Cerema

Qualifier les données géographiques
Un décryptage de la norme ISO 19157

Méthodes d'échantillonnage

La connaissance de la qualité des données, en sécurisant l'utilisateur, incite davantage à leur réutilisation.

Ce décryptage de la norme ISO 19157 a pour vocation de donner un cadre méthodologique pour qualifier les données lors de leur diffusion.

L'essor des données ouvertes et géolocalisées et la profusion d'usages existants et à venir nous rend tous progressivement producteur et utilisateur de données géographiques.

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 explicite les recommandations de la norme ISO 19157 pour constituer des échantillons en vue du contrôle qualité des données géographiques, ainsi que différentes méthodes d'échantillonnage spatial des données géographiques. L'approche concrète des producteurs et utilisateurs de données géographiques en sera facilitée.

Le lecteur se reportera à la fiche n° 4 « Éléments statistiques » détaillant les méthodes statistiques pour évaluer la taille significative d'un échantillon en fonction de la taille de la population à contrôler, et les valeurs de rejets en fonction d'une Limite d'Acceptation de la Qualité (LAQ).

1. Les définitions utilisées

L'annexe H de la norme ISO 19157 présente les recommandations pour définir des échantillons ainsi que des méthodes de qualification des données géographiques par échantillonnage.

Elle s'appuie sur la série de normes ISO 2859 et l'ISO 3951-1:2005 conçues pour une utilisation d'échantillons pour des contrôles de données, et elle présente des techniques d'échantillonnage spatial spécifiques aux données géographiques.

Dans la suite, les termes suivants, définis par la norme, seront utilisés.

Lot : il s'agit du lot de données à évaluer, pouvant par exemple être défini par une ou plusieurs thématiques géographiques sur un territoire.

Objet : unité minimale à contrôler

> **Exemple :** les objets « bâtiments » correspondant aux spécifications du produit seront contrôlés.

Strate : sous-emprise géographique du lot de données à qualifier présentant des caractères d'homogénéité la distinguant des autres strates. Les strates ne se chevauchent pas et peuvent constituer une partition complète du territoire. On constitue ainsi une partition géographique du lot de données à qualifier, suivant des caractéristiques homogènes.

> **Exemple :** « zone rurale » « zone urbaine ».

Échantillon : sous-ensemble représentatif du lot de données permettant de le qualifier avec un certain niveau de confiance.

2. Méthodes d'échantillonnage

Il est nécessaire de procéder à un échantillonnage lorsque l'évaluation exhaustive d'un lot de donnée est trop coûteux, en temps ou en ressources, ou matériellement impossible.

L'échantillon à constituer doit être suffisamment représentatif de la population d'objets géographiques du lot de données à qualifier.

Pour ce faire, on s'appuie sur différents **critères** et on définit une **stratégie d'échantillonnage**.

2.1 Les critères d'échantillonnage

Les principaux critères permettant de définir et assurer la représentativité d'un échantillon sont :

■ Le nombre d'objets

La taille de l'échantillon est exprimée généralement en pourcentage du nombre total d'objets pour un type donné.

> **Exemple :** l'échantillon contiendra 20 % du total des objets « ponts ».

■ La surface couverte

La taille de l'échantillon est exprimée en pourcentage de la surface totale du lot de données, notamment dans le cas d'objets surfaciques.

> **Exemple :** « l'échantillon couvrira une surface égale à 10 % de l'emprise du lot de données ».

■ L'emplacement

Les échantillons doivent être judicieusement répartis dans l'espace afin de représenter l'ensemble des objets et des strates existants dans le jeu de données.

La répartition des échantillons dans l'espace du lot de données peut se réaliser suivant différentes stratégies d'échantillonnage.

2.2 Stratégies d'échantillonnage

La stratégie d'échantillonnage des données géographiques repose sur deux aspects :

- la détermination des échantillons suivant un **effectif par type d'objet** ou suivant une **approche surfacique**. On peut également envisager une combinaison des deux.

> **Exemple :** 10 % des bâtiments », « Les bâtiments contenus dans N périmètres de 1 km² », « 100 bâtiments sélectionnés dans N périmètres de 1 km² ».

- **le caractère probabiliste plus ou moins aléatoire** avec lequel ces entités sont sélectionnées.

Pour ce second aspect, on admet une variété de stratégies de sélection allant d'un tirage purement aléatoire à une sélection au jugement de l'opérateur basée sur son expertise, en passant par des possibilités intermédiaires :

- purement aléatoire : les objets ou surfaces d'échantillonnage sont choisis aléatoirement ;
- semi-aléatoire (voir ci-dessous) ;
- aléatoire stratifié (voir ci-dessous) ;
- « au jugé » de l'opérateur : les objets ou surfaces d'échantillonnage sont choisis par l'expert qualité en se basant sur un savoir-faire spécialisé ou sur un jugement professionnel.

