expandium

ISO 9001:2008
Certified Company

# DATA SCIENCE @EXPANDIUM

**Nicolas Greffard PhD,** Data Scientist

# DATA SCIENCE

@EXPANDIUM

**Expandium**

- 11 ans

**Métier**

- Monitoring télécom

**Solutions**

- Software / Hardware => carte d'acquisition réseau + décodage/traitement en temps réel

**Clients**

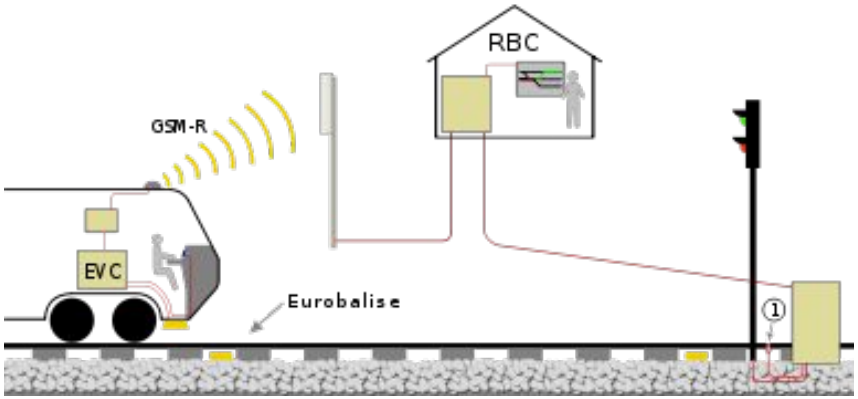
- Railway => 9 millions de lignes = 2 semaines
- Opérateurs publics (SFR etc…) => 9 milliards de lignes = 3 jours

# DATA SCIENCE

@EXPANDIUM

## GSM-R

- Permet aux trains de communiquer avec les postes de régulation du trafic ferroviaire & aux agents de conduite, de circulation et de maintenance de communiquer entre eux…
- ETCS (European Train Control System) : normalisation des protocoles GSM-R entre tous les pays européens (23 protocoles non compatibles différents..)



# DATA SCIENCE

@EXPANDIUM

## Données GSM-R

- TAC (mobile)
- DL/UL_rxlev (niveau de champ) : [-110,-43] db => [0,63] (- normale)
- DL/UL_rxqual (qualité de la voix) : [0, 7] (- power law)
- Nid_c Nid_bg (identifiant dernière balise rencontrée)
- distance_balise (distance à la dernière balise) [0,63]x50m
- lac ci (identifiant BTS)
- TA (timing advance, distance à la BTS) : [0, 63] x 550m
- timestamp
- nid_engine (identifiant locomotive)
- nid_operational (identifiant trajet, eg Paris=>Nantes@9h != Paris=>Nantes@11h)

# DATA SCIENCE

@EXPANDIUM

# DATA SCIENCE

@EXPANDIUM

## Reference Analytics

- Comparer chaque nouvel appel avec son appel de **référence**
- Faire remonter les différences significatives

## Méthodologie

- Algorithme en 2 temps
  - 1) Apprentissage de la topologie du réseau & sélection des appels de référence
  - 2) Comparaison des appels avec leur référence

## Restitution

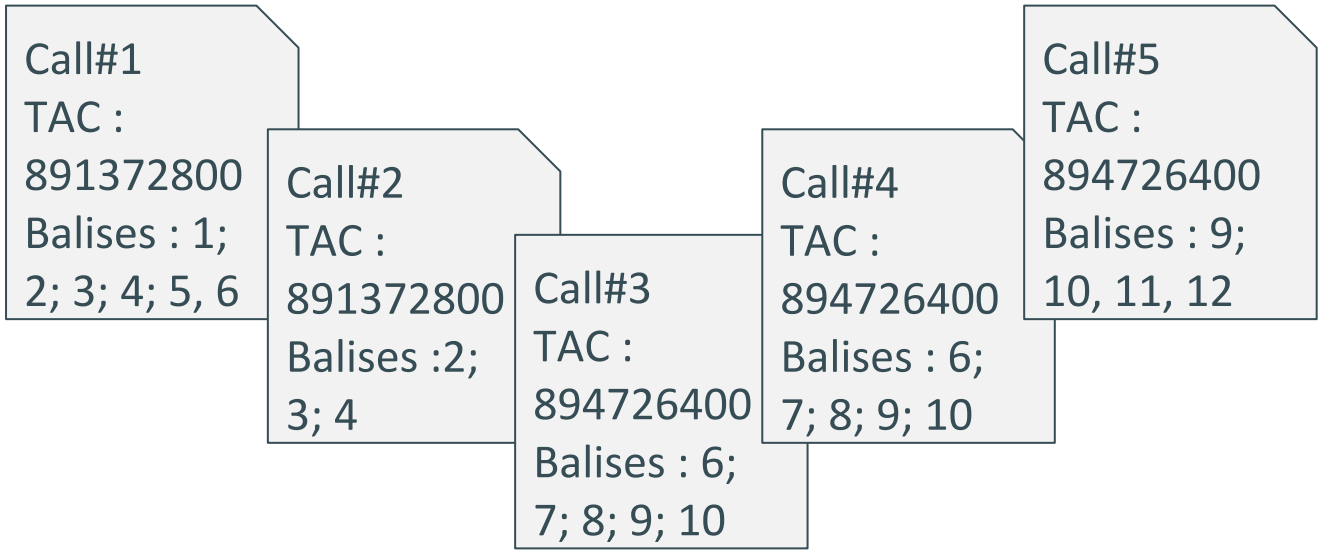
- Vue tabulaire des problèmes identifiés
- Graphe des mesures radio (appel + référence)

# SÉLECTION DE LA RÉFÉRENCE

## QATS REFERENCE ANALYTICS

**Fonctionnement**

- Lancé périodiquement (ex: toutes les deux semaines)
- 1) Regroupe les appels émis de trains parcourant les mêmes sections (et dans le même sens) et étant équipés d'équipements radio similaires (TAC)
  - Effectué avec une mesure de similarité entre les ensembles de balises ETCS traversées pendant l'appel (intersection normalisée des deux ensembles)

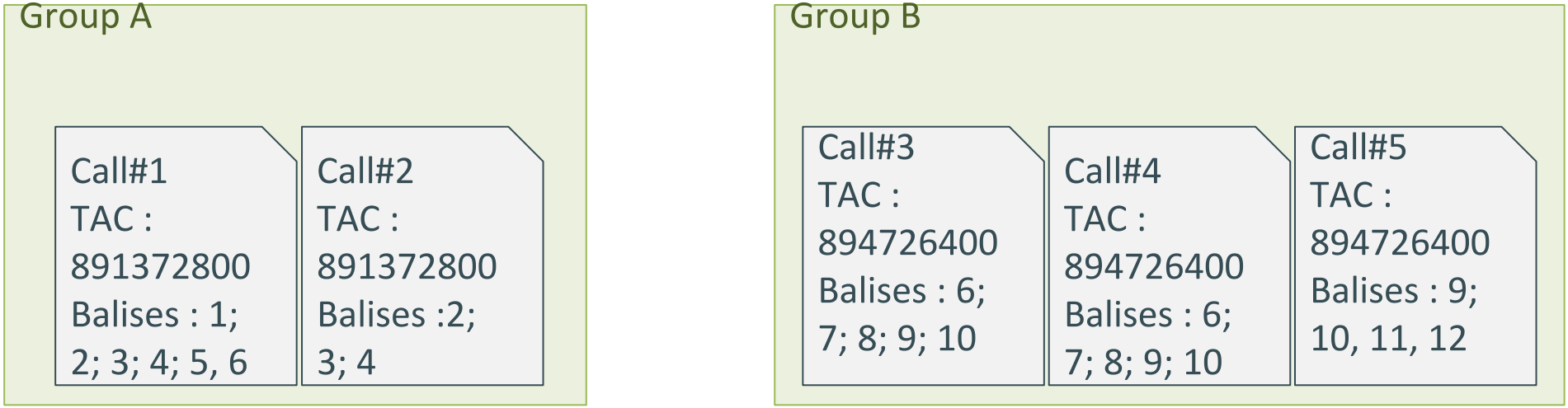


# SÉLECTION DE LA RÉFÉRENCE

## QATS REFERENCE ANALYTICS

**Fonctionnement**

- Lancé périodiquement (ex: toutes les deux semaines)
- 1) Regroupe les appels émis de trains parcourant les mêmes sections (et dans le même sens) et étant équipés d'équipements radio similaires (TAC)
  - Effectué avec une mesure de similarité entre les ensembles de balises ETCS traversées pendant l'appel (intersection normalisée des deux ensembles)



# SÉLECTION DE LA RÉFÉRENCE

QATS REFERENCE ANALYTICS

**Fonctionnement**

- 2) Pour chaque groupe, on cherche l'appel ayant les meilleurs performances radio
  - Grossièrement : $\max_{call}( \int ecdf(\text{Radio meas.}_{call}) + \log(\#\text{balises}_{call}) )$
  - ie : plus les mesures radio sont bonnes et plus l'appel est long (et couvre de distance) plus il a de chance d'être sélectionné comme référence pour son groupe

