

# Predicting the Environment of a Neighborhood: A Use Case for France

Nelly Barret<sup>1</sup><sup>a</sup>, Fabien Duchateau<sup>1</sup><sup>b</sup>, Franck Favetta<sup>1</sup><sup>c</sup> and Loïc Bonneval<sup>2</sup>

<sup>1</sup>LIRIS UMR5205, Université de Lyon, UCBL, Lyon, France

<sup>2</sup>Centre Max Weber, Université de Lyon, France

**Keywords:** Data Science, Machine Learning, Data Integration, Environment Prediction, Neighbourhood Study.

**Abstract:** Notion of neighbourhoods is critical in many applications such as social studies, cultural heritage management, urban planning or environment impact on health. Two main challenges deal with the definition and representation of this spatial concept and with the gathering of descriptive data on a large area (country). In this paper, we present a use case in the context of real estate search for representing French neighbourhoods in a uniform manner, using a few environment variables (e.g., building type, social class). Since it is not possible to manually classify all neighbourhoods, our objective is to automatically predict this new information.

## 1 INTRODUCTION

Data science is a recent discipline which aims at exploiting data for producing new knowledge. It is at the crossroads of data management, statistics, machine learning and visualization. It is also usually associated with Big Data, as the amount of available information grows exponentially every year. Two main tasks for data scientists involve data preparation and prediction on observations through the detection of patterns (Dhar, 2013). Numerous application domains benefit from this discipline: health, transport, environment, media, biology, astronomy – to only cite a few. Relating data science experiences enable to demonstrate the feasibility of such projects, the importance of main tasks such as data preparation and the quality of predictive models.


In this paper, we are interested in the **study of territories**, a long-standing research topic which has recently gained more attention with the emergence of *smart cities* (Caragliu et al., 2011). Many works in Digital Humanities aim at studying neighbourhoods, since this spatial delimitation enables the detection of local trends (Delmelle, 2015). Our work is also related to neighbourhoods in the context of real estate search, especially when moving to a new city (e.g., job transfer). Indeed, people who look for renting or buying an accommodation may not necessarily have


prior knowledge about their future city or area. Thus it is difficult to choose a suitable neighbourhood to live in. One may look for a vibrant neighbourhood with many pubs while other may prefer a quiet residential area close to schools and parks. Although several works were designed to tackle this challenge (Yuan et al., 2013; Tang and Sangani, 2015; Le Falher et al., 2015; Barret et al., 2019), most of them are either dedicated to a few cities, or the predicted criteria are limited to have a clear understanding of the neighbourhood. Thus, our objective is to characterize neighbourhoods according to a few criteria such as social class (e.g., lower, middle) or morphological position (e.g., urban, rural).


Our paper describes a **use case for predicting the environment of a neighbourhood** in France. To reach this goal, the main challenging processes of data science are required. We first introduce related work (Section 2). Next, we present our method for modelling, collecting and integrating data about neighbourhoods (Section 3), and we describe our choices for prediction (Section 4). In Section 5, preliminary experiment results are detailed and analysed. Section 6 concludes and highlights perspectives.

## 2 RELATED WORK

Multiple projects focus on studying neighbourhoods. A recent paper shows that the definition of the area perceived as neighbourhood is different according to the point of view (e.g., administrative, from locals,

<sup>a</sup> <https://orcid.org/0000-0002-3469-4149>

<sup>b</sup> <https://orcid.org/0000-0001-6803-917X>

<sup>c</sup> <https://orcid.org/0000-0003-2039-3481>

economic), and consequently, its delimitation are not completely fixed (Bonneval et al., 2019).

**Gathering data collection** is a critical issue and most projects or application do not detail this process. The website DataFrance<sup>1</sup> integrates data from diverse sources, such as indicators provided by the National Institute of Statistics (INSEE), geographical information from the National Geographic Institute (IGN) and surveys from newspapers for prices (L'Express).

**Comparison between neighbourhoods** is performed in many works. Cosine similarity and Jaccard metric are the most used methods to perform such comparison (Yu et al., 2016). They enable a direct computation of the similarity degree between two spatial areas described as vectors of values. For instance, authors of HoodSquare exploit Foursquare check-ins, place types (e.g., restaurant, office, entertainment) and temporal information to detect neighbourhood boundaries and similar areas (Zhang et al., 2013). The work from Le Falher et al. discovers similar neighbourhoods between cities (Le Falher et al., 2015). To reach this goal, they use classification algorithms applied on social networks data.

