A framework for evaluating urban land use mix from crowd-sourcing data

Luciano Gervasoni, Martí Bosch, Serge Fenet, Peter Sturm

A framework for evaluating urban land use mix from crowd-sourcing data

Luciano Gervasoni*, Marti Bosch†, Serge Fenet‡ and Peter Sturm§
*‡§Inria Grenoble – Rhône-Alpes, France
*§Univ Grenoble Alpes, Lab. Jean Kuntzmann, Grenoble, France
*§CNRS, Lab. Jean Kuntzmann, F-38000 Grenoble, France
†Ecole Polytechnique Fédérale de Lausanne
‡Univ. Lyon 1/INSA de Lyon, LIRIS, F-69622, Lyon, France
Email: *luciano.gervasoni@inria.fr, †marti.bosch@epfl.ch, ‡serge.fenet@liris.cnrs.fr, §peter.sturm@inria.fr

***Abstract*—Population in urban areas has been increasing at an alarming rate in the last decades. This evidence, together with the rising availability of massive data from cities, has motivated research on sustainable urban development.**

In this paper we present a GIS-based land use mix analysis framework to help urban planners to compute indices for mixed uses development, which may be helpful towards developing sustainable cities. Residential and activities land uses are extracted using OpenStreetMap crowd-sourcing data. Kernel density estimation is performed for these land uses, and then used to compute the mixed uses indices. The framework is applied to several cities, analyzing the land use mix output.

I. Introduction

The number of people living in cities has been increasing considerably since 1950, from 746 million to 3.9 billion in 2014 [1]. More than 66% of the world's population are projected to live in urban areas by 2050, against 30% in 1950. Even though the global rural population is expected to only slightly decline by 2050, the continuing population growth and urbanization are thus projected to add 2.5 billion people to the world's urban population by 2050. This situation brings new challenges on how to conceive cities that host such amounts of population in a sustainable way. This sustainability question should address several aspects, ranging from economical to social and environmental matters among others.

In this paper, we focus on the formalization of a measure of mixed use development or land use mix (LUM) in a city, i.e. how different urban land uses are located close enough to each other. Such an urban land use mix has been largely proven to contain beneficial outcomes in terms of sustainability. Indeed, dense and compact cities have been largely discussed in the literature [2], [3], [4], [5] in terms of sustainable development. While the desired degree of compactness remains an open question, a wide agreement exists on the positive impact of land use mix: it has been proven to be positive not only in terms of sustainable development, but also to contribute to several aspects such as societal, health and public transportation among others.

In order to aid urban planners, we propose a framework to capture land use mix by means of crowd-sourced data. The output is a Geographic Information System (GIS) containing the degree of land use mixture of cities. For this, land uses extraction is performed from the input data. Then, spatial statistics kernels are estimated, and used to define the final mix degree. Even though it is desirable to achieve a degree of land use mix as high as possible, in this study we do not perform numerical interpretations of its desired value in order to contribute to sustainable development, and we focus only on the computation of the spatialized land use mix map.

The manuscript is organized as follows. In Sect. II, an overview of the general context is presented, followed in Sect. III by an analysis of existing indicators. Sect. IV outlines the data used in our framework. The proposed framework is presented in Sect. V. Finally, we present an application to a few cities in Sect. VI, followed by the conclusions and a discussion of future work.

II. Context

A. Motivation

A clear steady migratory pattern from rural to urban areas can be observed through history, becoming more and more important since the nineteenth century. This process, that can be observed in all countries, has lead to a constant increase of urban population: now more than half of the human population lives in urban areas, and cities are consequently facing an ever-increasing population and struggle to continue ensuring essential services and quality of life. This concentration leads to new demands on how to conceive cities in a way that promotes sustainability and efficient resource consumption. Still, urban processes take place as a consequence of different interacting factors, linked between them in such a way that the resulting process is complex.

Given the increasing number of people living in cities, understanding the underlying complexity of these urban patterns is becoming a pressing issue. As a result of the increasing availability of massive amounts of urban data, it is now possible to analyze the ways citizens interact within cities. More importantly, understanding the complex feedback loop linking citizens' and cities' development could allow us to better solve future sustainability questions. Consequently, considerable efforts have been emerging recently, aiming to understand cities' sustainability and increase it by using newly-available data sources.

The present work was motivated by three main observations:

- First, as sustainability is becoming the central common issue in several scientific fields, we advocate that computer science tools, and more specifically data mining and model building applied to real world data, can help to better understand the dynamics of complex socioeconomic systems. These tools will allow a better comprehension and efficient decision making towards improved sustainability and resilience.
- Second, several measures of land use mix exist, see [6] and references therein. Most can be categorized in divisional or integral measures. The former are more expressive than the latter, delivering a single value for each divisional unit. However, all of them are sensitive to the scale used to aggregate data to compute the measure. In this sense, a desirable measure should be independent of the scale of analysis.
- Third, even if crowd-sourcing data is growing with both increasing precision and frequency, missing data will always occur to some degree, impacting the above measures. Hence, statistical tools are vital in this context, in order to better capture the cities' underlying features.

B. Brief historical background

Before the 20th century, mixed use was a natural trend in city development, since scarcity of transportation possibilities imposed geographical proximity constraints on the location of every-day activities. Early in the 1900s in the United States, zoning practices started to assign unique land uses, inducing segregation between residential and activities uses. This has been occurring particularly from the 1910s to the 1950s, where mixed use development was quite infrequent. During this period, segregated development was the norm. In later years, starting from the 60s and 70s, and after having been neglected for several decades, mixed uses started appearing again, bringing altogether its advantages in various aspects to the society.

More recently, since the late 90s, mixed use development has (re-)emerged as a major concept in the context of urban planning. The Congress for the New Urbanism ([7], http://www.cnu.org) campaigns for "pedestrian-friendly, and mixed-use" neighborhoods, and the Smart Growth Network [8] includes mixing uses as one of its ten principles. The concept of Transit-Oriented Development and several transit agencies also support the provision of a mix of land uses [9]. In order to attain this state of highly mixed development, both intensity and diversity of land uses are promoted, together with the integration of the segregated uses.

