**RC25116 (W1012-034) December 10, 2010**
**Computer Science**

# IBM Research Report

## Selecting Relevant Sensor Providers for Meeting "Your" Quality Information Needs

**George Tychogiorgos**
Electrical and Electronic Engineering
Imperial College
London SW7 2AZ, UK

**Chatschik Bisdikian**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

**Research Division**
**Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich**

# Selecting Relevant Sensor Providers for Meeting "Your" Quality Information Needs

George Tychogiorgos
Electrical and Electronic Engineering
Imperial College
London SW7 2AZ, UK
*g.tychogiorgos@imperial.ac.uk*

Chatschik Bisdikian
IBM Research
Thomas J. Watson Research Center
Hawthorne, NY 10514, USA
*bisdik@us.ibm.com*

***Abstract*—The systematic or on-demand deployment of tethered or untethered sensor networks raises the challenge for selecting the providers (i.e., sensor networks) supplying the most "relevant" sensory information. This paper considers the spatial relevancy of information and investigates determining the spatial relevancy of provided information relatively to the quality gradations of the desired information. The paper introduces an expansion-proof descriptor of desired or provided information and uses it to select the single, most relevant provider. Then it considers multi-provider aggregation in the context of two objectives: (*a*) selecting the minimum number of providers that cumulatively maximizes the information relevancy; and (*b*) considering a cost per provider, selecting the subset of providers that cumulatively maximizes the overall information relevancy subject to a budgetary constraint. The performance and robustness of the proposed solutions are studied both analytically and by simulation for a number of provider topologies.**

## I. Introduction

Consider the case where, say, a city agency needs to monitor air-quality (or, hazmat concentration levels, etc.) throughout the area of its authority. The agency would like to collect air-quality information at different quality levels, e.g., higher granularity in densely populated regions, and lower granularity at other regions. To collect the needed information, the agency uses sensors that it had deployed in the past. Unfortunately though, due to budget constraints and other logistics challenges, these sensors cover only portions of the area of interest. To supplement its information needs, the agency has decided to select and engage third-party fixed and mobile sensory information providers with whom it would create persistent or transient relations as necessary. The third-party providers could be other city agencies, private operators that, for example, monitor air-quality in public areas (parks, arenas, etc.), fleet operators whose fleet vehicles are equipped (for various reasons) with the necessary sensory devices, and even individuals whose smart-phones are capable of sensing air-quality conditions.

Research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

The above hypothetical (albeit not improbable) scenario exemplifies a trend where increased deployment and use of sensor networks is ushering a new era where information-rich solutions are becoming even more pervasive and integrated parts of our personal and professional lives. The emergence of the *Internet of Things* (IoT) [1] and *participatory sensing* [2] will further hasten the rate and ease with which information from tethered, untethered sensors, the Web, etc., will coalesce on demand to support our information needs.

There are undoubtedly several challenges in realizing the "city agency" scenario. They relate, and not only, to technology; system (HW/SW) architecture and design, operation, and management; regulatory constraints; and, this being a city agency, public perception. It is the purpose of this paper to study one of these challenges that of dealing with selecting information providers that supply the most *relevant* information for our (i.e., the user's) needs. Specifically, we seek to establish procedures by which we can compare information sources based on how relevant the information they produce is to the desired and sought after information.

To this end, we need to develop means to capture properties of the information against which relevancy can be assessed and metrics to capture the ensuing levels of relevancy. In [3], we proposed using the spatiotemporal properties of information for identifying (or at least narrowing down) the relevant information. These properties also serve as the basis for *quality of information* (QoI) metadata representing the physical context of the information [4]. Assuming compatible types for the sought and provided information, relevancy was measured by "how spatiotemporally close" a piece of information provided is to the information desired. Specifically, we defined and measured spatial relevancy by the degree of overlap between the region $R_p$ describing the coverage of sensory information from a provider and the region $R_d$ describing the coverage of sensor information desired by a user; we, likewise, defined and measured temporal relevancy.

As the number and variety of potential sources of information as well as the number of applications that depend on and search for them increases, the process of selecting the most relevant ones becomes more and more challenging. Furthermore, the fluidity of untethered sources (humans in participatory sensing, sensor-equipped vehicles, etc.) adds to

the challenge as application interested in information from a particular region, may need to seek for and bind repeatedly to new(er) relevant sources. These challenges have a three-fold impact: increased processing, storage and communication requirements; all elements of concern when considering resource-constrained sensor networks. The processing challenge is the obvious and direct one as more and more candidate sources have to be assessed and selected from. The other two are more subtle. The increases in the number of sources and applications will inadvertently result in an increase in the pertinent advertisements and exchange of metadata about (at least) the spatiotemporal and general QoI properties from the sources and/or desired by the applications. These metadata will also have to be stored at various nodes in the network.

There is an additional challenge that can further exacerbate all three previous challenges: *metadata expansion*. As more sources become available, new compound sources could (and would) be created as needed. For example, a new source reporting air-quality from the east (E) and north (N) regions areas of a city can be created by the combination of regional sources reporting air-quality from the portions of the E, NE, and N regions of the city. How should the spatiotemporal properties of the compound source be represented? The obvious way is to combine (e.g., take the union of) the corresponding metadata from each of the constituent sources. This will result in a more populous entry for these metadata. As more and more sources are compounded this will lead to the unbounded increase of the related metadata entry, which of course will create major management burdens regarding their processing, communication, and storage.

In this paper, we build upon our early work in [3] by considering the aforementioned multitude of operational challenges as sensory sources and applications that depend increases. The contributions in this paper are: (*a*) the introduction of QoI functions for describing the contextual desirability/quality of information; (*b*) the definition of a novel problem and a new metric regarding information relevancy based on the QoI functions; (*c*) the provision of finite, expansion-proof metadata descriptors for the QoI functions, using approximation techniques, such as spline surfaces; (*d*) the formulation of optimization problems for selecting a single or multiple relevant providers with our without constraints; and (*e*) the solution algorithms and study of these optimization problems.

The organization of the paper is as follows: Section II introduces relevancy and its QoI function based metric. Section III presents the expansion-proof description of QoI functions. Section IV introduces the multi-provider composition problem and studies pertinent optimization problems along with solution algorithms. Section V provides the numerical evaluation of our solutions for various provider topologies. Finally, Section VI summarizes the paper and provides concluding remarks along with related work.

## II. The Relevancy of Sensory Information

We start with a brief summary of relevancy from [3], and then we built upon it focusing on a quality influenced

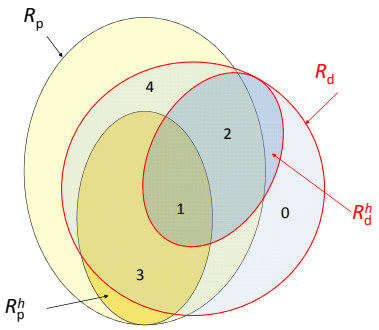


Fig. 1. Spatial properties of desired and provided sensor-originated information–regions are shown as ellipses for illustrative purposes only.

definition of it. We then present the problems at hand, and discuss solution approaches in the following sections. For ease of presentation, and without lack of generality, we focus only on spatial relevancy over two-dimensional regions. Extensions to 3-D (or 4-D) spatiotemporal volumes are possible, albeit at increased levels of notational (and computational) complexity.

