SANDIA REPORT

SAND2015-8099
Unlimited Release
Printed September, 2015

Preliminary Results on Uncertainty Quantification for Pattern Analytics

David J. Stracuzzi, Randy Brost, Maximillian Chen, Rebecca Malinas,
Matthew Peterson, Cynthia Phillips, David G. Robinson, Diane Woodbridge

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2015-8099
Unlimited Release
Printed September, 2015

Preliminary Results on Uncertainty Quantification for Pattern Analytics

David J. Stracuzzi Randy Brost Maximillian Chen
Rebecca Malinas Matthew Peterson Cynthia Phillips
David G. Robinson Diane Woodbridge

Abstract

This report summarizes preliminary research into uncertainty quantification for pattern analytics within the context of the Pattern Analytics to Support High-Performance Exploitation and Reasoning (PANTHER) project. The primary focus of PANTHER was to make large quantities of remote sensing data searchable by analysts. The work described in this report adds nuance to both the initial data preparation steps and the search process. Search queries are transformed from *does the specified pattern exist in the data?* to *how certain is the system that the returned results match the query?* We show example results for both data processing and search, and discuss a number of possible improvements for each.

Acknowledgment

This work was funded by the Sandia National Laboratories Laboratory Directed Research and Development (LDRD) program as a part of the Pattern Analytics to Support High-Performance Exploitation and Reasoning (PANTHER) grand challenge project. The authors thank Mary Moya and Mark Koch for their effort in describing the sources of error in synthetic aperture radar imagery. We also thank Kristina Czuchlewski for helpful discussion and patience as we felt our way through these murky research problems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This report summarizes research into uncertainty quantification for pattern analytics within the context of the larger Pattern Analytics to Support High-Performance Exploitation and Reasoning (PANTHER) project. One of PANTHER's primary goals is to make large quantities of sensor data searchable. The project successfully demonstrated methods for processing sensor data into a form suitable for representation as a geospatial semantic graph along with powerful graph search techniques. Semantic graphs support a variety of pattern search operations that highlight specific spatial and temporal relationships among objects in the data. As a result, data analysts can now efficiently search large data corpora for specific patterns of interest.

Finding objects and patterns in remote sensor data is only one part of the data analysis task, however. In order to support decision making, analysts and their customers must also assess their confidence in results, which includes consideration of uncertainty. In this report, we summarize two preliminary efforts to quantify the uncertainty associated with the analysis and search of remote sensing data. The first effort focuses on the geospatial semantic graphs by evaluating the match quality of returned search results and quantifying the uncertainty associated with the quality scores. This has the effect of transforming the analyst's search query from an existential one (does the pattern exist?) to one of quality and certainty (which matches are best and how certain are they?). The second effort then focuses on improving the semantic graph results by revisiting the data processing steps that produce the graph input to extract more information about the uncertainty in the underlying data. This additional information can then be propagated through the search process to provide more detailed assessments of the search results.

Analyzing the uncertainty in sensor data has proven to be a challenging research topic. In the following, we attempt to highlight some of the abandoned research paths and the simplifying assumptions made along the way in an effort to provide a more complete description of the problem space. We also attempt to highlight some of the most promising (in our opinion) near-term next steps and long-term research directions.

Chapter 2

Approximate Pattern Matching in Geospatial Semantic Graphs

Geospatial semantic graphs form one of the centerpieces of the PANTHER project. As such, our initial foray into uncertainty analysis focused on evaluating graph search results. The discussion below briefly summarizes the work reported by Stracuzzi et al. (2015) focusing primarily on the details omitted from the published article, such as abandoned research paths and possible expansions of the work. The discussion below also includes some images and tables that were omitted from the published article due to space considerations.

The published article identifies two main contributions. First, it provides a detailed examination of issues in uncertainty analysis for geospatial pattern search applications. Most of the identified issues have deep technical components that require research beyond both the published article and the work reported here. Many of these are reflected in the discussion of next steps and future work in Chapter 2.4. The second contribution is three distinct methods for computing match quality scores with uncertainty intervals. Each method relies on a different set of information and therefore provides a different set of strengths and weaknesses. We discuss these only briefly, reserving the details and performance results for the referenced article.

2.1 Problem Space and Use Case

The goal of this work is to improve the efficiency of sensor data analysts by providing information about the relative quality quality of candidate search results. Consider the following problem illustration. An overhead sensor array collects imagery from several different sources (such as optical and LiDAR) over tens of thousands of square kilometers of the Earth's surface. An analyst is then tasked with identifing all of the high schools located in that region. Clearly there exist better solutions for finding high schools, such as a web search, but the problem provides a reasonable proxy for a number of other mission-relevant tasks that do not have alternate solutions.

The simplest approach is for the analyst to perform a visual inspection of all the generated imagery. This can take a long time and may be error-prone, particularly for large quantities of data. Moreover, if we change the goal slightly into identifying *recently constructed* high schools, then the task gets more difficult as the analyst must compare old and new imagery. Depending on the size of the imaged area and the frequency of new construction starts and imaging passes, the problem can quickly become overwhelming.

Geospatial semantic graphs (Brost et al., 2014) improve the process by automating the search process. The sensor data is first processed into a set of image regions corresponding to primitive semantic objects such as buildings, grass, forest, and pavement. Semantic object extraction can be performed via a variety of methods, and the work reported here relies on the approach described in O'Neil-Dunne et al. (2013). The extracted objects effectively partition the image into disjoint regions, each with a semantic label. These regions then form the nodes of the semantic graph, while edges represent spatial relationships such as adjacency or distance. Each node and edge also includes a set of attributes, such as the semantic label, area, and perimeter of the associated object or distance between linked nodes.

Given the graph representation, the analyst can now query the graph for a high school pattern. Figure 2.1 shows an example. The template contains one node corresponding to the classroom building, with additional nodes for other objects typically associated with high schools, such as football fields, tennis courts, and parking lots. Each node in the template includes a set of attribute constraints, such as a minimum and maximum area. Thus, to query the graph, the analyst must first specify the pattern of interest by identifying relevant constraints on nodes, edges, and attributes.

For the purposes of this work we defined a loose query, so the returned matches included a number of false positives. The original query reported in Watson et al. (2014) correctly found the 12 public high schools, with only two false positives. As a test case for quality scoring and uncertainty analysis, such an accurate query design is undesirable for two reasons. First, tuning the query parameters required numerous iterations. However, quality scores

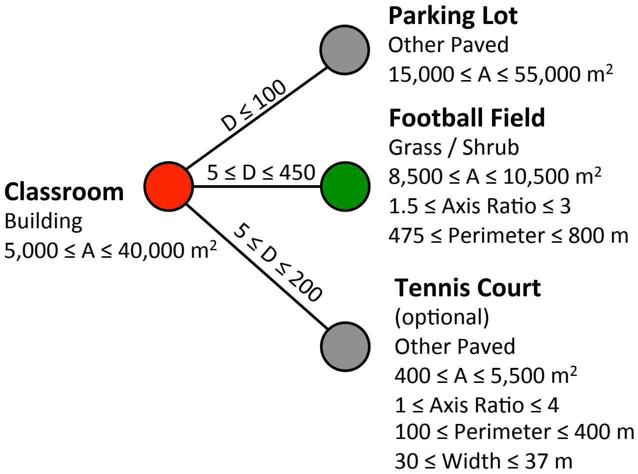


Figure 2.1. A simple high school query template.

and uncertainty are most informative during initial stages of this tuning process because they help the user to quickly identify true positives and narrow the scope of the search. Second, many queries may never produce such accurate results, so performance in the context of ambiguous results is critical. For example, a query designed to identify big-box retailers may not be able to separate them from supermarkets, mega-churches, and furniture warehouses.

After specifying the search template and querying the graph, the analyst receives back a list of candidate high school matches, each with associated node and edge attribute values, and a map indicating the location of each candidate. Table 2.1 shows a list of matches for the high school template for one United States county (true positives in bold font), while Figure 2.2 shows the associated map. The list of attribute values represents a clear reduction in the amount of data that the analyst must consider to locate the high schools. Likewise the map provides context for each candidate match without requiring that the analyst manually consider the entire content of the underlying imagery.

Nevertheless, a visual inspection of the attribute table does not reveal a simple pattern for distinguishing true high schools from false positives. The analyst must therefore manually look at the imagery for each candidate match to determine which are true positives. Manual verification is probably acceptable for the 40 results returned in the example. However, suppose the imagery and search spanned the entire United States, which has over 37,000 public and private secondary schools.[1] Then the analyst would be overwhelmed, particularly if the task included separating existing high schools from new.

The problem with Table 2.1 is that it provides no information about the relative quality of the individual candidate matches or of the level of uncertainty associated with that quality. The job of identifying candidate matches for the pattern of interest has gotten easier, but the problem of separating true positives from false positives and unknowns has not. In the remainder of this section, we discuss several desired solution properties and possible approaches, followed by a brief summary of likely next steps. For performance results, see Stracuzzi et al. (2015).

2.2 Desired Solution Properties

The primary goal in adding match quality scores and uncertainty estimates to semantic graph search is to make the analyst task of finding patterns in large data corpora as simple, efficient, and reliable as possible. The quality scores provide a measure of how well individual match candidates meet the specified search criteria, while the uncertainty intervals provide a measure of score reliability. Taken together, the scores and intervals can provide a basis for analysts to rank the match candidates and determine which candidates require more careful consideration. A viable solution must therefore satisfy a number of constraints, which we summarize below.

[1] From the U.S. Department of Education: `http://www2.ed.gov/about/offices/list/ovae/pi/hs/hsfacts.html`

ID	Classroom	Football Field			Parking	…	Distance		
	Area	Area	Axis	Perim	Area		C-FF	C-P	C-TC
1	8400	9638	2.44	564	19141		390	33	126
2	6397	9638	2.44	564	19141		337	0	82
3	**15265**	**9638**	**2.44**	**564**	**51464**		**37**	**0**	**50**
4	7203	8727	2.12	605	33720		219	0	108
5	39333	10293	1.6	518	52621		263	58	--
6	**11567**	**9801**	**2.15**	**616**	**36631**		**84**	**0**	**192**
7	12815	8711	2.59	584	24695		166	0	23
8	**22402**	**10083**	**2.16**	**506**	**40549**		**56**	**0**	**44**
9	6088	9092	2.15	596	35250		394	28	168
10	6188	10083	2.16	506	40549		103	78	--
11	6549	9092	2.15	596	15705		283	55	130
12	6438	9092	2.15	596	15705		293	75	111
13	7265	9092	2.15	596	15705		36	0	136
14	5314	9092	2.15	596	15705		132	88	115
15	7583	8767	1.65	767	19922		380	41	--
16	8340	9092	2.15	596	26468		22	28	125
17	6656	9092	2.15	596	26468		206	33	--
18	13049	8767	1.65	767	35640		422	69	162
19	10510	9092	2.15	596	26468		71	0	156
20	8662	9092	2.15	596	32336		363	0	32
21	**14011**	**9428**	**2.52**	**499**	**27136**		**88**	**0**	**132**
22	10839	9624	1.52	700	23279		447	0	169
23	13528	8811	2.47	567	50562		274	58	35
24	14285	8596	2.17	510	44017		290	0	82
25	**33281**	**9211**	**2.43**	**589**	**44762**		**274**	**0**	**--**
26	**17081**	**8811**	**2.47**	**567**	**50562**		**122**	**0**	**180**
27	5919	8811	2.47	567	36834		432	0	--
28	19519	9921	2.25	533	23682		280	0	20
29	6365	9921	2.25	533	23682		399	38	14
30	**17947**	**9921**	**2.25**	**533**	**21297**		**173**	**0**	**71**
31	11411	9904	2.3	499	16301		435	26	69
32	7522	9904	2.3	499	23798		199	29	170
33	**18938**	**9714**	**2.45**	**519**	**33207**		**118**	**0**	**76**
34	9402	9530	2.44	584	32196		314	18	--
35	**14802**	**9530**	**2.44**	**584**	**32196**		**95**	**0**	**26**
36	**13640**	**8957**	**2.36**	**773**	**41296**		**46**	**0**	**77**
37	**16136**	**10235**	**2.23**	**562**	**46364**		**75**	**0**	**23**
38	**16406**	**10016**	**2.22**	**510**	**46065**		**62**	**0**	**46**
39	9787	10016	2.22	510	46065		408	0	--
40	**14431**	**8989**	**2.44**	**593**	**21627**		**362**	**0**	**27**

Table 2.1. Match candidate attributes for the high school query for one county. Bolded entries indicate true high schools while the others represent false positives.