C. Impacts of mixed use development

In this section, we briefly outline the principal issues and effects related to sustainability within the presence of high/low mixed use development. Both direct and indirect relations between sustainability and mixed uses has been intensively studied in the literature, as quickly described below:

- *Urban sprawl:* Urban sprawl refers to the process of spatial expansion of the population moving away from central urban areas into sparse, mono-functional and usually car-dependent communities. This suburbanization, i.e. the creation of suburban areas, has been growing with an ever-increasing rate in the last century. As a consequence, low urban land use mix is one of the distinctive characteristics related to this urban sprawl phenomenon (i.e. residential or industrial sprawl).
  In the classical literature, the urban sprawl process has been largely linked to negative effects in terms of environment, health, society and economy [10]. More recently, and in conjunction with the appearance of the concept of sustainable development of cities, urban planners are increasingly taking into account the consequences of sprawling [11] – even if the negative consequences of sprawl have been already addressed early in the 70s.
  The lack of coordinated land use planning and its negative consequences start to appear as one of the key components of urban sprawl [12], [13], [14], [15].
- *Transportation:* In terms of public transportation, sprawling areas with low mixed use development are not sustainable [9]. In this direction, those sprawling areas with high dispersion involve the fact that large distances have to be driven for low demands, causing inefficiency in the transportation area.
- *Health:* It was shown that land use diversity has an impact on different types and amounts of physical exercises [16]. This means that urban design may directly impact the physical activity of citizens, which is an important health issue nowadays. Thus, a higher mix of uses is associated to more walking trips [16] and less obesity [17]. For the latter, it was claimed to be effective as health intervention.
- *Car dependency:* The choice of travel behavior, considering modal choice and distance traveled, has been strongly related to urban land use balance in [18], [19]. The presence of nearby commercial land uses was associated with both short commuting distances and low vehicle ownership rates [18], which is positive in environmental terms since the sources of pollutant emissions are reduced. Studies on designing urban forms that reduce vehicle dependence [20] determined that a lower automobile dependence requires a minimum value of "urban intensity", (i.e. residents and jobs per hectare), which is directly related to the urban land use mix concept.

III. Limits of existing indicators

Efforts on measuring land use mix abound in the literature. A comprehensive review is provided in [6], whereas each measure has its own strengths and weaknesses. The suitability of a certain measure of urban land use mix relies directly on two aspects. Firstly, it depends on the context of application, where the intended use is the major driver for the measure's behavior. Secondly, the input data characteristics condition the performance of the different measures, in aspects such as the presence of noise among others.

Urban land use mix measures are largely inspired by landscape ecology metrics, or spatial statistics analysis [21]. Several indices have been proposed, such as the Atkinson index, the Clustering index, the Dissimilarity index, the Exposure index and the Gini index. In addition, Shannon entropy-based metrics were also proposed.

The input data-sets are rarely discussed. They are normally obtained to study a certain region without further questioning. On the one hand, some of these data-sets are private, obstructing the reproducibility of the research. On the other hand, open data-sets exist, but are usually limited to certain regions. It is well known that results are sensitive to the input data-sets, and the usage of heterogeneous data-sets across the world may have led to a trend in developing ad-hoc measures which model the desired behavior for the input data-set, while eventually limiting results to that region of analysis.

Several works in the literature have to cope with the modifiable areal unit problem (MAUP), naturally produced by the aggregation of data. Outputs of measures depend on the chosen geography of division. The sensitivity of the results to the chosen aggregation strategy is undesirable in any application and is a major concern when trying to compare metrics. The usage of geo-localized input data, together with the estimation of a continuous surface by means of a kernel density estimation method allows to cope with the MAUP problem [22].

IV. Dataset

The OpenStreetMap[1] (OSM) is a collaborative project to create a free editable map of the world; it is a prominent example of volunteered geographic information (VGI). It is a knowledge collective that provides user-generated street maps [23]. Volunteers across the world share geographic information on OSM in various ways and are sometimes considered as "intelligent sensors" [24].

Since its creation, the project has been increasingly used across the world for a wide variety of purposes. Quality metrics have been proposed in [25], [26], [27], [28], followed by different quality assessments, in particular for different countries. For instance, it has been concluded that the quality is "fairly accurate" for England [29], and it is even shown that OSM data is superior to the official data-set for Great Britain Meridian 2. Thus, the previous work has been extended for France [30]. Studies focusing on the street network of Germany have been also been conducted in [31], where it is concluded that the data-sets can be considered complete in relative comparison to a commercial data-set.

In addition, the OSM data-set for Hamburg already covers about 99.8% of the street network [32] according to the surveying office of Hamburg. The latter study also remarks that "Besides the street network, the real advantage of the data-set is the availability of manifold points of interest". These points of interest allow for a deeper understanding of city dynamics, enriched with the provided location and embedded information. In China, the volume of points of interest has been increasing substantially, e.g. nine-fold in the period 2007-2013 [33].

In this work, we use OSM data for analyzing cities' land use mix, which leads to numerous advantages. Firstly, worldwide coverage is a huge asset. It is the first step towards cities metrics for further comparisons using homogeneous data-sets. Secondly, it is being continuously updated by means of crowdsourcing, allowing to adapt its information to the rapid urban changes present nowadays. In addition, there exists a very active community, which iteratively improves the data precision and completeness. Thirdly, OSM is Open data. The fact that data is freely available for everyone allows for coherence and valid comparison between different contributions in the field of urban planning. As well, results are reproducible for anyone, an important aspect for the community, towards improval of these tools. Finally, geo-localized data provide a great advantage in terms of spatial location specificity, granting the possibility of a finer analysis in comparison to gridded data for instance.

Currently, the biggest limitation related to OSM is still missing data. Even though it can occur quite frequently, it is reduced drastically in big cities where lots of contributors exist, coming from a growing community of crowd-sourcing.

In contrast, Land Use Land Cover (LULC) data has several limitations compared to OSM. For LULC, availability is generally restricted to country level boundaries. Then, even though resolution has been improving considerably, intensities of activity and residential uses are not captured. Further, the classification by means of using a single allocation of land use, and the aggregated land use categories provided are insufficient to infer cities' land use mix [16]. Last but not least, accounting for the increasing speed of change in urban dynamics, LULC data-sets can get outdated in relatively short time.

V. Method

In this section, we describe the computing pipeline that allows us to compute the continuous spatial representation of the land use mix from raw OSM data.

A. Data extraction

We explain here how to retrieve the different land uses which will be analyzed later to define the land use mix degree. The OSM data for a given geographical area, defined by its geographical bounding box, is retrieved in the shapefile format, obtained from *Mapzen Metro Extracts*[2]. These files are the result of the *osm2pgsql* process, where the OSM data is converted to postGIS-enabled PostgreSQL databases. From these files, points and polygons are used as input data for further processing.

