Exploring Budgeted Learning for Data-Driven Semantic Inference via Urban Functions

CHIDUBEM IDDIANOZIE, MICHELA BERTOLOTTO, AND GAVIN MCARDLE
School of Computer Science, University College Dublin, Dublin 4, Ireland
Corresponding author: Chidubem Iddianozie (chidubem.iddianozie@ucdconnect.ie)

ABSTRACT The performance of a machine learning algorithm is dependent on the quality of the available data for model development. However, in practical situations, the availability of the data is variable and can be limited. This limitation creates a budget problem for data-driven techniques and the objective in such situations is to develop the best model given the available data. In this article, we examine the budgeted learning problem for spatial data within the urban context. We demonstrate the effectiveness of a novel approach for inferring the attributes of spatial data when the data for the model is budgeted. This is achieved using urban functions - which describe the designated use of a geographical space - to infer the types of streets in a city. We evaluated the approach by comparing the performance of the model using the data in each urban function (the budget) against the results from the aggregate of all the functions (all data). The results indicate that with our model, individual urban functions are sufficient to infer the type attributes of streets.

INDEX TERMS Budgeted learning, machine learning, semantic inference, urban functions, OpenStreetMap, data quality.

I. INTRODUCTION

Typically, machine learning algorithms assume that the data which is necessary for building an accurate model is readily available at time of model development [1]. However, this is not always the case. For example, certain data might be unavailable, too expensive or just too difficult to collect. Furthermore, even if the model has access to all the data, it could be the case that not all the data will be actually useful to the model and its inclusion may reduce the model quality and inadvertently increase computational costs. Hence, the general problem is, given the data available, how can one determine the optimal subset of the data that will produce the best model? This is called a budgeted learning problem [1], [2].

A branch of data mining that has seen very practical applications is semantic inference which seeks to autonomously infer or predict the meaning of words and data. In the spatial domain, this field is largely explored using data-driven techniques to infer the descriptive elements of spatial objects which are called attributes. For example, a building could have factory as one of its attributes. Popular applications of semantic inference in the spatial domain include improving data quality [3], information retrieval [4], ontology learning [5], [6] among others. In this article, we focus on improving the data quality of urban street data using data-driven techniques.

It can be argued that the increased adoption of data-driven techniques for semantic inference is not unrelated to the spatial auto-correlation or dependency that exists between spatial objects [7]–[9]. To illustrate, it is unlikely to find a residential street close to an object attributed as a factory. This suggests that a model will be able to predict that a street is not a residential street if there are factory attributed objects close to the street. Such information about the spatial context in which the street object exists is called contextual information. However, a problem associated with embedding contextual information in machine learning and inference models is that there is a risk of using too much information which could inadvertently lead to reduced model quality as well as increased computational costs. Towards mitigating this, a naive approach may be a brute-force method which explores all possible combinations of the contextual information to determine the optimal set of objects and attributes of the data to learn from. However, this might be technically impracticable. There are hundreds, possibly thousands of
objects that exist in and across different spatial domains. Even if it were technically possible to determine the optimal set of objects and their attributes in reasonable time, there is the additional problem of guaranteeing that they will be available during model development. Thus, the problem is to determine the optimal set of objects and attributes that can produce the best model, given the available data and computational resources - a budget problem [2]. In this paper, we focus on the budget problem raised by data availability. Now, a peculiar and interesting characteristic of spatial objects is that they tend to be auto-correlated and are dependent on each other [7]–[10]. Consequently, this suggests that certain objects and attributes may be better at inferring the attributes of certain objects than others and adopting them could improve model performance and also save computational costs. Hence, this could offer a solution to the budget problem that exists in semantic inference.

In this article, we propose a solution to the budget problem by exploiting the notion of urban functions (See Fig. 1). Urban functions describe the designated use of a geographical space. By using nearby urban functions as contextual information for an object we can infer information about that object. As a use-case, we attempt to contribute to addressing the semantic data quality problem that exists in OpenStreetMap (OSM) using data-driven techniques. OSM is a crowd-sourced database that suffers from poor data quality [11]–[14] in some locations and addressing the quality of OSM data is still an open and relevant research problem, we discuss this in section II-A. We selected six cities from OSM and attempted to infer four distinct values for the attribute describing the type of street: tertiary, primary, secondary and residential. Our justification for choosing these particular street types is explained in section IV-A. We used the contextual information that exists for each street as the description of the street (in this case study, the contextual information is the objects, and their attributes, surrounding a street). This is analogous to features of a model in Machine Learning terminology. Having defined the problem formally, our methodology considers four urban functions: recreational, residential, utilities and commercial. We developed models for two cases (1) using each of the urban functions and (2) their aggregate. Our results show an overall mean F-score of 60% and a mean differential of $\approx 1.2\%$ between both cases. This suggests a benefit in adopting urban functions to solve the budget problem in data-driven semantic inference. To the best of our knowledge, this is the first attempt to formulate and propose a solution to the budget problem that exists in data-driven semantic inference. The approach we describe in this article is very practical and can be used in other areas of spatial data mining to address model optimization. The main contributions of this article are two-fold:

1) We formulate the budgeted learning problem for data-driven semantic inference and address it using the concept of urban functions.

2) We demonstrate a data-driven solution towards improving the semantic data quality of street types in OSM.

The remainder of this article is organized as follows: In section II, we provide a detailed description of the problem domain and review related work. The problem is formulated and defined in section III. Section IV outlines our methodology. The results are presented and discussed in section V. We make our conclusions and recommendations for future work in section VI.

