Modelling Metropolitan Activity through Abductive Reasoning on Geographic Space

Salvatore F. Pileggi
Department of Computer Science
University of Auckland (New Zealand)
f.pileggi@auckland.ac.nz

Robert Amor
Department of Computer Science
University of Auckland (New Zealand)
trebor@cs.auckland.ac.nz

***Abstract*—This paper proposes a novel approach to define the geographic space focusing on the integration of common geographical specification of the space with complex semantics aimed at a more active role of the space inside information processing tasks. Generic data (called *Activity*) is processed in the space model in order to retrieve a well defined behaviour of the interest parameters on the target spaces. A domain-specific semantic understanding of the whole data ecosystem allows one to overcome of unrealistic assumptions to switch to an effective reasoning on the space.**

***Keywords*—Behaviour Modelling, Semantic Reasoning, Geographic Information Systems, Big Data.**

I. Introduction

The major advances of the last few years in Information and Communication Technology (ICT) have radically changed the view and the understanding of most systems, as well as their applications inside the Information Society (IS). The unstoppable tendency towards Big Data [1], that exceed the processing capacity of conventional database systems or do not fit the structures of common database architectures, implies the need of alternative ways to process information in order to gain value. Apart from the problem of size, due to availability and popularity of sources (e.g. Social Media, Social Networks) and the great number of prosumers, data often move and change too fast. Furthermore, the heterogeneity of the information, both with the intrinsic complexity of certain content (e.g. Social Object [2]), implies in most cases a contextual meaning of data. Consolidated high performance techniques for the creation, distribution, use, integration and manipulation of information on specific sets of data are not adequate for this new evolving context. In function of the subjects, objects, problems and scopes, researchers are working on specific solutions in order to dynamically accommodate large sets of requirements. Data associated to large scale environments (e.g. cities) implicitly reflect complex social dynamics as well as significant economic, political, and cultural activities. Cities are the places where people live, work, spend most of their life and perform everyday activities, as individual and as part of a community. The integration of data and space models (GIS) provides a global data environment where the information can be associated to the space. GIS are being affected by the Big Data revolution as well as by other technological trends [3]. Common GIS models, mostly reflecting only a physical view of the space, are useful to support basic operations (e.g. geographical filtering or aggregation) though with limitations (e.g. interoperability [4]) but, due to the intrinsic limitations of the physical view of the space, play a poor or passive role in the critical phases of the data' lifecycle (e.g. analysis). The processing of data [5] is the key factor, inside IS, to gain competitiveness by using Information Technology (IT) in a creative and productive way aimed at the knowledge economy model. This paper proposes a novel approach to define the space that focuses on the integration of common physical views of the space with complex semantics aimed at a more active role of the space inside information processing tasks. More concretely, generic data (called *Activity*) is processed with the space model in order to retrieve a well defined behaviour of the interest parameters on the target spaces. Evidently, this kind of process can significantly vary in function of the target domain/application, as well as in function of the considered parameters. In this context, the activity is resulting from light processing of large scale data streams and it is considered exclusively from a quantitative point of view [6]. A qualitative analysis is interesting, and the object of ongoing studies, but is out of the scope of the paper. In this study, the nature of the parameters of objects of study does not affect the proposed model or the techniques for the analysis and the processing of the information. So just the most generic semantics is associated to the concept of activity in the style of "*something is happening there*", "*people are there doing something*" or "*this place is popular*" in a certain period of time under consideration. A domain-specific semantic understanding of a whole data ecosystem allows one to overcome unrealistic assumptions (e.g. the homogeneity of the space) to switch to an effective semantic reasoning on the space. The impact of the space inside data processing tasks can be significant and can have a critical role when, as in the case of Big Data, great amounts of implicit information have to be converted into explicit intelligible knowledge.

