A Framework for Measuring Urban Sprawl from Crowd-Sourced Data

Martí Bosch

▶ **To cite this version:**

Martí Bosch. A Framework for Measuring Urban Sprawl from Crowd-Sourced Data. Modeling and Simulation. 2016. hal-01401376

HAL Id: hal-01401376
https://inria.hal.science/hal-01401376

Submitted on 23 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Master Of Science in Informatics at Grenoble

A Framework for Measuring Urban Sprawl from Crowd-Sourced Data

Author:
Martí Bosch

Supervisors:
Serge Fenet
Peter Sturm

June 22, 2016

Contents

1 Introduction

The amount of people living in cities by 1800 was roughly around 3 percent of the world population. This number has increased dramatically during the last centuries, and currently it is estimated that one out of two people lives in cities. Furthermore, according to United Nations[1] 60 percent of the world population will live in cities by 2030.

This situation brings new challanges on how to conceive cities that host such amounts of population in a *sustainable* way while sacrificing as little as possible the inhabitants' *quality of life*. This *sustainability* should address to several aspects that can be classified as *economical*, *social* and *environmental*. The cities are and will be centers of *economical* activity, and thus should provide facilities for business, innovation and culture. Such *economical* development should benefit all the levels of the *social* hierarchy, preventing inequalities and social segregation. The *environmental* part concerns the efficient utilization of the resources as well as the minimization of the impact on the ecosystems that sourround such cities. A convenient public transportation network, the preservation of green areas, the recycling of waste or the use of renewable energies are some examples of means to reduce the cities' *environmental* impact.

Unluckily by taking a look at the current *megacities*, it can be easily observed that very few of them meet the *sustainability* characteristics formerly reviewed. This can be partly explained by the *urbanism pattern* that derives from the processes of *industrialization*. When a city experiences such *industrialization*, with the related *economic growth* and the *expansion of transportation networks*, the *middle class* tends to migrate towards the outskirts of the city, potentially to live in terraced houses sorrounded by green areas. Such a pattern is commonly referred to as *urban sprawl*, and it was first observed in London and Paris during the 19^{th} century. Cities like New York, Chicago went through this process during the early 20^{th} century, and so did most central and northern-European cities around 1970-1990. Nowadays *urban sprawl* might even be more prevalent in *developing countries*, as it is the case with Mexico City, Beijing, Delhi, Johannesburg or Cairo.

Contributions and Outline This work reviews first the existing literature on *urban sprawl* in Section 2, comparing the term's various definitions. Additionally, a set of measures proposed to gauge the phenomena are audited and brought to a common notation for better comprehension and comparability. Then, in Section 3 a selection of this measures is performed according to certain suitability criterias, and a framework able to collect crowd-sourced data and compute the elected measures is presented. Some analysis with the currently implemented measures over data corresponding to real cities is performed in Section 4.1. Some perspectives for future work are described in Section 5

[1] Lewis, Mark (2007-06-11). "Megacities Of The Future". Forbes. Retrieved 2011-11-30

2 Literature Review

The first reference to the term *urban sprawl* was made by Earle Draper, as part of a conference of *urban planners* of the southeastern United States in 1937 (Wassmer 2002; Nechyba and Walsh 2004). Ever since then the use of the term has been spreading to a wide range of domains, nevertheless a common consensual definition has not been adopted. This led to the current situation where *urban sprawl* is an ambiguous term that might be used in several domains, such as urbanism, geography, economics or more recently also remote sensing and data science.

2.1 Urban Sprawl: Timeline of a Loose Term

This section intends to chronologically review the literature on *urban sprawl* focusing on the descriptions of the term that are proposed, and separating among them what can be considered *causes*, *characteristics* and *consequences*.

Early Works: Characteristics and Speculation The first studies on the domain focus mostly on a series of *characteristics* that devise the term of *urban sprawl*: (Whyte 1958) considers it "scattered leapfrog development". A more exhaustive characterization is found in (Harvey and Clark 1965), where it is stated that *urban sprawl* is located at the "urban fringe", scattering around undeveloped or agricultural land, and "occurs in three major forms": (1) "low density continuous development", (2) "ribbon development" and (3) "leapfrog development".

On the other hand, (Clawson 1962; Bahl 1968; Archer 1973) identify *land speculation* as a *cause*, while also alluding to *scattered* and *leapfrog* development.

Diversely, (McKee and G. H. Smith 1972) discerns some other *causes*: the "love affair between people and metropolis is over." and that the "ideal place to live is now the suburbs", as well as "poor planning" or "haphazard expansion". It also points out a few *characteristics* as in "very low density development over a large area", "ribbon extending axially along the access routes of major urban areas", leapfrog development, and an excessive "consumption of land resources". Furthermore, (McKee and G. H. Smith 1972) contextualizes *sprawl* by associating it to "single family homes" in developed countries and to "squatter settlements" in less developed ones. The study also mentions as a *consequence* that "metropolitan populations do not grow as fast as the urban areas".

The 90s: Land Use Mix, Automobile Dependency, and Social Segregation In the decade of the 90s, most of the works kept hinting at the *characteristics* of *urban sprawl* reviewed formerly, notwithstanding the articles start to reveal a new magnitude: the *land use mix*. Lack of planning and coordination is pointed out as a *cause* in (Nelson and Duncan 1995) resulting in this fresh *characteristic* of dysfunctional mix of uses. An emphasis on the *land use mix* is also clear in (HUD, 1999).

The literature reivew of (R. H. Ewing 1995) lists *low density*, *strip*, *scattered* and *leapfrog* development as the classic patterns of *urban sprawl*, and associates to it the *poor accessibility* and lack of *open space*. It points to several *causes*: "subsidies for suburban development", "externalities" and "government regulation". As *consequences*, the study refers to "environmental deprivation", *automobile dependency* and traffic congestion, the *excessive costs* of providing public services as well as *environmental* effects such as air pollution and loss of farmland.

The *automobile dependency* is also considered a *consequence* of *urban sprawl* in other works as well: (Sierra Club, 1998) states that it "separates where people live from where they shop, work, recreate and educate - thus requiring cars to move between zones", while (PTCEC, 1998) refers to "automobile-dependent development pattern of hosuing, shopping centers and business parks".

Furthermore (Burchell et al. 1998) mentions land use types "segregated from one another" and also adduces *social segregation* as a side-effect of such a partition. Coetaneously, (Downs 1999) gathers 10 recurring traits of *urban sprawl* in the literature, which in addition to the lack of *land use mix* and the *automobile dependency* also feature *political* aspects like "fragmentation of powers over land use among many small localities", "no centralized planning or control of land-uses" and *socio-economic* facets such as "fiscal disparities among localities" or "reliance mainly on the trickledown or filtering process to provide housing to low-income households".

Focusing on the *land use* issue, (Rolf Pendall 1999) explores the influence of local *land use* controls, and its results indicate that adequate provisioning of public facilities and expensive housing (among other factors) discourage *urban sprawl* whereas jurisdictional fragmentation increases it, thus making the lack of coordinated *land use* planning a *cause* of the process.

Three more works from this decade must be remarked: first, (Nelson and Duncan 1995) provides one of the most broad definitions of *urban sprawl*, which in accordance to formerly mentioned studies, describes *urban sprawl* as "Unplanned, uncontrolled, and uncoordinated single-use development that does not provide for an attractive and functional mix of uses and/or is not functionally related to surrounding land uses and which variously appears as low density, ribbon or strip, scattered, leapfrog, or isolated development". Secondly, one of the most cited articles on the topic is (Reid Ewing 1997), which delineates the *characteristics* of *urban sprawl* in three forms: (1) "leapfrog or scattered" (2) "commercial strip" and (3) large "low-density or single use" and conversely "low accessibility" and "lack of functional open space". At last, (Gordon and Richardson 1997) examines the costs and benefits of compactness versus sprawl and concludes, in contradiction to most of the related literature, that compactness is not the most preferable and feasible planning goal.

The 21^{st} Century: Environment, Health and GIS Data Science The increase of the ease to access data through the internet together with significant advances on *data science* methods as well as computing capacity brought the

field to a data-driven era. An important part of the research focuses on defining quantitative techniques to measure *urban sprawl* through the processing Geographical Information Systems (GIS) data, such as images obtained from *remote sensing*. Another main research axis aims its attention to gathering data from a loose set of categories (socio-economics, environment, health, transportation...) and correlating it to *urban sprawl* indicators.

An extensive survey on the literature on the environmental impacts of *urban sprawl* is found in (Johnson 2001), where the author associates to the phenomenon a large set of ecological damages, such as *air pollution*, high *energy consumption*, *ecosystem fragmentation* with an excessive loss of *environmentally fragile lands*, *open space*, *farmland species diversity* as well as an increase of the natural risks. The topic is also an issue in environment conservation: (Beach 2003) considers it, together with population growth, as one of the main threats to the United States' coast, (Radeloff, Hammer, and Stewart 2005) assesses how metropolitan (*fringe*) and rural *sprawl* affect the surrounding forests, and (Blair 2004) evaluates that the occurrences of native bird communities is affected by the distribution and intensity of urban patches.

After ranking metropolitan areas of the United States by its measure of *urban sprawl*, (R Ewing, R Pendall, and D Chen 2002) suggests possible *consequences* by pointing out that the habitants of the most sprawled areas have higher *automobile dependency*, traffic fatalities, air *pollution*, as well as (after the ranking's posterior update (R. H. Ewing, Hamidi, and America 2014)) worse health conditions. On the other hand, (Song and Knaap 2004) reviews the growth management policies adopted at Portland's metropolitan area (Oregon) and its effects on several indicators of *urban sprawl*. Very similarly, (Arbury n.d.) presents an analogous study in the case of Auckland, New Zealand. Moreover, (Ludlow 2006) remarks that sprawl represents an issue in Europe too, and (Catalán, Saurí, and Serra 2008) depicts a detailed portrait of growth patterns in archetypal Mediterranean polycentric metropolitan regions (based in a case study of the Barcelona metropolitan region).

Several reports, such as those of (Frumkin 2002; McCann and Reid Ewing 2003; Sturm and Cohen 2004; Lopez 2004; Frumkin, Frank, and Jackson 2004), and remarkably (Reid Ewing, Schmid, et al. 2008) point to obesity and bad physical conditions as *consequences* of *urban sprawl*. Nevertheless, (Eid et al. 2008) reckons those works as politically biased and disputes that *cause-consequence* effect, indicating that the *correlations* are explained by the fact that obese people prefers to live in sprawled areas.

Previous definitions of *urban sprawl* are categorized in (Galster et al. 2001) as in: (1) by example, i.e. Los Angles, (2) by aesthetics, (3) cause of an externality, as for example car dependency, (4) consequence of an independent variable, such as poor planning, (5) patterns of development, (6) process of development over time. Most works are built on such previous characterizations, with the noteworthy exception of (Jaeger, Bertiller, Schwick, and Kienast 2010), which proposes a definition in which landscapes "suffer from urban sprawl" when "permeated by urban development", and the "degree of urban sprawl" is proportional to the built-up area, dispersion and (after posterior definition update

in (Jaeger and Schwick 2014)) land uptake per inhabitant".

In addition to the formerly mentioned research, a new line of research unveils strongly: analysis of GIS data, often obtained through remote sensing. The availability of time series of satellite images makes these methods extremely powerful to monitor patterns of urban growth and land use changes, in the USA: (Masek, Lindsay, and Goward 2000; Yang et al. 2003; Wilson et al. 2003; Sutton 2003; Hasse and Lathrop 2003; C. Wu 2004; Xian and Crane 2005; Yuan et al. 2005; Ji et al. 2006) ; fast-growing regions in Asia: (Yeh and Xia 2001; Sudhira, Ramachandra, and Jagadish 2004; Li and Yeh 2004; Xiao et al. 2006; Yu and Ng 2007; Jat, Garg, and Khare 2008; B Bhatta 2009) ; or presenting generalizations to better assess *urban sprawl* morphology and dynamics globally: (Nagendra, Munroe, and Southworth 2004; Huang, Lu, and Sellers 2007; Angel, Parent, and Civco 2007; Ba Bhatta, Saraswati, and Bandyopadhyay 2010).

