Introduction au traitement de données – partie 2

Mamadou Mbodj, Septembre 2024
Contact : mamadou.mbodj@ansd.sn

Cours : Traitement de données

(ENSAE 2024)

Classe : ISE 3

SCAN ME

Plan du cours

1. Définition du traitement de données
2. Cycle de vie des données
3. Types de données courants
4. Défis courants et potentielles solutions
5. Exercice Pratique

❖ Définition

Le traitement de données consiste en une série d'étapes permettant de transformer des données brutes en données exploitables. Cette phase est essentielle en statistique, et dans d'autres disciplines, car elle prépare les données pour l'analyse et la modélisation.

Exemple : Après une enquête, il faut obligatoirement appurer les données collectées avant toute analyse statistique avancée.

❖ Objectifs

Les principaux objectifs du traitement de données incluent :

- **Nettoyage :** Éliminer les erreurs et incohérences dans les données.
- **Transformation :** Préparer les données pour les techniques d'analyse (normalisation, standardisation, etc.).
- **Exploration :** Identifier des tendances, comme dans une analyse préliminaire de la distribution du niveau de revenu par région.

- **Collecte de données**
  - Enquêtes, recensement : collecte auprès des ménages ;
  - Données de trafic : applications, capteurs, etc.
  - Autres sources : satellites, téléphonie mobile, etc.

- **Exploration des données : statistiques descriptives**
  - Identification des patterns ou anomalies ;
  - Détection des valeurs aberrantes, des valeurs manquantes, des erreurs d'encodage et autres incohérences

- **Préparation et nettoyage des données**
  - Nettoyage : suppression doublons, correction erreurs de saisie, imputation des valeurs manquantes, correction des valeurs aberrantes, repondération, etc.
  - S'assurer de la cohérence et la qualité des données ;
  - Transformation : Standardisation, normalisation, ou création de nouvelles variables.

3. Types de données courants

Type de données	Sources potentielles	Anomalies fréquentes	Exemples de traitement
Données quantitatives continues	Enquêtes, capteurs, logs de transactions	Valeurs aberrantes (ex. : températures extrêmes), erreurs de saisie, valeurs manquantes	Imputation simple (ex. par la médiane ou moyenne) ou imputation multiple ; détection et gestion des valeurs aberrantes (ex. : remplacer par un percentile, capping)
Données quantitatives discrètes	Comptes d'objets ou d'événements, questionnaires	Comptes erronés, valeurs aberrantes, valeurs nulles	Imputation par valeurs similaires ; arrondi ou transformation en continu si nécessaire pour analyse (ex. : nombre d'enfants)
Données qualitatives nominales	Données démographiques (sexe, état civil), identifiants	Valeurs hors catégorie, fautes de saisie, doublons	Normalisation des catégories (ex. : standardiser sexe : F/H), gestion des doublons et des incohérences
Données qualitatives ordinales	Enquêtes de satisfaction, niveaux d'éducation	Incohérences dans les ordres, valeurs manquantes	Uniformisation des échelles ; imputation par la modalité la plus fréquente ou par régression
Données temporelles	Logs de systèmes, transactions, capteurs	Dates manquantes, incohérences temporelles (ex. : dates futures ou antérieures absurdes)	Normalisation du format (ISO 8601), remplacement par une valeur par défaut ou calcul d'une date relative, traitement des fuseaux horaires

3. Types de données courants

Type de données	Sources potentielles	Anomalies fréquentes	Exemples de traitement
Données géospatiales	GPS, systèmes d'information géographique, adresses	Coordonées erronées, absence de géolocalisation	Imputation par des données géographiques proches ; vérification de la validité (ex. : coordonnées dans une région précise)
Données textuelles	Formulaires, commentaires en ligne	Faute d'orthographe, répétitions, mots indésirables	Nettoyage textuel (stopwords, correction d'orthographe) ; lemmatisation ; etc.
Données catégorielles binaires	Données démographiques (oui/non, vrai/faux), enquêtes	Valeurs manquantes, faux positifs/négatifs, incohérences	Normalisation (ex. : tout en 0 et 1), transformation des incohérences en NA, suppression des valeurs extrêmes
Données images	Surveillance, médias sociaux, analyses d'objets	Mauvaise résolution, images floues, absence de métadonnées	Redimensionnement ; amélioration de la netteté ; ajout de métadonnées de base pour le suivi
Données audio/vidéo	Enregistrements d'interviews, contenus de surveillance	Bruit de fond, silences, désynchronisation	Filtrage du bruit ; segmentation en parties exploitables ; extraction de caractéristiques clés (transcription, fréquence)

❖ **Données quantitatives continues**

- **Défis courants :**
  - **Valeurs aberrantes :** Les valeurs extrêmes peuvent biaiser les résultats statistiques.
  - **Valeurs manquantes :** Des mesures manquantes peuvent limiter la robustesse des analyses.

- **Solutions potentielles**
  - **Détection des valeurs aberrantes :** Utiliser des méthodes graphiques (boxplot et histogrammes) ou des méthodes analytiques comme l'écart interquartile (IQR) ou la z-score pour identifier les valeurs extrêmes.
  - **Imputation des valeurs manquantes :** Appliquer des techniques d'imputation simples (basées sur la médiane, la moyenne, ou les méthodes de régression) ou multiples pour compléter les valeurs manquantes. Par exemple, si une valeur de température manque dans une série temporelle, on peut interpoler en utilisant la moyenne des valeurs précédentes et suivantes.

