RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

IGN
INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

CHANGER D'ÉCHELLE

Institut national de l'information géographique et forestière

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

IGN
INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

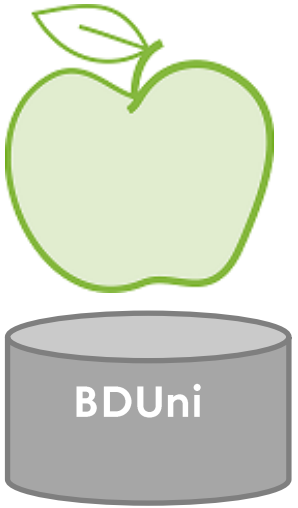
CHANGER D'ÉCHELLE

DÉTECTION D'ERREURS DANS LA BDUNI VIA ANALYSES STATISTIQUES

Sommaire

1. Introduction

Contexte



Réalisation Gabriel Bregand, pendant son alternance Bachelor Universitaire de Technologie Science des Données sous l'encadrement de Nicolas Py (IGN/DT-CE).

Contexte et objectif

- BDUni : big data.
- Des règles et processus d'assurance qualité sont déjà en place pour garantir la qualité des données.
- Problème identifié : des incohérences/erreurs subsistent malgré ces processus.

Notre démarche

- But : établir des **propositions** pouvant faire l'objet de nouvelles règles et/ou contrôles qualité
  (Non la création de ces règles et/ou contrôles.)

Rappels

Les 5 critères proposés par la norme ISO 19157

- **Cohérence logique**
  - la cohérence **conceptuelle**
  - la cohérence des **domaines** de valeur
  - la cohérence du **format**
  - la cohérence **topologique**
- **Exhaustivité**
  - **l'excédent**
  - **l'omission**
- **Précision thématique**
  - la justesse du **classement**
  - la justesse des **attributs non quantitatifs**
  - la précision des attributs **quantitatifs**
- **Précision de position**
  - la précision de **position absolue**
  - la précision de position **relative**
  - la précision de position de **données matricielles**
- **Qualité temporelle**
  - l'exactitude de la mesure temporelle
  - la cohérence temporelle
  - la validité temporelle.

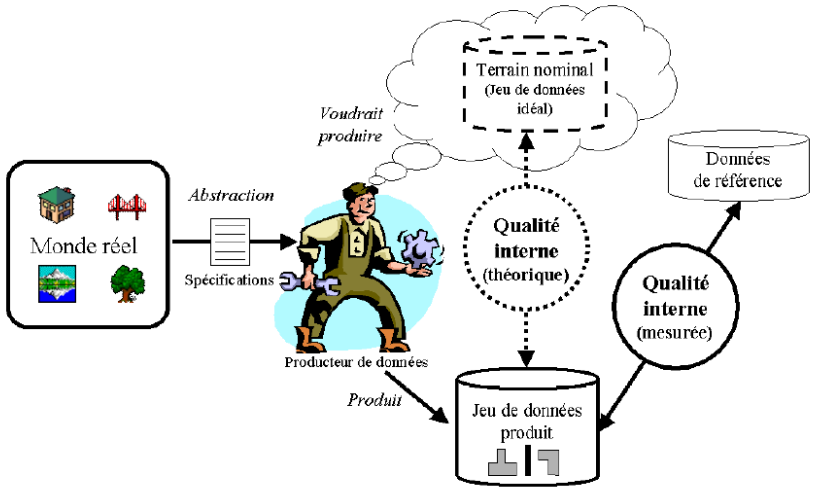


Figure 7 : Concepts de qualité interne et son évaluation

Rodolphe Devillers. Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales. Géographie. Université de Marne la Vallée, 2004. Français. ⟨tel-00008930⟩

Champ d'étude

Tables étudiées

- Bâtiment
- Tronçon de route
- ZAI

Types de variables étudiées

- INT
- FLOAT
- BOOLEAN
- LISTE ORDONNÉE / NON ORDONNÉE
- Géométrie

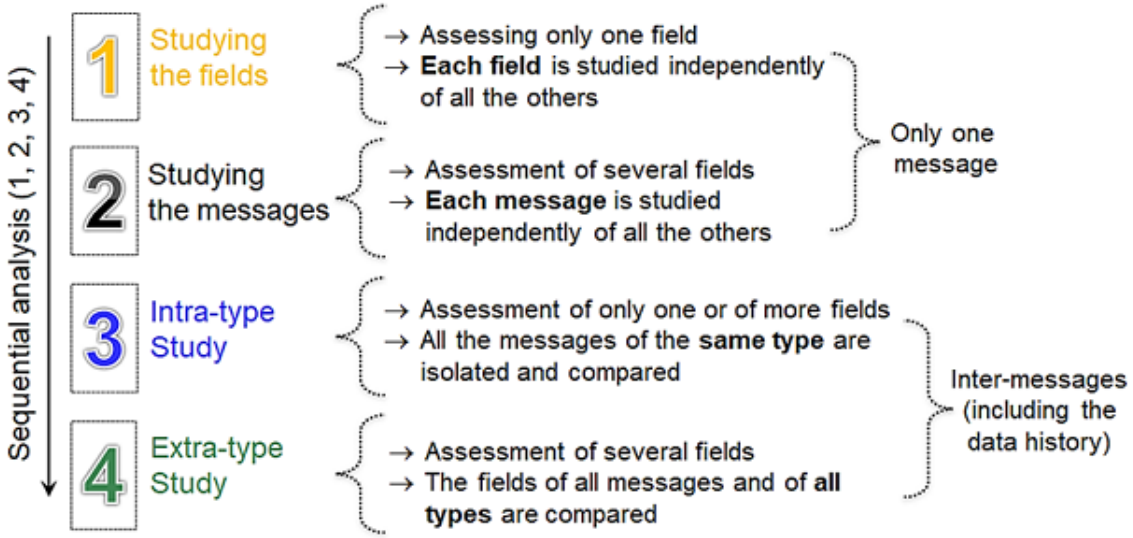


Figure 4.1. *A four-step analysis of integrity. For a color version of this figure, see www.iste.co.uk/batton/geographic1.zip*

