An Environmental Study of French Neighbourhoods*

Nelly Barret[1 [0000−0002−3469−4149]], Fabien Duchateau[1 [0000−0001−6803−917X]], Franck Favetta[1 [0000−0003−2039−3481]], Aurélien Gentil[2], and Loïc Bonneval[2]

[1] LIRIS UMR5205, Université de Lyon, UCBL, Lyon, France
nelly.barret@etu.univ-lyon1.fr, {fduchate, ffavetta}@liris.cnrs.fr
[2] Centre Max Weber, Université de Lyon, France
aurelien.gentil@univ-lyon2.fr, loic.bonneval@univ-lyon2.fr

Abstract. Neighbourhoods are key places for daily activities and many studies in social sciences, health or biology use this spatial concept as an impact factor. Conversely, the neighbourhood environment is rarely defined (e.g., in terms of landscape or main social class). In this paper, we propose six descriptive variables for this environment, and we provide a dataset of 270 annotated neighbourhoods. Next, we detail two methods (prediction and spatial computation) for describing environment of remaining neighbourhoods, and we show in our set of experiments an acceptable quality.

Keywords: Neighbourhood, Environment, Data Integration, Machine Learning, Prediction, Spatial computation.

1 Introduction

Neighbourhoods are a very common concept in studies from diverse domains such as health, social sciences, or biology. For instance, Japanese researchers investigated the relationships between social factors and health by taking into account not only behavioural risks, but also housing and neighbourhood environments [38]. In a British study, authors describe how living areas have an impact on physical activities, from which they determine a walkability index at the neighbourhood level for improving future urban planning [16]. Smarts cities also consider neighbourhoods as an ideal unit division for measuring urban quality [18]. Lastly, a survey describes the luxury effect, i.e., the impact of wealthy neighbourhoods on the surrounding biodiversity [27]. In addition to the essential role of the neighbourhoods, these examples also show that the definition of neighbourhood is subject to various interpretations [20, 24, 8], which may depend on the point of view (e.g. administrative, functional, economic). The definition (mainly in terms of borders) and description (features) of a neighbourhood is therefore a complex task.

* This paper is an extended version of a short paper published in the DATA 2020 proceedings [3]. 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.

When studying neighbourhoods, one of the challenges is to compare them according to some criteria. Different works have proposed solutions to tackle this issue. It is possible to exploit social networks data to detect similar neighbourhoods between cities [26] or in the same city [13, 41]. Accommodation advertisements (rental or buying) are also used for predicting price and neighbourhood characteristics [39]. In a neighbourhood search, researchers assume that they have users, so they can compare their profiles to annotated ones [40] or their original neighbourhood with regards to target ones [4]. However, most of these works have a limited scope (e.g., a few cities) or the multiple comparison criteria make it difficult to understand the differences between two neighbourhoods. In a recent paper, we have proposed to compare environments of neighbourhoods [3]. In this extended version, we provide a complete description of the environment variables, a few improvements for computing this environment as well as a new method (for the *geographical position* variable), and an updated experimental validation including two extra experiments.
In the context of the HiL project[3], we aim at studying the impact of the neighbourhood environment when people moves in another city, i.e., we plan to answer questions such as *"do they choose a similar environment?"* and *"how does their possible salary increase affect their choice of neighbourhood?"*. Indeed, the choice of a neighbourhood may be difficult, especially without any prior knowledge about the future city. One may look for a vibrant neighbourhood with many pubs while other may prefer a quiet residential area close to schools and parks. To reach this goal, it is necessary to characterize environment of neighbourhoods in a simple way (i.e., with a few attributes, such as type of buildings, location of the neighbourhood in the city, main social class, etc.). Such description is useful for comparing neighbourhoods, for instance in social science studies or when searching for accommodations. However, in large countries, there are too many neighbourhoods (e.g., about 50,000 for France) and it is not possible to manually describe environment for each of them. We tackle this problem using an exploratory methodology and machine learning.
This paper includes the following contributions:

- Description of a neighbourhood environment. Based on the literature in social sciences and on a survey of 155 individuals, we propose a list of six variables (each with a limited number of values) to define this neighbourhood;
- Dataset `mongiris`. We describe which and how data have been integrated for about 50,000 French neighbourhoods (e.g., number of bakeries, average income), and we provide 270 neighbourhoods annotated with their environment. The resulting dataset named `mongiris` is publicly available;
- Methods for computing environment. We present one method using machine learning to predict environment, with a focus on the selection of the most interesting features. The `predihood` tool used to configure classifiers and to visualize neighbourhoods on a map is publicly available. Another method enables the computation of the *geographical position* variable;
- Experimental validation. A set of experiments on the French territory demonstrates the benefits of our proposals, as well as possible clues for improvement.

[3] HiL project, `http://imu.universite-lyon.fr/projet/hil/`

We first introduce related work (Section 2). Next, we describe variables representing the environment of a neighbourhood (Section 3). Methods for computing environment are detailed in Section 4 while Section 5 presents and analyses experimental validation. Section 6 concludes and highlights perspectives.

2 Related work

Most works dealing with modelling (urban) environment falls in the energy and transportation domains [35]. Multiple projects focus on studying neighbourhoods, but they do not aim at defining and describing their environments. As explained in the literature, the concept of neighbourhood is difficult to describe due to various perceptions [34], and each work provides its own definition and borders for neighbourhoods [20, 24, 8].
A first category of works relies on social networks, which contain a wealth of geolocated information, especially tweets, likes or check-ins. In the Livehoods project, the goal is to discover city's dynamics from its resident's behaviour [13]. Spatial and social proximities are used as input features of a spectral clustering algorithm, and the evaluation compares the machine-learning based algorithm fed with 18 million check-ins against the neighbourhood description from 27 interviews. In the same fashion, Le Falher et al. discover similar neighbourhoods between cities [26]. To reach this goal, they use classification algorithms applied on social networks data, namely *Information Theoretic Metric Learning* and *Large Margin Nearest Neighbour*. These algorithms build a matrix with human activities occurring in places along with surrounding points of interest, and the classes come from Foursquare categories. Next, they use the Earth Mover distance to measure the effort for "transforming" one area into another one. The Hoodsquare approach from Zhang et al. aims at detecting boundaries and similar areas [41]. Each neighbourhood is described as a vector of features (e.g., place types, Foursquare check-ins, temporal information) and similarity metrics such as Cosine similarity compute a similarity score between two neighbourhoods. Location recommendation is another motivation [29]. Authors do not rely on the user point of view, but rather on the location to neighbourhood characteristics. They assume that locations in the same neighbourhood share more similar user preferences, and that locations in the same region may share similar user preferences, thus leading to a two-level matrix factorization solution.
Another category exploits profiles of inhabitants. Researchers in South Korea have proposed to find the most relevant neighbourhood and accommodation based on similar user profiles [40]. They have built a database of residents, which includes information such as household composition, budget, accommodation preferences and distance from home to work. A new profile is compared to existing ones using case-based reasoning, and recommendations are adjusted consequently. A recent project *my neighbourhood, my neighbours*[4] analyses the relationships between residents in their neighbourhood. About 2,500 inhabitants from various areas (city centres, urban and peri-urban) in two cities (Lyon and Paris) have answered a survey about their vision of their neighbourhood, city and profile. Descriptions of residents provide an overall qualification of each considered neighbourhood. Preliminary results show that the neighbour perception strongly varies according to density and to social characteristics (e.g., young people

[4] Project *mon quartier, mes voisins*, `http://mon-quartier-mes-voisins.site.ined.fr/`

include city centres while older ones constraint it to a few streets). Besides, they question the relevance of neighbourhoods in peri-urban areas. These results about neighbour representation confirm previous studies such as the one from Pan Ké Shon [32].

