Cerema

QUALIFIER LES DONNEES GEOGRAPHIQUES

Un décryptage de la norme ISO 19157

Critère Précision Thématique

La qualité des données est un frein avéré aux démarches d'ouverture des données. A contrario, connaître le degré de qualité de données détermine le degré de confiance qui peut leur être accordé et incite davantage à leur réutilisation.

L'essor des données ouvertes et géolocalisés et la profusion d'usages existant et à venir nous rend tous progressivement producteur et utilisateur de données géographiques. Cette évolution ne doit pas obérer la nécessaire adaptation au besoin.
Les activités régaliennes ou les politiques publiques s'appuient sur de l'information maîtrisée où la qualité des données produites ou utilisées devient un entrant indispensable. Pour autant, tout le monde ne dispose pas des moyens des producteurs institutionnels de données et il parait utile de fournir des recommandations et des méthodes plus adaptées au contexte de chacun, pour qualifier les données géographiques, communiquer sur les résultats obtenus voire savoir les interpréter. C'est l'objectif que s'est fixé le Cerema en proposant cette collection de fiches, à l'interface des productions et des usages.
Cette fiche propose d'expliciter un des critères qualité de la norme ISO19157 retenu dans cette collection de fiches, *la précision thématique*. Elle s'attache à qualifier le contenu des attributs portés par les objets ainsi que la justesse de classement des objets dans leur famille de rattachement. C'est un critère qualité important chaque fois que l'exploitation des données dépasse le simple fait d'identifier la seule présence d'un objet sur le territoire, que ce soit pour des représentations cartographiques ou pour des exploitations statistiques par exemple.
Il est donc nécessaire de maîtriser l'évaluation de la précision thématique et de connaître les indicateurs permettant de partager cette évaluation entre producteurs et utilisateurs.

1. Définition de la norme

La précision thématique se définit comme la précision des attributs quantitatifs, et la justesse des attributs non quantitatifs, du classement des entités et de leurs relations.

Elle se compose de trois éléments de qualité des données :

- **la justesse du classement** : comparaison des classes attribuées aux entités ou à leurs attributs par rapport à l'univers du discours (par exemple, monde réel ou données de référence) ;
- **la justesse des attributs non quantitatifs** : mesure permettant d'établir si

les valeurs d'un attribut non quantitatif sont correctes ou pas ;

- **la précision des attributs quantitatifs** : proximité de la valeur d'un attribut par rapport à la valeur vraie ou reconnue comme vraie.

Remarque : les attributs de datation font l'objet d'un autre critère de qualité (le critère « qualité temporelle»). Pour autant, la qualification des attributs de datation et plus particulièrement celle évaluant l'exactitude des mesures temporelles étant similaire à celles des attributs quantitatifs, elle sera traitée simultanément. Ainsi, les résultats obtenus pour cet élément de qualité intégreront l'ensemble de la qualification de la précision thématique.

2. Description des mesures possibles à réaliser

Les mesures diffèrent en fonction de l'élément de qualité des données à évaluer. Elles sont nécessairement disjointes les unes des autres.

2.1. Justesse du classement

Les mesures décrites dans ce paragraphe concourent toutes à établir la bonne classification des objets. Cela a du sens quand, dans un jeu de données, des confusions sont possibles entre classes d'objets (cette notion de classe d'objets peut être assimilée à une table ou une couche de données).

Exemple : soit la présence de 2 classes d'objets pour les bâtiments, l'une représentant les habitations, la seconde les bâtiments à but industriel ou commercial. L'affectation d'objets à tort dans l'une des deux classes sera évaluée par les mesures ci-dessous.

Par contre, dans le cas où il n'existe qu'une seule classe d'objets des bâtiments avec un attribut caractérisant la destination, l'évaluation de « la mauvaise classification » relèvera de l'élément « Justesse des attributs non quantitatifs » décrit au § 2.2.

Remarque : une erreur de justesse de classement ne doit pas être comptabilisée par ailleurs en excédent ou omission.

