Metrics for Evaluating Feature-Based Mapping Performance

Pablo Barrios, Martin Adams, *Senior Member, IEEE*, Keith Leung, *Member, IEEE*, Felipe Inostroza, *Member, IEEE*, Ghayur Naqvi, and Marcos E. Orchard, *Member, IEEE*

***Abstract*—In robotic mapping and simultaneous localization and mapping, the ability to assess the quality of estimated maps is crucial. While concepts exist for quantifying the error in the estimated trajectory of a robot, or a subset of the estimated feature locations, the difference between all current estimated and ground-truth features is rarely considered jointly. In contrast to many current methods, this paper analyzes metrics, which automatically evaluate maps based on their joint detection and description uncertainty. In the tracking literature, the optimal subpattern assignment (OSPA) metric provided a solution to the problem of assessing target tracking algorithms and has recently been applied to the assessment of robotic maps. Despite its advantages over other metrics, the OSPA metric can saturate to a limiting value irrespective of the cardinality errors and it penalizes missed detections and false alarms in an unequal manner. This paper therefore introduces the cardinalized optimal linear assignment (COLA) metric, as a complement to the OSPA metric, for feature map evaluation. Their combination is shown to provide a robust solution for the evaluation of map estimation errors in an intuitive manner.**

***Index Terms*—Map metric, simultaneous localization and mapping, mobile robots.**

I. Introduction

FUNDAMENTAL to any state estimation problem is the concept of estimation error. Solutions to robotic mapping and simultaneous localization and mapping (SLAM), in which usually the location of an unknown number of features should be estimated, are numerous offering various degrees of performance. Examples include classical methods such as recursive EKF SLAM [1], [2], multihypothesis (MH) FastSLAM [3], batch estimation methods such as GraphSLAM [4] and iSAM [5], [6] and random finite set (RFS) methods [7], [8]. Irrespective of the estimation methods used, while clear concepts exist for quantifying the error in the estimated pose or trajectory of a robotic vehicle [9] and/or a *subset* of the estimated feature locations within a single reference frame [2], the absolute difference between *all* currently estimated and ground-truth (GT) features in the map is rarely considered jointly. In SLAM, this is of equal importance to the vehicle trajectory estimate.

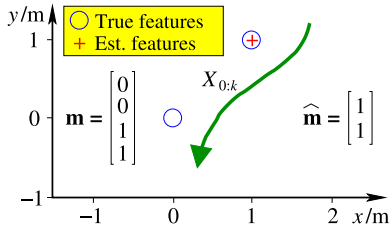


Fig. 1. Hypothetical scenario showing a fundamental inconsistency with vector representations of feature maps. If $\mathbf{m}$ is the GT map (blue circles), how should the error be assigned when the number of features in the map estimate, $\hat{\mathbf{m}}$, (red cross) is incorrect?

Previous methods, which have been used to compare estimated and GT maps, include the root mean squared (RMS) and normalized estimation error squared (NEES) metrics [10]–[13]. These metrics are only defined if the number of estimated and GT features are the same. In realistic scenarios, these numbers will differ, meaning that a *subset* of at least one of the map's features must be used.

To illustrate the dilemma in map quality estimation, recall that in vector-based SLAM formulations, the map is constructed by stacking features into a vector, and consider the simplistic scenario depicted in Fig. 1. There are two true features at (0,0) and (1,1) (blue circles), but due to a missed detection, the estimated map (red cross) comprises only one feature at (1,1). In such a situation, it is difficult to define a mathematically consistent error metric (Euclidean error, RMS, NEES) between the vectors $\mathbf{m}$ and $\hat{\mathbf{m}}$ since they contain a different number of features.

Examining the spatial errors between subsets of the estimated and GT features, for example, by "pruning" the vector with the higher dimensionality using map management heuristics, followed by data association (DA), then allows standard error metrics such as the average Euclidean, RMS, or NEES errors to be evaluated. This technique may illustrate the consistency of the spatial state of the selected features, but gives no indication of the quality of the estimate of the actual multifeature map

Manuscript received May 11, 2016; revised October 6, 2016; accepted November 1, 2016. Date of publication December 23, 2016; date of current version February 3, 2017. This paper was recommended for publication by Editor Carme Torras and Associate Editor Josep M. Porta upon evaluation of the reviewers' comments. This work was supported in part by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under Grant FA9550-15-1-0069, in part by the "Becas Conicyt–Doctorado Nacional, 2012 and 2014," in part by the Conicyt–Fondecyt project 1150930, in part by the CSIRO–CORFO Project 10CEII-9007-F1-L3-P2, and in part by the AMTC and Conicyt Anillo Project ACT 1120, Chile.

P. Barrios, F. Inostroza, G. Naqvi, and M. Orchard are with the Department of Electrical Engineering, Universidad de Chile, Santiago 837-0451, Chile (e-mail: pbarrios@ing.uchile.cl; finistro@ug.uchile.cl; ghayur.naqvi@ing.uchile.cl; morchard@u.uchile.cl).

M. Adams is with the Department of Electrical Engineering, Universidad de Chile, Santiago 837-0451, Chile, and also with the Advanced Mining Technology Center, Universidad de Chile, Santiago 837-0451, Chile (e-mail: martin@ing.uchile.cl).

K. Leung is with Applanix, Trimble, Richmond Hill, ON L4B 3B3, Canada (e-mail: kleung@applanix.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2016.2627027

state. The primary difficulty in mathematically defining map estimation error is caused by the differences between the estimated and true *number* of features, and the need to satisfy the four metric axioms [14].[1]

The quality of the estimated map is important in the robotic global planning phase. Missed detections could cause the robot to plan paths through a space that it believes is obstacle free. When the robot reaches a point near to the formerly missed detection, if it is then detected, it would be forced to replan its path, which in the worst case could require global replanning. If it still remains undetected, a collision may result. On the other hand, false alarms could prevent the robot from computing an optimal path, or in fact any path, which reaches its goal. Therefore, this paper suggests that a concise map estimation metric, which penalizes estimators for both feature detection as well as description (of which location is an example) errors, is important, but lacking in the literature.

In [15], it is noted that many robotic mapping performance measures are not true metrics, since in order to obtain such a measure, which is intuitive, at least one of the metric axioms is violated. To demonstrate the basic problem of precisely quantifying feature-based mapping error, Fig. 2 shows posterior map estimates from two different feature-based SLAM filters, referred to as SLAM Algorithms 1 and 2, respectively. A natural question is: Which estimate is closer to GT? Visual intuition is difficult due to the combination of missed detections, false alarms and spatial errors. Metrics to answer this fundamental question are addressed in this paper.

In the target tracking literature, a similar dilemma exists in which the performance of trackers must be assessed, which led to the development of multiobject[2] metrics, based on sets. It will be shown, that multiobject metrics developed in the target tracking community, which consider both multitarget state estimation cardinality as well as state errors and which obey the metric axioms, provide a basis for gauging feature-based maps in an intuitive manner. Set-based metrics are analyzed and compared for the evaluation of feature-based maps, and a new method, called the cardinalized optimal linear assignment (COLA) metric, is introduced, and compared to its optimal subpattern assignment (OSPA) counterpart. This paper demonstrates various mathematical properties of each metric, and their consequences on their "meaningful physical interpretations," when comparing maps under a common reference frame.

It should be noted that the metrics presented in this paper are not, in their current form, suitable for evaluating nonfeature-based maps, such as those produced by some SLAM algorithms. They are designed to evaluate estimated-to-estimated or estimated-to-GT maps, in the same frame of reference, containing any features, which can be defined in terms of their attributes/descriptions (e.g., geometric, color, semantic) in both maps under comparison.

[1]The four metric axioms can be defined as follows. Let $\mathcal{X}$ be an arbitrary, nonempty set. Then the function d is a metric iff: 1) $d(\mathbf{x},\mathbf{y}) \geq 0$, for all $\mathbf{x},\mathbf{y} \in \mathcal{X}$; 2) $d(\mathbf{x},\mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$, for all $\mathbf{x} \in \mathcal{X}$ (identity axiom); 3) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$, for all $\mathbf{x},\mathbf{y} \in \mathcal{X}$ (symmetry axiom); 4) $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$, for all $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{X}$ (triangle inequality axiom).

[2]In the tracking community, a vector-based estimator is often referred to as a *single-object* estimator, whereas the RFS form is referred to as a *multiobject* estimator.

This paper is an extension of [16], in which the COLA metric was first presented. It extends the analysis by deriving the COLA metric from the Wasserstein construction; providing a theoretical comparison of the way in which false alarms and missed detections are quantified by the OSPA and COLA metrics; deriving the map conditions which can cause them to differ; and providing more in depth results and conclusions demonstrating the complementary nature of the OSPA and COLA metrics for useful robotic map evaluation.

Section II discusses related work and Section III provides the mathematical definitions necessary to quantify feature maps, for metric-based evaluation. Section IV overviews the state-of-the-art OSPA metric and Section V introduces the COLA metric. Section VI, analyzes the two metrics under different theoretical conditions. Section VII discusses the applicability of mapping metrics when no GT is available, and finally, Section VIII analyzes various metrics under real-world mapping scenarios.

II. Related Work

Contrary to feature-based mapping approaches, occupancy grid methods use a predefined quantity of occupancy values and do not need to encapsulate cardinality estimation error—i.e., an estimated grid map with N cells is typically compared with a GT grid map, also with N cells. In these cases, error metrics for fixed dimension problems, such as a sum of the squared error [17], [18], or more complex errors such as "brokenness," [19] can be applied. Other work provides methods to gauge grid-based SLAM or mapping performance in robotic competitions such as search and rescue and RoboCup [20], [21].

In other mapping approaches, [22] presented the concept of assessing map quality through a binary classification of point cloud data by automatically labeling sections as plausible or suspicious, through the use of conditional random fields, but without relating the data to GT.

Various dataset websites such as [23] can compare SLAM algorithms based on the error in the corrected SLAM trajectory component, without analyzing the corresponding map. A metric based on the energy required to deform the estimated trajectory to its GT value was defined in [9]. This metric is based on an intuitive concept for comparing estimated and GT trajectories, and complies with the metric axioms. Two further trajectory-based error metrics, which gauge visual SLAM and odometry systems based on RGB-D camera data were presented in [24]. It will, however, be shown in this paper that these metrics can contradict principled multiobject metrics that evaluate the *map* accuracy of the SLAM algorithm, which produced both the trajectory and the map.

In [25], a method for comparing maps suggested the use of the Hausdorff metric. Although this metric has been successfully applied in numerous vision applications for gauging the similarity between pixelated images, it suffers various problems in the way it penalizes cardinality errors between estimated and GT maps, as will be demonstrated in this paper.

A feature-based map quality assessment method, based on searching for nearest neighbor (NN) equivalent features between an estimated and GT map was given in [26]. Map quality was assigned according to the number of associated point to point

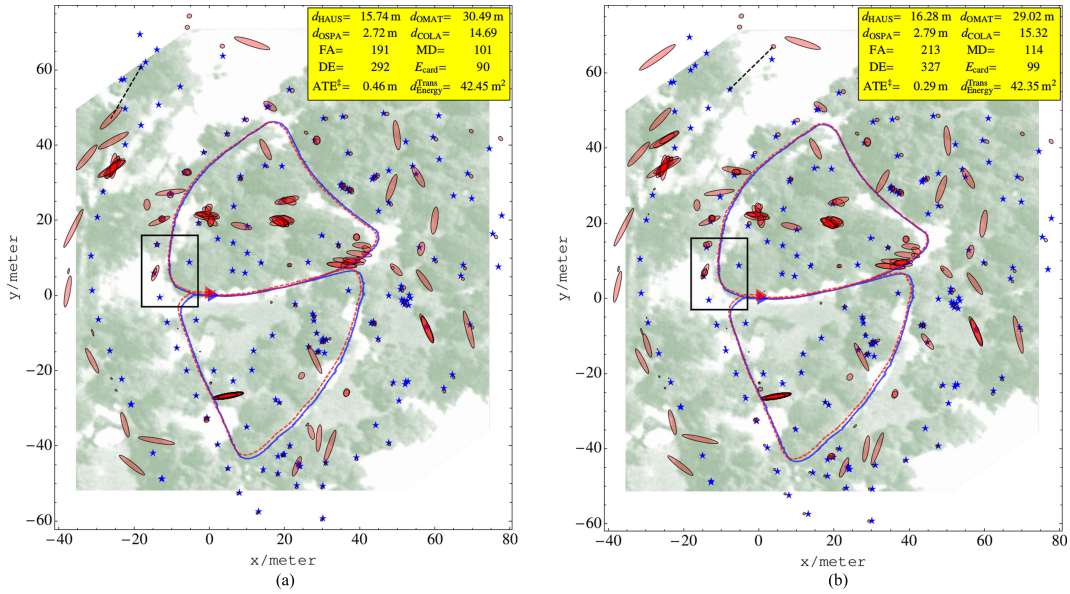


Fig. 2. GT and estimated trajectories from SLAM Algorithms 1(a) and 2(b). The GT feature locations (blue stars) are superimposed onto a satellite image of the park. The estimated maps produced by SLAM Algorithms 1(a) and 2(b) are shown as red "3-sigma" confidence interval ellipses. The dashed line in each figure represents the Hausdorff distance between each map set. "FA" = no. false alarms, "MD" = no. missed detections, "DE" = no. detection errors (= MD + FA) and $E_{\text{card}} = ||\widehat{\mathcal{M}}| - |\mathcal{M}||$. ATE‡ and $d^{\text{Trans}}_{\text{Energy}}$ give the trajectory-based RMS ATE and energy metric values respectively.

feature matches, however, no penalties for false alarms and missed detections were considered.

The need for feature map metrics has also been identified in publicly available datasets, such as the European Union's FP6 Rawseeds project [27], where corner-based performance metrics are suggested for comparing estimated and GT maps.