2.2 A Dynamic Process

In the previous section, a large set of studies were reviewed. Most of them focus on providing metrics to gauge the *causes* and *consequences* of *urban sprawl*, as well as the *characteristics at a given snapshot of time*. On the other hand, the lastly revised works on GIS data do already consider the *time dynamics* of the phenomena.

It must then be remarked that, as pointed out by (Torrens and Alberti 2000), *urban sprawl* is a *dynamic* process "at the forefront of dynamic urban growth". Nevertheless, many of the *urban sprawl* metrics "are themselves static", and to "examine sprawl in a truly dynamic fashion it may be necessary to employ a simulation model. These metrics could still be used to calibrate the model".

These simulations constitute themselves another line of research, where cities are modelled as *self-organizing* systems, as extensively described in (Portugali, Benenson, and Omer 1997; Portugali 2000). To appraise the dynamics of urban systems, a series of land use/cover change (LUCC) models have been proposed, often based on cellullar automata (CA) and multi-agent systems (MAS). Examples of works include (White and Engelen 1993; O'Sullivan and Torrens 2001; Batty 2007) the review of (Parker et al. 2003), and the integrated land use and transport models (LUTI) of (De La Barra 1989) and (Waddell 2002).

2.3 Dimensions of Urban Sprawl

The diversity of contexts in which *urban sprawl* is mentioned outlines a complexity that is embedded with the term. Consequently, it must be considered as a *multidimensional* phenomenon and in order to appraise it, a dimensional breakdown is a mandatory preliminary stage.

2.3.1 Decompositions from the Literature

Most of the dimensional decompositions of *urban sprawl* were drawn in the 21^{st} century, arguably because by then there was already a large volume of literature

on the subject, and the data and means for computing measures started to gain availability to researchers.

The work of (Galster et al. 2001) presents one of the most exhaustive and cited dissections of *urban sprawl*, distinguishing 8 dimensions: (1) *density* (of residential units), (2) *continuity*, (3) *concentration*, (4) *clustering*, (5) *centrality*, (6) *nuclearity*, (7) *mixed uses* (*land use mix*) and (8) *proximity*.

The previously reviewed works of (R Ewing, R Pendall, and D Chen 2002) and (R. H. Ewing, Hamidi, and America 2014) calculate aggregated indices of *urban sprawl* out of four dimensions: (1) *density*, (2) *land use mix*, (3) *centering* and (4) *accessibility*.

Focusing on the measurement of urban forms, (Tsai 2005) distinguishes two main types of *urban sprawl*, which are intensity-based and spatial pattern-based. A four-dimensional representation is proposed: (1) population *size*, (2) population *density* (intensity-based), (3) *evenness* of distribution (spatial pattern-based) and (4) *clustering* (spatial pattern-based).

In (Jaeger, Bertiller, Schwick, Cavens, et al. 2010) and its posterior amend (Jaeger and Schwick 2014), an indicator is proposed after a decay of *urban sprawl* into (1) fraction of developed area, (2) its *dispersion* and (3) utilization *density*.

Another intensive analysis of the *multidimensionality* of the term is found in (Arribas-Bel, Nijkamp, and Scholten 2011), where the authors present a hierarchical characterization: in the category of *urban morphology* there are the dimensions of (1) *scattering*, (2) *connectivity* and (3) *availability of open space*; and categorized by *internal composition* there is (4) *density*, (5) *decentralization* and (6) *land use mix*.

2.3.2 Meta Decomposition

Summarizing the formerly audited decompositions and most frequent terms used when referring to *urban sprawl* in Section 2.1, the phenomenon's *dimensions* can be listed as:

1. **Density** as the *low density development* term suggests. It is arguably the most mentioned characteristic of *urban sprawl*.

2. **Development's Distribution** deduced from the terms of *open space*, *leapfrog*, *scattered* or *segregated* development. The decompositions from Section 2.3.1 atomize the *dispersion* in different dimensions, such as *continuity*, *concentration*, *clustering*, *centrality*, and *nuclearity*. Nevertheless, they all respond to the same magnitude: how the areas with high concentration of activities are distributed. Furthermore, *clustering* already englobes *centrality* and *nuclearity* in a more general way: it gauges how activities tend to clump together, regardless if it is around one center (monocentric) or several ones (polycentric).

3. **Land use mix** (or absence of thereof) is often considered as a consequence of the lack of *coordinated planning*. In pursuance of a clear dimensional

decomposition, this characteristic can be seen as, how the land uses are distributed for a *given urban development*.

4. **Accessibility** after the associations of *urban sprawl* with *automobile dependency*. This is often attributed to *fringe* and *ribbon* development around suburban highways, however such characteristic is rather related to the *dispersion* (the former dimension) of the developed patches. Consequently, in order to separate these two dimensions, *accessibility* will rather denote how well the transportation network efficiently serves a *given urban development*.

Such a decomposition is practically the same as the four-factor one proposed by (R Ewing, R Pendall, and D Chen 2002) and (R. H. Ewing, Hamidi, and America 2014) (*activity centering* is a particular acception of *development distribution*). It is also very similar to the characterizations of (Tsai 2005) and (Jaeger and Schwick 2014), nevertheless these do not consider the *land use mix*, which several studies arguably consider as a key factor of *urban sprawl*.

2.4 Indicators in the Literature

Given the looseness of the term and its multi-dimensionality, a large set of indicators have been proposed to measure different characteristics of *urban sprawl*. This part will put together and audit the most common ones, according to the four dimensions highlighted in Section 2.3.2.

2.4.1 Density

Density has a very simple and intuitive definition. In spatial analysis, for a given magnitude f, its *density* will correspond to its value divided by the area a that f involves. In the context of *urban sprawl*, such a magnitude might correspond to the number of housing units, the number of jobs or the number of residents.

Works like (Galster et al. 2001; Malpezzi and Guo n.d.; R Ewing, R Pendall, and D Chen 2002; Reid Ewing, Rolf Pendall, and Don Chen 2003) and (R. H. Ewing, Hamidi, and America 2014) borrow from several statistics of the tract densities, such as percentile-based tract density or the extreme values. However, as (Tsai 2005) points out, such variables have been empirically shown to be highly correlated to the average density itself.

Density Gradients gauge how the density decays with the distance to the city centers. As (Torrens and Alberti 2000) remarks, this is a key factor in economics and *urban sprawl* since it influences the development of residential units at the urban fringe based on the housing prices and the commuting costs. The gradients are often characterized by the *inverse power function*, as in:

$$D(x) = D_0 x^{-\alpha} \tag{1}$$

where D_0 is the density at the center and α is the decay parameter as moving away from it (increasing x).

Very similarly, the same magnitude can be modelled with a *negative exponential function* such as:

$$D(x) = D_0 e^{-\lambda x} \tag{2}$$

λ being then the decay parameter.

According to (Torrens and Alberti 2000), the inverse power function of Equation 1 has "a tendency to over-predict" the density of "areas close to the CBD", whereas the negative exponential model of Equation 2 usually "does a poor job of predicting central densities" (Batty and Kim 1992).

2.4.2 Distribution

The distribution of urban development distinguishes two different facets: the *evenness* and the *clustering* (or centrality). The *evenness* determines how equitably a magnitude is distributed among the studied area. Nonetheless, this does not reveal any spatial relation among the sub-areas with high density of f, such as whether they are clustered or randomly distributed. The latter in fact constitutes the other facet, the *clustering* of high density sub-areas.

Most of the *distribution* measures are determined after an areal decomposition (i.e. grid or census tracts) of a given region. As (D. Smith 1975) exposes, their values often depend on such areal decomposition features as the shapes and sizes of sub-areas. This issue is described in detail in (De La Barra 1989), with the denomination of the *modifiable areal unit problem* (MAUP). Furthermore, (Openshaw 1991) states that this problem might induce unspecified influence on the results of the spatial analysis.

Notation The areal decomposition will be noted by Ω, which will be a set of $N = |\Omega|$ sub-areas. For any given magnitude β, its value in the i sub-area will be noted as β_i. The magnitude's capitalization B will represent its total value in all N sub-areas $B = \sum_\Omega \beta_i = \sum_{i=1}^{N} \beta_i$ whereas its mean will be noted as $\bar{\beta} = \frac{B}{N}$.

The letter a will be reserved to denote the area, whereas f will be used generically and can correspond to any occurence, such as points of interest (POIs). Most of the measures in the literature are proposed based on a particular magnitude (such as f = number of housing units) or a combination of magnitudes (such as f = number of inhabitants + number of jobs). Despite this, the equations in this section will be expressed in terms of a generic magnitude f, to allow more versatility in their semantics.

Measures of Evenness

Delta Index is an adaptation of the index of dissimilarity of (Massey and Denton 1988), initially formulated to reflect levels of residential segregation. In (Galster et al. 2001), it is tailored as:

$$\delta = \frac{1}{2}\sum_{i=1}^{N} |\frac{f_i}{F} - \frac{a_i}{A}| \tag{3}$$

which computes the proportion of f that is located in sub-areas a_i with above average density of f. Higher values of δ indicate that f is more concentrated in fewer sub-areas.

Gini Coefficient is the most commonly used measure of *economic inequality*, and can be adapted to determine the *equality distribution* among sub-areas as in:

$$G = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{|f_i - f_j|}{f_i} \tag{4}$$

ranging from 0 to 1, 0 corresponding to perfect equalty and 1 to one sub-area i concentrating all F. This adaptation is acknowledged as an indicator of *urban sprawl* in (Malpezzi and Guo n.d.) and (Tsai 2005).

Shannon's Entropy was first proposed in (Shannon 1948) in the context of information theory. For spatial analysis, it can be expressed as:

$$H = -\sum_{i=1}^{N} \frac{f_i}{F} ln(\frac{f_i}{F}) \tag{5}$$

whose value starts at 0, indicating a compact distribution of f, and approaches $ln(N)$ for very dispersed ones. For cities, a halfway mark $\frac{ln(N)}{2}$ is considered by some authors a sprawling threshold. Often H is rescaled into a 0 to 1 range, as in $H' = \frac{H}{ln(N)}$ and denominated *relative entropy*.

The first use in spatial analysis is found in (Theil 1967), and according to (Thomas 1981) it is better than other dispersal statistics since it is invariant to the areal decomposition and thus not affected by the MAUP. This claim is reaffirmed by (Yeh and Xia 2001; Ba Bhatta, Saraswati, and Bandyopadhyay 2010) when adopting it to measure *spatial dispersion*. It is also used in this sense in (Sudhira, Ramachandra, and Jagadish 2004; Li and Yeh 2004) and (Jat, Garg, and Khare 2008), among other works.

On the other hand, (Tsai 2005) discards choosing entropy to gauge dispersion since it cannot be determined for sub-areas of density 0 (a_i such that $f_i = 0$), and notes that those sub-areas do exist in metropolitan areas (i.e. parks).

Theil's Index (Theil 1967) defines one of the main measures of economic inequality, determined as:

$$T = \sum_{i=1}^{N} \frac{a_i}{A} ln(\frac{a_i/A}{f_i/F}) \tag{6}$$

which is adapted as a measure of *urban sprawl* in (Malpezzi and Guo n.d.).

Measures of Clustering, Centrality and Compactness Across the literature, the terms *clustering*, *centrality* and *compactness* are used to refer to the same semantics: to which degree the sub-areas with high-density are clumped closely or dispersed randomly.

Some of the reviewed measures assume the existence of a central business district (CBD), especially those that refer to the *centrality* term. Consequently, they might not be appropriate to assess cases of polycentric metropolitan areas.

Morans I (Moran 1950) proposed a measure of *spatial autocorrelation*, which given an areal decomposition of a region, is determined as:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(f_i - \bar{f})(f_j - \bar{f})}{\sum_{i=1}^{N}(f_i - \bar{f})^2} \tag{7}$$

with $w_{i,j}$ being a *weighting* between sub-areas i and j. It ranges from -1 to +1, where -1 corresponds to a "chessboard" pattern, 0 to random scattering and +1 to a strong clustering of high-density sub-areas.

It is proposed in (Torrens 2008) and (Tsai 2005) as a measure of *centrality* and *clustering* in the context of *urban sprawl*. Furthermore, (Tsai 2005) shows that I characterizes well all the following:

- number of centers (clusters of high density), as in *monocentric* (high I), *polycentric* (medium I) and *decentralised* (low I). This feature is also noted as *nuclearity* in (Galster et al. 2001)
- discontinuous or leapfrogging development
- strip development

It also argues that weighting $w_{i,j}$ with the *inverse distance* between i, j's centroids is more accurate than with *contiguity*, since *contiguity* only considers neighboring cells.