Group B

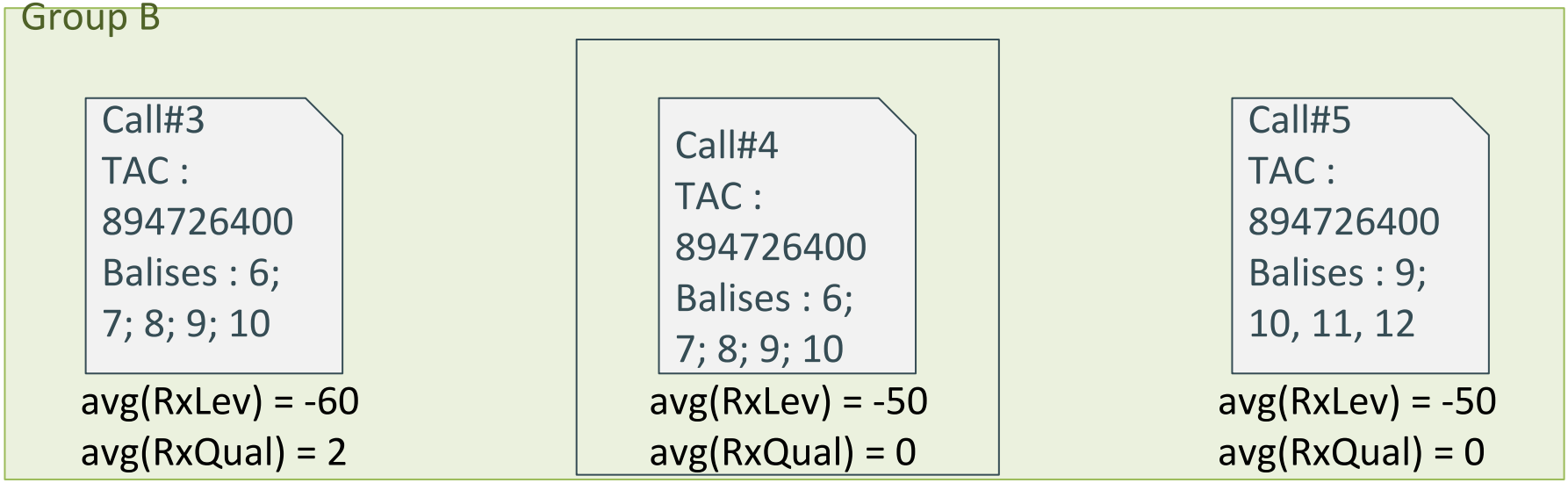
| Call#3<br>TAC :<br>894726400<br>Balises : 6;<br>7; 8; 9; 10 | Call#4<br>TAC :<br>894726400<br>Balises : 6;<br>7; 8; 9; 10 | Call#5<br>TAC :<br>894726400<br>Balises : 9;<br>10, 11, 12 |
|---|---|---|
| avg(RxLev) = -60<br>avg(RxQual) = 2 | avg(RxLev) = -50<br>avg(RxQual) = 0 | avg(RxLev) = -50<br>avg(RxQual) = 0 |

# SÉLECTION DE LA RÉFÉRENCE

## QATS REFERENCE ANALYTICS

### Fonctionnement

- 2) Pour chaque groupe, on cherche l'appel ayant les meilleurs performances radio
  - Grossièrement : $\max_{call}( \int ecdf(\text{Radio meas.}_{call}) + \log(\#\text{balises}_{call}) )$
  - ie : plus les mesures radio sont bonnes et plus l'appel est long (et couvre de distance) plus il a de chance d'être sélectionné comme référence pour son groupe



# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Détection d'anomalies

- Pour chaque nouvel appel, on extrait son appel de référence
- Pré-processing : alignement des mesures radios

| Rxlev | RxQual | balise | distance |
|---|---|---|---|
| -55 | 0 | 1 | 0 |
| -53 | 0 | 1 | 50 |
| -54 | 0 | 2 | 0 |
| -50 | 1 | 2 | 50 |
| -52 | 0 | 3 | 0 |
| -56 | 0 | 3 | 50 |
| -55 | 0 | 3 | 100 |

| Rxlev | RxQual | balise | distance |
|---|---|---|---|
| -53 | 0 | 1 | 0 |
| -51 | 0 | 1 | 50 |
| -52 | 0 | 1 | 100 |
| -49 | 0 | 2 | 0 |
| -51 | 0 | 2 | 50 |
| -57 | 0 | 3 | 0 |
| -53 | 0 | 3 | 50 |

# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Détection d'anomalies

- Pour chaque nouvel appel, on extrait son appel de référence
- Pré-processing : alignement des mesures radios

| Rxlev | RxQual | balise | distance |
|---|---|---|---|
| -55 | 0 | 1 | 0 |
| -53 | 0 | 1 | 50 |
| -54 | 0 | 2 | 0 |
| -50 | 1 | 2 | 50 |
| -52 | 0 | 3 | 0 |
| -56 | 0 | 3 | 50 |
| -55 | 0 | 3 | 100 |

| Rxlev | RxQual | balise | distance |
|---|---|---|---|
| -53 | 0 | 1 | 0 |
| -51 | 0 | 1 | 50 |
| -52 | 0 | 1 | 100 |
| -49 | 0 | 2 | 0 |
| -51 | 0 | 2 | 50 |
| -57 | 0 | 3 | 0 |
| -53 | 0 | 3 | 50 |

# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Détection d'anomalies

- Pour chaque nouvel appel, on extrait son appel de référence
- Pré-processing : alignement des mesures radios

| Rxlev | RxQual | balise | distance |
|---|---|---|---|
| -55 | 0 | 1 | 0 |
| -53 | 0 | 1 | 50 |
| -54 | 0 | 2 | 0 |
| -50 | 1 | 2 | 50 |
| -52 | 0 | 3 | 0 |
| -56 | 0 | 3 | 50 |

| Rxlev | RxQual | balise | distance |
|---|---|---|---|
| -53 | 0 | 1 | 0 |
| -51 | 0 | 1 | 50 |
| -49 | 0 | 2 | 0 |
| -51 | 0 | 2 | 50 |
| -57 | 0 | 3 | 0 |
| -53 | 0 | 3 | 50 |

# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

**Anomalies côté train**

- Nous avons deux séries de mesures radio échantillonnées aux mêmes endroits
- **Hypothèse :** Sachant que ces mesures proviennent approximativement d'équipements (TAC identiques) et de zones similaires (balises identiques), leurs distributions devraient être semblables
- Réfuter cette hypothèse pour une paire (appel, référence) pourrait indiquer une anomalie au niveau de l'équipement radio du train
- Elle peut être "facilement" vérifiée via un test statistique classique

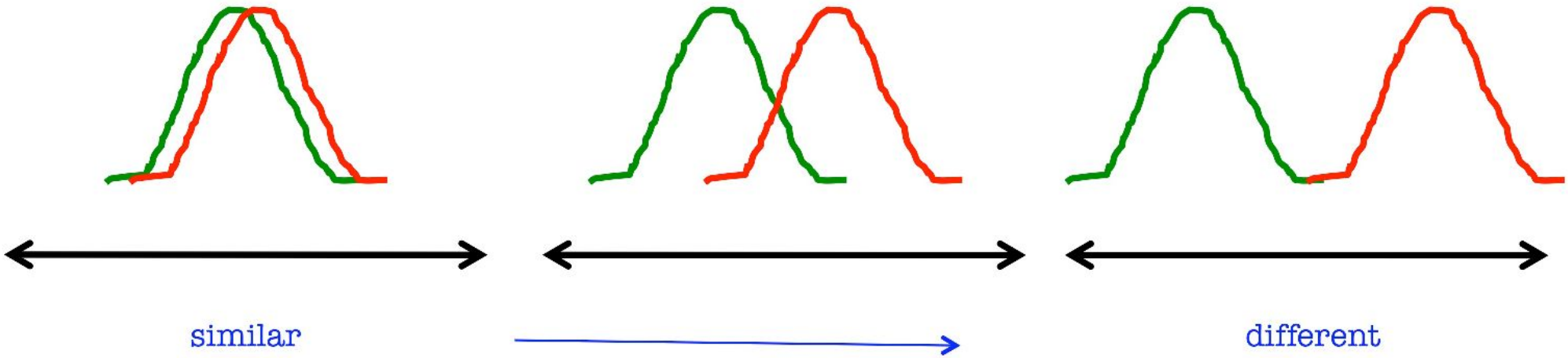
**Keywords :** T-test, Wilcoxon test, Friedman test

# DÉTECTION D'ANOMALIES

QATS REFERENCE ANALYTICS

**Anomalies côté train**

- Tests statistiques d'homogénéité : est-ce que deux échantillons proviennent de la même population ?



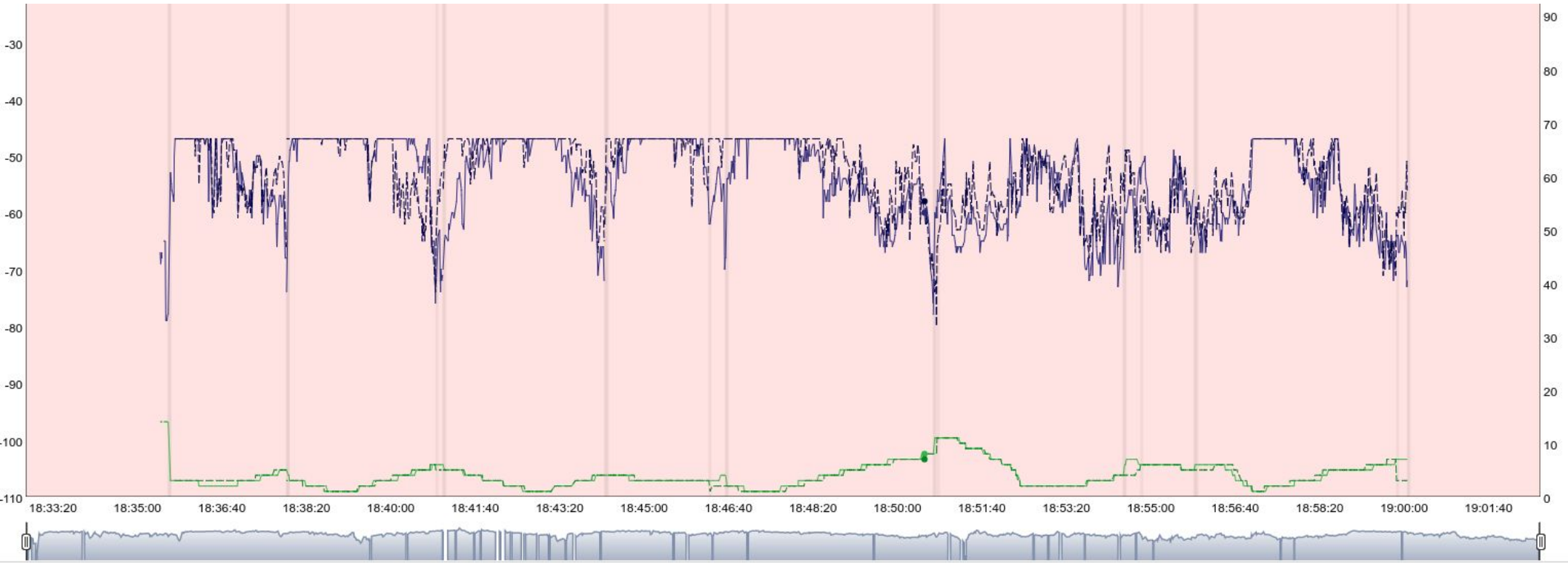

**Keywords :** T-test, Wilcoxon test, Friedman test

# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Anomalies côté train

- Comportement normal :
- Chevauchement des mesures radio pendant toute la durée de l'appel
- DL Rxlev

