

Standard de qualité des données

Base d'Accidents Corporels de la Circulation (BAAC)

Département 17 – Charente-Maritime

Année de référence : 2023

Lourdi Mendes

Master 2 Management des Systèmes d'Information
IAE La Rochelle

Projet de fin de cours – Qualité des données

Décembre 2025

Résumé

Ce document présente le standard de qualité défini pour les données BAAC 2023 du département 17 (Charente-Maritime). Sept règles de validation, réparties en trois dimensions (complétude, validité, cohérence), sont pondérées pour calculer un score global de qualité par fichier. Les résultats montrent une qualité globale excellente (97,29 %), avec un point d'attention sur la complétude du fichier Usagers (79,19 %).

Table des matières

1	Introduction	2
1.1	Contexte	2
1.2	Description des fichiers	2
1.3	Dimensions de la qualité des données	2
1.4	Traitement appliqué	2
1.5	Limites des données	3
1.6	Objectifs du système de validation	3
1.7	Périmètre de l'analyse	3
2	Méthodologie d'analyse	3
2.1	Architecture du système de qualité	3
2.2	Système de notation	4
2.3	Calcul du score global	4
3	Règles de validation implémentées	5
3.1	Règles de complétude	5
3.1.1	Règle C1 : complétude du champ	5
3.1.2	Règle C2 : taux de renseignement réel	5
3.2	Règles de validité	5
3.2.1	Règle V1 : validité du format date	5
3.2.2	Règle V2 : validité des valeurs numériques	6
3.2.3	Règle V3 : conformité aux valeurs autorisées	6
3.3	Règles de cohérence	6
3.3.1	Règle CO1 : unicité des enregistrements	6
3.3.2	Règle CO2 : cohérence temporelle des dates	6
4	Résultats de l'analyse	7
4.1	Vue d'ensemble	7
4.2	Analyse par fichier	7
4.2.1	Fichier Caractéristiques (99,97 %)	7
4.2.2	Fichier Lieux (96,91 %)	7
4.2.3	Fichier Véhicules (98,50 %)	7
4.2.4	Fichier Usagers (93,76 %)	8
4.3	Analyse comparative des dimensions	8
5	Interprétation et recommandations	9
5.1	Synthèse globale	9
5.2	Recommandations par priorité	9
5.2.1	Priorité 1 : améliorer la complétude du fichier Usagers	9
5.2.2	Priorité 2 : optimiser le fichier Lieux	9
5.3	Bilan de qualité	9
6	Méthodologie technique	9
7	Conclusion	10

1 Introduction

1.1 Contexte

La Base d'Accidents Corporels de la Circulation (BAAC) centralise l'ensemble des accidents corporels de la circulation routière en France. Cette étude se concentre sur les données du département 17 (Charente-Maritime) pour l'année 2023. La qualité de ces données est essentielle pour :

- l'analyse statistique des accidents au niveau départemental ;
- la définition des politiques de sécurité routière locales ;
- l'identification des zones à risque en Charente-Maritime ;
- le suivi des indicateurs de mortalité routière.

1.2 Description des fichiers

Les données BAAC 2023 sont composées de quatre fichiers CSV correspondant aux quatre tables de la base des accidents corporels de la circulation routière.

- **caract_2023.csv** : caractéristiques générales de l'accident (date, heure, localisation, conditions atmosphériques, luminosité, type de collision). Chaque ligne représente un accident unique identifié par `Num_Acc`.
- **lieux_2023.csv** : caractéristiques du lieu de l'accident (catégorie de route, régime de circulation, état de la surface, profil en long, infrastructure). Lié à la table caractéristiques par `Num_Acc`.
- **vehicule_2023.csv** : informations sur les véhicules impliqués (catégorie, manœuvre, point de choc, obstacle heurté, motorisation). Chaque véhicule est identifié par `id_vehicule`.
- **usagers_2023.csv** : informations sur les personnes impliquées (gravité des blessures, sexe, année de naissance, catégorie d'utilisateur, équipements de sécurité). Chaque usager est identifié par `id_usager`.

1.3 Dimensions de la qualité des données

L'évaluation de la qualité des données repose sur quatre dimensions fondamentales :

Pour la base BAAC 2023, les principales mesures de qualité sont :

- **Complétude** : taux de cellules renseignées par variable (ex. : coordonnées GPS souvent manquantes) ;
- **Unicité** : vérification de l'unicité des identifiants `Num_Acc`, `id_vehicule` et `id_usager` ;
- **Validité** : contrôle des valeurs dans les plages autorisées (ex. : gravité entre 1 et 4).