However, in the same study it is shown how I measures only the geographical 2-dimensional *dispersion*, without considering the equality of the distribution. This entails that in order to well assess *urban sprawl*, I shall be used in conjunction with *density* and *evenness of distribution* measures. Additionally, (Jelinski and J. Wu 1996) shows that I can be affected by the MAUP.

Geary's C (Geary 1954) defined another index of *spatial autocorrelation* as:

$$C = \frac{(N-1)}{2\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(f_i - f_j)^2}{\sum_{i=1}^{N}(f_i - \bar{f})^2} \tag{8}$$

which ranges from 0 indicating a positive autocorrelation to 2 indicating a negative one. A value of 1 means that there is no spatial autocorrelation whatsoever. The index can also be adjusted as in $C' = -(C-1)$ so that it matches Moran's I range.

Despite their similarity, Moran's I is a measure of global autocorrelation whereas C is more sensitive to local patterns. Moreover, (Tsai 2005) experiments with different urban forms and shows that Moran's I is more adequate to gauge *urban sprawl*, which coincides with the fact that C has a lesser presence in the topic's literature.

Suen's level of Scatter (Suen 1998) formulates the equation:

$$S_s = \sum_{i=1}^{N} \frac{f_i d_{c_i,C'}}{F} \tag{9}$$

which is also used in (Torrens and Alberti 2000). Here c_i is the centroid of the sub-area i, and $d_{c_i,C'}$ represents the distance from c_i to the weighted global centroid C'. To determine C' each sub-area i weighted by its f_i value as in $C' = \frac{1}{F}\sum_{i=1}^{N} f_i c_i$.

The existence of a CBD is not considered in S_s, and intuitively C' should respond well to polycentric cases given its weighting. Nonetheless, very few works borrow from this indicator, so there is no empirical evidence of how it responds.

Bertaud's ρ (Bertaud and Malpezzi n.d.) introduce a *compactness index* as:

$$\rho = \frac{\sum_{i=1}^{N} d_i \frac{f_i}{a_i}}{\rho_c} \tag{10}$$

where d_i is the distance of i's centroid to the CBD, and ρ_c corresponds to the converse measure for a *cylindrical* city with a circular base equal to the total area A and a constant height equal to the average density $\bar{f}$. Nevertheless the author admits that ρ might not have much significance in polycentric cities.

Galster's Clustering Coefficient (Galster et al. 2001) builds the following *clustering coefficient*:

$$G = \sum_{i=1}^{N} \frac{\sum_{s=1}^{S} |\frac{f_{i,s}}{a_{i,s}} - \frac{f_i}{a_i}|}{a_i} \tag{11}$$

where the index s corresponds to a sub-decomposition of the sub-area a_i into S further sub-areas $a_{i,s}$. The indicator aims to measure how "development is bunched to minimize occupied space" in each sub-area a_i.

Given that in the original work an areal squared decomposition is used, on it each a_i is decomposed into $S = 4$ equal sub-squares $a_{i,s}$. Notwithstanding the values of G are clearly very sensitive to the chosen decompositions.

Galster's Centralization Index (Galster et al. 2001) proposes a *centralization index* based on the work of (Massey and Denton 1988). It operates by iterating a series of concentric rings around the CBD, intending to capture how a magnitude f accumulates relative to the area as moving away from the CBD. Additionally, the study proposes to use the average distance of f to the CBD as a *centrality* indicator as well. However, both measures are strongly based on the assumption of the existence of a CBD.

2.4.3 Land Use Mix

A pervasive audit of the *land use mix* measures is found in (Song, Merlin, and Rodriguez 2013), where two main types of measures are defined: the *integral* ones, that operate over a whole area, and the *divisional* ones, which are built upon an areal decomposition.

The author also decomposes the semantics of the mixity of land uses into two different facets: the *quantity* of mix and the (geographical) *distance* or *proximity* of the most mixed areas. Such delineation is semantically akin to the two facets of *distribution*: the *evenness* and *clustering*. Often the *evenness* indicators of the *land use mix* do not borrow from an areal decomposition (*integral* measures), whereas the *clustering* ones are necessarily *divisional* measures.

On the other hand, another set of measures that will be reviewed are included in the landscape metrics of the noteworthy software FRAGSTATS (McGarigal and Marks 1995), in a framework of more general patch-like landscape analysis.

Notation In order to formally consider the land use, an extension to the notation from Section 2.4.2 will be introduced: the generic magnitude f will be categorized into different land uses as $f^{(k)}$, where k is one of the M land use types (i.e. k = residential units o4 k = shops). Analogously, $F^{(K)}$ will denote the magnitude's total value in all N sub-areas $F^{(k)} = \sum_{i=1}^{N} f^{(k)}$, and its mean is $\bar{f^{(k)}} = \frac{F^{(k)}}{N}$. If the sub-areal index i is not specified, $f^{(k)}$ will refer to the sum of the k use magnitudes over the N sub-areas as in: $f^{(k)} = \sum_{i=1}^{N} f_i^{(k)}$.

Note that the aggregation of all land uses in a given sub-area i as $\sum_{k=1}^{M} f_i^{(k)}$ does not make sense unless the values of the different land uses are comparable. For example, in a city with $F^{(act)} = 100$ activities and $F^{(res)} = 10000$ residential units and a sub-area i with $f_i^{(act)} = 2$ and $f_i^{(res)} = 1000$, the addition $f_i^{(act)} + f_i^{(res)} = 2 + 1000 = 1002$ does not give any interpretable information. Instead the relativization of those sub-areal quantities with respect to the city's

total, as in $\frac{f_i^{(act)}}{F^{(act)}} = 0.02$ and $\frac{f_i^{(res)}}{F^{(res)}} = 0.1$ might provide more comprehensible information, such as the predominance of residential development in the sub-area i.

Another convenience is the use of the letter p as in $p_i^{(k)}$ to denote the proportion of the land use k in the sub-area $i \in \Omega$. Then $P^{(k)}$ denotes the proportion of the land use k in the whole area. Such proportion might be defined in terms of area, so then $p_i^{(k)}$ corresponds to the amount of i's area destined to the land use k as in $p_i^{(k)} = \frac{a_i^{(k)}}{a_i}$. To do so, some works assume that a given sub-area i is destined to only one land use, but in such case the results might be too sensitive to the chosen areal decomposition. When the amounts of POIs of different land uses are comparable, $p_i^{(k)}$ might also be defined in terms of quantity of POIs of type k as in $p_i^{(k)} = \frac{f_i^{(k)}/F^{(k)}}{\sum_{l=1}^{M} f_i^{(l)}/F^{(l)}}$. Furthermore, the availability of data can also condition the definition of $p_i^{(k)}$. Nonetheless, the construction of the indicators in this section will generically use $p_i^{(k)}$ to allow more flexibility.

Measures of Evenness, Exposure Most of the metrics reviewed in this section are adaptations of the measures of *evenness of distribution* listed in Section 2.4.2, but with their summations defined over the land uses $k \in [1, M]$ instead of the areal decomposition Ω, turning them into *integral* measures according to the classification of (Song, Merlin, and Rodriguez 2013). Such measures are not sensible to the MAUP but ignore the geographical configuration of the uses.

On the other hand, some measures of this section do borrow from an areal decomposition Ω, and usually respond capture better some patterns that *integral* measures ignore. Nevertheless, the formulations of such measures tend to be more complex unless they are formulated for the mix of only two land uses (k, l).

Shannon' s Diversity Index relying on (Shannon 1948) communication entropy index, it can be modified to measure the diversity of "information" (or land uses) as in:

$$H_D = -\sum_{k=1}^{M} P^{(k)} ln(P^{(k)}) \tag{12}$$

where a H_D equals zero in case of absence of diversity, and approaches $ln(M)$ when the diversity is maximal. It can be adjusted to the 0 to 1 scope as in $H'_D = \frac{H_D}{ln(M)}$, in the same manner as for (5).

Herfindahl-Hirschman Index is a common measure employed often in economics to appraise marked concentration, and it is defined as:

$$HHI = \sum_{k=1}^{M} (P^{(k)})^2 \tag{13}$$

which ranges from $\frac{1}{M}$ when all land uses are equally present to 1 when there is only one use in the whole area.

For the most part, HHI behaves very similar to H_D of Equation 12, but HHI is more sensitive to the most prevalent use.

Simpson's Evenness Index is propounded by (Torrens 2008) as a measure of activity evenness as:

$$S_e = \frac{1 - \sum_{k=1}^{M} P^{(k)}}{1 - \frac{1}{M}} \tag{14}$$

which nears zero for (uneven) distributions dominated by one land use and reaches one for an area-proportional land use allocation. It is used also in (Arribas-Bel, Nijkamp, and Scholten 2011) as the indicator of land use mix.

Dissimilarity Index is presented by (Massey and Denton 1988) to reflect the levels of residential segregation. In the same way as in Equation 3, it can be adapted to measure to which degree the distribution of different land uses among sub-areas of Ω is similar to the same distribution in the whole area.

$$D^{(k,l)} = \frac{1}{2} \sum_{i=1}^{N} |p_i^{(k)} - p_i^{(l)}| \tag{15}$$

where $D^{(k,l)}$ denotes the dissimilarity of the land use k with respect to l. The index's values range from 0 to 1, indicating high and low levels of use mix respectively.

One of the interests of $D^{(k,l)}$ is that it weights each sub-area $i \in \Omega$ according to its $p_i^{(k)}$ and $p_i^{(l)}$ values, so greater land mix in areas of high density ponders more than in low density ones.

Exposure Index (Massey and Denton 1988) defines two *exposure* indices for residential segregation, considering only two types of residents: the minority k and the majority l. In this pretext, the indices are defined as:

$$E_{inter}^{(k,l)} = \sum_{i=1}^{N} \frac{f_i^{(k)}}{F^{(k)}} \frac{f_i^{(l)}}{F^{(k)} + F^{(l)}} \tag{16}$$

$$E_{isol}^{(k,l)} = \sum_{i=1}^{N} \frac{f_i^{(k)}}{F^{(k)}} \frac{f_i^{(k)}}{F^{(k)} + F^{(l)}} \tag{17}$$

where both range from 0 to 1, with higher values indicating higher exposure of k to l. The metric $E_{inter}^{(k,l)}$ refers to the interaction of the minority k with the majority l, whereas $E_{isol}^{(k,l)}$ refers to the isolation of the minority k.

The work of (Galster et al. 2001) provides an adaptation more suited for gauging *land use mix* in the context of *urban sprawl*:

$$E^{(k,l)} = \frac{NA}{F^{(k)}F^{(l)}} \sum_{i=1}^{N} \frac{f_i^{(k)} f_i^{(l)}}{a_i^2} \tag{18}$$

where an $E^{(k,l)} = 0$ means that k is not exposed to l ($f_i^{(k)}$ and $f_i^{(l)}$ are orthogonal, $\langle f_i^{(k)}, f_i^{(l)} \rangle = 0 \quad \forall i \in \Omega$). The value can reach the maximum when corresponding to areal density of one of the land uses ($\frac{f_i^{(k)}}{a_i}$ or $\frac{f_i^{(l)}}{a_i}$) in a cell i.

The three former metrics correspond semantically to measures of *exposure* (of k to l) rather than *evenness*. As reviewed in (Massey and Denton 1988) and (Song, Merlin, and Rodriguez 2013), such metrics depend "on the relative sizes of the two groups being compared while evenness measures do not". This means that when one of the land use types k or l is small over the whole area, *exposure* measures will always produce low values, whereas *evenness* ones might not.

Measures of Clustering and Proximity The formerly reviewed *land use mix* metrics are invariant to spatial permutations of the decomposition's sub-areas $i \in \Omega$. However, whether the areas with high diversity of land uses tend to be clumped together or randomly scattered in space is very relevant to *urban sprawl*. Although for *land use mix* this magnitude is often referred to as *proximity*, it actually corresponds to the more generalized concept of *clustering* or *spatial autocorrelation* (analogously to the *clustering* of Section 2.4.2).