II. Modelling Metropolitan Activity

Information processing on a large scale is a knowledge intensive set of tasks that applies complex techniques in order to generate the expected outcomes. In most cases, in order to assure their effectiveness and correctness, those processes require human-like processing of the data (or information). Sometimes the expectations about certain processes is a real challenge since it is implicitly assumed that a machine processes data better than humans would do, under the completely unrealistic assumption of human-scale for the information, complex data can be ambiguous for humans too, including when the context of the information is defined. Apart from those intrinsic limitations, the fact that big data is not always better data [6] is often not emphasized enough: large scale information allows an innumerable set of advantages (such as statistical analysis [7]) but introduces several problems (lack of accuracy and reliability for example [8]). Two different simple schema for modelling metropolitan activities on the space level are shown in fig. 1. They propose a similar conceptual structure since provide their output on the base of the same input model that includes:

- *Asserted Data*. These are the data/information to process. Asserted data can have a completely different meaning and function due to the context in which they are considered. For example, they could be synonymous of available or reliable data, the outcomes of some processing over big data, as well as any other fact input to the system.
- *Space Model*. This is the context to which asserted data are associated and in which asserted data are processed. The nature of this model determines the two different schemas for data/information processing considered in the paper. A space model that exclusively reflects a physical view of the space (fig. 1, left) assumes a data infrastructure in which data is associated to the space according to some logic on the model of a GIS. Filtering, federations and more in general geographic-aware computation (e.g.[9]) is well supported. Overcoming the physical view of the space and integrating it with a semantic perspective of the space[10] allows one to switch from a data infrastructure to an ecosystem of data: subjects and objects are defined according to a semantic approach including semantic properties as well as the relations among the concepts composing the model. The processing schema (fig. 1, right) is strongly affected because the extended capabilities in terms of expressivity can have direct and critical implications on the process itself (e.g. semantic reasoning[11]).
- *Rules*. Both data and information processing assume some kind of manipulation or change of the input to achieve some knowledge or goal. This process is normally driven by an extensive set of rules.

Whichever parameter of interest related to a certain space $y_{i \in S}(t)$ can be *asserted* or *not asserted*. In the first case, it is already one of the output for the process; in the second case it has to be calculated or inferred by the process. Modelling metropolitan activity mostly consists of defining the rules to deduce or calculate the values for not asserted activity parameters from the asserted ones. They have to be processed (fig. 1) assuming as input the asserted data in the context of the space they are related to.

A. Mathematical approach: Assuming the homogeneity of the space

The mathematical approach is based on a strictly physical view of the space. For instance, the space is a discrete environment that can be described at different levels of abstraction and detail. According to this approach, the specification of a space is exhaustively provided by:

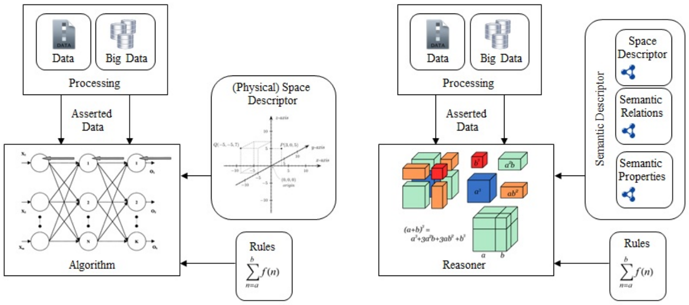

Fig. 1. Mathematical (left) and semantic (right) approach.

- *Space Partitioning*. The *Domain Space* (*S*) is assumed to be composed of a number of *subspaces* or *containers*. A-priori there are no constraints (e.g. disjointness) but specific applications can introduce some of them. Even though the semantics of the defined space are not specified, by adopting this approach spaces should have the same level of abstraction (e.g. districts or places). Mixed composition of spaces are evidently possible but, because of the simplicity of the space model, could resulting confusion and, so, are quite hard to match concrete scenarios. Examples of partitioning are showed in fig. 2 (panel 1,2 and 4).
- *Neighbour (N)*. Once the spaces composing *S* have been defined, they can be related to each other according to some relation. Since the mathematical approach reflects a physical view of the space, the context of a space is defined according to a position-based logic that just expresses the physical proximity among spaces. An example is shown in fig. 2 (panel 1).

