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

Or wondered whether the neighbourhood is suitable for you?
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)

Introduction and challenges
Concepts definition

- 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, ...)

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

<table>
<thead>
<tr>
<th>Building type</th>
<th>Usage</th>
<th>Landscape</th>
<th>Social class</th>
<th>Morphological</th>
<th>Geographical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing estates</td>
<td>Housing</td>
<td>Urban</td>
<td>Lower</td>
<td>Central</td>
<td>Centre</td>
</tr>
<tr>
<td>Mixed Towers</td>
<td>Shopping</td>
<td>Green areas</td>
<td>Lower middle</td>
<td>Urban</td>
<td>North</td>
</tr>
<tr>
<td>Housing subdivisions</td>
<td>Other activities</td>
<td>Forest</td>
<td>Middle</td>
<td>Peri-urban</td>
<td>North East</td>
</tr>
<tr>
<td>Houses</td>
<td></td>
<td>Countryside</td>
<td>Upper</td>
<td>Rural</td>
<td>East</td>
</tr>
</tbody>
</table>
Introduction and challenges
Concepts definition

- 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?
Introduction and challenges

Overview of Predihood

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

Introduction and challenges
- Context
- State of the art
- Concepts definition
- Overview of Predihood

The Predihood approach
- Data integration
- Representativeness
- Feature selection

Experimental validation
- Protocol
- Quality results

Conclusion
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)

```json
{
    "_id": "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.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>Expertise</th>
<th>France</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological position</td>
<td>rural</td>
<td>5%</td>
<td>68%</td>
</tr>
<tr>
<td>Landscape</td>
<td>forest countryside</td>
<td>17%</td>
<td>68%</td>
</tr>
<tr>
<td>Social class</td>
<td>middle, upper middle</td>
<td>82%</td>
<td>71%</td>
</tr>
<tr>
<td>Type of buildings</td>
<td>houses, towers, ...</td>
<td>68%</td>
<td>56%</td>
</tr>
<tr>
<td>Geographical position</td>
<td>north, south, ...</td>
<td>equally distributed</td>
<td></td>
</tr>
<tr>
<td>Usage</td>
<td>requires a specific analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some variables are biased due to the nature of data.
Predihood approach

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 $L^k_v$ with different sizes $k$, for each environment variable $v$
1 Introduction and challenges
- Context
- State of the art
- Concepts definition
- Overview of Predihood

2 The Predihood approach
- Data integration
- Representativeness
- Feature selection

3 Experimental validation
- Protocol
- Quality results

4 Conclusion
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

<table>
<thead>
<tr>
<th>Usage</th>
<th>LR</th>
<th>RF</th>
<th>KNN</th>
<th>SVC</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing</td>
<td>52.9</td>
<td>64.5</td>
<td>59.3</td>
<td>51.1</td>
<td>55.6</td>
</tr>
<tr>
<td>Shopping</td>
<td>52.6</td>
<td>61.2</td>
<td>63.8</td>
<td>49.6</td>
<td>59.6</td>
</tr>
<tr>
<td>Other activities</td>
<td>55.9</td>
<td>64.1</td>
<td>63.0</td>
<td>49.6</td>
<td>56.6</td>
</tr>
<tr>
<td></td>
<td>51.1</td>
<td>61.2</td>
<td>62.3</td>
<td>49.6</td>
<td>60.8</td>
</tr>
<tr>
<td></td>
<td><strong>57.8</strong></td>
<td>63.0</td>
<td><strong>60.8</strong></td>
<td>49.2</td>
<td><strong>56.3</strong></td>
</tr>
<tr>
<td></td>
<td><strong>56.3</strong></td>
<td><strong>64.9</strong></td>
<td><strong>62.2</strong></td>
<td>46.6</td>
<td><strong>61.1</strong></td>
</tr>
<tr>
<td></td>
<td>50.7</td>
<td>63.4</td>
<td><strong>60.8</strong></td>
<td>51.1</td>
<td><strong>58.2</strong></td>
</tr>
<tr>
<td></td>
<td><strong>53.7</strong></td>
<td>64.5</td>
<td>59.3</td>
<td>51.1</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Quality of prediction for usage variable (%).

Our lists of indicators improve accuracy for most algorithms.
Experimental validation

Quality results

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

*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
1 Introduction and challenges
- Context
- State of the art
- Concepts definition
- Overview of Predihood

2 The Predihood approach
- Data integration
- Representativeness
- Feature selection

3 Experimental validation
- Protocol
- Quality results

4 Conclusion
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
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