https://www.istegroup.com/fr/produit/limperfection-des-donnees-geographiques-1/

Document de Méthodologie

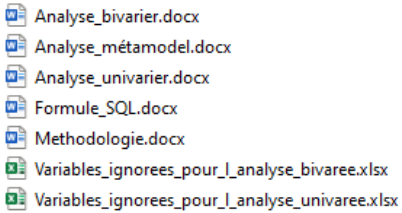
Objectif : Expliquer le processus général de vérification, réplicable sur les tables que nous n'avons pas étudiées.
Incluant :

- Étapes clefs de la réalisation du projet
- Les pistes d'étude que nous avons sélectionnées
- Explication de l'objectif des différents programmes conçus pour les analyses
- Les étapes principales de chaque type d'analyse

Documents d'Analyses

Objectif : Expliquer notre réflexion pour chaque type d'analyse sur les erreurs détectées
Incluant :

- Analyse, Réflexion, Ampleur de l'erreur
  - Graphique / tableaux de comptage
  - Requête SQL
- Proposition



2. Résultats


usage_1_COMPARE_hauteur
Distribution of usage_2
Résidentiel
Religieux
Industriel
Commerce et services
Annexe
Nature
Bac ou liaison maritime
Bretelle
importance
calcul de la moyenne avec 0 .
calcul de la moyenne sans 0 : 181.82
calcul de la médiane avec 0 : 118.500000
calcul de la médiane sans 0 : 118.600000
nombre de lignes : 2542005 (100.00%)
nombre de lignes sans 0 : 2513875 (98.89%)
nombre de lignes avec 0 : 441 (0.
nombre de lignes avec NA : 27689 (1.09%)
usage_2

Analyse métamodèle

Le métamodèle IGN est une base de donnée décrivant la structure (type, noms, contraintes, emploi dans quelles bases dérivées, description, spécifications de saisie, …) de la base de donnée de production (BDUni) et diffusion (BDTopo).
Nous avons réalisé 8 propositions

Exemple d'incohérence

Liste vide,
Liste avec au moins deux arguments NULL / "",
Liste avec plusieurs fois le même objet,

Les regex ne sont pas toujours respectés pour les variables de type numérique telles que :
Troncon_de_route

- restriction_de_poids_total
- restriction_de_poids_par_essieu
- restriction_de_longueur
- restriction_de_largeur
- restriction_de_hauteur

	id_type_attribut integer	id_theme integer	type_attribut_fr text	valeurs text[]
1	9	135	Liste	[null]
2	9	135	Liste	[null]
3	9	116	Liste	[null]
4	9	116	Liste	[null]
5	9	72	Liste	[null]
6	9	72	Liste	[null]
7	9	72	Liste	[null]
8	9	72	Liste	[null]
9	9	72	Liste	[null]

valeurs
text[]

{"",Antenne-relais,"Pylône météo","Tour hertzienne","",Chevalement,"Mât d'éclairage","Mât de mesure","Pylône de pont","Structure artificielle d'escalade","Tour de guet","","","",Campanile,"",Orato

{"","Aire de repos","Aire de service","","",Echangeur,Rond-point,"","","Arrêt touristique saisonnier","Arrêt routier ferroviaire","","Arrêt de fret","","","","Gare funiculaire","Gare RER","Gare TGV","Gare touri

{"","","",""}

{Eglise,"Eglise orthodoxe","Eglise simultanée","Centrale hydroélectrique","Centrale marémotrice","Centrale photovoltaïque","Centrale thermique","",Champignonnière,Cressonnière,Pépinière,"I

{"","Château fort",Tour,"",Bergerie,Cabane,"","","",Chalet,"Hôtel particulier",Immeuble,Lotissement,"Quartier urbain",Résidence,"","Eglise ruinée","Château ruiné","Tour ruinée","",Carbet,"Village détr

{Paravalanche,Rempart,"",Pêcherie,Tranchée,"Voie ferrée déposée","",Vanne,"","","","Mur de pierres sèches","","","",Culture en terrasse",Aqueduc,Passerelle,"","Pont isolé","Pont mobile","Pont subm

{"",Antenne-relais,"Pylône météo","","",Chevalement,"Mât d'éclairage","Mât de mesure","Pylône de pont","Structure artificielle d'escalade","Tour de guet","","","",Campanile,"",Oratoire,Statue,Vierge,"

{"",Epave,"","",Tourelle,"","",Saut,"",DFCI,"","","","",Bouée lumineuse","Tourelle lumineuse","","","","",Tourbière,"","",Abreuvoir,Puits,Réservoir,"","",""}

{"","Château fort",Tour,"",Bergerie,Cabane,"","","",Chalet,"Hôtel particulier",Immeuble,Lotissement,"Quartier urbain",Résidence,"","Eglise ruinée","Château ruiné","Tour ruinée","",Carbet,"Village détr