Summarizing, due to the lack of specific semantics, the view of a space is simple and can be defined as in eq. 1. y_s is the activity parameter associated to the space *s*.

$$s \in S \quad \Rightarrow s(N(s,k), y_s), \quad k \in S \tag{1}$$

Each non-asserted activity value can be calculated according to eq. 2 which, due to the lack of semantics, represents the only set of rules applicable to the whole space domain *S*. Each related space provides an independent contribution (eq. 2).

$$y_i(t) = y_k - \alpha(x,t) \exists N(i,k) \quad \wedge \quad y_i \notin Asserted(y) \tag{2}$$

The overall activity value is obtained by collecting *k* single contributions according to the eq. 3. *w* parameters are used to allow different weights for the different contributions.

$$y_i(t) = \frac{\sum_k w_k (y_k - \alpha(x,t))}{k} \tag{3}$$
$$\exists N(i,k) \quad \wedge \quad y_i \notin Asserted(y), \quad w_k = const$$

Eq. 3 is implemented by alg. 1. This simple algorithm processes the context information for each space starting from the spaces associated with asserted data (that cannot be

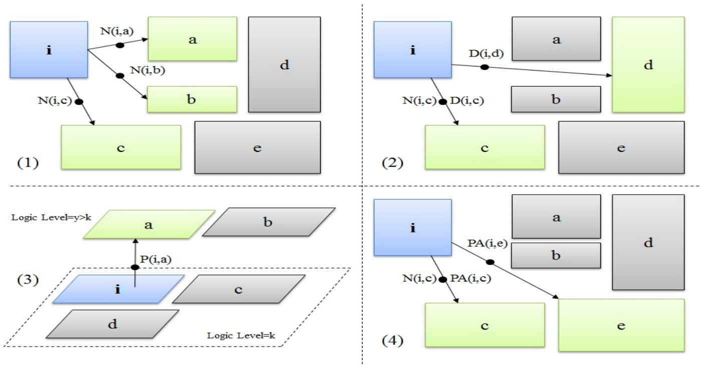

Fig. 2. Examples of relations.

modified in this version of the algorithm). The function α is associated with the *continuity* of the space: lower values of α determine soft changes in the activity parameter for contiguous spaces (high continuity); on the contrary, high values mean strong changes among contiguous spaces (low continuity). In the context of this work the α function is simplified and defined as positive constant value.

Algorithm 1 Processing Rules

```
$Rules_S \quad is \quad eq(\ 2)$
**for** $each \quad i \in S$ **do**
    $RULES_S$
**end for**
```

Regardless of the complexity of α, the greater limitation of the mathematical approach resides in the model itself: focusing exclusively on a physical view of the space intrinsically implies the assumption of full homogeneity for the target spaces. In other words, the model, due to the poor context information, does not allow one to distinguish among spaces that are considered all the same from a computational point of view even though they are not the same in reality. As discussed in the next section, this results in poor capabilities in terms of analysis.

B. Semantic approach: Reasoning on the Space

The Semantic approach extends the previous one by introducing further capabilities in terms of representation and analysis. More concretely, the semantic understanding of the space integrates previous concepts with the following:

- *Extended Space Partitioning.* It is assumed the whole space is composed of multiple layers or levels of abstraction. A space can be associated with a concrete level of abstraction through the parameter *la*. Two different kind of space (with two different levels of abstraction) are considered in this study case: the *Public Space* (*pa*=0) and the *District* (*pa*=1). An example of extended space partitioning is shown in fig. 2 (panel 3).
- *Dependent (D).* This relation is used to specify a dependence of a space with another. It is often confused with the N relation (as previously described) even though it has a completely different nature and semantics: there is no relation at all between *D* and physical proximity even though it is quite common to have a dependency amongst close spaces. Furthermore, in contrast to *N*, *D* is not a bilateral relation. An example is showed in fig. 2 (panel 2).
- *Parent (P).* This relation establishes a logic link among spaces associated with different logic levels (from a lower to a higher one). For example, public spaces can be related to their districts through this relation. An example is showed in fig. 2 (panel 3). The parent relation is, for instance, a logic relation: it can reflect a physical inclusion (the lower space being physically in the higher one) but it has not to have. In fact it is common to define logic containers that have no physical relation with the spaces included but semantic relations established by some property or parameter. Evidently the relation is not reversible.
- *Other Relations.* The semantic relations included in a specific model are usual to be considered domain/application specific. However there are at least two relations that are generic enough to be included in most models: *Pair (PA)* and *Inclusion (I)*. The first one (*PA*) establishes a semantic similitude between two spaces on the basis of a concrete set of properties or parameters. An example of pair spaces is showed in fig. 2 (panel 4). It plays a critical role when processing semantic properties (out of the scope of the paper). *Inclusion* expresses a physical inclusion of a space (normally at a lower level) with another one. It is used to provide details inside spaces or complex compositions of them. These two relations are important to have a picture of the whole model but have a minor role in the use case proposed in the paper.
- *Semantic Properties (SP).* These are used in order to characterize the different spaces according to a certain classification or behaviour. The use cases shown in the paper focus on semantic relations. In real applications, properties play a role similar to relations since they provide a further set of input for the reasoners. Apart from their use, the most relevant difference at the level of model between properties and relations consists in the fact that properties are normally associated to a single space and relations usual involve two or more spaces (and eventually properties).