Since the direct comparison of vectors of features, estimated by SLAM algorithms, is meaningless, recent work in SLAM has suggested that a collection of map features can and should be modeled as a set, rather than a vector [7], [8]. Indeed the mathematical definitions of the four metric axioms apply to a *set* of vectors and it will be demonstrated in this paper that if a GT set-based map $\mathcal{M}_k$ and its set-based estimate $\widehat{\mathcal{M}}_k$, which vary with discrete time k, are modeled as *finite sets* of feature location vectors, then a mathematically consistent notion of estimation error is possible, even when the number of estimated and GT features differ. This is because, the "distance," or error between sets, is a well understood concept. Examples include the Hausdorff distance and more recent metrics defined in the multitarget tracking literature, such as the optimal mass transfer (OMAT) [28] and OSPA [14] distances.

The pioneering work by [14] introduced the OSPA metric, based on the concepts that it should:

1) be a metric on the space of finite sets;
2) have a natural (meaningful) physical interpretation;
3) capture cardinality errors and state errors meaningfully;
4) be easily computed.

This philosophy is continued in this paper, by analyzing current as well as the proposed COLA set-based metrics.

III. Map Set Definitions

Modeling a robotic feature map as a set $\mathcal{M}_k$, provides a general model, since the vectors within each set can contain any (spatial and/or description) information of relevance to the type of feature to be estimated. Throughout this paper, the GT map set $\mathcal{M}_k$ is considered to contain $\mathfrak{m}_k$ vectors $\mathbf{m}^i_k$, $1 \leq i \leq \mathfrak{m}_k$. For ease of notation and explanation, and without loss of generality, $\mathbf{m}$ will be referred to as a *spatial* variable and the time index k is now dropped. Therefore, the GT map $\mathcal{M}$ is the set $\{\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^{\mathfrak{m}}\}$, where $\mathbf{m}^i \in \mathbb{R}^N \ \forall i : 1 \leq i \leq \mathfrak{m}$. The estimated map $\widehat{\mathcal{M}}$ is considered to contain $\widehat{\mathfrak{m}}$ vectors modeling the location of map objects, i.e., $\widehat{\mathcal{M}} = \{\widehat{\mathbf{m}}^1, \widehat{\mathbf{m}}^2, \ldots, \widehat{\mathbf{m}}^{\widehat{\mathfrak{m}}}\}$, where $\widehat{\mathbf{m}}^i \in \mathbb{R}^N \ \forall i : 1 \leq i \leq \widehat{\mathfrak{m}}$. N is the dimension of the general feature space. Note that $\widehat{\mathfrak{m}}$ is itself *an estimate* of $\mathfrak{m}$, and therefore, in general, $\widehat{\mathfrak{m}} \neq \mathfrak{m}$ ($|\widehat{\mathcal{M}}| \neq |\mathcal{M}|$).

IV. OSPA Metric

In 2008, Schuhmacher *et al.* [14] introduced the OSPA metric, which obeys the metric axioms and improves most of the problems of the OMAT metric, as will be shown in Section VI. It has recently gained popularity in gauging target tracking [29]–[32] as well as robotic mapping performance [7], [8].

A. Definition of the OSPA Metric

The OSPA metric is based on the inner metric

$$d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)}) = \min(c, d(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})) \tag{1}$$

where $d(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})$ is any metric distance (e.g., Euclidean, Mahalanobis, Hellinger) between $\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)}$ and $c > 0$ is the cutoff parameter. The assignment $\sigma(i)$ can be determined via an optimal assignment method, such as the Hungarian method, [33], [34]. Then, the OSPA metric $d^{(c,p)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}})$ with power p, for $\widehat{\mathfrak{m}} > \mathfrak{m}$, is defined as

$$d^{(c,p)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}) = \left(\frac{1}{\widehat{\mathfrak{m}}} \min_{\sigma} \sum_{i=1}^{\widehat{\mathfrak{m}}} d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})^p \right)^{1/p} \tag{2}$$

$$= \left(\frac{1}{\widehat{\mathfrak{m}}} \left(\min_{\sigma} \sum_{i=1}^{\mathfrak{m}} d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})^p + c^p(\widehat{\mathfrak{m}} - \mathfrak{m}) \right) \right)^{1/p} \tag{3}$$

σ is a permutation of the set $\{1,\ldots,\mathfrak{m}\}$ which minimizes $(\sum_{i=1}^{\mathfrak{m}} d^{(c)}(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})^p)$ and $1\leq p<\infty$. If both sets are empty, $\mathfrak{m}=\widehat{\mathfrak{m}}=0$, $d_{\text{OSPA}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})=0$. For $\mathfrak{m}>\widehat{\mathfrak{m}}$, the metric is defined as $d_{\text{OSPA}}^{(c,p)}(\widehat{\mathcal{M}},\mathcal{M})$.

From (3), the OSPA metric yields a measure of the difference between $\widehat{\mathcal{M}}$ and $\mathcal{M}$ in units of distance. From (3) and (1), $d_{\text{OSPA}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})$ has minimum value zero and saturates to a maximum value c for all $\mathcal{M}$ and $\widehat{\mathcal{M}}$. The effect of, and concepts for selecting c and p are discussed in [14].

V. COLA Metric

The OSPA metric saturates to c when all of the localization errors are larger than c, and is then, insensitive to the size of the map cardinality error. Further, it will be shown that it penalizes missed detections and false alarms in an unequal manner and that its evaluations of mapping errors can be very sensitive to the choice of c. Due to the above limitations, this section explains a new set-based metric for feature map evaluation.

A. Definition of the COLA Metric

The COLA metric can be derived from the Wasserstein construction as shown in Appendix A. For $\widehat{\mathfrak{m}}\geq\mathfrak{m}$, it is defined as

$$d_{\text{COLA}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})=\left(\min_{\sigma}\sum_{i=1}^{\widehat{\mathfrak{m}}}\left(\frac{d^{(c)}(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})}{c}\right)^p\right)^{1/p} \tag{4}$$

$$=\left(\min_{\sigma}\sum_{i=1}^{\mathfrak{m}}\left(\frac{d^{(c)}(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})}{c}\right)^p+(\widehat{\mathfrak{m}}-\mathfrak{m})\right)^{1/p} \tag{5}$$

where, σ, p, and c carry the same definitions and ranges as in the OSPA metric and $d^{(c)}(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})$ is defined in (1). For $\mathfrak{m}>\widehat{\mathfrak{m}}$, the metric is defined as $d_{\text{COLA}}^{(c,p)}(\widehat{\mathcal{M}},\mathcal{M})$. Appendix B proves that $d_{\text{COLA}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})$ is a true metric.

B. Intuitive Explanation of the COLA Metric

Whereas the OSPA metric has the units of localization error (i.e., distance), the COLA metric has the units of the cardinality error (i.e., no units). In contrast to the OSPA metric, when the distance between an assigned feature i and feature $\sigma(i)$ *decreases* to c it changes from a cardinality error to a fractional cardinality error $(d^{(c)}(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})/c)^p$. Although the difference between the OSPA and COLA metrics may seem trivial, Sections VI and VIII will demonstrate significant differences in the intuitive behavior of the COLA metric over its OSPA counterpart, when evaluating feature maps.

As in the OSPA metric, for $\widehat{\mathfrak{m}}>\mathfrak{m}$, the first term of the right-hand side (RHS) of (5) again determines individual assignments between all $\mathfrak{m}$ of the feature location vectors within $\mathcal{M}$ and a subset of dimension $\mathfrak{m}$ of the feature vectors within $\widehat{\mathcal{M}}$. Now, however, due to (1), this term is fractional if $d(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})<c$ or unity otherwise, in which case it is effectively added as a single cardinality error on to the pure cardinality error $(\widehat{\mathfrak{m}}-\mathfrak{m})$ on the RHS of (5). In contrast to the OSPA metric, it can be seen from (5) and (1) that $d_{\text{COLA}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})$ has minimum value zero and maximum value $(\widehat{\mathfrak{m}})^{1/p}$ if $\widehat{\mathfrak{m}}>\mathfrak{m}$ or $(\mathfrak{m})^{1/p}$ otherwise.

1) Interpreting the Components of the COLA Metric: In [14], it was shown that the OSPA error contains two components, which separately account for localization and cardinality errors. This also applies to the COLA metric with individual components $d_{\text{LOC}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})$ and $d_{\text{CARD}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})$ as

$$d_{\text{LOC}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})=\left(\min_{\sigma}\sum_{i=1}^{\mathfrak{m}}\left(\frac{d^{(c)}(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})}{c}\right)^p\right)^{1/p} \tag{6}$$

$$d_{\text{CARD}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})=(\widehat{\mathfrak{m}}-\mathfrak{m})^{1/p}. \tag{7}$$

Similarly to the note in [14], the functions $d_{\text{LOC}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})$ and $d_{\text{CARD}}^{(c,p)}(\mathcal{M},\widehat{\mathcal{M}})$ themselves are not strict metrics on the space of finite subsets and the decomposition of the metrics should not usually be necessary for gauging mapping performance. Equations (6) and (7), however, can be evaluated to provide extra information regarding the contributions of description and cardinality mapping errors.

The selection of the COLA metric's parameters c and p, and their physical interpretation, must also be addressed.

2) Effect of p: In a similar manner to the OSPA metric [14], as p increases, the emphasis on localization errors diminishes. Therefore, the COLA metric also becomes more unforgiving to cardinality errors for higher values of p. Based on the COLA metric form, given in (4), Appendix C shows that for the same value of c, the COLA metric is also ordered with respect to p, however, contrary to the OSPA metric

$$d_{\text{COLA}}^{(c,p_1)}(\mathcal{M},\widehat{\mathcal{M}})\geq d_{\text{COLA}}^{(c,p_2)}(\mathcal{M},\widehat{\mathcal{M}})\text{ for }1\leq p_1<p_2<\infty. \tag{8}$$

Choosing $p=1$ also makes the COLA metric behave in a relatively simple manner, in that it is then composed of the sum of cardinality and localization errors, which can be interpreted as fractional cardinality errors. Selecting $p=2$ also makes the COLA metric more in-line with other metrics, which are often $L2$-norms, and this value will be used throughout this paper.

3) Effect of c: Analysis of the COLA metric form in (5) shows that as $c\to\infty$, the COLA metric becomes only sensitive to cardinality errors. Therefore, as in the OSPA metric, increasing c emphasizes the cardinality errors. However, contrary to the OSPA metric, from (5)

$$d_{\text{COLA}}^{(c_1,p)}(\mathcal{M},\widehat{\mathcal{M}})\geq d_{\text{COLA}}^{(c_2,p)}(\mathcal{M},\widehat{\mathcal{M}})\text{ for }1\leq c_1<c_2<\infty. \tag{9}$$

Similarly to the OSPA metric [14], c should be chosen based on "What distance (e.g., how many meters) the designer wants to penalize a false or missing estimate," which in any application should significantly aid its practical choice. It should also be noted that the single object metric $d^{(c)}(\mathbf{m}^i,\widehat{\mathbf{m}}^{\sigma(i)})$ does not have to be a distance (e.g., Euclidean) metric—but could be the Mahalanobis (statistical) distance. In this case, c could be chosen as a validation gate, corresponding to a probability interval within a chi-square test, within which an estimated feature is considered to correspond to a GT feature. Note the link here with classical DA methods. c can be thought of as a gate. If the "distance" between an estimated and its assigned GT feature is

more than c, it contributes maximum value (a single cardinality error) to (5), equivalent to being "unassociated."

4) Setting a Threshold for the COLA Metric: A natural question which arises in the use of the COLA metric is: What threshold should be used to establish whether or not an estimated map is good? In contrast to the OSPA metric, the COLA metric does not saturate to a limiting value (c). In the case of $p = 1$, the COLA metric yields the total cardinality error between two set-based maps, which can be made up of fractional cardinality errors (assigned features which do not exactly coincide in terms of their attributes) and integer cardinality errors, due to a difference in the number of estimated and GT features. Therefore, a threshold to determine the difference between a good and bad map can be determined in terms of the number of effective cardinality errors, or outliers, one is willing to tolerate for a given application.

C. Computational Complexity of the COLA Metric

Since metrics are often determined offline, after experiments have terminated, during the testing phases of mapping or SLAM algorithms, their computational complexity is usually not of major concern. During the normal online use of such an algorithm, a GT (or other reference) map would typically not be available, and a metric would not be required. However, if in a particular application, the metric is to continuously gauge mapping/SLAM performance during algorithm execution, its complexity is important. For the COLA and OSPA metrics, their computational complexities are equivalent and dependent on the assignment method used to determine $\sigma(i)$. The Jonker and Volgenant algorithm is an efficient assignment method, which is, in general, faster than the Hungarian method, and has worst case cubic complexity in the dimension of the distance matrix $d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})^p$, yielding a computational complexity of $\mathcal{O}(\max(\mathfrak{m}, \widehat{\mathfrak{m}})^3)$ [14], [34], [35]. Other computational speed-ups of the Hungarian method are available, which are reported to reduce the execution time in linear assignment problems by up to 90% [36]. Recently, further optimizations of the Jonker and Volgenant algorithm have been reported [37].

VI. Analyses of Each Metric's Interpretation

The physical interpretation of the OSPA and COLA metrics is now analyzed and compared for particular and general cases of the maps $\mathcal{M}$ and $\widehat{\mathcal{M}}$, to highlight the usefulness of each metric in assigning a meaningful score to map estimators. Schuhmacher *et al.* [14] demonstrated theoretical map scenarios, which showed that the Hausdorff and OMAT metrics, although sensitive to spatial errors in the estimated map, are both insensitive to the cardinality error between $\mathcal{M}$ and $\widehat{\mathcal{M}}$. In those theoretical scenarios, it was shown that the OSPA metric provides intuitive results in terms of its sensitivity to both spatial and cardinality errors, and in [16], it was shown that the COLA metric provides intuitive results also. Therefore, in this section, in order to highlight differences between the OSPA and COLA metrics, and to study realistic mapping results, scenarios will be analyzed in which at least one of the maps is empty; one of the maps contains multiple features which are imbalanced[3] with respect to the other map; and one of the maps contains outliers. An analysis of their penalizations of false alarms and missed detections will be provided as well as the theoretical conditions under which they can disagree in their comparisons of mapping performance.