{Paravalanche,Rempart,"",Pêcherie,Tranchée,"Voie ferrée déposée","",Vanne,"","","","Mur de pierres sèches","","","",Culture en terrasse",Aqueduc,Passerelle,"","Pont isolé","Pont mobile","Pont subm

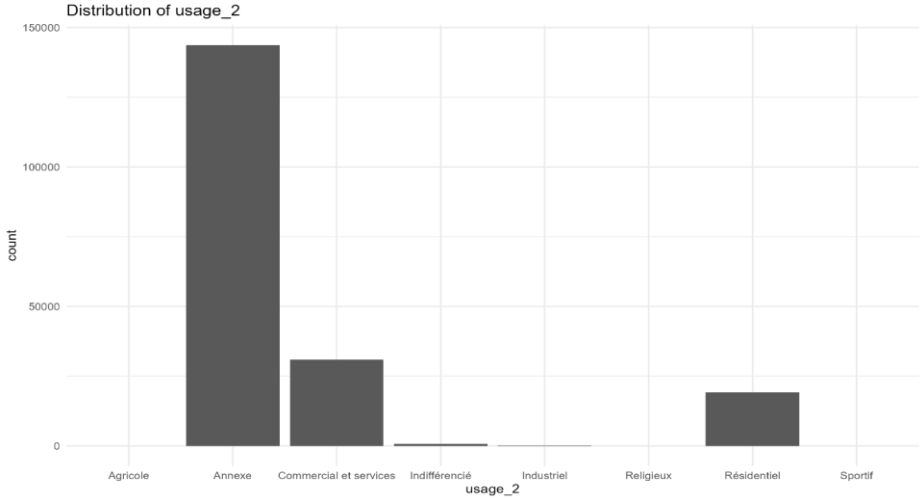
{Paravalanche,Rempart,"",Pêcherie,Tranchée,"Voie ferrée déposée","",Vanne,"","","","Mur de pierres sèches","","","",Culture en terrasse",Aqueduc,Passerelle,"","Pont isolé","Pont mobile","Pont subm

	id_theme integer	id_classe integer	id_attribut integer	valeurs text[]
1	211	212	1649	{"Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire"
2	269	303	430	{"Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire"
3	269	271	430	{"Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire"
4	53	55	430	{"Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire","Administratif ou militaire"

Analyse univariée

Nous avons réalisé 25 propositions suite à des comptages et des représentations graphiques.

Ces propositions concernent les trois tables étudiées, avec un total de 18 variables citées



Exemple d'incohérence :
Table : Bâtiment
Variable : Usage_2

Usage 2 = « Indifférencié »

Spécifications de saisie : Non utilisé.

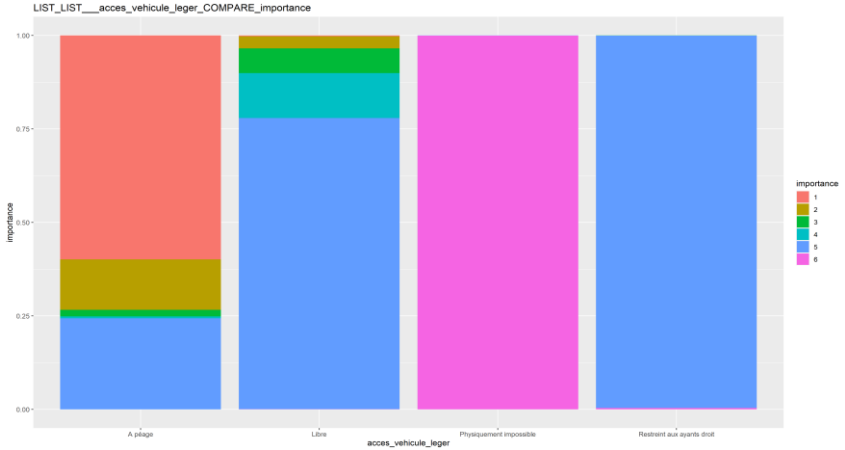
	usage_2 character varying	count bigint
1	Indifférencié	13303

Proposition
Remplacer les usages_2 indifférencié par des NULL

Analyse bivariée

Nous avons réalisé 13 propositions suite à des comptages et tables de contingence et leur représentations graphiques.

Ces propositions concernent les trois tables étudiées, avec autant de propositions que de combinaisons de variables.



Exemple d'incohérence :
Table : Tronçon de route
Variable : acces_vehicule_leger / importance

Assurance qualité : 'Accès véhicule léger' = "Restreint aux ayant droits" ou "Impossible" / Importance = 5 ou 6

Proposition

Les tronçons de route avec l'attribut Accès véhicule léger = « Physiquement impossible » devrait être d'importance = « 6 » ou « 5 » ou avoir leur attribut Accès véhicule léger modifié.

	acces_vehicule_leger	importance	n	pct	lbl
1	A péage	1	1497	5.990396e-01	59.9%
2	A péage	2	337	1.348539e-01	13.5%
3	A péage	3	44	1.760704e-02	1.8%
4	A péage	4	13	5.202081e-03	0.5%
5	A péage	5	608	2.432973e-01	24.3%
6	Libre	1	1853	2.248759e-03	0.22%
7	Libre	2	26708	3.241223e-02	3.24%
8	Libre	3	54323	6.592517e-02	6.59%
9	Libre	4	99075	1.202352e-01	12.02%
10	Libre	5	641712	7.787672e-01	77.88%
11	Libre	6	339	4.114028e-04	0.04%
12	Physiquement impossible	2	5	4.081699e-05	0.0041%
13	Physiquement impossible	3	5	4.081699e-05	0.0041%
14	Physiquement impossible	5	11	8.979738e-05	0.0090%
15	Physiquement impossible	6	122477	9.998286e-01	99.9829%
16	Restreint aux ayants droit	2	2	5.456579e-05	0.0055%
17	Restreint aux ayants droit	3	4	1.091316e-04	0.0109%
18	Restreint aux ayants droit	4	10	2.728290e-04	0.0273%
19	Restreint aux ayants droit	5	36505	9.959621e-01	99.5962%
20	Restreint aux ayants droit	6	132	3.601342e-03	0.3601%