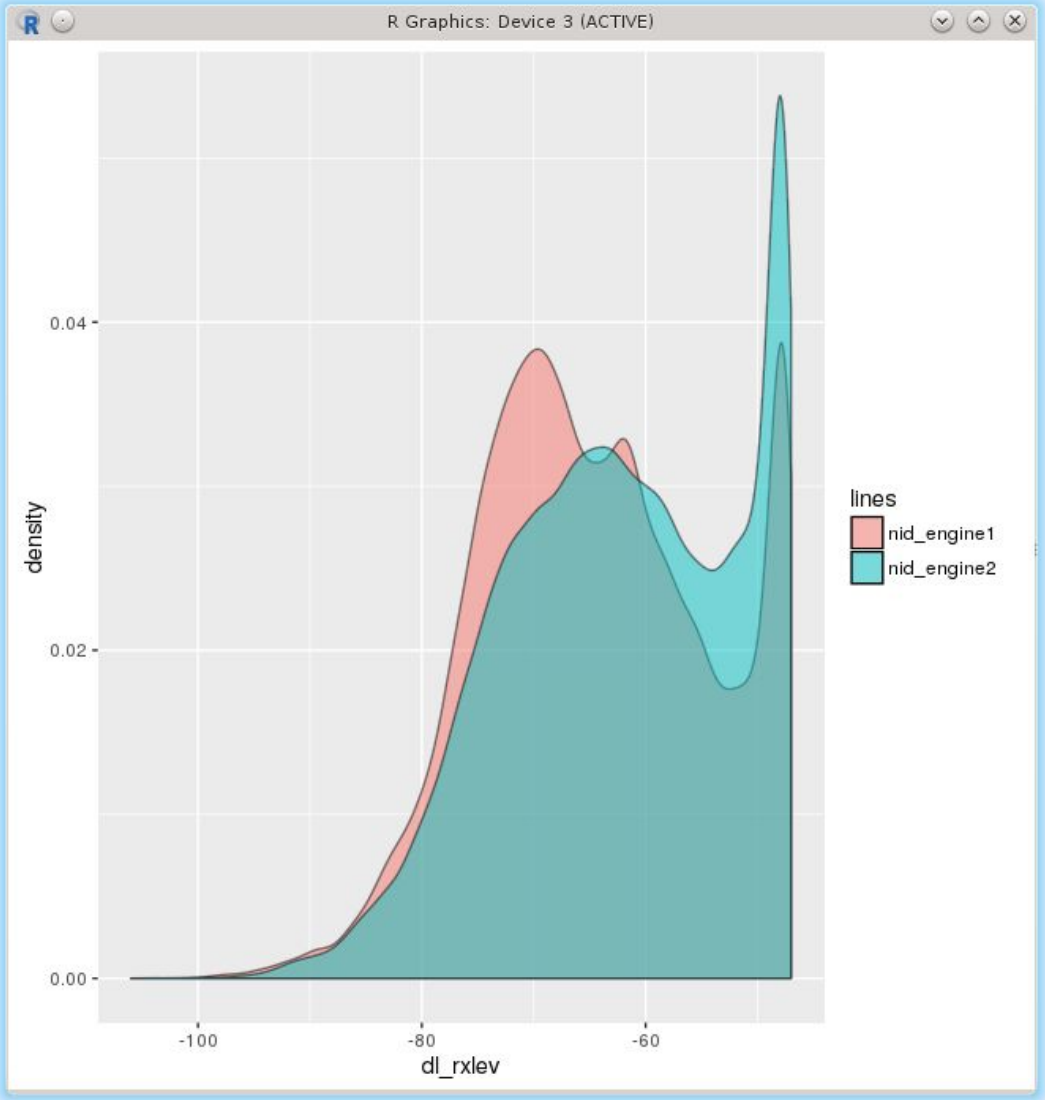


# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Anomalies côté train

- Ce que l'on attend :
- Distributions similaires
- Chevauchement des distributions



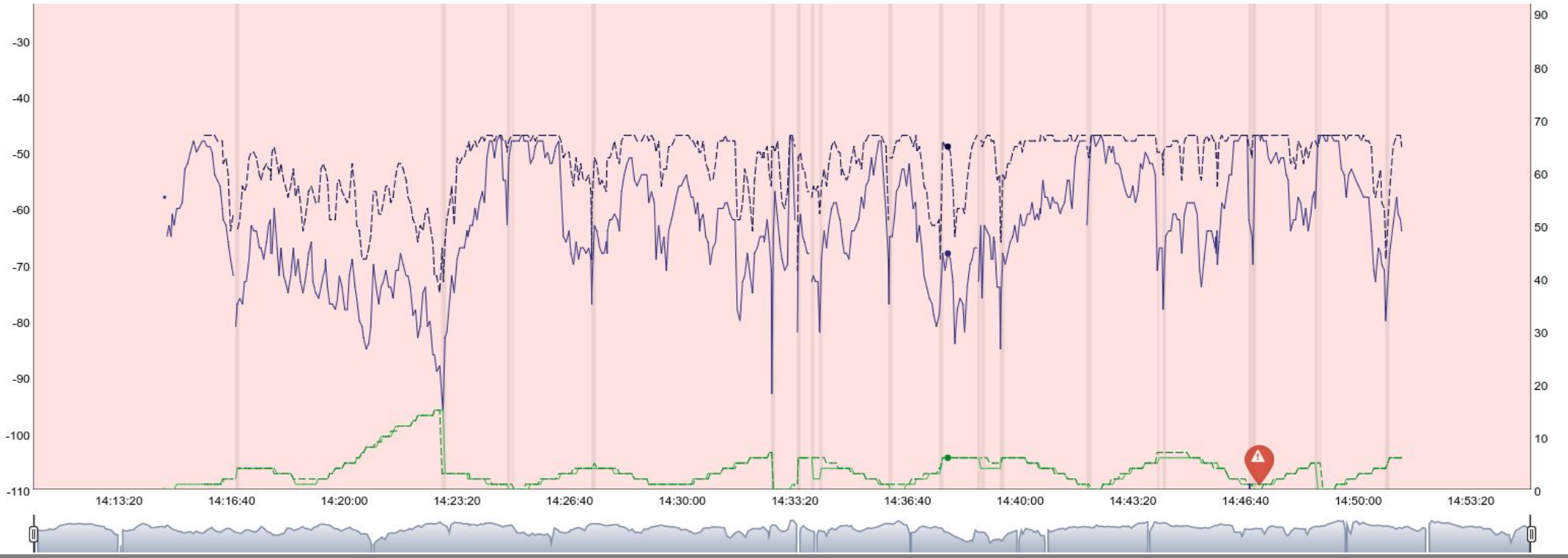

*Distribution de la densité de probabilité (DL Rxlev)*

# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Anomalies côté train

- Ce que l'on identifie comme comportement anormal :
- Différences importantes sur l'intégralité de l'appel
- DL Rxlev



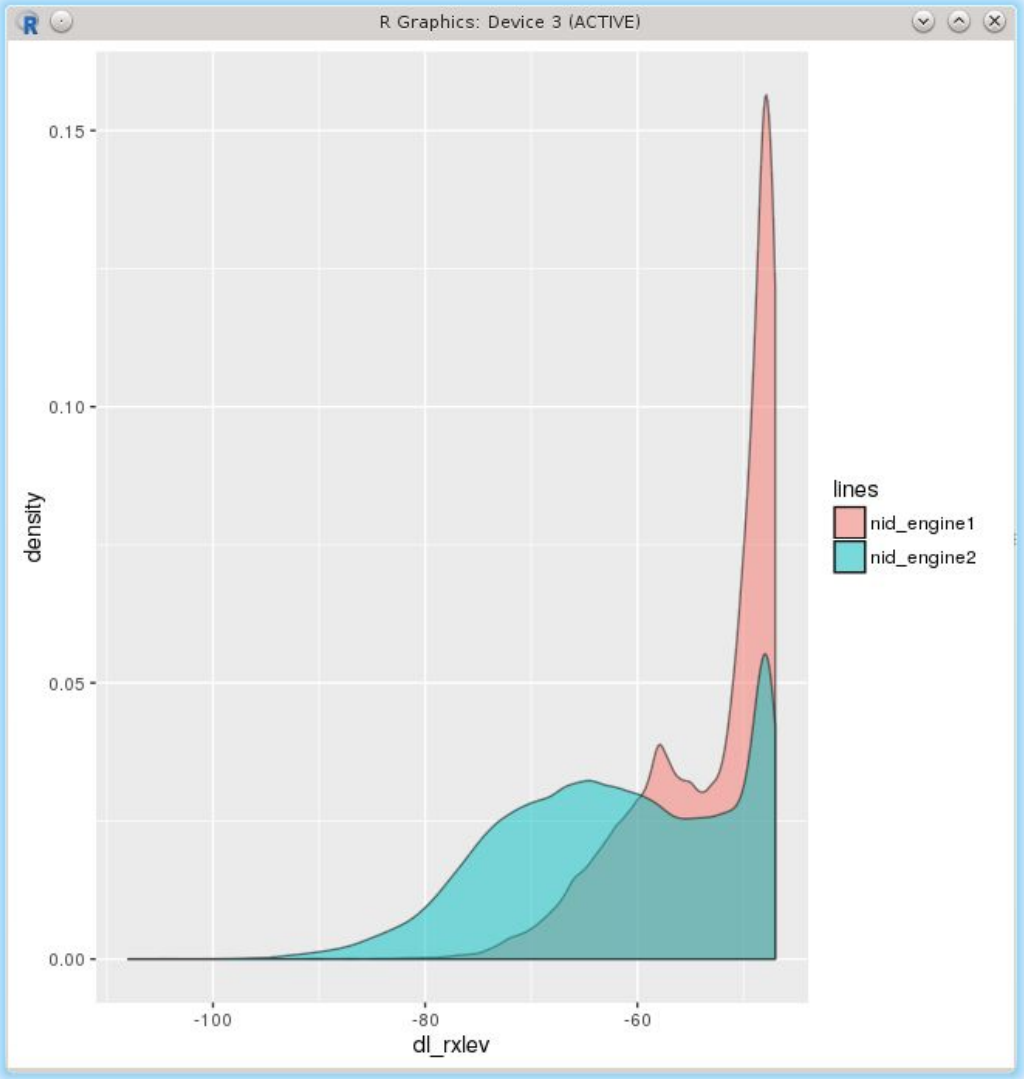
# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