### A. Background on Spatial Relevancy Metrics

In [3], we (implicitly) defined spatial relevancy as the degree of spatial overlap that there exists between the information we seek and the information we are provided, e.g., the coverage of the sensor networks supplying the sensor data feeds that an application taps to. Consequently, we defined the metric $r_s$ of *spatial relevancy* as:

$$r_s(R_d, R_p) = \frac{f\big(\mathcal{A}[R_d, R_p]\big)}{f\big(\mathcal{A}[R_d, R_d]\big)}; \tag{1}$$

where $R_d$ is a description of the *desired* spatial properties of the information sought and $R_p$ are those for the information *provided*; mnemonically, $R$ could stand for *region*, but not necessarily. $\mathcal{A}$ maps a correlation of the spatial properties of, say, $R_d$ and $R_p$ to the nonnegative reals (mnemonically, $\mathcal{A}$ could stand for the area of the overlapping regions $R_d$ and $R_p$). Finally, $f(\cdot)$ is a non-negative, non-decreasing function of its argument, such as $f(x) = x$. The denominator $f\big(\mathcal{A}[R_d, R_d]\big)$ plays the role of a normalization coefficient so that $r_s \in [0, 1]$.

As an example, consider Figure 1 showing the spatial coverage of the desired information $R_d$ in the regions enclosed by the red lines, and the sensor-generated information $R_p$ in colored region; ignore the superscript $h$ for the moment. $\mathcal{A}[R_p, R_c]$ represents the area overlap between regions $R_d$ and $R_p$, then, assuming $f(x) = x$, then:

$$r_s(R_d, R_p) = \frac{area[R_d \cap R_p]}{area[R_d]}. \tag{2}$$

### B. Generalizing Spatial Relevancy and the QoI Functions

By adding gradations in the desirability or quality of information across the regions $R_d$ and $R_p$, we can generalize the "overlap" principle in (2) and, consequently, the information spatial relevancy definition and metric. The superscript $h$ in Figure 1 stands for higher quality or desirability in contrast to

regular ones. Specifically (going beyond [3]), let $\omega = (x, y)$ be a point in a two dimensional region $R$ and let

$$q_d : \omega \in R_d \rightarrow [0, 1], \quad \text{with} \int_{R_d} q_d(\omega)\, d\omega < \infty \tag{3}$$

be a *desired QoI function* describing the quality of the *desired* information related to point $\omega$. For example, at point $\omega_0$, a detection application "desires" to receive information about event occurrences that have probability of correct detection $z_0$ (i.e., $q_d(\omega_0) = z_0$), or the concentration of air-pollutants at that point with accuracy $\epsilon_0$. The range of $q_d$ could be the entire real line, but we assume that it is expressed in relative terms and normalizable with values closer to 1 representing higher desired information quality levels. By convention, we set $q_d(\omega) = 0$, for all points $\omega$ outside the desired region $R_d$. We can define the *provided* (or *provider*) *QoI function* $q_p(\omega)$ on a set $R_p$ in a completely analogous fashion to $q_d(\cdot)$ and $R_d$.

According to one operational mode, sensor-enabled applications may "announce" their information needs by broadcasting their desired QoI function $q_d$ and its support region $R_d$; interested providers may then respond to the application in kind. According to another operational mode, providers may "advertise" their sensing capabilities by broadcasting their function $q_p$ and its support region $R_p$; applications can sift through these advertisements and select appropriate providers.

These (or other) operational modes are beyond the scope of this paper. We are concerned only with the fact that an application ends up with a collection of QoI functions $q_p$ from providers. Based on them and its own QoI function $q_d$, it assesses their relevancy to its information needs, ranks them accordingly, and chooses an appropriate one (or ones).

To this end, we extend the relevancy metrics of the previous subsection and write

$$r_s^v(q_d, q_p) = \frac{\int_{R_d \cap R_p} v\big(q_p(\omega); q_d\big)\, d\omega}{\int_{R_d} v_d\big(q_d(\omega)\big)\, d\omega}, \tag{4}$$

where $v\big(\cdot; q_d\big)$ is a (non-negative) *value* function that represents the value the sensor-enabled application gains in executing its task when it uses information of quality $q_p(\omega)$ at point $\omega$, and $v_d(\cdot) \stackrel{\text{def}}{=} v(\cdot; q_d)$. The notation $v\big(\cdot; q_d\big)$ implies that, in general, the value function can be expressed in relative to $q_d$, as was done, in a different context, with the *QoI satisfaction index* in [5]. The denominator in (4), which is assumed to be finite, plays the role of a normalization factor so that $r_s^v(q_d, q_p) \in [0, 1]$. We silently assume that an application gains nothing extra if it receives information of higher quality than what it asked for, and, hence, for each $\omega \in R_d \cap R_p$: $v\big(q_p(\omega); q_d\big) \in [0, v_d\big(q_d(\omega)\big)]$. If the latter is not the case, one may need to appropriately redefine the normalization role of the denominator; we do not consider the latter case in this paper.

To reflect intuition, the value function is (selected) such that the relevancy metric exhibits an increasing trend with $q_p$. Specifically, if there are two providers $a$ and $b$ with $a$ "closer" to the desired needs of the application than $b$, i.e., their QoI functions satisfy:

$$\|[q_p - q_p^a]^+\| \leq \|[q_p - q_p^b]^+\|, \text{ where } x^+ \stackrel{\text{def}}{=} \max(0, x), \tag{5}$$

according to some function norm operator defined over the set $R_d$, e.g., the $l_2$ norm, then:

$$\int_{R_d \cap R_p^a} v\big(q_p^a(\omega); q_d\big)\, d\omega \geq \int_{R_d \cap R_p^b} v\big(q_p^b(\omega); q_d\big)\, d\omega. \tag{6}$$

For the numerical results later in the paper, we will use "$\min$" as the value function; "$\min$" satisfies all behaviors expected from a value function. In this case, we write (we will drop the superscript $v$ for brevity):

$$r_s(q_d, q_p) = \frac{\int_{R_d \cap R_p} \min\big\{q_p(\omega), q_d(\omega)\big\}\, d\omega}{\int_{R_d} q_d(\omega) d\omega}. \tag{7}$$

Note there might be alternative interpretations of (4), such as probability expectations of some form, or the conditional or relative entropy of the desired information in the presence of the provided information [6]. There could be some operational complications that these interpretations may introduce, such as the need for a priori knowledge or on-demand computation of joint or conditional probability densities between entities (the providers and the applications) that had no prior kinship to each other. Nonetheless, in principle, these alternative interpretations do not alter the fundamentals of advertising desired or provided QoI functions and making provider selections based on them, which is the premise of this paper.