The whole set of relations for *S* defines the *Semantic Relations (SR)*. Semantic Properties are generically referred as a set *SP* of properties associated with the considered space. According to this approach the vision at a single space is extended as in eq. 4.

$$s \in S \quad \Rightarrow s(pa_s, SR(s,\_), SP(s), y_s) \tag{4}$$

The reasoner processes the context information for each space starting from spaces associated with asserted data (that cannot be modified) exactly as the algorithm previously proposed (alg. 2). But it can reason on semantic relations and apply different rules and function for semantic matchings. That is a strong added value in the context of a complex environment like the space. In this paper, two different sets of rules are considered for the semantic approach. The first one is eq. 2 changing N for D. The second one is defined as in eq. 5 and 6.

$$\begin{aligned} y_i(t) = y_k - \beta(x,t)\alpha(x,t) \\ \exists D(i,k) \quad \wedge \quad y_i \notin Asserted(y) \end{aligned} \tag{5}$$

$$\begin{aligned} y_i(t) = \frac{\sum_k w_k(y_k - \beta(x,t)\alpha(x,t))}{k} \\ \exists D(i,k) \quad \wedge \quad y_i \notin Asserted(y), \quad w_k = const \end{aligned} \tag{6}$$

β is a parameter associated with the *homogeneity* of upper spaces. It reflects an estimation of how much the spaces related (through *P*) are semantically similar. It is strongly in contrast with the previous approach that implicitly assumes a whole and homogeneous upper space. In this context the β function is simplified and defined as a constant value. This is a strong difference with respect to the physical view that assumes the logic space is a whole and, consequently, the definition of static patterns. An example of semantic matching is provided by alg. 3. It allows the reasoner to understand if it is processing spaces inside the same upper container eventually applying a different set of rules.

Algorithm 2 Reasoning on Space

```
$Rules_a \quad is \quad eq(\ 2), \quad N \leftarrow D$
$Rules_b \quad is \quad eq(\ 5)$
for $each \quad i \in S$ do
    if $Eval(SR(i,\_), C)$ then
        $RULES_a$
    else
        $RULES_b$
    end if
end for
```

Algorithm 3 An example of Semantic Matching

```
$Function \quad Eval$ $(SR(i,k), C)$: Boolean
$C_1 \quad \leftarrow \quad \exists P(i,p), \quad p \in S$
$C_2 \quad \leftarrow \quad \exists P(k,p), \quad p \in S$
if $\exists p : (C_1 \wedge C_2)$ then
    return true
else
    return false
end if
$EndFunction$
```