**Anomalies côté train**

- Comportement anormal :
- Divergence des distributions
- Conclusion :
  - nid_engine2 a bien plus de chances

d'être sujet à de mauvaises mesures radio




*Distribution de la densité de probabilité (DL Rxlev)*

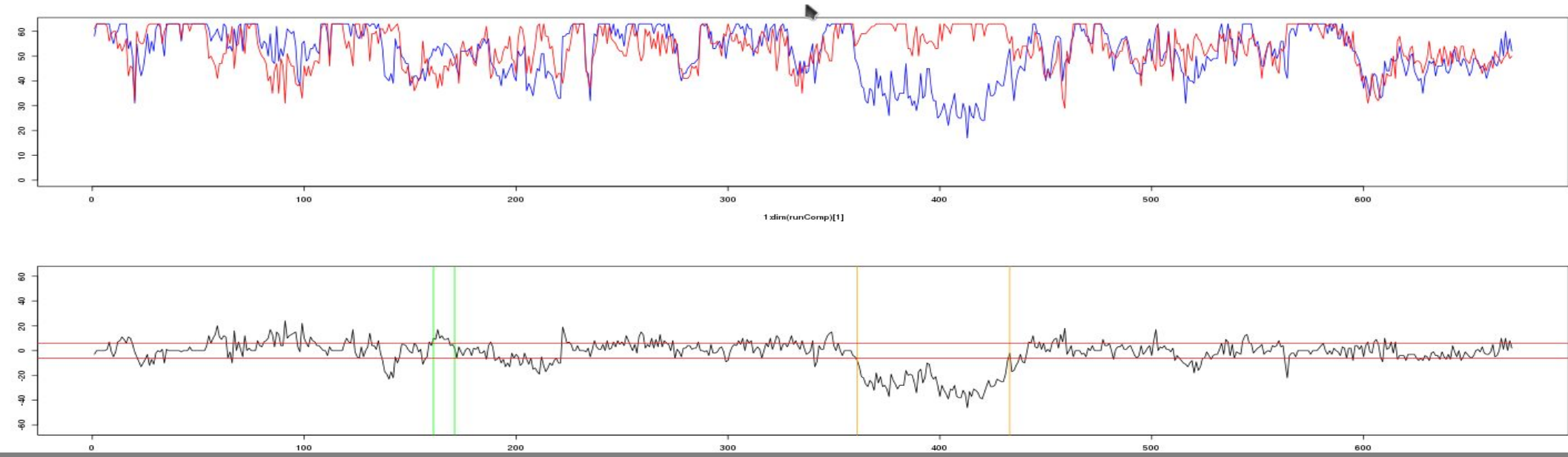
# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

**Anomalies côté cellule**

- Si il n'y a pas d'anomalie côté train, l'analyse se poursuit :
- **Hypothèse :** Puisque les mesures radio suivent globalement la même distribution, une différence locale peut être imputée à un problème lié à une cellule

- Un algorithme en une passe permet d'identifier ces **zones d'intérêt**



# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

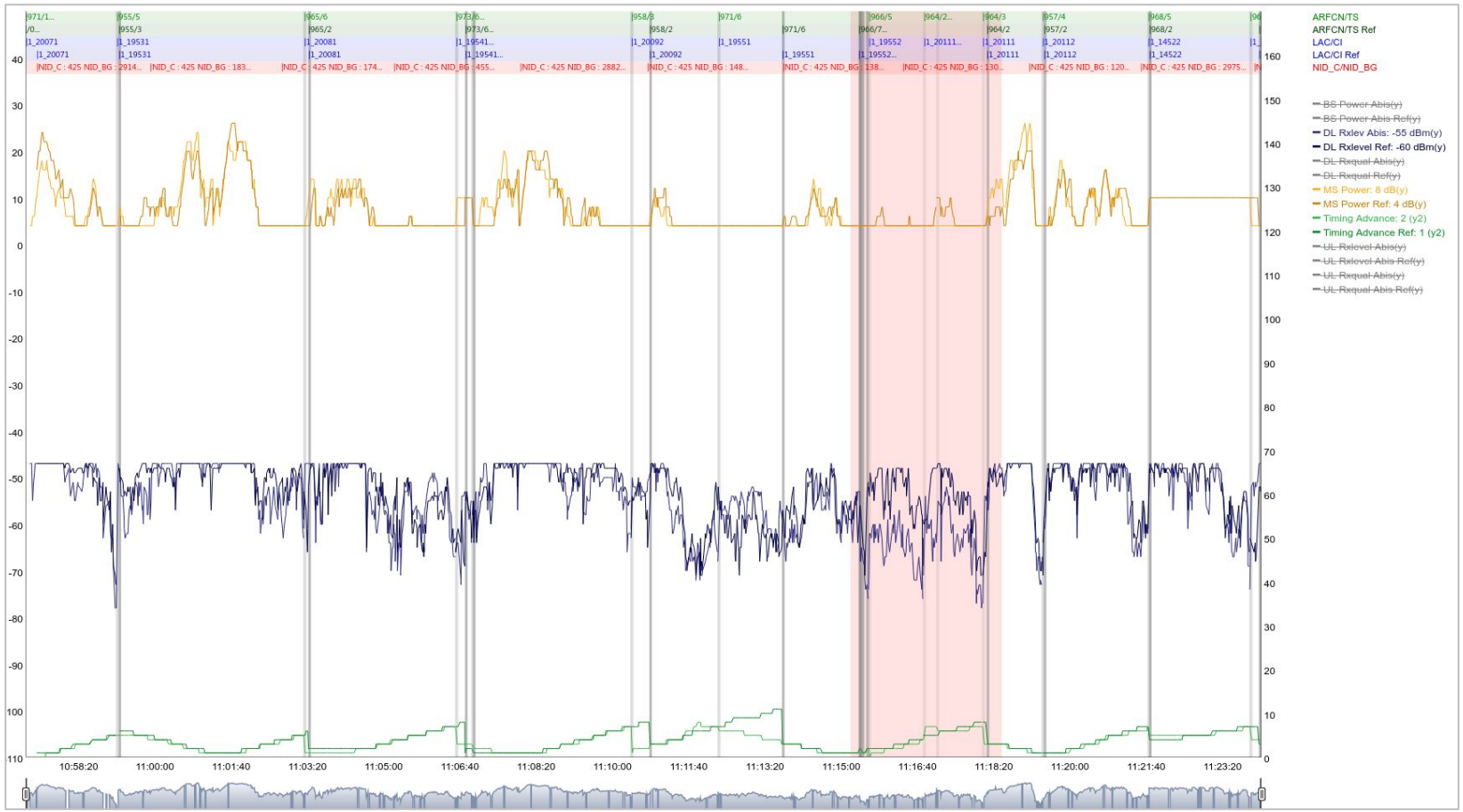
**Anomalies côté cellule**

- Pour chaque zone d'intérêt:
  - On applique le même test statistique que précédemment sur les sous ensembles de données correspondants à la cellule incriminée sur les deux appels
  - Si le test échoue (ie: les distributions divergent) alors une anomalie côté cellule est remontée

# DÉTECTION D'ANOMALIES

QATS REFERENCE ANALYTICS

## Anomalies côté cellule



# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Anomalies côté cellule

# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

**Autres anomalies**

- Certaines zones d'intérêt couvrent de nombreuses cellules
- Certaines durent très longtemps ou représentent une amplitude de divergence très élevée
- La cause de ces anomalies peut dépendre de plusieurs facteurs externes (météo, croisement de trains,etc...)
- On souhaite faire ressortir les anomalies les plus critiques :
  --> problème de classification à une classe