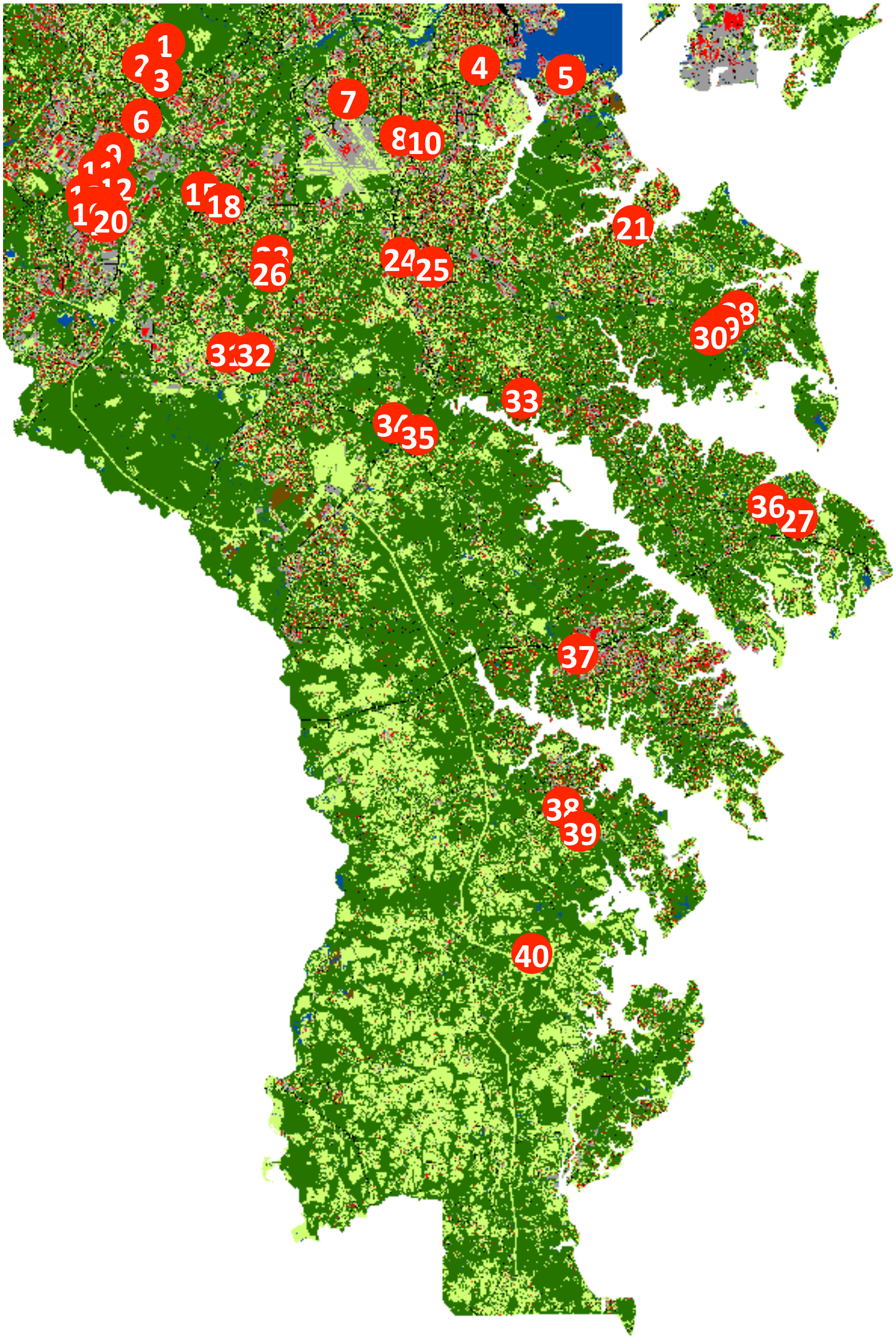


Figure 2.2. Locations of match candidates for the high school query overlaid onto map showing primitive semantic objects.

1. *Efficient Computation:* We expect that quality and uncertainty will need to be calculated very frequently due to large numbers of matches and due to repeated queries that make incremental improvements to the search template.

2. *Domain Knowledge Incorporation:* Domain knowledge, whether from labeled examples of the target pattern or from elicited expertise, can have a substantial impact on both quality assessment and uncertainty. Methods should incorporate as much information as possible into candidate evaluation, yet still function if no such information is available. Likewise, incomplete or missing data should be tolerated to the extent possible.

3. *Practical Knowledge Elicitation and Data Labeling:* As a corollary to the previous constraint, methods should not depend on information that is difficult to obtain. For example, analysts cannot make accurate guesses at complicated distributions during an elicitation, nor can they label hundreds of examples.

4. *Intuitive Interpretation:* Semantic graph users will not necessarily be experts in statistics or computing. For example, scores and uncertainties should display a monotonic response to changes in the query. Likewise, quality scores can be viewed as similarity or distance functions between the search template and candidate match, so they should satisfy the triangle inequality.

5. *Capture prioritization:* Many factors influence the relative quality of match candidates, but some are more important than others. For example, one high school candidate may include a perfect football field, parking lot, and tennis court, but no classroom building. A second candidate may have all four elements present, but each may be individually suboptimal. The classroom building is critical to the definition of a high school, so the latter match candidate should score strictly higher than the former.

6. *Quality scores and uncertainties should be independent of the semantic representation.* Different sensors support different levels of semantic abstraction. For example, optical imagery may support recognition of buildings as primitives, while synthetic aperture radar (SAR) may separately identify walls, roofs, and shadows which indicate a building when observed in a specific pattern. The quality score and uncertainty intervals should not change for these two cases if the underlying data is comparable and indicates an equivalent match.

7. *Results should be independent of the order in which template components are evaluated.* For example, the quality scores should not change if the classroom building gets evaluated before or after the football field. Note however that a fundamental change to the template structure, such as using the football field as the main point of reference instead of the classroom building, might change the resulting scores and uncertainties.

8. *Quality scoring and uncertainty calculations should support varying levels of contribution from template components.* In many applications, some components of the search template may be less important than others. For example, not all high schools have tennis courts, so the lack of a tennis court should not force a candidate to a low score.

In other cases, the number of component matches may be important. For example, large numbers of fuel storage tanks increase the quality of refinery candidates.

9. *Solutions should accurately evaluate rare items.* In some applications, analysts must identify rare patterns in data. Overwhelming numbers of highly scored false positives, and any low-scoring true positives, are detrimental to analyst performance.

10. *Solutions should generalize across domains.* Assumptions made by estimation methods should be theoretically justified. Tunable modeling parameters that require empirical adjustment for each application are undesirable.

Ideally, analysts would focus their effort and attention on the candidates that have the highest uncertainty. In practice, this means that analysts should be able to rely on the automated scores for matches that score either very high or very low with low uncertainty and give manual consideration only to candidates with intermediate or highly uncertain scores. Figure 2.3 shows an example. Analysts should be able to minimize the attention paid to the match candidates *High School 1* and above because they all have high quality scores and low uncertainty. Likewise, they should be able to ignore all of the candidates *Salon and Spa* and below based on the low scores and small uncertainty intervals. The remaining 11 candidates have either middling scores or large uncertainty intervals, and so require analyst attention. These 11 represent a substantial reduction from the initial set of 40 candidates.

Note that the quality scores do not perfectly separate the true high schools from false positives. This is acceptable (and anticipated), but may require additional performance

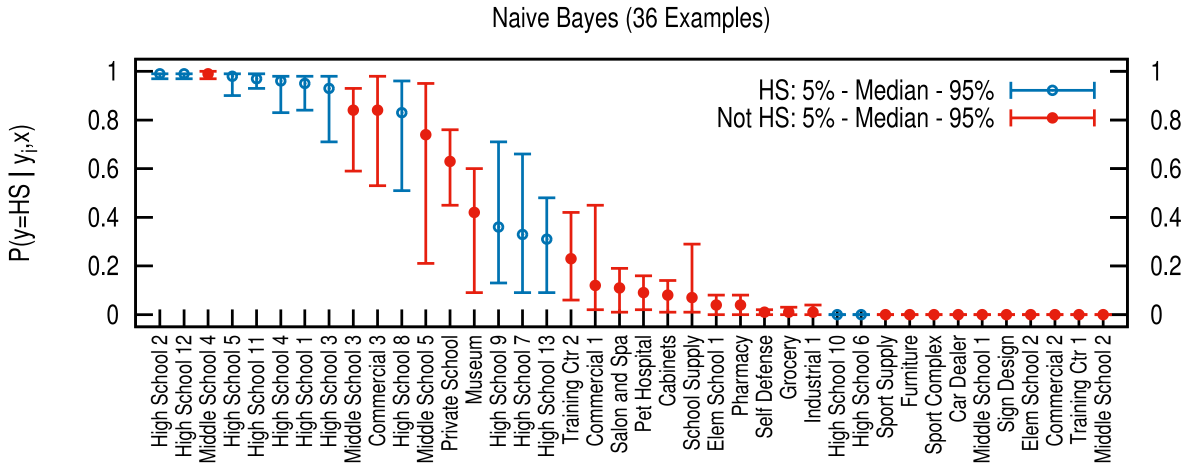


Figure 2.3. High school search results ranked by quality score with uncertainty intervals.

characterization to engender analyst trust. For example, statements such as *80% of candidates with quality score above 0.8 represent true positives* help to provide the additional information needed to support analyst goals. Stracuzzi (2015) discusses performance characterization along with several other analyst community needs in a companion report on uncertainty analysis in national security missions.

2.3 Methods

In the following, we discuss the methods that we considered for determining match quality scores. We do not revisit the three methods, elicitation-based beta distributions, naïve Bayes, and the distance-based quality metric, discussed in detail by Stracuzzi et al. (2015). However, we do discuss three other methods not pursued, two of which warrant future consideration.

Bayesian networks (Pearl, 1988) model complex joint probability distributions by taking advantage of known dependencies among the variables. Each node in a Bayesian network represents a conditional probability table that describes the effects on a variable by its dependencies. Given a set of observations, inference algorithms can estimate the probability of any other variables in the network. For semantic graphs, the observations are the attributes associated with each match candidate, such as the area and perimeter of nodes in the semantic graph, and the goal is to model the probability that the observed values indicate a match to the search query, such as a high school. To evaluate a candidate match, an inference algorithm walks the graph, incorporating the observed data and updating the conditional probabilities. The result is a posterior conditional probability that the candidate subgraph matches the query given the data.

Bayesian networks carry three advantages over traditional statistical modeling; they (a) tolerate missing data, such as from a sensor error, (b) can incorporate online user feedback into the inference process, and (c) model the query structure and dependencies. The primary disadvantage is that they require large amounts of *labeled* training data to populate the conditional probability tables. Consequently, they cannot evaluate queries for rare patterns, previously unobserved patterns, or patterns for which no labeled data is available. Bayesian networks also rely heavily on the correctness of the assumed dependencies among variables. In practice determining these can be challenging, and the assumption that the search query will correctly capture these is very strong. Finally, although Bayesian networks produce a distribution over the possible values that a categorical variable may take (high school or not, for example), they are not typically used to quantify the uncertainty of the distribution, called credibility intervals, on the category probabilities. Determining how to produce them remains an important prerequisite for applying Bayes nets to the match quality and uncertainty problem. The naïve Bayes method outlined in Stracuzzi et al. (2015) used a Monte Carlo sampling technique to estimate the uncertainty intervals, but that approach would be impractical for the computationally expensive probability estimation methods used by full Bayesian networks.

The referenced article explores one possible distance-based quality metric in depth. That metric is based on a simple distance function that maps individual attribute values to distances based on how well the match candidate's attributes adhere to the constraints specified in the search template. The method requires the analyst to differentiate between *preferred* and *allowable* attribute values, with the latter indicating a range of suboptimal but still acceptable attribute values. The extra specification amounts to a form of elicitation, and is ultimately somewhat brittle in that the analyst ultimately must specify hard constraints on the range of attribute values.

An alternative to the distance mapping function, earth mover's distance (Rubner et al., 1998), also known as the Wasserstein metric (Wasserstein, 1969) or Mallows' distance (Mallows, 1972), computes the amount of "effort" required to convert the set of attribute values observed in the candidate match into the target set specified in either the query template or a set of exemplars. Greater differences between the observed values and the template require more effort. More specifically, earth mover's computes a distribution over likely conversion efforts. This distribution can then be used to associate confidence intervals with mean similarity values. Importantly, earth mover's distance extends naturally to the case in which the observed attribute values are represented as distributions with uncertainty intervals instead of the point estimates used in the preceding discussion (see Chapter 3). Open questions with respect to the earth mover's distance include identifying sufficiently efficient implementations for the semantic graphs application and establishing (or removing the assumption of) the distance measures used to indicate the relative cost of moving values from one bin (area) to another (perimeter).

Finally, one remaining avenue for exploration concerns the Dempster-Shafer theory of evidence (Shafer, 1976). This mathematical framework combines evidence from multiple sources to compute an uncertainty interval instead of a distribution over possible values. It is typically applied to complex decision-making problems, including sensor fusion. In this framework, the posed search question becomes *how much evidence can we find that is consistent with the given query pattern* while the response becomes *at least x and at most y*. The difference between the two indicates uncertainty, for example due to ambiguous and conflicting evidence. Though well-developed, the Dempster-Shafer framework is quite complex, both computationally and conceptually.

