

Analyse de la distribution des noms de rues en France

Gomez Malaurie

POM 2022, UCBL

Encadrement par Fabien Duchateau
et Franck Favetta

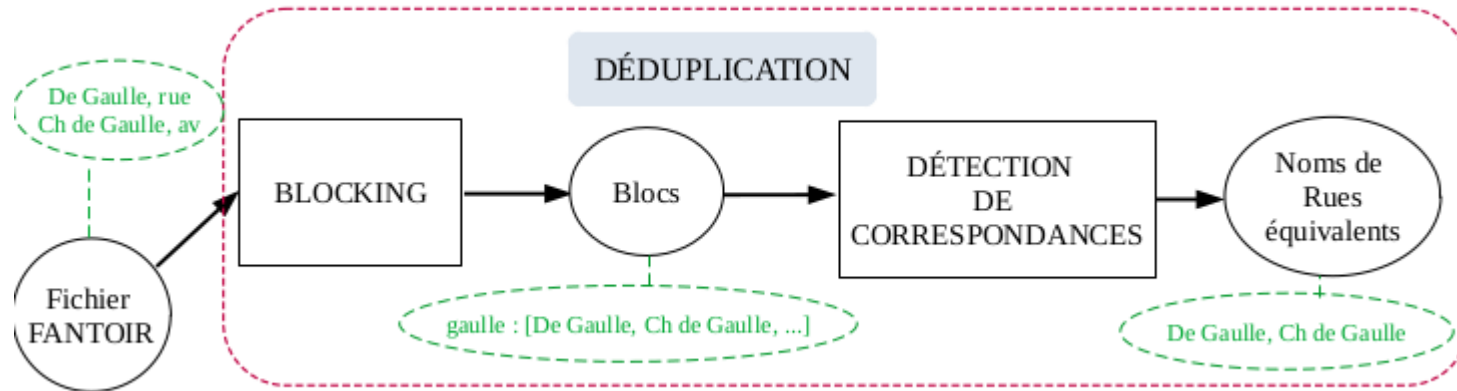
Contexte et problématique

- **Types** de noms de rues variés (e.g., rue de la liberté, rue Louise Michel, rue de la gare) [1]
 - Comment étudier les noms de rues sur une région ou un pays ?
- Challenges
 - **Déduplication** afin de détecter les noms équivalents [2]
 - **Classification** afin de catégoriser les noms (e.g., personnes, lieu)