To simplify the ensuing analyses here, $d(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})$ will be the Euclidean distance metric. To demonstrate their generality, actual SLAM performance evaluations in Section VIII will apply the COLA and OSPA metrics, in which $d(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})$ is the Mahalanobis distance, allowing the incorporation of estimated feature covariance information, as often provided by SLAM algorithms.

A. Non Empty Set Versus Empty Set

Consider a GT map $\mathcal{M} = \emptyset$ and its estimate $\widehat{\mathcal{M}} = \{\widehat{\mathbf{m}}^1, \ldots, \widehat{\mathbf{m}}^{\widehat{\mathfrak{m}}}\}$ or vice versa. This corresponds to either a region containing no GT features or an empty estimated map, which could occur when no features are detected. The Euclidean, Hausdorff, and OMAT metrics are all undefined in this case, since both sets must be nonempty. Meanwhile, the OSPA metric is given by

$$d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = \left(\frac{1}{\widehat{\mathfrak{m}}} c^p (\widehat{\mathfrak{m}} - 0)\right)^{1/p} = c \tag{10}$$

and in this case, the COLA metric yields

$$d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = (\widehat{\mathfrak{m}} - 0)^{1/p} = \widehat{\mathfrak{m}}^{1/p}. \tag{11}$$

Both metrics demonstrate a desirable asset since a metric should be defined when one of the sets is empty. However, the COLA metric can be considered to provide a more intuitive result. Irrespective of the difference in cardinality, the OSPA metric gives the same score (c) and is insensitive to this difference, whereas the COLA metric increases with $\widehat{\mathfrak{m}}$. For $p = 1$, the COLA metric increases linearly with $\widehat{\mathfrak{m}}$ which is the true value of the cardinality error in this case.

B. Multiple GT and Estimated Features

Consider the COLA metric's performance with GT map $\mathcal{M} = \{\mathbf{m}^1, \ldots, \mathbf{m}^{\mathfrak{m}}\}$ and estimated map

$$\begin{aligned}\widehat{\mathcal{M}} = \{&\widehat{\mathbf{m}}^{1,1}, \ldots, \widehat{\mathbf{m}}^{1,q}, \ldots, \widehat{\mathbf{m}}^{\mathfrak{m}-1,1}, \ldots, \widehat{\mathbf{m}}^{\mathfrak{m}-1,q}, \\ &\widehat{\mathbf{m}}^{\mathfrak{m},1}, \ldots, \widehat{\mathbf{m}}^{\mathfrak{m},q-s}\}\end{aligned} \tag{12}$$

(i.e., the GT landmarks $\mathbf{m}^1$ to $\mathbf{m}^{(\mathfrak{m}-1)}$ have in their neighborhoods q estimates whereas GT landmark $\mathbf{m}^{\mathfrak{m}}$ has in its neighborhood $q - s$ estimates, making the estimated map imbalanced

[3]The concept balanced and imbalanced maps was addressed in [14]. The term "balanced" map refers to each feature in the GT map having in its proximity the same number of estimated (possibly > 1) features. Although an exactly balanced map is unlikely to occur in practice, it demonstrates the effect of an estimator producing multiple feature estimates per true feature, which is possible in high clutter scenarios, under which map management routines can initiate false features. The ability of a metric to provide intuitive assessments in such situations is therefore important. The term "imbalanced" is defined in Section VI-B.

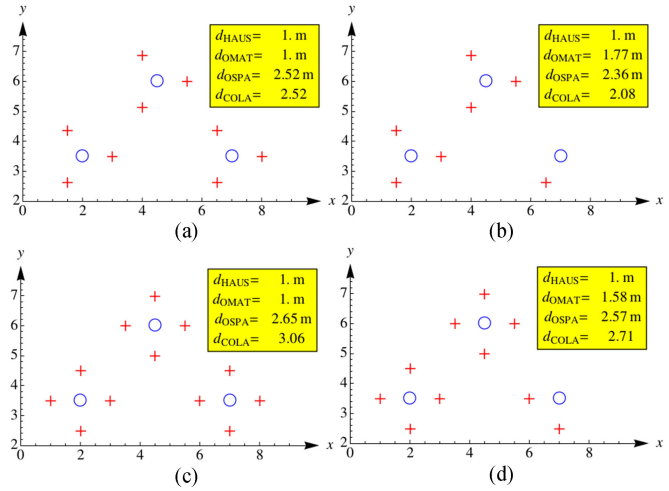


Fig. 3. Metric performances when cardinality errors exist. (a) and (c) are balanced maps while (b) and (d) are imbalanced. The distance between the center of each GT landmark (blue circles) and its neighboring estimates (red crosses) is 1 m. For calculating d_{OSPA} and d_{COLA}, $c = 3.00$ m and $p = 2$.

[14]). The subset $\mathcal{M}' = \{\mathbf{m}^1, \ldots, \mathbf{m}^{\mathfrak{m}-1}\} \in \mathcal{M}$ is balanced and

$$\begin{aligned} d(\mathbf{m}^i, \widehat{\mathbf{m}}^{i,l}) &= d \leq c,\ 1 \leq i \leq \mathfrak{m}-1,\ 1 \leq l \leq q \\ d(\mathbf{m}^i, \widehat{\mathbf{m}}^{j,l}) &> d \quad \forall \quad i \neq j, l \\ d(\mathbf{m}^{\mathfrak{m}}, \widehat{\mathbf{m}}^{\mathfrak{m},l}) &= d \leq c, \quad 1 \leq l \leq q-s \\ d(\mathbf{m}^i, \mathbf{m}^j) &> 2d \quad \forall \quad i \neq j. \end{aligned} \tag{13}$$

The general imbalanced scenario described in (12) yields a COLA metric value

$$d^{(c,p)}_{\text{COLA:imbal}}(\mathcal{M}, \widehat{\mathcal{M}})^p = \left(\mathfrak{m}\left(\frac{d}{c}\right)^p + (\mathfrak{m}(q-1) - s)\right)^{1/p}. \tag{14}$$

Comparing the COLA metric values for the cases of balanced ($s = 0$) and imbalanced maps, it can be seen that $d^{(c,p)}_{\text{COLA:bal}}(\mathcal{M}, \widehat{\mathcal{M}})^p \geq d^{(c,p)}_{\text{COLA:imbal}}(\mathcal{M}, \widehat{\mathcal{M}})^p$ for $s \geq 0$ and vice versa for $s < 0$, which complies with intuition. This can be seen in Fig. 3 since the estimates in Fig. 3(b) and (d) are better than those of Fig. 3(a) and (c) because they have less cardinality errors but the same spatial errors. In general, and in the examples portrayed in Fig. 3, these cases are not reflected intuitively in the Hausdorff and OMAT metric values.

It should also be noted that, in contrast to the OSPA metric, true to its nature of having units of cardinality error, for $p = 1$, $d^{(c,p)}_{\text{COLA:bal}}(\mathcal{M}, \widehat{\mathcal{M}})^p - d^{(c,p)}_{\text{COLA:imbal}}(\mathcal{M}, \widehat{\mathcal{M}})^p = s$, yielding the exact cardinality error between the maps.

C. Outliers

Consider an estimated map with one outlier, i.e., $\mathcal{M} = \{\mathbf{m}^1, \ldots, \mathbf{m}^{\mathfrak{m}}\}$ and $\widehat{\mathcal{M}} = \{\widehat{\mathbf{m}}^1, \ldots, \widehat{\mathbf{m}}^{\mathfrak{m}+1}\}$, and assume that every single GT landmark has a perfect estimate, i.e., $d(\mathbf{m}^i, \widehat{\mathbf{m}}^i) = 0$ for $1 \leq i \leq \mathfrak{m}$. From (3), the OSPA metric is

$$d^{(c,p)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}})^p = c\left(\frac{1}{\mathfrak{m}+1}\right)^{1/p}. \tag{15}$$

If $\mathfrak{m} \to \infty$, the OSPA metric $d^{(c,p)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}})^p \to 0$ distance units, giving an intuitive result.

In the case of the COLA metric,

$$d^{(c,p)}_{\text{COLA}}(\mathcal{M}, \widehat{\mathcal{M}})^p = 1 \tag{16}$$

which, since this metric yields cardinality, as opposed to average distance units, can also be considered to be intuitive. In this case, there is a single outlier, and the COLA metric correctly reports it. It should be noted here that the COLA metric is somewhat unforgiving to cardinality errors. For example, if a mapping algorithm estimates a large number of perfectly located estimates, with just one false alarm, the COLA metric always penalizes the algorithm, even though as $\mathfrak{m} \to \infty$, the algorithm can be argued to be approaching perfection. In this sense, the OSPA metric behaves more intuitively. It should be noted, however, that this does not cause any problems in the use of the COLA metric, when estimated maps are compared to the *same GT, or other algorithm's map*, as carried out at the end of mapping/SLAM tasks, since the same strict penalization of cardinality errors is applied to all estimates. However, when assessing mapping performance *during* SLAM execution, as more of the GT map passes through the field(s) of view of the sensor(s), the COLA metric would never lower its value due to previous outliers, even if the map estimator approaches perfection later on. This would not allow a useful online evaluation of mapping performance. In this case, due to the per feature averaged distance nature of the OSPA metric, it provides the most intuitive evaluation of the time varying estimated map.

D. Penalization of False Alarms and Missed Detections

A theoretical analysis will now demonstrate that the OSPA and COLA metrics penalize over and underestimated cardinality errors (false alarms and missed detections) in a significantly different manner. Consider a GT map $\mathcal{M} = \{\mathbf{m}^1, \ldots, \mathbf{m}^{\mathfrak{m}}\}$, and two map estimates, $\widehat{\mathcal{M}}_1 = \{\widehat{\mathbf{m}}^1, \ldots, \widehat{\mathbf{m}}^{\mathfrak{m}-\mathfrak{a}}\}$ and $\widehat{\mathcal{M}}_2 = \{\widehat{\mathbf{n}}^1, \ldots, \widehat{\mathbf{n}}^{\mathfrak{m}+\mathfrak{a}}\}$, where $\mathfrak{m} \geq \mathfrak{a}$. Estimate $\widehat{\mathcal{M}}_1$ underestimates the map size by $\mathfrak{a}$ and estimate $\widehat{\mathcal{M}}_2$ overestimates it by the same amount, $\mathfrak{a}$. Suppose also that $\widehat{\mathcal{M}}_1$ and a subset of $\widehat{\mathcal{M}}_2$, have no localization errors, i.e., $\widehat{\mathbf{m}}^{\sigma(i)} = \mathbf{m}^i$ for $1 \leq i \leq \mathfrak{m}-\mathfrak{a}$ and $\widehat{\mathbf{n}}^{\pi(i)} = \mathbf{m}^i$ for $1 \leq i \leq \mathfrak{m}$, where $\sigma(i)$ and $\pi(i)$ are the assignments used in comparing maps $\widehat{\mathcal{M}}_1$ with $\mathcal{M}$ and $\widehat{\mathcal{M}}_2$ with $\mathcal{M}$, respectively. The OSPA errors for both cases are

$$d^{(c,p)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}_1) = c\left(\frac{\mathfrak{a}}{\mathfrak{m}}\right)^{1/p} \tag{17}$$

$$d^{(c,p)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}_2) = c\left(\frac{\mathfrak{a}}{\mathfrak{m}+\mathfrak{a}}\right)^{1/p} \tag{18}$$

implying that, independent of parameters c and p, the OSPA metric penalizes the estimate with the missed detections *more* than the estimate with the same number of false alarms.

On the other hand, the COLA metric for both cases yields

$$d^{(c,p)}_{\text{COLA}}(\mathcal{M}, \widehat{\mathcal{M}}_1) = d^{(c,p)}_{\text{COLA}}(\mathcal{M}, \widehat{\mathcal{M}}_2) = \mathfrak{a}^{1/p} \tag{19}$$

implying that the COLA metric penalizes both estimates by the same amount. Arguably, the COLA metric can be considered to

provide a more intuitive error estimate of both mapping errors, since neither the OSPA nor the COLA metrics are designed to judge them in an unequal manner. The consequences of missed detections and false alarms on robot navigation are application specific, and judging one type of error to be worse than the other is beyond the scope of these metrics alone.

E. Can the COLA and OSPA Metrics Differ?

This section demonstrates the conditions under which the COLA and OSPA metrics disagree. Consider a GT map $\mathcal{M}$ with cardinality $\mathfrak{m}$ and its estimates $\widehat{\mathcal{M}}_1$ and $\widehat{\mathcal{M}}_2$ with cardinalities $\widehat{\mathfrak{m}}_1$ and $\widehat{\mathfrak{m}}_2$, respectively, where

$$\widehat{\mathfrak{m}}_1 \leq \mathfrak{m} < \widehat{\mathfrak{m}}_2 \tag{20}$$

$$\text{and} \quad d_{\text{OSPA}}^{(c,p)}(\widehat{\mathcal{M}}_1, \mathcal{M}) \geq d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_2). \tag{21}$$

The COLA metric values for both cases are then

$$d_{\text{COLA}}^{(c,p)}(\widehat{\mathcal{M}}_1, \mathcal{M}) = \frac{\mathfrak{m}^{1/p}}{c} d_{\text{OSPA}}^{(c,p)}(\widehat{\mathcal{M}}_1, \mathcal{M}) \tag{22}$$

$$d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_2) = \frac{\widehat{\mathfrak{m}}_2^{1/p}}{c} d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_2). \tag{23}$$

For the metrics to disagree, it is necessary that

$$d_{\text{COLA}}^{(c,p)}(\widehat{\mathcal{M}}_1, \mathcal{M}) < d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_2) \tag{24}$$

which requires that

$$\frac{d_{\text{OSPA}}^{(c,p)}(\widehat{\mathcal{M}}_1, \mathcal{M})}{d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_2)} < \left(\frac{\widehat{\mathfrak{m}}_2}{\mathfrak{m}}\right)^{1/p}. \tag{25}$$

A similar analysis shows that if

$$\mathfrak{m} \leq \widehat{\mathfrak{m}}_1 < \widehat{\mathfrak{m}}_2 \tag{26}$$

$$\text{and} \quad d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_1) \geq d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_2) \tag{27}$$

the metrics can also disagree if

$$\frac{d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_1)}{d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}_2)} < \left(\frac{\widehat{\mathfrak{m}}_2}{\widehat{\mathfrak{m}}_1}\right)^{1/p}. \tag{28}$$

Hence, disagreement between both metrics is only possible if (20), (21) and (25) or (26), (27), and (28) are simultaneously satisfied, as will be demonstrated in Section VIII-C.