Les trois premières de ces stratégies de sélection sont dites probabilistes par le fait que la sélection des entités en vue de la constitution d'un échantillon répond à une probabilité déterminée.

■ Échantillonnage orienté entité

L'échantillonnage orienté entité sélectionne les objets de l'échantillon en se basant sur leurs attributs non spatiaux, donc sans tenir compte de leur localisation dans l'espace.

Exemple 1 : « toutes les zones 'pâturage' de la base occupation du sol ».

Exemple 2 : « toutes les routes départementales de plus de deux voies ».

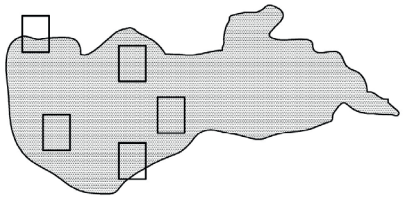
Un échantillon peut ainsi être déterminé en supposant que l'ensemble du lot de données présente une répartition spatiale homogène de ses caractéristiques.

■ Échantillonnage orienté surface

À l'inverse de la précédente, une stratégie d'échantillonnage orienté surface entraîne la sélection des objets échantillonnés uniquement suivant des considérations spatiales. Les unités d'échantillonnage peuvent être des surfaces géographiques existantes (Exemple : des emprises communales) ou de toute autre nature, ou bien choisies aléatoirement. Elles ne doivent pas s'intersecter, pour ne pas comptabiliser plusieurs fois les mêmes objets, et donc les éventuels caractères de non qualité qu'ils portent.

Exemple 1 : « tous les objets intégralement contenus dans les zones naturelles d'intérêt écologique, faunistique et floristique (ZNIEFF) afin d'évaluer leurs attributs ».

Exemple 2 : unités d'échantillonnage surfaciques géométriques réparties aléatoirement :



■ Échantillonnage orienté surface et entité

Il s'agit d'une combinaison des deux précédentes stratégies d'échantillonnage. L'échantillonnage orienté surface est suivi d'un échantillonnage orienté entité dans chaque subdivision surfacique.

Exemple : « toutes les zones 'pâturage' de la base OCS, appartenant à une ZNIEFF ».

■ Échantillonnage simple aléatoire

Les objets ou surfaces d'échantillonnage sont choisis aléatoirement. Toutes les sélections possibles d'objets présentent ainsi la même probabilité.

Cette méthode d'échantillonnage est pertinente lorsque la population est géographiquement homogène, c'est-à-dire sans tendance spatiale ou thématique importante.

Exemple : « sur 1574 objets ponts, on en sélectionne aléatoirement 100 pour contrôler qu'ils sont exactement positionnés au croisement de deux objets (route, voie ferrée, cours d'eau) ».

Cette méthode souffre du défaut que l'échantillon sélectionné peut ne représenter qu'une fraction de l'emprise spatiale du lot de données.

Exemple : « ayant été choisis aléatoirement 80 % des objets se situent dans la zone sud-ouest, les autres zones sont sous-représentées ».

Dans ce cas, un échantillonnage semi-aléatoire ou stratifié donnera de meilleurs résultats.

■ Échantillonnage semi-aléatoire

Il s'agit d'un échantillonnage aléatoire **guidé par un systématisme** pour tout ou partie de l'échantillon, et des règles de sélection pour l'éventuel complément de l'effectif de l'échantillon.

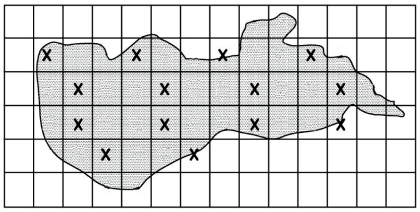
Exemple 1 : « on constitue l'échantillon en sélectionnant un objet tous les dix objets » ou « on sélectionne une dalle toutes les cinq dalles de 1 km² en coordonnées Lambert 93 »

Exemple 2 : un prestataire de topographie doit évaluer la justesse des attributs d'objets géographiques de différents types appartenant au plan corps de rue simplifié (PCRS). Il contrôle un échantillon d'objets sélectionnés **aléatoirement**, et contrôle des objets de différents types situés à **proximité** du premier échantillon, de façon à atteindre dans chaque type l'effectif d'objets à contrôler. Le contrôle à proximité [présentant un caractère semi-aléatoire] permet une réduction des coûts de déplacement et de contrôle sur le terrain.