Contagion (Turner et al. 1989) delineates an index for spatial analysis that corresponds to the probability that two randomly chosen adjacent sub-areas belong to the same class. It is constructed as:

$$CONT = 1 + \frac{1}{2ln(M)} \sum_{k=1}^{M} \sum_{k=1}^{M} P^{(k)} \frac{g_{k,l}}{g_k} ln(P^{(k)} \frac{g_{k,l}}{g_k}) \tag{19}$$

with $g_{k,l}$ representing the number of adjacencies between sub-areas of land uses k and l, and $g_k = \sum_{l=1}^{M} g_{k,l}$ the total number of adjacences of the land use k with other land uses. Values of zero indicate maximum disaggregation (small dispersed sub-areas of different land uses), whereas the maximum value of one corresponds to a large cluster of sub-areas of a single land use.

Although not considered explicitly in its formulation, Ω is implicit in Equation 19 through the adjacencies $g_{k,l}$. Additionaly, the indicator is built on the assumption that sub-areas are only of one hard land use type. This might not be a problem for studies like (Torrens and Alberti 2000; Torrens 2008) that propose to work at a pixel level, nevertheless it might be a problem for grid

divisions or census tracts. Besides, the value is though very sensitive to the areal decomposition Ω: for the same urban development pattern, $CONT$ can change significantly depending on the sub-areas' sizes.

With data about land covers other than urban development (i.e. forest, agricultural, often obtained through remote sensing), C might be used as an index of distribution of development, like the ones of Section 2.4.2.

Clustering Index (Massey and Denton 1988) intoduced a *clustering* metric for residential segregation among a minority k and a majority l as in:

$$C^{(k,l)} = \frac{\sum_{i=1}^{N} p_i^{(k)} \sum_{j=1}^{N} w_{ij} f_i^{(k)} - \frac{F^{(k)}}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}}{\sum_{i=1}^{N} p_i^{(k)} \sum_{j=1}^{N} w_{ij} (f_j^{(k)} + f_j^{(l)}) - \frac{F^{(k)}}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}} \tag{20}$$

where the weighting w_{ij} among sub-areas i and j originally corresponds to a negative exponential of the distance between i and j respective centroids d_{ij} as in $w_{ij} = e^{-d_{ij}}$.

In (Song, Merlin, and Rodriguez 2013) it is readjusted to appraise *land use mix* in the context of *urban sprawl* considering two land use types: residential and non-residential.

2.4.4 Accessibility

The *accessibility* facet is influenced by *urban sprawl* to an important extent. A *low-density* development, with a given *distribution* and *land uses* certainly conditions the possibilities of providing good *accessibility* to the residents and the services.

Some research on the topic uses scalar indicators such as average commuting times as measures of *accessibility*, nevertheless such measures do not offer any analytic approach that permits the modelling of the phenomena. A remarkable review of *accessibility* measures for *urban sprawl* is done by (Torrens and Alberti 2000). Those measures will be listed in this section.

Gravity Models are built upon ideas of Newtonian physics, and are expressed as:

$$A_{i,j} = \frac{h_i h_j}{(d_{i,j})^{\alpha}} \tag{21}$$

where h_i and h_j are respectively the capacities of the origin i to generate a trip, and of j to receive one. The distance $d_{i,j}$ between i and j has a weighting mechanism represented by α that discourages long trips.

Utility Models are based on discrete-choice models and are determined by the choice that a given individual makes out of a set of available options. These measures tend however to require a large amount of data in order to build a comprehensive utility model for a set of individual prototypes.

Isochronic Measures gauge the intuition of how many different POIs can be reached from a starting point i in a time cost lower than certain threshold τ.

3 A Framework for Measuring Sprawl

3.1 Data Sources

As pointed out in (Torrens and Alberti 2000), data is one of the main bottlenecks when it comes to define the extent of a study on *urban sprawl*. The availability of data can vary dramitically among different regions of the world, so that it constitutes already a constraint on the scope of studies on *urban sprawl*: very few intend to replicate analysis around many diverse manifestations of the phenomena. In fact, as audited in Section 2, most of the works determine their measures of *sprawl* over areas in the United States, most probably due to the greater availability of data.

Nonetheless, the latter literature brought *urban sprawl* to a more transversal geographical context, specially those built on data obtained through remote sensing. Moreover, a new generation of sources of data is emerging: crowd-sourced data. There are several considerations to have in mind when working with crowd-sourced data, such as the commonness of missing data or reliability of the user inputs.

Despite such concerns, the amount of crowd-sourced data has been continuously increasing during the last years, and it is reasonably expected to become soon one of the main sources of data for many fields of research. Besides, among others, quality metrics and user reporting and contributor reviews are intended to improve the data's soundness. At last but not least, one of the most important advantages of this data is the ease of access to it.

There exist already significant sources of crowd-sourced data that can be useful in order to assess different axes of *urban sprawl* at a world scale. Although at this point this work only uses data from OpenStreetMap (OSM), the following sub-sections will comment on sources that can be relevant (and used in the future) for extensions of this project.

3.1.1 OpenStreetMap

The OpenStreetMap (OSM) is a collaborative project to create a free editable map of the world, which results to be a prominent example of volunteered geographic information (VGI). It is a knoweledge collective that provides user-generated street maps (Haklay and Weber 2008). Volunteers across the world share geographic information to OSM in various ways, also considered as "intelligent sensors" (Goodchild 2007).

Since its creation, the project has been increasingly used across the world for a wide variety of purposes. Quality metrics have been proposed in (Forghani and Delavar 2014; Barron, Neis, and Zipf 2014; Mooney, Corcoran, and Winstanley 2010; Fan et al. 2014), followed by different quality assessments, in particular for different countries. For instance, it has been concluded that the quality is

"fairly accurate" for England (Haklay 2010), and it is even shown that OSM data is superior to the official dataset for Great Britain Meridian 2. Thus, the previous work has been extended for France (Girres and Touya 2010). Studies focusing on the street network of Germany have been also been conducted in (Neis, Zielstra, and Zipf 2011), where it is concluded that the data-sets can be considered complete in relative comparison to a commercial dataset. In addition, the OSM data-set for Hamburg already covers about 99.8% of the street network (Over et al. 2010) according to the surveying office of Hamburg. The latter study also remarks that "Besides the street network, the real advantage of the dataset is the availability of manifold points of interest". These POIs allow for deeper understanding of cities dynamics, enriched with the provided location and the embedded information.

Extraction of information The documentation of the service is provided in the OSM Wiki. The OSM data is usually represented as a tree of XML-like named elements with key:value tags. Such elements might represent different kinds of roads, POIs, or land uses[2].

Overpass API The OSM data can be retrieved through the read-only Overpass API which supports its own query language[3].

Mapzen Metro Extracts A collection of OSM data is maintained by Mapzen Metro Extracts and can be retrieved for a set of cities in form of the GIS data Shapefiles. The service bypasses the OSM to PostGIS conversion of the data so the user can directly handle it with commonly used GIS software, such as ArcGIS or the open-source QGIS.

3.1.2 Other Data Sources to Consider

Transitland The site Transitland gathers transit data from many sources to offer an API which users can query for routes, stops or schedules of different transportation means and operators. Nevertheless it does not have many contributions outside the United States yet.

Bike Share Map Data about bicycle systems is collected and displayed in The Bike Share Map for more than a hundred cities around the world.

[2]Tag specifications of OSM: roads (`http://wiki.openstreetmap.org/wiki/Key:highway`); POIs (`http://wiki.openstreetmap.org/wiki/Key:amenity`, `http://wiki.openstreetmap.org/wiki/Key:shop`, `http://wiki.openstreetmap.org/wiki/Buildings`); and land uses (`http://wiki.openstreetmap.org/wiki/Key:landuse`)

[3]Overpass Query Language: `http://wiki.openstreetmap.org/wiki/Overpass_API/Overpass_QL`

3.2 Formalization of the Framework

3.2.1 Notation

In addition to the notation previously defined when reviewing the indicators in the literature in Section 2.4, the following conventions will be adopted:

- **Region:** $S = \{(lon, lat) \in B\} \subset \mathbb{R}^2$ where *lon* and *lat* mean *latitude* and *longitude* respectively, and B is the *region boundary*. This will correspond to the metropolitan area of the city in question. In the same way as in Section 2.4, Ω will denote the areal decomposition of S into N sub-areas.
- **Road Graph:** $G = (V, E)$ in the *primal form*, where *nodes* $v \in V$ represent intersections and are geo-referenced as in $v = (lon, lat)$ and each *edge* $e = uv = (u, v) \in E$ represents a *directed* path between the nodes $u \in V$ and $v \in V$. There is also a *distance* function $d : E \longrightarrow \mathbb{R}^+$ and a *travel cost* function $c : E \longrightarrow \mathbb{R}^+$ to gauge accessibility more precisely.
- **Points of Interest (POIs):** $P = \{p_1, ..., p_M\}$ where every *point of interest (poi)* is of the form $p = (cat, lon, lat) \quad \forall p \in P$ and is categorized by the *cat* component and geo-referenced through (lon, lat).

3.2.2 Obtaining Data

Given a region of interest S, two data structures must be constructed:

i A *road graph* $G = (V, E)$ with the *nodes* that are located inside the region S as in $v = (lon, lat) : (lon, lat) \in S \quad \forall v \in V$. The *distance* function for an edge $e = uv$, $d(e)$ can be determined using the *Haversine formula*[4] given that u and v are geo-referenced. For the *travel cost* function of the edge $c(e)$, its value will be computed as a product of e distance $d(e)$ and a velocity coefficient γ that will depend on what kind of road does the edge represent according to e's OSM tags [2].

ii A set of *POIs* P such that each *poi* is located inside the region S as in $p = (cat, lon, lat) : (lon, lat) \in S \quad \forall p \in P$. The *categories* will be divided into two main axes: *activities* and *residential*, according to p's OSM tags[2].

3.3 Considerations when Choosing Measures

3.3.1 Suitability

Measuring *urban sprawl* is a complex task which in certain situations is only feasible under very subjective assumptions. An extensive series of suitability criteria for such indicators is defined in (Jaeger, Bertiller, Schwick, and Kienast 2010). This specifications include ease of interpretation, simplicity and other mathematical blueprints.

[4] See `http://en.wikipedia.org/wiki/Haversine_formula`

Some works often borrow from a large set of indicators that are strongly correlated among themselves, and use statistical techniques such as principal component analysis (PCA) in order to build measures. Such an approach can give interpretable results but it does not provide any formal model to assess *urban sprawl* outside of the data that has been used for the analysis.

On the other hand, there is a clear interdependence between the dimensions of *urban sprawl*. For example, a given *distribution of development* clearly constraints the *accessibility*. Furthermore, the *accessibility* is also most likely related to the *density* and the allocation of *land uses*. This has to be taken into consideration, and can mean that the aggregated indicators might not permit a clear interpretation of the phenomena.

3.3.2 Data

Several works in the literature, such as (Torrens and Alberti 2000) and (Jaeger, Bertiller, Schwick, and Kienast 2010), remark that the availability of data is one of the main bottlenecks that limit the potential extent of studies of *urban sprawl*. The choice of crowd-sourced data is principally justified by the following reasons:

- crowd-sourced data is often of open access, so the measures that are determined out of it can be commented, compared, contrasted or improved by the research community more easily
- there exist a very active community of developers that craft libraries that interact with the data sources APIs, and process its data into easily manipulable formats
- the amount of crowd-sourced data as well as its reliability already has, and is expected to improve greatly

In the context of this framework, the two main data structures mentioned in Section 3.2.2, G and P can be easily obtained through OSM. For the case of P's categories, the OSM tag specification[2] allows for many categorical classifications, however for this first version only two main axes are considered: activities and residential. This choice is for the sake of simplicity, which is very subjective, but it can be conveniently modified if desired.