The last category relies on objective criteria that characterize neighbourhoods. The study from Tang et al. compares Airbnb announcements in San Francisco to determine their price and neighbourhood location [39]. The features include structured information (e.g., type of accommodation, number of rooms), bag of words (most frequent terms in the announcement), word class (among nine predefined classes such as nature, nightlife or culture), text sentiment and visual characteristics. In the VizLIRIS prototype, users may search for and visualize ideal neighbourhoods in the context of job transfers [4]. Hundreds of features are available to describe each area, such as the number of transportation means, the average income or inhabitants classified per socio-professional class. Distance-based algorithms (e.g., KMeans) are used for recommending neighbourhoods similar to selected ones while clustering algorithms enable the detection of similar neighbourhoods in a given area. In addition to scientific literature, many online applications produce neighbourhood recommendations, as described in the following list (centred on France and non exhaustive). The website DataFrance[5] integrates data from diverse French sources, such as indicators provided by the National Institute of Statistics[6] (INSEE), geographical information from the National Geographic Institute[7] (IGN) and surveys from newspapers for prices (L'Express[8]). The search for neighbourhoods which satisfy user criteria is performed manually through the interface. Kelquartier[9] describes the main French cities using quantitative criteria (e.g., average income, density of schools, density of shops). A manual search for neighbourhoods includes tens of criteria about the area (e.g., density of restaurants, schools), about real estate (e.g., building seniority, ratio of landlords) and about inhabitants (e.g., income, age, type of household). Home in Love[10], vivroù[11] and Cityzia[12] 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[13] is a collaborative website for evaluating French cities. Users can give a score (out of 10) for each of the ten categories, from healthcare to security or culture. Although limited to the city or district level, user comments frequently include mentions of neighbourhood, which may be useful for a (manual) assessment of the quality of a neighbourhood. Social science works also highlight the double perception of the neighbourhood, either from the inside (residents' perception) or from the outside (objective criteria) [17, 2].

[5] DataFrance, `http://datafrance.info/`
[6] INSEE, `http://www.insee.fr/en/`
[7] IGN, `http://www.ign.fr/`
[8] L'express, `http://www.lexpress.fr/`
[9] Kelquartier, `http://www.kelquartier.com/`
[10] Home in Love company (in French), `http://homeinlove.fr/`
[11] Vivroù, `http://www.vivrou.com/`
[12] Cityzia, `http://www.cityzia.fr/`
[13] Ville idéale, `http://www.ville-ideale.fr/`

Our contribution differs from existing works on several points. First, some works are limited to a few cities, which is not possible when studying population's trajectories. Indeed, rural migration is still very active, thus requiring a description of all areas. Relying on social data implies prior analysis in order to avoid bias (e.g., over-represented class of people or activities). User profiles are an interesting direction, but it requires a long and costly study to collect all necessary information (which, moreover, people may not be willing to provide). Criteria are not directly available for describing the environment of a neighbourhood, although some of them provide an insight (e.g., the average income is a clue for determining social class, but it is not sufficient and often relative to a local context). Besides, too many criteria makes it difficult both for obtaining a simple representation of the area and for comparing and understanding the choice of a neighbourhood. Finally, there is no work which aim at associating both perceptions of the neighbourhood (inside and outside), and it is a real challenge to automate what qualitative studies are able to do, but at larger scale.

3 Environment of a neighbourhood

As previously explained, the notion of neighbourhood varies according to the point of view, making it more difficult to define representative criteria. Besides, most studies provide description about quality of life (e.g., security, health), which may include bias and subjectivity. In order to obtain a simple description of neighbourhoods, social science researchers have studied information about residents and their neighbourhoods and they have extracted a list of six descriptive variables for neighbourhoods. Our proposal focuses on France, but could be applied to similar countries.

3.1 Neighbourhood definition

Neighbourhoods have a different definition according to usage [20, 8, 2]. For instance, geographers mainly rely on natural borders while inhabitants have a less precise vision of the boundaries. Voting and cadastral definitions are typically used by administrative employees. Historical or economical divisions may also impact the definition.
In our context, we have chosen a small division unit of the French territory named IRIS[14] to represent our neighbourhoods. They are produced by the National Institute of Statistics (INSEE) and are considered of good quality due to their frequent updates and wide use by many organizations. An IRIS usually includes between 2,000 up to 5,000 residents, and consequently are rather small in cities while their size increases in the countryside. These units are constrained by geographic and demographic criteria and their borders are easily identifiable and stable in the long term. There exist three types of IRIS: housing (accounting for around 90% of the dataset), activities or business (e.g., industrial area, university campus), and miscellaneous (e.g., parks, forests). The French territory is split into 49,800 IRIS. In the rest of the paper, we use the term neighbourhood instead of IRIS. Indeed, although they are defined as statistical division units, IRIS are considered as a reliable approximation of the perceived neighbourhood:

[14] IRIS definition, `http://www.insee.fr/en/metadonnees/definition/c1523/`

in outside urban areas, they are usually similar to the town, and in large cities, they enable to estimate the diversity of people (and environment elements) which are met in daily activities [31]. Last, one of the objectives of this paper is to check whether an approach based at the IRIS level is sufficient to simulate the perceived neighbourhood.

3.2 Methodology for defining environment

We have analysed data from a company[10] specialized in accompanying the search of an accommodation during job transfers. The dataset is not available for confidentiality reasons. At the time of writing, it includes 155 customers (each representing an household), thus 310 locations (previous accommodation before the job transfer, and the new one). Some customers came from other countries (no neighbourhood information), and several neighbourhoods were redundant (e.g., several people moving to the same village close to a large industrial factory). Our dataset results in a total of 270 distinct neighbourhoods. For each customer, hundreds of information are available in various categories:

- Personal information (names, birthdate, household composition, etc.);
- Work information (label, socio-professional category, salary, etc.) both for the previous job and the new one;
- Tax information, which may explain some situations, for instance when people have other incomes than salaries;
- Address of the previous accommodation, and address of the new accommodation (which was discovered using the company's services). From these data, we deduce the neighbourhoods;
- Expenses (credit, rental, monthly bills, etc.);
- Accommodation (description from real-estate agencies, type of heat-system, presence of amenities such as garden, parking or swimming-pool, shared equipments, etc.);
- Profile (optional and filled in by customers);
- Ideal accommodation (optional and filled in by customers);
- Expectations about the future neighbourhood (optional and filled in by customers);
- A narrative analysis, written by social science researchers, about the life style and story of the household (e.g., marriage, type of neighbourhood), including assumptions about the job transfer (social trajectory) and about accommodation search.

The main idea is to study the environment of these households, mainly in terms of social, material and natural aspects. Our methodology is independent from households' profiles, and it includes the exploitation of household data and data from Google Maps, namely aerial and street views, photographs of buildings, urban furniture, parked cars, type and brand of shops, leisures and park areas. The virtual exploration of a neighbourhood enables a detailed observation of the environmental surroundings of each address, in a variable radius. The closest environment (400 meters radius) is always analysed, and if needed, further exploration is performed[15]. A careful attention was taken about

[15] The scale differences for analysing the environment make it difficult to automate the process, hence one of the objectives of this paper.

seasons (e.g., less green areas visible in winter) as well as the date of street views (no more than four years). A description of the neighbourhood based on these observations are stored so that objective and fine-grained comparisons between two addresses are possible at very small scale (usually smaller than the neighbourhood). Finally, this exploration provides means of comparison between both accommodations of the same customer for studying residential choices. Note that this manual observation step takes several hours (per address / neighbourhood) when rigorously performed.
From this analysis, neighbourhoods were characterized and classified into six categories, built using an inductive process (popular in social sciences). The various decisions which led to this classification involve (a part of) arbitrary choices and subjectivity. But our methodology consists in classifying using very detailed and meticulous observation and interpretation, rather than relying on more "objective" data produced by different providers. Three additional verifications were performed to comfort our decisions, namely investigation with related social science works, consistency with external data sources about neighbourhoods (DataFrance[5] and KelQuartier[9]) and consistency between departure neighbourhood and arrival neighbourhood of an household (given their situation). For this last point, an initial question deals with the comparison of both neighbourhoods. Indeed, the context of job transfer implies that employees usually search for a similar neighbourhood to the one they come from, mainly for securing the residence change [36]. If two neighbourhoods are not similar, researchers check how information about the household (e.g., salary increase, children) and/or the city (e.g., moving from a costly city to a small town) could explain the differences between both neighbourhoods. Redundant neighbourhoods (where different people originate or arrive) were also exploited as a verification means. This methodology provides a solid background for defining the environment of a neighbourhood, which is presented in the next section.