2.4 Future Work

The discussion section of Stracuzzi et al. (2015) highlights several issues that require further study. The most important of these concerns extending all of the methods to handle uncertainty in the geospatial boundary locations and semantic labels. Chapter 3 below describes preliminary results on extracting the needed uncertainty information from the data along with initial thoughts on extending the semantic graph representation to use it. Other directions for additional research include improvements to computational efficiency and the knowledge elicitation process. See the referenced journal article for more detailed discussion.

A second, graph-related improvement relates to double counting of search results. In many cases, multiple candidate matches cover a single region, often sharing component nodes. Separating true positives from false positives is difficult in this case, as evidenced by the confusion between high schools and co-located middle schools. In practice, it may be sufficient to recognize that multiple matches refer to a single region, and then direct the analyst to the region, possibly by reporting only the highest probability or score for the associated set of candidates.

A more fundamental expansion of the work involves establishing a feedback loop between the analyst and the computational analysis. For example, when an analyst considers the candidate high school matches shown in Table 2.1, he may identify some as true positives or false negatives. This information could be used to improve both the labeling process used to generate input for the semantic graphs and the matching process used to identify candidates. Doing so requires propagating semantic and geospatial labeling information backward from abstract labels such as *high school* to primitive labels such as *pavement*.

One approach to supporting this type of backward information propagation relies on structure in the match evaluation process. The high school search is defined in terms of objects like classrooms and football fields. However, these semantic classes are not extracted from the raw data. Instead, we have buildings and grassy areas. Currently all of our methods collapse the high school search template to depend directly on buildings or grassy areas and evaluate candidates as a single, unstructured collection of nodes and attributes. This simplifies the scoring and uncertainty estimation into a combination over a set of features.

Unstructured evaluation can lose important subtleties, however. For example, a candidate that represents a horse racing track may get a good evaluation because every feature except the "football field" area matches perfectly (the infield of a horse track resembles an oversized football field). In practice, many queries naturally break into a hierarchy of components, such as the football field and tennis courts. These can be evaluated independently of the larger high school, forcing higher level features such as the football field to better conform. Stracuzzi and Könik (2008) discuss an approach to this issue that would pair naturally with naïve Bayes.

Chapter 3

Data Integration Under Uncertainty

The work discussed in Chapter 2 makes an initial effort at calculating the uncertainty associated with quality scores. In particular, the probabilistic methods capture concept variance, which corresponds to the number of non-target objects that satisfy the search pattern. In the high school example, this includes a number of middle schools and shopping centers among others. Conversely, the distance-based metric captures a crude form of uncertainty in the data by assuming a known standard deviation to the errors associated with boundary locations. From there, the method samples a range of possible attribute values such as area and uses the results to establish uncertainty intervals for the distance (quality) scores.

Both methods fail to adequately capture the uncertainty due to the source data, such as uncertainties that stem from integrating multiple data sources. Explicit quantification of how well two data sources agree, how frequently a given pattern appears across a region, or how well a given set of observations fit a target pattern can strongly impact subsequent decision making. Ultimately the user's ability to make decisions based on analytic results determines the value of both the analytic framework and the data itself. Put another way, data and analysis are valuable to the extent that they reduce uncertainty with respect to some consequential decision.

The large volume and velocity of available data means that early analytical steps, such as integration of multimodal data, need to be largely automated. For our purposes, multimodal data refers to data from multiple sources (not limited to imagery), at varying resolutions and geographical and temporal coverage. Integration in this context refers to combining data sources to produce a more complete view of the state and activity in the sensed area than any one source could produce alone (data fusion is a subset of integration). As a result, uncertainty analysis also plays a substantial role in determining the quality and success of the integration (Simonson et al., 2007). The work reported in this section takes the view that these questions can best be answered by performing an end-to-end uncertainty analysis, starting from the raw data and propagating through the most complex search patterns and results.

In this section we develop a framework for uncertainty analysis in remote sensing domains, and provide some initial application results. The framework needs to account for the integration of data produced by multiple sensor modalities and needs to support higher-level geospatial pattern analysis, such as the geospatial semantic graphs discussed by Brost et al. (2014). Broadly speaking, we hypothesize that the analysis results can be used to assess the relative value of individual sensors and collections. More specifically, the target framework should provide information about the uncertainty contributions of individual sensors and identify the value of including new data from external sensors. For example, we would like to be able to identify which sensor contributes the majority of uncertainty to a result, whether a result's uncertainty is a product of disagreement among multiple sensors, or whether adding data from a new source will improve results. In the long run, this type of uncertainty analysis may also provide insights into sensor tasking.

3.1 Sensor Data Analysis

Our primary technical objective is to develop consistent and uniform methods for analyzing data from multiple sensor modalities. Current approaches to sensor data integration typically follow one of two paths. In the first, data is co-registered and fused to form a rich feature vector describing each sampled point. These feature vectors are then segmented and labeled into objects relevant to the intended analysis. An example of this would be fusing optical and LiDAR data collects over the same region to form a land cover map. In the second path, sensor modalities are initially treated separately. Objects of interest are segmented from the raw data and labeled using the most appropriate algorithms for the given sensor source. Later, the results from different modalities are merged, which often requires specialized algorithms for co-registering the data, resolving segmentation and labeling differences, and combining any available uncertainty information. For example, subsequently collected synthetic aperture radar (SAR) data could be used to identify changes in the aforementioned land cover map. Importantly, the SAR data would typically be processed using different algorithms and parameters from optical and LiDAR data collected for the same geographical region.

We call both approaches "non-uniform" because the details of the fusion and merging steps depend on the data sources. Some specialization is unavoidable; feature constraction depends heavily on sensor physics. However, when the process of detecting predictive features is dependent on sensor combinations, such as during sensor fusion, then the problem becomes combinatorial. Thus, while achieving state of the art classification accuracies may require deriving sensor-specific features, this should occur on a sensor-specific basis and should not need to be revisited when other sensors change. In other words, sensor exploitation occurs for each individual sensor, not in a combined, multi-sensor space.

In the following, we provide a brief survey of the many analytic steps associated with combining and extracting semantic regions from sensor data. Each of these steps can introduce new uncertainties to the result, in addition to the uncertainty associated with the data itself. The steps listed below should be viewed neither as a strict ordering nor as a single monolithic process. Certain steps may be unnecessary in some situations, while in other cases, some steps may be visited multiple times. In general, we assume that the process of integrating multiple data sources will be incremental, meaning that each step will produce an output that may serve as input to a subsequent step. However, the boundaries between individual steps listed below are not well defined. For example, registration and fusion can be performed simultaneously in some cases. Likewise, segmentation and classification can be performed in any order or simultaneously.

The analysis process is also likely to be iterative in nature. Both individual algorithms and sequences of them may need to be revisited to achieve the desired results, particularly in light of results achieved later in the process. This is a significant complication from the perspective of developing an automated framework for integrating and analyzing multi-source data. However, it also implies an opportunity to propagate both the results and objectives of high-level analytical processes, such as pattern search, back through earlier stages, such as the integration of raw data sources. Such *top-down* guidance may prove useful.

1. *Registration* aligns multiple data modalities such that any given location or object is comparably described by each. For example, LiDAR collects data about location height, while optical sensors collect data about color. To use both data sources together, they must be transformed into a single coordinate system and scale. Registration is itself a multi-step process, for which Goshtasby (2005) provides a detailed summary and survey of associated methods. The literature is packed with methods for interpolating among data points to align image pixels, most of which are specialized for the specific data sources considered. Relatively few methods attempt to produce uncertainty estimates. Simonson et al. (2007) and Domokos et al. (2008) both discuss uncertainty analyses for binary images, arguing that converting to binary greatly simplifies the registration process.

2. *Sensor fusion* synthesizes new features from multiple, co-registered data modalities. As with registration, the literature is dominated by domain- and application-specific methods. Among general methods,Cressie and Johannesson (2008) introduce fixed-rank kriging (FRK), a framework for performing fusion on large geospatial data set. Nguyen (2009) and Braverman et al. (2010) provide a detailed discussion of issues in fusion of multi-source remote sensing data and generalize FRK to the multi-source domain. Nguyen also discusses the relationship between kriging and other methods, such as geographically weighted regression (an interpolation method) and multi-scale methods, such as Bayesian hierarchical models.

   One common computational concern is that large data sets can quickly overwhelm basic algorithm implementations. Sun and Genton (2012) reviews several approaches for handling large geospatial data sets, including methods based on maximum likelihood estimation, Bayesian methods, and kriging. Xu et al. (2015) extend the Bayesian approach to the spatiotemporal domain.

3. *Segmentation and classification*, though distinct processes, are tightly linked. Segmentation partitions an image in distinct, approximately homogeneous regions or objects while classification assigns a label or other semantic tag to the partitions. A huge number of methods for performing these tasks have been studied (Russ, 2011), though relatively few consider uncertainty at all. A 15 year meta-analysis of the remote sensing image classification literature showed that, for a variety of reasons, performance

has not improved appreciably (Wilkinson, 2005). Wilkinson also notes that lack of consideration of labeling utility and uncertainty may be a cause of poor performance.

Segmentation and classification can be done in either order. One option is to first group individual pixels (or other derived local feature values), such as by clustering, into regions that are then featurized and classified as a single object. Lelandais et al. (2014) discuss segmentation of multimodal medical images, distinguishing between uncertainty due to sensor noise and imprecision due to sensor limitations. Saad et al. (2010) adds the notion of using shape and appearance priors derived from expert segmentations (also in the medical imaging domain). Shi et al. (2012) discusses a graph-based method for combining both registration and segmentation in medical images while estimating uncertainty. Lizarazo and Elsner (2009) discusses a method for producing fuzzy or overlapping segments, though they do not explicitly calculate uncertainty.

The alternative approach is to first classify individual pixels (or other local feature vectors) and then merge adjacent pixels with the same label into uniform regions. This usually requires some adjustment to account for noisy and spuriously labeled pixels. Gonçalves et al. (2008) shows one method for incorporating uncertainty analysis into this approach. Martin et al. (2006) discuss methods for evaluating classification and segmentation results given that the ground truth itself is uncertain.

4. *Spatial region attribution* concerns the process of deriving properties of the regions from the segmentation and classification, such as area, perimeter, and distance to other regions. Deriving these values from the uncertain segment boundaries or class labels raises a new set of issues. For example, Fonte et al. (2013) discusses the relationship between classification uncertainty, which can be estimated directly from the available data, and classification accuracy, which cannot. Similarly, Canters (1997) discusses the relationship between the uncertainty of region area estimates and the methods used to classify the region.

5. *Pattern matching and similarity metrics* focus on the problem of determining how well a selection of image objects and their associated attributes match a target pattern. Incorporating uncertainty into semantic graphs raises several issues related to matching and similarity. In particular, the semantic labels represent a crude approximation of the meaning assigned to image objects. The probabilistic and uncertainty information associated with the image objects takes into account at least some of the context surrounding the objects, but this will typically be lost by the semantic label. As a result, even sophisticated semantic matching methods that attempt to match patterns based on their conceptual interpretations will tend to ignore important information. Gallagher (2006) reviews a variety of structural and semantic pattern matching techniques. Although some of these methods allow for the possibility that relationships among objects may be probabilistic, most do not consider that the objects themselves may exhibit uncertainty in their labels or attributes. Likewise, Lin (1998) develops an information-theoretic definition of similarity, but incorporation of uncertainty would require extension of the framework.

3.2 Technical Approach

Our approach is based on the hypothesis that by choosing the correct representation prior to data integration and fusion, much of the specialization currently required for sensor integration can be avoided. Specifically, we will produce a probabilistic representation of both the semantic labels and the geospatial extent of the sensor-specific semantic objects. However, while many methods focus on extracting the most likely labels and segment boundaries from the methods, we assert that the underlying distributions contain far more valuable information. We also differentiate the proposed work from the common practice of measuring the probability of detection (of a given feature) and false alarm, and then using receiver-operating characteristics (ROC curves) as a measure of detection performance. In practice, this approach only speaks to how well the algorithm performs across a large sample of data, not the uncertainty associated with the algorithm's individual decisions.

3.2.1 Unsupervised Segmentation and Labeling

Our approach builds on the Bayesian pixel classification methods surveyed by (Falk et al., 2015). Generally speaking, the data samples (such as the pixels shown in Figure 3.1a) are first probabilistically clustered. The clusters then establish the set of categories to which individual samples are assigned, and we represent each sample as a probabilistic mixture of the possible categories as shown in Figure 3.1b. We use the term *category* to refer to the unsupervised clusters, and *class* to refer to the semantic tags used to label the categories. Classes can be assigned to each unsupervised cluster either manually by the user, or through a supervised labeling process.

