Statistics and Visualizations for Assessing Class Size Uncertainty

Emmanuelle Beauxis-Aussalet

La fameuse pipe, me l'a-t-on assez reprochée ! Et pourtant, pouvez-vous la bourrer ma pipe ?
Non, n'est-ce pas, elle n'est qu'une représentation.
Donc si j'avais écrit sous mon tableau « Ceci est une pipe », j'aurais menti !

- René Magritte

The famous pipe, how people reproached me for it! And yet, can you stuff my pipe? No, it is just a representation, is it not? So had I written on my picture « This is a pipe », I would have lied!

- René Magritte

SIKS Dissertation Series No. 2019-01
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN-13 978-90-393-7084-1

Cover images:
L'interpretation des rêves, René Magritte, 1927 (front).
La clé des songes, René Magritte, 1930 (back).
These paintings, contemporary with *La trahison des images* (*The treachery of images*, 1929), discuss the limitations of all forms of representation, as these fail to convey reality itself. There is more to reality than what our senses, languages or arts may represent. Magritte aimed at preserving this complexity through surrealism. *"Le Surréalisme, c'est la connaissance immédiate du réel" ("Surrealism is the immediate knowledge of reality")*. Similarly, there is more to reality than what our datasets and artificial intelligence models may represent. These technological limitations are the matter of this dissertation. May the art of Magritte bring to the reader's attention deep underlying problems that this dissertation humbly aims at addressing. *"Le monde et son mystère ne se refait jamais, il n'est pas un modèle qu'il suffit de copier" ("The world and its mystery never remakes itself, it is not a model which copying suffices").*

Statistics and Visualizations for Assessing Class Size Uncertainty

Statistiek en Visualisaties voor het Vaststellen van Onzekerheid in Klassenfrequenties
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 28 januari 2019 des ochtends te 10.30 uur

door

Emmanuelle Morgane Aude Lucie Beauxis-Aussalet

geboren op 1 augustus 1983 te Parijs, Frankrijk

Promotor: Prof.dr. H.L. Hardman

This thesis was partly accomplished with financial support from the NWO institute Centrum Wiskunde & Informatica (CWI), the European Union Seventh Framework Programme (FP7), and Amsterdam Data Science (ADS).

Contents

Chapter 1

Introduction

Classification technologies are increasingly pervasive in our societies and impact our professional and personal lives. For instance, classification systems are used in domains such as medical diagnosis, information retrieval, fraud detection, loan default prediction, or natural language processing. Handling classification uncertainty is a crucial challenge for supporting efficient and ethical systems. For instance, providing understandable uncertainty assessments to stakeholders is necessary for conducting **responsible data science**, i.e., for controlling **accuracy** and **fairness**, and achieving **transparency**[1].

This thesis addresses uncertainty issues that pertain to estimating class sizes. We focus on the perspective of end-users with little or no expertise in machine learning, who are interested in numbers of objects per class, i.e., class sizes. Such users may analyse the patterns in class sizes, but may not seek to retrieve individual objects of particular classes. We aim at **enabling end-users of classification systems to conduct uncertainty-aware and scientifically-valid analysis of class sizes**.

Our research is motivated by a practical use case of computer vision for monitoring fish populations, implemented within the Fish4Knowledge project[2]. Monitoring animals in their natural habitats allows scientists to study population sizes and behaviors, and phenomena such as reproduction or migration. It also provides evidence on how environmental conditions and human activities impact animal populations, whether in positive or negative ways. In our era facing major environmental challenges, monitoring wild animal populations provides key information to assess the needs for protecting natural habitats.

[1] Dutch initiative for Responsible Data Science: www.responsibledatascience.org (van der Aalst et al. 2017)

[2] Website of the Fish4Knowledge project: www.fish4knowledge.eu

Deploying human observers to study animals in their natural environment involves significant costs that limit the extent of such studies. Human observers may also disturb animals and interfere with their natural behaviors, so that observations can be biased (e.g., animals may avoid areas where observers are present). In contrast, deploying cameras instead of human observers offers opportunities to reduce such costs and biases.

Computer vision systems can classify animals' species or behaviors, and the class sizes provide a means to monitor the sizes of animal populations. However, such application requires rigorous assessments of the uncertainty issues that impact the classification results. Without assessing the uncertainty, no scientific conclusions can be drawn on the animal populations. This is a challenge we aim to address in this thesis.

Hence we investigate **how to support end-users' understanding of class size uncertainty**, in the context of in-situ video monitoring of animal populations. From the specific use case within the Fish4Knowledge project, we derive generalizable methods for:

- **Assessing the uncertainty factors and the uncertainty propagation** that result in high-level errors and biases in class size estimates.
- **Visualizing classification uncertainty** when evaluating classification systems, and interpreting class size estimates.
- **Estimating the magnitude of classification errors in class size estimates**.

1.1 The Fish4Knowledge project

The Fish4Knowledge project[3] delivered computer vision tools for studying fish populations (Figure 1.1). The project used 9 fixed underwater cameras (Figures 1.2 and 1.3) to continuously monitor Taiwanese coral reef ecosystems during 3 years. It produced 87 thousand hours of video, in which 1.4 billion fish were detected. Observations were collected over continuous periods of time (e.g., observing populations over complete days, seasons and years) and with limited disturbance from the data collection devices. The resulting dataset is highly valuable for studying fish populations in their natural environment.

The project delivered computer vision software able to differentiate fish and non-fish objects in individual video frames (Figure 1.4), track individual fish across video frames, and recognize up to 23 fish species (Figures 1.5 and 1.6). Our research contributed to developing visualization tools for exploring the computer vision results and their uncertainties. Our results provided tools and methods for conducting uncertainty-aware analyses of the fish populations.

[3] Book: R. B. Fisher *et. al*, Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data. Springer (2016). Teaser: https://www.youtube.com/watch?v=AFV-FiKUFyI (Boom et al. 2012).

Figure 1.1: Example fish species monitored within the Fish4Knowledge project.

Figure 1.2: Locations of the Fish4Knowledge cameras in southern Taiwan.

The Fish4Knowledge project was funded by the European Union Seventh Framework Programme FP7 (grant 257024) and lasted 3 years from 2012 to 2015. It included research teams from Edinburgh University (United Kingdoms), Catania University (Italy), National Centre for High Performance Computing (Taiwan), Academia Sinica (Taiwan), and CWI (the Netherlands).

1.2 Interpreting computer vision results

Computer vision technologies contrast with traditional practices, such as experimental fishing or diving observations, as the information collected and the uncertainty issues are different. Computer vision is based on visual information, such as contour,

contrast, colour histograms or textures, while ecology research is based on biological characteristics, such as species, size, age or behavior. It is challenging to derive the biological information from the visual information: the high-level information needs of ecologists may not be fully addressed, or may not be addressed with the required reliability.

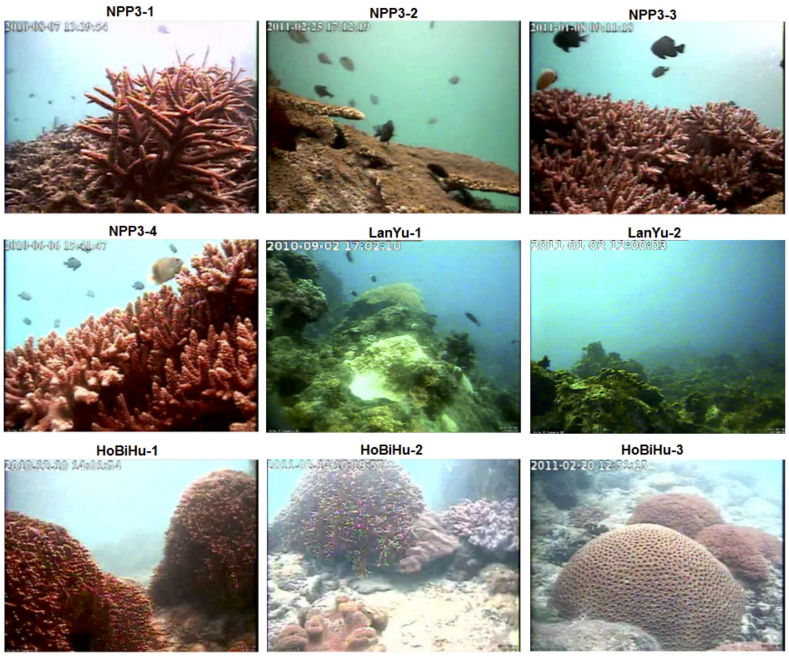

Figure 1.3: Views from the cameras deployed within the Fish4Knowledge project.

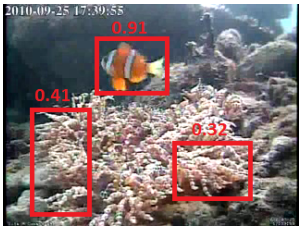

Figure 1.4: Classification of fish and non-fish objects.

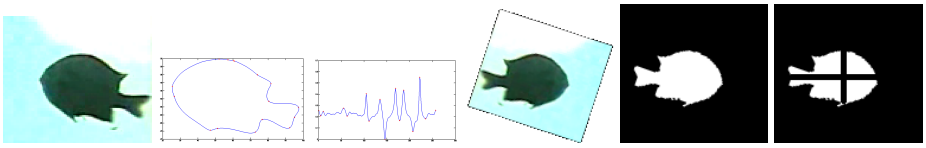

Figure 1.5: Description of the visual features (e.g., contour, orientation, body parts)

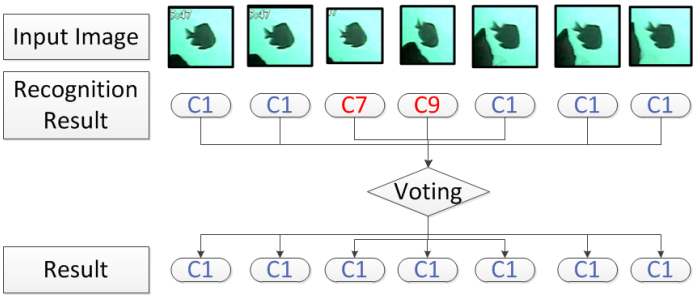

Figure 1.6: Classification of fish species (e.g., into classes C1, C7, or C9) using all images along fish trajectories.

The classification of objects appearing in the videos is inherently imperfect. **Many underlying factors can impact the magnitude of classification errors**. For example, video images of poor quality yield more errors than high-quality images (Figure 1.7). Computer vision systems typically use pipelines of classifiers, and **uncertainty can propagate from one classifier to another**. For example, if fish are not detected in all video frames, their trajectories are misidentified (Figure 1.8 and 1.9). If fish trajectories are discontinued, individual fish are counted as two separate fish and the resulting class sizes are over-estimated. If fish trajectories contain non-fish objects misclassified as fish, the classification of such fish into species has increased chances of errors. Ultimately, the classification errors impact the high-level information provided to ecologists. For example, **the population sizes can be over- or under-estimated** (e.g., if fish are not detected, if non-fish objects are classified as fish, or if fish species are misclassified).

It is crucial to communicate the uncertainties that computer vision results can carry. As scientists, ecologists are required to investigate and disclose the potential sources of uncertainty and, where possible, to estimate the resulting errors and biases. These are prerequisites for drawing scientifically valid interpretations of computer vision and classification data. Hence the perspective of ecologists is particularly relevant for researching the means to assess and communicate class size uncertainty, and to enable accountable classification systems.

1.3 Analysing class sizes

Our use case concerns **users of classification systems who study population sizes estimated as class sizes**. **The users have no technical expertise in classification technologies, yet need to assess the uncertainty issues**. They need to assess whether the class sizes are representative of the actual populations. Within the Fish4Knowledge project, for instance, ecologists use computer vision to classify fish into species. They need to draw scientific conclusions on the class sizes, yet have no expertise in the biases that classification and computer vision entail. From this particular use

case, this thesis develops generalizable methods and tools for **supporting end-user understanding of computer vision and classification uncertainty**, and **enabling uncertainty-aware and scientifically-valid analyses of class sizes**.

Analysing class sizes is a common task in domains other than ecology. For example, geologists can analyse land coverage from satellite images, e.g., by classifying image pixels into forest, sand, urban areas or other types of land. In this case, the numbers of pixels per class, i.e., the class sizes, evaluate areas of types of land. In the medical domain, when classifying the pixels of images of cancerous tissues, class sizes estimate the sizes of tumors.

Analysing class sizes is also common with technologies other than computer vision. For instance, when classifying the topics of texts, class sizes measure the frequency at which topics are discussed. Within the financial sector, when classifying borrowers' potential defaults, class sizes estimate the risks associated with loan portfolios.

Hence, our generic use case concerns **the analysis of class sizes and their uncertainties**. For this use case, the high-level uncertainty concerns **how class sizes drawn from classification systems may differ from the actual class sizes in the real world**.

1.4 Research questions

We first explore the specific topic of monitoring animal populations using computer vision, before addressing the more generic topic of assessing class size uncertainty.

Question 1: **What high-level information needs and uncertainty requirements in marine ecology research can be addressed with computer vision systems?**

As computer vision technologies are relatively new in marine ecology, we need to establish which high-level tasks and information needs can or cannot be addressed, and which types of uncertainty are acceptable. This is the topic of our first research question, addressed in *Chapter 2 - User Information Requirements*.

Figure 1.7: Example of low quality images collected within the Fish4Knowledge project. From left to right: encoding error, murky water, dirt on the lens.

Figure 1.8: Example of objects that are difficult to classify into fish or non-fish objects.

Figure 1.9: Uncertainty propagation yielding tracking error. One fish was not detected in a video frame. This fish trajectory was misinterpreted (green line). The missing fish image was replaced with one from the nearby fish.

Question 2: **What information on classification errors is required for end-users to establish informed trust in classification results?**

Providing information on classification errors may improve users' trust and acceptance of classification systems. Without sufficient information on classification errors, users' trust or mistrust of classification results may be uninformed. We need to establish the information that support user understanding of uncertainty issues, and informed decision when interpreting classification results. This is the topic of our second research question, addressed in *Chapter 3 - Establishing Informed Trust*.

Question 3: **When applying computer vision systems for population monitoring, what uncertainty factors can arise from computer vision systems, and from the environment in which systems are deployed?**
Question 4: **How uncertainty assessment methods address the combined effect of uncertainty factors?**

With the insights from our initial research questions, we can develop a comprehensive overview of the underlying factors that contribute to the high-level uncertainty when estimating population sizes. To enable transparent and accountable computer vision systems for population monitoring, we must consider how uncertainty propagates within the pipeline of classification algorithms. We must also consider the uncertainty that arises from the conditions under which the computer vision system is deployed. This is the topics of our third and fourth research question, addressed in *Chapter 4 - Uncertainty Factors and Assessment Methods*.

Question 5: **How can we estimate the magnitudes of classification errors in end-results?**

Key uncertainty factors are not fully addressed by existing assessment methods. In particular, we identify missing methods for estimating the magnitudes of classification errors that can be expected in classification results. Test sets are used to measure the rates of classification errors. Such error rates intend to represent the classification errors to expect in future applications. However, end-users are not provided with formal methods to estimate the magnitude of errors in classification results, using error rates measured with test sets. This is the topic of our fifth research question, addressed in *Chapter 5 - Estimating Classification Errors*.

Question 6: **How can visualization support non-expert users in understanding classification errors?**

It is not trivial to understand how the magnitudes of classification errors can bias class size estimates. The end-users who must assess such classification errors may have no expertise in classification. Without understanding the implications of classification errors, end-users cannot perform uncertainty-aware interpretations of class sizes. Hence we focus on the means to support non-experts' understanding of classification errors. We investigate simplified visualization designs that enable non-experts to choose classifiers, use simple tuning parameters, and understand the magnitude of errors to expect in future classification results. This is the topic of our sixth research question, addressed in *Chapter 6 - Visualization of Classification Errors*.

Question 7: **How can interactive visualization tools support the exploration of computer vision results and their multifactorial uncertainties?**

Conducting uncertainty-aware class size analyses does not only involve classification errors. Other uncertainty factors must be considered, such as those identified through our research question 3. Hence we investigates comprehensive user interfaces that provide complete information on computer vision results and their uncertainties. This is the topic of our last research question, addressed in *Chapter 7 - Visualization for Monitoring Uncertain Population Sizes*.

With these research questions, we address the needs of scientists dealing with the multiple uncertainty factors of computer vision systems for population monitor. Beyond this specific use case, our research questions investigate fundamental visualization and statistical methods for tackling classification uncertainty.

1.5 Scope

The computer vision technologies included in our scope are those developed within the Fish4Knowledge project. These technologies did not include measurements

of fish body size, or other numerical data such as speed. The Fish4Knowledge system provides classification data (i.e., categorical data) that describe the types and sizes of fish populations. Hence, our scope concerns uncertainty issues related to classification problems, such as estimating the misclassifications that can occur between each class. Uncertainty issues inherent to computer vision are considered from the perspective of their impact on classification results provided to end-users.

Our research does not concern the development or improvement of computer vision or classification technologies. We do not aim at reducing the uncertainty in computer vision or classification results. Instead we aim at enabling end-users to understand the uncertainty, to account for the uncertainty when analysing computer vision and classification data, and to draw uncertainty-aware conclusions.

Our scope does not concern uncertainty related to sampling methods, e.g., related to the number and locations of the video samples, and cameras deployed in the ecosystem. Handling such uncertainty is highly dependent on the specific studies conducted by ecologists, who have the domain knowledge to elicit the appropriate methods for handling issues with sampling the ecosystem. However, our scope includes sampling issues that are related to computer vision and classification technologies, i.e., regarding the sampling of groundtruth sets used to train and test classifiers and computer vision algorithms. As we do not aim at improving the computer vision and classification technologies, we do not investigate methods for sampling or selecting the groundtruth *training sets* (used to train classifiers and computer vision algorithms).

However, we are concerned with the groundtruth *test sets* that are used to estimate the classification errors. Test sets intend to represent the errors to expect in future applications, and are crucial for assessing classification uncertainty. Thus we consider sampling issues such as representativity, scarcity and error rate variance.

1.6 Thesis overview

Our preliminary user studies elicit the user information needs (Chapter 2) with a particular focus on information needs w.r.t. classification uncertainty (Chapter 3). From these studies, we derive the uncertainty issues of concern to end-users, and the related uncertainty assessment methods (Chapter 4). We then introduce new methods for estimating the numbers of errors in classification results, and for correcting the ensuing biases in class size estimates (Chapter 5). Finally, we investigate new visualization tools for assessing classification errors (Chapter 6) and for analysing population sizes and their uncertainties (Chapter 7). We conclude by discussing the implications of our results (Chapter 8).

Chapter 2 - User Information Requirements

We establish the scope of high-level information that can be provided by computer vision systems for the scientific study of animal populations. We study the applica-

tion domain by interviewing marine ecologists. Typical data collection techniques are compared to derive generic information needs. After interviewing computer vision experts, **we identify the information needs that can or cannot be addressed by video monitoring techniques**. Finally, the **uncertainty issues inherent to each data collection technique are discussed, and high-level requirements for uncertainty assessment are identified**.

Chapter 3 - Establishing Informed Trust

We investigate the information on uncertainty issues that support end-users in developing informed uncertainty assessments. Our second user study explores **how information about classification errors impacts users' understanding, trust and acceptance of the computer vision system**. We collect users' feedback on uncertainty factors other than classification errors, and discuss the relationships between user (mis)understanding of uncertainty, trust and acceptance of the system. Our conclusions highlight **unfulfilled information needs requiring additional uncertainty assessments, and high-level user-oriented information that uncertainty assessments must provide**.

Chapter 4 - Uncertainty Factors

We identify key uncertainty factors that must be considered for enabling scientifically valid analyses of computer vision results. We focus on in-situ video monitoring technologies such as those implemented within the Fish4Knowledge system, which provides counts of individuals per class of species, and uses fixed underwater cameras without stereoscopic vision. Our scope includes uncertainty factors beyond the computer vision system, arising from the in-situ environment in which the system is deployed (e.g., camera placement and fields of view). After specifying the typical computer vision system and deployment conditions, the uncertainty factors are elicited from interviews of marine ecologists and computer experts. We then identify the **interactions between uncertainty factors, and how uncertainties propagates to high-level information**. Finally, we identify the uncertainty assessment methods that are applicable or that are missing.

Chapter 5 - Estimating Classification Errors

We identify methods for estimating the numbers of errors in classification results, using error measurements performed with test sets. These methods can **provide unbiased estimates of class sizes** and do not primarily aim at identifying which specific items are misclassified. **Class sizes can be corrected to account for the potential False Positives and False Negatives in each class**. We review existing *bias correction* methods from statistics and epidemiology, and investigate their applicability for computer vision classifiers. We then extend the *bias correction* methods to estimating the number of errors between specific classes. We identify the unaddressed case of

disjoint test and target sets, which impacts the variance of *bias correction* and *error estimation* results. **We introduce 3 new methods:**

- **The Sample-to-Sample method estimates the variance of *bias correction* and *error estimation* results for disjoint test and target sets.**
- **The Ratio-to-TP method uses atypical error rates that have properties of interest for estimating the variance of *error estimation* results.**
- **The Maximum Determinant method uses the determinant of error rates**, encoded as a confusion matrix, **as a predictor of the variance of *error estimation* results, prior to applying the classifier to target sets.**

Chapter 6 - Visualization of Classification Errors

We introduce a **simplified design for visualizing classification errors**, i.e., the errors measured on a groundtruth test set and typically encoded in confusion matrices. We avoid the display of error rates which can be misinterpreted. Our design rationales select **raw numbers of errors as a basic yet complete metric, and simple barcharts where several visual features distinguish the actual and assigned classes**. We present a user study that compares our simplified visualization to well-established visualizations (ROC curve and confusion matrix with heatmap). **We identify the main difficulties that users encountered with the visualizations and with understanding classification errors**, depending on user's background knowledge.

Chapter 7 - Visualization Tool for Exploring Uncertain Class Sizes

We introduce a **comprehensive visualization tool that enables end-users to monitor population sizes, and to investigate uncertainties** in specific subsets of the data. We introduce an interaction design for exploring population sizes, as well as the underlying uncertainty factors (e.g., quality of video footage, classification errors of computer vision algorithms). We present a user study that investigates the interface design, and how it supports user awareness of uncertainty. **We highlight the factors that facilitated or complicated the exploration of the data and its uncertainties**, and in particular, **how users may be unaware of important uncertainty factors**. We conclude with recommendations for improving the design of such interfaces.

1.7 Thesis contributions

Our research results contribute to enabling the scientific study of animal populations based on computer vision. Our results contribute to a broader range of applications dealing with uncertain computer vision and classification data. They inform the design of comprehensive uncertainty assessment methods and tools.

Empirical contributions

➜ **Domain analysis of computer vision for video monitoring animal populations** (Chapter 2).
❑ Typical use cases are synthesized (Section 2.2), establishing key high-level information needs (Section 2.3), data collection methods (Section 2.4) and uncertainty concerns (Section 2.5).
❑ The synthesis highlights high-level information needs that can be addressed with computer vision (Table 2.5) and uncertainty issues they entail (Table 2.6).

➜ **User behaviors towards trust, acceptance, information needs, and understanding of uncertainty** (Chapter 3).
❑ Mechanisms underlying the development of informed trust and acceptance of classification systems are reported (Section 3.4).
❑ Information needs about uncertainty issues are identified (Section 3.5).

➜ **Applicability of methods for estimating classification errors and biases in class size estimates** (Chapter 5).
❑ *Error estimation* methods from statistics and epidemiology domains are successfully applied to the domain of machine learning classification (Section 5.2).
❑ Issues with existing *error estimation* methods are demonstrated (Section 5.2.4): sensitivity to stable or varying class proportions, and limited sample sizes (e.g., small datasets yield high error rate variance).
❑ Applicability to estimating the error composition in class size estimates is demonstrated, i.e., detailing the numbers of errors between all possible combination of classes (Section 5.3).

➜ **Applicability of methods for estimating the variance of classification error estimates** (Chapter 5).
❑ The variance estimation solution provided by our Sample-to-Sample method is empirically validated.
Its compatibility with *error estimation* methods, and applicability to disjoint test and target sets are demonstrated (Section 5.4).
❑ Existing methods for estimating the variance of *error estimation* results are shown to be inapplicable if test and target sets are disjoint (Section 5.4.5). Such case is common in machine learning, but did not concern the initial application domains of *error estimation* methods.

➜ **Factors impacting user understanding of classification errors and their visualization** (Chapter 6).
❑ Users' issues when interpreting classification errors using visualization supports are reported. The influence of users' prior knowledge is considered (Section 6.7).
❑ The report establishes issues with the complexity of technical concepts and terminology, and how visualization features address or aggravate them.

➜ **Factors impacting user understanding of uncertainty issues when exploring computer vision results with interactive visualization** (Chapter 7).
❑ Usability issues with the Fish4Knowledge user interface are reported (Section 7.3.1).
❑ Issues with visual features and dataset features are distinguished, e.g., choice of metrics to display, and style of display (Section 7.3.2).
❑ Recommendations are elicited for improving the interface's support of user-awareness of uncertainty (Section 7.3.3).

Theoretical contributions

➜ **Model of uncertainty factors pertaining to computer vision for monitoring animal populations** (Chapter 4).
❑ The model comprehends uncertainty factors arising from computer vision and classification systems, or from the environment in which systems are deployed (Section 4.1).
❑ Uncertainty issues are synthesized as a combination of uncertainty factors (Section 4.2).
❑ The interactions between uncertainty factors are described (Section 4.3).

➜ **Sample-to-Sample variance estimation** (Chapter 5).
❑ The distribution of rate estimators is specified for the case of disjoint datasets, i.e., for rates measured in one dataset and used as estimators of rates in disjoint datasets. Datasets are disjoint but sampled from the same population. For instance, such estimators can represent rates of classification errors in target sets using error rates measured in disjoint test sets (Section 5.4.1).

➜ **Maximum Determinant variance prediction** (Chapter 5).
❑ The hypothesis that the determinants of error rate matrices are predictors of classification errors' variance is conjectured. (Section 5.5).
❑ The type of error rate (e.g., FP Rate or Ratio-to-TP) and numbers of classes are shown to influence the predictive power (Table 5.3).
❑ Future work is required for establishing theory and validating the prediction method (Section 5.6).

Methodological contributions

➜ **Guidelines for comprehensive and user-oriented uncertainty assessments** (Chapter 2).
❑ Methodological steps are proposed for establishing the uncertainty factors and uncertainty assessment methods that address end-users' needs (Section 2.7).

➜ **Methods for estimating classification errors in end-results** (Chapter 5).
❑ Error estimation method are established for binary problems, combining the

Misclassification method, Sample-to-Sample method, and Fieller's theorem (Sections 5.4.3 and 5.4.4).

➜ **Metric for estimating classification errors in end-results, and for normalizing the visualization of classification errors** (Chapters 5 and 6).
❑ Ratio-to-TP error rates ($^{FN}/_{TP}$) support alternative methods for estimating and predicting classification errors in end-results (i.e., in target sets). Prediction methods require future work for establishing theory (Section 5.3.1).
❑ Ratio-to-TP error rates supports normalized visualization of classification errors. Such normalization is of interest for illustrating the impact of varying class proportions, and for facilitating the comparisons of False Positives and False Negatives (Section 6.4, Figure 6.8).

Artifact contributions

➜ **Visualization of classification errors for non-expert end-users** (Chapter 6).
❑ The visualization of confusion matrices is simplified with Classee barcharts, designed to facilitate non-experts' understanding of classification error (Section 6.4).
❑ The design is applicable to binary and multiclass problems.
❑ The design provides alternative to ROC and Precision/Recall curves, and includes additional information of interest to end-users (Section 6.2, Table 6.2).
❑ Open source visualization components and web interface are delivered (http://classee.project.cwi.nl).

➜ **User interface for exploring computer vision results and their uncertainties** (Chapter 7).
❑ The Fish4Knowledge user interface is delivered to ecologists and the general public. It provides access to the computer vision results collected within the Fish4Knowledge project (Section 7.2).
❑ The interface supports the exploration of fish population sizes and key uncertainty factors (Table 7.1).
❑ The interface design is applicable to multidimensional data exploration, and multifactorial uncertainty assessment. Reuse has been experimented with the SightCorp emotion recognition system (Section 7.4, Figure 7.28).

1.8 Publications

The research presented in this PhD thesis is based on the following publications:

Bastiaan J. Boom, Phoenix X. Huang, Cigdem Beyan, Concetto Spampinato, Simone Palazzo, Jiyin He, Emma Beauxis-Aussalet, Sun-In Lin, Hsiu-Mei Chou, Gayathri Nadarajan, Yun-Heh Chen-Burger, Jacco van Ossenbruggen, Daniela Giordano, Lynda Hardman, Fang-Pang Lin, Robert B. Fisher. **Long-Term Underwater Camera Surveillance for Monitoring and Analysis of Fish Populations**. Workshop on Visual

observation and Analysis of Animal and Insect Behavior (VAIB) at ACM Multimedia Conference. 2012. Mentioned in **Chapter 1**.

Emma Beauxis-Aussalet, Lynda Hardman. **User Information Needs**. Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data. Springer. 2016. Reported in **Chapter 2**.

Concetto Spampinato, Emma Beauxis-Aussalet, Simone Palazzo, Cigdem Beyan, Jacco van Ossenbruggen, Jiyin He, Bas Boom, Phoenix X. Huang. **A Rule-Based Event Detection System for Real-Life Underwater Domain**. Machine Vision and Applications 25(1). 2014. Mentioned in **Chapter 2**.

Emma Beauxis-Aussalet, Lynda Hardman: **Understanding Uncertainty Issues in the Exploration of Fish Counts**. Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data. Springer. 2016. Reported in **Chapter 3**.

Emma Beauxis-Aussalet, Elvira Arslanova, Lynda Hardman, Jacco van Ossenbruggen. **A Case Study of Trust Issues in Scientific Video Collections**. International Workshop on Multimedia Analysis for Ecological Data (MAED) at ACM Multimedia Conference. 2013. Reported in **Chapter 3**.

Emma Beauxis-Aussalet, Lynda Hardman. **Multifactorial Uncertainty Assessment for Monitoring Population Abundance using Computer Vision**. IEEE Conference on Data Science and Advanced Analytics (DSAA). 2015. Reported in **Chapter 4**.

Emma Beauxis-Aussalet, Lynda Hardman. **Extended Methods to Handle Classification Biases**. IEEE Conference on Data Science and Advanced Analytics (DSAA). 2017. Reported in **Chapter 5**.

Bastiaan J. Boom, Emma Beauxis-Aussalet, Lynda Hardman, Robert B. Fisher. **Uncer-tainty-Aware Estimation of Population Abundance using Machine Learning**. Multimedia Systems 22(6). 2016. Mentioned in **Chapter 5**.

Emma Beauxis-Aussalet, Elvira Arslanova, Lynda Hardman. **Supporting User Understanding of Classification Errors**. ACM European Conference on Cognitive Ergonomics (ECCE). 2018. Reported in **Chapter 6**.

Emma Beauxis-Aussalet, Elvira Arslanova, Lynda Hardman. **Supporting User Understanding of Classification Errors (Extended Versions)**. CWI Technical Report No. IA-1801. 2018. Reported in **Chapter 6**.

Emma Beauxis-Aussalet, Lynda Hardman. **Simplifying the Visualization of Confusion Matrix**. Belgian-Dutch Conference on Artificial Intelligence (BNAIC). 2014. Reported in **Chapter 6**.

Medha Katehara, Emma Beauxis-Aussalet, Bilal Alsallakh. **Prediction Scores as a Window into Classifier Behavior**. NIPS Symposium on Interpretable Machine Learning. 2017. Mentioned in **Chapter 6**.

Emma Beauxis-Aussalet, Elvira Arslanova, Lynda Hardman. **Supporting Non-Experts' Awareness of Uncertainty: Negative Effects of Simple Visualizations in**

Multiple Views. ACM European Conference on Cognitive Ergonomics (ECCE). 2015. Reported in **Chapter 7**.

Emma Beauxis-Aussalet, Lynda Hardman. **Multi-Purpose Exploration of Uncertain Data for the Video Monitoring of Ecosystems**. EuroGraphics Workshop on Visualization in Environmental Sciences (EnvirVis) at EuroVis Conference. 2015. Reported in **Chapter 7**.

Emma Beauxis-Aussalet, Lynda Hardman: **Appendix I: User Interface and Usage Scenario**. Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data. Springer. 2016. Reported in **Chapter 7**.

Sabine Theis, Christina Brohl, Matthias Wille, Peter Rasche, Alexander Mertens, Emma Beauxis-Aussalet, Lynda Hardman, Christopher M. Schlick: **Ergonomic Considerations for the Design and the Evaluation of Uncertain Data Visualizations**. Springer Conference HCI International. 2016. Mentioned in **Chapter 7**.

Emma Beauxis-Aussalet, Simone Palazzo, Gayathri Nadarajan, Elvira Arslanova, Concetto Spampinato, Lynda Hardman. **A Video Processing and Data Retrieval Framework for Fish Population Monitoring**. International Workshop on Multimedia Analysis for Ecological Data (MAED) at ACM Multimedia Conference. 2013. Mentioned in **Chapter 7**.

Project deliverables by the author are:

Emma Beauxis-Aussalet, Lynda Hardman, Jacco van Ossenbruggen. **D2.1 User Information Needs**. 2011. Reported in **Chapter 2**.
URL: http://groups.inf.ed.ac.uk/f4k/DELIVERABLES/Del21.pdf

Emma Beauxis-Aussalet, Lynda Hardman. **D2.2 User Scenarios and Implementation Plan**. 2012. Reported in **Chapter 7**.
URL: http://groups.inf.ed.ac.uk/f4k/DELIVERABLES/F4K_Del2-2_v3-9.pdf

Emma Beauxis-Aussalet, Jiyin He, Concetto Spampinato, Baastian J. Boom, Jacco van Ossenbruggen, Lynda Hardman. **D2.3 Component-based prototypes and evaluation criteria**. 2013. Reported in **Chapter 7**.
URL: http://groups.inf.ed.ac.uk/f4k/DELIVERABLES/F4KDel23.pdf

Emma Beauxis-Aussalet, Elvira Arslanova, Jacco van Ossenbruggen, Lynda Hardman. **D2.4 Advanced User Interface and component-based evaluation**. 2013. Reported in **Chapter 7**.
URL: http://groups.inf.ed.ac.uk/f4k/DELIVERABLES/D2.4.pdf

Emma Beauxis-Aussalet, Tiziano Perrucci, Lynda Hardman. **D2.5 UI components integrated into end-to-end system**. 2013. Reported in **Chapter 7**.
URL: http://groups.inf.ed.ac.uk/f4k/DELIVERABLES/D2.5.pdf

Emma Beauxis-Aussalet, Elvira Arslanova, Lynda Hardman. **D6.6 Public Query Interface**. 2013. Reported in **Chapter 7**.
URL: http://groups.inf.ed.ac.uk/f4k/DELIVERABLES/F4KDel66.pdf

Chapter 2

User Information Requirements

To inform the design of computer vision systems for population monitoring, we must investigate the domain of application. We must establish the high-level tasks that ecologists seek to perform, and the high-level information required to perform these tasks. Then, we can identify which high-level information can be provided by computer vision systems, and which high-level tasks can be addressed.

Our investigations of the application domain include users' concerns with uncertainty issues. We aim at developing comprehensive information requirements, concerning not only the types of information needed to perform end-users' tasks, but also the types of uncertainty that are acceptable. This chapter, which addresses our first research question: *What high-level information needs and uncertainty requirements in marine ecology research can be addressed with computer vision systems?* (Section 1.4).

To elicit the user requirements that computer vision can address, we need to account for constraints from both the technology and the application domain. Hence we interviewed both computer vision experts and domain experts (Section 2.1). From interviews with marine ecologists, we draw an overview of the domain of application, including typical use cases(Section 2.2), high-level tasks and information needs (Section 2.3), data collection techniques(Section 2.4), and uncertainty issues (Section 2.5). Supplemented with feedback from computer vision experts, our domain analysis highlights the tasks and information needs that computer vision can address, and key uncertainty issues of concern. From these findings, we discuss the applicability of computer vision systems such as the Fish4Knowledge system (Section 2.6) and elicit guidelines for developing comprehensive uncertainty assessment methods that address end-user needs (Section 2.7).

2.1 Interviews with stakeholders

We investigated the domain of application, and the potential use cases for computer vision systems, through series of interviews with marine ecologists and computer vision experts. This iterative process us allowed to develop a comprehensive understanding of user needs and technical issues. Conducting the interviews iteratively allow unforeseen information requirements and uncertainty issues to emerge. Including feedback from computer vision experts was crucial to complement the interviews of marine ecology experts. Ecologists were not acquainted with the technical constraints of computer vision, and therefore could not envision all potential limitations and uncertainty issues. Computer vision experts were able to indicate uncertainties related to specific high-level information needs, and low-level technical features of computer vision technologies. To help ecologists familiarize themselves with computer vision technologies, we used user interface and visualization prototypes that provided tangible examples of the computer vision capabilities.

Marine ecology experts were recruited from universities and research centres within research teams studying fish populations in their natural environment. Ecologists were interviewed in three studies. Our first study consisted of semi-structured interviews exploring existing practices in marine ecology research. The questionnaire (Table 2.1) was followed with additional free-form questions collecting additional insights on the working environment, existing data analysis practices, uncertainty issues, and interest in video monitoring systems. The results are reported in this chapter. Our second and third studies included visualization and interface prototypes, and are reported in Chapters 3 and 7.

We first recruited 3 senior marine ecologists who answered the first-step questionnaire during phone calls lasting 45 minutes to 1 hour[1]. The interview details are available in the Fish4Knowledge Deliverable 2.1[2].

To explore user needs in more detail, we recruited 9 additional ecologists who answered the first-step questionnaire in face-to-face interviews[3]. These interviews lasted 45 minutes to 1 hour, and were conducted under the presence of two user interface experts[4].

Computer vision experts' feedback was collected at the Fish4Knowledge project meetings, twice a year during 3 years. The general setup consisted of presenting the high-level ecologists' needs drawn from our user studies, and then discussing the means to address them and the potential uncertainty issues. Marine ecology experts were also present at most of the meetings, for a complete feedback loop mediated by the team in charge of the Fish4Knowledge user interface[5]. The group of computer

[1]These participants included 2 professors from Academica Sinica (Taiwan) and Aristotle University of Thessaloniki (Greece), and 1 senior researcher from Oxford University (UK).

[2]http://groups.inf.ed.ac.uk/f4k/DELIVERABLES/Del21.pdf

[3]These participants included 8 senior researchers and 1 master student from Wageningen University (The Netherlands).

[4]The interviewers are myself and a 9-month PdEng intern expert in user experience.

[5]The user interface team included 1 professor, 1 associate professor, 1 postdoctoral researcher, 1 PhD student (myself) and 1 PdEng intern from CWI (The Netherlands).

vision experts included 9 researchers from Catania University (Italy) and Edinburgh University (UK)[6]. The marine ecology experts attending the meetings included at least one Professor from Academia Sinica in Taiwan, with decades of experience in researching the marine ecosystem targeted by the Fish4Knowledge project.

The Fish4Knowledge project description: This project aims at realizing a video analysis tool dedicated to the study of undersea ecosystems. Fixed underwater cameras continuously record videos that are automatically analysed to detect fish species and behaviours.
1. Briefly, what are your scientific research goals and topics of interest? (if relevant, please name biological patterns, processes or models implied)
2. What information, data or measures do you need to fulfil your goals?
3. How do you collect relevant data (manual methods as well as automated)? What trust or reliability issues do you encounter?
4. What tools do you use to process and analyse those data? What issues do you encounter while using those tools?
5. What would be the 20 most important questions you would ask the Fish4Knowledge tool?

Table 2.1: Questions of the semi-structured interview of marine ecology experts (Section 2.1).

2.2 Population monitoring use cases

From the interviews with ecologists, we identify typical use cases of data collection practices for fish population monitoring, and the uncertainty issues they entail (Table 2.2). The use cases are drawn from 11 out of the 12 interviews we conducted, as 1 interview did not provide sufficient information about the data collection practices of the participant. The use cases are synthesized by grouping together ecologists who share the same high-level topics of study and data collection methods. The use cases summarize ecologists' usual practices, uncertainty issues, and potential applications of computer vision systems, as mentioned during the interviews.

Case 1 - Video at single point (1 participant). The team based in The Netherlands studies Caribbean reef fish, e.g., the distribution of specific species and their variations over time (e.g., population dynamics and migrations). They use baited stereoscopic cameras to count fish, identify their species and evaluate their size. They use vessels to collect video samples at single-point locations that cover the areas and periods of interest. They manually identify single fish, without duplicates, by analyzing only one frame per video sample. They select the frame with the most fish. The uncertainties caused by occlusions are resolved by browsing other video frames. Their existing method is satisfactory, but the manual image analysis is time-consuming. They would potentially use video analysis tools for automatically counting fish

[6]The computer vision experts included 1 professor, 1 associate professor, 2 senior researchers and 5 PhD students.

and identifying species, with the same sampling method using the most dense frame. The uncertainty issues introduced by video analysis are easily accepted because the cost reduction is important.

Case 2 - Video in transects (3 participants). The team based in The Netherlands studies North Sea deep-water corals and seabed ecosystems, e.g., the distribution of species in the various deep sea habitats, and the related trophic systems (i.e., food chain). They use cameras held by a line just above the seabed, and moved in transects (lines) within the areas of interest. A laser measures the exact distance between the camera and the seabed. It serves to calibrate the measurement of fish size. They manually identify each organism and habitat features (e.g., rocks), and measure their size. The organisms are very sparse and noticeable on the empty seabed surface, but they encounter uncertainties with respect to species identification and cryptic (hidden or camouflaged) organisms. A video browsing tool allows them to manually extract object size by using the size measured in pixels and the camera to seabed distance. The observations and measures are manually collected in spreadsheet files. Their existing method is satisfactory, but the manual image analysis is extremely time-consuming and the vessel is very expensive. They would potentially use a video analysis tool for automatically identifying objects in their video collection, or for designing cheaper data collection techniques.

Case 3 - Diving along transects (1 participant). The team based in The Netherlands focuses on commercial fisheries. They study the abundance, distribution, and trophic systems of the Philippines' coral reef fish, and their vulnerability to fishing. They collect diving observations along transects at varying depth. Video cameras are used for backup purposes and occasional refinements of the live observations. The analysis of the diving notes and videos is entirely manual. They encounter uncertainty issues with the missed detection, since many organisms occur simultaneously. They usually approximate the number of fish in dense fish groups with many overlaps. The observable species are different depending on the depth, and it requires an extensive taxonomic knowledge and sample collection to cover their diversity. The data collection technique is satisfactory but costly and time-consuming, which limits the quantity of samples. They would potentially use video analysis tools for browsing the video collection, or for designing new data collection techniques.

Case 4 - Experimental fishery (1 participant). The team based in Greece studies population dynamics, trophic systems, reproduction and physiology of pelagic fish living in the Aegan Sea. They sample and dissect fish from experimental fisheries, as commonly practiced in the marine biology domain. They collect fish at single-point locations or following a stratified sampling method. Fish dissection provides precise identification of look-alike species, and precise measurements of age, fertility and feeding habits. They encounter uncertainty regarding the replicability of fish catch. Fish catches performed under the same conditions (e.g., one after the other, releasing and re-catching fish) provide highly variable results. This issue is difficult overcome,

and may require collecting large numbers of samples. This data collection technique is costly but satisfactory. Their acceptance of our tool is low because: i) video analysis cannot supply all the data they need, ii) they need a different sampling of the areas of interest, and iii) video analysis introduces uncertainties they can avoid with their existing method.

Case 5 - Commercial fishery (2 participants). Their separate teams, based in The Netherlands, conduct similar studies of population dynamics in the North Sea. They collect fish counts from commercial fisheries, as practiced for decades in the marine biology domain. The large amount of available data supports the study of population dynamics, migration and reproduction. Commercial fisheries target only specific species, and onboard fisherman may not report the bycatches of not commercialized fish species and often misidentify unusual species. Thus uncertainty issues arise due to the uneven or biased sampling of species, areas, depths and environmental conditions. However, the large amount of collected data allows statistical methods to overcome the uncertainty issues. This data collection technique is satisfactory, but could be complemented by video analysis tools for compensating the sampling biases.

Case 6 - Diving at single points and transects (2 participants). Their separate teams, based in Taiwan and The Netherlands, conduct similar studies of coral reef ecosystems. They study population dynamics, interactions between species (trophic systems, reproduction), migration patterns, and vulnerability to environmental changes. They collect fish counts, species identification and approximate fish size from diving observations. They collect data at single-point locations or in transects. They encounter uncertainty issues regarding missed detections, multiple detection of single fish, species misidentification, and some species are likely to avoid divers, thus biasing the collected data. These issues are tackled by statistical methods (e.g., ANOVA) and by comparing data from different sources. They would potentially use video analysis tools to reduce data collection costs, and for collecting larger numbers of samples.

Case 7 - Video and commercial fishery (1 participant). The team based in The Netherlands studies population dynamics and the vulnerability of the Wadden Sea fish to fisheries. They collect data from industrial waste of commercial fisheries. This data collection technique is its early stage of development. It uses common CCTV cameras to record individuals falling out of the nets, or being discarded during industrial fish sorting. They manually count fish and identify species, while they are developing video analysis software to address this task. With their video analysis tool, they encounter uncertainty issues regarding the misidentification of species and non-fish objects. This is due to the speed at which fish pass by the camera during industrial processes.

	Data Collection Technique	Sampling Method	Uncertainty Issues	Interest in Computer Vision
Case 1	**Video Images**: baited stereoscopic camera, manual image analysis	Single-point locations	Avoid detecting the same fish multiple times. Few overlaps in fish groups.	To avoid manual image analysis.
Case 2	**Video Images**: lighted camera held close to deep sea floor, at a constant calibrated distance from seabed, and manual image analysis	Transects (i.e., along a virtual line)	Rare misidentification of species. Cryptic organisms may remain undetected.	To avoid manual image analysis. To reduce expensive use of scientific vessels.
Case 3	**Diving Observations** with handheld camera for backup purposes	Transects (at varying depths)	Species misidentification. Some species hide from divers. Overlaps in fish groups.	To analyze existing videos. To avoid diving.
Case 4	**Experimental Fishery** with fish dissection	Single-point locations or transects	Variability of fish catch albeit identical experimental conditions.	Excluded, due to unsupported measurements and uncertainty issues
Case 5	**Commercial Fishery**: data from the North-Sea fish market	Dependent on commercial fisheries	Variability of fish catch. Targets only commercial species. Misidentifies uncommon species.	To compensate the biases in the market-dependent sampling conditions
Case 6	**Diving Observations**	Single-point locations or transects	Species misidentification. Some species are hiding from divers. Overlaps in fish groups.	To avoid diving.
Case 7	**Video Images & Commercial Fishery**: onboard video monitoring of fish discarded during fish processing	Dependent on equipment available onboard	Misidentification of species and non-fish objects.	Experimented in 2013, needs improvement.

Table 2.2: Summary of 7 typical use cases of fish population monitoring for ecology research.

	Research Topic				Information Need			
	Population Dynamics	Migration	Reproduction	Trophic Systems	Fish Count	Species Recognition	Body Size	Other
Case 1	x	x			x	x	x	
Case 2	x			x	x	x	x	Other organisms
Case 3	x			x	x	x	x	
Case 4	x	x	x	x	x	x	x	Weight, Bone size, Stomach content, Chemicals
Case 5	x	x	x		x	x	x	Weight
Case 6	x		x	x	x	x	x	Behavior
Case 7	x				x	x		

Table 2.3: High-level information needs drawn from the use cases in Section 2.2 and Table 2.2.

2.3 High-level information needs

We aim at identifying widespread information needs that concern a broad range of research topics in marine ecology. Identifying the most essential user needs informs the design of computer vision systems that address a broad range of applications within marine ecology. Thus we report the information needs and research topics that are most common amongst the ecologists we interviewed. We analyze the 7 use cases introduced in Section 2.2 (Table 2.3) and examples of information seeking tasks collected from ecologists (Table 2.4). We identify 4 key research topics (population dynamics, migration, reproduction, trophic systems) and 4 key information needs (fish counts, species recognition, behavior recognition, body size).

		fish count	species recog.	behavior recog.	body size
1	How many species appear and their abundance and body size in day and night including sunrise and sunset period.	x	x		x
2	How many species appear and their abundance and body size in certain period of time (day, week, month, season or year). Species composition *[set of species and relative population sizes]* change within one period.	x	x		x
3	Give the rank of above species, i.e., list them according to their abundance or dominance. How many percent are dominant (abundant), common, occasional and rare species.	x	x		
4	Fish colour pattern change and fish behaviour in the night for diurnal fish and in daytime for nocturnal fishes.	x	x	x	
5	Fish activity within one day (24 hours).	x	x	x	
6	Feeding, predator-prey, territorial, reproduction (mating, spawning or nursing) or other social or interaction behavior of various species.	x	x	x	
7	Growth rate of certain species for a certain colony or group of observed fish.	x	x		x
8	Population size change for certain species within a single period of time.	x	x		
9	The relationship of above population size change or species composition change with environmental factors, such as turbidity, current velocity, water temperature, salinity, typhoon, surge or wave, pollution or other human impact or disturbance.	x	x		
10	Immigration or emigration rate of one group of fish inside one monitoring station or one coral head.	x	x		
11	Solitary, pairing or schooling behavior of fishes. *[these behavior have different meanings depending on species]*	x	x	x	
12	Settle down time or recruitment season *[when species stop migrating and start reproducing]*, body size and abundance for various fish.	x	x		x
13	**In certain area or geographical region, how many species could be identified or recognized easily and how many species are difficult. The most important diagnostic characteristics to distinguish some similar or sibling species** *[species which look-alike]*.		x		
14	Association *[co-occurrence]* among different fish species or fish-invertebrates.	x	x		
15	Short term, mid-term or long term fish assemblage *[co-occurrence]* fluctuation at one monitoring station or comparison between experimental and control stations in MAP. *[MPA: Marine Protected Area]*	x	x		
16	**Comparison of the different study result between using diving observation or underwater real time video monitoring techniques. Or the advantage and disadvantage of using this new technique.**	x	x	x	x
17	**The difference of using different camera lens and different angle width.**	x	x	x	x
18	**Is it possible to do the same monitoring in the evening time.**	x	x	x	x
19	**How to clean the lens and solve the biofouling problem.**				
20	**Hardware and information technique problem and the possible improvement based on current technology development and how much cost they are.**				
21	What is the average body size for species X? How many percent of fish are small, normal or big?	x	x		x
22	What is the number of fish in area X for indicative species related to pollution? *[for species which absence is likely due to pollution]*	x	x		
23	What is the distribution and number of fish for indicative species of factor X? *[for species which presence or absence is likely due to the factor of interest (e.g., water acidity)]*	x	x		
24	What is the analysis of factor X impact, using pattern of indicative data Y? *[Indicative data include fish counts and behavior observations for indicative species, i.e., species that are known to react to factor X]*	x	x	x	
25	What are the areas and periods of time of species X migrations?	x	x		
26	What are the areas and periods of time of species X SPAGS? *[SPAGS:Spawning Agregation Sites, where fish gather to reproduce]*	x	x	x	
27	What are the SPAGS periods in area Y?	x	x	x	x

Table 2.4: Information seeking tasks that ecologists would perform with the Fish4Knowledge system. *The tasks are reported using participants' own words, in the order they were mentioned, when answering question 5 in Table 2.1. The texts in [...] explain the concepts from the marine ecology domain. The tasks in bold refer to uncertainty or technical issues. The last 4 columns identify high-level information needs (discussed in Section 2.3). The tasks were collected from one participant of Case 6 (tasks 1-20) and from the participant who was not included in the use cases (tasks 21-27).*

All the use cases and information seeking tasks require information on **fish counts** and **species recognition** (e.g., fish count per species) except tasks 13, 19 and 20 which concern uncertainty issues (Table 2.4). With this information, ecologists can investigate how many fish occur in specific time periods and locations (i.e., fish *abundance*), what are their species, what is the species distribution and density over areas, what is the proportion of each species in the overall population (i.e., *species composition*), or what is the total number of species (i.e., *species richness*).

Ecologists are also interested in information on **fish body size** and **behavior recognition** (9 and 10 tasks in Table 2.4, respectively). From fish body size, ecologists derive fish age and maturity, as well as reproductive cycles (e.g., presence of offspring). From fish behavior (e.g., mating, feeding, nursing, aggression), ecologists derive fish maturity and reproductive cycles too, but also seasonal cycles and trophic systems. Few mentions of behavior recognition occurred when ecologists were asked to describe their current data collection practices, and behaviors cannot be directly observed from fishery data (*Cases 4*, *5* and *7*). Ecologists' interest in behaviors emerged when asked what would be the most important tasks they would perform with the Fish4Knowledge system (Table 2.4). Such computer vision system was deemed promising for observing behaviors without disturbance from divers.

From information on fish counts, species, behaviors and body size, ecologists can study **population dynamics**, i.e., how species distributions evolve over time, locations or environmental conditions. For instance, monitoring population dynamics can support the study of ecosystems' typology (e.g., types of habitat, distributions of animal and plant species, food chains and predator/prey relationships), the study of differences between ecosystem (e.g., before and after seasonal changes, or events such as typhoons, pollutions or construction works), or the study of species life cycles (e.g., daily routines, reproduction, migration and maturity phases). With information on fish counts, species, behaviors and body size, ecologists can also study three main phenomena influencing population dynamics: **migration**, **reproduction**, and **trophic systems** (i.e., food chains describing which species feed on which species).

Each topic of study requires specific information (Tables 2.3 and 2.5) but all require at least information on fish counts per species. For instance, population dynamics concerns the relative sizes of species populations over time periods and locations (i.e., species distributions). Migrations, reproduction and trophic systems require the recognition of fish species, as these phenomena are species-dependent (e.g., each species has specific time periods or locations for migrations, reproductions or feeding behaviors).

Additional information is of interest for studying underlying phenomenom that impact **migrations**. For example, information on fish age (estimated from body size, or otolith bone size) supports investigations of relationships between migration and reproduction. Chemicals in fish bodies or surrounding waters, or other environmental information such as temperature or pressure, support investigations of relationships between migration and environmental conditions.

The topic of **reproduction** can be studied using only fish counts and species identification, given that ecologists can rely on prior knowledge of the typical repro-

duction sites and periods. For example, changes in fish population sizes occurring at these known sites and locations can be assumed to be related to reproduction. However, information on fish body size and behavior provides more reliable evidence of reproduction cycles (e.g., time period and locations) and more information on the demographic characteristics of fish populations.

The topic of **trophic systems** is more difficult to study using only fish counts and species identification. Provided with the species composition, i.e., the distribution of fish per species, ecologists can infer the potential food chains. However, such inference must rely on prior knowledge of typical feeding behaviors of each fish species, and of other available nutrients (e.g., sea weed or plankton species). Information on fish behaviors and on stomach contents are of particular interest for providing evidence of the food chains in the ecosystems. While observing fish behaviors informs ecologists on predator-prey and foraging mechanisms, analysing stomach contents informs ecologists on actual diets resulting from these behaviors.

	Fish Count	**Species Recognition**	**Behavior Recognition**	**Body Size**
Research Topic				
Population Dynamics	mandatory	mandatory	optional	important
Migration	mandatory	mandatory	optional	optional
Reproduction	mandatory	mandatory	important	important
Trophic Systems	mandatory	mandatory	important	important
Data Collection Technique				
Experimental Fishery (*Case 4*)	+	+/++[1]	-	+
Commercial Fishery (*Cases 5*, *7*)	+	+	-	+
Diving Observation (*Cases 3*, *6*)	+	+	++	+
Manual Image Analysis (*Cases 1*, *2*, *3*, *7*)	+	+	+	-/+ [2]
Computer Vision	+	+	-/+[3]	-/+ [2]

The signs indicate whether data collection techniques: **-** *cannot supply the information,* **+** *can supply the information,* **++** *can supply the most precise information.*

[1] *Fish dissection, sometimes performed after experimental fishing, is the most accurate technique for differentiating fish species that are visually similar.*

[2] *Information supplied if stereoscopic vision or calibrated distance camera-background are available.*

[3] *The state-of-the-art does not fully address the wide scope of fish behavior variety.*

Table 2.5: Information required for the main topics of study, and ability of data collection techniques to provide this information.

2.4 Data collection techniques

From the 7 use cases of ecology research on fish populations (Section 2.2), we identify 4 well-established data collection techniques: experimental fishery (i.e., sampling fish stock), commercial fishery data, diving observations, and manual image analysis (Section 2.4.1, Table 2.5). To provide reliable information, data collection techniques must be applied with appropriate sampling methods. We outline sampling methods that are usually applied by ecologists (Section 2.4.2) and discuss sampling strategies for video data collection (Section 2.4.3). Finally, we outline ecologists' rationales for selecting appropriate data collection and sampling methods (Section 2.4.4).

2.4.1 Well-established data collection methods

Experimental fishery - Scientific vessels are used to catch fish at specific sampling locations and time periods, with calibrated nets or fish traps (*Case 4*). Ecologists can then perform measurements (e.g., from fish dissection) that include information unavailable with other data collection techniques, such as fish weight, bone size (e.g., *otolith* precisely indicating fish age), stomach content (e.g., to study trophic systems), traces of chemicals (e.g., from pollution), or the presence of fish eggs (e.g., to study reproduction cycles).

Commercial fishery - Data can be collected onboard commercial vessels, by ecologists (*Case 7*) or by non-scientific personnels of fishery companies (*Case 5*). The latter involves trust issues and potential biases due to the person in charge of collecting the data, e.g., lack of expertise with rare species, inconsistent practices between observers (Kraan et al. 2013). Commercial fishery data have the advantage of offering large coverage of marine areas, but at the disadvantage of targeting only commercial species.

Diving observation - Divers can collect information on fish counts and species recognition, and can observe a variety of fish behaviors (*Cases 3* and *6*). Data can be collected by individual divers, or in teams who compare their observations to limit human biases. Observations are collected within fixed areas (e.g., delimited with frames or ropes) or along transects (i.e., predetermined path on the sea floor covering a representative part of the ecosystem). Cryptic and benthic species (camouflaged or living on the seabed) are better sampled as they are unlikely to be caught in fishing nets. However, diving observations carries uncertainty as human observers disturb natural fish behaviors and can make mistakes, e.g., depending on their diving experience, or difficulties with the fish species or ecosystems (e.g., fast or small fish, overlaps in fish groups, fish fleeing divers, inaccessible locations). Such human biases are difficult to quantify. To address them, ecologists collect data repeatedly and use well-specified consistent protocols.