**Prediction and recommendation** are main objectives when working with neighbourhoods. The study from Tang et al. compare Airbnb announcements in San Francisco to determine their price and neighbourhood (Tang and Sangani, 2015). The VizLIRIS application uses machine learning to detect similar areas in France, which is convenient when moving to a new location (Barret et al., 2019). Besides, it includes a grouping functionality for displaying similar neighbourhoods in a selected area or city. In South Korea, finding the most relevant neighbourhood and accommodation is based on similar user profiles (Yuan et al., 2013). For instance, the household composition and the distance from home to work are part of these profiles. Case-based reasoning is performed to associate a new profile to existing ones, and thus to adjust recommendations. Finally, the objective of the Livehoods project is to deduce city's dynamics from its resident's behaviour (Cranshaw et al., 2012). Experiments in Pittsburgh, using 18 million check-ins and validated by 27 interviews, have confirmed that municipal districts have a different shape than a representation based on their usage.

**Several applications**, which are closer to the context of this paper, produce neighbourhood recommendations. The following list is non exhaustive and centred on France. Kelquartier<sup>2</sup> describes the main French cities using quantitative criteria (e.g., average income,

density of schools, density of shops). Home in Love<sup>3</sup>, vivrou<sup>4</sup> and Cityzia<sup>5</sup> are more oriented towards users as they take into account itineraries (e.g., from and to work) or life style. All aim at recommending the most relevant neighbourhood(s). Finally, ville-ideale<sup>6</sup> is a collaborative website for evaluating French cities. Users give a score (out of 10) for each of the ten categories, from healthcare to security or culture. Although not at a fine-grained level, user comments may include mentions of neighbourhood, which is useful for a (manual) estimation of its quality.

**Positioning.** Our contribution differs from existing works on several points. First, some works are limited to a few cities, which is not possible in a context of job transfer. Our approach should work for a whole country. Although exploiting user profiles is interesting, they are not always available. Relying on social data implies prior analysis in order to avoid bias (e.g., over-represented class of people or activities). Most approaches focus on life quality while our goal is to describe environment. Finally, neighbourhoods may be described using tens or hundreds of criteria, which makes it difficult both for obtaining a simple representation of the area and for explaining or proving justifications about recommendations. Figure 1 summarizes the main steps of our proposition, from data preparation (concepts definition, data gathering and integration, presented in Section 3) and prediction (representativeness, feature selection and algorithm execution, described in Section 4).

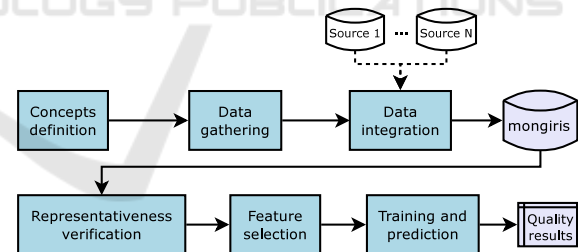


Figure 1: The main steps of our approach.

### 3 DATA PREPARATION

In his study about data science, Donoho estimates that data scientists spend 80% effort to prepare data (Donoho, 2017). Indeed this process consists in determining the main concepts to be used, gathering and extracting information that describe these concepts and integrating them into a single database.

<sup>3</sup><http://homeinlove.fr/>

<sup>4</sup><http://www.vivrou.com/>

<sup>5</sup><http://www.cityzia.fr/>

<sup>6</sup><http://www.ville-ideale.fr/>

<sup>1</sup><http://datafrance.info/>

<sup>2</sup><http://www.kelquartier.com/>

### 3.1 Concepts Definition

The first concept that needs to be defined is the neighbourhood (Bonneval et al., 2019). The definition of this spatial object clearly depends on the point of view: administrative people could refer to a voting or cadastral definition while geographers could rely on natural borders. Residents also have their own boundaries in mind and an economic point of view can span over several neighbourhoods. Given these constraints, we have chosen to use **IRIS<sup>7</sup> as neighbourhoods**, i.e., small division units of the French territory defined for statistical purposes (with about 2,000 residents, thus mainly small-sized in cities and wider in rural areas). They are defined by INSEE, the National Institute of Statistics, which ensures a certain data quality and frequent updates. In addition to their official nature, the main advantage is that each IRIS includes many indicators such as the average income, the number of bakeries, the number of buildings built before 1950 or the percentage of residents per socio-economic category. These indicators could serve as features for the prediction part. There are around 49,800 IRIS in France, with an average of 550 indicators per IRIS. In the rest of this paper, we use either IRIS or neighbourhood with the same meaning.