### C. Problem Scope Definition

Ideally, communicating and manipulating general functions such as $q_d$ and $q_p$ defined over general sets $R_p$ and $R_d$ in order to calculate the relevancy metric in (4) requires unpredictable (if not infinite) accuracy, storage, and computational resources. Operationally, this is impossible and these functions will be communicated via a collection of QoI metadata approximately describing $R_i$ and the corresponding $q_i$, $i \in \{d, p\}$. Geospatial descriptions of regions, based on various types of polygon representations, provide for the boundaries of regions [7]. These descriptions are typically used to decide topological relationships such as when querying whether a point $p$ or a region $A$ is internal, external, at the boundary, or intersecting another region $B$.

Owed to the fact that we are dealing with region intersections, such as $R_d \cap R_p$, topological relationships play role in our case as well. But our queries are not topological in nature, at least not in the traditional sense. Our objective is to order and select providers based on relevancy assessed over the intersection of support sets for the QoI functions $q_d$ and $q_p$ and of course their values, see (4), all described by QoI metadata. Note that these QoI metadata will be communicated and stored at recipient nodes, e.g., the application node, or a provider registry, and the QoI functions $q_p$ (and their support sets) could be the result of aggregation from constituent $q_p$'s. Thus, for example, if the QoI metadata entry in a provider registry table may accommodate only up to $M$ elements, then, this number

is bound to be exceeded if metadata for compound provider are simply the union of the metadata of the constituent providers. Hence, owed to the latter fact, a predictable structure for these metadata will also be required.

Hence the information relevancy problem at hand is summarized as follows:

- *Summarize the QoI functions of providers and applications through finite-sized, expansion-proof descriptors (metadata).* Using these descriptions,
- *Assess the relevancy of providers to an application's needs.* Using these assessments,
- *Select one or more of the providers to satisfy the application's needs given selection criteria, such as the most relevant provider, or the most relevant collection of providers given a budget (e.g., energy, cost) constraint.*

In the next section, we consider the single relevant provider sub-problem, while the multi-provider sub-problem is considered at a later section. We will use splines to satisfy the need for expansion-proof description of the QoI functions.

## III. Expansion-proof QoI Function Description

As stated before, due to the generality of QoI functions $q_d$ and $q_p$ their communication, storage, and processing requirements may be quite unpredictable which has severe implication in managing system resources effectively. Hence, it would be desirable to describe them in a way that ensures predictable utilization of system resources while acknowledging their role in the process of selecting the most relevant information providers to serve an application's needs based on the relevancy metric $r_s^v$ in (4).

To this end, we present a way to describe QoI functions using a collection of data points of finite size $M$; we will refer to these as the *metadata*.[1] The size $M$ is a design parameter trading-off between efficiency with accuracy in describing quality functions. Our numerical results later on show that ranking relevant providers remains robust with respect to $M$. Hereafter, we will write $q$ for a QoI function in general; whenever necessary, we will specify $q_d$ or $q_p$.

### A. Spline-based QoI Function Description

*Splines* are piecewise polynomial curves which are differential up to a prescribed order [8]. A *B-spline* has the property that every spline of a given polynomial degree can be expressed as linear combination of a set of B-splines of the same degree. The *B-spline surfaces* are the result of the tensor product of B-spline curves, where a tensor product surface is generated by:

$$p(x,y) = \sum_{i=1}^{K} \sum_{j=1}^{L} B_i(x) B_j(y) \alpha_{ij}, \tag{8}$$

with $B_i(\cdot)$ and $B_j(\cdot)$ independent spline curves that form a basis and $\alpha_{ij}$ the spline control points.

[1]These could be part of a bigger collection of QoI metadata [4].

The construction of the B-spline curves $B_i(\cdot)$ and $B_j(\cdot)$ is a two-pass process (one for each variable) and is based on the calculation of the so-called *knot* vectors and *control points* $\alpha_{ij}$ [9]. The design parameters of the method are the size of the knot vectors, $n_{knot}^x$ and $n_{knot}^y$, and the spline order along each direction, $order_x$ and $order_y$. The spline order is in essence the order of the polynomial used for the approximation. The input of the approximation procedure is the sample matrix of the QoI function, $\tilde{q}$, along with the sampling vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.[2] The resulting finite description of $q$ consists of $P$ parameters, the knot vectors of size $n_{knot}^x$ and $n_{knot}^y$, and a matrix of size $(n_{knot}^x - order_x) \times (n_{knot}^y - order_y)$ containing the *control points*, $\alpha_{ij}$. Thus, the finite description of $q$ will be $P = n_{knot}^x + n_{knot}^y + (n_{knot}^x - order_x) \cdot (n_{knot}^y - order_y)$ points. These $P$ parameters are necessary to be communicated so that the other parties can generate the approximated sample points, $\hat{q}$.

Due to their smooth, differentiable behavior, and ease of construction, splines and spline surfaces have been studied long and are popular in approximating single- and multivariate functions. Because they can be described by a finite number of points, they are also our preferred approximation choice in describing QoI functions $q$. Since, we define $q$ on $\mathbb{R}^2$, we make use of spline surfaces as in (8) and use the aforementioned $P$ control points and knots to describe it. Increasing the number of *knots* of the approximation or the order of the *spline*, and, hence, the number of *control points*, and eventually $P$, would give better approximation results. However, the simulation results presented later on show the efficiency of the method even for low order approximations.

With regard to our "city agency" scenario earlier, we assume that the $P$ parameters to determine the spline approximations are known to all providers that do business with the city. Specifically,

- Providers encode their $q_p$ functions using spline surfaces defined by the $P$ parameters. Being good citizens, they also calculate the minimum rectangular containing the desired region $R_p$. This requires three additional $(x, y)$ points, and it is used to speed-up the search process. The "city agency" may encode its $q_d$ likewise.
- The city agency distributes a CfP (Call for Providers) along with its $P(+6)$ parameters of its own $q_d$ function and collects responses from providers.
- The agency quickly filters out any provider $p$ whose minimum rectangle containing the corresponding $R_p$ does not intersect with the minimum rectangle for $R_q$;
- It uses the $P$ parameters to approximate the $q_p$ of any remaining provider by generating $B_i(\cdot)$ and $B_j(\cdot)$, based on the knot vectors for variables $x$ and $y$, respectively, and (8); and, finally,
- It determines each provider's relevancy using the approximated $q_p$ and $q_d$ QoI functions in (4) and the providers are then ordered accordingly.

[2]The QoI function $q$, the sample matrix $\tilde{q}$ and the sampling vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are connected by: $\tilde{q}(i,j) = q(\boldsymbol{x}(i), \boldsymbol{y}(j))$.

