

Extraction des données des articles de PLU

Rapport technique – Version du 27/09/2018

Numen Digital

Un des objectifs du projet SmartPLU est l'extraction automatique de données réglementaires dans une liste d'articles de PLU sélectionnés. Afin de valider cette extraction automatique par des méthodes de *machine learning*, un corpus d'apprentissage annoté a été constitué par les opérateurs de Numen Digital. Une fois ce corpus constitué, deux méthodes d'extraction ont été comparées :

- Des modèles purement syntaxiques à base d'expressions régulières, définis manuellement ;
- Des modèles à base de réseaux de neurones profonds, qui permettent un apprentissage automatique à partir de données annotées et qui sont généralement les plus performants lorsque la quantité de données d'entraînement est importante.

Nous présentons dans ce rapport une étude sur l'extraction d'information par ces deux types de modèles sur un corpus de 2490 PLU représentant 12 777 articles de PLU.

1. Constitution d'un corpus annoté

2. Constitution d'un corpus annoté

L'utilisation de techniques d'apprentissage automatique (*machine learning*) nécessite la constitution d'un corpus annoté permettant à la fois d'apprendre les modèles et de les évaluer. Ce corpus doit être suffisamment important car une partie des exemples est utilisée pour l'apprentissage du modèle (*train set*) et une partie **disjointe** de la première est utilisée pour l'évaluation du modèle (*test set*). Le modèle est donc évalué sur des exemples qu'il n'a jamais vus lors de son apprentissage, on peut ainsi évaluer sa capacité à *généraliser* à de nouveaux cas. Le corpus doit être annoté, c'est-à-dire que la tâche que l'on souhaite faire apprendre à la machine doit au préalable avoir été réalisée par des opérateurs humains sur tous les exemples.

Dans notre cas, la tâche cible est l'extraction de données sur un certain nombre d'articles. Cette extraction a donc été réalisée par les opérateurs de Numen sur 2490 PLU. Les expériences présentées dans ce rapport ont été conduites sur l'intégralité de ce corpus.

Les informations à extraire étaient les suivantes :

N'article	Valeur	Description
B1_BANDE	Numérique x >0	
ART_5	Numérique	Minimum parcellaire en m ²
ART_6	Numérique	Distance minimale des constructions par rapport à la voirie en mètre
ART_71	<ul style="list-style-type: none"> ➤ 0 ➤ 1 ➤ 2 	<p>Si dans l'article 7, il y a une règle sur l'implantation en limite séparative :</p> <ul style="list-style-type: none"> ➤ 0 : s'il n'y a aucun retrait imposé par rapport aux limites séparatives ➤ 1 : s'il y a une règle qui impose un retrait de l'implantation par rapport aux limites séparatives ➤ 2 : s'il y a une règle qui impose un retrait par rapport aux limites séparatives mais sur un côté seulement
ART_72	Numérique	<p>Distance minimale des constructions par rapport aux limites séparatives en mètre</p> <p>Si 71=0, 72 = -99</p> <p>Si 71 = 1 ou 2, cette valeur existe</p>
ART_73	Numérique	Distance minimale des constructions par rapport à la limite séparative de fond de parcelle en mètre
ART_74	<ul style="list-style-type: none"> ➤ 1 ➤ 2 	<p>Indique la distance minimale des constructions par rapport à la limite séparative relative à la hauteur de bâtiment</p> <p>1 : retrait même valeur que la hauteur H</p> <p>2 : retrait = moitié de la hauteur H/2</p>
ART_8	Numérique	Distance minimale des constructions par rapport aux autres sur une même parcelle en mètre