Manual image analysis - Images are widely used as a means of observation. Cameras can be used at fixed or moving locations, with or without baits attracting fish (*Case 1*). They can be oriented toward the open sea, or toward the sea floor for observing benthic ecosystems. For the latter, calibrating a fixed distance between cameras and sea floor allows the measurement of fish body size (*Cases 2* and *7*). Stereoscopic vision (i.e., the use of pairs of cameras) is another technique for estimating fish body size. Body size is derived by classifying image pixels as inside or outside a fish contour (a classification task called segmentation). Divers also use handheld cameras, at fixed locations or moved along transects (*Case 3*). Otherwise, cameras can be dragged by boats or embarked on remote controlled vehicles (e.g., BRUV, Baited Remote Underwater Video systems). Image analysis is mainly performed manually, as automatic image analysis with computer vision are not supported with well-established methods for handling uncertainty and technological issues. However, computer vision has raised interest as a promising cost-effective technique (Harvey

et al. 2001, Cappo et al. 2004, Hetrick et al. 2004, Langlois et al. 2006, Lowry et al. 2012, Shafait et al. 2016).

2.4.2 Sampling methods

Sampling error is a crucial source of uncertainty. Only subsets of ecosystems are actually observed, and conclusions drawn on overall ecosystems based on limited sets of samples are inherently uncertain. In the case of computer vision system, collecting video samples caries specific uncertainty and may require specific sampling methods. To inform the design of sampling methods applicable to computer vision systems, we discuss the methods usually applied by the ecologists we interviewed.

Sampling methods are designed to target specific conditions: ecosystems, habitats, environmental conditions, time periods, locations, species or behaviors of interest. The choice of sampling methods depends on the topic of study and the scientific requirements of the research. For instance, to study migration it can be necessary to collect samples over large areas and time periods (e.g., multiple years).

Samples are collected within subsets of the locations and time periods of interests. In the marine ecology domain, the observable populations can greatly vary depending on the time periods and locations. To remain representative of the ecosystem of interest, the sampling methods must account for natural cycles and habitat topologies. For instance, the sampled time periods need to account for the hours of the day (e.g., some species appear in morning, evening or at night, for feeding) and the seasons of the year (e.g., some species migrate or reproduce in spring). The sampled locations can be fixed points (i.e., *single-point locations* in Table 2.2) or predetermined path covering a representative part of the ecosystem (i.e., *transects* in Table 2.2). The sampled locations must represent ecosystems' components, e.g., the types of habitats and their proportional land coverage. Samples are often collected in each part of the ecosystems, proportionally to their geographical coverage, and aggregated using the *stratified sampling* method (Cochran 2007).

Sampling methods provide repeated measurements to account for their variance. Estimating *sample variance* (i.e., the variance between measurements collected in each sample) contributes to the interpretation of the patterns observed in the collected data. Well-founded statistical methods (e.g., ANOVA) account for sample variance to compute the probability that the patterns observed in the data occurred by chance, and may not be representative of the actual fish populations. For example, population sizes can differ between two time periods or two species, but the difference may not be significant due to high sample variance. Such statistical methods are essential for ecology research, as they support the scientific validity of conclusions drawn on fish populations. Statistical methods can also be applied to estimate the overall population sizes in the overall ecosystem, and the variance of such estimate. However, the relative trends in fish populations often provide sufficient information for assessing population dynamics, without needing to estimate overall population sizes for specific areas.

2.4.3 Impact of video technologies on sampling methods

Special attention must be paid to the spatio-temporal coverage of video samples. The spatial coverage depends on the placement and orientation of cameras, on the type lens, and on the image resolution. The placement and orientation of cameras target specific habitats, and impact the species that are likely to be observed. The type of lens and the image resolution impact the depth and width of fields of view. These modify the areas and volumes within which fish populations can be observed, as well as the quality of observations (e.g., small fish in the background can be unrecognizable).

Estimating the spatial coverage of video samples is essential to the design of sampling methods, and to the analysis of the collected data (e.g., to study fish density). But estimating the spatial coverage of a camera is a difficult task. For instance, it requires controlling the distance within which information collection is possible, or is reliable enough (e.g., for detecting small fish). The information quality depends not only on the depth of field of view (i.e., on camera lens) but also on other environmental factors (e.g., lighting, water turbidity) and on the capabilities of the computer vision software (e.g., how the software performs with low image quality). Finally, when baits are used, estimating the area covered by cameras is more subtle. The strength and direction of currents modify the areas in which animals can sense the bait, and thus the spatial coverage of video samples (Taylor et al. 2013).

Regarding the temporal coverage of video samples, the use of fixed cameras that continuously monitor fish populations is an important paradigm shift. It contrasts with common data collection techniques that perform measurements during limited time periods. Their temporal coverage concerns a small set of preselected time periods, and the measurements performed within a time period are intended to represent of all the species living in the environment. With video monitoring systems such as Fish4Knowledge, the temporal coverage is very large, covering all time periods when there is sunlight. Ecologists do not need to extrapolate the fish populations that would occur in time periods for which no sample is available. Instead, the can assume that the fish populations occur in the videos samples at their natural frequency.

2.4.4 Choice of data collection and sampling method

Each data collection technique has its own advantages and disadvantages, and no single method fits all types of ecology research. The most important information needs are addressed by a choice of data collection techniques, as summarized in Table 2.5. The requirements for selecting a data collection technique comprise constraints on the types of ecosystem to access, the time periods for performing the study, the human and material resources available, the funding for acquiring and maintaining equipments, the information that needs to be collected, the measurements' potential errors and biases, and on the uncertainties that are acceptable. Uncertainty issues are crucial for choosing a data collection technique. Given alternative methods that can collect the information of interest, analysing uncertainty issues allow stakehold-

ers to understand the tradeoffs of each data collection technique. For example, a method may be faster or cheaper but entails unacceptable uncertainty. In other cases, a method may limit uncertainty but entails additional costs that are not worthwhile compared to alternatives.

	Experimental Fishery	**Commercial Fishery**	**Diving Observation**	**Manual Image Analysis**	**Computer Vision**
Benthic species	-[1]	-[1]	=	=	=
Sedentary species	-[1]	-[1]	=	=/+[2]	=/+[2]
Schooling species	=	=	-/+	-/+	-/+[2]
Small fish	-/=[3]	-/=[3]	-/=[4]	-/=[4]	-/=[4]
Shy species	-	-	-/=[5]	-/=[6]	-/=[6]
Cryptic species	-	-	=	-	-
Look-alike species	=	=	-/+	-/+	-/+
Rare species	=	-	=	=	-/=[7]
Herbivorous or carnivorous species	-/=[8]	=	=	-/=[8]	-/=[8]

The signs indicate whether parts of ecosystems are likely to be **+** *over-represented,* **=** *neither under- nor over-represented,* **-** *under-represented.*

[1] *Considering that the destructive use of trawl nets is not an option.*

[2] *Species living in coral heads often swim in and out of the camera field of view, which may yield over-estimated fish counts.*

[3] *Large granularity of nets' and fish traps' mesh can let small fish slip through.*

[4] *Small fish may not be visually detectable from a large distance.*

[5] *Cloaking procedures can allow the observation of shy fish.*

[6] *With handheld cameras, some species flee from divers.*

[7] *The recognition of all rare species may not be possible due to lack of ground-truth images.*

[8] *Baits, if used, can attract either herbivorous or carnivorous species.*

Table 2.6: Main biases with species that are potentially under- or over-estimated by data collection techniques.

2.5 Biases of data collection techniques

All data collection techniques carry uncertainty issues and can yield errors and biases in the collected data, e.g., some species are potentially over- and under-represented. For example, cryptic species camouflaged amongst corals are typically under-represented because they are more difficult to detect. Data collection techniques are thus always *selective*, i.e., specific parts of ecosystems and specific species can entail a different magnitude of errors than the the rest of the data, while other parts are measured with lower and consistent levels of errors.

From comparative studies of data collection techniques (Trevor et al. 2000, Harvey et al. 2001, Cappo et al. 2004, Lowry et al. 2012)) and from our interviews with ecologists, we identified nine types of fish species that are particularly susceptible to biases depending on the data collection technique. Table 2.6 summarizes the potential biases entailed by the common data collection techniques discussed in Section 2.4.

Benthic species - Organisms living on the seafloor are under-estimated in experimental or commercial fishery data. Fish nets are usually cast in the open sea (i.e., *pelagic zone*) where the species usually living in the seafloor (i.e., the *benthic zone*) are rarely found. Trawl nets dragging the seafloor can collect samples of benthic species, but this fishing technique dramatically destroys benthic ecosystems and thus is usually excluded for scientific purposes.

Sedentary species - Sedentary species living in the same rocks or coral heads, rather than circulating across larger areas, are less likely to swim in the open sea and thus to be sampled through fishery. Computer vision potentially over-estimates sedentary species because they are likely to repeatedly swim in and out of the camera field of view. Hence single individuals may be repeatedly counted. For instance, with the Fish4Knowledge system, we observed over-estimation of the sedentary species *Dascyllus reticulatus*.

Schooling species - Species living in groups can be under- or over-estimated through diving observation, manual image analysis and computer vision. Fish in a *school* (i.e., school) occlude each other, and individual fish are likely to swim in and out of cameras' field of view. With computer vision, the number of fish in a school can be either under-estimated due to occlusion, or over-estimated due to repeated occurrences of the same individuals. With diving observations and manual image analysis, humans need to interpret the overall size of the school and can subjectively over- or under-estimate the number of fish.

Note: To overcome biases with sedentary and schooling species, the ecologists from the *Case 1* of our first user study (Section 2.2) count the fish appearing in only one frame of the video footage. However, this method is likely to further under-estimate rare species, since the chance they appear in one single frame is lower than the chance they appear in the complete set of frames. Further, this method disables the analysis of visual features over several frames (e.g., fish trajectories) which can be necessary for recognizing fish behavior, and for identifying species for which swimming behavior is more discriminative than visual appearance.

Small fish - Detecting small species or offspring is difficult for all data collection techniques in Table 2.5. Small fish are difficult to detect and recognize if they are too far away from divers or cameras (e.g., depending on visual acuity and fish body sizes). In the case of diving observations, manual image analysis and computer vision, this type of bias is limited if observations are performed within small depths of field of view. With large depths of field of view (e.g., observing the open sea), ecologists need to consider that small fish are sampled only in a limited range around cameras or divers.

Shy species - Some species flee boats and divers as they detect their sounds, movements (especially that of bubbles from divers), and sometimes their chemicals (sensing underwater chemicals is comparable to sensing smells). Ecologists overcome with cloaking procedures, such as using no-bubble diving equipment (e.g., rebreather) and

allowing time for shy species to come back after divers settled in. Cameras are non-intrusive and are well-suited for observing shy species, unless divers or boats are too nearby.

Cryptic species - Cryptic species (e.g., camouflaged) are difficult to detect for both computer vision software and human observers. Cryptic species are very likely to be under-estimated, and ecologists need to apply specific methods for studying them. For instance, divers carefully scrutinize sea floors or coral heads, or use of toxicants forcing the fish to leave their camouflaged position. Data collection based on imagery is not suitable for their study. Cryptic species are often benthic species, and are thus likely to be under-estimated by commercial and experimental fisheries.

Look-alike species - Species that look-alike are difficult to detect for both computer vision software and human observers. Ecologists may rely on specific expertise to differentiate look-alike species. For instance, the species behaviors or body sizes may differ.

Rare species - Ecologists are trained to target and recognize rare species, so they can collect unbiased measurements from experimental fisheries, diving observations and manual image analysis. Commercial fishery and computer vision potentially under-estimate rare species. Computer vision software may not recognize species for which there are insufficient image samples to train the recognition algorithm. Uncommon species may not be recognized and recorded in commercial fishery data. But these uncommon species may be frequent enough for collecting sufficient image samples to train computer vision algorithms to recognize them.

Herbivorous or carnivorous species - Baits attract only the species that feed on the materials used as baits. Thus specific types of bait attract specific species, which may be over-estimated while other species are under-estimated. Baits may, however, be of particular interest for sampling species that would otherwise remain largely unobserved (e.g., rare, shy or crptic species), or for limiting the duration, thus the costs, of data collection.

2.6 Implications for the Fish4Knowledge system

Computer vision systems can address essential user information needs with two basic functionalities: detecting fish in video images, and classifying their species. Such information supports the study of four key topics in ecology research: population dynamics, migration, reproduction and trophic systems (Section 2.3). The Fish4Knowledge system was able to provide information on fish counts and species. Other information needs (e.g., behavior recognition, body size) could not be addressed due to technical limitations (e.g., no stereoscopic vision, or ground-truth collection issues) and were excluded from the scope of our research. The lack of information on fish behavior particularly impacts the study of trophic systems. Classification software can be developed to recognize fish behaviors (Spampinato et al.

2014). However, it is challenging to differentiate the large variety of fish behaviors, and collect sufficient groundtruth data for each behavior of interest.

Computer vision systems entail uncertainty issues due software components (e.g., errors from the classification software), hardware components (e.g., camera settings), and ecosystems in which computer vision systems are deployed (e.g., light conditions, visibility). End-users require that these uncertainty issues are assessed (e.g., information seeking tasks in bold in Table 2.4, Section 2.2). Furthermore, biases can arise due the characteristics of fish species (Section 2.5). The Fish4Knowledge system uses cameras without bait, at fixed positions, not held by divers, and that can be positioned to observe benthic zones and coral heads. These settings can limit potential biases with benthic, sedentary, shy, herbivorous and carnivorous species. Yet, biases are still at stake with sedentary, schooling, cryptic, look-alike and rare species, as well as small fish.

2.7 Requirements for accountable classification systems

Our study of the marine ecology domain provide insights on the uncertainty issues pertaining to computer vision systems for monitoring animal populations. Further investigations are required to elicit comprehensive scopes of uncertainty factors (e.g., depending on specific system features and application conditions) and to identify the high-level impacts of uncertainty. Such uncertainty assessments must eventually provide end-users with practical information on the uncertainty that pertains to the specific datasets they are using. This section draws guidelines for conducting such uncertainty assessment that specifically address end-users needs (e.g., rather than the needs of technology experts who seek to improve computer vision systems).

The Fish4Knowledge computer vision system delivers classification data where class sizes represents population sizes, e.g., from specific species or behaviors. Hence we focus on classification systems, beyond the domains of computer vision and ecology. We thus draw high-level user requirements for supporting uncertainty-aware analysis of class sizes. We do not intend to provide fully exhaustive requirements, however, we aim at providing essential guidelines for enabling accountable classification systems for monitoring class sizes.

2.7.1 Identify the application conditions

Uncertainty arises from the interactions between the classification system and its application conditions, e.g., how the system is deployed and in which ecosystem. To identify the uncertainty issues pertaining to specific applications, it is necessary to first specify the internal characteristics of the classification system, and the external characteristics of the environment in which the system is deployed.

Requirement 1-a - Specify the components underlying classification system: *The pipeline of interoperating components within the classification system must be specified.*

For describing the pipeline of components (e.g., classification software), the specifications must include i) the execution sequence of the components; ii) the data inputted and outputted by each component, describing how uncertainty can propagate along the pipeline of components.

Requirement 1-b - Specify the application conditions: *The external environment in which the system is deployed, and the material characteristics of the system implementation must be specified.*

In the case of computer vision systems, the specifications must include i) the cameras and their technical features (e.g., lens, frame rate, resolution); ii) the real-world environment observed through the cameras, including the kind of events that are expected to occur, whether desirable (e.g., fish populations of interest) or undesirable (e.g., dirt on the lens or occlusions by floating object). The Human-Computer system that enables end-users to process the classification data must also be specified, as supporting uncertainty-aware data analyses cannot be achieved if end-users have no access to complete and understandable information on uncertainty. The specifications must include i) the end-users prior knowledge and skills, their goals, their high-level information needs, and the data analysis tasks they intend to perform; ii) the working environment of end-users, the interface used to access the classification data, and other information sources used to perform the high-level tasks, including human collaborators or other information systems.

These requirements are consistent with prior work considering that uncertainty arises from three information processing steps (Pang et al. 1997): *data collection* (e.g., the conditions in which systems are deployed to collect data, requirement 1-b), *data processing* (e.g., the pipeline of software components, requirement 1-a), and *data interpretation* (e.g., the Human-Computer system, requirement 1-b).

We address these requirements in Chapter 4 for population monitoring systems such as the Fish4Knowledge system. The application conditions at the *data interpretation* level, i.e., regarding the Human-Computer system, are investigated in Chapters 3, 6 and 7.

2.7.2 Identify the uncertainty factors

A large variety of issues can arise depending on systems' technical features and application conditions. Uncertainty can arise from low-level factors (e.g., image features and quality) but needs to be described in terms of the high-level impacts on the data analysed by end-users. To identify the relevant lower-level factors of uncertainty, it is necessary to relate the low-level factors to the higher-level impacts on user tasks.

Requirement 2-a - Identify the high-level impacts: *The misinterpretations that can occur if uncertainty is not considered when interpreting classification data must be identified.*

In the case of population monitoring, the misinterpretations include, e.g., considering that the class sizes are representative of the true population sizes, while the class sizes can under- or over-estimate the actual populations (e.g., due to random or systematic classification errors); or considering that the trends observed in class sizes (e.g., over time periods or locations) are representative of the actual trends in population sizes, while the observed and actual trends can differ (e.g., due to biases arising from varying image quality).

Requirement 2-b - Identify the uncertainty factors: *The chain of phenomenon that can yield discrepancies between facts and information provided to end-users must be identified.*

The factors of uncertainty arise from technical issues within the classification system, and from the environment in which the system is deployed. Thus addressing requirement 2-b must rely on the specifications provided by requirements 1-a and 1-b. In the case of fish population monitoring, the uncertainty factors from the classification system include, e.g., errors in detecting fish and non-fish objects, or errors in recognizing species and behaviour. The uncertainty factors from the in-situ deployment conditions include, e.g., lens biofouling, water turbidity or low light, which increase the chances of errors from the classification system. The uncertainty propagation, i.e., the uncertainty accumulated though interactions between uncertainty factors, must also be specified. For example, fish detection errors are propagated to species recognition algorithms, and increase the chances that species are misclassified. Finally, uncertainty factors also arise from the way information is provided to end-users, e.g., if key information are difficult to access or understand.

We address these requirements in Chapter 4 for population monitoring systems such as the Fish4Knowledge system. The uncertainty factors at the end-user level, i.e., when end-users interpret the classification data, are investigated in Chapters 3, 6 and 7.

2.7.3 Identify the uncertainty measurements

Given the scope of uncertainty factors (Requirement 2-a), end-users need to estimate the resulting uncertainty in high-level information (Requirement 2-b). The characteristics of uncertainty factors can be measured for each factor separately. However, end-users are particularly concerned with measuring their combined impact resulting from uncertainty propagation. To deal with the high-level impacts of multiple uncertainty factors, it is necessary to identify i) how each uncertainty factor can impact other uncertainty factors; ii) the metrics and methods that can specify each component's uncertainty; and iii) the methods that can estimate the combined uncertainty resulting from the interactions between uncertainty factors.

Requirement 3-a - Identify factor-specific measurements: *The characteristics of each uncertainty factor's impact on high-level information or on other uncertainty factors, and the means to measure these characteristics, must be identified.*

Factor-specific measurements aim, for example, at describing the characteristics of image quality that impact the classification errors. Image features that do not impact the classification uncertainty or the end-results are of no concern.

Requirement 3-b - Identify uncertainty propagation measurements: *The means to estimate the combined uncertainty in high-level information, resulting from interactions between uncertainty factors, must be identified.*

Uncertainty propagation measurements aim, for example, at describing the magnitude of classification errors as a function of image quality features. In the case of population monitoring, uncertainty propagation measurements eventually describe the potential noise and bias in class sizes (i.e., random or systematic discrepancies between class sizes and true population sizes). Uncertain propagation measurements must match the high-level information that users are analysing. For example, if users are analysing trends in class sizes, e.g. populations' growth rates, then the uncertainty propagation measurements must express uncertainty in terms of growth rates (e.g., providing confidence intervals for growth rates rather than population sizes).

We address these requirements in Chapter 4 where we identify existing and missing uncertainty assessment methods for computer vision systems such as the Fish4Knowledge system, and in Chapter 5 where we introduce factor-specific measurements addressing classification uncertainty.

2.7.4 Estimate uncertainty in end-results

End-users need to interpret the uncertainty in the specific datasets they are analysing. Each dataset has specific characteristics which can vary across datasets and impact the uncertainty. For instance, datasets can be drawn from videos with different image quality. To estimate the uncertainty in specific datasets in particular, it is necessary to i) identify the datasets' characteristics that can impact the uncertainty factors; ii) measure uncertainty in controlled conditions, with datasets whose characteristics are representative of the potential end-usage datasets; and ii) estimate uncertainty in specific data subsets by accounting for their specific characteristics.

Requirement 4-a - Identify the typical characteristics of end-usage datasets: *The possible characteristics of end-usage datasets, e.g., the range of feature values, must be identified.*

The uncertainty measurements must cover the potential conditions that can be encountered in practice when applying the classification system. Thus the possible values of datasets characteristics must be identified. For instance, the range of image quality features must be identified. The datasets characteristics to consider are those impacting the uncertainty factors, identified by requirements 3-a.

Requirement 4-b - Measure uncertainty in controlled conditions: *Uncertainty measurements must be performed for the most typical characteristics of uncertainty factors.*

Given the uncertainty measurement methods identified through requirements 3-a and 3-b, uncertainty measurement must be performed in controlled conditions that represent the potential end-usage conditions identified through requirement 4-a. For example, classification errors must be measured for the typical characteristics of image quality, e.g., for the potential values of contrast and luminosity.

Requirement 4-c - Assess uncertainty in specific datasets: *Uncertainty in specific sets of classification data must be estimated using the uncertainty measurements in controlled conditions, and the specific characteristics of the dataset.*

Uncertainty in specific sets of end-results must be estimated using the uncertainty measurements performed in controlled conditions, provided by requirement 4-a. For example, classification errors can be estimated using groundtruth evaluations performed on test sets. However, the uncertainty measurements in controlled conditions may not exactly match those of other datasets. For instance, the rates of classification errors can randomly vary across datasets. Hence, estimating uncertainty in end-results using test set *samples* also carries uncertainty, e.g., due to sample variance, which must also be estimated. For instance, error rate variance must be estimated.

Requirement 4-d - Communicate uncertainty to end-users: *The uncertainty in classification results must be communicated to end-users, in an comprehensive, understandable and accessible manner.*

Uncertainty assessment can only be achieved if end-users are provided with relevant and understandable information, that enable end-users to comprehend the impact of uncertainty factors on their data analysis task. *"Data science can only be effective if people trust the results and are able to correctly interpret the outcomes."* (van der Aalst et al. 2017). End-users who are not experts in classification or computer vision may require specific visualization and user interface support.

In this thesis, we do not address requirements 4-a to 4-b as we do not aim at describing the particular characteristics of a single classification system. Requirement 4-c is address in Chapter 5, which methods for estimating numbers of classification errors in specific end-usage datasets. Requirement 4-d is addressed in:

- Chapter 3 where we investigating the uncertainty information of interest to end-users, and its impact on users' trust,
- Chapter 6 where we investigate visualizations that communicate classification errors,
- Chapter 7 where we investigate the Fish4Knowledge interface design that conveys comprehensive information on multifactorial uncertainty.

2.8 Conclusion

This chapter provides an overview of the domain of population monitoring for marine ecology research. Our analysis outlines the potential applications of computer vision for this domain, and answers our first research question: *What high-level information needs and uncertainty requirements in marine ecology research can be addressed with computer vision systems?*

Key high-level information needs are identified: fish counts, species recognition, behavior recognition, fish body size. They address four main topics of research: population dynamics, migration, reproduction and trophic systems. The most essential information needs are fish counts and species recognitions. This information supports all four topics of study (Table 2.5). Information on fish behaviors and body sizes are important for studying reproduction and trophic systems. Fish body size is also of interest for describing the age groups underlying population dynamics.

Information on fish count and species recognition can be provided by computer vision systems that integrate classification software, e.g., for detecting fish and non-fish objects (binary classification) and recognizing fish species (multiclass classification). Computer vision systems can estimate fish body size if appropriate hardware is implemented, e.g., stereoscopic vision, or calibrated fields of view. Recognizing fish behaviors does not required specific hardware, as classification software can address this problem. However, it is challenging to address the variety of fish behaviors: their characteristics differ depending on each species, and collecting groundtruth datasets for each behavior of interest is tedious and costly.

We outline uncertainty issues that are inherent to marine ecology research, and that computer vision systems compound. Uncertainty issues include sampling errors (Section 2.4.3) and biases arising from the characteristics of fish species (Table 2.5). Further investigations are required for establishing more comprehensive uncertainty assessments and related user information needs. We thus proposed a set of high-level requirements that provide guidelines for addressing the information needs of end-users dealing with the multiple uncertainty issues of computer vision and classification systems (Section 2.7). These requirements provided directions for the remainder of the research presented in this thesis.

The user needs and domain requirements we present in this chapter inform the design of computer vision systems for a broad range of applications within marine ecology research. We provide insights for eliciting functionalities that address important user requirements, depending on the topics of research and the characteristics of ecosystems and species of interest. These findings informed the design of the Fish4Knowledge system and its user interface.

Chapter 3

Establishing Informed Trust

To support informed trust and acceptance of classification systems, end-users must be provided with sufficient information on the classification errors that such systems entail. End-users must be aware of the types of errors (e.g., False Positives and False Negatives), their magnitudes, and their impact on classification results. Without such information, end-users may mistrust or misinterpret classification results.

This chapter investigates users' understanding of classification errors, and its impact on users' trust and acceptance of classification systems. We highlight mechanisms that underlie informed or uninformed trust and acceptance. Our findings inform the design of methods and tools for supporting user awareness of classification uncertainty, and answer our second research question: *What information on classification errors is required for end-users to establish informed trust in classification results?* (Section 1.4).

Our investigations are conducted within the context of the Fish4Knowledge project, where classification techniques are used to detect fish and recognize their species. We investigate how information about classification errors (Section 3.1) delivered with different levels of detail (Section 3.2) can impact users' *understanding*, *trust* and *acceptance* of classification systems (Section 3.3). We also investigate which *information needs* about classification uncertainty remain unfulfilled.

We observe that users' trust and acceptance can remain relatively high regardless of the information delivered on classification errors, or the actual understanding of this information (Section 3.4). Detailing the types and magnitudes of classification errors can increase users' trust and acceptance, unless users' skepticism increase together with their understanding of the classification errors. User information needs on classification uncertainty are broader and additional uncertainty assessments are required, regarding classification errors and other uncertainty factors (Section 3.5).

3.1 Errors in binary classification

Our study introduced users to a basic classification algorithm: the *Fish Detection* algorithm that identifies fish occurring in video images. Fish detection is an interesting classification task for our investigations because it is in-between lower- and higher-level tasks of the computer vision system:

- It is impacted by lower-level uncertainty factors (e.g., image quality, segmentation errors) which users may also wish to investigate.
- It serves as the basis for higher-level computer vision algorithms (e.g., the *Species Recognition* algorithm that classifies fish into species). Thus understanding *Fish Detection* uncertainty is required for understanding how uncertainty propagates to higher-level information.
- It is simpler to evaluate for users with no technical expertise because it deals with only two classes (i.e., binary classification of *fish* or *non-fish* objects) while other classification algorithms may involve numerous classes (e.g., 23 classes for the *Species Recognition* algorithm of the Fish4Knowledge project).

3.2 Experimental setup

We recruited 15 marine ecology experts as described in Chapter 2 (Section 2.1 p.18). Six participants, who also completed our first study, performed the experiment at their workplace while monitored by two user interface experts (Elvira Arslanova and myself). The other 9 participants performed the same experiment remotely through an online interface, without being observed by the experimenters. The experimental interface was the same for all participants.

The interface presented short tutorials that gradually explained the *Fish Detection* algorithm and the method used to measure its classification errors (i.e., groundtruth evaluation with test sets). The tutorials were organized in three tabs:

- The **introduction tab** described the video collection and the groundtruth test set and training set (Fig. 3.1).
- The **video analysis tab** presented an evaluation of the *Fish Detection* algorithm (Fig. 3.2 left).
- The **application tab** presented an example of the *Fish Detection* results, using synthetic data representing fish counts and seasonal trends over one full year (Fig. 3.2 right).

The technical concepts were gradually introduced in 3 steps, with dedicated tutorials. At each step, the *video analysis* and *application tabs* introduced additional information and technical concepts, while the *introduction tab* remained identical. The tutorials provided examples of *Fish Detection* results and errors which were all

drawn from simulated data. Using simulated data allowed us to control the error magnitudes, which were relatively high in order to expose participants to significant levels of uncertainty. The technical concepts were explained as follows:

- **Explanations at Step 1 - Errors in fish counts**: The *Fish Detection* algorithm learns the fish appearance using a groundtruth *training set*, i.e., a set of videos in which fish are manually detected. The number of fish detected by the *Fish Detection* algorithm may not match the actual number of fish appearing in the videos, i.e., *fish counts can be over- or under-estimated*. The difference between actual and automatic fish counts can be measured using a groundtruth *test set* distinct from the *training set*.
  The *video analysis tab* compared fish counts from *Fish Detection* software and *test set*. The *application tab* presented an example of *Fish Detection* results, showing fish counts and seasonal trends over one full year. It provided an extrapolation of the errors to expect, assuming the magnitude of errors remains as in the test set.

- **Explanations at Step 2 - Types of errors**: There are two types of errors. *False Negatives* are fish that were not detected, and *False Positives* are non-fish objects that were detected as fish.
  The *video analysis tab* detailed the comparison of actual fish counts and *Fish Detection* results by showing the numbers of False Positives, True Positives and False Negatives. The *application tab* extended the extrapolation of errors in the data example by adding estimates for the False Positives.

- **Explanations at Step 3 - Balancing the types of errors**: The tradeoff between *False Negatives* and *False Positives* can be controlled using a *threshold* parameter, e.g., increasing the threshold decreases the False Positives but increases the False Negatives.
  The *video analysis tab* showed the numbers of False Positives, True Positives and False Negatives for 4 different values of the *threshold* parameter. The *application tab* extended the simulated example of the *Fish Detection* results by showing the different fish counts that would be obtained if using the *threshold* values presented in the *video analysis tab*.

At each step, a questionnaire evaluated the impact of the information that was introduced (Table 3.1). At each question, participants could provide feedback in free form text. The questionnaire investigated which **information needs** regarding uncertainty remained unfulfilled, and measured user **understanding** of the information presented in the tutorial, user **trust** in the computer vision system, and user **acceptance** of the system and its uncertainty.

With this experimental setup, we observed how the technical information, and its *understanding* by users, impacted the *trust* and *acceptance* of the fish detection algorithm, and the fulfilment of *information needs*. In the next Section 3.3, we specify the concepts of *trust*, *acceptance*, *understanding* and *information needs*, and the method used to measure them with our questionnaire.

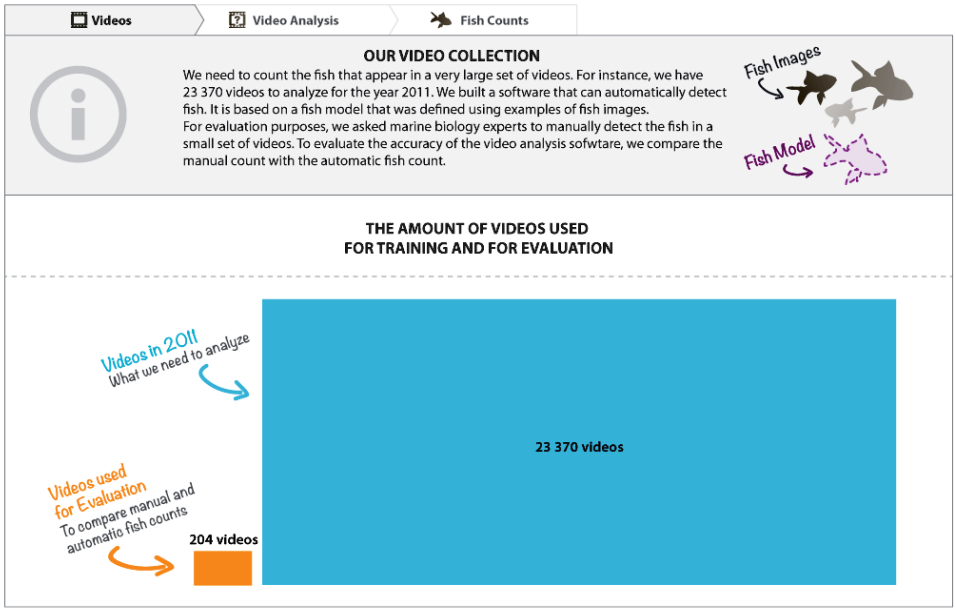

Figure 3.1: Interface tab introducing the *Fish Detection* software (Steps 1, 2, 3).

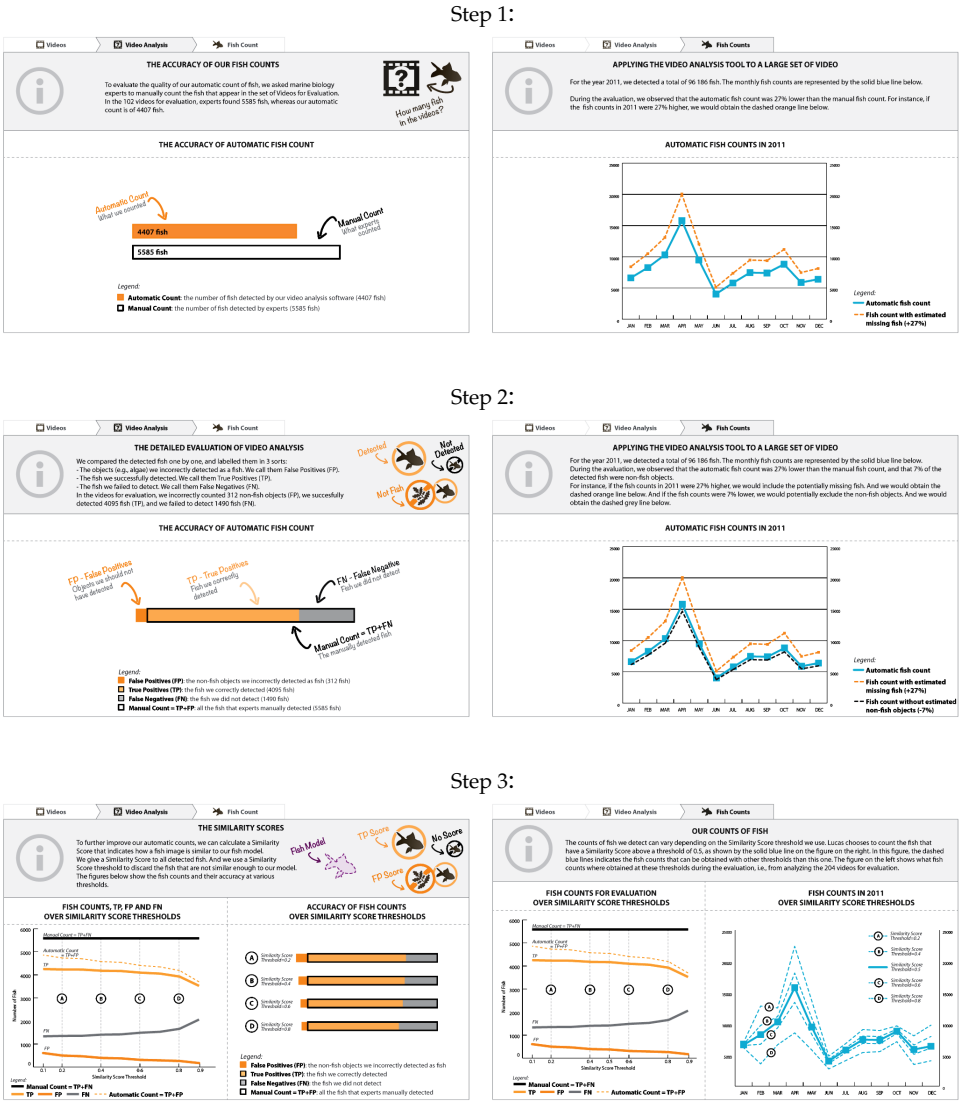

Figure 3.2: Interface tabs showing the *Fish Detection* errors measured with a *test set* (left), corresponding levels of errors for the complete dataset (right, top and middle), or alternative results obtained using different parameter settings (bottom right).

Step 1 - Information on Uncertainty: Errors in fish counts			
Q1	What is this trend? How likely is it to be the same in reality? *[asked 4 times with 4 different trends, Appendix A, Figure A.3 p.190]*		**T**
Q2	Can this explain the difference between manual and automatic counts: i) The automatic fish count is likely to contain non-fish objects (e.g., rocks) that are incorrectly considered as being a fish. ii) When one single fish swims in and out of the camera's field of view, it is counted several times by the video analysis software. It is also counted several times by the experts that manually count the fish. iii) The automatic fish count is likely to miss some fish that are not detected at all.		**U**
Q3	i) Some videos may be missing due to errors during the recording of the video. ii) Some videos may be of very poor quality due to video encoding errors. iii) Some videos may be of very poor quality due to dirt or algae on the camera lens. iv) Some videos may not be analyzed at all due to video processing errors. v) The camera's field of view may have changed (e.g., due to strong current). vi) For the large collection of videos for the year 2011, some fish counts may include more non-fish objects, in a much greater proportion than for the videos used for evaluation. vii) For the large collection of videos for 2011 some fish counts may miss more non-detected fish, in a much greater proportion than for the videos used for evaluation.	Can we encounter these errors? Do you want to evaluate the importance of these errors?	**U** **I**
Q4	Which is the most accurate version of the software? *[asked twice with different datasets, Appendix A, Fig. A.1 p.189]*		**U**
Q5	Which fish count would you choose to use for studying the variations of fish counts over time? *[with or without extrapolation of classification errors]*		**I**
Q6	A-i) This software is suitable for counting fish. A-ii) The automatic fish counts produced by the software are as good as the fish counts that marine biology experts could produce. A-iii) The accuracy of the software is good enough to be used for the scientific study of trends in fish abundance. A-iv) I would like to use the video analysis software to count fish.		**A**
	T-i) The software uses an appropriate method for analyzing the videos and counting fish. T-ii) The system correctly handles the errors it produces. T-iii) The automatic fish counts are trustworthy.		**T**
	U-i) I fully understood the explanations given about the video analysis software. U-ii) I fully understand how the video analysis software works. U-iii) I know how the errors produced by the video analysis software can influence the results of my scientific study of fish counts. U-iv) I understand how to handle the errors that were produced by the video analysis software and minimize their influence on my scientific research.		**U**
	I-i) The software is transparent about its possible errors. I-ii) The given explanations contained enough information for understanding how the video analysis software works. I-iii) I would need more explanations about how the software works. I-iv) It is easy to understand how the video analysis software works. I-v) I was interested in the explanations given about how the video analysis software works.		**I**
Step 2 - Information on Uncertainty: Types of Errors (FP, FN)			
Q2	Does it influence the number of False Positives (FP), True Positives (TP), and/or False Negatives (FN): i) Some versions of the software are more likely to detect non-fish objects (e.g., seaweed) as being a fish. ii) Some versions of the software are more likely to correctly detect the fish in the videos. iii) Some versions of the software are more likely to miss the detection of some fish in the videos.		**U**
	Is it possible that A>B, A=B and/or A<B: iv) We compare A) the number of False Positives (FP); and B) the number of False Negatives (FN). v) We compare A) the manual fish count; and B) the sum of True Positives (TP) and False Negatives (FN). vi) We compare A) the manual fish count; and B) the automatic fish count.		**U**
Q1 and Q4-5: *Same as Step 1, Q4 asked thrice (Appendix A, Figure A.2 p.189)*			
Step 3 - Information on Uncertainty: Balancing the Types of Errors (FP, FN)			
Q2	Does it influence the number of False Positives (FP), True Positives (TP), and/or False Negatives (FN): i) Some thresholds are more likely to discard non-fish objects (e.g., seaweed) that were detected as being a fish. ii) Some thresholds are more likely to include non-fish objects in the fish counts. iii) Some thresholds are more likely to incorrectly discard fish that were correctly detected.		**U**
	Is it possible that A>B, A=B and/or A<B: iv) We compare the number of True Positives (TP) for A) a threshold = 0.2; and B) a threshold = 0.6. v) We compare the number of False Positives (FP) for A) a threshold = 0.2; and B) a threshold = 0.6. vi) We compare the number of False Negatives (FN) for A) a threshold = 0.2; and B) a threshold = 0.6.		**U**
Q1 and Q6: *Same as Step 1*			

Table 3.1: Questionnaire investigating the relationships between user **Trust** (**T**, last column) in the video system, **Acceptance (A)** of the system and its uncertainty, **Understanding (U)** of the technical features and sources of uncertainty, and the satisfaction of **Information Needs (I)** on uncertainty issues.

3.3 Trust, acceptance, understanding & information needs

We investigate user information needs w.r.t. uncertainty issues in order to **support informed trust and acceptance of classification systems** such as the Fish4Knowledge system. This section provides definitions for the concepts of Trust, Acceptance, Understanding and Information Needs, and introduces the means we used to measure them.

The definition of trust from (Madsen and Gregor 2000) and (McAllister 1995, p.25) can be adapted to our context as: *"The extent to which a user is confident in, and willing to [use] the [video analysis system]"*. We define **Trust** as the *confidence in* the video analysis system, and **Acceptance** as the *willingness to use* the video analysis system.

Both (Madsen and Gregor 2000) and (McAllister 1995) consider that trust is based on *affect-based* and *cognition-based* components. **Understanding** is a cognition-based component defined as *"the [user] can form a mental model and predict future system behavior"* (Madsen and Gregor 2000, p.11) with a focus on the perceived user understanding (i.e., users self-appraisal of their understanding). We retain this approach and also consider the actual user understanding (i.e., the correct understanding of the technical concepts).

User understanding may be correct but incomplete, as crucial information may be unknown (e.g., information on classification errors or other uncertainty issues). Fulfilling the **Information Needs** on uncertainty issues is necessary to assess the system's Reliability (e.g., *"the system [may not] provide the advice required to make [a] decision"*) and Technical Competence (e.g., *"the advice the system produces [may not be] as good as that which a highly competent person could produce"*) which are the two other cognition-based components of trust.

Affect-based components of trust (*Faith* and *Personal Attachment*) are excluded from this study because classification and computer vision systems are new to our target users. Such systems are not part of our users' practices, thus these users could not develop Personal Attachment to the system nor rely on Faith when using it.

A number of models and scales are used to measure trust in different computational systems (Artz and Gil 2007). However, they concern decision aid systems rather than computer vision systems. Hence we designed a questionnaire (Table 3.1) that addressed our context, and included 5 questions adapted from (Madsen and Gregor 2000) (Q6-A-ii and -iv, Q6-T-iii, Q6-I-vii and -viii).

Some questions were identical at each step of the experiment and others were specific to each step. The step-specific questions (Q2-4) evaluated the *Actual Understanding* of the tutorials. Quantitative measurements were derived from the numbers of correct and incorrect answers. The step-invariant questions (Q1, Q6) evaluated users' *Acceptance, Trust, Perceived Understanding* and *Information Needs*. Quantitative measurements were derived from participants' agreement to statements about the system. The levels of agreement were indicated using Likert-scales with gradual values, where the neutral answer "Neither agree or disagree" scored 0.

Questions Q3 and Q5 collected qualitative feedback and did not contribute to the quantitative measurements. Participants' oral and written feedback complemented our interpretation of the quantitative measurements. For instance, the feedback showed that some participants had a generally good understanding of the uncertainty issues, but gave wrong answers to questions which concepts or terminology were misinterpreted. The detailed participants' answers are given in Appendix A.

3.4 Impact of introducing classification error assessments

We analyse the evolution of users' *Trust* and *Acceptance* over the steps of the experiment (Section 3.4.1) and how it relate to users' *Understanding* of the system and unfulfilled *Information Needs* (Section 3.4.2).

3.4.1 Trust and Acceptance

At the first step of the experiment, 9 participants had *Trust* in the system, i.e., above neutral (Fig. 3.3 top). Other participants remained rather neutral, e.g., neither trusting or distrusting the system. Participants' *Acceptance* of the system did not necessarily match their Trust. Of the 9 trusting participants, 3 expressed neutral or negative Acceptance of the system. Of the 6 more skeptical participants, 2 expressed rather positive Acceptance of the system.

Low Acceptance was related to participants' need for further information on the uncertainty factors[1] and on the variability of classification errors[2] before the system may be deemed suitable for scientific research. High Acceptance despite limited Trust was related to users' acknowledgement that uncertainty is unavoidable with any data collection technique[3] and that computer vision has high potential (e.g., to lower the costs of data collection).

Along the next steps of the experiment, participants' *Acceptance* of the system remained relatively unchanged for most participants, increasing for 2 participants and decreasing for 1 participant. It indicates that Acceptance may rely on factors other than the information provided on classification errors (e.g., on other unfulfilled Information Needs[4]). However, providing details on the classification errors improved some participants' Trust and Acceptance, especially when detailing the errors to expect in the classifier's output [5]. Although all participants were willing to

[1] Participant P4, Step 1, Question Q5: *"I think I will be to understand why we lost 27% [of the fish to detect, due to classification errors]"*.

[2] Participant P5, Step 1, Question Q4: *"Maybe data on several runs and standard deviation of those runs will help to really see which [classifier] is better"*, *"The important is to see how good are the methods giving consistent counts"*.

[3] Participant P3, Step 1, Question Q4: *"Experts might have missed fish too"*.

[4] Participant P5, Step 2, Question Q1: *"The new information don't really solve the doubts expressed before"*.

[5] Participant P12, Step 1, Question Q1: *I'm very convinced about the new line added to the graph with the fish count with estimated non-fish object"*.

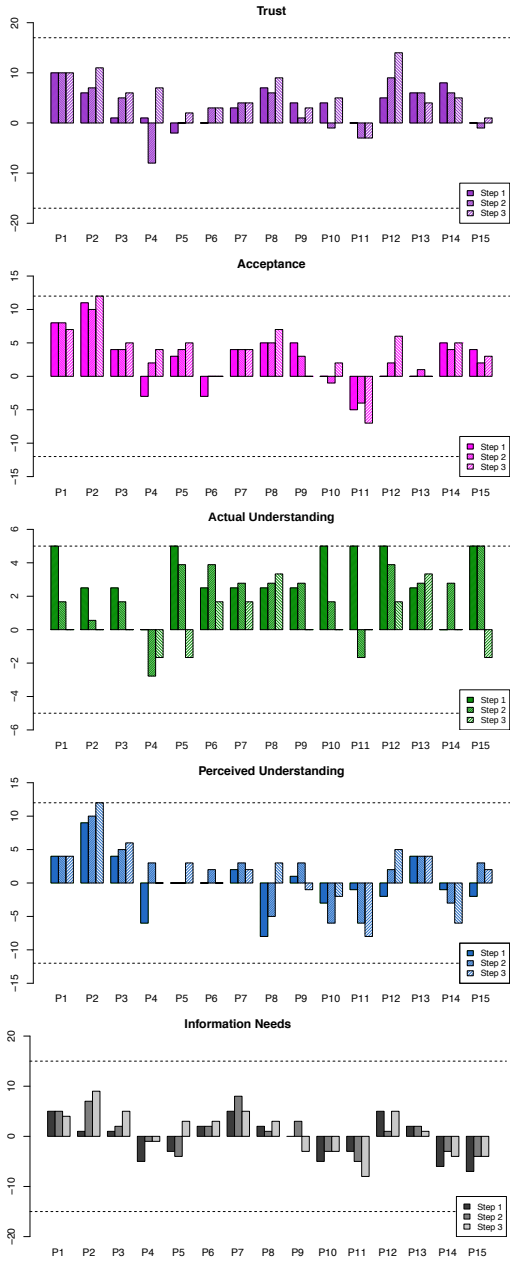

Figure 3.3: Measurements collected for participants P1 to P15. *Measurements are expressed as the sum of Likert-scale values for all questions related to the same concepts (Trust, Acceptance, Actual Understanding, Perceived Understanding, Information Needs). Dashed lines indicate the highest and lowest possible scores. Actual Understanding was measured from different numbers of questions at each step. To support comparisons, the scores were normalized to range from -5 to 5.*

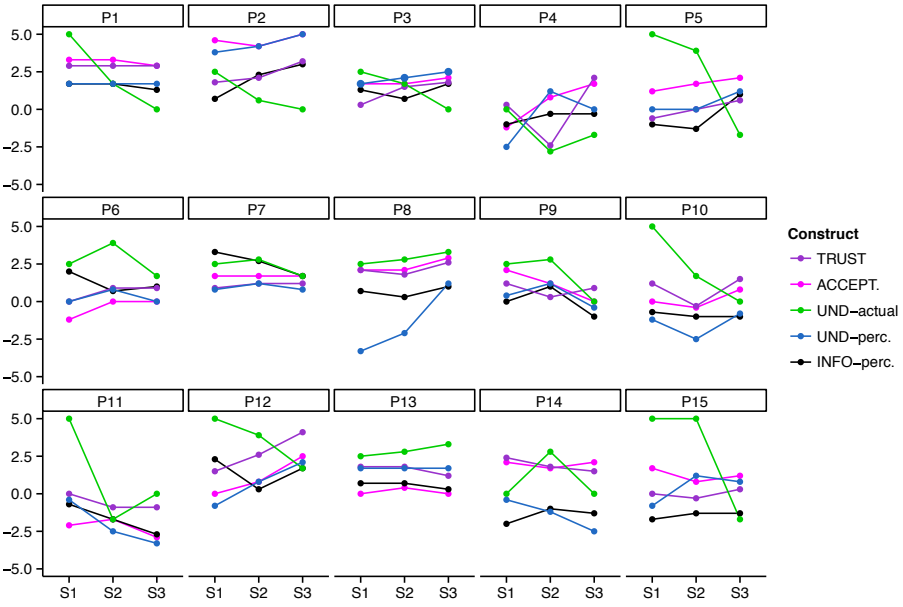

Figure 3.4: Comparison of measurements for participants P1 to P15 at step S1 to S3. *Measurements are expressed as the sum of Likert-scale values for all questions related to the same concepts (Trust, Acceptance, Actual Understanding, Perceived Understanding, Information Needs) and normalized to range from -5 to 5 (to support comparisons).*

use such estimations of errors in classification end-results, they expressed concerns regarding the variability of the underlying error rates[6].

The evolution of participants' *Trust* was consistent with the evolution of their Acceptance, except for participant P4. For most participants, the trends followed the same direction (i.e., gradual increase or decrease, or relative stability). However some participants' Trust has first decreased then increased, especially participant P4. This pattern can be explained by analysing participants' Understanding and Information Needs.

3.4.2 Understanding and Information Needs

Participants' Trust and Acceptance decreased due to either good or poor understanding of the information provided on the classification errors. With a good understanding, participants gained awareness of the uncertainty issues, and were thus more skeptical about using or trusting the system. On the contrary, with a poor understanding, participants struggled to comprehend the classification errors, or the

[6] Participant P5, Step 2, Question Q4: *"The error seems constant all over the trend. But that may not be the case and I want to know when that happens"*. Participant P7, Step 2, Question Q4: *"It is always better to have an estimate of possible error margins"*.

breadth of other issues, and were less confident in the system. However, after developing an understanding of the uncertainty issues and their impact on their data analysis goals, some participants envisioned methods to further address these issues. Their Acceptance increased accordingly, despite their initial skepticism, as they were willing to use the system to conduct more experiments on methods to handle uncertainty.

Participants' *Perceived Understanding* did not necessarily match their *Actual Understanding*. Participants were not always aware that they misunderstood some of the technical concepts about classification errors. For example, participants often misunderstood the test set as being drawn from diving observation (instead of the manual analysis of video footage). In many cases, participants' Perceived Understanding remained low because their *Information Needs* were largely unfulfilled. Participants needed more information on the classification algorithm, and the impact of other uncertainty factors (e.g., image quality, small or occluded objects, fish camouflage). Without such information, some participants considered that their understanding of the system and its classification errors was critically incomplete.

For 8 participants, the levels Actual Understanding did not match those of other measures. Their Actual Understanding decreased while their Trust, Acceptance and Perceived Understanding increased or remained relatively unchanged. It indicates that participants' assessment of the system can rely on factors other than the information provided on the classification errors.

Despite the misunderstandings of the information provided on classification errors, participants' written feedback show that they seek to build informed Trust and Acceptance, and that they developed an understanding of other uncertainty factors.

We conclude that **further Information Needs regarding uncertainty must be fulfilled for users to build informed Trust in the system. High Acceptance of the system may not entail that users Understand the uncertainty issues, nor that the Information Needs about uncertainty are fulfilled.** Users can be willing to use the system despite its uncertainty as:

- Dealing with uncertainty is at the core of users' common practices.
- The system's potential benefits is worth developing the necessary uncertainty assessment methods.
- Using the system is necessary to experiment with the uncertainty assessment methods.

3.5 Unadressed information needs

This section describes participants' unaddressed information needs w.r.t. uncertainty issues. These information needs were derived from users' written feedback (Appendix A, Tables A.5 to A.7). They concern classification errors (Section 3.5.1) and other uncertainty factors (Section 3.5.2). The information needs we identified are synthesized in Table 3.2.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
Classification errors															
Explanation of Terminology	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Causes of Errors				x							x	x			
Errors for Each Species		x		x	x					x	x	x			
Human Errors & Groundtruth Quality			x	x						x	x		x		
Error Rate Variability				x	x	x	x					x			
Classification Errors in End-Results	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Other uncertainty factors															
Domain Knowledge					x	x	x			x	x				
Duplicated Individuals				x				x			x	x			
Image Quality			x				x			x					
Missing Videos			x	x		x	x			x		x			x
Field of View & Sampling Validity		x		x		x	x	x		x	x	x			

Table 3.2: Information needs on uncertainty, derived from ecologists' feedback.

3.5.1 Information on classification errors

Explanation of terminology: Classification errors must be explained carefully. The technical concepts are likely to be overwhelming and misunderstood by users who have no prior knowledge of classification. Further, the classification terminology may conflict with the terminology in the domain of application. For example, the terms *accuracy* and *precision* have different definitions in the classification and ecology domains (Fig. 3.5).

We assumed that the terms groundtruth test set, True Positive, False Positive and False Negative are confusing for non-experts. Hence our questions often replaced the technical terms with common terms (e.g., *"missed fish"* instead of False Negatives). However, such simplified terminology did not ensure that answers were correct (e.g., answers to questions Q2-i to -iii were mostly incorrect). For example, the term *"manual fish count"*, used instead of *"fish count from the test set"*, was often misunderstood as fish counts from diving observations instead of counts from the manual analysis of video footage (e.g., question Q2-ii at Step 1, which answers were thus excluded).

The terminology issues may limit user understanding of the classification uncertainty. At Step 1, only 5 participants correctly answered all the questions evaluating the Actual Understanding of the tutorials (Q2-3). However, all participants understood the visualization of classification errors (Fig. 3.2) and correctly answered all the questions where the information was visualized (Q3). At Step 2, only 1 participant correctly answered all the questions measuring the Actual Understanding, but 10 participants correctly answered all the questions where the information was visualized. Hence visualization may facilitate user understanding of classification errors and help overcoming the terminology issues.

We conclude that terminology issues must not be overlooked. Replacing technical terms by common terms does not ensure a solution to terminology issues. However, visualization is a promising solution to support user understanding, which is investigated further in Chapter 6.

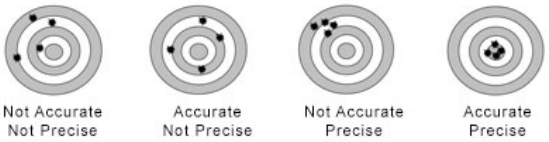

Figure 3.5: Meaning of the terms *Accuracy* and *Precision* for marine ecologists. These differ from their meaning in the classification domain where Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ and Precision = $\frac{TP}{TP+FP}$. *Illustration from the National Oceanic and Atmospheric Administration website, http://www.noaa.gov.*

Causes of errors: Several participants sought to understand what causes the system to misclassify the fish and non-fish objects. Participants needed to understand which application conditions (e.g., kinds of fish, camera settings) can yield high uncertainty[7]. Hence explaining the causes of errors should include explanations of the ecosystem's characteristics that can interfere with the classification algorithms (e.g., lens biofouling, occlusions due to rocks or dense groups of fish).

Errors for each species: Several participants required that classification errors are measured at the species level, i.e., for the species recognition classifier. Participants also required that fish detection errors are estimated for each species separately, in order to assess whether some species yield more errors than others[8] and how fish detection errors may vary as species composition vary (i.e., the relative species population sizes). Without such information, the fish detection errors were of limited interest. However, most participants were interested in the classification errors at the fish detection level. Some participants were even willing to balance the False Positives and False Negatives using the tuning parameter introduced at Step 3[9].

We conclude that assessing uncertainty propagation is a key information need. Users need to assess what variations of objects' features (e.g., their species) affect the classification errors, and how errors from one classifier can impact the errors of another classifier (e.g., how *Fish Detection* error impact *Species Recognition* errors).

Human errors & groundtruth quality: Several participants mentioned that humans too make errors when classifying fish, and may produce different manual fish counts[10]. It impacts the quality of the groundtruth which is produced by different humans. Some participants requested that the human errors in the groundtruth be

[7] Participant P4, Step 1, Question Q5: *"Why we lost 27% [of the fish to detect]"*. Participant P11, Step 1, Question Q6: *"You don't explain how the software is counting the fish. Does it react on movement?"*. Participant P3, Step 1, Question Q3: *"Video blocked by an object"*. Participant P11, Step 1, Question Q2: *"Some fishes if they swim too far away from the camera and could not be detected by software especially when the water visibility is not good [...] especially when camera lens has biofouling problem"*.

[8] Participant P11, Step 1, Question Q2: *"Smaller body size fish or cryptic fish may not be detected"*. Participant P2, Step 1, Question Q3: *"Benthic fish can be missed"*.

[9] Participant P10, Step 2, Question Q4: *"Rather have the non-fish selections removed from my data-set [less FP] then have more fish in my count [less FN]"* .