3.3.3 Quasi-Continuous Surfaces

Given a set of data points, the Kernel Density Estimation (KDE) interpolates a continuous surface through a given kernel (e.g. a *Gaussian/Normal density function*). As (Torrens and Alberti 2000) expounds, there can be advantages of surface smoothing when measuring urban sprawl as it allows, for a sub-region, the legitimate inference of its magnitude density based on the observations in the neighbouring regions. A visualization of how a KDE over S region's POIs P looks is displayed in Figure 1

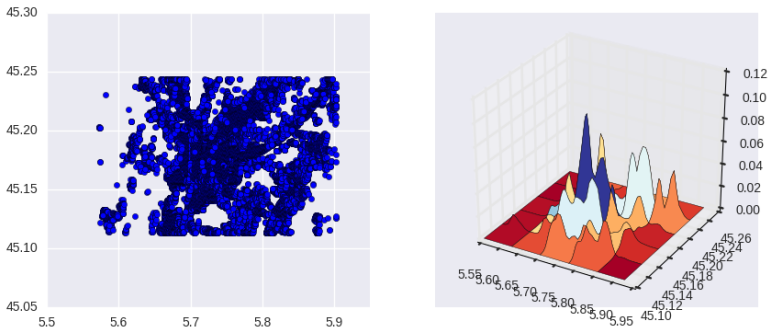

Figure 1: KDE inference of density for a given distribution of *residential* POIs. Location: *Grenoble, France*

To determine the indicator's values given the S region's categorized POIs P, and areal decomposition, the proposed framework offers two options:

(i) for each category k take the number of POIs that fall into each cell i as $f_i^{(k)}$ as the magnitude to measure, or

(ii) compute a KDE $\psi^{(k)}$ for each category k, and for each cell i take its average as the magnitude to measure $f_i^{(k)}$ (determined as $f_i^{(k)} = \frac{1}{a_i}\int_i \psi^{(k)} da$)

The option (ii) can both be convenient when the missing data is distributed spatially in an even way, the inference of density through a KDE function can be reasonable. See Figure 2 for an example of this situation.

Figure 2: KDE inference of density for an even spatial distribution of the missing *residential* POIs. In such a situation, it is reasonable to assume that in such a grid there are more residential POIs given the OSM-rendered gray color and the existing roads. Location: *Chandigarh, India*

However, in some other situations such as regions with very dispersed POIs, it does not make sense to introduce density since the areas around a POI are

actually undeveloped. Furthermore, such an interpolation might impede the spotting of *urban sprawl* manifestations such as *leapfrog* development. An illustration of these conditions is shown in Figure 3.

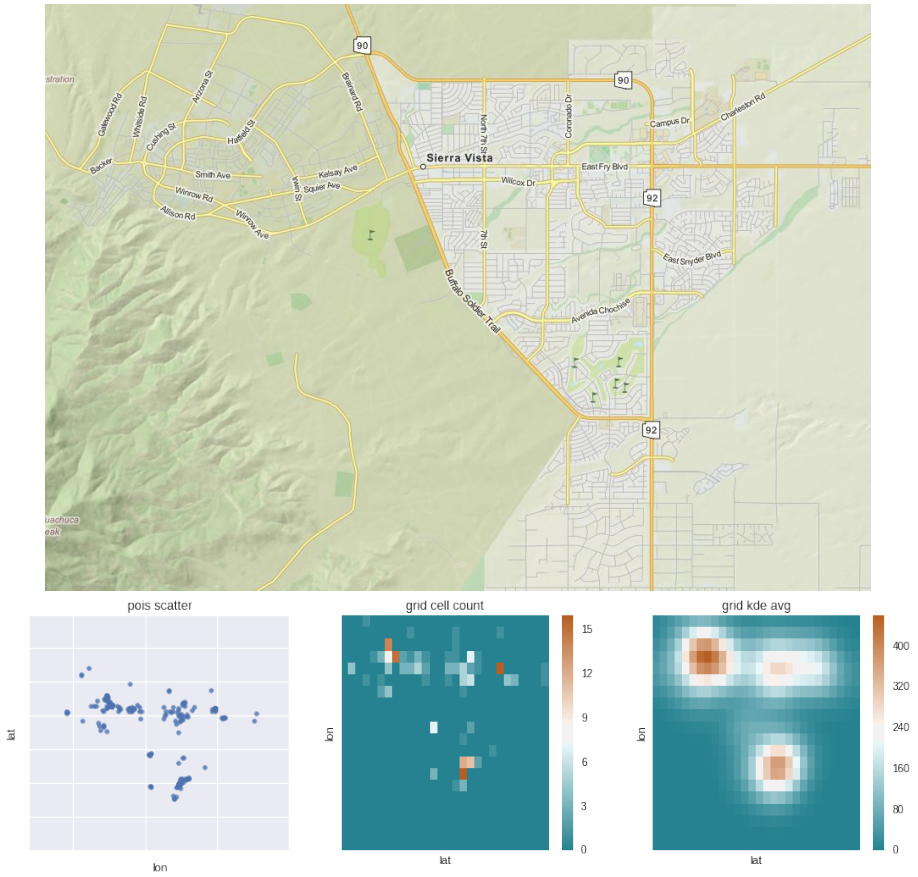

Figure 3: Above there is the OSM rendering of the region. Below there is, from left to right: (1) the spatial distribution of *activity* POIs, (2) number of *activity* POIs inside every cell of a given decomposition Ω, (3) KDE inferred average density *activity* POIs of each Ω's cell. For such a dispersed distribution, interpolating a continuous density function might pollute the data and indicate less *urban sprawl*. Location: *Sierra Vista, Arizona, USA*

In any case, it must be accounted that interpolation of density, by its own nature will smooth the evenness and consequently the values of the correspondent measures listed in Section 2.4.2 and Section 2.4.3.

3.4 Proposed Indicators

Bearing in mind the measures of *urban sprawl* reviewed in Section 2.4 and the considerations detailed in Section 3.3, this section proposes a set of indicators conceived to gauge the different dimensions of *urban sprawl*.

3.4.1 Density

As reviewed in Section 2.4.1, the density itself is empirically strongly correlated to other related measures such as statistics on tract densities. Consequently, the following three straightforward indicators of density will be adopted:

- **Activity Density** deduced from the OSM activity POIs P that lay inside the region S. The coverage of activities is reasonabliy satisfactory in all the major cities.
- **Residential Units Density** determined out of the OSM residential POIs P for the region S. It might not be a very reliable indicator since the current contributions in this category vary dramatically among cities.

3.4.2 Development's Distribution

The measures used to gauge the *distribution* of urban development will be determined separately for both the activity and residential land uses. Each one of the *distribution*'s facets will be assigned one indicator.

Evenness After the revision of some indicators of equality of distribution in Section 2.4.2, the chosen indicator for this magnitude is:

- **Shannon Entropy H** determined by Equation 5, with $f_i^{(k)}$ being the average KDE density of the land use k in the sub-area i. The choice is justified after its empirically-proven robust behaviour towards variations of areal decompositions, as well as lower computational cost $O(N)$ (instead of $O(N^2)$ of the Gini).

One of the advantages of the famework's KDE option described in Section 3.3.3, is that with a KDE-based $f_i^{(k)}$, H is calculable since $f_i^{(k)} > 0$ for any i and k. The same does not happen when using the POIs count as the indexes' magnitude $f_i^{(k)}$ because a given sub-area i can have zero occurrences of POIs of the category k (i.e. a park i without residential units will have $f_i^{(res)} = 0$). Note that ignoring such sub-areas would prevent the indicators of detecting patterns of *sprawl* such as *leapfrog* or *discontinuous* development.

Clustering Considering the literature review of *clustering*, *compactness* and *centrality* indicators from Section 2.4.2, and after (Tsai 2005) exhaustive inspection of its behaviour, the measure used for this facet will be:

- **Moran's I** calculated as in Equation 7 with $f_i^{(k)}$ as the number of POIs of category k in the sub-area i, and with the inverse distance as weighting function. Its computational cost $O(N^2)$ can be though an issue to consider for large regions.

3.4.3 Land Use Mix

The measures of *land use mix* audited in Section 2.4.3 are rather similar to those that asses the spatial *distribution*. This is arguably because the semantics of both dimensions are very similar and can also be divided into *evenness* and *clustering* (or *compactness*).

Evenness Thera are two important considerations to mention to justify the choice of this measure, both having to do with the data extracted from OSM. The first, is that such data does not permit a *hard* decomposition of the land uses, with each sub-area i corresponding to one and only one land use k. The second is that the data is classified into two main categories: activities and residential. Taking into account these concerns, the measures from Equation 12 and Equation 14 do not seem appropriate, and so the following indicator will be employed:

- **Exposure Index E** calculated as in Equation 18 which given that only two land uses are considered, it only needs to be determined once as $E_{k,l}$ reveals the exposure of land use k to l (in this context k =activities, l =residential) and it is symmetric.

Clustering Reckoning still with the former data considerations, it can be observed that the Contagion index of Equation 19 does not appear to be convenient.

The idea is to conceive a new indicator based on Geary's C (Equation 8), that will be noted as Φ. Given an areal decomposition Ω, Φ iterates over the combination set of pairs of sub-areas $\binom{\Omega}{2}$. For each pair $(i,j) \in \binom{\Omega}{2}$, a term $\phi_{i,j}$ will be determined, fulfilling the following specifications:

1. mixity of land uses *inside a given sub-area* i should always increase the global land use mixity Φ
2. mixity of land uses *for adjacent or neighbouring sub-areas* i, j should always increase the global land use mixity as well Φ

Consequently, each term $\phi_{i,j}$ will be of the form:

$$\phi_{i,j} = \phi_{intra}(i) + \phi_{intra}(j) + w_{i,j}\phi_{inter}(i,j) \tag{22}$$

where ϕ_{intra} will be a function that satisfies the first specification, whereas ϕ_{inter} fulfils the second one. The weighting $w_{i,j}$ for $\phi_{inter}(i,j)$ will increase more the global mixity Φ for i and j that are close to each other. With the

adequate choice of such functions, the indicator Φ should be able to gauge to which extent zones with high mixity of land uses are situated closely. Such magnitude is often referred to as *proximity* in the spatial analysis literature.

3.4.4 Accessibility

Given that OSM can provide the road graph G as described in Section 3.2.2, the measures of *accessibility* should be built over this information.

A first approach can be to apply Equation 21 to an areal decomposition Ω, the trip's origin being then a sub-area i and the destination another sub-area j. The capacities of generating and receiving trips could be determined by the zones' intensity of land uses, with some function σ depending on all the combinations of $f_i^{(k)}$, as in $\sigma(f_i^{(act)}, f_i^{(res)}, f_j^{(act)}, f_j^{(res)})$.

Other topological properties of G might also reveal information relevant to *accessibility*, such as the G's efficiency ε (ratio between the euclidean distance between two points and the actual distance when traversing G's corresponding edges) or the centrality indices of G's nodes V. For example, Figure 4 shows an illustration of how G's nodes' $v \in V$ centrality is correlated to its interpolated activity density $f_v^{(act)}$ (which gives another case where the KDE pre-processing of Section 3.3.3 might be useful).

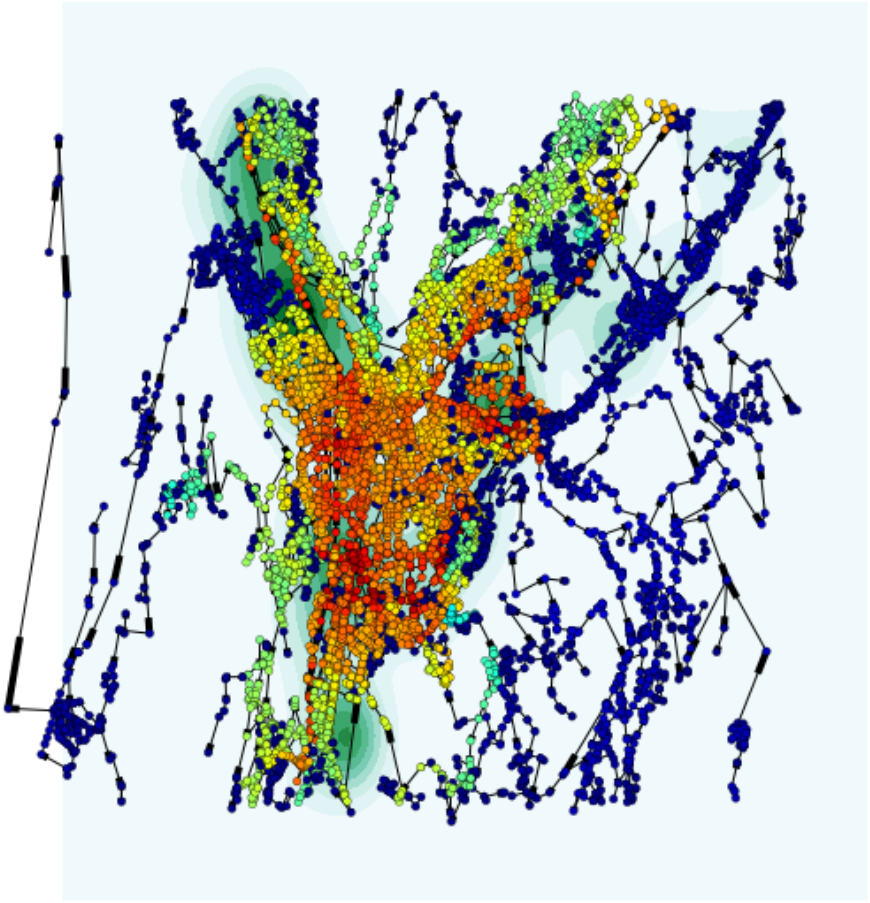

Figure 4: Road Graph $G = (V, E)$ with nodes colored according to their *closeness centrality* value, and plotted over the activity KDE $\psi^{(act)}$. There is a significant 0.591955 Pearson's correlation between the *closeness centrality* of each $v \in V$ and its interpolated activity KDE $f_v^{(act)}$. Location: *Grenoble, France*

4 Experiments

4.1 Chosen Dataset

A set of cities available in the Mapzen Metro Extracts (Section 3.1.1) have been processed within the proposed famework in order to compute some of the already implemented indicators. The datasets have previously been manually explored in OSM in order that the missing data is not a determinant factor of the analysis' results.