Analyses multivariées ad hoc

Exemples d'analyse sur la table bâtiment, utilisation des liens évidents entre variables :

- Comparaison de tous les attributs liés avec le Z
  - Altitude minimale sol
  - Altitude minimale toit
  - Altitude maximale toit
  - Altitude maximale sol
- Utilisation du bâtiment
  - Nature
  - usage_1
  - usage_2
  - aussi comparaison avec la présence ou non de logement

Nature	Usage 1	Usage 2	Nombre	Suspicion
Indifférenciée	Indifférencié		20628334	
Indifférenciée	Indifférencié	Agricole	2	
Indifférenciée	Indifférencié	Annexe	264	
Indifférenciée	Indifférencié	Commercial et services	110	
Indifférenciée	Indifférencié	Industriel	3	
Indifférenciée	Indifférencié	Résidentiel	647	
Indifférenciée	Indifférencié	Sportif	2	
Arène ou théâtre antique	Indifférencié	Commercial et services	1	
Chapelle	Indifférencié	Commercial et services	2	
Chapelle	Indifférencié	Résidentiel	1	
Château	Indifférencié	Résidentiel	3	
Eglise	Indifférencié	Religieux	1	
Industriel, agricole ou commercial	Indifférencié	Agricole	1	
Industriel, agricole ou commercial	Indifférencié	Annexe	121	
Industriel, agricole ou commercial	Indifférencié	Commercial et services	202	
Industriel, agricole ou commercial	Indifférencié	Industriel	2	
Industriel, agricole ou commercial	Indifférencié	Résidentiel	139	
Industriel, agricole ou commercial	Indifférencié	Sportif	1	
Moulin à vent	Indifférencié	Annexe	1	
Silo	Indifférencié	Agricole	1	
Tour, donjon	Indifférencié	Industriel	1	
Tribune	Indifférencié	Commercial et services	3	
Indifférenciée	Agricole	Indifférencié	1	
Indifférenciée	Annexe	Indifférencié	63	
Indifférenciée	Commercial et services	Indifférencié	1083	
Indifférenciée	Industriel	Indifférencié	2	
Indifférenciée	Religieux	Indifférencié	11	
Indifférenciée	Résidentiel	Indifférencié	1252	
Indifférenciée	Sportif	Indifférencié	44	
Industriel, agricole ou commercial	Agricole	Indifférencié	513	
Industriel, agricole ou commercial	Annexe	Indifférencié	3	
Industriel, agricole ou commercial	Commercial et services	Indifférencié	507	
Industriel, agricole ou commercial	Industriel	Indifférencié	299	
Industriel, agricole ou commercial	Résidentiel	Indifférencié	10	
Arène ou théâtre antique	Indifférencié	Indifférencié	1	
Indifférenciée	Indifférencié	Indifférencié	8434	
Industriel, agricole ou commercial	Commercial et services	Commercial et services	2	
Industriel, agricole ou commercial	Indifférencié	Indifférencié	768	
Chapelle	Commercial et services		36	Très probablement pas d'erreur mais l'usage reste suspect par rapport à la nature
Eglise	Commercial et services		30	
Indifférenciée	Agricole	Sportif	1	Usage Agricole et Sportif semble peu probable
Indifférenciée	Sportif	Agricole	1	
Industriel, agricole ou commercial	Agricole	Sportif	1	
Industriel, agricole ou commercial	Religieux	Commercial et services	3	Usage_1 religieux ne semble pas coincidé avec la nature mais possible
Industriel, agricole ou commercial	Religieux		22	
~~Industriel, agricole ou commercial~~	~~Résidentiel~~		~~57099~~	
Industriel, agricole ou commercial	Sportif	Annexe	9	
Industriel, agricole ou commercial	Sportif	Industriel	4	
Industriel, agricole ou commercial	Sportif	Résidentiel	17	
Industriel, agricole ou commercial	Sportif		1418	
Chapelle	Résidentiel		16	Possible mais peu probable possible erreur.
Eglise	Résidentiel		15	
Château	Annexe	Commercial et services	3	Qu'un château serf d'annexe à un autre bâtiments semble suspect Annexe Définition : Petit bâtiment à vocation d'annexe au sens fiscal (garage externe, abri...).
Château	Annexe	Résidentiel	193	
Château	Annexe		990	
Château	Commercial et services	Annexe	26	
Château	Indifférencié	Annexe	4	
Château	Résidentiel	Annexe	2047	
Tour, donjon	Agricole		6	Potentiellement Nature = Silo
~~Chapelle~~	~~Sportif~~		~~4~~	~~Connaissez-vous l'histoire de la chapelle des arts martiaux à Aumale ?~~
~~Eglise~~	~~Sportif~~		~~1~~	
Silo	Agricole	Résidentiel	101	Silo : « Sélection : Le silo est exclusivement destiné aux produits agricoles. » Donc l'usage résidentiel ou commercial et services semble être des erreurs
Silo	Commercial et services		27	
Silo	Industriel	Commercial et services	7	
Silo	Industriel	Résidentiel	1	
Silo	Résidentiel	Industriel	1	
Silo	Résidentiel		18	

Type ERREUR	Graviter ERREUR
Nature Indifférencié et usage Indifférencié	
usage_1 indifférencié et usage_2 != NULL	
usage_2 = Indifférencié	
Usage_1 = Usage_2	
Peut suspect	
Assez suspect	
Très suspect	
Faux selon les Spéc	

Analyses multivariées par création de modèles

Objectif :

Prédire une variable cible $\hat{Y}$ à partir des autres variables X et la comparer avec la valeur réelle Y dans la base de données.