The clustering process therefore creates a probabilistic segmentation and classification of the data (image or otherwise). Taken together, these label distributions also indicate a

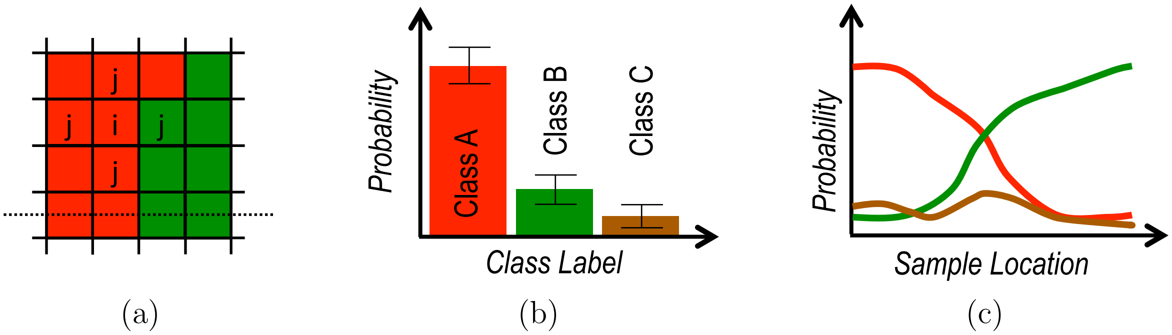


Figure 3.1. Data model used to determine class mixture probabilities. Panels show (a) a typical set of image pixels, (b) pixel i's label as a mixture of all possible labels, and (c) a view of the pixel label probabilities as a function of location.

distribution over possible segment boundary locations. Specifically, if we take a slice through the samples (dashed line through Figure 3.1a), the resulting probability distributions indicate the change in label probabilities relative to sample location, as shown in Figure 3.1c. Thus our method provides all of the information needed to establish the geospatial attributes as probability distributions.

More formally, consider the regularly spaced lattice indexed by $i = 1, \ldots, N$ depicted in Figure 3.1a. Each index refers to a specific pixel which may belong to one of K categories or states. For example, these categories may refer to land cover categories or pixel colors. The true state, z_i, for each index is not directly observable and is referred to as a latent variable to be predicted in the course of the analysis. However, available to us are noisy observations y_i regarding pixel i, corresponding for example to RGB values for optical images, or height values for LiDAR. The state of each pixel is a random variable conditioned on the observations and the assumed model. Our goal is the posterior estimation of the model parameters and therefore the true state conditioned on the observations.

To simplify our analysis, we assume that the state of pixel i is related only to the pixels in its immediate neighborhood. Notationally, we will refer to the neighboring pixels in toto with index j as $i \sim j$ and the neighborhood of pixels will be referred to as δ_i . The model can be easily extended to include pixels beyond the one spatial unit shown in Figure 3.1a with the only restriction being the resulting computational burden.

Given an image with N pixels, the observed state of the image $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ can be described by a mixture model

$$g(\mathbf{y}|\boldsymbol{\theta}) \approx \prod_{i=1}^{N} \sum_{k=1}^{K} \lambda_k f(y_i|\theta_k), \tag{3.1}$$

where $\lambda_k > 0$ with $\sum_k \lambda_k = 1$ represents a weight vector, $\boldsymbol{\theta}$ is a vector of parameters, and f are specified probability functions. In the case of a Gaussian mixture model where $\phi_j(y_i|\mu_j, \sigma_j)$ is the Normal density function, then $\boldsymbol{\theta}$ corresponds to the parameters $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ and $f = \phi_j$, yielding

$$g(y|\boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{j=1}^{K} \lambda_j \phi_j(y_i|\mu_j, \sigma_j). \tag{3.2}$$

The number of possible states K may or may not be known a priori. When known, a finite mixture model is sufficient to represent the distribution of pixel values in Equation 3.1. Otherwise, an infinite mixture model can be specified via a Dirichlet process prior (see Falk et al., 2015, for a review of possible approaches) and used to estimate the number of categories from the data.

The unobservable, true state of pixel i, z_i, is a multinomial random variable represented by an indicator vector with K possible individual states, $z_i = (z_{i1}, z_{i2}, \ldots, z_{iK})$. Each element of the pixel state vector can take on values $\{0,1\}$ if and only if the state of the pixel, $z_{ik}=1$ belongs to state k.

$$g(z_i|z_{\delta_i},\mathbf{y},\boldsymbol{\theta}) \sim \text{Multinomial}(1;\omega_{i1},\omega_{i2},\ldots,\omega_{iK}) \tag{3.3}$$

where ω_{ik} represents the posterior probability that pixel i belongs to class k.

We assume that the observations are drawn from a hidden Markov random field (MRF) with a general Gibbs distribution (Besag, 1974). The particular Gibbs model assumed here is the Potts model (Potts, 1952),

$$g(z|\beta) = \frac{\exp[\beta \sum_{i\sim j}\delta(z_i - z_j)]}{C(\beta)}\ , \tag{3.4}$$

where the indicator function δ explicitly limits the pixels of interest to those in the neighborhood of pixel i as discussed above, and $C(\beta)$ is a normalizing constant. The parameter β represents the strength of the similarity between neighboring pixels. A zero value indicates that the pixels are independent of their neighbors, while increasing values of β indicate increasing similarity between neighbors and higher likelihood that they belong to the same categories.

In a homogeneous Potts model the parameter β is a constant, while for an inhomogeneous model the parameter varies over the image. In the homogeneous case, the parameter can be considered a random variable, such as $\beta \sim Uniform(\zeta=0,\gamma=3)$. Alternatively, it may be a deterministic value that varies across the image. In the inhomogeneous case, for example: $\beta = 1-(|g_i-g_j|/\theta)$ where the g_i, g_j are the pixel color vectors and

$$\frac{1}{N}\frac{1}{n_i}\sum_{i=1}^{N}\sum_{i\sim j}|g_i-g_j|$$

is the average difference in image color values with n_i being the size of the pixel neighbor set (four, in the case of Figure 3.1a).

3.2.2 A Note on Generality

Our assumption of a Gaussian mixture model in Equation 3.2 may at first appear strong. We have no reason to expect that the category parameters follow a Normal distribution. Moreover, different data sources are likely to be characterized by different distributions. However, neither estimation of the unobservable state associated with each pixel in Equation 3.3, use of the Potts model in Equation 3.4, nor parameter estimation techniques such as Markov Chain Monte Carlo (MCMC) or Expectation Maximization (EM) exploit the Gaussian assumption. The combination of the assumptions of a Gaussian mixture models and a Multinomial distribution for pixel categories therefore provides a convenient mathematical vehicle for segmenting and classifying data sources without substantially influencing the outcomes.

Given the weakness of our modeling assumptions, how can we add available background information into an analysis? The traditional Bayesian approach is to alter the priors used by parameter estimation methods such as MCMC or EM. The priors serve as a bias that nudge the parameter search in a particular direction. A related alternative is to bias the starting point of the search by setting the initial parameter values. A third approach is to use the sample neighborhood, δ_i. Some data sources may respond better to larger or smaller neighborhoods, or to neighborhoods with different shape. Biasing the parameter search carries a certain risk — if the added bias is incorrect then the search may converge to a suboptimal solution. Such failures can often be overcome with sufficient training data and computation time, however.

3.2.3 Characterizing Uncertainty

The mathematical model described above supports characterization of uncertainty on two different levels. First, the category probability distributions associated with each sample (image pixel) identify uncertainty in the segmentation and labeling. For example, using Equation 3.3, we can construct maps showing the probability of each category over the entire image. The center of Figure 3.2 shows the probability maps for two categories derived from a small section of an image that was segmented into six categories, while the color bars on the outside of the image show the mapping between the colors and category probability values. Areas with the darkest colors indicate low uncertainty in the category identity, while areas with light colors for both categories indicate high uncertainty (the pixels could plausibly belong to either category). The probability maps therefore provide far more information about the data and analysis results than more traditional approaches, which would simply select and report only the most likely category. Importantly, the information about the confusion among different categories that probability maps capture so thoroughly can be only crudely approximated with methods such as confusion matrices, which rely only on the precision and recall of the most likely category.

The second level of uncertainty characterization considers the uncertainty in the category probability estimates for each data sample. These correspond to the error bars, called credibility intervals, shown in Figure 3.1b. While the segmentation uncertainty described above captures the degree to which the mathematical model can sort the individual samples into categories, the credibility intervals capture the degree to which different model parameterizations influence the results. More specifically, if we view parameter estimation as an optimization problem, then an issue arises when the optimization function has a large, flat minima. Many different parameterizations may be equivalent with respect to goodness-of-fit, yet result in substantially different category probability distributions for the data samples. To characterize these differences, we sample many parameterizations that are equivalent with respect to the optimization criteria and then calculate and record the resulting probabilities.

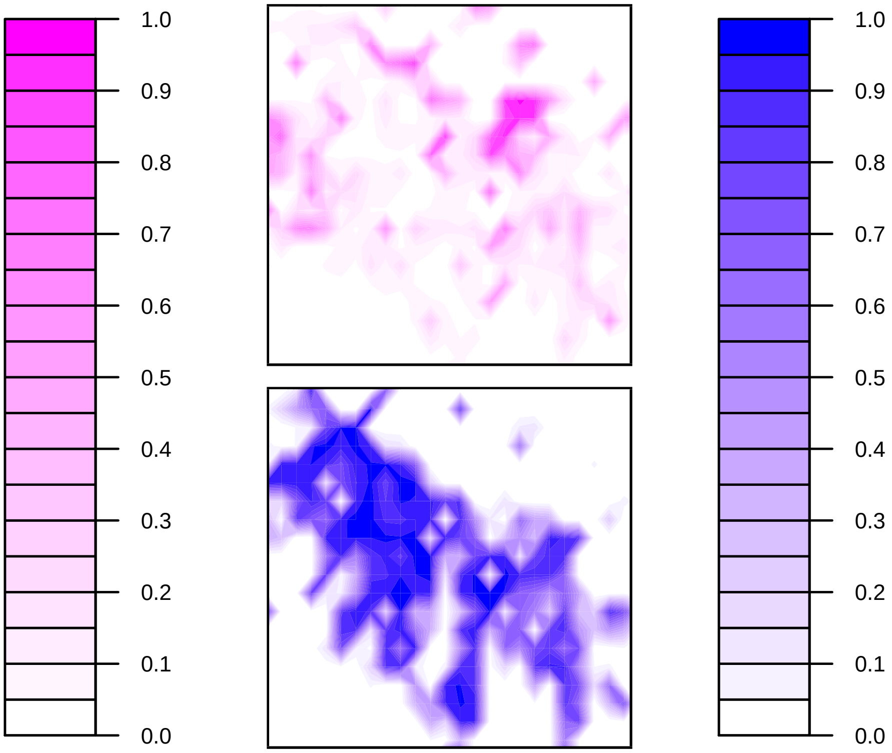


Figure 3.2. Two coincident category probability maps; sidebars show the mapping between colors and probabilities.

These probabilities then define the credibility distribution.

3.2.4 Combining Multiple Data Sources

Multiple data modalities can be combined in at least two ways. First, the data samples can be co-registered and merged into a vector that describes each data point. For example, the RGB values of an optical image can be merged with the LiDAR height value into a vector of four values for each pixel. After optionally applying fusion to the resulting vector, the vectors can be segmented and classified using the mixture model described above. The convenience of simply concatenating the co-registered samples from different data sources follows from the conjugate nature of the Normal and Multinomial density function. In effect, the added data simply re-weights the ω_{ik} in Equation 3.3. A second approach applies the segmentation and classification method to each data source separately. The resulting probability distributions (one per sample or pixel) are then co-registered and “convolved” — the exact operator is a point of future work — to generate combined distributions that contain information from both sensors.

The category probability distributions and associated uncertainty intervals can also be used to evaluate the value of various data sources. For example, suppose that we use optical imagery to produce the land cover classifications used to populate the geospatial semantic graphs described in Chapter 2. Now suppose that we acquire LiDAR data for the same geospatial region and add it to the analysis. Ideally, we would like to characterize the degree to which the added data improved our land cover classification. Traditionally, we would simply test the performance of the resulting classification with respect to accuracy or precision and recall or some other metric.

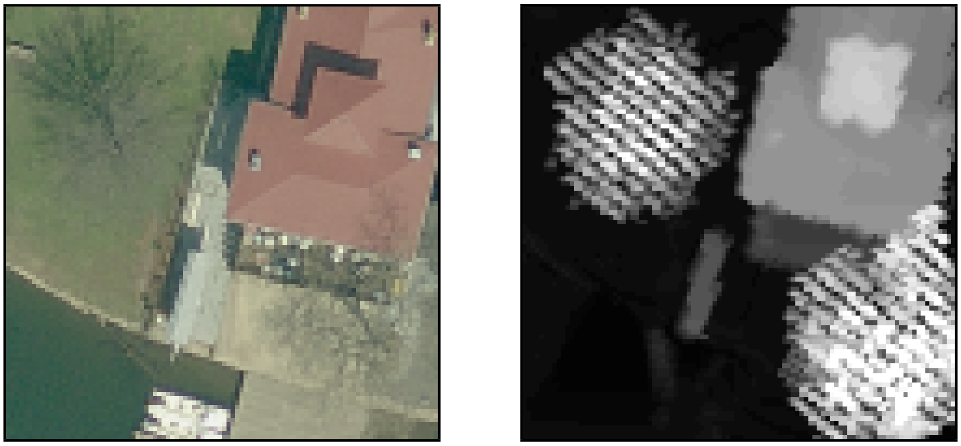
However, given the probability distributions and uncertainty intervals described above, we can also evaluate the changes in probability distributions and uncertainties *for each semantic class*. The result is a far more informative characterization of the changes made by the added sensor modality. For example, the added sensor may not substantially change the resulting land cover classifications or region boundaries, but may reduce uncertainty in both region labels and geospatial boundaries. More traditional tests such as precision and recall curves or confusion matrices cannot identify benefits such as reduced uncertainty.