III. An application scenario: Public Spaces inside Metropolitan Ecosystems

In this section the model is applied to a simple scenario resulting from composing public spaces inside metropolitan ecosystems. A public space is a social space that is generally open and accessible to people. A low scale example is proposed: a set of spaces (*pa*=0) is defined in the context of a simple understanding of a metropolitan area (*pa*=2) including districts (*pa*=1) in which spaces are defined (or are associated to). A physical view of the space (as previously defined in the paper) is proposed as an application of common techniques. It is compared with the corresponding approach by using semantic reasoning on the space. A physical analysis of the space inside a semantic reasoner can be obtained simply assuming $\beta = 1$. Six different scenarios, characterized by increasing complexity, are proposed in fig. 3. Two asserted data (y_A=8 and y_G=10) respectively for spaces *A* and *G* are provided as input for the reasoner. Fig. 4 proposes a comparison between the distributions obtained by applying the two approaches to those scenarios, assuming $\beta = 2$ for the semantic approach. As shown, the two techniques produce exactly the same results for the first scenario. That is because of its simplicity (fig. 3): spaces are defined all in the context of the same district, so the reasoner is unable to recognize and process the context of the information. In this kind of scenario, the physical and semantic view of the information are the same. Once the complexity of the context is progressively introduced (from scenario 2 to 6), the capabilities of a physical perspective are evidently limited and the information is processed by an actor that is looking at the ecosystem as a simple "observer": it looks at the space, it can see but cannot understand what is is watching. On the contrary, the extended analysis provided by the semantic approach allows an extensive understanding of the space and its context. Consequently, the reasoner can process the data according to the semantics of the target space providing a context-aware output. In fig. 5 average values of the distributions are considered as a function of the parameter β that can vary inside a range of values. Apart from the already mentioned limitations of the physical approach ($\beta = 1$), detected patterns propose a regular decrease: reasoners are able to detect different districts and distribute values among neighbour spaces according to the homogeneity (β) associated to the upper spaces. Apart from scenario 1 (already discussed), there is another clear exception to the dominant pattern (scenario 4). In this paper we implicitly assumed the complexity of a scenario as directly related to the number of spaces and relations defined. This is a good approximation, acceptable in most case studies, but not a proper formal pattern. In fact, according to this view of the complexity of the space, scenario 4 should propose average values between the scenario 3 and the scenario 5. But the obtained values are mostly in line with the pattern of scenario 3 (a bit higher in the considered range). This is because (fig. 3) scenario 4 is directly derived from scenario 3 introducing the space *H* and its context. This space is not really adding complexity to the overall space but it is just extending the space in one direction. So the expected relaxing of the values of distributions is not detected. On the contrary, there is a little contraction of values: the difference of pattern between scenarios 3 and 4 is mostly explained by the relations of the space *H* that is directly dependent on the space *B* and influences the space *E*. The previously proposed scenarios represents the progressive characterization of a real environment corresponding to a small section of the city of Auckland (New Zealand). Due to the generality of the relations adopted to semantically describe the space, the resulting perspective has not a specific focus and mostly reflects the point of view of a generic observer (e.g. a citizen). The public spaces of the scenario 6 are showed in fig. 6. Public spaces are painted by using different colors in function of the activity associated to each of them: the gray is associated to a no available activity value; yellow/orange tones indicate a low activity, as well as the red scale marks medium (light red) or high (dark red) activity. The map up in the figure shows