Ávila is a medieval town of the historical region of Castille, in Spain. The city's oldest part is surrounded by prominent walls. Such an architectural setting can be expected to affect the distribution of the POIs. An OSM rendering of the city's area and the distribution for the two categories of POIs is shown in Figure 5.

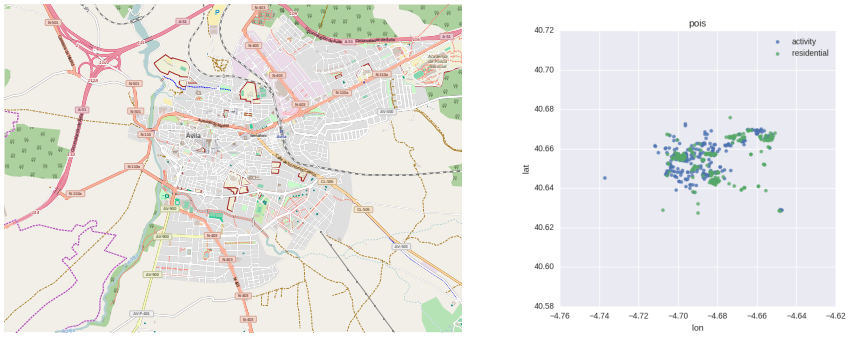

Figure 5: Left: OSM rendering. Right: categorized POIs distribution. Location: *Ávila, Spain*

Chandigarh is one of the early planned cities in the post-independence India, and it is well known for its urbanism after the master plan conducted by the remarkable Swiss-French architect and urban planner Le Corbusier. The city is divided into 60 sectors of 800 x 1200 meters, where residential units and activities such as stores, facilities or offices are distributed in an even way. Additionally, the urban pattern assures a fluid traffic circulation as well as a significant availability of green spaces. The OSM rendering and POIs are displayed in Figure 6.

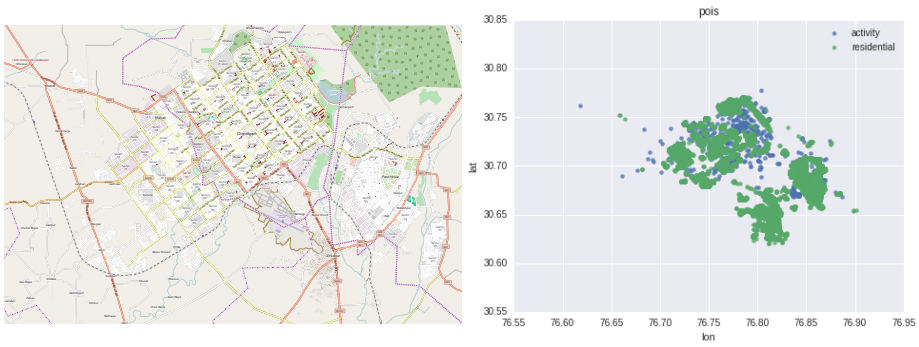

Figure 6: Left: OSM rendering. Right: categorized POIs distribution. Location: *Chandigarh, India*

Dresden is an ancient city in the Free State of Saxony (currently Germany) which has had several important remodellations throughout the history. The most remarkable one corresponds to the transformation of the city after the destruction that it suffered during the World War II bombings. After the war, an industrial period came, and with it, the urban growth came too. Specially the satellite towns experienced important urban development in the form of single residential units, that ended up creating a low-density scattered housing zone in the city's outskirts. According to (Ludlow 2006), it is one of the most *sprawled* cities in Europe. The OSM rendering and POIs are shown in Figure 7.

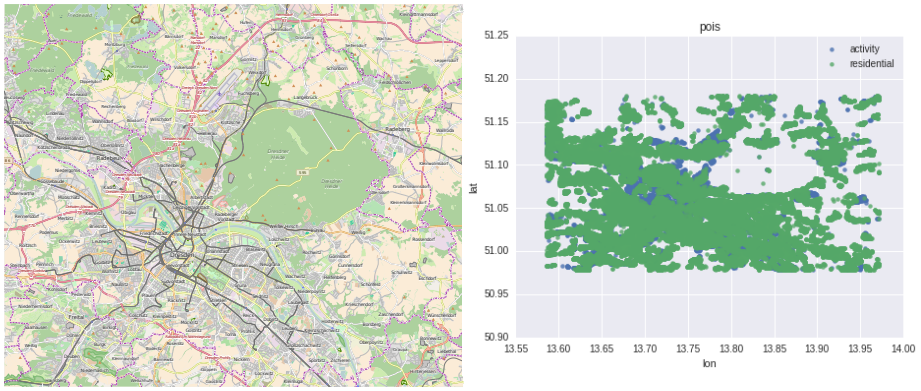

Figure 7: Left: OSM rendering. Right: categorized POIs distribution. Location: *Dresden, Germany*

Grenoble is one of the major cities located in the Alps mountain range (in the French part). Due to its location, it is often referred to as "The Capital of the Alps". Its first settlements by Gallic tribes date back from 43 BC. The city has lived several periods of important growth, the latter ones being in 1925 with the International Exhibition of Hydropower and Tourism, and after the Xth Olympic Winter Games in 1968. Nowadays, the city is developed around its historical center, clumped with its neighbouring municipalities in valley, and then encompassed by small towns and ski resorts that are scattered over the surrounding mountains. In such settling, the topography of the region can be expected to condition the patterns of urban development. Its OSM rendering and POIs are shown in Figure 8.

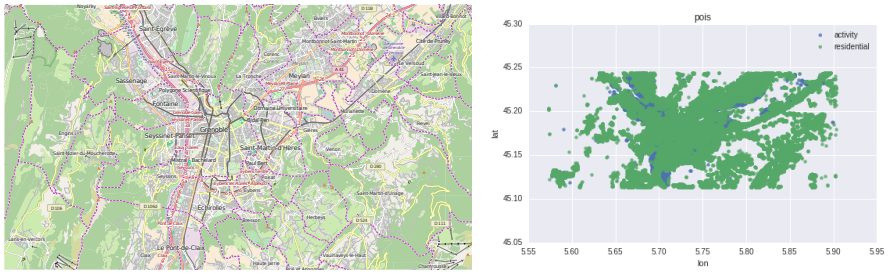

Figure 8: Left: OSM rendering. Right: categorized POIs distribution. Location: *Grenoble, France*

Raleigh is the capital of the state of North Carolina, in the United States of America. It is one of the cities that has encountered greatest sub-urban growth after the opening of the Research Triangle Park in 1959, which is now one of the largest research centers in the world. The term "Triangle" comes from the fact that three cities are part of that research pole: Raleigh, Durham and Chapel Hill. The low density and scatteredness of the sub-urban development of the three cities has practically connected them contiguously. According to (R. H. Ewing, Hamidi, and America 2014), it is one of the most *sprawled* metropolitan areas in the United States. Its spatial disposition is illustrated in Figure 9 through the OSM rendering and the POIs distribution.

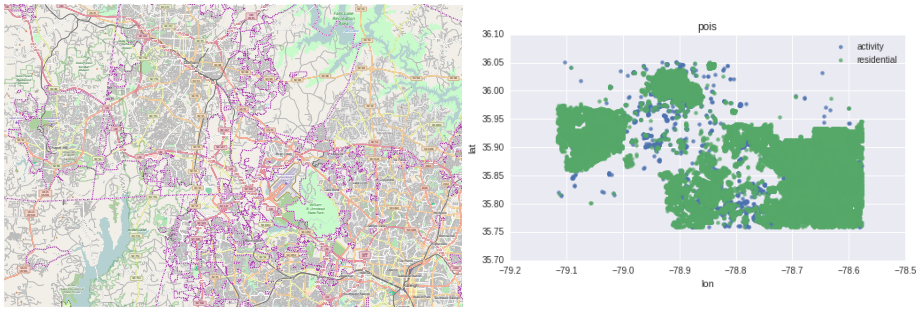

Figure 9: Left: OSM rendering. Right: categorized POIs distribution. Location: *Raleigh, North Carolina, USA*

4.2 Results

The values of the indicators designated in Section 3.4 are listed for each of the cities in Table 1. For each column, the values are determined corresponding to POI occurrences for the activity "act" and residential "res" categories separately.

Several squared grid decompositions with different step sizes each have been

tested until the values showed a variation of less than a 5% variation in the last 3 samples.

City	Area (km^2)	POI count		Moran's I		Entropy H'^5	
		act	res	act	res	act	res
Ávila	40.9252	291	326	0.0563	0.0211	0.8654	0.7726
Chandigarh	474.3383	825	32848	0.0249	0.0597	0.7964	0.8293
Dresden	605.4000	21654	96645	0.0649	0.0399	0.9409	0.9696
Grenoble	376.4705	11260	84443	0.0547	0.0862	0.8697	0.9354
Raleigh	1575.2812	6718	126871	0.0147	0.1053	0.9389	0.9522

Table 1: City values for the POI count, Moran's I and Relative Shannon's Entropy H' indicators.

Considering the cities aesthetics and their descriptions provided in Section 4.1, it seems reasonable to assume that Ávila and Chandigarh will show a lesser extent of *urban sprawl* than Dresden, Grenoble and Raleigh. It has to be observed though, that the data corresponding to the "POI count" column does not vary proportionally to the city sizes, which probably indicates that some of the cities have a greater contribution coverage of their activities and residential POIs. Such factor could greatly influence the results, so it has to be taken into consideration before exposing strong claims based on the indicators' values.

The values for the Moran's I are relatively close to 0 in all the entries, which reveals that the POIs distributions follow a random scattering. There do not seem to be clear conclusions to be extracted, not when comparing between cities nor when comparing between the "act" and "res" categories. With the exception of Ávila, that has very few POIs, Moran's I seems to actually be correlated to the amount of pois of each category, with I showing higher values as more POIs are present. Further studies with cities of similar "Area" and "POI count" but different distribution patterns are certainly promising to accept or refuse that I is strongly influenced by the number of POIs. In any case, the patterns of POIs' dispersion seem to be too complex to evaluate just with a scalar value such as I.

On the other hand, the values of the relative entropy H' do clearly respond more accordingly to the expected results. First, Ávila shows smaller values of H', which can reasonabily be assumed that is related to the presence of a medieval wall that concentrates the development in smaller parts of the city. The iconic urban planning of Chandigarh also seems to have contributed to a more uniform distribution of the POIs, as indicated by the also relatively small H's. The cities that were assumed to be *sprawled*, Dresden and Raleigh do show noticeable higher dispersion of the POIs of both categories. Furthermore, the H' index also captured a very particular pattern of the city of Grenoble, where activity POIs are concentrated in the valley (lower H') and the residential ones are more dispersed among the surrounding mountain ranges.

5 Future Work

5.1 Comprehensive Characterization of Cities through Urban Sprawl

One of the experiments that are intended to be conducted through the proposed framework intends to determine to what extent *urban sprawl* is related to aspects of the cities more transversal than urban morphology and use composition, for example socio-economic (car ownership, segregation, crime) or environmental factors.