Note that if $\widehat{\mathfrak{m}}_1,\ \widehat{\mathfrak{m}}_2 \leq \mathfrak{m}$, both metrics always agree.

VII. When GT is Unavailable

There are cases in SLAM in which reliable GT maps are unavailable or incomplete, such as dense visual-based maps. This paper therefore analyzes metrics which compare *all* of the features estimated by a mapping or SLAM algorithm, either with those produced by another algorithm, or GT. In the absence of GT, a possibility for evaluation is to select the most robust known algorithm X, and then, if the mapping performance of algorithms Y and Z are to be compared, determine which is closer to X. Further, when GT is unavailable, the triangular inequality can be exploited under certain conditions, to at least provide bounds for the mapping error between estimated and GT maps. Assume that two estimated maps $\widehat{\mathcal{M}}_1$ and $\widehat{\mathcal{M}}_2$ are available and that the GT map is $\mathcal{M}$. If it is known that estimated map $\widehat{\mathcal{M}}_2$ is closer to GT than estimated map $\widehat{\mathcal{M}}_1$ then

$$d(\widehat{\mathcal{M}}_2, \mathcal{M}) = \alpha d(\widehat{\mathcal{M}}_1, \mathcal{M}) \text{ where } 0 \leq \alpha < 1. \tag{29}$$

Any chosen true metric $d(\widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2)$, which compares sets $\widehat{\mathcal{M}}_1$ and $\widehat{\mathcal{M}}_2$ can then be used to determine upper bounds on the error between $\widehat{\mathcal{M}}_1$ and $\mathcal{M}$ and $\widehat{\mathcal{M}}_2$ and $\mathcal{M}$. From the triangular inequality

$$d(\widehat{\mathcal{M}}_1, \mathcal{M}) \leq d(\widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2) + d(\widehat{\mathcal{M}}_2, \mathcal{M}). \tag{30}$$

From (29) and (30)

$$d(\widehat{\mathcal{M}}_1, \mathcal{M}) \text{ and } d(\widehat{\mathcal{M}}_2, \mathcal{M}) \leq \left(\frac{1}{1-\alpha}\right) d(\widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2) \tag{31}$$

meaning that upper bounds on the mapping error between $\widehat{\mathcal{M}}_1$ and $\mathcal{M}$ and $\widehat{\mathcal{M}}_2$ and $\mathcal{M}$ can be determined in terms of the error between the map estimates of the two algorithms.

If a third map estimation algorithm is introduced which yields $\widehat{\mathcal{M}}_3$, the above analysis can be extended to give

$$d(\widehat{\mathcal{M}}_3, \mathcal{M}) \leq d(\widehat{\mathcal{M}}_3, \widehat{\mathcal{M}}_1) + \left(\frac{1}{1-\alpha}\right) d(\widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2). \tag{32}$$

Note that since a GT map $\mathcal{M}$ is not required in the RHS of (31) and (32), the map feature representations are not restricted to ones knowledge of GT, and can include dense (e.g., visual) feature representations. Note here the importance of the chosen map metric being a true metric, such that the 4 axioms are obeyed. It is also desirable that α be as small as possible.

From here on, comparisons will be referred to as being between estimated and GT maps, although the concepts equally apply to comparisons of multiple map estimates.

VIII. Results—Evaluations With Real SLAM Maps

In this section, the performance of the Hausdorff, OMAT, OSPA, and COLA metrics, as well as the SLAM trajectory energy metric of [9] and absolute trajectory error (ATE) metric of [24], will be analyzed by comparing their ability to score real estimated SLAM results in a physically meaningful manner. The effect of varying c will also be further elucidated.

Fig. 2 shows the SLAM estimates from two different SLAM algorithms,[4] which were designed to estimate 2-D vehicle trajectories and maps corresponding to the x, y location of the centers of tree trunks. Each algorithm is referred to as "SLAM Algorithm 1" and "SLAM Algorithm 2" and each result is superimposed onto a satellite image of the area[5] to show the tree coverage and spacing corresponding to the vehicle's "eight-shaped" trajectory. The GT trajectories (blue lines) were obtained via manual scan-matching[6] due to the lack of reliable GPS in the

[4]The estimated SLAM solutions are based on MH-FastSLAM [3] and Rao–Blackwellized-PHD-SLAM [7].

[5]The experiments were carried out in Parque O'Higgins, Santiago, Chile.

[6]This corresponds to manually identifying points in successive scans from identifiable tree trunks, and determining the corresponding vehicle displacement to align such points within the global coordinate system.

environment, and the red dashed lines represent the estimated trajectories from each SLAM algorithm. The blue stars represent the GT features (tree trunk center locations), again obtained through independent, manual scan matching procedures.[7] The red ellipses are centered at the estimated feature locations and represent the covariances for each estimated feature. The ellipses shown correspond to "three-Sigma ellipses," which from two degree-of-freedom Chi-squared tables, correspond to a theoretical probability mass within each ellipse of 0.99. Hence, $d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})$ in (3) can be the Mahalanobis distance such that

$$d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)}) = \min\left(c, \sqrt{(\mathbf{m}^i - \widehat{\mathbf{m}}^{\sigma(i)})^T (\mathbf{P}^i)^{-1} (\mathbf{m}^i - \widehat{\mathbf{m}}^{\sigma(i)})}\right) \tag{33}$$

where $c = 3.00$ and $\mathbf{P}^i$ is the sum of the error covariance submatrices corresponding to estimated feature $\widehat{\mathbf{m}}^{\sigma(i)}$ and GT feature $\mathbf{m}^i$. In this analysis, it is assumed that the error associated with all GT features is zero[8] and that the estimated feature covariance values are available from the SLAM estimator. In each figure, the Hausdorff (d_{HAUS}), OMAT (d_{OMAT}), OSPA (d_{OSPA}), COLA (d_{COLA}), numbers of false alarms (FA), missed detections (MD), detection errors (DE = FA + MD), and estimated cardinality errors ($E_{\text{card}} = ||\widehat{\mathcal{M}}| - |\mathcal{M}||$) are provided. In all of the experiments, $p = 2$, which according to [14], yields smooth distance curves, and is commonly used in other metrics, such as the L2 distance.

A. Performance of the Trajectory Energy Metric

The energy metric proposed in [9], which calculates the energy required to deform the estimated (red dashed) trajectory $\widehat{\mathbf{X}}_{0:k}$ to the GT (blue solid) trajectory $\mathbf{X}_{0:k}$, can be calculated based on the trajectories shown in Fig. 2. In Fig. 4, the values of the translational component of the energy metric, $d^{\text{Trans}}_{\text{Energy}}(\mathbf{X}_{0:k}, \widehat{\mathbf{X}}_{0:k})$ over time, given the GT trajectory, during each SLAM run are shown. The metric is evaluated based on the average energy needed to deform adjacent components of nearby poses between the estimated and GT trajectories, to determine the local consistency of each trajectory as suggested in [9]. The final values of $d^{\text{Trans}}_{\text{Energy}}(\mathbf{X}_{0:k}, \widehat{\mathbf{X}}_{0:k})$, and in fact the ATE metric, for SLAM Algorithms 1 and 2 suggests that SLAM Algorithm 2 slightly outperforms SLAM Algorithm 1. Interestingly, this contradicts the final OSPA, COLA, and Hausdorff metric evaluations, shown in Fig. 2 and also contradicts most of the time varying OSPA and COLA metric values recorded during the SLAM runs, shown in Fig. 5 for two different values of c. This demonstrates the necessity of gauging mapping as well as trajectory performance.

[7]The GT map is also of limited precision since the laser data, used for scan matching, is prone to range errors of up to 5 cm, and more importantly, the determination of the centers of the trees was prone to errors of up to 20 cm. Importantly, these distance errors are significantly less than the average mapping distance errors and the GT map is guaranteed to contain all circular objects detectable by the algorithm used here.

[8]In the target tracking literature, [32] applied the OSPA metric based on a Hellinger distance metric, in which the GT target covariances were replaced with their Cramer–Rao lower bound values.

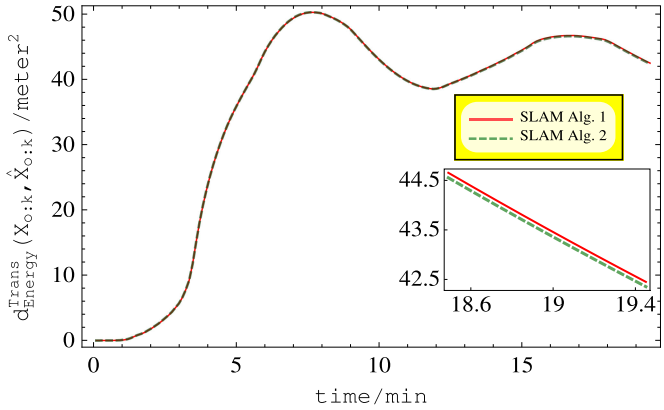


Fig. 4. Translational component of the energy metric applied to SLAM Algorithm 1 (Red line) and SLAM Algorithm 2 (Green dashed line).

B. Performance of the Multiobject Metrics

Referring to Fig. 5, it is important to note that both the OSPA and COLA map error evaluations are based on the time varying number of features which have theoretically been covered by the vehicle's sensor's field of view (GT) and those which have been estimated at each time. Setting the Mahalanobis distance-based cutoff parameter $c = 0.25$ in both metrics, Fig. 5(a) shows that during most of the SLAM trial, the OSPA metric for both algorithms saturates to its maximum value $c = 0.25$. This indicates that most of the estimated spatial errors are larger than 0.25 m within both SLAM estimates. It should be noted, however, that under these circumstances, the OSPA metric fails to give any indication as to which algorithm is superior, even in terms of the detection (map cardinality) errors. Conversely, however, the corresponding COLA metric plots [see Fig. 5(b)] show that although the first 8 min of each SLAM trial perform similarly, after this time, it is possible to define which algorithm performs better (usually SLAM Algorithm 1) despite most estimates having Mahalanobis distances larger than $c = 0.25$. This is due to the COLA metric's ability to continue gauging cardinality errors, even when the OSPA metric saturates.

In Fig. 5(c) and (d), the OSPA and COLA metrics are again plotted versus time, during the same SLAM experiments, but this time with a cutoff parameter $c = 3.0$. Now, as shown in [14], the values of the OSPA metric increase for both SLAM algorithms. However, now that more features are assigned to (gated with) GT values, the curves in Fig. 5(c) remain unsaturated, and provide a significant difference in their judgment of SLAM Algorithms 1 and 2, at all times. Note that as time increases, the effective field of view of the sensed area has increased. The COLA metric, in Fig. 5(d), shows its tendency to increase as it is unforgiving to an increase in cardinality error. On the other hand, the OSPA metric, in Fig. 5(c) shows its averaging nature as the number of erroneous feature estimates increases with increasing sensor coverage.

Table I shows the number of gated features (Gtd.), Missed Detections (MD) and False Alarms (FA) and d_{COLA} and d_{OSPA} based on a validation gate $c = 3.00$. d_{HAUS}, d_{OMAT}, the NEES mapping error and the trajectory $d^{\text{Trans}}_{\text{Energy}}$, and ATE errors are also given. The NEES metric, undefined for the full estimated and

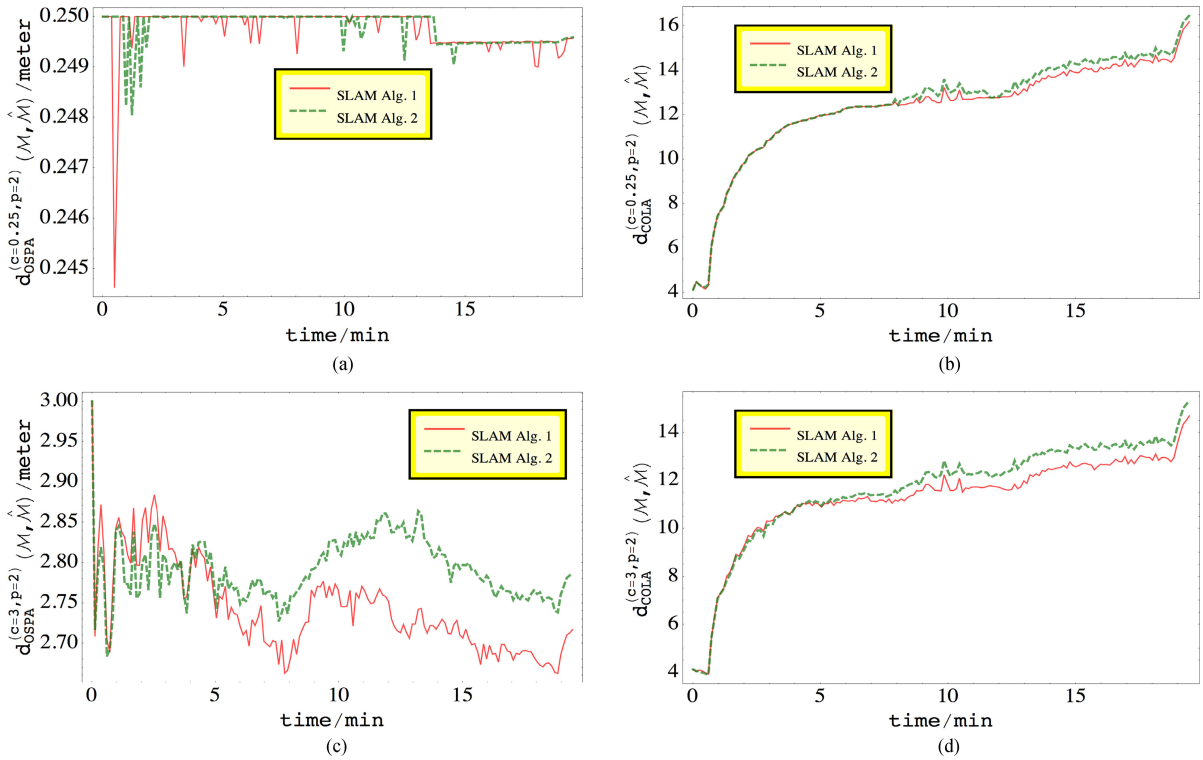


Fig. 5. OSPA and COLA metrics versus time for SLAM Algorithms 1 and 2, with differing values of c. (a) OSPA metric versus time ($c = 0.25$ and $p = 2$). (b) COLA metric versus time ($c = 0.25$ and $p = 2$). (c) OSPA metric versus time ($c = 3.00$ and $p = 2$). (d) COLA metric versus time ($c = 3.00$ and $p = 2$).