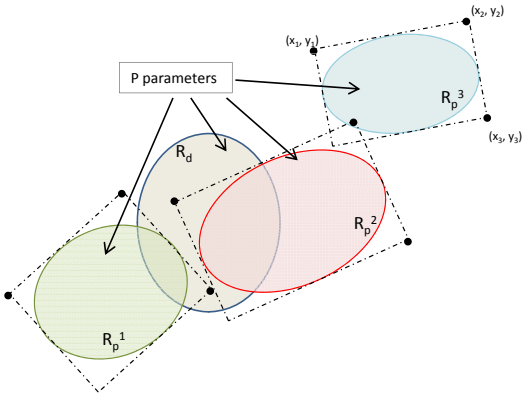


Fig. 2. Example of multiple desired/provided regions $R$ and containing rectangles.

With respect to Figure 2, B-splines are used to generate $P$ parameters describing $q_i$ in region $R_i$, $i \in \{d, p\}$. Three additional points $\{(x_i, y_i); i = 1, 2, 3\}$ are also used to describe the minimum rectangle containing these regions.

Note, that whether $(a)$ the CfP contains just the $P$ parameters, just the 6 rectangle parameters, or all $P + 6$ parameters; or $(b)$ a provider precalculates its $q_p$ approximation, post-calculates it based on the CfP, e.g., use its $P$ points to describe $q_p$ only in the region of interest and not on the entire $R_p$; or $(c)$ the agency and the providers communicate with each other directly or through a proxy/broker in the middle; or $(d)$ only providers need to encode their $q_p$; or $(e)$ ..., are left for future investigation. Here we focus only on the fundamental structures and procedures of the relevancy assessment on top of which all the other choices can be considered and evaluated.

Clearly, using *B-splines* to approximate $q$ is an effective method for generating QoI metadata, but is not the only way. Other approaches were considered as well such as sampling the $R$ regions or quantizing the $q$ values. Due to space limitations we do not present these other cases here. Nonetheless, we have found spline approximations quite general and effective.

## IV. Multi-Provider Consideration

While it is possible that a single provider may suffice in satisfying an application's needs, it is quite likely that it will not. In this case, it would be desirable to be able to judiciously select a number of providers that cumulatively provide the most relevant information.

Using our finite-size, expansion-proof metadata principle, in this section, we consider the composition of sensory information providers and the selection of the most appropriate set of providers based on criteria such as maximum coverage and maximum aggregate geospatial relevancy for a given constraint. In the context of the city agency scenario, this may correspond to the case that the city agency will have to select the most appropriate providers given a budget constraint.

In general, we assume an application with $q_d$ and $R_d$ representing its desired QoI function and corresponding region. There is also a set $\mathcal{P}$ of providers of size $|\mathcal{P}| = N$ with $q_p^i$ and $R_p^i$, $i \in \{1, \ldots, N\}$, the corresponding provider QoI functions and regions. In the following subsections, we consider two cases: $(a)$ the no-cost case, where we seek to find the minimum number of providers that satisfy the application needs without any budgetary constraints; $(b)$ the cost case, where engaging providers comes at a cost and applications have budgetary constraints. In both cases, we will first formulate a model for the problem and then consider a solution for it.

### A. Maximum Relevancy with Minimum Providers and No-cost

We start with the case of selecting the minimum number of providers that can cover as much of the desired region as possible while attaining as high quality as possible. To this end, let $\mathbf{I} = [I(1), \ldots, I(N)]$ be the provider *selection indicator* vector, with $I(i) = 1$ if provider $i$ is selected, and $= 0$ otherwise. Additionally, let the aggregate provider region $R_p^{\mathbf{I}}$ be the union of all the selected provider regions, i.e.,

$$R_p^{\mathbf{I}} = \bigcup_{i=1}^{N} I(i) \cdot R_p^i. \tag{9}$$

The selection of the appropriate set of providers to maximize the coverage of the desired region with no cost can be modeled by the following optimization problem $\Pi_{nc}$:

**Problem $\Pi_{\mathbf{nc}}$**: For $I(i) \in \{0, 1\}$, $i \in \{1, \ldots, N\}$,

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{N} I(i), \quad \text{such that, } \forall \omega \in R_d \cap R_p^{\mathbf{I}}: \\ &(1) \quad \sum_{i:\omega \in R_d \cap R_p^i} I(i) \geq 1; \text{ and} \\ &(2) \quad \max_{i:\omega \in R_d \cap R_p^i} \left[ I(i) \cdot q_p^i(\omega) \right] = \max_{i:\omega \in R_d \cap R_p^i} \left[ q_p^i(\omega) \right]. \end{aligned} \tag{10}$$

Constraint (1) is a *coverage* constraint that states that for each point $\omega \in R_d$ covered by one or more providers, at least one of them will be selected. Constraint (2) is a *preference* constraint that states that the provider with the highest QoI at a point $\omega$ shall be chosen. Note that this model allows the selection of providers that overlap at some points, however, it assures that the best provider at each point is among the selected ones. Therefore, the formulation is implicitly maximizing the aggregated spatial relevancy.

Problem $\Pi_{nc}$ is a generalization of the *set covering* problem [10] on three dimensions (each 2*D* point $\omega$ is also associated with quality value $q_d(\omega)$) and for unity costs. The set covering problem relates to finding the minimum number of sets whose union includes all points of the "universe." It is modeled by the following *integer programming* formulation: For $I(s) \in \{0, 1\}$, for all $s \in S$,

$$\text{minimize } \sum_{s \in \mathcal{S}} c(s) \cdot I(s), \text{ such that } \sum_{s:e \in s} I(s) \geq 1, \tag{11}$$

for all elements $e \in \mathfrak{U}$, where $\mathfrak{U}$ is the universe of points, $\mathcal{S}$ is a family of subsets of $\mathfrak{U}$ and $c(s)$ is the cost associated with set $s$ in $\mathcal{S}$. The *set covering* problem is one of Karp's 21 *NP*-complete problems [11]. Therefore, the $\Pi_{nc}$ problem

is *NP*-complete as well and, hence, there is no polynomial-time algorithm that solves it. The most efficient algorithm solving (approximately) the set covering problem is a *greedy* algorithm that is based on the following simple operation: At every iteration, choose the set that contains the largest number of uncovered elements. The algorithm terminates when all elements are contained in the sets already selected.

Based on the aforementioned iterative operation, we propose a solution to problem $\Pi_{nc}$ described by Algorithm 1 which, at each iteration, selects the most appropriate subset of providers that maximize the total relevancy with respect to the desired information, which is described by the QoI function $q_d$. Because of the possibility that $R_d \cap R_p^{\mathbf{I}}$ contains infinitely many points, the selection criterion at each iteration is not the number of points contained in each set of providers but, instead, the increase in the spatial relevancy metric. Thus, the provider that results to the largest increase in the aggregate relevancy is chosen at each iteration and the algorithm terminates when none of the remaining providers can increase the aggregate relevancy further.

