Prédiction de l'environnement d'un quartier

Stage de Master 2

Nelly Barret
Université Claude Bernard Lyon 1

Encadrée par Fabien Duchateau et Franck Favetta

26 juin 2020

Contexte du stage

Le projet Home in Love

- Objectif : aider à la recherche immobilière dans le cadre de la mobilité professionnelle
- Projet pluridisciplinaire débuté en 2017

Contexte scientifique

- Comment qualifier simplement l'environnement d'un quartier ?

Objectif du stage

- Prédire l'environnement d'un quartier par apprentissage supervisé

État de l'art

- **Recommandation de logements** [YLKK13]
  - 3 critères : localisation, prix, unité urbaine
  - Utilisation d'une ontologie et du *case-based reasoning*

- **Recommandation de quartiers** [LWSM14]
  - Prise en compte du voisinage
  - Similarité entre quartiers (matrice) et régions (*clustering*)

- **Comparaison manuelle de quartiers** : datafrance.info
  - 5 critères : éducation, santé, services, commerces, loisirs
  - Classement des communes et visualisation cartographique

Yuan, X. et al.: Toward a user-oriented recommendation system for real estate websites. Information Systems (2013).

Liu, Y. et al.: Exploiting geographical neighborhood characteristics for location recommendation. Conference on Information and Knowledge Management (2014).

État de l'art

- **Recommandation de logements** [YLKK13]
  - 3 critères : localisation, prix, unité urbaine
  - Utilisation d'une ontologie et du *case-based reasoning*

- **Recommandation de quartiers** [LWSM14]
  - Prise en compte du voisinage
  - Similarité entre quartiers (matrice) et régions (*clustering*)

- **Comparaison manuelle de quartiers** : datafrance.info
  - 5 critères : éducation, santé, services, commerces, loisirs
  - Classement des communes et visualisation cartographique

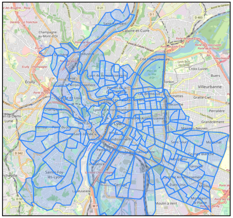
Recommandation de quartiers basée sur une expertise sociologique

---

Yuan, X. et al.: Toward a user-oriented recommendation system for real estate websites. Information Systems (2013).

Liu, Y. et al.: Exploiting geographical neighborhood characteristics for location recommendation. Conference on Information and Knowledge Management (2014).

Prérequis



Quartier [HL07]

- Selon l'INSEE, un IRIS $\rightarrow$ zone de 2000 à 5000 habitants
- En France, 50 000 IRIS

Indicateurs INSEE

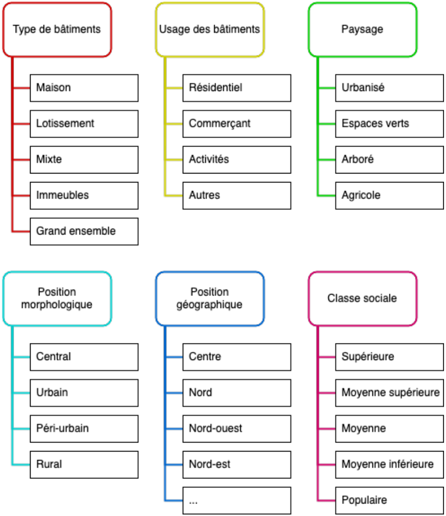
- 641 indicateurs
- Nombre de restaurants, population entre 18 et 25 ans, ...

Variables d'environnement

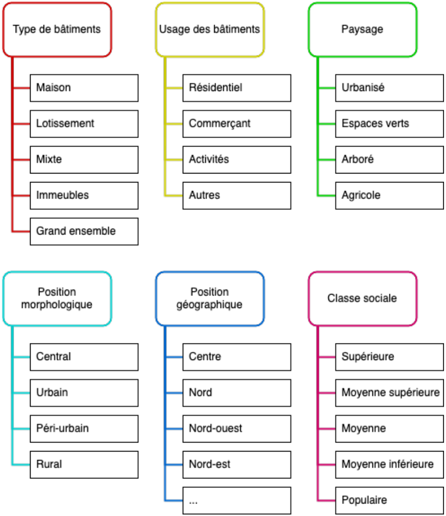
- Description simplifiée d'un quartier par 6 critères
- Résultat d'une analyse qualitative des sociologues sur 300 IRIS

Humain-Lamoure, A.L.: Le quartier comme objet en géographie. La Découverte (2007).