To clarify the representation of neighbourhoods, we have collaborated with sociologists who have defined **six environment variables<sup>8</sup>**, each with its possible values. They are summarized in Table 1. *Type of building* represents the most common buildings in the neighbourhood (from large housing complexes to individual houses). *Usage* describes local activities while *landscape* defines the quantity of surrounding natural elements. *Social class* denotes the degree of wealth. *Morphological position* indicates the distance of an IRIS from a city centre. And the *geographical position* (nine values) stands for the direction towards the city centre of the closest city<sup>9</sup>. The objective of these variables is to facilitate the description of a neighbourhood in the context of a real estate search, and to enable the comparison of neighbourhoods in social sciences studies for instance. By investigating scientific literature, IRIS data, online information such as <http://www.kelquartier.com/> and street views, social science researchers have annotated so far 270 IRIS using these six variables. Note that this manual

<sup>7</sup><http://www.insee.fr/en/metadonnees/definition/c1523/>

<sup>8</sup>In machine learning, variables usually represent features while outcome is referred to as target values and classes. However, we keep the term *variables* for classes in this paper for consistency with social sciences.

<sup>9</sup>Geographical position can bear implicit knowledge, e.g., East part of cities were traditionally poorer due to industrial pollution coming from West winds.

assessment requires at least 1 to 2 hours (per IRIS) when done properly.

### 3.2 Data Gathering

The main notions of neighbourhood and environment variables have been defined. A second challenge deals with the **collection of relevant data about neighbourhoods**, which results from various discussions with social science researchers. In the era of smart cities, open data become more and more available (Ojo et al., 2015). However, it is still necessary to check for data quality (e.g., provenance, frequency of updates, usage in other projects). Our choice of IRIS as neighbourhoods was also supported by the fact they come with many indicators about population, buildings, shops, leisures, education, etc. First, each neighbourhood includes 17 descriptive characteristics (identifier, IRIS name, city name, postcode, administrative department, administrative region, type, etc.). These indicators are mostly useful for visualization. The remaining hundreds of indicators are either quantities (e.g., number of bakeries, of elementary schools, of tennis courts), unit quantities (e.g., average income, average income for the agricultural class), socio-economic coefficients (e.g., Gini coefficient<sup>10</sup>, S80/S20 ratio<sup>11</sup>), percentages (e.g., percentage of unemployed people, percentage of fiscal households) or string values (e.g. notes about incomes). In addition, each IRIS has a geometry (i.e., list of coordinates delimiting a polygon), which is useful for cartographic visualization. From this geometry, it is possible to compute the surface of the neighbourhood, an important feature. Indeed, cities are usually made of small IRIS while rural areas have larger IRIS.

Once the data sources have been identified and data extracted (using dumps, API, queries), **data needs to be cleaned** because it may contain anomalies, inconsistencies or missing values. This process is often referred to as data cleaning or data wrangling (Donoho, 2017). In our case, a few IRIS have incorrect boundaries (e.g., overlapping edges in their geometries) and they have been corrected using GIS tools. Another problem is related to unknown values, which are globally solved during the next step.

### 3.3 Data Integration

Relevant data sources have been identified, but they are still scattered around and heterogeneous. Data integration aims at **centralizing merged data** and pro-

<sup>10</sup>[http://en.wikipedia.org/wiki/Gini\\_coefficient](http://en.wikipedia.org/wiki/Gini_coefficient)

<sup>11</sup><https://www.insee.fr/en/metadonnees/definition/c1666>

Table 1: Environment variables and their possible values.

