Qualité d'une base de données géographique : concepts et terminologie

Benoît David, Pascal Fasquel

BI
INSTITUT GÉOGRAPHIQUE NATIONAL
BULLETIN D'INFORMATION DE L'IGN

Qualité d'une base de données géographique : concepts et terminologie

67

BI

BULLETIN D'INFORMATION DE L'IGN

Qualité d'une base de données géographique : concepts et terminologie

67

Qualité d'une base de données géographique : concepts et terminologie

01 Classement IGN-SR : **970004/R-LIV**

02 Type : **LIV**

03 Nom et adresse du directeur de la publication :
Jacques Poulain
IGN - Direction Technique
2-4 avenue Pasteur
F-94165 Saint-Mandé cedex

04 Auteurs :
Benoît David & Pascal Fasquel
IGN - Direction Technique
2-4 avenue Pasteur
F-94165 Saint-Mandé cedex

05 Réalisé à :
IGN
Direction Technique
Service de la Recherche
2-4 avenue Pasteur
F-94165 Saint-Mandé cedex

07 Date d'édition : **1997/2**

14 Numéro du volume : **67**

15 Nom de la collection : **Bulletin d'Information de l'IGN**

Résumé

Ce numéro définit les concepts et la terminologie utilisés pour décrire la qualité d'une base de données géographique vectorielle (BDG). Il a été rédigé pour les BDG produites par l'IGN (principalement la BDTOPO®, la BDCARTO® et GÉOROUTE®), mais peut s'appliquer à d'autres BDG similaires.

Il a été rédigé à l'IGN dans le cadre du projet Qualité des Bases de Données entre 1994 et 1996.

Mot-clé principal : information géographique

Mots-clés : qualité - spécification - terrain nominal - assurance - mesure - contrôle - généalogie - cohérence logique - exhaustivité - actualité - précision de position - précision sémantique - BDTOPO - BDCARTO - GEOROUTE.

SOMMAIRE

Chapitre 1

Introduction

1) Définition

Ce document définit les concepts et la terminologie utilisés pour décrire la qualité d'une base de données géographique vectorielle (BDG). Il a été rédigé pour les BDG produites par l'IGN (principalement la BDTOPO®, la BDCARTO® et GÉOROUTE®), mais peut s'appliquer à d'autres BDG similaires.

2) Domaine d'application

Ce document a été validé par l'IGN et y est appliqué. Il est destiné aussi bien à l'IGN pour :

1. les spécifications de produit des BDG, dont les spécifications de qualité,
2. les plans d'assurance qualité,
3. les contrôles de validation et de recette,
4. la formation,
5. les prestations de conseil en information géographique,
6. la recherche et le développement,

qu'aux utilisateurs des données et des documents sur ces données.

Les concepts définis dans ce document sont en particulier utilisés dans les spécifications de produit des différentes bases de données pour définir les seuils de qualité, et dans les rapports de mesure et de contrôle qualité qui seront réalisés pour la validation et la recette des bases de données.

Ce document ne définit ni les spécifications de qualité des bases de données, ni les spécifications de processus des contrôles qualité.

Ce document ne traite pas des méthodes d'échantillonnage utilisées pour estimer les caractéristiques de la qualité.

3) Définition de la qualité d'une base de données géographique

Les bases de données géographiques sont des produits récents et coûteux car elles sont complexes à produire et à utiliser. Il est important, aussi bien pour les producteurs que pour les utilisateurs, d'évaluer l'aptitude de ces bases de données à satisfaire les besoins des utilisateurs. Cette aptitude correspond à la qualité du produit, définie dans la norme ISO 8402 comme *" l'ensemble des propriétés et caractéristiques d'un produit ou service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites "*.

La qualité définie ci-dessus suppose donc la donnée d'un produit et d'une application de ce produit. Cependant, une base de données géographique est généralement destinée à des utilisateurs multiples et divers. C'est le cas des bases de données produites par l'IGN. Pour évaluer la qualité d'une base de données par rapport à une application, il est nécessaire de procéder en deux étapes : d'une part, d'évaluer l'**adéquation de la spécification de la base à un besoin exprimé ou implicite**, et, d'autre part, d'évaluer la **conformité de la base à sa spécification**. La première étape dépend de l'application considérée et est indépendante des données de la base, la seconde est indépendante des besoins.

Dans ce document, on considère la mesure et le contrôle qualité d'une BDG indépendamment des applications utilisant cette base et la qualité est donc considérée comme la conformité de la base à sa spécification de produit.

L'adéquation de la spécification à l'application est de la responsabilité de l'utilisateur ou de celui qui recommande un produit à un utilisateur pour un besoin donné.

4) Origine

A la date de rédaction du présent document, de nombreux travaux de normalisation traitant du même sujet sont en cours, dont notamment ceux du Comité Européen de Normalisation (CEN) dans le cadre des Comités Technique 287 (cf [CEN/TC287/prEN 287008][1], document soumis à enquête CEN par le TC 287) et 278 (cf. GDF 2.0 [CEN/TC278/N356]) et ceux de l'Organisation Internationale de Normalisation (ISO) dans le cadre du Comité Technique 211 (cf [ISO/TC211/WG3/N13]). La rédaction du présent document tient compte de ces travaux ainsi que des concepts issus du document du CNIG (en référence 8).

5) Références normatives

1) *(ISO 3534-1)* : norme ISO 3534-1 - Statistiques - Vocabulaire et symboles - Partie 1 : Probabilités et termes statistiques généraux - 47 pages (décembre 1993).
2) *(ISO 3534-2)* : norme ISO 3534-2 - Statistiques - Vocabulaire et symboles - Partie 2 : Maîtrise statistique de la qualité - 33 pages (décembre 1993).
3) *(ISO 8402)* : norme ISO 8402 - Management de la qualité et assurance de la qualité - 50 pages (avril 1994).
4) *(NF X 06-044)* : norme NF X 06-044 de l'AFNOR - Application de la statistique - Traitement des résultats de mesure, détermination de l'incertitude associée au résultat final - 31 pages (mai 1984).
5) *(SDTS)* : Spatial Data Transfer Standard (SDTS) - Federal Information Processing Standard (FIPS) 173, USA - 200 pages (July 1992).
6) *(FGDC)* : Content Standards for Digital Geospatial Metadata - Federal Geographic Data Committee (FGDC), USA (June 1994).
7) *(EDIGéO)* : norme EDIGéO (Echange de Données Informatisées dans le domaine de l'information GéOgraphique) - 382 pages (août 1992).
8) *(CNIG)* : document du Groupe de travail "Qualité des Données Géographiques Échangées" du CNIG - 22 pages (septembre 93).

Les définitions énoncées dans la suite de ce document sont parfois issues de normes existantes. Dans ce cas, à la suite du terme **Définition**, annonçant une définition, figure une référence à la norme dont est issue cette définition. Cette référence correspond aux noms en italique de la liste de normes présentées ci-dessus et renvoie à cette liste. Par exemple, l'indication *(CNIG)* fait référence au document en référence normative 8.

[1] Les inscriptions entre crochets renvoient à la bibliographie en fin de document.

Chapitre 2

Modèle de données de référence

Ce document s'applique aux bases de données géographiques vectorielles et le modèle de données de référence choisi est celui de la norme EDIGéO.

La norme EDIGéO considère que l'information géographique se modélise à l'aide d'éléments appelés primitives, objets, attributs et relations.

Les primitives sont de trois types : nœud, arc et face et permettent de décrire la géométrie d'un objet géographique.

Les objets géographiques sont de deux types :

- ♦ les objets géographiques simples qui s'appuient sur une ou plusieurs primitives et qui sont :
  - de type ponctuel s'ils sont constitués d'un ou plusieurs nœuds,
  - de type linéaire s'ils sont constitués d'un ou plusieurs arcs,
  - de type surfacique s'ils sont constitués d'une ou plusieurs faces.
- ♦ les objets géographiques complexes qui sont composés soit d'objets géographiques simples, soit d'objets géographiques complexes, soit d'objets des deux types.

Les attributs sont des caractéristiques ou des propriétés déterminées :

- d'un objet,
- d'une primitive,
- d'une relation.

Les relations qui sont des liens entre objets, entre primitives ou entre objets et primitives, sont de deux types :

- les relations de construction
  Exemple : La relation de composition entre une région et les départements de cette région
- les relations sémantiques
  Exemple : La relation de jumelage entre deux communes

***Note**:* Les objets partageant des caractéristiques communes sont regroupés en ensembles appelés dans la suite de ce document : classe.

Nature des attributs

Dans ce document, certains concepts, telles que la précision sémantique et l'exhaustivité (section 2.3 du chapitre 5), dépendent de la nature des attributs. Il est donc nécessaire de préciser les natures possibles d'un attribut.

• Quantitatif / Qualitatif

On distingue les attributs qualitatifs des attributs quantitatifs selon les valeurs prises par ces attributs.

Un attribut est dit quantitatif lorsque les valeurs possibles de cet attribut peuvent être évaluées numériquement en utilisant une unité de mesure bien définie servant de référence et permettant de faire des comparaisons et des opérations algébriques (addition, soustraction, moyenne, écart-type...) sur les valeurs.

Un attribut est dit qualitatif si on ne peut évaluer numériquement, en utilisant une unité de mesure bien définie servant de référence, la différence entre deux valeurs possibles de cet attribut ou si cette évaluation n'a pas de signification propre. Un attribut qualitatif peut être nominal (s'il n'existe pas d'ordre entre les valeurs possibles de l'attribut) ou ordinal (si un tel ordre existe).

• Énuméré / Non énuméré

On dit qu'un attribut est énuméré si l'ensemble de ses valeurs est un ensemble fini.

Exemples :

1. l'attribut classement administratif d'un tronçon de route est qualitatif énuméré.
2. l'attribut nom de la commune est qualitatif non énuméré.
3. l'attribut nombre de voies d'un tronçon de route est quantitatif énuméré dans le schéma conceptuel de données (ou SCD) de la BDCARTO®.
4. l'attribut altitude des courbes de niveau est quantitatif non énuméré dans le SCD de la BDTOPO®.

Note *:* L'attribution rigoureuse d'une nature à un attribut sera, dans la suite de ce document, nuancée en fonction des traitements qui doivent être appliqués à cet attribut. La nature quantitative énumérée d'un attribut, notamment, ne sera plus utilisée dans la suite de ce document. De tels attributs seront traités, suivant le nombre de valeurs possibles, comme des attributs qualitatifs énumérés ou des attributs quantitatifs non énumérés.

Exemples :

1. le nombre de voies d'une route pourra être traité comme un attribut qualitatif énuméré (effectuer des calculs statistiques, de moyenne et d'écarts types sur cet attribut ne présente pas un intérêt colossal).
2. l'attribut population d'une commune, si l'on suppose que ce nombre est un entier compris entre 0 et 5 millions, est de nature quantitative énumérée. Pour cet attribut, des calculs statistiques peuvent être pertinents vu le nombre important de valeurs que peut prendre cet attribut. Il sera traité comme un attribut quantitatif non énuméré.

• La classification, suivant la nature des attributs, adoptée dans ce document est la suivante :

<table>
<tr><td rowspan="2">qualitatif</td><td>non énuméré
Exemple : nom d'une commune</td></tr>
<tr><td>énuméré
Exemple : niveau hiérarchique d'une entité administrative</td></tr>
<tr><td>quantitatif
Exemple :
température du sol</td><td></td></tr>
</table>

Notes :

1) On rappelle que, sauf si le schéma conceptuel de données l'interdit, un attribut peut ne pas avoir de valeur. Dans ce cas, on dira que la valeur de l'attribut n'est pas déterminée.

La signification de cette absence de valeur doit être définie dans la spécification du produit. On distingue trois cas différents dans lesquels la valeur de l'attribut n'est pas déterminée :

- sans objet : l'absence de valeur est expliquée par la spécification de produit et constitue un cas particulier dans lequel l'attribut ne correspond à aucune valeur. Par exemple, l'attribut "nombre de voies" d'un tronçon de route peut être sans objet dans le cas d'un tronçon de route multi-chaussées si le nombre de voies est défini pour chacune des chaussées par d'autres attributs.

- n'existe pas : l'absence de valeur correspond à une absence d'information à représenter. Par exemple, certains itinéraires portent un toponyme alors que d'autres n'en portent pas.

- inconnu : il existe une valeur à représenter mais cette valeur n'est pas enregistrée dans la BDG, pour une raison quelconque.

Les deux premiers cas devraient normalement correspondre à des valeurs particulières pour l'attribut alors que le dernier cas est spécifique car, il correspond à un défaut d'information et donc à une non-conformité. Il serait donc souhaitable que l'absence de valeur corresponde toujours au dernier cas.

2) Pour alléger certains tableaux, la notation suivante sera adoptée : **Classe.Attribut**. Ceci désigne un attribut et la classe sur laquelle s'applique cet attribut.

Exemple : **Monument.Toponyme** indique l'attribut "Toponyme" de la classe "Monument".

Chapitre 3

Concepts généraux pour la définition de la qualité d'une B.D. géographique

1) Jeu de données

Définition : Unité de volume des données pour laquelle est estimée la qualité.

Notes :

1) Ceci ne signifie pas que toutes les caractéristiques exprimant la qualité (cf chapitre 5) sont mesurées sur tout le jeu de données. La quantité de données sur laquelle est, en pratique, mesurée une de ces caractéristiques est un échantillon (cf définition 2). La qualité du jeu de données est alors une extrapolation des caractéristiques mesurées sur les échantillons.