3.3 Six variables for environment neighbourhood

Social science researchers followed the previously mentioned methodology to define 6 environment variables, whose goal is to facilitate the description and the comparison of neighbourhoods. These variables are summarized in Table 1 along with their list of values. *Building type* refers to the most common buildings in the neighbourhood. *Usage* represents local activities and *landscape* defines the space conceded to green areas. *Social class* stands for the stratification of a population according to position in the social hierarchy. *Morphological position* can be seen as the relationship to centrality. *Geographical position* denotes the direction towards the city centre of the closest city.
Let us now provide more details about each variable:

Building type. We have distinguished five categories. *Large housing estates* are composed of similar residential towers, such as social housing or winter sports apartments. A neighbourhood classified as *buildings* usually stands for areas with heterogeneous buildings. *Mixed* neighbourhoods are typically found in cities and combine other possible values. Contrary to *housing subdivisions*, individual *houses* (both in cities and rural areas) are heterogeneous in terms of construction period or architecture;

Table 1: Environment variables and their possible values.

Environment variable	Values (comments)
Building type	Large housing estates (homogeneous tower conglomerate) Mixed (both buildings and houses) Buildings (heterogeneous) Housing subdivisions (homogeneous) Houses (heterogeneous)
Usage	Residential area (few local shops) Shopping (areas with many local shops) Other activities (mixed zones with factories, companies and some houses, usually outside cities)
Landscape	Urban (high density of buildings, near absence of green areas) Green areas (built area, but with some natural spaces) Forest (high density of green areas or forests) Countryside (crop fields and natural areas)
Social class	Lower Lower middle Middle Upper middle Upper
Morphological position	Central Urban (in the main town, but not in the centre) Peri-urban (at the periphery of the city) Rural (area further than urban and peri-urban areas)
Geographical position (9 different values)	Centre North North East East ...

Usage. Three main categories enable to classify the studied environments. In a *residential area* neighbourhood, there is almost no shop. They are usually found at the periphery of urban areas or in housing subdivisions. *Shopping* areas, usually in city centres, are marked by a high density of local shops. The last category (*other activities*) corresponds in general to areas located at the borders between urban and peri-urban, with houses surrounded by companies, large commercial zones and factories. This variable could include more categories (e.g., arts, education, nightlife, work), as in the Hoodsquare project [41]. However, in our context of job transfer (with many people searching in peri-urban areas), the main goal is to distinguish the usage in peri-urban areas (and not necessarily those in urban zones). Besides, adding more general categories would not be sufficient as users are interested in specific types of point of interest (e.g., kindergarten or elementary school, bakery, organic shops);

Landscape. Four types of sceneries have been identified according to the density of surrounding green areas and plants as well as their natural state. Neighbourhoods classified as *urban* imply a quasi-absence of plants. *Green areas* offers a significant

presence of parks, gardens or tree alleys, but they are delimited and maintained. *Forest* are wooded neighbourhoods where green areas are strongly visible. Finally, the *countryside* value includes agricultural and farming spaces, as well as natural zones with few buildings (mountains, vast forests);

Social class. This variable is one of the most studied in social sciences [9, 23, 31]. In our context, we defined 5 groups of social class, ranging from *lower*, *lower middle* and *middle* up to *upper middle* and *upper*. To perform this classification, we rely on various revealing clues from our observations of the people living in the studied neighbourhoods: architectural aspects of buildings, position in the city, type and brand of local shops, type and brand of parked cars, configuration of outdoor spaces, etc. Social class is certainly the environment variable which involves the most difficult interpretation. The difference between a middle-upper area and an upper one, or at the other end of the social hierarchy, between a lower and a lower-middle neighbourhood, may not be visible from the external view of a neighbourhood, especially in urban and central which includes more social diversity [14]. However, and although the choices between close classes may be tenuous, we note that our observations were clearly sufficient to determine the main trend (i.e., rather lower, rather middle or rather upper), which are critical in peri-urban settings;

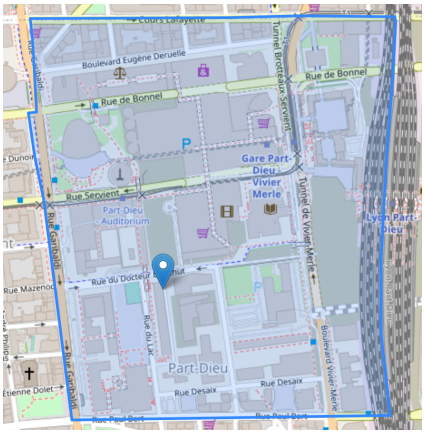
Morphological position. The morphology criterion has been divided into four values: *central*, *urban*, *peri-urban* and *rural*, each denoting a placing according to the centrality of a geographic area. Neighbourhoods in each category may share trends or characteristics. This variable also enables sociologists to study phenomena such as rural flight and urban planning.

Geographical position. This variable indicates the direction towards the city centre of the closest city. This is an essential information for peri-urban neighbourhoods. Indeed, a central morphologic area may not be geographically centred (e.g., the main shopping and service area of the peri-urban town Vénissieux is central, but the city centre of the largest city Lyon is located north). Urban areas were not built at random, and people with a similar lifestyle tend to live in the same neighbourhood. For instance, it is well-known that East districts were poorer due to industrial pollution coming from West winds [37]. Thus, detecting the geographical position can help when analysing population in neighbourhoods. In rural areas, this variable represents the direction of the closest large city. Note that our work does not take into account poly-centralities (i.e., secondary cities, which may be as attractive as the largest city for surrounding peri-urban towns).

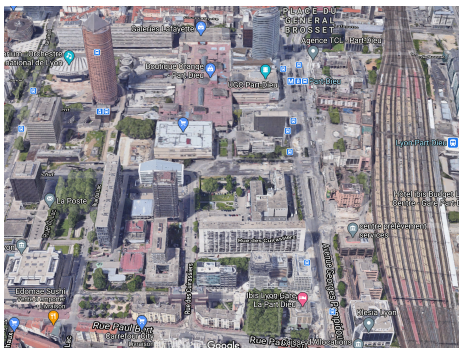
3.4 An example of environment in Lyon Part Dieu

To illustrate these environment variables, let us describe the neighbourhood *Part Dieu*, in the city centre of Lyon, France. In the INSEE data, this neighbourhood is identified with code *693830301* and is located in Lyon 3rd district. Its activity type is A, which stands for a commercial, services, or industrial area (i.e., which includes more workers than residents). Figure 1 depicts views of this neighbourhood: the left picture shows the border of the neighbourhood while the right one provides an aerial view. We notice that it includes the main railway station of the city, business towers including a

large mall, movie theatre, the main library, the auditorium, hotels and some important administrative buildings.



(a) Delimitation of the neighbourhood.



(b) Aerial view.

Fig. 1: The *Part Dieu* neighbourhood, Lyon, France (source Google Maps).

Social science researchers have described the environment of *Part Dieu* neighbourhood as shown in Table 2. The building type is mainly composed of buildings, and its usage is indeed dedicated to shopping as the first floor of most buildings is dedicated to merchant activities. Although we notice some trees in the aerial view, the perception is clearly urban in the neighbourhood. The last variables are easier to check, since this neighbourhood is obviously central (in the city) and in the centre of the closest city. Researchers also provided a short description about the accommodation in its area. For the household living in *Lyon Part Dieu*, the comment is *"Building from the 1930's, along the Jean-Jaurès avenue, close to many shops"*.

Table 2: Environment variables for neighbourhood *Part Dieu*, in Lyon.