II. BACKGROUND

A. BUDGETED LEARNING AND SPATIAL SEMANTIC INFERENCE

It is a common assumption that the data which is necessary to develop a machine learning model is readily available, however this is not always the case [1]. Depending on the situation, it could be the attribute values or other class labels that are missing. Budgeted learning is concerned with the lack of attribute values describing an object. The budgeted learning problem arises in a hypothetical situation where a learner $L$ is provided with the class labels but some or all of the attribute values are not given. In this case, the learner has to acquire the attribute values at a cost $C = \{c_1, \ldots, c_n\}$. This problem is defined formally as follows: Given a budget $B \in \mathbb{N}$ and an instance $I = (X, Y)$. Where $X = \{x_1, \ldots, x_n\}$ and $Y$ is a finite set of class labels, a learner $L$ can acquire a feature $x_i$ at a cost $c(x_i) \in C_i$ iff $C_i \leq B$. The objective of the learner is to produce the best model within the budget provided. Perhaps, the first formal definition of budget learning problem was by Lizotte et al. [2]. The authors formalized this problem in the context of maximizing the success of clinical trials within a budget. Subsequently, their work has been extended.
to fields such as mobile computing, advertisement placement optimization and learning with crowd-sourced data [1], [15], [16]. While the problems in these fields are peculiar to them as well as their approaches, their similarity lies in the fact that they all seek to determine and exploit the best set of features that maximizes reward. Like most learning problems, the budgeted learning problem is characterised by exploration and exploitation. We discuss both characteristics in the context of budgeted learning. Exploration is concerned with determining the set of features (in our case, the features are the attributes of the spatial objects in proximity to a street which is considered as the context of the street) which produce the best models. However, at the start of model development, the learner has no idea what these attributes are and has to discover them. Exploitation deals with acquiring more of those attributes in order to maximize reward and by effect achieve the objective. Now, given that there could be an infinite number of features, the problem is to find a trade-off between exploration and exploitation such that the features used are representative of the problem domain while at the same time producing the best model.

Semantic inference is a branch of geospatial artificial intelligence (geoAI) and has been beneficial towards tackling problems such as traffic management, disease spread, improving data quality among others [17]. In the case of improving data quality, it has been used to enrich the semantic quality of OSM. OSM is the most successful example of Volunteered Geographical Information (VGI). This is an open geo-spatial database where users can freely make changes to objects. This project has been acclaimed to provide a more complete or up-to-date view of a place than authoritative data-sets [12]. It has also been invaluable in scenarios such as disaster management [18], [19]. However, the open nature of this project calls into question the quality of the data contained therein, with respect to accuracy and completeness [11], [13]. Nonetheless, research has shown that OSM data is of very high quality in urban centres [11], [13], [14], [20]. This would explain why companies such as Facebook and Strava use OSM to power some of their services [21]. That said, the problem now is such that the quality of the data in OSM is non-uniform across different spatial domains. However, if models can be trained to infer the semantics of a spatial object, then the quality of OSM data can be improved. This fact has encouraged some works in this direction.

One of the first practical attempts was by Jilani et al. [22], [23], where the authors inferred all the street type semantics of OSM using machine learning methods. More recently, Iddianozie and McArdle successfully extended this attempt by building transferable models [24]. While the results from both works were impressive, neither of them used contextual information to build their models. Contextual information refers to the knowledge about the context in which an object exists. Given that spatial data exhibits heterogeneity and dependence with respect to space, we argue that the use of spatial context may improve the models. For example, knowing that factories are rarely found on residential streets is valuable information for inferring street types. However, there are some issues that are raised as a result of using contextual information. Firstly, there could be an extremely large number of objects around an object to be inferred. This is problematic, because for machine learning algorithms to produce acceptable results, they require the data for model development to be available to the machine [1], [2]. This raises an issue that is particularly relevant to the spatial domain, as the availability of objects vary from domain to domain. This is further amplified when you consider the issue of data completeness associated with crowd-sourced databases such as OSM, especially from the perspective of transferable models [24]–[27]. Secondly, even if all the data were available, it is the case that computational time and space are limited and not all objects will be beneficial to the model. Hence, the problem will be to determine the best objects for the model. Therefore, we can say that the use of contextual information for data-driven semantic inference raises a budget problem.

B. THE NATURE OF CITIES, URBAN FUNCTIONS AND THE OPENSSTREETMAP CASE

Cities are one of the most complex man-made systems, exhibiting characteristics such as non-linearity, feedback loops, emergence, dependence and constant evolution [28]–[30]. Scott and Storper [31] posited that an attempt to propose a single all-encompassing concept of the city can be futile or misleading. However, the nature of cities can be described or understood using the notion of urban functions. Urban functions refer to the designated and recognized use of a place [32], [33]. For example, an apartment building and a playground have a residential and recreational function respectively. Urban functions exhibit spatial inter-dependence and coherence [34], [35], which is a derivative of the complex nature of cities, brought about by the interaction between objects [28].

In OSM, the semantic description of objects are denoted using tags, where every tag has a key and a value [36]. An example is a tag key: building which has church, apartments and school as some of its tag values. Though the attributes encoded in the tag values of the different tag keys vary, the tag values can be grouped into urban functions (see Fig. 2). For example, a detached house and a movie theatre belongs to the residential and recreational urban functions respectively. However, certain objects could exhibit dual-functions, for example the movie theatre could be grouped under a commercial function because it generates income. In this article, we assume the primary function of every spatial object to be the only function.

III. GENERAL PROBLEM STATEMENT

Given that data-driven methods have proven to be beneficial for inferring spatial semantics, we consider a multi-class classification problem where we need to infer the types $Y = \{y_1, \ldots, y_k\}$ of streets in a domain $D$ (or city) based on
and variable, and recognizing that there exists a budget $B$ street. Thus, we propose to minimize Equation 1 as follows, the objective is to build $n$ number of streets in the city. Given that in the attribute vector $s$, semantic inference of road types is possible when imposing of all data). Thus, in our experiments we seek to show that is the extremely large attribute vector for the $i$th street in the set of streets $S = \{s_1, \ldots, s_d\}$ which is an extremely large $n$–dimensional attribute vector. In our case, contextual information is the type of buildings and amenities around a given street.