We first perform a classification of both the points of interest (POIs) and the polygons based on residential or activity uses. It is done following the OSM Wiki[3]. Points and polygons are

[1] http://www.openstreetmap.org/

[2] https://mapzen.com/data/metro-extracts/

[3] https://wiki.openstreetmap.org/

Key	Value
Activities classification	
amenity	bar, pub, restaurant, biergarten, cafe, fast_food, food_court, ice_cream, pub, restaurant, college, kindergarten, library, public_bookcase, school, music_school, driving_school, language_school, university, fuel, bicycle_rental, bus_station, car_rental, taxi, car_wash, ferry_terminal, atm, bank, bureau_de_change, baby_hatch, clinic, dentist, doctors, hospital, nursing_home, pharmacy, social_facility, veterinary, arts_centre, brothel, casino, cinema, community_centre, fountain, gambling, nightclub, planetarium, social_centre, strip-club, studio, swingerclub, theatre, animal_boarding, animal_shelter, courthouse, coworking_space, crematorium, dive_centre, dojo, embassy, fire_station, gym, internet_cafe, marketplace, police, post_office, townhall
shop	*
building	commercial, office, industrial, retail, warehouse, cathedral, chapel, church, mosque, temple, synagogue, shrine, civic, hospital, school, stadium, train_station, transportation, university, public, kiosk, garage, garages, hangar, stable, cowshed, digester
leisure	adult_gaming_centre, amusement_arcade, beach_resort, dance, hackerspace, ice_rink, pitch, sports_centre, stadium, summer_camp, swimming_area, water_park
landuse	commercial, industrial, retail, port, quarry, salt_pond, construction, military, garages
Residential classification	
building	hotel, farm, apartment, apartments, dormitory, house, residential, retirement_home, terrace, houseboat, bungalow, static_caravan, detached
Other land use classification	
landuse	cemetery, landfill, railway, water, reservoir, basin, allotments, conservation, farmland, farmyard, forest, grass, greenfield, greenhouse_horticulture, meadow, orchard, pasture, peat_cutting, plant_nursery, recreation_ground, village_green, vineyard

TABLE I
LAND USES CLASSIFICATION.

associated to a certain land use type according to their input information as denoted in Table I.

In OSM data-sets a great quantity of polygons are tagged as buildings, without any additional information. In such cases, the polygons' land uses are inferred as detailed in Table II. First, polygons containing a defined key for land use are sorted in residential, activity, and other uses (e.g. forest, water). Let P_R, P_A and P_O denote the three sets of polygons associated with these land uses. All polygons tagged as buildings are processed to estimate their containing land use. For a polygon P, the land use is estimated as follows:

$$LU(P) = LU(p) \quad if\ \exists p \text{ with } \min_{p \in P_R \cup P_A \cup P_O} A(p)\ ,\ P \subseteq p$$

where $A(\cdot)$ denotes a polygon's area, and $LU(\cdot)$ its land use. Note that polygons with different land uses might overlap. We consider as relevant the information contained in the smallest encompassing polygon whose land use is known.

Sometimes, on the contrary, it might happen that a polygon which needs to be inferred is not contained in any polygon with a defined land use value. A residential purpose is then assumed for these polygons. This hypothesis is made in order to somewhat counter the fact that in OSM data-sets, still relatively few residential tags exist. In the future, this hypothesis may be dropped.

Key		Classification
leisure = *Activity*		Activity
amenity = *Activity*		Activity
shop = *Activity*		Activity
building = *Activity*		Activity
building = Residential		Residential
building = yes	landuse = *Activity*	Activity
	landuse = Residential	Residential
	landuse = Null	To be inferred
landuse = Residential	building = Null	Residential inferring
landuse = *Activity*	building = Null	Activity inferring
landuse = Other	building = Null	Null inferring

TABLE II
PROCEDURE FOR EXTRACTING LAND USE FROM OSM DATA.

Experimentally, it has been observed that polygons which are uniquely tagged as buildings (i.e. no other complementary information exists) correspond in great part to residential buildings. Firstly, it is important to note that otherwise, the building polygon would have been tagged with complementary information (e.g. any activity classification). Secondly, there exists no land use tag under the area of the building, which is related to a non-residential land use (i.e. natural or industrial land uses).

Later, all residential buildings with a computed squared footage smaller than 12 squared meters, are filtered out. This is done due to the high probability that such buildings are a false positive in the classification, rather than a true residential building.

Finally, the polygons are converted to POIs by computing their centroid. The objective is to obtain a full distribution of geo-referenced points which contain information on residential or activity land uses, throughout a city. The results of this procedure are depicted in Fig. 1 for London, England, where transparency is used to clearly distinguish highly concentrated zones.

B. Kernel Density Estimation

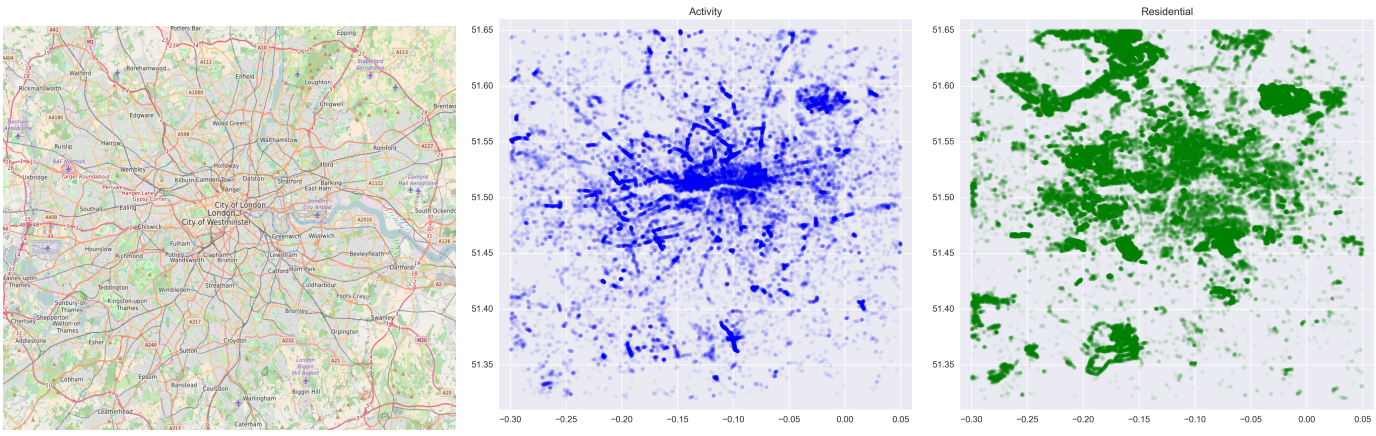
In this section we describe the process of performing a Kernel density estimation (KDE), a statistical process for density estimation, for both activity and residential uses. As stated in [34], the KDE is a well-recognized technique for visualizing and analyzing complex and technical data in a clear and understandable way to non-mathematicians.