[10] Participant P4, Step 1, Question Q2: *"Different experts' count will be different"*. Participant P13, Step 1, Question Q3: *"Inter-observer differences"*. Participant P3, Step 1, Question Q4: *"Experts might have missed fish too"*. Participant P11, Step 1, Question Q2: *"Certainly different divers may have different results. That is a bias by different observers"*.

evaluated[11] or that the magnitudes of classification errors be compared with the variability of human observations[12].

We conclude that users need information on i) the human errors and disagreements when producing the groundtruth; and ii) the impact that such groundtruth uncertainty can have on classification errors and their measurement.

Image quality: Questions Q3-ii and -iii investigated the issue of poor image quality due to dirt accumulating on camera lenses (Q3-ii) and encoding errors that can erase sections of the images (Q3-ii). These image quality issue may increase the number of classification errors. Almost all participants acknowledged this issue, but only 3 participants requested more information about it. Participants may consider that low image quality occurs randomly or rarely and are thus negligible. However, lens fouling may consistently lower image quality for long periods of time (i.e., until lenses are cleaned) and encoding errors can significantly increase the classification errors (e.g., we observed extreme peaks of False Positives). Other image quality issues may systematically modify the error rates (e.g., water turbidity due to environmental events such as typhoon, low light at dawn and dusk, colour bias due to algae bloom).

We conclude that users need explanations to understand the importance image quality. For example, users need to know which image quality issues can impact the error rates, the magnitudes at which error rate may differ, and how frequently, randomly or systematically the image quality issues can occur. Further, when analysing classification results, users should be provided with information on the quality of the images from which the results were drawn.

Error rate variability: Questions Q3-vi and -vii investigated users' concerns for error rate variability: the rates of False Positives (Q3-vi) or False Negatives (Q3-vii) may vary across datasets, thus the error rates measured from the test set may differ from the error rates in other datasets. Almost all participants acknowledged this issue but only one participant requested more information about it. However, in the text feedback 4 other participants requested information on error rate variability but using a different terminology (e.g., *"standard deviation"*, *"error margin"*)[13].

We conclude that providing estimates of error rate variability is a relevant information need to address. This issue is investigated in Chapter 5.

Classification errors in end-results: Question Q5 at Step 1 and 2 investigated users' need for estimating the number of classification errors in the end-results. Such estimation can be performed using the error rates measured for the groundtruth test

[11] Participant P10, Step 1, Question Q4: *"We have to take human as well as computer errors into account"*.

[12] Participant P11, Step 1, Question Q3: *"We should have both data, one from software count and another from divers count, and then make a comparison study"*.

[13] Participant P5, Step 1, Question Q4: *"Maybe data on several runs and standard deviation of those runs will help to really see which [classifier] is better"*. Participant P5, Step 2, Question Q5: *"The error seems constant all over the trend. But that may not be the case and I want to know when that happens"*. Participant P7, Step 2, Question Q5: *"It is always better to have an estimate of possible error margins"*. Participant P4, Step 1, Question Q4: *"Because of interference [there] may not [be] much difference [between the classifiers' results]"*, *"interference"* concerned the variability of error rates (e.g., random variations due to sample variance, and systematic variations due to biasing factors such as species composition or image quality).

set, assuming they are representative of the error rates in the end-results. Such estimation is particularly relevant for assessing the trends in population sizess[14]. For example, a class size may increase due to an increase of classification errors, while in reality this class size is not increasing. All participants but one required such estimation of errors in end-results, sometimes with a focus on either False Negatives or False Positives[15]. The remaining participant was skeptical because the extrapolation method must be verified, e.g., to account for the error rate variability [16].

We conclude that estimating the classification errors in end-results is a key information needs that must be addressed for providing accountable classification systems. This information need is related to the need for estimating *error rate variability*. The error estimation relies on the assumption that error rates in the test sets are similar to error rates in the end-results. If error rates differ, the error estimation is inaccurate. This problem is investigated in Chapter 5.

3.5.2 Information on other uncertainty factors

Domain knowledge: Several participants stated that their trust in the system's results needs to be rooted in prior knowledge of the ecosystem, its species, and the usual trends in fish populations[17]. Participants also needed to compare the classification results with results obtained from a well-accepted and trusted technique, such as diving observations[18]. The need for prior knowledge requires information beyond the scope of what classification and computer vision systems can provide. Thus we did not include this information need in the scope of user needs we address in this thesis.

Duplicated individuals: Question Q2-ii at Step 1 investigated the issue of individual fish that swim in and out of the field of view, and are thus detected several times by the classification system. The text feedback showed that ecologists usually try to identify unique individuals, and avoid counting them several times[19]. Further feedback from ecologists indicated that the chances of repeatedly counting individuals depend on the species' swimming behaviors. For example, the chances

[14] Participant P4, Step 2, Questions Q4-5: *"The trend is the focus, not the numbers"*.

[15] Participant P10, Step 1, Question Q5: *"It's relevant to know how much errors in the estimates you have, especially if you want to use the data for further analysis!"*. Participant P12, Step 2, Question Q1: *"I'm very convince about the new line [...] with estimated non-fish objects"*.

[16] Participant P11, Step 1, Question Q5: *"First of all, I should know how you estimate the missing fish and whether it is reasonable or not"* . Participant P11, Step 2, Question Q5: *"We should do some evaluation on the accuracy [Fig. 3.5] of video analysis"* .

[17] Participant P5, Step 1, Question Q1: *"Can I say that this likely to be what is happening there? No I can't without background information on location, species composition, etc."*.

[18] Participant P11, Step 1, Question Q3: *"We should have both data, one from software count and another from divers count, and then make a comparison"*.

[19] Participant P4, Step 1, Question Q2: *"The expert will not repeat count for the same fish"*. Participant P11, Step 1, Question Q2: *"Diver can judge whether the fish swim out and in the camera field is the same or different individual"*. Participant P12, Step 1, Question Q2: *"When doing the fish count manually it is more likely that the same fish has not been recorded several times"*.

of individuals swimming in and out of the field of view are higher for schooling species and sedentary species living in coral heads.

We conclude that estimating the chances of duplicates for each species is a key information need. This information allow users to assess potential biases in the classification results. For example, sedentary species may have the largest population size in the classification results, while in reality they represent a small number of individuals that live in the coral heads in front of the camera.

Missing videos: Questions Q3-i and -iv investigated the issue of video footages that are missing or unusable due to issues when encoding or processing the videos. Missing videos reduce the number of video samples used to monitor which fish populations, and thus reduce the external validity of the conclusion drawn from computer vision data. It also impacts the comparisons of fish counts drawn from different sets of videos: the more videos the more fish, and the more representative the trends. Almost all participants acknowledged this issue, but only half of them requested further information about it. Participants may consider that missing videos occurs randomly or rarely, and are thus negligible. However, technical incidents may interrupt the monitoring of significant time periods or locations.

We conclude that the number of available videos needs to be provided to end-users, with information on their location and time periods. This information allow users to estimate the sample size, and the uncertainty that may result from small or unequal samples.

Field of view and sampling validity: Questions Q3-v investigated issues with static cameras' field of view that can shift over time, e.g., due to strong current or lens cleaning operations. Almost all participants acknowledged this issue, and 7 participants requested more information about it. The feedback also mentioned further issues with the fields of view, e.g., accidental occlusions, parts of the ecosystem that are over- or under-represented, or the size of the areas within the field of view[20]. We conclude that the cameras' field of view is a key uncertainty issue, as it strongly impacts the validity and consistency of the sampling method. Users need information on the parts of ecosystems that are observed or not. Users also need information of how the fields of view of static cameras have shifted over time. These shifts can be inspected manually, by browsing the video footages. The shifts can also be detected automatically by developing dedicated computer vision algorithms.

3.6 Conclusion

This chapter reports mechanisms underlying the development of informed trust and acceptance of classification systems. We identify information needs that support the development informed trust and acceptance of classification systems, and answer

[20] Participant P12, :*"The range of view.. especially if you want to compare the videos. For instance, when coral is blocking the view of the cameras. Also, the position of the cameras, because you can miss certain reef associated fish species when the cameras are pointing a bit upwards"*.

our second research question: *What information on classification errors is required for end-users to establish informed trust in classification results?*.

Users' trust and acceptance of classification systems may not be supported by their actual understanding of classification errors. Users may not be aware that they do not fully understand the types of classification errors, or their impact on end-results (i.e., users' perceived and actual understanding may not match). To support user understanding, particular attention must be paid to the technical terminology used to describe the classification errors. The visualizations used in our experiment offered promising support for improving user understanding of classification errors. This finding motivates the development of simplified visualizations that address the needs of non-expert end-users, presented in Chapter 6.

Users may accept classification systems without trusting them nor understanding their errors. This behavior arises from users' interest in the opportunities that such systems provide, e.g., for collecting information that would otherwise be unavailable or costly. Users may also accept uncertain classification systems for experimental purposes, e.g. to develop uncertainty assessment methods that fit their requirements. In contrast, users may correctly understand the classification errors, deem their magnitudes acceptable, and yet not trust or accept classification systems. This behavior is due to information needs on uncertainty issues that remain largely unfulfilled.

Several uncertainty factors must be considered to develop informed trust and acceptance of classification systems (Table 3.2). Providing measurements of classification errors drawn from test sets does not address all these uncertainty factors. For instance, underlying factors can impact the magnitudes of classification errors, e.g., the image quality. Test sets can contain human errors and misrepresent the applications conditions, e.g., the image quality. Error rates may systematically or randomly vary between datasets, e.g., depending on image quality or sample variance. Hence error measurements drawn from test sets may not represent the errors in end-usage datasets. However, end-users require such estimation of classification errors in end-results. Thus statistical methods are required to assess the reliability of such classification errors estimates, e.g., accounting for error rates' variability between test sets and end-usage datasets.

These findings inform our model of uncertainty factors pertaining to computer vision systems for population monitoring, presented in Chapter 4. They also motivate the development of methods for estimating the classification errors in end-results, presented in Chapter 5.

Chapter 4

Uncertainty Factors and Assessment Methods

In Chapters 2 and 3 we identified ecologists' concerns for uncertainty issues arising from the computer vision system and its deployment conditions. From these insights, this chapter synthesizes key uncertainty factors of concerns to end-users of computer vision systems for population monitoring. A model of interactions among uncertainty factors, and ensuing uncertainty propagation, is derived. The model provides guidelines for reviewing how uncertainty assessment methods address the uncertainty factors and uncertainty propagation of concern to end-users.

Uncertainty factors are identified from the perspective of a core task in ecology research, identified in Chapter 2: the analysis of population sizes, e.g. populations from specific species or exhibiting specific behaviors. For instance, analysing population sizes over time periods and locations supports the study of migration or reproduction (Section 2.3).

We consider computer vision systems that classify individuals occurring in video footage into classes representing the populations of interest, e.g., a class can represent a species or a behavior. Ecologists can then analyze the numbers of individuals per class, i.e., the class sizes. For instance, within the Fish4Knowledge system, class sizes represent population sizes of different fish species.

To assess the validity of video-based estimations of population sizes, we must consider how uncertainty propagates through the computer vision system and its components (e.g., the pipeline of classification components). We must also consider the uncertainty that arises from the application conditions (e.g., from the environment in which the system is deployed). These requirements are identified in Chapter 2 (requirements 1-a and 1-b, Section 2.7.1, p.32).

We first specify the typical computer vision system and application conditions we consider (Section 4.1). We then describe the uncertainty factors arising from the computer vision system, the deployment conditions, or both (Section 4.2). Finally, we analyse the interactions between uncertainty factors and how uncertainty propagates into high-level information (Section 4.3). This model of uncertainty factors addresses requirements 2-a and 2-b in Chapter 2 (Section 2.7.2, p.33) and our third research question: *When applying computer vision systems for population monitoring, what uncertainty factors can arise from computer vision systems, and from the environment in which systems are deployed?*

We conclude our analysis of uncertainty factors by discussing the applicable uncertainty assessment methods (Section 4.4) and discuss uncertainty factors unaddressed in the literature (highlighted in Figure 4.2, p.65). This overview of uncertainty assessment methods addresses requirements 3-a and 3-b in Chapter 2 (Section 2.7.3, p.34) and partially addresses our fourth research question: *How uncertainty assessment methods address the combined effect of uncertainty factors?*

Our model of uncertainty factors, and overview of uncertainty assessment methods, synthesize the insights we collected in Chapters 2 and 3. These insights are drawn from interviews with marine ecology experts and computer vision experts (introduced in Chapter 2, Section 2.1, p.18). Involving experts from both system and application domains limited the issue of *"framing problems such that the context fits the tacit values of the experts and/or fits the tools, which experts can use to provide a solution to the problem"* (Walker et al. 2003) as the *locations* of uncertainty may lie beyond those considered by a single domain of expertise (i.e., computer vision or marine ecology).

4.1 Sources of uncertainty

Uncertainty arises from the interactions of different factors, depending on the technologies employed by the system and the conditions in which the system is deployed. *"Different forms of uncertainty are introduced into the pipeline as data are acquired, transformed, and visualized"* (Pang et al. 1997) and *"uncertainty gets transformed as data moves through the analytics process"* (Correa et al. 2009). We thus consider sources of uncertainty from both 1) the computer vision system, i.e., arising at the *data processing* step and 2) the deployment conditions, i.e., arising at the *data collection* step (Pang et al. 1997). The *context* of the system (e.g., the deployment conditions) is of particular concern since *"external driving forces [can] have an influence on the system and its performance"* (Walker et al. 2003).

Hence this section describes the main elements of the computer vision system (Section 4.1.1) and its deployment conditions (Section 4.1.2). Defining such *"logical structure of a generic system model within which it is possible to pinpoint the various sources of uncertainty"* is essential for identifying the *locations* of uncertainty (Walker et al. 2003). For instance, uncertainty can be *located* in each component of the computer vision system.

4.1.1 Computer vision system

We consider a computer vision system that uses classification algorithms to monitor the sizes of different classes of animal populations. The classes can represent animal species (e.g., fish species in the Fish4Knowledge project) or behaviors (e.g., preying, mating). Our scope does not include other measurements such as body sizes, requiring other technologies than those used in the Fish4Knowledge project.

Computer vision systems may apply different kinds of algorithm (e.g., SVM, Bayes, GMM) and low-level feature extraction methods (e.g., Fourier descriptor, Gabor filter, Histogram of Oriented Gradients, Moment Invariants). Regardless of the kind of algorithms and feature descriptors, the computer vision systems we consider perform 3 main high-level tasks: binary classification (e.g., detect individuals), tracking (e.g., follow individuals across video frames), and multiclass classification (e.g., recognize species or behaviors). We focus on a typical pipeline of algorithms that performs such classification and tracking tasks (Figure 4.1). This pipeline was, for example, deployed within the Fish4Knowledge project (Fisher et al. 2016, Beauxis-Aussalet et al. 2013).

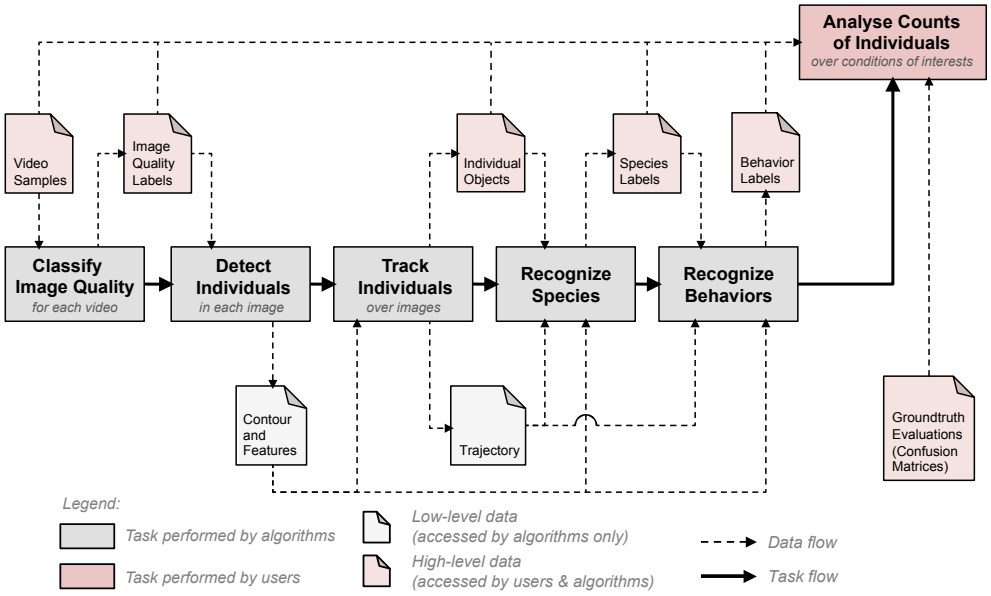

Figure 4.1: Typical pipeline of computer vision components, each introducing potential uncertainty (BPMN notation).

Our scope of algorithms excludes low-level sub-processing algorithms that are not directly related to the end-user's task of analysing class sizes. For instance, algorithms which *Detect Individuals* use lower-level segmentation algorithms that classify each pixel as being within or outside an object contour. Imperfect segmentation influences the uncertainty of higher-level algorithms, but measuring segmentation errors does

not directly contribute to assessing the errors in the class size estimates. However, other use cases may require the estimation of such segmentation errors (e.g., land coverage estimated from satellite images, where segmentation detects types of lands, and class sizes represent area sizes).

The use cases captured by the computer vision system in Figure 4.1 may have chosen alternative implementation strategies. For instance, *Recognize Species* may be performed before *Track Individuals*, as species labels can be used by the tracking algorithm. This would impact how uncertainty propagates in the system. The key uncertainty factors would remain unchanged, but their interactions and related uncertainty propagation would differ.

Our pipeline of algorithms relies on two important conditions that, if inapplicable, can introduce additional uncertainty factors.

1. The system processes continuous video streams that are sequenced in video clips of equal duration, called *video samples*. For instance in the Fish4knowledge project, the video streams are split into 10-minute samples. Considering video samples of equal duration simplifies the uncertainty assessment.

2. Image quality is assessed for each video sample, and classified into several categories. The next classification algorithms use this information to apply different parameters depending on the image quality (e.g., correcting exaggerated green colors in case of algae bloom, or low contrasts at dawn and dusk). Image quality could be measured with continuous values (e.g., blur score) or within parts of each image (segmentation). Opportunities of such approaches are worth being investigated in future work.

In the system we consider, after classifying the image quality of video samples, individuals are detected in each video frame (binary classification). Object features (e.g., contour, texture) are extracted, normalized depending on image quality, and made available to other algorithms. The tracking algorithm identifies the trajectory of individuals over each video frame. The species recognition algorithm classifies each individual into a species (multiclass classification) considering all images along the individual's trajectory. Finally, the trajectory, species and features of individuals are used to classify their behaviors (multiclass classification, although multi-label approaches are relevant but not considered here).

4.1.2 In-situ system deployment

In-situ video monitoring involves dispatching cameras in the ecosystem of interest, as well as setting servers to host the computer vision system and process the videos. The computations executed on the servers may fail, resulting in missing data or video samples. The ecosystems' environment is subject to changes of light (e.g., low contrast and skewed colors at dawn and dusk) or weather conditions (e.g., storms yielding murky waters). These can impact the image quality and degrade the camera setup (e.g., dirt on the lens, camera breakdown, camera displacement).

The cameras' features (e.g., frame rates, resolution, lenses) also impact the image quality, as well as the breadth and depth of the field of view. Their placement in the ecosystem, and the coverage of their field of view, can also impact the image quality. For instance, cameras with large depths of view can observe distant thus fuzzy objects.

The geographical or topological locations of the cameras are crucial for implementing a correct sampling of ecosystems and populations of interest: the monitored ecosystems' components (e.g., habitats, sources of food or shelter) greatly impact the species and behaviors that can be observed. The monitored time periods are also crucial: seasonal and daily cycles greatly impact the species and behaviors that can occur at specific locations (e.g., nocturnal species, seasonal behaviors like mating).

Depending on the sampling strategy, and the types of species or behaviors of interest, end-users can choose between static or moving cameras (e.g., handheld by divers, or trawled by boats), operating a different depth, altitudes or habitats, with distinct or overlapping fields of view (e.g., stereoscopic vision), oriented towards an open view or a specific ecosystem element (e.g., rocks or coral heads), with or without devices designed to attract or repel individuals of interests (e.g., bait, light, noise). The cameras can also be deployed in artificial, experimental environment (e.g., fish tanks, zoos).

Amongst the variety of potential application setups, we focus on setups that consist of static cameras, with fixed and distinct fields of views (e.g., no stereoscopic vision) that continuously record videos in long-term time periods (e.g., several years), and that are deployed in natural habitats without any device to attract or repel specific populations (e.g., no bait). The types of camera may vary (e.g., lenses, frame rates) and we consider their impacts in terms of fields of view and image quality.

4.2 Uncertainty factors

This section describes key uncertainty factors that arise from i) the computer vision system (Section 4.2.1); ii) the system's deployment conditions (Section 4.2.2); and iii) both the computer vision system and its deployment conditions (Section 4.2.3). Overall, 12 key uncertainty factors are identified, as summarized in Table 4.1.

4.2.1 Uncertainty factors from the computer vision system

Within the computer vision domain, uncertainty factors are often investigated from the perspective of the underlying algorithms, focusing on uncertainties specific to particular machine learning techniques (Csurka et al. 1997, Zhu and Wu 2004, Spampinato et al. 2012, Senge et al. 2014). Here we consider the algorithms as black boxes and focus on higher-level uncertainty, i.e., the uncertainty in the class sizes provided to end-users.

The computer vision system may produce 4 types of high-level errors:

- **Object Detection Errors** concern the erroneous detections of individuals in each video frame, i.e., undetected individuals (False Negatives) and other objects identified as individuals of interest (False Positives).
- **Tracking Errors** concern the misidentification of individuals' trajectories across multiple frames, i.e., splitting, merging or intertwining trajectories of different individuals (Spampinato et al. 2012).
- **Species Recognition Errors** concern individuals that are classified into a species they do not actually belong to.
- **Behavior Recognition Errors** concern individuals that are classified into a behavior they are not actually exhibiting.

The image quality of video samples impacts the appearance of objects, and thus the visual features extracted by computer vision algorithms and used to recognize animals, species and behaviors. Hence **Image Quality** has a direct impact on the 4 types of computer vision errors we consider.

Computer vision algorithms use groundtruth training sets to learn to detect individuals, species or behaviors, but also to track individuals and to detect image quality. Groundtruth is typically manually annotated by experts, but is often crowdsourced by non-experts (He et al. 2013). Hence **Groundtruth Quality** is essential to control the errors in computer vision results. Scarcity, unrepresentative views of objects, unrepresentative image quality, or labelling errors in groundtruth may yield error-prone computer vision software.

4.2.2 Uncertainty factors from the in-situ system deployment

This source of uncertainty is usually not in the scope of evaluations performed in the computer vision and classification domains. Evaluations of computer vision and classification algorithms are intended to be valid for most applications, and are abstracted from case-specific application conditions. However, errors and biases in the algorithms' results can be significantly influenced by several uncertainty factors arising from the application conditions.

Time-varying environmental conditions (e.g., lighting, turbidity, biofouling) or camera features (e.g., lens, resolution) can lower the **Image Quality**. The placement of cameras and their **Field of View** can target specific habitats. Thus the Fields of View can under-represent species living in other habitats, or over-represent animal behaviors occurring in these habitats. The Fields of view can also modify the chances of **Duplicated Individuals** (e.g., targeting a feeding zone may increase the number of individuals moving back and forth, thus in and out the field of view), and the chances of obtaining low *Image Quality* (e.g., in shade- or turbidity-prone locations). The numbers of cameras may not provide sufficient **Sampling Coverage**. Finally, computational issues with the servers executing the computer vision algorithms can yield **Fragmentary Processing** (e.g., missing videos).

Factor	*Description*
Uncertainty factors due to computer vision system (Section 4.2.1)	
Groundtruth Quality	Groundtruth items may be scarce, represent the wrong animals, odd animal appearances (i.e., odd feature distributions).
Object Detection Errors	Some individuals may be undetected, and other objects may be erroneously detected as individuals of interest.
Tracking Errors	Trajectories of individuals tracked over video frames may be split, merged or intertwined.
Species Recognition Errors	Some species may not be recognized, or confused with another.
Behavior Recognition Errors	Some behaviors may not be recognized, or confused with another.
Uncertainty factors due to in-situ system deployment (Section 4.2.2)	
Field of View	Cameras may observe heterogeneous ecosystems, and over- or under-represent species, behaviors or objects features. Fields of view may be partially or totally occluded, cover heterogeneous area sizes, and shift from their intended position.
Fragmentary Processing	Some videos may be yet unprocessed, missing, or unusable (e.g., encoding errors).
Duplicated Individuals	Individuals moving back and forth are repeatedly recorded. Rates of duplication vary among species behaviors and *Fields of view*.
Sampling Coverage	The numbers of video samples may not suffice for end-results to be statistically representative.
Uncertainty factor due to both system and in-situ system deployment (Section 4.2.3)	
Image Quality	Lighting, water turbidity, contrast, resolution or fuzziness may impact the magnitude of computer vision errors.
Noise & Bias	Computer vision errors may be random (noise) or systematic (bias). Biases may emerge from a combination of factors (*Image Quality*, *Field of View*, *Duplicated Individuals*, *Object Detection Errors*, *Species* & *Behavior Recognition Errors*). Additional biases arise from *Duplicated Individuals* and heterogeneous *Fields of View*.
Uncertainty in Specific Datasets	Uncertainty in specific sets of computer vision results depend on the specific characteristics of the datasets (e.g., distribution of image quality) which impact the magnitude of *Noise and Bias*.

Table 4.1: Key uncertainty factors in computer vision systems for population monitoring.

4.2.3 Uncertainty factors from both system and in-situ deployment

Image Quality is a factor of uncertainty that impacts the computer vision algorithms, and that is impacted by the in-situ deployment conditions. Beside image quality, we identified two other uncertainty factors arising from both the computer vision system and the deployment conditions. Assessing these high-level uncertainty factors is necessary for conveying the uncertainty propagation to end-users.

When analysing class sizes, ecologists are concerned with differentiating stochastic errors (noise) from systematic errors (bias). Such **Noise and Bias** arise from a combination of factors that may yield class size estimates that are lower or higher than their true values. Errors from the computer vision algorithms (*Object Detection, Tacking, Species Recognition* and *Behavior Recognition Errors*) may yield class sizes that over- or under-estimate specific populations. For example, two similar species can be often confused for one another. Species appearing at dawn and dusk, where dim natural light degrades the image quality, have higher chances of being misclassified. Such under- or over-estimation of class sizes may be random (yielding noise) or systematic (yielding biases).

The levels of *Noise and Bias* may differ depending on the specific subsets of computer vision data. The chances of computer vision errors may vary, e.g., depending on the image quality or the object features in the data subset. The placement of cameras may create additional biases. The *Fields of View*, *Duplicated Individuals* and *Sampling Coverage* modify the chances that specific species or behaviors appear on the videos. For example, *Fields of View* observing the open sea, with no foreground coral head, are not likely to collect samples of species or behaviors that usually occur on specific coral heads. Hence the specific cameras from which the data subset is collect impact the chances of over- or under-estimating the population sizes.

Thus for deriving the **Uncertainty in Specific Datasets**, end-users must account for the specific characteristics of each dataset. They need to assess:

- The proportion of *Image Quality* in the dataset, e.g., to infer the magnitude of computer vision errors given the errors measured with groundtruth test sets from each image quality.
- How the *Fields of View* impact the chances of *Duplicated Individuals* and the completeness of *Sampling Coverage*, as these potentially under- or over-estimate some species or behaviors.
- How *Fragmentary Processing* of the video samples impact the *Sampling Coverage*.

4.3 Uncertainty propagation

The uncertainty factors interact with each other, yielding a complex scheme of uncertainty propagation (Figure 4.2). We describe these interactions (Section 4.3.1) and discuss their impact on the high-level information provided to end-users (Section 4.3.2).

4.3.1 Interactions between uncertainty factors

Each computer vision algorithm is impacted by the errors of the algorithms previously applied. In systems such as the Fish4Knowledge system (Figure 4.1) *Object Detection Errors* impact *Tracking Errors* as missing individuals (False Negatives) and other objects (False Positives) can yield erroneous interpretations of trajectories.

Species Recognition Errors are impacted by both *Object Detection* and *Tracking Errors*, as False Positives (e.g., non-fish objects) may be attributed a species, and species recognition suffers from intertwined trajectories merging individuals from different species. *Behavior Recognition Errors* are impacted by *Species Recognition Errors* as behavior features are species-specific (e.g., one speed indicates predator/prey behaviors for one species, but is a neutral movement for another).

The *Fields of View* impact the kind of ecosystems observed by each camera. It also impacts the chances of *Duplicated Individuals*, e.g., observing coral heads is more likely to yield overestimation of sedentary species than observing the open sea. The depth of *Field of View* impacts the size of the monitored areas, hence the *Sampling Coverage*. The initial *Sampling Coverage* of the set of cameras can be reduced by the *Fragmentary Processing* of the videos, i.e., due to unprocessed or missing videos.

The depth of *Field of View* further impacts the *Image Quality* as resolution and fuzziness are poorer for distant backgrounds than foregrounds. *Image Quality* is further impacted by the *Field of View* as some cameras may be placed in area where low light, turbidity or bio-fouling are more likely to occur. Different types of *Image Quality* can yield different levels of *Object Detection Errors*, *Species Recognition Errors* and *Behavior Recognition Errors*, and thus potential *Noise and Biases*. Hence the *Fields of View* can under- or over-represent species, behaviors and ranges of image quality, thus influencing the potential *Noise and Biases*.

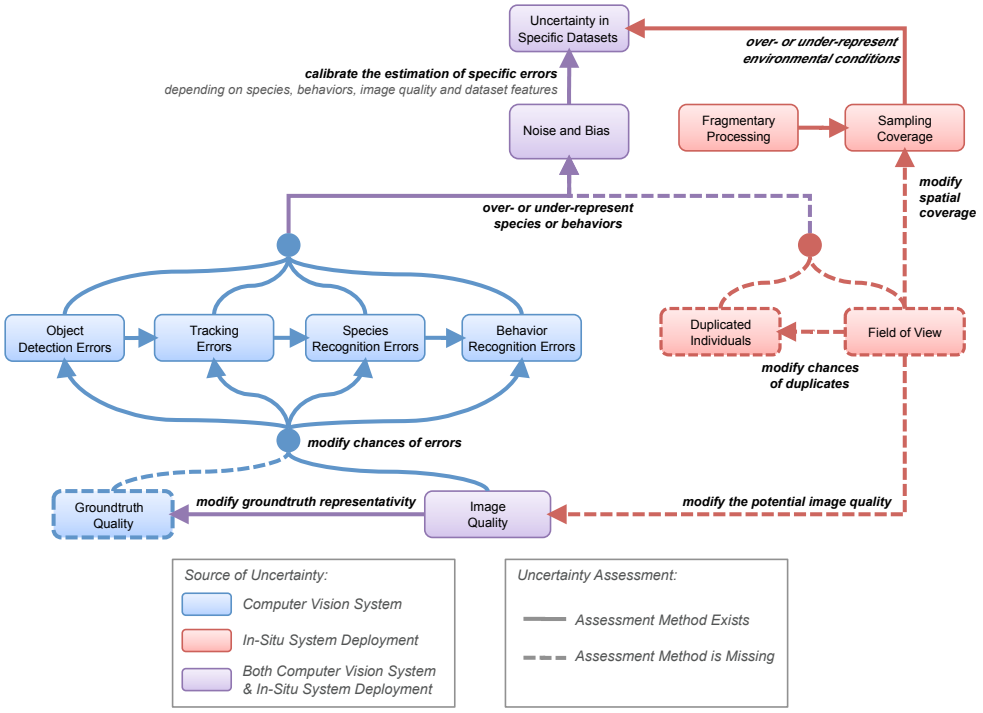

Figure 4.2: Interactions among the uncertainty factors in Table 4.1

The *Groundtruth Quality* depends on how image samples are representative of the possible *Image Quality*. The groundtruth needs to contain samples of the possible object appearances (e.g., different angles), but also samples that represent the variations of object appearances depending on image quality and low-level image features (e.g., variability of shapes or colors).

4.3.2 High-level impact

The interactions between uncertainty factors propagates uncertainty to the high-level information provided to end-users, i.e., the class size estimates. The class size estimates may not be representative of the actual population sizes in the ecosystem. We discuss how the sampling method (Section A) or the computer vision errors (Section B) can both yield unrepresentative class sizes.

A. Sampling validity

Inappropriate sampling methods can yield class sizes that are not representative of the ecosystem of interest, even if the computer vision system makes no error. For example, ecologists might seek to study the relative species distribution (e.g., which species are dominant or rare) while the cameras observe habitats where some species are not likely to occur. Further, the videos may be sampled in time periods where some species or behaviors are likely to occur, and others are not (e.g., depending on daily cycles of species behaviors). Such inappropriate spatio-temporal coverage can under- or over-estimate specific populations. Additionally, individuals from sedentary species may be repeatedly observed as they swim in and out of the fields of view (i.e., over-estimation). Finally, too few video samples impact the statistical validity of the observed class sizes, e.g., the findings on population sizes may not be generalizable.

B. Computer vision errors

Computer vision errors can yield class size estimates that differ from the actual content of the videos. Species and behaviors can be over- or under-estimated, randomly or systematically, as the propagation of computer vision errors result in *Noise and Bias* in class size estimates.

For example, for a particular class, the magnitude of computer vision errors can be random, yielding noisy class sizes misestimated by +/- 10% with average number of errors close to 0. For another class, the magnitude of of computer vision errors can be systematic, yielding biased class sizes over-estimated by +10% on average.

The magnitude of biases depends on the classes of objects that co-occur in the videos (e.g., class A is over-estimated when class B also occurs, as class B objects are often misclassified as class A) and on the quality of images and object appearances (e.g., class sizes are under-estimated due to unrecognised blurry or occluded objects). Hence uncertainty due to the *Fields of View* and *Duplicated Individuals* propagates to

higher-level *Noise and Bias* in class sizes, as they modify the chances of observing specific species, behaviors, image quality, and object viewpoints such as occluded objects.

Finally, the *Noise and Biases* due to computer vision errors propagates to the *Uncertainty in Specific Datasets*. The magnitudes of noise and bias is specific to each set of video samples, as it depends on the species, behaviors and and image quality occurring in the videos.

4.4 Uncertainty assessment methods

We investigate how to assess the combined impact of uncertainty factors on the high-level population sizes estimated by computer vision systems. We review how uncertainty assessment methods address the uncertainty factors and uncertainty propagation of concern to end-users. This review, synthesized in Figure 4.2, allows to identify uncertainty issues unaddressed in the literature and requiring future research. For instance, uncertainty assessment methods may address system engineers' concerns rather than end-users' concerns, e.g., by assessing individual system components in isolation, thus not addressing the uncertainty propagating to and from the components.

We focus on uncertainty related to the information processing techniques of computer vision systems. Uncertainty related to sampling techniques is excluded because it depends on the specificity of ecosystems (e.g., the 3-dimensional land topology) and on the related sampling strategies (e.g., stratification may be required). We discuss how to measure computer vision errors (Section 4.4.1) and how to measure the impact of in-situ deployment conditions, i.e., how camera setup modifies the chances of computer vision errors (Section 4.4.2).

4.4.1 Measuring computer vision errors

We review uncertainty assessment methods that can assess the computer vision errors in end-results. We highlight that assessment methods do not directly address the uncertainty propagation in pipelines of classification components, nor the impact of groundtruth uncertainty.

Tacking errors - The computer vision algorithms we consider are primarily classification algorithms except for tracking algorithms that identify single individuals across several video frames. Tracking algorithms have specific error metrics, such as rates of correct tracking from one frame to another, or rates of incorrect individuals within single trajectories (Spampinato et al. 2012). These metrics are excluded from our scope because the user task of analysing class sizes does not directly concern analysing trajectories. The impact of tracking errors on class sizes must be considered, but in terms of classification errors and numbers of errors in class sizes, rather than numbers of errors within individual trajectories.

Groundtruth quality - Classification errors are typically measured by using groundtruth *test sets*, e.g., sets of items that are manually classified. Manual and automatic classifications are compared, typically by using confusion matrices. Each classifier is usually evaluated separately, using a specific test set independent of other classifiers' test sets.

This well-established approach relies on the assumption that groundtruth test sets do not contain any errors. However, in practice groundtruth datasets are manually classified and humans can make errors, ambiguous objects may not be identifiable with full certainty (e.g., in fuzzy images). Existing assessment methods, such as Cohen's kappa, can measure the *agreement* between the humans that produced the groundtruth (e.g., *agreement* occur when humans classify the same item in the same class). The lower the agreement, the higher the chances of error in the groundtruth. Such approach assesses the *Groundtruth Quality*, however, it does not estimate the number of errors in the groundtruth.

Future work is needed to estimate the number of errors in groundtruth datasets, e.g., using the *agreement* measures. Such error estimation is required to refine the measurements of classification errors, and account for the potential errors in groundtruth test sets. For example, an object classification may be correct but evaluated as an error because the groundtruth test set contains an error, and assigns the wrong class to that object.

Uncertainty propagation - Class sizes obtained through a pipeline of classification algorithms are impacted by the combined errors of each classifier. For measuring the classification errors that propagate in the pipeline of classifiers into the class size estimates, the test sets used to evaluate each classifier must be must be representative of the potential errors of the previous classifiers. For example, *Object Detection Errors* can be measured after tracking is performed, rather than before. Errors can be measured for each object trajectory, rather than for each object occurrence in individual video frames. Such *Object Detection Errors* should be measured with test sets that consists of the results of the previous algorithms that *segment*, *detect* and *track* individuals in each video frame. The test sets should include examples of segmentation and tracking errors. However, some object trajectories may be ambiguous, e.g., if half of trajectory's images are fish and the other half are non-fish. Such example of tracking errors should be included in the test set, but they are difficult to label as error or not. Ideally such ambiguous trajectories must remain very rare in the tracking results.

To continue assessing uncertainty propagation with a consistent test set, the *Species Recognition Errors* should be measured with a test set that is representative of the *Object Detection Errors*. For example, such test set should include examples of False Positives objects (e.g., trajectories of non-fish objects detected as fish), trajectories containing False Positives, and trajectories containing individuals from different species.

Test sets that represent the errors of previous classifiers can be difficult to collect. Examples of computer vision errors can be difficult to label, e.g., low-quality

images are also difficult for human to recognize, and trajectories containing many tracking errors may warrant no clear species label. Furthermore, measuring the errors that propagate from the previous algorithms' errors can require additional classes that represent the errors from previous algorithms.

For example, estimating *Species Recognition Errors* would require only one additional class to represent the False Positives from *Object Detection Errors* (e.g., non-fish objects). This additional class allows to measures the *Species Recognition Errors* arising from the False Positives in *Object Detection Errors*. Estimating *Behavior Recognition Errors* would require many additional classes: one class for each possible species misclassifications (e.g., items can be from Species A and misclassified as Species B, thus increasing the chances of misclassifying the behaviours).

On top of representing the combined classification errors, the test sets should also represent the potential *Image Quality*. It is difficult to collect examples of all possible combinations of classification errors and image quality, and the resulting confusion matrices can be difficult to analyse by end-users.

Classification noise and bias - Confusion matrices do not easily convey the uncertainty in specific class size estimates. Confusion matrices can have many cells (i.e., n^2 cells for classifications into n classes). End-users need to analysis all cells, and associate them row-wise and column-wise, which can be tedious and error-prone. For example, to derive the errors in a class size, end-users need to read the n cells within the same row or column, and sum them to derive the total number of errors. It is thus complex to estimate the *Noise and Bias* due to classification errors, e.g., to assess how class sizes are over- or under-estimated. Simplified visualization tools for assessing the potential *Noise and Biases* due to classification errors are thus addressed in Chapter 6.

Resulting class size uncertainty - It is complex to derive the *Uncertainty in Specific Datasets*, i.e., the classification errors in specific class size estimates. For instance, the uncertainty in specific class size estimates is not directly conveyed by confusion matrices.

The errors measured in test sets can differ from the errors in specific end-usage datasets, called *target sets*. For instance, the class distribution, i.e., the relative class sizes, can differ between test and target sets. This impacts the magnitude of classification biases. For example, if Species A is more prevalent in the target set, it yields more misclassifications between Species A and other species, thus different magnitudes of biases. Methods for assessing *Errors in Specific Datasets*, arising from *Noise and Biases* due to classification errors, are addressed in Chapter 5.

4.4.2 Measuring the impact of deployment conditions

The literature does not offer well-established methods for assessing the impact of uncertainty arising from the conditions in which computer vision systems are deployed. Computer vision research usually focuses on generic uncertainty assessment abstracted from specific application conditions. We highlight that methods for assess-

ing the biases arising from *Duplicated Individuals* and heterogeneous *Fields of View* are largely unaddressed. However, we identify methods for dealing with *Fragmentary Processing*.

Duplicated individuals - Future work is needed to develop methods for measuring *Duplicated Individuals*, e.g., depending on species, behaviors and *Fields of View*. Such measurements are required for assessing over-estimations of species that often move in and out certain fields of view. Such measurements should also account for schooling behaviors (i.e., swimming in group) where individuals can be duplicated as well as occluded.

It can be difficult to collect groundtruth data to assess *Duplicated Individuals*. It is difficult for humans to identify single individuals swimming in and out of the fields of view, and to estimate the total number of individuals in a group. Diving observations on cameras' site may provide groundtruth data, as the total sizes of fish groups can be estimated by experienced divers. However, divers can make mistake, and they interfere with the natural environment thus observing different fish behaviors.

Fields of view - Uncertainty assessment methods are also missing for issues with shifting *Fields of View*. For systems using fixed cameras, the fields of view may gradually vary over time, e.g., due to typhoons, strong currents, or maintenance operations such as lens cleaning. Detecting and measuring the shifts of fields of view is not sufficient to estimate the resulting uncertainty: the changes in the *Sampling Coverage* must also be measured (e.g., the sizes and types of areas within the fields of view).

Fragmentary processing - Assessments methods for handling uncertainty due to *Fragmentary Processing* (e.g., missing videos) are easier to establish given that video samples are of equal duration. Such assessment methods can rely on counting the numbers of video samples, and estimating the class sizes per video sample. We recommend using video samples of equal duration, and of duration that is long enough to avoid too many split trajectories at the beginnings and ends of the samples (e.g., 10 minutes for the Fish4Knowledge project). Otherwise, handling *Fragmentary Processing* is more complicated.

Class sizes can be drawn from different numbers of video samples, making comparisons difficult (e.g., increasing class sizes can be due to increasing numbers of samples, or to actual increases of population sizes). To compare class sizes drawn from different numbers of video samples, we first consider the case of video samples collected from the same camera. We propose to use *mean class sizes per video sample* estimated with equation (4.1). Mean class sizes can be compared, e.g., over time periods, even if drawn from different numbers of video samples. However, mean class sizes drawn from scarce samples are less representative of actual population sizes. Such uncertainty with the temporal *Sampling Coverage* can be assessed by computing the *variance of mean class sizes* with equation (4.2).

$$\overline{C}_k = \frac{C_k}{N_k}$$

$\overline{C}_k$: Mean class size per video sample for class c observed from camera k
C_k : Number of individuals classified in class c (class size) observed from camera k
N_k : Number of video samples collected from camera k (4.1)

$$V\left(\overline{C}_k\right) = \sum_{i=1}^{V_k} \frac{\left(\overline{C}_k - C_{kt}\right)^2}{N_k}$$

$V\left(\overline{C}_k\right)$: Variance of mean class size $\overline{C}_k$ (4.1) for class c observed from camera k.
N_k : Number of video samples collected from camera k
C_{kt} : Number of individuals classified in class c observed from camera k in a single video sample t (i.e., representing time unit t)

(4.2)

To analyse class sizes drawn from different cameras, *mean class sizes per video sample* must be estimated with equation (4.3). It would be incorret to divide the class sizes by the total number of video samples for all cameras, i.e., $\sum_k C_k / \sum_k N_k$. For example, with 2 video samples recorded simultaneously from different cameras, and observing 100 and 50 fish, the total number of fish that occurred during this time period is 150, not 150/2. This approach assumes that cameras have no overlapping fields of view, and are placed sufficiently far away from each others, so that the same individuals are not recorded several times by different cameras. If these assumptions are violated, a different approach must be considered.

$$\overline{C} = \sum_k \overline{C}_k$$

C : Mean class size per video sample for class c observed from all cameras
$\overline{C}_k$: Mean class size per video sample (4.1) for class c observed from camera k

(4.3)

Estimating the variance of mean class sizes over several cameras (4.3) can be difficult. As sums of random variables, their variance is given by equations (4.4). It requires estimating the covariance between mean class sizes $\overline{C}_k$ (4.2) drawn from different cameras k. We consider that mean class sizes $\overline{C}_k$ and $\overline{C}_{k'}$ covary along individual time units t, and specify their covariances in equation (4.5).

$$V\left(\overline{C}\right) = \sum_k V\left(\overline{C}_k\right) + \sum_k \sum_{k' \neq k} Cov\left(\overline{C}_k, \overline{C}_{k'}\right)$$

$V\left(\overline{C}\right)$: Variance of mean class size $\overline{C}$ (4.3) for class c observed from all cameras
$V\left(\overline{C}_k\right)$: Variance of mean class size $\overline{C}_k$ (4.2) for class c observed from camera k
$Cov\left(\overline{C}_k, \overline{C}_{k'}\right)$: Covariance of mean class sizes $\overline{C}_k$, $\overline{C}_k$ for class c observed from cameras k, k'

(4.4)

$$Cov\left(\overline{C}_k, \overline{C}_{k'}\right) = \sum_t \frac{\left(\overline{C}_k - C_{kt}\right)\left(\overline{C}_{k'} - C_{k't}\right)}{N_t}$$

$\overline{C}_k$: Mean class size (4.2) for camera k
$\overline{C}_{k'}$: Mean class size (4.2) for camera k'
C_{kt} : Class size for camera k and time unit t
$C_{k't}$: Class size for camera k' and time unit t (i.e., from the video sample corresponding to time unit t)
N_t : Number of time units t from which class sizes are drawn

(4.5)

Such approach assumes that all video samples are recorded simultaneously over time units t of equal duration, and available for all cameras k and all time units t. For example, the Fish4Knowledge system uses 10-minute time units, and all video samples are recorded over the same time units. For instance, video samples are all recorded from 08:00 to 08:10, then 08:10 to 08:20, and so on. No video sample is recorded from 08:05 to 08:15. If one cameras provides a video sample for a time unit t (e.g., 08:00 to 08:10 on Jan. 1st 2012) then all cameras must provide a video sample for that time unit. Otherwise, equation (4.5) cannot be computed.

Such issues with missing video samples can be addressed with imputation methods. Alternative methods exist and must be chosen depending on the application requirements (Little and Rubin 2014). However, these methods do not address heterogeneous time units, i.e., video samples having differemt beginning and end times (e.g., some videos are recorded from 08:00 to 08:10, and others from 08:05 to 08:15).

4.5 Conclusion

This chapter synthesized uncertainty issues of concerns to end-users, which we identified in Chapters 2 and 3. The remainder of this thesis addresses uncertainty factors that are introduced in this chapter: methods to estimates numbers of classification errors in specific datasets (*Uncertainty in Specific Datasets*, Chapter 5), simplified visualizations for communicating classification errors and biases to non-expert end-users (classification *Noise and Bias*, Chapter 6), and user interface for assessing the multiple uncertainty factors (Chapter 7).

We conclude this chapter by highlighting how uncertainty factors impact end-users' tasks (Section 4.5.1) and the uncertainty assessment methods of interest to end-users (Section 4.5.2).

4.5.1 Impacts of uncertainty factors

We introduced a model of key uncertainty factors pertaining to computer vision system for population monitoring. The model provides foundations for assessing class size uncertainty from the perspective of end-users, and answers our third research

question: *When applying computer vision systems for population monitoring, what uncertainty factors can arise from computer vision systems, and from the environment in which systems are deployed?*

Uncertainty factors arising from the computer vision system result in classification errors that can be random (yielding noise) or systematic (yielding biases). Classification bias threatens the validity of the resulting estimations of population sizes, and trends in population sizes. Classification biases may misrepresent the proportions of the different populations (e.g., species composition) and yield deceptive increases or decreases of population sizes.

Image quality (e.g., blurry images yield more errors) and groundtruth quality (e.g., unrepresentative groundtruth yield more errors) are also of concern. For instance, within the Fish4Knowledge project, the low contrast of images recorded at dusk can increase the number of *Species Recognition Errors* as fish colors are not distinguishable while being an important discrimination factor between species of similar body shapes.

Uncertainty factors arising from the deployment conditions can have significant impacts on the population estimates. Besides impacting the sampling validity, and the statistical validity of class size estimates, the deployment conditions can significantly impact the noise and biases in class size estimates. For instance, the cameras' fields of view modify the chances that class sizes are over- or under-estimated. The fields of view impact the image quality, thus the magnitudes of computer vision errors. The fields of view also impact the chances to repeatedly detect the same individuals, thus the over-estimation of specific species (e.g., living or feeding within the fields of view).

4.5.2 User-oriented assessment methods

End-users require assessments of the uncertainty that results from the multiple uncertainty factors. This chapter briefly reviewed the applicable assessment methods, and partially answered our fourth research question: *How uncertainty assessment methods address the combined effect of uncertainty factors?*

Unaddressed factors - Three key uncertainty factors are not addressed in the literature: the heterogeneity and shifts of camera's fields of view (e.g., impacting the sampling validity), the impact of duplicates (e.g., individuals that are repeatedly detected can greatly over-estimate specific populations depending on their behaviors), and the errors in groundtruth datasets. Methods to assess groundtruth quality are available but do assess the uncertainty propagating to the classification results, e.g., to the class size estimates.

Existing methods - Classification errors can be assessed with well-establish assessment methods. However, these methods do not assess the uncertainty that prop-

agates along pipelines of classifiers. Nor do they estimate the errors in classification end-results, but in test sets only. Methods to infer the numbers of classification errors in end-results are developed in Chapter 5. To assess uncertainty propagation along pipelines of classifiers, we propose to use test sets that represent the errors of the previous classifiers, and include additional classes to represent the propagated errors. However such an approach can be challenging in practice, as extensive groundtruth and many additional classes may be needed.

Finally, assessing the impact of varying numbers of video samples (e.g., fragmentary processing) can rely on computing averages and variances of population sizes, as in equations (4.1)-(4.5). This approach is compatible with the methods introduced in Chapter 5, which can refine the population sizes that are averaged in equations (4.1)-(4.5).

Chapter 5

Estimating Classification Errors

Classification errors can yield biased class sizes, and threaten the validity of class size estimates. For instance, class size estimates can be systematically over- or under-estimated due to systematic confusions of specific classes (Chapter 4, Section 4.2.3, p.63). Well-established methods can assess classification errors by using groundtruth test sets, and measuring error rates such as Precision, TP Rate, Accuracy, or F-measures. These methods do not estimate classification errors in end-results but in test sets only (Chapter 4, Section 4.4.1, p.69). However, end-users require estimations of classification errors in end-results, as estimating potential classification errors and biases is required to establish informed interpretation of classification data (Chapter 3, Section 3.5.1, p.52).

This chapter investigates methods for estimating classification errors in end-results by using error measurements from test sets. Our results address requirement 4-c in Chapter 2 (*Assess uncertainty in specific datasets*, p.36) and answer our fifth research question: *How can we estimate the magnitudes of classification errors in end-results?*

After introducing the problems we address (Section 5.2), we review **existing methods** for estimating unbiased class sizes (Section 5.2). We then provide additional methods for:

- Estimating the **number of errors between specific classes** (Section 5.3)
- **Estimating the variance** of error estimation results (Section 5.4)
- **Predicting the variance** of error estimation results, i.e., before classifiers are applied (Section 5.5)

We discuss the applicability of these error estimation methods (Section 5.6) and research directions to refine them (Section 5.7). Finally, we underline higher-level implications of our findings (Section 5.8). For instance, variance and bias issues in error estimation problems also concern classifier assessment problems. If estimating classification errors is uncertain, so is assessing classifiers' suitability for end-users' task.

5.1 Introduction

The statistics and epidemiology domains devised *bias correction* methods that can estimate classification biases in specific sets of end-results (Tenenbein 1972, Grassia and Sundberg 1982, Shieh 2009, Buonaccorsi 2010). Given estimates of the classification errors, i.e., drawn from test sets, unbiased class sizes can be derived. This approach does not identify which individual items are misclassified.

These bias correction methods are applicable to machine learning classifiers, but are seldom considered except for land coverage estimation (Card 1982, Hay 1988, van Deusen 1996, Foody 2002). However, bias correction methods are of interest for a large range of use cases, e.g., for analysing class sizes, class probabilities and class distributions. Without estimating potential classification biases, e.g., with bias correction methods, no scientific conclusion can be drawn from classification data.

We investigate the application of existing bias correction methods to classification problems with machine learning software (i.e., to assess the *Uncertainty in Specific Datasets*, Chapter 4, p.69). We show that these methods can **reduce biases in class size estimates**. However, we highlight cases where the bias correction methods can yield high result variance or increased biases (Section 5.2).

We extend the application of bias correction methods to estimating numbers of errors in classification results, i.e., **detailing the error composition**, for instance, within the items misclassified as class y how many truly belong to class x. Estimating the error composition describes the quality of classification data beyond accuracy or precision. We introduce an alternative method for estimating the error composition, called **Ratio-to-TP method**. It provides exactly the same result as one extended bias correction method, but has properties of interest (Section 5.3).

We show that the **variance of error estimation results** can be critical and is crucial to estimate. For instance, with small datasets the variance magnitude can exceed the bias magnitude, thus applying error estimation methods may worsen the initial biases.

Variance estimation methods exist for uses cases where test sets are randomly sampled within classification end-results (Tenenbein 1972). However, machine learning classifiers are usually evaluated using test sets that are distinct from the end-usage datasets to which classifiers are applied, called *target sets*. For instance, all potential target sets may not be known when classifiers are evaluated. For disjoint test and target sets, existing variance estimation methods describe the overall population from which test and target sets are sampled (Grassia and Sundberg 1982, Shieh 2009, Buonaccorsi 2010, van Deusen 1996). If applied to describing the target set itself, they provide biased estimates.

We thus introduce the **Sample-to-Sample method** that addresses the case of disjoint test and target sets, and estimates the variance of error estimation results that pertain to specific target sets. The Sample-to-Sample method estimates the variance at the level of the error rate estimator, which must account for the class sizes in both test and target sets. From error rates' variance estimates, we derive well-bounded confidence intervals for the error estimation results in binary problems. Multiclass

problems are more complex to formalize algebraically, but may be addressed with bootstrapping or simulation (Section 5.4).

End-users may prefer classifiers that minimize the variance of error estimation results. However, **predicting the variance of error estimation results** is difficult when the characteristics of potential target sets are unknown (e.g., the class sizes). To address this case, we postulate that the determinant of error rate matrices can predict the variance of error estimation results. We derive the **Maximum Determinant method** for predicting which classifier yields the least variance when applying error estimation methods, without knowledge of the potential target sets. Initial results are promising but future research is needed to establish theory, e.g., to specify the effects of class sizes and proportions, number of classes, and error rate magnitudes (Section 5.5).

The methods presented in this chapter rely on the assumption that error rates do not vary *systematically* between test and target sets, but may vary *randomly*. If **feature distributions** (e.g., class models) differ between test and target sets, bias ensues. We illustrate this domain adaption problem and its critical impact on error estimation results. However, domain adaptation problems that concern shifts in *class prior probability* can be addressed with the methods presented in this chapter (Section 5.6).

The methods presented in this chapter are demonstrated empirically, with real and synthetic data. We discuss the need for establishing theory and guidelines for choosing the methods to apply depending on the characteristics of test and target sets (Section 5.8).

5.2 Existing bias correction methods

As introduced in Section , the statistics and epidemiology domains devised *bias correction* methods that can estimate unbiased class sizes (Tenenbein 1972, Grassia and Sundberg 1982, Shieh 2009, Buonaccorsi 2010). Unbiased class sizes, or class proportions[1], can be estimated without identifying which individual items are misclassified. These bias correction methods address end-users' need for assessing the *Uncertainty in specific datasets* due to classification *Noise and Bias* (Chapter 4, Section 4.4.1, p.69).

Bias correction methods are based on error rates measured in **test sets**, i.e., sets of items whose actual class is known (also called groundtruth, gold standard, validation or calibration set). The error rates are assumed to be the same in *target sets*, i.e., for the datasets to which classifiers are applied in practice (also called unlabelled, real-life or end-usage data).

Two bias correction methods exist:

- The **Reclassification method** (Buonaccorsi 2010), also called inverse calibration (Katila 2006), ratio method (Hay 1988) or double sampling (Tenenbein 1972). It requires equal class proportions in test and target sets (Section 5.2.1).

[1] Class size divided by total number of items to classify, also considered as class probability.

- The **Misclassification method** (Buonaccorsi 2010), also called classical calibration (Katila 2006), matrix inversion method (Hay 1989), or PERLE (Beauxis-Aussalet and Hardman 2015)[2]. It is robust to varying class proportions (Section 5.2.2).

The Misclassification method yields a larger results variance than the Reclassification method, as noted by Shieh (2009) and shown in Figure 5.1 (p.81). Thus, when possible, it is preferable to use the Reclassification method, using test sets which class proportions are similar to the target set. However, this is often impossible in machine learning problems, as class proportions may vary over target sets or are unknown when test sets are collected.

We specify the two error estimation methods using the notation in Table 5.1 and the following variables:

n_{xy} : Number of items that actually belong to class x and are classified as class y
$n_{x.}$: Actual class size for class x (i.e., number of items that actually belong to class x)
$n_{.x}$: Output class size for class x (i.e., class size estimated by the classifier)
$n_{..}$: Total number of items in the dataset

		Actual Class				**Estimated Class Size**
		class 1	**class** 2	...	**class** x	
Output Class	**class** 1	n_{11}	n_{21}	...	n_{x1}	$n_{.1}$
	class 2	n_{12}	n_{22}	...	n_{x2}	$n_{.2}$
	...	...	...	...	...	...
	class x	n_{1x}	n_{2x}	...	n_{xx}	$n_{.x}$
Actual Class Size		$n_{1.}$	$n_{2.}$	...	$n_{x.}$	Total $n_{..}$

Table 5.1: Confusion matrix and notation.

The variables for the target set are denoted with the prime symbol, to distinguish them from the variable related to the test set. With prime symbol, n' concerns the target set. Without prime, n concerns the test set. For example, $n_{1.}$ is the actual size of class1 in the test set, and $n'_{1.}$ the actual class size in the target set.