2) Une base de données n'est pas toujours homogène en qualité. Certaines parties d'une base peuvent avoir un niveau de qualité satisfaisant et d'autres peuvent être non conformes à la spécification de produit. Ainsi, afin d'éviter les inconvénients de cette hétérogénéité, il peut être préférable d'estimer séparément les qualités de différents jeux de données appartenant à la base plutôt que d'avoir une valeur moyenne de la qualité sur toute la base.

Exemple : L'ensemble des données d'un département de GÉOROUTE® peut constituer un jeu de données. La qualité est alors estimée pour chaque département, mais pas globalement sur la France.

3) Un jeu de données est constitué d'une ou plusieurs populations (au sens statistique) sur lesquelles sont estimées les caractéristiques de la qualité (cf. chapitre 5).

Exemple: Si le jeu de données est l'ensemble des données d'un département de GÉOROUTE®, les tronçons de route de ce département peuvent être une population de ce jeu de données.

2) Échantillon

Définition *(ISO 3534-1):* Une ou plusieurs unités d'échantillonnage prélevées dans une population et destinées à fournir des informations sur cette population.

Exemple : Si le jeu de données est un département et si l'on s'intéresse à la population des tronçons de route, un échantillon peut être l'ensemble des tronçons de route inclus dans une ou plusieurs communes de ce département.

Notes :

1) Par mesure d'économie, entre autres, la mesure ou le contrôle de qualité sont généralement effectués sur des échantillons différents de la population complète. Les résultats obtenus sur un échantillon sont ensuite extrapolés à la population, ce qui comporte un risque d'erreur d'évaluation.

2) Dans le cadre de la qualité d'une base de données géographique, un échantillon est un sous ensemble du jeu de données sur lequel sont effectuées des mesures ou des contrôles de qualité.

3) A chaque population et à chaque caractéristique exprimant la qualité (cf chapitre 5) que l'on veut mesurer, peut correspondre un échantillon. Pour certaines mesures, un échantillon sera toute la population; pour d'autres, ce sera une partie de cette population.

4) Le choix d'échantillons est une étape primordiale dans l'appréciation de la qualité d'un jeu de données. La définition de méthodes d'échantillonnage n'est pas développée dans ce document.

3) Spécification

Définition *(ISO 8402) :* Document qui prescrit les exigences auxquelles le produit ou le service doit se conformer.

Note *:* Dans le reste de ce document, le terme "spécification" est utilisé pour spécification de produit et non spécification de processus ou de saisie. La spécification de produit inclut le schéma conceptuel des données.

Exemple : La spécification de la BDTOPO® indique que les routes sont décrites par leur axe.

4) Univers

Définition *:* Ce que l'on cherche à représenter de façon plus ou moins détaillée.

Note *:* L'appellation "univers" est préférable à l'appellation "terrain réel" car cette dernière évoque immanquablement le monde physique. Or ce que l'on désire représenter, le point de départ n'est pas toujours le monde physique. Cela peut être, par exemple, un graphe de circulation comme dans le cas de GÉOROUTE®.

5) Terrain nominal

Définition *:* Image de l'univers, à une date donnée, à travers le filtre défini par la spécification de produit.

Exemple : Soit l'univers représenté sur la figure 1a et le terrain nominal associé à cet univers et à la spécification (simplifiée) de GÉOROUTE®, représenté sur la figure 1b.

Figure 1a : une partie de l'univers

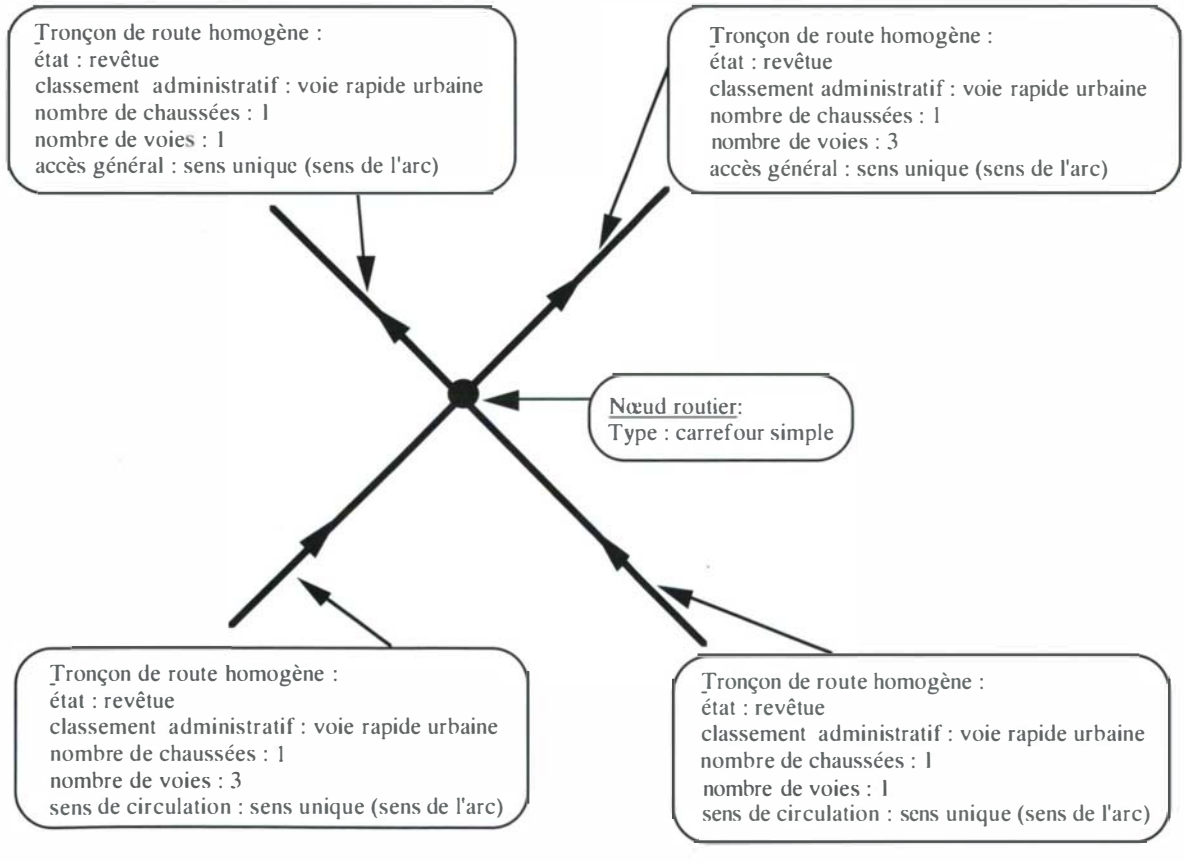

Figure 1b : le terrain nominal correspondant

Notes :

1) On aurait pu employer le terme "univers nominal" au lieu de terrain nominal car le mot "terrain" évoque le monde physique. Cependant, "terrain nominal", couramment employé, semble s'être débarrassé de cette connotation physique et l'introduction d'un nouveau terme pourrait, au contraire, semer une certaine confusion. Dans ce document, terrain nominal sera utilisé.

2) Pour une spécification donnée, il n'y a pas unicité du terrain nominal en raison de l'imprécision de cette spécification (causée par des défauts de spécification ou par l'impossibilité d'y prévoir tous les cas particuliers de l'univers, cf la définition 15 de ce chapitre). En particulier, si la spécification prévoit que les informations possèdent une ancienneté maximale avant mise à jour, alors cette dernière introduit de l'incertitude sur le terrain nominal (cf chapitre 5).

3) **La qualité d'un jeu de données est définie par l'écart entre ce jeu de données et le terrain nominal.**

6) Mesure de la qualité

Définition *:* Ensemble des actions de mesures, d'examens, d'essais, de calibrage d'une ou plusieurs caractéristiques d'un jeu de données qui permet d'évaluer l'écart entre le jeu de données et le terrain nominal.

Note : Les concepts et la terminologie définis dans ce document sont principalement orientés vers la mesure de la qualité d'une BDG.

7) Contrôle qualité

Définition *(ISO 8402) :* Ensemble des actions de mesures, d'examens, d'essais, de calibrage d'une ou plusieurs caractéristiques d'un produit et de comparaisons aux exigences spécifiées en vue d'établir leur conformité.

Notes :

1) La principale différence entre "contrôle qualité" et "mesure de la qualité" tient à l'existence d'une spécification de qualité et d'une procédure de rejet des jeux de données non conformes, c'est-à-dire aux "exigences spécifiées en vue d'établir la conformité du produit". Si ces exigences ne sont pas définies, le contrôle est impossible et seule la mesure peut être effectuée.

2) Les exigences spécifiées en vue d'établir la conformité du produit doivent être exprimées à l'aide des paramètres de qualité décrits au chapitre 5.

8) Assurance qualité

Définition *(ISO 8402):* Ensemble des actions préétablies et systématiques réalisées au fur et à mesure de la production et qui seront nécessaires pour donner la confiance appropriée en ce qu'un produit satisfera aux exigences données relatives à la qualité.

9) Source de saisie

Définition *:* Source à partir de laquelle sont saisies les données qui vont constituer le jeu de données.

Exemples :

1. Photographie aérienne.
2. Carte existante.
3. Univers lorque l'opérateur va sur le terrain pour recueillir les données.
4. Autre base de données, géographique ou non.

Note *:* Lorsque les informations ne sont pas toutes disponibles sur la même source, plusieurs sources de saisie doivent être utilisées.

Exemple : Les sources de saisie de la BDTOPO® sont les photographies aériennes (exploitées dans la phase de restitution photogrammétrique) complétées par l'univers (phase dite de "complétement" sur le terrain). S'y ajoutent le cadastre, des informations administratives de l'INSEE...

10) Données de contrôle

Définition *:* Ensemble de données qui permet de mesurer ou de contrôler la qualité d'un jeu de données par comparaison avec un échantillon de ce jeu de données.

Notes *:*

1) Le terrain nominal est une entité abstraite commode à présenter de façon théorique. Cependant, le terrain nominal n'étant pas directement accessible, la mesure pratique de chaque caractéristique de la qualité ne peut se faire que par rapport à des données de contrôle. Les données de contrôle utilisées pour mesurer une certaine caractéristique de la qualité forment donc une estimation (estimation partielle) "proche" du terrain nominal (c'est-à-dire dont on sait qu'elle représente une partie du terrain nominal de façon plus fidèle que le jeu de données produit).

Exemple : Pour l'appréciation de certains paramètres qualité du réseau routier de la BDCARTO®, le réseau routier de la BDTOPO® peut être utilisé comme données de contrôle. Pour certains paramètres du réseau ferré, des données de contrôle peuvent être des données de la SNCF.

2) La qualité d'un jeu de données est donc estimée par l'écart entre ce jeu de données et des données de contrôle. Ceci suppose donc de pouvoir faire correspondre les éléments du jeu de données avec leurs homologues dans les données de contrôle choisies.

3) Les données de contrôle sont parfois appellées “la référence”.

4) Les données de contrôle doivent être utilisées en parallèles avec la spécification, notamment lors de l'appariement (cf. chapitre 4). En effet, par exemple, il est possible qu'un objet ponctuel P du jeu de données dont l'homologue est ponctuel dans le terrain nominal, ait un homologue surfacique dans la référence. Il convient, alors de combiner la référence et la spécification pour choisir un point de cette surface comme homologue de P. Les termes " données de contrôle " et " référence " sont, dans la suite de ce document, utilisés dans ce sens étendu de combinaison avec la spécification.

11) Source de contrôle

Définition *:* Source à partir de laquelle sont saisies les données de contrôle.

Notes *:*

1) Les données de contrôle sont généralement issues de plusieurs sources de contrôle, chacune couvrant une partie (que ce soit une partie spatiale ou sémantique) du terrain nominal.

Exemple : L'univers est une source de contrôle pour le contrôle géométrique des feuilles BDTOPO®.

2) Le choix de sources de contrôle est capital dans la procédure de mesure ou de contrôle de la qualité d'un jeu de données.

La figure 2 illustre les notions décrites précédemment.