Variables d'environnement

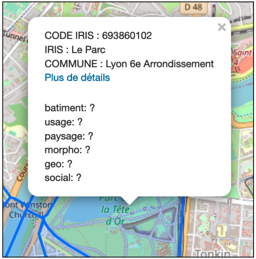


Variables d’environnement

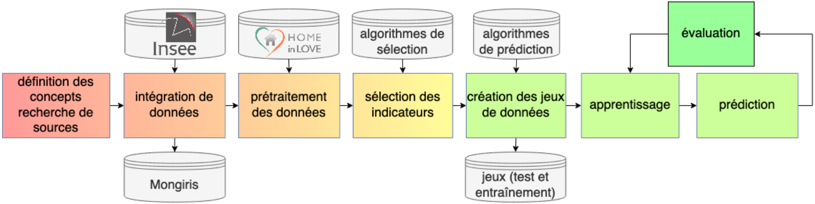


Variables d'environnement

- Type de bâtiments
  - Maison
  - Lotissement
  - Mixte
  - Immeubles
  - Grand ensemble
- Usage des bâtiments
  - Résidentiel
  - Commerçant
  - Activités
  - Autres
- Paysage
  - Urbanisé
  - Espaces verts
  - Arboré
  - Agricole
- Position morphologique
  - Central
  - Urbain
  - Péri-urbain
  - Rural
- Position géographique
  - Centre
  - Nord
  - Nord-ouest
  - Nord-est
  - ...
- Classe sociale
  - Supérieure
  - Moyenne supérieure
  - Moyenne
  - Moyenne inférieure
  - Populaire



L'approche Predihood



Nguyen, G., et al.: Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. AI Review (2019).

Pré-traitement des données

IRIS expertisés

- Transformation des adresses en IRIS
- Traitement des valeurs inconnues : médiane des valeurs
- Traitement des valeurs erronées : semi-automatique

Indicateurs INSEE

- Traitement des valeurs inconnues : médiane des valeurs
- Normalisation : densité de population

Représentativité des IRIS expertisés

Variable	Catégorie	Expertise	France
Position morphologique	rural	5%	68%
Paysage	rural agricole	17%	68%
Classe sociale	moyenne moyenne$^+$	82%	71%
Type de bâtiments	collectif	68%	44%
Position géographique	nord, sud, ...	équitablement répartie	
Usage des bâtiments	nécessite une analyse particulière		

Tableau : L'analyse des IRIS expertisés montre que la plupart des variables d'environnement comportent un biais.

Sélection des indicateurs

Objectif : générer des listes d'indicateurs utiles à la prédiction

Étapes :

- Filtrage des indicateurs
  - 17 sont descriptifs (code postal, code IRIS...)
  - 208 sont trop spécifiques (nombre de courts de tennis couverts)
  - 59 sont non renseignés
- Sélection d'un sous-ensemble parmi 363 indicateurs restants
  - Combinaison d'algorithmes
  - Approche alternative : distribution des indicateurs

Résultat : plusieurs listes de k indicateurs pour chaque variable d'environnement v, notée L_v^k

Sélection par combinaison d'algorithmes

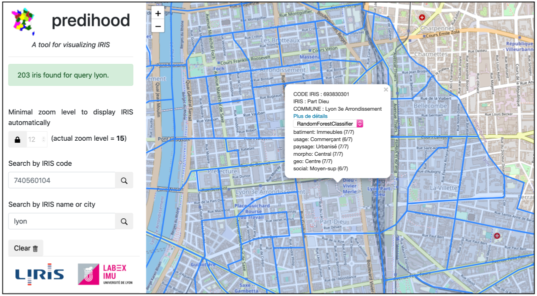
- Matrice de corrélation : supprimer les indicateurs 100% corrélés
- Algorithmes RF et ET : classer les indicateurs par importance
- Prise en compte de la diversité des catégories d'indicateurs

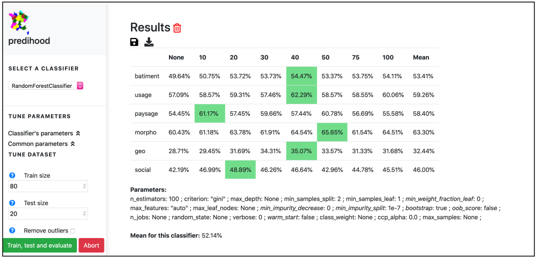
Algorithme 1 : Sélection des indicateurs pertinents pour la prédiction

Entrée : liste d'indicateurs $\mathcal{I}$, liste des variables d'environnement $\mathcal{V}$
Sortie : listes d'indicateurs L_v^k

$C \longleftarrow$ matriceCorrelation($\mathcal{I}$).where(corr $= 1$);
$\mathcal{I} \longleftarrow \mathcal{I} - C$;
for $k \in [10, 20, 30, 40, 50, 75, 100]$ **do**
  for $v \in \mathcal{V}$ **do**
    $L_v \longleftarrow \emptyset$;
    $F_v^{ET} \longleftarrow$ top-k(ET.rank_features($\mathcal{I}$), k);
    $F_v^{RF} \longleftarrow$ top-k(RF.rank_features($\mathcal{I}$), k);
    $F \longleftarrow F_v^{ET} \cup F_v^{RF}$;
    for $f \in F$ **do**
      $p_f \longleftarrow$ parent(f);
      if $p_f \in F$ **then**
        $p_f.score \longleftarrow p_f.score + f.score$;
        $F \longleftarrow F - \{f\}$;
    $L_v^k \longleftarrow F$;