More specifically, at each iteration $t$, the aggregate region $\mathcal{S}$ of the already selected providers $\mathcal{F}$, i.e., $\mathcal{S} = \cup_{k\in\mathcal{F}} R_p^k$, is merged with the new candidate region $R_p^i$. Then, the relevancy of the aggregated QoI function $q_p^{i,\mathcal{F}}(\omega)$ (see shortly) is calculated for all candidate providers $i$ and the one in the already selected set $\mathcal{F}$. Consequently, the provider leading to the highest aggregate relevancy ($V^t$), is selected, until there is no further increase in the total relevancy.

**Algorithm 1** – Aggregate Relevancy

1: Initialize: $\mathcal{F} = \emptyset$, $\mathcal{S} = \emptyset$, $\mathcal{P} = \{1, \ldots, N\}$, $t = 1$ and $V^0 = 0$;
2: Set: $\mathcal{F}_i^t = \mathcal{F} \cup \{i\}$, $\mathcal{S}_i^t = \mathcal{S} \cup R_p^i$ for all providers $i \in \mathcal{P}$;
3: Calculate *spatial relevancy*, $r_s^t\big(q_d(\omega), q_p^{i,\mathcal{F}}(\omega)\big)$, for all regions $\mathcal{S}_i^t$ using equation (4);
4: $k = \arg\max_i \big\{ r_s^t\big(q_d(\omega), q_p^{i,\mathcal{F}}(\omega)\big)\big\}$; let $V^t$ the corresponding maximum value of $r_s^t(\cdot)$;
5: **if** $V^t = V^{t-1}$ **then**
6: STOP;
7: **else**
8: Set: $\mathcal{F} \leftarrow \mathcal{F}_k^t$, $\mathcal{S} \leftarrow \mathcal{S}_k^t$; $\mathcal{P} \leftarrow \mathcal{P} \setminus \{k\}$;
9: Go to step 2 with $t \leftarrow t + 1$;
10: **end if**

In step 3 of Algorithm 1, we use the *aggregated* QoI function $q_p^{i,\mathcal{F}}(\omega)$ which represents the collective behavior of the already selected providers (in the set $\mathcal{F}$) and the new candidate provider $i$ at the point $\omega \in \mathcal{S}$. Specifically, given two providers $i$ and $j$ with $q_p^k$ and $R_p^k$, $k \in \{i, j\}$, their respective QoI functions and provider regions, their combined QoI function $q_p^{i,j}$ is defined on $R_p^i \cup R_p^j$ where $q_p^{i,j}(\omega) = h\big(q_p^i(\omega), q_p^j(\omega)\big)$; recall that $q_p$ is set to 0 outside its region $R_p$. The transformation $h$ produces another QoI function from the constituent QoI functions which reflects how the quality of fused information is assessed. For example, if the accuracy of a measurement from provider $i$ at a point $\omega$ is 3% and from provider $j$ is 5%, the aggregated quality from the two providers could be the best of the two, i.e., 5%, i.e., "$h \equiv \max$." We use the latter example $h$ in our numerical results later on, thus for $\omega \in R_p^i \cup R_p^j$, we will use:

$$q_p^{i,j}(\omega) \stackrel{\text{def}}{=} h\big(q_p^i(\omega), q_p^j(\omega)\big) = \max\big\{q_p^i(\omega), q_p^j(\omega)\big\}. \quad (12)$$

Algorithm 1 can be implemented in polynomial time. At each iteration, the algorithm determines the optimal provider to select, but this may not necessarily lead to the optimal overall solution, which is similar to how the *greedy* algorithm behaves for the set covering problem.

The scenario described in this section did not take into account a possible cost for using third party sensory information. Problem formulation $\Pi_{nc}$ and its solution in Algorithm 1 identify the best subset of providers that maximize the aggregate spatial relevancy of information independently of the cost. Next we consider an additional model formulation that takes this cost into account when choosing the optimal provider set.

### *B. Maximum Relevancy with Budget Constraints*

Assuming that nothing comes for free, the city agency will have to face the realities of budgetary constraints sooner or later. In this case, suppose the city agency's budget is $B$ and the provider $i$'s cost is $c_i$, $i = 1, \ldots, N$. The cost $c_i$ could be a flat rate that the provider charges or a contracted price reflective of the attained relevancy $r_s(q_d, q_p^i)$; we will not delve further on this issue. Thus, we are now interested in finding the optimal set of providers that will maximize the spatial relevancy of the provided information subject to the budget constraint $B$. Again, this case can be modeled by a combinatorial optimization problem. Specifically, let again $I(i)$ be the 0-1 indicator variables for selecting provider $i$ and $\mathbf{I}$ the corresponding vector. Thus, the formulation of the optimization problem in this case will be:

**Problem $\Pi_{\mathbf{bg}}$**: For $I(i) \in \{0, 1\}$, $i \in \{1, \ldots, N\}$,

$$\text{maximize } r_s\big(q_d, q_p^{\mathbf{I}}\big), \text{ such that } \sum_{i=1}^{N} I(i) \cdot c_i \le B, \quad (13)$$

where $r_s\big(q_d, q_p^{\mathbf{I}}\big)$ is the relevancy of a "super-provider" with a QoI function aggregated from the providers indicated by selection vector $\mathbf{I}$, as discussed earlier in relation to (12), and defined on $R_p^{\mathbf{I}}$ in (9). We note that in $\Pi_{bg}$ the increase of the relevancy when adding a specific provider $i$ does not only depend on $i$ alone but on the already selected providers as well. In the case that the providers already selected are offering good enough quality on all points $\omega$ in $R_p^i$, adding provider $i$ may not increase the relevancy attained.

Problem $\Pi_{bg}$ is a generalization of the 0-1 *knapsack* problem [12] where the value of each item is a function of the items already selected to be included in the knapsack. For example, adding a lighter in the knapsack may reduce (even to zero) the subsequent value of a box of matches. This is captured with the use of $q_p^{\mathbf{I}}$ as a function of the vector $\mathbf{I}$. The 0-1 knapsack

problem is an *NP*-hard optimization problem which means that there is no algorithm that finds the optimal solution in polynomial time. The *greedy* algorithm would need to check all $2^N$ different combinations between the $N$ providers, prune those that do not satisfy the available *budget* and then choose the combination that maximizes the aggregate relevancy. A *dynamic programming* algorithm has been proposed that solves the problem in *pseudo-polynomial* time. The algorithm splits the main problem into smaller subproblems and stores some of the intermediate results for later use to speed up the calculation of the main problem [12]. However, this algorithm can not be directly applied to $\Pi_{bg}$ since the aggregated spatial relevancy when adding one provider depends also on the selection of other providers, as explained earlier. Nonetheless, Algorithm 2 has been developed to solve $\Pi_{bg}$ using the same idea of storing intermediate results.