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		...

viding an uniform query access (Halevy et al., 2006). IRIS data is spread in tens of CSV files (one for population, another one for education, etc.), produced at different periods and by different persons. Besides, they are not organized or structured in the same manner (different interpretation of a concept, label heterogeneity, grouping or splitting of IRIS, etc.). Several processes are required when integrating data. First schema or ontology matching needs to be performed in order to detect correspondences between concepts or metadata (Bellahsène et al., 2011). In our context, data sources have almost no overlapping and this matching task is manual. For instance, renaming headers in CSV files solves label heterogeneity. Another important process is record linkage or data matching (Christen, 2012; Shen et al., 2015), which consists in detecting equivalent information (e.g., tuples, entities, values). It avoids redundancies and facilitates merging. Although IRIS have an identifier, the fact that some of them were merged or split between two files was a challenge. A script has been written to detect missing IRIS and changes in specific attributes (IRIS name, city name), and the decision to discard or add an IRIS was manual.

During integration, we have also **created a new attribute** labelled *grouped indicators*, which reflects the content of a neighbourhood with a higher level of abstraction. For example, the grouped indicator *health* sums up the number of doctors, pharmacies, hospitals, etc. Local commerces (which exclude large supermarkets) aggregates the number of bakeries, butcheries, open markets, etc. In total, thirteen grouped indicators have been defined and added as features for each IRIS. The surface of polygons is also computed during this step. Unknown values have been replaced by the median score of the column: zero values are not acceptable (specific meaning that an IRIS does not have a given feature) and the average is more sensitive to outliers. The last issue is the difference of units and meaning between indicators (e.g., quantities, percentages, quantiles). Some classification algorithms require comparable information. Social science collaborators suggested that population and population density were the most relevant normalization factors. Both the size and the number of residents have an impact on the characteristics of

an IRIS (e.g., two neighbourhoods may have 5,000 residents, but one of them is a large rural area around a village while the other is a small city area). Consequently, **all indicators have been normalized** according to the population density.

Since IRIS are spatial objects, we have chosen the GeoJSON format<sup>12</sup> to store them. In the end, we obtain a **consolidated MongoDB database** named *mongiris*<sup>13</sup>. It contains 49,800 French neighbourhoods fully covering the country along with integrated data to describe them. A Python API is also provided to facilitate the querying of the database (e.g., retrieve an IRIS from its code, get a list of all neighbours).

## 4 PREDICTION

Relevant data has been collected and aggregated into the *mongiris* database. It contains about 49,800 neighbourhoods, among which 270 have been expertised (i.e., their six environment variables have been filled in by social science researchers). The objective is to predict values of the environment variables for the remaining neighbourhoods. We face two main challenges: the former is the number of expertised IRIS (270) with regards to the total number (49,800), only 0.6%, so it is important to check whether the annotated neighbourhoods are sufficiently representative of the complete set. The latter deals with the high number of indicators (550 in average for each IRIS), which may negatively impact prediction due to overfitting. This section presents our solutions to tackle these challenges.

### 4.1 Representativeness

Given the ratio between annotated IRIS and the remaining ones (0.6%), we perform a quick analysis to check for representativeness. Indeed, prediction results may also be explained when the quantity of examples is not sufficient to represent the whole dataset. Yet, it is very difficult to compute this representativeness and we have chosen to study some of our environment variables.

<sup>12</sup><http://geojson.org/>

<sup>13</sup><http://gitlab.liris.cnrs.fr/fduchate/mongiris>

The **morphological variable** indicates whether a neighbourhood is inside or far from a city. According to the IRIS definition<sup>8</sup>, 16100 IRIS were constructed using cities with more than 10,000 inhabitants and most towns with more than 5,000 residents. To cover the rest of the territory, one IRIS was created for each remaining town, resulting in more than 33,000 additional neighbourhoods. If we consider that these sparsely populated places are rural areas, we estimate that 68% of IRIS are rural in the whole dataset. However, 14 out of the 270 annotated IRIS are considered as rural, thus representing only 5%. This difference can be easily explained by the fact that in our context of job transfer, people tend to leave small towns to the benefit of larger cities. Thus there is a bias for this morphological variable.

The **landscape variable** is closely related to the morphology. We roughly assume that urban and green areas are usually found in cities while forestry and countryside are mainly linked to rural. These last two categories describe 46 of our annotated IRIS, thus representing 17%. We are very far from the 68% expected in France, so this variable is biased too.

The **social class variable** is difficult to analyse, especially because the class definitions are not clear. In France, 59% of households belong to the middle class, including lower and upper middles (Bigot et al., 2011). A commonly accepted definition of middle class consists of incomes ranging from 70% to 150% of the median income, which represent 71% of IRIS. Since 82% of our annotated neighbourhoods are in the middle class, this denotes a slight bias towards the whole dataset.