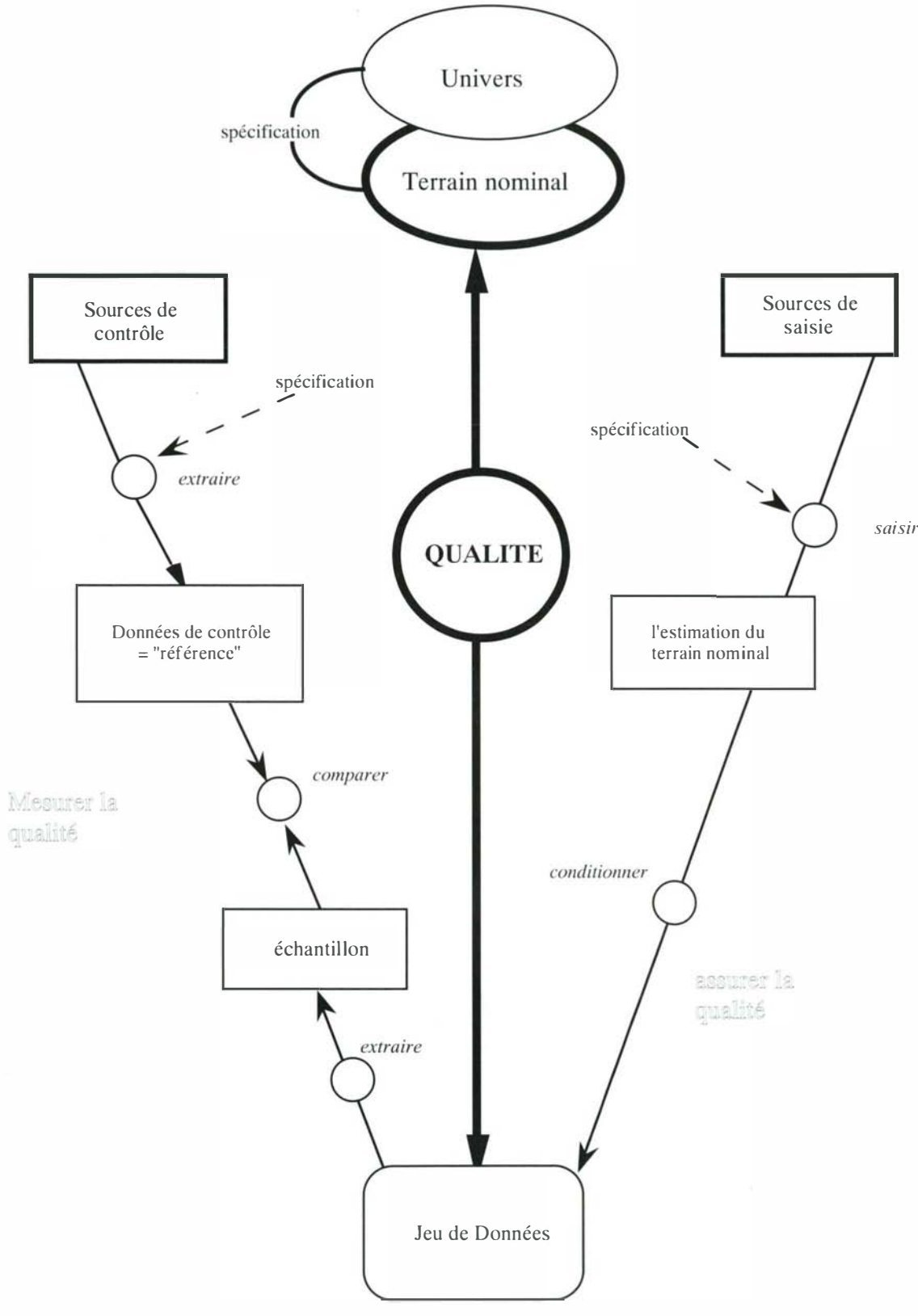

Figure 2: la qualité d'un jeu de données

12) Erreur de mesure

Définition *(NF X 06-044):* Résultat d'un mesurage moins valeur vraie de la grandeur mesurée.

Note *:* Le terme mesurage signifie : "ensemble d'opérations ayant pour but de déterminer la valeur d'une grandeur" (NF X 06-044).

Les trois définitions suivantes décrivent les différentes origines d'erreurs que l'on peut rencontrer dans la représentation d'une information géographique.

13) Faute (ou erreur parasite)

Définition *(NF X 06-044) :* Erreur souvent grossière qui résulte d'une exécution incorrecte du mesurage.

Exemples :

1. Un gros oubli lors d'une saisie.
2. Un bogue dans un programme.
3. Si l'on imagine des tirs sur une cible, une faute serait le résultat d'un tir complétement en dehors de la cible alors que de nombreux autres tirs étaient dans la cible (cf figure 3).

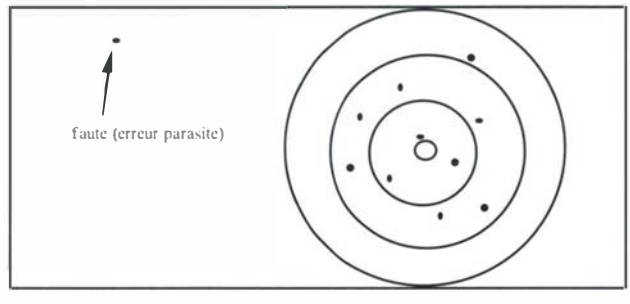

Figure 3 : faute

14) Incertitude d'une mesure de saisie

Hormis les fautes, les mesures restent imparfaites car diverses erreurs s'immiscent dans les mesures. Ces erreurs sont de deux types : les erreurs aléatoires et les erreurs systématiques.

Définition *(NF X 06-044):*

1) L'erreur aléatoire est la composante de l'erreur qui, lors de plusieurs mesurages[1] du même mesurande[2], varie de façon imprévisible.

2) L'erreur systématique est la composante de l'erreur qui, lors de plusieurs mesurages du même mesurande, reste constante ou varie d'une façon prévisible.

Exemples :

1. L'erreur due aux frottements de l'index d'un appareil de mesure est une erreur aléatoire.
2. Une erreur de graduation d'un appareil est une erreur systématique

[1] cf la note de la définition 12

[2] cf note 3 de la définition 14

Notes *:*

1) Les erreurs systématiques et leurs causes peuvent être connues (on applique alors une correction aux résultats), ou inconnues.

2) Les erreurs aléatoires peuvent être modélisées statistiquement.

3) Le terme mesurande employé dans les définitions précédentes signifie grandeur soumise à un mesurage (NF X 06-044).

4) Dans la suite de ce document, l'erreur systématique est chiffrée par un **biais**, défini ci-dessous.

Définition *:* Différence entre l'espérance d'une mesure (ou d'une estimation) d'une grandeur et la valeur nominale de cette grandeur (cf définition 16)

15) Incertitude du terrain nominal

Définition *:* Incertitude introduite par le processus permettant de passer de l'univers au terrain nominal. Elle s'exprime par l'écart qui peut exister entre deux occurrences possibles du terrain nominal pour une même spécification de produit et un même univers.

Notes *:*

1) Le terrain nominal n'est pas unique pour une spécification de produit donnée car, aussi précise soit-elle, la spécification ne peut pas être parfaite. Elle laisse toujours lors de la collecte des données une place à l'interprétation humaine donc à diverses possibilités de représentation des données.

Exemples :

1. Si la spécification précise de représenter les forêts par des zones, la définition de la limite de ces zones n'est pas forcément très précise.
2. Quel est le point final d'un chemin "se perdant" dans une forêt ?
3. Si la spécification définit une ancienneté maximale pour une information, on introduit de l'incertitude sur le terrain nominal (voir la définition de l'actualité dans le chapitre 5).

2) Un processus de mesure de cette incertitude du terrain nominal pourrait être envisagé, par exemple en réalisant plusieurs saisies d'une même zone par différents opérateurs utilisant la même spécification de produit et en mesurant les écarts entre ces diverses saisies.

3) Il convient donc de distinguer les véritables erreurs, c'est-à-dire les manifestations de différences entre ce qui est produit (jeu de données produit) et ce qui est convenu (terrain nominal), de l'incertitude du terrain nominal et donc de ne pas comptabiliser en erreur ce qui relève de cette incertitude.

16) Exactitude

Définition *:* Etroitesse de l'accord entre une mesure ou une estimation d'une grandeur et la valeur nominale de cette grandeur.

Notes *:*

1) Le terme valeur nominale signifie : valeur qui sert de référence pour une comparaison et qui résulte, soit d'une valeur théorique ou établie, fondée sur des principes scientifiques (valeur vraie), soit d'une valeur expérimentale adoptée comme vraie.

2) L'exactitude est chiffrée par une erreur moyenne quadratique.

3) Il convient de ne pas confondre exactitude et précision (cf définition 17).

17) Précision

Définition : Etroitesse de l'accord entre une mesure ou une estimation et l'espérance de cette mesure ou de cette estimation.

Notes :

1) La précision dépend uniquement de la distribution des erreurs aléatoires et n'a aucune relation avec la valeur nominale (cf note 1 de la définition 16).

2) La précision est chiffrée par un écart-type.

3) Il ne faut pas confondre dans ce document le terme précision tel qu'il est défini ci-dessus, qui s'applique à une mesure ou à une estimation et qui est soit un nombre, soit un intervalle et les termes précision sémantique et précision géométrique (cf chapitre 5) qui sont des paramètres de qualité composés d'un ensemble d'indicateurs.

4) **Précision** et **exactitude** sont souvent synonymes dans le langage courant. Pourtant, en statistiques et dans tout processus de mesure, il convient de les distinguer. En effet, précision et exactitude ne sont identiques qu'en l'absence de biais.
La figure 4, décrivant les résultats d'une série de tirs sur une cible, illustrent cette différence.

L'exactitude mesure les fluctuations des valeurs d'une série de mesures autour de la valeur nominale (ou valeur "vraie"); la précision mesure les fluctuations de cette série de mesures autour de son espérance; le biais mesure l'écart entre l'espérance de la série de mesures et la valeur nominale.

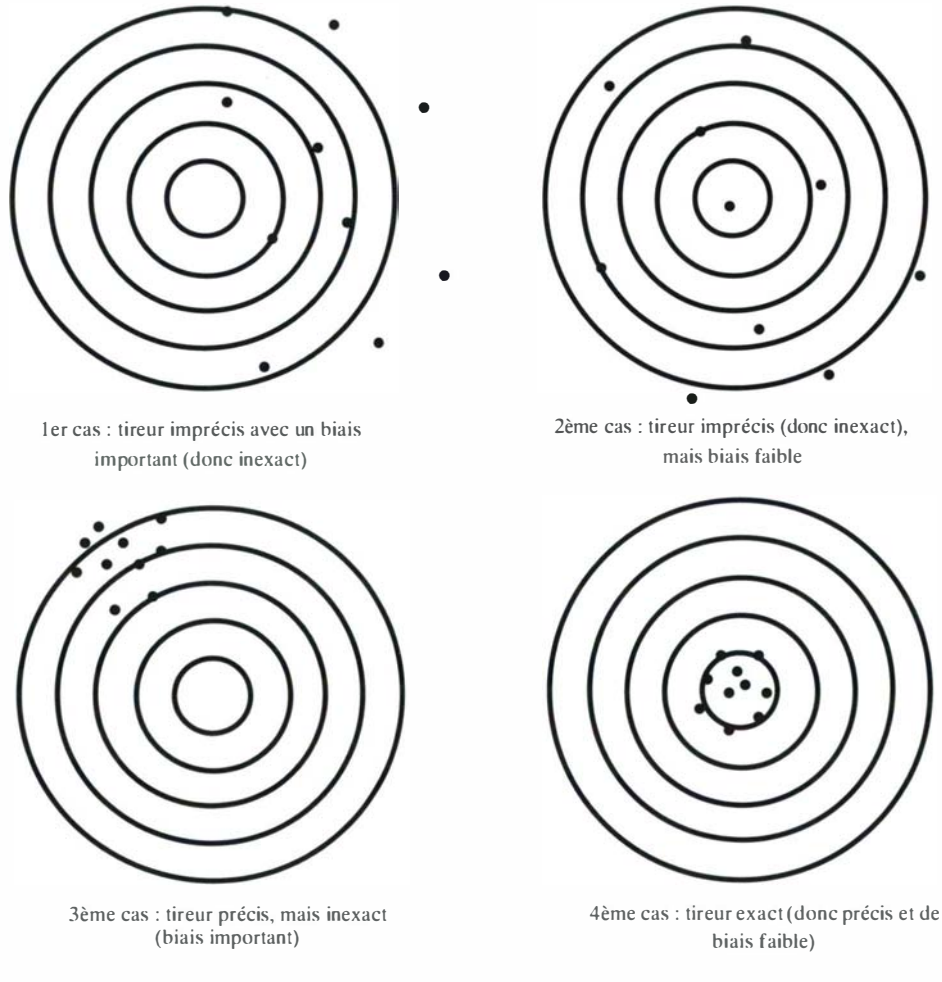

Figure 4 : précision, exactitude et biais

Chapitre 4

Préliminaires : étape et outil nécessaires à l'expression de la qualité d'une B.D. géographique

1) Appariement des objets

Les définitions exposées dans la suite de ce document supposent la possibilité d'associer les objets du jeu de données aux objets des données de contrôle par des critères principalement géométriques. Cette association constitue l'appariement des objets du jeu de données aux objets des données de contrôle. Si cette opération est immédiate dans le cas de données raster (où on apparie directement les pixels de même coordonnées), elle n'est pas évidente dans le cas du mode vecteur. Or, l'appariement des objets influe sur la répartition des erreurs (cf figures 5 à 7).

Soit le terrain nominal représenté sur la figure 5.

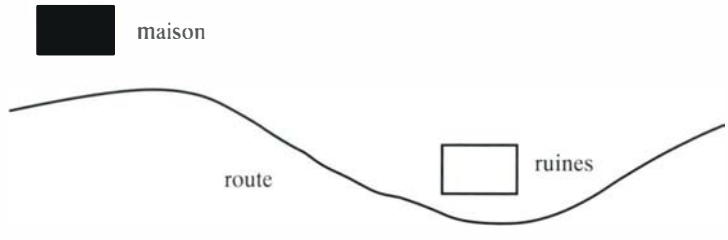

Figure 5 : terrain nominal

Si les données sont représentées comme sur la figure 6, la maison de la figure 5 et celle de la figure 6 ne sont pas appariées. Les erreurs, au nombre de deux sont une erreur de précision sémantique (maison au lieu de ruines) et une erreur d'exhaustivité (maison manquante) (pour plus de précisions sur les notions de précision sémantique et d'exhaustivité, se reporter à la section 2.3 du chapitre 5).