TABLE I
DETECTION ERRORS AND MAP METRIC RESULTS FOR SLAM ALGORITHMS 1 AND 2 ($c = 3.00$ AND $p = 2$)

Alg.	$\widehat{\mathfrak{m}}$	Gtd.	MD	FA	d_{COLA}	d_{OSPA} [m]	d_{HAUS} [m]	d_{OMAT} [m]	NEES*	NEES†	$d^{\text{Trans}}_{\text{Energy}}$ [m^2]	ATE‡ [m]
1	263	72	101	191	14.69 (1)	2.72 (1)	15.74 (1)	30.49 (2)	231.33 (2)	3.213(2)	42.45 (2)	0.464 (2)
2	272	59	114	213	15.32 (2)	2.79 (2)	16.28 (2)	29.02 (1)	159.34 (1)	2.701(1)	42.35 (1)	0.285 (1)

$\widehat{\mathfrak{m}}$ = number of estimated features, "Gated" = number of gated features, "MD" = number of Missed Detections, "FA" = number of False Alarms. *The NEES metric was calculated only based on the number of gated features, ignoring detection errors. †The NEES metric averaged per gated feature. ‡RMS values were used for the ATE metric. The bracketed numbers give the ranking by each metric.

GT maps since their cardinalities are different, has been calculated based purely on the gated features (for $c = 3.00$). The NEES metric can only be calculated for the 59 gated, out of the 272 estimated, features in SLAM Algorithm 2, and ignores the remaining estimates, even though they constitute detection errors. For SLAM Algorithm 1, the NEES metric gave a higher value, purely because more features were gated (72). Note that both $d^{\text{Trans}}_{\text{Energy}}$, the NEES and ATE metrics, often used in gauging SLAM performance, favor SLAM Algorithm 2, disagreeing with the multiobject metrics d_{COLA}, d_{OSPA}, and d_{HAUS}.

In light of Table I, SLAM Algorithm 1 has less detection errors and more gated features and d_{COLA}, d_{OSPA}, and even d_{HAUS} all confirm its mapping superiority over SLAM Algorithm 2.

C. Sensitivity of the OSPA and COLA Metrics With Respect to c

To highlight the sensitivity of the OSPA and COLA metrics to parameter c, the estimated and GT submaps in the boxes in Fig. 2(a) and (b) are analyzed for various values of c. Comparing the smaller submaps simplifies the assessment of the COLA and OSPA metrics' *intuitive performances*. In these boxes, the number of GT features is $|\mathcal{M}| = 4$, the number of estimates for SLAM Algorithm 1 is $|\widehat{\mathcal{M}_1}| = 3$ and that for the SLAM Algorithm 2 is $|\widehat{\mathcal{M}_2}| = 5$. Under these conditions, it will be shown that there are values of c where (20), (21), and (25) are satisfied, meaning that the OSPA and COLA metrics can disagree.

Fig. 6(a) and (b) shows the values of the OSPA and COLA metrics, for the submaps within Fig. 2(a) and (b), for $p = 2$ and varying c. As shown in [14], d_{OSPA} increases with c, whereas, as shown in (9), d_{COLA} decreases. The discontinuities in the gradients of the curves in Fig. 6(a) and (b) occur when estimated features become gated. The metric evaluations of SLAM Algorithms 1 and 2, based on this submap, now follow for three values of c.

1) Fig. 7(a) and (d) shows the performances of SLAM Algorithms 1 and 2, and their metric values, when $c = 0.15$. Each estimated feature, labeled f_1, f_2, etc., is surrounded by its validation gate ellipsoid, corresponding

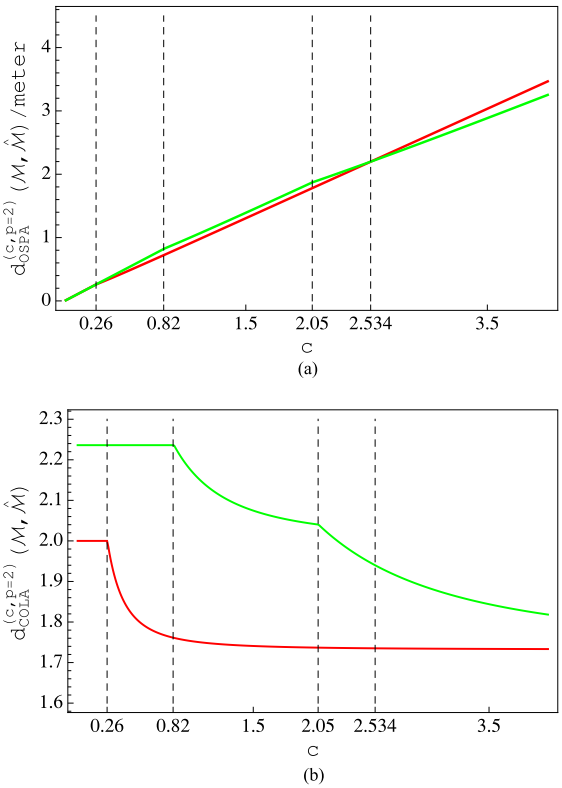


Fig. 6. OSPA and COLA metrics versus c. The red and green curves shows the metric values for SLAM Algorithms 1 and 2 respectively. (a) OSPA metric versus c, for $p = 2$. (b) COLA metric versus c, for $p = 2$.

to the region in which the Mahalanobis distance $\sqrt{(\mathbf{m}^i - \hat{\mathbf{m}}^{\sigma(i)})^T (\mathbf{P}^i)^{-1} (\mathbf{m}^i - \hat{\mathbf{m}}^{\sigma(i)})} < c$ (33). Note that for $c = 0.15$, these ellipses are barely visible, but will be for larger values of c. Since all of the GT features, labeled g_1, g_2, etc., are ungated, the map estimation result of SLAM Algorithm 1 constitutes four missed detections (g_1, g_2, g_3 and g_4) and three false alarms (f_1, f_2, and f_3), i.e., a total of seven detection errors. The result of SLAM Algorithm 2 constitutes four missed detections (g_1, g_2, g_3, and g_4) and five false alarms (f_1, f_2, f_3, f_4, and f_5), i.e., a total of nine detection errors. For both SLAM Algorithms 1 and 2, OSPA saturates to its cutoff value $c = 0.15$ m and is unable to differentiate between the two. The COLA metric (intuitively) penalizes SLAM Algorithm 2 more than SLAM Algorithm 1, showing its dependence on the cardinality of the larger set.

2) Fig. 7(b) and (e) shows the case when $c = 2.20$. Note the significantly larger validation gates, centered on each estimate. Two feature estimates (f_2 and f_5) are gated by SLAM Algorithm 2, implying a total of two missed detections (g_3 and g_4) and three false alarms (f_1, f_3, and f_4). Interestingly, this matches the total number of detection errors exhibited by SLAM Algorithm 1 (three missed detections (g_2, g_3, and g_4) and two false alarms (f_1 and f_3)). For SLAM Algorithm 1, the COLA metric has almost settled to a steady-state value wrt, c, as indicated by the red curve in Fig. 6(b).
3) Fig. 7(c) and (f) show the case when $c = 3.00$. In the case of SLAM Algorithm 1, its estimate f_2 that was gated when $c = 2.20$ becomes statistically better localized (due to a more tolerant, larger validation gate), however, three missed detections (g_2, g_3, and g_4) and two false alarms (f_1 and f_3) remain. For SLAM Algorithm 2, two of its previously gated estimates (f_2 and f_5) become statistically better localized as c increases, still leaving two missed detections (g_3 and g_4) and three false alarms (f_1, f_3, and f_4), i.e., still the same cardinality error as SLAM Algorithm 1. Interestingly, as verified in Fig. 6, the OSPA metric now reverses its decision, favoring SLAM Algorithm 2 over SLAM Algorithm 1, in contrast to the COLA metric. Note that only in this case are (20), (21), and (25) simultaneously satisfied, indicating disagreement between the OSPA and COLA metrics. It can be seen from Fig. 6, that the OSPA metric decisions are very sensitive to small changes in c, when features are gated. Initially, for $c \leq 0.26$, $d^{(c,p=2)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}) = c$, and as c increases, a small number of features start to gate, causing the OSPA metric to only lower its value very slightly below c. Since $d^{(c,p=2)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}_1)$ and $d^{(c,p=2)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}_2)$ are both very similar, and close to c for most values of c, the slight gradient changes (see Fig. 6) which occur when features gate, result in nonintuitive changes in the mapping performance decisions of the OSPA metric. This problem is avoided in the COLA metric, as can be seen in Fig. 6(b).

D. Unequal Penalization of False Alarms and Missed Detections

In the limit, as $c \to \infty$, the three estimates in SLAM Algorithm 1 will gate with three of the four GT features. Similarly, for SLAM Algorithm 2, four of its five estimates will all gate with the four GT features. Therefore, (3) shows that for SLAM Algorithm 1, $d^{(c,p=2)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}_1) \to (1/\sqrt{4})c$ and for SLAM Algorithm 2, $d^{(c,p=2)}_{\text{OSPA}}(\mathcal{M}, \widehat{\mathcal{M}}_2) \to (1/\sqrt{5})c$, demonstrating the averaging ability of the OSPA metric, since for SLAM Algorithm 1, one out of four of the GT features remains unassigned and for SLAM Algorithm 2, one out of five of the estimated features remains unassigned. This is also a particular case of the analysis presented in Section VI-D, with $\mathfrak{m} = 4$ and $\mathfrak{a} = 1$, in which (17) and (18) confirm the above values, showing the OSPA metric's lower penalization of false alarms compared with missed detections.

For the COLA metric, (5) shows that as $c \to \infty$, under a similar analysis, $d^{(c,p=2)}_{\text{COLA}}(\mathcal{M}, \widehat{\mathcal{M}}_1)$ and $d^{(c,p=2)}_{\text{COLA}}(\mathcal{M}, \widehat{\mathcal{M}}_2) \to \sqrt{1}$, again demonstrating the COLA metric's nature of gauging cardinality, rather than averaged distance error, and its ability to equally penalize missed detections and false alarms.

The coarse dashed lines in Fig. 7 indicate the Hausdorff distances. The OMAT metric disagrees with all the other metrics. Note that it made assignments only within the considered region.

E. Results With the "CityTrees10k" Dataset

To test the intuitive behavior of each metric with a publicly available dataset, "CityTrees10k" was chosen, due to its available GT map [5]. This dataset comprises 10 000 simulated robot movements and 100 GT features (simulated trees) which can be

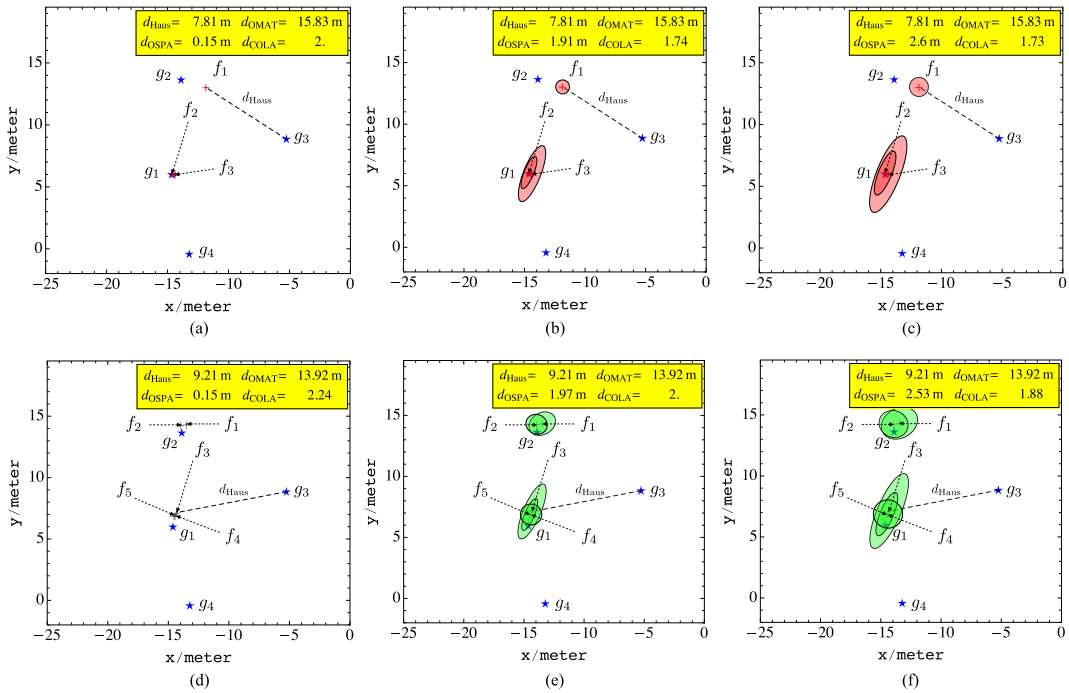


Fig. 7. Metric values for $p = 2$ and three different values of the cutoff parameter c. (a) SLAM Algorithm 1, $p = 2$ and $c = 0.15$. (b) SLAM Algorithm 1, $p = 2$ and $c = 2.20$. (c) SLAM Algorithm 1, $p = 2$ and $c = 3.00$. (d) SLAM Algorithm 2, $p = 2$ and $c = 0.15$. (e) SLAM Algorithm 2, $p = 2$ and $c = 2.20$. (f) SLAM Algorithm 2, $p = 2$ and $c = 3.00$.