Building type	Buildings
Usage	Shopping
Landscape	Urban
Social class	Upper middle
Morphological	Central
Geographical	Centre

3.5 Statistics and representativeness

To conclude this section, we provide statistics about the classifications of our 270 neighbourhoods and their representativeness with regards to the whole country.
Table 3 depicts the value distribution per variable. We note that households either live in buildings (apartments), houses or mixed areas, and half of them are located in residential areas according to the *usage* variable. The *social class* variable is over-represented by the middle and upper middle classes. The type of *landscape* is typical from cities (urban and green areas), which is consistent not only with the morphological situation focused around the centre and urban areas, but also with the job transfer context. Finally the *geographical position* shows some favoured directions such as Centre, East, North and South.

Table 3: Statistics of neighbourhoods according to environment variables.

Building type	
Large housing estates	11
Mixed	89
Buildings	91
Housing subdivisions	34
Houses	45

Social	
Lower	11
Lower middle	13
Middle	89
Upper middle	124
Upper	33

Usage	
Residential area	145
Shopping	91
Others	34

Landscape	
Urban	102
Green areas	122
Forest	24
Countryside	22

Morphological	
Central	89
Urban	74
Peri-urban	93
Rural	14

Geographical	
Centre	58
North	43
North East	17
East	55
South East	18
South	34
South West	8
West	26
North West	11

Next, we check how representative our dataset of 270 annotated neighbourhoods is with regards to the total number of French neighbourhoods (49,800). It only represents 0.6% of this total, which may not be sufficient for machine learning algorithms used for predicting environment of the remaining neighbourhoods.

Building type and usage. Both variables are difficult to verify, as there are few additional information about the composition of neighbourhoods.

Morphological position. This variable indicates whether a neighbourhood is inside or far from a city. Based on the INSEE methodology for constructing their division unit[14], one third of the neighbourhoods (16,100) are found in cities with more than 10,000 inhabitants and most towns with more than 5,000 inhabitants. Remaining locations (33,700) were considered as sparsely populated and a single unit is affected to each one. Assuming these small towns are rural areas, they account for 68% of the whole dataset, while our annotated dataset only includes 5% of rural neighbourhoods. This difference is easily understandable due to our context of job transfers, mainly to the benefit of big cities.

Landscape. This variable is closely related to the morphological position. If we assume that forest and countryside are representative of rural areas, they obtain a representation of 17%, which is disconnected from the 68% expected in France but consistent with the 5% previously mentioned rural neighbourhoods.

Social class. Classification of the population according to wealth is not an easy task, and many studies have their own definitions and categories. According to Bigot et al. [7], the French population includes 59% of households belong to the middle class (in a broad meaning, thus encompassing lower and upper middles). This middle class is defined with incomes ranging from 70% to 150% of the median income (1,750 euros for 2014), which corresponds to 71% of neighbourhoods in the country. Conversely, our dataset includes 82% of middle class neighbourhoods.

Geographical position. Although this variable is more balanced, a few directions appear more frequently (e.g., South, East, North). Providing an explanation for these cases requires more research in social sciences. The centre value is the most represented, since it is correlated with the central morphological position. A comparison with the whole dataset is difficult: computing the direction of the closest city for all areas depends on several parameters, as shown in Section 4.4.

To summarize, we have identified six environment variables from a dataset of 270 annotated neighbourhoods. Compared to the full set of neighbourhoods, the *morphological* and *landscape* variables are biased, and the *social class* variable includes a small bias too. The dataset is not representative of the whole country as it focuses on moving employees. Besides, we are less interested in urban neighbourhoods, contrary to similar works which mainly study old or sensitive neighbourhoods, thus promoting peri-urban study. Even without a fair representativeness, our dataset is interesting as it provides a diversity of environments. However, identified biases may have an impact for predicting environment of any neighbourhood, as described in the next section.

4 Computing environment

To compare resident's moves between neighbourhoods, it is important to annotate the environment of all neighbourhoods. Manual annotation is a time-consuming process (see Section 3.2), thus an automatic solution is preferable. We propose two approaches for computing the environment of a neighbourhood: prediction (using machine learning algorithms) and spatial computation (only for the *geographical position* variable). Figure 2 depicts the whole process. First, data description aims at gathering and collecting relevant data sources about neighbourhoods. They are integrated into a merged database named *mongiris*. From this point, the top approach predicts the environment by selecting relevant features and applying machine learning training and testing while the alternative approach enables the computation of the *geographical* variable.

4.1 Data description

A predictive approach requires features to build a model and classify instances. The adoption of Open Data principles has led to a wealth of information available in different data sources [1]. Following is a description of considered data sources:

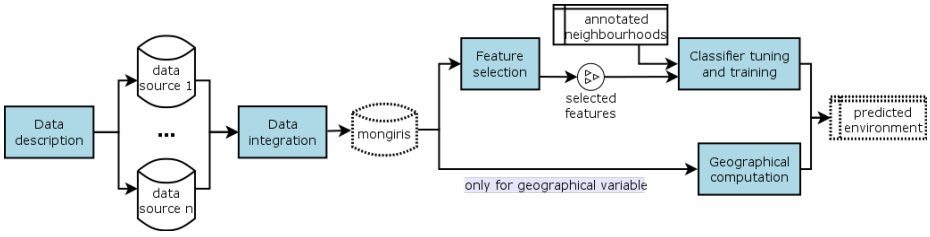


Fig. 2: Process for computing environment of any neighbourhood.

- **IRIS data.** As explained in Section 3.1, we selected the unit division IRIS as neighbourhoods, a choice also supported by the fact that they come with many indicators about population, buildings, shops, leisures, education, etc. These data are updated every 4 years, and we use the 2016 version. First, each neighbourhood includes 17 descriptive information (identifier, 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 buildings built before 1950, of tennis courts), unit quantities (e.g., average income, average income for the agricultural class), coefficients (e.g., Gini coefficient[16], S80/S20 ratio[17]), percentages (e.g., percentage of unemployed people, percentage of fiscal households) or string values (e.g. notes about incomes);
- **Spatial data.** Each neighbourhood 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 either as an indicator[18] or for normalizing other data;
- **Prices.** This information is valuable in the sense that it can leverage several environment variables. For instance, costly accommodations are typically found in richer neighbourhoods, situated in the city centres or close to remarkable locations such as green areas or historic buildings. Prices in large cities are usually higher than in small cities, thus peri-urban neighbourhoods of large cities may have comparable prices to central areas in small cities. Local context is therefore needed to wisely exploit prices. In addition, this kind of sensible or monetizable information is usually incomplete or rarely available for free. It is available as open data at a higher cartographic level (administrative department), which is not sufficiently accurate to be useful. In DataFrance[5], prices are only available for 600 cities, based on a newspaper survey. Real estate agencies own such data, but it may be biased (e.g., towards a specific type of accommodation), incomplete (e.g., specific to a region) and/or confidential. The recent DVF project[19] provides prices of all sold buildings per year, but it requires more work for deduplication (sales are filled in at the parcel

[16] Gini coefficient, `http://en.wikipedia.org/wiki/Gini_coefficient`
[17] S80/20 ratio, `http://www.insee.fr/en/metadonnees/definition/c1666`
[18] Neighbourhoods in cities tend to be small while those in rural areas have a larger size.
[19] Accommodation prices in France, `http://app.dvf.etalab.gouv.fr/`

level, but not at the accommodation level) and exploitation (issues related to local context, spatial conflicts, management of the annual sales) [12]. Besides there is no database about rents. Currently, we have not included any price information;
- **List of points of interest (POI)**. IRIS data only provide a number of shops, but not their names. Yet, the presence of a given brand may convey information about the neighbourhood. For instance, an organic shop is usually found in middle or upper class neighbourhoods. Providers could be private companies (e.g., Bing Maps, Here) or collaborative projects (e.g., Open Street Map), but there may be limitations (e.g., data usage policies, numbers of daily queries, management of obsolete locations and updated ones). Although the GeoAlign tool could be used to gather POI with a high degree of completeness [5], it requires a deeper study to select the most relevant brands.