We wish to build a learner $L(S) = Y$ which learns a mapping between the street types (Y) and the contextual information ($s_i$) for every street. Hence, the cost $C$ of building $L$ can be expressed as follows:

$$C(L) = \sum_{s=1}^{d} \left( \sum_{i=1}^{n} c_i(x_i) \right)$$  

subject to $C(L) \leq B$. $c_i(x_i)$ is the cost of acquiring an element in the attribute vector $s_i$ (for a single street) and $d$ is the number of streets in the city. Given that $n$ is extremely large and variable, and recognizing that there exists a budget $B \in \mathbb{N}$, the objective is to build $L$ using $S_{red}$, where $S_{red} \subseteq S_i$. Recall that $s_i$ is the extremely large attribute vector for the $i$th street. Thus, we propose to minimize Equation 1 as follows:

$$C^*(L) = C(\Phi(L))$$

Here, $\Phi$ is a function that produces a reduced version of $s_i$ as $s_{rel}$, where the size of $s_{rel}$ is significantly smaller than $n$. In this article, we perform this reduction by using four urban functions. Here, the original set ($s_i$) is split into four sets which each correspond to the defined urban functions.

Now, instead of acquiring all elements of $s_i$, we hypothesize that $L$ can be built with acceptable results using only objects from a single urban function.

In this article, we associate a unit cost for the use of each $x_i$, we recognize a budget $B$ which is less than $C(L)$ (the cost of all data). Thus, in our experiments we seek to show that semantic inference of road types is possible when imposing a budget via the urban functions. Our experiment to infer and label road types considers two cases: a budgeted case (using a single urban function per inference) and an unconstrained one (the aggregate of the urban functions). We present and discuss the results of these experiments in section V.

**IV. METHODOLOGY**

We have determined that a budget problem exists and formulated the problem as described in section III. Also, we have made the assumption that only a finite set of features can be correlated to any class label and as such provide the best information about that label. In this paper, a class label refers to a street type. Our goal in this article is to demonstrate that given a finite set of labels $Y = \{y_1, \ldots, y_k\}$ and a very limited budget $B \in \mathbb{N}$, a learner model can produce acceptable results using a subset of a variable but extremely large vector ($s_i = \{x_1, \ldots, x_n\}$) which holds the contextual information for a street in the set of streets $S = \{s_1, \ldots, s_d\}$ (See Equation 2).

The experimental set-up we designed to achieve our goal is made up of two key components of the methodology: the input data which is the contextual information for road types and the algorithm to build $L$, the learner. We discuss the components of the experiment in this section.

**A. DATA**

The data we use in this article are collected from OSM. We consider six cities from two continents and four countries. The cities are Chicago, Munich, Frankfurt, Manchester, Rome and Charlotte. We decided to use multiple cities because we believe that for our proposed solution to be practical, we have to demonstrate how the solution scales across different spatial domains. Furthermore, some of the difficulties associated with working with spatial data, especially with regards to knowledge discovery are related to heterogeneity and uncertainty [9], [10], [35] and these are encountered using this multi-city approach. Hence, we are of the opinion that this should be the standard for approaches that seek to develop methods that can be used across multiple domains. It will be beneficial to show how theoretical or empirical approaches scale to different domains. To the best of our knowledge, we are the first to tackle data-driven semantic inference, specifically with regards to improving data quality at this scale. We selected cities from urban areas in Europe and North America as research has shown that OSM is sufficiently accurate in these areas [11], [20]. The data we collected can be divided into two categories: street networks and contextual information. We discuss both in the next sections.

**1) STREET NETWORKS**

The street networks in this study represent drivable streets from the cities under consideration. OSM recognizes other types of streets such as footpaths, cycle paths among others. However, the drivable streets are more heterogeneous in nature than the other types of streets. For example, a drivable street can be a primary or a secondary street but a footpath has no subdivisions. This characteristic of drivable streets suggests that they are more susceptible to suffer from
poor semantic quality on OSM as contributors may struggle to perform correct sub-classifications. Furthermore, drivable streets rank higher in the hierarchy of streets in most cities and since they form a large portion of the data available on OSM, a solution to improve its quality should be a priority [11], [24].

OSM recognizes multiple classifications of drivable streets. However, not all of them are important or globally used. In [24], the authors identified the most relevant and frequently used drivable street types as: Motorway, Trunk, Primary, Secondary, Unclassified, Residential, Motorway-Link. Our analysis of these street types show that the distribution of the types across different domains varies hugely. For example, a given type may occur 10,000 times while another type could occur just 50 times. In fact, in certain cities, there are some OSM recognized street types that do not occur there at all. See Fig. 4 for details on the street type distribution across the cities under consideration.

This phenomenon raises a serious issue with regards to having a balanced data set and producing an unbiased model. Hence, we had to determine the optimal number of types given the cities under consideration that will ensure a balanced data set for the input to the model which will ultimately infer these street types. These types are tertiary, primary, secondary and residential. These four types represent the number of streets that we determine will ensure that the model does not become biased to a particular street type during model development for all the cities. From Fig. 4, one can see that the residential streets occur more frequently than other street types. We suspected that this may be a quality issue that arises from over editing or duplication of street objects on OSM [11], [13]. However, we handled this by checking for cases of multiple objects with the same geo-coordinates. Also, ensuring that the data set used was balanced mitigates this issue.