As a dynamic programming algorithm, Algorithm 2 is trading memory space for time. In other words, it splits the problem into smaller subproblems, stores their solutions into memory, and, then, uses them to calculate the solution of the main problem. Algorithm 2 iteratively constructs the $N \times B$ matrix $\mathbf{Values}$, whose entry $Values[i, b]$ is the maximum aggregate spatial relevancy of the first $i$ providers for a budget $b$; the corresponding provider selections reside in the indicator vector $\mathbf{I}_i^b$. Entry $Values[N, B]$ stores the maximum aggregate spatial relevancy of all providers for budget $B$, which is the optimal solution for $\Pi_{bg}$ and the optimal provider selection will reside in the vector $\mathbf{I}_N^B$. As mentioned earlier, $\Pi_{bg}$ is an extended 0-1 knapsack problem with variable item value. Therefore, lines 6-9 of Algorithm 2 calculate the spatial relevancy (i.e., the "value") of the specific selection vector $\mathbf{I}$. The spatial relevancy of vectors $\mathbf{I}$ that have already been calculated at earlier iterations are evoked from memory. This has a significant impact in accelerating the algorithm. Moreover, lines 11-21 of the algorithm determine whether selecting a new provider will result in higher aggregate spatial relevancy, in which case the provider is selected, or not.

The dynamic programming algorithm for the 0-1 knapsack problem has complexity of $O(nB)$, where $n$ is the number of items and $B$ the available budget. In the worst case, Algorithm 2 will calculate the spatial relevancy $r_s\big(q_d, q_p^{\mathbf{I}}\big)$ at each iteration, which needs $O(N)$ time. Therefore, the absolutely worst case time complexity of Algorithm 2 is $O(N^2B)$, where $N$ is the total number of providers. Regarding the memory requirements, in the worst case, it is necessary to store the matrix $\mathbf{Values}$ of size $N \times B$, the relevancy values $r_s(q_d, q_p^{\mathbf{I}})$ for each selection vector $\mathbf{I}$, which are in total $\min\{2^N, N \times B\}$, and the optimal selection vector $\mathbf{I}_i^b$ of size $N$ for the $N \times B$ iterations of the algorithm. However, the implementation of the algorithm can be accelerated both in time and memory requirements significantly in two ways. First, instead of examining all $N$ providers the algorithm can be run only for those intersecting with the desired QoI function. The intersection operation will be run only once, at the beginning of the process, and can be implemented in linear time. Then, instead of iterating for all values in the range $[0, B]$, we can calculate the *greatest common divisor* $gcd$ of $c_i$, $i = 1, \cdots, N$ and $B$ and then run the algorithm in the range $[0, B/gcd]$ with costs $c_i/gcd$, $i = 1, \cdots, N$.

**Algorithm 2** – Budget Constrained Aggregate Relevancy

```
1:  for i = 1 to N do
2:    for b = 0 to B do
3:      if c_i ≤ b then
4:        I = I_{i-1}^{b-c_i}; where: I_0^{b-c_i} ≝ 0 and I_{i-1}^0 ≝ 0;
5:        I(i) = 1;
6:        if r_s(q_d, q_p^I) not calculated then
7:          Calculate r_s(q_d, q_p^I) using (4);
8:        else
9:          Get r_s(q_d, q_p^I) from memory;
10:       end if
11:       if r_s(q_d, q_p^I) > Values[i-1, b] then
12:         Values[i, b] = r_s(q_d, q_p^I);
13:         I_i^b = I;
14:       else
15:         Values[i, b] = Values[i-1, b];
16:         I_i^b = I_{i-1}^b;
17:       end if
18:     else
19:       Values[i, b] = Values[i-1, b];
20:       I_i^b = I_{i-1}^b;
21:     end if
22:   end for
23: end for
```

## V. Numerical Results

The QoI metadata construction scheme and the spatial relevancy calculation procedure were implemented in MATLAB. The analysis performed, shown next, verify the efficiency of the proposed schemes.

### A. Single-provider Spatial Relevancy

The objective of the single-provider study is assessing the robustness of the spline-based, finite-size approximation of QoI functions in ordering providers according to their relevancy to a desired QoI function.

Due to the ease by which they can flexibly approximate several shapes with respect to orientation, flatness, peak(s), etc., we constructed QoI functions using mixtures of the Gaussian density functions. The parameters of these shapes included their relative position on the plane, their maximum value and the number of Gaussian functions mixed. Note that the resulting QoI function does *not* posses the properties of a density function. These functions were approximated using B-splines (with $M = P+6$ parameters) which then used to calculate the relevancy of providers, see (7), and order them accordingly.

With regard to the regions $R$, we considered two topology cases: (*a*) a *rural* topology where the desired and the various provider regions are dispersed in an area, see figure 3; and (*b*) a *urban* topology where the desired and the various provider regions line-up along city streets (the "Manhattan street"

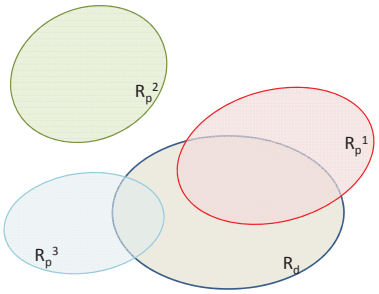


Fig. 3. The rural topology case.

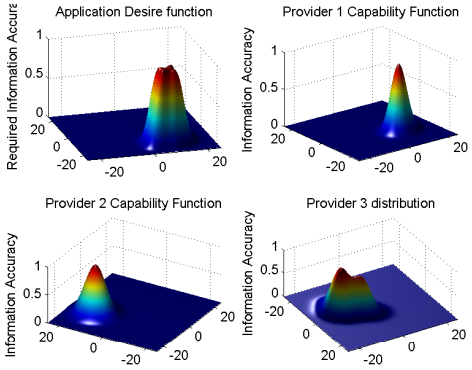


Fig. 4. QoI example functions for the rural topology.

topology), see figure 5. Figures 4 and 6 show QoI example functions for the two topology cases.

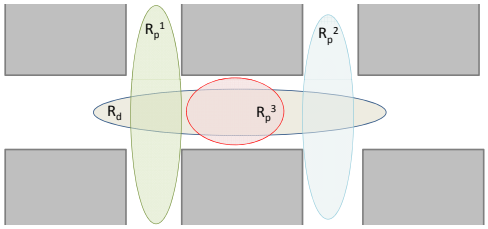


Fig. 5. The urban (Manhattan street) topology case.

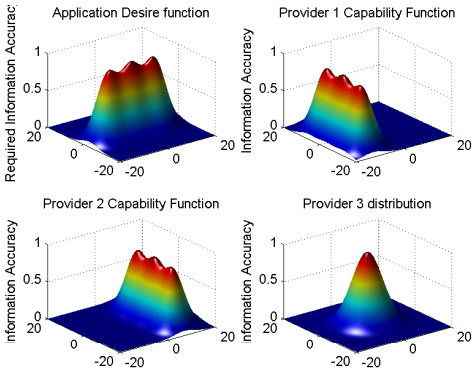


Fig. 6. QoI example for the urban topology.