These data sources provide heterogeneous models, formats and semantics, thus an integration step and a quality check is required.

4.2 Data Integration

Relevant data about neighbourhoods are extracted from identified data sources (using dumps, API, queries), but they need to be merged into a single model. Since we manipulate spatial objects, we have chosen the GeoJSON format[20] to store neighbourhoods and their features.

Data integration is a common task [22]. Spatial data is stored according to the OGC standard Geometry Model [19] while IRIS features are scattered in tens of CSV files (one for population, another one for education, etc.), produced at different periods, by different persons and using various concept representations. Thus, data may contain anomalies, inconsistencies or missing values and need to be cleaned through data cleaning or data wrangling processes [15].

First, we have performed a manual schema matching step [6], i.e. the detection of corresponding attributes between data sources. There is no need to use a dedicated tool since the attributes' overlapping is limited and renaming headers in CSV files solves most label heterogeneity issues. The next step is record linkage or data matching [11], which consists in detecting equivalent information (e.g., tuples, entities, values) between data sources, mainly in order to avoid duplicates in the merged database. Each IRIS has its own identifier, but the following modifications may occur from one data source to another: some IRIS were simply missing (e.g., no information about education in this neighbourhood), several IRIS were merged into a single new one or split into smaller units (e.g., due to diverse federations of towns[21], more than 1,250 in 2020). We developed a Python script to enable the detection of these challenging modifications, based on names (both IRIS and city) comparisons and area juxtaposition.

During the integration of data sources into a single database, we computed the surface of polygons. A few neighbourhoods have incorrect boundaries such as overlapping

[20] GeoJSON format, `http://geojson.org/`

[21] Federation of towns, `http://en.wikipedia.org/wiki/Communes_of_France#Intercommunality`

edges in their geometries and they have been corrected using GIS tools. Moreover, there may be some unknown values (e.g., no information about the number of florists in a small town). These values have been replaced by the median score of the column: zero values are not acceptable (already a specific meaning, i.e., a neighbourhood does not have a given feature) and the average is more sensitive to outliers. Another issue is the difference of units and meaning between indicators (e.g., quantities, percentages, quantiles). Some classification algorithms require comparable information. Social science researchers 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 a neighbourhood (e.g., two areas 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. Lastly, we have created a new attribute labelled *grouped indicators*, which reflects the characteristics 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) aggregate the number of bakeries, butcheries, open markets, etc. In total, 30 grouped indicators have been defined and added as features for each neighbourhood.
In the end, we obtain a consolidated MongoDB database named `mongiris`[22]. It contains 49,800 French neighbourhoods fully covering the country. Each of them includes an average of 550 raw indicators from data sources and 30 grouped indicators. A Python API is also provided to facilitate the querying of the database (e.g., retrieve a neighbourhood from its code, get a list of all surrounding neighbourhoods of a given one).

4.3 Predicting environment

Our neighbourhoods include a number of indicators, and they can be used for predicting the environment of any neighbourhood. One of the issue is the high number of indicators (550 in average), which may degrade the performance of machine learning techniques due to over-fitting. Indeed, Lillesand et al. have established that a reasonable number of features f is given by the formula `10f > n > 100f`, with n the size of training data [28]. In our context, we have 270 annotated examples, thus we should use between 3 and 27 indicators as features.
To solve this problem, we reduce the number of features as follows. Descriptive features (17) such as city or neighbourhood names are removed, as well as indicators that are either empty or filled in with the same value[23] (59). INSEE indicators can also be very detailed. For instance, one field counts the number of "tennis courts", a second one stands for the number of "tennis courts with at least one covered", and another one about the number of "tennis courts with at least one lighted". A hierarchy of all indicators has been semi-automatically built, and 213 over-detailed ones have been removed (only "tennis courts" is kept as feature in the previous example). Most neighbourhoods still have many features (362 remaining for those with the maximum of 647). Next, we study

[22] Mongiris database, `http://gitlab.liris.cnrs.fr/fduchate/mongiris`
[23] Indicators from INSEE may not be filled in (empty or default value), especially for data provided by local communities (small towns may not have the resources to manage this task).

the correlation between indicators using the Spearman coefficient [30]. When a pair of indicators obtains 100% correlation, we discard the one which is the most detailed (i.e., at lower levels in our hierarchy).
A last option for reducing the number of indicators is to produce lists of selected features for each variable. Feature importance is a popular method to reach this goal [21], but it may promote the same category of indicators (e.g., population, incomes) to the detriment of category diversity. We therefore propose Algorithm 1, an algorithm based on existing feature selection techniques. It first generates ranked lists of features (lines 3 and 4) based on the Extra Trees (ET) and Random Forest (RF) techniques. The output of these algorithms are merged, and the resulting table sorted with indicators at the higher level of our hierarchy ranked first (lines 5 and 6). To avoid strong impact of a single category, an indicator is removed if its parent is already in the list (lines 8 to 10), else it is added in the resulting set F' (line 12). Merged indicators are then sorted by score (line 13). 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.

Algorithm 1: Selection of relevant features (adapted from [3]).

input : set of indicators I, set of variables $\mathcal{V}$
output: lists of features L_v^k

1 **for** $v \in \mathcal{V}$ **do** /* for each environment variable */
2 $L_v, F' \longleftarrow \emptyset$;
3 $F_v^{ET} \longleftarrow$ ET.rank_features(I) ; /* selected features of Extra Trees */
4 $F_v^{RF} \longleftarrow$ RF.rank_features(I) ; /* selected features of Random Forest */
5 $F \longleftarrow F_v^{ET} \cup F_v^{RF}$;
6 $F \longleftarrow sort(F)$; /* sort from general to specific w.r.t. hierarchy */
7 **for** $f \in F$ **do**
8 $p_f \longleftarrow$ parent(f);
9 **if** $p_f \in F'$ **then**
10 $p_f.score \longleftarrow p_f.score + f.score$; /* boost parent score in F' */
11 **else**
12 $F' \longleftarrow F' + \{f\}$; /* add feature in F' */
13 $F' \longleftarrow sort(F')$; /* sort by descending score */
14 **for** $k \in [10, 20, 30, 40, 50, 75, 100]$ **do** /* generate lists of various size */
15 $L_v^k \longleftarrow$ top-K(F', k);

When features have been selected, the next step is the prediction using machine learning. We are in a classification problem[24] since the objective is to classify a neighbourhood according to the possible values of an environment variable. Thus we generate one instance of a classifier per variable. The main issue is to choose a relevant classifier and

[24] Predicting all variables at the same time is a multi-output classification problem, which is more complicated to manage and more adapted to correlated classes.

to correctly tune its parameters (e..g, thresholds, weights, distance metric), which have a considerable impact on the achieved quality [25]. Due to the complexity of adjusting these parameters, we have developed the `predihood`[25] tool to ease this task.
This machine learning based method is general (i.e., applicable to all variables), but the geographical variable can be directly computed using cartographic systems, as presented in the next part.

4.4 Computing geographical variable

Rather than predicting the direction of the closest city, it is possible to directly compute this value. Let us describe this idea. Starting from a given neighbourhood, we search for a large city by iteratively increasing the search radius. When a large city is found, two representative points for the neighbourhood and the city are calculated, and the direction between these points is then computed.
Several questions arise from this idea. First, how to define a large city? And how to compare the surface of a neighbourhood with the one of a city? How to deal with the centre value, which does not include a clear definition? To enable some flexibility in our approach with regards to these questions, various parameters were introduced:

- MAX_DISTANCE is the maximum radius to search within;
- MIN_CITY is the minimum number of neighbourhoods so that the city can be considered as a large one;
- DISTANCE_CENTRE is the distance below which a neighbourhood is considered in the centre of the found city;
- INCREASE_RADIUS is the distance to be added to the search radius at each iteration;
- REF_POINTS is a method for representing the neighbourhood and the city, either by their centroids or by their nearest points;
- ANGLE_DIRECTIONS is the choice of angles, either *45°* (for all directions) or *30°-60°* (small angles for the four major cardinal points and higher value for corners such as NW, NE, SE and SW, in order to better reflect human estimation) .