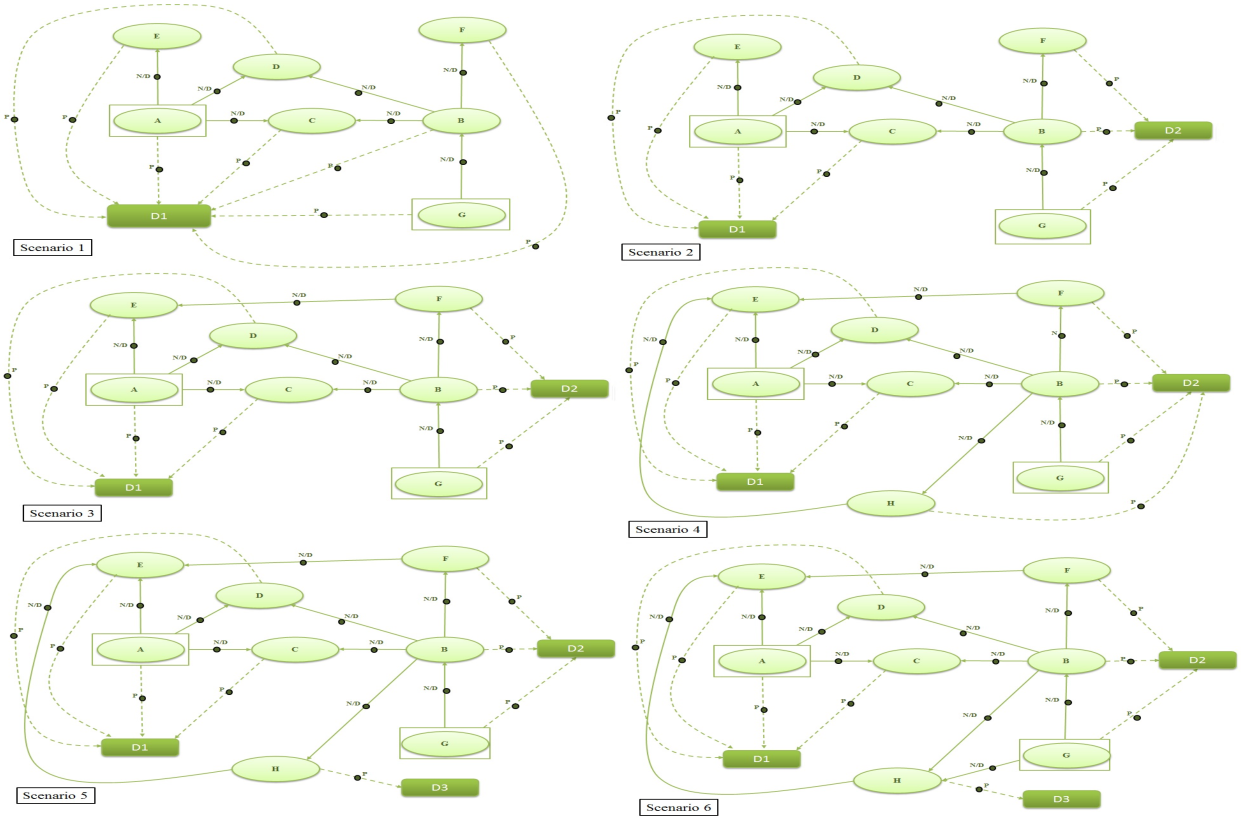

Fig. 3. Scenarios.