The KDE infers a density function f, given an i.i.d. sample $x_1, x_2, ..., x_n$ from the corresponding probability distribution. The kernel density estimator is defined as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\|x - x_i\|}{h}\right), \tag{1}$$

with kernel $K(\bullet)$ and h a smoothing parameter denominated bandwidth. In our case, x_i are two-dimensional vectors containing the latitude and longitude coordinates.

Then, given a set of data points, the KDE interpolates a continuous surface using a given kernel (defined e.g. by a Gaussian/normal function). This procedure has two major advantages, namely spatial smoothing and spatial interpolation. Both aspects are relevant in our context, where missing data



(a) (b) (c)

Fig. 1. (a) OpenStreetMap of London, England. (b) and (c) depict, respectively, the extracted Activity and Residential uses.

can occur. Thus, centers of activities and residential purposes can be smoothed, while interpolation is done locally to cope with missing data.

As described before, the KDE is controlled by a weighting function K and the bandwidth h. For the weighting function, we assume that both residential and activity uses have a normally distributed spatial influence. For instance, the impact of an activity is maximal in its original location, and reduces with increasing distance, according to the normal distribution, which is the default choice in most works using KDE.

The bandwidth is associated with the distance of influence. It is important to mention that several techniques for an automatic computation of the bandwidth exist in the literature, but in several applications the desire is to reduce complexity during the estimation at the expense of precision. In our case, we are mostly interested in capturing information (i.e. residential and activities points) within local neighborhoods. It is thus natural to relate the bandwidth to the extent of walkable distances.

In [35], it is suggested that 400 meters correspond to the distance an average American will walk rather than drive. According to [36], this distance is also considered as the greatest distance someone is likely to walk to a transit station. Similarly, in [37] the implementation of an average destination within a distance of between 400 and 450 meters is promoted. Nonetheless, some studies promote walkable distances slightly higher than the one adopted, as for example in [38] that studied the association between physical activity and the mixture of destinations located within 400 and 1500 meters of residents' homes.

In our context, and for the purpose of favoring the neighborhoods with high uses mix, we decided to adopt the suggested value of 400 meters. This value defines the spatial bandwidth during the density estimation procedure.

As mentioned before, the KDE is a very efficient tool to smooth spatial data and interpolate locally-missing data. However, in the case of strict geographical borders like frontiers or coastal regions, it can lead to an undesirable over-smoothing and the estimation of non-existent information. Such a case will be studied in detail in further work.

As done in [6], we chose to adopt two land use types: residential and activities (i.e. related to non-residential uses, such as shopping, leisure, etc.). Consequently, all geo-localized POIs from the procedure described above, are used to compute one KDE for each category: residential and activities KDE. The probability density function is then evaluated in a grid of points covering the region of analysis with any resolution that a user could want. In practice, we construct a grid with a step of 100 meters. This value was set in order to aid the visualization of the land uses densities on the different neighborhoods within a city. Finally, the computed grid points are normalized for each computed KDE.

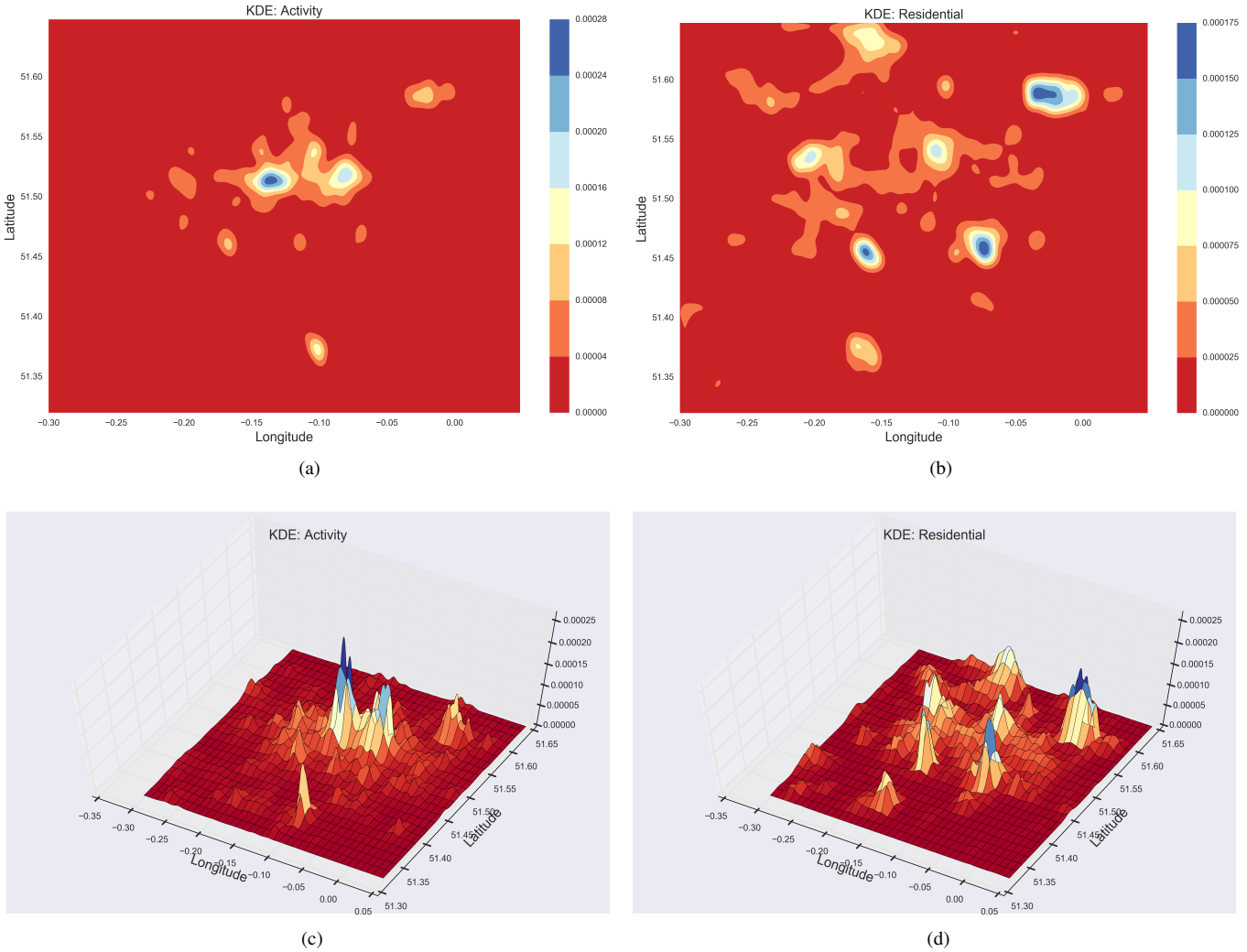
The resulting KDE's are depicted in Fig. 2, applied to the city of London, England.