Dans le système de score présenté ci-après, l'unicité est intégrée à la dimension *Cohérence* (règle CO1).

1.4 Traitement appliqué

Le nettoyage des données a comporté les opérations suivantes :

Dimension	Définition	Calcul
Complétude	Données renseignées	$\frac{\text{Cellules non nulles}}{\text{Total cellules}} \times 100$
Unicité	Absence de doublons	$\frac{\text{Lignes distinctes}}{\text{Total lignes}} \times 100$
Validité	Respect des formats attendus	$\frac{\text{Valeurs conformes}}{\text{Total valeurs}} \times 100$
Cohérence	Absence de contradictions	Vérification des règles métier

TABLE 1 – Dimensions de la qualité des données

- **Valeurs manquantes** : les cellules vides, les points (.) et les valeurs -1 ont été remplacés par des valeurs nulles, conformément à la documentation BAAC de l’ONISR.
- **Recodage des variables** : les variables codées numériquement (gravité, luminosité, météo, etc.) ont été remplacées par leurs libellés explicites pour faciliter l’analyse.

1.5 Limites des données

- Les accidents non déclarés aux forces de l’ordre ne figurent pas dans la base.
- L’indicateur « blessé hospitalisé » n’est plus comparable aux années antérieures à 2018.
- Les coordonnées GPS ne sont pas renseignées pour tous les accidents.

1.6 Objectifs du système de validation

Le système de validation vise à :

1. détecter les anomalies et incohérences dans les données ;
2. quantifier la qualité globale des fichiers BAAC ;
3. identifier les champs nécessitant une attention particulière ;
4. faciliter le nettoyage et l’amélioration des données.

1.7 Périmètre de l’analyse

Périmètre géographique : département 17 (Charente-Maritime)

Périmètre temporel : année 2023

Fichiers analysés :

2 Méthodologie d’analyse

2.1 Architecture du système de qualité

Le système évalue la qualité selon trois dimensions complémentaires :

Fichier	Contenu
caract_2023.csv	Caractéristiques générales des accidents
lieux_2023.csv	Informations sur les lieux des accidents
vehicule_2023.csv	Données relatives aux véhicules impliqués
usagers_2023.csv	Informations sur les usagers impliqués

TABLE 2 – Fichiers BAAC analysés

Dimension 1 : COMPLÉTUDE Mesure le taux de renseignement des champs. Une donnée complète contient le minimum de valeurs manquantes. Poids : 30 %
Dimension 2 : VALIDITÉ Vérifie que les valeurs respectent les formats attendus et les contraintes métier (plages de valeurs, listes autorisées). Poids : 40 %
Dimension 3 : COHÉRENCE Assure la cohérence logique des données (absence de doublons, cohérence temporelle). Poids : 30 %

2.2 Système de notation

Chaque règle produit un taux de conformité compris entre 0 et 1. Ce taux est ensuite classé selon trois seuils :

Taux	Statut	Signification
$\geq 95 \%$	✓ Excellent	Qualité excellente, pas d'action requise
$[80 \%, 95 \%[$	△ À améliorer	Qualité satisfaisante, amélioration souhaitable
$< 80 \%$	× Critique	Qualité insuffisante, action corrective nécessaire

TABLE 3 – Grille d'évaluation de la qualité

2.3 Calcul du score global

Le score global de qualité est calculé comme une moyenne pondérée des trois dimensions :

$$\text{Score}_{\text{global}} = 0,30 \times \text{Score}_{\text{complétude}} + 0,40 \times \text{Score}_{\text{validité}} + 0,30 \times \text{Score}_{\text{cohérence}} \quad (1)$$

où chaque score dimensionnel est la moyenne arithmétique des taux de ses règles :

$$\text{Score}_{\text{dimension}} = \frac{1}{n} \sum_{i=1}^n \text{Taux}_{\text{règle}_i} \quad (2)$$

3 Règles de validation implémentées

3.1 Règles de complétude

3.1.1 Règle C1 : complétude du champ

Règle C1 : Complétude du champ
Objectif : mesurer le taux de renseignement effectif de chaque colonne. Formule : $\text{Taux}_{\text{complétude}} = \frac{\text{Nb lignes sans valeur manquante}}{\text{Nb total de lignes}} \quad (3)$ Valeurs considérées comme manquantes : — NaN : valeurs nulles natives de Python/Pandas ; — -1 : code BAAC pour « non renseigné » ou « inconnu ». Application : toutes les colonnes du jeu de données.