L'échantillonnage en grille constitue un autre exemple d'échantillonnage semi-aléatoire. La position initiale d'une grille est déterminée de manière aléatoire et les échantillons sont constitués à des intervalles régulièrement espacés dans l'espace, suivant les cellules de la grille.

Cette méthode est pratique et facile à mettre en œuvre pour couvrir intégralement l'emprise du lot de données à contrôler.

Exemple 3 : l'échantillon est constitué en retenant tous les objets contenus dans une dalle sur trois :



Exemple 4 : l'échantillon est constitué de 10 communes choisies aléatoirement dans chaque département.

■ Échantillonnage aléatoire stratifié

L'échantillonnage stratifié nécessite que la population d'objets soit répartie en sous-ensembles géographiques indépendants. Une strate présente des caractères d'homogénéité la distinguant des autres strates. Les strates ne se chevauchent pas et peuvent constituer une partition complète du territoire.

Exemple 1 : différents types de paysages peuvent constituer les strates d'un lot de données. Pour le contrôle des adresses postales, on choisit de distinguer et délimiter les milieux « agglomération », « ville », « péri-urbain », « village », « rural », en adaptant les contrôles à chacune de ces situations.

Exemple 2 : au sein d'un PLU intercommunal, on évaluera séparément la précision géométrique des zonages d'urbanisme suivant qu'ils aient été numérisés sur fond cadastral PCI vecteur, BDParcellaire, ou BDParcellaire image.

Cette stratégie d'échantillonnage fournira pour chaque strate une meilleure exactitude des estimations statistiques : moyenne, variance, etc.

3. Considérations statistiques

3.1 Ressources

Les tableaux figurant dans la fiche Méthode « Éléments statistiques » fournissent les informations relatives à l'effectif minimal de l'échantillon selon :

- **la taille du jeu de données** ;
- **le niveau de rejet associé** quand une ambition de qualité a été précisée dans les spécifications, appelée « limite d'acceptation de la qualité » (LAQ) ;
- **le niveau de confiance recherché** correspondant à des valeurs généralement comprises entre 95 % et 99 % de confiance.

Ces méthodes s'appliquent soit à la recherche d'éléments conformes/non-conformes (cas des critères exhaustivité et précision thématique), soit dans un contrôle qualité faisant intervenir un écart-type (cas du critère précision de position).

3.2 Mise en garde

En cas de non-respect des règles de taille et représentativité d'un échantillon, les taux mesurés d'éléments manquants dans le(s) échantillon(s) ne peuvent pas être comparés directement à la limite d'acceptation de la qualité (LAQ) pour l'ensemble du lot de données. (cf. la fiche n 4 « Éléments statistiques »).

4. Choisir son échantillonnage

4.1 Processus théorique

Ce processus repose sur cinq étapes :

❶ Définir les objets et/ou thématiques et/ou emprises à contrôler en fonction des spécifications de produit et des exigences de qualité, par exemple en fonction des limites d'acceptation de la qualité spécifiées, qui permettent de définir la taille de l'échantillon.

❷ S'il n'est pas homogène, découper le lot de données en sous-lots homogènes, quand c'est possible.
Un lot de données est supposé homogène sur le plan de sa qualité lorsque ces trois conditions sont réunies :

- les données alimentant le processus de production étaient de qualité homogène ;

Exemple : tout le lot a été saisi sur fond orthophotographique, toutes les images avaient la même résolution terrain, il n'y a pas eu de saisie sur fond cartographique de moindre précision.

- les systèmes de production (matériels, logiciels, compétences des opérateurs) étaient constants ;

Exemple : toutes images aériennes sont issues de la même caméra, les orthophotographies étaient réalisées suivant le même processus, avec la même version de spécifications.

- les causes potentielles de non-conformités étaient constantes pour l'ensemble du lot de données.

❸ Constituer les échantillons, en suivant les stratégies décrites au paragraphe 2.2 précédent.

Le tableau reprend certains cas fréquents de manière synthétique.

Caractéristiques du lot de données	Stratégie conseillée
Lot de données présentant une répartition homogène dans l'espace	Échantillonnage aléatoire simple
Lot de données étendu dont on souhaite couvrir l'intégralité	Échantillonnage semi-aléatoire
Lot de données hétérogène, à diviser en strates homogènes	Échantillonnage aléatoire stratifié

❹ Effectuer un tirage aléatoire des échantillons à contrôler (si leur nombre le justifie...).

❺ Contrôler tous les éléments des échantillons sélectionnés selon les mesures choisies et au regard des spécifications qualité (lorsqu'elles existent).