The **geographical variable** has a more balanced distribution. There are around 25 IRIS for most values. But a few directions, such as *centre*, *south* and *north*, have twice more IRIS. The case of *centre* is understandable due to its correlation with the central morphology, but more research in social science is needed to explain others.

Variables **building type** and **usage** cannot be directly studied.

## 4.2 Feature Selection

In the popular book from Lillesand et al., authors propose that "a rule about the relationship between training sample size  $n$  and data dimensionality  $p$  is that  $n$  lies between  $10p$  and  $100p$ " (Lillesand et al., 2015). In our context, we have 270 samples and an average of 550 indicators, while a **reasonable number of features** should be between 3 and 27. To tackle this challenge (i.e., too many features), we first remove indicators that are not useful for machine learning:

17 descriptive characteristics (e.g., city name, IRIS name), 59 empty or invariant features, and 213 over-detailed indicators (e.g., only "tennis courts" is kept as feature while "tennis courts with at least one covered" and "tennis courts with at least one lighted" are discarded). The 647 original indicators have been reduced to 362 features (55% of the original). Yet, this number is still quite high.

The next option for removing features is to **check their correlation** (Bruce and Bruce, 2017). Indeed, when two features are strongly correlated, they lead to the same trend during learning. Figure 2 depicts a correlation matrix for the 362 indicators, computed using Spearman coefficient (Mukaka, 2012). The darker a point is, the least correlation between the corresponding two indicators. Only a few of them are not much correlated with the others (dark lines), and the majority have a strong correlation (white areas). Indicators with a perfect correlation to another one are deleted.

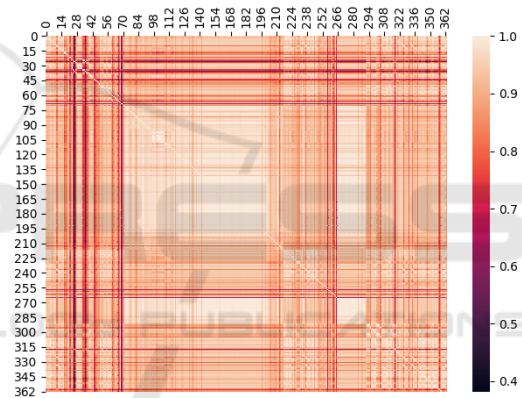


Figure 2: Correlation matrix between  $362 \times 362$  indicators.

A third idea is to **select the most relevant indicators** (for each variable) using feature importance techniques (Guyon and Elisseeff, 2003). Algorithm 1 illustrates this process by generating ranked lists of features (lines 3 and 4) based on algorithms Extra Trees (ET) and Random Forest (RF). The output of these algorithms are merged, and to avoid strong impact of a category of indicators (e.g., related to population), an indicator is removed if its parent is already in the list (lines 6 to 11). In the end, we obtain several list of features noted  $L_v^k$  which contain the most  $k$  relevant indicators for variable  $v$ . We have chosen to retain several lists containing from 10 to 100 indicators due to the complexity of prediction.

When features have been selected, the learning process can be run, as shown in the next section.

Algorithm 1: Selection of relevant features.

---

```

input : set of indicators  $I$ , set of variables  $\mathcal{V}$ 
output: lists of features  $L_v^k$ 
1 for  $v \in \mathcal{V}$  do
2    $L_v \leftarrow \emptyset$ ;
3    $F_v^{ET} \leftarrow \text{ET.rank\_features}(I)$ ;
4    $F_v^{RF} \leftarrow \text{RF.rank\_features}(I)$ ;
5    $F \leftarrow F_v^{ET} \cup F_v^{RF}$ ;
   /* sort, specific to general */
6    $F \leftarrow \text{sort}(F)$ ;
7   for  $f \in F$  do
8      $p_f \leftarrow \text{parent}(f)$ ;
9     if  $p_f \in F$  then
10       $p_f.\text{score} \leftarrow p_f.\text{score} + f.\text{score}$ ;
11       $F \leftarrow F - \{f\}$ ;
12  for  $k \in [10, 20, 30, 40, 50, 75, 100]$  do
13  |  $L_v^k \leftarrow \text{top-K}(F, k)$ ;

```

---

## 5 EXPERIMENTS

This section covers our preliminary experiments.