ART_9	Valeur $0 \leq x \leq 1$	<p>Coefficient d'emprise au sol, ratio entre 0 et 1</p> <p>Si valeur exprimée en %, convertir en ratio</p> <p>Ex : $5\% = 5/100 = 0,05$; $85\% = 85/100 = 0.85$</p>
ART_10T	<ul style="list-style-type: none"> ➤ 1 ➤ 2 ➤ 3 ➤ 4 ➤ 5 ➤ 6 ➤ 7 ➤ 8 ➤ 9 ➤ 10 ➤ 11 	<p>Unité de mesure de la hauteur des bâtiments</p> <ul style="list-style-type: none"> ➤ 1 : si par rapport au nombre de niveau R ➤ 2 : en m du sol au faîtage ➤ 3 : en m par rapport à la hauteur plafond ➤ 4 : en m du sol au point le plus haut ➤ 5 : en m par rapport à la hauteur de façade à l'égout ➤ 6 : en m par référence à la hauteur NGF hors édifices ➤ 7 : en m par rapport à la hauteur à la côte du trottoir ➤ 8 : en m par rapport au point le plus haut hors cheminées ➤ 9 : en m par rapport au point le plus haut hors cheminées, ouvrages techniques ➤ 10 : en m du sol à l'acrotère ➤ 11 : en m par rapport au point le plus haut tout inclus ➤ 12 : lorsque les références de mesure ne figurent pas dans cette liste <p>S'il y a d'autres références de mesure mentionnées dans le document (ex : du sol par rapport à la sablière), saisir ce critère dans la colonne « Explication anomalie », toujours insérer la valeur dans la colonne « Valeur » et capturer au lasso le texte de l'article.</p>
ART_10	<p>Pour $10T = 1$, valeur = R + nombre entier</p> <p>Pour le reste : Numérique</p>	<p>Hauteur maximale autorisée.</p> <p>Les valeurs pour chaque unité de 10T</p>
ART_12	Numérique	Nombre de places par logement
ART_13	Valeur $0 \leq x \leq 1$	<p>Partie minimale d'espaces libres de toute construction exprimée par rapport à la surface totale de la parcelle.</p> <p>Si valeur exprimée en %, convertir en ratio</p>

		Ex : 5%= 5/100 = 0,05 ; 85%= 85/100 = 0.85
ART_14	Valeur 0 =< x =< 1	Coefficient d'occupation du sol. Si valeur exprimée en %, convertir en ratio Ex : 5%= 5/100 = 0,05 ; 85%= 85/100 = 0.85
N ' a r t i c l e	Vale ur	Descriptions
B 1 - B A N D E	Num ériqu e x >0	
A R T - 5	Num ériqu e	Minimum parcellaire en m ²
A R T - 6	Num ériqu e	Distance minimale des constructions par rapport à la voirie en mètre
A R T - 7 1	<ul style="list-style-type: none"> ➤ 0 ➤ 1 ➤ 2 	<p>Si dans l'article 7, il y a une règle sur l'implantation en limite séparative :</p> <ul style="list-style-type: none"> ➤ 0 : s'il n'y a aucun retrait imposé par rapport aux limites séparatives ➤ 1 : s'il y a une règle qui impose un retrait de l'implantation par rapport aux limites séparatives ➤ 2 : s'il y a une règle qui impose un retrait par rapport aux limites séparatives mais sur un côté seulement
A R T - 7 2	Num ériqu e	<p>Distance minimale des constructions par rapport aux limites séparatives en mètre</p> <p>Si 71=0, 72 = -99</p> <p>Si 71 = 1 ou 2, cette valeur est existe</p>

A R T - 7 3	Num ériqu e	Distance minimale des constructions par rapport à la limite séparative de fond de parcelle en mètre
A R T - 7 4	➤ 1 ➤ 2	Indique la distance minimale des constructions par rapport à la limite séparative relative à la hauteur de bâtiment 1 : retrait même valeur que la hauteur H 2 : retrait = moitié de la hauteur H/2
A R T - 8	Num ériqu e	Distance minimale des constructions par rapport aux autres sur une même parcelle en mètre
A R T - 9	Vale ur 0 =< x =< 1	Coefficient d'emprise au sol, ratio entre 0 et 1 Si valeur exprimée en %, convertir en ratio Ex : 5%= 5/100 = 0,05 ; 85%= 85/100 = 0.85
A R T - 1 0 T	➤ 1 ➤ 2 ➤ 3 ➤ 4 ➤ 5 ➤ 6 ➤ 7 ➤ 8 ➤ 9 ➤ 10 ➤ 11	Unité de mesure de la hauteur des bâtiments ➤ 1 : si par rapport au nombre de niveau R ➤ 2 : en m du sol au faîtage ➤ 3 : en m par rapport à la hauteur plafond ➤ 4 : en m du sol au point le plus haut ➤ 5 : en m par rapport à la hauteur de façade à l'égout ➤ 6 : en m par référence à la hauteur NGF hors édifices ➤ 7 : en m par rapport à la hauteur à la côte du trottoir ➤ 8 : en m par rapport au point le plus haut hors cheminées ➤ 9 : en m par rapport au point le plus haut hors cheminées, ouvrages techniques ➤ 10 : en m du sol à l'acrotère ➤ 11 : en m par rapport au point le plus haut tout inclus ➤ 12 : lorsque les références de mesure ne figurent pas dans cette liste S'il y a d'autres références de mesure mentionné dans le document (ex : du sol par rapport à la sablière), saisir ce critère dans la colonne « Explication anomalie », toujours insérer la valeur dans la colonne