The existing error estimation methods estimate actual class sizes $n'_{x.}$ in target sets, given the known output class sizes $n'_{.x}$ and numbers of error n_{xy} measured in the test set. We present the error estimation methods in terms of class size estimates $n'_{x.}$ rather than class proportions $n'_{x.}/n'_{..}$ as in the literature, the latter being easily derived from the former.

5.2.1 Reclassification method

The Reclassification method is based on error rates that use output class sizes $n_{.y}$ as denominators, e.g., precision in binary problems. Assuming equal error rates in test

[2]We introduced the PERLE method in this former publication, at a time we had without prior knowledge of similar prior work. Hence the PERLE method was incorrectly introduced as a new method.

and target sets (i.e., $\widehat{e'_{xy}} = e_{xy}$), actual class sizes are estimated with equation (5.1). This assumption is violated, and the method is not applicable, if class proportions differ between test and target sets (Section 5.2.4).

$$e_{xy} = \frac{n_{xy}}{n_{.y}} \qquad \widehat{n'_{xy}} = e_{xy} n'_{.y} \qquad \widehat{n'_{x.}} = \sum_y e_{xy} n'_{.y} \tag{5.1}$$

Variance estimates $V\left(\widehat{n'_{x.}}\right)$ are provided by Tenenbein (1972) for test sets randomly sampled within target sets, using a weighted sum to account for the sample size of both test and target sets.

5.2.2 Misclassification method

The Misclassification method is based on error rates that use actual class size $n_{x.}$ as denominator, e.g., recall in binary problems (5.2). Assuming equal error rates in test and target sets (i.e., $\widehat{\theta'_{xy}} = \theta_{xy}$), actual class sizes are estimated with equation (5.2), i.e., solving a system of linear equations (Beauxis-Aussalet and Hardman 2015).

$$\theta_{xy} = \frac{n_{xy}}{n_{x.}} \qquad \begin{pmatrix} \widehat{n'_{1.}} \\ \widehat{n'_{2.}} \\ \vdots \\ \widehat{n'_{x.}} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{x1} \\ \theta_{12} & \theta_{22} & \dots & \theta_{x2} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1x} & \theta_{2x} & \dots & \theta_{xx} \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \vdots \\ n'_{.x} \end{pmatrix} \tag{5.2}$$

Variance estimates $V\left(\widehat{n'_{x.}}\right)$ are provided by Grassia and Sundberg (1982) for test sets that are randomly sampled within target sets, and with similar class proportions $n_{x.}/n_{..} \simeq n'_{x.}/n'_{..}$. The case of disjoint test and target sets with different class proportions is addressed by Shieh (2009) and Buonaccorsi (2010) for estimating the characteristics of the overall populations from which both test and target sets are sampled.

5.2.3 Application

We demonstrate the applicability of bias correction methods to classification problems in the machine learning domain. We apply Reclassification and Misclassification methods to open-source datasets from the UCI repository. To demonstrate issues with results variance (Shieh 2009), we select datasets with smaller to larger class sizes. To demonstrate issues with class proportions Buonaccorsi (2010), we split the datasets into test and target sets of different class proportions (Table 5.2).

We randomly sample test sets of predefined sizes (Table 5.2), and consider the remaining items as target sets. We draw 100 random splits to show the variance and bias in the initial classification results, and in the error estimation results (Figure 5.1). To maximize the sizes of test and target sets, we do not select distinct training sets

but use 10-fold cross validation. We applied a common classification technique: a Naive Bayes classifier (from the Weka platform).

Dataset	**Test Set Size** $n_{x.}$	**Target Set Size** $n'_{x.}$
Iris	$n_{1.}$=25 $n_{2.}$=20 $n_{3.}$=30	$n'_{1.}$=25 $n'_{2.}$=30 $n'_{3.}$=20
Ionosphere	$n_{1.}$=63 $n_{0.}$=150	$n'_{1.}$=63 $n'_{0.}$=75
Segment	$n_{1.,3.,5.,7.}$=210 $n_{2.,4.,6.}$=110	$n'_{1.,3.,5.,7.}$=120 $n'_{2.,4.,6.}$=220
Ohscal	$n_{0.}$=471 $n_{1.}$=433 $n_{2.}$=124 $n_{3.}$=125 $n_{4.}$=275 $n_{5.}$=205 $n_{6.}$=738 $n_{7.}$=339 $n_{8.}$=490 $n_{9.}$=613	$n'_{0.}$-$n'_{9.}$=400
Waveform	$n_{1.}$=600 $n_{2.}$=900 $n_{3.}$=1200	$n'_{1.}$=1092 $n'_{2.}$=753 $n'_{3.}$=455
Chess	$n_{1.}$=1000 $n_{0.}$=500	$n'_{1.}$=669 $n'_{0.}$=1027

Table 5.2: Datasets used for experiments in Figure 5.1 Source: UCI Repository (*https://archive.ics.uci.edu/ml/datasets.html*).

The Reclassification method yields biased results (i.e., median results differ from actual class sizes) because class proportions differ between test and target sets. The Misclassification method is unbiased but yields larger variance than the Reclassification method.

5.2.4 Discussion

The Misclassification method is unaffected by changes in class proportions because its error rates θ_{xy} involve items belonging to the same true class, unlike the error rates e_{xy} of the Reclassification method, as shown in equations (5.3)- (5.4).

$$\text{Class proportions in binary problems:} \quad \frac{n'_{x.}}{n'_{..}} = \alpha \frac{n_{x.}}{n_{..}} \qquad \frac{n'_{y.}}{n'_{..}} = \beta \frac{n_{y.}}{n_{..}} \qquad \alpha, \beta \in \mathbb{R}_{<0}$$

$$\text{Assuming proportional errors:} \quad n'_{xy} = \alpha\, n_{xy} \quad \text{and} \quad n'_{yy} = \beta\, n_{yy}$$

$$\text{With unequal class proportions } \alpha \neq \beta\text{:} \quad n'_{.y} = n'_{xy} + n'_{yy} = \alpha n_{xy} + \beta n_{yy} \neq \alpha n_{.y}$$

$$\theta'_{xy} = \frac{\alpha n_{xy}}{\alpha n_{x.}} = \theta_{xy} \qquad e'_{xy} = \frac{\alpha n_{xy}}{\alpha n_{xy} + \beta n_{yy}} \neq e_{xy} \tag{5.3}$$

$$\text{Class proportions in multiclass problems:} \quad \frac{n'_{z.}}{n'_{..}} = \alpha_z \frac{n_{z.}}{n_{..}} \qquad n'_{zy} = \alpha_z\, n_{zy} \qquad \alpha_z \in \mathbb{R}_{<0}$$

$$\text{If } \exists \text{ classes } \zeta_x,\ \zeta_z \text{ with } \alpha_x \neq \alpha_z \text{ then:} \quad n'_{.y} = \textstyle\sum_z n'_{zy} = \sum_z \alpha_z n_{zy} \neq \alpha_x n_{.y}$$

$$\theta'_{xy} = \frac{\alpha_x n_{xy}}{\alpha_x n_{x.}} = \theta_{xy} \qquad e'_{xy} = \frac{\alpha_x n_{xy}}{\sum_z \alpha_z n_{zy}} \neq e_{xy} \tag{5.4}$$

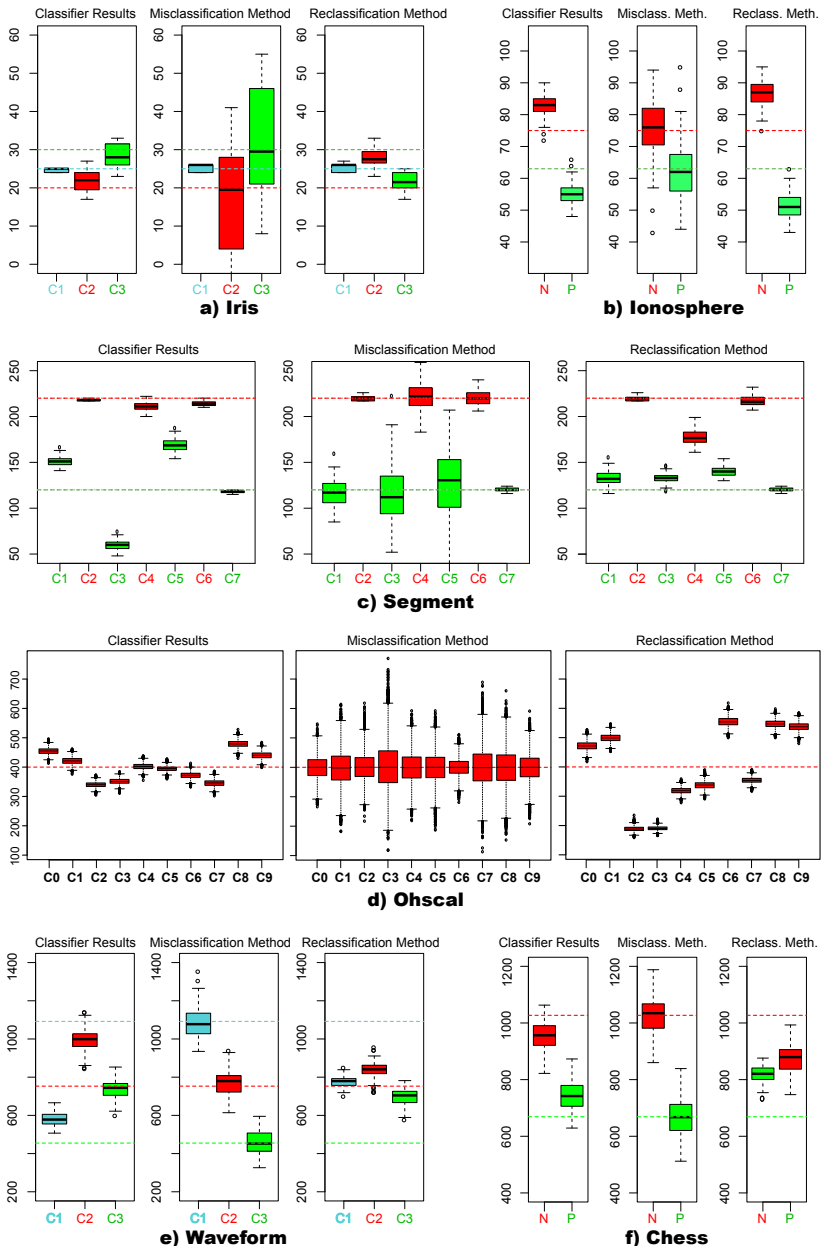

Figure 5.1: Class sizes provided by the raw classifier output (left graphs), error estimation with the Misclassification method (middle graphs) and Reclassification method (right graphs). *Boxplots show the median, 50%, 95% quartiles for 100 randomly sampled test and target sets. Horizontal dashed lines indicate actual class sizes, and colors indicate the related class (e.g., green boxplots with median values on green dashed lines indicate unbiased results).*

The Misclassification method yields significantly higher variance than the Reclassification method. The latter uses a simple linear sum of random variable $n'_{.y}$ $\boldsymbol{e_{xy}}$ while the former uses a matrix inversion. Cramer's rule (Kosinski 2001) shows that the random variables $\boldsymbol{\theta_{xy}}$ are involved several times in the denominator and numerator of a fraction, hence the higher variance (e.g., as the estimator is not linear).

If the test or target sets are small, or changes in class proportions are not significant, the variance of the Misclassification method may introduce more bias than the Reclassification method or the initial classification results (Figure 5.1-a to -d). Combining both methods does not reduce the variance (e.g., estimate $n'_{x.}$ with the Misclassification method, subsample the test set with similar class proportions $n_{x.}=\alpha n'_{x.}\ \forall x$, and apply Reclassification method using the resampled test set). Demonstration is omitted for brevity but reproducible with code in Section 5.9.

We conclude that existing bias corrections methods are applicable to machine learning classification problems. However, these applications must consider issues with results variance and changes in class proportions. If class proportions differ between test and target sets, the Misclassification method must be applied and the Reclassification method is inappropriate. If class sizes are relatively small, and class proportions do not differ, the Reclassification method is preferable. With the Misclassification method, variance issues have significant impact even when class sizes are not scarce (e.g., even with class sizes of several hundreds items, Figure 5.1-d to -f). Future work must investigate the guidelines for choosing the bias correction method to apply (or not) depending on test and target sizes, magnitude of changes in class proportions, and magnitude of classification biases.

5.3 Error composition

The methods presented in Section 5.2 can refine class size estimates. However, end-users may require more details on the errors between specific classes, e.g., in an output of $n'_{.y}$ items classified as class y, how many items n'_{xy} actually belong to class x. Such estimates of the error composition are of interest for describing the quality of classification results. We thus apply the methods presented in Section 5.2 to estimating the number of errors n'_{xy} between all possible combination of classes.

The Reclassification and Misclassification methods are easily extended to estimate n'_{xy} as in equation (5.5), using the error rates and class size estimates specified in equations (5.1)-(5.2).

$$\widehat{n'_{xy}} = e_{xy}\, n'_{.y} \qquad\qquad \widehat{n'_{xy}} = \theta_{xy}\widehat{n'_{x.}} \tag{5.5}$$

5.3.1 Ratio-to-TP method

We introduce an alternative method called Ratio-to-TP (Section 5.3.1). It provides exactly the same estimates as the Misclassification method, and is impacted by the

same variance magnitude[3]. However, it uses different error rates whose properties of interest are discussed in Sections 5.3.3 and 5.5.

The Ratio-to-TP method is based on atypical error ratios r_{xy} that use True Positives n_{xx} as denominators, as shown in equation (5.6), with $r_{xx} = 1$ and assuming $n_{xx} \neq 0$. Assuming equal error ratios in test and target sets (i.e., $\widehat{r'_{xy}} = r_{xy}$), we can construct the system of linear equations (5.7). The system's solution estimates the True Positives n'_{xx} in the target set, from which the number of errors n'_{xy} and true class sizes $n'_{x.}$ are easily derived, as shown in equation (5.6).

$$r_{xy} = \frac{n_{xy}}{n_{xx}} \qquad \begin{pmatrix} \widehat{n'_{11}} \\ \widehat{n'_{22}} \\ \vdots \\ \widehat{n'_{xx}} \end{pmatrix} = \begin{pmatrix} 1 & r_{21} & \dots & r_{x1} \\ r_{12} & 1 & \dots & r_{x2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1x} & r_{2x} & \dots & 1 \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \vdots \\ n'_{.x} \end{pmatrix} \qquad \begin{aligned} \widehat{n'_{xy}} &= r_{xy}\, \widehat{n'_{xx}} \\ \widehat{n'_{x.}} &= \textstyle\sum_y \widehat{n'_{xy}} \end{aligned} \tag{5.6}$$

$$n'_{.y} = \sum_x n'_{xy} = \sum_x n'_{xx} r'_{xy} \qquad \begin{Bmatrix} n'_{.1} = & n'_{11} & + & n'_{22}\, r'_{21} & + & \dots & + & n'_{xx}\, r'_{x1} \\ n'_{.2} = & n'_{11}\, r'_{12} & + & n'_{22} & + & \dots & + & n'_{xx}\, r'_{x2} \\ \dots = & \dots & + & \dots & + & \dots & + & \dots \\ n'_{.x} = & n'_{11}\, r'_{1x} & + & n'_{22}\, r'_{2x} & + & \dots & + & n'_{xx} \end{Bmatrix} \tag{5.7}$$

5.3.2 Application

We verify the applicability of the Ratio-to-TP method (5.6) and extension of Misclassification methods (5.5). We applied these methods using the same experimental setup as in Section 5.2.3. Both methods result in the same estimates[4], which are unbiased but with potentially high variance due to random differences between θ_{xy} and θ'_{xy} (Figure 5.2). We conclude that the potentially high variance is a challenge for estimating both $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$.

5.3.3 Discussion

The error rate matrix $\mathbf{M_r}$=$\begin{pmatrix} 1 & r_{2x} & \dots \\ r_{x2} & 1 & \dots \\ \dots & \dots & \dots \end{pmatrix}$ of the Ratio-to-TP method has all diagonal values equal to 1. It offers a simple condition to ensure its invertibility (i.e., that its determinant $|\mathbf{M_r}| \neq 0$) which is required for the Ratio-to-TP method to be applicable. Under condition (5.8) $\mathbf{M_r^\intercal}$ is diagonally dominant, thus invertible, and since $|\mathbf{M_r}| = |\mathbf{M_r^\intercal}|$ then $\mathbf{M_r}$ is also invertible. Setting a threshold t for all error rates $r_{xy,x\neq y} < t$ can ensure that the condition (5.8) is satisfied. $\mathbf{M_r}$ is always invertible under condition (5.9) where c is the number of classes (e.g., for 3-class problems t=0.5, 4-class t=0.33, 5-class t=0.25). It is also possible that $\mathbf{M_r}$ is invertible even if the condition is not met.

[3] Demonstration omitted for brevity but reproducible with code in Section 5.9.

[4] Demonstration is omitted but reproducible with code in Section 5.9.

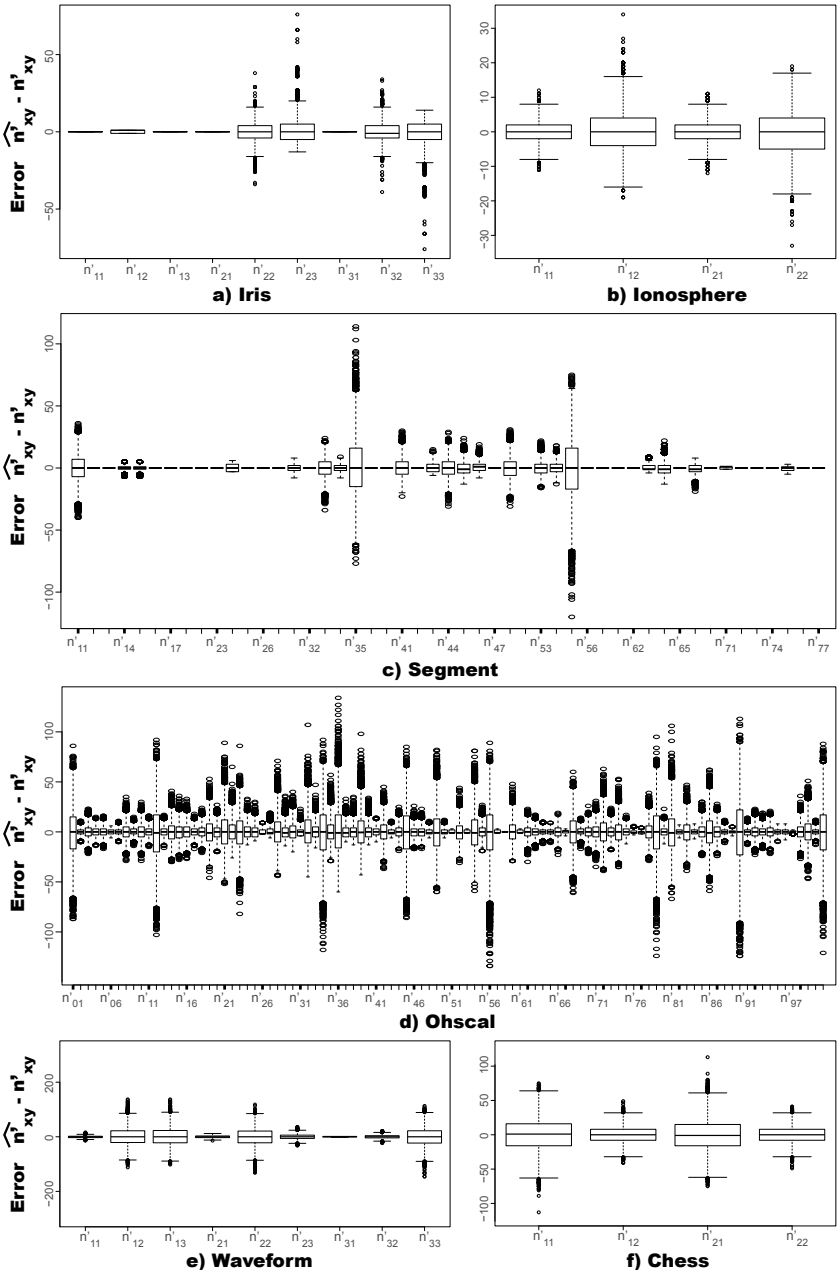

Figure 5.2: Evaluation of estimated $\widehat{n'_{xy}}$, showing the absolute error $n'_{xy} - \widehat{n'_{xy}}$ for 10^4 pairs test and target sets sampled as in Section 5.2.3.

$$\mathbf{M_r} = \begin{pmatrix} 1 & r_{21} & \dots & r_{x1} \\ r_{12} & 1 & \dots & r_{x2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1x} & r_{2x} & \dots & 1 \end{pmatrix} \qquad |\mathbf{M_r}| \neq 0 \text{ if for all class } x \sum_{y,y\neq x} r_{xy} < 1 \tag{5.8}$$

$$\text{Given } c \text{ number classes, if all } r_{xy,y\neq x} < \frac{1}{c-1} \text{ then } \sum_{y,y\neq x} r_{xy} < (c-1)\frac{1}{c-1} = 1$$

$$\text{Thus } |\mathbf{M_r}| \neq 0 \text{ if all } r_{xy,y\neq x} < \frac{1}{c-1} \tag{5.9}$$

The Misclassification method also requires its error rate matrix $\mathbf{M}_\theta = \begin{pmatrix} \theta_{11} & \theta_{21} & \dots \\ \theta_{12} & \theta_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix}$ to be invertible, but the Ratio-to-TP method offers a simple threshold condition to guarantee its matrix invertibility. We empirically observed that error rate matrices $\mathbf{M_r}$ and $\mathbf{M}_\theta$ drawn from the same test set were either both invertible, or both non-invertible. Future work is needed to establish if the threshold condition (5.9) ensuring the invertibility $\mathbf{M_r}$ also ensures the invertibility of $\mathbf{M}_\theta$.

We conclude that the Ratio-to-TP and Misclassification methods are applicable to estimating the detailed number of error between specific classes. However, these methods' results entail potentially high variance, which challenges the estimation of both $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$. Hence it is crucial to provide variance estimation methods. The Ratio-to-TP method uses error ratios r_{xy} that follow a Cauchy distribution, in contrast to θ_{xy} which follows a binomial distribution. Estimating the variance $V(r_{xy})$ is more complex, as the variance of the Cauchy distribution is undefined. Hence we focus on error rates θ_{xy} to estimate the variance of $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$.

5.4 Sample-to-Sample method

As mentioned in Sections 5.2 and 5.3, the Misclassification method entails potentially high variance. Hence providing variance estimation is crucial to support user awareness of the uncertainty in class size and error estimates from the Misclassification method. Existing variance estimation methods do not address the case of disjoint test and target sets (Section 5.2). We address this case by introducing the Sample-to-Sample method.

The Sample-to-Sample method estimates the variance of $\widehat{\theta'_{xy}}$, $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$ for the target set S', using measurements from the disjoint test set S (i.e., $S \cap S' = \emptyset$). We first approximate the variance of the $\widehat{\theta'_{xy}}$ estimator (Section 5.4.1) and validate our approach using known $n'_{x.}$ (Section 5.4.2). The method is then evaluated in practice with unknown $n'_{x.}$ by using estimated $\widehat{n'_{x.}}$ instead (Section 5.4.3). The method performs well for estimating the variance of $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$ in binary problems. Multiclass problems require future work investigating bootstrapping techniques, or simulations using Sample-to-Sample estimates of $\widehat{V}\left(\widehat{\theta'_{xy}}\right)$ (Section 5.4.5).

5.4.1 Error rate estimator

We focus on the estimator $\widehat{\theta'_{xy}}=\theta_{xy}$ for the unknown target set error rate θ'_{xy} based on the known error rate θ_{xy} in a disjoint test set. Test and target sets are assumed to be randomly sampled from the same population $n^*_{x.}\to\infty$ with error rate θ^*_{xy}. For test and target sets sampled with $n_{x.}$ and $n'_{x.}$ items, the expected value and variance of $\boldsymbol{\theta_{xy}}$ and $\boldsymbol{\theta'_{xy}}$ are given in equation (5.10) (Cochran 2007).

$$E[\boldsymbol{\theta_{xy}}] = E[\boldsymbol{\theta'_{xy}}] = \theta^*_{xy} \qquad V(\boldsymbol{\theta_{xy}}) = \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n_{x.}} \qquad V(\boldsymbol{\theta'_{xy}}) = \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n'_{x.}} \tag{5.10}$$

The estimator $\widehat{\boldsymbol{\theta'_{xy}}}=\boldsymbol{\theta_{xy}}$ yields the mean squared error in equation (5.11), which notation omits the subscripts, e.g, $\boldsymbol{\theta}=\boldsymbol{\theta_{xy}}$.

$$\begin{aligned}
MSE(\widehat{\boldsymbol{\theta}'}) &= E\Big[\big(\boldsymbol{\theta}-\boldsymbol{\theta}'\big)^2\Big] = E\Big[\big(\boldsymbol{\theta}-E[\boldsymbol{\theta}]+E[\boldsymbol{\theta}]-\boldsymbol{\theta}'\big)^2\Big] \\
&= E\Big[\big(\boldsymbol{\theta}\ -\ E[\boldsymbol{\theta}]\big)^2 + 2\big(\boldsymbol{\theta}\ -\ E[\boldsymbol{\theta}]\big)\big(E[\boldsymbol{\theta}]-\boldsymbol{\theta}'\big) + \big(E[\boldsymbol{\theta}]-\boldsymbol{\theta}'\big)^2\Big] \\
&= E\Big[\big(\boldsymbol{\theta}-E[\boldsymbol{\theta}]\big)^2\Big] - 2E\Big[\big(\boldsymbol{\theta}-E[\boldsymbol{\theta}]\big)\big(\boldsymbol{\theta}'-E[\boldsymbol{\theta}']\big)\Big] + E\Big[\big(\boldsymbol{\theta}'-E[\boldsymbol{\theta}']\big)^2\Big] \\
&= V(\boldsymbol{\theta}) - 2Cov(\boldsymbol{\theta},\boldsymbol{\theta}') + V(\boldsymbol{\theta}') \\
&\quad Cov(\boldsymbol{\theta},\boldsymbol{\theta}')=0 \text{ since Test Set}\cap\text{Target Set}=\emptyset \text{ and } \boldsymbol{\theta},\boldsymbol{\theta}' \text{ i.i.d., thus:} \\
MSE(\widehat{\boldsymbol{\theta}'_{xy}}) &= V(\boldsymbol{\theta_{xy}}) + V(\boldsymbol{\theta'_{xy}})
\end{aligned} \tag{5.11}$$

Following the results in equation (5.11), the Sample-to-Sample method considers that the estimator $\widehat{\boldsymbol{\theta'_{xy}}} = \boldsymbol{\theta_{xy}}$ is approximately distributed as in equation (5.12).

$$\widehat{\boldsymbol{\theta'_{xy}}} \sim N\Big(\theta_{xy}, V(\boldsymbol{\theta_{xy}}) + V(\boldsymbol{\theta'_{xy}})\Big) \tag{5.12}$$

Including variance component from both test and target sets is consistent with our empirical observations in Figure 5.3, where the sample size of either test or target sets impact the variance magnitude. Comprehensive evaluations of the Sample-to-Sample methods are presented in Sections 5.4.2 to 5.4.4.

5.4.2 Evaluation of error rate estimator

We evaluate the Sample-to-Sample estimates in equation (5.12) by simulating binary datasets and drawing confidence intervals for $\widehat{\boldsymbol{\theta'_{01}}}$. We focus on a single class 0 and ignore class 1, i.e., we simulate only n_{0y} and n'_{0y}. We draw 68% rather than 95% confidence level for a better verification of over-estimated intervals (e.g., one interval's coverage may be slightly higher than 95% but significantly higher than 68%). To estimate $V(\boldsymbol{\theta'_{01}})$ we use the known $n'_{0.}$ and apply Sample-to-Sample (5.12) using in equation (5.13). Further evaluations address realistic cases where $n'_{x.}$ is unknown (Sections 5.4.3 and 5.4.4).

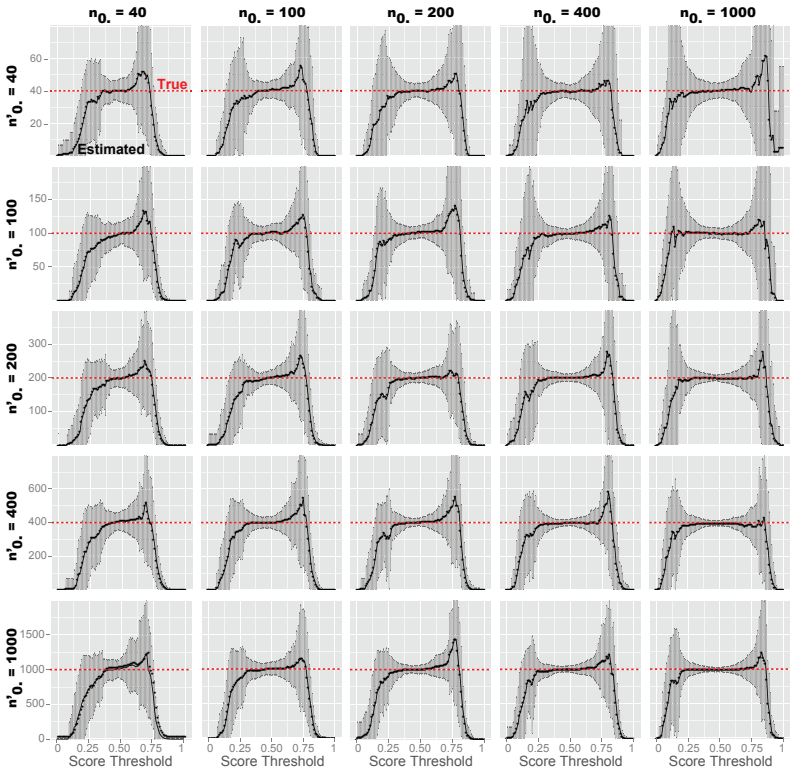

Figure 5.3: Results of Misclassification method for simulated data, showing results variance for different sample sizes of test and target sets. Score thresholds (x axis) are used to assign class 0 or 1, and simulate different magnitudes of error rate, as explained in Figure 5.4. Class sizes $\widehat{n'_{0.}}$ (y axis) are estimated for 10^4 pairs of test and target sets randomly sampled with score probability and class proportions in Figure 5.4, and for thresholds selected with granularity 0.01. We randomly sampled 100 test sets, and then randomly sampled 100 distinct target sets for each test set. This approach is realistic, as in practice one test set is used for several target sets. Unbiased means $\widehat{n'_{0.}}$ (black line) are close to true $n'_{0.}$ (red line) unless test sets are too small and error rate too close to 0 or 1 (e.g., when extreme thresholds yield few observations with $n_{xy} \approx$ a few items).

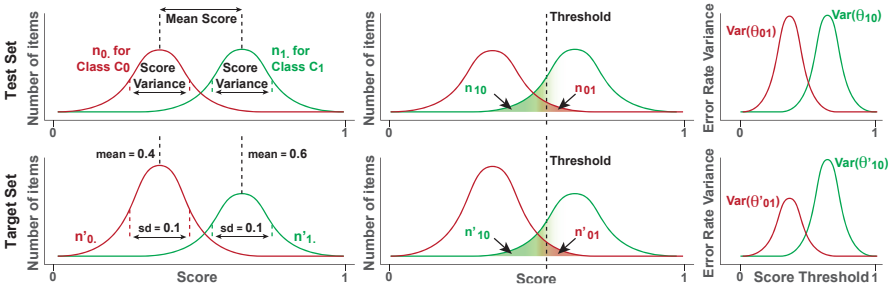

Figure 5.4: Specification of classification problem in Figure 5.3. Left: score distribution with means μ_0=μ'_0=0.4 for class 0, μ_1=μ'_1=0.6 for class 1, and σ_x=σ'_x=0.1. Middle: example of score threshold and related errors n_{01}, n_{10}. Right: error rate variance over thresholds. $V(\theta'_{01})$<$V(\theta'_{10})$ because we use $n'_{0.}$=$2n'_{1.}$ and $n_{0.}$=$n_{1.}$ to obtain different class proportions in test and target sets.

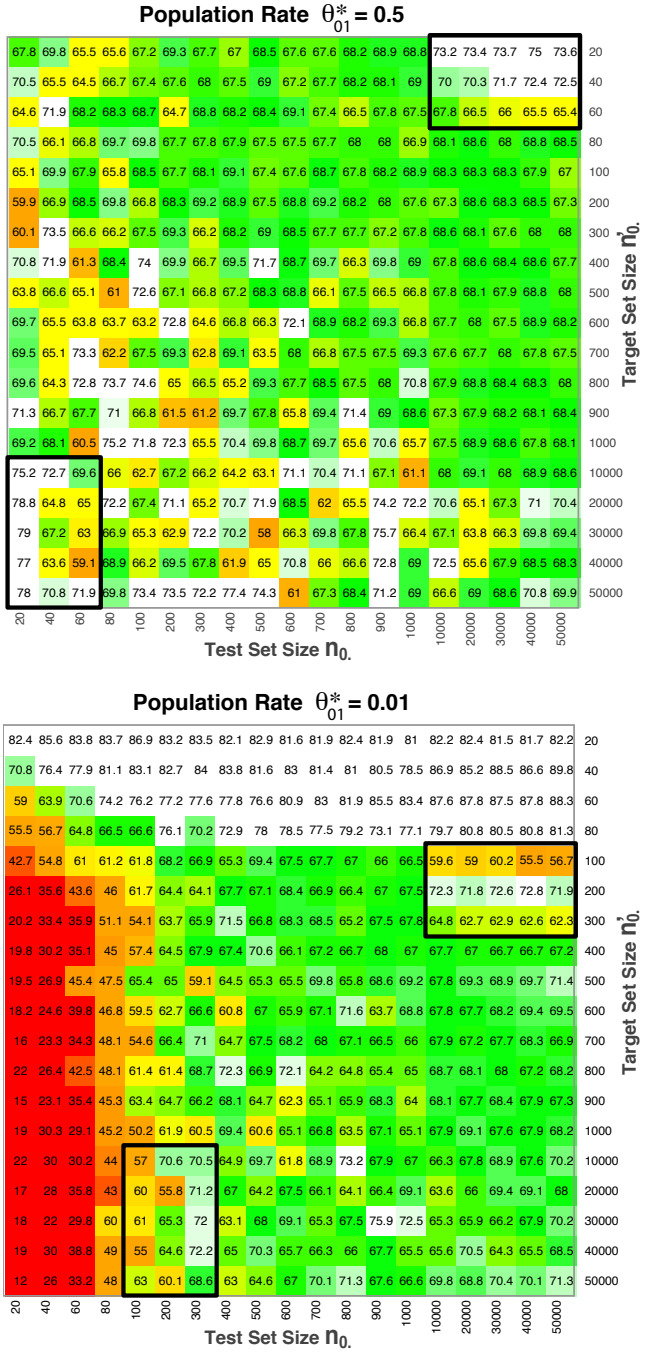

Figure 5.5: Evaluation of Sample-to-Sample using known $n'_{x.}$ to derive $V(\widehat{\theta'_{xy}})$ and draw 68% confidence intervals for $\widehat{\theta'_{xy}}$. The cells show the percentage of intervals containing true θ'_{01} for a total of 10^4 tests. Green cells have correct coverages ≈68%, red indicates too small coverages, white indicates too large coverages.

We sample 100 test sets of sizes $n_{0.} \in \{20, ..., 50\,000\}$ randomly drawn from an infinite population with $\theta^*_{01} \in \{0.01, 0.5\}$. For each test set, we measured θ_{01} and use equations (5.12-5.13) to draw confidence intervals for $\widehat{\theta'_{01}}$ in target sets of sizes $n'_{0.} \in \{20, ..., 50\,000\}$. For each interval, we randomly sample 100 target sets with the same population rate θ^*_{01}.

$$V\left(\widehat{\boldsymbol{\theta'_{01}}}\right) = \frac{\theta_{01}(1-\theta_{01})}{n_{0.}} + \frac{\theta_{01}(1-\theta_{01})}{n'_{0.}} \tag{5.13}$$

The graph cells in Figure 5.5 show the percentage of θ'_{xy} contained in confidence intervals derived using the Sample-to-Sample method. The confidence intervals achieve the desired confidence level, except when sample sizes $n_{x.}$ and $n'_{x.}$ are too small w.r.t. error rates θ^*_{xy} (in bottom graph only, e.g., $n_{xy} \approx 1$ item, same as the biases observed in Figure 5.3), or w.r.t. each other ($n_{x.} \ll$ or $\gg n'_{x.}$, black contours). The interval coverage varies more if $n_{x.} < n'_{x.}$ (lower left triangle of graphs) but mean coverage is correct (e.g., for θ^*_{xy}=0.5, in lower left triangle μ=68.1% and σ=4, otherwise μ=68.3% and σ=1.5).

5.4.3 Application to estimating class sizes

We evaluate the Sample-to-Sample method applied to estimating confidence intervals for the target class sizes $\widehat{n'_{x.}}$ resulting from the Misclassification method in binary problems. As in Section 5.4.2, we simulate 100 test sets and 100 target sets for each test set, with sizes $n_{x.}, n'_{x.} \in \{300, 500, 1000, 2000\}$, drawn from populations with θ^*_{xy} specified in equation (5.14).

$$\begin{pmatrix} \theta^*_{00} & \theta^*_{10} \\ \theta^*_{01} & \theta^*_{11} \end{pmatrix} \in \left\{ \begin{pmatrix} .9 & 0 \\ .1 & 1 \end{pmatrix}, \begin{pmatrix} .9 & .1 \\ .1 & .9 \end{pmatrix}, \begin{pmatrix} .9 & .2 \\ .1 & .8 \end{pmatrix}, \begin{pmatrix} .8 & .2 \\ .2 & .8 \end{pmatrix} \right\} \tag{5.14}$$

Confidence intervals are estimated using Fieller's theorem, as by Shieh (2009). We express the results of the Misclassification method as ratios in equation (5.15), assuming $1-\widehat{\theta'_{01}}-\widehat{\theta'_{10}} \neq 0$. Fieller's theorem applies to ratios of correlated random variables A/B, e.g., $A = n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}$ and $B = 1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}$. The variance and covariance of A and B are detailed in Section 5.9. For the estimator $\widehat{\theta'_{xy}} = \theta_{xy}$, we use the variance estimate in equation (5.16), derived from the Sample-to-Sample method, and using the results of the Misclassification method $\widehat{n'_{x.}}$ as estimates of the unknown $n'_{x.}$.

$$\widehat{n'_{0.}} = \frac{n'_{.0} - \widehat{\theta'_{10}}\, n'_{..}}{1 - \widehat{\theta'_{01}} - \widehat{\theta'_{10}}} \qquad \widehat{n'_{1.}} = \frac{n'_{.1} - \widehat{\theta'_{01}}\, n'_{..}}{1 - \widehat{\theta'_{01}} - \widehat{\theta'_{10}}} \tag{5.15}$$

$$\widehat{V}\left(\widehat{\boldsymbol{\theta'_{xy}}}\right) = \frac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}} \tag{5.16}$$

The results in Figure 5.6 show that the Sample-to-Sample method provides accurate confidence intervals for $\widehat{n'_{x.}}$. For each model in equation (5.14), the mean and variance of intervals' coverage are respectively: μ=68.1% σ=0.7, μ=68.2% σ=0.7, μ=68.2% σ=0.7, μ=68.2% σ=0.7.

These results are obtained without rounding the estimated $\widehat{n'_{x.}}$ nor the confidence limits. If these are rounded, the intervals are slightly biased and over-estimated. For instance, with our experiment setup, the coverage approximatively varied by ±3% for 68% intervals with μ=69.1%, and ±1% for 95% intervals with μ=95.6%.

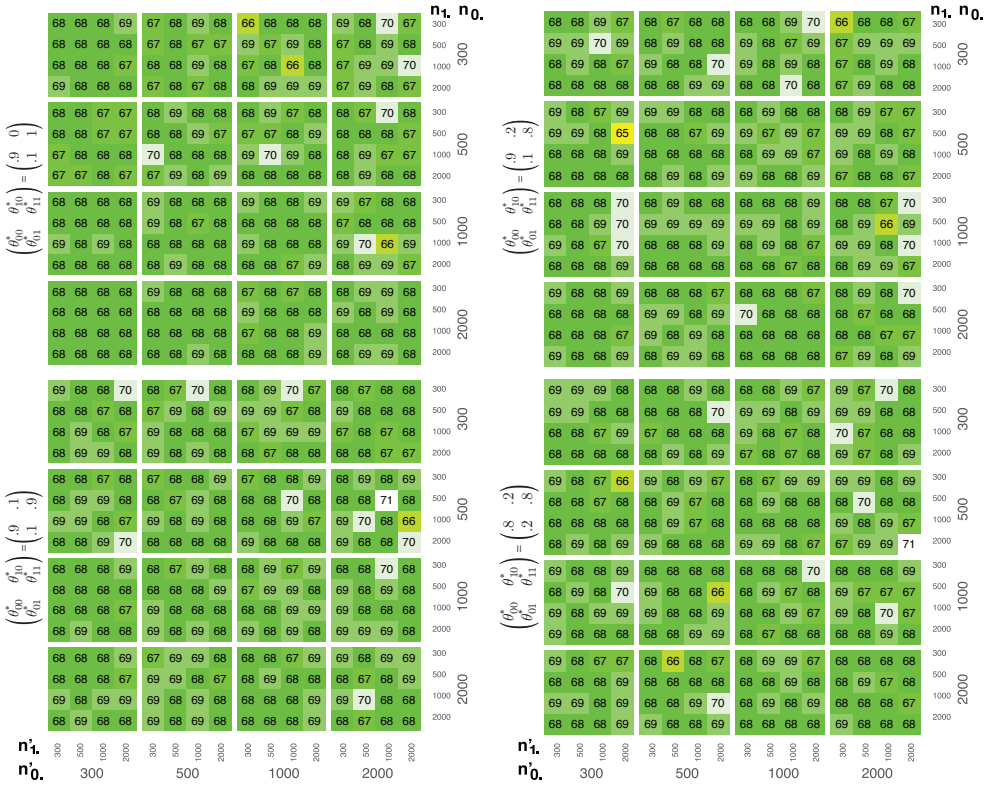

Figure 5.6: Results of Sample-to-Sample applied to estimating confidence intervals for $\widehat{n'_{x.}}$. The intervals accurately include the desired percentage of actual class size $n'_{x.}$ (68%, green cells). Test and target datasets are randomly sampled with sizes on columns and rows. The cells show the % of intervals that contained $n'_{0.}$ for a total of 10^4 tests (the % are rounded for clarity).

5.4.4 Application to estimating error composition

We evaluate the Sample-to-Sample method applied to estimating confidence intervals for the results $\widehat{n'_{xy}}$ of the extended Misclassification method (Section 5.3). As in Section 5.4.3, Fieller's theorem is applied with the same experimental setup, to derive confidence intervals for $\widehat{n'_{01}}$ instead of $\widehat{n'_{0.}}$. In this case, using the result of equation (5.5), $A=\widehat{\boldsymbol{\theta'_{01}}}\left(n'_{.0}-\widehat{\boldsymbol{\theta'_{10}}}\,n'_{..}\right)$ in equation (5.15). The variance and covariance of A and B are detailed in Section 5.9.

Instead of drawing a graph as Figure 5.6, we report the mean and variance of interval coverage for each model in (5.14), respectively: μ=68.0% σ=0.7, μ=68.1% σ=0.8, μ=68.2% σ=0.7, μ=68.3% σ=0.7. It shows that the Sample-to-Sample method provides accurate confidence intervals for $\widehat{n'_{xy}}$.

5.4.5 Discussion

In this section, we discuss how the Sample-to-Sample method contributes to **prior work** focusing on class proportions, and applications of the **Reclassification methods** (p.92). We then introduce future work of interest for addressing **multiclass problems** (p.94), and investigating the potential impact of **number of classes** (p.94) on variance magnitude.

Prior work - The Sample-to-Sample approach is applicable to prior work focusing on estimating class proportions, e.g., $\pi'_x=n'_{x.}/n'_{..}$ (Shieh 2009, Buonaccorsi 2010). This prior method is restated for class 0 in equation (5.17) using our notation.

The main difference with the Sample-to-Sample approach is how test and target set sizes $n_{x.}$, $n'_{x.}$ are considered for the variance estimation. The Sample-to-Sample approach accounts for test and target sets' class sizes $n_{x.}$, $n'_{x.}$ to estimate the error rate variance $V\left(\widehat{\theta'_{xy}}\right)$. The prior method uses only test sets' class sizes $n_{x.}$ to estimate the error rate variance $V\left(\widehat{\theta'_{xy}}\right)$. Target sets' class sizes $n'_{x.}$ are used only for estimating the variance of class proportions $n'_{.y}/n'_{..}$ from the classifier output (i.e., prior to applying the Misclassification method).

With the prior approach from Shieh (2009) and Buonaccorsi (2010), the variance of the numerator and denominator in equation (5.17) is estimated with equation (5.18). Then, Fieller's theorem can be applied as detailed in additional materials (Section 5.9).

$$\widehat{\pi_0} = \frac{\widehat{n'_{0.}}}{n'_{..}} = \frac{n'_{.0}/n'_{..} - \theta_{10}}{1-\theta_{01}-\theta_{10}} \tag{5.17}$$

$$\begin{aligned} V\left(n'_{.0}/n'_{..} - \theta_{10}\right) &= \frac{n'_{.0}/n'_{..}\left(1-n'_{.0}/n'_{..}\right)}{n'_{..}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}} \\ V\left(1-\theta_{01}-\theta_{10}\right) &= \frac{\theta_{01}(1-\theta_{01})}{n_{0.}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}} \end{aligned} \tag{5.18}$$

The results of equations (5.17)-(5.18) are reproduced in Figure 5.7, using variables similar to prior evaluation (Shieh 2009): $n_{x.} \in \{25, 50, 125, 250\}$, $n'_{x.} \in \{50, 125, 250, 500\}$, θ_{01}=0.1, θ_{10}=0.2. When used for estimating target sets' class proportions $\widehat{\pi'_x} = n'_{x.}/n'_{..}$, the prior method is biased for many values of $n_{x.}$ and $n'_{x.}$.

The prior method was designed for estimating the class proportions $\widehat{\pi^*_x} = n*_{x.}/n*_{..}$ in the overall population from which test and target sets are randomly sampled. We show that this prior variance estimation method is not applicable for estimating the class sizes or proportions of target sets, i.e., $\widehat{n'_{x.}}$ or $\widehat{\pi'_x} = n'_{x.}/n'_{..}$ as in Figure 5.7.

The bias in Figure 5.7 can be corrected with the Sample-to-Sample method, considering no variance for the initial class proportion $n'_{.y}/n'_{..}$ as shown in equation (5.19). The corrected results in Figure 5.8 have a small bias when sample sizes $n_{x.}$ and $n'_{x.}$ are small w.r.t. the error rates (i.e., yielding small numbers of errors n_{xy}, n'_{xy} where variations of ±1 item can yield significant error rate variations, as mentioned in Figure 5.3). Estimates drawn using the larger sample sizes in Figure 5.6 are unbiased with mean coverage μ=68.2%, σ=0.7. These results show that the Sample-to-Sample method is suitable for estimating target sets' class proportions π'_x=$n'_{x.}/n'_{..}$.

$$\widehat{V}_{corrected}\left(n'_{.0}/n'_{..} - \theta_{10}\right) = \frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}} \tag{5.19}$$

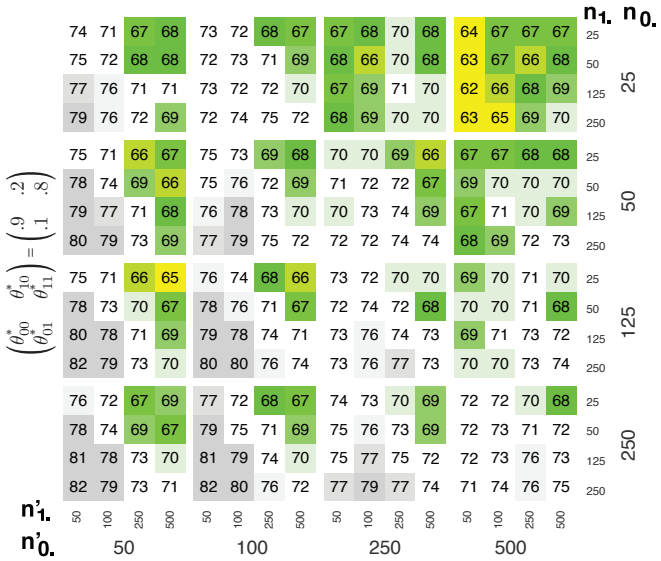

Figure 5.7: Confidence intervals drawn using prior work by Shieh (2009) and Buonaccorsi (2010). The intervals are biased and tend to include a too large percentage of actual class sizes $\widehat{n'_{x.}}$ (too large intervals, white to grey cells). Test and target datasets are randomly sampled with sizes on columns and rows. The cells show the percentage of intervals that contained π'_0=$n'_{0.}/n'_{..}$ for a total of 10^4 tests per cell (percentages are rounded for clarity).

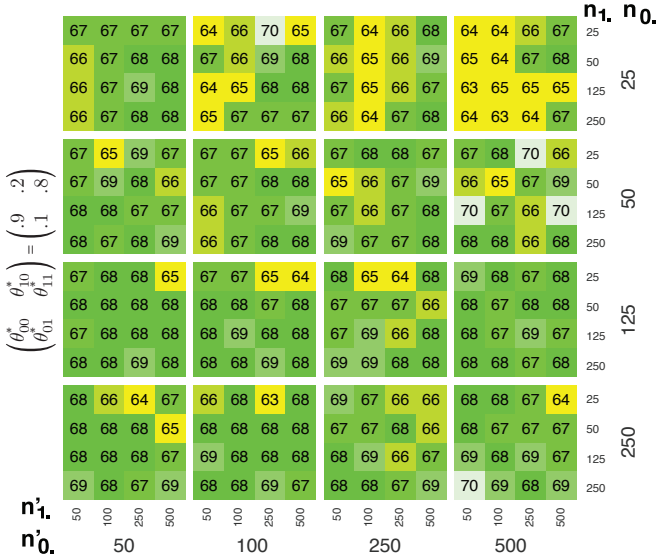

Figure 5.8: Results of Sample-to-Sample method used to correct the bias in Figure 5.7. The intervals accurately include the desired percentage of actual class size $n'_{x.}$ (68%, green cells). However, bias may occur if class sizes are scarce (e.g., around 25 items, yellow cells) as variations of a few errors can have significant impacts on the resulting error rates.

Reclassification method - The Sample-to-Sample approach is not always applicable to the Reclassification method, i.e., to the error rate estimator $\widehat{e'_{xy}} = e_{xy}$ (equation (5.1) p.79). Equations (5.10) and (5.11) do not apply when class proportions differ between test and target sets, as $\boldsymbol{e'_{xy}}$ and $\boldsymbol{e_{xy}}$ are not identically distributed. As shown in equations (5.3)-(5.4) p.80, their denominators depend on the class proportions. If class proportions differ between test and target sets, the denominators differ and are not proportional to the numerators, yielding different distributions. Thus their expected values differ $E[\boldsymbol{e_{xy}}] \neq E[\boldsymbol{e'_{xy}}]$ and equations (5.10) and (5.11) do not apply.

However, stable class proportions is a prerequisite for the reclassification to be applicable, as otherwise bias ensues (Section 5.2.4). With equal class proportions, the Sample-to-Sample approach can be applied with equations (5.20)-(5.22) (equation (5.21) omits the subscripts, e.g, $\boldsymbol{e}$=$\boldsymbol{e_{xy}}$). However, class proportions may vary randomly due to sample variance. Thus error rate variance should consider the variance of its denominator (i.e., $V(n_{.y}) = \sum_x V(n_{xy})$ with $Cov(n_{xy}, n_{zy})$=0 since Class $x \cap$ Class $z = \emptyset$) which is ignored in equation (5.20).

$$\text{iif } \forall x,\ E\left[\frac{\boldsymbol{n_{x.}}}{\boldsymbol{n_{..}}}\right] = E\left[\frac{\boldsymbol{n'_{x.}}}{\boldsymbol{n'_{..}}}\right] :$$

$$E[\boldsymbol{e_{xy}}] = E[\boldsymbol{e'_{xy}}] = e^*_{xy} \quad V(\boldsymbol{e_{xy}}) = \frac{e^*_{xy}(1-e^*_{xy})}{n_{.y}} \qquad V(\boldsymbol{e'_{xy}}) = \frac{e^*_{xy}(1-e^*_{xy})}{n'_{.y}} \tag{5.20}$$

$$MSE(\widehat{\boldsymbol{e}'}) = E\left[\left(\boldsymbol{e}-\boldsymbol{e}'\right)^2\right] = V(\boldsymbol{e}) - 2Cov(\boldsymbol{e},\boldsymbol{e}') + V(\boldsymbol{e}')$$

$$Cov(\boldsymbol{e},\boldsymbol{e}') = 0 \text{ since Test Set} \cap \text{Target Set} = \emptyset \text{ and } \boldsymbol{e},\boldsymbol{e}' \text{ independent, thus:} \quad (5.21)$$

$$MSE(\widehat{\boldsymbol{e}'_{xy}}) = V(\boldsymbol{e}_{xy}) + V(\boldsymbol{e}'_{xy})$$

$$\widehat{\boldsymbol{e}'_{xy}} \sim N\Big(e_{xy}, V(\boldsymbol{e}_{xy}) + V(\boldsymbol{e}'_{xy})\Big) \qquad \widehat{V}(\boldsymbol{e}_{xy}) = \frac{e_{xy}(1-e_{xy})}{n_{.y}} \qquad \widehat{V}(\boldsymbol{e}'_{xy}) = \frac{e_{xy}(1-e_{xy})}{n'_{.y}} \quad (5.22)$$

Multiclass problems - Classification problems with more than 3 classes are not easily solved as fractions of random variables using Cramer's rule, as in equation (5.15) p.89. Thus Fieller's theorem is not easy to apply. Sarrus' rule applies to 3-class problems, providing a solution that can be expressed as ratios using Cramer's rule. However, applying Fieller's theorem to the resulting ratios remains complex.

Bootstrapping methods are thus recommended for multiclass problems (Buonaccorsi 2010). Monte Carlo simulations are also of interest. Datasets can be simulated using error rate variance from the Sample-to-Sample method, using equation (5.12). Future work should investigate Monte Carlo simulations, and compare their results to bootstrapping methods.

Number of classes - Future work could investigate whether the number of classes impacts the variance of the Misclassification method. For instance, problems with larger numbers of classes may entail larger magnitudes of variance (for problems with similar error rates and class size magnitudes).

According to Cramer's rule, the results of the Misclassification method are fractions of two matrix determinants (Kosinski 2001), as shown in equation (5.23) for 4-class problems. The matrices are composed of random variables $\boldsymbol{\theta}_{xy}$ and output class sizes $n'_{.y}$.

The Laplace expansion shows that matrices' determinants are weighted sums of sub-matrices' determinant, as shown in equation (5.24) for 4-class problems. The variables $\boldsymbol{\theta}_{xy}$ are used several times in these sub-matrices. As the variables $\boldsymbol{\theta}_{xy}$ are duplicated in the sub-matrices, their variances $V(\boldsymbol{\theta}_{xy})$ may have increased impact on the variance of the determinants, and thus on the results of the Misclassification method (as mentioned in Section 5.2.4). Problems with larger numbers of classes involve more sub-matrices, and thus more duplicated variables $\boldsymbol{\theta}_{xy}$. Thus we can expect that the larger the number of classes, the higher the variance of the Misclassification method's results.

$$\widehat{n'_{1.}} = \frac{\begin{vmatrix} n'_{.1} & \theta_{21} & \theta_{31} & \theta_{41} \\ n'_{.2} & \theta_{22} & \theta_{32} & \theta_{43} \\ n'_{.3} & \theta_{23} & \theta_{33} & \theta_{43} \\ n'_{.4} & \theta_{24} & \theta_{34} & \theta_{44} \end{vmatrix}}{\begin{vmatrix} \theta_{11} & \theta_{21} & \theta_{31} & \theta_{41} \\ \theta_{12} & \theta_{22} & \theta_{32} & \theta_{43} \\ \theta_{13} & \theta_{23} & \theta_{33} & \theta_{43} \\ \theta_{14} & \theta_{24} & \theta_{34} & \theta_{44} \end{vmatrix}} \qquad \widehat{n'_{2.}} = \frac{\begin{vmatrix} \theta_{11} & n'_{.1} & \theta_{31} & \theta_{41} \\ \theta_{12} & n'_{.2} & \theta_{32} & \theta_{43} \\ \theta_{13} & n'_{.3} & \theta_{33} & \theta_{43} \\ \theta_{14} & n'_{.4} & \theta_{34} & \theta_{44} \end{vmatrix}}{\begin{vmatrix} \theta_{11} & \theta_{21} & \theta_{31} & \theta_{41} \\ \theta_{12} & \theta_{22} & \theta_{32} & \theta_{43} \\ \theta_{13} & \theta_{23} & \theta_{33} & \theta_{43} \\ \theta_{14} & \theta_{24} & \theta_{34} & \theta_{44} \end{vmatrix}} \quad \dots \quad (5.23)$$

$$\begin{vmatrix}\theta_{11} & \theta_{21} & \theta_{31} & \theta_{41}\\ \theta_{12} & \theta_{22} & \theta_{32} & \theta_{43}\\ \theta_{13} & \theta_{23} & \theta_{33} & \theta_{43}\\ \theta_{14} & \theta_{24} & \theta_{34} & \theta_{44}\end{vmatrix} = \theta_{11}\begin{vmatrix}\theta_{22} & \theta_{32} & \theta_{43}\\ \theta_{23} & \theta_{33} & \theta_{43}\\ \theta_{24} & \theta_{34} & \theta_{44}\end{vmatrix} - \theta_{12}\begin{vmatrix}\theta_{21} & \theta_{31} & \theta_{41}\\ \theta_{23} & \theta_{33} & \theta_{43}\\ \theta_{24} & \theta_{34} & \theta_{44}\end{vmatrix} + \theta_{13}\begin{vmatrix}\theta_{21} & \theta_{31} & \theta_{41}\\ \theta_{22} & \theta_{32} & \theta_{43}\\ \theta_{24} & \theta_{34} & \theta_{44}\end{vmatrix} - \theta_{14}\begin{vmatrix}\theta_{21} & \theta_{31} & \theta_{41}\\ \theta_{22} & \theta_{32} & \theta_{43}\\ \theta_{23} & \theta_{33} & \theta_{43}\end{vmatrix} \tag{5.24}$$

5.5 Maximum Determinant method

The Sample-to-Sample method introduced in Section 5.4 can assess the variance of classification error estimates for specific datasets. However, when comparing classifiers to select the optimal classifier for their tasks, end-users are not interested in classifiers' performance for one specific dataset but for a variety of potential datasets (e.g., unknown target sets). The characteristics of the target sets (e.g., target class sizes) may be unknown when comparing classifiers. In this case, the Sample-to-Sample method cannot be applied, and end-users cannot assess which classifier may yield the smallest variance when estimating class sizes and numbers of errors.