**Keywords : Artificial Neural Network, AutoEncoder**

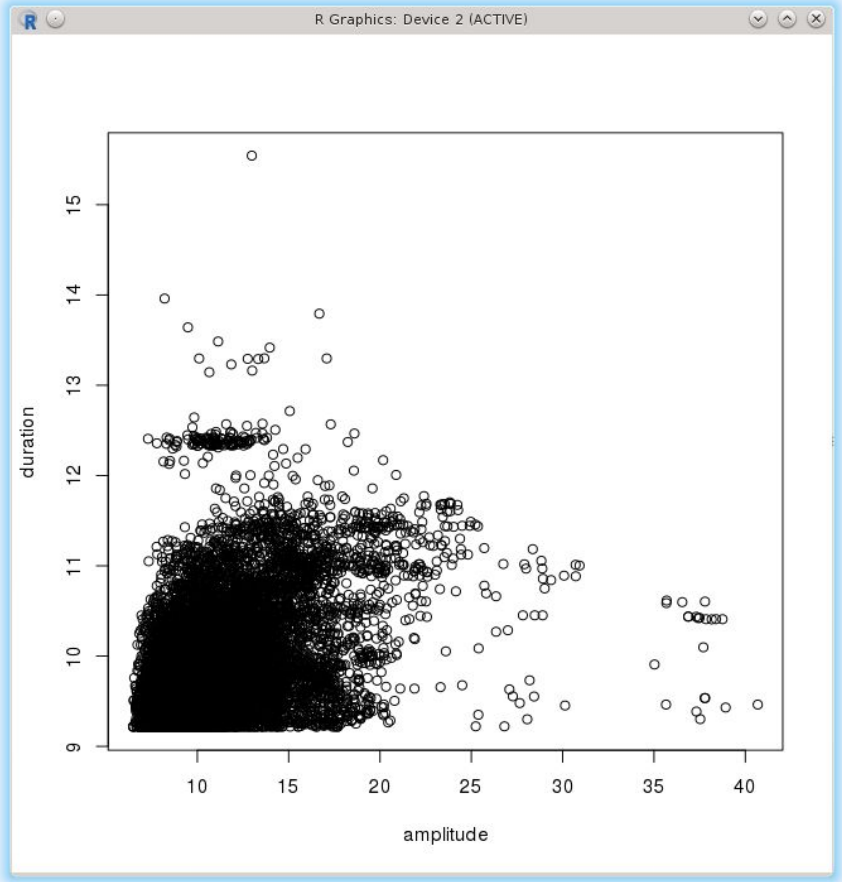
# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Autres anomalies

On caractérise chaque zone d'intérêt :

- sa durée
- son amplitude
- la fréquence d'anomalie du TAC correspondant
- la fréquence d'anomalie du nid_engine correspondant
- la fréquence d'anomalie de la cellule correspondante
- le ratio du temps passé à l'arrêt pendant la durée de l'anomalie radio



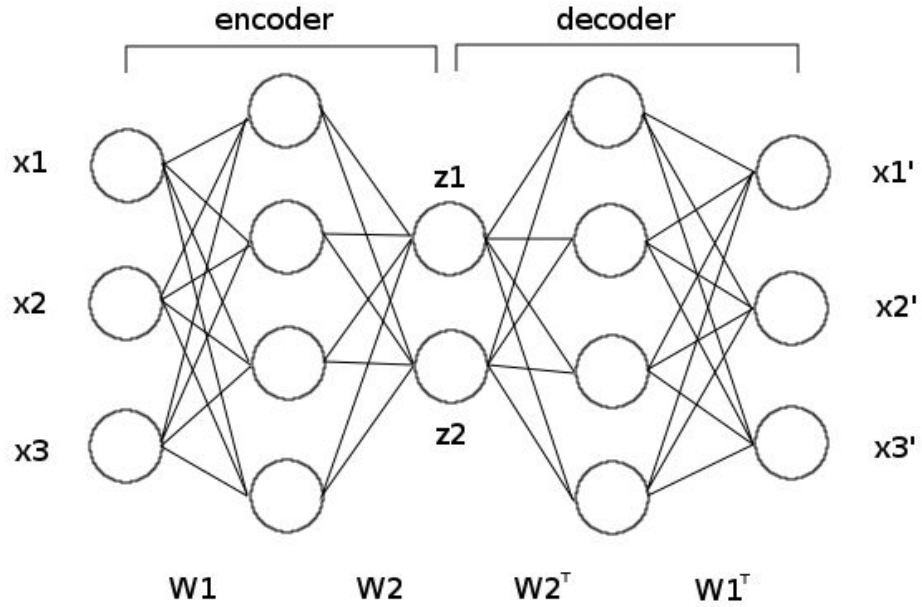
# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Autres anomalies

On utilise un auto-encoder

- de la famille des réseaux de neurones




- topologie en forme de sablier
- on calcule l'erreur de reconstruction et non de prédiction
- ici les neurones (z1,z2) contiennent une version "compressée" de l'information de (x1,x2,x3)

# DÉTECTION D'ANOMALIES

QATS REFERENCE ANALYTICS

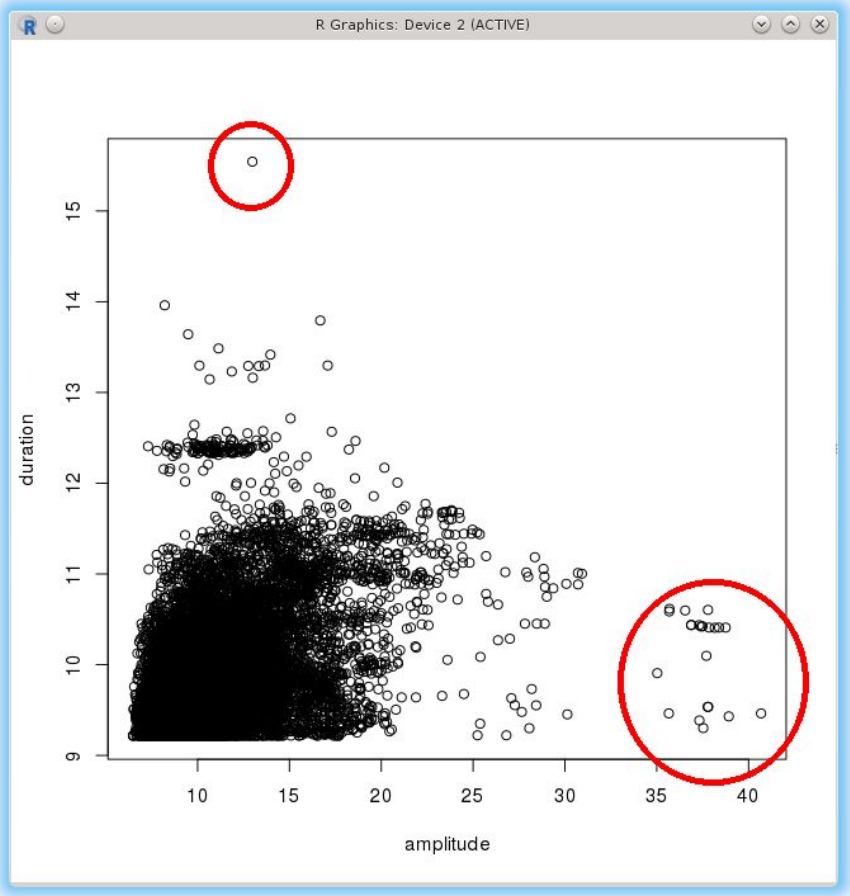
**Autres anomalies**

Auto-encoder

- Apprend les comportements normaux / attendus
- Identifie les comportements anormaux / nouveaux

**E.g. :**

- 1ère anomalie d'un équipement donné
- Durée particulièrement élevée
- Amplitude particulièrement élevée

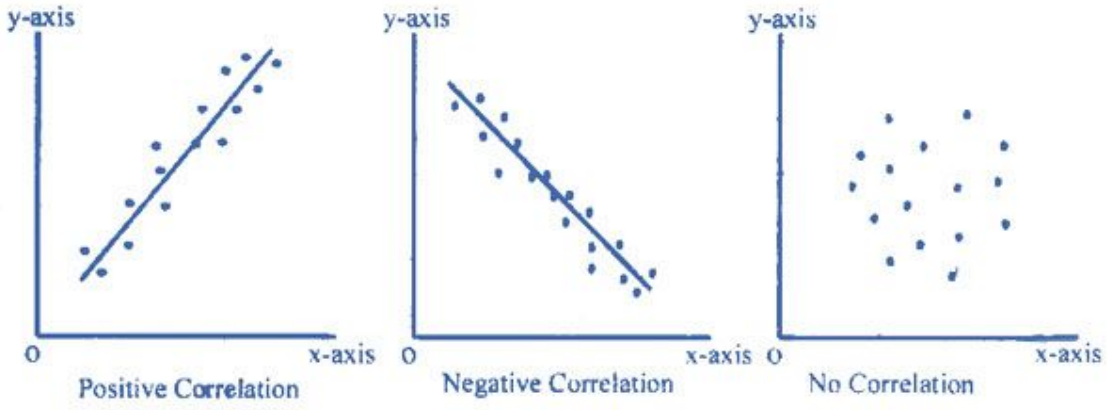


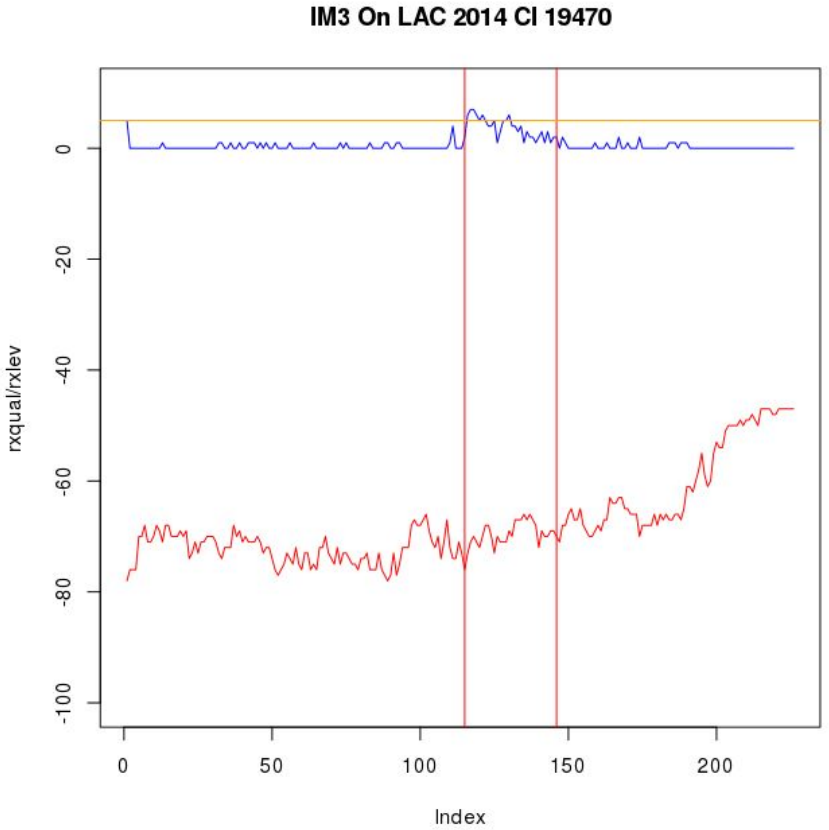
# DÉTECTION D'ANOMALIES

QATS REFERENCE ANALYTICS

## Interférences

- Anomalie spécifique
- Très mauvais Rx QUAL non explicable par un mauvais Rx LEV
- Rien de sorcier :
- Seuils  & non corrélation croisée négative