		« Valeur » et capturer au lasso le texte de l'article.
A R T — 1 0	Pour 10T = 1, valeur = R + nombre entier r Pour le reste : Numérique	Hauteur maximale autorisée. Les valeurs pour chaque unité de 10T
A R T — 1 2	Numérique	Nombre de places par logement
A R T — 1 3	Valeur 0 =< x =< 1	Partie minimale d'espaces libres de toute construction exprimée par rapport à la surface totale de la parcelle. Si valeur exprimée en %, convertir en ratio Ex : 5%= 5/100 = 0,05 ; 85%= 85/100 = 0.85
A R T — 1 4	Valeur 0 =< x =< 1	Coefficient d'occupation du sol. Si valeur exprimée en %, convertir en ratio Ex : 5%= 5/100 = 0,05 ; 85%= 85/100 = 0.85

Un formulaire de saisie spécifique aux PLU, présenté sur la Figure 1, a été configuré pour permettre aux opérateurs de saisir les différentes informations :


- type de zone
- numéro d'article
- valeur

- texte source
-

ARTICLE U 1-6: Implantation des constructions par rapport aux voies et emprises publiques

Les installations techniques d'intérêt public devront être implantées à une distance comptée horizontalement de tout point de la construction à édifier au point le plus proche de la limite séparative, au moins égale à 1 mètre.

3. En dehors de ce cas, en l'absence de toute indication contraire figurée sur le plan de zonage précisant la marge de reculement des constructions, toutes les constructions ou installations seront implantées à une distance minimum de 4 mètres de l'alignement actuel ou futur des voies communales ou départementales, si l'élargissement est prévu.



Propriétés du document

Main **rule**

rule

Intitulé de la zone	Code règle	Valeur	Texte source
U1	ART_5	-88	ARTICLEU1-5 SuperficieminimaldesterrainsSansobjet
U1	ART_6	1	ARTICLEU1-6 Implantationdesconstructionsparrapportau

Figure 1: masque de saisie manuelle

Extraction des articles par expressions régulières

La première étape pour la mise en place des différentes méthodes d'extraction des valeurs de règles est l'extraction du texte des articles et l'identification de leur numéro. Nous avons réalisé cette extraction des articles par expression régulière. Un exemple d'expression régulière d'extraction du titre d'un article est :

```
reg = 'ARTICLE|Article|ART |Art |Art.| article UB |^(?!ARTICLE|Article|ART|Art.) U\w?\w? [1-9]'
```

L'évaluation de l'extraction des articles par expressions régulières a été menée par comparaison aux valeurs annotées par les opérateurs. Pour chaque article, 3 cas sont possibles :

- l'article extrait par l'expression régulière correspond à un article annoté : l'extraction est correcte (en vert sur la Figure 2)
- l'article n'a pas été trouvé dans le PLU par l'expression régulière : article non extrait (en jaune sur la Figure 2)
- un article été extraite par l'expression régulière mais il ne correspond à aucun article annoté: article en plus (en rose sur la Figure 2)

La répartition des différentes erreurs possibles pour chaque article analysé est présentée sur la Figure 2.