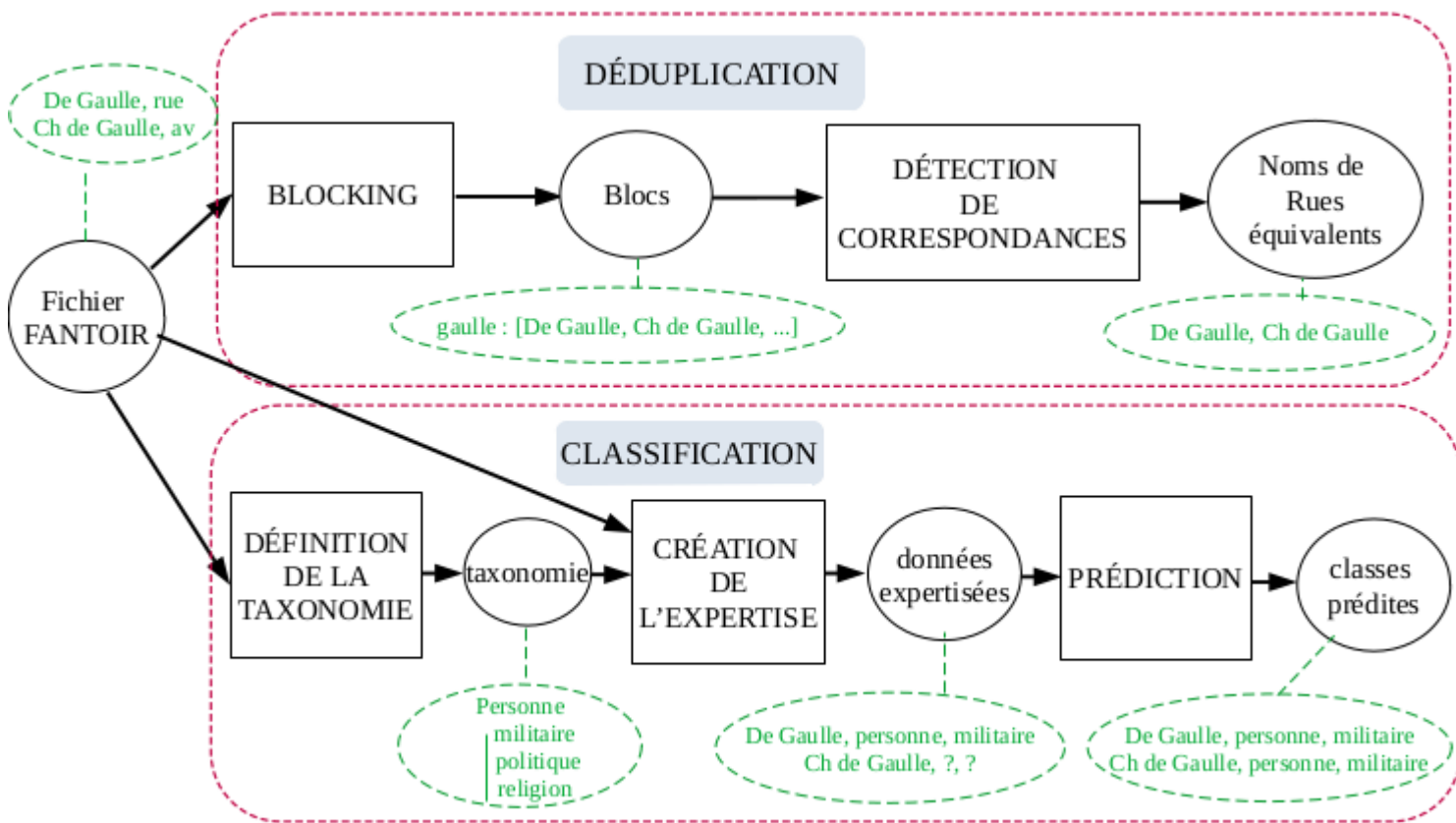
[1] Dominique Badariotti. Les noms de rue en géographie. plaidoyer pour une recherche sur les odonymes. In Annales de Géographie, t. 111, n°625, 2002., 2002

[2] Christen, P. (2012). The data matching process. In Data matching (pp. 23-35). Springer, Berlin, Heidelberg. 2