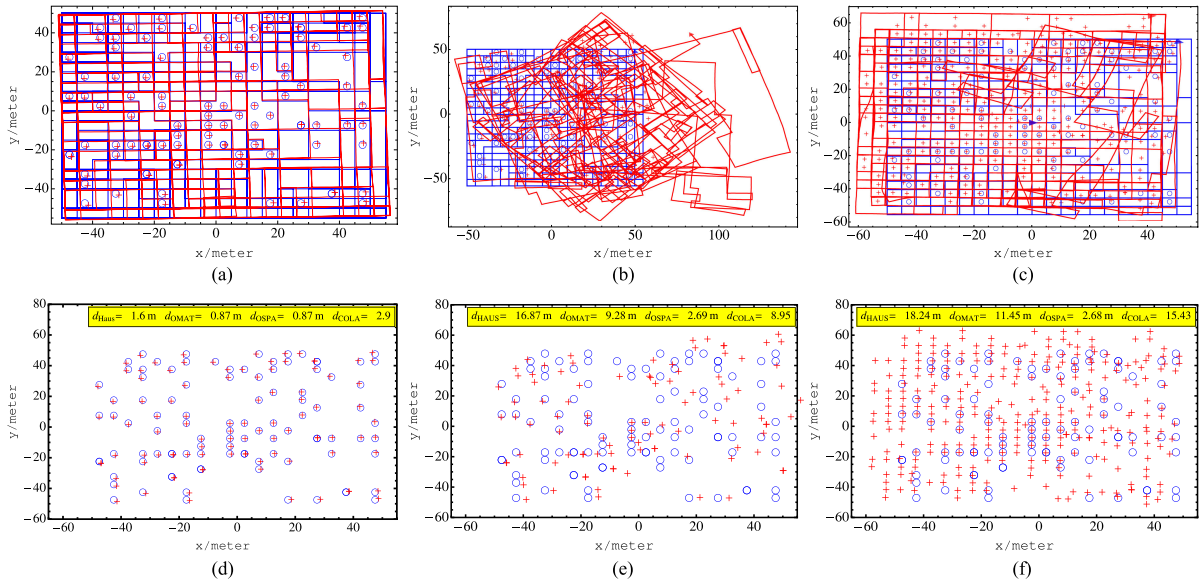


Fig. 8. "CityTrees10k" SLAM and mapping results using the g^{2}o solver. Results show given DA, a limited number of g^{2}o iterations and automatic NN DA based on [4]. Blue lines/circles: GT trajectory/map, red lines/circles: corresponding estimates. (a) SLAM Algorithm 3: g^{2}o SLAM given DA. (b) SLAM Algorithm 4: g^{2}o before convergence. (c) SLAM Algorithm 5: g^{2}o with NN DA failures. (d) SLAM Algorithm 3: Estimated and GT maps. (e) SLAM Algorithm 4: Estimated and GT maps. (f) SLAM Algorithm 5: Estimated and GT maps.

detected within a field of view of 10 m around the robot. g^{2}o graph optimization SLAM solutions [38], with perfect (given) DA (SLAM Algorithm 3), before convergence (SLAM Algorithm 4) and with estimated NN DA (SLAM Algorithm 5) are shown in Fig. 8. In Fig. 9, the results of iSAM [5], [6] with perfect (given) DA (SLAM Algorithm 6) and NN DA failures (SLAM Algorithm 7) are shown.

Fig. 8(a) (SLAM Algorithm 3) shows the SLAM result based on perfect DA (provided with the dataset) using the g^{2}o solver. Fig. 8(d) shows the corresponding GT and estimated maps only. Fig. 8(b) (SLAM Algorithm 4) and (e) shows the SLAM and mapping results, respectively, with perfect DA, but before complete convergence of the g^{2}o solver, yielding a distorted trajectory. As shown in Fig. 8(e), due to perfect DA, all of the feature locations are reasonably close to GT, since they were estimated within the less erroneous, initial part of the trajectory. Fig. 8(c) (SLAM Algorithm 5) and (f) shows the SLAM and mapping results, respectively, of allowing the g^{2}o solver to converge, when

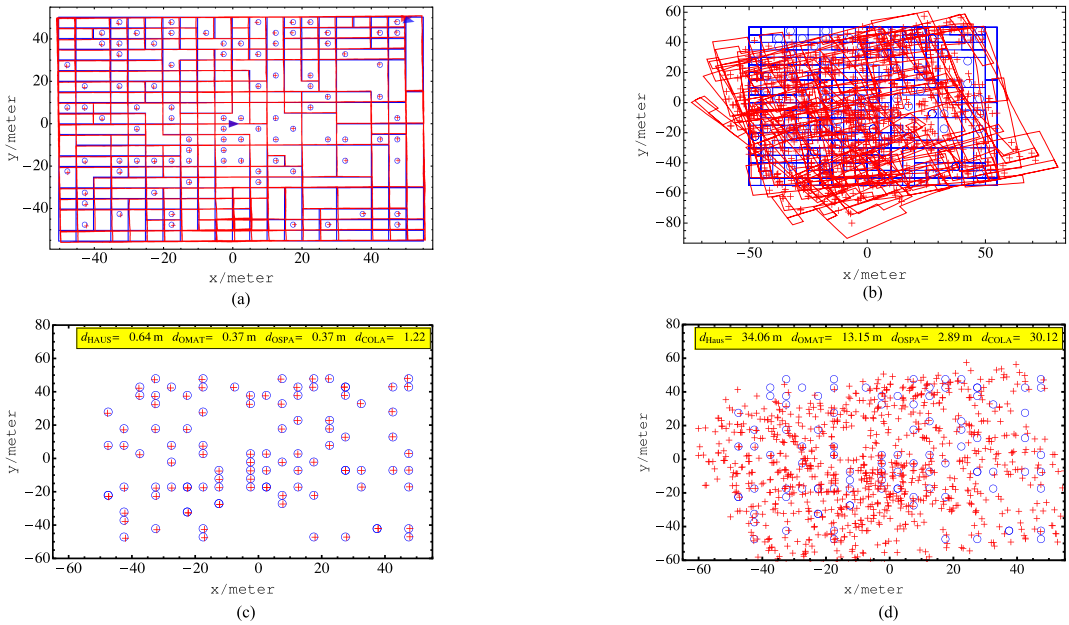


Fig. 9. "CityTrees10k" dataset SLAM and mapping results based on iSAM. Results show given DA and NN DA based on [4]. Blue lines/circles: GT trajectory/map, red lines/circles: corresponding estimates. (a) SLAM Algorithm 6: iSAM SLAM given DA. (b) SLAM Algorithm 7: iSAM SLAM NN DA failures. (c) SLAM Algorithm 6: Estimated and GT maps. (d) SLAM Algorithm 7: Estimated and GT maps.

replacing the given (perfect) DAs with an automated NN DA algorithm, based on a Euclidean gating distance of 3 m. The NN DA routine of [4] was used in this case.

Fig. 9(a) shows the SLAM result of iSAM with given DA and Fig. 9(c) the corresponding estimated and GT maps. When perfect DA is replaced with the same NN DA routine of [4], the SLAM result in Fig. 9(b) results, with its corresponding estimated and GT maps shown in Fig. 9(d).

Table II summarizes the metric evaluations of Figs. 8 and 9. SLAM Algorithm 6 appears to perform best, as reflected by all the metrics. Some interesting differences occur in each metric's ranking of SLAM Algorithms 4 and 5. Intuitively, SLAM Algorithm 5 may initially appear superior than SLAM Algorithm 4, since the trajectory appears to bear a closer relationship to GT, as indicated in Fig. 8(c), and the Trajectory Energy Metric $d^{\text{Trans}}_{\text{Energy}}$ in Table II. However, SLAM Algorithm 5 has committed multiple DA errors, causing it to vastly overestimate the number of map features ($\widehat{\mathfrak{m}} = 299$) [compare Fig. 8(f) and (e)]. From a mapping perspective, intuition dictates that SLAM Algorithm 5 is worse, with many more false alarms than the map produced by SLAM Algorithm 4, and, contrary to the OSPA metric, the COLA metric correctly reflected this. The reason that OSPA, nonintuitively reports a lower mapping error can be seen from (3), where the averaging nature of the OSPA metric requires division by $\widehat{\mathfrak{m}}$. Since in SLAM Algorithm 4, [see Fig. 8(e)] $\widehat{\mathfrak{m}} = 100$, a higher OSPA value results than for SLAM Algorithm 5 [see Fig. 8(f)], with $\widehat{\mathfrak{m}} = 299$. Note that in the case of SLAM Algorithms 4 and 5, (20), (21), and (25) are obeyed, indicating that disagreement between the COLA and OSPA metrics will result.

For comparison, and since no feature covariance information was used, the sum of squared Euclidean errors (SSEE) values are given in Table II which, as in the case of the NEES metric, can only be calculated based on the number of gated features. Although SLAM Algorithm 5 has gated more features (68) than SLAM Algorithm 4 (35), the sum of their localization errors is lower, resulting in a lower SSEE value. Since the SSEE metric ignores detection errors, it disagrees with the COLA metric.

In Table III, the results of 100 comparisons of the results of running Algorithms 4 and 5 on the CityTrees10k data set are shown. Each entry a_{ij} in the table corresponds to the number of times each evaluated metric in row i and column j disagreed. Interestingly, out of the 100 comparisons made, 55 of the map evaluations made by the COLA metric disagreed with the trajectory energy metric, showing that the presented results in Fig. 8 are not an isolated case. It is also interesting to note that, even both of the tested trajectory based metrics $d^{\text{Trans}}_{\text{Energy}}$ and ATE, actually disagree with each other 8 times out of the 100 comparisons. Interestingly, the number of times that the Hausdorff map metric disagreed with the $d^{\text{Trans}}_{\text{Energy}}$ or the ATE trajectory metrics is also high (52 and 44 times respectively). Until now, the Hausdorff metric has been widely used and accepted as a valid mapping metric for feature maps, and yet when used to compare maps, it can disagree with trajectory based metrics.

IX. Summary

This paper suggests the following points for selecting a metric which can intuitively gauge feature map accuracy:

1) Single object metrics, such as the Euclidean distance, RMS, or NEES, can only gauge the error between maps, or subsets of maps, with the same cardinality. When feature cardinalities differ, multiobject metrics offer a solution.
2) Metrics which gauge the error in the SLAM trajectory, do not necessarily provide intuitive assessments of the full SLAM error. It was demonstrated that the multiobject map metrics, which provide intuitive assessments of

TABLE II
DETECTION ERRORS AND MAP METRIC RESULTS FOR SLAM ALGORITHMS 3 TO 7 ($c = 3.00$ AND $p = 2$)

Algorithm	$\widehat{\mathfrak{m}}$	Gtd	MD	FA	d_{COLA}	d_{OSPA} [m]	d_{HAUS} [m]	d_{OMAT} [m]	SSEE* [m^2]	SSEE† [m^2]	$d^{\text{Trans}}_{\text{Energy}}$ [m^2]	ATE‡ [m]
3	100	100	0	0	2.90 (2)	0.87 (2)	1.60 (2)	0.87 (2)	75.91 (3)	0.759 (2)	1.066 (2)	0.120 (2)
4	100	35	65	65	8.95 (3)	2.69 (4)	16.87 (3)	9.28 (3)	136.54 (4)	3.901 (5)	39.214 (5)	28.550 (5)
5	299	68	32	231	15.43 (4)	2.68 (3)	18.24 (4)	11.45 (4)	64.73 (2)	0.952 (3)	10.974 (3)	7.292 (3)
6	100	100	0	0	1.22 (1)	0.37 (1)	0.64 (1)	0.37 (1)	13.47 (1)	0.135 (1)	0.893 (1)	0.093 (1)
7	975	86	14	889	30.12 (5)	2.89 (5)	34.06 (5)	13.15 (5)	162.32 (5)	1.887 (4)	16.157 (4)	11.243 (4)

*The SSEE metric was calculated only based on the number of gated features, ignoring detection errors. †The SSEE average per gated feature. ‡RMS values were used for the ATE metric. The bracketed numbers give the ranking by each metric.

TABLE III
TABLE SHOWING THE NUMBER OF TIMES EACH METRIC DISAGREED WITH ANOTHER METRIC WHEN COMPARING 100 DIFFERENT RESULTS FROM SLAM ALGORITHMS 4 AND 5

Disagreement [%]	d_{COLA}	d_{OSPA} [m]	d_{HAUS} [m]	d_{OMAT} [m]	SSEE* [m^2]	SSEE† [m^2]	$d^{\text{Trans}}_{\text{Energy}}$ [m^2]	ATE‡ [m]
d_{COLA}		13	3	10	46	81	55	47
d_{OSPA} [m]			14	13	41	68	42	36
d_{HAUS} [m]				7	47	80	52	44
d_{OMAT} [m]					50	75	47	41
SSEE* [m^2]						41	55	51
SSEE† [m^2]							50	56
$d^{\text{Trans}}_{\text{Energy}}$ [m^2]								8

mapping errors, can significantly disagree with the trajectory evaluation metrics.