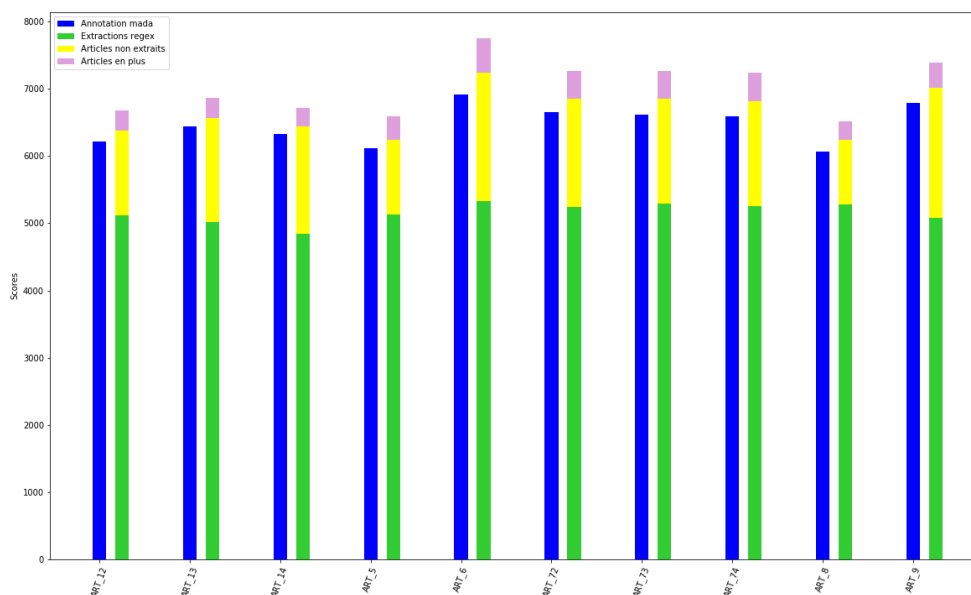


Figure 2: Extraction des articles par expressions régulières

Les valeurs et taux correspondants à la Figure 2 sont les suivants :

Article	Nombre annotés	Nombre extraits	Nombre en plus	Pourcentage extraits
5	6119	5222	354	84 %
6	6915	5451	516	77 %
72	6648	5285	411	79 %
73	6611	5291	401	80 %
74	6585	5277	419	80 %
8	6068	5313	278	87 %
9	6795	5103	373	75 %

12	6212	5201	305	82 %
13	6445	5105	304	79 %
14	6328	4872	266	77 %

Les taux d'extraction correcte varient entre 75% et 87% suivant les articles et se situent principalement autour de 80%. Le nombre d'extractions en plus est assez faible et peut correspondre à des articles oubliés par les annotateurs. Les extractions manquantes correspondent à des articles dont la désignation n'a pas été capturée par l'expression régulière et donc à des titres d'articles avec des formulations atypiques.

Une fois les articles extraits, nous avons testé et comparé deux méthodes pour l'extraction des valeurs : une méthode manuelle basée sur des expressions régulières et une méthode automatique par apprentissage automatique (*machine learning*). Le schéma de comparaison de ces deux méthodes est présenté sur la Figure 3. Les deux méthodes sont présentées et évaluées dans les sections suivantes.

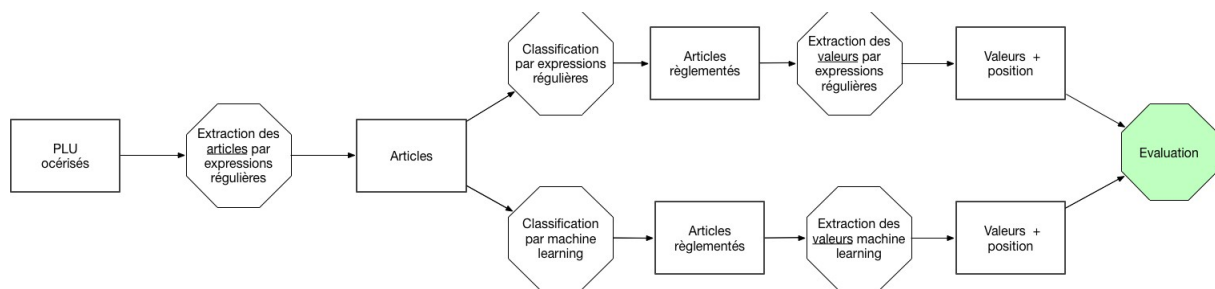


Figure 3: Comparaison de deux méthodes d'extraction : expressions régulières et machine learning