=> Aka prédire une colonne à l'aide des autres colonnes

$\Rightarrow$ Si $\hat{Y}$ diffère de Y, est que la confiance dans le modèle est élevée, alors incohérence potentielle

Principes méthodologiques :

- Utilisation de modèles non paramétriques
- Evaluation intrinsèque, pas de création de donnée de vérité terrain

Implications et difficultés méthodologiques :

Données IGN de bonne qualité, donc avec peu d'incohérence détectables (~sensibilité du modèle à détecter des incohérences)

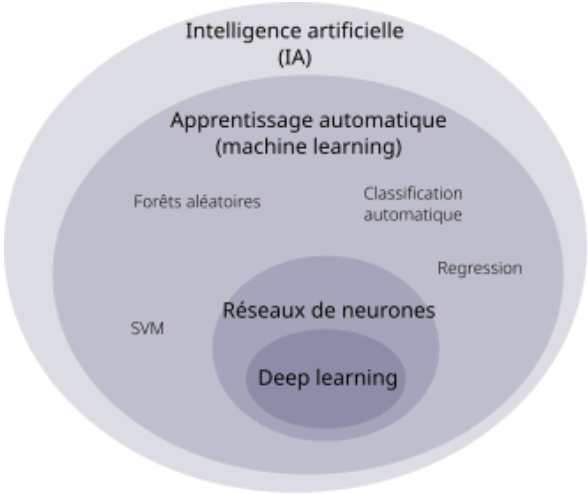
Données IGN contenant de nombreux NA

Il faut que le modèle soit confiant dans ses prédictions

Création de multiples modèles pour la même table de données, avec un hyperparamétrage pour chacun

Augmentation des attributs par calculs de descripteurs de la géométrie, afin de considérer la géométrie dans les analyses (ex, pour les segments d'un tronçon, longueur moyenne, angle min+moy+max entre segments, …)

Soin dans l'échantillonnage, représentatif

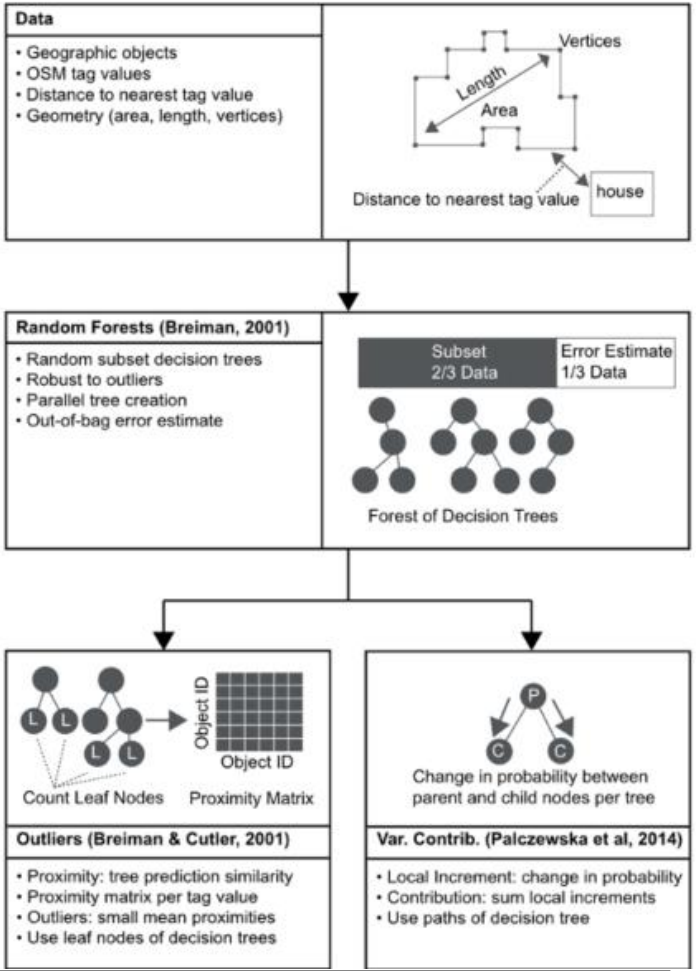


https://fr.wikipedia.org/wiki/Apprentissage_automatique

Random forest

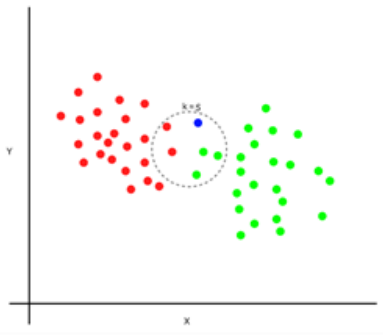
Pour une première approche d'application de modèle statistique sur nos données, nous avons retenu la méthode des forêts aléatoires (Random Forest) pour plusieurs raisons :