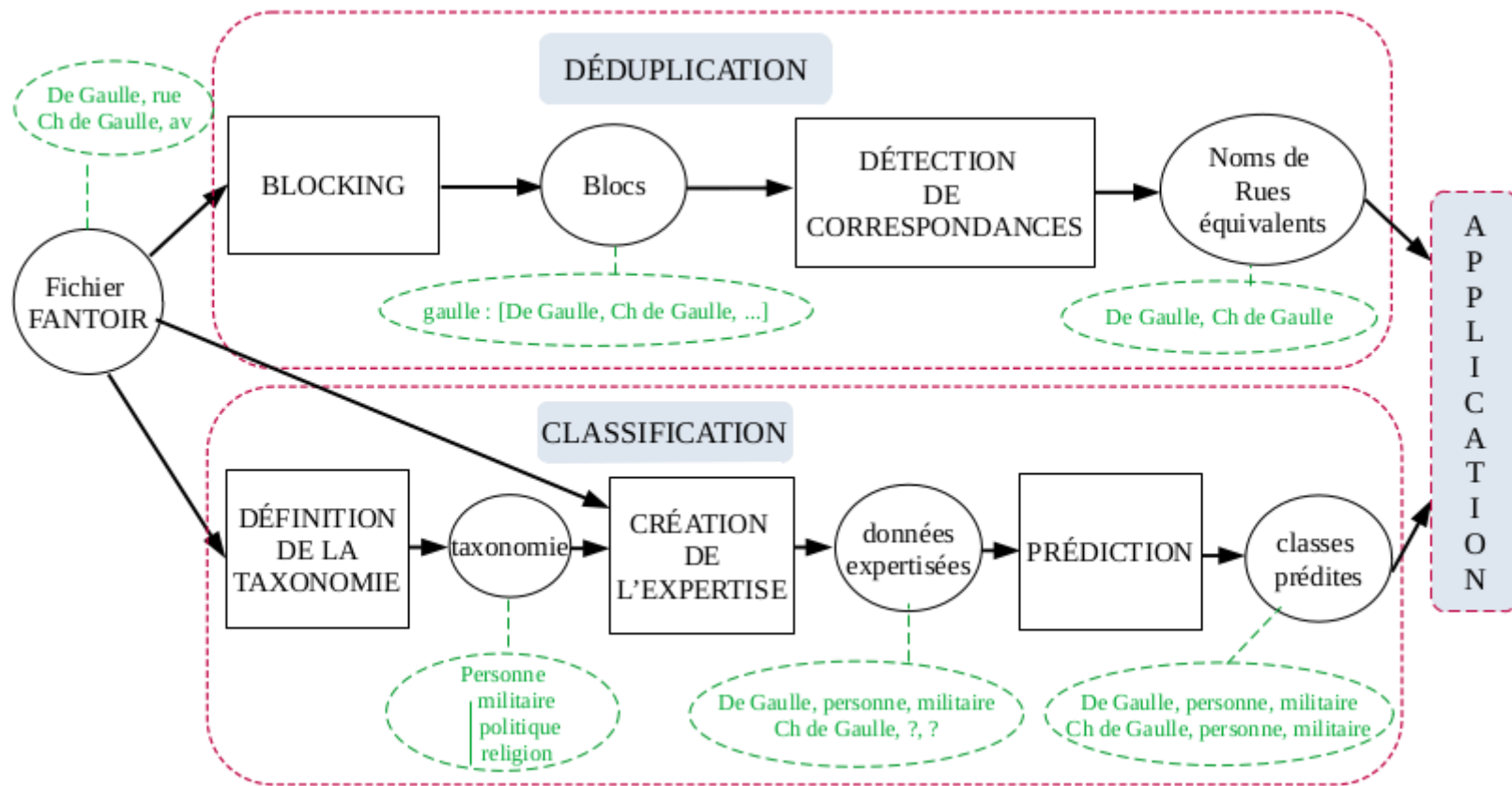
Aperçu de notre proposition



Aperçu de notre proposition

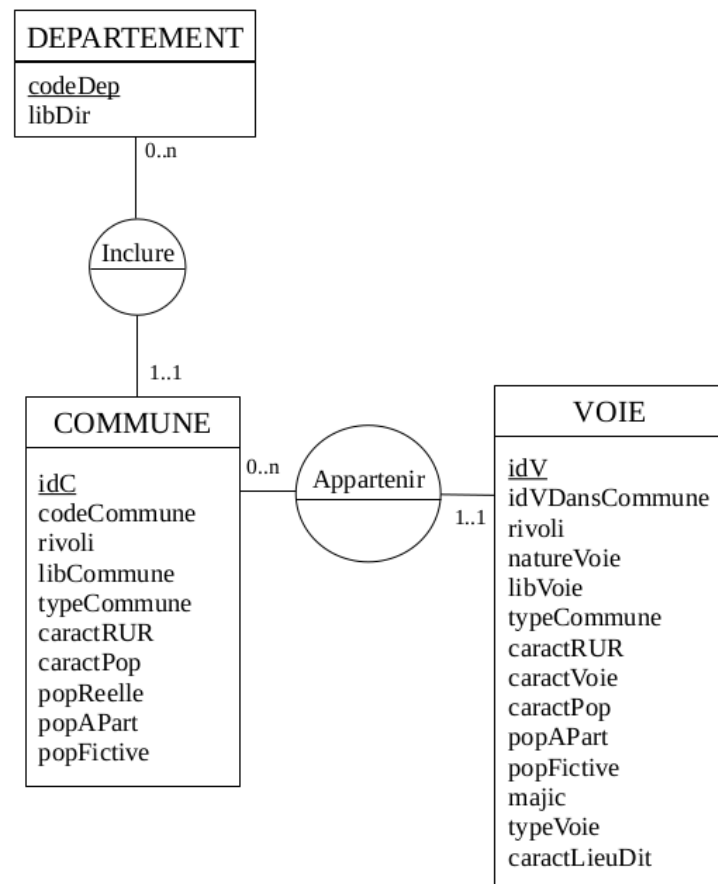


Aperçu de notre proposition



Déduplication - Données

- Fichier **FANTOIR**
 - 9 millions d'entrées
 - 1 Go (décompressé)
- **Modélisation et migration** des données dans une base de données Postgresql



Déduplication - Blocking

- Création de **blocs de noms de rues** à comparer plus finement
- Solution choisie
 - Clé de blocking = présence d'un **mot commun**
 - **Prétraitements** pour optimiser les résultats
 - Ex : « De Gaulle » et « Ch de Gaulle » → « Gaulle » en commun
→ gaulle : [« De Gaulle », « Ch de Gaulle »]

Déduplication - Appariement

Détection de noms de rues en **doublon**

- utilisation de 3 mesures de similarité terminologiques (Levenshtein, Jaro-Winkler et Jaccard) et 1 combinaison
- Différentes configurations

	“de gaulle”, “ch de gaulle”
Levenshtein 1	0.75
Levenshtein 2	0.75
Jaro-Winkler	0.85
Jaccard 1	0.77
Jaccard 2	0.73

Déduplication – Évaluation de la qualité

- Expérimentations sur un jeu de données **expertisé** manuellement (850 correspondances)

	f-mesure	précision	rappel	seuil	nb résultats
levenshtein1	0.83	0.96	0.73	0.81	646
levenshtein2	0.83	0.96	0.73	0.81	647
jaroWinkler	0.82	0.96	0.72	0.95	640
jaccard1	0.78	0.83	0.74	0.87	761
jaccard2	0.89	0.95	0.84	0.63	751
levenshtein1 + jaccard2	0.84	0.98	0.72	0.72	630

Résultats de chaque mesure avec les seuils permettant d'obtenir la meilleure f-mesure.

Déduplication – Évaluation du blocking

- Expérimentations sur un jeu de données **expertisé** manuellement (800 correspondances)