4. Classification des articles réglementés/non réglementés

L'analyse du contenu des PLU nous a permis d'identifier une caractéristique importante des articles : une partie non négligeable d'entre eux n'est pas réglementée. Dans ce cas, il n'y a pas de valeur à extraire. La formulation indiquant qu'un article n'est pas réglementé étant assez stable entre les différents PLU analysés, il nous a semblé intéressant de mettre en place une première classification pour détecter si un article est réglementé ou non. Si l'article n'est pas réglementé, il n'y a pas de valeur à extraire.

Pour la classification, nous avons testé deux types de modèles : un modèle basé sur des expressions régulières similaires à celles présentées dans les sections précédentes et un modèle par apprentissage automatique que nous présentons dans cette section.

Le modèle par apprentissage automatique que nous avons choisi est le système *fasttext*¹ de Facebook qui se base sur une représentation distribuée du texte (*word embedding*) et une classification par réseaux de neurone. Nous avons choisi ce système car il obtient généralement de bons résultats en classification, il combine une représentation distribuée du texte et un classifieur par réseau de neurones et est relativement rapide à entraîner. Ce système a été entraîné sur les articles annotés manuellement par les opérateurs de Numen.

Lors de l'entraînement, le système prend en entrée le texte de l'article et la classe annotée : article réglementé ou non réglementé. Le système apprend alors automatiquement quels sont les mots et les formulations qui permettent de classer un texte dans l'une ou l'autre des

1

catégories. Lors de l'évaluation, le système doit prédire la classe pour des articles qu'il n'a pas vu lors de la phase d'entraînement. Nous avons séparé les données en deux ensembles disjoints pour l'entraînement (111 369 articles) et pour le test (30032 articles).

L'évaluation consiste alors à comparer les prédictions du système aux annotations manuelles et à calculer le taux de prédictions correctes (*accuracy*). Nous avons mené l'étude sur un sous-ensemble des articles annotés pour des contraintes de temps de préparation des données. Cependant, les résultats de cette étude sont généralisables aux autres articles. Les résultats par article sont présentés dans le tableau ci-dessous :

Article	Taux de classification correct	
	Expression Régulière	Machine Learning
6	54%	57%
8	95%	98%
9	91%	96%
12	74%	95%
14	98%	99%

Les modèles par *machine learning* permettent dans tous les cas de dépasser le taux de classification obtenu par les expressions régulières. Il n'est donc pas nécessaire de constituer manuellement un ensemble de règles pour séparer les articles non règlementés des articles règlementés.

5. Extraction des valeurs de règles

Une fois le texte des articles localisés dans les PLU, identifiés avec leur numéro de règle et classé en règlementé/non règlementé, il faut extraire du texte la valeur recherchée. Nous

avons testé deux méthodes pour réaliser cette extraction : une méthode basée sur l'écriture manuelle de règles sous forme d'expressions régulières et une méthode basée sur l'apprentissage automatique d'un réseau de neurones.

Extraction par expression régulière

Nous avons défini manuellement pour chaque article en ensemble de règles d'extraction composé d'expressions régulières afin d'extraire la valeur recherchée.

Un exemple d'expression régulière d'extraction de la valeur article 6, une fois le texte extrait est donné ci-dessous :

```
r1 = 'un recul minimal de|avec un retrait \'au moins|un retrait d\'au moins|un recul de|minimum de|retrait minimum est de'
r2 = r1 +'|mètres minimum par rapport|doit être implantée l[\'|']alignement.|une distance minimale de|au moins égale|'
r2 += 'doivent être implant|Les constructions doivent être édifiées'
r3 = r2 +'|Toute construction nouvelle doit être|est en retrait d\'au moins|mètres au moins de l\'alignement|'
r3 += 'Les constructions peuvent s\'implanter en limite|minimum|t s[\'|']implant'
reg_rules = r3 + '|s\'implanter avec un recul|un retrait compris|mètres'

reg_rulues_extraction = '(\d+|\d+.\d+|\d+)( )?m|distance( minim.{,5})? (de|d\'au moins) (\d+|\d+.\d+|\d+,\d+)|'
reg_rules_extraction += 'au moins égale (\d+|\d+.\d+|\d+,\d+)|inférieure (\d+|\d+.\d+|\d+,\d+)|'
reg_rules_extraction += '|minimum de (\d+|\d+.\d+|\d+,\d+)|un recul de (\d+|\d+.\d+|\d+,\d+)'
```