Pour évaluer la justesse du classement on pourra recourir aux mesures suivantes.

Nombre d'entités classées de manière incorrecte

C'est une mesure brute simple caractérisée par :

- **définition** : nombre d'entités classées de manière erronée
- **description** : comptage simple des erreurs
- **type de valeur** : nombre entier

Taux de classement erroné

C'est une mesure relative caractérisée ainsi :

- **définition** : nombre d'entités classées de manière incorrecte par rapport au nombre d'entités devant figurer dans la classe
- **description** : comptage simple des erreurs ramené à l'effectif contrôlé (base entière ou échantillon)
- **type de valeur** : nombre réel (taux)

Taux global de bon classement ou taux d'accord global

C'est une mesure relative caractérisée ainsi :

- **définition** : nombre d'entités correctement classées par rapport au nombre total d'entités contrôlées pour l'ensemble des classes.
- **description** : comptage du nombre d'entités sur la diagonale de la matrice (MCM) ramené à l'effectif contrôlé (base entière ou échantillon)
- **type de valeur** : nombre réel (taux)

Remarque : Cette mesure n'est pas présente dans la norme mais est ajoutée car elle représente l'indicateur synthétique le plus simple et le plus robuste.

Matrice de classement erroné (MCM)

Appelée également "Matrice de confusion", c'est une mesure brute plus élaborée qui « affine » l'approche par ***Nombre d'entités classées de manière incorrecte*** caractérisée ainsi :

- **définition** : matrice indiquant le nombre d'éléments de la classe i classés sous la classe j.
- **description** : la matrice de classement erroné est une matrice quadratique comportant autant de lignes et de colonnes que de classes contrôlées. Chaque cellule (i,j) de la matrice contient le nombre d'objets de la classe i classées dans la classe j.

MCM(i,j)= [nombre d'éléments de la classe i classés dans la classe j]

Ainsi, les éléments de la diagonale contiennent le nombre d'objets correctement classés pour chaque classe, et les éléments extérieurs à la diagonale contiennent le nombre d'erreurs de classement (confusions entre deux classes).

- **type de valeur** : nombres entiers (dénombrement)

Réalité terrain (classe i)	Jeu de données (résultat du classement : classe j)			
	Habitation	Bâti industriel ou commercial	Bâti remarquable	Total
Habitation	2152	25	1	2178
Bâti industriel ou commercial	185	651	5	841
Bâti remarquable	2	13	154	169
Total	2339	689	160	3188

Exemple de matrice de classement erroné des entités

Matrice relative de classement erroné (RMCM)

Deux matrices "relatives" également appelées "Matrices de confusion relatives" découlent de la matrice originale (MCM). Elles présentent en pourcentage les bons classements et les erreurs (selon le dénominateur pris en compte, soit le total des entités, en ligne ou en colonne). Cette mesure est élaborée ainsi :

- **définition** : matrice indiquant le nombre d'éléments de la classe i classés sous la classe j, divisé soit (1) par le nombre d'éléments de la classe i, soit (2) par le nombre d'éléments de la classe j, exprimé en pourcentages.
- **description** : ces deux RMCM, et les indicateurs qui en sont déduits, sont complémentaires.
  (1) met en relief la précision « producteur » pour chaque classe au regard des omissions (ou déficits)
  (2) caractérise la précision « utilisateur » pour chaque classe au regard des commissions (ou excédents).

RMCM1(i,j)= [nombre d'éléments de la classe i classés sous la classe j]/[nombre d'éléments de la classe i] 100*

RMCM2(i,j)= [nombre d'éléments de la classe i classés sous la classe j]/[nombre d'éléments de la classe j] 100*

- **type de valeur** : nombre réel (taux)