5.1.1 A Dataset of City Indicators

As mentioned before, the avability of data conditions the extend of any urban analysis. This constraint must be considered when defining a dataset X for this experiment, in two parts that are dependent on each other:

1. The dataset X should feature a large-enough number of cities m so the conclusions of further analysis are not too specific to X
2. A set of d indicators relevant to the context of the study must be available *for each* city $\vec{x_i} \in X$

There is then a compromise between (1) and (2), since the larger the m, the harder it is to find reliable data on the d indicators. Conversely, if the analysis wants to consider a large number of indicators d, there will be a smaller number of cities m that have available data for all the indicators.

Global Power City Index the Global Power City Index (GPCI)[6] provides a considerable number of reliable indicators that are relevant to this study ($d = 23$ indicators have been selected), which are listed in Table 2.

[6] See `http://www.mori-m-foundation.or.jp/english/ius2/gpci2/` for more information on the GPCI series

ID	Description
1	Total Unemployment Rate
2	Total Working Hours
3	Level of Satisfaction of Employees with Their Lives
4	Average House Rent
5	Price Level
6	Number of Murders per Population
7	Disaster Vulnerability
8	Life Expectancy at Age 1 & 60
9	Number of Medical Doctors per Population
10	Population Density
11	Percentage of Renewable Energy Used
12	Percentage of Waste Recycled
13	CO2 Emissions
14	Density of Suspended Particulate Matter (SPM)
15	Density of Sulfur Dioxide (SO2), Density of Nitrogen Dioxide (NO2)
16	Water Quality
17	Level of Green Coverage
18	Number of Runways
19	Density of Railway Stations
20	Punctuality and Coverage of Public Transportation
21	Commuting Convenience
22	Transportation Fatalities per Population
23	Taxi Fare

Table 2: GPCI Indicators relevant to the study

Numbeo is a large database of user contributions which provides a set of indicators classified into seven main axes: (i) cost of living, (ii) property prices, (iii) crime, (iv) health care, (v) pollution, (vi) traffic, and (vii) quality of life. It provides an extensive API[7] with a free academic license. The provided information might be used in complemention of the GPCI indicators.

Building the Dataset given the choice of the GPCI indicators from Table 2, there are $m = 40$ cities that appear in all the editions of GPCI that have been accessible to this study, which are listed in Table 3.

[7]See http://www.numbeo.com/common/api.jsp for a detailed specification of Numbeo's API

ID	City	ID	City	ID	City	ID	City
1	Amsterdam	11	Frankfurt	21	Milan	31	Singapore
2	Bangkok	12	Fukuoka	22	Moscow	32	Stockholm
3	Barcelona	13	Geneva	23	Mumbai	33	Sydney
4	Beijing	14	Hong Kong	24	New York	34	Taipei
5	Berlin	15	Istanbul	25	Osaka	35	Tokyo
6	Boston	16	Kuala Lumpur	26	Paris	36	Toronto
7	Brussels	17	London	27	San Francisco	37	Vancouver
8	Cairo	18	Los Angeles	28	Sao Paulo	38	Vienna
9	Chicago	19	Madrid	29	Seoul	39	Washington
10	Copenhagen	20	Mexico City	30	Shanghai	40	Zurich

Table 3: Cities featured in all GPCI editions

The combination of d indicators and m cities yields the dataset X that will be considered in further sections. In such dataset, each city is represented by a vector $\vec{x} = (x_1, ..., x_d)$ where each component x_j represents the j-th indicator of the city $\vec{x} \in X$.

5.1.2 Clustering Cities

The number of samples of X, m can be reduced into a smaller dataset X' with $k << m$ prototypes through clustering techniques. The work of (Arribas-Bel, Nijkamp, and Scholten 2011) presents a similar analysis (in the context of European cities only) through the self-organizing maps (SOM) algorithm which performs both (i) a reduction of the dimensionality d and (ii) a reduction of the number of samples m.

With the dataset X of Section 5.1.1, a first sample-clustering has been performed with the k-means algorithm for several desired number of prototypes k, and with 1000 different initial centroid random choices. The strength of the clustering results has been evaluated through the silhouette score for $4 < k < 10$, with the best classification found for $k = 4$ and corresponding to the clusters shown in Table 4

ID	Characteristics	Cities
(a)	mainly European cities	Amsterdam, Barcelona, Berlin, Brussels, Copenhagen, Frankfurt, Fukuoka, Geneva, Madrid, Milan, Osaka, Stockholm, Vancouver, Vienna, Zurich
(b)	developed mega-cities from different continents	London, New York, Paris, Seoul, Singapore, Tokyo
(c)	mega-cities of developing regions	Bangkok, Beijing, Cairo, Hong Kong, Istanbul, Kuala Lumpur, Mexico City, Mumbai, Sao Paulo, Shanghai, Taipei
(d)	mainly North-American cities	Boston, Chicago, Los Angeles, Moscow, San Francisco, Sydney, Toronto, Washington D.C.

Table 4: results out of 1000 randomly initialized k-means clustering for $k = 4$ (silhouette average of 0.200395072433)

5.1.3 Linking City Indicators and Sprawl Measures

The objective of this part is to see to what extend the measures of *urban sprawl* proposed in Section 3.4 are related to the cities' d indicators.

The idea is to explore which combination of indicators city can precisely determine to which of theprototypes (a, b, c or d of the clustering results of Table 4) the city relates better. Such exploration can be formulated as a *supervised learning* problem.

The available dataset will be represented with $S = \{(\vec{u_i}, y_i) \quad \forall i \in [1, m]\}$ where for each city i:

- its r *urban sprawl* measures are determined as explained in Section 3.4, so $\vec{u_i} = (u_{i,1}, ..., u_{i,r})$ is known
- it belongs to one of the k clusters obtained in Table 4 $y_i \in [1, k]$

With a training set $S_{train} \subset S$, the aim is to learn a multi-class classifier of the form $h_w : U \to K$ where:

- U is the vector space defined by the r *urban sprawl* measures' values
- $K = \{1, ..., k\}$ is the set of k possible classes that correspond to the clusters of Table 4

Then, given the *urban sprawl* measured values of the training set S_{train}, the trained classifier h_w should be able to determine to which cluster the city belongs for the remaining $S_{test} = S \setminus S_{train}$. Furthermore, the weight vector

w associated to the classifier h_w should reveal clues about which *urban sprawl* measures are more determinant for the cluster choice.

It must be remarked however that the current state of the OSM's contributions do not allow to draw very reliable conclusions, since the coverage of the POIs might vary dramatically among the cities of Table 3.

6 Conclusions

This work performs an extensive literature review for the loose an ambiguous phenomena of *urban sprawl*, in an effort to delineate the quantifiable aspects of it. Several measures are audited in order to compute indicators that can help assessing the phenomena for a configuration given at certain time instance. It is remarkable how many of the proposed measures are built on very questionable assumptions, so their revision in a common notation might help spot these weaknessess.

Choices of appropriate measures are justified, and for the elected ones, its computation is included in the framework. The framework allows flexibility in the calculation of the formulas and includes a kernel density estimation (KDE) preprocessing that can help in some cases to overcome missing data. Furthermore, the KDE preprocessing also permits the calculation of Entropy measures since the zones of zero density are smoothed.

The framework's operation over user-contributed and open-access data such as OpenStreetMap (OSM), allows the research community to collaboratively work on the subject in a common frame, and intends to reduce the ambiguity and looseness of the research on *urban sprawl*. It is a novel contribution that is not present in the reviewed literature. The current coverage of OSM presents still important lacks of data in most of the world's big cities. However, the contributions have been continuously increasing, and there exists an additional large amount of literature in other aspects of crowd-sourced data that can help ensuring its reliability.

The first experimentations conducted seem to indicate that most of the scalar *urban sprawl* indicators might not be easy to interpret, as there is too much information to be extracted from just one value. Nevertheless the flexibility of the framework paves the way for many possible automated experiments based on OSM data, and the inclusion of additional data sources as well as the implementation of new measures can make the framework a very important contribution to the research of *urban sprawl*.

References

Angel, Shlomo, Jason Parent, and Daniel Civco (2007). "Urban sprawl metrics: an analysis of global urban expansion using GIS". In: *Proceedings of ASPRS 2007 Annual Conference, Tampa, Florida May*. Vol. 7. 11. Citeseer.

Arbury, Joshua. "From Urban Sprawl to Compact City–An analysis of urban growth management in Auckland". In:

Archer, Raymon Walter (1973). "Land speculation and scattered development; failures in the urban-fringe land market". In: *Urban Studies* 10.3, pp. 367–372.

Arribas-Bel, Daniel, Peter Nijkamp, and Henk Scholten (2011). "Multidimensional urban sprawl in Europe: A self-organizing map approach". In: *Computers, Environment and Urban Systems* 35.4, pp. 263–275.

Bahl, Roy W (1968). "A land speculation model: the role of the property tax as a constraint to urban sprawl". In: *Journal of Regional Science* 8.2, pp. 199–208.

Barron, Christopher, Pascal Neis, and Alexander Zipf (2014). "A comprehensive framework for intrinsic OpenStreetMap quality analysis". In: *Transactions in GIS* 18.6, pp. 877–895.

Batty, Michael (2007). *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*.

Batty, Michael and Kwang Sik Kim (1992). "Form follows function: reformulating urban population density functions". In: *Urban studies* 29.7, pp. 1043–1069.

Beach, Dana (2003). "Coastal sprawl: The effects of urban design on aquatic ecosystems". In: *of the United States, Pew Oceans Commission 2002*. Citeseer.

Bertaud, Alain and Stephen Malpezzi. "The spatial distribution of population in 35 World Cities: the role of markets, planning and topography". In:

Bhatta, B (2009). "Analysis of urban growth pattern using remote sensing and GIS: a case study of Kolkata, India". In: *International Journal of Remote Sensing* 30.18, pp. 4733–4746.

Bhatta, Ba, S Saraswati, and D Bandyopadhyay (2010). "Urban sprawl measurement from remote sensing data". In: *Applied geography* 30.4, pp. 731–740.

Blair, Robert (2004). "The effects of urban sprawl on birds at multiple levels of biological organization". In: *Ecology and Society* 9.5, p. 2.

Burchell, Robert W et al. (1998). *The costs of sprawl-revisited*. Project H-10 FY'95.

Catalán, Bibiana, David Saurí, and Pere Serra (2008). "Urban sprawl in the Mediterranean?: Patterns of growth and change in the Barcelona Metropolitan Region 1993–2000". In: *Landscape and urban planning* 85.3, pp. 174–184.

Clawson, Marion (1962). "Urban sprawl and speculation in suburban land". In: *Land economics* 38.2, pp. 99–111.

Club, Sierra (1998). *Sprawl: The Dark Side of the American Dream*. Sierra Club. URL: https://books.google.fr/books?id=LIAinQEACAAJ.

De La Barra, Tomas (1989). *Integrated land use and transport modelling. Decision chains and hierarchies*. 12.

Downs, Anthony (1999). "Some realities about sprawl and urban decline". In: *Housing policy debate* 10.4, pp. 955–974.

Eid, Jean et al. (2008). "Fat city: Questioning the relationship between urban sprawl and obesity". In: *Journal of Urban Economics* 63.2, pp. 385–404.

Ewing, R, R Pendall, and D Chen (2002). "Measuring sprawl and its impact: Smart growth America [M/OL]". In:

Ewing, Reid (1997). "Is Los Angeles-style sprawl desirable?" In: *Journal of the American planning association* 63.1, pp. 107–126.

Ewing, Reid H (1995). "Characteristics, causes, and effects of sprawl: A literature review". In: *Urban Ecology*. Springer, pp. 519–535.

Ewing, Reid H, Shima Hamidi, and Smart Growth America (2014). *Measuring sprawl 2014*.

Ewing, Reid, Rolf Pendall, and Don Chen (2003). "Measuring sprawl and its transportation impacts". In: *Transportation Research Record: Journal of the Transportation Research Board* 1831, pp. 175–183.

Ewing, Reid, Tom Schmid, et al. (2008). "Relationship between urban sprawl and physical activity, obesity, and morbidity". In: *Urban Ecology*. Springer, pp. 567–582.

Fan, Hongchao et al. (2014). "Quality assessment for building footprints data on OpenStreetMap". In: *International Journal of Geographical Information Science* 28.4, pp. 700–719.