❖ Données quantitatives discrètes

○ Défis courants

- **Incohérences et erreurs de comptage :** Des erreurs de saisie peuvent surévaluer ou sous-évaluer les comptes.
- **Absence de granularité :** Parfois, une variable discrète n'est pas suffisamment précise (ex. : nombre d'enfants par foyer sans spécifier l'âge).

○ Solutions potentielles

- **Vérification des valeurs aberrantes :** Par exemple, pour le nombre d'enfants, éliminer les valeurs absurdes (ex. : un ménage avec plus de 20 enfants).
- **Recodage et agrégation** : Pour simplifier l'analyse, les données discrètes peuvent être agrégées ou regroupées (ex. : groupes d'âge).

❖ Données qualitatives nominales

○ Défis courants

- **Valeurs hors catégorie** : Des fautes de frappe ou de saisie peuvent entraîner des valeurs incohérentes.
- **Catégories redondantes** : Dans les enquêtes, des catégories proches peuvent se chevaucher.

○ Solutions potentielles

- **Standardisation des catégories** : Harmoniser les valeurs pour qu'elles correspondent aux catégories définies. Par exemple, dans une variable “Sexe”, transformer toutes les valeurs “Femme”, “femme” et “F” en une seule valeur.
- **Gestion des catégories rares** : Grouper les catégories ayant peu de valeurs sous une seule étiquette (ex. : “Autres”). Encore faudrait-il traiter les “Autres”. Nous y reviendrons.

- **Données qualitatives ordinales**
  - **Défis courants**
    - **Incohérences d'échelle** : Différentes échelles peuvent être utilisées pour des variables similaires (ex. : échelle de satisfaction sur 1-5 vs. 1-10).
    - **Absence d'ordre défini** : Parfois, l'ordre logique des valeurs est mal renseigné.
  - **Solutions potentielles**
    - **Uniformisation des échelles** : Recalculer les réponses pour qu'elles utilisent la même échelle (ex. : convertir une échelle de 1 à 10 en une échelle de 1 à 5).
    - **Codage des valeurs ordinales** : Attribuer des scores aux catégories (ex. : "Faible" = 1, "Moyenne" = 2, "Élevée" = 3) afin de faciliter les analyses.

❖ Données catégorielles binaires

○ Défis courants :

- **Incohérences de codage :** Par exemple, utiliser 0/1, Oui/Non, Vrai/Faux dans la même colonne (surtout en cas d'ajout de données).
- **Manque de valeurs explicites :** Parfois, des valeurs NA ou nulles sont confondues avec une valeur binaire.

○ Solutions :

- **Recodage standard :** Convertir toutes les valeurs en une norme choisie (ex. : 1 « un » pour Oui, 0 « zéro » pour Non).
- **Imputation des valeurs manquantes :** Remplacer les valeurs manquantes par une estimation basée sur les autres colonnes de l'échantillon (ex. : dans une enquête, "Non réponse" pour une variable binaire peut être recodée en 0).
- **Ajout d'une catégorie explicite pour NA :** Si les valeurs manquantes ont un sens particulier (ex. : "Non réponse"), en faire une catégorie explicite.

❖ Données temporelles

○ Défis courants

- **Incohérences de format** : Les dates peuvent être dans différents formats (ex. : JJ/MM/AAAA vs. MM/JJ/AAAA).
- **Incohérences de fuseaux horaires** : Des données provenant de différents fuseaux peuvent être mal alignées.
- **Valeurs manquantes ou incorrectes** : Des dates erronées ou hors d'ordre peuvent altérer les analyses.

○ Solutions potentielles

- **Conversion au format ISO 8601** : Uniformiser toutes les dates dans ce format (AAAA-MM-JJ) pour faciliter les calculs et comparaisons.
- **Gestion des fuseaux horaires** : Utiliser les données de localisation pour ajuster les heures, surtout pour les analyses de séries temporelles multinationales.
- **Interpolation** : Dans une série chronologique, utiliser des méthodes d’interpolation linéaire pour combler les dates manquantes, ou exclure les lignes problématiques si nécessaire.

❖ Données textuelles

○ Défis courants

- **Orthographe et incohérences** : La diversité linguistique et les fautes de frappe peuvent affecter l'analyse.
- **Présence de stopwords et de caractères spéciaux** : Ces éléments augmentent la taille des données et diminuent la pertinence des analyses.

○ Solutions potentielles

- **Nettoyage textuel** : Enlever les stopwords, les ponctuations et homogénéiser la casse (ex. : transformer tous les mots en minuscules).
- **Stemming et lemmatisation** : Transformer les mots en leur forme racine ou canonique pour réduire la variabilité (ex. : "chats", "chatte" et "chat" deviennent tous "chat").
- **Correction orthographique automatique** : Utiliser des bibliothèques de traitement du langage pour corriger les fautes d'orthographe dans les réponses textuelles.

5. Exercice Pratique