Sur la base de la MCM présentée précédemment, les deux RMCM apparaissent ainsi :

réalité terrain (classe i)	jeu de données (résultat du classement = classe j)			total :		Précision "producteur"	OMISSIONS : taux de déficit
	habitation	bâti industriel ou commercial	bâti remarquable				
habitation	**98,81%**	1,15%	0,05%	100	→	**98,81%**	1,19%
bâti industriel ou commercial	22,00%	**77,41%**	0,59%	100	→	**77,41%**	22,59%
bâti remarquable	1,18%	7,69%	**91,12%**	100	→	**91,12%**	8,88%
total :	121,99%	86,25%	91,76%				

Exemple de matrice relative "producteur" de classement erroné des entités

réalité terrain (classe i)	jeu de données (résultat du classement = classe j)			total :
	habitation	bâti industriel ou commercial	bâti remarquable	
habitation	**92,01%**	3,63%	0,63%	96,26%
bâti industriel ou commercial	7,91%	**94,48%**	3,13%	105,52%
bâti remarquable	0,09%	1,89%	**96,25%**	98,22%
total :	100	100	100	
	↓	↓	↓	
Précision "utilisateur"	92,01%	94,48%	96,25%	
COMMISSIONS : taux d'excédent	7,99%	5,52%	3,75%	

Exemple de matrice relative "utilisateur" de classement erroné des entités

Coefficient kappa

C'est une mesure relative élaborée caractérisée ainsi :

- **définition** : coefficient destiné à mesurer le degré d'accord d'affectation aux classes en supprimant la part de hasard et de subjectivité.
- **description** : Il s'agit d'une formule permettant de réduire l'analyse d'une matrice de classement erroné à un seul chiffre sans pondération de gravité entre les erreurs de classement.

$$k = \frac{N * \sum_{i=j}^{r} MCM(i,i) - \sum_{i=1}^{r} \left(\sum_{j=1}^{r} MCM(i,j) * \sum_{j=1}^{r} MCM(j,i) \right)}{N^2 - \sum_{i=1}^{r} \left(\sum_{j=1}^{r} MCM(i,j) * \sum_{j=1}^{r} MCM(j,i) \right)}$$

Où r est le nombre de classe considérées. La formule peut effrayer mais elle se calcule facilement, le plus complexe et le plus long étant de remplir les cellules de la matrice de classement erroné.

- **type de valeur** : nombre réel.

Le coefficient kappa prend une valeur comprise entre -1 et 1. Plus la valeur du kappa est élevée, plus la concordance est forte :

- La valeur 1 indique un classement parfait
- La valeur 0 exprime que le classement est aléatoire
- La valeur -1 correspond au cas où tous les classements sont erronés.

A titre d'illustration, dans l'exemple ci-dessus de la matrice de classement erroné portant sur la justesse de classement entre classes de bâtiments, le coefficient kappa obtenu est de 0.83.

Il est utile de préciser que ce coefficient kappa n'est pas toujours un indicateur approprié pour mesurer la justesse d'un classement : en effet, dans certaines situations, son comportement est imprévisible, celui-ci se montrant notamment plus adapté lorsque les classes sont relativement indépendantes les unes des autres (soit avec un risque limité de confusion entre les classes). Par ailleurs, cet indicateur "réagit" mal quand une (ou plusieurs) classe(s) est(sont) peu représentée(s) dans le lot de données.

2.2. Justesse des attributs non quantitatifs

Il s'agit ici d'évaluer la qualité des attributs non quantitatifs – aussi dénommés attributs qualitatifs. On entend par attribut qualitatif tout attribut qui peut prendre une valeur dans une liste, éventuellement très grande, et sur lequel on n'effectue pas de calcul (somme, moyenne) par opposition aux attributs quantitatifs. Ce sont les attributs qui permettent de définir une catégorisation (type d'occupation du sol, classement administratif, toute codification, les noms, etc.).