- **Apprentissage automatique**
- **Gestion d'un grand nombre de variables explicatives (existantes ou calculées)**
- **Adapté aux individus avec des variables NULL**
- **Prédictions basées sur des arbres décisionnels**
- **Facilité d'optimisation des paramètres**



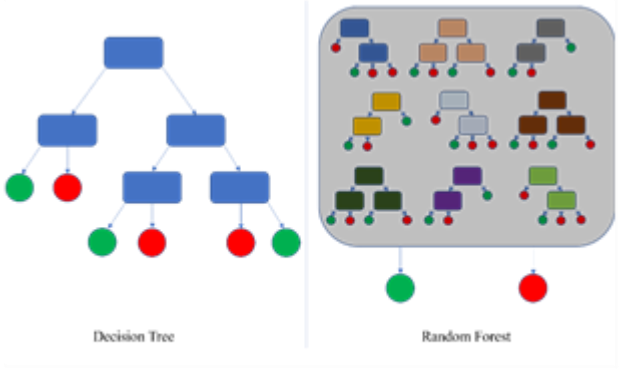
Étude de Référence pour Orienter Notre Approche

Outlier Detection in OpenStreetMap Data using the RandomForest Algorithm and Variable Contributions
https://github.com/rrwen/poster-gisci-osmol



Exemple

https://pub.towardsai.net/understanding-k-nearest-neighbors-a-simple-approach-to-classification-and-regression-e4b30d37f151



https://fr.wikipedia.org/wiki/For%C3%AAt_d%27arbres_d%C3%A9cisionnels#/media/Fichier:Decision_Tree_vs._Random_Forest.png

KNN: La Méthode des k plus proches voisins est une méthode d'apprentissage non paramétrique et supervisé qui utilise la **distance** entre les objets pour prédire un la modalité d'un objet. Pour connaître la modalité d'un objet, on calcule quelles sont les k points les plus proches de lui dans un espace à n dimensions et on lui attribue la modalité majoritairement présente parmi les k plus proches voisins.

Random Forest: ensemble de prédictions d'arbres de décision. Random Forest s'appuie sur plusieurs arbres construits à partir d'un échantillon aléatoire des données fourni au model (un rééchantillonnage) et une sélection aléatoire des variables.

Bonus: score d'importance des variables -> variables dont on pourrait suggérer de faire particulièrement attention à leur qualité.

Isolation Forest, similaire au Random Forest, repose sur l'hypothèse que plus un objet nécessite de divisions pour être isolé, plus il est suspecté d'être un outlier. Ce modèle ne se base pas sur la prédiction finale, mais uniquement sur la longueur du chemin dans les arbres pour estimer si l'objet est un objet utile.

Bonus:
•L'importance des variables par impureté (Gini importance) : calculée à partir de la réduction d'impureté induite par chaque variable dans les arbres de décision.
•L'importance par permutation : qui mesure la dégradation de la performance du modèle lorsque les valeurs d'une variable sont aléatoirement réorganisées, fournissant ainsi une estimation plus robuste de son utilité réelle.

Hyperparamétrage recquis
•n_estimators, nombre d'arbres dans la forêt
•max_depth, nombre de de nœuds max par le quelle peut passer un objet
•min_samples_split, Le nombre minimum d'objets pour qu'une feuille se re sépare en nœuds
• …

Exemple, tronçon de route, nature (1|3)

Classification Report :

	precision	recall	f1-score	support
Bretelle	0.60	0.92	0.73	8808
Chemin_Pratique	1.00	1.00	1.00	56068
Rond-point	0.96	0.98	0.97	28008
Route à chaussée	0.98	0.88	0.93	55792
Sentier_Escalier	1.00	1.00	1.00	36472
Type autoroutier	0.95	0.97	0.96	14852
accuracy			0.96	200000
macro avg	0.92	0.96	**0.93**	200000
weighted avg	0.97	0.96	0.96	200000

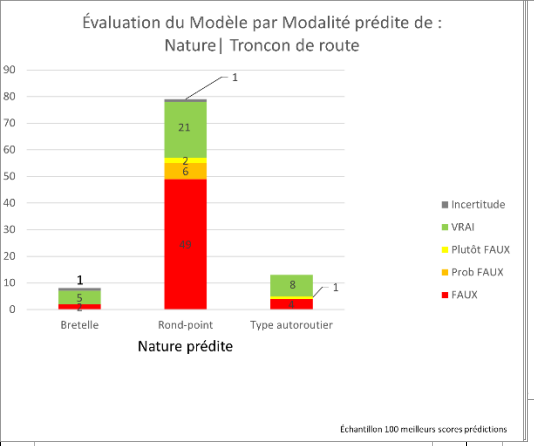
Bonnes performances globales, avec un F1-score macro moyen de 0.93, ce qui indique qu'il parvient à bien prédire, en moyenne, toutes les classes, même les moins représentées.

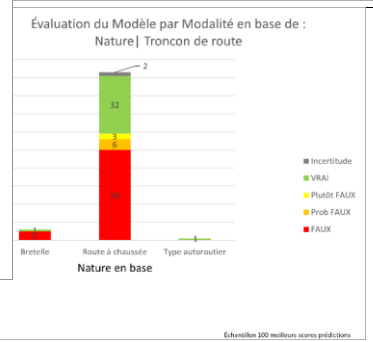
L'accuracy de 96 % signifie que 96 % des objets contenus dans le jeu de données test ont été correctement prédits. Pour la classe Bretelle, le rappel est élevé (0.92) mais la précision est plus faible (0.60) : le modèle retrouve bien la plupart des vraies bretelles, mais en prédit aussi à tort.