Several possible approaches to evaluating the contribution of a given sensor are available. For the purposes of this work, we simply calculated the difference in probability at each pixel. However, several measures based on information theory may be more appropriate and more informative. These include Kullback-Liebler divergence (Kullback and Leibler, 1951), Shannon entropy (Shannon, 2001), Jensen-Shannon divergence (Lin, 1991), Hellinger distance (Nikulin, 2001), variation of information (Kraskov et al., 2005), and the Earth Mover's distance (Rubner et al., 1998). Each computation produces a subtly different measure of the difference between two probability distributions, and identifying which of these is most appropriate for our application remains a point of future work.

3.3 Results

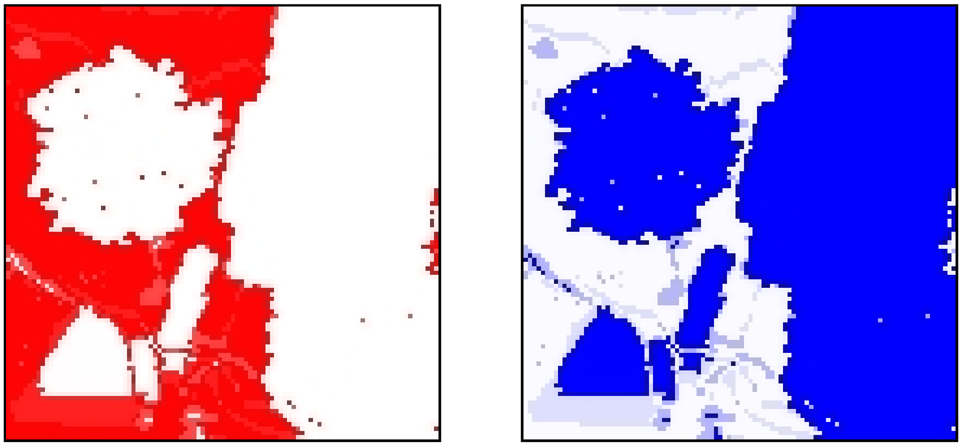
We tested the described methods using optical and LiDAR data collected from a small area near Philadelphia. The two sources, shown in Figure 3.3, were first co-registered. The diagonal lines in LiDAR image of the trees (panel b) appear to be an artifact of the the data collection combined with the uneven surface provided by the tree branches and foliage. We then applied the mixture model to each data source separately using only the individual pixels as input data (no neighborhood pixels). We used MCMC to estimate the model parameters, and the model determined the number of classes based on the data with a preset maximum of six. We used the mixture models to calculate the category probabilities, but extracting the credibility intervals remains a point of future work.

Figure 3.4 shows the result on the LiDAR data in which the model identified two categories corresponding to *tall* and *short*. The shade of color in the image corresponds to the category probability values for each pixel as they did in Figure 3.2. Notice that while most of the probabilities are very near to zero or one, some areas, such as the water, show greater uncertainty. High uncertainty over the water is expected for LiDAR, as the water tends to absorb the infrared sensor beam.



(a) (b)

Figure 3.3. Source optical (a) and LiDAR (b) imagery used for experiment.



(a) (b)

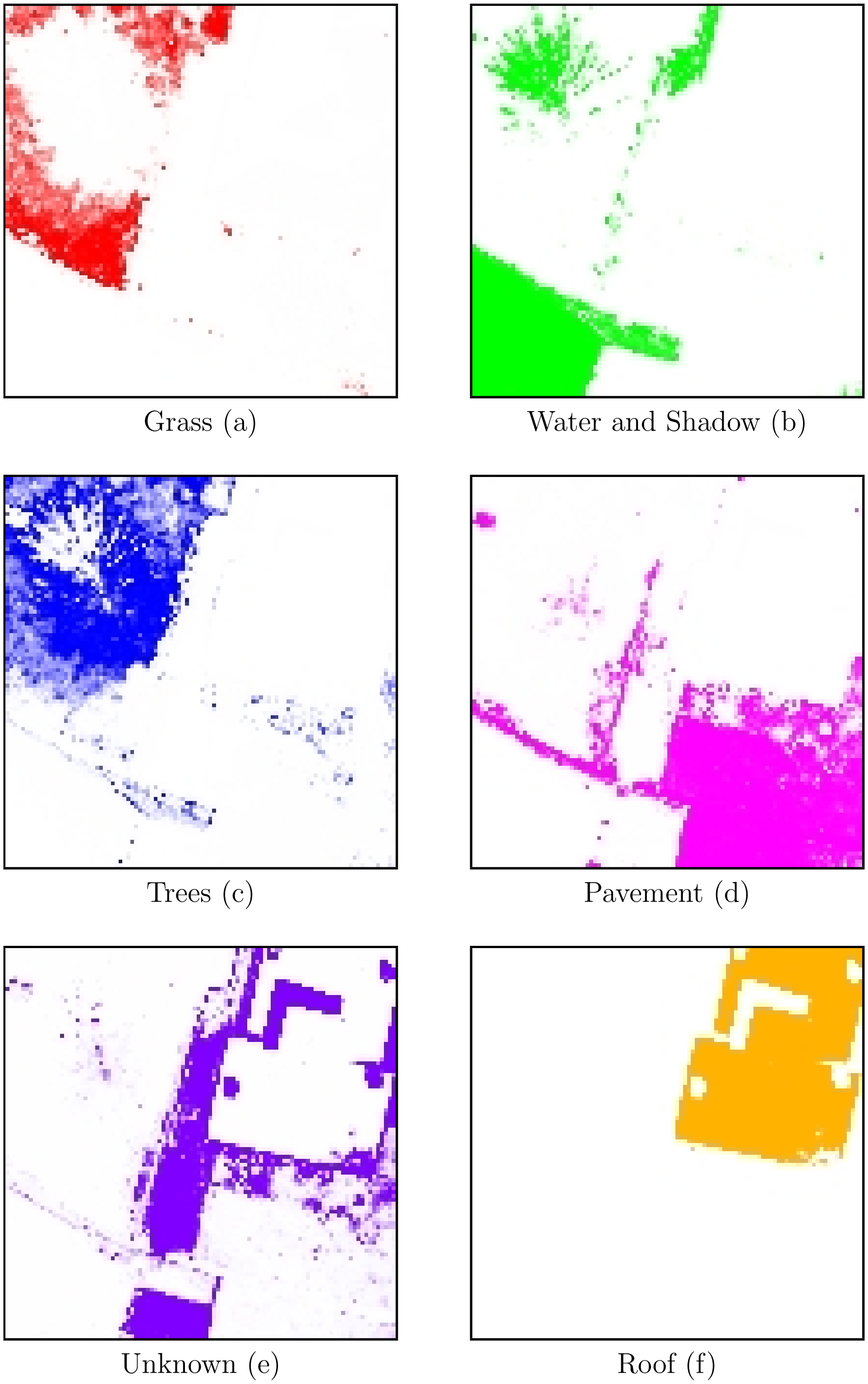
Figure 3.4. Probability maps for LiDAR imagery without the Potts model. Color intensity indicates probability, with darker colors indicating high probability while shades near white indicate low probability. Image (a) corresponds approximately to short things while (b) corresponds to tall.

Figure 3.5 shows the six categories that resulted from the optical imagery. To facilitate comparison among results, the category colors and names shown below each panel will remain consistent throughout the remainder of the report (except for LiDAR-only results). The optical imagery produces much more uncertainty and confusion among the categories. For example, the red category (panel a) corresponds to grass, but shows high uncertainty around the edges of the tree (due in part to the lack of leaves on the tree). Similarly, the green category (panel b) corresponds to a combination of shadow and water, and it shows substantial confusion with the tree itself, which is captured in the blue category (panel c). Other categories, such as the roof (panel f), show almost no uncertainty. In this case, the low uncertainty follows from the distribution of color through the source imagery in Figure 3.3 (a). The only red pixels in the scene belong to the roof; this type of uniformity is atypical.

Figure 3.6 shows the the six categories that resulted from the combination of optical and LiDAR imagery. In this case, the feature vector describing each pixel consisted of the red, green, and blue color values plus the LiDAR height value. One of the most salient effects of combining the data sources is the overall reduction in segmentation uncertainty. The overlap between the grass (panel a) and tree (panel c) categories is largely eliminated. A second major effect is the reclassification of the tree in the lower right corner of the images from pavement (panel d) to tree (panel c). The optical imagery alone cannot distinguish between a tree with no leaves and the pavement below it. Figure 3.7 highlights the impact of including LiDAR with the optical image by showing the difference between the probability maps from Figures 3.5 and 3.6.

Figures 3.8 through 3.11 show the probability maps for LiDAR, optical, combined, and difference respectively, this time using the Potts model with the four-neighborhood illustrated in Figure 3.1a. Several important differences between the two sets of results stand out. First, the model produced a more fine-grained segmentation of the LiDAR imagery (five versus two categories) shown in Figure 3.8. The diagonal lines in source image over the trees are captured in (panels b and c). Also note that the tall center section of the building roof was segmented separately from the other, lower sections, though it's unclear as to why the tall roof section was segmented together with low areas of the ground in panel d. The second clear impact of using the Potts model is reduced uncertainty in both the optical and combined results shown in Figures 3.9 and 3.10 respectively. In both cases, almost all of the segment boundaries transition almost immediately from very high probability to very low probability for a given category.

Figure 3.12 illustrates the changes in category probability enacted by using the pixel neighborhood functions. Many of the uncertainties indicated in the combined result without the Potts model, such as in the upper left corner of Figures 3.6a–c have been removed in Figure 3.10. This is a strict improvement. Also beneficial is the reduction in the number of pixels categorized as unknown (panel e). Conversely, some changes represent classification errors. For example, a large section of pavement in the lower middle of panel d has been remapped to grass (panel a). Similarly, some of the roof sections have been remapped to pavement (panels d and f) or unknown (panels e and f). This shift in roof categorization



Grass (a)

Water and Shadow (b)

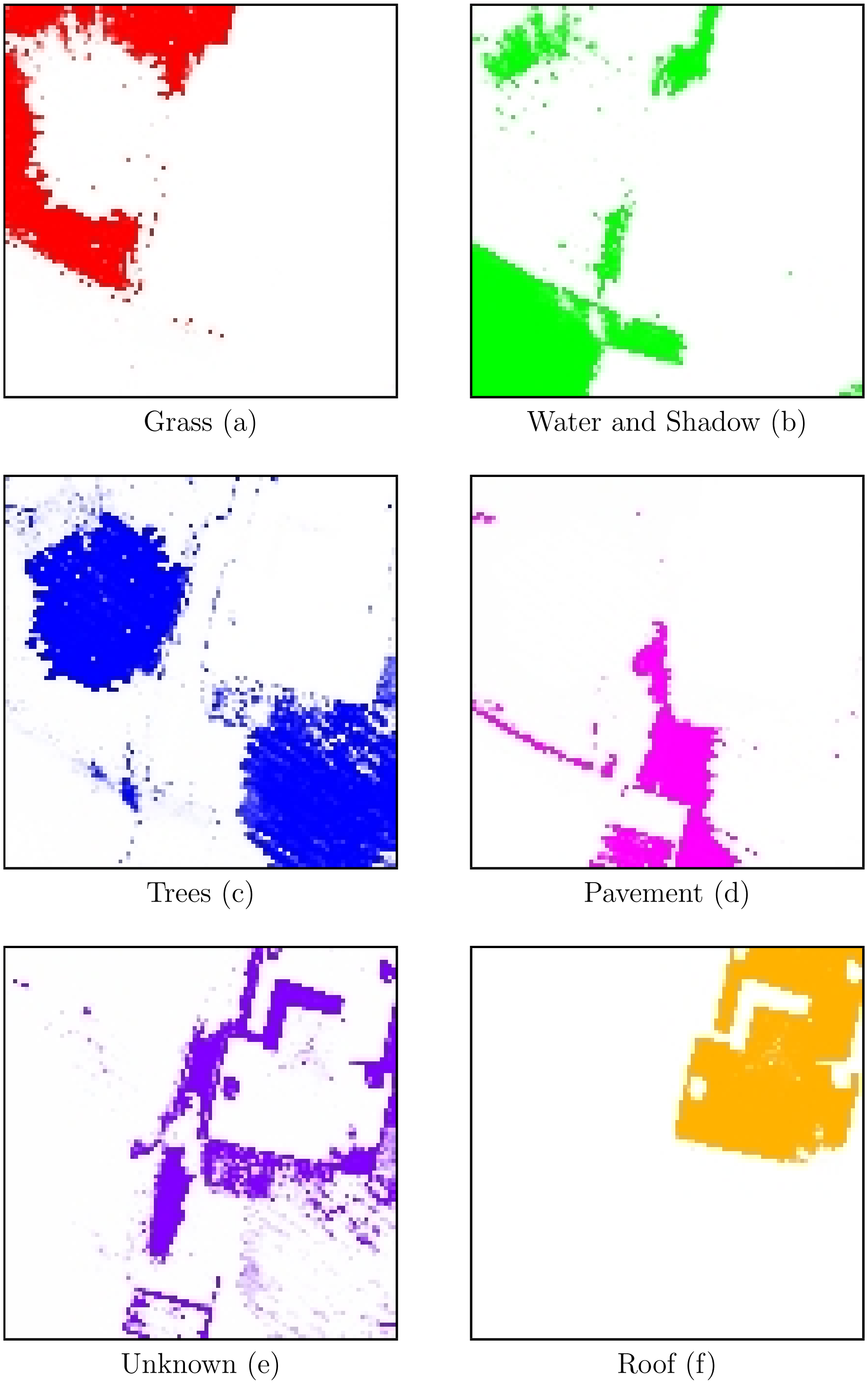
Trees (c)

Pavement (d)

Unknown (e)

Roof (f)

Figure 3.5. Probability maps for optical imagery without the Potts model. Manually identified semantic names shown below each panel.