Nombre de valeurs d'attribut incorrectes

- **définition** : nombre total de valeurs d'attribut erronées dans la partie concernée du jeu de données
- **description** : comptage de toutes les valeurs d'attribut quand la valeur est incorrecte
- **type de valeur** : nombre entier

Taux de valeurs d'attribut incorrectes

- **définition** : nombre de valeurs d'attributs signalées comme étant incorrectes par rapport au nombre total de valeurs.
- **type de valeur** : nombre réel (taux)

Taux de valeurs d'attribut correctes

- **définition** : nombre de valeurs d'attributs correctes par rapport au nombre total de valeurs.
- **type de valeur** : nombre réel (taux)

2.3. Exactitude des attributs quantitatifs

Les mesures décrites dans ce paragraphe s'adressent à l'ensemble des attributs caractérisant une quantité (nombre d'habitants, hauteur de pylône, surface d'une parcelle, etc.) représentée nécessairement par un nombre entier ou réel. Elles concernent également les attributs de datation comme indiqué au §.1. Ce sont des attributs sur lesquels il est envisageable d'effectuer des calculs statistiques.

On y retrouve des principes mathématiques similaires aux mesures de la précision de position[1] :

Incertitude de la valeur d'attribut avec un seuil de signification de 95 %[2]

- **définition** : demi-longueur de l'intervalle défini par une limite supérieure et inférieure, dans laquelle la valeur véritable se situe avec une probabilité de 95%
- **variantes** : plusieurs mesures sont construites sur le même principe avec des exigences croissantes : incertitude de la valeur avec un seuil de signification de 50%, 68,3 %, 90 %, 99 % ou encore 99,8 %
- **type de valeur** : nombre réel

L'incertitude est toujours exprimée dans l'unité de l'attribut. C'est également le cas pour les attributs temporels.

3. Indicateurs retenus

On retient du chapitre précédent que les mesures proposées par la norme pour qualifier la précision thématique sont nombreuses. On en retiendra les plus pertinentes, compréhensibles et faciles à mettre en œuvre.

3.1. Justesse de classement

La **matrice de classement erroné** (MCM) ou **matrice de confusion** est la mesure la plus complète. Sans demander de surcroît de travail par rapport au *« dénombrement d'entités clas-*

1 Voir la fiche n°10 « Précision de position ».

2 Ce choix de 95 % de niveau de confiance s'appuie sur les éléments statistiques *(cf. fiche méthode Eléments statistiques §6)* où la vraie valeur recherchée est située dans un intervalle de confiance centré sur la moyenne des mesures avec un intervalle de plus ou moins deux fois l'écart-type (2σ).

sées de manière incorrecte » ou au *« taux global de bon classement »*, elle donne une vision synthétique et laisse la possibilité de juger de la gravité des confusions entre classes. C'est évidemment la solution préconisée en cas d'un rapportage détaillé avec une représentation complexe de la qualité thématique d'un lot de données.

Le **taux global de bon classement** (ou **taux d'accord global**) représente l'indicateur synthétique, issu de la MCM, le plus simple et le plus robuste. Bien que globalisant l'information liée aux confusions entre classes, à condition que l'échantillon soit représentatif du lot de données, il permet d'obtenir une première approche du niveau de qualité thématique.

Le **coefficient kappa** représente un autre indicateur synthétique, et reste une possibilité de mesure de la justesse de classement, Mais n'étant pas adapté à toutes les situations, l'utilisateur doit être averti des possibilités et limites de cet indicateur.

.

Ces deux approches, taux global de bon classement (ou taux d'accord global) et coefficient kappa, présentent le mérite d'être similaires aux solutions retenues pour d'autres critères comme l'exhaustivité, tout en étant bien conscient que cette globalisation des mesures unitaires gomme les éventuelles disparités entre les classes.

Dans tous les cas, évaluer la justesse de classement ne sera pertinent que s'il existe des risques de confusion entre classes d'objets.

> *Exemple : pour la Directive Inondation, un risque de classement erroné se présente entre les enjeux d'ordre économique (un bâtiment de bureau) et ceux d'ordre culturel (bâtiment classé). Par contre, il n'y a pas de risque de confusion entre un enjeu ponctuel et une zone d'aléa.*

La norme ISO 19157 n'impose pas de limite au nombre de classes à traiter au sein d'une même matrice de classement erroné (MCM).