3.1.2 Règle C2 : taux de renseignement réel

Règle C2 : Taux de renseignement réel
Objectif : quantifier spécifiquement l'utilisation des codes manquants BAAC dans les données brutes. Formule : $\text{Taux}_{\text{renseignement}} = \frac{\text{Nb lignes sans code } -1}{\text{Nb total de lignes}} \quad (4)$ Application : colonnes numériques uniquement.

3.2 Règles de validité

3.2.1 Règle V1 : validité du format date

Règle V1 : Validité du format date
Objectif : vérifier que les champs de type date respectent un format valide et lisible. Formule : $\text{Taux}_{\text{validité}} = \frac{\text{Nb dates valides}}{\text{Nb valeurs non nulles}} \quad (5)$ Application : colonnes de type date.

3.2.2 Règle V2 : validité des valeurs numériques

Règle V2 : Validité des valeurs numériques
Objectif : détecter les valeurs numériques aberrantes hors de la plage attendue. Formule : $\text{Taux}_{\text{conformité}} = \frac{\text{Nb valeurs dans } [\text{min}, \text{max}]}{\text{Nb valeurs non nulles}} \quad (6)$ Intervalle de validation : $[0, \text{max}_{\text{observé}} \times 2]$

3.2.3 Règle V3 : conformité aux valeurs autorisées

Règle V3 : Conformité aux valeurs autorisées
Objectif : vérifier que les variables à modalités définies respectent la nomenclature BAAC. Formule : $\text{Taux}_{\text{conformité}} = \frac{\text{Nb valeurs autorisées}}{\text{Nb valeurs non nulles}} \quad (7)$ Application : champs dotés d'une nomenclature BAAC.

3.3 Règles de cohérence

3.3.1 Règle CO1 : unicité des enregistrements

Règle CO1 : Unicité des enregistrements
Objectif : détecter les lignes entièrement dupliquées dans le jeu de données. Formule : $\text{Taux}_{\text{unicité}} = \frac{\text{Nb lignes uniques}}{\text{Nb total de lignes}} \quad (8)$ Application : ensemble du jeu de données.

3.3.2 Règle CO2 : cohérence temporelle des dates

Règle CO2 : Cohérence temporelle
Objectif : vérifier que les dates sont cohérentes avec le contexte d'analyse (année 2023). Formule : $\text{Taux}_{\text{cohérence}} = \frac{\text{Nb dates en 2023 ou nulles}}{\text{Nb total de lignes}} \quad (9)$ Application : colonnes de type date validées.

4 Résultats de l'analyse

4.1 Vue d'ensemble

L'analyse de qualité des données BAAC 2023 pour le département 17 révèle des scores de qualité globaux supérieurs à 93 % pour l'ensemble des fichiers.

Fichier	Score de qualité	Statut
Caractéristiques	99,97 %	✓ Excellent
Lieux	96,91 %	✓ Excellent
Véhicules	98,50 %	✓ Excellent
Usagers	93,76 %	△ À améliorer
Moyenne globale	97,29 %	✓ Excellent

TABLE 4 – Scores de qualité globaux par fichier

4.2 Analyse par fichier

4.2.1 Fichier Caractéristiques (99,97 %)

Le fichier des caractéristiques présente le meilleur score de qualité.

Dimension	Score
Complétude	99,89 %
Validité	100,00 %
Cohérence	100,00 %
Score global	99,97 %

TABLE 5 – Détail du fichier Caractéristiques

Points forts :

- validité et cohérence parfaites (100 %) ;
- taux de complétude quasi optimal (99,89 %) ;
- aucun doublon détecté ;
- toutes les dates sont cohérentes avec l'année 2023.

4.2.2 Fichier Lieux (96,91 %)

Points forts : validité et cohérence parfaites.

Axe d'amélioration : la complétude (89,69 %) se situe dans la zone « à améliorer » ; certains champs de localisation sont insuffisamment renseignés.

4.2.3 Fichier Véhicules (98,50 %)

Points forts : validité et cohérence parfaites ; complétude proche du seuil d'excellence (94,99 %).

Dimension	Score
Complétude	89,69 %
Validité	100,00 %
Cohérence	100,00 %
Score global	96,91 %

TABLE 6 – Détail du fichier Lieux

Dimension	Score
Complétude	94,99 %
Validité	100,00 %
Cohérence	100,00 %
Score global	98,50 %

TABLE 7 – Détail du fichier Véhicules

4.2.4 Fichier Usagers (93,76 %)

Le fichier des usagers présente le score le plus faible et se classe dans la catégorie « à améliorer ».