Forghani, Mohammad and Mahmoud Reza Delavar (2014). "A quality study of the OpenStreetMap dataset for Tehran". In: *ISPRS International Journal of Geo-Information* 3.2, pp. 750–763.

Frumkin, Howard (2002). "Urban sprawl and public health." In: *Public health reports* 117.3, p. 201.

Frumkin, Howard, Lawrence Frank, and Richard J Jackson (2004). *Urban sprawl and public health: Designing, planning, and building for healthy communities*. Island Press.

Galster, George et al. (2001). "Wrestling sprawl to the ground: defining and measuring an elusive concept". In: *Housing policy debate* 12.4, pp. 681–717.

Geary, Robert C (1954). "The contiguity ratio and statistical mapping". In: *The incorporated statistician* 5.3, pp. 115–146.

Girres, Jean-François and Guillaume Touya (2010). "Quality assessment of the French OpenStreetMap dataset". In: *Transactions in GIS* 14.4, pp. 435–459.

Goodchild, Michael F (2007). "Citizens as sensors: web 2.0 and the volunteering of geographic information". In: *GeoFocus* 7, pp. 8–10.

Gordon, Peter and Harry W Richardson (1997). "Are compact cities a desirable planning goal?" In: *Journal of the American planning association* 63.1, pp. 95–106.

Haklay, Mordechai (2010). "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In: *Environment and planning B: Planning and design* 37.4, pp. 682–703.

Haklay, Mordechai and Patrick Weber (2008). "Openstreetmap: User-generated street maps". In: *Pervasive Computing, IEEE* 7.4, pp. 12–18.

Harvey, Robert O and William AV Clark (1965). "The nature and economics of urban sprawl". In: *Land Economics* 41.1, pp. 1–9.

Hasse, John E and Richard G Lathrop (2003). "Land resource impact indicators of urban sprawl". In: *Applied geography* 23.2, pp. 159–175.

Housing, United States. Dept. of and Urban Development (1999). *State of the Cities – 1999*. DIANE Publishing. ISBN: 9781428966482. URL: `https://books.google.fr/books?id=6oWVutUNre4C`.

Huang, Jingnan, Xi X Lu, and Jefferey M Sellers (2007). "A global comparative analysis of urban form: Applying spatial metrics and remote sensing". In: *Landscape and urban planning* 82.4, pp. 184–197.

Jaeger, Jochen AG, Rene Bertiller, Christian Schwick, Duncan Cavens, et al. (2010). "Urban permeation of landscapes and sprawl per capita: New measures of urban sprawl". In: *Ecological Indicators* 10.2, pp. 427–441.

Jaeger, Jochen AG, Rene Bertiller, Christian Schwick, and Felix Kienast (2010). "Suitability criteria for measures of urban sprawl". In: *Ecological Indicators* 10.2, pp. 397–406.

Jaeger, Jochen AG and Christian Schwick (2014). "Improving the measurement of urban sprawl: Weighted Urban Proliferation (WUP) and its application to Switzerland". In: *Ecological indicators* 38, pp. 294–308.

Jat, Mahesh Kumar, P K_ Garg, and Deepak Khare (2008). "Monitoring and modelling of urban sprawl using remote sensing and GIS techniques". In: *International journal of Applied earth Observation and Geoinformation* 10.1, pp. 26–43.

Jelinski, Dennis E and Jianguo Wu (1996). "The modifiable areal unit problem and implications for landscape ecology". In: *Landscape ecology* 11.3, pp. 129–140.

Ji, Wei et al. (2006). "Characterizing urban sprawl using multi-stage remote sensing images and landscape metrics". In: *Computers, Environment and Urban Systems* 30.6, pp. 861–879.

Johnson, Michael P (2001). "Environmental impacts of urban sprawl: a survey of the literature and proposed research agenda". In: *Environment and Planning A* 33.4, pp. 717–735.

Li, Xia and Anthony Gar-On Yeh (2004). "Analyzing spatial restructuring of land use patterns in a fast growing region using remote sensing and GIS". In: *Landscape and Urban planning* 69.4, pp. 335–354.

Lopez, Russ (2004). "Urban sprawl and risk for being overweight or obese". In: *American Journal of Public Health* 94.9, pp. 1574–1579.

Ludlow, David (2006). "Urban sprawl in Europe: the ignored challenge". In:

Malpezzi, Stephen and Wen-Kai Guo. "Measuring "sprawl": alternative measures of urban form in US metropolitan areas". In:

Masek, JG, FE Lindsay, and SN Goward (2000). "Dynamics of urban growth in the Washington DC metropolitan area, 1973-1996, from Landsat observations". In: *International Journal of Remote Sensing* 21.18, pp. 3473–3486.

Massey, Douglas S and Nancy A Denton (1988). "The dimensions of residential segregation". In: *Social forces* 67.2, pp. 281–315.

McCann, Barbara A and Reid Ewing (2003). "Measuring the health effects of sprawl: A national analysis of physical activity, obesity and chronic disease". In:

McGarigal, Kevin and Barbara J Marks (1995). "FRAGSTATS: spatial pattern analysis program for quantifying landscape structure." In:

McKee, David L and Gerald H Smith (1972). "Environmental diseconomies in suburban expansion". In: *The American Journal of Economics and Sociology* 31.2, pp. 181–188.

Mooney, Peter, Padraig Corcoran, and Adam C Winstanley (2010). "Towards quality metrics for OpenStreetMap". In: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, pp. 514–517.

Moran, Patrick AP (1950). "Notes on continuous stochastic phenomena". In: *Biometrika* 37.1/2, pp. 17–23.

Nagendra, Harini, Darla K Munroe, and Jane Southworth (2004). "From pattern to process: landscape fragmentation and the analysis of land use/land cover change". In: *Agriculture, Ecosystems & Environment* 101.2, pp. 111–115.

Nechyba, Thomas J and Randall P Walsh (2004). "Urban sprawl". In: *The Journal of Economic Perspectives* 18.4, pp. 177–200.

Neis, Pascal, Dennis Zielstra, and Alexander Zipf (2011). "The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011". In: *Future Internet* 4.1, pp. 1–21.

Nelson, A.C. and J.B. Duncan (1995). *Growth management principles and practices*. Planners Press, American Planning Association. ISBN: 9780918286925. URL: `https://books.google.fr/books?id=cmpPAAAAMAAJ`.

Openshaw, Stan (1991). "A spatial analysis research agenda". In:

O'Sullivan, David and Paul M Torrens (2001). "Cellular models of urban systems". In: *Theory and Practical Issues on Cellular Automata*. Springer, pp. 108–116.

Over, M et al. (2010). "Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany". In: *Computers, Environment and Urban Systems* 34.6, pp. 496–507.

(Pa.), 21st Century Environment Commission (1998). *Report of the Pennsylvania 21st Century Environment Commission*. Pennsylvania 21st Century Environment Commission. URL: `https://books.google.fr/books?id=J-i2NwAACAAJ`.

Parker, Dawn C et al. (2003). "Multi-agent systems for the simulation of land-use and land-cover change: a review". In: *Annals of the association of American Geographers* 93.2, pp. 314–337.

Pendall, Rolf (1999). "Do land-use controls cause sprawl?" In: *Environment and Planning B: Planning and Design* 26.4, pp. 555–571.

Portugali, Juval (2000). *Self-organization and the city*. Springer Science & Business Media.

Portugali, Juval, Itzhak Benenson, and Itzhak Omer (1997). "Spatial cognitive dissonance and sociospatial emergence in a self-organizing city". In: *Environment and Planning B: Planning and Design* 24.2, pp. 263–285.

Radeloff, Volker C, Roger B Hammer, and Susan I Stewart (2005). "Rural and suburban sprawl in the US Midwest from 1940 to 2000 and its relation to forest fragmentation". In: *Conservation biology* 19.3, pp. 793–805.

Shannon, C.E. (1948). "A Mathematical Theory of Communication". In:

Smith, D.M. (1975). *Patterns in Human Geography: An Introduction to Numerical Methods*. David & Charles. ISBN: 9780844807645. URL: `https://books.google.fr/books?id=43m1PwAACAAJ`.

Song, Yan and Gerrit-Jan Knaap (2004). "Measuring urban form: Is Portland winning the war on sprawl?" In: *Journal of the American Planning Association* 70.2, pp. 210–225.

Song, Yan, Louis Merlin, and Daniel Rodriguez (2013). "Comparing measures of urban land use mix". In: *Computers, Environment and Urban Systems* 42, pp. 1–13.

Sturm, Roland and Deborah A Cohen (2004). "Suburban sprawl and physical and mental health". In: *Public health* 118.7, pp. 488–496.

Sudhira, HS, TV Ramachandra, and KS Jagadish (2004). "Urban sprawl: metrics, dynamics and modelling using GIS". In: *International Journal of Applied Earth Observation and Geoinformation* 5.1, pp. 29–39.

Suen, I-Shian (1998). "Measuring sprawl: a quantitative study of residential development pattern in King County, Washington". PhD thesis. University of Washington.

Sutton, Paul C (2003). "A scale-adjusted measure of urban sprawl using nighttime satellite imagery". In: *Remote Sensing of Environment* 86.3, pp. 353–369.

Theil, H. (1967). *Economics and information theory*. Studies in mathematical and managerial economics. North-Holland Pub. Co. URL: `https://books.google.fr/books?id=VVNVAAAAMAAJ`.

Thomas, R.W. (1981). *Information Statistics in Geography*. Concepts and techniques in modern geography. Geo Abstracts. ISBN: 9780860940906. URL: `https://books.google.fr/books?id=jK5rPgAACAAJ`.

Torrens, Paul M (2008). "A toolkit for measuring sprawl". In: *Applied Spatial Analysis and Policy* 1.1, pp. 5–36.

Torrens, Paul M and Marina Alberti (2000). "Measuring sprawl". In:

Tsai, Yu-Hsin (2005). "Quantifying urban form: compactness versus' sprawl'". In: *Urban studies* 42.1, pp. 141–161.

Turner, Monica G et al. (1989). "Effects of changing spatial scale on the analysis of landscape pattern". In: *Landscape ecology* 3.3-4, pp. 153–162.

Waddell, Paul (2002). "UrbanSim: Modeling urban development for land use, transportation, and environmental planning". In: *Journal of the American Planning Association* 68.3, pp. 297–314.

Wassmer, Robert W (2002). "An economic perspective on urban sprawl". In: *California: California Senate Office of Research*.

White, Roger and Guy Engelen (1993). "Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns". In: *Environment and planning A* 25.8, pp. 1175–1199.

Whyte, W.H. (1958). *The Exploding Metropolis: A Study of the Assault on Urbanism and how Our Cities Can Resist it*. A Doubleday Anchor book. Doubleday. URL: `https://books.google.fr/books?id=BiMpOAAACAAJ`.

Wilson, Emily Hoffhine et al. (2003). "Development of a geospatial model to quantify, describe and map urban growth". In: *Remote sensing of environment* 86.3, pp. 275–285.

Wu, Changshan (2004). "Normalized spectral mixture analysis for monitoring urban composition using ETM+ imagery". In: *Remote Sensing of Environment* 93.4, pp. 480–492.

Xian, George and Mike Crane (2005). "Assessments of urban growth in the Tampa Bay watershed using remote sensing data". In: *Remote Sensing of Environment* 97.2, pp. 203–215.

Xiao, Jieying et al. (2006). "Evaluating urban expansion and land use change in Shijiazhuang, China, by using GIS and remote sensing". In: *Landscape and urban planning* 75.1, pp. 69–80.

Yang, Limin et al. (2003). "Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data". In: *Photogrammetric Engineering & Remote Sensing* 69.9, pp. 1003–1010.

Yeh, Anthony Gar-On and Li Xia (2001). "Measurement and monitoring of urban sprawl in a rapidly growing region using entropy". In: *Photogrammetric engineering and remote sensing* 67.1, pp. 83–90.

Yu, Xi Jun and Cho Nam Ng (2007). "Spatial and temporal dynamics of urban sprawl along two urban–rural transects: A case study of Guangzhou, China". In: *Landscape and Urban Planning* 79.1, pp. 96–109.

Yuan, Fei et al. (2005). "Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat remote sensing". In: *Remote sensing of Environment* 98.2, pp. 317–328.