To address this issue, the Maximum Determinant method is a promising approach. The method aims at predicting which classifier may yield the smallest variance when applying the error estimation methods, without requiring information on the potential target sets.

5.5.1 Determinants as variance predictors

The Maximum Determinant method focuses on the determinant $|\mathbf{M}|$ of error rate matrices, i.e., $|\mathbf{M}_\theta| = \left|\begin{smallmatrix}\theta_{11} & \theta_{21} & \dots\\ \theta_{12} & \theta_{22} & \dots\\ \dots & \dots & \dots\end{smallmatrix}\right|$ for the Misclassification method, or $|\mathbf{M_r}| = \left|\begin{smallmatrix}1 & r_{21} & \dots\\ r_{12} & 1 & \dots\\ \dots & \dots & \dots\end{smallmatrix}\right|$ for the Ratio-to-TP method. According to Cramer's rule, the results of the misclassification and Ratio-to-TP methods are fractions of two matrix determinants $\widehat{n'_{x.}} = \frac{|\mathbf{A}|}{|\mathbf{M}|}$ (Kosinski 2001). The fraction's denominator is the determinant of the error rate matrix $|\mathbf{M}_\theta|$ or $|\mathbf{M_r}|$. If the determinant $|\mathbf{M}| \to 0$ then $\widehat{n'_{x.}} \to \infty$.

For a small determinant $|\mathbf{M}| \to 0$, a variation $|\mathbf{M}^+| = |\mathbf{M}| + \delta$ can yield a large variation in $\widehat{n'_{x.}}$ as $\widehat{n'_{x.}} \to \infty$. For a larger determinant $|\mathbf{M}| \gg 0$, the same variation $|\mathbf{M}^+| = |\mathbf{M}| + \delta$ yields a smaller variation in $\widehat{n'_{x.}}$.

Hence the Maximum Determinant method postulates that the larger the difference $\big||\mathbf{M}| - 0\big|$ the smaller the variance $V(\widehat{n'_{x.}})$. This approach allows to compare classifiers' error rate matrices to predict which classifier may yield the least variance when estimating the classification errors in target sets. However, this approach is only applicable if the error rate matrices to compare are drawn from the same test set.

5.5.2 Application

We present an initial evaluation of the Maximum Determinant method, which results are shown in Figure 5.9 and Table 5.3. We use the same datasets as in Section 5.2.3. To sample several target sets for the same test set, we use smaller sample sizes than in Section 5.2.3 (i.e., in Table 5.3, $n_{x.}+n'_{x.}<n^*_{x.}$ where $n^*_{x.}$ is the total number of items available for class x). We sample 1000 test sets and measure their matrix determinants $|\mathbf{M}_\theta|$ and $|\mathbf{M_r}|$. For each test set, we sample 100 distinct target sets and compute the variance $V(\widehat{n'_{x.}})$ over the target sets. We visualize the relationship between the variance $V(\widehat{n'_{x.}})$ and the matrix determinants $|\mathbf{M}_\theta|$ or $|\mathbf{M_r}|$ (Figure 5.9) and compute their correlation (Table 5.3).

From Figure 5.9, we observe that $V(\widehat{n'_{x.}})$ seems to be a linear function of $|\mathbf{M}|$. From Table 5.3, we observe that the negative correlation between $|\mathbf{M}_\theta|$ and $\sum_x V(\widehat{n'_{x.}})$ or $\sum_x \sum_y V(\widehat{n'_{xy}})$ is consistent with the hypothesis that high determinants yield lower variance $V(\widehat{n'_{x.}})$ and $V(\widehat{n'_{xy}})$.

The observed correlation is significant for multiclass datasets, and less significant but consistent for binary datasets (i.e., negative or null). Hence the Maximum Determinant method may not be relevant for some binary problems.

				Correlation of $\lvert\mathbf{M}_\theta\rvert$ and $\sum$Var		**Correlation of $\lvert\mathbf{M_r}\rvert$ and $\sum$Var**	
	Dataset	**Test Set $n_{x.}$**	**Target Set $n'_{x.}$**	$\mathbf{V}(\widehat{n'_{x.}})$	$\mathbf{V}(\widehat{n'_{xy}})$	$\mathbf{V}(\widehat{n'_{x.}})$	$\mathbf{V}(\widehat{n'_{xy}})$
Test T1	Iris	n_{1-2}=20 n_3=15	n_{1-3}=25	-0.81	-0.79	**-0.91**	**-0.89**
	Ionosphere	n_1=50 n_0=50	n_1=50 n_0=100	**-0.35**	**-0.13**	-0.21	-0.01
	Segment	n_{1-7}=100	$n_{1,3,5,7}$=100 $n_{2,4,6}$=200	**-0.83**	**-0.81**	-0.79	-0.76
	Ohscal	n_{0-9}=400	n_{0-4}=100 n_{5-10}=200	-0.72	-0.52	**-0.75**	**-0.64**
	Waveform	n_{1-3}=300	n_1=300 n_2=600 n_3=900	**-0.53**	**-0.40**	-0.16	-0.08
	Chess	n_1=300 n_0=500	n_1=1000 n_0=500	-0.01	0.08	0	0.08
Test T2	Iris	n_{1-2}=10 n_3=15	n_{1-3}=25	-0.79	-0.77	**-0.89**	**-0.87**
	Ionosphere	n_1=30 n_0=30	n_1=50 n_0=100	**-0.36**	**-0.12**	-0.23	0.01
	Segment	n_{1-7}=50	$n_{1,3,5,7}$=100 $n_{2,4,6}$=200	**-0.83**	**-0.81**	-0.78	-0.75
	Ohscal	n_{0-9}=200	n_{0-4}=100 n_{5-10}=200	-0.71	-0.53	**-0.75**	**-0.65**
	Waveform	n_{1-3}=200	n_1=300 n_2=600 n_3=900	**-0.49**	**-0.35**	-0.18	-0.10
	Chess	n_1=200 n_0=300	n_1=1000 n_0=500	-0.01	0.08	0	0.09
Test T3	Iris	n_{1-3}=25	n_{1-2}=10 n_3=15	-0.24	-0.24	**-0.35**	**-0.34**
	Ionosphere	n_1=50 n_0=100	n_1=30 n_0=30	**-0.80**	**-0.64**	-0.75	-0.58
	Segment	$n_{1,3,5,7}$=100 $n_{2,4,6}$=200	n_{1-7}=50	-0.72	-0.71	**-0.77**	**-0.74**
	Ohscal	n_{0-4}=100 n_{5-10}=200	n_{0-9}=200	-0.68	-0.49	**-0.72**	**-0.59**
	Waveform	n_1=300 n_2=600 n_3=900	n_{1-3}=200	**-0.61**	**-0.46**	-0.16	-0.08
	Chess	n_1=1000 n_0=500	n_1=200 n_0=300	-0.33	-0.16	**-0.34**	**-0.17**

Table 5.3: Results of Maximum Determinant method

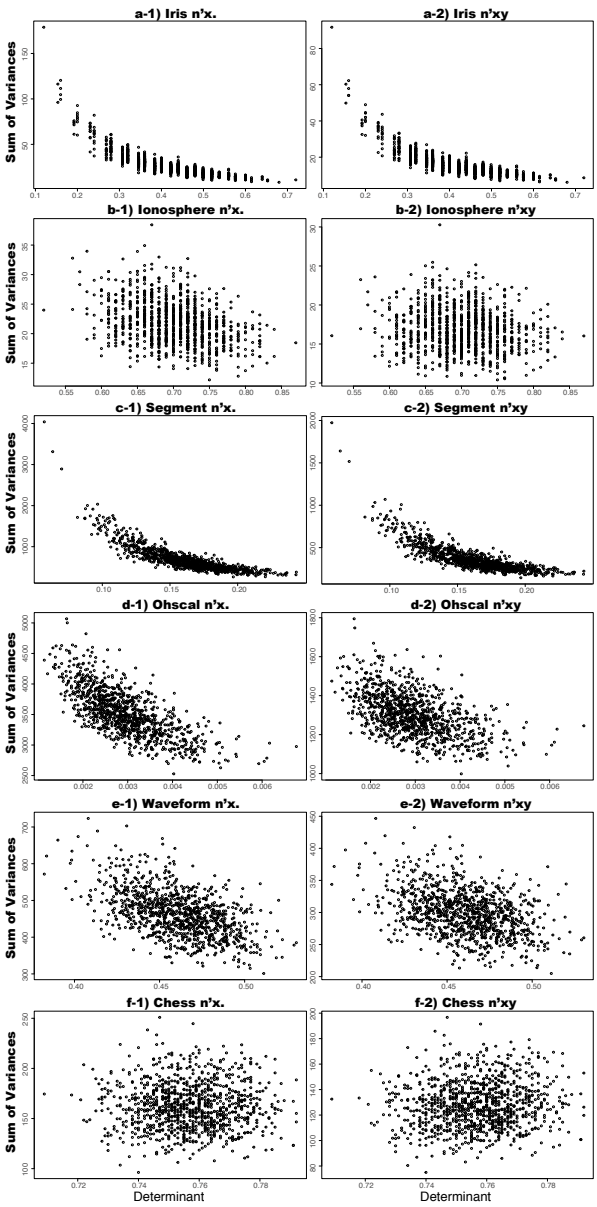

Figure 5.9: Results of Maximum Determinant method, applied using $|\mathbf{M}_\theta|$ (Misclassification method) and the datasets of test T1 in Table 5.3. The x-axis shows $|\mathbf{M}_\theta|$, and the y-axis $\sum_x V(\widehat{n'_{x.}})$ (left graphs) and $\sum_x \sum_y V(\widehat{n'_{xy}})$ (right graphs). Each dot represents a test set for which 10^2 target sets are sampled. The summation of variance may explain the exponentiality in graph -a) -c).

5.5.3 Discussion

The initial results are promising for multiclass problems. Error rate matrices from either the Misclassification method $|\mathbf{M}_\theta|$ or Ratio-to-TP method $|\mathbf{M_r}|$ are shown to correlate with the variance of the error estimation methods, e.g., correlation coefficients between -0.49 to -0.91 for $\sum_x \widehat{V(n'_{x.})}$. However, the observed correlations may not hold for cases unaddressed in our initial evaluation.

Future work is required for establishing theory and identifying the problem's variables and impacts, e.g., to answer questions such as:

- What are the parameters of the functions $f(|\mathbf{M}_\theta|)=\widehat{V(n'_{x.})}$ and $f(|\mathbf{M_r}|)=\widehat{V(n'_{x.})}$?
- Are there binary problems for which the Maximum Determinant method is irrelevant?
- In which cases is $|\mathbf{M}_\theta|$ or $|\mathbf{M_r}|$ a better predictor?
- Given error rate matrices for alternative classifiers c_1 and c_2, and their determinant $|\mathbf{M_{r,c1}}|$, $|\mathbf{M_{r,c2}}|$, $|\mathbf{M}_{\theta,\mathbf{c1}}|$, $|\mathbf{M}_{\theta,\mathbf{c2}}|$. Are the determinants' order of magnitude consistent whether using error rate θ or r, i.e., does $|\mathbf{M}_{\theta,\mathbf{c1}}| < |\mathbf{M}_{\theta,\mathbf{c2}}|$ imply that $|\mathbf{M_{r,c1}}| < |\mathbf{M_{r,c2}}|$?
- Do smaller test sets with a higher matrix determinant yield less variance than larger test sets with a lower determinant?
- Is it recommend to draw alternative split of the groundtruth into test and training sets, and select the split yielding the highest matrix determinant? This approach requires training classifications models several times, i.e., with each alternative training set to draw the corresponding error rate matrices.
- How to refine the Maximum Determinant prediction by using include information on the potential target set, e.g. ranges of potential class sizes or feature distributions?

The Maximum Determinant method is based on a postulate that is not established at this stage. However, inspecting the determinant of error rate matrices is nonetheless of interest to assess how error estimation results may vary. For instance, with determinants close to zero, the Misclassification method may not be recommended and the Reclassification method may be preferred (if class proportions remain unchanged between test and target sets).

5.6 Applicability issues

The methods presented in this chapter are applicable under specific conditions. For instance, if class proportions differ between test and target sets, the Reclassification method is biased. However, the Misclassification method yields potentially high result variance. The variance is higher when the class sizes are smaller, either in test or target sets. Thus if datasets are small with limited variations of class proportions, the Reclassification method may be preferable to the Misclassification method.

Future work is required to establish guidelines for choosing appropriate error estimation methods. Examples of issues to consider when assessing a method's applicability are given in Table 5.4. Impractical cases can be identified, e.g., when error estimation results are unrealistic (Section 5.6.1). Otherwise, test set representativity must be assessed (Section 5.6.2) as impractical cases ensue with small test sets or varying feature distributions (Section 5.6.3).

	Test and target sets characteristics				
	Error rates differ	**Class proportions differ**	**Feature distributions differ**	**Small class size**	**Overlap** Test∩Target $\neq \emptyset$
Applicability w.r.t.:	*Bias issues*			*Variance issues*	
Reclassification method Section 5.2.1, p.78	No	No	No	No if $n_{x.}$<~100	Yes
Misclassification method Section 5.2.2, p.79	No	Yes	No	No if $n_{x.}$<~500	Yes
Ratio-to-TP method Section 5.3.1, p.82	No	Yes	No	No if $n_{x.}$<~500	Yes
Sample-to-Sample method Section 5.4, p.85	No	Yes	No	?	No
Maximum Determinant method Section 5.5, p.95	No	Yes	No	?	?
Logistic Regression method Section 5.7.2.A, p.106	Yes	No	Yes	?	Yes
Bayesian method Section 5.7.2.B, p.106	No	Yes	Yes	?	Yes

Table 5.4: Method applicability depending on dataset characteristics (the last two methods are introduced in Section 5.7, p.105)

5.6.1 Impractical cases

We identify practical issues that may arise when applying error estimation methods, and that indicate that the methods may not be applicable:

- The methods may provide **negative estimates of** $n'_{x.}$ **and** n'_{xy} (p.100)
- The methods use error rate matrices that require **matrix invertibility** (p.100)
- The methods may have **negative effects** and worsen the initial classification bias (p.100)
- The methods may yield critical result variance if applied to **small class sizes** (p.101)

Negative estimates $n'_{x.}$ or n'_{xy}

The Misclassification and Ratio-to-TP methods can yield negative estimates $\widehat{n'_{x.}} < 0$ or $\widehat{n'_{xy}} < 0$, and the Sample-to-Sample method can yield confidence intervals with negative lower bounds. However, it happened rarely in our experiments, usually with scarce class sizes or extreme error rates (e.g., $\theta \to 0$ or 1).

Negative estimates are easily handled for binary problems, i.e., if $\widehat{n'_{0.}}$<0, set $\widehat{n'_{1.}}$ to $\widehat{n'_{1.}} + \widehat{n'_{0.}}$, and $\widehat{n'_{0.}}$ to 0. Future research is required to handle negative estimates in multiclass problems, e.g., by using the linear combination (5.25). More importantly, negative estimates indicate that the methods may not be applicable.

Matrix invertibility

The Misclassification method is not applicable when the determinant of the error rate matrix is zero. For instance, in binary problems, the determinant is zero iff $\theta_{01} + \theta_{10} = 1$. Such cases occur with random classifiers (i.e., $\theta_{01} = \theta_{10} = 0.5$), or with classifiers performing worse than random for one class and inversely proportional for the other class (e.g., $\theta_{01} = 0.8$ and $\theta_{10} = 0.2$). Such cases imply classifiers performing very poorly, and are thus impractical. For multiclass problems, future work is required to specify the cases where the determinant of the error rate matrix is zero, and their practical implications (e.g., poorly performing classifiers).

Negative effect

Random error rate variations can worsen the initial classification bias when applying error estimation methods, especially with the Misclassification method. This issue is addressed by Shieh (2009) with a method balancing the uncorrected classifier output $n'_{.y}$ and estimated $\widehat{n'_{x.}}$ in a linear combination, e.g., fitting the α parameter in (5.25). This approach is of interest for future work.

$$\widehat{n'_{x.,combined}} = \alpha\, \widehat{n'_{x.}} + (1 - \alpha)\, n'_{.y} \tag{5.25}$$

Small datasets

High variance is particularly critical for small datasets (Figure 5.1). Furthermore, biases may occur if class sizes are scarce, e.g., if $n_{x.}, n'_{x.}, n_{xy}$ or n'_{xy} are less than a few items (Figure 5.3, p.87 and Section 5.4.2, p.89). Further research is needed to identify the data sizes for which error estimation methods are not recommended, or linear combinations (5.25) are preferable (e.g., depending on error rate magnitudes). Cases where small n_{xy} yield error rate $\theta_{xy} \to 0$ or 1 should also be investigated (e.g., higher error rates may be preferable).

From the evaluation in Figure 5.1 (p.81), we observe that variance issues are critical when class sizes contain less than 500 items for the Misclassification method, or less than 100 items for the Reclassification method. Many applications cannot afford the costs of collecting extensive test sets, e.g., with more than 500 items per class. Hence applying error estimation methods may not be practical in many cases.

These issues with variance and insufficient test set sizes question the applicability of error estimation methods. More importantly, they question the representativity of classifier evaluation in general. If error measurements from test sets are not reliable enough to estimate the numbers of errors in target sets (i.e., due to high variance when applying error estimation methods), then the test sets do not provide reliable descriptions of the errors to expect when applying classifiers. **We stress that error rate variance significantly impacts the reliability of classifier assessments and is, however, largely overlooked when assessing classifiers.**

5.6.2 Test set representativity

The error estimation methods presented in this chapter rely on the assumption that test sets are representative of the target sets[5]. We discuss key test set characteristics that impact their representativity:

- **Test set size**, as random differences between test and target set error rates increase as test set size decreases (p.101)
- **Sampling method**, as test sets must represent the feature distributions and must be disjoint from target sets (p.102)

Test set size

Test set size is critical for test sets to be representative of classifier error rates. If test set size decreases, the test set error rates may differ further from the target set error rates, and the variance of error estimation results increases (Section 5.4). Small test sets are especially critical with extreme error rates (e.g., $\theta \to 0$ or 1) as small variations of ± 1 error can greatly impact the resulting error rates.

[5] In particular, the methods assume that test and target set error rates converge asymptotically to the same value as test and target set sizes increase.

Hence it is recommended to maximize test set sizes, e.g., by using cross-validation, unsupervised classification, or reduced training set sizes. Future work is required to establish strategies for maximizing the test set size, e.g., depending on the availability of groundtruth, and classifier characteristics.

For instance, reducing training set sizes may increase the classifier's error rates, e.g., as class models may become imprecise. However, increasing classifier's error rates may reduce the variance of error estimation results, as the numerator increases in $V(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n}$, and the risk of extreme error rates (e.g., $\theta \to 0$ or 1).

Although unintuitive, these results suggest that **classifiers with higher error rates** but tested with a larger test set (which yields lower variance), **may be preferable to classifiers with lower error rates** but tested with a smaller test set (which yields higher variance). However, increased error rates may yield classifiers approaching random classifiers, which may worsen variance issues (i.e., as determinants of error rate matrices tend to zero $|\mathbf{M}| \to 0$, Section 5.5 p.95).

Test set sampling

The sampling methods used to collect test sets must be carefully designed to ensure that test sets are representative of the potential target sets. Prior work dealt with test sets that are randomly sampled within the target set (i.e., for classifiers applied to a single target set) which ensures test sets' representativity. However, in machine learning problems, test sets are often disjoint from the target sets. Several issues arise if test sets are not sampled within a single target set, beside variance issues addressed in Section 5.4. For instance, class proportions may differ between test and target sets, and the Reclassification method may be inapplicable. Otherwise, error rates may systematically vary between test and target sets when:

- Training sets are used as test sets. The resulting error estimations may be biased, as in Saerens et al. (2001) with the Misclassification method. Our experiments in Sections 5.2, 5.3 and 5.5 use cross-validation where test and training sets are not strictly separated. No bias was observed in our experiments, however future work is required to investigate the impact of using cross validation, or strictly separated test and training sets.
- Quality improvement methods designed for training sets (e.g., reducing noise or excluding outliers) are applied to test sets.
- Test and target sets have different feature distributions, e.g., if target sets are of lower data quality (e.g., low image quality in computer vision).

For example, with the Fish4Knowledge system, varying feature distributions can occur if target sets contain many low quality images, while test set image quality is more balanced. The feature distributions of each class may systematically differ, e.g., different colors and contours due to lower contrast or fuzziness. Target sets with lower image quality may have higher error rates than test sets with higher image quality.

If feature distributions systematically vary between test and target sets, the assumption of equal error rates may be violated, and **none of the error estimation methods we presented may be applicable**. The critical impact of varying feature distributions is demonstrated in Section 5.6.3, and potential solutions are discussed in Section 5.7.

5.6.3 Varying feature distributions

Classifiers typically use feature distributions to build models of each class, i.e., describing the characteristics of the objects to classify. If feature distributions differ between test and target sets, error rates may differ too (e.g., if a target set has more low-contrast images, more images may be misclassified). This may worsen the classification biases when applying the bias correction methods introduced in Section 5.2. Figure 5.10 shows examples where a single feature is used, a score as in Figure 5.3 (p.87). Small variations of the feature distribution have created significant biases.

Hence **varying feature distributions are critical and must be assessed prior to applying bias correction** (Section 5.2) **and error estimation methods** (Section 5.3). For instance, the differences between the feature distributions of test and target sets may be assessed using distance metrics such as Mallows distance (Levina and Bickel 2001).

If test and target sets have similar class proportions, their joint feature distributions can be directly compared (i.e., their global feature distributions joining together item sets from all classes). If test and target sets have different class proportions, their joint feature distributions differ even if feature distributions are identical at the level of each class (e.g., even if all items from the same class have the same features). In this case, feature distributions must be compared for each class separately, i.e., comparing the feature distributions of the $n_{x.}$ and $n'_{x.}$ items that actually belong to the same class.

However, the actual classes of target set items are unknown. Hence, selecting the $n'_{x.}$ items actually belonging to class x is impossible, and only the $n'_{.x}$ items classified into class x are known. The $n'_{.x}$ items classified into class x may be used to approximate the feature distribution for class x. However, this feature estimator may be biased since the $n'_{.x}$ items classified as class x may include items that actually belong to other classes and exhibit different feature distributions. To address this issue, methods can be developed to identify the misclassified items (as discussed in Section 5.7.3) and exclude them when comparing class-specific feature distributions.

In addition to impacting the applicability of error estimation methods, **issues with varying feature distributions question the representativity of classifiers' evaluation in general**. If test set feature distributions do not support the estimation of the numbers of errors in target sets, then the test sets do not provide reliable descriptions of the errors to expect when applying classifiers.

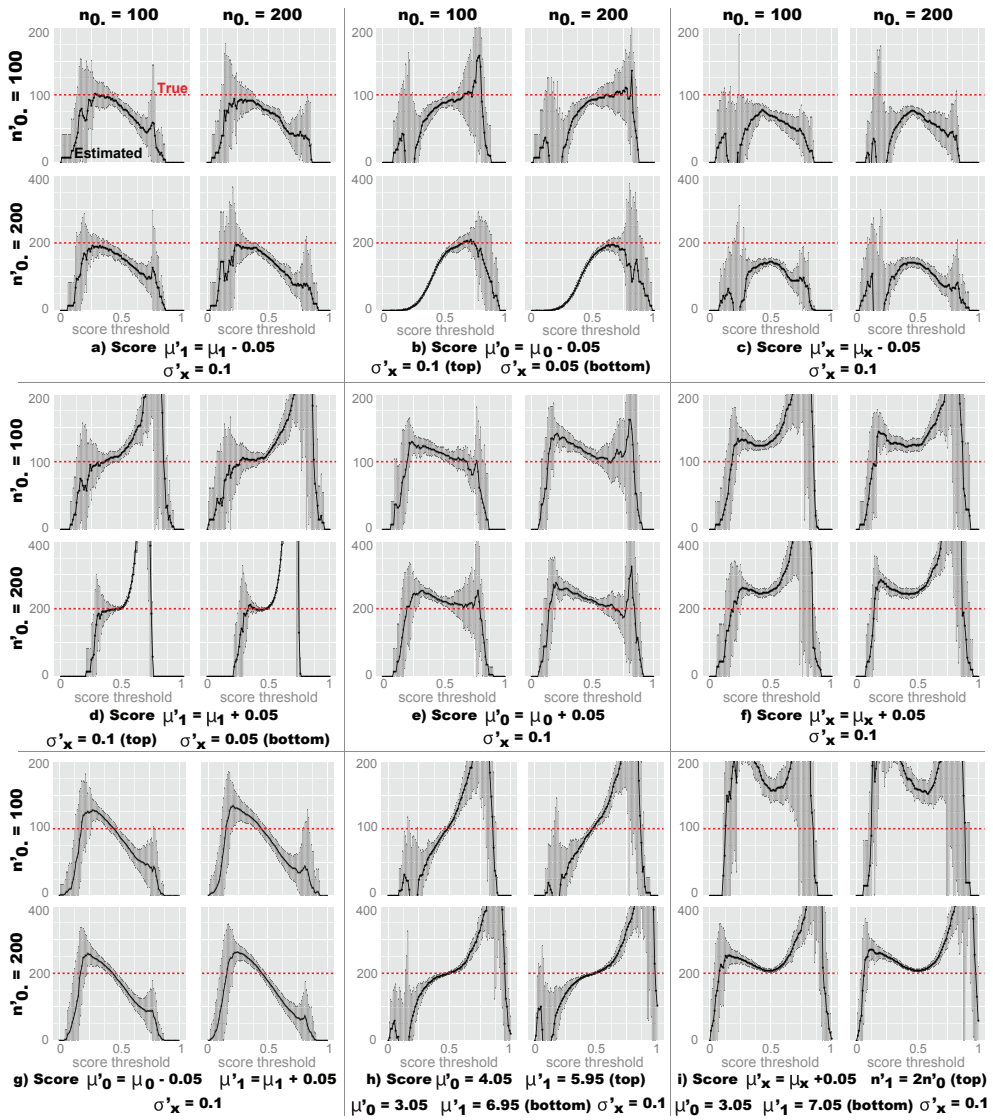

Figure 5.10: Results of Misclassification method for simulated data with varying feature distribution. As in Figure 5.3, a score threshold (x axis) is used to assign class 0 or class 1 to the items to classify. Class sizes $\widehat{n'_{0.}}$ (y axis) are estimated for 10^4 pairs of test and target sets. Test sets are randomly sampled with class proportions $n_{0.}$=$n_{1.}$, mean scores μ_0=0.4 for class 0, μ_1=0.6 for class 1, and score variance σ_x=0.1. Target sets are sampled from score distributions that differ from the test sets, with μ'_x=$\mu_x \pm 0.05$ and variance $\sigma'_x \in \{0.05, 0.1\}$, and with class proportions $n'_{0.}$=$2n'_{1.}$. Lower graphs -h) and -i) illustrate additional cases where μ'_x=$\mu_x \pm 0.1$.

Further work is required to develop methods for handling varying feature distributions. For binary classifiers providing threshold parameters, for example, the results in Figure 5.10 suggest that thresholds averaging the mean scores (i.e., $(\mu_0 + \mu_1)/2$ in our simulations) may minimise the biases due to varying feature distributions (and the variance in any case, as suggested in Figure 5.3, p.87). Future work may develop methods to derive optimal thresholds that are specifically adapted to the target sets, depending on their feature distributions.

Furthermore, information on test and target set feature distributions can be used to develop refined error estimation methods. This can be done with discrete or continuous approaches, as discussed in Section 5.7.

5.7 Future work

The methods discussed in the chapter are relatively unexplored within the machine learning domain. Thus our work opens several perspectives for future work, addressing problems such as:

- Correcting critical biases due to varying feature distributions (mentioned in Section 5.6.3) using discrete or continuous approaches (Sections 5.7.1 and 5.7.2)
- Identifying which individual items are misclassified, i.e., to refine the class assigned to individual items (Section 5.7.3).

5.7.1 Discrete approaches

As discussed in Section 5.6.3, and shown in Figure 5.10, if feature distributions differ between test and target sets (e.g., if the target set has lower image quality), error estimation methods may not be applicable as the equal error rate assumption may be violated.

Within the Fish4Knowledge project, this issue was addressed with a discrete approach. Discrete *types* of image quality are identified, and error rates are estimated for each type of image quality. Error estimation methods can then be applied using error rates measured for each image quality (Beauxis-Aussalet and Hardman 2015).

This discrete approach implies that test sets must be collected for each type of image quality. However, it may be difficult to collect sufficiently large test sets, e.g., containing examples for each species observed with each image quality. If test sets are small for each combination of species and image quality, the variance of error rates and error estimation results may increase significantly, and applying error estimation methods may not be appropriate.

5.7.2 Continuous approaches

Continuous approaches should be investigated in future work. They may improve the error estimation results, and address the issues identified in Section 5.6.3, without requiring to partition test sets into many discrete combinations of classes and features.

Instead, for example, linear models can be fitted to represent error rates as a function of feature distribution. This approach is discussed in Section B.

A. Logistic Regression method

We developed a continuous approach for handling varying feature distributions by fitting logistic regression models that represent error rates as a function of similarity measures (Boom et al. 2016). Similarity measures represent how similar an item is to a class model. They can be provided by certain classifiers, for each item and each class model. This Logistic Regression method is explained in the tutorials provided in additional materials (Section 5.9.3, Figures 5.11-5.12).

This prior work requires equal class proportions between test and target set, i.e., extending the Reclassification method. Future work is needed to develop methods for the case where class proportions differ between test and target sets, e.g., extending the Misclassification method.

B. Bayesian statistics method

Bayesian statistics may offer solutions for developing continuous approach compatible with the Misclassification method. Within the Bayesian framework, varying class proportions are equivalent to varying in *class prior probabilities*. The Misclassification method can estimate target sets' *class prior probabilities*, while Bayesian statistics can refine each item's class probabilities using its feature distributions. Hence Misclassification and Bayesian methods can be combined to address issues with varying class proportions (Section 5.6.3).

This approach is illustrated in equations (5.26)-(5.28), with variables defined in Tables 5.5). The Misclassification method is used to estimate class prior probabilities, providing the estimates in equation (5.28).

A similar approach is introduced (Saerens et al. 2001) without investigating the results' variance. Future work could investigate applications of the Sample-to-Sample method for estimating the results' variance of equation (5.28).

$$P(\zeta_x|F_i) = \frac{1}{P(F_i)}\ P(F_i|\zeta_x)\ P(\zeta_x) \qquad \text{thus} \qquad \begin{pmatrix} P(\zeta_1|F_i) \\ P(\zeta_2|F_i) \\ \dots \\ P(\zeta_x|F_i) \end{pmatrix} = \frac{1}{P(F_i)} \begin{pmatrix} P(F_i|\zeta_1) \\ P(F_i|\zeta_2) \\ \dots \\ P(F_i|\zeta_x) \end{pmatrix} \begin{pmatrix} P(\zeta_1) \\ P(\zeta_2) \\ \dots \\ P(\zeta_x) \end{pmatrix} \tag{5.26}$$

$$\begin{pmatrix} \widehat{P'}(\zeta_1) \\ \widehat{P'}(\zeta_2) \\ \dots \\ \widehat{P'}(\zeta_x) \end{pmatrix} = \begin{pmatrix} P(\zeta_{1\to1}|\zeta_1) & P(\zeta_{2\to1}|\zeta_2) & \dots & P(\zeta_{x\to1}|\zeta_x) \\ P(\zeta_{1\to2}|\zeta_1) & P(\zeta_{2\to2}|\zeta_2) & \dots & P(\zeta_{x\to2}|\zeta_x) \\ \dots & \dots & \dots & \dots \\ P(\zeta_{1\to x}|\zeta_1) & P(\zeta_{2\to x}|\zeta_2) & \dots & P(\zeta_{x\to x}|\zeta_x) \end{pmatrix}^{-1} \begin{pmatrix} P'(\zeta_{\to1}) \\ P'(\zeta_{\to2}) \\ \dots \\ P'(\zeta_{\to x}) \end{pmatrix} \tag{5.27}$$

$$\begin{pmatrix} \widehat{P'}(\zeta_1|F_i) \\ \widehat{P'}(\zeta_2|F_i) \\ \dots \\ \widehat{P'}(\zeta_x|F_i) \end{pmatrix} = \frac{1}{P'(F_i)} \begin{pmatrix} P(F_i|\zeta_1) \\ P(F_i|\zeta_2) \\ \dots \\ P(F_i|\zeta_x) \end{pmatrix} \begin{pmatrix} P(\zeta_{1\to1}|\zeta_1) & P(\zeta_{2\to1}|\zeta_2) & \dots & P(\zeta_{x\to1}|\zeta_x) \\ P(\zeta_{1\to2}|\zeta_1) & P(\zeta_{2\to2}|\zeta_2) & \dots & P(\zeta_{x\to2}|\zeta_x) \\ \dots & \dots & \dots & \dots \\ P(\zeta_{1\to x}|\zeta_1) & P(\zeta_{2\to x}|\zeta_2) & \dots & P(\zeta_{x\to x}|\zeta_x) \end{pmatrix}^{-1} \begin{pmatrix} P'(\zeta_{\to1}) \\ P'(\zeta_{\to2}) \\ \dots \\ P'(\zeta_{\to x}) \end{pmatrix} \quad (5.28)$$

$P(\zeta_x)$	Probability that an item truly belongs to class x (*prior probability*)
$P(\zeta_{\to y})$	Probability that an item is classified into class y
$P(\zeta_{x\to y}\|\zeta_x)$	Probability that an item is classified into class y, given that it truly belongs to class x
$P(F_i)$	Probability that an item exhibits the set of features F_i
$P(F_i\|\zeta_x)$	Probability that an item exhibits the set of features F_i, given that it truly belongs to class x
$P(\zeta_x\|F_i)$	Probability that an item truly belongs to class x, given its set of features F_i (*posterior probability*)
$n_{xy,i}$	Number of items truly belonging to class x, classified into class y, and exhibiting the set of features F_i
$n_{xy,.}$	Number of items truly belonging to class x, classified into class y, and exhibiting any kind of features
$n_{x.,.}$	Number of items truly belonging to class x, classified into any class, and exhibiting any kind of features
$n_{..,.}$	Number of items truly belonging to any class, classified into any class, and exhibiting any kind of features (i.e., total number of items)

$$P(\zeta_x) = \frac{n_{x.,.}}{n_{..,.}} \qquad P(\zeta_{\to y}) = \frac{n_{.y,.}}{n_{..,.}} \qquad P(\zeta_{x\to y}|\zeta_x) = \frac{n_{xy,.}}{n_{x.,.}}$$

$$P(F_i) = \frac{n_{..,i}}{n_{..,.}} \qquad P(\zeta_x|F_i) = \frac{n_{x.,i}}{n_{..,i}} \qquad P(F_i|\zeta_x) = \frac{n_{x.,i}}{n_{x.,.}}$$

Table 5.5: Definitions of variables used in equations (5.26)-(5.28). Variables using prime symbols, e.g., $P'(\zeta_x)$, refer to target sets. Without prime symbols, variables refer to the test set.

The feature probabilities $P(F_i)$ and $P(F_i|\zeta_x)$ may not be drawn from the discrete but impractical approach in Table 5.5. This discrete approach requires collecting test set items that represent all the possible sets of features F_i. Instead, linear models can be fit on the features measured in test and target sets, to derive continuous feature probabilities.

5.7.3 Identify the misclassified items

Given the estimated error composition (Section 5.3), i.e., the numbers of errors n'_{xy}, methods can be derived for identifying the misclassified items individually, and correcting their assigned class. Probabilistic classifiers such as Bayesian classifiers are of interest to address this problem. Classifiers providing similarity measures, i.e., representing how items are similar to classes' model, are also of interest.

For example, lets consider the error estimation results estimating that n'_{xy} items are misclassified into class y while belonging to class x, and items' probabilities of class membership, e.g., $\widehat{P'}(\zeta_x|F_i)$ in equation (5.28). Within the items classified as class y, we can select the n_{xy} items with the highest probability of belonging to class x.

Alternatively, provided with similarity measures, we can select items with the highest similarity to class x model.

However, the problem is not as simple as it may seems from this example because a single item can have a high probability of class membership (or similarity) for several classes. For example, when selecting the n_{xy} and n_{zy} items with the highest chances of belonging to classes x and z, the same items may be selected for both classes x and z.

5.8 Conclusion

We demonstrated the applicability of existing *error estimation* methods to machine learning classification problems (Section 5.2). We extend existing methods, designed to estimate unbiased class sizes, to estimating numbers of errors in target sets, and introduce an alternative method called Ratio-to-TP (Section 5.3). Given the $n'_{.y}$ items classified as class y, the extended methods estimate how many n'_{xy} items truly belong to class x. Such estimation of the error composition describes classification uncertainty beyond accuracy and metrics such as precision or False Positive rate (Section 5.3).

The results of error estimation methods are subject to potentially high variance due to random error rate variations. For small datasets, the variance magnitude is critical and applying error estimation methods may worsen the initial biases. To address such issues, we introduced a novel variance estimation method called Sample-to-Sample. We demonstrate that for disjoint test and target sets, variance estimation must account for the class sizes in both test and target sets. The Sample-to-Sample method provides accurate confidence intervals describing the variance of error estimation results (Section 5.4).

Finally, we introduce a promising method for predicting the variance of error estimations without prior knowledge of the potential target sets. We observe correlations between the determinant of error rate matrices and the variance of error estimation results. We thus postulate that determinants of error rate matrices are predictors of error estimation variance. If validated in future work, this predictor can be used to compare and choose classifiers that minimize the error estimation variance (Section 5.5).

This chapter addressed requirement 4-c in Chapter 2 (*extrapolate uncertainty in specific datasets*, p.36) and answered our fifth research question: *How can we estimate the magnitudes of classification errors in end-results?*.

The methods we introduced can assess the *Noise and Bias* due to classification errors, and the *Uncertainty in Specific Datasets* which are key uncertainty factors identified in Chapter 4. They are compatible with the methods we developed to handle *Fragmentary Processing*, another key uncertainty factor identified in Chapter 4 (i.e., they provide class size estimates that can be used within the metrics in equations (4.1)-(4.5), p.71).

We identified conditions that can impact the applicability of error estimation methods, i.e., class sizes (in both test and target sets), error rate magnitudes, number of classes, and *random* or *systematic* variations of error rates, class proportions and feature distributions. These findings inform the choice of methods depending on the use case at hand.

However, future work is required to formally identify inapplicable cases and quantify the test and target set characteristics that invalidate error estimation methods (Section 5.6). The most critical applicability issues concern small test or target sets and varying feature distributions. To address the latter, directions for future work are identified (Section 5.7).

We underline that **issues with the applicability of error estimation methods question the representativity of classifier evaluation in general. If test sets do not support the estimation of the numbers of errors in target sets, then the test sets do not provide reliable descriptions of the errors to expect when applying classifiers.** For instance, variance issues are critical with small datasets. However, error rate variance is seldom considered when assessing classifiers.

5.9 Additional materials

5.9.1 Code

The R code used to apply and evaluate the methods described in this paper is available online, free of use: *https://github.com/emma-cwi/classification_error*

5.9.2 Application of Fieller's theorem

Fieller's theorem (Fieller 1954) defines the confidence intervals limits $[\ell^-, \ell^+]$ for a ratio of correlated random variables A/B as (5.29), with z=1 for 68% confidence level.

$$\ell^{\pm} = \frac{(\mu_A\mu_B - z^2\sigma_{A,B}) \pm \sqrt{(\mu_A\mu_B - z^2\sigma_{A,B})^2 - (\mu_A^2 - z^2\sigma_A^2)(\mu_B^2 - z^2\sigma_B^2)}}{\mu_B^2 - z^2\sigma_B^2} \tag{5.29}$$

In Section 5.4.3, for estimating $\widehat{n'_{0.}}$, $A = n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}$ and $B = 1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}$ (equation (5.15) p.89). The mean, variance, covariance of A, B are detailed below, knowing that $\widehat{\boldsymbol{\theta'_{01}}}$ and $\widehat{\boldsymbol{\theta'_{10}}}$ are independent with null covariance.

$$\mu_B = E\left[1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}\right] \qquad \widehat{\mu_B} = 1 - \theta_{01} - \theta_{10}$$

$$\sigma_B^2 = V\left(1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}\right) = V\left(\widehat{\boldsymbol{\theta'_{01}}}\right) + V\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$

$$\widehat{\sigma_B^2} = \frac{\theta_{01}(1-\theta_{01})}{n_{0.}} + \frac{\theta_{01}(1-\theta_{01})}{\widehat{n'_{0.}}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}}$$

$$\mu_A = E\left[n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right] \qquad \widehat{\mu_A} = n'_{.0} - \theta_{10}\, n'_{..}$$

$$\sigma_A^2 = V\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right) = n'^2_{..}\, V\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$

$$\widehat{\sigma_A^2} = n'^2_{..}\left(\frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}}\right)$$

$$\sigma_{A,B} = Cov\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..},\ 1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}\right) = n'_{..}\, V\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$

$$\widehat{\sigma_{A,B}} = n'_{..}\left(\frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}}\right)$$

In Section 5.4.4, for estimating $\widehat{n'_{01}}$, $A = \widehat{\boldsymbol{\theta'_{01}}}\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right)$. B remains unchanged. Their mean, variance, covariance are detailed below. The covariance of products of random variables is drawn from Bohrnstedt and Goldberger (1969).

$$\mu_A = E\left[\widehat{\boldsymbol{\theta'_{01}}}\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right)\right] \qquad \widehat{\mu_A} = \theta_{01}\left(n'_{.0} - \theta_{10}\, n'_{..}\right)$$

$$\sigma_A^2 = E\left[\widehat{\theta'_{01}}\right]^2 V\left(n'_{.0} - \widehat{\theta'_{10}}\, n'_{..}\right) + E\left[n'_{.0} - \widehat{\theta'_{10}}\, n'_{..}\right]^2 V\left(\widehat{\theta'_{01}}\right) + V\left(\widehat{\theta'_{01}}\right) V\left(n'_{.0} - \widehat{\theta'_{10}}\, n'_{..}\right)$$

$$\widehat{\sigma_A^2} = \theta_{01}^2 n'^2_{..} \widehat{V}\left(\widehat{\theta'_{10}}\right) + \left(n'_{.0} - \theta_{10}\, n'_{..}\right)^2 \widehat{V}\left(\widehat{\theta'_{01}}\right) + n'^2_{..} \widehat{V}\left(\widehat{\theta'_{01}}\right) \widehat{V}\left(\widehat{\theta'_{10}}\right)$$

$$\sigma_{A,B} = n'_{..}\left(Cov\left(\widehat{\theta'_{01}}\,\widehat{\theta'_{10}}, \widehat{\theta'_{01}}\right) + Cov\left(\widehat{\theta'_{01}}\,\widehat{\theta'_{10}}, \widehat{\theta'_{10}}\right)\right) - n'_{.0} V\left(\widehat{\theta'_{10}}\right)$$

$$Cov\left(\widehat{\theta'_{xy}}\,\widehat{\theta'_{yx}}, \widehat{\theta'_{xy}}\right) = E\left[\widehat{\theta'_{xy}}\right] Cov\left(\widehat{\theta'_{yx}}, \widehat{\theta'_{xy}}\right) + E\left[\widehat{\theta'_{yx}}\right] Cov\left(\widehat{\theta'_{xy}}, \widehat{\theta'_{xy}}\right)$$
$$= E\left[\widehat{\theta'_{yx}}\right] V\left(\widehat{\theta'_{xy}}\right)$$

$$\widehat{\sigma_{A,B}} = n'_{..}\left(\theta_{10} \widehat{V}\left(\widehat{\theta'_{01}}\right) + \theta_{01} \widehat{V}\left(\widehat{\theta'_{10}}\right)\right) - n'_{.0} \widehat{V}\left(\widehat{\theta'_{01}}\right)$$

With $\widehat{V}\left(\widehat{\theta'_{xy}}\right) = \dfrac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \dfrac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}}$ (Sample-to-Sample)

In Section 5.4.5, for estimating $\widehat{\pi_0} = \widehat{n'_{0.}}/n'_{..}$ by Shieh (2009) and Buonaccorsi (2010), $A = n'_{.0}/n'_{..} - \widehat{\theta'_{10}}$ (equation (5.17) p.91). B remains unchanged. The mean, variance, covariance used by Shieh (2009) and Buonaccorsi (2010) are restated below.

$$\mu_A = n'_{.0}/n'_{..} - \theta_{10}$$

$$\sigma_A^2 = \frac{n'_{.0}/n'_{..}\left(1 - n'_{.0}/n'_{..}\right)}{n'_{..}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}}$$

$$\sigma_B^2 = \frac{\theta_{01}(1-\theta_{01})}{n_{0.}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}}$$

$$\sigma_{A,B} = \frac{\theta_{10}(1-\theta_{10})}{n_{1.}}$$

5.9.3 Tutorials explaining the Logistic Regression method

As discussed in Section 5.7.2, error estimation methods can be refined using the feature distributions of the items to classify. The Logistic Regression method uses a linear model (i.e., logistic regression) to represent error rate distributions as a function of similarity measures provided by the classifiers. Similarity measures represent how similar an item is to a class model, i.e., how an item's features are similar to the class model's features. The Logistic Regression method is explained in Boom et al. (2016) and in the tutorials shown in Figures 5.11 and 5.12.

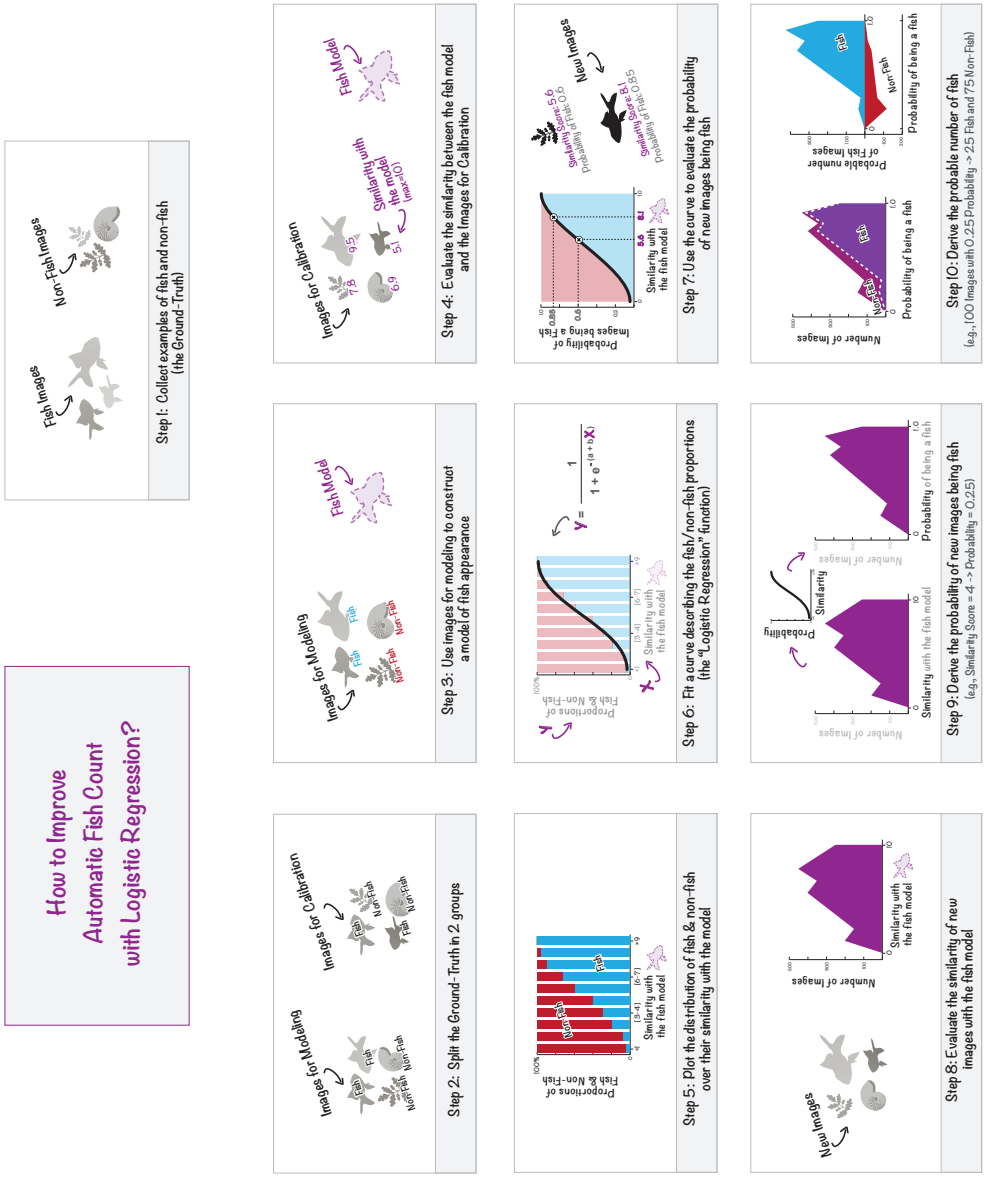

Figure 5.11: Logistic regression methods for binary problems, designed by the author of this thesis (Boom et al. 2016).

Figure 5.12: Logistic regression methods for multiclass problems, designed by the author of this thesis (Boom et al. 2016).

Chapter 6

Visualization of Classification Errors

Classifiers are applied in many domains where errors have significant implications, e.g., medicine, security, eScience. However, end-users may not always understand classification errors and their impact (Chapter 3, Section 3.4.2, p.48). Existing error visualizations primarily address the needs of classification experts who aim at improving classifiers. These visualizations may not address the specific needs of end-users, especially those with limited expertise in classification technologies. We thus investigate visualizations that address the needs of non-expert end-users, and answer our sixth question: *How can visualization support non-expert users in understanding classification errors?* (Section 1.4).

We first introduce end-users requirements (Section 6.1) and identify information needs that pertain to either end-users or developers (Section 6.2). We then discuss existing visualizations of classification errors and the end-users' or developers' needs they address (Section 6.3). We introduce a visualization design named Classee (Figures 6.1-6.4), that aims at addressing specific needs of end-users (Section 6.4). We evaluate this design with users from three levels of expertise, and compare it to ROC curves and confusion matrices (Section 6.5). From the quantitative results, we discuss users' performance w.r.t. the type of visualization and users' level of expertise (Section 6.6). From the qualitative results, we identify key difficulties with understanding the classification errors, and how visualizations address or aggravate them (Section 6.7).

Abbr.	*Correctness*	*Prediction*	*Definition*
FP	**False**	**Positive**	Object classified into the *Positive* class (i.e., as the class of interest) while actually being *Negative* (i.e., belonging to a class other than the *Positive* class).
TP	**True**	**Positive**	Object correctly classified into the *Positive* class.
FN	**False**	**Negative**	Object classified into the *Negative* class while actually belonging to the *Positive* class.
TN	**True**	**Negative**	Object correctly classified into the *Negative* class.

Table 6.1: Definition of FP, TP, FN, TN.

6.1 End-user requirements

To support end-users' understanding of classification errors, visualizations must provide accessible information requiring little to no prior knowledge of classification technologies. The information provided must be relevant for end-users' data analysis tasks, e.g., clarifying the practical implications of classification errors without providing unnecessary details. This requirement was identified in Chapter 2 (requirement 4-d, p.36).

User information needs primarily concern the estimation of numbers of errors to expect in classification end-results, for each class of interest (Chapter 3, Section 3.5.1, p.50). Users also expressed concerns regarding error variability, i.e., random variance due to random differences among datasets, as well as systematic error rates differences due to lower data quality. Our findings in Chapter 5 confirmed users' concerns, as we demonstrated that random and systematic differences among datasets significantly impact the magnitude of errors to expect in classification end-results.

Our findings in Chapter 5 also demonstrated that class proportions (i.e., the relative magnitudes of class sizes) impact the magnitudes of errors. In particular, one class's size directly impacts the magnitude of its False Negatives, i.e., items that actually belong to this class but are classified into another class. The larger the class, the larger the False Negatives it generates. These misclassified False Negatives are also False Positives from the perspective of the class into which they are classified. The transfer of items *from* their actual class (as False Negatives) *into* their predicted class (as False Positives) is the core mechanism of classification errors.

To understand the impact of classification errors, it is crucial to assess the *error directionality*, i.e., the actual class *from* which errors originate, and the predicted class *into* which errors are classified. Error directionality reflects the two-fold impact of classification errors: items are *missing* from their actual class, and are *added* to their predicted class.

Hence end-user-oriented visualizations of classification errors must address 5 key requirements:

- **R1: Provide the magnitude of errors for each class**.
- **R2: Provide the magnitude of each class size**.

- **R3: Detail the error directionality**, i.e., errors' true and predicted classes, and the magnitude of errors for all combinations of true and predicted classes.
- **R4: Estimate how the errors measured in test sets may differ from the errors that actually occur when applying the classifier to another dataset**, e.g., considering random error rate variance, and bias due to lower data quality or varying feature distributions.
- **R5: Omit unnecessary technical details**, e.g., about the underlying classification technologies, especially details not related to estimating the errors in classification end-results.

	Task			**Visualization**		
	Improve Model and Algorithm	Tune Classifier	Estimate Errors in End-Results	Confusion Matrix	Precision-Recall and ROC curves	Classee
Target Audience						
End-Users		X	X			X
Developers	X	X		X	X	X
Low-Level Metric						
Raw Numbers	X	X	X	X		X
ROC-like Error Rates in Equation (6.1)	X	X	X		X	X
Precision-like Error Rates in Equation (6.2)	X	X	X[1]		X	X
Accuracy (6.3)	X	X				X
AUC	X				X	X[2]
High-Level Information						
Total Number of Errors	X	X		X	X	X
Errors over Tuning Parameter	X	X			X	X
Errors over Object Features	X		X[3]			X[4]
Error Composition	X	X	X	X	X	X
Class Proportions		X	X	X		X
Class Sizes		X	X	X		X

[1] *If class proportions are equal (Chapter 5).*
[2] *Barcharts' areas show information similar to AUC.*
[3] *Features distributions can be used to tune error estimates (Boom et al. 2016), and verify issues with varying distributions (Chapter 5).*
[4] *Objects' features can be used as the x-axis dimension.*

Table 6.2: Relationships among users, tasks, information needs, metrics and visualizations

6.2 Information needs

We identified key information needs through interviews of machine learning experts and end-users, reported in Chapters 2 and 3, and synthesized in Chapter 4. We found that the needs of developers and end-users have key differences and overlaps (Table 6.2).

Developers often seek to optimise classifiers on all classes and all types of error (e.g., limiting both FP and FN). They often use metrics that summarize the errors over all classes, e.g., accuracy shown in equation (6.3). For example, they measure the Area Under the Curve (AUC) (Fawcett 2006) to summarise all types of errors (FN and FP) over all possible values of a tuning parameter. This approach is irrelevant for end-users who apply classifiers that are already tuned with fixed parameter values (Requirement R5, Section 6.1).

Metrics that summarize all types of errors for all classes (e.g., AUC, Accuracy) fail to convey *"the circumstances under which one classifier outperforms another"* (Drummond and Holte 2006), e.g., for which classes, class proportions (e.g., rare or large classes, Requirement R2), error directions (e.g., the composition of errors between all possible classes, Requirement R3) and values of the tuning parameters. These characteristics are crucial for end-users: specific classes and types of errors can be more important than others; class proportions may vary in end-usage datasets; and optimal tuning parameters depend on the classes and errors of interest, and on the class proportions in the datasets to classify.

End-users are also interested in extrapolating the errors in their end-usage datasets (e.g., within the objects classified as class Y how many truly belong to class X?). Such extrapolation depends on class sizes, class proportions and error directions, and can be refined depending on the features of classified objects as discussed in Chapter 5 (Requirement R4).

6.3 Related work

Existing visualizations - Recent work developed visualizations to improve classification models (Liu et al. 2017, Krause et al. 2017, Elzen and Wijk 2011), e.g., using barcharts (Ren et al. 2017, Alsallakh et al. 2014). They are algorithm-specific (e.g., applicable only to probabilistic classifiers or decision trees) but end-users may need to compare classifiers based on different algorithms. These comparisons are easier with algorithm-agnostic visualizations, i.e., using the same representations for all algorithms, and limiting complex and unnecessary information on the underlying algorithms (Requirement R5, Section 6.1).

Confusion matrices, ROC curves and Precision-Recall curves are well-established algorithm-agnostic visualizations (Fawcett 2006) but they are intended for machine learning experts and simplifications may be needed for non-experts (e.g., understanding ROC curve's error rates may be difficult, especially for multiclass data). Furthermore, ROC curves and Precison-Recall curves omit the class sizes although

this is a crucial information for understanding the errors to expect in classification end-results (Requirement R2).

Cost curves (Drummond and Holte 2006) are algorithm-agnostic and investigate specific end-usage conditions (e.g., class proportions, costs of errors) but they are also complex, intended for experts, omit class sizes (Requirement R2), and do not address multiclass data. The non-expert-oriented visualizations in Micallef et al. (2012), Khan et al. (2015) use simpler trees, grids, Sankey or Euler diagrams, but are illegible with multiclass data due to multiple overlapping areas or branches.

Choice of error metrics - Different error metrics have been developed and their properties address different requirements (Sebastiani 2015, Hossin and Sulaiman 2015, Sokolova and Lapalme 2009). Error metrics are usually derived from the same underlying data: numbers of correct and incorrect classifications encoded in confusion matrices, and measured with a *test set* (a data sample for which the actual class is known). These raw numbers provide simple yet complete metrics. They are easy to interpret (no formula involved) and address most requirements for reliable and interpretable metrics, e.g., they do not conceal the impact of class proportions on error balance, and have known values for *perfect, pervert* (always wrong) and *random* classifiers (Sebastiani 2015). These values depend on the class sizes in the test set, which is not recommended in Sebastiani (2015). However, raw numbers convey the class sizes, omitted in rates, but needed to assess the class imbalance and statistical significance of error measurements (Requirement R2). These are crucial for estimating the errors to expect in end-usage applications, as discussed in Chapter 5.