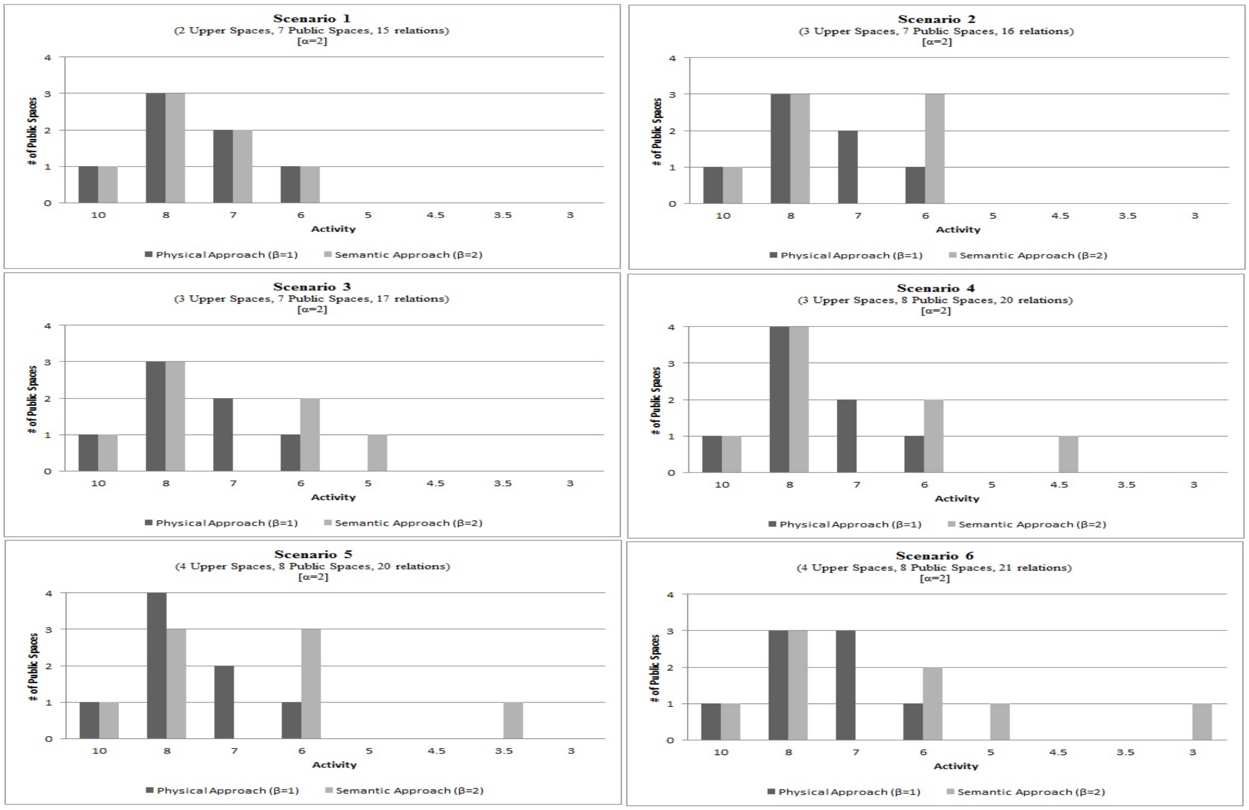

Fig. 4. A quantitative comparison between views.

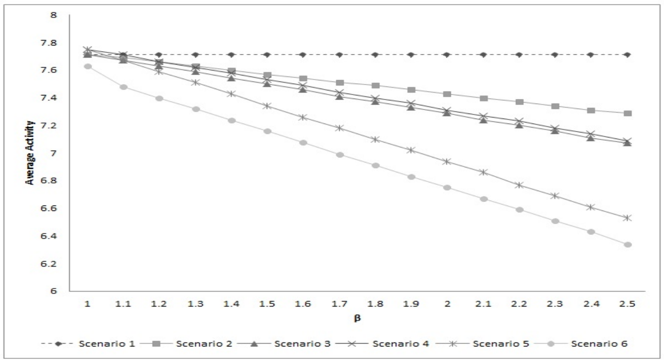

Fig. 5. Average values of distributions in function of β.

the input to the reasoner. This input represents a quantitative estimation of the activity according to complex data sources (social networks in this case). As shown in the picture is not possible to associate a significant activity value for all the spaces, even when they are very central and popular places (as in this case). The lack of information does not mean poor activity. More realistically, the information is implicit and strongly distributed inside contents. In fact, the input picture is not realistic at all, though it is provided by using real data. The same approach is followed for low level activities detected. They could be reliable results but it's also possible that most relevant information was not explained by using direct data processing. Only medium/high activity detected by direct observations are considered as assertions (input data) for the reasoner. Details about the technique used to process social data are out of the scope of the paper. The output of the reasoner is shown in fig. 6 for both the mathematical (bottom left) and the semantic (bottom right) approach. The capabilities of the reasoner to understand the context information are extremely limited if only the physical space is processed. As shown in the picture, the reasoner propagates the activity in the whole space according to a linearly decreasing pattern. On the contrary, the semantic approach allows the reasoner to propagate based on a effective analysis of the context space.

IV. Main Limitations

The effective application of this approach mostly depends on the capability to define a domain-specific view of the physical space in a formal instance of the model (including the reference vocabulary, semantic relations and properties) as well as the rules to process it. This assumption can be considered as fully realistic in most cases where domain specialists are involved in the process. This is, for example, the case of social studies aimed at the understanding and evaluation of complex dynamics and behaviours.