### 5.1 Experiment Protocol

We use the popular scikit-learn library for machine learning (Pedregosa et al., 2011). We are in a classification problem and five scikit-learn algorithms have been used<sup>14</sup>: Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbours (KNN), Support Vector Classification (SVC), and AdaBoost (AB). Many parameters have an impact in machine learning (Jordan and Mitchell, 2015), and we tested several configurations (e.g., weights, maximum depth in trees, number of neighbours, distance metric) to retain the best one. The annotated neighbourhoods are split into 80% training data and 20% evaluation data with cross-validation, as recommended in the literature (Bruce and Bruce, 2017). We use accuracy as quality metric, i.e. the fraction of correct predictions. It is common to have outliers in data, especially when working with real-world data. They can be detected using *Isolation Forest* algorithm for example, but it requires a manual approbation (i.e., verifying that the distribution confirms the outlier). Although outliers may decrease accuracy, we have a tiny supervised dataset and removing outliers means even less supervised data in our context.

<sup>14</sup>Other algorithms such as Stochastic Gradient Descent or Nearest Centroid have been tested, but they mostly follow the same trend or achieve insufficient accuracy.

## 5.2 Results

The main objective is to correctly predict the values for each environment variable of an IRIS. Recall that we have generated, for each variable, lists of top-10, up to top-100 features. Tables 2 to 7 provides the accuracy score (percentage) computed for each variable using different algorithms. In these tables, the baseline list  $I$  stands for all indicators (i.e., no feature selection) while  $L^k$  represents a list of  $k$  selected features. The underlined scores indicates the best result for an algorithm (i.e., by column). A **bold score** means that the corresponding list of features achieves a better score than the list  $I$ . The **highlighted cells** correspond to the best score in the whole table.

**Table 2** presents the quality results for the *building type* variable. Without feature selection, quality spans from 36% to 57%. Smaller lists enable an improvement over list  $I$  (e.g.,  $L^{20}$ ). The best score is achieved by RF with list  $L^{20}$ .

Table 2: Prediction quality for variable *building type*.

	LR	RF	KNN	SVC	AB
$I$	46.6	57.0	55.2	45.5	36.5
$L^{10}$	44.3	<b>59.3</b>	<u>57.8</u>	44.7	<b>41.7</b>
$L^{20}$	<u>49.2</u>	<b>60.0</b>	<b>56.3</b>	43.6	<b>43.6</b>
$L^{30}$	45.1	<b>58.9</b>	<b>55.9</b>	43.6	32.1
$L^{40}$	46.2	<b>59.3</b>	54.8	43.2	27.6
$L^{50}$	46.6	<b>58.9</b>	54.8	45.5	32.4
$L^{75}$	44.3	<b>58.2</b>	55.2	<u>45.9</u>	32.0
$L^{100}$	43.6	57.0	55.2	45.5	36.5

**Table 3** shows prediction quality for the *usage* variable. The scores without selection is tighter, between 50% and 65%. A few of the smallest lists perform better than the baseline one, but without significant improvement. RF obtains the best result with list  $L^{50}$ .

Table 3: Prediction quality for variable *usage*.

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

**Table 4** provides accuracy scores for the *landscape* variable. Similarly to previous results, small lists are able to improve quality over list  $I$  with three algorithms. SVC obtains the same score whatever the list.

Table 4: Prediction quality for variable *landscape*.

	LR	RF	KNN	SVC	AB
<i>I</i>	53.7	60.8	59.6	47.7	50.3
$L^{10}$	48.1	<b>62.7</b>	59.6	47.7	<b>51.8</b>
$L^{20}$	51.5	<b>63.0</b>	<b>60.4</b>	47.7	<b>52.6</b>
$L^{30}$	50.3	60.8	<b>61.9</b>	47.7	<b>52.5</b>
$L^{40}$	49.2	<b>62.7</b>	<b>61.5</b>	47.7	49.2
$L^{50}$	47.7	<b>61.5</b>	<b>61.1</b>	47.7	48.1
$L^{75}$	52.6	<b>62.3</b>	59.3	47.7	48.5
$L^{100}$	<b>56.3</b>	60.8	59.6	47.7	50.3

Table 5: Prediction quality for variable *social class*.