4.2 Recommandations diverses

Les « non-conformités en cascade » sont traitées comme un phénomène unique.
Au cas où des objets présentent un systématisme de non-conformité, il convient de les regrouper et les considérer comme un phénomène unique provoquant un défaut systématique.

Exemple : si « une portion de route est codée en chemin » et «si le chemin permet une circulation à 90km/h » : il s'agit d'une seule et même anomalie liée à l'erreur de transcodage de la route en chemin.

Lorsque la méthode d'échantillonnage fait intervenir des surfaces géographiques, il convient de définir les règles à propos de l'inclusion des objets partiellement contenus dans ces surfaces. Exemple : si on sélectionne les objets « intersectant la surface », il conviendra pour la justesse du critère d'exhaustivité de s'assurer qu'ils ne seront pas également sélectionnés dans une autre surface d'échantillonnage.

Ce qu'il faut retenir

Lorsque l'effectif est peu important et les exigences qualité particulièrement élevées, on cherchera à contrôler toute la population d'objets géographiques.

- **Exemple** : La collectivité territoriale sera particulièrement vigilante à contrôler tous les zonages d'urbanisme figurant sur le PLU dématérialisé avant toute publication sur le Géoportail de l'urbanisme, entraînant des effets juridiques.

Hormis ce cas de figure particulier, la qualification des données géographiques procède classiquement par échantillonnage, en particulier pour des raisons de coût.

Compte-tenu de sa caractéristique spatiale, contrairement aux pratiques d'échantillonnage de l'industrie suivant les normes en vigueur, la qualification des données géographiques ne se satisfait pas pleinement des tirages aléatoires.

En effet, les sélections de zones aléatoires présentent le risque que certains types de zones géographiques échappent au contrôle.

En dehors des échantillons constitués de façon semi-aléatoire pour couvrir une large emprise suivant un systématisme défini, on identifiera préalablement les éventuelles strates du lot de données : ce sont des sous-populations homogènes, distinctes et géographiquement réparties.

On établira ainsi une typologie et des échantillons représentatifs des données géographiques au regard des critères à qualifier.

Les résultats des tests de qualification seront ensuite redressés suivant les proportions relatives des strates.

- **En effet** : 95 % des objets d'une strate représentant 10 % du lot de données ne représentent que 9,5 % de l'effectif total.

Enfin, en cas d'utilisation d'une méthode d'échantillonnage pour estimer la qualification d'un lot de données, on ne peut pas extrapoler directement les résultats de la qualité de l'échantillon à l'intégralité du jeu de données.

Par exemple, le taux estimé d'éléments manquants dans les échantillons ne peut pas être comparé directement à la limite d'acceptation de la qualité pour l'ensemble du lot de données avec un certain niveau de confiance. On se référera sur ce point aux notions statistiques décrites dans la fiche n° 4 « Éléments statistiques ».

Série de fiches « Qualifier les données géographiques »

Contributeurs

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

Rédacteurs
Yves Bonin (Cerema Méditerranée), Arnauld Gallais (Cerema Ouest).

Contributeurs
Mathieu Rajerison, Silvio Rousic (Cerema Méditerranée).

Relecteurs
Benoît David (Mission information géographique MTES/CGDD), Stéphane Rolle (CRIGe PACA), Magali Carnino (DGAC), Stéphane Lévêque (Cerema Territoires et ville).

Maquettage
Cerema Territoires et ville
Service édition

Impression
Jouve
Mayenne

Date de publication 2017
ISSN : 2417-9701
2017/59

Contact

accueil.dtectv@cerema.fr

Boutique en ligne : catalogue.territoires-ville.cerema.fr

La collection « Connaissances » du Cerema

Cette collection présente l'état des connaissances à un moment donné et délivre de l'information sur un sujet, sans pour autant prétendre à l'exhaustivité. Elle offre une mise à jour des savoirs et pratiques professionnelles incluant de nouvelles approches techniques ou méthodologiques. Elle s'adresse à des professionnels souhaitant maintenir et approfondir leurs connaissances sur des domaines techniques en évolution constante. Les éléments présentés peuvent être considérés comme des préconisations, sans avoir le statut de références validées.

Aménagement et développement des territoires - Ville et stratégies urbaines - Transition énergétique et climat - Environnement et ressources naturelles - Prévention des risques - Bien-être et réduction des nuisances - Mobilité et transport - Infrastructures de transport - Habitat et bâtiment

Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement - www.cerema.fr
Cerema Territoires et ville - 2 rue Antoine Charial - 69426 LYON cedex 03 - Tél. +33 (0)4 72 74 58 00
Siège social : Cité des mobilités - 25 av. François Mitterrand - CS 92803 - 69674 BRON Cedex - Tél. +33 (0)4 72 14 30 30