Algorithm 2 presents our approach for computing the direction of the largest city given an input neighbourhood. Previously mentioned parameters appear in small capital letters. The recursive function *compute_city* (lines 1 to 9) is in charge of returning the largest city by an iterative search. It collects all neighbourhoods in the considered area, and groups them according to city postcode. The city with the highest number of neighbourhoods is extracted from this counting, and its number of neighbourhoods is compared to a threshold value to decide whether it is a sufficiently large city. If not, the function calls itself by incrementing the search radius. The main procedure starts at line 10. It first computes the reference points of both the neighbourhood and its city, so that it checks whether the former is in the city centre (lines 11 to 14). If not, a search for large city τ is performed (line 15), and its reference point is also calculated (line 17). To compute the direction between both reference points, we compute the difference between their coordinates (lines 18 and 19). To use the arctangent function, we first check

[25] Predihood tool, `http://gitlab.liris.cnrs.fr/fduchate/predihood`

that both points are not located on the same longitude (lines 20 to 23), and we compute the angle between both points (line 24). As the arctangent function returns values in the range [-90°, +90°], the function *get_direction_from_angle* performs some adjustments (e.g., adding 180 for dials on the West side), and it returns the cardinal direction according to the choice of angles (line 25).

Algorithm 2: Computation of geographical variable.

input : neighbourhood η
output: direction

```
1  function compute_city(point, radius)
2      if radius > MAX_DISTANCE then                    /* no large city found */
3          return ∅;
4      N ⟵ get_neighbours(point, radius);
5      c ⟵ extract_largest_city(N) ;          /* neighbourhoods grouped by city */
6      nb_c ⟵ count_neighbourhoods(c);
7      if nb_c > MIN_CITY then  /* number of neighbourhoods above threshold */
8          return c;
9      return compute_city(point, radius + INCREASE_RADIUS);

10 c_η ⟵ get_city(η) ;                                 /* city of the neighbourhood */
11 p_η ⟵ get_ref_point(η, METHOD) ;     /* reference point of neighbourhood */
12 p_{c_η} ⟵ get_ref_point(c_η, METHOD) ;              /* reference point of city */
13 if Δ(p_η, p_{c_η}) < DISTANCE_CENTRE then     /* neighbourhood in city centre */
14     return 'Centre';
15 τ ⟵ compute_city(p_η, 1000) ;          /* large city for the neighbourhood */
16 if τ ≠ ∅ then                                  /* a large city has been found */
17     p_τ ⟵ get_ref_point(τ, METHOD) ;     /* reference point of large city */
18     δ_y ⟵ p_η.y − p_τ.y;
19     δ_x ⟵ p_η.x − p_τ.x;
20     if δ_x = 0 and δ_y > 0 then              /* dial N, avoids divide by zero */
21         return 'North';
22     if δ_x = 0 and δ_y < 0 then              /* dial S, avoids divide by zero */
23         return 'South';
24     angle ⟵ 180 × arctan(δ_y/δ_x)/π ; /* angle in [-90°, +90°], East at 0 */
25     direction ⟵ get_direction_from_angle(angle, ANGLE_DIRECTIONS);
26     return direction;
```

This algorithm is generic due to various parameters, and we show in the next section how their tuning affects the overall quality.

5 Experimental validation

Three experiments are presented in this section: quality results of the prediction with various classifiers (Section 5.1), without rural neighbourhoods (Section 5.2), and quality results of the geographical computation (Section 5.3).
For experiments based on machine learning, we use the popular scikit-learn library for machine learning [33]. The annotated neighbourhoods are split into 80% training data and 20% evaluation data, as recommended in the literature [10]. We use accuracy as quality metric, i.e. the fraction of correct predictions, which is the average quality obtained by 10 runs.
An open discussion concludes this section.

5.1 Predicting with different classifiers

In this first experiment, the main objective is to correctly predict the values for each environment variable of a neighbourhood (Section 4.3). We have used 5 scikit-learn algorithms[26]: 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 [25], and we tested several configurations (e.g., weights, maximum depth in trees, number of neighbours, distance metric) to retain the best one. We also measure the impact of the proposed lists for selecting indicators (see Section 4.3). Tables 4 to 9 provide 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 indicate 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 4 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 5 shows prediction quality for the *usage* variable. The scores without selection range from 51.1% to 64.5%. A few of the smallest lists perform better than the baseline one, but without significant improvement. RF obtains the best results with list L^{50}.
Table 6 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 of features.
Table 7 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) and population characteristics (number of students, employees, farmers, unemployed, etc.).

[26] Other algorithms such as Stochastic Gradient Descent or Nearest Centroid have been tested, but they mostly follow the same trend or achieve insufficient accuracy.

Table 4: Prediction quality for variable *building type* (from [3]).

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

Table 5: Prediction quality for variable *usage* (from [3]).

	LR	RF	KNN	SVC	AB
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

Table 6: Prediction quality for variable *landscape* (from [3]).

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

Table 7: Prediction quality for variable *social class* (from [3]).

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

Table 8 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 9 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. As shown in Section 5.3, these results can be improved by computing the value of the geographical variable instead of predicting it.

To conclude this experiment, we note that best scores range from 33% for geographical position and 50% for *social class* to 60-65% for the remaining four variables. Although algorithms 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. 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. 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

Table 8: Prediction quality for variable *morphological* (from [3]).

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

Table 9: Prediction quality for variable *geographical* (from [3]).

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

algorithms were not suitable, for instance SVC requires many features (best results with all indicators or with largest lists of features).

5.2 Removing rural neighbourhoods

In Section 3.5, we have shown that there was a bias in the dataset due to rural neighbourhoods. This experiment aims at measuring their impact. The 14 rural areas (around 5% of annotated neighbourhoods) have been removed from the dataset, and new accuracy scores (percentage) are shown for the Random Forest classifier in Table 10. Note that results are similar with other classifiers. Numbers inside parenthesis represent the gain or loss compared to the whole dataset of 270 neighbourhoods, and **bold values** highlight the gains. Without rural neighbourhoods, one could expect that classifiers now focus on distinguishing urban areas and thus improve accuracy. On the contrary, their absence involves a decrease in terms of overall quality, up to 13% in some cases. A possible explanation is that these neighbourhoods are easy to predict. Indeed, they usually include indicators with missing values or specific values (e.g., low density, less shops and restaurants, high number of farmers). However, the morphological variable acts as an exception with improved results (up to 7.6%). This is mainly due to the thin border between rural and peri-urban neighbourhoods, which disappear in this setup. This experiment finally confirms the benefit of the feature selection process, since most lists of selected features achieve better results than the list I and smaller lists (L^{10} to L^{30}) tend to minimize the loss.

5.3 Computing geographical variable

In Section 4.4, we have presented a method for calculating the direction of the closest city (i.e., value of the geographical environment variable). This experiment aims at checking whether the proposed method is efficient in terms of accuracy. Remind that we identified six important parameters to compute the direction of the closest city. Preliminary experiments have shown that three of them could be fixed: the REF_POINTS is either set to *centroid* or *nearest points*, but the latter value has problems in case of close locations. The ANGLE_DIRECTIONS parameter can be tuned to *45°* or *30°-60°*,

Table 10: Prediction quality without rural neighbourhoods (RF classifier).