Ces règles sont constituées manuellement par essai et erreur, ce qui peut prendre plusieurs jours. Ici encore, pour des contraintes de temps de préparation des données et de définition des règles, nous avons évalué le système sur un sous-ensemble des articles. Les performances de chacun de ces modèles sont résumées dans le tableau suivant :

Article	Taux correct			Taux de non règlementé
	Classification	Extraction	Total	
5	99 %	21 %	88 %	87 %
6	54 %	60 %	59 %	20 %
72	31 %	87 %	80 %	18 %
73	98 %	26 %	96 %	98 %

74	84 %	89 %	86 %	51 %
8	95 %	83 %	91 %	65 %
9	91 %	92 %	91 %	73 %
12	74 %	67 %	71 %	54 %
13	96 %	84 %	92 %	67 %
14	98 %	77 %	96 %	87 %

La colonne « classification correcte » reprend les performances en classification (règlementé/non règlementé) décrites dans la section précédente. La colonne « extraction correcte » donne le taux d'extraction par expressions régulières correctement extrait du texte par comparaison à l'annotation manuelle. La colonne « Total » indique le taux d'identification de valeur de règle correcte après classification et extraction. Le taux d'articles non règlementés est donné en dernière colonne. Les performances d'extraction après classification sont généralement assez bonnes, dans 7 cas sur 10 supérieures à 86%.

Extraction par apprentissage automatique

L'extraction des valeurs des règles dans les PLU se rapproche de la tâche d'extraction d'entités nommées bien connue dans la communauté de traitement automatique de la langue naturelle. Dans le cas d'extraction d'entités nommées, à partir d'un texte, il faut extraire les noms de personnes, d'organisations, de lieux, etc. Dans notre cas, il nous faut extraire des entités numériques. Nous avons donc décidé d'entraîner un modèle d'extraction d'entités nommées proposé par la librairie `spacy`² sur nos données. Ce modèle est actuellement parmi les plus performants et le plus simples à mettre en œuvre.

Le modèle d'extraction proposé par `spacy` se base sur les principes suivants :

- Apprentissage de plusieurs représentations distribuées (*word embeddings*) pour chaque mot, basée sur la forme normalisée du mot, le préfixe, le suffixe et sa composition (majuscule/minuscule, numérique/alphabétique). Les représentations sont combinées pour obtenir la représentation de chaque mot.
- Concaténation des *embeddings* de mots pour prendre en compte le contexte
- Mécanisme d'attention pour identifier les caractéristiques importantes

2

- Prédiction de la nature de l'entité

Ces principes sont illustrés sur la Figure 4.

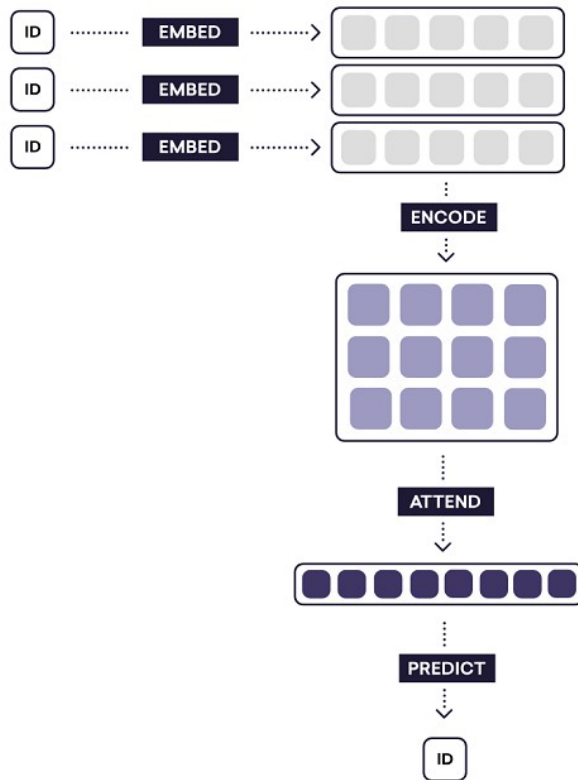


Figure 4: Principe de l'extraction d'entités avec le modèle Spacy.

