Predicting the environment of a neighbourhood: a use case for France

Nelly Barret, Fabien Duchateau, Franck Favetta, Loïc Bonneval

LIRIS, Université de Lyon, France
Centre Max Weber, Université de Lyon, France

July 8, 2020

Introduction and challenges

Context

Have you ever searched for an accommodation in a city you don't know?

Or wondered whether the neighbourhood is suitable for you?

Introduction and challenges

Context

Have you ever searched for an accommodation in a city you don't know?

Or wondered whether the neighbourhood is suitable for you?

Challenges:

- How to simply describe the environment of a neighbourhood?
- How to predict the environment of a neighbourhood?

Introduction and challenges

State of the art

Three categories:

- Gathering data: HomeInLove.com, DataFrance.info, KelQuartier.com
- Comparison between neighbourhoods [ZNSM13], [FGM15]
- Prediction and recommendation [CSHS12], [YLKK13]

Our Predihood proposition for predicting neighbourhood's environment:

- Based on a social study
- Focus on environment instead of life quality
- Prediction for a whole country (currently France)

Zhang, et al., *Hoodsquare: Modeling and recommending neighborhoods in location-based social networks*, SC, 2013.

Le Falher, et al., *Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities*, CWSM, 2015.

Cranshaw, et al., *The livehoods project: Utilizing social media to understand the dynamics of a city*, CWSM, 2012.

Yuan, et al., *Toward a user-oriented recommendation system for real estate websites*, IS (2013).

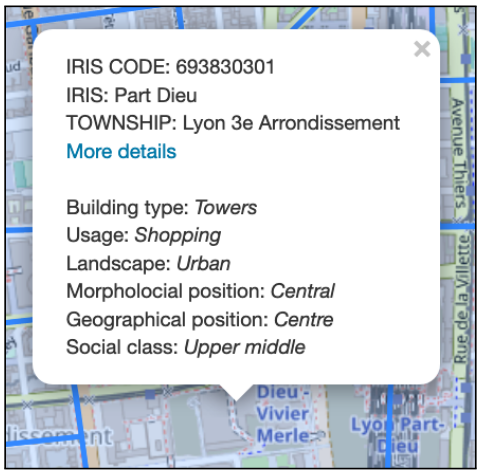
- Neighbourhood
  - Variable definition according to the point of view
  - Use of IRIS units, defined by the National Institute of Statistics
  - Each IRIS includes 650 indicators (e.g. number of bakeries, average income, ...)

IRIS in Paris.

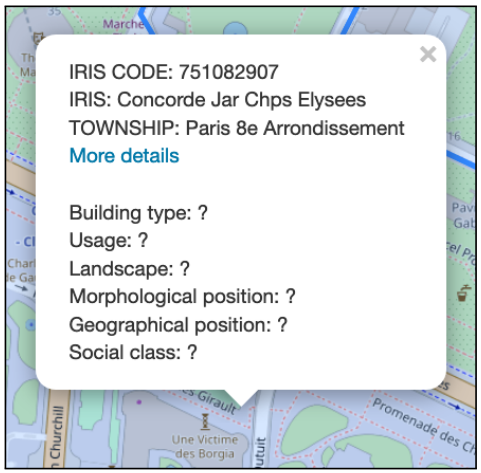
- Six environment variables
  - Describe simply the environment of a neighbourhood
  - Defined by social sciences researchers

Building type	Usage	Landscape	Social class	Morphological	Geographical
Housing estates	Housing	Urban	Lower	Central	Centre
Mixed	Shopping	Green areas	Lower middle	Urban	North
Towers	Other activities	Forest	Middle	Peri-urban	North East
Housing subdivisions		Countryside	Upper middle	Rural	East
Houses			Upper		...

- 50,000 IRIS for the whole country
- 300 IRIS annotated (with environment variables) by social sciences researchers

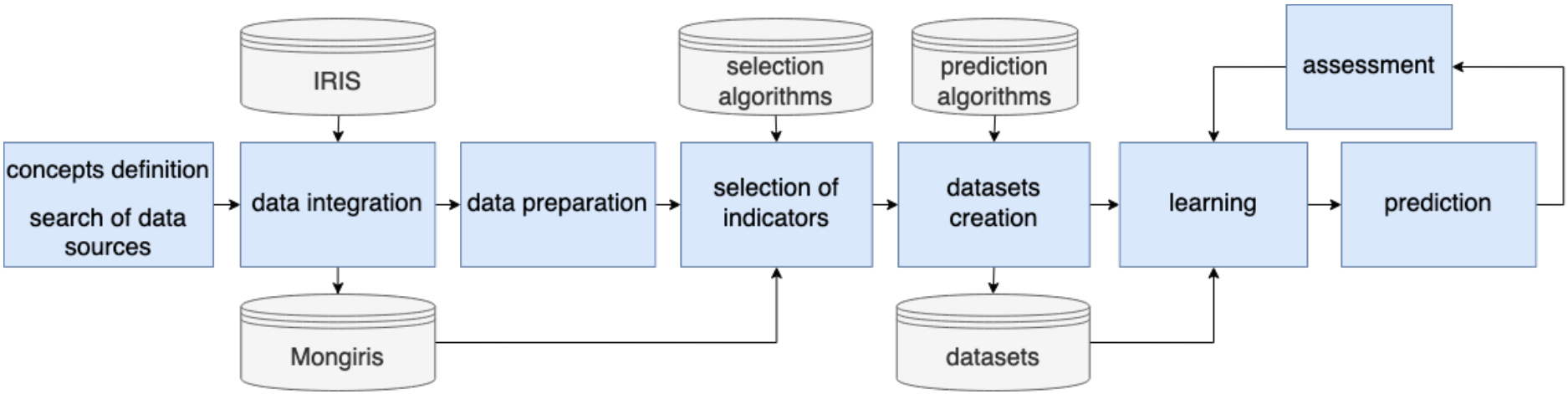


An annotated IRIS.