Figure 6 : jeu de données 1

Si les données sont représentées comme sur la figure 7, il y a toujours deux erreurs, mais elles sont de nature différente de celles du cas précédent. En effet, il y a une erreur d'exhaustivité (ruines manquantes) et une erreur géométrique (la maison de la figure 7, cette fois appariée à celle de la figure 5, est décalée de la distance D) (pour plus de précision sur les erreurs géométriques, se reporter à la section 2.2 du chapitre 5)

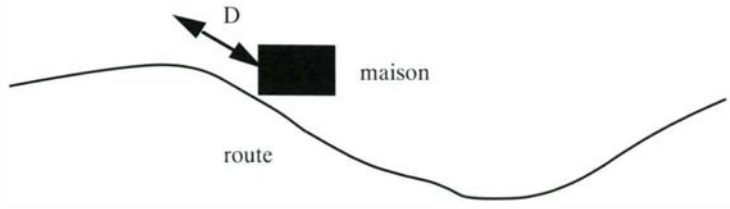

Figure 7 : jeu de données 2

Cependant, il existe entre ces deux cas une certaine similarité. En effet, la répartition des erreurs dépend de la distance D. Si D est "petite", les maisons du terrain nominal et du jeu de données sont appariées, l'erreur est géométrique. Si D est "grande", les maisons ne sont pas appariées, l'erreur est une erreur d'exhaustivité.
La valeur maximale de D permettant de trancher doit être définie (elle peut, par exemple, être identique à 10 fois la précision (statistique) géométrique ponctuelle planimétrique).

Il est, par conséquent, indispensable, lorsqu'on mesure la qualité, de décrire précisément la méthode d'appariement utilisée.

2) Précisions d'estimation

Certains paramètres de qualité décrits dans le chapitre 5 sont des estimations calculées sur des échantillons. Il convient donc d'accompagner ces estimations d'une évaluation de la fiabilité de ces estimations, qui peut être chiffrée par un **intervalle de confiance** associé à **un niveau de confiance** ou par **la précision de l'estimation**.

2.1) Pour l'exactitude (précision géométrique)

L'exactitude est estimée par les formules suivantes :

Si on dispose d'un échantillon de n points du terrain nominal et des points homologues dans le jeu de données et si $X_{i,TN}$, $Y_{i,TN}$, $Z_{i,TN}$, $X_{i,JD}$, $Y_{i,JD}$, $Z_{i,JD}$ représentent respectivement les coordonnées X, Y, Z d'un point de l'échantillon du terrain nominal et les coordonnées de son homologue dans le jeu de données.

$$EMQ_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,TN} - X_{i,JD}\right)^2}$$

$$EMQ_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_{i,TN} - Y_{i,JD}\right)^2}$$

$$EMQ_Z = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z_{i,TN} - Z_{i,JD}\right)^2}$$

$$EMQ_{XY} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,TN} - X_{i,JD}\right)^2 + \left(Y_{i,TN} - Y_{i,JD}\right)^2}$$

Les précisions d'estimation correspondantes sont respectivement estimées par :

$$\sigma_{EMQ_x} = \frac{EMQ_X}{\sqrt{2n}}$$

$$\sigma_{EMQ_y} = \frac{EMQ_Y}{\sqrt{2n}}$$

$$\sigma_{EMQ_z} = \frac{EMQ_Z}{\sqrt{2n}}$$

$$\sigma_{EMQ_{xy}} = \frac{EMQ_{XY}}{2\sqrt{n}}$$

2.2) Pour une proportion

Pour un échantillon de taille N issue d'une population, si T est une proportion (T est compris entre 0 et 1) estimée sur cet échantillon alors la demi longueur de l'intervalle de confiance au niveau de confiance de 1-α, associée à l'estimation est donnée par la formule :

$$p = k_\alpha \sqrt{\frac{T(1-T)}{N}}$$

où k_α est tel que, si X représente une variable aléatoire de loi normale centrée réduite :

$$\Pr(|X| \succ k_\alpha) = 1 - \alpha$$

et $\sqrt{\frac{T(1-T)}{N}}$ est la précision d'estimation du taux T.

Les précisions d'estimations des taux formulées ci-dessus sont approximatives et ne sont valables que si les effectifs auxquels elles se rapportent sont supérieurs à 30 et si le taux de sondage est inférieur à 10%.

Le nombre 30 provient de la limite de validité de l'approximation asymptotique d'une loi binômiale par une loi normale. De plus l'utilisation de la loi binômiale n'est valable que dans le cas d'une population infinie ou lorsque les tirages des éléments de l'échantillon sont effectués indépendamment (tirage avec remise). Lorsque la population est finie et que les tirages ne sont pas indépendants (tirage sans remise) la loi hypergéométrique doit être utilisée ce qui complique sensiblement les calculs d'intervalle de confiance. Cependant, lorsque le taux de sondage est inférieur à 10%, on peut substituer la loi binômiale à la loi hypergéométrique.

Notes :

1) Ces formules seront utilisées dans les matrices de confusion ou d'absence (cf 2.3 du chapitre 5). Si les conditions ci-dessus (taux de sondage inférieur à 10% et effectif supérieur à 30) ne sont pas respectées, la demi longueur de l'intervalle de confiance sera remplacée dans ces matrices par un point d'interrogation.

2) Lorsque la taille N de l'échantillon est faible (quand même supérieure à 2) ou que le taux de sondage f est quelconque (en particulier supérieur à 10%) la précision de l'estimation du taux T peut être estimée par la formule :

$$\sqrt{(1-f)\frac{T(1-T)}{N-1}}$$

Néanmoins, ceci ne permet pas d'en déduire un intervalle de confiance simple pour T.

Chapitre 5

Expressions de la qualité d'une B.D. géographique

Dans ce chapitre sont définis les termes exprimant la qualité d'une base de données géographique vectorielle. Ces termes sont répartis en trois parties, 1) la généalogie, 2) les paramètres de qualité et 3) l'actualité. La généalogie et les paramètres de qualité correspondent respectivement à des expressions qualitatives et quantitatives de la qualité à un instant donné, l'actualité traduit le décalage entre le jeu de données produit et le terrain nominal à un instant ultérieur.

1) Généalogie (expression qualitative)

Définition *:* Description de l'histoire des données. Elle fournit une description des sources et méthodes d'acquisition des données, des opérations appliquées sur ces données et des organismes responsables.

Notes:

1) Les opérations appliquées sur les données sont des transformations (changement de système de référence, passage du mode vecteur au mode maillé...) ou des mises à jour.

2) L'information de généalogie est essentiellement qualitative. Elle fournit une indication sur la qualité globale d'un jeu de données.

3) Certaines informations de généalogie peuvent être confidentielles car elles décrivent le processus de production de la base de données.

Informations à fournir

Les informations proposées ci-après ne sont que des indications, c'est-à-dire qu'elles ne sont ni exhaustives ni obligatoires. Le niveau de description des informations de généalogie ne peut être défini dans ce document, il dépend essentiellement du produit et doit être défini dans la spécification de produit.

- **pour une source :**
  - **description de la source**
  - **date de création**
  - **organisme responsable**
  - **référence à une spécification de produit**

Exemple : Photographies aériennes prises le 20 mai 1993 par le Service des Activités Aériennes de l'IGN. Description de la mission : altitude de vol XXX mètres, focale utilisée XXX millimètres, type d'avion XXX. Application de la spécification 251.6 du SAA du 24 septembre 1984.

♦ **pour les données :**

- **zone couverte**
- **classes concernées**
- **objectif de la création**
- **date de saisie**
- **méthode de saisie**
- **organisme responsable**
- **date de validation**
- **les sources dont sont issues les données**

Exemple : Les données ont été produites par l'IGN afin de réaliser une cartographie au 1/10000 des risques d'incendies en Corse. A partir des photographies aériennes décrites précédemment, les données qui ont été restituées en utilisant les stations CASIMIR, le 2 juillet 1993, sont les suivantes : limites des zones boisées, cours d'eau.... Le processus de production est conforme à la norme ISO 9002.

♦ **pour une opération :**

- **description de l'opération**
- **date d'application (début et durée de l'application)**
- **équation et paramètres de l'opération**
- **organisme responsable**
- **référence à une spécification**
- **données initiales**
- **données finales**
- **données en paramètres**
- **référence à des versions de logiciels et de matériels**

Exemple : Changement de système de référence réalisé le 10 juin 1994 par l'IGN : passage du WGS 84 à l'ITRF 93, en utilisant les équations et paramètres suivants : paramètres de rotation (xxx, yyy, zzz), de translation (ttt, uuu, vvv) et d'échelle eee.

2) Paramètres de qualité (expression quantitative)

2.1) Cohérence logique

Définition: Degré de cohérence interne des données selon les règles de modélisation et les règles inhérentes à la spécification de produit du jeu de données. Les règles pour la cohérence logique peuvent être de deux types :

1) Les règles de formatage du jeu de données dont le respect rend possible la lecture des données, au sens de la lecture des fichiers contenant les données.

2) Les contraintes d'intégrité définies dans le schéma conceptuel des données ou issues de la spécification de produit.

Exemples :

1. l'existence et l'unicité de l'identifiant des éléments du jeu de données,
2. une contrainte de domaine (toute route a une largeur comprise entre 1 et 40 mètres, un attribut doit avoir une valeur remplie (non indéterminée)),
3. une contrainte géométrique et topologique de graphe planaire,
4. les bâtiments du jeu de données doivent avoir une superficie supérieure à 10 m^2.

Note *:* Les règles du premier type doivent évidemment être impérativement respectées, sinon les données ne peuvent pas être lues.

Informations à fournir :

- **description des violations (référence aux règles enfreintes),**
- **nombre d'irrespects aux règles,**

  Note : En supposant que les données soient lisibles, chaque violation des contraintes d'intégrité imposées est comptabilisée. On obtient donc le nombre d'irrespects aux règles.

- **la taille de l'échantillon.**

Note: Dans la suite de ce document, les paramètres de qualité ne sont définis que sur des entités logiquement cohérentes.

Exemples :

1. Si un attribut doit impérativement avoir une valeur, tout objet du jeu de données n'ayant pas de valeur pour cet attribut n'est pas pris en compte dans la mesure de l'exhaustivité et de la précision sémantique (cf partie 2.3 dans ce chapitre).
2. Si la spécification de produit précise que seules les impasses de plus de 100 mètres doivent figurer dans le jeu de données, alors toute impasse de moins de 100 mètres présente dans le jeu de données n'est pas un excédent (cf. partie 2.3 de ce chapitre), mais une erreur de cohérence logique.

2.2) Précision géométrique

Définition *:* Estimation de la fluctuation des écarts entre les positions nominales (positions dans le terrain nominal) et les positions contenues dans le jeu de données.

***Notes*:**

1) La précision géométrique se décompose en deux types de précision :

- la précision de position composée des précisons de position ponctuelle et linéaire (cf 2.2.1 et 2.2.2)
- la précision de forme (cf 2.2.3)

2) La position sur le terrain nominal est mesurée sur la source de contrôle choisie, la mesure de l'écart s'appuie sur un appariement.

3) Il est de coutume de séparer précision altimétrique et planimétrique car, en pratique (par des méthodes "classiques"), les coordonnées planimétriques et altimétriques sont déterminées séparément, et les systèmes de référence utilisés sont distincts. Dans ce document, sauf mention contraire, nous ne les distinguerons pas car dans les deux cas l'estimation de la précision se fait selon la même méthode (les éventuelles différences seront signalées). Les mêmes définitions peuvent s'appliquer si les trois coordonnées (X, Y et Z) sont estimées dans le même système de référence.

4) La précision géométrique peut être fournie pour une classe d'objets, un regroupement de classes ou une partie de classe définie par un critère de sélection.

2.2.1) Précision de position ponctuelle

Définition : Précision géométrique appliquée à des objets ponctuels (chaque objet est représenté par un point sur le terrain nominal et dans le jeu de données).

Exemples : Précision géométrique d'angles de bâtiments ou d'angles de terrains de sport.

Notes :

1) Un objet ponctuel dans terrain nominal peut avoir un homologue non ponctuel dans les données de contrôle (par exemple, un château d'eau de la BDTOPO® peut être représenté par une surface sur un levé terrain). Dans ce cas, on choisira un centre de la surface (centre de gravité, centre du rectangle englobant...) pour représenter l'homologue de l'objet ponctuel du terrain nominal dans les données de contrôle.

2) La précision de position ponctuelle n'est pas exprimée par une unique valeur, mais par un ensemble d'indicateurs.

Informations à fournir (indicateurs de la précision de position ponctuelle) :

Les trois premières valeurs doivent être exprimées dans l'unité du système de coordonnées

- **La moyenne des erreurs** (en X, en Y ou en Z)

  Ces moyennes fournissent une estimation du **biais**.

- **La grille régulière du biais régionalisé** circonscrite au jeu de données :

  Définition : Grille régulière, définie sur le jeu de données, où, en chaque nœud, est calculé un biais (en X, Y ou Z), dit biais régionalisé. Le biais en un nœud est estimé par la moyenne des erreurs (en X, Y ou Z) sur les points plus proches de ce nœud que de n'importe quel autre nœud.

- **L'exactitude** estimée par la moyenne quadratique des erreurs (EMQ) en X, en Y, en Z, en XY (cette dernière étant appelée EMQ planimétrique) ou en XYZ. Celle-ci doit être accompagnée de sa **précision d'estimation** (cf 2.1 chapitre 4).

- **La taille de l'échantillon** sur lequel sont effectués chacun des calculs précédents.

- **Le taux de rejet** :

  Définition : Rapport du nombre de valeurs d'un échantillon dépassant une certaine tolérance, sur le nombre de valeurs de l'échantillon.

Note : La précision de position ponctuelle peut aussi être exprimée sous la forme d'une tolérance associée à un taux de rejet. Cette expression simplifie l'usage de certaines méthodes d'échantillonnage en contrôle qualité.

2.2.2) Précision de position linéaire

Définition : Précision géométrique appliquée à des objets linéaires (chaque objet est représenté par une ligne sur le terrain nominal et dans le jeu de données).

Exemples : Précision géométrique de routes, de cours d'eau.

Notes :

1) Un objet linéaire du terrain nominal peut avoir un homologue non linéaire dans les données de contrôle (par exemple, une route peut être représentée par une bande, donc une surface dans les données de contrôle). Dans ce cas, on choisira une ligne contenue dans la bande (l'axe, par exemple) pour représenter l'homologue de l'objet linéaire du terrain nominal dans les données de contrôle.