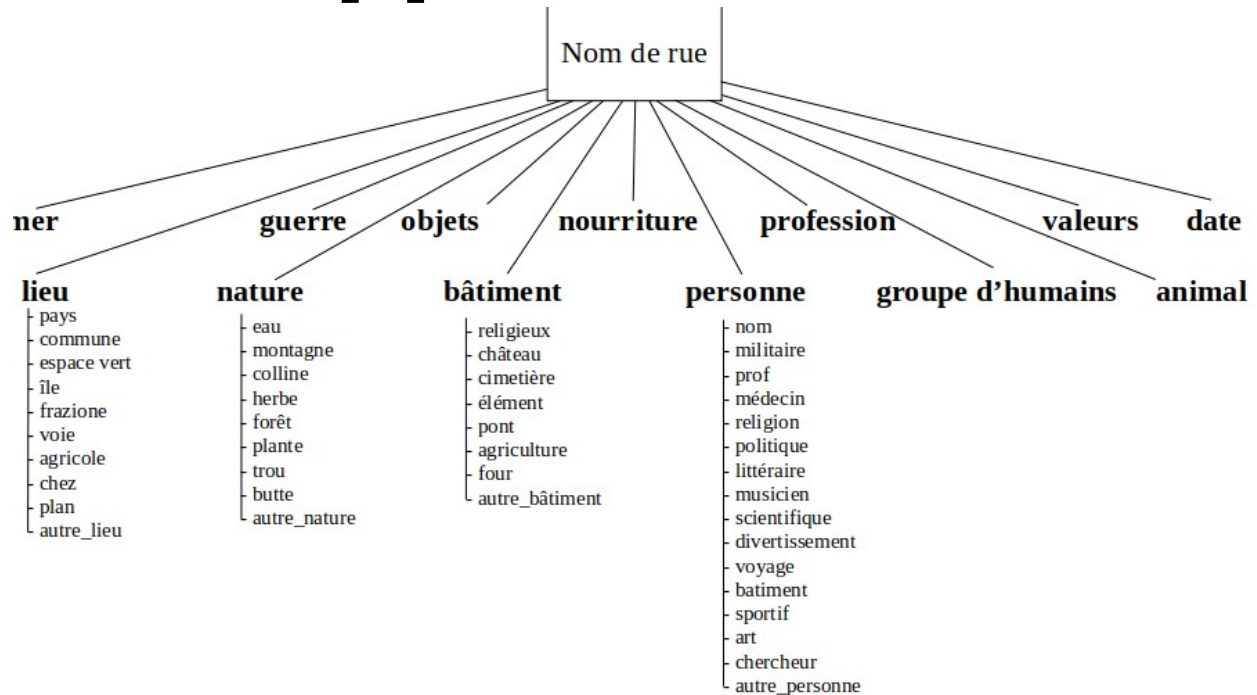
Avec blocking (22 503 comparaisons)		f-mesure	précision	rappel	seuil	nb résultats	temps
	levenshtein1	0.83	0.96	0.73	0.81	646	34.8 s
	levenshtein2	0.83	0.96	0.73	0.81	647	36.3 s
	jaroWinkler	0.82	0.96	0.72	0.95	640	33.6 s
	jaccard1	0.78	0.82	0.74	0.87	761	33.1 s
	jaccard2	0.89	0.94	0.84	0.63	751	33.2 s
Sans blocking (10 591 503 comparaisons)		f-mesure	précision	rappel	seuil	nb résultats	temps
	levenshtein1	0.65	0.58	0.73	0.81	1060	677 s
	levenshtein2	0.65	0.58	0.73	0.81	1062	1037 s
	jaroWinkler	0.67	0.63	0.72	0.95	981	337 s
	jaccard1	0.42	0.29	0.75	0.87	2161	138 s
	jaccard2	0.69	0.58	0.84	0.63	1212	170 s

Résultats de chaque mesure avec et sans blocking

Classification - Taxonomie

- **Catégorisation des rues [1]**

- 13 classes
- 45 sous-classes



[1] Dominique Badariotti. Les noms de rue en géographie. plaidoyer pour une recherche sur les odonymes. In Annales de Géographie, t. 111, n°625, 2002., 2002

Classification - Expertise

- Utilisation de **Wikidata**
 - Correspondances entre les classes de la taxonomie et les types Wikidata
 - Nature > montagne → wdt::Montagne, wdt::Sommet, wdt::Volcan
- Utilisation du **blocking**
 - Utilisation de noms de rues contenant les mots les plus fréquents
 - Nature > trou → « fosse », « creux »
- 2302 exemples annotés

Classification - Méthodes

Méthode	Non binaire (générale)	Binaire	Classe	Sous-classe
gen-classes	X		X	
bin-classes		X	X	
gen-sc	X			X
combi-classes	X	X	X	

Les différentes approches utilisées pour la classification

- 19 critères utilisés (nom de la rue, nature de la voie, longueur, ...)

Classification - Expérimentation

	Avec tous les critères	Critères = [nomRue]
LinearSVC	0.72	0.72
SGDClassifier	0.66	0.71
AdaBoostClassifier	0.39	0.29
MultinomialNB	0.53	0.62
KNeighborsClassifier	0.59	0.63
SVC	0.62	0.7
DecisionTreeClassifier	0.66	0.54

Résultats de la classification parmi toutes les classes (gen-classes) avec différents classifieurs et deux ensembles de critères

