regression Modeling of Time to Event Data, 2nd Edition*. John Wiley & Sons, Inc.

James, G., D. Witten, T. Hastie, et R. Tibshirani (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer.

James, G., D. Witten, T. Hastie, et R. Tibshirani (2014). *An Introduction to Statistical Learning with Applications in R*. Springer.

Jones, E., T. Oliphant, P. Peterson, et al. (2001–). SciPy : Open source scientific tools for Python. [Online ; accessed 2017-02-10].

Rosenberger, M. (2011). Future challenges to wheel detection and axle counting. *SIGNAL+ DRAHT 9*.

Rosenberger, M. et F. Pointner (2015). High availability : definition, influencing factors and solutions.

Schroeder, B. et G. A. Gibson (2007). Disk failures in the real world : What does an mttf of 1,000,000 hours mean to you ? In *FAST*, Volume 7, pp. 1–16.

Wei, C.-l., C.-c. Lai, S.-y. Liu, et al. (2010). A fiber bragg grating sensor system for train axle counting. *IEEE Sensors Journal 10*(12), 1905–1912.

## Summary

This paper proposes an approach to analyze operation records of axle counters, a core part of railway infrastructure. Our aim is to introduce an efficient way to automatically extract knowledge regarding failures of such devices.

As the data provided does not contain a ground truth regarding causes of failures, failure information and their causes should be extracted from underlying relations between recorded events. After a data pre-processing step, the recorded events are clustered with respect to the relationships that can be highlighted among them. As a result, classes of events can be highlighted, from which a classification system can be proposed.

Beyond this specific application, the approach is a novel way to tackle reliability analysis problems.