2) Si l'on considère une ligne comme un ensemble de points, on peut penser utiliser les indicateurs du 2.2.1 sur certains de ces points pour estimer la précision de position linéaire. Or, cette démarche, consistant à appliquer la méthode du **contrôle ponctuel** à des **objets linéaires** pose plusieurs problèmes qui ont été soulevés dans [Abbas, 94]. En particulier, il est en pratique souvent impossible de déterminer sur la référence les homologues des points d'une ligne complexe du jeu de données (et donc d'appliquer proprement la méthode du contrôle ponctuel).

Informations à fournir (indicateurs de la précision de position linéaire)

Le **contrôle linéaire** tel qu'il est exposé dans [Abbas, 94] est fondé sur l'exploitation de la **distance de Hausdorff** sur des couples de lignes homologues (cf figure 8) et fournit deux indicateurs de la qualité planimétrique d'un jeu de données :

- **taux d'accord entre le tracé du jeu de données et le tracé de la référence**

  Définition : Pourcentage de tracé du jeu de données en accord avec le tracé de la référence après **régularisation** des contours.

  Note : La régularisation des contours consiste à couper les détails discordants, c'est-à-dire hors d'une certaine tolérance, sur les contours homologues, tout en comptabilisant les longueurs de ces parties coupées.

- sur les parties en accord, **une estimation de l'EMQ planimétrique** (et sa précision d'estimation).

  Note : Le calcul de l'EMQ planimétrique (linéaire) fait intervenir les composantes de la distance de Hausdorff d'un contour du jeu de données au contour homologue, mais également les composantes de la distance de Hausdorff entre des **simulations** du jeu de données et de la référence. Ce calcul n'est pas détaillé dans ce document, pour plus de détails, se référer à [Abbas, 94].

M_1 : point du contour 1

M_2 : point du contour 2

$d_1 = \underset{M_1}{\text{Max}}\ \underset{M_2}{\text{Min}} \{ d(M_1, M_2) \}$

$d_2 = \underset{M_2}{\text{Max}}\ \underset{M_1}{\text{Min}} \{ d(M_1, M_2) \}$

$d_H = \text{Max}(d_1, d_2)$

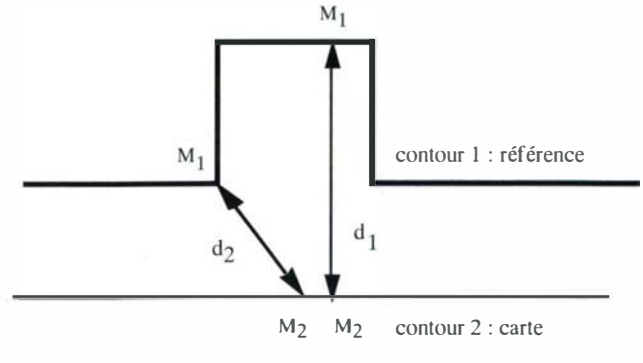

Figure 8 : distance de Hausdorff

2.2.3) Précision de forme

Définition *(CNIG) :* Qualification des éléments géométriques (distances, surfaces, courbures, volumes...) construits à partir de coordonnées connues, et dont la précision de position planimétrique et/ou altimétrique est par ailleurs évaluée. Elle donne l'écart entre la valeur de ces éléments sur la source de contrôle et la valeur issue du jeu de données.

***Notes**:*

1) La précision de position relative est la précision de forme sur la distance entre deux points.

2) La précision de forme suppose que les éléments géométriques ne soient pas des attributs spécifiques. En particulier, si la courbure d'une courbe n'est pas issue d'un calcul à partir des éléments qui décrivent cette courbe, mais est présente en tant qu'attribut de la courbe, il s'agit alors d'une information sémantique.

3) Ne pouvant prévoir dans ce document quelles seront les caractéristiques de forme utiles pour l'appréciation de la qualité d'un jeu de données, il est indispensable de préciser ces caractéristiques dans la spécification de produit du jeu de données.

Informations à fournir :

- **description de la caractéristique appréciée,**
- **valeur,**
- **unité,**
- **taille de l'échantillon,**
- **taux de rejet.**

2.3) Exhaustivité et précision sémantique

Définitions *(CNIG) :*

1) L'exhaustivité est la conformité de la présence ou de l'absence des éléments du jeu de données par rapport au terrain nominal. Elle s'attache à des objets, des attributs ou des relations.

2) La précision sémantique est la conformité des valeurs des éléments du jeu de données avec les valeurs de leurs homologues dans le terrain nominal. Elle porte sur la classification des objets, la codification des attributs et les relations entre les objets.

Note: Dans ce document, l'exhaustivité et la précision sémantique sont regroupées dans la même sous-partie car les informations de qualité qu'elles transmettent sont très liées.

L'exhaustivité et la précision sémantique s'expriment différemment pour la classification des objets, pour la codification des attributs au sein d'une même classe ou pour une relation.

2.3.1) Classification des objets

Soient :

- N_i et n_i les nombres d'objets de la classe C_i du terrain nominal et du jeu de données respectivement.
- N_{0j} et N_{i0} représentant respectivement le nombre d'objets de la classe C_j du jeu de données sans homologue dans le terrain nominal (objets en trop dans le jeu de données) et le nombre d'objets de la classe C_i du terrain nominal sans homologue dans le jeu de données (objets manquants dans le jeu de données).
- N_{ij} le nombre d'objets de la classe C_i du terrain nominal appariés à un objet de la classe C_j du jeu de données.

Informations à fournir :

Pour l'exhaustivité :

- Le **taux de déficit** et le **taux d'excédent** définis et calculés de la façon suivante :

  Définition *:*

  1) Le taux de déficit de la classe C_i, noté T_{Ci0}, est le rapport N_{i0}/N_i (si $N_i=0$, alors $N_{i0}=0$ et par convention $T_{Ci0}=0$).

  2) Le taux d'excédent de la classe Cj, noté T_{C0j}, est le rapport N_{0j}/n_j (si $n_j=0$, alors $N_{0j}=0$ et par convention $T_{C0j}=0$).

Pour la précision sémantique :

- Le **taux de confusion** et le **taux d'accord** définis et calculés de la façon suivante :

  Définition *:*

  1) Le taux de confusion entre les classes C_i et C_j (avec $i \neq j$), noté T_{Cij}, est le rapport N_{ij}/N_i (si $N_i=0$, alors $N_{ij}=0$ et par convention $T_{Cij}=0$).

  2) Le taux d'accord de la classe C_i, noté T_{Cii}, est le rapport N_{ii}/N_i (si $N_i=0$, alors $N_{ii}=0$ et par convention $T_{Cii}=1$).

***Notes*:**

1) Au lieu d'être exprimés en nombre d'objets, les taux précédents peuvent être exprimés en mesure. Par exemple, il peut être plus évocateur pour des tronçons de route d'exprimer ces taux en termes de longueur qu'en termes de nombre de tronçons. L'expression des taux en mesure devient nécessaire lorsque plusieurs objets du terrain nominal (respectivement du jeu de données) correspondent à un unique objet du jeu de données (respectivement du terrain nominal).

2) Les taux peuvent aussi être exprimés en associant un certain poids à chaque objet. Ces poids doivent être définis dans la spécification ou imposés par la méthode de sondage adoptée.

3) On peut remarquer que pour une même classe du terrain nominal la somme des taux d'accord, de confusion et de déficit relatif à cette classe vaut 1: $\forall i \in \{1,\ldots,p\}, T_{Ci0} + \sum_{j\in\{1,\ldots,p\}} T_{Cij} = 1$

4) Ces taux peuvent être calculés pour des regroupements de classes ou des parties de classe définies par un critère de sélection.

5) Les taux sont généralement exprimés en pourcentage.

• Les effectifs de chaque classe de l'échantillon et de la partie du terrain nominal correspondante.

Note : Ces informations peuvent être représentées sous la forme d'une matrice de confusion pour la classification des objets définie ci-après.

Définition *:* Une matrice de confusion pour la classification des objets est un tableau dans lequel figurent les taux de confusion, d'accord, d'excédent et de déficit relatifs aux classes (ou regroupement de classes ou partie de classe d'après la note 3 ci-dessus) du terrain nominal et du jeu de données, ainsi que les effectifs de ces classes et les demi-longueurs des intervalles de confiance des estimations de ces taux (qui dépendent des précisions d'estimation de ces taux).

Exemple : La figure 9 fournit un exemple imaginaire de matrice de confusion pour la classification des objets.

Matrice de confusion pour la classification des objets
Les taux sont exprimés en pourcentage par rapport au: nombre d'objets
Le niveau de confiance est de 95%

jeu de données / terrain nominal	Chemin 95	Sentier 50	Allée 41	Piste cyclable 8	Néant		
Chemin 100	90% ±5,9%	10% ±5,9%	0% 0%	0% 0%	0% 0%	Σ=100	*éléments de confusion*
Sentier 50	10% ±8,3%	70% ±12,7%	0% 0%	0% 0%	20% ±11%	Σ=100	
Allée 40	0% 0%	0% 0%	100% 0%	0% 0%	0% 0%	Σ=100	
Piste cyclable 10	0% ?	0% ?	0% ?	80% ?	20% ?	Σ=100	
Néant	0% 0%	10% ±8,3%	2,4% ±4,6%	0% ?		excédents	
	éléments de confusion				déficits	**accord**	

Figure 9 : Exemple de matrice de confusion pour la classification des objets

Clés de lecture :

- dans la deuxième colonne, à la première ligne de la matrice (taux de confusion de la classe Chemin (du terrain nominal) avec la classe Sentier (du jeu de données)) le taux de 10% signifie que 10% des objets de la classe Chemin du terrain nominal (donc 10 objets) ont un homologue dans le jeu de données qui appartient à la classe Sentier. Toutes les autres cases de la matrice de confusion s'interprétent de la même façon, à l'exception de la dernière ligne,
- dans la deuxième colonne, à la dernière ligne (taux d'excédent en sentier) le taux de 10% signifie que 10% des objets de la classe Sentier du jeu de données (donc 5 objets) n'ont pas d'homologue dans le terrain nominal, donc sont en trop. Les autres taux de la dernière ligne (taux d'excédents) s'interprétent de la même façon.

2.3.2) Codification des attributs au sein d'une même classe

Au sein d'une même classe C on ne considère plus que les objets "correctement appariés", c'est-à-dire les objets de C dans le terrain nominal appariés à un objet de C dans le jeu de données et, on désigne par N le nombre de ces objets.

Informations à fournir

Pour l'exhaustivité

- Le **taux de déficit sur un attribut** et le **taux d'excédent sur un attribut** définis et calculés de la façon suivante :

Définitions *:*

1) Le taux de déficit sur l'attribut V de la classe C est le rapport du nombre d'objets de C correctement appariés dont la valeur de V est déterminée dans le terrain nominal alors que la valeur pour V de son homologue n'est pas déterminée, sur N.

2) Le taux d'excédent sur l'attribut V de la classe C est le rapport du nombre d'objets de C correctement appariés dont la valeur de V n'est pas déterminée dans le terrain nominal alors que la valeur pour V de son homologue est déterminée, sur N.

Pour la précision sémantique

- Le **taux d'absence sur un attribut** et le **taux de présence sur un attribut** définis et calculés de la façon suivante :

Définitions *:*

1) Le taux d'absence sur l'attribut V de la classe C est le rapport du nombre d'objets de C correctement appariés dont la valeur de V ainsi que celle de son homologue ne sont pas déterminées, sur N.

2) Le taux de présence sur l'attribut V de la classe C est le rapport du nombre d'objets de C correctement appariés dont la valeur de V ainsi que celle de son homologue existe, sur N.

Notes :

1) La signification de ces quatre taux dépend de la signification associée à la valeur indéterminée pour l'attribut (voir note 1 du chapitre 2). En particulier, si la valeur indéterminée correspond au cas "inconnu" alors la valeur est toujours déterminée dans le terrain nominal et donc le taux d'excédent et le taux d'absence sont toujours nuls.

2) La somme des taux d'absence, de présence, de déficit et d'excédent relatifs à un attribut V d'une classe C vaut 1.

3) On rappelle que seuls des objets cohérents participent à l'évaluation de l'exhaustivité et de la précision sémantique. Ainsi, s'il est interdit qu'une valeur ne soit pas déterminée, pour l'attribut V de la classe C, un objet du jeu de données n'ayant pas de valeur pour V ne se comptabilise pas comme un déficit sur l'attribut V, mais comme une erreur de cohérence logique (cf 2.1).

4) Au lieu d'être exprimés en nombre d'objets, les taux précédents peuvent être exprimés en mesure (longueur, surface...) ; ils peuvent aussi être éventuellement pondérés (cf notes 1 et 2 de la définition des taux de confusion et d'accord).

5) Ces taux peuvent être calculés pour des regroupements de classes ou des parties de classe définies par un critère de sélection.

6) Ces informations peuvent être représentées sous la forme d'une matrice d'absence pour un attribut définie comme suit :

Définition : Pour un attribut d'une classe d'objets, la matrice d'absence est un tableau dans lequel sont représentés les taux de déficit, d'excédent, d'absence et de présence sur cet attribut, ainsi que les demi-longueurs des intervalles de confiance associés à ces taux.

Exemple: La figure 10 fournit un exemple imaginaire d'une matrice d'absence pour un attribut.