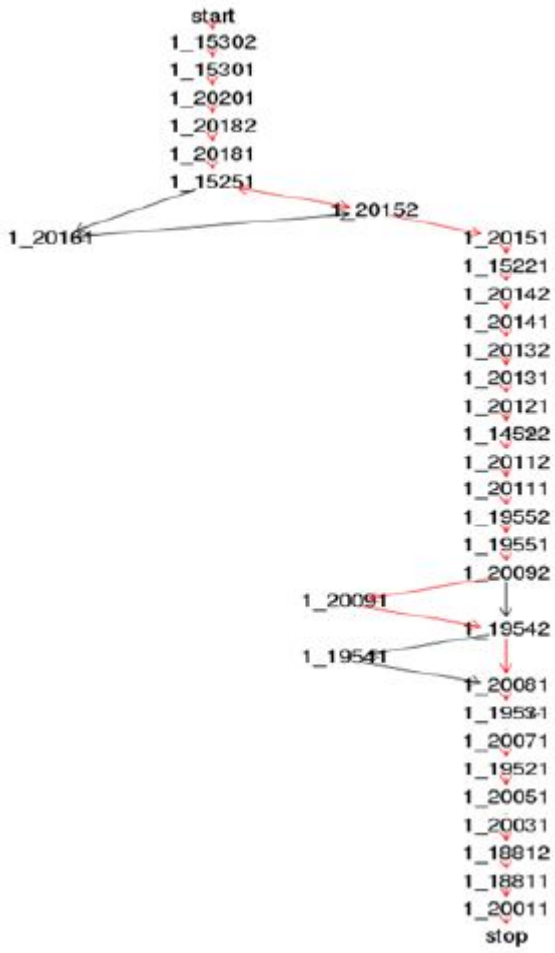



# DÉTECTION D'ANOMALIES

## QATS REFERENCE ANALYTICS

### Anomalies de handovers

- Handover : passage d'une cellule à une autre
- Topologie du réseau : séquence pré-établie
- Déviation de cette séquence : anormale




- Maintien de la matrice de transition :
- Nouveau lien ou faible probabilité = anormal
- ping-pong = anormal

# APPLICATION GSM-R

## QATS REFERENCE ANALYTICS

**Résumé**

- Détection d'anomalies statistiques à différents niveaux de granularité
- De liens improbables dans des graphes

**Techniquement**

- Package R maison
- lancé via cron / script / api rest
- déployé directement en packages debian ou via une image docker


Quid de la partie MNO ?

# MNO

## Généralités

- Données : xdr (transaction - un message réseau )
- Plusieurs xdrs pour un SMS / appel (initialisation, contenu, fermeture)
- D'autres xdrs : fonctionnement interne du réseau / location update
- Entre 3 et 4 milliards de xdrs par jour

- 1 xdr = plus de 140 champs (principalement vides)
- champs les plus intéressants dans notre cas :
- timestamp / calling_number / called_number / imsi / xdr_type / xdr_duration

- Objectif : Détection d'abus (exploitation des vulnérabilités de l'architecture SS7) et détection de fraudes (ping callback / spam / etc..)

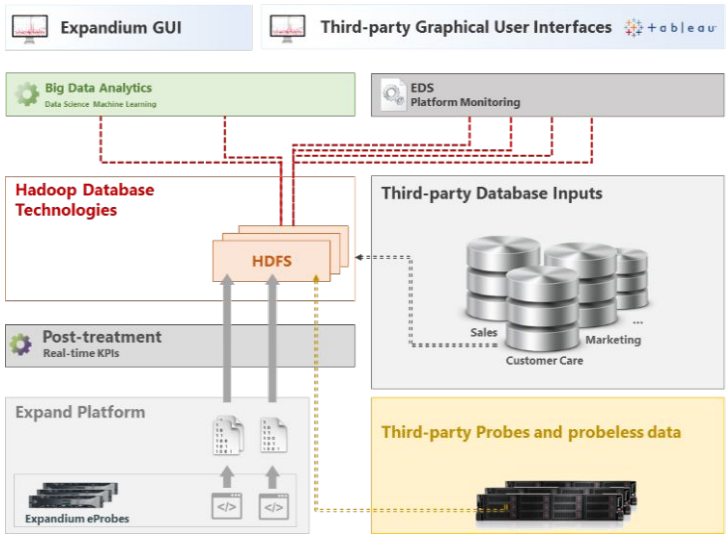
# MNO

## Architecture

- Stack Big Data classique
- HDFS => SPARK => H2O.ai

- Scalable State of the Art Machine Learning Algorithm

- Plusieurs POCs développés en interne

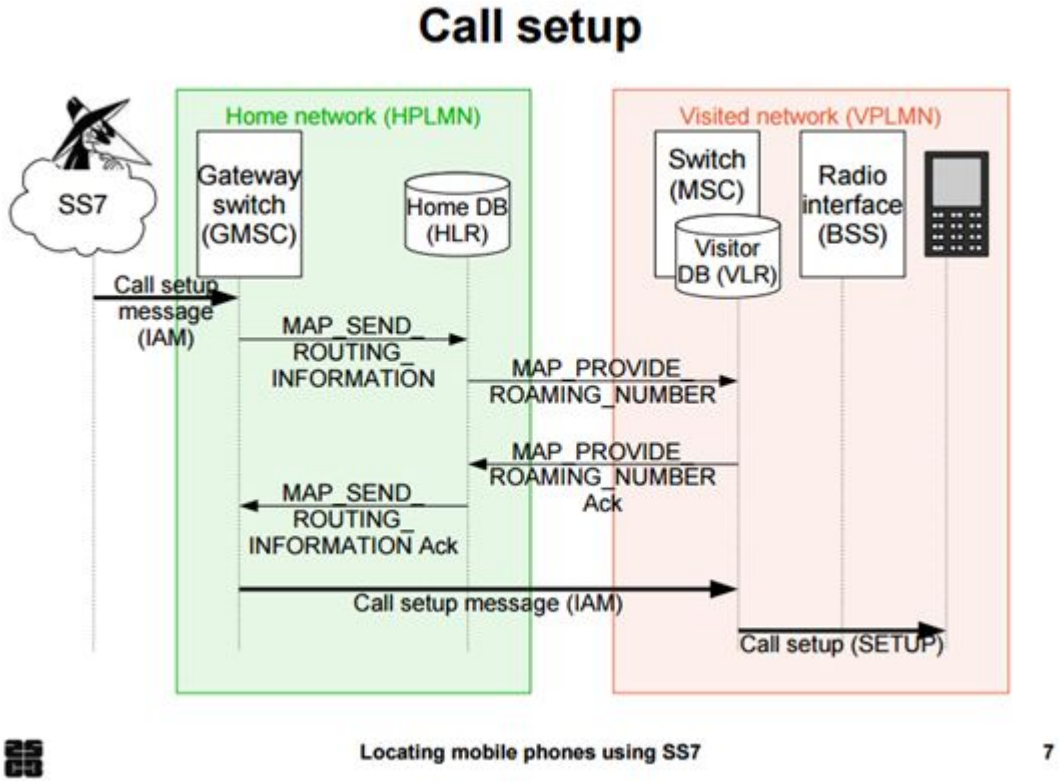


# MNO

## Failles SS7

- Architecture SS7 (Signalling System #7) : ensemble des protocoles de signalisation téléphonique permettant d'établir et de libérer des appels entre fixes et mobiles

- Vulnérabilités permettant notamment : la géolocalisation approximative d'un utilisateur / le re-routing d'appels