Using raw numbers of errors, we focus on conveying basic error rates in equations (6.1)-(6.2) where n_{xy} is the number of objects actually belonging to class x and classified as class y (i.e., errors if $x \neq y$), $n_{x.}$ is the number of objects actually belonging to class x (actual class size), and $n_{.y}$ is the number of objects classified as class y (predicted class size). Accuracy is a widely-used metric summarizing errors over all classes, as shown in (6.3) where n_{xx} is the number of objects correctly classified as class x, and $n_{..}$ is the total number of objects for all classes. We also consider conveying accuracy, and focus on overcoming its bias towards large classes (Hossin and Sulaiman 2015) and missing information on class sizes (Requirement R2) and error directionality, e.g., high accuracy can conceal significant errors for specific classes (Requirement R3).

$$\text{Error rates w.r.t. actual class size (e.g., ROC curves):} \quad \frac{n_{xy}}{n_{x.}} \tag{6.1}$$

$$\text{Error rates w.r.t. predicted class size (e.g., Precision):} \quad \frac{n_{xy}}{n_{.y}} \tag{6.2}$$

$$\text{Accuracy:} \quad \frac{\sum_x n_{xx}}{n_{..}} \quad \text{e.g., for binary data:} \quad \frac{TP+TN}{TP+TN+FP+FN} \tag{6.3}$$

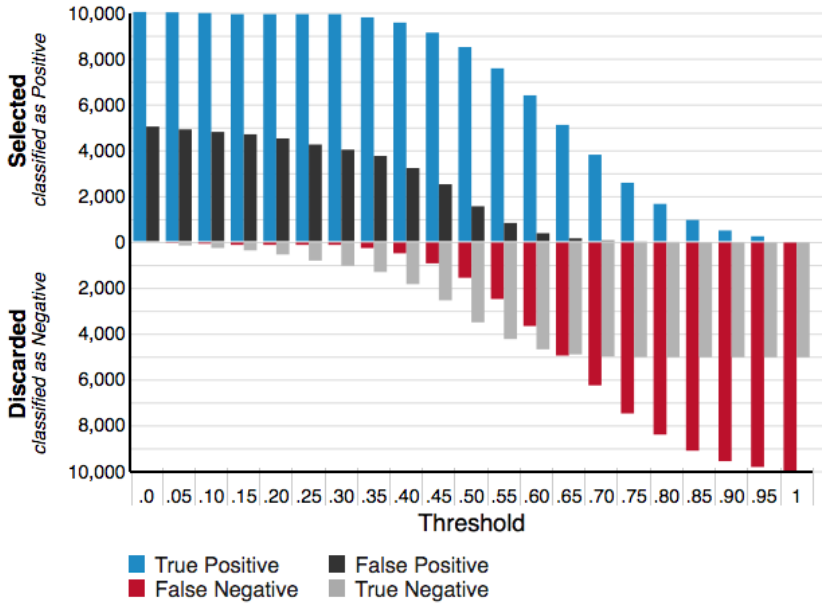

Figure 6.1: Classee visualization of classification errors for binary data.

6.4 Classee visualization

The Classee project simplified the visualization of classification errors by using ordinary barcharts and raw numbers of errors (Figure 6.1-6.4). The *actual* class and the error types are differentiated with color codes: vivid colors if the *actual* class is positive (blue for TP, red for FN), desaturated colors if the *actual* class is negative (grey for TN, black for FP). The bars' positions reinforces the perception of the actual class, as bars representing items from the same actual class are staked on each other into a continuous bar: TP above FN, and FP above TN (Figure 6.2 left). The zero line distinguishes the *predicted* class: TP and FP are above the zero line, FN and TN are below (Figure 6.2 right).

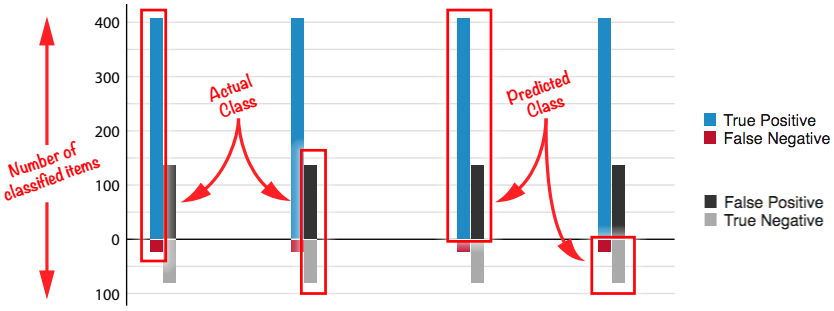

Figure 6.2: Bars representing the actual and predicted classes.

For binary data (Figure 6.1), objects from the same actual class are stacked in distinct bars: TP above FN for the positive class, and FP above TN for the negative class. Basic error rates can easily be interpreted visually (Figure 6.3). ROC curve's error rates in equation (6.1) are visualized by comparing the blocks within continuous bars: blue/red blocks for TP rate, black/grey blocks for FP rate. Precision-like rates in equation (6.2) are visualized by comparing adjacent blocks on each side of the zero line: blue/black blocks for Precision, red/grey blocks for False Omission Rate. Accuracy, i.e., equation (6.3), can be interpreted by comparing blue and grey blocks against red and black blocks, which is more complex. However, it overcomes key issues with accuracy (Hossin and Sulaiman 2015) by showing the error balance between FP and FN, and potential imbalance between large and small classes. The visualization also renders information similar to Area Under the Curve (Fawcett 2006) as blue, red, black and grey areas can be perceived.

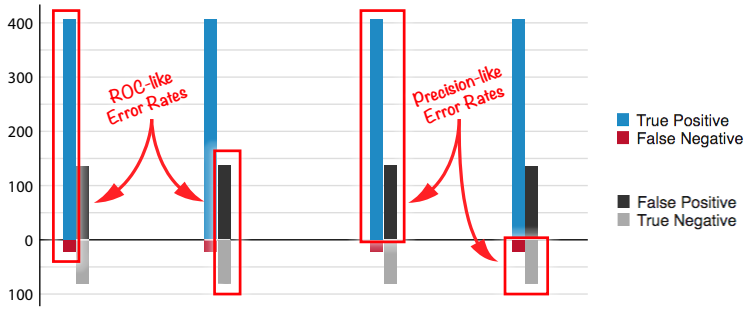

Figure 6.3: Bars showing basic error rates in equations (6.1)-(6.2).

Perceiving ROC-like rates (6.1) requires comparing *divided* and *adjacent* blocks. It can lower perception accuracy (Talbot et al. 2014) compared to unadjacent blocks in Ren et al. (2017) (TP rates rendered with separated TP and FN blocks) or (Alsallakh et al. 2014) (FP rates rendered with separated TN and FP blocks). However, Classee shows *part-to-whole* ratios while (Talbot et al. 2014) researched *part-to-part* ratios, and suggests that perceiving *part-to-whole* is more intuitive and effective. Further, Classee lets users compare the positions of bar extremities to the zero line, and perceiving positions is more accurate than perceiving relative bar lengths (Cleveland and McGill 1984). Precision-like rates (6.2) are perceived using *aligned* and *adjacent* blocks. It supports more accurate perceptions (Talbot et al. 2014, Cleveland and McGill 1984) compared to divided unadjacent blocks in Ren et al. (2017), Alsallakh et al. (2014).

For multiclass data (Figure 6.4), errors are shown for each class in a one-vs-all reduction, i.e., considering one class as the positive class and all other classes as the negative class, and so for all classes (e.g., for class x, $\mathrm{FP} = \sum_{y \neq x} n_{yx}$ and $\mathrm{TN} = \sum_{y \neq x} \sum_{z \neq x} n_{yz}$). TN are not displayed because they are typically of far greater magnitude, especially with large numbers of classes, which can reduce other bar sizes to illegibility. TN are also misleading as they do not distinguish correct and incorrect classifications (e.g., n_{zz} and $n_{yz,y \neq z}$). Without TN, FP are stacked on TP which shows the Precision for each class.

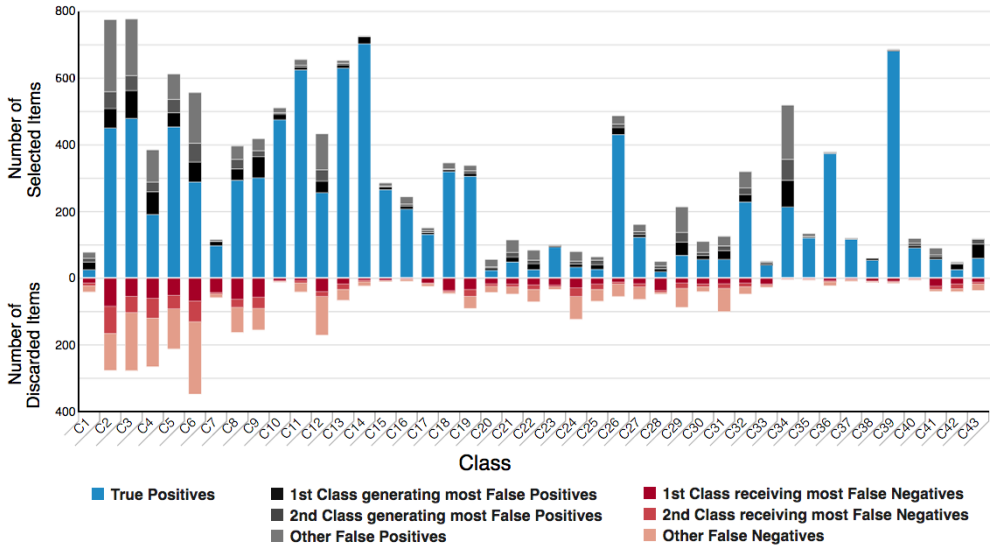

Figure 6.4: Classee visualization of classification errors for multiclass data.

Basic error rates can easily be interpreted visually (Figure 6.3), using the same principles as for binary classification. ROC curve's error rates in equation (6.1) are visualized by comparing the blue and red blocks (representing the actual class, Figure 6.5 left). Precision-like rates in equation (6.2) are visualized by comparing the blue/black blocks (representing the predicted class, Figure 6.5 middle).

Accuracy can be interpreted by comparing all blue blocks against either all red blocks, or all black blocks (the sum of errors for all red blocks is the same for all black blocks, as each misclassified object is a FP for its predicted class and a FN for its actual class). Users can visualize the relative proportions of correct and incorrect classifications, although the exact equation of accuracy (6.3) is harder to interpret. However, Classee details the errors between each class, which are omitted in accuracy.

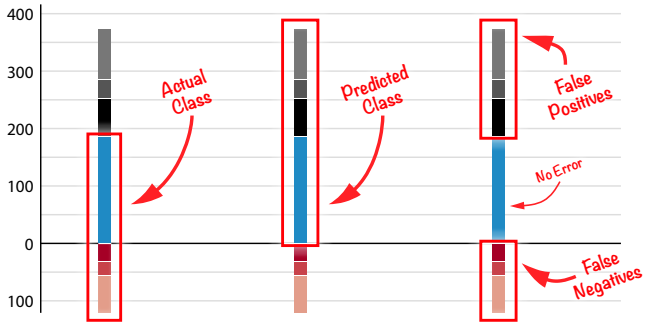

Figure 6.5: Bars representing the actual and predicted classes.

Compared to (Ren et al. 2017) stacking TP-FP-FN in this order, Classee stacking facilitates the interpretation of TP rates (6.1) and actual class sizes by showing continuous blocks for TP and FN (Figure 6.5 left). Compared to chord diagrams in (Alsallakh et al. 2014) encoding error magnitudes with surface sizes, Classee uses bar length to support more accurate perceptions of error magnitudes (Cleveland and McGill 1984).

Inspecting the error directionality, i.e., the magnitude of errors between specific classes, is crucial for understanding the impact of errors in end-results (Requirement R3, Section 6.1). Users need to assess the errors between specific classes and their *directionality* (i.e., errors *from* an actual class are misclassified *into* a predicted class). If errors between two classes are of significant magnitudes, it creates biases in the end-results (Chapter 5). For example, errors from large classes can result in FP of significant magnitude for small classes that are thus over-estimated. Such biases can be critical for end-users' applications.

Hence Classee details the error composition between actual and predicted classes. The FP blocks are split in sub-blocks representing objects from the same actual class. The FN blocks are also split in sub-blocks representing objects classified into the same predicted class. To avoid showing too many unreadable sub-blocks, Classee shows the 2 main sources of errors in distinct sub-blocks and merges the remaining errors in a 3rd sub-block. The FP sub-blocks show the 2 classes from which most FP actually belong, and the remaining FP as a 3rd sub-block. The FN sub-blocks show the 2 classes into which most FN are classified, and the remaining FN as a 3rd sub-block. Future implementations could let users control the number of sub-blocks to display, and the *boxes* in Ren et al. (2017) may improve their rendering.

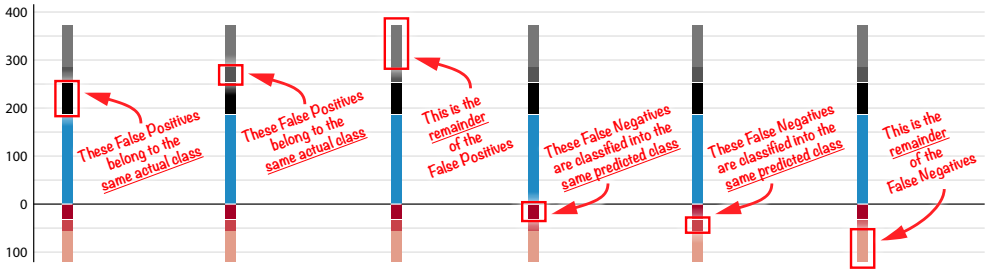

Figure 6.6: Bars representing the actual and predicted classes.

Users can select a class to inspect its errors (Figure 6.7). It shows which classes receive the FN and generate the FP. The FN sub-blocks of the selected class are highlighted within the FP sub-blocks of their predicted class. The FP sub-blocks are highlighted within the FN sub-blocks of their predicted class. Users can identify the error *directionality*, i.e., they can differentiate *Class X objects misclassified into Class Y* and *Class Y objects misclassified into Class X* (e.g., in Figure 6.7, objects from class C6 are misclassified into C34, but not from C34 into C6). Future implementations could also highlight the remaining FN and FP merged in the 3rd sub-blocks.

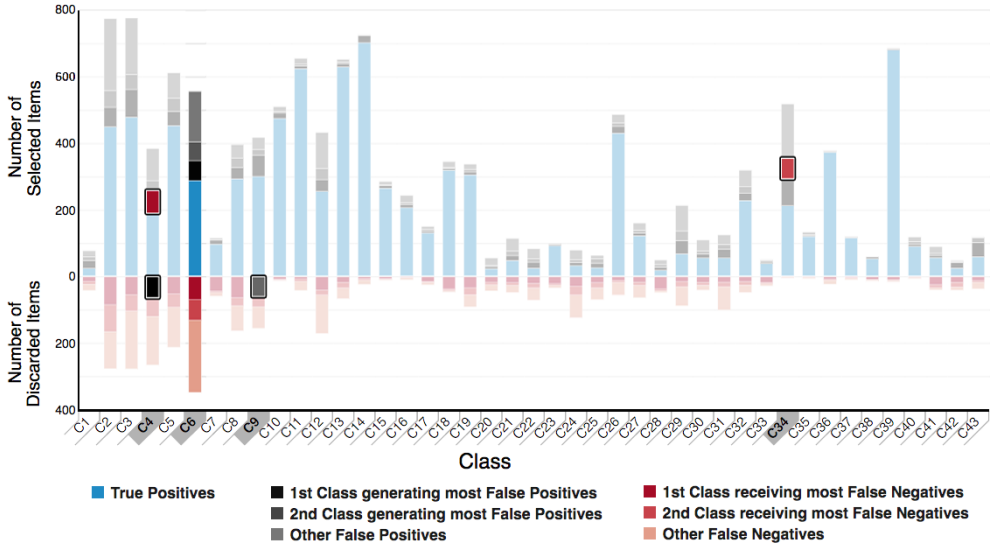

Figure 6.7: Rollover detailing the errors for a specific class.

Large classes (with long bars) can hinder the perception of smaller classes (with small bars). Thus we propose a normalised view that balances the visual space of each class (Figure 6.8). Errors are normalised on the TP of their actual class as n_{xy}/n_{xx} (i.e., dividing $^{FN}/_{TP}$ and reconstructing the FP blocks using the normalised errors $^{FN}/_{TP}$). Although unusual, this approach aligns all FP and FN blocks to support easy and accurate visual perception (Talbot et al. 2014, Cleveland and McGill 1984). It also reminds users of the impact of varying class proportions: the magnitude of errors change between normalised and regular views, as they would change if class proportions differ between test datasets (from which errors were measured) and end-usage datasets (to which classifiers are applied). It is also the basis of the Ratio-to-TP method that estimate the numbers of errors to expect in classification results (Chapter 5, Section 5.3, p.82).

Color choices - Classee uses blue rather than green as in Alsallakh et al. (2014) to address colorblindness (Tidwell 2010) while maintaining a high contrast opposing warm and cold colors. Compared to class-specific colors in Ren et al. (2017) which can clutters the visualization to illegibility, e.g., with more than 7 classes (Murch 1984), Classee colors can handle large numbers of classes.

Following the *Few Hues, Many Values* design pattern (Tidwell 2010), sub-blocks of FN and FP use the same shades of red and black. The shades of grey for FP may conflict with the grey used for TN in binary classification. The multiclass barchart does not display TN and its shades of grey remain darker. Thus color consistency issues are limited, and we deemed that Classee colors are a better tradeoff than adding a color for FP (e.g., yellow in Alsallakh et al. (2014)).

As a result, the identification of *actual* and *predicted* classes is reinforced by the interplay of three visual features: position (below or above the zero line for the

predicted class, left or right bar for the actual class), color hues (blue/red if the actual class is positive), and color (de)saturation (black/grey if the actual class is negative).

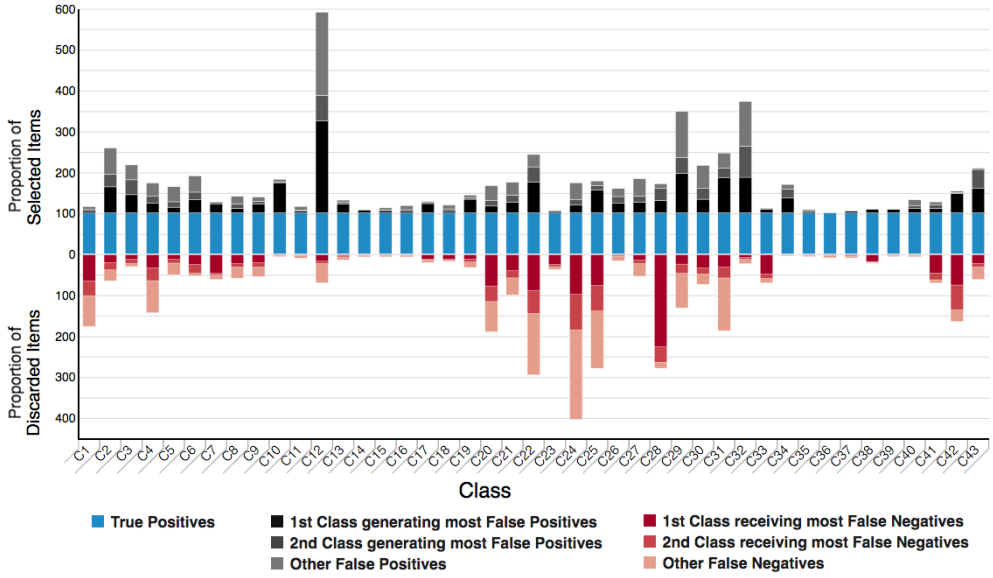

Figure 6.8: Normalized view with errors proportional to True Positives

6.5 User experiment

We evaluated Classee and investigated the factors supporting or impeding the understanding of classification errors. We conducted in-situ semi-structured interviews with a think-aloud protocol to observe users' *"activity patterns"* and *"isolate important factors in the analysis process"* (Lam et al. 2012). We focused on evaluating the *Visual Data Analysis and Reasoning* rather than *User Performance* (Lam et al. 2012) as our primary goal is to ensure a correct understanding of classification errors and their implications. We conducted a qualitative study that informs the design of end-user-oriented visualization, and is preparatory to potential quantitative studies. We included a user group of mathematicians to investigate how mathematical thinking impacts the understanding of ROC curves and error metrics. Such prior knowledge is a component of the *Demographic Complexity* interacting with the *Data Complexity*, and thus impacting user cognitive load (Huang et al. 2009).

The 3 user groups represented three types of expertise: 1) practitioners of machine learning (4 developers, 2 researchers), 2) practitioners of mathematics but not machine learning (5 researchers, 1 medical doctor), and 3) practitioners of neither machine learning, mathematics nor computer science (including 1 researcher). A total of 18 users and 2 users per condition (3 groups x 3 visualizations x 2 users) was sufficient to yield significant observations, as we repeatedly identified key factors impacting user understanding.

The 3 experimental visualizations compared the simplified barcharts to two well-established alternatives: ROC curve and confusion matrix (Figure 6.9-6.11). ROC curves are preferred to Precision-Recall curves which exclude TN and do not convey the same information as the barcharts. All visualizations used the same data and users interacted only with one kind of visualization. This between-subject study accounts for the learning curve. After interacting with a first visualization, non-experts gain expertise that would bias the results with a second visualization.

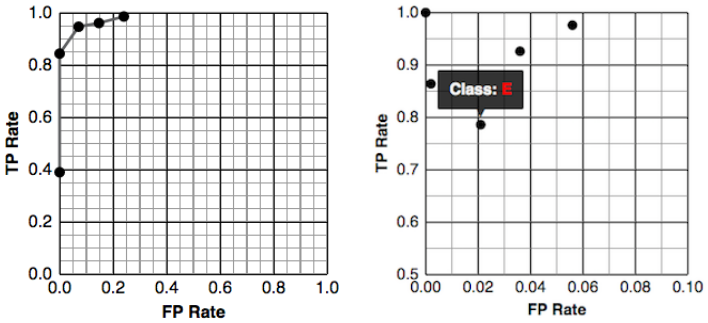

Figure 6.9: ROC curves used for binary and multiclass data.

Threshold	True Positive	True Negative	False Positive	False Negative
0.2	1647	1160	367	22
0.4	1605	1302	225	64
0.6	1581	1419	108	88
0.8	1408	1527	0	261
1	651	1527	0	1018

Figure 6.10: Confusion table for binary data.

Automatic Classification

Actual Class	A	B	C	D	E	True Total
A	122	0	3	0	0	= 125
B	0	110	0	0	0	= 110
C	6	0	113	0	3	= 122
D	8	0	0	95	7	= 110
E	12	0	14	1	99	= 126
Classifier Total	= 148	= 110	= 130	= 96	= 109	

Automatic Classification

Actual Class	A	B	C	D	E	True Total
A	122	0	3	0	0	= 125
B	0	110	0	0	0	= 110
C	6	0	113	0	3	= 122
D	8	0	0	95	7	= 110
E	12	0	14	1	99	= 126
Classifier Total	= 148	= 110	= 130	= 96	= 109	

Figure 6.11: Confusion matrices used for tasks T2-7 to T2-9.

For binary data, classification errors were shown for 5 values of a tuning parameter called a selection *threshold*. Confusion matrices for each threshold were shown as a table (Fig. 6.10) with rows representing the thresholds, and columns representing TP, FN, TN, FP. The table included heatmaps reusing the color coding of the barcharts. The color gradients form the default heatmap template from D3 library were mapped on the entire table cells' values, which is not optimal. Each column's values have ranges that largely differ. Thus the color gradients may not render the variations of values within each column, as the variations are much smaller than the variations within the entire table. Hence color gradient should be mapped within each column separately.

For multiclass data, the confusion matrix also included a heatmap with the same color coding. The diagonal showed TP in blue scale. A rollover on a class showed the FP in dark grey scale and the FN in red scale (Figure 6.11 right). If no class was selected, red was the default color for errors (Figure 6.11 left). The ROC curves to multiclass data displayed a single dot per class, rather than complex multiclass curves. The option to normalize barchart (Figure 6.8) was not included, to focus on evaluating the basic barchart using raw numbers of errors.

The 15 user tasks were in two parts, for binary and multiclass data (Table 6.3). Each part started with a tutorial explaining the visualization and the technical concepts. This could be displayed anytime during the tasks. For binary problems, it explained TP, FN, FP, TN and the threshold parameter to balance FN and FP. For multiclass problems, it explained class-specific TP, FN, FP, TN in one-vs-all reductions, and that FN for one class (the actual class) are FP for another (the predicted class). The explanations of the technical concepts were the same for all users and visualizations. Only the explanations of the visualization differed.

The tasks used synthetic data that predefined the right answers. To assess user awareness of uncertainty, users had to indicate their confidence in their answers. User confidence should match the answer correctness (e.g., low confidence in wrong answers). The response time was measured, but without informing users to avoid *Time Complexity* and stress impacting user cognitive load (Huang et al. 2009). The task complexity targeted 3 levels of data interpretation, drawn from Situation Awareness (Endsley 1995). Level 1 concerned the understanding of individual data (e.g., a number of FP). Level 2 concerned the integration of several data elements (e.g., comparing FP and FN). Level 3 concerned the projection of current data to predict future situations (e.g., the potential errors in end-usage applications). To facilitate users' learning process, the tasks were performed from Level 1 to 3.

Compared to the 3 levels of *Task Complexity* in Huang et al. (2009), our level 1 introduces a lower level of complexity. Our level 2 has less granularity and encompasses all 3 levels in Huang et al. (2009). Our level 3 introduces a higher level of complexity related to extrapolating unknown information (e.g., the errors to expect when applying classifiers to end-usage datasets). Our level 3 also introduces *Domain Complexity*, e.g., it concerns different application domains in tasks T1-4 to -6. The domain at hand can influence user answers. To channel this influence, tasks T2-5

to -9 are kept domain-agnostic, and T1-4 to -6 involve instructions that entail unambiguously right answers, and the same data and reasoning as previous tasks T1-1 to -3.

ID	*Level*	*Question*	*Right Answer*
Step 1 - Binary Classification			
T1-1	L1	Which threshold produces the most False Positives (FP)?	0.2
T1-2	L1	Which threshold produces the most False Negatives (FN)?	1
T1-3	L2	Which threshold produces the smallest sum of False Positives (FP) and False Negatives (FN)?	0.6
T1-4	L3	Choose the most appropriate threshold for person authentication? (*Task presentation tells users to limit FP*)	0.8 or 1
T1-5	L3	Choose the most appropriate threshold for detecting cancer cells? (*Task presentation tells users to limit FN*)	0.2
T1-6	L3	Choose the most appropriate threshold for detecting paintings and photographs? (*Task presentation tells users to limit both FP and FN*)	0.6
Step 2 - Multiclass Classification			
T2-1	L1	Which class has lost the most False Negatives (FN)?	Class E
T2-2	L1	Which class has the most False Positives (FP)?	Class A
T2-3	L2	Which class has the fewest False Positives (FP) and False Negatives (FN)?	Class B
T2-4	L3	Which statement is true? 1) Objects from Class A are likely to be classified as Class E. 2) Objects from Class E are likely to be classified as Class A. 3) Both statements are true. 4) No statement is true.	Statement 2
T2-5	L3	Which statement is true? 1) The number of objects in Class A is likely to be under-estimated (lower than the truth). 2) The number of objects in Class A is likely to be over-estimated (higher than the truth). 3) The number of objects in Class A is likely to be correctly estimated (close to the truth).	Statement 2
T2-6	L3	Which statement is true? 1) The number of objects in Class D is likely to be under-estimated (lower than the truth). 2) The number of objects in Class D is likely to be over-estimated (higher than the truth). 3) The number of objects in Class D is likely to be correctly estimated (close to the truth).	Statement 1
T2-7	L3	Imagine that you are particularly interested in Class D. Choose the classifier that will make the fewest errors for Class D.	Classifier 1
T2-8	L3	Imagine that you are particularly interested in Class A. Choose the classifier that will make the fewest errors for Class A.	Classifier 2
T2-9	L3	Imagine that you are interested in all the classes. Choose the classifier that will make the fewest errors for all Classes A to E	Classifier 2

Table 6.3: Tasks of the experiment.

Quantitative feedback was collected with a questionnaire adapted from SUS method to evaluate interface usability (Brooke 1996) (Table 6.4). Users indicated their agreement to positive or negative statements about the visualizations, e.g., disagreeing with negative statements is a positive feedback.

F1-1, F2-1	I would like to **use** the visualization **frequently**.
F1-2, F2-2	The visualization is unnecessarily **complex**.
F1-3, F2-3	The visualization was **easy to use**.
F1-4, F2-4	I would **need the support** of an expert to be able to use the visualization.
F1-5, F2-5	Most people would **learn** to use the visualization **quickly**.
F1-6, F2-6	I felt very **confident** using the visualization.
F1-7, F2-7	I would need to **learn a lot more** before being able to use the visualization.

Table 6.4: Feedback questionnaire

6.6 Quantitative results

We discuss user prior knowledge (Figure 6.12), user performance between visualizations (Figure 6.13) and user groups (Figure 6.14). User performance is considered improved if i) wrong answers are limited; ii) confidence is lower for wrong answers and higher for right answers; and iii) user response time is reduced. Finally, we review the quantitative feedback (Figure 6.15). The detailed participants' answers are given in Figure 6.20 (p.141).

Figure 6.12: Profiles of study participants.

The prior knowledge of math experts often included TP, FN, FP, TN as these are involved in statistical hypothesis testing (Figure 6.12). Machine learning experts knew the technical concepts well, except a self-taught practitioner who was only familiar to terms related to his daily tasks, e.g., *Accuracy* but not *ROC Curve* or *Confusion Matrix*. This participant, who was in charge of implementing, integrating and testing classifiers, mentioned *"Clients only ask for accuracy"* but did not recall its formula. Two other machine learning experts were unfamiliar with either Precision-Recall or ROC curves, and related formulas, because their daily tasks involved only one of these.

Machine learning practitioners use different approaches for assessing classification errors, using specific metrics or visualizations. They may not recall the meaning and formulae of unused metrics, or even metrics used regularly. Some metrics are

not part of their routine, but may be relevant for specific use cases or end-users. Hence experts too can benefit from Classee since i) Remembering error rate formulae is not needed as rates are visually reconstructed; ii) Both ROC-like or Precision-like rates can be visualized, i.e., equations (6.1)-(6.2); and iii) Accuracy can also be interpreted, i.e., by comparing the relative proportions of errors (FP and FN in red and black bars) and correct classifications (TP in blue bars, TN in grey bars for binary data). Classee also shows the error composition (i.e., which specific classes are often confused) and class sizes. It supports machine learning experts tasks of tuning and improving classifiers (Table 6.2).

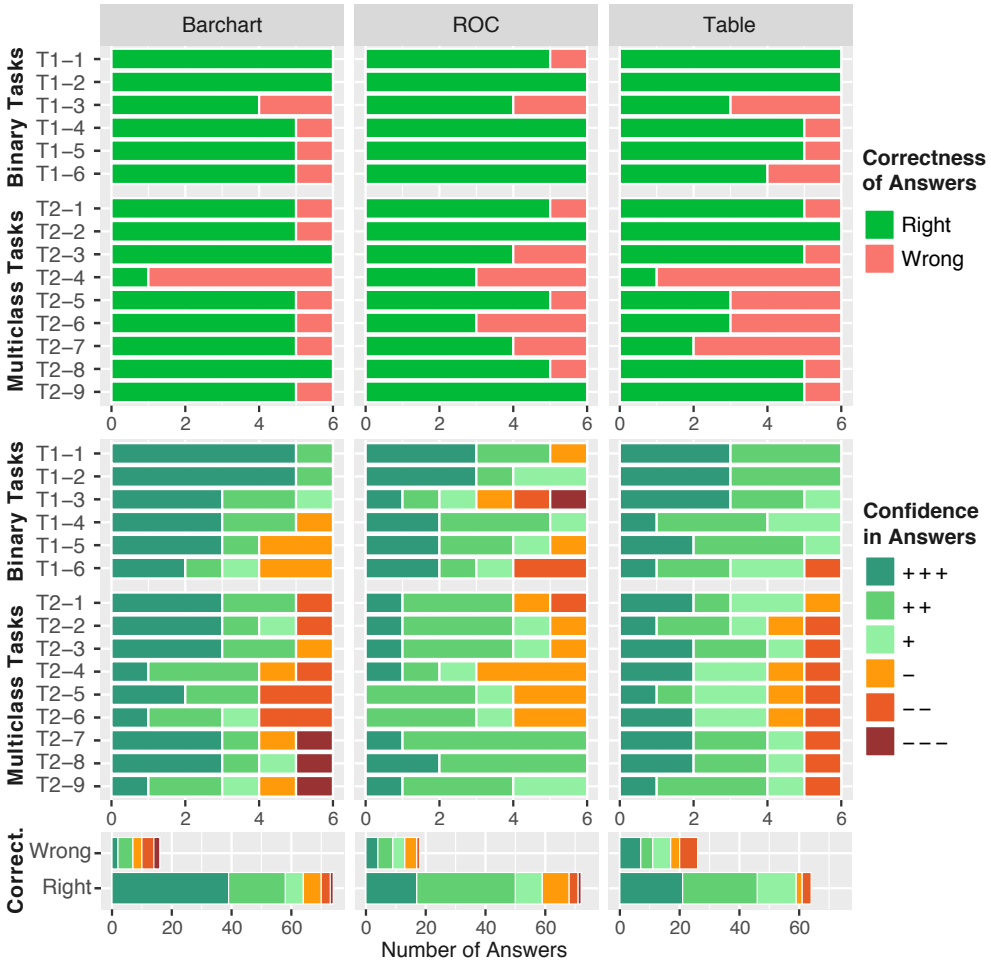

Figure 6.13: Task performance per visualization.

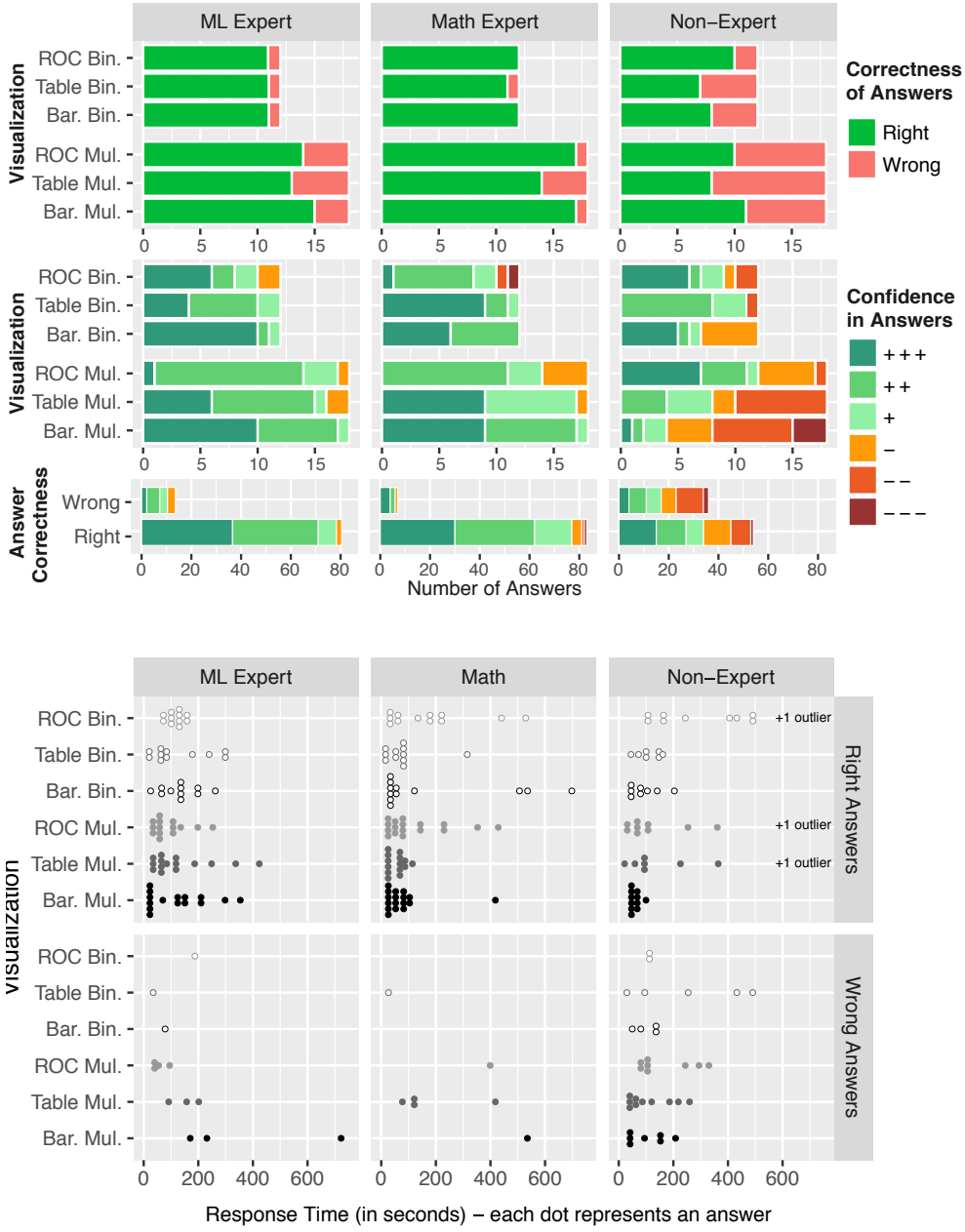

Figure 6.14: Task performance per user group.

With binary data, the number of **wrong answers** differed between tasks T1-1 to -3 and T1-4 to -6 while both sets of tasks entail the same answers and use the same dataset (Figure 6.13 top). Tasks T1-4 to -6 involved extrapolations for end-usage applications. These tasks introduced *Domain Complexity* (Huang et al. 2009) and the tasks' description had increased *task discretion* (less detailed instructions provided to users) thus increasing the cognitive load (Gill and Hicks 2006). The increased task discretion had an important impact as users spent significant efforts relating the terms TP, FN, FP, TN to the real objects they represent (e.g., intruders are FP). With barcharts, user **confidence** better matched answer correctness (lower for wrong answers, higher for right answers) and so for all user profiles (Figure 6.14). Machine learning and math experts gave almost no wrong answers regardless of the visualization, but were more confident with barcharts than ROC curves (and than tables for machine learning experts). Non-experts gave more wrong answers and were over-confident with tables, but with barcharts and ROC curves their lower confidence indicates a better awareness of their uncertainty.

User **response time** was lower with barcharts (Figure 6.14 bottom) except for machine learning experts. Their response time was equivalent for all visualizations but varied less with ROC curves, possibly because this graph was most familiar.

With multiclass data, **wrong answers** were limited until task T2-4 (Figure 6.13 top). Answers were mostly wrong from task T2-4 onwards, as task complexity increased to concern extrapolations of errors in end-results. With barcharts, wrong answers were scarce after T2-4, e.g., after users have familiarized with the graph, but remained high with other graphs. Machine learning and math experts were more **confident** with barcharts (Figure 6.14 middle) but non-experts were under-confident. Yet their **response time** decreased with barcharts, and was as fast as machine learning and math experts (Figure 6.14 bottom).

User feedback was collected twice, after the tasks for binary and multiclass data, with the same questionnaire (Table 6.4). **At the user profile level** (Figure 6.15 top), for binary data, non-experts and machine learning experts had the most negative feedback for ROC curves. Math experts had equivalent feedback for all visualizations. For multiclass data, confusion matrices had the most negative feedback from non-experts and math experts. ROC-like visualizations had the most positive feedback from all profiles. **At the question level** (Figure 6.15 middle), for binary data, barcharts had the most positive feedback on the design *complexity* (F1-2). ROC curves had the most negative feedback for *frequent use* and *need for support* (F1-1, -4). For multiclass data, confusion matrices received negative feedback at all questions, especially for *confidence* and *need for training* (F2-6, -7).

One barchart user gave the lowest possible feedback to almost all questions. This user disliked math and any form of graph (*"Ah! I hate graphs!"*, *"I hate looking at graphs, it's too abstract for me"*) and was particularly reluctant to *frequently using* the graphs (F1-1, F2-1). However, this user's performance was excellent with barcharts for binary data: only right answers with high confidence, and positive feedback especially on the *learnability* (F1-2, *"The graph is easy, even I can use it"*).

Besides this participant, barcharts had the most positive feedback for *frequent use*, *usability* and *need for training* (F2-1, -3, -7). ROC curves had the most positive feedback on *complexity* and *learnability* (F2-2, -5) but its apparent simplicity (only 5 dots on a grid) may conceal underlying data complexity, leading to wrong answers (Figure 6.13).

Over all questions (Figure 6.15 bottom), for binary data the most negative feedback was observed for ROC curves. The feedback was equivalently positive for barcharts and tables. For multiclass data, the most negative feedback was observed for confusion matrices. The feedback was equivalently positive for barcharts and ROC visualizations, excluding the barchart user especially averse to any data visualization.

Users wondered if the feedback also concerned the explanations, hence the results may not represent only the visualization. Other limitations concern the small number of users, and user tendency to avoid either average or extreme feedback (*"I'm not the kind of person having strong opinions"*). More detailed and generalizable insights on the usability are elicited from our qualitative analysis of user interviews.

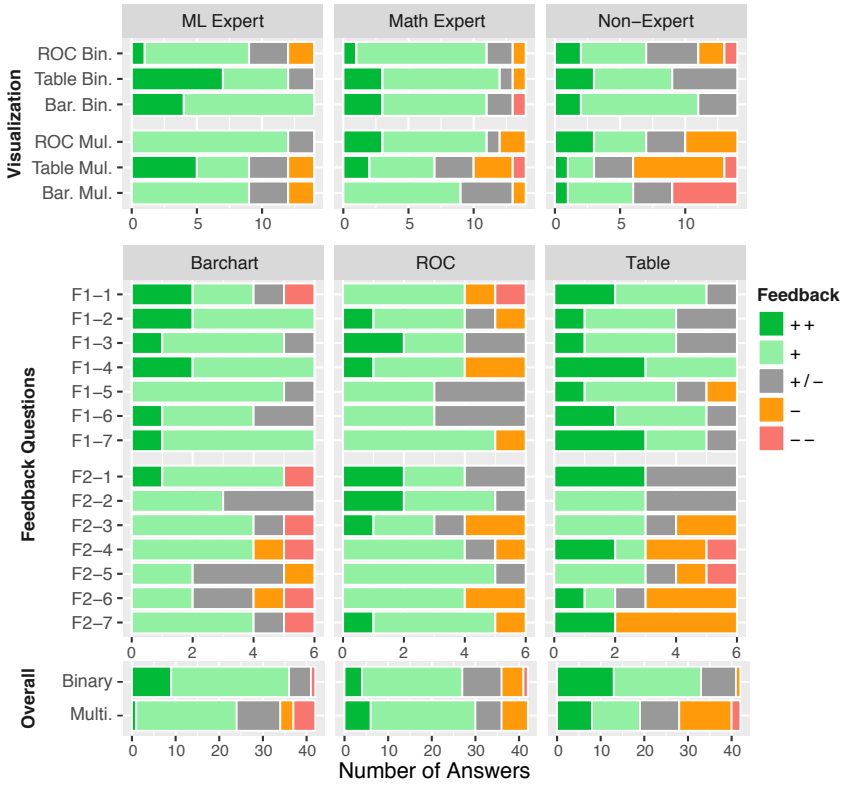

Figure 6.15: User feedback.

6.7 Qualitative analysis

To identify the factors influencing user understanding of classification errors, we analysed user comments and behaviours by transcribing written notes of the interviews. To let the factors emerge from our observations, we first proceeded with *grounded* coding (no predefined codes). We then organized our insights into themes and proceeded to *a priori* coding (predefined codes). We identified 3 key difficulties that are independent of the visualizations:

- The terminology (e.g., TP, FN, FP, TN are confusing terms);
- The error directionality (e.g., considering both FN and FP);
- The extrapolation of error impact on end-usage application (e.g., a class may be over-estimated).

We report these difficulties and how the visualizations aggravated or addressed them.

Terminology - The basic terms TP, FN, FP, TN were difficult to understand and remember (*"In 30 minutes I'll have completely forgotten"*). Twelve users (66%) mentioned difficulties with these terms, including machine learning experts. The terms *Positive/Negative* were often misunderstood as the actual class (instead of the predicted class) especially when not matching their applied meaning (*"Cancer is the positive class, that's difficult semantically"*). Users were also confused by the unusual syntax (*"Positive and Negative are usually adjectives but here they are nouns, it's confusing"*) and the association of antonyms (e.g., False and Positive in FP, *"False is for something negative"*) and synonyms (e.g., *"The words are so close"* with True and Positive in TP, *"I understand that FN are not errors"* because Negative and False is a logical association). Users misinterpreted the terms *True* and *False* as representing the actual or predicted class, and both are incorrect. Some users suggested adverbs to avoid such confusion (*"Falsely"*, *"Wrongly"*). To cope with the semantic issues, users translated the technical terms into more tangible terms, using concrete examples (*"Falsely Discarded"*, *"False face"*). A machine learning expert requested short acronyms (e.g., *TP* for *True Positive*). A non-expert suggested icons as another form of abbreviation (*"like a smiley"* Figure 6.16). This user preferred labels mentioning the actual class first (using *Negative/Positive*) then the errors (using *True/False*).

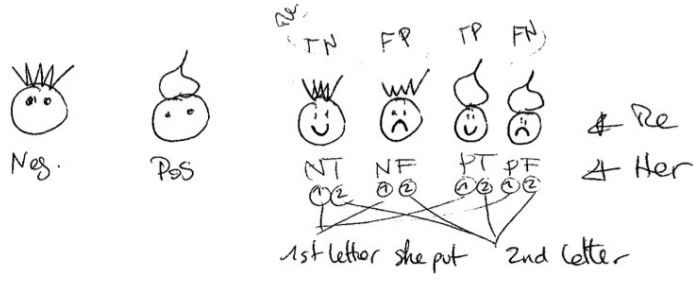

Figure 6.16: User-suggested icons for TP, FN, FP, TN. *Drawn by the interviewer following user's instructions in post-experiment discussions. User-suggested labels are below the icons. Usual labels were later added above.*

The terminology of legends and explanations can yield difficulties (*"You could make the text more clear"*). The terms *Select* and *Discard* in our tutorials and legends can be at odds with their application (*"Discarding objects may be confusing if both classes are equally important"*). The term *true* in its common meaning ("*true* class", "*truly* belong to [class x]") conflicts with its meaning in TP, TN and must be avoided.

Math experts were often familiar with TP, FN, FP, TN as these are involved in statistical hypothesis testing. Machine learning experts knew the technical terms well, except a self-taught practitioner who was only familiar to terms used in daily tasks, e.g., *Accuracy* but not *ROC Curve* or *Confusion Matrix*. This user mentioned *"Clients only ask for accuracy"* but did not recall its formula. Two other machine learning experts were unfamiliar with either Precision-Recall or ROC curves, as their daily tasks involved only one of these. Hence machine learning practitioners may not recall the meaning and formula of unused metrics, or even metrics used regularly. Some metrics are not part of their routines, but may be relevant for specific use cases or end-users. Hence experts too can benefit from Classee since i) remembering error rate formulae is not needed as rates are visually reconstructed; ii) both ROC-like or Precision-like rates can be visualized (6.1)-(6.2); and iii) accuracy can also be interpreted.

Error Directionality - Users need to distinguish the actual and predicted classes of errors, and the direction of errors *from* an actual class classified *into* a predicted class. Ten users (56%) from all profiles had difficulties with error directions, e.g., confusing FP and FN (*"Oh my FP were FN, why did I switch!"*). With binary data, users may not understand how the tuning parameter influence errors in both directions, e.g., decreasing FN but increasing FP (*"I put a high threshold so that there's no error [FP, FN] in the results"*, *"High threshold means high TP and TN"*). With multiclass data, users may not understand that FN for one class are FP for another, and that errors for class x concern both errors with predicted class x and actual class x (e.g., not considering both FN and FP).

Terminology issues complicated user understanding of error directionality, e.g., the terms *Positive/Negative* could mean both the actual or predicted class. Some users intuitively interpreted these terms as the predicted class, others as the actual class. Users often used metaphors and more tangible terms to clarify the error directionality (*"The destination class"*, *"We steal [the FP] from another class"*). The terms *Selected* and *Discarded*, although using a tangible metaphor, can be misunderstood as the actual class (*"The class that must be selected"*) yielding misinterpretations of error directionality.

Extrapolation of Errors in End-Usage Applications - Users needed additional information to extrapolate the classification errors in end-usage applications (*"It's impossible to deduce a generality"*, *"How can I say anything about the rest of the data?"*). More information on the consequences of error was needed to decide which errors are tolerable (*"There can be risks in allowing FP, additional tests have further health risks"*, *"No guidance on how to make the tradeoff"*). Users questioned whether the error measurements are representative of end-usage conditions, regarding potential changes

in class sizes and error magnitudes (*"Assuming class proportions are equal"*, *"This is a sample data, another sample could have some variations"*). They also wondered about additional sources of uncertainty, such as changes in object features or the presence of other classes (*"Will it contain only paintings and photographs?"*) and their impact on the algorithm (*"How does the classifier compute the problem"*). The lack of context information decreased user confidence, e.g., when assessing if a class is likely to be over- or under-estimated.

ROC Curve - It is unusual to visualize line charts where both x- and y-axes represent a rate, and where thresholds are a third variable encoded on the line. It is more intuitive to represent thresholds on the x-axis and rates on the y-axis, with distinct lines for each rate (as a user suggested). Non-experts primarily relied on text explanations to perform the tasks (e.g., reading that low thresholds reduce FP, then checking each dot's threshold to find the lowest). Only machine learning and math experts were comfortable with interpreting the data visually (*"My background makes me fluent in reading ROC curves visually"*, *"I don't use formulas, I compare the dots with each other without reading the values"*).

Error rate formulae were difficult to understand and remember, even for experts (*"Formulas are still confusing, and still require a lot of thinking"*). All users but one needed to reexamine the equations and their meaning many times during the tasks. It increased their response time and impacted their confidence (*"To be sure I'll need to read it again"*). Some users interpreted the rates as numbers of errors, for a simpler surrogate metric. Otherwise, without the numbers of errors, class sizes and potential imbalance are unknown, and it aggravates the difficulties with extrapolating the errors in end-results, e.g., it is impossible to assess the balance of errors between large and small classes (*"Unknown ratio of Positive/Negative"*,*"Assuming class proportions are equal"*). The error composition (how many objects from class X are confused with class Y) is unavailable for multiclass data. Some users noticed the lack of information (*"There's not enough information, errors can come from one class or another"*, *"Assuming the destination class is random"*) but others failed to notice, even for one task that was impossible to answer without knowing the error composition.

Error rates' ambiguous labels aggravated the terminology issues. The rates have actual class sizes as denominators (6.1) but the term *Positive* in *TP* and *FP rate* refers to the predicted class. It misled users in considering that both rates have the predicted class size as denominator, e.g., misinterpreting TP rate (6.1) as Precision (6.2). This is consistent with (Khan et al. 2015) where misinterpretations were more frequent with denominators than numerators, and with (Hoffrage et al. 2015) where a terminology specifying the denominator of probabilistic metrics improved user understanding. A user suggested to replace TP rate by the opposite FN rate (1 - TP rate). It is more intuitive that both rates focus on errors (rather than on correct TP), and by mentioning both *Positive* and *Negative* labels, it may indicate that the denominators differ. Yet the terminology remains confusing as it fails to indicate the rate's denominator. Longer labels could clear ambiguities but may be tedious to read.

Thus ROC curves aggravated the difficulties with the terminology and error directionality, because error rate labels are ambiguous and fail to clarify the denominator. They also aggravated the difficulties with extrapolating errors in end-results because their rates fail to provide the required information, and end-users may fail to notice this limitation.

Confusion Matrix - It is unusual to interpret rows and columns as in confusion matrices, e.g., tables are usually read row per row. Users needed to reexamine the meaning of rows and columns many times during the tasks. It was difficult to remember if they represent the actual or predicted class, which aggravated the difficulties with error directionality. By confusing the meaning of rows and columns, all users but one confused FN and FP. By reading the table either row by row, or column by column, users did not consider both FN and FP (including 2 machine learning experts). The experimental visualization included large labels *Actual Class* and *Automatic Classification* to specify the meaning of rows and columns, but further clarification was needed. Row and column labels showed only the class names (e.g., *Class A*, *Class B*). It was confusing because the list of labels was identical for rows and columns. Labels could explicitly refer to the actual or predicted class, e.g., *Actual Class A*, *Classified as Class B*. One user suggested icons to provide concise indications of the meaning of rows and columns. Another suggested animations to show the relationships of rows or columns and the error directionality, e.g., a rollover on a cell shows an arrow connecting it *from* its actual class *to* its predicted class.

Thus confusion matrices aggravated the difficulties with error directionality because the visual features do not differentiate actual and predicted class. Users must rely on row and column labels, and terminology issues can arise (e.g., if the labels only mention the class names). Color codes and heatmaps can help differentiating FP from FN, but only when a class is selected (errors are FP or FN from the perspective of a specific class) and heatmaps support less accurate perceptions of magnitudes (Cleveland and McGill 1984). Difficulties with extrapolating the errors in end-results were also aggravated because errors are not easy to compare, i.e., users need to relate cells at different positions in the matrix.

Classee - The histograms were intuitive and quickly understood, especially for binary problems (*"This you could explain to a 5-year-old"*). For multiclass problems, it was unusual to interpret histograms where two blocks can represent the same objects. Indeed errors are represented twice: in red FN blocks for their actual class, and in black FP blocks for their predicted class. When a class is selected (Figure 6.7), highlighting the related FP and FN blocks helped users to understand the error directionality (*"Highlight with rollover helps understanding how the classifier works"*) but clarifications were requested (*"You could use an arrow to show the correspondence between FP and FN"*, Figure 6.17). Animations may better show the related FN and FP (e.g., FN blocks moving to the position of their corresponding FP blocks).

Once users familiarized with the duplicated blocks, Classee supported a correct understanding of error directionality, and answers were rarely wrong (*"It's something to get trained on"*, *"Once you get used to it, it's obvious"*). Difficulties remained with

confusion matrices and ROC curves, as misunderstandings of FP and FN remained frequent. Classee better clarified the error directionality with visual features that clearly distinguish actual and predicted classes (*"I like the zero line, it makes it more visual"*). These also reduced the difficulties with the technical terminology and its explanation (*"Explanations are more difficult to understand than the graph"*, *"We usually say it's easier said than done, but here it's the opposite: when you look at the graph it's obvious"*) even though multiclass legends were unclear (*"What do you mean with 1st class and 2nd class?"*). Classee was more tangible and self-explanatory (*"I see an object that contains things"*) and non-experts were more confident than they expected (*"I am absolutely sure but I should be wrong somewhere, I'm not meant for this kind of exercise"*, *"It sounds so logical that I'm sure it's wrong"*).

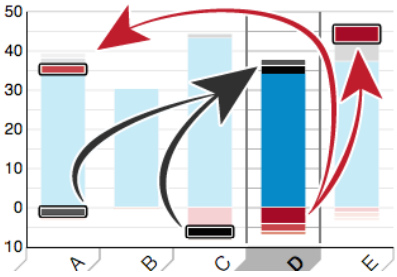

Figure 6.17: User-suggested animation with arrows

Extrapolating the errors in end-results was also easier with Classee. Using numbers of errors provides complete information while ROC curves conceal the class sizes (*"You get more insights from the barchart"*). Confusion matrices also use numbers of errors, but are more difficult to interpret (cell values are difficult to compare, rows or columns can be omitted or misinterpreted). Class sizes and error balance were easier to visualize with Classee (*"Here the grey part is more important than here"*, *"Histograms are more intuitive"*).

Thus Classee limited the difficulties with extrapolating errors in end-results because its metrics and visual features are more tangible and intuitive, and they provide complete information (including class sizes and error balance). Classee also limited the difficulties with the terminology and error directionality by using visual features that clearly distinguish actual and predicted classes. Yet error directionality can be further clarified for multiclass data by adding interactive features to reinforce the correspondence of FP and FN (e.g., animations) and choose the details to display (e.g., error composition for more than 2 classes, or for specific classes).

After the experiment, we introduced the alternative visualizations. Most users preferred Classee, especially after using the other graphs (*"It's easier, I can see what I was trying to do"*, *"This is what I did in my mind to understand the threshold"*). Two machine learning experts preferred Classee, others preferred the familiar confusion matrix or ROC curve (*"You get more insights from the barchart, but ROC curve I read it in a glimpse"*) or would use both confusion matrix and Classee as they complement each other with overview and details.

6.8 Conclusion

We identified issues with the terminology, the error directionality (objects *from* an actual class are misclassified *into* a predicted class) and the extrapolation of error impacts in end-usage applications. To address these issues, labels and visual features must reinforce the identification of actual and predicted classes, e.g., using domain terminology and tangible representations (animations, icons).

Error metrics have crucial impacts on user cognitive load. With error rates, users may overlook missing information (e.g., class sizes) and misinterpret the denominators, which is worsened by terminology issues. Raw numbers of errors are simpler to understand, but are difficult to analyse with confusion matrices.

Classee successfully addressed these issues. Its use of numbers of errors encoded in histograms is more tangible and self-explanatory, and supports accurate perceptions of error magnitudes and class sizes (Requirement R1-3, Section 6.1). The combination of 3 visual features that distinguish the actual and predicted class (position, color hue, color saturation) clarified the error directionality. It helped overcome the terminology issues while providing complete information for choosing and tuning classifiers, and for extrapolating errors in end-usage applications.

Multiclass problems remain particularly difficult to visualize. All three experimental visualizations involve unusual representations in otherwise common graphs. ROC curves have rates on both axes, confusion matrices are read both column- and row-wise, and Classee has duplicated blocks representing the same errors (as FN or FP). In our evaluation, Classee was the easiest to learn and familiarize with, but its legends and interactions should be improved (e.g., with animations highlighting the error directionality).