Matrice d'absence pour l'attribut : **Monument.Toponyme**
Taux exprimés en % par rapport au nombre d'objets
Echantillon de 100 objets
Niveau de confiance de 95%

jeu de données / terrain nominal	non déterminée	déterminée
non déterminée	25% ±8,5%	10% ±5,9%
déterminée	15% ±7%	50% ±9,8%

Figure 10 : exemple de matrice d'absence pour un attribut

Clé de lecture :

Le taux de 25% dans la case correspondant à la première ligne et première colonne de la matrice d'absence signifie que 25% des objets considérés, donc 25 objets, n'ont pas de valeur dans le terrain nominal pour l'attribut Toponyme et leur homologue n'a pas non plus de valeur pour ce même attribut. Les autres cases s'interprétent de la même façon.

Autres informations à fournir :

- **La taille de l'échantillon**, c'est-à-dire le nombre d'objets sur lesquels ont été effectués chacun des calculs précédents.

- **Les autres informations à fournir dépendent de la nature de l'attribut et sont détaillées dans les pages suivantes (de 2.3.2.1) "Attribut qualitatif énuméré" à 2.3.2.3) "Attribut quantitatif").**

2.3.2.1) Attribut qualitatif énuméré

On considère ici que l'attribut V peut prendre les valeurs V_1, V_2,...V_p. On peut dans ce cas préciser les taux de déficit et d'excédent en fonction des valeurs de V et le taux de présence en taux de confusion et d'accord sur les valeurs de l'attribut. Tous ces taux sont alors regroupés dans une matrice de confusion.

Exemple : La figure 11 fournit un exemple imaginaire de matrice de confusion pour un attribut qualitatif énuméré.

Matrice de confusion pour l'attribut qualitatif énuméré :
Tronçon de route.Vocation
Les taux sont exprimés en pourcentage par rapport au nombre d'objets
Le niveau de confiance est de 95%

jeu de données / terrain nominal	Chemin 94	Bretelle 54	Escalier 47	Passerelle 16	Non déterminée 9		
Chemin 100	90% ±5,9%	5% ±4,3%	0% 0%	0% 0%	5% ±4,3%	$\Sigma(T_{C1k})=100$, k=0,1,...p	*éléments de confusion*
Bretelle 50	2% ±2,7%	90% ±5,9%	0% 0%	4% ±3,8%	4% ±3,8%	$\Sigma(T_{Cik})=100$, k=0,1,...p	
Escalier 50	6% ±4,6%	4% ±3,8%	90% ±5,9%	0% 0%	0% 0%	$\Sigma(T_{Cjk})=100$, k=0,1,...p	
Passerelle 10	0% ?	0% ?	0% ?	100% ?	0% ?	$\Sigma(T_{Cpk})=100$, k=0,1,...p	
Non déterminée 10	0% 0%	3,7% ±3,7%	4,2% ±3,9%	25% ?	22,2% ?	**excédents**	
	éléments de confusion				**déficits**	**accord**	

Figure 11: exemple de matrice de confusion pour un attribut qualitatif énuméré

Clés de lecture :

- dans la deuxième colonne, à la première ligne de la matrice (taux de confusion entre la valeur Chemin (du terrain nominal) avec la valeur Bretelle (du jeu de données)) le taux de 5% signifie que 5% des objets "correctement appariés" ayant la valeur Chemin dans le terrain nominal (donc 5 objets) ont un homologue (dans le jeu de données) qui possède la valeur Bretelle. Toutes les autres cases de la matrice de confusion s'interprétent de la même façon, à l'exception de la dernière ligne.
- dans la deuxième colonne, à la dernière ligne (taux d'excédent sur la valeur Bretelle) le taux de 3,7% signifie que 3,7% des objets "correctement appariés" ayant la valeur Bretelle dans le jeu de données (donc 2 objets) ont un homologue dans le terrain nominal dont la valeur de l'attribut Vocation n'est pas déterminée. Les autres taux de la dernière ligne (taux d'excédents) s'interprétent de la même façon.

2.3.2.2) Attribut qualitatif non énuméré

Le taux de présence se décompose ici en deux taux :

Définitions :

1) Le taux d'accord qui est le rapport du nombre d'objets de C correctement appariés dont la valeur de V correpond à celle de son homologue, sur N.

2) Le taux de désaccord qui est le rapport du nombre d'objets de C correctement appariés dont la valeur de V diverge de celle de son homologue.

Ces taux peuvent être représentés dans la matrice d'absence de l'attribut comme dans l'exemple de la figure 9.

Notes :

1) Pour l'attribut V (qualitatif non énuméré) de la classe C, la somme du taux d'accord et du taux de désaccord est identique au taux de présence.

2) Ces taux peuvent être calculés pour des regroupements de classes ou des parties de classe définies par un critère de sélection.

3) Ces taux peuvent être représentés dans la matrice d'absence de l'attribut comme dans l'exemple de la figure 12.

Matrice d'absence pour l'attribut : **Commune.Nom**
Taux exprimés en pourcentage par rapport au nombre d'objets.
Echantillon de 100 objets
Niveau de confiance de 95%

jeu de données / terrain nominal	non déterminée	déterminée
non déterminée	10% ±5,9%	5% ±4,27%
déterminée	15% ±7%	accord : 50% ±9,8% 70% ±9% désaccord : 20% ±7,8%

Figure 12 : exemple de matrice d'absence pour un attibut qualitatif non énuméré.

Clé de lecture : cette matrice s'interprète comme la matrice d'absence présentée en début de 2.3.2

2.3.2.3) Attribut quantitatif

Pour les objets de C correctement appariés ayant une valeur pour l'attribut V appariés à un objet (du jeu de données) ayant une valeur pour V, on fournit :

- **la moyenne des écarts,**
- **l'erreur moyenne quadratique,**

- **l'écart-type,**
- **la taille de l'échantillon,**
- **le taux de rejet.**

Note *:* Ce type de calcul peut ne présenter qu'un intérêt restreint pour certains attributs de nature quantitative (nombre de voies d'une route). Dans ce cas, on donnera plutôt les valeurs correspondant aux attributs de nature qualitative énumérée.

2.3.3) Relation

On se restreint à étudier les relations entre entités si les entités liées par une relation dans le terrain nominal possèdent des homologues dans le jeu de données (les entités en excédent ou en déficit entraîneront forcément des liens incorrects).
Soient E l'ensemble des couples d'entités vérifiant une relation R, E_{TN} et E_{JD} définis respectivement par l'ensemble E du terrain nominal et l'ensemble E du jeu de données.

- **Le taux d'accord sur une relation**

Définition *:* Le taux d'accord sur la relation R est le rapport du nombre de couples de E_{TN} que l'on peut mettre en correspondance avec un couple de E_{JD}, sur le nombre de couples de E_{TN}.

Note La mise en correspondance d'un couple est la mise en correspondance des deux champs du couple.

- le **taux de déficit sur une relation** et le **taux d'excédent sur une relation** définis de la façon suivante :

Définitions *:*

1) Le taux de déficit est le rapport du nombre de couples de E_{TN} que l'on ne retrouve pas dans E_{JD}, sur le nombre de couples de E_{TN}.

2) Le taux d'excédent est le rapport du nombre de couples de E_{JD} que l'on ne retrouve pas dans E_{TN}, sur le nombre de couples de E_{JD}.

- **La taille de l'échantillon**

Note *:* Les attributs de relations se traitent de la même manière que les attributs d'objets (cf 2.3.2).

2.3.4) Note sur la partie "Exhaustivité et précision sémantique"

Les informations de qualité relevant de la sémantique (précision sémantique et exhaustivité) peuvent également être représentées sous la forme de diagrammes tels que celui de la figure 13. Cette représentation compacte et facilement accessible possède, néanmoins, l'inconvénient de cacher certaines informations : les confusions sont regroupées en "données incorrectes" et les précisions d'estimation ne figurent pas.

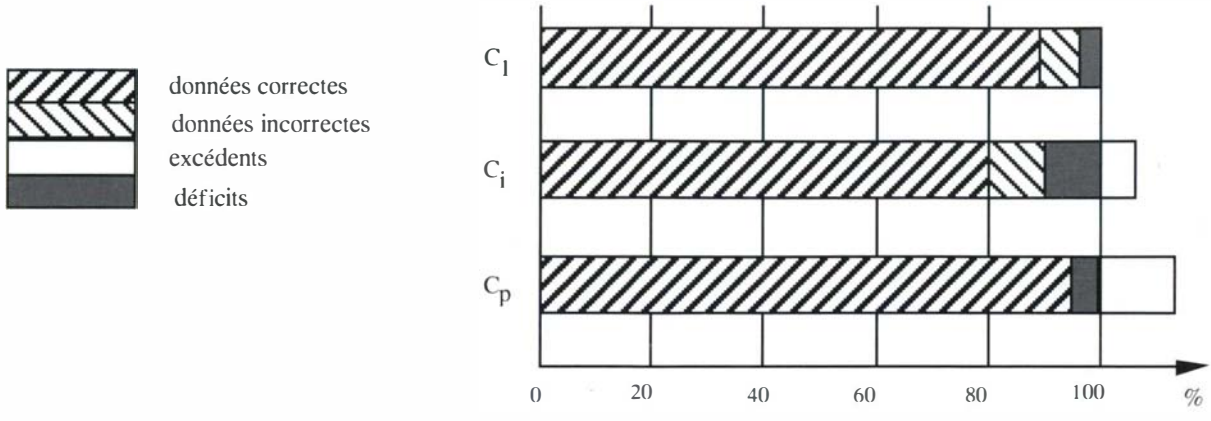

Figure 13 : diagramme de précision sémantique et d'exhaustivité

Sur la figure 13, C_i signifie un ensemble de données regroupées selon une caractéristique (soit une classe d'objets, soit une partition d'une classe selon une valeur d'attribut, etc...).

Exemple : Le diagramme de la figure 14 est la traduction de l'exemple du 2.3.1 Classification des objets, de la figure 10, selon le modèle défini ci-dessus.

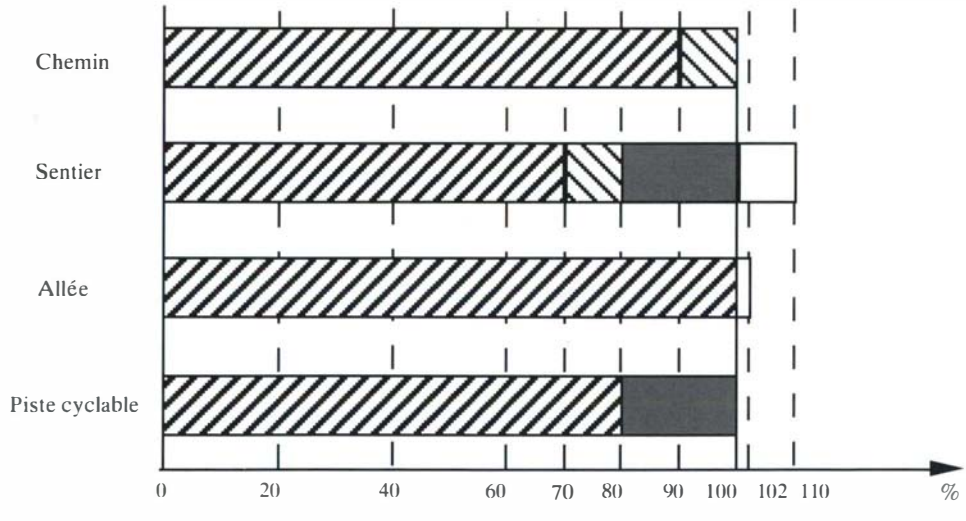

Figure 14 : exemple de diagramme

2.4) Qualité spécifique

Définition *(CNIG) :* Ce paramètre permet de transmettre des informations, concernant la qualité, non prévues par les critères précédents.

***Notes**:*

1) On ne doit recourir à la qualité spécifique que dans le cas où les autres paramètres sont insuffisants pour décrire l'information de qualité recherchée.

2) Pour des produits spécifiques, tels que les MNT, d'autres critères de qualité que ceux présentés jusqu'ici doivent être pris en compte, ils peuvent donc s'intégrer dans ce paramètre de qualité spécifique.

Informations à fournir :

- **description de la caractéristique appréciée,**
- **valeur de la caractéristique,**
- **unité,**
- **taille de l'échantillon,**
- **taux de rejet.**

Exemple : Pour les toponymes, on distingue souvent les fautes d'orthographe des autres fautes. Pour pouvoir utiliser cet indicateur, il est nécessaire de définir une mesure pour ce type de fautes.

3) Actualité

Jusqu'à présent (dans ce document) l'aspect évolutif du terrain nominal a été ignoré.

Définition : Décalage entre un jeu de données et le terrain nominal à une date T. Elle décrit en quelque sorte la "fraîcheur" des données. Elle se chiffre par la précision sémantique et l'exhaustivité du jeu de données à la date T.

Notes :

1) L'évaluation de l'actualité à la date T ne peut se faire rigoureusement que par des mesures de la précision sémantique et de l'exhaustivité des données à la date T. Or ceci suppose d'effectuer une mesure ou un contrôle qualité du jeu de données. Ainsi, entre deux mises à jour, seul l'utilisateur peut estimer l'actualité des données qu'il possède pour savoir si ces dernières sont toujours correctes au moment où il veut les utiliser. Cependant, l'utilisateur ne veut pas avoir la charge d'une mesure ou d'un contrôle qualité coûteux. Il a donc besoin d'informations annexes lui permettant de se faire une idée de l'actualité de son jeu de données. Ces informations sont définies dans cette partie.