Now, we discuss the four street types as described on OSM as follows:

- **Tertiary**: These are major streets, connecting towns and major city streets. The Kreisstraßen in Germany and a C street in the U.K.
- **Primary**: Usually streets that link larger towns. The Bundesstraße (national street) in Germany or the highest-level street in urban centers. In the U.K, it would be the non-primary A street with black and white signs. And the primary highway in the U.S.
- **Secondary**: This is called the secondary highway in the U.S and a B street in the U.K with black and white signs. In Germany, these are regional streets or Landesstraße.
- **Residential**: When the streets are dotted with residences or lead into housing blocks, the residential description is used.

2) CONTEXTUAL INFORMATION

The second aspect of the data collection deals with the contextual information that exists in each city. Our goal is to infer street types on a spatial network represented as an indirect topological graph $G$, defined as $G = (N, E)$, where $N$ represents the streets and $E$ are the edges which link streets to each other. In this case, the edges denote the intersections between the streets. The spatial context, which is the collection of objects within the proximity of a street is collected as contextual information and used to train the model. See Fig. 5 for a graphical description of streets and their contextual information. To determine the context, we used the object tag descriptions on OSM [36]. These tags describe the characteristics of a spatial object. For example, an OSM tag could classify a building as a residential building.

One of the strengths of OSM is the quantity of semantic information it can hold for spatial objects, this allows for very descriptive information on objects. In OSM, semantic information is grouped into tag types and each tag has

---

1) Best viewed in colour

---

FIGURE 3. A workflow depicting the experimental methodology.

FIGURE 4. Frequency of street types in the cities under consideration. It is seen that residential streets consistently occurs more frequently than any other street type, with it’s frequency many orders of magnitude greater than other types. 

Secondary: This is called the secondary highway in the U.S and a B street in the U.K with black and white signs. In Germany, these are regional streets or Landesstraße.

Residential: When the streets are dotted with residences or lead into housing blocks, the residential description is used.
different values, where the tag values represent the attribute of a spatial object. For the purpose of this article, we considered five (5) tag types and thirty-one (31) values because they were the most used and relevant for the experiments [36]. The five tag types are building, amenity, landuse, shop and service. The tag values we used are presented in the second column of Table 3. The selection comprised mostly of the tag values which were intuitively relevant to the problem domain. The next step was to construct the contextual information for each node of $G$. In this paper, we achieved this using the multiple fixed buffer approach which involves superimposing multiple buffers over each street segment with varying radii $R = \{r_1, \ldots, r_c\}$. This is depicted in Fig. 8. The areas of the buffers is represented as a set $A = \{a_1, \ldots, a_c\}$. The first buffer is a polygon which encloses the street segment, while every other buffer is a polygon with a hole (see Fig. 7). The area of each buffer is determined with a radius $r$ and the count of objects in each buffer is collected and mapped to each node to obtain the contextual information for each node. It is worthy to mention that this ensures that the contextual information collected in each buffer is not duplicated. The same $A$ is used for all the cities. Our procedure incrementally performs this mapping by taking as input each node(street) in the street network ($G$), the areas of the buffers ($A$) and returns the contextual information for this street as an attribute vector ($s_i$).

When complete, the output of this procedure is a matrix $M$ where each row corresponds to a street and the columns hold the count of each attribute with respect to the buffers ($A$). This means that the columns of $M$ will correspond to the count of each $x_i \forall x_i \in A$, where $x_i$ is an element in the attribute vector $s_i$. To illustrate, a street ($s_1$) can have the number of residential houses between 0 and 500 metres and the number of restaurants between 500 metres and 1 kilometre as two of its features. The final stage of this mapping process involves grouping the columns of $M$ into the four urban functions: recreational, residential, utilities and commercial. Figure 6 shows the proportions of these urban functions. The definitions of these functions and the tag-value to function allocations are described in tables 1 and 3. The allocation of these tags to function was done manually.

### B. Model Development

The next stage of the methodology is to determine an algorithm to solve the budget problem (refer to section III). We make our choice of algorithm mostly on the inherent nature of spatial data. The characteristics of spatial data which includes non-linearity, uncertainty and heterogeneity means that the optimal learner will be one that can adequately overcome or model them [9], [10]. This fact already rules out the suitability of any linear model for this problem.

We propose the use of the AdaBoost ensemble algorithm to address this problem [37]. Ensemble methods are a learning paradigm that are based on the simple but powerful notion of wisdom of crowds. The main hypothesis behind ensemble methods is that a contingent of weak learners will produce a learner that performs better than any one weak learner. There are two popular variations of ensemble methods: bagging and boosting [38]. Bagging is essentially a paradigm where the average of the results produced from the weak learners are used to make the final prediction.
Boosting however, adaptively improves the individual learners such that at each iteration the current learner tries to improve on the performance of the previous learner [37], [38]. This ensures that at the end of the learning process, the final learner is a highly improved version of every previous learner (see Fig. 9). An example of a boosted ensemble method is the AdaBoost algorithm. Assuming, we have a set of input variables \( X = \{x_1, \ldots, x_n\} \) where \( x_i \in \mathbb{R} \), we wish to build a learner \( F(X) \), such that when a new input \( X \) is given, we can assign it to a label in \( Y = \{y_1, \ldots, y_K\} \). The AdaBoost algorithm iteratively combines many classifiers which produces a result that is at least better than any single classifier \( f(X) \) [39], this procedure can be described succinctly as follows:

\[
F(X) = \text{sign} \left( \sum_{m=1}^{M} \theta_m f_m(X) \right)
\]

where \( f_m(X) \) represents each weak learner, \( \theta_m \) is the corresponding coefficients of each learner, \( M \) is the number of learners, the final classifier \( F(X) \) is a weighted sum of each classifier. This procedure was developed by Freund and Shapire [39] to address the generic binary classification problem, where \( y = \{-1, +1\} \). Thus, applications to multi-class problems had to be decomposed to multiple binary classification problems. Zhu et al. [37] developed an algorithm that extended the original idea by [39] to multi-classification problems directly without a need for decomposing the problems into binary cases. The exponential loss function for each learner was expressed statistically in [40] thus:

\[
L(y, f) = e^{-yf(x)}
\]