	Building	**Usage**	**Landscape**	**Social**	**Morphological**	**Geographical**
$\boldsymbol{I}$	43.4 (-13.6)	54.1 (-10.4)	58.7 (-2.1)	40.6 (-10.5)	58.0 (-1.7)	29.9 (-3.7)
$\boldsymbol{L}^{10}$	48.4 (-10.9)	56.6 (-4.6)	59.4 (-3.3)	44.1 (-2.5)	61.9 **(+1.9)**	33.5 **(+3.6)**
$\boldsymbol{L}^{20}$	51.2 (-8.8)	57.6 (-6.5)	59.8 (-3.2)	44.1 (-2.5)	66.2 **(+5.0)**	33.5 **(+2.2)**
$\boldsymbol{L}^{30}$	50.5 (-8.4)	58.4 (-2.8)	62.3 **(+1.5)**	42.7 (-6.9)	64.7 **(+3.5)**	30.3 (-1.4)
$\boldsymbol{L}^{40}$	50.5 (-8.8)	56.9 (-6.1)	60.9 (-1.8)	40.2 (-11.6)	64.7 **(+3.5)**	33.1 **(+0.3)**
$\boldsymbol{L}^{50}$	49.1 (-9.8)	55.1 (-9.8)	60.1 (-1.4)	39.9 (-8.2)	66.9 **(+7.6)**	30.6 (-1.5)
$\boldsymbol{L}^{75}$	49.8 (-8.4)	57.6 (-5.8)	60.1 (-2.2)	39.5 (-8.6)	62.6 **(+1.8)**	29.6 (-3.2)
$\boldsymbol{L}^{100}$	47.7 (-9.3)	57.3 (-7.2)	59.1 (-1.7)	38.8 (-12.3)	60.8 **(+1.1)**	29.9 (-3.7)

but the second option outperforms in all tests. In addition, the INCREASE_RADIUS parameter, which does not affect much the results, is set to 2 kilometres. Table 11 provides accuracy results when varying the remaining three parameters. A single run (270 neighbourhoods) takes about five minutes.

Table 11: Accuracy for variable *geographical* according to parameters maxd (MAX_DISTANCE in kms), minc (MIN_CITY) and dcent (DISTANCE_CENTRE in kms).

maxd	**minc**	**dcent**	**Acc.**
30	15	0.5	34.0
30	15	1.0	35.6
30	15	1.5	**38.9**
30	15	2.0	**37.7**
30	20	0.5	33.6
30	20	1.0	35.2
30	20	1.5	**38.1**
30	20	2.0	**37.3**
30	25	0.5	30.8
30	25	1.0	32.4
30	25	1.5	**34.8**
30	25	2.0	**34.0**

maxd	**minc**	**dcent**	**Acc.**
40	15	0.5	32.8
40	15	1.0	34.4
40	15	1.5	**37.3**
40	15	2.0	**36.4**
40	20	0.5	32.8
40	20	1.0	34.4
40	20	1.5	**37.3**
40	20	2.0	**36.4**
40	25	0.5	30.8
40	25	1.0	32.4
40	25	1.5	**34.8**
40	25	2.0	**34.0**

maxd	**minc**	**dcent**	**Acc.**
50	15	0.5	31.6
50	15	1.0	32.4
50	15	1.5	**34.8**
50	15	2.0	**34.0**
50	20	0.5	31.6
50	20	1.0	32.4
50	20	1.5	**34.8**
50	20	2.0	**34.0**
50	25	0.5	30.0
50	25	1.0	30.8
50	25	1.5	**33.2**
50	25	2.0	**32.4**

We observe that quality decreases as the maximum distance grows. In France, a large city is typically found within 40 kilometres. As for the minimal number of neighbourhoods to be considered as a large city, 15 is the best trade-off because quality slightly decreases with higher numbers. Another comment is that a higher DISTANCE_CENTRE value (bold scores) achieves better results than smaller distances. This means that centred neighbourhoods are usually up to 2 kilometres around the city's reference point. The best score is 38.1% (green cell), which is slightly above the best scores of the predictive approach (33%). In addition, detailed results enable a better understanding of the complexity for this variable. First, several issues are related to the **city definition**. Indeed, it is based on a minimum number of neighbourhoods, which is a hard thresh-

old value and thus not adaptative to different situations. We also found cases in which two cities were roughly at the same distance of the considered neighbourhood: our algorithms selects the one with the highest number of neighbourhoods, but experts may have considered other elements such as direct roads, natural obstacles, etc. Last, a frequent case deals with neighbourhoods inside cities: the returned direction depends on the DISTANCE_CENTRE parameter, either *centre* when the neighbourhood is below the threshold value, or one of the eight cardinal points otherwise. In small cities, a high threshold value tends to incorrectly return *centre*. Two other issues are related to the **subjectivity of the expertise**. When manually checking for the direction, experts may not take into account the whole surface of both city and neighbourhoods (thus resulting in predicting the next value of the dial, e.g., South East instead of East). With an algorithm (using centroids and a dial divided in 8 parts), results are consistent between them, but may still be inaccurate due to lack of human perception (e.g., a city which includes a large forest inside its borders). Besides, Google Maps may give focus to small towns (e.g, *La Chaise-Dieu*, a touristic village of 700 inhabitants, appear at the same level of bigger cities with tens of thousand residents). Finally, the last problems concern the **three major French cities** (Paris, Lyon, Marseille), which are divided into boroughs. Inside one borough, values may be different (between *centre* and another direction) which makes the computation more difficult. Next, neighbourhoods in surrounding cities of Paris, Lyon and Marseille are usually sufficiently populated to be elected as large city, and if the choice between the major city and the smaller surrounding cities is quite clear for a human, our algorithm may be wrong because it stops when the closest city satisfies the minimum number of neighbourhoods. All these identified issues show that these results could still be improved.

5.4 Discussion

These results are promising and show that an approach based on statistical division units (IRIS) provides an acceptable quality with regards to neighbourhood perception. Improvements are still possible, especially by addressing the representativeness issues presented in Section 3.5 or by programming heuristics to enhance the computation of geographical position (e.g., around major cities). The number of annotated neighbourhoods is also limited due to the time-consuming manual annotation (as explained in Section 3.2), which may negatively impact the results.

We finally provide answers to research questions about the trends of people who moves to another city in the context of job transfers. Households from our dataset mostly belong to the middle and upper-middle social classes, and they usually stay in this category, which means that their choice of neighbourhood should not drastically change. Half of them were occupants prior to moving, one fourth owned their accommodation while the remaining fourth was hosted (mostly students about to leave their parent's home). After the move, a large majority ends up as occupants, either because they need time to discover their new city before buying, or because their new job is temporary, or they are first-time workers. Households were more or less fairly divided into the 5 types of buildings (except for under-represented *large housing estates*), but half of them resides in buildings after the move. This can be explained by the weight of first-time

workers, who may not afford to live in houses or mixed areas. In a similar fashion, residents tend to leave residential neighbourhoods in favour of shopping zones, an expected situation due to the job transfer context. This trend is confirmed by the morphological position, since rural and peri-urban and even urban become less attractive to the benefit of central areas (which doubles its score). Landscape is partly correlated to morphological position and neighbourhood usage, and urban areas, which accounts for one third of the shares before moving, represent half of the neighbourhoods in the end. This comparison between start and arrival neighbourhoods enables a better understanding of residential choices in this context.

6 Conclusion

In this paper we first present a new set of six variables for describing the environment of neighbourhoods. They were derived from a manual observation of various elements such as customer's information, aerial views and raw indicators about the neighbourhood. Due to the job transfer context, most of the 270 annotated neighbourhoods were located in peri-urban areas, which results in a bias compared to the 49,800 neighbourhoods in France. The main challenge was to check whether this manual observation process could be automated to describe neighbourhood environment of the whole country. We first integrated different data sources into the single database `mongiris`[22]. Next, we proposed two approaches for computing environment variables, based on machine learning techniques and on a spatial computation of the *geographical position* variable. The former approach, implemented in the `predihood`[25] tool, requires a specific pre-process for selecting a subset of indicators while the latter involves different parameters to take into account open questions such as the definition of a large city. Our experimental validation confirms that it is possible to use statistical unit divisions as neighbourhoods and to predict their environment with an acceptable quality. Yet, this computation is still a difficult task and our results could be improved.