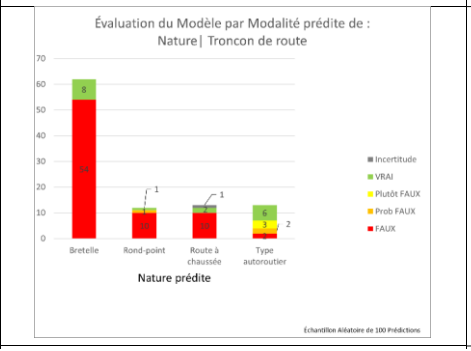
Cela peut entraîner un nombre important de fausses prédictions, et mérite donc une attention particulière.

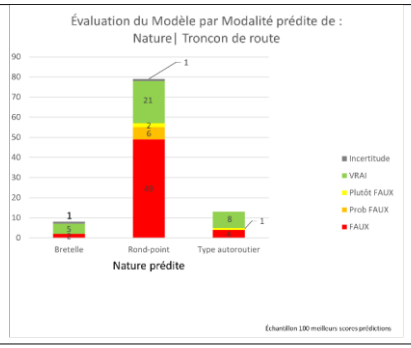
Sur la BDD de validation : accuracy de 0.9597 qui confirme que le modèle généralise bien et qu'il n'y a probablement pas de surapprentissage.

Exemple, tronçon de route, nature (2|3)







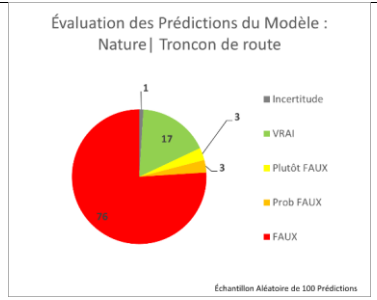


Parmi les objets où nature prédite != nature en base, on sélectionne:

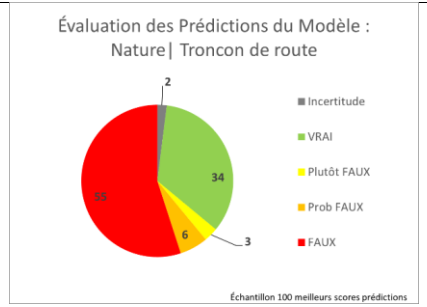
- Un échantillon aléatoire de 100 objets
- Les 100 prédictions où la confiance est la plus élevée

Puis on vérifie et juge l'incohérence détectée

100 objets choisis aléatoirement,



100 objets ayant les predict_poutlier_scores (predict_proba) les plus élevés



Tableaux de comptage des 100 objets pour lesquels la valeur prédite par le modèle (tirée aléatoirement) différait de celle présente en base.

Pour chacun, une vérification manuelle a permis de qualifier l'erreur:

(1) le modèle avait raison de prédire une valeur différente (la base était erronée)

(2) le modèle s'était trompé (la base était correcte)

(3) le cas reste indécidable (pas de conclusion tranchée):

Prédite / Base	Bretelle	Rond-point	Route à chaussée	Type autoroutier
Bretelle		0|0|0	1|4|1	1|1|0
Rond-point	1|0|0		1|6|0	0|0|0
Route à chaussée	2|52|0	1|11|0		5|6|0
Type autoroutier	5|2|0	0|0|0	0|0|0	

Total erreur détectée = 17

Tableaux de comptage des 100 objets pour lesquels la valeur prédite par le modèle (avec une forte confiance) différait de celle présente en base.

Pour chacun, une vérification manuelle a permis de qualifier l'erreur:

(1) le modèle avait raison de prédire une valeur différente (la base était erronée)

(2) le modèle s'était trompé (la base était correcte)

(3) le cas reste indécidable (pas de conclusion tranchée):

Prédite / Base	Bretelle	Rond-point	Type autoroutier
Bretelle		0|2|0	1|3|0
Route à chaussée	4|2|1	21|55|1	7|2|0
Type autoroutier	1|0|0	0|0|0	

Total erreur détectée = 34

Exemple, tronçon de route, nature (3|3)

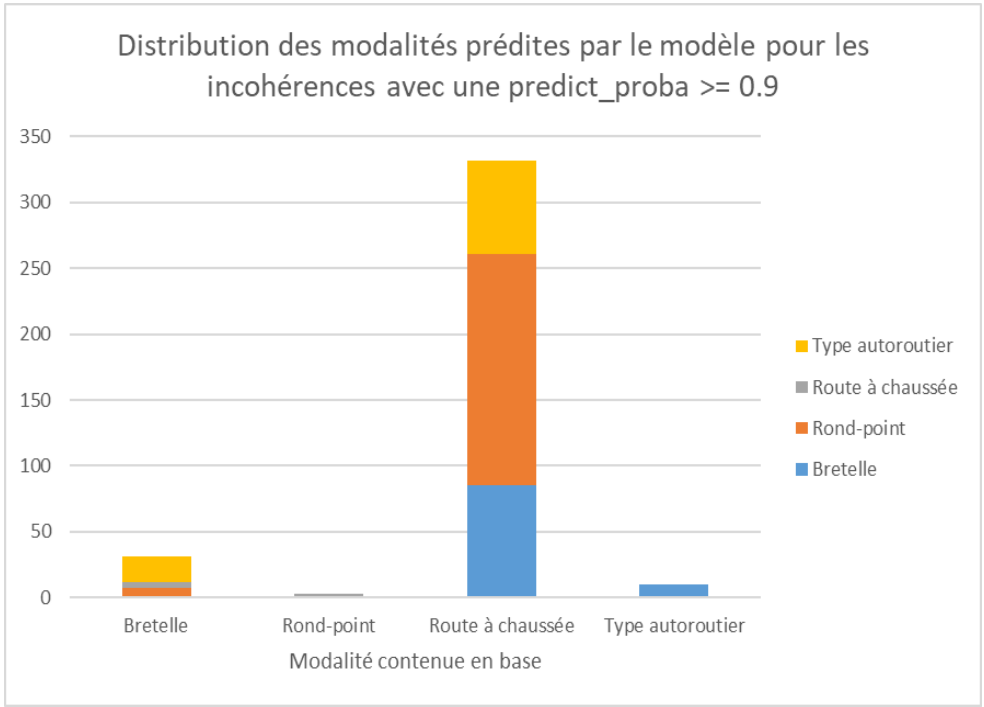
Environ 4% de l'échantillon montre une incohérence [valeur prédite / valeur observée].