This is implemented iteratively as a stage-wise additive model

\[
F(X) = \arg \max_k \sum_{m=1}^{M} \theta_m \cdot I(f_m(X) = k)
\]

Which is very similar to the original algorithm [39], except for the inclusion of \( \log(K - 1) \) in the weight update procedure of \( m \), now given as

\[
\theta_m = \log \left( \frac{1 - \text{err}_m}{\text{err}_m} \right) + \log (K - 1)
\]

And,

\[
\text{err}_m = \frac{1}{n} \sum_{i=1}^{n} w_i I(y_i \neq f_m(x_i)) \frac{1}{n} \sum_{i=1}^{n} w_i
\]

where, \( w_i = 1/n, i = 1, 2, \ldots, n \) is the observation weights assigned to each \( x_i \) at the start of each iteration. We refer
C. Iddianozie et al.: Exploring Budgeted Learning for Data-Driven Semantic Inference via Urban Functions

FIGURE 9. A description of how the AdaBoost algorithm works. The learner is incrementally boosted at each iteration, where the wrongly classified points from the last iteration are prioritized and the weights assigned to them are adjusted [41].

TABLE 2. Summary of the training data distribution by class. The values are in thousands. The training set was perfectly balanced before model development.

| City       | Size of dataset | Tertiary | Primary | Secondary | Residential |
|------------|-----------------|----------|---------|-----------|-------------|
| Chicago    | 46.8            | 4.31     | 3.73    | 6.43      | 31.2        |
| Munich     | 19.5            | 1.31     | 0.60    | 3.1       | 13.7        |
| Frankfurt  | 12.1            | 1.1      | 0.96    | 1.41      | 7.4         |
| Manchester | 17.4            | 1.45     | 0.65    | 0.37      | 12.5        |
| Rome       | 57.7            | 9.0      | 2.5     | 4.6       | 37.4        |
| Charlotte  | 35.5            | 3.3      | 1.8     | 2.1       | 26.4        |

the interested reader to [39], [40] and [37] for more details. We adapt the version of AdaBoost described by [37] where the weak learner chosen is the decision tree. Decision trees are non-linear which is ideal for the problem domain, but they exhibit high variance and can be easily biased to a particular class or label. However, this weakness will be mitigated using the iterative adaptive boosting of the AdaBoost algorithm.

For the purpose of this article, we build models for each urban function to classify street types.

We proceed to build each model by receiving as input a matrix of street nodes \( M \), the AdaBoost algorithm [37] and a reward. Recall that \( M \) is the two-dimensional matrix that holds the street nodes and the count of their corresponding contextual information with respect to each of the buffers. Now it is of the form \( M = (X, Y) \), where \( X \) is the contextual information and \( Y = \{y_1, \ldots, y_k\} \) with \( k = 4 \) in this case for each of the street types under consideration. \( M \) is adjusted to ensure that the distributions of each class \( Y \) are perfectly balanced. We achieved this by down-sampling (without replacement) all the classes to the size of the smallest occurring class size. The implementation uses 80% of \( M \) for training and 20% for testing. We present the class distributions of the training set before the balancing in Table 2. For every iteration, the set is shuffled and randomly selected without replacement. We initialized \( d \) and \( m \) to run from 1 to \( \sqrt{|X|} \) and \( |X| \) respectively, where \( |X| \) is the size of \( X \). The reward function is the F-score. The algorithm is used to classify the type of each street in the testing set using each the urban function individually followed by an aggregation of all 4 urban functions.

V. RESULTS AND DISCUSSION

A. MODEL RESULTS

Recall that the street types considered were primary, secondary, tertiary and residential. For evaluation, we focus on the F-score to measure the performance of our methodology. We chose this measure because it offers a holistic view of model performance. F-score is the harmonic mean of precision and recall, where precision is the proportion of correctly retrieved points to the total retrieved points and recall is the proportion of correctly retrieved points to the total points in the data set. Thus, the F-score offers a balanced view of both metrics where a perfect score is 1, that is precision and recall are both at 100%. As this is a multi-class classification problem, we believe that it is the most suitable metric to evaluate the performance of the methodology with regards to predicting the street types given the urban functions and the domain (i.e city). This is particularly important considering that if a data-driven model were to be used to improve the data quality, then it will be imperative to understand the overall effectiveness of the model for each street type. The F-score is defined as follows:

\[
F\text{-score} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}
\]

where,

\[
\text{Precision} = \frac{(\text{True Positives})}{(\text{True Positives} + \text{False Positives})},
\]

\[
\text{Recall} = \frac{(\text{True Positives})}{(\text{True Positives} + \text{False Negatives})}
\]

The results are presented in Figure 10, where the F-scores of the model for each city are represented with regards to the urban functions used by AdaBoost and the street type being inferred. Each line in each graph in this figure represents an individual urban function or the aggregate of all four urban functions. The x-axis represents the street types while the
TABLE 3. Allocation of OSM tag values to urban functions.

| Function     | Tag values Allocated                                      |
|--------------|-----------------------------------------------------------|
| Recreational | restaurant, cafe, bar, pub, recreation ground            |
| Residential  | house, apartments, residential, garage, detached, terrace, driveway |
| Utilities    | parking, bench, school, fuel, bank, bicycle parking, parking entrance, library, community centre, railway |
| Commercial   | industrial, commercial, marketplace, farmland, retail, department store, convenience, supermarket, mall |

FIGURE 10. Plot of F-scores achieved on each of the cities. Each line represents the results from each urban function, except for the black line which is the aggregate of all urban functions.