C. Computation of land use mix

In this section we focus on the computation of a land use mixture measure, given the input densities estimated for residential and activity uses respectively. The goal of this measure is to determine to what extent the spatial configuration of land uses is well distributed in a city, as well as to assess the co-occurrence of both residential and activity uses in neighborhoods. From a global point of view, it is desirable to achieve a high degree of land use mix in most neighborhoods.

As expressed in [6], measures for land use mix must address two underlying concepts: distance and quantity. Distance describes the proximity between different land uses. We compute this closeness, or influence of a certain land use on another one, according to the walkability distance, as denoted in Sect. V-B. The quantity is modeled in the density estimation procedure, given the input of geo-localized activity and residential spots.

In accordance with the notion of distance and quantity, the land use mix measurement must depict the regions where a



(a) (b)

(c) (d)

Fig. 2. Estimated KDE's for London, England, for (a), (c) activity and (b), (d) residential uses.

high use mix exists. By contrast, it will also point out the poorly distributed usages with low diversity of land uses.

As expressed before, this framework intends to be an aid for urban planners, and thus the visualization of geo-spatial land use mixture is of paramount importance to provide support during the decision-making procedure. In this context, this framework is also designed to expose the importance of the different sub-regions within a city. The grid meshes computed from KDE results allow for determining the different sub-region's importance in terms of land use intensity. This provides a great aid for urban planners on the need for improving and planning mixed use development at particular locations.

In our first results, the land use intensity is illustrated using a bubble plot as shown in Fig. 4. The intuition behind this is to help visualize the land use mix in the core regions of a city, where land use intensity, either residential, activity, or both, is high. Other methods fail to highlight this important aspect, which is crucial in distinguishing important sub-regions relative to others.

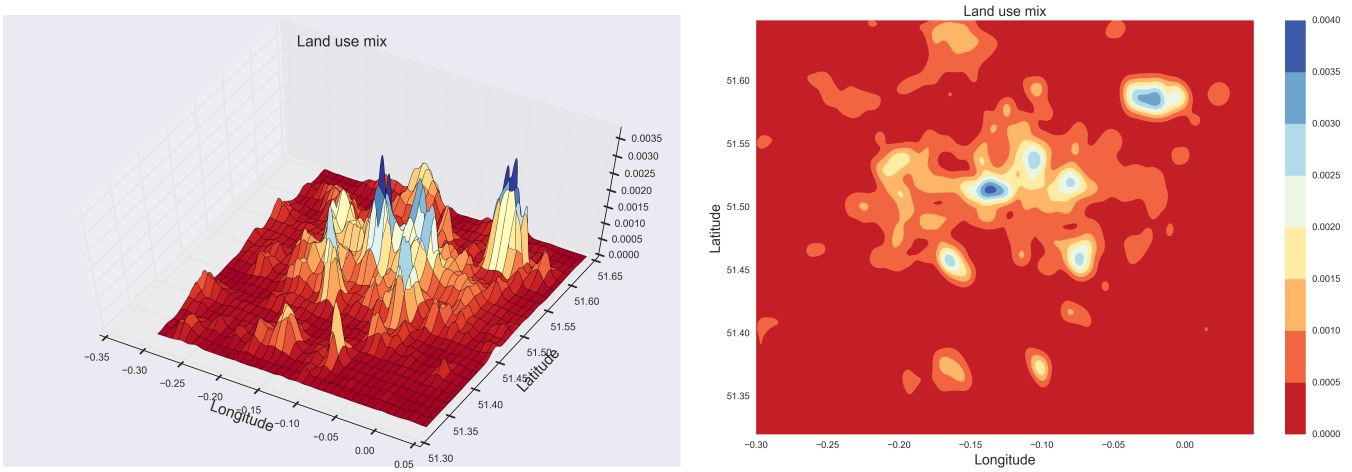
In this work we chose to compute the Entropy Index (EI), as done in [39], to evaluate the degree of land use mix for every point along the two spatial dimensions of the KDE's. Still, our method allows to use any kind of index computation, given the local context of analysis built from OSM data.

Let P_j be the percentage of each land use type, obtained from the respective KDE's. The amount of land uses k employed equals the number of KDE's computed (here, two, although with increasing completeness of OSM data, this may increase):

$$EI = \frac{-\sum_{j=1}^{k} P_j \; ln(P_j)}{ln(k)} \tag{2}$$

The Entropy Index output will be between 0 and 1, where the higher the value, the higher the mixture of land uses.

Fig. 3 presents the computed land use mix for London, England, while Fig. 4 shows a bubble plot which better conveys local land use intensities. A further analysis of these results is presented in Sect. VI.



(a) (b)

Fig. 3. Land use mixture for London.

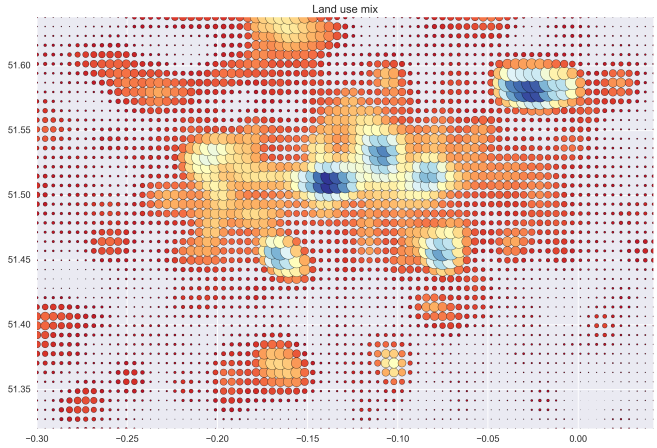


Fig. 4. Bubble plot of LUM for London. Each point's size denotes the local land use intensity.