3) The Hausdorff and OMAT metrics are undefined when one map set is empty and were shown in [14] to be insensitive to cardinality errors, giving non-intuitive map assessments.
4) The OSPA metric is an averaged distance metric. When assessing mapping performance during SLAM execution, as more of the GT map passes through the field(s) of view of the sensor(s), this metric provides the most intuitive evaluation of the time varying estimated map (see Section VI-C).
5) When comparing multiple estimated maps against a fixed GT map, the OSPA metric has some disadvantages, such as its saturation to parameter c when no features are gated, irrespective of the cardinality errors, and its higher penalization of missed detections over false alarms.
6) When comparing multiple estimated maps with a fixed GT map, the COLA metric solves some of the problems associated with the OSPA metric. In particular, it is able to distinguish between map qualities, even if no features are gated (it does not saturate) and it equally penalizes missed detections and false alarms.
7) Since the OSPA metric often yields values close to c, small changes in c can result in the metric changing its decision when comparing different maps to GT. This problem is avoided with the COLA metric.

Multiobject metrics also have possible applications as principled measures for point cloud registration and general DA, which take into account data differences (e.g., object location) *and* their cardinalities and, therefore, provide an interesting avenue for future research. For example, this could be useful in reinitializing transformations between the reference frames of estimated and GT submaps.

APPENDIX A
DERIVATION OF THE COLA METRIC

Consider two probability densities $f(\mathcal{M})$ and $g(\widehat{\mathcal{M}})$ in the n-Euclidean space. By definition, the Wasserstein distance calculates the similarity between f and g and is given by

$$d_w^p(f,g) := \min_h \left(\int \int d(\mathbf{m}, \widehat{\mathbf{m}})^p \, h(\mathbf{m}, \widehat{\mathbf{m}}) d\mathbf{m} d\widehat{\mathbf{m}} \right)^{1/p} \tag{34}$$

where $1 \leq p < \infty$, $d(\mathbf{m}, \widehat{\mathbf{m}})$ is any distance metric between $\mathbf{m}$ and $\widehat{\mathbf{m}}$, and $h(\mathbf{m}, \widehat{\mathbf{m}})$ is any joint distribution whose marginals are $f(\mathbf{m})$ and $g(\widehat{\mathbf{m}})$ such that

$$\int h(\mathbf{m}, \widehat{\mathbf{m}}) d\widehat{\mathbf{m}} = f(\mathbf{m}), \quad \int h(\mathbf{m}, \widehat{\mathbf{m}}) d\mathbf{m} = g(\widehat{\mathbf{m}}). \tag{35}$$

Consider the marginal distributions in (35) to be

$$f(\mathbf{m}) = \frac{1}{\mathfrak{m}} \sum_{i=1}^{\mathfrak{m}} \delta(\mathbf{m} - \mathbf{m}^i), \quad g(\widehat{\mathbf{m}}) = \frac{1}{\widehat{\mathfrak{m}}} \sum_{j=1}^{\widehat{\mathfrak{m}}} \delta(\widehat{\mathbf{m}} - \widehat{\mathbf{m}}^j) \tag{36}$$

where δ is the Dirac delta function. This implies that the joint distribution $h(\mathbf{m}, \widehat{\mathbf{m}})$ is given by

$$h(\mathbf{m}, \widehat{\mathbf{m}}) = \sum_{i=1}^{\mathfrak{m}} \sum_{j=1}^{\widehat{\mathfrak{m}}} c_{i,j} \delta(\mathbf{m} - \mathbf{m}^i) \delta(\widehat{\mathbf{m}} - \widehat{\mathbf{m}}^j). \tag{37}$$

The derivation of the COLA metric follows a similar procedure as the OSPA metric [14]. A unique assignment coefficient

$$c_{i,j} = \delta_{j,\sigma(i)} / \max\{\mathfrak{m}, \ \widehat{\mathfrak{m}}\} \tag{38}$$

which satisfies (37), is used where $\sigma(i)$ is a permutation of the larger set and $\delta_{j,\sigma(i)} = 1$ iff $j = \sigma(i)$ and 0 otherwise.

Consider a GT map $\mathcal{M}' = \{\mathbf{m}^1, \ldots, \mathbf{m}^{\mathfrak{m}}\}$, and its estimated map $\widehat{\mathcal{M}} = \{\widehat{\mathbf{m}}^1, \ldots, \widehat{\mathbf{m}}^{\widehat{\mathfrak{m}}}\}$. The COLA metric can then be de-

rived from (34) by replacing

$$d(\mathbf{m}^i, \widehat{\mathbf{m}}^j) = \frac{(\max\{\mathfrak{m},\ \widehat{\mathfrak{m}}\})^{1/p}}{c} d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^j). \tag{39}$$

Note that if $d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^j)$ is a metric, then $d(\mathbf{m}^i, \widehat{\mathbf{m}}^j)$ is also guaranteed to be a metric as required in (34). For $\widehat{\mathfrak{m}} \geq \mathfrak{m}$, replacing $c_{i,j}$ in (37) with (38) and then substituting the result, together with (39) into (34) yields a new metric

$$d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = \left(\min_{\sigma} \sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})}{c} \right)^p \right)^{1/p} \tag{40}$$

where $d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})$ is defined in (1), with cutoff parameter c. For $\mathfrak{m} \geq \widehat{\mathfrak{m}}$, the metric is defined as $d_{\text{COLA}}^{(c,p)}(\widehat{\mathcal{M}}, \mathcal{M})$.

Appendix B
Proof That $d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}})$ is a Metric

The proof that $d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}})$ is a metric follows a similar procedure to the proof that $d_{\text{OSPA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}})$ is a metric in [14]. $d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) \geq 0$ for all $\mathcal{M}, \widehat{\mathcal{M}}$ because metric $d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\mathrm{j}}) \geq 0$ for all $\mathbf{m}^{\mathrm{i}}$ and $\widehat{\mathbf{m}}^{\mathrm{j}}$. Similarly $d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = 0$ iff $\widehat{\mathcal{M}} = \mathcal{M}$ - proof: From (5), if $d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = 0$, then

$$\min_{\sigma} \sum_{i=1}^{\mathfrak{m}} \frac{d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})^p}{c^p} = -(\widehat{\mathfrak{m}} - \mathfrak{m}) \leq 0 \tag{41}$$

since (5) is defined for $\widehat{\mathfrak{m}} \geq \mathfrak{m}$. The RHS of (41) therefore implies that $\widehat{\mathfrak{m}} = \mathfrak{m}$ (since the LHS is positive), meaning that $d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)}) = 0 \quad \forall i$. $d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = d_{\text{COLA}}^{(c,p)}(\widehat{\mathcal{M}}, \mathcal{M})$ because $d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\mathrm{j}})$ satisfies the symmetry property. It remains to be verified that the triangle inequality is satisfied.

Consider the set $\widehat{\mathcal{N}} = \{\widehat{\mathbf{n}}^1, \ldots, \widehat{\mathbf{n}}^{\widehat{\mathfrak{n}}}\}$, with cardinality $\widehat{\mathfrak{n}} \in \mathbb{N}_0$. Consider the following sets of dummy points $\mathcal{U} = \{\mathbf{u}^{\mathrm{i}}\}_{\mathrm{i} \in \mathbb{N}_0}$ and $\mathcal{V} = \{\mathbf{v}^{\mathrm{j}}\}_{\mathrm{j} \in \mathbb{N}_0}$ in $\mathbb{R}^N$ where

$$d(\mathbf{u}^{\mathrm{i}}, \mathbf{x}) \geq c,\ d(\mathbf{v}^{\mathrm{i}}, \mathbf{x}) \geq c,\ d(\mathbf{u}^{\mathrm{i}}, \mathbf{v}^{\mathrm{j}}) \geq c \quad \forall\, \mathbf{x},\ \mathrm{i}, \mathrm{j}.$$

Case 1: $(\mathfrak{m} \leq \widehat{\mathfrak{m}} \leq \widehat{\mathfrak{n}})$. In order to raise the cardinality of sets $\mathcal{M}$ and $\widehat{\mathcal{M}}$ to $\widehat{\mathfrak{n}}$, consider the dummy points:

$$\mathbf{m}^{\mathfrak{m}+\mathrm{i}} = \mathbf{u}^i, \quad 1 \leq \mathrm{i} \leq \widehat{\mathfrak{n}} - \mathfrak{m} \tag{42}$$

$$\widehat{\mathbf{m}}^{\widehat{\mathfrak{m}}+j} = \mathbf{v}^j, \quad 1 \leq \mathrm{j} \leq \widehat{\mathfrak{n}} - \widehat{\mathfrak{m}}. \tag{43}$$

Then, choose $\sigma, \tau \in \Pi_{\widehat{\mathfrak{n}}}$ such that

$$\min_{\pi \in \Pi_{\widehat{\mathfrak{n}}}} \sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\pi(\mathrm{i})})}{c} \right)^p = \sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\sigma(\mathrm{i})})}{c} \right)^p \tag{44}$$

$$\min_{\pi \in \Pi_{\widehat{\mathfrak{n}}}} \sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\widehat{\mathbf{n}}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\pi(\mathrm{i})})}{c} \right)^p = \sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\widehat{\mathbf{n}}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\tau(\mathrm{i})})}{c} \right)^p \tag{45}$$

$$\therefore d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = \left(\min_{\pi \in \Pi_{\mathfrak{m}}} \sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\pi(\mathrm{i})})}{c} \right)^p \right)^{1/p} \tag{46}$$

$$\leq \left(\min_{\pi \in \Pi_{\mathfrak{m}}} \sum_{i=1}^{\mathfrak{m}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\pi(\mathrm{i})})}{c} \right)^p + (\widehat{\mathfrak{n}} - \mathfrak{m}) \right)^{1/p} \tag{47}$$

$$\leq \left(\min_{\pi \in \Pi_{\widehat{\mathfrak{n}}}} \sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\pi(\mathrm{i})})}{c} \right)^p \right)^{1/p} \tag{48}$$

$$\leq \left(\sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\sigma(\mathrm{i})}) + d^{(c)}(\widehat{\mathbf{n}}^{\sigma(\mathrm{i})}, \widehat{\mathbf{m}}^{\tau(\sigma(\mathrm{i}))})}{c} \right)^p \right)^{1/p} \tag{49}$$

$$\leq \left(\sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\sigma(\mathrm{i})})}{c} \right)^p \right)^{1/p} + \left(\sum_{i=1}^{\widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\widehat{\mathbf{n}}^{\sigma(\mathrm{i})}, \widehat{\mathbf{m}}^{\tau(\sigma(\mathrm{i}))})}{c} \right)^p \right)^{1/p} \tag{50}$$

$$\leq d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{N}}) + d_{\text{COLA}}^{(c,p)}(\widehat{\mathcal{N}}, \widehat{\mathcal{M}}). \tag{51}$$

In (47), $\widehat{\mathfrak{n}} - \mathfrak{m}$ dummy points were added to the set $\mathcal{M}$ yielding (48). In (48), the triangular inequality on the metric $d^{(c)}$ and the application of (44) and (45) resulted in (49). Finally, Minkowski's inequality yielded (50).

Case 2: $(\mathfrak{m}, \widehat{\mathfrak{n}} \leq \widehat{\mathfrak{m}})$. In order to raise the cardinality of sets $\mathcal{M}$ and $\widehat{\mathcal{N}}$ to $\widehat{\mathfrak{m}}$, consider the dummy points

$$\mathbf{m}^{\widehat{\mathfrak{m}}-\mathrm{i}+1} = \mathbf{u}^i, \quad 1 \leq \mathrm{i} \leq \widehat{\mathfrak{m}} - \mathfrak{m} \tag{52}$$

$$\widehat{\mathbf{n}}^{\widehat{\mathfrak{m}}-\mathrm{j}+1} = \mathbf{u}^j, \quad 1 \leq \mathrm{j} \leq \widehat{\mathfrak{m}} - \widehat{\mathfrak{n}} \tag{53}$$

where $d(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\mathrm{i}}) = 0, \quad \max(\mathfrak{m}, \widehat{\mathfrak{n}}) \leq i \leq \widehat{\mathfrak{m}}$.

Then, choose $\sigma, \tau \in \Pi_{\widehat{\mathfrak{m}}}$ such that

$$\min_{\pi \in \Pi_{\mathfrak{m} \vee \widehat{\mathfrak{n}}}} \sum_{i=1}^{\mathfrak{m} \vee \widehat{\mathfrak{n}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\pi(\mathrm{i})})}{c} \right)^p =$$

$$\min_{\pi \in \Pi_{\widehat{\mathfrak{m}}}} \sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\pi(\mathrm{i})})}{c} \right)^p = \sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{n}}^{\sigma(\mathrm{i})})}{c} \right)^p \tag{54}$$

$$\min_{\pi \in \Pi_{\widehat{\mathfrak{m}}}} \sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\widehat{\mathbf{n}}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\pi(\mathrm{i})})}{c} \right)^p = \sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\widehat{\mathbf{n}}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\tau(\mathrm{i})})}{c} \right)^p \tag{55}$$

$$d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = \left(\min_{\pi \in \Pi_{\widehat{\mathfrak{m}}}} \sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathrm{i}}, \widehat{\mathbf{m}}^{\pi(\mathrm{i})})}{c} \right)^p \right)^{1/p} \tag{56}$$

where $\mathfrak{m} \vee \widehat{\mathfrak{n}} = \max(\mathfrak{m}, \widehat{\mathfrak{n}})$. Therefore, finally

$$\leq \left(\sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathfrak{i}}, \widehat{\mathbf{n}}^{\sigma(\mathfrak{i})}) + d^{(c)}(\widehat{\mathbf{n}}^{\sigma(\mathfrak{i})}, \widehat{\mathbf{m}}^{\tau(\sigma(\mathfrak{i}))})}{c} \right)^p \right)^{1/p} \tag{57}$$

$$\leq \left(\sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\mathbf{m}^{\mathfrak{i}}, \widehat{\mathbf{n}}^{\sigma(\mathfrak{i})})}{c} \right)^p \right)^{1/p} + \left(\sum_{i=1}^{\widehat{\mathfrak{m}}} \left(\frac{d^{(c)}(\widehat{\mathbf{n}}^{\sigma(\mathfrak{i})}, \widehat{\mathbf{m}}^{\tau(\sigma(\mathfrak{i}))})}{c} \right)^p \right)^{1/p} \tag{58}$$

$$\leq d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{N}}) + d_{\text{COLA}}^{(c,p)}(\widehat{\mathcal{N}}, \widehat{\mathcal{M}}). \tag{59}$$

In (56), the triangular inequality on the metric $d^{(c)}$ and the application of (54) and (55) resulted in (57). Again, Minkowski's inequality yielded (58).