We applied the single-provider relevancy method to these example cases using different values of $M$, the maximum number of B-spline parameters. Then, the spatial relevancy metric was calculated through the B-spline approximation and was compared against the actual spatial relevancy of the providers using their original QoI function. We studied: ($a$) the estimation error as a function of $M$; and ($b$) the effect of this error on ordering providers according to spatial relevancy. Note that the latter is what we are ultimately interested in. Specifically, the goodness of the approximation is judged not in absolute terms (which is a comparison over a continuum of values) but rather over an ordering outcome (which is a comparison over a finite set of alternatives).

The measurements presented in Figures 7 and 8 illustrate the robustness of the method with regard to this objective. As the top plot in Figures 7 and 8 show, the estimation error for the spatial relevancy of each provider is relatively low even when using around 100 parameters for the QoI function approximation. More importantly, there are no misordering effects even when the spatial relevancy of some providers is almost identical, as in the case of providers 1 and 2 for the urban topology case. This is indicated in the bottom plot in Figure 8 by the fact that the red and blue lines do not intersect; an intersection would mean a change in the relative order of provider relevancy.

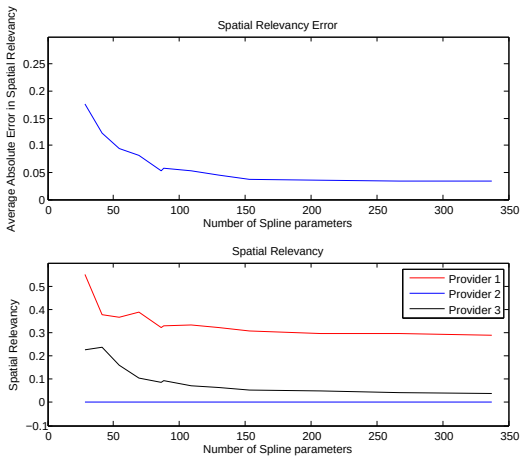


Fig. 7. Spatial relevancy for the rural topology.

We used (7) to compute the provider relevancy. A couple of comments are in order regarding these computations. The intersection of the desired and provided regions was computed using the fast algorithm to determine the intersection of convex polygons described in [13]. Due to the requirement for convex regions, the *convex hull* of (the generally) non-convex $R$ regions was calculated before applying the algorithm. However, this operation does not affect the value of the spatial relevancy.

The B-splines approximation and the spatial relevancy metric of (7) involved a sampling process of the continuous QoI function. We made use of the B-spline generation algorithm provided in MATLAB which minimized the squared-error at the sampling points; these were uniformly spaced on the respective regions $R$. This uniform sampling technique used during the two scenarios gave sufficiently good approximations even for the case that the special relevancy of providers

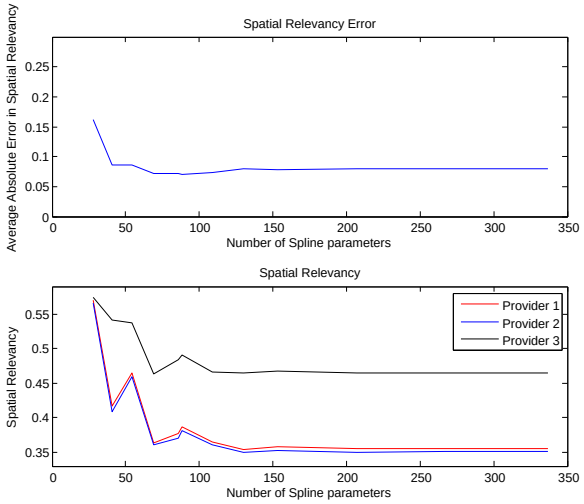


Fig. 8. Spatial relevancy for the urban topology.

was almost identical. However, as with most approximation methods, we would expect that non-uniform sampling would have improved performance especially for QoI functions that experience regions of significant and/or abrupt changes. An example of such a case would be a QoI function for which, given a region $R$, $q(\omega) = 1$ for $\omega \in R$ and 0 otherwise. In such a case, dedicating more samples around the boundaries of region $R$ would yield better results. Note that testing the performance of the spline-generation algorithms themselves is beyond the scope of this paper. We chose B-splines as a flexible, convenient and well-studied means to address our problem of describing QoI functions with finite, expansion-proof collection of parameters. Through our analysis study we confirmed that they are also a very effective aid in ranking relevant providers.

Building upon this procedure of calculating spatial relevancy for single providers, in the next subsection we will present the simulation results of the two algorithms proposed for the multi-provider composition problem.

### B. Multi-provider Spatial Relevancy

The two algorithms proposed to solve the *multi-provider composition* problems with or without the budget constraint were also simulated in MATLAB environment. Again, the QoI functions used were mixtures of varying number of Gaussian density functions, randomly scaled and placed on the two-dimensional plane. Figure 9 shows an example case, where the desired QoI function is colored in blue, and 9 providers are colored in red, cyan and green.

The proposed algorithms are based on pseudopolynomial heuristics to solve *NP-Hard* problems. These algorithms were adjusted to accommodate our objectives regarding the spatial relevancy of providers. Hence, the objective of our simulation study was the assessment of their effectiveness in selecting the right providers that satisfy problem $\Pi_{nc}$ and $\Pi_{bg}$ in Section IV. The assessment is performed by comparing the solutions

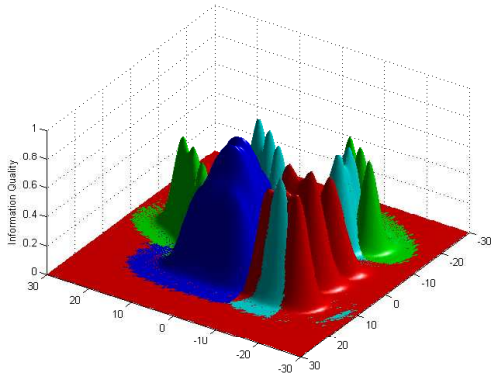


Fig. 9. QoI example functions for the multi-provider case.

and execution time of the proposed algorithms against those from the exhaustive search algorithm. For the no-cost case, the latter calculates the spatial relevancy of all $(2^N - 1)$ different combinations between the $N$ providers and the selection of the best one according to (10). For the budget constraint case, the exhaustive search algorithm includes the comparison of all *feasible* combinations, i.e., those with a total cost less or equal to the budget, and the selection of the optimal one among them according to (13).

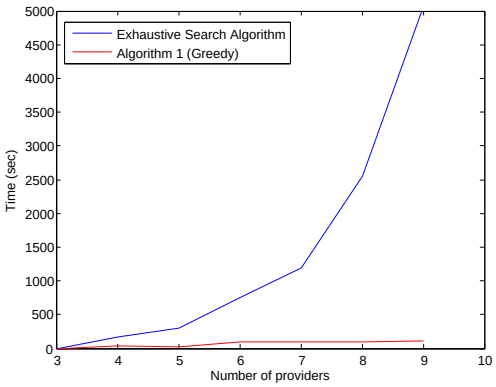


Fig. 10. Execution times for Algorithm 1 and exhaustive search.