	LR	RF	KNN	SVC	AB
<i>I</i>	44.4	51.1	42.1	45.5	36.5
$L^{10}$	43.6	46.6	<b>43.9</b>	44.7	<b>41.7</b>
$L^{20}$	39.1	46.6	<b>45.1</b>	43.6	<b>43.6</b>
$L^{30}$	41.4	49.6	<b>45.1</b>	43.6	32.1
$L^{40}$	39.1	<b>51.8</b>	<b>46.6</b>	43.2	27.6
$L^{50}$	42.1	48.1	<b>44.3</b>	45.5	32.4
$L^{75}$	<b>45.1</b>	48.1	<b>44.0</b>	<b>45.9</b>	32.0
$L^{100}$	40.7	51.1	42.1	45.5	36.5

**Table 5** depicts quality results for the *social class* variable. The lists of selected features, either small or large depending on the algorithm, allows a better quality in a few cases. The best score is slightly above 50%, which shows that this variable is difficult to predict. Yet, many features describe incomes (median, per decile), population characteristics (number of students, employees, farmers, unemployed, etc.).

**Table 6** details quality obtained for the *morphological position*. The  $L^{10}$  list mainly wins against the baseline list, except with SVC which achieves similar scores (44%) whatever the features.

Table 6: Prediction quality for variable *morphological*.

	LR	RF	KNN	SVC	AB
<i>I</i>	46.6	59.7	58.2	44.7	45.8
$L^{10}$	<b>48.5</b>	<b>60.0</b>	<b>60.8</b>	44.0	<b>49.9</b>
$L^{20}$	44.0	<b>61.2</b>	<b>58.5</b>	44.4	<b>48.5</b>
$L^{30}$	39.2	<b>61.2</b>	58.2	44.4	<b>48.8</b>
$L^{40}$	33.5	<b>61.2</b>	<b>58.6</b>	44.4	<b>50.7</b>
$L^{50}$	36.1	59.3	57.4	44.4	<b>46.2</b>
$L^{75}$	41.3	<b>60.8</b>	57.1	44.7	<b>49.2</b>
$L^{100}$	43.2	59.7	58.2	44.7	45.8

**Table 7** is dedicated to *geographical position*. Scores are far lower than for other variables (33% as best value), which is not surprising given the *a-priori* irrelevant indicators for this prediction. Still, small lists mostly perform better than the baseline.

We conclude this experimental section with a discussion. Best scores range from 33% for geographical position and 50% for *social class* to 60-65% for the remaining four variables. Although al-

Table 7: Prediction quality for variable *geographical*.

	LR	RF	KNN	SVC	AB
<i>I</i>	22.0	<b>33.6</b>	27.2	25.0	15.6
$L^{10}$	<b>25.3</b>	29.9	<b>27.6</b>	24.6	<b>21.9</b>
$L^{20}$	<b>26.1</b>	31.3	<b>29.5</b>	<b>25.3</b>	<b>20.1</b>
$L^{30}$	<b>26.1</b>	31.7	<b>28.3</b>	<b>27.2</b>	<b>17.5</b>
$L^{40}$	<b>29.1</b>	32.8	<b>28.3</b>	24.6	<b>17.1</b>
$L^{50}$	<b>25.0</b>	32.1	27.2	23.8	<b>19.0</b>
$L^{75}$	<b>24.6</b>	32.8	27.2	25.0	<b>17.9</b>
$L^{100}$	<b>24.6</b>	<b>33.6</b>	27.2	25.0	15.6

gorithms obtain different scores with the baseline list, their results mainly improve by a few percent (in average per column) when using other lists of features, which could demonstrate that current indicators are not sufficient or useful. These results are promising, but still require improvements, especially regarding the representativeness issues presented in Section 4.1. It was not possible to predict on a sufficient number of new IRIS because the manual verification is time-consuming, as explained in Section 3.1. Among the ten algorithms and configurations we have tested so far, Random Forest seems to be the most interesting in our context because it achieves all best scores. Some algorithms were not suitable, for instance SVC requires many features (best results with all indicators or with largest lists of features). Our algorithm for feature selection has also proven useful, since many lists outperform the baseline (whatever the algorithm or variable). Lists of 20 up to 50 features are particularly effective. However, the improvement is not significant (a few percent at best compared to baseline). On the contrary, larger lists (top-100) usually provide the same quality as the baseline.