APPENDIX C
PROOF OF INEQUALITY (8)

For $|\widehat{\mathcal{M}}| \geq |\mathcal{M}|$, the COLA metric is defined as

$$d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = \left(\min_{\sigma} \sum_{i=1}^{|\mathcal{M}|} \left(\frac{d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\sigma(i)})}{c} \right)^p \right)^{1/p}. \tag{60}$$

If π is the permutation set which achieves the minimization of the COLA metric, then

$$d_{\text{COLA}}^{(c,p)}(\mathcal{M}, \widehat{\mathcal{M}}) = \left(\sum_{i=1}^{|\mathcal{M}|} \left(\frac{d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\pi(i)})}{c} \right)^p \right)^{1/p}. \tag{61}$$

Consider $1 \leq p_1 \leq p_2$ and the substitutions

$$x_i = \frac{d^{(c)}(\mathbf{m}^i, \widehat{\mathbf{m}}^{\pi(i)})}{c}, \quad s = \left(\sum_{i=1}^{|\mathcal{M}|} x_i^{p_1} \right)^{1/p_1}, \quad y_i = \frac{x_i}{s} \leq 1 \tag{62}$$

$$\therefore y_i^{p_2} \leq y_i^{p_1} \tag{63}$$

$$\frac{\sum_{i=1}^{|\mathcal{M}|} x_i^{p_2}}{s^{p_2}} \leq \frac{\sum_{i=1}^{|\mathcal{M}|} x_i^{p_1}}{s^{p_1}} \tag{64}$$

$$\frac{\sum_{i=1}^{|\mathcal{M}|} x_i^{p_2}}{\left(\sum_{i=1}^{|\mathcal{M}|} x_i^{p_1} \right)^{p_2/p_1}} \leq 1 \tag{65}$$

$$\left(\sum_{i=1}^{|\mathcal{M}|} x_i^{p_2} \right)^{1/p_2} \leq \left(\sum_{i=1}^{|\mathcal{M}|} x_i^{p_1} \right)^{1/p_1} \tag{66}$$

and substituting for x_i from (62) into (66) yields

$$d_{\text{COLA}}^{(c,p_2)}(\mathcal{M}, \widehat{\mathcal{M}}) \leq d_{\text{COLA}}^{(c,p_1)}(\mathcal{M}, \widehat{\mathcal{M}}). \tag{67}$$

ACKNOWLEDGMENT

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright notation thereon. The authors acknowledge the invaluable advice from B.-N. Vo.

REFERENCES

[1] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Proc. Auton. Robot Veh.*, 1990, pp. 167–193.
[2] M. W. M. G. Dissanayake, P. Newman, S. Clark, H.F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," *IEEE Trans. Robot. Autom.*, vol. 17, no. 3, pp. 229–241, Jun. 2001.
[3] J. Nieto, J. Guivant, E. Nebot, and S. Thrun, "Real time data association for FastSLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, Sep. 2003, vol. 1, pp. 412–418
[4] S. Thrun and M. Montemerlo, "The GraphSLAM algorithm with applications to large-scale mapping of urban structures," *Int. J. Robot. Res.*, vol. 25, nos. 5/6, pp. 403–429, 2005.
[5] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Trans. Robot.*, vol. 24, no. 6, pp. 1365–1378, Dec. 2008.
[6] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Int. J. Robot. Res.*, vol. 31, no. 2, pp. 216–235, 2011.
[7] J. Mullane, B. N. Vo, M. D. Adams, and B. T. Vo, "A random-finite-set approach to Bayesian SLAM," *IEEE Trans. Robot.*, vol. 27, no. 2, pp. 268–282, Apr. 2011.
[8] C. S. Lee, D. E. Clark, and J. Salvi, "SLAM with dynamic target via single cluster PHD filtering," *IEEE Sel. Topics Signal Process.*, vol. 7, no. 3, pp. 543–552, Jun. 2013.
[9] R. Kümmerle *et al.*, "On measuring the accuracy of SLAM algorithms," *Auton. Robots*, vol. 27, no. 4, pp. 387–407, Nov. 2009.
[10] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation With Applications to Tracking and Navigation*. Hoboken, NJ, USA: Wiley, 2001.
[11] M. Walter, R. Eustice, and J. Leonard, "Exactly sparse extended information filters for feature based SLAM," *Int. J. Robot. Res.*, vol. 26, no. 4, pp. 335–359, Apr. 2007.
[12] J. A. Castellanos, R. Martinez-Cantin, J. D. Tardos, and J. Neira, "Robocentric map joining: Improving the consistency of EKF-SLAM," *Robot. Auton. Syst.*, vol. 55, no. 1, pp. 21–29, 2007.
[13] G. Hu, S. Huang, and G. Dissanayake, "Evaluation of pose only SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 3732–3737.
[14] D. Schuhmacher, B. T. Vo, and B. N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 86, no. 8, pp. 3447–3457, Aug. 2008.
[15] S. Schwertfeger, "Robotic mapping in the real world: Performance evaluation and system integration," Ph.D. dissertation, School of Eng. Sci., Jacobs Univ., Bremen, Germany, 2012. [Online]. Available: http://d-nb.info/1035268981/34.
[16] P. Barrios, G. Naqvi, K. Leung, and F. Inostroza, "The cardinalized optimal linear assignment (COLA) metric for multi-object error evaluation," in *Proc. 18th Int. Conf. Inf. Fusion*, Jul. 2015, pp. 271–279.
[17] H. Moravec and A. E. Elfes, "High resolution maps from wide angle sonar," in *Proc. IEEE Int. Conf. Robot. Autom.*, Mar. 1985, pp. 116–121.
[18] J. Mullane, M. Adams, and W. S. Wijesoma, "Robotic mapping using measurement likelihood filtering," *Int. J. Robot. Res.*, vol. 2, no. 28, pp. 172–190, 2009.
[19] A. Birk, "A quantitative assessment of structural errors in grid maps," *Auton. Robots*, vol. 28, no. 2, pp. 187–196, 2010.
[20] B. Balaguer, S. Balakirskyand, S. Carpin, and A. Visser, "Evaluating maps produced by urban search and rescue robots: Lessons learned from robocup," *Auton. Robots*, vol. 27, no. 4, pp. 449–464, 2009.
[21] J. Pellenz and D. Paulus, "Mapping and map scoring at the robo-cup rescue competition," *Quant. Perform. Eval. Navig. Solutions Mobile Robots*, vol. 46, pp. 47–66, 2008.
[22] M. Chandran-Ramesh and P. Newman, "Assessing map quality using conditional random fields," in *Field and Service Robotics (Springer Tracts in Advanced Robotics)*. New York, NY, USA: Springer, 2008, pp. 35–48.
[23] J. Sturm. *RGB-D SLAM dataset and benchmark*, 2014. [Online]. Available: http://vision.in.tum.de/data/datasets/rgbd-dataset.

[24] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vilamoura, Portugal, Oct. 2012, pp. 573–580.
[25] F. Amigoni, S. Gasparini, and M. Gini, "Good experimental methodologies for robotic mapping: A proposal," in *Proc. Int. Conf. Robot. Autom.*, Apr. 2007, pp. 4176–4181.
[26] A. I. Wagan, A. Godil, and X. Li, "Map quality assessment," in *Proc. 8th Workshop Perform. Metrics Intell. Syst.*, 2008, pp. 278–282.
[27] European Union FP6. *The rawseeds project*. [Online]. Available: http://www.rawseeds.org/rs/methods/view//10.
[28] J. R. Hoffman and R. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Trans. Syst., Man, Cybern.*, vol. 34, no. 3, pp. 327–336, May 2004.
[29] K. Ganstrom, C. Lundquist, and O. Orguner, "Extended target tracking using a Gaussian-mixture PHD filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 4, pp. 3268–3286, Oct. 2012.
[30] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3452–3457, Jul. 2011.
[31] D. F. Crouse, P. Willett, M. Guerriero, and L. Svensson, "An approximate minimum MOSPA estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 3644–3647.
[32] S. Nagappa, D. E. Clark, and R. Mahler, "Incorporating track uncertainty into the OSPA metric," in *Proc. 14th Int. Conf. Inf. Fusion*, Jul. 2011, pp. 1–8.
[33] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logist. Quart.*, vol. 2, nos. 1/2, pp. 83–97, 1955.
[34] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. New York, NY, USA: Dover, 1998.
[35] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, 1987.
[36] M. B. Wright, "Speeding up the Hungarian algorithm," *Comput. Oper. Res.*, vol. 17, no. 1, pp. 95–96, 1990.
[37] W. Jones, A. Chawdhary, and A. King, "Optimising the Volgenant-Jonker algorithm for approximating graph edit distance," *Pattern Recognit. Lett.*, vol. 84, pp. 1–9, Aug. 2016.
[38] R. Kuemmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g^2o: A general framework for graph optimization," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3607–3613.

Pablo Barrios received the Bachelor's degree in physics from the Department of Physics, Universidad de Chile, Santiago, Chile, in 2009. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering, Faculty of Physical and Mathematical Sciences, Universidad de Chile.

His research interests include multiobject metrics, SLAM performance evaluation, scan matching, and data association.

Martin Adams (SM'08) received the Graduate degree in engineering science from University of Oxford, Oxford, U.K, in 1988 and the D.Phil. degree from the Robotics Research Group, University of Oxford, in 1992.

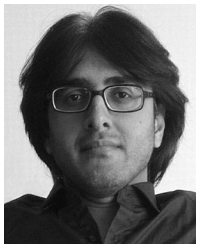
He is a Professor of electrical engineering in the Department of Electrical Engineering, Universidad de Chile, Santiago, Chile, and is currently the Chair of the IEEE Robotics and Automation Chilean Chapter. He is also a Principle Investigator in the Advanced Mining Technology Centre, Universidad de Chile. He continued his research in autonomous robot navigation as a Project Leader at the Institute of Robotics, Swiss Federal Institute of Technology, Zurich, Switzerland. He was employed as a Guest Professor and taught control theory in Buchs, St. Gallen, Switzerland, from 1994 to 1995. From 1996 to 2000, he served as a Research Scientist in robotics and control, in the field of semiconductor assembly automation, with the European Semiconductor Equipment Centre, Switzerland. From 2000 to 2010, he was an Associate Professor at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research work focuses on autonomous robot navigation, space situational awareness, sensing, sensor data interpretation, and control. He has been the Principle Investigator and Leader of many robotics and industrially sponsored projects and has served as an Associate Editor on various journal and conference editorial boards.

Keith Leung (M'06) received the B.A.Sc. degree in 2005 and the M.A.Sc. degree in 2007 from University of Waterloo, Waterloo, ON, Canada, in mechanical engineering. He received the Ph.D. degree in 2012 from University of Toronto, Toronto, ON, Canada.

He was a member of the Autonomous Space Robotics Laboratory and the Flight Systems Control Laboratory, Institute for Aerospace Studies. From 2012 to 2016, he was a Postdoctoral Researcher with the Advanced Mining Technology Center and Department of Electrical Engineering, Universidad de Chile. His area of expertise is in state estimation and robot navigation, as well as simultaneous localization and mapping. He is currently a vision processing, perception, and robotics expert with Applanix (Trimble) Richmond Hill, ON, Canada.

Felipe Inostroza (M'11) received the Graduate degree in engineering science in 2013 and the Master's of Electrical Engineering degree in 2015, both from Universidad de Chile, Santiago, Chile.

He is a doctoral student in the Department of Electrical Engineering, Universidad de Chile. His research interests focus on robotics in general and simultaneous localization and mapping in particular. He is currently studying the effects of including detection statistics into the SLAM problem.



Ghayur Naqvi received the B.S. degree in computer science from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2007. He received the M.S. degree in computer science from the School of Science and Technology, Örebro University, Örebro, Sweden, in 2012.

He is currently a Research Student in electrical engineering with Universidad de Chile, Santiago, Chile. His research interests include text mining, big data, machine learning, and business intelligence.

Marcos E. Orchard (M'06) received the B.S. degree in 1999 and the Civil Industrial Engineering degree with an electrical major in 2001 from Catholic University of Chile, Santiago, Chile. He received the M.S. and Ph.D. degrees from Georgia Institute of Technology, Atlanta, GA, USA, in 2005 and 2007, respectively.

He is an Associate Professor in the Department of Electrical Engineering, Universidad de Chile, Santiago, and was part of the Intelligent Control Systems Laboratory, Georgia Institute of Technology. His current research interest include the design, implementation, and testing of real-time frameworks for fault diagnosis and failure prognosis, with applications to battery management systems, mining industry, and finance. He has published more than 50 papers in his areas of expertise, which include statistical process monitoring, parametric/nonparametric modeling, and system identification. His research work at the Georgia Institute of Technology was the foundation of novel real-time fault diagnosis and failure prognosis approaches based on particle filtering algorithms.