L'entraînement du modèle d'extraction d'entités nécessite de fournir, lors de l'entraînement, les textes ainsi que les positions (en caractères) des entités à extraire. Pour notre tâche, nous avons donc identifié la position des entités annotées par les opérateurs dans le texte des articles. Les modèles ont été entraînés et testés pour chaque article avec les quantités de données suivantes, en nombre d'articles :

Article	Nombre d'exemples	
	Entrainement	Test
6	5035	1168
8	2408	1006
9	3446	1467

12	1630	723
14	1128	451

Certains articles ont été écartés de l'apprentissage automatique pour des raisons de contraintes de temps de préparation des données et aussi pour des raisons du type d'annotation. En effet, les opérateurs ayant annoté les valeurs à extraire, mais pas les positions (pour des raisons d'efficacité), certaines valeurs étaient difficiles à localiser avec précision pour entraîner les modèles. Nous avons préféré focaliser les expériences sur les articles dont les valeurs étaient correctement localisées.

Comparaison des modèles manuels et automatiques

Nous présentons dans cette section une comparaison des résultats d'extraction des valeurs par les deux types de modèles testés : les modèles manuels à base d'expressions régulières et les modèles automatiques à base de réseaux de neurones. La comparaison est menée sur 5 articles sélectionnés et les taux sont calculés sur un ensemble de test indépendant de l'ensemble d'entraînement (défini pour le modèle automatique) par comparaison aux valeurs annotées par les opérateurs. Les résultats sont présentés dans le tableau suivant :

Article	Modèle	Taux correct		
		Classification	Extraction	Total
6	Manuel	54%	60%	59%
	Automatique	57%	77%	70%
8	Manuel	95%	83%	91%
	Automatique	98%	83%	96%
9	Manuel	91%	92%	91%
	Automatique	96%	76%	91%
12	Manuel	74%	67%	71%

	Automatique	95%	75%	90%
14	Manuel	98%	77%	96%
	Automatique	99%	69%	97%

Le tableau reprend les performances des deux modèles en classification (règlementé/non règlementé), donne les performances en extraction sur les articles règlementés et le taux d'extraction correcte lorsque l'on combine classification et extraction. Les performances du modèle par apprentissage sont toujours supérieures ou égales aux performances du modèle manuel.

6. Conclusion

Cette étude présente l'extraction automatique des valeurs de règles de construction dans les PLU numérisés. Nous avons montré que des modèles par apprentissage automatiques pouvaient être entraînés et permettre d'obtenir des résultats supérieurs à des modèles manuels à base de règles.

Cette étude est une première étape dans l'extraction automatique de données dans les PLU. Les perspectives d'amélioration et de développement sont les suivantes :

- Analyse d'erreur pour proposer des pistes d'améliorations des modèles actuels : une étude précise des causes d'erreur permettra d'identifier les faiblesses de modèles actuels et de proposer des améliorations.
- Reprendre les articles écartés : certains articles ont été écartés de l'étude pour des raisons d'ambiguïté d'annotation ou de valeur à extraire. Ces ambiguïtés peuvent être levées et les articles concernés peuvent aussi être traités automatiquement.
- Estimer des taux d'extraction en automatisation partielle : les mesures réalisées dans cette étude se placent dans le cadre d'une automatisation totale. Cependant, pour obtenir des taux d'erreur faibles, il est souvent nécessaire de combiner des prédictions automatiques et une vérification manuelle sur les cas incertains pour la machine. Le traitement est alors en automatisation partielle, le travail des opérateurs venant en complément de la machine.