We envision different perspectives to this work. First, we have shown that the number of examples is low (less than 1% of the dataset) and not sufficiently heterogeneous. Increasing and varying the number of examples could therefore help in improving the quality. Designing heuristics for spatial computation or using different classifiers than the ones provided by scikit-learn are also clues for achieving a better quality. Another possibility could be the generation of a bigger synthetic dataset, which share similarities with the 49,800 neighbourhoods. The `mongiris` database includes hundreds of indicators for each neighbourhood. Other data sources such as the prices of sold accommodations or the type/brand of specific points of interest were presented but not integrated due to the need of enhanced thinking. Using new indicators and applying our feature selection algorithm could reveal whether they are useful for the prediction. Besides, indicators from INSEE are updated every couple of years. Observing the dynamics of a few indicators could reveal trends about the environment of neighbourhoods (e.g., evolution of unemployed people). Other application domains may have different needs about the environment (e.g., pollution degree, stopover possibilities for migratory birds), and a last perspective is to discuss with researchers and practitioners from other fields to adapt the description of the environment.

References

1. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Government Information Quarterly **32**(4), 399–418 (2015)
2. Authier Jean-Yves, Bacque Marie-Hélène, G.P.F.: Le quartier. Enjeux scientifiques, actions politiques et pratiques sociales. La Découverte (2007), https://www.cairn.info/le-quartier--9782707150714.htm
3. Barret, N., Duchateau, F., Favetta, F., Bonneval, L.: Predicting the environment of a neighborhood: a use case for france. In: International Conference on Data Management Technologies and Applications (DATA). pp. 294–301. SciTePress (2020)
4. Barret, N., Duchateau, F., Favetta, F., Miquel, M., Gentil, A., Bonneval, L.: À la recherche du quartier idéal. In: Extraction et Gestion des Connaissances. p. 429–432 (2019)
5. Barret, N., Duchateau, F., Favetta, F., Moncla, L.: Spatial entity matching with geoalign. In: ACM GIS SIGSPATIAL. p. 580–583. ACM (2019)
6. Bellahsène, Z., Bonifati, A., Rahm, E.: Schema matching and mapping. Springer (2011)
7. Bigot, R., Croutte, P., Müller, J., Osier, G.: Les classes moyennes en europe. Le CRÉDOC, Cahier de recherche **282** (2011)
8. Bonneval, L., Duchateau, F., Favetta, F., Gentil, A., Jelassi, M.N., Miquel, M., Moncla, L.: Étude des quartiers : défis et pistes de recherche. In: Extraction et Gestion des Connaissances (2019), http://dahlia.egc.asso.fr/atelierDAHLIA-EGC2020.html
9. Bourdieu, P.: What makes a social class? On the theoretical and practical existence of groups. Berkeley journal of sociology **32**, 1–17 (1987)
10. Bruce, P., Bruce, A.: Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly (2017), https://books.google.fr/books?hl=fr&lr=&id=ldPTDgAAQBAJ
11. Christen, P.: Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media (2012)
12. Coulondre, A.: Ouvrir la boîte noire des marchés du logement. Métropolitiques (2018), https://metropolitiques.eu/Ouvrir-la-boite-noire-des-marches-du-logement.html
13. Cranshaw, J., Schwartz, R., Hong, J., Sadeh, N.: The livehoods project: Utilizing social media to understand the dynamics of a city. In: ICWSM Conference (2012)
14. Dennis, G.: The American Class Structure. New York Wadsworth Publishing (1998)
15. Donoho, D.: 50 years of data science. Journal of Computational and Graphical Statistics **26**(4), 745–766 (2017). https://doi.org/10.1080/10618600.2017.1384734, https://doi.org/10.1080/10618600.2017.1384734
16. Frank, L.D., Sallis, J.F., Saelens, B.E., Leary, L., Cain, K., Conway, T.L., Hess, P.M.: The development of a walkability index: application to the neighborhood quality of life study. British journal of sports medicine **44**(13), 924–933 (2010)
17. Galster, G.: On the nature of neighbourhood. Urban studies **38**(12), 2111–2124 (2001)
18. Garau, C., Pavan, V.M.: Evaluating urban quality: Indicators and assessment tools for smart sustainable cities. Sustainability **10**(3), 575 (2018)
19. GIS, O.: Consortium Inc. Opengis simple features specification for SQL (1999), http://www.gismanual.com/relational/99-049_OpenGIS_Simple_Features_Specification_For_SQL_Rev_1.1.pdf
20. Guest, A.M., Lee, B.A.: How urbanites define their neighborhoods. Population and Environment **7**(1), 32–56 (1984), https://doi.org/10.1007/BF01257471
21. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research **3**(3), 1157–1182 (2003)
22. Halevy, A., Rajaraman, A., Ordille, J.: Data integration: the teenage years. In: Proceedings of the 32nd international conference on Very large data bases. pp. 9–16. VLDB Endowment (2006), http://portal.acm.org/citation.cfm?id=1164127.1164130

23. Hoyt, H.: The structure and growth of residential neighborhoods in American cities. Scholarly Press (1972)
24. Jenks, M., Dempsey, N.: Defining the neighbourhood: Challenges for empirical research. The town planning review pp. 153–177 (2007)
25. Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
26. Le Falher, G., Gionis, A., Mathioudakis, M.: Where Is the Soho of Rome? Measures and Algorithms for Finding Similar Neighborhoods in Cities. ICWSM **2**, 3–2 (2015)
27. Leong, M., Dunn, R.R., Trautwein, M.D.: Biodiversity and socioeconomics in the city: a review of the luxury effect. Biology Letters **14**(5), 20180082 (2018)
28. Lillesand, T., Kiefer, R.W., Chipman, J.: Remote sensing and image interpretation. John Wiley & Sons (2015)
29. Liu, Y., Wei, W., Sun, A., Miao, C.: Exploiting geographical neighborhood characteristics for location recommendation. In: Conference on Information and Knowledge Management. p. 739–748 (2014). https://doi.org/10.1145/2661829.2662002, https://doi.org/10.1145/2661829.2662002
30. Mukaka, M.M.: A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal **24**(3), 69–71 (2012)
31. Oberti, M., Préteceille, E.: La ségrégation urbaine. La Découverte (2016)
32. Pan Ké Shon, J.L.: La représentation des habitants de leur quartier: entre bien-être et repli. Économie et statistique **386**(1), 3–35 (2005)
33. 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., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
34. Reibel, M.: Classification approaches in neighborhood research: Introduction and review. Urban Geography **32**(3), 305–316 (2011). https://doi.org/10.2747/0272-3638.32.3.305, https://doi.org/10.2747/0272-3638.32.3.305
35. Salim, F.D., Dong, B., Ouf, M., Wang, Q., Pigliautile, I., Kang, X., Hong, T., Wu, W., Liu, Y., Rumi, S.K., et al.: Modelling urban-scale occupant behaviour, mobility, and energy in buildings: A survey. Building and Environment **183**, 106964 (2020)
36. Sigaud, T.: Accompagner les mobilités résidentielles des salariés: l'épreuve de l'entrée en territoire. Espaces et sociétés **162**(3), 129–145 (2015)
37. Tabard, N.: Des quartiers pauvres aux banlieues aisées: une représentation sociale du territoire. Economie et statistique **270**(1), 5–22 (1993)
38. Takada, M., Kondo, N., Hashimoto, H.: Japanese study on stratification, health, income, and neighborhood: study protocol and profiles of participants. Journal of epidemiology **24**(4), 334–344 (2014)
39. Tang, E., Sangani, K.: Neighborhood and price prediction for san francisco airbnb listings (2015), cs229.stanford.edu/proj2015/236_report.pdf
40. Yuan, X., Lee, J.H., Kim, S.J., Kim, Y.H.: Toward a user-oriented recommendation system for real estate websites. Information Systems **38**(2), 231–243 (2013). https://doi.org/https://doi.org/10.1016/j.is.2012.08.004, http://www.sciencedirect.com/science/article/pii/S0306437912001081
41. Zhang, A.X., Noulas, A., Scellato, S., Mascolo, C.: Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In: Social Computing. pp. 69–74. IEEE (2013)