Interface Predihood





	None	10	20	30	40	50	75	100	Mean
batiment	49.64%	50.75%	53.72%	53.73%	54.47%	53.37%	53.75%	54.11%	53.41%
usage	57.09%	58.57%	59.31%	57.46%	62.29%	58.57%	58.55%	60.06%	59.26%
paysage	54.45%	61.17%	57.45%	59.66%	57.44%	60.78%	56.69%	55.58%	58.40%
morpho	60.43%	61.18%	63.78%	61.91%	64.54%	65.65%	61.54%	64.51%	63.30%
geo	28.71%	29.45%	31.69%	34.31%	35.07%	33.57%	31.33%	31.68%	32.44%
social	42.19%	46.99%	48.89%	46.26%	46.64%	42.96%	44.78%	45.51%	46.00%

Parameters:
n_estimators: 100 ; criterion: "gini" ; max_depth: None ; min_samples_split: 2 ; min_samples_leaf: 1 ; *min_weight_fraction_leaf*: 0 ; max_features: "auto" ; max_leaf_nodes: None ; *min_impurity_decrease*: 0 ; *min_impurity_split*: 1e-7 ; *bootstrap*: true ; *oob_score*: false ; n_jobs: None ; random_state: None ; verbose: 0 ; *warm_start*: false ; class_weight: None ; ccp_alpha: 0.0 ; max_samples: None ;

Mean for this classifier: 52.14%

Validation expérimentale

Protocole

Niveau national

- Sélection de 5 algorithmes : *Logistic Regression* (LR), *Random Forest* (RF), *K-Nearest Neighbours* (KNN), *Support Vector Classification* (SVC) et *AdaBoost* (AB)
- Calcul de la précision pour chaque variable selon les 7 listes et les 5 algorithmes

Niveau communal

- Analyse qualitative des résultats prédits pour Lyon
- Effectuée par les sociologues

Résultats au niveau national

	LR	RF	KNN	SVC	AB
$\mathcal{I}$	52.9	64.5	59.3	<u>51.1</u>	55.6
L^{10}	52.6	61.2	**<u>63.8</u>**	49.6	**59.6**
L^{20}	**55.9**	64.1	**63.0**	49.6	**56.6**
L^{30}	51.1	61.2	**62.3**	49.6	**60.8**
L^{40}	**<u>57.8</u>**	63.0	**60.8**	49.2	**56.3**
L^{50}	**56.3**	**<u>64.9</u>**	**62.2**	46.6	**<u>61.1</u>**
L^{75}	50.7	63.4	**60.8**	51.1	**58.2**
L^{100}	**53.7**	64.5	59.3	51.1	55.6

Tableau : Qualité de prédiction pour la variable *usage*

- Les listes améliorent la précision de la majorité des algorithmes

Résultats au niveau national

Variable d'environnement	$\mathcal{I}$	L^k
Type de bâtiments	57%	60% (L^{20})
Usage des bâtiments	64%	65% (L^{50})
Paysage	61%	63% (L^{20})
Classe sociale	51%	52%(L^{40})
Position géographique	34%	33% (L^{40})
Position morphologique	60%	61% ($L^{20,30,40}$)

Tableau : Qualité de prédiction pour l'algorithme Random Forest

- Random Forest obtient les meilleurs résultats
- La sélection permet une meilleure explicabilité des résultats

Conclusion et perspectives

Approche Predihood

- Algorithmes pour la prédiction de l'environnement (sélection des indicateurs et représentativité des données)
- Interfaces pour la visualisation des quartiers et le paramétrage d'algorithmes

Perspectives

- Sélectionner les indicateurs selon leur distribution
- Augmenter le jeu de données (avec les sociologues)
- Calculer la position géographique
- Intégrer de nouvelles sources de données (e.g. points d'intérêt)

Deuxième proposition

Distribution des indicateurs

Algorithme de sélection à partir de la distribution des indicateurs :

- Sélection des indicateurs discriminants
- Indépendant des variables d'environnement

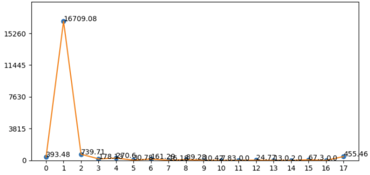


Figure: La Doua

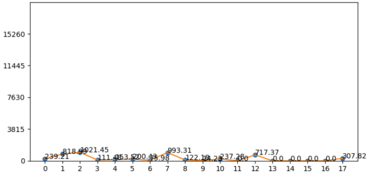


Figure: Saint-Cyr au Mont d'or