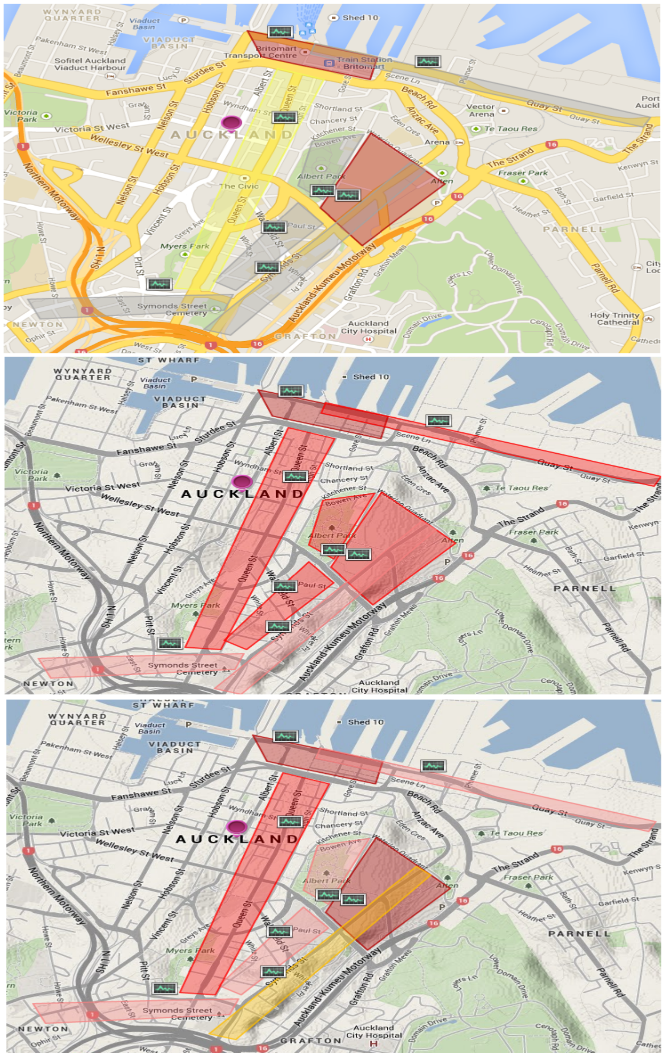

Fig. 6. Real data input on scenario 6 (top). Output by using a mathematical (bottom left) and a semantic (bottom right) approach.

References

[1] Michael Katina and Keith W. Miller, "Big data: New opportunities and new challenges," *IEEE Computer*, vol. 46, no. 6, pp. 22–24, 2013.

[2] Salvatore F. Pileggi, Carlos Fernandez-Llatas, and Vicente Traver, "When the social meets the semantic: Social semantic web or web 2.5," *Future Internet*, vol. 4, no. 3, pp. 852–864, 2012.

[3] Salvatore F. Pileggi and Robert Amor, "Addressing semantic geographic information systems," *Future Internet*, vol. 5, no. 4, pp. 585–590, 2013.

[4] MiguelngeI Manso and Monica Wachowicz, "Gis design: A review of current issues in interoperability," *Geography Compass*, vol. 3, no. 3, pp. 1105–1124, 2009.

[5] Christian Bizer, Peter A. Boncz, Michael L. Brodie, and Orri Erling, "The meaningful use of big data: four perspectives - four challenges," *SIGMOD Record*, vol. 40, no. 4, pp. 56–60, 2011.

[6] Danah Boyd and Kate Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication and Society*, vol. 15, no. 3, pp. 662–679, 2012.

[7] Bruce Ratner, *Statistical modeling and analysis for database marketing: effective techniques for mining big data*, CRC Press, 2003.

[8] Iman Saleh Wei Tan, M. Brian Blake and Schahram Dustdar, "Social-network-sourced big data analytics," *Internet Computing, IEEE*, vol. 17, no. 5, pp. 62–69, 2013.

[9] Wei Gao Jiming Chen Jialu Fan, Yuan Du and Youxian Sun, "Geography-aware active data dissemination in mobile social networks," in *2010 IEEE 7th International Conference on Mobile Adhoc and Sensor Systems (MASS)*, 2010.

[10] Salvatore F. Pileggi and Robert Amor, "Mansion-gs: semantics as the n-th dimension for geographic space," in *International Conference on Information Resource Management (Conf-IRM 2014)*, 2014.

[11] Luciano Serafini and Andrei Tamilin, "Drago: Distributed reasoning architecture for the semantic web," in *ESWC*, 2005, pp. 361–376.