The y-axis is the F-score of the model/learner built using each urban function to infer each street type. In Table 4, each row is representative of an urban function, the last row is the aggregate of the urban functions. The values on each column is the mean F-score of the learner using that urban function (or their aggregate) for inferring a street type across all the cities in the study. In Table 5, each table represents the performance of each urban function as well as their aggregate. Each row in the tables represent a city. The columns of each table hold the mean evaluation score for that urban function (or their aggregate) over the four street types inferred for that city. We present the precision, recall, F-score and accuracy in each of these columns. The F-score values in these tables mirror Figure 10.

B. DISCUSSION

From Figure 10, it can be seen that the difference between the results produced by the urban functions and their aggregate in all the cities is insignificant. This is also seen in Tables 5a -5e, where for each city, the mean F-score of each urban function is tightly clustered around a shared mean. Interestingly, the results show that there is no particular urban function that performs better for a street type than the other. Generally, this can be seen in Table 4, where the results from the best-performing street (i.e Primary) and the worst-performing street (i.e Residential) seems consistent across the urban functions and their aggregate. This may be counter-intuitive, because for example, the initial assumption may be that the residential urban functions will perform better for residential streets. Furthermore, these results clearly show that more data is not always better for data-driven methodologies [42].

TABLE 4. Cumulative Average F-scores produced by the urban functions for the street types in all cities.

| Function     | Tertiary | Primary | Secondary | Residential |
|--------------|----------|---------|-----------|------------|
| Recreational | 0.573    | 0.68    | 0.583     | 0.472      |
| Utility      | 0.585    | 0.694   | 0.605     | 0.497      |
| Residential  | 0.598    | 0.708   | 0.612     | 0.5        |
| Commercial   | 0.589    | 0.705   | 0.612     | 0.508      |
| All-features | 0.609    | 0.714   | 0.622     | 0.513      |
Table 5. Table of experimental results for the average precision, recall, f-score and accuracy for each city. Here, the average performance of each urban function over the street types is presented. The results shown here are the averages taken after five runs. The standard deviations computed for all the cities fall between 0.01 and 0.02.

| City   | Recreational          |          |          |          |          | Utility      |          |          |          |          |          |
|--------|-----------------------|----------|----------|----------|----------|-------------|----------|----------|----------|----------|----------|
|        | precision | recall   | f-score  | accuracy |          | precision   | recall   | f-score  | accuracy |          |          |
| Chicago| 0.4345    | 0.4283   | 0.4293   | 0.4285   |          | 0.4685    | 0.461   | 0.4606   | 0.461    |          |          |
| Munich | 0.603     | 0.6085   | 0.6045   | 0.6085   |          | 0.6366    | 0.6345  | 0.6333   | 0.6345   |          |          |
| Frankfurt | 0.6565 | 0.653    | 0.6534   | 0.653    |          | 0.6705    | 0.6725  | 0.67    | 0.6725   |          |          |
| Manchester | 0.739 | 0.742   | 0.7399   | 0.742    |          | 0.7587    | 0.762   | 0.7568   | 0.762    |          |          |
| Rome   | 0.5282    | 0.5325   | 0.5289   | 0.5325   |          | 0.5251    | 0.528   | 0.5259   | 0.528    |          |          |
| Charlotte | 0.5118 | 0.5035   | 0.5062   | 0.5035   |          | 0.534     | 0.5205  | 0.5246   | 0.5205   |          |          |

![Image](https://example.com/image.png)

As a matter of fact, in some cases, it is seen that the result of the aggregate is worse than that of each urban function. For example, in figure 10, the commercial and residential urban functions outperform the aggregate for inferring the primary street type.

In other words, knowing the distribution of commercial or residential objects around streets will predict the primary street type better than knowing the distribution of all the objects. This confirms the position of Schieder et al. [42] on spatial data mining approaches, that more data is not always required to make good predictions.

These results strongly suggest that in a situation where the entire feature set (i.e. the aggregate of the urban functions) is not available for model development - a budget problem - one of the urban functions can adequately infer the street types. The significance of these results is further heightened when transferability of models are considered. For example, it will be a practical solution to develop a model in one domain that can improve the quality of another domain in OSM [24]. Nonetheless, this is not only applicable for improving data quality. With the proliferation of spatial big data, data collection and processing techniques are faced with a variant of the budgeted learning problem. In such scenarios, efficient methodologies are sought for model development and validation [43]. This could be addressed using urban functions or a similar dis-aggregation of data.

Furthermore, it can be seen from Figure 10 and Tables 5a -5e that Chicago has the worst results on average compared to other cities. We posit that this can be explained by the spatial heterogeneity which exists across different spatial domains. For example, Chicago is the only one of the six cities that is planned [28]. This is evidenced by the grid-like block model of the city [24], [44]. This fact, which may have resulted in the spatial clustering of urban functions but not of the street types (at least on OSM), may explain the poor capacity of the model to infer the street types. We plan to investigate the relationship between spatial distributions of the functions and the street types in later work. In the same vein, it is seen that the recreational function exhibits the worst performance for all street types in all cities (Fig. 10). This may be misinterpreted to be as a result of it having the least number of OSM tag values (five) as depicted in Table 3. However, this is not the case because it is also seen that the residential function corresponding to seven tag values is not the next worst-performing urban function. In fact, it is the best urban function in many scenarios, even out-performing the aggregate (see Frankfurt - residential and primary streets, Rome and Charlotte - tertiary and primary streets in Fig. 10). Again, we posit that this may depend on spatial heterogeneity.
In summary, we see that our results suggest that using urban functions is an adequate solution to the budget problem that exists in our example of data-driven semantic inference. This is particularly important when you consider that the general assumption of machine learning algorithms is that the data to be used for model development is readily available and comes at constant cost. However, this is untrue, especially in the case of crowd-sourced spatial databases. An example is OSM, which still suffers from variable data quality. Now, the use of contextual information has been proposed to improve the quality of OSM [12], [22], but we raise the following questions with respect to this proposal: What data is good enough? and Is the data available? With regards to the first question, approaches could be tempted to use any or all data available. But, as our results show, using all the available data does not guarantee the best results. Feature selection techniques are used in some machine learning domains to choose appropriate parameters by examining the relative importance of features for model development. This does not represent a budget in terms of data availability as feature selection may assume that all data is readily available. However, feature selection does help if time or processing needs are to be budgeted as feature reduction will impact these metrics. There is also a need to ensure that the data which is suitable for addressing the problem in one domain, will be available in other domains. In this regard, our results show that an urban function could serve as an alternative, in the event that the required feature set is not available. The budgeted learning problem has been demonstrated based on the availability of data and features according to urban functions. In practice, there are various budgets. In OSM, there is variable data quality and completeness. In such cases the inference approach described here can be used with an urban function which is most complete.