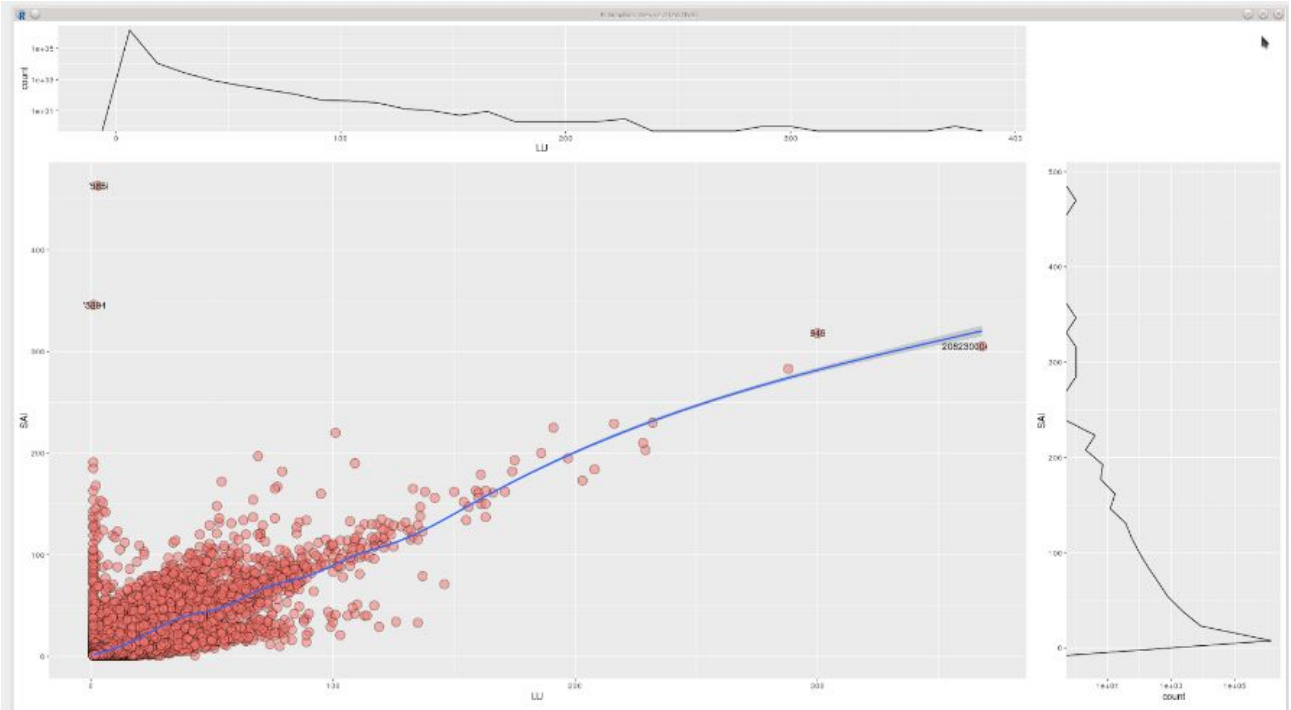
# MNO

## Failles SS7

- Nombre important de Send Authentification Info (SAI) par rapport au nombre de Location Update (LU)

- Tentative de spoofing ?

- Analyse sur 140k IMSI
- 10% des xdr MAP/INAP sur 10 jours (14m de xdrs)



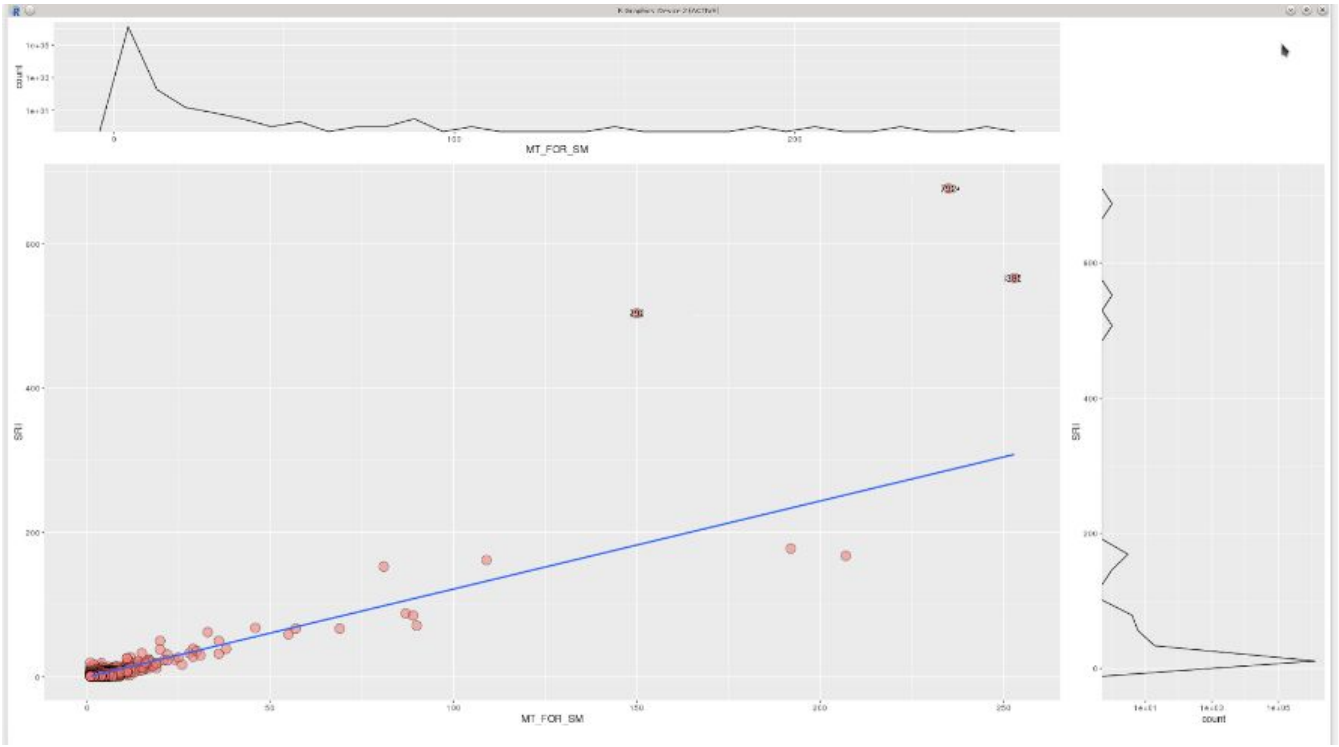
# MNO

## Failles SS7

- Nombre important de Send Rooting Info (SRI) par rapport au nombre de SMS_MT_FORWARD_SM

- Tentative de spoofing / de géo-localisation ?

- Analyse sur 140k IMSI
- 10% des xdr MAP/INAP sur 10 jours (14m de xdrs)



# MNO

## Comportements suspects

- Les seuils / probabilités permettant d'identifier les comportements précédents (ie : distance à la droite de régression) ne suffisent pas pour des comportements potentiellement frauduleux plus complexes :
  - Indirect / direct ping callback
  - Démarchage mobile (bloctel)

- Notre approche (inspirée de Becker et al. 2010) :
  - Calcul contenu de profils d'utilisation (par imsi)
  - Machine learning pour discriminer entre un profil OK et un profil de fraudeur
  - Ensemble Learning / Deep Learning

# MNO

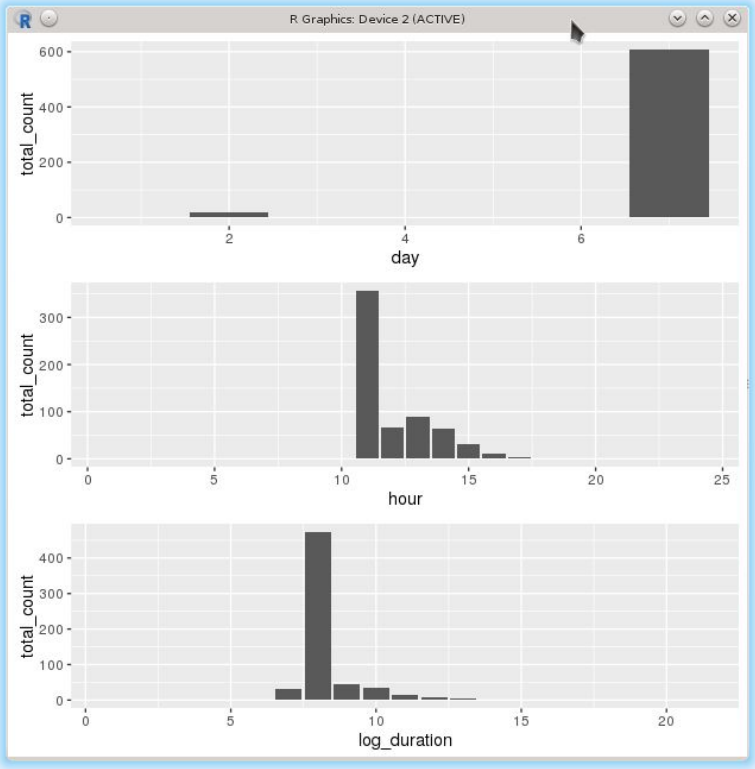
## Signatures d'utilisateurs

- Profils/Signatures : mis à jour par batch de 10-100k xdr via exponentially weighted moving average (EWMA) : Xn = Dc * θ + Xp * (1-θ)
- Xn : nouveau profil, Xp : ancien profil, Dc : nouvelles données; θ : paramètre contrôlant la vitesse à laquelle les anciennes données sont "oubliées"
- Données contenues dans chaque profil :

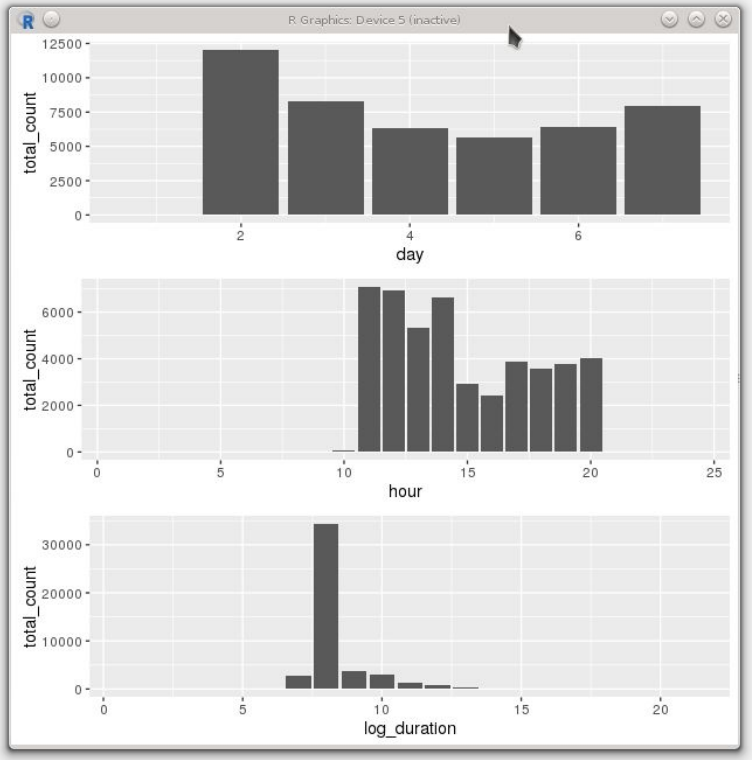
```
calling_number total_count offhour_weekday offhour_weekend onhour_weekend
        1:     427******         149      0.03179982      0.02015144      0.7221188
          onhour_weekday logtime_0 logtime_1 logtime_2 logtime_3 logtime_4 logtime_5
        1:       14.91852         0         0         0         0         0         0
          logtime_6 logtime_7 logtime_8 logtime_9 logtime_10 logtime_11 logtime_12
        1:         0         0        14         9        45        35        26
          logtime_13 logtime_14 logtime_15 logtime_16 logtime_17 logtime_18 logtime_19
        1:          6         0         0         0         0         0         0
          logtime_20 dow_0 dow_1 dow_2 dow_3 dow_4 dow_5 dow_6 hod_0 hod_1 hod_2 hod_3
        1:          0     0     5    32    41    40    19     0     0     0     0     0
          hod_4 hod_5 hod_6 hod_7 hod_8 hod_9 hod_10 hod_11 hod_12 hod_13 hod_14
        1:     0     0     0     0     7    19     15     14      8      5     12
          hod_15 hod_16 hod_17 hod_18 hod_19 hod_20 hod_21 hod_22 hod_23
        1:     14     12     12      3      0      0      0      0      0
          mean_convtime_offhour_weekday mean_convtime_offhour_weekend
        1:                      77816.28                      2011.799
          mean_convtime_onhour_weekend mean_convtime_onhour_weekday
        1:                     91875.45                     128481.2
```