2) Ainsi qu'il a été précisé dans les sections 5 et 15 du chapitre 3, l'imperfection de la spécification introduit une incertitude sur le terrain nominal et il faut distinguer les erreurs (différences entre le jeu de données et le terrain nominal) de cette incertitude. Il convient d'y prendre particulièrement garde si la spécification de produit d'une base de données prévoit une ancienneté maximale pour une information avant sa mise à jour. Par exemple, si une route apparaît dans l'univers le 1/1/95 et que la spécification d'une base de données prévoit que le décalage maximum entre l'univers et cette base de données est de six mois, alors si cette base est datée du 1/3/95, elle peut comporter ou non cette route sans qu'il y ait d'erreur. Ce décalage maximum (ou ancienneté maximale) est une incertitude du terrain nominal qui admet ou non l'absence ou la présence de la route.

• Date de validation

Définition : Pour un ensemble de données, la date de validation est la date la plus récente à laquelle cet ensemble de données a été validé, c'est-à-dire considéré comme correct.

Notes:

1) Cette date est, à l'origine, la date de collecte des données (date de prise de vues aériennes...). La modification de cette date peut se produire lors d'une mise à jour (elle devient alors la date de dernière mise à jour) ou lors d'un contrôle attestant que l'ensemble de données est toujours valide.

2) La date de validation doit être accompagnée de l'ensemble de données pour lequel elle s'applique. Il faut donc préciser les classes d'objets concernées par cette date, ainsi que l'étendue spatiale de ces objets.

3) Cette date ne correspond pas à la date de mise en circulation des données. Un laps de temps plus ou moins long peut séparer la collecte des données et leur mise en service.

• Date de validité

Définition *:* Pour une certaine donnée, la date de validité de cette donnée est la date à laquelle elle apparaît dans le terrain nominal.

Note: Généralement, la date de validité est inconnue et on se contente de la date de validation. Cependant, pour certaines informations, telles que les informations administratives, la date de validité est plus intéressante que la date de validation.

• **Date de péremption (ou de fin de validité)**

Définition *:* Date indiquant à partir de quand un ensemble de données n'est plus valide.

Note : Cette date n'est connue que pour un nombre restreint de données, comme, par exemple, la date d'ouverture d'une autoroute en construction. Par conséquent, dans la majorité des cas, la date de péremption ne peut être communiquée.

• **Taux d'évolution**

Définition *:* Pour une certaine donnée, le taux d'évolution de cette donnée est la probabilité que cette donnée change par unité de temps, cette dernière étant à préciser. Cette probabilité est estimée par :

$$\frac{\text{nombre de changements entre T et T0}}{T - T0}$$

où $T\text{-}T_0$ est une durée exprimée dans l'unité de temps et où un changement est soit une "destruction", soit une "création", soit une "modification".

Notes*:*

1) Les données dont on calcule le taux d'évolution ne sont pas nécessairement des éléments du jeu de données produit. Ce sont également des informations complémentaires permettant d'apprécier l'actualité du jeu de données.

Exemples :

1) Au 1er janvier 1990, un jeu de données comportant, entre autres, les habitations d'une agglomération est produit. Il est mis à jour pour la première fois le 1er janvier 1995. Pendant cette période de cinq ans, la collecte d'informations diverses a été réalisée régulièrement. En particulier, on a pu constater un taux de permis de construire de 5% sur cette agglomération. Ce taux peut être fourni avec le jeu de données mis à jour et donne une indication sur ce que sera peut être l'évolution des habitations après le 1er janvier 1995 dans l'agglomération.

2) Au 1er janvier 1994, la classe arbre du terrain nominal sur une commune est constituée de 500 objets. Au 1er janvier 1995, le terrain nominal comporte 550 objets de la classe arbre pour cette même commune. On a recensé 100 arbres en plus, 50 en moins et 50 modifications (par exemple dûes à une augmentation sensible de l'attribut taille d'un arbre) par rapport au 1er janvier 1994. En un an, il y a donc eu 200 changements, donc un pourcentage de changements de 200/500 = 40% en un an. Si l'on choisit comme unité de temps le mois, le taux d'évolution est donc de (40/12)% = 3,33%.

2) Le calcul pratique d'un taux d'évolution ne sera pas toujours possible ou sera très approximatif et peu fiable. Cependant, au fur et à mesure des mises à jour ou de la collecte, un affinement de cette estimation sera possible. La fiabilité d'un taux d'évolution dépend en grande partie du niveau d'agrégation de l'information. Si localement la variation temporelle d'un taux d'évolution risque d'être très importante, elle peut être beaucoup moins sensible à un niveau plus global. Par exemple, si le taux d'évolution de la surface forestière dans une commune peut varier énormément dans le temps à cause d'incendies, cette variation peut être faible au niveau départemental.

• **Politique d'entretien des données**

Définition *:* Organisation des mises à jour qui sont effectuées sur les données. Les types de mises à jour sont :

- pas de mise à jour,
- mise à jour sur demande d'un utilisateur,
- mise à jour continue. Le délai maximum entre le changement dans l'univers et la disponibilité du jeu de données doit être défini,
- mise à jour périodique. Les périodes peuvent varier pour différents types d'entités, d'attributs ou de relations ainsi que pour différentes régions de l'ensemble des données. La durée des différentes périodes doit être définie,
- mise à jour déclenchée par un volume donné de changements.

Chapitre 6

Etude d'un exemple

Dans ce chapitre, nous exposons un exemple illustrant certaines notions de qualité présentées précédemment. Ces notions sont celles de précision sémantique, d'exhaustivité et de cohérence logique. Vu la petite taille de l'univers et son aspect caricatural (pas de système de coordonnées), la précision géométrique n'est pas abordée dans cet exemple.

Pour cela nous utiliserons comme point de départ (comme univers) la figure 15 :

- le chiffre sous le symbole de l'arbre indique la hauteur de cet arbre,
- le chiffre dans la maison indique le nombre d'habitants de cette maison,
- une maison peut posséder un nom qui est dans ce cas inscrit à côté de la maison.

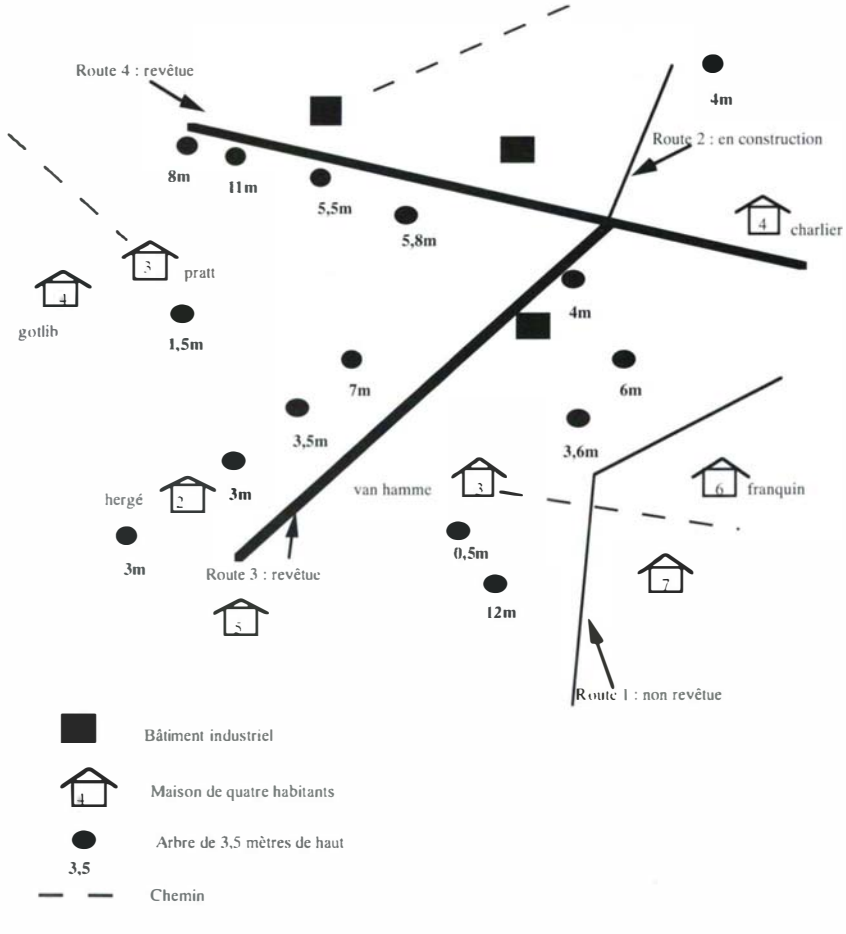

Figure 15 : une représentation de l'univers

La spécification de produit est décrite ci-dessous :

Le schéma conceptuel contient cinq types d'entité qui sont :

Bâtiment industriel

Maison
nom : chaine de caractères
nombre d'habitants : entier

Arbre
hauteur : {1 : de 1 à 3m, 2 : de 3 à 5m, 3 : de 5 à 10m, 4 : plus de 10m}

Chemin

Route
Etat : {revêtue, non revêtue, en construction}

- Chaque intersection 2D doit donner naissance à un nœud.
- Les arbres d'une hauteur inférieure à 1 mètre ne doivent pas être pris en compte.
- Lorsque la hauteur d'un arbre est inconnue, l'attribut "hauteur" d'un arbre ne prend pas de valeur.
- L'attribut état d'une route doit avoir une valeur.
- Lorsque le nom d'une maison est inconnue, l'attribut "nom" de la maison ne prend pas de valeur.
- Lorsque le nombre d'habitants d'une maison est inconnu, l'attribut "nombre d'habitants" de la maison ne prend pas de valeur.

Cette spécification définit le terrain nominal de la figure 16 :

(On remarque que l'arbre de hauteur 0,5 mètre n'apparaît pas dans le jeu de données conformément au critère de sélection sur la hauteur des arbres).

Les fautes d'orthographe dans les toponymes ne seront pas distinguées des autres erreurs sémantiques.

Supposons que le jeu de données produit soit celui représenté sur la figure 17.

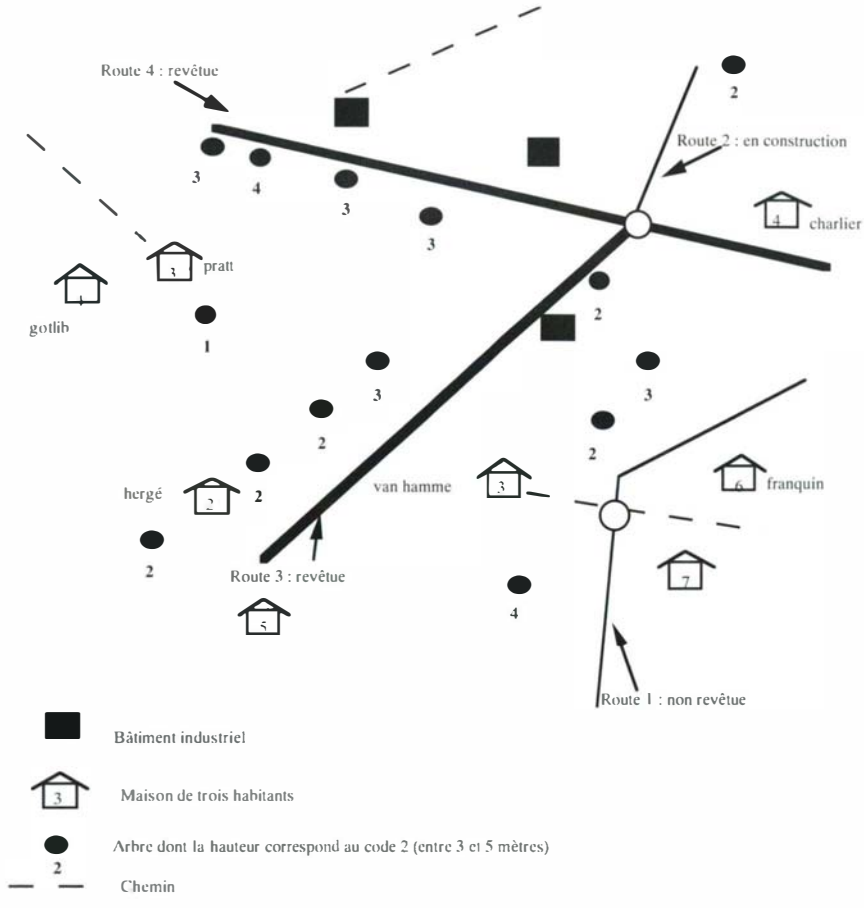

Figure 16 : le terrain nominal

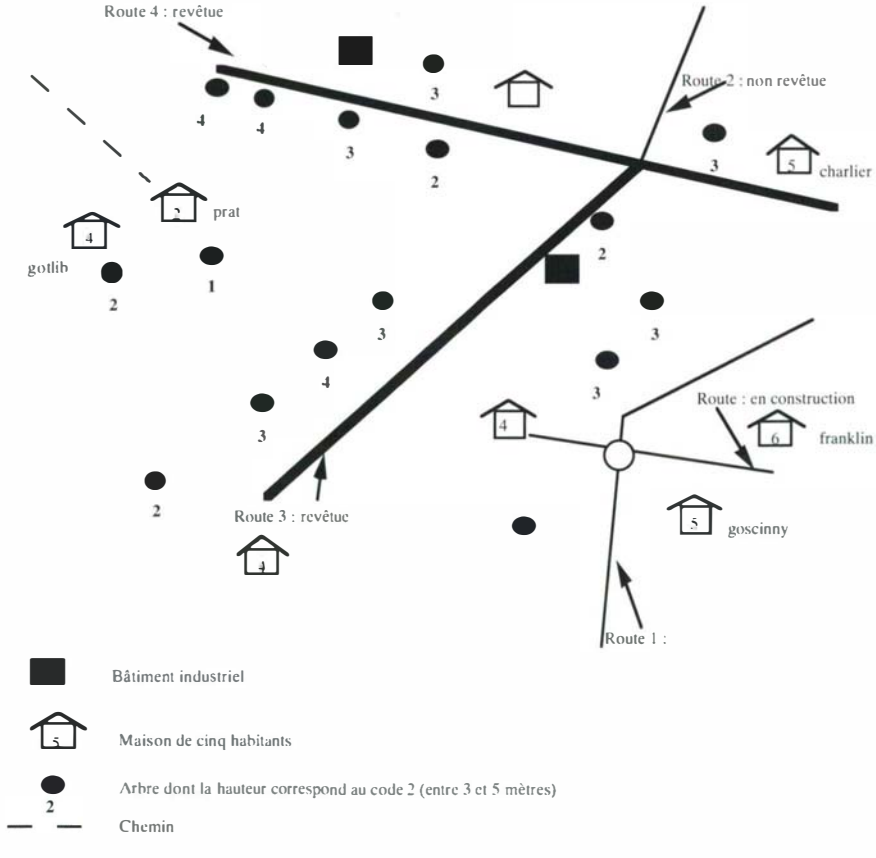

Figure 17 : le jeu de données

Un certain nombre d'erreurs se sont immiscées dans le jeu de données. La liste de ces erreurs est :

1) 2 arbres de hauteur comprise entre 5 et 10 mètres sont en excédent, de même qu'un arbre de hauteur comprise entre 3 et 5 mètres. Par contre, il manque un arbre de hauteur comprise entre 3 et 5 mètres.