Environ 5% des incohérences [valeur prédite / valeur observée] sont de predict_proba >= 0.9.

Si on extrapole à l'ensemble de la base, cela représente environ 0.2 %, soit près de 40 000 objets pouvant être détecter comme potentiellement porteur d'anomalie (avec une forte 'confiance' du modèle).

Si lors de cette inférence on obtient la même proportion d'objet réellement porteur d'anomalie/erreur que lors de la vérification manuelle sur l'échantillon des 100 meilleures prédictions, soit 1/3 des incohérences qui seraient des erreurs, cela pourrait permettre d'identifier environ 13 500 corrections sur l'attribut nature.

La majorité des incohérences concerneraient des objets de nature routes à chaussée (1 ou 2 chaussées), qui pourraient être de nature : type autoroutier, bretelle, ou rond-point.



Autres exemples

Tronçon de route
Nombre de voies

Tableaux de comptage des 100 objets pour lesquels la valeur prédite par le modèle (tirée aléatoirement) différait de celle présente en base.

Pour chacun, une vérification manuelle a permis de qualifier l'erreur:

(1) le modèle avait raison de prédire une valeur différente (la base était erronée)

(2) le modèle s'était trompé (la base était correcte)

(3) le cas reste indécidable (pas de conclusion tranchée) :

Prédite \ Base	0	1	2	3	4
0		0|0|0	0|0|0	0|0|0	0|0|0
1	0|0|0		4|33|3	0|0|0	0|0|0
2	1|1|0	4|4|1		6|15|0	0|0|0
3	0|0|0	1|0|0	7|3|1		1|0|1
4	0|0|0	0|0|0	1|1|0	6|5|0	

Total erreur détectée = 31

Tableaux de comptage des 100 objets pour lesquels la valeur prédite par le modèle (avec une forte confiance) différait de celle présente en base.

Pour chacun, une vérification manuelle a permis de qualifier l'erreur:

(1) le modèle avait raison de prédire une valeur différente (la base était erronée)

(2) le modèle s'était trompé (la base était correcte)

(3) le cas reste indécidable (pas de conclusion tranchée) :

Prédite \ Base	0	1	2	3	4
0		0|0|0	1|0|0	0|0|0	0|0|0
1	1|2|0		13|9|2	0|0|0	0|0|0
2	6|2|0	26|11|0		0|2|0	0|1|0
3	0|0|0	2|0|0	3|2|0		4|6|0
4	0|0|0	0|0|0	0|0|0	0|1|0	

Total erreur détectée = 56

1/3 des incohérences analysées sans tenir compte de la confiance du modèle pourrait amener à une modification/correction en base.

Cela pourrait permettre d'identifier environ 83 000 corrections.

La majorité des incohérences concerneraient des objets avec 1 ou 2 voies, qui pourraient avoir en réalité généralement + ou - 1 voies.

Bâtiments
Nature

BATIMENT0000000325134734	BATIMENT0000000291128278	BATIMENT0000000328614133
En base : Indifférenciée Prédit : Industriel, agricole ou commercial	En base : Eglise Prédit : Chapelle	En base : Industriel, agricole ou commercial Prédit : Indifférenciée

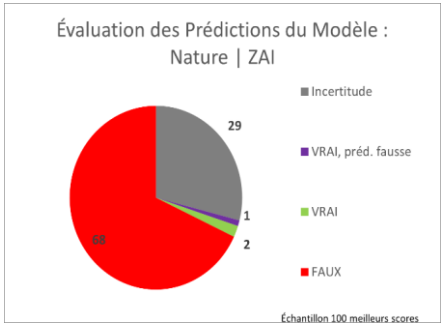
1/3 des incohérences analysées en prenant en compte la confiance du modèle pourrait amener à une modification/correction en base.

Cela pourrait permettre d'identifier environ 150 000 corrections.

La majorité des incohérences sont entre « Indifférenciée » et « Industriel, agricole ou commercial », suivi d'incohérence entre « Eglise » et « Chapelle ».

Le découpage du bâti par la cadastre crée des géométrie génératrices de perturbations pour le modèle.

ZAI



Pas de détection efficace, y compris
En se restreignant à un score de confiance élevé des prédicitions
En écartant les géométries fictives de 5*5m.

Difficultés sur la bâti:

		
BATIMENT0000000325134734 En base : Indifférenciée Prédit : Industriel, agricole ou commercial	BATIMENT0000000291128278 En base : Eglise Prédit : Chapelle	BATIMENT0000000328614133 En base : Industriel, agricole ou commercial Prédit : Indifférenciée