# MNO

## Profils



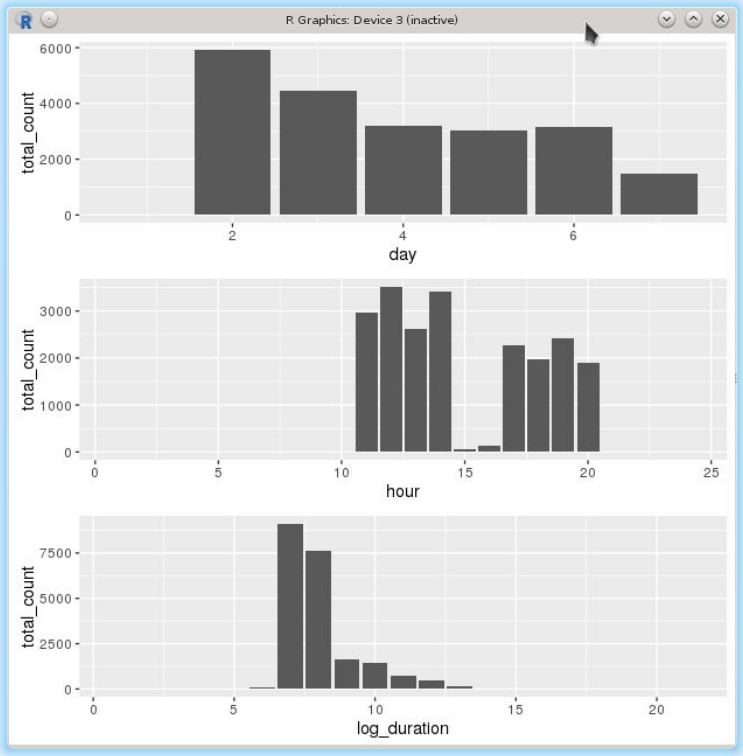

Utilisateur “normal” avec un grand nombre d’appels le samedi matin



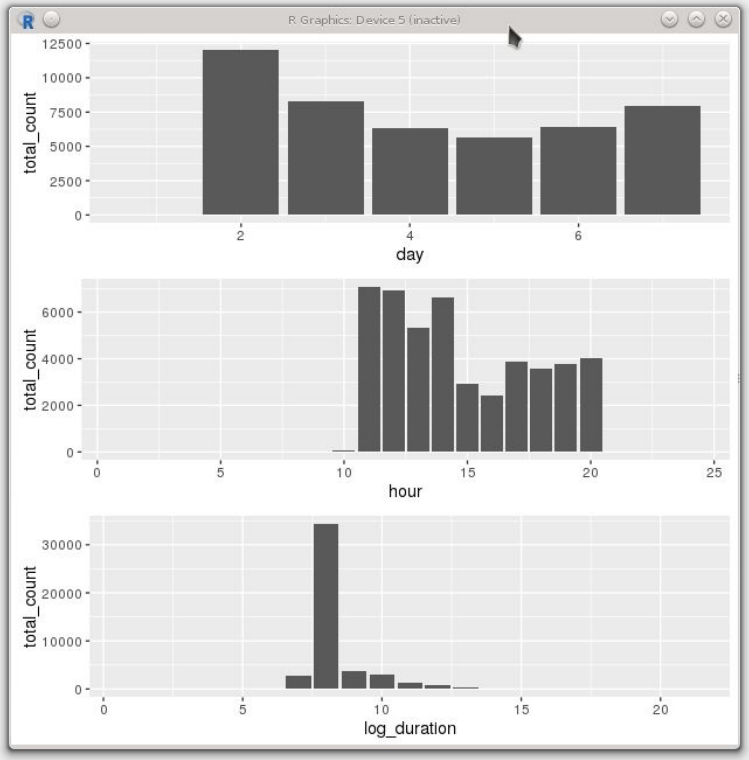

Centrale d’appel qui appelle toute la journée tous les jours sauf le dimanche

# MNO

## PROFILS




Centrale d'appel qui appelle toute la journée tous les jours sauf le dimanche et qui autorise les pauses le midi ?

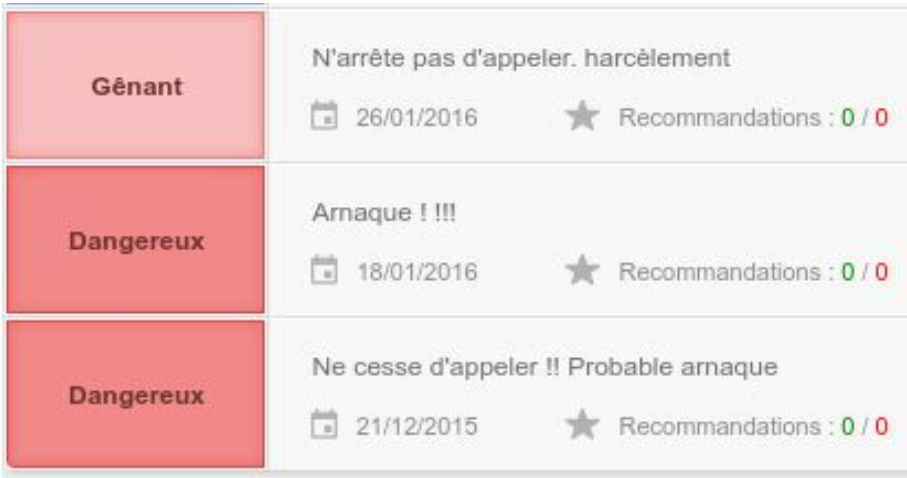



Centrale d'appel qui appelle toute la journée tous les jours sauf le dimanche

# MNO

## Classification

- Méthodologie de validation :
  - Apprentissage sur 10% des xdr sur 1 semaine
  - 10 numéros étiquetés comme “fraudeurs” (sur-échantillonnés)
  - Les autres sont étiquetés comme normaux
  - données traitées par batchs de 100k
  - Tests sur un autre échantillon de 10%
  - Résultats (sensibilité / spécificité)



# MNO

## Classification

- Algorithmes de classification (avec sparkling water : Spark + H2O.ai)
  - Réseau de neurones
    - 7 hidden layers; fully connected; 100 neurones chacun
  - Converge vers
    - Sensitivity : TP/(TP+FN) = 76%
    - Specificity : TN/(TN+FP) = 99.99%

  - Méthode ensembliste
    - GBM / GLM / Random Forest / Deep Learning (200x200)
  - Converge vers
    - Sensitivity : TP/(TP+FN) = 87%
    - Specificity : TN/(TN+FP) = 99.99%

  - Lors des premières itérations (profils vierges) le Deep Learning seul arrive à identifier certains profils fraudeurs tandis que l'approche ensembliste non

# MNO

## Perspectives

- Problème :
  - Etiquetage manuel
  - Comment rendre l'algorithme plus adaptatif ?
- Solution : feedback loop via le 33700


| calling_number | dialled_number | c_number | start_time | stop_time | xdr_type_txt | root_failure_txt | msisdn | imsi |
|---|---|---|---|---|---|---|---|---|
| 336 | 33700 | | 24 Oct. 2016 09:05:06:663 | 24 Oct. 2016 09:05:06:717 | Mo forward sm | Success | 3368 | |
| 3368 | 33700 | | 24 Oct. 2016 09:05:06:652 | 24 Oct. 2016 09:05:06:675 | Mo forward sm | Success | 3368 | |
| | | | 24 Oct. 2016 09:05:04:305 | 24 Oct. 2016 09:05:04:420 | Send authentification info | Success | | |
| 332 | 3365 | 254 | 24 Oct. 2016 09:03:49:833 | 24 Oct. 2016 09:04:21:047 | SIP Invite | Success | | |
| | | 336 | 24 Oct. 2016 09:03:49:749 | 24 Oct. 2016 09:03:49:754 | Provide roaming number | Success | | |
| | 3360 | 3365 | 24 Oct. 2016 09:03:49:629 | 24 Oct. 2016 09:03:49:819 | Send routing info | Success | 336 | |
| | | | 24 Oct. 2016 09:03:09:812 | 24 Oct. 2016 09:03:09:997 | Send authentification info | Success | | |

- Via quelques requêtes SQL on peut directement récupérer la liste des numéros qui ont entraîné un appel vers le 33700
- Si plusieurs personnes le font et que notre algo le classe comme "fraudeur" alors on le considère comme tel et on re-train le modèle
- Quid du reinforcement learning ?

# Merci pour votre attention.

# Et si ces problématiques vous intéressent : Expandium recrute

For more information, go to www.expandium.com

Like us on Facebook

LinkedIn Follow Us on LinkedIn