Grass (a)

Water and Shadow (b)

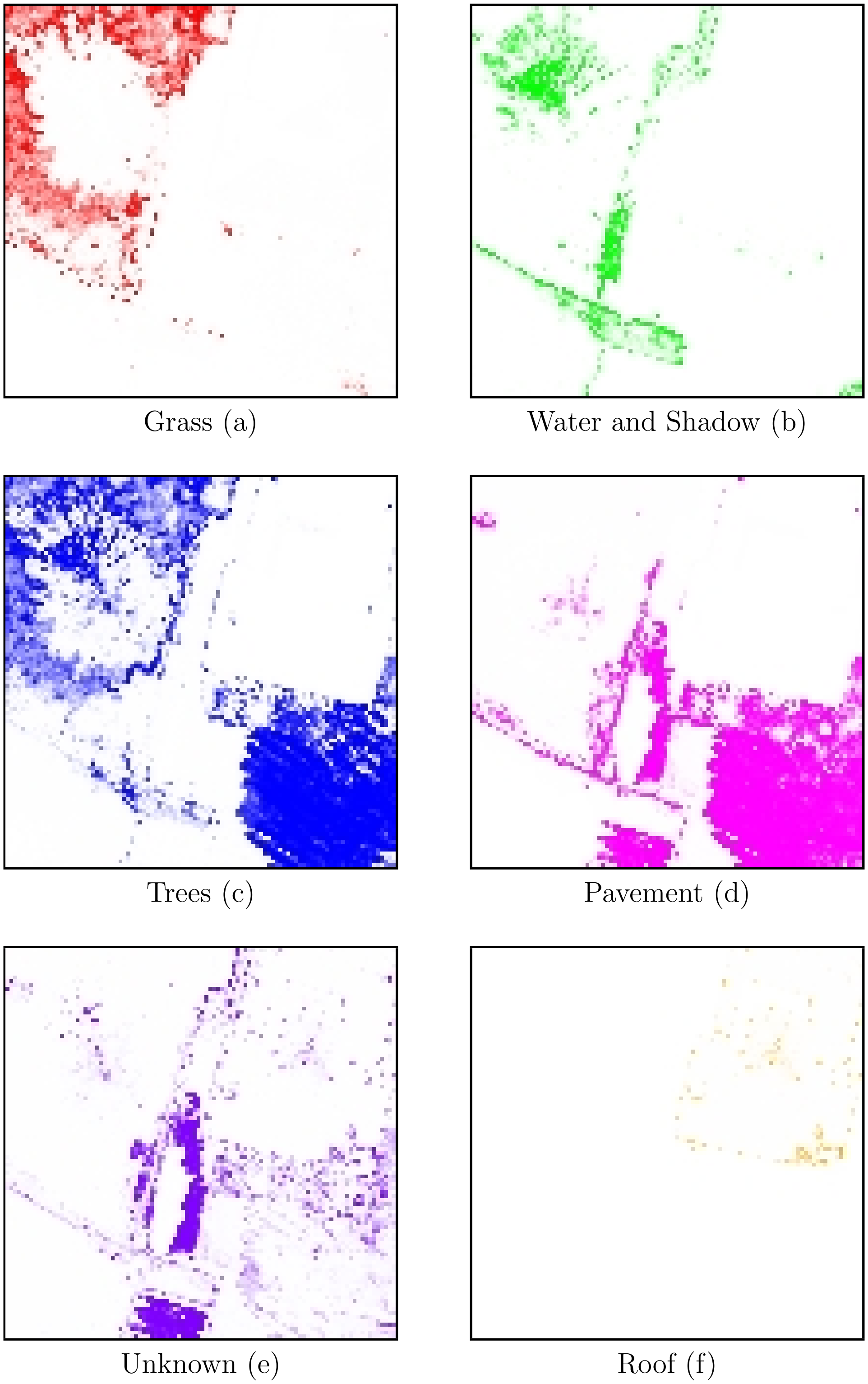
Trees (c)

Pavement (d)

Unknown (e)

Roof (f)

Figure 3.6. Probability maps for the combined imagery without the Potts model. Manually identified semantic names shown below each panel.



Grass (a)

Water and Shadow (b)

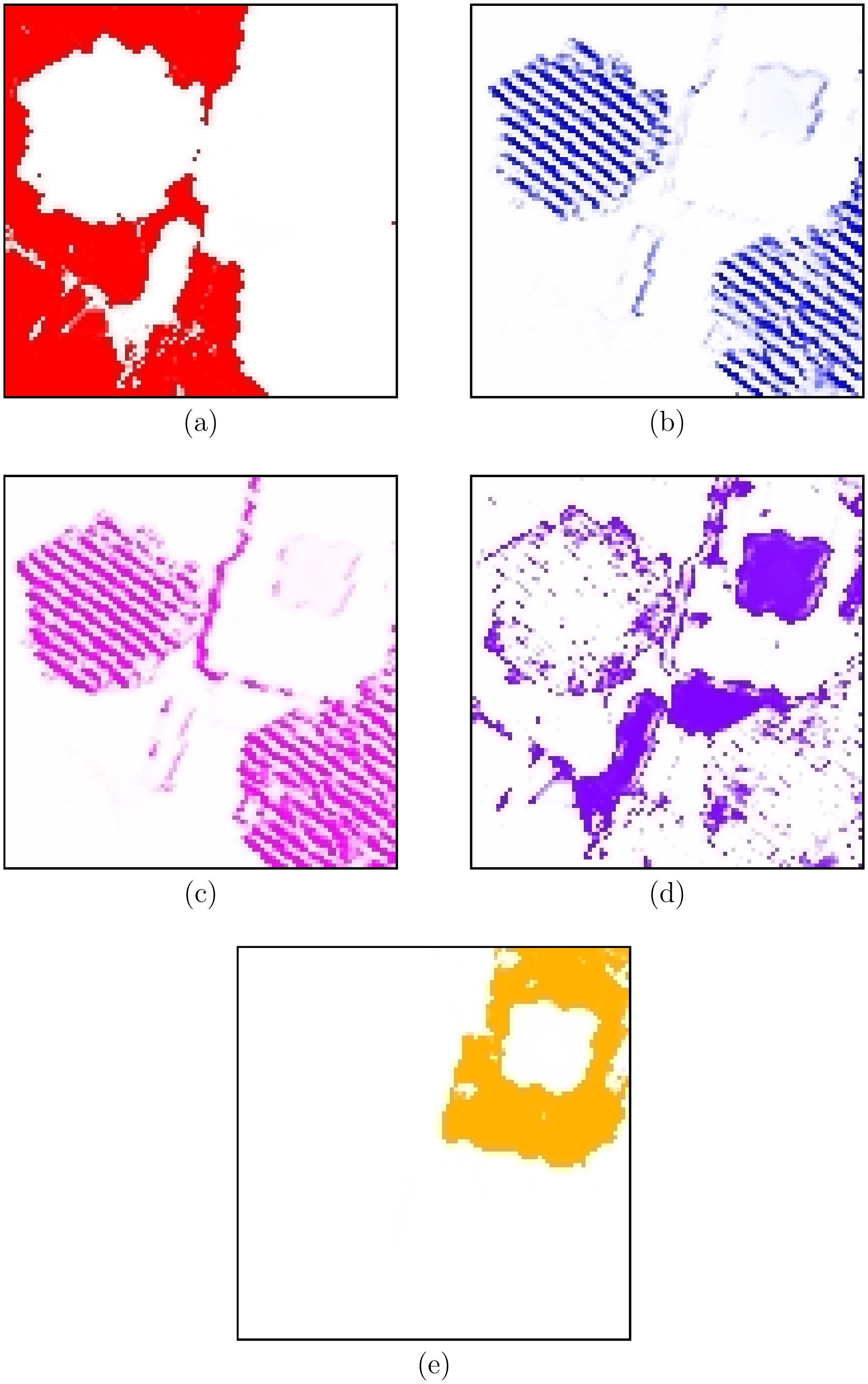
Trees (c)

Pavement (d)

Unknown (e)

Roof (f)

Figure 3.7. Probability difference maps showing the value of adding LiDAR to optical without the Potts model. Manually identified semantic names shown below each panel.



(a)

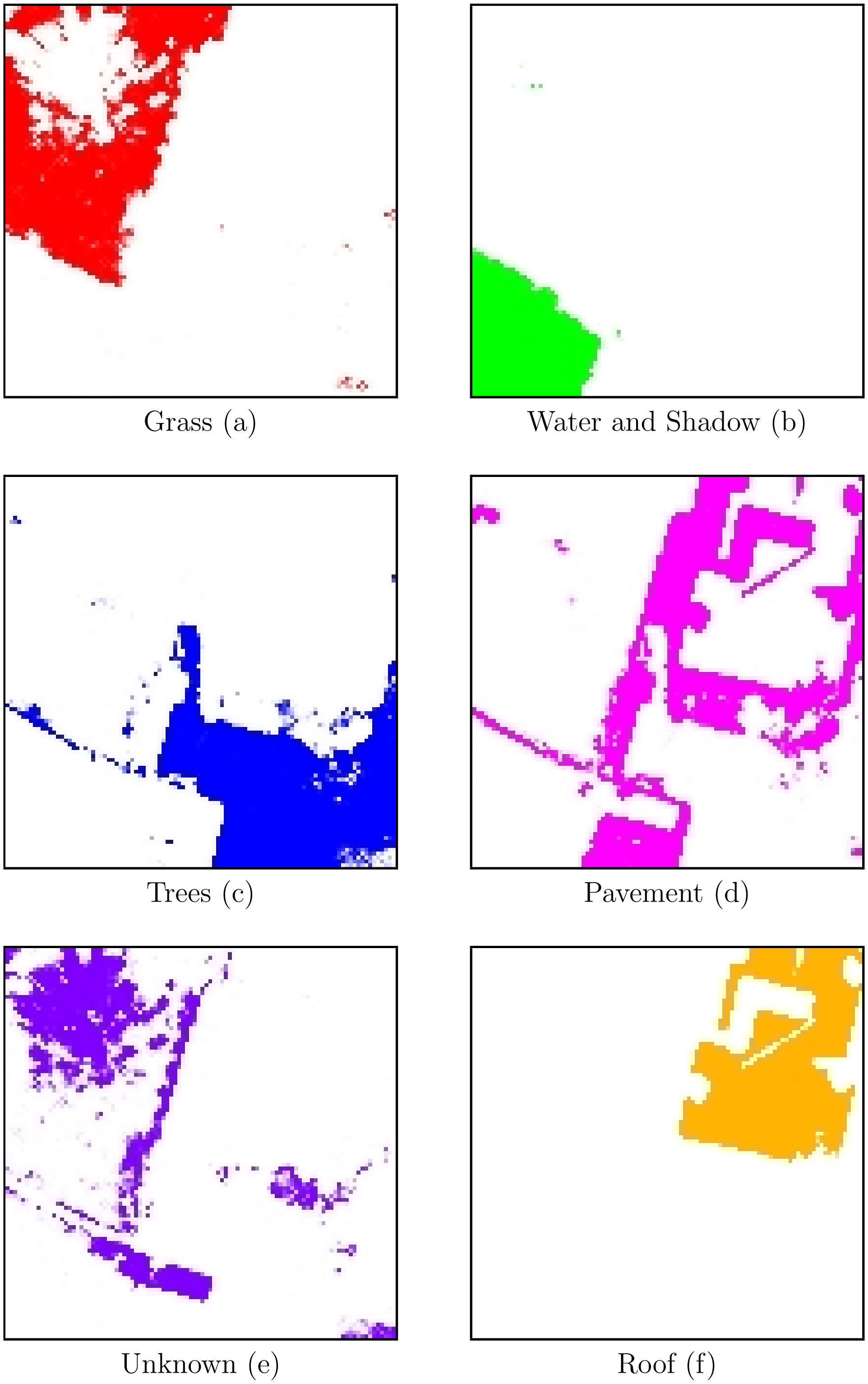
(b)

(c)

(d)

(e)

Figure 3.8. Probability maps for LiDAR imagery with the Potts model. Categories relate to height rather than semantic classes.