VI. CONCLUSION AND FUTURE WORK
In this article, we have presented a novel formalization of the budgeted learning problem in data-driven semantic inference. We adopted the semantic data-quality problem of OSM as a use-case and proposed a methodology to address this problem via urban functions. Urban functions describe the designated use of a geographical space, and given that they are inherently heterogeneous, we hypothesized that they could be used to address the budget problem. Our approach inferred the street type attribute for street objects from six cities. We evaluated the model results using F-score. Our results show that urban functions serve as an adequate solution to the budgeted learning problem for data-driven semantic inference. Furthermore, the use of a reduced feature set for modelling introduces cost and processing time efficiencies.

This work is timely with the ever increasing use of data-driven techniques for mining spatial data. The approach we describe in this article is practical and can be used to address the problem of model optimization in spatial data mining. A limitation of this work lies in our definition of the urban functions. Spatial objects can have more than one function but in this article, we have assumed that all objects have a single function and this may limit the suitability of the functions for certain cases. Also, we have not considered other Machine Learning algorithms. For future work, we plan to develop a methodology that leverages the spatial spread of functions to infer suitability in different scenarios. Artificial Neural Networks (ANNs) and other kernel-based methods such as Support Vector Machines (SVMs) have shown promise for multi-label classification techniques and will be investigated and contrasted to the AdaBoost technique described in this article.

ACKNOWLEDGMENT
The authors would like to thank the anonymous reviewers whose comments greatly improved the quality of this publication.

REFERENCES
[1] K. Deng, Y. Zheng, C. Bourke, S. Scott, and J. Masciale, “New algorithms for budgeted learning,” Mach. Learn., vol. 90, no. 1, pp. 59–90, Jan. 2013.
[2] D. J. Lizotte, O. Madani, and R. Greiner, “Budgeted learning of Naïve-Bayes classifiers,” in Proc. 19th Conf. Uncertainty Artif. Intell., San Mateo, CA, USA: Morgan Kaufmann, 2002, pp. 378–385.
[3] P. Corcoran, M. Jilani, P. Mooney, and M. Bertolotto, “Inferring semantics from geometry: The case of street networks,” in Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst., 2015, p. 42.
[4] Y. Hu, “Geospatial semantics,” 2017, arXiv:1707.03550. [Online]. Available: http://arxiv.org/abs/1707.03550
[5] M. Belguiz, I. Tomljenovic, T. Lampingthammer, T. Blaschke, and B. Hõfe, “Ontology-based classification of building types detected from airborne laser scanning data,” Remote Sens., vol. 6, no. 2, pp. 1347–1366, Feb. 2014.
[6] K. Janowicz, S. Scheider, T. Pehle, and G. Hart, “Geospatial semantics and linked spatiotemporal data–past, present, and future,” Semantic Web, vol. 3, no. 4, pp. 321–332, 2012.
[7] W. R. Black, “Network autocorrelation in transport network and flow systems,” Geograp. Anal., vol. 24, no. 3, pp. 207–222, Sep. 2010.
[8] T. Cheng, J. Haworth, and J. Wang, “Spatio-temporal autocorrelation of road network data,” J. Geograph. Syst., vol. 14, no. 4, pp. 389–413, Oct. 2012.
[9] B. Jiang, “Geospatial analysis requires a different way of thinking: The problem of spatial heterogeneity,” Geojournal, vol. 80, no. 1, pp. 1–13, Feb. 2015.
[10] S. Shekhar, M. R. Evans, J. M. Kang, and P. Mohan, “Identifying patterns in spatial information: A survey of methods,” WIREs Data Mining Knowl. Discov., vol. 1, no. 3, pp. 193–214, May 2011.
[11] M. Haklay, “How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets,” Environ. Planning B, Planning Des., vol. 37, no. 4, pp. 682–703, Aug. 2010.
[12] M. Jilani, P. Corcoran, and M. Bertolotto, “A multi-layer CRF based methodology for improving crowdsourced street semantics,” in Proc. 11th ACM SIGSPATIAL Int. Workshop Comput. Transp. Sci. (IWCTS), 2018, pp. 29–38.
[13] P. Mooney, P. Corcoran, and A. C. Winstanley, “Towards quality metrics for OpenStreetMap,” in Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (GIS), 2010, pp. 514–517.
[14] A. Mashhadi, G. Quattrone, L. Capra, and P. Mooney, “On the accuracy of urban crowd-sourcing for maintaining large-scale geospatial databases,” in Proc. 8th Annu. Int. Symp. Wikis Open Collaboration (WikiSym), 2012, p. 15.
[15] B. Nushi, A. Singla, A. Krause, and D. Kossmann, “Learning and feature selection under budget constraints in crowdsourcing,” in Proc. 4th AAAI Conf. Hum. Comput. Crowdsourcing, 2016, pp. 1–10.
[16] F. Nan and V. Saligrama, “Adaptive classification for prediction under a budget,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4727–4737.
[17] T. VoPham, J. E. Hart, F. Laden, and Y.-Y. Chiang, “Emerging trends in geospatial artificial intelligence (geaAI): Potential applications for environmental epidemiology,” Environ. Health, vol. 17, no. 1, p. 40, Dec. 2018.
[18] M. Eckle and J. A. P. D. Albuquerque, “Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes,” in Proc. ISCRAM, 2015, pp. 1–9.