L'analyse d'une grande matrice, avec de très nombreuses cellules dont une grande partie présentent des valeurs nulles, n'est certes pas aisé. Toutefois, il existe des moyens d'exploiter ce type de matrice en générant les indicateurs par groupe de plusieurs classes, ce qui revient en quelque sorte à réduire la taille de la matrice.

Par contre, il est recommandé de générer les matrices de classement erroné (MCM) rigoureusement selon chaque thématique, pour les seules classes où les risques de confusion sont réels, quitte à multiplier le nombre de matrices si nécessaire.

> *Exemple : dans un lot de données relatif à l'analyse multirisques sur un territoire, on utilisera une matrice relative de classement erroné pour évaluer la justesse de classement des aléas et une seconde matrice pour évaluer la justesse de classement des enjeux.*

Même s'il apparaît en premier dans la norme ISO 19157, le sous-critère *justesse de classement* n'est pas le plus important de la précision thématique, car il ne concerne qu'un nombre limité de jeux de données. Il faut en effet que ce dernier présente la particularité d'être composé d'un grand nombre de classes d'objets avec des finalités proches et des géométries similaires. L'expérience montre que ce cas de figure reste rare, les modélisations privilégiant plutôt des regroupements des classes proches avec l'ajout d'attributs caractérisant les typologies. Ainsi, la justesse des attributs non quantitatifs apparaît comme le sous-critère fondamental.

La qualification du lot de données sera indiquée :

- **prioritairement par le taux global de bon classement (ou taux d'accord global)**, qui représente l'indicateur synthétique le plus simple et le plus robuste : défini comme la somme des objets bien classés, divisée par le nombre total d'objets contrôlés sur l'ensemble des classes ;
- **secondairement, sera fournie la MCM, matrice de classement erroné ou matrice de confusion**, qui représente la mesure de base à partir de laquelle tous les indicateurs peuvent être calculés.

  Remarque : la fourniture d'un indicateur synthétique comme le taux global de bon classement, ou le coefficient kappa, ne dispense pas de fournir des matrices qui permettront, en fonction de l'usage, d'apprécier finement la justesse de classement et les classes concernées.

3.2. Justesse des attributs non quantitatifs

Les trois mesures proposées par la norme ISO 19157 demandent le même effort. L'approche

positive milite pour retenir préférentiellement le **taux de valeurs d'attribut correctes**.

Remarque : la mesure s'applique pour chaque attribut à qualifier. La norme ISO 19157 ne propose pas de mesure agrégeant l'ensemble des mesures sur les attributs d'une même classe d'objets, empêchant de fait une appréciation rapide de la précision thématique de la classe.

Il est proposé de ramener la justesse des attributs non quantitatifs d'une classe d'objets à la moyenne non pondérée des taux de valeurs d'attribut correctes pour l'ensemble des attributs évalués.

Par extension, on affectera une note globale à un lot de données constitué de plusieurs classes d'objets, en appliquant le même principe de calcul de la moyenne non pondérée des taux de valeurs d'attribut correctes pour l'ensemble des attributs de toutes les classes d'objets évaluées. Cette globalisation, comme toute moyenne, efface les disparités éventuelles entre attributs voire entre classes mais offre l'avantage de ramener à une seule note l'ensemble des mesures qualité effectuées sur les attributs non quantitatifs.

3.3. Exactitude des attributs quantitatifs

La mesure d'« **incertitude de la valeur d'attribut avec un seuil de signification** » s'impose comme étant la seule admise par la norme.
On adoptera préférentiellement le **seuil de signification de 95 %.** Ce seuil est un bon équilibre, car il permet d'avoir une confiance suffisante aux résultats tout en réduisant l'effort pour les obtenir.
Pour rappel, ce sous-critère regroupe également l'évaluation des attributs de datation.