Classification - Expérimentation

	Accuracy	True positives	False positives	True negatives	False negatives
Lieu	0.91	65	4	359	37
Nature	0.91	109	6	313	35
Batiment	0.94	43	6	395	20
Personne	0.91	59	7	363	35
Date	1	9	0.1	454	1
Animal	0.98	0.26	10	454	2
Profession	0.98	0	0	456	8
Guerre	0.99	0	0	458	6
Nourriture	0.99	0	0	462	3
Objet	0.99	3	0	455	6
Mer	0.99	0	0	460	5
Groupe d'humains	0.99	0	0	461	4
Emotion	1	0	0	461	4
Valeur	1	0	0	463	1
Couleur	1	0	0	463	2

Résultats de la classification en binaire sur les classes (bin-classes) avec le classifieur LinearSVC et tous les critères

Développement d'une application web

Rues de France

Méthodologie

General

Lieu

Nature

Batiment

Personne

Date

Animal

Profession

Guerre

Nourriture

Objet

Mer

Groupe d'humains

Région

Département

Recherche commune

☒ LinearSVC

☐ SGDClassifier

Chercher

Développement d'une application web

Rues de France

Méthodologie

General

Lieu

Nature

Batiment

Personne

Date

Animal

Profession

Guerre

Nourriture

Objet

Mer

Groupe d'humains

Auvergne-Rhône-Alpes

RHONE (69)

LYON 2EME

☒ LinearSVC

☐ SGDClassifier

Chercher

Classes les plus représentées

personne (150 instances)

nature (47 instances)

batiment (9 instances)

lieu (3 instances)

date (1 instances)

profession (1 instances)

personne

nature

ADELAIDE PERRIN

ALPHONSE FOCHIER

AMEDEE BONNET

AMPERE

ANDRE MURE

ANTOINE DE SAINT EXUPERY

ANTOINE DELANDINE

ANTOINE RIBOUD

ANTOINE RIVOIRE

ANTOINE SALLES

ANTOINE VOLLON

ANTONIN GOURJU

ANTONIN PONCET

AUGUSTE COMTE

Conclusion

- Analyse des noms de rues en France
 - Détection des **noms de rues équivalents** → ~80% de f-mesure, blocking performant
 - **Catégorisation** des noms de rues → ~70% de prédiction correcte
 - **Interface** de présentation des résultats par niveau géographique
- Perspectives d'amélioration
 - Tests avec d'autres **combinaisons** de mesures pour la déduplication
 - Utilisation du **multi-classes** pour la prédiction
 - **Représentation cartographique** sur l'interface et enrichissement avec d'autres statistiques

Annexe – Critères de classification

Nom du critère	Explications	Valeurs possibles	Origine	Présence de valeurs nulles
nomRue	Nom de la rue	Chaîne de caractères	FANTOIR	non
nbChar	Nombre de caractères dans le nom de la rue	[0, n]	Calculs sur le nom de la rue	non
nbMots	Nombre de mots dans la nom de la rue	[0, n]	Calculs sur le nom de la rue	non
natureVoie	Nature de la voie	[rue, che, rte, av, ...]	FANTOIR	oui
typeVoie	Type de la voie	[voie, ensemble immobilier, lieu-dit, pseudo-voie, voie provisoire]	FANTOIR	non
occBD	Nombre de fois ou le nom exact de la rue est présent dans la base de données	[0, n]	Calculs sur la base de données	non
long	Longueur de la voie.	[0, n]	Source externe	oui ~ 50 %
occMotsBd	Nombre de fois ou les mots présent dans le noms de la voie apparaissent dans la base (fait à partir du fichier de blocking sur la France).	[0, n]	Fichier de blocking	non

Critères principaux utilisés pour la classification

Annexe – Résultats classification

- Méthode **gen-classes**, avec LinearSVC et tous les critères : 0.77 d'exactitude
- Méthode **gen-sc** avec LinearSVC et tous les critères : au mieux 0.8 d'exactitude

	Accuracy	Positives	Negatives
Lieu	0.80	82	20
Nature	0.81	118	26
Batiment	0.84	54	10
Personne	0.55	65	52
Objet	0.95	9	0.5

Résultats de la classification parmi toutes les sous-classes d'une classe (gen-sc) avec le classifieur LinearSVC et tous les critères