Our findings inform the design of visualization tools that support end-users' understanding of classification errors, and answer our sixth question: *How can visualization support non-expert users in understanding classification errors?*

We identified factors that support or hinder the assessment of *Noise and Bias* due to classification errors, the resulting *Uncertainty in Specific Datasets*. These are key uncertainty factors identified in Chapter 4. The Classee visualizations we introduced address the assessment of classification *biases*. Future work must investigate the means to assess *noise*, i.e., how errors may randomly vary among datasets to which classifiers are applied (Requirement R4, Section 6.1). Random error variance can be estimated with the Sample-to-Sample method introduced in Chapter 5. Classee visualizations can be used to display variance, e.g., as in Figures 6.18-6.19.

Variance visualization partially addresses the assessment of the *Uncertainty in Specific Datasets*, i.e., estimating the errors to expect in classification end-results (Requirement R4).As identified in Chapter 4 and demonstrated in Chapter 5, the classification errors may largely vary depending on changes in feature distributions (e.g., lower data quality). Hence assessing the *Uncertainty in Specific Datasets* requires the visualization of non-random error variations, depending on datasets' feature distributions.

Visualizing classification errors as a function of varying feature distributions is

complex, but Classee visualizations provide basic templates to address this problem. For example, with binary data, Classee visualization can display the feature values as the x-axis and corresponding errors are the y-axis (i.e., instead of the threshold parameter in Figure 6.1). For multiclass data, the x-axis of Classee visualization can also used to represent the feature values (i.e., instead of the classes in Figure 6.4). However, in this case the graph can only display the errors for a single class, omitting information on the relative class sizes and error directionality (Requirement R2-3). Thus future work is required to design visualizations for exploring the relationships between classification errors and feature distributions.

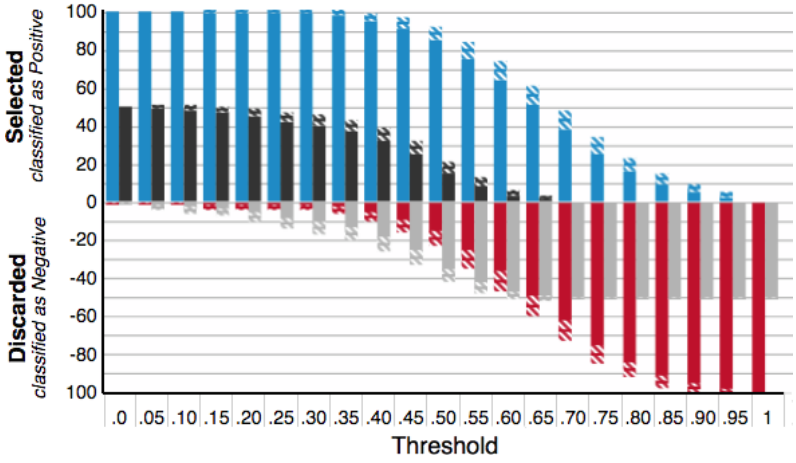

Figure 6.18: Visualization of error variance, avoiding error bars (Correll and Gleicher 2014)

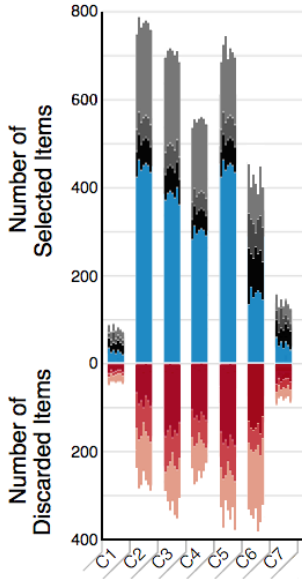

Figure 6.19: Visualization for variance for stacked barcharts, splitting the data in 10 subsamples and juxtaposing them (as stacking variance is mathematically incorrect).

Figure 6.20: Answers' correctness and confidence for each participant

Chapter 7

Visualization Tool for Exploring Uncertain Class Sizes

End-users of computer vision systems for population monitoring are provided with classification results, where class sizes represent population sizes. To draw valid conclusions on the population sizes (e.g., whether population sizes actually increase or decrease as their surrogate class sizes), end-users have to deal with several uncertainty factors. These uncertainty factors arise from computer vision systems and their classification components, and from the environment in which systems are deployed, as identified in Chapters 2 and 3, and synthesized in Chapter 4. End-users must be aware of the uncertainty factors, and their impact on the computer vision results, as identified in Chapter 2 (requirement 4-d, p.36). The information provided on uncertainty factors must be accessible and understandable, as end-users may have little to no expertise in computer vision and classification technologies. As a consequence, it is challenging to enable end-users to make informed decisions when analysing computer vision results. The impact of uncertainty may be misunderstood, uncertainty factors may be overlooked, and end-users may not ever be aware of their lack of information or misunderstanding, as identified in Chapters 3 and 6.

This chapter investigates an interface for visualizing computer vision results and their multiple uncertainty factors, and addressing the needs of non-expert end-users. The interface supports the exploration of **multidimensional computer vision results** (e.g., exploring the distribution of population sizes over multiple dimensions such as time periods or locations) and **multifactorial uncertainty issues** (e.g., exploring uncertainty arising from classification errors or from image quality). The interface was developed within the Fish4Knowledge project (Section 1.1, p.2). It provides access to the end-results of the Fish4Knowledge system[1].

[1] Fish4Knowledge user interface: http://f4k.project.cwi.nl.

We first discuss related work on uncertainty visualization, and on assessing user awareness of uncertainty (Section 7.1) before describing the interface design (Section 7.2). Then, we describe the user experiment we conducted (Section 7.3). The experiment evaluated user awareness of uncertainty, and correctness of data interpretation. We analyze the usability issues that users encountered, and the factors impacting the perception of uncertainty. We compare how users' perception of uncertainty was impacted by the data features (e.g., the level of uncertainty) or the visualization features.

Our user study answers our seventh research question: *How can interactive visualization tools support the exploration of computer vision results and their multifactorial uncertainties?* We identify factors that impact user understanding of uncertainty when exploring computer vision results with interactive visualization. We identify successful interaction principles, and design issues requiring improvement. We provide recommendations for improving the interface design, and for prioritizing the information that must be most salient to end-users.

7.1 Related work

We identify insights from the visualization literature that guided the design of our interface (Section 7.1.1). Then, we discuss usability issues identified in the literature and that are relevant from our use case of non-experts dealing with complex and uncertain information (Section 7.1.2). Finally, we discuss insights from the situation awareness domain, as its considerations of users' information processing issues and awareness of uncertainty provided guidelines for designing our user study (Section 7.1.3).

7.1.1 Visualizing multidimensional and uncertain data

Visualizations of multidimensional data often rely on multiple views (Wang Baldonado et al. 2000). Uncertainty is itself multidimensional: it is of various types depending on the sources of uncertainty (Correa et al. 2009, Thomson et al. 2005) and techniques for uncertainty visualization represent uncertainty as extra dimensions of canonical graphical representations (Griethe and Schumann 2006, Pang et al. 1997). Hence multiple views offer solutions for visualizing the multiple dimensions of computer vision results and their complex multifactorial uncertainty.

Non-expert users need *contextual information* that explain the visualized data (Heer et al. 2008). For instance, the applied data filters should be displayed at all times, and propagated to all views of the interface (Elias and Bezerianos 2011). Propagating the filters constrains interfaces to display views of the same dataset, and limits the expressibility of multiple views. However, it increases the usability as it helps manipulating data attributes. Limiting expressibility in favour of usability is reasonable for our audience of non-expert users (Tang et al. 2004). Thus we applied

Source code: https://sourceforge.net/projects/fish4knowledgesourcecode/files/User%20Interface/

the recommendation from Elias and Bezerianos (2011), i.e., data filters are displayed at all times and propagated to all views.

7.1.2 Usability issues

is a major challenge for our audience of users who are not experts in classification and computer vision. Several issues with non-experts' understanding of visualization have been identified by Grammel et al. (2010):

- Translating questions into data attributes (e.g., selecting data filters of interest).
- Constructing visualizations (e.g., mapping data attributes into visual templates).
- Interpreting visualizations.

The core difficulty for users is *converting concepts* of different natures:

- The concepts involved in data attributes (e.g., categorical or numerical data).
- The concepts involved in a user's mental model (e.g., the meaning and implications of data attributes).
- The concepts involved in visual features (e.g., the geometry of the graphical representations).

Issues with *converting concepts* of different natures relates to issues with:

- Manipulating data attributes (Grammel et al. 2010).
- Identifying the visualizations that are most relevant to users' specific tasks (Griethe and Schumann 2006).
- Locating pieces of information and characterizing their relationships (Amar et al. 2005, Wang Baldonado et al. 2000, Shneiderman 1996).

Misinterpretations are frequently caused by information overload, memory loss and users' limited working memory (Wickens and Carswell 1997). Memory-loss (i.e., forgetting information that was previously perceived) is related to the delays between receiving the information and using it. Such delays can be due to intermediate interactions with the system, or confusing layouts where locating information is tedious or involves trial and error. Working memory and memory loss are crucial in our case.

Our users are not familiar with computer vision systems and their uncertainties. The systems' end-results (e.g., class sizes) and the information on uncertainty issues are unusual. It is challenging to combine and interpret such unusual information. It may overload users' working memory and yield memory loss, which can be worsened by issues with manipulating the visualization interface. We took these issues into account when designing the Fish4Knowledge interface, and investigated the occurrence of these issues during our user study.

7.1.3 Situation awareness

Within the Situation Awareness domain, Endsley (1988a) distinguishes 3 levels of end-users' information processing tasks that are similar to the tasks involved with

interpreting visualizations:

1. **Perception** of cues: occurs when information is simply read without further interpretation or correlation.

2. **Comprehension**: concerns the integration and assessment of multiple pieces of information.

3. **Projection**: concerns the forecast of unknown situations (e.g., future events, interpretations of uncertain data).

The first two levels echo the visual analytic tasks of *locating* and *associating* information (Amar et al. 2005, Shneiderman 1996, Wang Baldonado et al. 2000). The third level is particularly relevant for uncertain information, as it concerns unknown situations. For example, the unknown situations may concern the exact population sizes for which only uncertain estimates are available.

Methods for evaluating Situation Awareness rely on the usage of *probes* (i.e., predefined states of the system in which users are emerged) and consider the uncertainty in users' knowledge of a situation (Jousselme et al. 2003, McGuinness 2004). Probes can be used to expose users to specific interface layouts, prior to letting users interact with the system. Probes allow to evaluate separately the layout design and the interaction design. The tasks to perform while immersed in a probe can target a specific level of information processing (Perception, Comprehension, Projection). These levels can be introduced gradually to allow users to familiarize themselves with the interface and the information on classification results and uncertainty.

7.2 User interface

This section discusses the design of the user interface (Figure 7.3, p.150) that was developed within the Fish4Knowledge project. The interface addresses two high-level user information needs identified in Chapter 2:

- **Estimating the sizes of fish populations for specific species, time periods and locations**. Populations sizes are estimated by a computer vision system using a pipeline of classification components, described in Chapter 4 (Section 4.1.1, p.59). The resulting classification data represents each species with a class. The numbers of items per class (i.e., the class size) represents the population sizes.

- **Assessing the uncertainty of the population size estimates**. The interface addresses 7 of the uncertainty factors identified in Chapter 4 and shown in Table 7.1.

We first discuss the design rationale (Section 7.2.1) before describing the interface (Section 7.2.2) and its usage scenario (Section 7.2.3).

In UI	*Factor*	*Description*
Uncertainty due to computer vision system		
	Groundtruth Quality	Groundtruth items may be scarce, represent the wrong animals, odd animal appearances (i.e., odd feature distributions).
X	**Object Detection Errors**	**Some individuals may be undetected, and other objects may be detected as individuals of interest.**
X	Tracking Errors	Trajectories of individuals tracked over video frames may be split, merged or intertwined.
X	**Species Recognition Errors**	**Some species may not be recognized, or confused with another.**
	Behavior Recognition Errors	Some behaviors may not be recognized, or confused with another.
Uncertainty due to in-situ system deployment		
X	**Field of View**	**Cameras may observe heterogeneous ecosystems, and over- or under-represent species, behaviors or objects features. Fields of view may be partially or totally occluded, and shift from their intended position.**
X	**Fragmentary Processing**	**Some videos may be yet unprocessed, missing, or unusable (e.g., encoding errors).**
X	**Duplicated Individuals**	**Individuals moving back and forth are repeatedly recorded. Rates of duplication vary among species behaviors and *Fields of view*.**
X	**Sampling Coverage**	**The numbers of video samples may not suffice for end-results to be statistically representative.**
Uncertainty due to both computer vision system and in-situ system deployment		
X	**Image Quality**	**Lighting, water turbidity, contrast, resolution or fuzziness may impact the magnitude of computer vision errors.**
	Noise and Bias	Computer vision errors may be random (noise) or systematic (bias). Biases may emerge from a combination of factors (*Image Quality, Field of View, Duplicated Individuals, Object Detection Errors, Species* & *Behavior Recognition Errors*). Additional biases arise from *Duplicated Individuals* and heterogeneous *Fields of View*.
	Uncertainty in Specific Datasets	Uncertainty in specific sets of computer vision results depend on the specific characteristics of the datasets (e.g., distribution of image quality) which impact the magnitude of *Noise and Bias*.

Table 7.1: Scope of uncertainty factors addressed in the Fish4Knowledge User Interface.

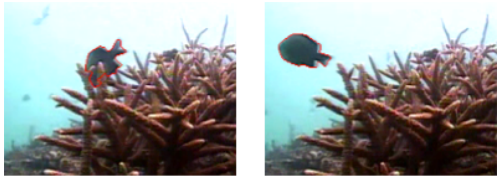

Figure 7.1: Example of certainty scores indicating the species classification uncertainty, i.e., the potential *Species Recognition Errors*. The scores are attributed to each fish occurrence, and measure the similarity between a fish occurrence and its species model (as learned by the classification algorithm). The higher the score, the more certain is the species recognition.

7.2.1 Design rationale

Our design decisions address three challenges:

- **C1 - Unfamiliar technology**. Users have to deal with technologies that are relatively novel in their domain. They need to understand what data can be extracted by computer vision and classification technologies, the limitations of such technologies, and the implications for their data analysis tasks. It demands significant cognitive effort, as reported in Chapters 2 and 6.

- **C2 - Multifactorial uncertainty**. Users have to deal with multiple factors of uncertainty, occurring at different information processing steps (Table 7.1). The resulting complexity is a major challenge.

- **C3 - Heterogeneous goals**. Users have a variety of research goals, introduced in Chapter 2. Users may need to apply specialized data analysis and visualization methods, which may not be addressed with a one-size-fits-all visualization tool.

To address these challenges, we aim at providing a generic user interface that allows users to familiarize themselves with the data and its uncertainty (*C1*). We aim at supporting the exploration of the multidimensional data and uncertainty, while limiting visual and cognitive complexity. We thus use simple graphs and handle multidimensionality with multiple views (*C2*). We target generic data analysis tasks, e.g., *Retrieve, Filter, Determine Range, Correlate Information* (Amar et al. 2005), and exclude advanced data mining and statistical methods (*C3*).

General layout - To explore the uncertainty factors at each information processing step, we organize information in tabs that represent the information processing steps (e.g., from video collection, to video analysis, to data visualization and interpretation). The tabs guide users through the information processing technologies (*C1, C2*) and provide contextual information about the data, as recommended for non-expert users (Heer et al. 2008). The rule of *Diversity*, i.e., separate different types of information, inspired the organization of information into tabs (Wang Baldonado et al. 2000).

Data filtering - As recommended by Elias and Bezerianos (2011), data filters are displayed at all times and propagated to all views and tabs of the interface (when relevant as some tabs do not display data). Propagating filters to all tabs and views follows the rule of *Consistency*, i.e., make the interface consistent (filters are displayed the same way) and the state of the interface consistent (same filters are applied, same data subsets are displayed) (Wang Baldonado et al. 2000).

Data filters are selected using widgets, where each widget represents a specific data dimension. The organization of multidimensional filters into widgets follows the rule of *Decomposition*, i.e., create manageable chunks (Wang Baldonado et al. 2000). The widgets are displayed on-demand to avoid information overload and cluttered interface. A textual summary of the selected filters is always displayed.

Data visualization - The interface provides interactive data visualizations in a dedicated tab (the *Visualization* tab) and within the filter widgets. The widgets display small histograms (Figure 7.2). The filter values are discrete, and each selectable value is displayed on the histogram x-axis. Users can click on a histogram to select the data subset represented by the histogram, and filter out the remaining data.

Widgets' y-axis represent the same dimension as the y-axis of the main graph in the Visualization tab. Users can select the dimension to display on the y-axis, for example, a species' population size. In this case, the widget to filter data from specific cameras shows the species distribution over the cameras, i.e., over the cameras' geographical locations.

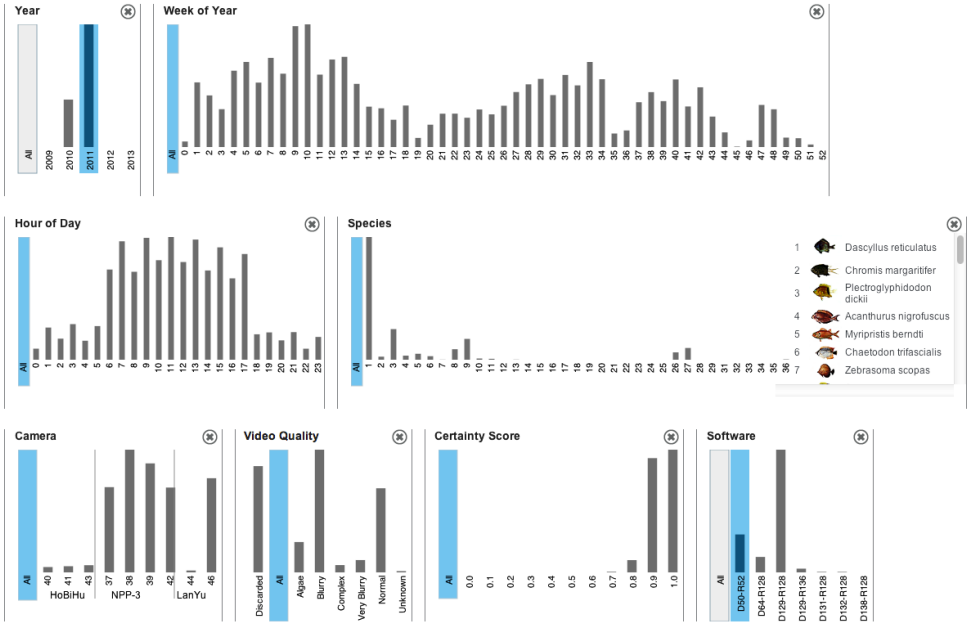

Figure 7.2: The filter widgets that let users select the dataset of interest, e.g.., the time periods, camera locations, or fish species. The histograms provide an overview of the data distributions. The y-axis represents the same dimension as the main graph of the Visualization tab (Figure 7.3).

Using the same dimension on all y-axes (i.e., widgets and main visualization) follows the rules of *Consistency* and *Complementarity* (Wang Baldonado et al. 2000), i.e., expose the relationships between the data dimensions while limiting the interface complexity (*C1, C2*). The interface can display data distributions on all different dimensions (i.e., using the widgets) thus showing the relationships between data dimensions. For instance, a population size (y-axis dimension) is influenced by the population distributions over species or camera locations (x-axis dimensions).

In the Visualization tab (Figure 7.3), users can select the dimensions displayed in the x- and y-axes. Users can also change the type of graph, while keeping the same axes' dimensions. Two graphs offer a third dimension: stacked charts (Figures 7.11-7.12) and boxplots (Figure 7.15) which use a third dimension to break down data into stacked subsets, or subsamples for boxplots. Users can select visualization variants by swapping the graph axes and the type of graph. This navigation design synthesizes features from ManyEyes (Viegas et al. 2007) (swap axes) and Tableau [2] (swap graph templates).

As visualizations are modified (e.g., users select x-axis, then the y-axes, then the data filters) and each modification is propagated to all graphs of the interface. The filters and y-axis are also propagated to the Video tab, i.e., the widgets remain the

[2]Tableau Software, http://www.tableausoftware.com

same in both tabs. The consistency of graph modifications aims at limiting issues with *context switching* which can yield *memory loss*.

Swapping axes and graphs, and displaying widgets on demand offer a large scope of possible data associations and comparisons, while limiting cluttered display and information overload (e.g., as widgets are opened and closed on demand). This is desired in a context where users pursue a variety of research goals (*C3*) while being unfamiliar with the data (*C1*).

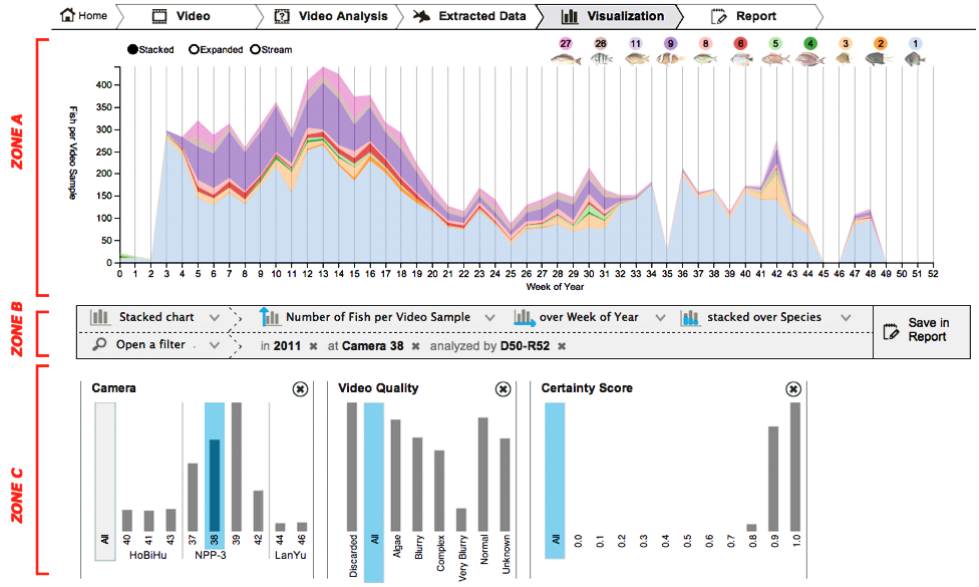

Figure 7.3: The Fish4Knowledge interface is organized in 5 tabs (above Zone A). The Visualization tab lets user explore the classification results, and is organized in 3 zones. **Zone A** contains the main graph, e.g., showing the population sizes for each species. **Zone B** contains a menu that lets users control the type of graph displayed in Zone A, and the data dimensions represented on the main graph's axes. It also recaps the filters in use, and lets users cancel the filters and open the filter widgets. **Zone C** contains filter widgets, each representing a specific data dimension. The widgets provide 2 functionalities: select filter parameters (e.g., data from cameras 38 at location NPP-3) and overview the data distributions over the widgets' data dimension (e.g., camera locations, video quality). The set of available widgets is shown in Figure 7.2. The Video tab (upper left tab) is also organized in 3 zones and reuses the same widgets, filters and menu (Figure 7.4).

7.2.2 Interface design

This section describes the 5 tabs of the Fish4Knowledge interface (above Zone A in Figure 7.3): the *Video* tab (Section A), the *Video Analysis* tab (Section B), the *Extracted Data* tab (Section C), the *Visualization* tab (Section D) and the *Report* tab (Section E). The tabs reflect the information processing steps: data collection (*Video* tab), data

processing (*Video Analysis* tab), and data interpretation (*Extracted Data, Visualization* and *Report* tabs). The first three tabs guide users through the computer vision system components, and the uncertainty factors they entail. The other two provide tools for interpreting the computer vision results.

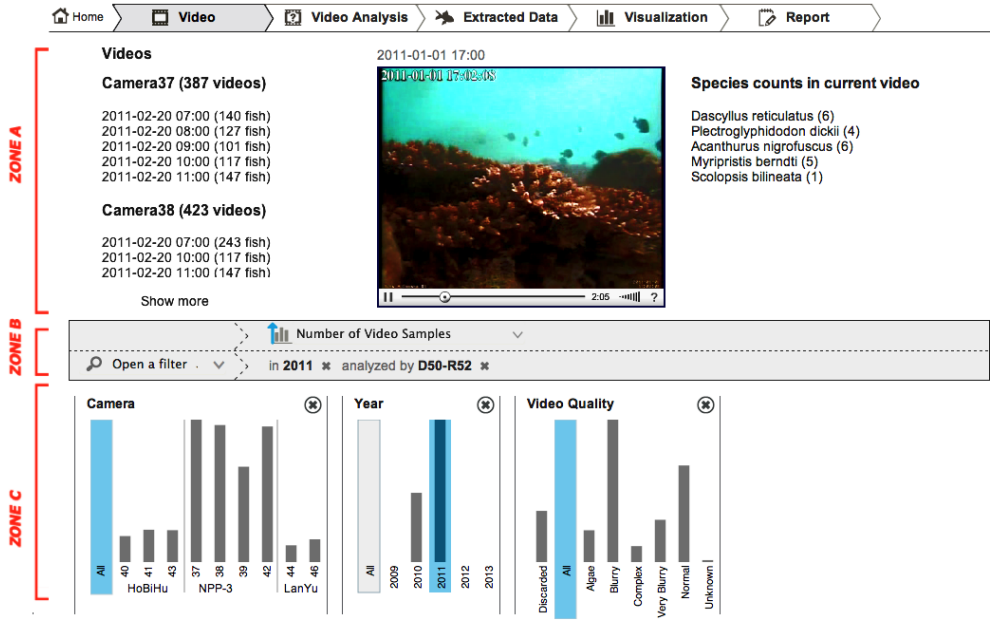

Figure 7.4: The Video Tab lets users explore the video footage (Zone A) and the numbers of video samples over different dimensions (Zone C). Users can display numbers of video as the y-axis of the widgets' histograms, and open the widgets of interest (using menus in Zone B)

A. The Video tab

The Video tab supports direct browsing of the 10-minute video footages that were processed by the computer vision system (Figure 7.4). It contains filter widgets for selecting the set of videos of interest (e.g., collected at specific locations and time periods). With this tab users can inspect the video data collection conditions: which ecosystems are observed, with which field of view and image quality (e.g., blurry images, algae bloom yield green and murky water). Videos of the same image quality can be filtered by using the *Video Quality* widget.

This tab partially supports the assessment of uncertainty issues with *Fields of View*, *Duplicated Individuals*, *Sampling Coverage* and *Image Quality* (Table 7.1). Although no quantitative measurement of the uncertainty is provided, browsing the video footage offers valuable means to visually assess these uncertainty factors. Users can visually assess biases due to low *Image Quality* and inadequate or shifting *Fields of View*. The latter impacts the chances of *Duplicated Individuals* and the geographical *Sampling*

Coverage. Uncertainty can be further assessed by exploring the number of video samples, i.e., if displayed as y-axis of filter widgets, per *Image Quality*, camera (i.e., *Field of Views*) or time period (i.e., temporal *Sampling Coverage*), e.g., using the widgets in Figure 7.4.

B. The Video Analysis tab

The Video Analysis tab provides explanations of the video processing steps, and visualizations of computer vision errors. It explains basic technical concepts needed for understanding computer vision uncertainty. The Overview sub-tab provides explanations of the main video processing steps (Figure 7.5). The Fish Detection, and Species Recognition sub-tabs provide visualizations of the classification errors when detecting fish and non fish objects (Figure 7.6) and when classifying the fish species (Figure 7.7). The prototypes developed for the Fish4Knowledge project can be improved by including the visualization introduced in Chapter 6, and the menu used in the Video and Visualization tabs (Figure 7.8).

The Workflow sub-tab provides on-demand video processing (Figure 7.9). Users can request the analysis of specific videos (e.g., from time periods and cameras of interest) with specific component versions (e.g., with the fewest classification errors for the species of interest). It serves either for processing videos that were not yet analyzed, or for experimenting with different versions of the video analysis components (e.g., to check the robustness of observations).

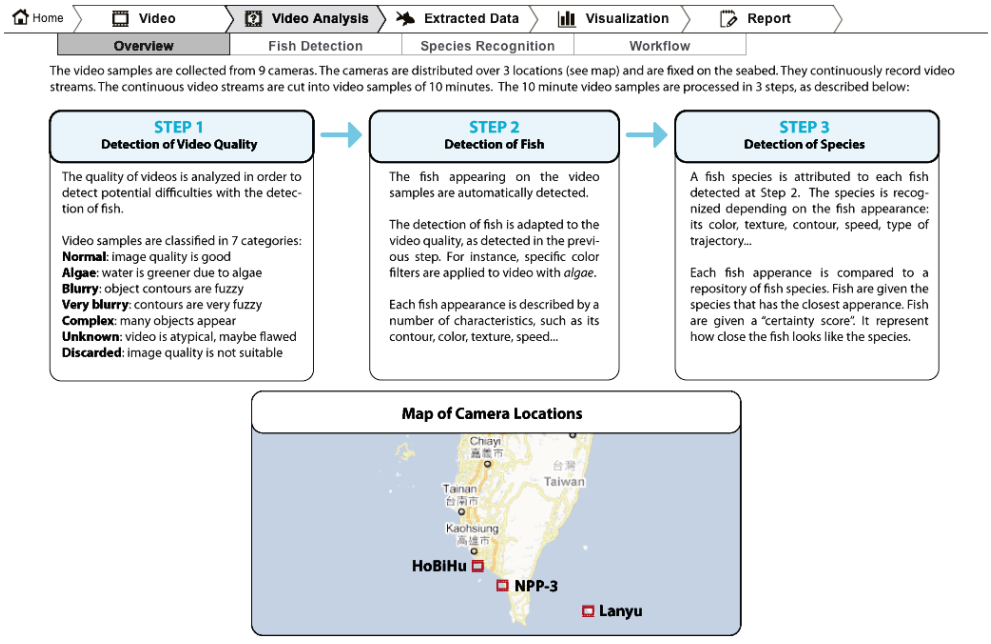

Figure 7.5: The Video Analysis Tab - Overview sub-tab explains the video processing steps.

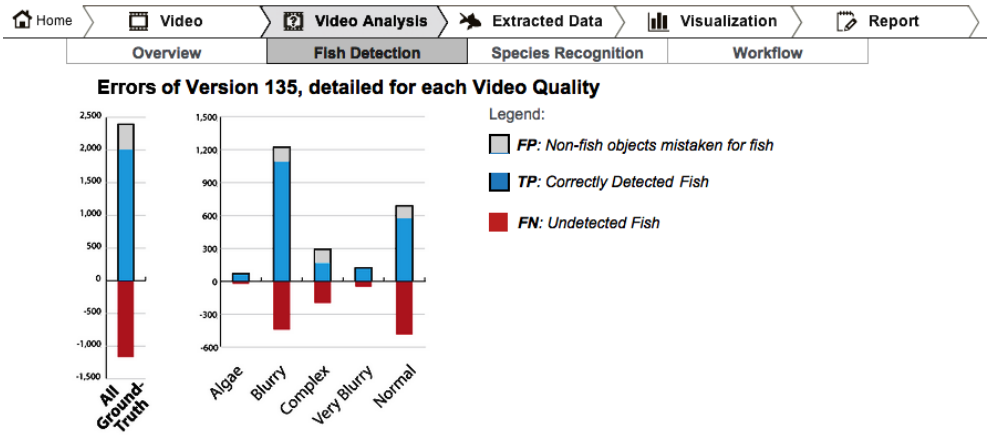

Figure 7.6: The Video Analysis Tab - Fish Detection sub-tab provides simplified visualizations of classification errors for the Fish Detection algorithm (detecting fish and non-fish objects). Errors are detailed for each type of video quality, and each version of the Fish Detection algorithm.

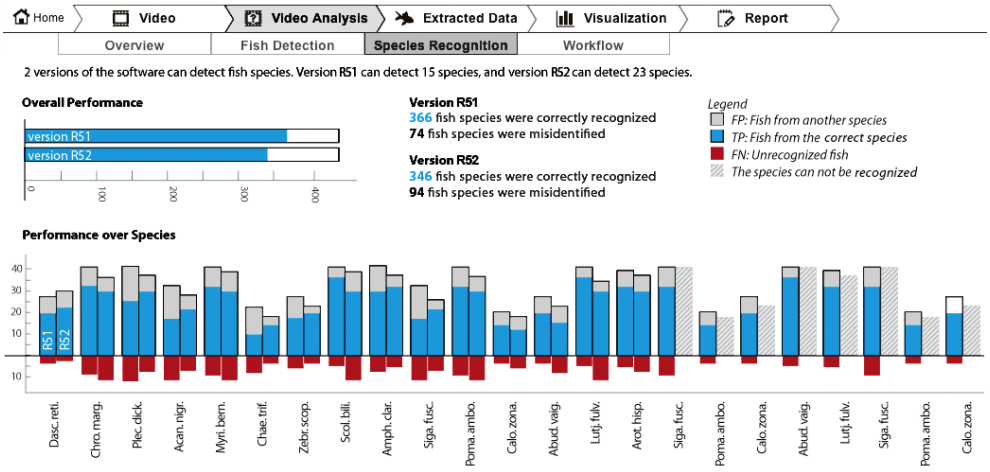

Figure 7.7: The Video Analysis Tab - Species Recognition sub-tab provides clear and simple visualizations of classification errors for the Species Recognition components. Errors are detailed for each species, and each version of the Species Recognition algorithm. The algorithm version R52 can recognize fewer species than the R51 version, as the R52 algorithm focuses on recognising the most important species (from ecologists' point of view), hence excluding the recognition of less important species (which are rare species in the environment studied by ecologists).

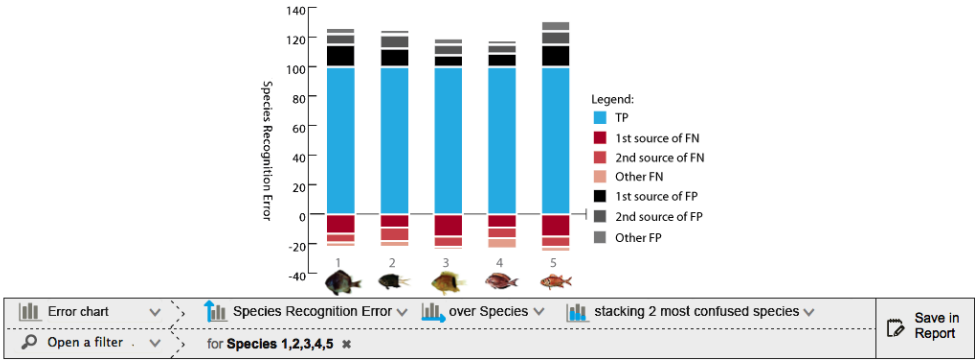

Figure 7.8: Alternative for the Video Analysis Tab - Species Recognition sub-tab using Classee visualization (Chapter 6) and menus from the Video and Visualization tabs (Zone B in Figures 7.3 and 7.4).

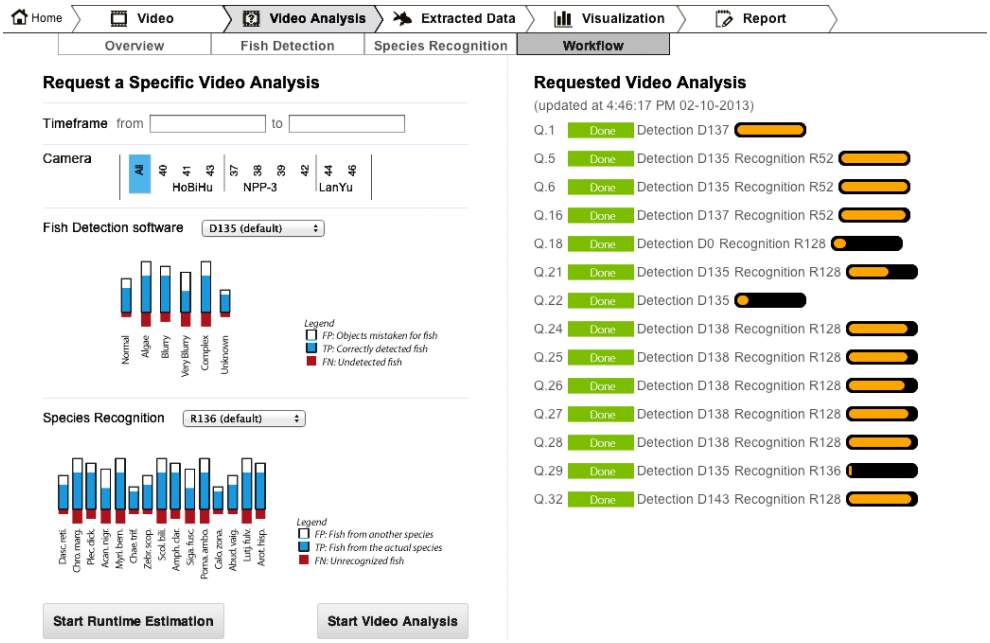

Figure 7.9: The Video Analysis Tab - Workflow sub-tab supports user requests for specific video processing tasks to be executed by the computer vision system. The interface shows the classification errors of the component versions that users plan to use.

The Video Analysis tab supports the assessment of uncertainty issues with *Object Detection Errors* and *Species Recognition Errors* (Figures 7.6 to 7.7). Future work is required to enable the assessment of *Object Tracking Errors*, *Groundtruth Quality*, as no well-established method are available for assessing these uncertainty factors, and their high-level impact on the *Noise and Bias* in the classification results (Chapter 4, Section 4.4.1, p.67).

Assessing *Groundtruth Quality* can be enabled by letting users display *numbers of groundtruth items* as the y-axis of widgets and main graph in the Visualization tab. However, each classification component (e.g., object detection, tracking and species recognition) may be evaluated with their own groundtruth. In this case, the *numbers of groundtruth items* must be displayed for each groundtruth set. This requires adding several dimensions that can be displayed as y-axes, which increases user cognitive load. By visualizing the groundtruth size, groundtruth scarcity and imbalance may be assessed. However, the uncertainty propagation to the classification end-results is not addressed and requires future work.

C. The Extracted Data tab

The Extracted Data tab specifies the characteristics of the information extracted from the video footage, i.e., the data dimensions. It explains 4 metrics provided for describing fish populations, and that can be displayed as the y-axis of the widgets and main graph in the Visualization tab:

- Number of Fish
- Number of Video Samples (e.g., to check for missing videos and assess *Fragmentary Processing*)
- Mean Number of Fish per Video Sample (e.g., to compensate for missing videos, introduced in Chapter 4, Section 4.4.2, equation (4.1) p.71)
- Number of Species (e.g., for studying *species richness*, a user information need identified in Chapter 2, Section 2.3, p.22).

This tab shows the aspects of fish populations that can be monitored with the computer vision system. The overview of the data dimensions helps identify the information that is relevant for users' research goals (Challenge C3, Section 7.2.1) and the functionalities for filtering and visualizing datasets of interest. Future work can investigate more elaborate tutorials, e.g., improved textual and visual explanations (e.g., animations, or comic book style tutorials as in Figures 5.11-5.12, p.112).

D. The Visualization tab

The Visualization tab, shown in Figure 7.3 p.150, provides a means of exploring the computer vision results, e.g., the class sizes representing fish population sizes. The layout is organized in 3 zones. *Zone A* contains the main graph, *Zone B* provides widgets for filtering data, and *Zone C* controls the widget display and the adaptation of the main graph to specific user needs.

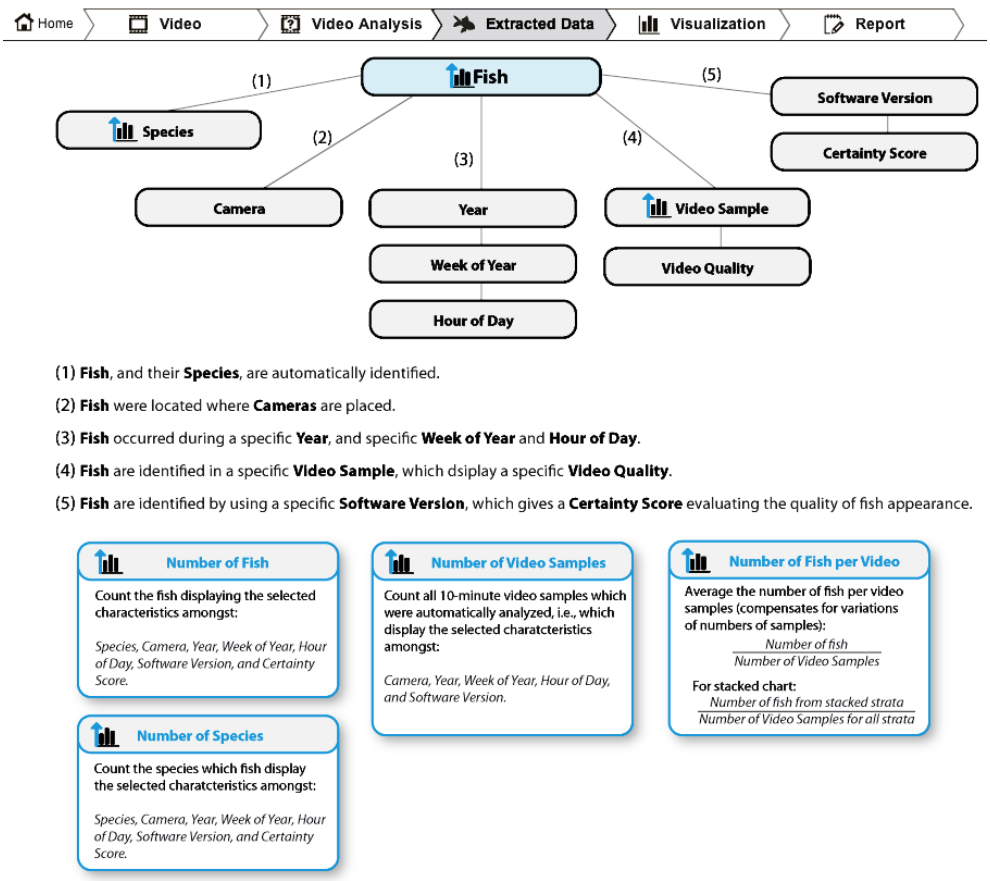

Figure 7.10: The Extracted Data tab provides a schema of the data dimensions, and explanations of the y-axis metrics.

In *Zone B*, users can specify what the axes of the main graph represent. For instance, while the y-axis represents numbers of fish, the x-axis can represent their distribution over weeks of the year (Figure 7.3) or hours of the day (Figures 7.11-7.12). Users can also select other types of graph, e.g., stacked chart or boxplot, leading to the display of dedicated menus for adapting the visualization further. For instance fish counts can be stacked by species (Figure 7.11) or by camera (Figures 7.12).

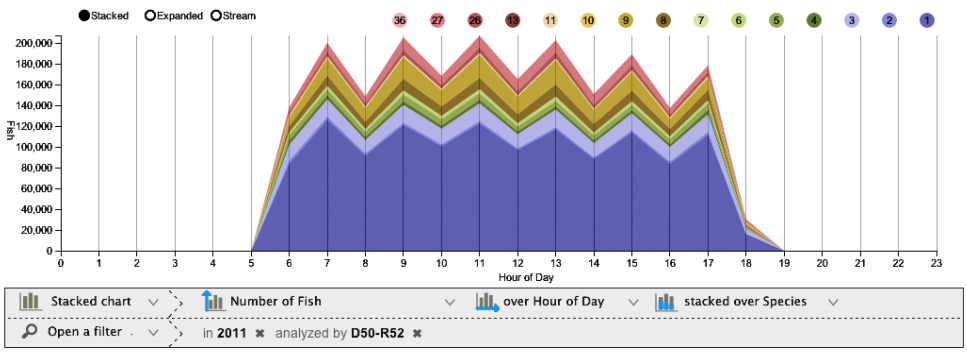

Figure 7.11: Visualizations of fish counts stacked by species.

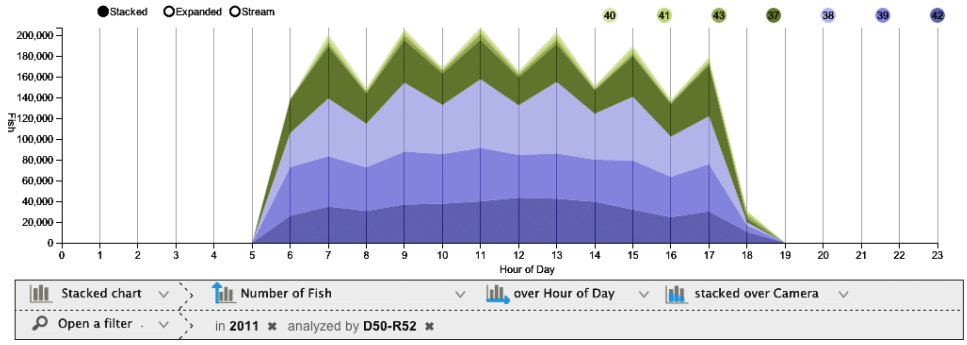

Figure 7.12: Visualizations of the same fish counts as in Figure 7.11 but stacked by camera.

Zone C contains filter widgets for selecting datasets of interest, and the widgets' histograms provide an overview of the dataset over several dimensions. Filter widgets are displayed on-demand, using the lower left menu in Zone B. There are widgets for each dimension of the data, e.g., Year, Week of Year and Hour of Day of fish occurrence, Camera, Species, Video Quality, Software Version, and Certainty Score (Figure 7.2).

A summary of the applied filters is provided in *Zone B*. To limit information overload, unused filters (e.g., all species, all cameras) are not mentioned in the summary. The widgets' histograms display the same y-axis as the main graph, and the same dataset. For instance, in Figure 7.3 both the graph of *Zone A* and the histograms of *Zone C* display mean numbers of fish per video sample. Both use the same dataset, e.g., of videos processed by algorithms' version *D50-R52*, occurring in 2011 at Camera 38 (and belonging to all species, certainty scores, image quality, weeks of year and hours of day). The *Camera* widget uses a dataset from all cameras, and highlights which camera is selected (Camera 38).

The *Visualization tab* supports the exploration of uncertainty due to *Fragmentary Processing*, *Sampling Coverage*, *Image Quality* and *Species Recognition Errors* (Table 7.1).

The type of *Image Quality* is detected for each video sample during pre-processing (i.e., before recognizing the objects occurring in the videos). This uncertainty factor can be assessed by visualizing the distribution of video samples over the types of image quality. *Species Recognition Errors* can be assessed using the Certainty Score widget. Certainty scores and types of image quality are data dimension that can be displayed as the main graph's x-axis, and filtered using dedicated widgets. However, filtering by certainty scores must be used with care as it can introduce biases. For example, when filter out low certainty scores, most fish may be filtered out and the selected data may not represent the actual fish populations.

Fragmentary Processing, i.e., missing videos, impacts the temporal and geographical *Sampling Coverage*. Videos can be missing due to camera maintenance, encoding errors, or unfinished processing queues. Users can explore the number of video samples available for each data dimension, i.e., by selecting number of videos as y-axis, and opening widgets or modifying the main graph's x-axis (Figures 7.13). Users can also explore how variations in numbers of video impact the class sizes, i.e., by switching the y-axis between numbers of fish and numbers of videos. Finally users can explore normalized population sizes, abstracted from variations in video numbers, i.e., by displaying the mean number of fish per video as the graph's y-axis (Figure 7.14 using equation (4.1) from Chapter 4, p.71).

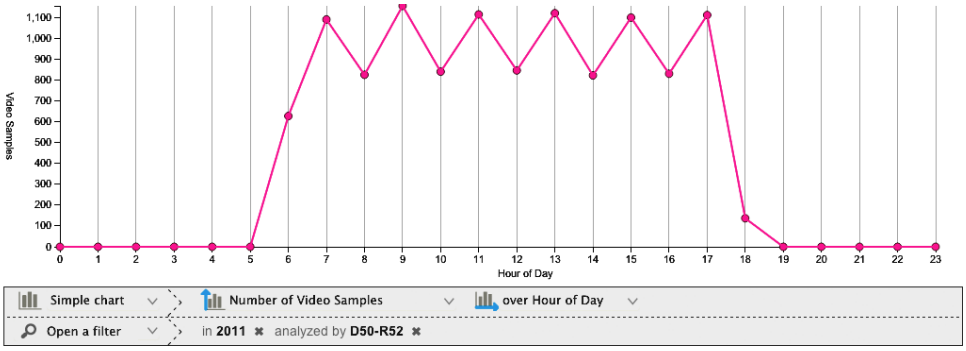

Figure 7.13: Visualization of the number of videos from which fish counts in Figures 7.11-7.12 were extracted. The number of videos has a direct impact on the absolute fish counts: the more the videos, the higher the fish counts. The results shown in this Figure are an example of *Fragmentary Processing* issues (Table 7.1). *The variations in numbers of videos shown in this Figure are due to the batch processing strategy. To process videos over an entire year, and obtain preliminary results, batches of videos are processed for every uneven hour of the year (e.g., all videos recorded at 07:10, 09:10, 11:10 etc..) and then other batches of videos for every even hour. This process is repeated until the entire video collection is processed.*

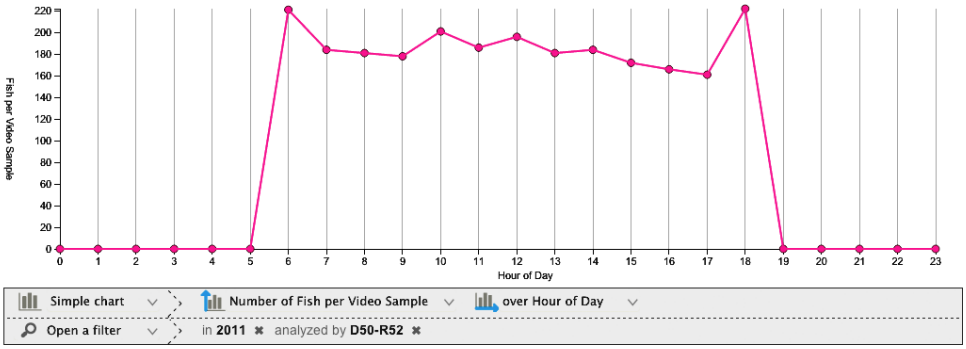

Figure 7.14: Visualization of the mean number of fish per video. It balances the impact of heterogeneous numbers of videos on absolute fish counts shown in Figures 7.11-7.12.

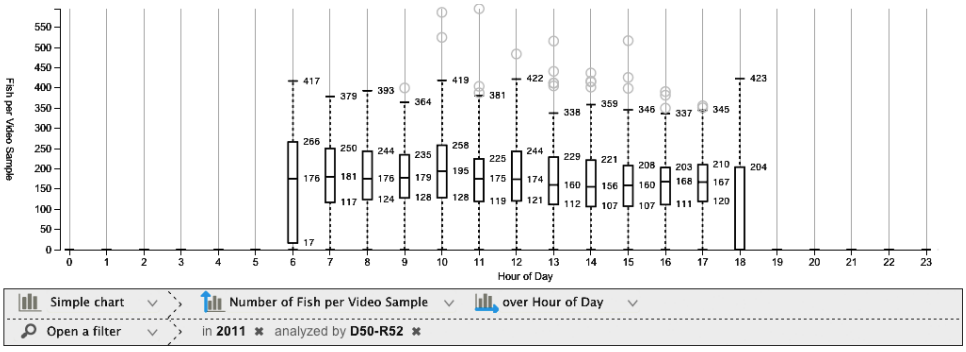

Figure 7.15: Visualization showing the variance of mean number of fish per video. The dataset selected in Figure 7.14 is sub-sampled each week of the year. The boxplot shows how fish counts for each hour of the day vary over the weeks of the year.

E. The Report tab

The Report tab supports manual grouping and annotation of graphs created in the Visualization tab (Figure 7.16). Visualizations can be added to and removed from a *report*, and their interpretation can be described with free-form text. Using the *Download* button, users can save the report they are currently working on. Downloaded reports consist of a text file containing a list of parameters. They can be stored or shared with other users as any kind of text file. To visualize a downloaded report, users can upload the parameter files with the *Upload* button of the *Report* tab. With this tab, users can document their data exploration and interpretation process, and their findings.

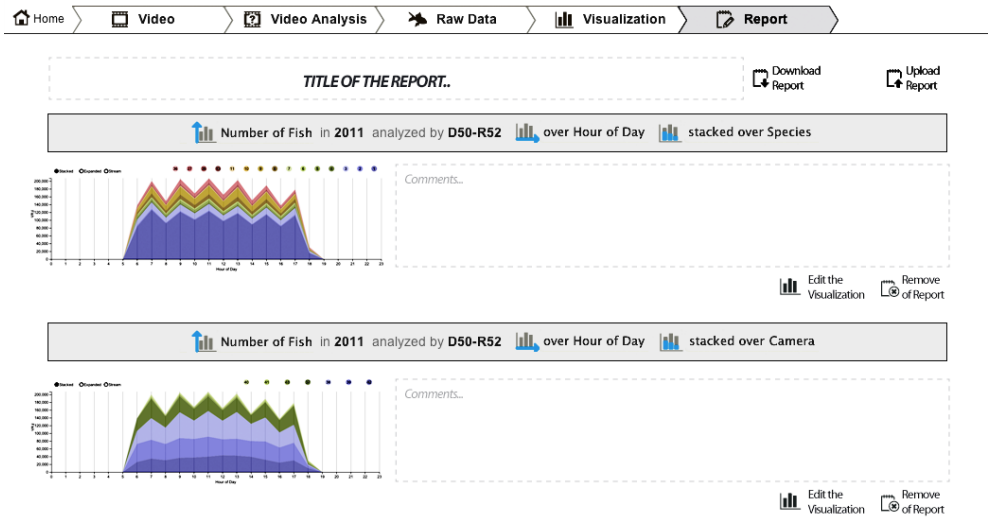

Figure 7.16: The Report tab showing two visualizations (i.e., Figures 7.11-7.12) that are saved in a report, together with their specifications (i.e., filters and displayed dimensions are recapped above the graphs). Users can comment their findings in free-form text (i.e., using the text fields on the right of the graphs).

7.2.3 Usage scenario

This section describes typical interactions involved in the analysis of population sizes (i.e., the analysis of class sizes) and two uncertainty issues: *Species Recognition Errors* and *Fragmentary Processing* (Table 7.1). More detailed usage scenarios are given in Appendix 1 of the Fish4Knowledge book (Beauxis-Aussalet and Hardman 2016).

The usage scenario described in this section focuses on the interaction and layout design evaluated in Section 7.3. The usage scenario is illustrated with screenshots of the user interface prototype that was used to conduct the user study. The prototype was later refined, based partially on the results of this evaluation, which explains the small differences with the interface presented in Section 7.2.2.

A. Exploring issues with Fragmentary Processing

When analyzing the fish populations in Figure 7.17, users may wonder if the population sizes drop in weeks 35 and 45 due to missing videos. Using the *Y Axis* menu in *Zone B*, they can display the numbers of videos from which fish counts were extracted (Figure 7.18). As no videos were processed for Week 45, no insight can be drawn on fish populations in this period. Considering the high variability of video numbers, visualizing mean numbers of fish per video is preferable to visualizing absolute fish counts. The *Y Axis* menu provides this visualization (Figure 7.19).

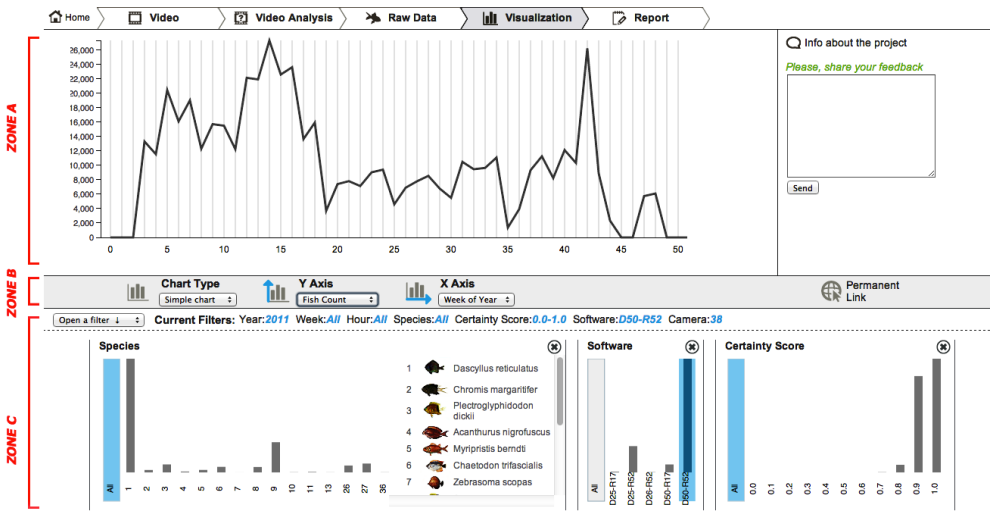

Figure 7.17: The Visualization tab as shown for the first probe of the user study, and needed for answering questions 1-4.

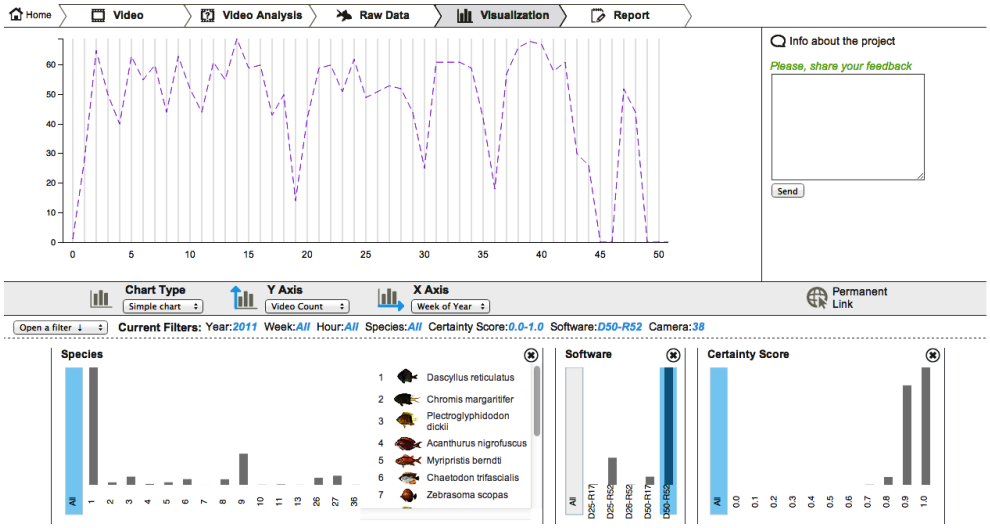

Figure 7.18: Visualization for exploring the impact of Fragmentary Processing, and needed for answering questions 5 and 7 of the user study (Section 7.3).