Grass (a)

Water and Shadow (b)

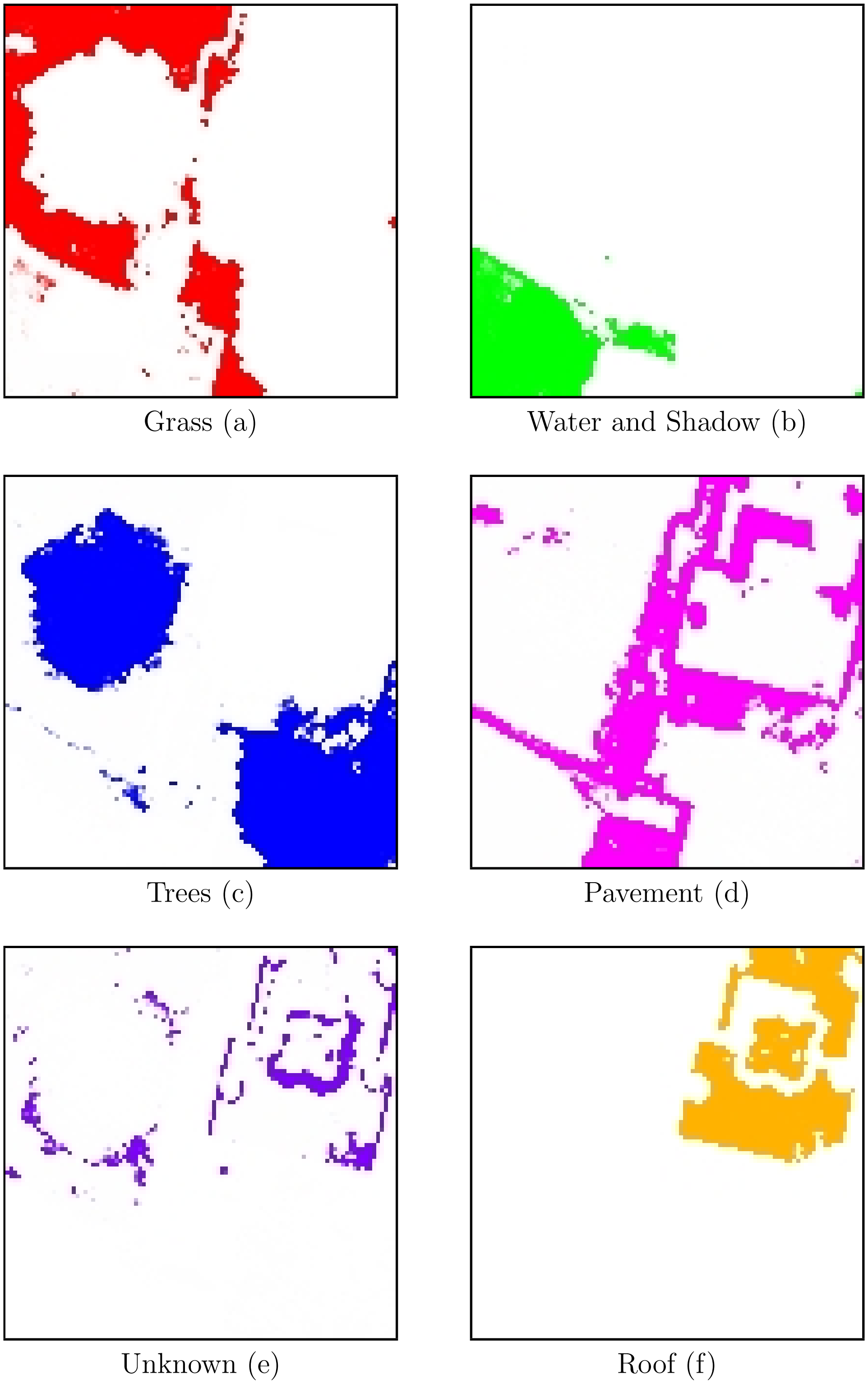
Trees (c)

Pavement (d)

Unknown (e)

Roof (f)

Figure 3.9. Probability maps for optical imagery with the Potts model. Manually identified semantic names shown below each panel.



Grass (a)

Water and Shadow (b)

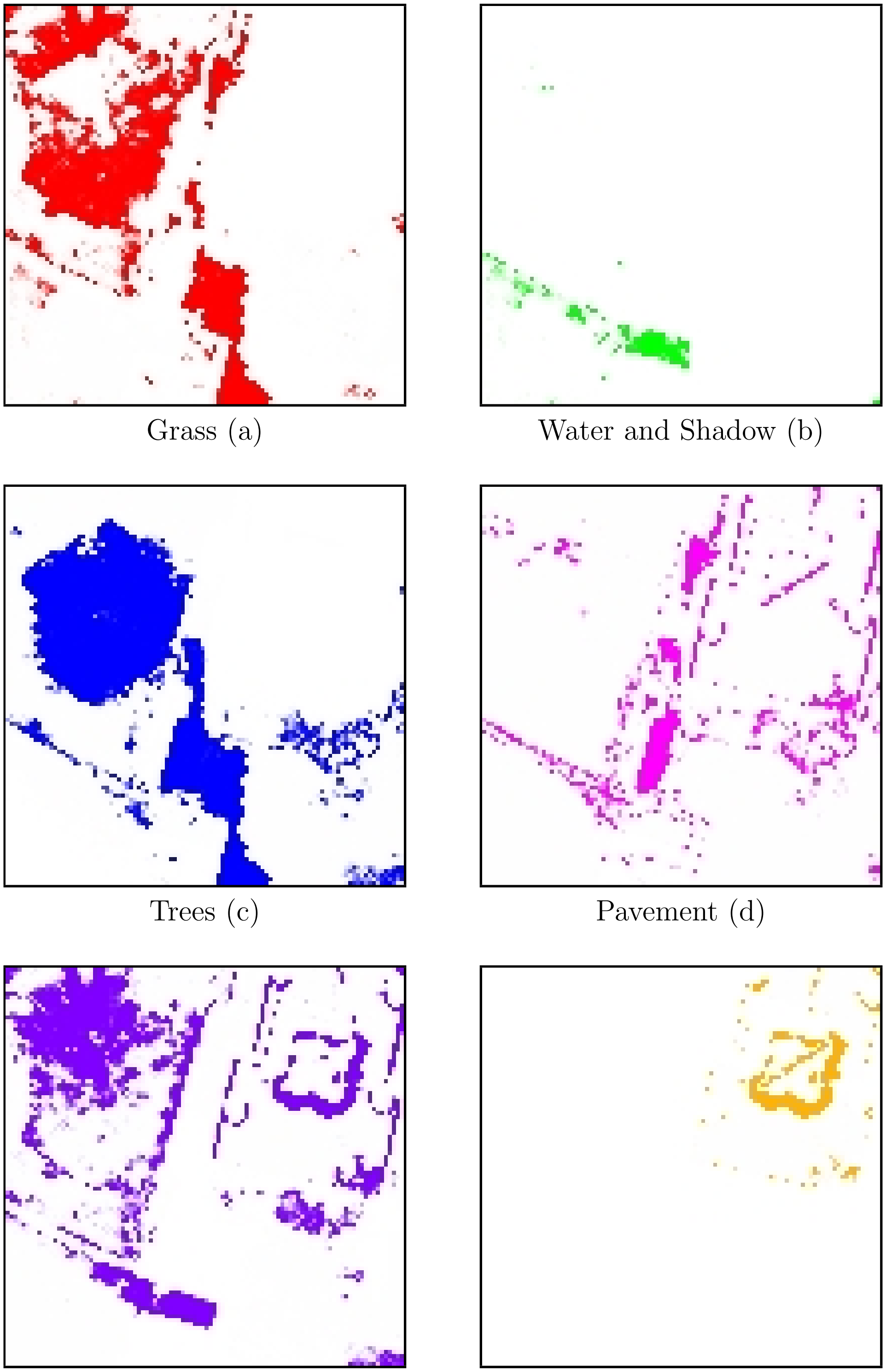
Trees (c)

Pavement (d)

Unknown (e)

Roof (f)

Figure 3.10. Probability maps for the combined imagery with the Potts model. Manually identified semantic names shown below each panel.



Grass (a)

Water and Shadow (b)

Trees (c)

Pavement (d)

Unknown (e)

Roof (f)

Figure 3.11. Probability difference maps showing the value of adding LiDAR to optical using the Potts model. Manually identified semantic names shown below each panel.

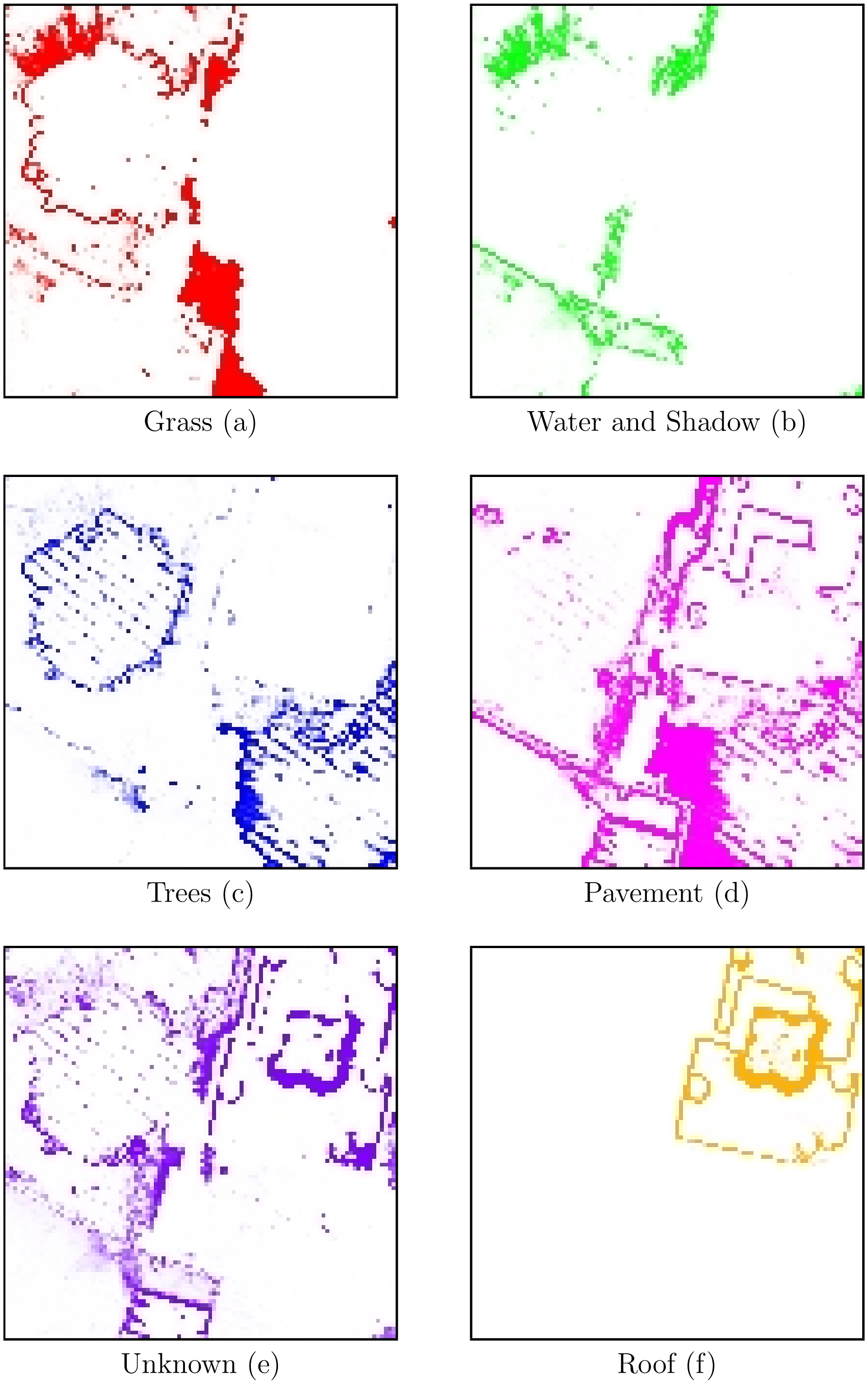


Figure 3.12. Probability difference maps comparing models with and without Potts. Manually identified semantic names shown below each panel.

appears to follow from the unexpected clustering of tall roof sections with sections of the ground.