4. Méthode de contrôle

4.1. Généralités

Mesurer la précision thématique d'un lot de données consiste à évaluer :

- la justesse de classement des objets en établissant des matrices relatives de classement erroné ;

  Rappel : La justesse de classement ne doit être mesurée qu'entre classes où existe réellement un risque de confusion amenant à une erreur de classement. Si on admet aisément de contrôler entre elles deux classes « Routes » et « chemins », il semble inutile voire contre-productif de vouloir évaluer le classement entre les classes « routes » et « bâtiments ».
- la justesse des attributs non quantitatifs en exprimant des taux de valeurs d'attribut correctes ;
- l'exactitude des attributs qualitatifs par la mesure de leur incertitude avec un seuil de signification donné.

Les méthodes employées pour qualifier la précision thématique étant proches de celles relatives à la qualification de l'exhaustivité du jeu de données il sera avantageux de mener ces contrôles en parallèle.
À l'instar de la qualification de l'exhaustivité, les méthodes diffèrent selon que l'on dispose ou pas d'un jeu de données de référence.

4.2. Existence d'une référence utilisable en tant que source de contrôle

Si l'on dispose d'un jeu de données ou d'un document de référence, on réalisera ces mesures par rapport au terrain nominal exprimé via ce jeu de données ou document de référence.
Cette référence peut prendre toute forme faisant foi dans son domaine même si elle ne provient pas d'une source publique ou officielle : base de données, tableur, listing, définition administrative, arrêté, site internet de référence, etc.
En complément des éléments de qualité, on rappellera systématiquement :

- la référence utilisée en tant que source de contrôle (producteur, date, spécifications, etc.)
- la méthode utilisée dans le cas d'un contrôle par échantillonnage, ainsi que la taille de l'échantillon et les effectifs de chaque classe de l'échantillon.

4.3. Réalisation des contrôles

Le contrôle de la précision thématique consistera en requêtes successives opérées sur le lot de données à fin de comptages comparatifs entre le jeu de données à évaluer et le jeu de données de référence.

Les comptages sont effectués par classe d'objets pour la mesure de la justesse de classement, et par attribut d'objets pour la justesse et l'exactitude des attributs.

En cas de contrôle sur un échantillon, les règles standard d'échantillonnage s'appliquent *(cf. fiche méthode échantillonnage)*.

4.4. Absence de source de contrôle

Dans le cas où aucune base de données de référence n'est connue ou accessible, il reste possible d'exploiter *(cf. la fiche Critère Exhaustivité)* : le contrôle terrain, le contrôle à dire d'expert, l'interprétation de la généalogie des données et l'exploitation des spécifications. Parmi ces méthodes, la seule permettant de fournir des éléments de qualité objectifs est bien sûr le contrôle terrain, mais il s'agit de la plus coûteuse.

5. REPRESENTATION - NOTATION

La norme ISO 19157 préconise le rapportage des sous-critères de précision thématique soit sous forme de métadonnées conformes à la norme ISO 19115, soit sous forme d'un rapport qualité autosuffisant dont la structure est laissée au libre choix du rédacteur.

En complément des éléments de qualité, on rappellera systématiquement le contexte du contrôle qualité :

- **le jeu de données de référence** utilisé le cas échéant en tant que source de contrôle et ses caractéristiques : producteur, date, spécifications, etc.
- **la méthode utilisée** dans le cas d'un contrôle par échantillonnage, ainsi que la taille de l'échantillon et les effectifs de chaque classe de l'échantillon.
- **Les objectifs de qualité** portant sur la précision thématique dans le cas où ils ont été précisés dans les spécifications de produit.

5.1. Représentation sous forme complexe

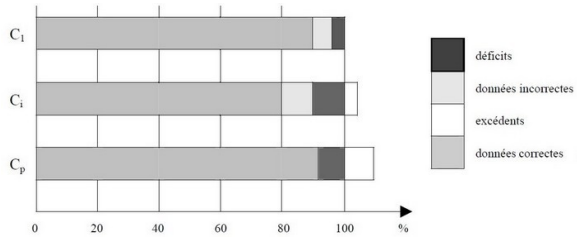
La forme complexe consiste à présenter dans le détail les contrôles et résultats obtenus. S'adressant prioritairement aux producteurs de données et aux commanditaires, elle trouve tout son intérêt quand des spécifications existent *(cf. fiche méthode « modes de représentation »)*.