Dimension	Score
Complétude	79,19 %
Validité	100,00 %
Cohérence	100,00 %
Score global	93,76 %

TABLE 8 – Détail du fichier Usagers

Points forts : validité et cohérence parfaites (100 %), aucun doublon détecté.

Point d'attention : avec 79,19 %, la complétude passe juste sous le seuil critique de 80 %. Certains champs relatifs aux usagers sont insuffisamment renseignés : c'est la priorité d'amélioration.

4.3 Analyse comparative des dimensions

Observations :

- **Validité** : score parfait (100 %) pour tous les fichiers ;
- **Cohérence** : score parfait (100 %) pour tous les fichiers ;
- **Complétude** : seule dimension variable selon les fichiers (de 79,19 % à 99,89 %).

Dimension	Caractéristiques	Lieux	Véhicules	Usagers
Complétude	99,89 %	89,69 %	94,99 %	79,19 %
Validité	100,00 %	100,00 %	100,00 %	100,00 %
Cohérence	100,00 %	100,00 %	100,00 %	100,00 %

TABLE 9 – Comparaison des dimensions par fichier

5 Interprétation et recommandations

5.1 Synthèse globale

L’analyse de qualité des données BAAC 2023 du département 17 révèle une **qualité globale excellente**, avec un score moyen de **97,29 %**.

Points forts majeurs :

1. **Validité parfaite** : 100 % sur tous les fichiers, aucune valeur aberrante détectée ;
2. **Cohérence parfaite** : 100 % sur tous les fichiers, aucun doublon, toutes les dates cohérentes ;
3. **Structure conforme** : les données respectent strictement les nomenclatures BAAC.

5.2 Recommandations par priorité

5.2.1 Priorité 1 : améliorer la complétude du fichier Usagers

Constat : complétude de 79,19 %, sous le seuil critique de 80 %.

Actions concrètes :

- identifier les champs les moins renseignés ;
- sensibiliser les services de saisie à l’importance de ces champs ;
- mettre en place des contrôles de saisie obligatoire.

Impact attendu : une complétude supérieure à 90 % ferait passer le score du fichier de 93,76 % à plus de 95 % (statut « Excellent »).

5.2.2 Priorité 2 : optimiser le fichier Lieux

Constat : complétude de 89,69 %.

Actions concrètes :

- vérifier le renseignement des informations géographiques ;
- améliorer la complétude des données de localisation.

5.3 Bilan de qualité

6 Méthodologie technique

Outils : Python (Pandas) pour le traitement et le calcul des indicateurs, Power BI pour le tableau de bord de suivi.

Étapes de prétraitement :

1. filtrage des données sur le département 17 ;

Aspect	Évaluation
Exploitabilité	Les données sont exploitables pour des analyses statistiques.
Fiabilité	Aucune anomalie critique de validité (100 %).
Complétude	Bonne dans l'ensemble, à renforcer sur les fichiers Usagers et Lieux.
Cohérence	Parfaite : aucun doublon ni incohérence temporelle.

TABLE 10 – Bilan synthétique de la qualité

2. conversion des codes manquants BAAC (-1) en valeurs nulles ;
3. validation des formats de dates ;
4. détection des doublons.

Jointures effectuées :

- utilisation de la clé Num_Acc pour lier tous les fichiers ;
- ajout de la variable **dep** dans tous les fichiers par jointure ;
- vérification de l'intégrité référentielle.

7 Conclusion

L'analyse de qualité des données BAAC 2023 pour le département 17 (Charente-Maritime) démontre une **qualité globale excellente**, avec un score moyen de **97,29 %**.

Résultats clés :

- **3 fichiers sur 4** présentent un score excellent ($> 95 \%$) ;
- **validité et cohérence parfaites** (100 %) sur l'ensemble des fichiers ;
- **aucune anomalie de validité** détectée ;
- données **exploitables** pour des analyses statistiques et décisionnelles.

Axe d'amélioration principal : renforcer la complétude du fichier Usagers (79,19 % → objectif $> 90 \%$).

Le système de validation développé permet d'objectiver la qualité des données et fournit des indicateurs précis pour piloter cette qualité dans le temps. Les données du département 17 constituent une base solide pour des analyses approfondies sur la sécurité routière en Charente-Maritime.

Projet réalisé par Lourdi Mendes dans le cadre du Master 2 Management des Systèmes d'Information – IAE La Rochelle

« La qualité des données est le fondement de toute analyse fiable et de toute décision éclairée en matière de sécurité routière. »