D. Implementation

The framework's infrastructure is built over a Python stack and is composed of the following modules:

- `extract_uses.py` given the points and polygons shapefiles from the Mapzen Metro Extract, the extraction of their uses is done according to Section V-A. The geometry of the polygons is processed with the Shapely library, and the final output is a Pandas [40] DataFrame with the geo-referenced categorized POIs `pois_df`.
- `kde.py` given `pois_df`, it determines the KDE for both activity `kde_act` and residential `kde_res` types using the Statsmodels [41] library.
- `loaders.py` manages a local Hadoop Distributed File Store (HDFS) through the [42] library. Such an HDFS stores for each studied *city* the results of the expensive computations of `extract_uses.py` and `kde.py`. When for a given *city* such information is not stored in the local HDFS, it proceeds as follows:
  1) A query for the corresponding *city*'s shapefiles is issued to Mapzen
  2) When the shapefiles are received, it calls `extract_uses.py` and `kde.py` in order to compute `pois_df`, `kde_act` and `kde_res`
  3) The results of the computation are stored in the local HDFS under the corresponding *city* key.
- `measures.py` determines the implemented indicators out of the `pois_df`, `kde_act` and `kde_res` (i.e. the entropy of Equation (2)) using Numpy arrays [43].
- `plots.py` generates the desired plots from `pois_df`, `kde_act` and `kde_res` (i.e. longitude-latitude scatter-plots, KDE plots...) using the Matplotlib library [44].
- `city_analysis.py` is the Python class to interact with the Jupyter Server. To start an analysis for a given *city*, it will ask for the *city*'s information to `loaders.py`. Once such information is available, the class manages the calls to the methods of `measures.py` and `plots.py` in order to get the outputs interactively.

The implementation is represented in Figure 5. The source code of the framework is publicly available at [4]. A web-interface is also planned to be released to allow non computer-savvy, urban planners among others, to evaluate the framework online.

VI. Application

A. Grenoble, France

The city of Grenoble, France, is an ancient medium-size European city. It is located at the foot of the French Alps, and possesses a distinctive Y shape because of the surrounding mountain ranges of the Chartreuse, Belledonne and Vercors. The agglomeration of this city, containing a population size of $665,000$, was processed and its LUM was computed. A

[4] https://github.com/martibosch/landusemix

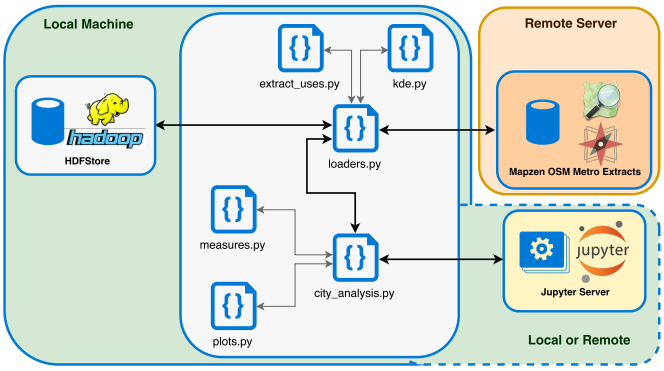


Fig. 5. Implementation of the framework

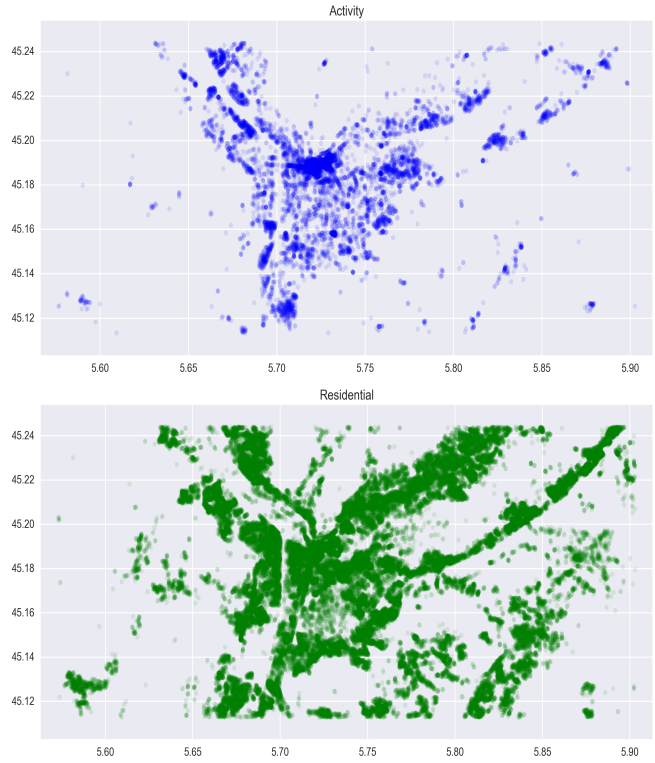


Fig. 6. Extracted points for Grenoble, France.

total number of $33,055$ points and $118,207$ polygons were processed in 10 minutes. As a result, $11,258$ activities and $84,445$ residential points were extracted. Both residential and activity KDE's were computed in 32 minutes, and the final land use mix was rapidly computed in 2 seconds.

The different outputs are depicted in Fig. 7. In the activity's KDE, the high concentration of activities lying around the historical center can be observed. This pattern can be frequently seen in ancient European cities, where the historical centers contain the biggest concentration of activities due to a long historical process.

Regarding residential uses, it is noticeable that the highest concentration can also be found around the historical center, in agreement with the mentioned historical process. For this type of cities, urban development is the result of a long process dating back to long before the invention of the car. Consequently, one can find a good land use mix around the cities' historical center, as depicted visually in the different land use mix plots.

The results also exhibit clearly a particular observable structure: as the outskirts of the city are mainly composed of half-mountainous landscape, several residential collections can be found in areas remote enough from the existing activity centers, providing a good quality of life and access to nature, but still easily accessible by car or public transportation. The structures at $5.58E/45.13N$ and $5.84E/45.14N$ match respectively with the East side of the plateau of the Vercors and the valley of Vaulnaveys. Both are notoriously beautiful residential areas with few activities apart from arts and crafts shops and farms.