[19] S. Scholz, P. Knight, M. Eckle, S. Marx, and A. Zipf, “Volunteered geographic information for disaster risk reduction—The missing maps approach and its potential within the red cross and red crescent movement,” Remote Sens., vol. 10, no. 8, p. 1239, 2018.

[20] S. Camboim, J. Bravo, and C. Sluter, “An investigation into the completeness of, and the updates to, OpenStreetMap data in a heterogeneous area in Brazil,” ISPRS Int. J. Geo-Inf., vol. 4, no. 3, pp. 1366–1388, Aug. 2015.

[21] OpenStreetMap. (2019). Major OpenStreetMap Consumers. Accessed: Feb. 27, 2019. [Online]. Available: https://wiki.openstreetmap.org/wiki/Major_OpenStreetMap_Consumers

[22] M. Jilani, P. Corcoran, and M. Bertolotto, “Automated highway tag assessment of OpenStreetMap road networks,” in Proc. 22nd ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (SIGSPATIAL), 2014, pp. 449–452.

[23] M. Jilani, P. Corcoran, and M. Bertolotto, “Automated quality improvement of road network in OpenStreetMap,” in Proc. Agile Workshop (Action Interact. Volunteered Geograph. Inf.), 2013, p. 19.

[24] C. Iddianozie and G. McArdle, “A transfer learning paradigm for spatial networks,” in Proc. 34th ACM/SIGAPP Symp. Appl. Comput. (SAC), 2019, pp. 659–666.

[25] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” J. Big Data, vol. 3, no. 1, p. 9, Dec. 2016.

[26] N. Segev, M. Harel, S. Mannor, K. Crammer, and R. El-Yaniv, “Learn on source, refine on target: A model transfer learning framework with random forests,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 9, pp. 1811–1824, Sep. 2017.

[27] J. Gao, W. Fan, J. Jiang, and J. Han, “Knowledge transfer via multiple model local structure mapping,” in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2008, pp. 283–291.

[28] C. Alexander, “A city is not a tree,” in The Urban Design Reader. Evanston, IL, USA: Routledge, 2013, pp. 172–186.

[29] C. Roth, S. M. Kang, M. Batty, and M. Barthélémy, “Structure of urban movements: Polycentric activity and entangled hierarchical flows,” PloS ONE, vol. 6, no. 1, Jan. 2011. Art. no. e15923.

[30] J. Portugali, “What makes cities complex?” Complex., Cognition, Urban Planning Des., pp. 3–19, 2016.

[31] A. J. Scott and M. Storper, “The nature of cities: The scope and limits of urban theory,” Int. J. Urban Regional, vol. 39, no. 1, pp. 1–15, Jan. 2015.

[32] A. T. Crooks, A. Croitoru, A. Jenkins, R. Mahabir, P. Agouris, and A. Stefanidis, “User-generated big data and urban morphology,” Built Environ, vol. 42, no. 3, pp. 396–414, Oct. 2016.

[33] W. Tu, J. Cao, Y. Yue, S.-L. Shaw, M. Zhou, Z. Wang, X. Chang, Y. Xu, and Q. Li, “Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns,” Int. J. Geograph. Inf. Sci., vol. 31, no. 12, pp. 2331–2358, Dec. 2017.

[34] B. Zhang, L. Zhang, T. Guo, Y. Wang, and F. Chen, “Simultaneous urban region function discovery and popularity estimation via an infinite urbanization process model,” in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2018, pp. 2692–2700.

[35] A. Crooks, D. Pfeifer, A. Jenkins, A. Croitoru, A. Stefanidis, D. Smith, S. Karagiozoglu, A. Efentakis, and G. Lamprianidou, “Crowdsourcing urban form and function,” Int. J. Geograph. Inf. Sci., vol. 29, no. 5, pp. 720–741, May 2015.

[36] OpenStreetMap. (2019). OpenStreetMap Taginfo. Accessed: Mar. 30, 2019. [Online]. Available: https://taginfo.openstreetmap.org

[37] T. Hastie, S. Rosset, J. Zhu, and H. Zou, “Multi-classboost,” Statist. Interface, vol. 2, no. 3, pp. 349–360, 2009.

[38] J. R. Quinlan, “Bagging, boosting, and c4.5,” in Proc. AAAI/IAAI, vol. 1, Aug. 1996, pp. 725–730.

[39] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of online learning and an application to boosting,” J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, Aug. 1997.

[40] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors),” Ann. Statist., vol. 28, no. 2, pp. 337–407, Apr. 2000.

[41] B. Marsh, “Multivariate analysis of the vector boson fusion higgs boson,” Tech. Rep., 2016. [Online]. Available: https://www.researchgate.net/publication/306054843_Multivariate_Analysis_of_the_Vector_Boson_Fusion_Higgs_Boson/stats

[42] S. Scheider, F. O. Ostermann, and B. Adams, “Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis,” Future Gener. Comput. Syst., vol. 72, pp. 11–22, Jul. 2017.

[43] J.-G. Lee and M. Kang, “Geospatial big data: Challenges and opportunities,” Big Data Res., vol. 2, no. 2, pp. 74–81, Jun. 2015.

[44] G. Boeing, “Urban spatial order: Street network orientation, configuration, and entropy,” Appl. Netw. Sci., vol. 4, no. 1, p. 67, 2019.