Pour la justesse de classement, cela consiste à fournir les différentes matrices relatives de classement erroné. Pour les attributs non quantitatifs, il revient de communiquer les taux de valeurs d'attributs corrects pour tous les attributs évalués. Enfin, pour les attributs quantitatifs, on indiquera les mesures d'incertitude des valeurs d'attribut obtenues au seuil de signification de 95 %.

Cas particulier : dans certaines situations, le dénombrement d'objets sera avantageusement remplacé par des comptages dans une unité de mesure plus parlante (longueur, surface, etc.)

> *Exemple : il peut être pertinent d'exprimer des confusions de classification de zones d'occupation des sols plutôt en surfaces (km²) qu'en nombre d'objets.*

Il existe, pour cette forme de représentation complexe, une alternative proposée par l'IGN (voir illustration ci-après) qui présente le double mérite d'agréger la représentation des deux critères *exhaustivité* et *précision thématique* en proposant une représentation sous forme de diagramme qui reste relativement détaillée bien que compacte. Ainsi, les informations qualité du niveau sémantique sont figurées simultanément, ce qui donne une vision rapide pour chaque classe d'objets.



Dans ce diagramme, *Ci* représente une caractéristique évaluée, une classe d'objets (pour la justesse de classement), ou un attribut de classe d'objets (pour la justesse et l'exactitude).

Ce diagramme synthétique présente pour chaque classe ou attribut d'objets le pourcentage de classements d'objets (respectivement de valeurs d'attribut) corrects et incorrects, ainsi que le taux d'objets (respectivement de valeurs d'attributs) en excédent et en déficit.

5.2. Représentation sous forme simplifiée

Il est difficile de réduire l'ensemble des éléments de qualité du critère de précision thématique à une seule note du fait de l'hétérogénéité des éléments de qualité d'une part, et du fait de l'importance de ce critère d'autre part. Il est donc proposé de retenir des préconisations différentes en fonction des sous-critères qualité.

En premier lieu, se pose la question des deux sous-critères *justesse de classement* et *justesse des attributs non quantitatifs*, que l'on pourrait qualifier de « qualitatifs » et qui induisent des impacts similaires sur les exploitations des

données. Les deux apportent une information sur la catégorisation des objets qu'elle se traduise sous forme de la bonne affectation à une classe d'objets ou de la bonne sélection d'une valeur d'attribut dans une liste. C'est le choix de la modélisation qui va influencer le sous-critère à évaluer selon que le lot de données se compose de nombreuses classes d'objets présentant des confusions possibles ou, à l'inverse, peu de classes avec des attributs permettant de retrouver des sous-populations. Comme déjà évoqué, la deuxième situation est la plus fréquente et la justesse de classement sera un sous-critère rarement évalué.

Pour éviter, dans la représentation simplifiée, de mettre en avant un indicateur non noté dans la plupart des cas, il est proposé d'en définir un nouveau, hybride, qui correspondra, en fonction des lots de données, soit au seul taux de valeurs d'attribut correctes, soit à un mix avec le taux global de bon classement.

Afin d'éviter toute confusion avec les mesures préconisées par la norme, cet indicateur est intitulé **taux de classification/catégorisation**.

La note synthétique retenue est alors issue de la moyenne du taux global de bon classement et du taux de précision des attributs qualitatifs (lui-même résultant de la moyenne non pondérée des taux de valeurs d'attribut correctes pour l'ensemble des attributs évalués).

On applique le même raisonnement, que l'on évalue les attributs d'une classe d'objets ou de plusieurs.

On retient une échelle similaire à l'exhaustivité, dont les seuils ont été choisis en relation avec des exemples de situations proposés par des services utilisateurs représentant bien une réalité terrain.