B. *London, England*

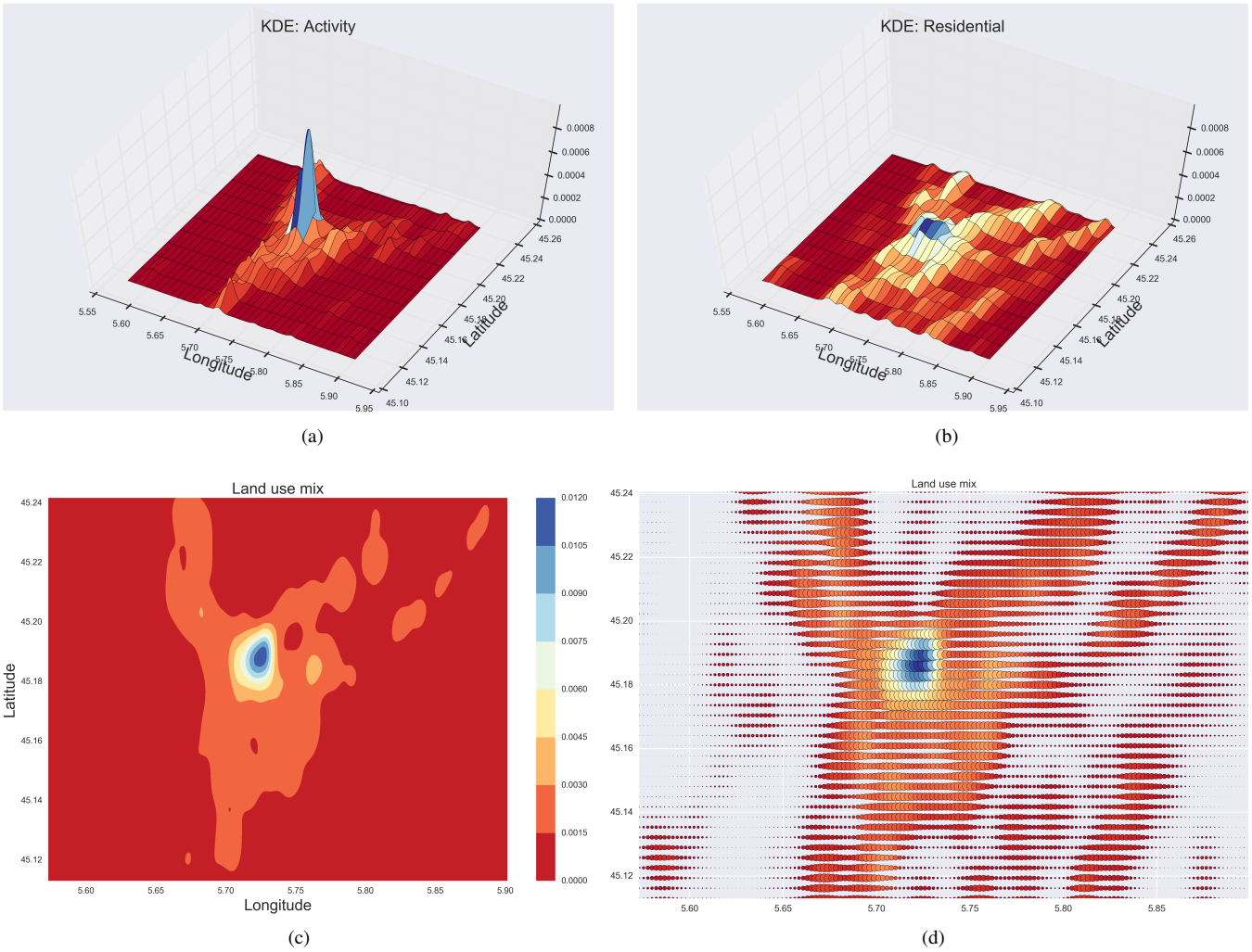
Another evaluation was performed on the city of London. As a result, a total number of $80,702$ and $238,626$ points were extracted for activity and residential uses respectively. The results of this procedure are depicted above in Fig. 1. The KDEs were then computed, as shown in Fig. 2. The LUM is shown in Fig. 3 and Fig. 4.

London city went across big changes across history, a process which molded the structure in which the current 8.674 million people live. In Fig. 2, the City of London and Soho are depicted as the areas containing the highest density of activities. Regarding residential land use, East Dulwich, the Higham Hill, and the area lying between the Wandsworth Common and the Clapham Common were found to be the most intense.

The Croydon Vision 2020 is an urban planning program seeking to develop Croydon as a hub of living, retailing, culture and business in the context of a local development framework. An important concentration of activities exists in this area, while an important residential area lies nearby, to its west. Our framework can detect that a relatively good mix of uses already exists for this region, as depicted in Fig.3 at the longitude and latitude values of -0.100594 and 51.376495 respectively. Further, mixed-used development is acknowledged in this urban program. Thus, its mixed use development is expected to further improve in the near future with a higher integration of different land uses.

VII. Conclusions and Future Work

In this paper, we presented a framework for capturing mixed uses development in cities. It uses crowd-sourcing data from OpenStreetMap to extract the geo-localized land uses. Due to the universality of this data source, we are able to process any geographical area in the world, as long as sufficient data are available in OSM. A Kernel Density Estimation is performed for each of the land uses, outputting the spatial distribution of its land uses. Based on this representation, a measure of land use mix is then calculated using the Entropy Index. The GIS output that results, shows enriched information for



(a) (b) (c) (d)

Fig. 7. The framework's results for Grenoble, France. (a), (b) stand for Activity and Residential KDE's, while (c) and (d) stand for LUM.

urban planners, supporting and aiding the decision-making procedure.

The framework was applied on the cities of London and Grenoble. Both activities and residential densities estimation were validated by means of inspecting local neighborhoods with a high concentration of the different land uses. Afterwards, their spatial LUM distribution was computed, allowing for an easy visualization concentrating on the mixed uses values, and showing the LUM relative importance in terms of land use intensity.

Future work includes integrating the LUM output for measuring the urban sprawl phenomenon and performing numerical interpretations of desirable mixed use values. We will also study the potential integration to transportation models, where land use mix correlation with the activities and residential uses can help improve demand estimation.

In addition, further investigation can be done by means of analyzing the different types of activities, which in this work are classified only as activities. Rich information provided from OSM data allows to divide the extracted activities into differing classifications such as shop, leisure, amenities, commercial and industrial among others. Finally, the estimation of LUM can be refined by taking into account, besides their location, the accessibility between different land uses, which is partly conditioned by the transportation infrastructure.

References

[1] "World urbanization prospects: The 2014 revision, highlights," United Nations, Population Division, Department of Economic and Social Affairs, Tech. Rep., 2014.

[2] P. Gordon and H. W. Richardson, "Are compact cities a desirable planning goal?" *Journal of the American Planning Association*, vol. 63, no. 1, pp. 95–106, 1997.

[3] M. Breheny, "The compact city and transport energy consumption," *Transactions of the Institute of British Geographers*, pp. 81–101, 1995.

[4] R. Burgess, "The compact city debate: A global perspective," in *Compact cities: Sustainable urban forms for developing countries*. Spon Press: London, UK, 2000, pp. 9–24.

[5] M. Neuman, "The compact city fallacy," *Journal of Planning Education and Research*, vol. 25, no. 1, pp. 11–26, 2005.