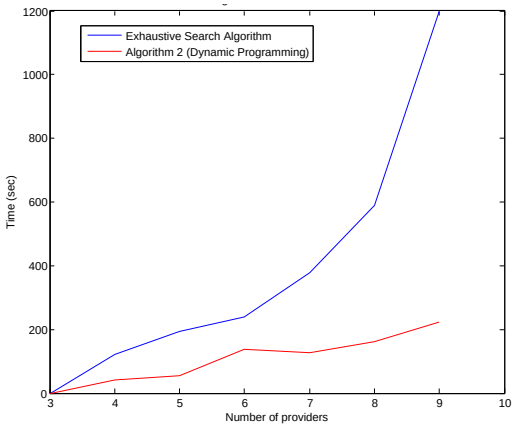


Fig. 11. Execution times for Algorithm 2 and exhaustive search.

Figures 10 and 11 show the comparison of the execution time between the proposed algorithms and the exhaustive method in each case. The simulations were run on a 2.4GHz dual core Windows PC with 4GB of RAM. For all cases studied, the solutions that the proposed algorithms arrived at were the same as the ones given by the exhaustive search methods, which of course are the optimal ones. As expected, the execution time of the exhaustive algorithms increases exponentially as the number of providers increases, while algorithms 1 and 2 need almost linear time. The execution of the proposed algorithms has also been accelerated by a mechanism of pruning providers not intersecting with the desired QoI function. In such cases, these providers are removed from the rest of the process with the result of further reducing the number of combinations examined.

## VI. Concluding remarks

In this paper, we introduced a novel problem area for sensor networks that of identifying and selecting sensory-information *relevant* providers based on their sensing capabilities in relation to an application's information needs along spatial dimensions. The focus in this paper has been in the spatial domain for ease of presentation, but temporal extensions are also possible. This problem will become more and more prominent as the number of providers increases and their sensing capabilities change in the spatiotemporal domain, such as when using wireless and mobile sensor networks operating over a multi-administrative domains, e.g., vehicle-mounted sensors, participatory sensors, etc. Within this area we derived a relevancy metric based on the concept of QoI functions that describe the desirability or quality levels of the information desired or produced at a given location. We then developed a finite, expansion-proof technique based on B-spines to describe QoI functions and use those to advertise desired and provided sensing capabilities. The use of expansion-proof descriptors raises from the need for predictable resource usage (e.g., storage, communications) especially when considering composite providers built from the aggregation of other "regional" providers. Finally, we have formulated related optimization problems and proposed efficient algorithms for selecting the best single, or multiple collection of providers that are most relevant to our needs given various constraint objectives.

To the best of our knowledge this is the first endeavor in this area, and there is no prior work closely related to this paper. There is, however, prior literature that inspired and influenced our research. Specifically, supplementing our own cited work on QoI, [14] discusses quality metadata describing geospatial information. Ref. [7] provides an extensive review of the models for spatio-temporal information databases and related queries. Ref. [15] describes a process for merging topological maps where the possibility of the unbounded increase of metadata/parameters becomes evident. Granted, our case is not equivalent to merging topological maps, yet the underlying problem of metadata explosion still exists whenever we compose behaviors (the QoI functions) defined over different spatiotemporal horizons. See also [16] which deals with building and manipulating maps described by simple rectangles. Ref. [17] considers summarizing 2D shapes via a bounded number of parameters. These shapes could correspond to our regions $R$ and, thus, the proposed approach in [17] could serve as an alternative to our B-spline approach. We do not discount the latter approach and could have been used in our paper as well. However, given that we ultimately pursue a comparison and selection of relevant providers, we found the use of the B-spline approach more flexible. Finally, our inspiration in using splines comes from [18] which considers the explosion of time-decaying security metadata of documents produced by the combination of contributing documents.

Future work in this novel area, may include the study of the various architectural aspects related to QoI function advertisements eluded earlier in the paper as well as the explicit consideration of time-varying QoI functions that could result by system impediments, such as loss of sensors, and fluidity of sources, such as in participatory sensing.

## References


[1] N. Gershenfeld, R. Krikorian, and D. Cohen, "The Internet of Things," *Scientific American*, vol. 291, no. 4, pp. 76–81, October 2004.

[2] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *World Sensor Web Workshop (in ACM Sensys'06)*, Boulder, CO, USA, October 31 2006.

[3] C. Bisdikian, J. Branch, K. K. Leung, and R. I. Young, "A letter soup for the quality of information in sensor networks," in *IEEE Information Quality and Quality of Service (IQ2S'09) Workshop (in IEEE PerCom'09)*, Galveston, TX, USA, Mar. 2009.

[4] C. Bisdikian, L. M. Kaplan, M. B. Srivastava, D. J. Thornley, D. Verma, and R. I. Young, "Building principles for a quality of information specification for sensor information," in *12th Intl Conf. on Information Fusion (FUSION'09)*, Seattle, WA, USA, July 2009.

[5] C. H. Liu, C. Bisdikian, J. W. Branch, and K. K. Leung, "Qoi-aware wireless sensor network management for dynamic multi-task operations," in *IEEE SECON'10*, Boston, MA, USA, June 2010.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2006.

[7] N. Pelekis, B. Theodoulidis, I. Kopanakis, and Y. Theodoridis, "Literature review of spatio-temporal database models," *The Knowledge Engineering Review*, vol. 19, no. 3, pp. 235–274, September 2004.

[8] H. Prautzsch, W. Boehm, and M. Paluszny, *Bezier and B-Spline Techniques*. Springer-Verlag, 2002.

[9] C. de Boor, *A Practical Guide to Splines*. Springer, 2001.

[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. MIT Press, 2009, pp. 85–103.

[11] R. Karp, *Reducibility Among Combinatorial Problems*. Plenum Press, 1972, pp. 85–103.

[12] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, 1990.

[13] J. O'Rourke, C.-B. Chien, T. Olson, and D. Naddor, "A new linear algorithm for intersecting convex polygons," *Computer Graphics and Image Processing*, vol. 19, no. 4, pp. 384–391, 1982.

[14] R. Devillers, Y. Bdard, and R. Jeansoulin, "Multidimensional management of geospatial data quality information for its dynamic use within gis," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 2, pp. 205–215, February 2005.

[15] W. H. Huang and K. R. Beevers, "Topological map merging," *The Int'l J. of Robotics Research*, vol. 24, no. 8, pp. 601–613, August 2005.

[16] W. Xue, Q. Luo, L. Chen, and Y. Liu, "Contour map matching for event detection in sensor networks," in *ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'06)*, Chicago, IL, USA, June 2006.

[17] J. Hershbergery, N. Shrivastava, and S. Suriz, "Summarizing spatial data streams using clusterhulls," in *8th Wksp on Algorithm Engineering and Experiments (ALENEX'06)*, Miami, FL, USA, Jan. 2006.

[18] M. Srivatsa, D. Agrawal, and S. Reidt, "A metadata calculus for secure information sharing," in *16th ACM Conference on Computer and Communications Security (CCS'09)*, Chicago, IL, USA, Nov. 2009.