Classification/catégorisation	Note sur 5
De [95 % à 100 %]	5
De [90 % à 95 %[	4
De [75 % à 90 %[	3
De [50 % à 75 %[	2
Taux < 50 %	1

Il reste une difficulté pour traiter la représentation simplifiée de **l'exactitude des attributs quantitatifs**. En effet, les informations quantitatives sont très variées et potentiellement définies dans des unités différentes, ce qui rend délicat tout « mélange » des résultats obtenus.

Par exemple, pour une classe d'objets des arbres d'alignement urbains comprenant les 2 attributs quantitatifs « âge » et « hauteur » évalués respectivement avec les incertitudes de 1,5 an ± 2,1 ans et 4,8 m ± 3,2 m, vouloir réduire la justesse des attributs quantitatifs à une seule valeur n'a pas de sens.

Il est proposé, bien que cette solution soit peu satisfaisante, de ne pas fournir, en représentation simplifiée, de note pour l'exactitude des attributs quantitatifs. On privilégiera, en parallèle de la notation, une information littérale des attributs présentant des incertitudes importantes du point de vue de celui qui évalue.

Cela pourra se faire un peu à l'image de ce qui est préconisé pour le sous-critère « cohérence conceptuelle » (*cf. fiche cohérence logique*) où l'on recommande de lister les anomalies constatées en distinguant les problèmes critiques (erreur rédhibitoire) des avertissements (uniquement portés à titre indicatif). Comme pour tous les critères, il importe de commenter chaque note attribuée en fournissant la description de la méthode utilisée.

6. Ce qu'il faut retenir

Le critère de précision thématique est très important. Il est dans la plupart des cas nécessaire de le contrôler. La directive Inspire recommande son utilisation dans 6 thèmes sur les 34 référencés. La norme propose trois sous-critères et neuf mesures. Nous en retiendrons une seule par sous-critère :

- **la justesse de classement** des objets exprimée dans un taux global de bon classement (ou taux d'accord global), indicateur synthétique qui sera accompagné de la matrice de classement erroné (ou matrice de confusion) ;
- **l'exactitude des attributs quantitatifs** exprimée en mesurant leur incertitude avec un seuil de signification ;

- **la justesse des attributs non quantitatifs** exprimée par des taux de valeurs d'attribut correctes.

Les contrôles portant sur l'aspect « sémantique » du lot de données (exhaustivité et précision thématique) étant assez proches dans leur méthodologie, leur réalisation et leur rapportage, il est recommandé de les mener en parallèle.

En présence de données de référence, ce critère peut être très simple à mesurer. En revanche, en leur absence, son évaluation devient plus complexe et difficile. On aura alors recours aux différentes méthodes préconisées comme le contrôle terrain et le recours aux avis d'experts en prenant soin de toujours évaluer le taux d'incertitude qui s'attache à une telle évaluation.

La représentation simplifiée sera limitée aux deux premiers sous-critères dont la réduction au travers d'une seule note garde du sens. L'information sur l'exactitude des attributs quantitatifs se fera sous forme littérale si elle s'avère nécessaire.

Enfin, la note affectée à ce critère devra toujours être accompagnée des éléments de méthode utilisés pour sa détermination.

Fiche réalisée sous la coordination de Gilles Troispoux et Bernard Allouche (Cerema Territoires et ville)

Rédacteurs
Yves Bonin (Cerema Méditerranée), Arnauld Gallais (Cerema Ouest), Benoît Ségala (consultant)

Contributeurs
Mathieu Rajerison, Silvio Rousic, (Cerema Méditerranée)

Relecteurs
Benoît David (Mission information géographique MEEM/CGDD), Stéphane Rolle (CRIGe PACA), Magali Carnino (DGAC), Stéphane Lévêque (Cerema Territoires et ville), Yvan Bédard (Professeur Honoraire à l'université Laval, CEO d' Intelli[3])

Contacts
Bernard.Allouche@cerema.fr
Boutique en ligne : catalogue.territoires-ville.cerema.fr