## 6 CONCLUSION

In this paper, we have presented a use case for predicting the environment of any neighbourhood in France through six descriptive variables. We have studied the representativeness of our 270 annotated neighbourhoods compared to the whole set of 49,800 to detect some bias. Due to a large quantity of available indicators, we also selected a subset for each variable. Our experiments show that (small) lists generated with our feature selection perform better than a learning using all indicators. However, the overall prediction scores need to be improved before predicting at a larger scale.

The main perspective is to achieve better prediction results. Our first results could be exploited by social science researchers in order to facilitate the manual annotation of neighbourhoods, thus increasing the amount of examples. Another possibility could be the

generation of a bigger synthetic dataset, which share similarities with the 49,800 neighbourhoods. We also plan to integrate prices and points of interest (that could reflect the nature of a neighbourhood, for instance an organic shop is usually found in middle or upper class neighbourhoods). A fourth perspective is the correlation between variables, which are not totally independent. For instance, a rural area has more chances to be classified as countryside and to host houses. The prediction of a given variable could impact the classification of others, especially the most difficult ones such as geographical or social. Finally, we plan to release a tool named `predihood` for letting researchers implement and test their classification algorithms on our dataset.

## ACKNOWLEDGEMENTS

This work has been partially funded by LABEX IMU (ANR-10-LABX-0088) from Université de Lyon, in the context of the program "Investissements d'Avenir" (ANR-11-IDEX-0007) from the French Research Agency (ANR), during the HiL project<sup>15</sup>.

## REFERENCES

- Barret, N., Duchateau, F., Favetta, F., Miquel, M., Gentil, A., and Bonneval, L. (2019). À la recherche du quartier idéal. In *EGC*, page 429–432.
- Bellahsene, Z., Bonifati, A., and Rahm, E. (2011). *Schema matching and mapping*. Springer.
- Bigot, R., Crouette, P., Müller, J., and Osier, G. (2011). Les classes moyennes en europe. *Le CRÉDOC, Cahier de recherche*, 282.
- Bonneval, L., Duchateau, F., Favetta, F., Gentil, A., Jelassi, M. N., Miquel, M., and Moncla, L. (2019). Étude des quartiers : défis et pistes de recherche. In *EGC*.
- Bruce, P. and Bruce, A. (2017). *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly.
- Caragliu, A., Del Bo, C., and Nijkamp, P. (2011). Smart cities in europe. *J. of urban technology*, 18(2):65–82.
- Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.
- Cranshaw, J., Schwartz, R., Hong, J., and Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. In *AAAI Conference on Weblogs and Social Media*.
- Delmelle, E. C. (2015). Five decades of neighborhood classifications and their transitions: A comparison of four us cities, 1970–2010. *Applied Geography*, 57:1 – 11.
- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.
- Donoho, D. (2017). 50 years of data science. *J. of Computational and Graphical Statistics*, 26(4):745–766.
- Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. of Machine Learning Research*, 3(3):1157–1182.
- Halevy, A., Rajaraman, A., and Ordille, J. (2006). Data integration: the teenage years. In *VLDB*, pages 9–16.
- Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.
- Le Falher, G., Gionis, A., and Mathioudakis, M. (2015). Where Is the Soho of Rome? Measures and Algorithms for Finding Similar Neighborhoods in Cities. *ICWSM*, 2:3–2.
- Lillesand, T., Kiefer, R. W., and Chipman, J. (2015). *Remote sensing and image interpretation*. Wiley & Sons.
- Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71.
- Ojo, A., Curry, E., and Zeleti, F. A. (2015). A tale of open data innovations in five smart cities. In *Int. Conf. on System Sciences*, pages 2326–2335. IEEE.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *J. of Machine Learning Research*, 12:2825–2830.
- Shen, W., Wang, J., and Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *TKDE*, 27(2):443–460.
- Tang, E. and Sangani, K. (2015). Neighborhood and price prediction for san francisco airbnb listings.
- Yu, M., Li, G., Deng, D., and Feng, J. (2016). String similarity search and join: a survey. *Frontiers of Computer Science*, 10(3):399–417.
- Yuan, X., Lee, J.-H., Kim, S.-J., and Kim, Y.-H. (2013). Toward a user-oriented recommendation system for real estate websites. *Information Systems*, 38(2):231–243.
- Zhang, A. X., Noulas, A., Scellato, S., and Mascolo, C. (2013). Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In *Social Computing*, pages 69–74. IEEE.

<sup>15</sup><http://imu.universite-lyon.fr/projet/hil/>