An unknown environment.

How to automatically annotate the environment of a neighbourhood?



Overview of the Predihood approach, based on the CRISP methodology. [NDB+19]

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

Predihood approach

Data integration

- Replace unknown values by the median value of the indicator
- Normalize indicators with population density
- Compute grouped indicators (aggregation of attributes)
- Store consolidated data into MongoDB (*mongiris* database)


```
{
   "_id":"ObjectId("    "5be32b2cf3f0b960b1f83643"    ")",
   "geometry":{ },
   "type":"Feature",
   "properties":{
      "IRIS":"2907",
      "NOM_IRIS":"Concorde Jar Chps Elysees",
      "grouped_indicators":{ },
      "TYP_IRIS":"D",
      "DEP":"75",
      "raw_indicators":{ },
      "INSEE_COM":"75108",
      "NOM_COM":"Paris 8e Arrondissement",
      "REG":"11",
      "CODE_IRIS":"751082907",
      "GRD_QUART":"7510829",
      "LAB_IRIS":"3",
      "LIBIRIS":"Concorde Jar Chps Elysees",
      "MODIF_IRIS":"0",
      "TRIRIS":"750401",
      "UU2010":"00851"
   }
}
```


Attributes for the Concorde area in Paris.

Predihood approach

Representativeness

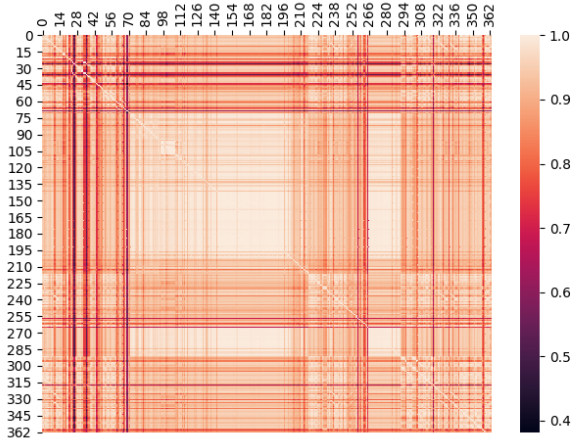
Variable	**Category**	**Expertise**	**France**
Morphological position	rural	5%	68%
Landscape	forest countryside	17%	68%
Social class	middle upper middle	82%	71%
Type of buildings	houses, towers, ...	68%	56 %
Geographical position	north, south, ...	equally distributed	
Usage	requires a specific analysis		

Some variables are biased due to the nature of data

Predihood approach

Feature selection

1. Filter indicators (i.e. descriptive, too detailed, fully empty)
2. Remove indicators that are 100% correlated (Spearman correlation matrix)
3. Rank features by importance and select the k best
4. Integrate the diversity of indicators by merging child with parent



Seven lists L_v^k with different sizes k, for each environment variable v

Experimental validation

Protocol

1. 5 algorithms: *Logistic Regression* (LR), *Random Forest* (RF), *K-Nearest Neighbours* (KNN), *Support Vector Classification* (SVC) and *AdaBoost* (AB)
2. Tuning each algorithm by testing different configurations
3. Set cross-validation with 80/20 distribution
4. Compute accuracy for each environment variable, each list of selected indicators and each algorithm

Experimental validation

Quality results

Usage
Housing
Shopping
Other activities

	LR	RF	KNN	SVC	AB
$\mathcal{I}$	52.9	64.5	59.3	51.1	55.6
L^{10}	52.6	61.2	**63.8**	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}	**57.8**	63.0	**60.8**	49.2	**56.3**
L^{50}	**56.3**	**64.9**	**62.2**	46.6	**61.1**
L^{75}	50.7	63.4	**60.8**	51.1	**58.2**
L^{100}	**53.7**	64.5	59.3	51.1	55.6

Quality of prediction for usage variable (%).

Our lists of indicators improve accuracy for most algorithms

Experimental validation

Quality results

Environment variable	$\mathcal{I}$	L^k
Building type	57%	60% (L^{20})
Usage	64%	65% (L^{50})
Landscape	61%	63% (L^{20})
Social class	51%	52%(L^{40})
Geographical position	34%	33% (L^{40})
Morphological position	60%	61% ($L^{20,30,40}$)

Quality of prediction for the Random Forest algorithm.

- Random Forest obtains the best results for every variable
- Feature selection allows a better interpretation of results

Conclusion

Our Predihood approach for predicting the environment of any French neighbourhood through 6 descriptive variables:

1. Representativeness of our annotated neighbourhoods
2. An algorithm for selecting lists of indicators for each variable
3. Experimental validation showing the benefits of our selection
4. Interfaces for visualizing IRIS and for configuring prediction algorithms

Conclusion

Our Predihood approach for predicting the environment of any French neighbourhood through 6 descriptive variables:

1. Representativeness of our annotated neighbourhoods
2. An algorithm for selecting lists of indicators for each variable
3. Experimental validation showing the benefits of our selection
4. Interfaces for visualizing IRIS and for configuring prediction algorithms

Perspectives:

- Increase the amount of data by using Predihood for annotating
- Integrate new data sources (e.g. points of interest, prices, ...)
- Study correlation between environment variables
- Compute geographical position instead of predicting it