2) Un chemin manque.

3) Une maison a pris la place d'un bâtiment industriel.

4) L'attribut état de la route 1 n'a pas de valeur.

5) Une route a pris la place d'un chemin.

6) Il manque une maison.

7) La valeur de l'attribut état de la route 2 n'est pas la bonne.

8) Il manque un nœud à l'intersection des routes 2, 3 et 4.

9) La valeur de l'attribut hauteur d'un arbre n'est pas déterminée dans le jeu de données alors qu'elle vaut 4 mètres dans le terrain nominal.

10) Il y des confusions sur les valeurs de l'attribut hauteur des arbres :

- deux arbres de hauteur comprise entre 3 et 5 mètres sont devenus des arbres de hauteur comprise entre 5 et 10 mètres,
- un arbre de hauteur comprise entre 3 et 5 mètres est devenu un arbre de hauteur de plus de 10 mètres,
- un arbre de hauteur comprise entre 5 et 10 mètres est devenu un arbre de hauteur de plus de 10 mètres,
- un arbre de hauteur comprise entre 5 et 10 mètres est devenu un arbre de hauteur comprise entre 3 et 5 mètres.

11) La maison nommée "van hamme" dans le terrain nominal n'a plus de nom dans le jeu de données. Au contraire, la maison nommée "goscinny" dans le jeu de données n'avait pas de nom dans le terrain nominal.

12) Les maisons nommées "franquin" et "pratt" sont nommées dans le jeu de données respectivement par "franklin" et "prat".

13) Certains nombres d'habitants sont inexacts.

La priorité dans le comptage des erreurs doit porter sur les erreurs de cohérence logique, puisque conformément à ce qui a été présenté dans la partie "Cohérence logique" les mesures de précision sémantique et d'exhaustivité ne se font que sur des objets logiquement cohérents.

Cohérence logique :

Il y a en tout deux violations aux règles de cohérence logique :

- l'attribut état de la route 1 n'a pas de valeur.
- il manque un nœud à l'intersection des routes 2, 3 et 4.

Précision sémantique et exhaustivité :

Dans les matrices représentées dans la suite de ce document, figurent les taux exprimés en pourcentage et (sauf pour les zéros) également sous la forme de quotient afin de bien montrer comment ils sont calculés. Les précisions d'estimation de ces taux ne sont pas calculés du fait des effectifs faibles de chaque classe.

1) Classification des objets

Les confusions, accords, déficits et excédents concernant la classification des objets sont représentés sur la matrice de confusion de la figure 18.

Matrice de confusion pour la classification des objets
Les taux sont exprimés par rapport au: nombre d'objets

jeu de données / terrain nominal	Chemin 1	Route 4	Arbre 16	Bâtiment industriel 2	Maison 8	Néant		
Chemin 3	1/3=33,3%	1/3=33,3%	0%	0%	0%	1/3=33,3%	Σ=1	*éléments de confusion*
Route 3	0%	3/3=100%	0%	0%	0%	0%	Σ=1	
Arbre 14	0%	0%	13/14=93%	0%	0%	1/14=7%	Σ=1	
Bâtiment industriel 3	0%	0%	0%	2/3=66,6%	1/3=33,3%	0%	Σ=1	
Maison 8	0%	0%	0%	0%	7/8=87,5%	1/8=12,5%	Σ=1	
Néant	0%	0%	3/16=18,7%	0%	0%		excédents	
	éléments de confusion					déficits	accord	

Figure 18 : matrice de confusion sur la classification des objets

L'effectif de 3 (et non 4) pour la classe route du terrain nominal provient du fait qu'on ne se préoccupe que d'objets logiquement cohérents, ce qui n'est pas le cas de la route 1 (cf. ci-dessus).

2) Codification des attributs

Dans les effectifs des matrices de confusion sur la valeur des attributs ne figurent que les objets d'une classe ayant un homologue dans cette même classe.

• Pour l'attribut hauteur de la classe arbre qui est de nature qualitative énumérée, les résultats sont représentés sur la figure 19 :

Matrice de confusion pour l'attribut qualitatif énuméré :
Arbre.Hauteur
Les taux sont exprimés en pourcentage par rapport au nombre d'objets

jeu de données / terrain nominal	**1:[1,3[1**	**2:[3,5[3**	**3:[5,10[5**	**4:≥10 3**	**Non déterminée 1**		
1:[1,3[1	1/1=100%	0%	0%	0%	0%	$\Sigma(T_{C1k})=1$, k=0,1,...p	*éléments de confusion*
2:[3,5[5	0%	2/5=40%	2/5=40%	1/5=20%	0%	$\Sigma(T_{Cik})=1$, k=0,1,...p	
3:[5,10[5	0%	1/5=20%	3/5=60%	1/5=20%	0%	$\Sigma(T_{Cjk})=1$, k=0,1,...p	
4:≥10 2	0%	0%	0%	1/2=50%	1/2=50%	$\Sigma(T_{Cpk})=1$, k=0,1,...p	
Non déterminée 0	0%	0%	0%	0%	0%	**excédents**	
	éléments de confusion				***déficits***	**accord**	

Figure 19 : matrice de confusion sur les valeurs de l'attribut hauteur de la classe arbre

• Pour l'attribut état de la classe route qui est de nature qualitative énumérée, les résultats sont représentés sur la figure 20 :

Matrice de confusion pour l'attribut qualitatif énuméré :
Route.Etat
Les taux sont exprimés en pourcentage par rapport au nombre d'objets

jeu de données / terrain nominal	**revêtu 2**	**non revêtue 1**	**en construction 0**		
revêtue 2	2/2=100%	0%	0%	$\Sigma(T_{C1k})=1$, k=0,1,...p	*éléments de confusion*
non revêtue 0	0%	0/0=100%	0%	$\Sigma(T_{Cik})=1$, k=0,1,...p	
en construction 1	0%	1/1=100%	0%	$\Sigma(T_{Cjk})=1$, k=0,1,...p	
	éléments de confusion			**accord**	

Figure 20 : matrice de confusion sur les valeurs de l'attribut état de la classe route

Dans cette matrice la valeur indéterminée n'apparaît pas puisqu'elle est interdite par les contraintes logiques.

• Pour l'attribut nom de la classe maison qui est de nature qualitative non énumérée, les résultats sont présentés sur la matrice d'absence de la figure 21 :

Matrice d'absence pour l'attribut : **Maison.Nom**
Taux exprimés en pourcentage par rapport au nombre d'objets : 7 objets

jeu de données / terrain nominal	non déterminée	déterminée
non déterminée	1/7=14,3%	1/7=14,3%
déterminée	1/7=14,3%	accord : 2/7=28,5% 4/7=57,1% désaccord : 2/7=28,6%

Figure 21 : matrice d'absence pour l'attribut nom de la classe maison

• Pour l'attribut nombre d'habitants de la classe maison qui est de nature quantitative :

- la moyenne des erreurs est -2/7 habitant,

- l'erreur moyenne quadratique est : $\sqrt{\frac{8}{7}} = 1{,}07$ habitants,

- la taille de l'échantillon est de 7 objets de la classe maison,

- le taux de rejet est 0.

3) Remarque

Dans la partie "Précision sémantique et exhaustivité" plusieurs notes indiquent que les différents taux que l'on chiffre peuvent être calculés pour des regroupements de classes ou des parties de classe définies par un critère de sélection. La suite de cet exemple illustre ce dernier point.

La figure 22 représente la matrice de confusion sur la classification des objets de la classe arbre répartie suivant les valeurs de l'attribut hauteur d'un arbre. Cette matrice se distingue de la matrice de confusion sur les valeurs de l'attribut hauteur de la classe arbre représentée par la figure 20. En effet contrairement à cette dernière, les taux sont calculés par rapport à tous les objets de la classe arbre du terrain nominal (ou tous les objets de la classe arbre du jeu de données en ce qui concerne les taux d'excédent), et pas par rapport à l'ensemble des objets ayant un homologue dans la classe arbre. On trouve dans cette matrice les taux de déficit et d'excédent en arbre par valeur de l'attribut hauteur et non plus les taux de déficit et d'excédent sur la valeur de l'attribut hauteur.
De même à l'intersection de la ligne des objets du terrain nominal ayant une valeur indéterminée pour l'attribut hauteur et de la colonne des objets du jeu de données ayant une valeur indéterminée pour l'attribut hauteur, ne se trouve pas le taux d'absence sur l'attribut hauteur (qui vaut 0), mais le taux d'accord de l'ensemble des objets de valeur indéterminée pour l'attribut hauteur (qui vaut 1 car il n'y a pas d'arbre de hauteur indéterminée dans le terrain nominal).

Matrice de confusion pour la classification des objets
Les taux sont exprimés par rapport au: nombre d'objets

jeu de données / terrain nominal	1:[1,3[1	2:[3,5[4	3:[5,10[7	4:≥10 3	Non déterminée 1	Néant		
1:[1,3[1	1/1=100%	0%	0%	0%	0%	0%	Σ=1	*éléments de confusion*
2:[3,5[6	0%	2/6=33,3%	2/6=33,3%	1/6=16,6%	0%	1/6=16,6%	Σ=1	
3:[5,10[5	0%	1/5=20%	3/5=60%	1/5=20%	0%	0%	Σ=1	
4:≥10 2	0%	0%	0%	1/2=50%	1/2=50%	0%	Σ=1	
Non déterminée 0	0%	%0	0%	0%	0/0=100%	0%	Σ=1	
Néant	0%	1/4=25%	2/7=28,5%	0%	0%		excédents	
	éléments de confusion					déficits	accord	

Figure 22 : matrice de confusion pour la classe arbre répartie selon les valeurs de l'attribut hauteur.

Bibliographie

[Abbas, 94] : “Base de données vectorielles et erreur cartographique : problèmes posés par le contrôle ponctuel; une méthode alternative fondée sur la distance de Hausdorff : le contrôle linéaire”, Thèse de doctorat en informatique, 1994.

[Abbas, 94] I. Abbas, P. Grussenmeyer, P. Hottier, “Contrôle planimétrique d’une base de données vectorielles : une nouvelle méthode basée sur la distance de Hausdorff : la méthode du contrôle linéaire”, Séminaire qualité de l’interprétation des images de télédétection pour la cartographie, Grignon (France), septembre 1994.

[CEN/TC287/prEN287008] : “Geographic information - Data description - Quality”, projet de norme européenne prEN 287008 du TC 287 du CEN, 54 pages, juillet 96 (traduit en français).

[CEN/TC278/N356] : “First draft prENV - Geographic Road Database - GDF for road traffic and transport telematics”, document N356 du groupe de travail numéro 7 du TC 278 du CEN, novembre 1994.

[ISO/TC211/WG3/N13] : "Geographic Information - Quality ", document provisoire du groupe de travail 3 du TC 211 de l'ISO, février 1996.

[Veillet, David, 95] : I. Veillet, B. David, "Description de la qualité d'une base de données géographique" Bases de données et systèmes d'information pour l'environnement, juin 1995, Versailles (France), pages 115-130, INRIA / Ministère de l'environnement.

[Le Men, Jamet, 94] : "Evaluation de la qualité d'interprétation d'images SPOT en occupation du sol" - Séminaire qualité de l'interprétation des images de télédétection pour la cartographie, Grignon (France).

[Veillet, Leconte, 94] I. Veillet, G. Leconte, “Evaluation de la qualité de la BDTOPO : l’approche IGN”, Séminaire qualité de l’interprétation des images de télédétection pour la cartographie, Grignon (France), septembre 1994.

Index

Les numéros de page sont ceux des pages où le terme est défini. Si, dans cet index, pour un même terme existent plusieurs numéros de page, c'est que le terme présente diverses définitions suivant le contexte. Par exemple, le "taux de déficit" est défini pour la classification des objets, la codification des attributs et les relations.

T

U

V

Directeur de la Publication : Jacques Poulain
Rédacteur en Chef : Serge Motet

Imprimé à l'Institut Géographique National

136 bis rue de Grenelle
F-75700 Paris 07 SP

Tél. 01.43.98.80.00

Imprimerie de l'Institut Géographique National
Dépot légal 2ème trimestre 1997
N° d'édition : 177 - n° d'impression : 198


INSTITUT GEOGRA
PHIQUE NATIONAL
I
G
N