Finally, notice that use of the Potts model has moved the boundaries of certain prominent image features, such as trees, roof, and the water/land boundary by several pixels. This raises several questions about performance evaluation. Which of the two results is better? How would we acquire sufficiently accurate ground truth results to distinguish between the two? More importantly, does the difference have a practical impact on analyst decision making? All three questions are somewhat domain and sensor dependent, yet have impact that is more than simply academic. Using the pixel neighborhoods increases computation, so we only want to apply it in the cases for which it improves results substantially.

Regardless of whether we include the Potts model or not, both analyses show a clear benefit of including the LiDAR data. The trees in particular are not well categorized by the the optical imagery alone. Moreover, both methods experienced several smaller improvements in segmentation and categorization as illustrated by the large number of colored pixels in Figures 3.7 and 3.11. These results provide an important proof of concept for the idea that a detailed analysis of uncertainty in data segmentation and categorization can provide insight into the value of individual data sources. Assessing the value of data sources, and the ability to determine which data sources have the power to improve an analysis, are becoming increasingly important in the context of the ever-increasing volume and availability of data.

3.4 Future Work

A broad variety of future work is needed in data integration under uncertainty. For example, although the basic mathematical model should generalize well to a variety of data sources and applications, a number of improvements are possible. One such improvement may include replacing the Gaussian distribution in Equation 3.2 with a Dirichlet distribution, which may better support estimation of the Multinomial parameters. Likewise, we used MCMC methods to estimate the model parameters, but a variety of different methods are possible, including expectation-maximization (Dempster et al., 1977) and variational Bayes (Jordan et al., 1999). Understanding the speed and performance tradeoffs of these methods is important for making our approach robust to many different data sources and data qualities.

Calculating the change in information due to the addition of a new data source is also a critical direction for future work. The probability difference maps shown above provide a primitive view into the impact of LiDAR on the optical imagery, but a variety of more theoretically well-founded methods are available, such as Shannon entropy (Shannon, 2001) and Earth Mover's distance (Rubner et al., 1998) as noted earlier. Success in this area could provide important new tools for conducting exploratory data analysis, and for optimizing data collection methodologies. For example, calculating the value of data and information could support stochastic optimization over data collection methodologies: sensor platform, modality, and resolution. Also important is calculating the credibility intervals for the category probabilities estimated at each data sample. This will provide important information on the uncertainty in the model parameters and the overall stability of the solution. Calculations on the value of data and information could also include the credibility intervals.

Computational efficiency is another avenue for future research. Calculating full probability distributions and the associated credibility intervals requires more computation and keeping track of many more parameter sets than methods that only determine the most likely category for each pixel or segment. Efficient implementations of the above algorithms therefore becomes critical to working with even modest sized data sets. More generally, we need to pair our methods with known techniques for scaling up to very large data sets, such as tiling (breaking oversized data, such as images, into overlapping tiles). As noted in the discussion of combining multiple data sources, we also hypothesize that we can derive great efficiency by processing each data source separately and then "convolving" the resulting probability distributions at each pixel to combine data sources. From the perspective of exploratory data analysis, in which identifying the most informative data sources is a critical step, this approach could save substantial time and computation over the traditional approach of combining data sources first, and then applying the models.

Incorporating the probabilistic land cover maps and associated uncertainties into the geospatial semantic graph representation is also a key area of future work. The pixel-level analysis described above requires several significant extensions to the existing region-based representation.

1. The high-resolution pixel representations need to be converted to a smaller number of aggregate discrete objects suitable for representation in a graph. This is the classic image segmentation problem, but made more complex by the multi-valued pixel information resulting from multiple categories such as shown in Figure 3.10. These multiple category probability maps can produce segment regions that may belong to multiple categories. One promising approach would segment the image into cells of roughly homogeneous probability mixtures. For example, a cell might correspond to an area in which class A has singularly high probability, whereas an adjacent cell might have moderate probability of both classes A and B. These could correspond to separate nodes in the graph, with suitable edges defining adjacency and other relationships.

2. The graph representation needs to adopt hierarchically structured node representations, as discussed in Chapter 2.4, to support viewing combinations of the probabilistically homogenous cells as primitive semantic objects. For example, suppose a query seeks a node of class A with minimum size constraint. No one cell alone meets the constraint, but the combination of one cell (class A) and its adjacent cell (class A or B) does have sufficient area.

3. The graph search and quality scoring algorithms need to represent and propagate the probability values from the fundamental nodes up through the hierarchy to estimate match quality. In the preceding example, the combined node would need to represent the effect of high confidence in class A over one portion of its area, and reduced confidence in class A over another portion. The result then needs to be propagated

up through the query structure and combine with other regions, for example from grassy area to football field (with requirements on shape and area) to high school (with multiple other regions, such as parking lots).

4. The graph search algorithms will need to be extended to manage the combinatorial complexity that would result from representing basic semantic objects as combinations of multiple cells with varying probability mixture. Under such an approach, many alternatives might exist for selecting which cells to include or exclude to estimate a basic semantic object. Thus, combinatorics becomes a major concern.

Finally, visualization of results requires substantial consideration. The probability maps used above, with or without isopleths, offer one possible way to visualize the modeling results. Better methods for comparing pairs or sets of category probability maps are clearly necessary, as visually comparing two separate images is very difficult. More generally, the extension of probability maps to credibility intervals is not clear. Likewise, methods for visualizing the value of information provided by added data sources and the associated changes to the credibility intervals are also needed.

Chapter 4

Conclusions

Analysis of uncertainty, both in the context of primitive data and its higher-level graph representations, provides fertile ground for improving data analysis techniques and results. The preceding discussion demonstrated how uncertainty analysis can be used to provide analysts with better information on which to base decisions and highlighted a number of directions in which the demonstrated methods can be improved and expanded. The most important lesson of the reported work is that by calculating and retaining the probability distributions and associated uncertainty intervals that underpin many modern machine learning and statistical data analysis methods, we provide far more information to data analysts than is otherwise available. Restating this point as a critique of common practice: *reporting only point estimates leaves a huge amount of information about the data on the table.* Clearly additional effort is needed to determine how best to present this new probabilistic information. Nevertheless, we have demonstrated how probability and uncertainty distributions can be used to practical advantage.

References

Besag, J. (1974). Spatial interaction and statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, 36(2):192–236.

Braverman, A., Nguyen, H., Cressie, N., Katzfuss, M., Olsen, E., Wang, R., Machalak, A., and Miller, C. (2010). Geostatistical data fusion for remote sensing applications. In *Earth Science Technology Forum*, Arlington, VA.

Brost, R. C., McLendon III, W. C., Parekh, O., Rintoul, M. D., Strip, D., and Woodbridge, D. M.-k. (2014). A computational framework for ontologically storing and analyzing very large overhead image sets. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on Analytics for Big Geospastial Data (BigSpatial)*, Dallas, TX. ACM Press. Also available as Sandia Report number SAND2014-17167C.

Canters, F. (1997). Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. *Photogrammetric Engineering and Remote Sensing*, 63(4):403–414.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *JOURNAL of the Royal Statistical Society, Series B (Statistical Methodology)*, 70(1):209–226.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Domokos, C., Kato, Z., and Francos, J. M. (2008). Parametric estimation of affine deformations of binary images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 889–892, Las Vegas, NV. IEEE Press.

Falk, M., Alston, C., McGrory, C., Clifford, S., Henron, E., Leonte, D., Moores, M., Walksh, C., Pettitt, A., and Mengerson, K. (2015). Recent bayesian approaches for spatial analysis of 2-D images with application to environmental modeling. *Environmental and Ecological Statistics*, 22.

Fonte, C., Apolinário, J., and Gonçalves, L. M. (2013). Assessing the spatial variability of classification accuracy using uncertainty information. In *Proceedings of the 16th AGILE Conference on Geographic Information Science*, Leuven, Belgium. Springer.

Gallagher, B. (2006). Matching structure and semantics: A survey on graph-based pattern matching. In *Papers from the AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*. AAAI Press.

Gonçalves, L. M., Fonte, C., Júlio, E. N., and Caetano, M. (2008). A method to incorporate uncertainty in the classification of remote sensing images. In *Proceedings of the*

8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, pages 179–185, Shanghai, China. World Academic Press.

Goshtasby, A. A. (2005). *2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications*. John Wiley and Sons, Inc., Hoboken, NJ.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–223.

Kraskov, A., Stgbauer, H., Andrzejak, R., and Grassberger, P. (2005). Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):278.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):7986.

Lelandais, B., Gardin, I., Mouchard, L., Vera, P., and Ruan, S. (2014). Dealing with uncertainty and imprecision in image segmentation using belief function theory. *International JOURNAL of Approximate Reasoning*, 55.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin. Morgan-Kaufmann.

Lin, J. (1991). Divergence measures based on shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145151.

Lizarazo, I. and Elsner, P. (2009). Improving urban land cover classification using fuzzy image segmentation. *Transactions on Computational Science*, VI, LNCS 5730:41–56.

Mallows, C. (1972). A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 43(2):508515.

Martin, A., Laanaya, H., and Arnold-Bos, A. (2006). Evaluation for uncertain image classification and segmentation. *Pattern Recognition*, 39:1987–1995.

Nguyen, H. (2009). *Spatial Statistical Data Fusion for Remote Sensing Applications*. PhD thesis, University of California at Los Angeles, Los Angeles, CA.

Nikulin, M. (2001). Hellinger distance. In Hazewinkel, M., editor, *Encyclopedia of Mathematics*. Springer.

O'Neil-Dunne, J. P. M., MacFaden, S. W., Royar, A. R., and Pelletier, Keith, C. (2013). An object-based system for LiDAR data fusion and feature extraction. *Geocarto International*, 28(3):227–242.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, San Mateo, CA.

Potts, R. (1952). Some generalized order-disorder transitions. *Proceedings of the Cambridge Philosophy Society*, 48:106–109.

Rubner, Y., Tomasi, C., and Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*, Bombay, India. IEEE Press.

Russ, J. C. (2011). *The Image Processing Handbook*. CRC Press.

Saad, A., Hamarneh, G., and Möller, T. (2010). Exploration and visualization of segmentation uncertainty using shape and appearance prior information. *IEEE Transactions on Visualization and Computer Graphics*, 16:1366–1375.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ.

Shannon, C. (2001). A mathematical theory of communication. *SIGMOBILE Mobile Computing and Communications Review*, 5(1):355.

Shi, W., Zhuang, X., Wolz, R., Simon, D., Tung, K., Wang, H., Ourselin, S., Edwards, P., Razavi, R., and Rueckert, D. (2012). A multi-image graph cut approach for cardiac image segmentation and uncertainty estimation. In *Statistical Atlases and Computational Models in the Heart (STACOM), LNCS 7085*, pages 178–187, Nice, France. Springer.

Simonson, K. M., Drescher Jr., S. M., and Tanner, F. R. (2007). A statistics-based approach to binary image registration with uncertainty analysis. *IEEE Transations on Pattern Analysis and Machine Intelligence*, 29(1):112–125.

Stracuzzi, D. J. (2015). Report from the workshop on data science and uncertainty quantification in national security missions. Technical Report (under review), Sandia National Laboratories.

Stracuzzi, D. J., Brost, R. C., Phillips, C. A., Robinson, D. G., Wilson, A. G., and Woodbridge, D. M. (2015). Computing quality scores and uncertainty for approximate pattern matching in geospatial semantic graphs. *Statistical Analysis and Data Mining*. Also available as Sandia Report number SAND2015-5820J.

Stracuzzi, D. J. and Könik, T. (2008). A statistical approach to incremental induction of first-order hierarchical knowledge bases. In Železný, F. and Lavrač, N., editors, *Proceedings of the Eighteenth International Conference on Inductive Logic Programming, LNAI 5194*, pages 279–296, Prague, Czech Republic. Springer-Verlag.

Sun, Y. and Genton, M. G. (2012). Geostatistics for large datasets. In Porcu, E., Montero, J. M., and Schlather, M., editors, *Advances and Challenges in Space-Time Modelling of Natural Events*, volume 207 of *Lecture Notes in Statistics*. Springer.

Wasserstein, L. (1969). Markov processes over denumerable products of spaces describing large systems of automata. *Problems of Information Transmission*, 5:4752.

Watson, J.-P., Strip, D. R., McLendon III, W. C., Parekh, O., Diegert, C., Martin, S., and Rintoul, M. D. (2014). Encoding and analyzing aerial imagery using geospatial semantic graphs. Sandia Report SAND2014-1405, Sandia National Laboratories.

Wilkinson, G. G. (2005). Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):433–440.

Xu, G., Liang, F., and Genton, M. G. (2015). A bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. *Statistica Sinica*, 25(1):61–79.

DISTRIBUTION:

1	MS 0899	Technical Library, 9536 (electronic copy)

v1.40

Sandia National Laboratories