[6] Y. Song, L. Merlin, and D. Rodriguez, "Comparing measures of urban land use mix," *Computers, Environment and Urban Systems*, vol. 42, pp. 1–13, 2013.

[7] M. Leccese and K. McCormick, *Charter of the new urbanism*. McGraw-Hill Professional, 2000.

[8] "This is smart growth," Smart Growth Network, Tech. Rep., 2006. [Online]. Available: www.smartgrowthamerica.org/documents/this_is_smart_growth.pdf

[9] R. Cervero, C. Ferrell, and S. Murphy, "Transit-oriented development and joint development in the United States: A literature review," *TCRP Research Results Digest*, no. 52, 2002.

[10] G. D. Squires, *Urban sprawl: Causes, consequences, & policy responses*. The Urban Insitute, 2002.

[11] M. P. Johnson, "Environmental impacts of urban sprawl: a survey of the literature and proposed research agenda," *Environment and Planning A*, vol. 33, no. 4, pp. 717–735, 2001.

[12] A. Nelson and J. Duncan, *Growth management principles and practices*. Planners Press, American Planning Association, 1995. [Online]. Available: https://books.google.fr/books?id=cmpPAAAAMAAJ

[13] R. H. Ewing, "Characteristics, causes, and effects of sprawl: A literature review," in *Urban Ecology*. Springer, 1995, pp. 519–535.

[14] "Sprawl: The dark side of the American dream," Sierra Club, Tech. Rep., 1998. [Online]. Available: http://vault.sierraclub.org/sprawl/report98/report.asp

[15] "State of the cities – 1999," United States Department of Housing and Urban Development, Tech. Rep., 1999. [Online]. Available: http://eric.ed.gov/?id=ED438358

[16] H. E. Christian, F. C. Bull, N. J. Middleton, M. W. Knuiman, M. L. Divitini, P. Hooper, A. Amarasinghe, and B. Giles-Corti, "How important is the land use mix measure in understanding walking behaviour? Results from the RESIDE study," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 8, no. 1, p. 1, 2011.

[17] L. D. Frank, M. A. Andresen, and T. L. Schmid, "Obesity relationships with community design, physical activity, and time spent in cars," *American Journal of Preventive Medicine*, vol. 27, no. 2, pp. 87–96, 2004.

[18] R. Cervero, "Mixed land-uses and commuting: evidence from the american housing survey," *Transportation Research Part A: Policy and Practice*, vol. 30, no. 5, pp. 361–377, 1996.

[19] K. Kockelman, "Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1607, pp. 116–125, 1997.

[20] P. Newman and J. Kenworthy, "Urban design to reduce automobile dependence," *Opolis*, vol. 2, no. 1, 2006.

[21] G. L. Raines, "Description and comparison of geologic maps with FRAGSTATS – a spatial statistics program," *Computers & Geosciences*, vol. 28, no. 2, pp. 169–177, 2002.

[22] H. A. Carlos, X. Shi, J. Sargent, S. Tanski, and E. M. Berke, "Density estimation and adaptive bandwidths: a primer for public health practitioners," *International Journal of Health Geographics*, vol. 9, no. 1, p. 1, 2010.

[23] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[24] M. F. Goodchild, "Citizens as sensors: web 2.0 and the volunteering of geographic information," *GeoFocus*, vol. 7, pp. 8–10, 2007.

[25] M. Forghani and M. R. Delavar, "A quality study of the OpenStreetMap dataset for Tehran," *ISPRS International Journal of Geo-Information*, vol. 3, no. 2, pp. 750–763, 2014.

[26] C. Barron, P. Neis, and A. Zipf, "A comprehensive framework for intrinsic OpenStreetMap quality analysis," *Transactions in GIS*, vol. 18, no. 6, pp. 877–895, 2014.

[27] P. Mooney, P. Corcoran, and A. C. Winstanley, "Towards quality metrics for OpenStreetMap," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2010, pp. 514–517.

[28] H. Fan, A. Zipf, Q. Fu, and P. Neis, "Quality assessment for building footprints data on OpenStreetMap," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 700–719, 2014.

[29] M. Haklay, "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets," *Environment and Planning B: Planning and Design*, vol. 37, no. 4, pp. 682–703, 2010.

[30] J.-F. Girres and G. Touya, "Quality assessment of the French OpenStreetMap dataset," *Transactions in GIS*, vol. 14, no. 4, pp. 435–459, 2010.

[31] P. Neis, D. Zielstra, and A. Zipf, "The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011," *Future Internet*, vol. 4, no. 1, pp. 1–21, 2011.

[32] M. Over, A. Schilling, S. Neubauer, and A. Zipf, "Generating web-based 3D city models from OpenStreetMap: The current situation in Germany," *Computers, Environment and Urban Systems*, vol. 34, no. 6, pp. 496–507, 2010.

[33] X. Liu and Y. Long, "Automated identification and characterization of parcels with OpenStreetMap and points of interest," *Environment and Planning B: Planning and Design*, vol. 43, no. 2, pp. 341–360, 2016.

[34] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.

[35] F. Atash, "Redesigning suburbia for walking and transit: Emerging concepts," *Journal of Urban Planning and Development*, vol. 120, no. 1, pp. 48–57, 1994.

[36] L. Aultman-Hall, M. Roorda, and B. W. Baetz, "Using GIS for evaluation of neighborhood pedestrian accessibility," *Journal of Urban Planning and Development*, vol. 123, no. 1, pp. 10–17, 1997.

[37] "Liveable neighbourhoods: a Western Australian Government sustainable cities initiative," Western Australian Planning Commission, Department for Planning and Infrastructure, Tech. Rep., 2007. [Online]. Available: http://www.planning.wa.gov.au/dop_pub_pdf/LN_Text_update_02.pdf

[38] G. R. McCormack, B. Giles-Corti, and M. Bulsara, "The relationship between destination proximity, destination mix and physical activity behaviors," *Preventive Medicine*, vol. 46, no. 1, pp. 33–40, 2008.

[39] Y. Song and G.-J. Knaap, "Measuring the effects of mixed land uses on housing values," *Regional Science and Urban Economics*, vol. 34, no. 6, pp. 663–680, 2004.

[40] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, 2010, pp. 51–56.

[41] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 57–61.

[42] F. Alted and M. Fernández-Alonso, "Pytables: processing and analyzing extremely large amounts of data in python," *PyCon2003. April*, pp. 1–9, 2003.

[43] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[44] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in science and engineering*, vol. 9, no. 3, pp. 90–95, 2007.