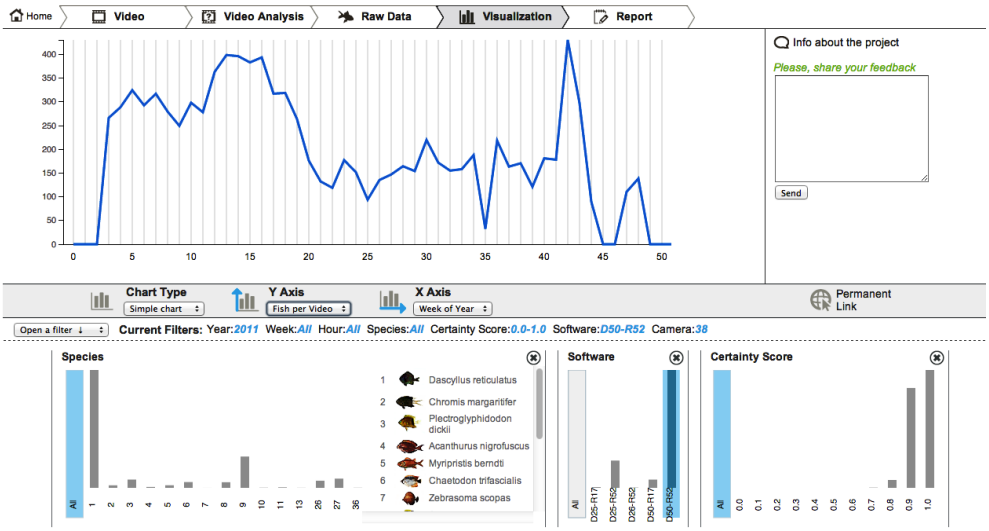

Figure 7.19: Visualization for exploring the impact of Fragmentary Processing, and needed for answering questions 6 and 7 (Section 7.3).

B. Exploring classification uncertainty

Considering that the trends in fish population sizes are not due to varying numbers of videos, users can question the reliability of the species recognition. The widget *Certainty Score* shows the quality of fish appearances (Figure 7.1). The more fish with high certainty scores, the more reliable the species recognition. Users can use the certainty scores to estimate potential biases due to species recognition errors. For example, Figures 7.20-7.21 compare the classification uncertainty for species 1 and 2. Higher certainty scores are observed for species 2, thus its recognition is likely to be more reliable than for species 1. Similarly, users can compare the certainty scores for week 35 with other weeks.

The classification uncertainty can be further detailed using the Video Analysis tab. However, this tab was excluded from our user study in Section 7.3, as we focus on evaluating the Visualization tab. The evaluation of the visualization of classification errors displayed in the Video Analysis tab is discussed in Chapter 6. The user study presented in this chapter focuses on the usage of certainty scores as an alternative metric of classification uncertainty.

In future work, the classification uncertainty should be represented in the Visualization tab by applying the methods for estimating classification errors in end-results introduced in Chapter 5. However, this chapter focuses on evaluating the multiple view design of the Visualization tab, the menus to modify the graph's axes, and the usage of certainty scores. The Fish4Knowledge user interface was implemented before we developed the error estimation methods in Chapter 5. These error estimation methods were not implemented within the Fish4Knowledge user interface, because additional uncertainty assessment methods are required to measure the impact of

errors from the Tracking component (Section 4.4.1, p.67). The challenge of assessing the uncertainty that propagates from the Tracking component to high-level class sizes remains unaddressed (Section 8.1.1, p.178). Therefore, the Fish4Knowledge user interface could only provide measurements of the Species Recognition Errors at the level of individual fish images, and not at the level of entire fish trajectories.

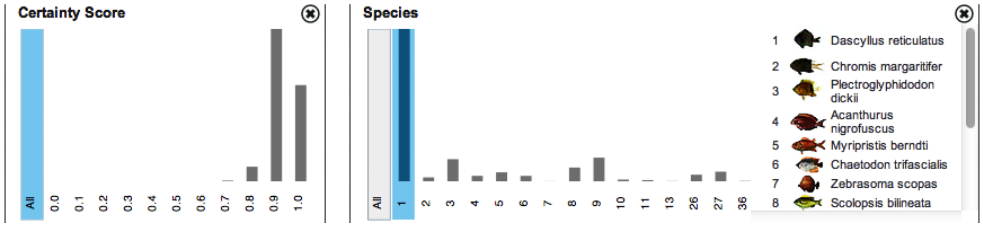

Figure 7.20: Widgets showing the certainty scores for species 1.

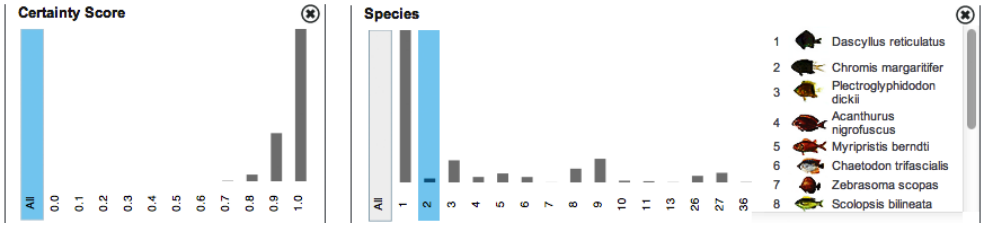

Figure 7.21: Widgets showing the certainty scores for species 2, and needed for answering question 18 of the user study (Section 7.3).

C. Comparing class sizes

For comparing the population sizes for each species, users have several options:

- Display the *Species* widget (e.g., Figure 7.21 right)
- Select *Species* as the dimension represented by the main graph's x-axis.
- Select *Stacked chart* in the *Chart Type* menu, and select *Species* as the dimension used for decomposing the fish counts (e.g., Figure 7.3 p.fc).

7.3 Evaluation

This section reports a user study that evaluates how the interface design supports user awareness of uncertainty. The study focuses on the Visualization tab and aims at identifying usability issues with the interface layout and interaction design. The study also investigates how providing certainty scores (Figure 7.1, indicating potential *Species Recognition Errors*), numbers of videos (Figure 7.13, showing potential *Fragmentary Processing*), and mean number of fish per video (Figure 7.14, compensation potential *Fragmentary Processing*) address user information needs on these uncertainty factors (Table 7.1).

7.3.1 Experimental setup

We recruited 10 marine ecologists from the research community in Taiwan. A 20-minute tutorial introduced the interface and the concept of certainty score. Participants learned the interactions needed to perform the usage scenario described in Section 7.2.3:

- Display visualizations with numbers of fish, numbers of video samples, or mean number of fish per video.
- Display visualizations using simple chart or stacked chart.
- Use filter widgets to select datasets of interest.
- Use filter widgets to compare fish distributions.

They also learned how to watch videos in the *Video* tab, since the participants of our previous user studies (Chapters 2 and 3) recurrently requested to check the footage.

Then, we asked participants to perform tasks, following a framework inspired by situation awareness methods. We exposed participants to 3 *probes*, i.e., predefined states of the interface with preselected filters and graph options. The interface showed real data from the Fish4Knowledge system. Participants were asked a total of 20 questions(Table 7.3). Participants indicated their confidence in their answers using a 5-grade scale (*Very Low, Low, Moderate, High* or *Very High* confidence).

The questions dealt with various complexity of *Fact* assessment (levels F1 to F3), *Uncertainty* evaluation (levels U1 to U3), and *Interaction* with the interface (levels I1 to I2), specified in Table 7.2. Related questions are given in Table 7.3. Levels F1-3 refer to levels of situation awareness postulated by Endsley (1988b). Levels U1-3 and I1-2 were created for our use case. Dealing with uncertainty (U2-3) implies dealing with complex facts (F3), as assessing uncertainty requires extrapolation. Thus all questions from levels U2-3 are also from level F3, and task complexity is synthesised in 4 levels F1, F2, U2, U3.

Fact Assessment	
F1 Perception	Read one single piece of information.
F2 Comprehension	Compare several pieces of information.
F3 Projection	Extrapolate unknown information from the given information.
Uncertainty Assessment	
U1 Conclusive	Only one answer is entirely true.
U2 Ambivalence	Several answers are valid. Sufficient information is provided to inform users' answers.
U3 Assumption	Several answers are valid. Insufficient information is provided to inform users' answers.
User Interaction	
I1 No Interaction	No manipulation of the interface is needed.
I2 Exploration	Manipulations of the interface are needed.

Table 7.2: Levels of complexity of the questionnaire.

The questionnaire was designed to draw attention to two uncertainty factors: *Species Recognition Errors* and *Fragmentary Processing* (Table 7.1). Issues with Fragmentary Processing were emphasized in questions Q5 and Q11-13. These questions required participants to inspect the numbers of video samples, and were asked before the questions requiring users to inspect the fish population sizes. Prior to question Q5 users dealt with absolute fish count and later with mean fish count per video sample, hence showing the effect Fragmentary Processing. Question Q13 explicitly examines the suitability of sampling size for scientific research.

In the following questions, we investigate whether participants acquired awareness of uncertainty issues due to Fragmentary Processing. Participants were not explicitly asked to inspect the numbers of video samples, and to use mean fish count per video instead of absolute fish counts. We consider that participants who do not inspect this information have not acquired the desired awareness of uncertainty.

Guiding participants' attention may artificially enhance their awareness of Fragmentary Processing. However this was desired both a priori, as Fragmentary Processing is an unfamiliar concept, and a posteriori, considering participants' poor reactivity to this awareness factor.

Usability issues and wrong answers were reported. Under uncertainty (levels U2-3), answers such as *"I don't know"* were considered as one of the possible valid answers, and were not considered as wrong answers.

This experimental setup allowed to observe:

- How users interact with the visualization when seeking information (e.g., using the widgets' overviews or the main graph).
- The usability issues that arise with either the layout or the interactions (e.g., with questions of levels I1-2, with or without interactions).
- How user confidence varies among the levels of information complexity (levels F1, F2, U2, U3).
- The quality of user awareness of uncertainty (e.g., confidence should be low with high uncertainty or wrong answers).

The small numbers of participants and questions for each condition (levels F1-F3, U1-U3, I1-I2) may not represent the general population, and our results may not be generalizable. However, our experiment is suitable for identifying major usability issues, and for eliciting recommendations for refining the means to support user awareness of uncertainty.

	Question	*Complexity*
Probe 1 (Fig. 7.17)		
Q1	What is the number of fish for the week 12?	F1 U1 I1
Q2	For which cameras are we counting the fish?	F1 U1 I1
Q3	Which week of the year has the most fish?	F2 U1 I1
Q4	At which period of the year can we observe the highest fish count?	F3 U2 I2
Q5	How many videos were analyzed for the week 12?	F1 U1 I2
Q6	What is the number of fish per video for the week 12?	F1 U1 I2
Q7	What is the fish abundance for the week 45?	F3 U2 I2
Q8	Which week of the year has the highest number of fish per video?	F2 U1 I1
Q9	What is the period of the year for which the fish population is the most abundant?	F3 U2 I1
Q10	Is it the same period of time for the camera 37?	F3 U3 I2
Probe 2 (Fig. 7.22)		
Q11	Is the number of video samples constant over hours of the day?	F2 U1 I1
Q12	Is the number of video samples constant over weeks of the year?	F2 U1 I2
Q13	Is the amount of video samples suitable for scientific research?	F3 U3 I1
Q14	Which is the most abundant species in HoBiHu?	F2 U1 I1
Q15	Which camera has the most abundant fish population from the species 2 (Chromis Margaritifer)?	F3 U3 I2
Q16	Do fish from species Chromis Margaritifer generally have high certainty scores?	F2 U1 I2
Q17	Is the abundance of species 2 (Chromis Margaritifer) lower than species 1 (Dascillus Reticulatus)?	F2 U1 I1
Q18	Is it because the video analysis may not correctly detect the species 2 (Chromis Margaritifer)	F3 U3 I2
Probe 3 (Fig. 7.23)		
Q19	Is there a correlation in the occurrence of fish from species 9, 26 and 27 over weeks of the year? (considering the entire dataset, for all time periods and all cameras)	F3 U3 I2
Q20	Is there a correlation in the occurrence of fish from species 9, 26 and 27 over hours of the day?	F3 U3 I2

Table 7.3: The questionaire of the user study. *Questions were provided in English and Chinese, and were translated by a native Chinese speaker from the Fish4Knowledge project.*

		User 1			User 2			User 3			User 4			User 5			User 6			User 7			User 8			User 9			User 10		
	Level	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.	Err.	Conf.	Usa.
Q1	F1 I1		5			3			5			5			5			5			5			5			4			5	
Q2	F1 I1		5			5			5			5			5			5			5			4			4			5	
Q3	F2 I1		5			5	X		5			5			5			5			5			5			4			5	
Q4	U2 I1		3			4			5			5			4			3			4		W	5			3			4	
Q5	F1 I2		4			4			5			5		W	5		W	4		W	3		W	2			4			3	
Q6	F1 I2		4			4	X		4			4			5			4			4			1	X		4			5	
Q7	U2 I2	W	5		W	5		W	4		W	5		W	5		W	3	X	W	5		W	3	X	W	2		W	5	
Q8	F2 I1		5			5			5			5			5			5			5		W	5	X		4			5	X
Q9	U2 I1		3			4			5			5			5			4			5			4	X		4			4	
Q10	U3 I1		4		W	4	X	W	5		W	4	X	W	5	X		5			5			4		W	4	X	W	5	X
Q11	F2 I1		5		W	5		W	3			5			5		W	4			5			4			4			4	
Q12	F2 I2		4			5		W	4			5			5			5			5		W	3	X		4			4	
Q14	F2 I1		5			5			5			5			5			5			5			5			4			5	
Q15	U3 I2		4			4			5			4			3			4			2		W	4	X		4			3	
Q16	F2 I2		4			3			4			5			5			5			5			5			4			5	
Q17	F2 I1		4			4			5			5		W	5		W	5			5		W	3	X		3			5	
Q18	U3 I2	W	1		W	3		W	4			5		W	2			5			4			4			3		W	2	
Q19	U3 I2		4			4			4			5			4			4			5			3			3			5	
Q20	U3 I2		4			3			4			5			5			5			5			4		W	2	X	W	2	X

Table 7.4: Detail of incorrect answers (Err.), user confidence (Conf.) and usability issues (Usa.).

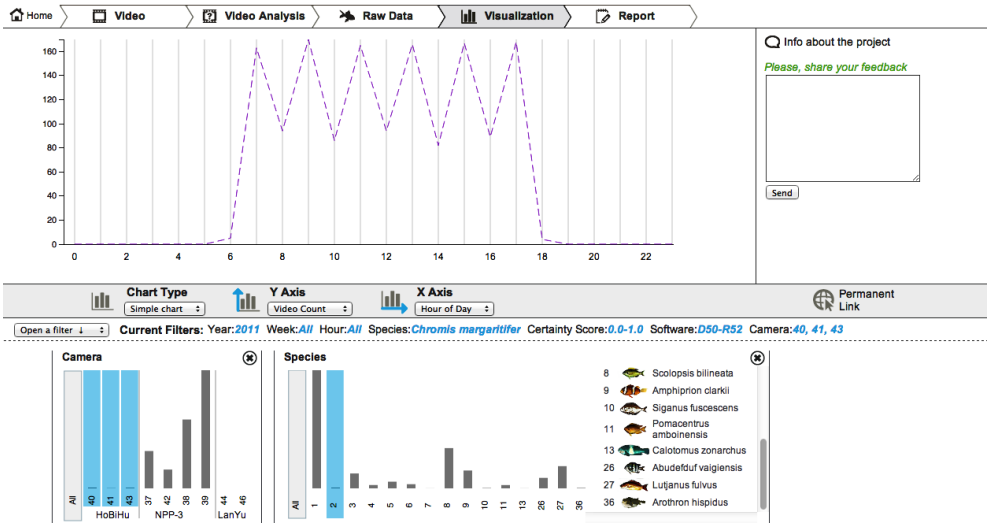

Figure 7.22: The second probe of the experiment.

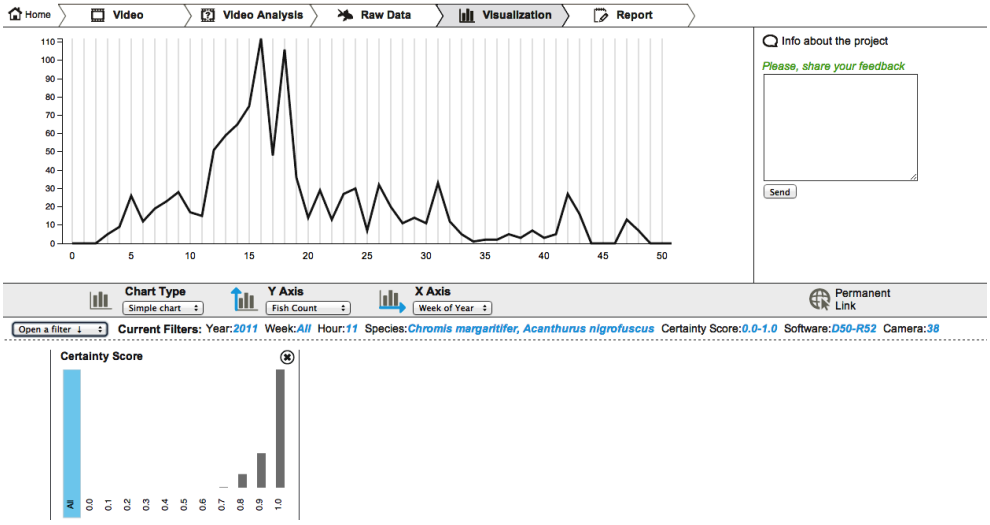

Figure 7.23: The third probe of the experiment.

7.3.2 Experiment results

The results are detailed in Table 7.4 and summarized in Figure 7.24. Question Q13 was discarded since answer correctness is ambiguous: the most precise answer is *"It depends on research goals"* as replied by one single user. To analyse participants' answers, we partitioned questions into groups containing distinct questions and representing task complexity (F1, F2, U2, or U3), interaction complexity (I1 or I2), answers' validity (Right or Wrong), and usability (Issue or No issue). With these groups of questions, we can observe the impact of tasks and interface complexity on participants' awareness of uncertainty.

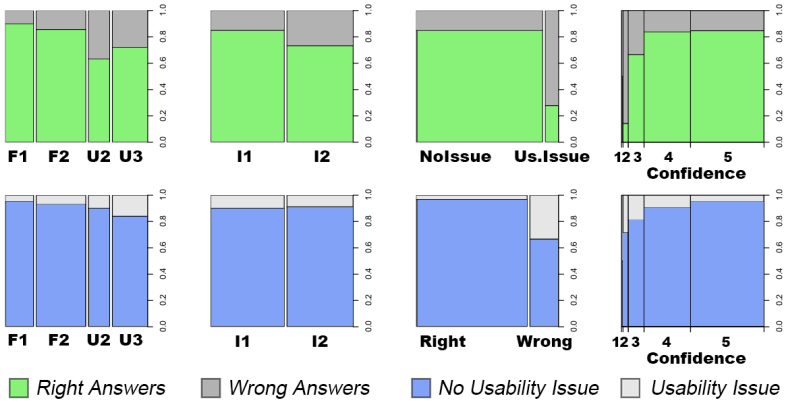

Figure 7.24: Proportions of right and wrong answers (top) and usability issues (bottom) for each question groups (x-axes).

Participants' confidence in their answers is shown in Figure 7.25. Participants were generally highly confident, even when answers were wrong or uncertainty was high. Level 5 is often the default answer, but some participants consider level 4 as the default, making comparisons difficult, e.g., for *Participant 4*, level 4 is weak confidence while being the optimal confidence for *Participant 9*.

To compare participants' confidence, we focus on *confidence drifts* (i.e., relative changes in confidence) rather than absolute confidence levels. For instance, confidence drifts are calculated by 1) averaging each participant's confidence for groups F1 and F2 distinctively; 2) subtracting each participant's average confidence to get the participant's confidence drift between groups F1 and F2.

We analyse confidence drift between question groups (Figure 7.26):

- The groups **F1-F2, F2-U2,** and **U2-U3** represent questions with increasing information complexity.
- The groups **I1-I2** represent questions involving interaction or not.
- The groups **Right-Wrong** (or R-W) represent questions which answers were right or wrong.

- The groups **NoIssue-Us.Issue** represent questions for which no usability issue occurred, or were usability issues were identified by the interviewers. The usability issues are reported in Section 7.3.3.

We also distinguish the effect of uncertainty and user interface, represented with the following question groups:

- The groups **Certain-Uncertain** (F1∪F2 against U2∪U3) represent questions impacted by uncertainty issues or not.
- The groups **Certain,I1-Certain,I2** (Certain∩I1 against Certain∩I2) represent questions involving interaction or not, while uncertainty issues must be considered.
- The groups **Uncertain,I1-Uncertain,I2** (Uncertain∩I1 against Uncertain∩I2) represent questions involving interaction or not, while no uncertainty issues need to be considered.

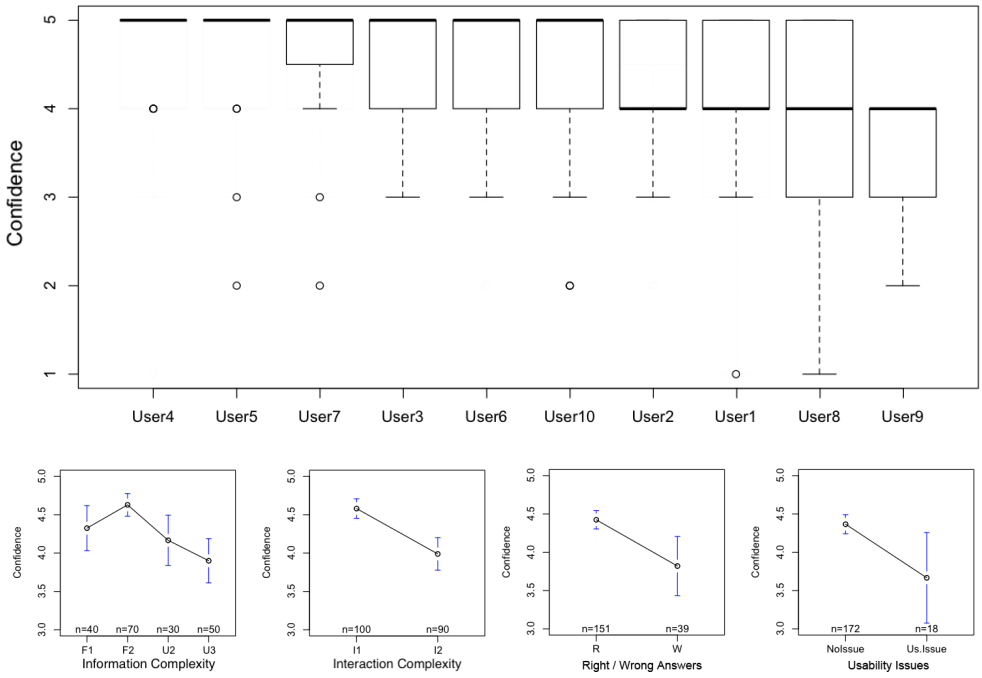

Figure 7.25: Confidence levels for all questions (upper boxplots) and groups of questions (lower line charts, with mean +/- 2 standard deviations).

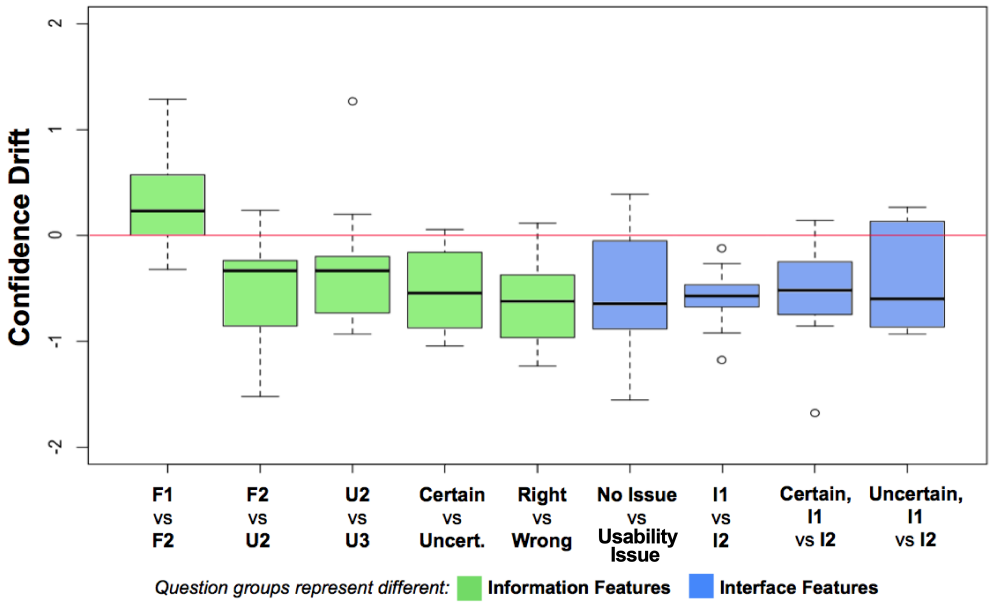

Figure 7.26: Confidence drifts per question groups.

Except for the groups *F1-F2*, increasing question complexity yielded a decrease in participants' confidence. Their **confidence decreased whether the complexity arise from the information features**[3] **or from the interface features**[4].

However, the statistical significance of the observed confidence drifts is not established. Using Welch t-test (compensating for the unequal variance shown in Figure 7.25), we tested the confidence drifts of each participant. For instance, we selected the answers of a single participant. We aggregated this participant's confidence in the answers to questions from group F1 and F2. We then applied Welch tests to assess the statistical significance of the difference in the participant's confidence between the groups F1 and F2 (i.e., the probability that the observed confidence drift occurred by chance, while there is no actual confidence drift but just random variations of the participant's confidence).

The resulting p-values are generally much greater than 0.05 (Figure 7.27). The number of cases where $p < 0.05$ happened with a frequency of around 0.05, and thus may be due to random effects. Each user's confidence had mostly the same value (e.g., level 5 or 4 by default, Figure 7.25 top). Participants' confidence was rarely lower than their default confidence levels. Thus, in general, participants' confidence levels do not differ significantly between question groups.

[3] i.e., between the question groups *F2-U2, U2-U3, Certain-Uncertain*, and *Right-Wrong*

[4] i.e., between the question groups *I1-I2, Us.Issue-NoIssue, CertainI1-CertainI2*, and *UncertainI1-UncertainI2*

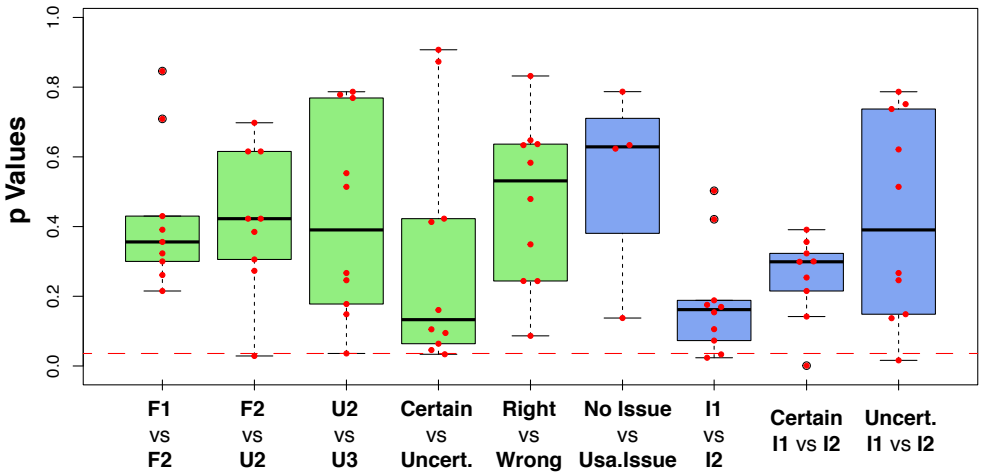

Figure 7.27: Results from Welch t-tests (each point represents a user). *T-tests were skipped if a user's confidence was equal for all answers, i.e., for groups F1-F2 (1 user), F2-U2 (1 user), CertainI1-CertainI2 (1 user). T-tests were skipped for group Us.Issue-NoIssue if a user had no usability issues (6 users).*

However, two observations give credence to the conclusion that **uncertainty and interactivity had similar effects on participants' confidence**:

- Observation O1: Except for groups *F1-F2*, confidence consistently decreased with questions' complexity. If the effect was random, confidence drifts would show as many increases than decreases.

- Observation O2: Confidence drifts are the most significant for the groups *I1-I2*, with the lowest variance and p-values (Figure 7.27), and median drift similar to that of the groups *Certain-Uncertain* (Figure 7.26).

We noticed that wrong answers and usability issues had an important effect on participants' confidence (Figure 7.25), but are outlying conditions (low numbers of observations, Figure 7.24). Further, wrong answers and usability issues often occurred together. Thus we repeated the analysis on right answers with no usability issue, and obtained similar observations (O1-O2).

We conclude that either interacting with the visualization, or analysing uncertain data, had similar effects on users' perception of uncertainty. This biases user awareness of uncertainty: low user confidence may not assess the strength of users' data interpretation, but may reflect difficulties with using the interface.

7.3.3 Interpretation and recommendations

This section discuss the insights drawn from the qualitative analysis of participants' answers and behaviors when interacting with the interface.

Over-confidence - Participants' confidence was generally high, even for wrong answers and uncertain information. Over-confidence may be due to the presence of observers during the task, inducing a will to perform well (Hawthorne effect). Participants may feel the need to perform well, thus to express only sure answers. We recommend that studies of user awareness of uncertainty **give strong incentives for users to express their low confidence**. For example, the 5-grade Likert scale may be reduced to a single checkbox for users to indicate when they are not fully confident.

Fragmentary Processing - Users overlooked uncertainty due to Fragmentary Processing. No spontaneous *Projection* (F3) of possible scarcity of video samples occurred. For instance, questions Q7, 10, 19, and 20 did not show numbers of videos, and most users did not spontaneously investigate potential imbalances in numbers of videos. Hence answers were fortuitously correct, and no right answers were given to Q7 that concerned a time period for which no video samples were available. However, *Perception* (F1) and *Comprehension* (F2) of numbers of videos were correct (Q5, 11, 12).

Experimental setup - Issues with overlooking Fragmentary Processing may be related to the experimental setup. The terminology may be ambiguous, e.g., *"What is the fish abundance?"* (Q7) may be interpreted as a need for raw fish count (i.e., simply reading the graph, instead of modifying it to check the numbers of of video samples). Further, the early prototype used for the experiment provided widget histograms that could only display raw fish counts, not the mean number of fish per video sample. Thus it may seem that raw fish count is the main metric for fish abundance. It may have deflected users' attention from the mean number of fish per video, and the potential issues with Fragmentary Processing.

Choice of metrics - Fragmentary Processing issues are similar to sampling size issues, e.g., insufficient number of samples, a well-known concern in marine ecology. However, Fragmentary Processing is specific to computer vision, and not assimilated by ecologists. They may expect video stream processing to be continuous, rarely missing videos. Hence we recommend that Fragmentary Processing issues are always made explicit. **Raw fish counts can be misleading, and by default, should not be displayed**. Mean number of fish per video could be displayed together with an indication of the sampling size, e.g., encoded as an extra visual dimension such as transparency, or by showing confidence intervals. Boxplots, a type of graph available but not investigated in this study, can also show sampling size, e.g., encoded in their width (McGill et al. 1978). It may prevent memory-loss (e.g., forgetting the numbers of videos).

Ultimately, the metrics used to represent population sizes should integrate the estimation of classification errors introduced in Chapter 5. The number of fish averaged over video sampled should be the *corrected* number of fish resulting from Chapter 5's methods. Displaying variance estimates is also crucial to uncertainty awareness. The variance estimates must considered both the variance related to the number of video samples (equations (4.2) and (4.4), p.71), and the variance of

Chapter 5's methods (related to class sizes of the test set and of the dataset under analysis, Section 5.4, p.85).

This approach provides more complete information on the uncertainty in the computer vision end-results. However, it requires users to deal with highly complex information, involving several layers of uncertainty assessment and rather complex equations. Yet users must understanding the underlying methods that provided such results, otherwise they cannot make informed decisions when interpreting the population sizes. Hence future work is required to investigate the means to explain and visualize such complex information combining the uncertainty assessment *Fragmentary Processing* and *Noise and Bias* due to classification errors.

Certainty scores - No users spontaneously considered the certainty scores, which are unfamiliar and complex. However, some users spontaneously noticed uncertain factors not included in the scope of the tutorial and questions: issues with Fields of View, Duplicated Individuals and differentiating Biases from Noise (Table 7.1). Assessing these uncertainty issues may be more important to build end-users' trust and confidence than providing certainty scores. Furthermore, using certainty scores as filters may introduce biases in end-users' data interpretation. For instance, the sizes of populations recognized with high certainty scores only may not be representative of the actual population sizes. If the large majority of fish have average or low certainty scores, very high certainty scores are conceptually close to being outliers. Hence, **certainty scores did not demonstrate opportunities for supporting users' uncertainty awareness**.

Learning curve and usability - The increase in confidence observed between F1 and F2 questions, although not statistically signifiant (p=0.07, Welch t-test), suggests an effect of the learning curve. Overcoming slightly higher complexity may induce a sentiment of higher level of expertise. User confidence may be reinforced because they gained experience with the interface, while the information they had dealt with does not justify increased confidence.

Similarly, difficulties interacting with the interface, and using unfamiliar functionalities, may reduce user confidence. As users need to learn the interaction features, they may not be confident that their interactions with the interface, and thus the obtained information, are correct.

We thus recommend to provide tutorials and memos that summarize the basic uncertainty exploration steps needed for valid data analysis. These should be easily accessible from the user interface, for quick checks while interacting with the data.

Filter widgets - Predefined filters of the 3rd probe were often overlooked, probably because participants did not select the filters up themselves. However, it suggests potential **attention tunnelling issues with the layout design**. Users' attention may be directed to more salient features of the interface, e.g., the main graph, rather than the selected filters. In the next version of the interface, filters were reinforced and highlighted in *Zone B* (Figure 7.3). The latest version of Zone B (e.g., Figure 7.13 p.158) describes both the main graph and the filters in natural language, and serves as a title for the main graph.

The dimensions not used for filtering (e.g., all years, all certainty scores, Figure 7.23) can saturate users' working memory, and are no longer displayed in the refined interface. Participants tried to click on the filter summary, thus we added interactions for resetting the filters (cross buttons next to each dimension in the filter summary, e.g., Figure 7.13 p.158).

Interaction and layout design - The interaction design for manipulating the widgets and the main graph was welcomed and easily understood (*"It is very nice, I can display anything I want."*). Participants used either the widgets' histograms and the main graph when appropriate. It suggests that **our interaction design is reasonable, while our layout design raised most of the usability issues**.

We recommend that uncertainty is always salient in the interface. It may complicate the layout design, yet it may be the best tradeoff regarding the high risk of misinterpretation. **Our design of simple graphs in multiple views is intuitive and quickly understood. However, it may over-simplify data exploration at the cost of concealing the uncertainties. Over-simplification may enhance attention tunnelling, memory loss and over-confidence.**

7.4 Conclusion

We presented a design for visualizing multidimensional and uncertain computer vision results. We evaluated the interactive design for exploring the multiple dimensions and uncertainty factors of the data. It aims at limiting information overload and interface cluttering, while facilitating the exploration of data dimensions with flexible visualizations. It supports preliminary data analyses for a wide range of potential usage of the dataset, which may be achieved with specialized data analysis techniques. Our design for preliminary data exploration can help users to familiarize themsleves with novel datasets, and identify issues and uncertainties that may impact further data analyses.

Our interaction design was found intuitive and easy to understand, although the dataset was unfamiliar to users. The layout and interaction principles can integrate information on the uncertainty factors presented in Chapter 4, the error estimation methods presented in Chapter 5, and the visualizations of classification errors presented in Chapter 6. Thus the interface design addresses the requirement 4-d *Communicate uncertainty to end-users*, identified in Chapter 2 (p.36).

Our design can contribute to similar use cases, possibly within domains other than marine ecology. For instance, the interface template was reused for a demonstration of the SightCorp company's classification system (Figure 7.28). However, for commercial applications, the information provided on uncertainty factors require a different approach than for scientific applications. For instance, marketing strategies and the impact on customers' trust must be carefully considered.

Our evaluation method, inspired by Situation Awareness, allowed us to distinguish issues of either layout or interaction design. This evaluation methodology can be applied to other evaluations of interactive visualizations.

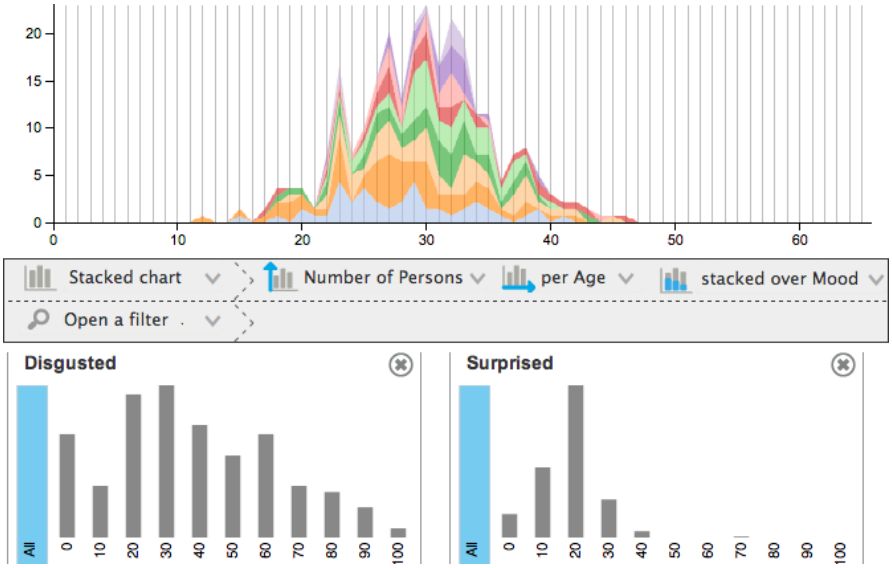

Figure 7.28: Reuse of the interface for the SightCorp company, where class sizes represent human emotions recognized by the company's computer vision system (https://sightcorp.com/).

Our main finding is that user confidence is generally high, and subjectively influenced by the interactions with the visualization: interaction complexity had effects similar to uncertainty itself. Using simple graphs with multiple views achieves high intuitiveness but may have negative effects on user awareness of uncertainty. The intuitiveness of the graphs and interactions may have contributed to overconfidence through a sentiment of mastering the interface and its information, which led to overlooking uncertainty issues. Furthermore, uncertainty assessment requires the visualization of several graphs within multiple views, which may result in attention tunnelling and memory loss, and induce misinterpretations and unawareness of crucial information on uncertainty.

We derive two main recommendations for improving the visualization interface and supporting user awareness of uncertainty:

- **Salient and persistent display uncertainty measurements**. The main visualization (e.g., used to explore population sizes) should always display indications of the uncertainty issues, e.g., encoded with visual features such as transparency or boxplots. However, detailed information on uncertainty assessment methods should be displayed in dedicated tabs, to avoid cluttering the main visualization.

- **Exclude display of uncertainty-agnostic metrics**. The main visualization should not display metrics that exclude all information on uncertainty issues. For example, to account for missing video samples, raw fish counts should not

be displayed, and mean fish count per video sample are preferable. However, uncertainty-agnostic metrics are of interest for explaining the uncertainty assessment methods (e.g., how mean fish count per video sample are calculated). Hence, uncertainty-agnostic metrics should be displayed in the interface tabs explaining the uncertainty assessment methods.

These findings contribute to answering our seventh research question: *How can interactive visualization tools support the exploration of computer vision results and their multifactorial uncertainties?* We introduce a layout design (using tabs for exploring the data processing steps) and an interaction design (swapping the dimensions represented by the graphs' axes) that provide support for exploring multidimensional and uncertainty datasets, in the domain of computer vision and classification and beyond. However, several challenges for supporting user awareness of uncertainty remain unaddressed. In particular, future work is required for designing tutorials and explanations, as discussed in Chapter 8.

Chapter 8

Conclusion

This thesis investigated key uncertainty issues that impact end-users of computer vision and classification systems, and the means to assess the resulting uncertainty. We identified high-level user requirements in the domain of computer vision for population monitoring (Chapter 2). We collected insights on end-users' development of informed trust in classification results provided by computer vision technologies (Chapter 3). From these insights, we identified key uncertainty factors of concerns to end-users (Chapter 4). We then developed uncertainty assessment methods and tools that address end-users' concerns: statistical methods for estimating classification errors in end-results (Chapter 5), visualizations for assessing classification errors (Chapter 6), and an interactive visualization for exploring computer vision results and their multiple uncertainty factors (Chapter 7).

To conclude this thesis, we reflect upon higher-level insights we gained on approaches to addressing uncertainty issues from the perspective of end-users. Addressing end-user requirements may challenge uncertainty assessment practices (Section 8.1). Nonetheless, we recommend to develop a common framework for assessing classification errors, addressing the concerns of both end-users and developers (Section 8.2). Such an endeavour requires the development of end-users' classification literacy. Hence, finally, we discuss the need for developing classification literacy in the general public (Section 8.3).

8.1 Practical challenges with end-users' requirements

Whether tuning or using classifiers, **both developers and end-users share the need for estimating the errors to expect in practical applications**. In practice, errors can arise from each software component integrated in the classification system (Chapter 4). Assessing the combined errors introduced by each component of a classifi-

cation system is a first challenge (Section 8.1.1). In practice, the characteristics of end-user datasets impact the magnitude of errors to expect (Chapter 5). Accounting for the potential differences between test sets and end-user datasets is a second challenge (Section 8.1.2).

8.1.1 Challenges with assessing error propagation

To assess the errors of classification results drawn from a pipeline of classification components, **test sets must be representative of the errors that propagate from one classification component to the next**.

Example 1

Uncertainty propagation with two classifiers:

*In computer vision systems, **binary classification components** (e.g., differentiate moving objects from background objects) **often precede multiclass components** (e.g., detect the type of objects). This pipeline of classification components results in the combination of errors from binary and multiclass classifiers. For instance, background elements may be misclassified as objects of interest (i.e., False Positives of the binary classification components), and then incorrectly assigned a type of object (i.e., False Positives propagated to the multiclass classification component). To assess the uncertainty propagation, **the test set of the multiclass classification component should represent the errors from the binary classification component**, e.g., by including an additional class to represent the binary classification errors (Section 4.4.1, p.68).*

Using test sets that are representative of the error propagation is a challenge. Ideally, each classification component should be trained and tested with datasets that represent the errors to expect from the previous component. In practice, however, classification components are often trained and tested using distinct groundtruth datasets, disconnecting the components from each other's potential errors. Unfortunately, **training and testing pipelines of classifiers using distinct datasets for each classifier may not provide end-users with representative uncertainty assessments**, nor optimal classification systems.

Measuring the errors that propagate in pipelines of classifiers requires that classifiers are tested with datasets that represent the errors from the previous classifiers. Whether classifiers are trained with different datasets is secondary, and rather concerns the tuning of classifiers' parameters.

In any case, **it is challenging to collect test sets that represent the errors of each classification component**. For instance, classification components may be developed separately. Examples of other components' errors may not be known when a classifier is developed, e.g., if classification components are developed at the same time, or for several potential pipelines. Furthermore, the combined errors of classification components may be critically ambiguous. When manually classifying the test sets, **humans may not be able to decide or agree on the true class of ambiguous objects**.

Example 2

Specifying the propagated errors:

With computer vision, after classifying the objects appearing in each video frame, tracking algorithms can detect the trajectory of individual objects across video frames. Erroneous trajectories may contain objects of different classes. Such mixed-class trajectories may not be confidently considered as belonging to a single class. Hence, ***it is challenging to measure the errors that propagate from the tracking components to the next classification components****.*

Assessing the errors that propagate along a pipeline of classification components challenges uncertainty assessment practices. Using test sets that are representative of other components' errors entail several issues. When training or tuning classifiers, developers may not be able to consider other components' errors, e.g., the other components may be under development and their errors unknown. Otherwise, it may be complicated or costly to manually label each component's errors.

Assessing the entire pipeline of algorithms as a black box may be easier than assessing each component individually. This approach requires test sets that are representative of the combined errors from each component of the pipeline. For example, with a pipeline comprising a binary classifier followed by a multiclass classifier (e.g., Example 1 above), measuring the combined errors requires the test set to include all the classes of the multiclass component and a class representing the Negative class of the binary component. Such test sets may be challenging to collect, e.g., for more complex pipelines. The test sets collected to evaluate pipelines as black boxes must represent the errors of each components within the black box, and thus, may as well be used to assess each component individually.

8.1.2 Challenges with assessing the errors in specific end-results

End-users may apply classifiers to datasets that differ from the test sets. For instance, end-user datasets may have different class proportions (e.g., small or empty classes) and feature distributions (e.g., lower data quality). Such differences between test and target sets threaten the validity of error estimations (Chapter 5). Accounting for the potential differences between test sets and end-user datasets challenges uncertainty assessment practices.

Example 3

Inconsistency between test and target sets:

End-user datasets may have different class proportions than the test set, e.g., some classes may be much larger and others may be empty. end-user datasets may also be of lower quality in recurring situations, for instance, lower image quality at dawn and dusk. Image quality may impact specific classes more than others, e.g., some classes may be empty as they are entirely misclassified at dawn. ***Both class sizes and data quality impact the magnitude of errors to expect in end-user datasets****.*

It is challenging to specify the potential characteristics of end-users datasets, e.g., the distribution of potential class sizes, data quality or class features. In particular, class sizes or feature distributions may co-vary, and such covariance between depen-

dent variables cannot be captured by a single test set. It is also challenging to refine error estimations to account for the characteristics of end-user datasets (e.g., using error rates specific to lower quality images, or other features).

Beside cases where test sets and end-user datasets differ significantly (e.g., due to altered feature distributions), **random differences among datasets can yield significant variance when estimating the classification errors** in end-user datasets (Section 5.2.4, p.80). For instance, when estimating the number of errors to expect in a specific dataset, the variance magnitude may show that error estimations are unreliable, e.g., with extremely large confidence intervals.

Error estimations have particularly high variance when class sizes are small, either in test sets or end-user datasets (e.g., even with class sizes of several hundreds of items, Figure 5.1, p.81). Hence **variance issues must be considered when assessing classification errors** (e.g., using the Sample-to-Sample method introduced in Section 5.4, p.85).

The challenges with estimating errors in specific datasets are also challenges with evaluating of classifiers: **if the test sets do not support reliable estimations of the errors to expect in specific datasets, then the test sets do not provide reliable assessments of classifier performance**. This is a first rationale for developing a unified uncertainty assessment framework, encompassing both end-users' and developers' tasks of tuning classifiers and estimating the errors to expect in specific end-results.

8.2 Unified classification assessment framework

When assessing classifiers, end-users and developers may have different goals and approaches (Section 6.2, p.118). For instance, developers are primarily concerned with reducing the errors of classifiers. Developers can tune classifiers to reduce their errors, e.g., by setting parameters that are typically complex and unfamiliar to end-users.

End-users are primarily concerned with the errors that may occur when applying classifiers to their specific datasets (e.g., would this classifier or parameter setting reduce the errors for the most important classes?). **Without understanding the errors to expect in practical applications, end-users cannot assess which classifier or tuning parameters suit them best**.

Developers may be provided with training and test datasets, but may not be provided with information on the potential end-user datasets (e.g., would users apply classifiers to datasets with potentially small or empty classes, or altered feature distributions?). **Without understanding the datasets to expect in practical applications, developers cannot fine-tune classifiers' tuning parameters**.

Hence both end-users and developers need to consider the potential characteristics of end-user datasets. As discussed in Section 8.1, this entails several practical challenges:

- End-users may apply classifiers within a pipeline of classification components,

hence **end-user datasets may include errors introduced by the previous classification components**. Collecting test sets that represent the errors of previous classification components can be costly and complicated.

- End-users may apply classifiers to datasets where **classes may have different feature distributions in recurring situations** (e.g., lower data quality) **which bias error estimations** (e.g., low image quality due to reduced natural light yields higher error rates than those measured with test sets). Specifying the variations of feature distributions can be costly and complicated, as is collecting test sets that represent the potential feature distributions (e.g., how the feature values may covary).
- Even if end-users apply classifiers to datasets that are highly similar to the test sets, **variance issues with random variations from the test sets may yield unreliable error measurements** (e.g., numbers of errors estimated with extremely large confidence intervals).

Collecting test sets that represent the variety of potential end-user datasets is a major practical challenge. Even if this challenge cannot be addressed, we advocate new paradigms for uncertainty assessments that can bridge the gap between end-users and developers, i.e., between error measurement in test sets and error estimations in end-user datasets.

We argue that **end-users should take part in classifier tuning**, joining forces with developers by using a unified uncertainty assessment framework (Section 8.2.1). To develop such uncertainty assessment framework, we advocate that **error measurements should be mapped to datasets' feature values** (Section 8.2.2), and that **the variance of error measurements should be systematically considered** (Section 8.2.3).

8.2.1 Tuning classifiers in collaboration with end-users

Tuning classifier parameters aims at minimising classification errors when applying the classifiers to end-user datasets. Developers can focus on reducing the errors for specific classes (e.g., the classes yielding most errors) or class features (e.g., the features yielding the most errors). Developers should address end-users' priorities, and **reduce errors for the classes and features that are most *important* to end-users**, e.g., the most frequent or the most valuable.

Hence **tuning classifiers requires both technical expertise** (i.e., knowledge of the relationships between classifiers' parameters, feature distributions, and resulting errors) **and domain expertise** (i.e., knowledge of the classes and feature distributions that are the most important or the most frequent). Enabling end-users to collaborate with developers for tuning classifiers can overcome issues with test sets' limitations. Extensive test sets are required to capture the potential variations of class proportions and feature distributions. even if extensive test sets may not be collected, **end-users can use their domain expertise to guide developers attention to the most important or frequent situations**.

Furthermore, classification technologies have become largely available in integrated libraries and frameworks (e.g., R, Python, Weka, RapidMiner). **Classification components are readily accessible to a public with little to no expertise in the underlying classification algorithms**. The main tasks when implementing such classifiers consist of training and tuning existing components, rather than developing new classification algorithms. In this context, classifiers disseminate to a public who may have more expertise in the application domain than in the classification technologies. Thus a **public with the expertise of end-users rather than developers may be increasingly tasked with tuning classifiers**. Hence the demand may increase for unified classification assessment frameworks, addressing both end-user-oriented and developer-oriented concerns. This trend also entails an increasing need for developing the classification literacy of domain experts (Section 8.3).

8.2.2 Mapping error rates and feature distributions

The errors to expect in classification results are largely impacted by the feature distributions of the datasets to classify. When estimating the number of errors in classification results, existing error estimation methods are largely biased when feature distributions differ between the test sets and the classified datasets (Section 5.6.3, p.103). However, specifying the relationships between error rates and feature distributions can support refinements of error estimations (Section 5.7, p.105).

Hence **the magnitude of errors to expect should be estimated as a function of the feature distributions**. However, it may be practically impossible to collect test sets for each combination of classes and feature values. For instance, *"Because [Machine Learning (ML) models] typically have a large number of inputs, it is not possible to thoroughly test even simple models, which leaves open the question of how a ML model will perform in a given situation."* [1]

Nonetheless, the **relationships between error rates and feature distributions can be estimated, e.g., using linear models** such as those introduced in the Logistic Regression methods (Section 5.7.2.A, p.106). Provided with estimations of error rates for the feature distributions of interest, end-users may better assess the classification uncertainty, and decide of the most important ranges of features for which errors must be reduced when tuning classifiers.

Furthermore, error measurements are often summarized into a single metric that encompasses errors occurring in different situations. Developers often measure the errors that occur over several parameter settings by using the Area Under the Curve (AUC) metric, i.e., by plotting classification errors as a function of parameter settings, and calculating the surface under the resulting curves (e.g., an ROC curve). We argue that **AUC is not informative for end-users** who apply classifiers with only one parameter setting, not all possible parameter settings.

[1] Report from Informatics Europe and ACM Europe Technology Policy Committee (EUACM), *When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making*, 2018 (p.10). URL: *http://www.informatics-europe.org/component/phocadownload/category/10-reports.html?download=74*

Mapping error rate estimations as a function of feature values can refine AUC-like summary measures. **It is more informative for end-users to plot errors as a function of feature values, and to measure the areas under these curves**. This measurement represents the errors that occur over several feature values. This approach is representative of the situations that may occur when applying classifiers, i.e., as the class features may vary but not the parameter settings.

8.2.3 Uncovering variance issues

When estimating the number of errors in classification results, existing error estimation methods are unbiased if end-user datasets are similar to the test sets (e.g., with similar error rates). However, **random error rate variations can have a significant impact if class sizes are small, either in test sets or in end-user datasets**. The variance of error estimations can be critical e.g., even for class sizes of several hundreds of items Chapter 5, (Section 5.2.4 and Figure 5.1, p.81).

Hence we recommend that the variance of error measurements (i.e., in test sets) and error estimations (i.e., in end-user datasets) always be estimated. For example, when visualizing classification errors measured in test sets, confidence intervals should be drawn to represent how much error rate may randomly vary. Although simple to compute, **error rate variance is largely overlooked when assessing classifiers**. For example, ROC curves and Precision-Recall curves are seldom provided with confidence intervals for the error rates they display.

The variance or error rates measured in test sets (i.e., random variations across potential test sets) can be easily estimated using basic frequentist methods, e.g. for error rates θ and class size n, $V(\theta) = {}^{\theta(1-\theta)}/_{n}$. When training and tuning classifiers, information on the potential target sets may be unknown. In such cases, target sets can be naively considered to be the same size as the test set. Using the Sample-to-Sample method (Section 5.4, p.85), the variance of error estimations can be naively estimated, e.g., for error rates θ' in end-user datasets, $V_{naive}(\widehat{\theta'}) = {}^{\theta(1-\theta)}/_{n} + {}^{\theta(1-\theta)}/_{n} = 2V(\theta)$.

Variance estimates should be available when assessing classifiers and tuning their parameters. Developers tend not to consider error variance when tuning classifiers, because **the same test set is used to measure the errors entailed by the parameter settings, thus error variance is highly similar over the parameter settings**. However, measuring variance can help direct attention to classes that exhibit high variance, as the errors from these classes may need to be reduced in priority.

Further strategy can be investigated for reducing critical variance issues. For instance, the size of the training set may be reduced to increase the size of the test set. Such an approach may increase the error rate, but reduce the error rate variance (as the test set is larger). **Estimating error rate variance is required to investigate the gain and loss when choosing the sizes of training and test sets**.

Further variance issues arise when assessing classification errors for specific feature distributions. When assessing the errors that are specific to certain feature

values, feature-specific error rates may be directly measured from the test set. If test sets do not contain examples of the feature values of interest, feature-specific error rates can be estimated, e.g., using linear models (Section 8.2.2). In this case, variance issues concern not only random variance across similar datasets (e.g., $V(\theta) = {}^{\theta(1-\theta)}/_{n}$ and $V(\widehat{\theta'}) = {}^{\theta(1-\theta)}/_{n} + {}^{\theta(1-\theta)}/_{n'}$) but also the variance of the feature-specific error estimation (e.g., with linear models, the squared errors between actual and predicted error rates).

Future work is required to investigate methods for deriving error rates to expect for specific feature values, and for assessing the variance of such feature-specific error estimation. **Estimating and interpreting the variance error estimations is a complex but necessary task, required for tuning classifiers and for assessing the errors to expect in specific datasets**. To enable end-users to harness such complex but necessary uncertainty assessment, it is necessary to develop end-users' classification literacy.

8.3 Developing classification literacy

Classification technologies already pervade many facets of society, impacting most professional domains (e.g., data-driven organization) as well as our personal life (e.g., information retrieval, social media, insurance, health). These technologies, together with the availability of extensive data collection techniques, open new perspectives and valuable opportunities. However, uncertainty issues remain a major challenge. **Uncertainty can be readily overlooked** as innovative applications thrive, and economic opportunities may prevail upon technological shortcomings. Yet, if left unaddressed, **uncertainty issues can put end-users' interests in jeopardy, and have direct sociological or economical impacts**.

Many application domains face crucial uncertainty issues, we give only a few examples here. When applying machine learning classification for scientific research (e.g., the Fish4Knowledge project) uncertainty issues compromise the scientific validity of data interpretations. When classifying patients or cell images for medical diagnosis, classification errors can leave health issues undetected. When applying machine learning to predict inmates' recidivism, biased prediction systems can yield inadequate or discriminatory decisions when granting or postponing inmate release. When predicting the risks associated with loan applicants, biased prediction systems can yield risky or unjust decisions when granting loans and calculating interest rates. When automatically classifying goods within factories, undetected defects can put consumers at risk, and discarded but flawless goods yield direct economic loss.

Addressing uncertainty issues is crucial for developing trustworthy classification systems. For instance, the topic of trustable and explainable machine learning has gained interest in recent years. However, existing work mainly targets an audience of experts, e.g., engineers or researchers seeking to improve machine learning systems, or exploring the sources of error. While **most efforts are spent on uncovering how uncertainty arises in low-level machine learning algorithms, higher-level**

uncertainty assessment from the perspective of end-users remains largely unaddressed. In particular, little support is provided to domain experts with no expertise in machine learning.

It is crucial that non-expert end-users be provided with tools and methods for understanding classification uncertainty. End-users need to be aware of the practical implications of classification errors for their specific applications. Considering the crucial impacts of uncertainty, e.g., regarding safety, economic, legal or moral implications, end-users must comprehend the uncertainty issues and make informed decisions when choosing, tuning and using classification systems. These are the goals, from a high-level perspective, to which this PhD thesis contributes.

Our research contributes to developing classification literacy with requirements for end-user-oriented uncertainty assessment (Chapter 2), insights on end-users' behaviours when assessing classification errors (Chapters 3, 6 and 7), visualization tools for assessing classification errors and other uncertainty factors (Chapters 6 and 7), and statistical methods for estimating classification errors in end-results (Chapter 5). However, many issues remain unaddressed.

First, **uncertainty assessment methods are unavailable for key uncertainty factors**, e.g., error estimations under varying feature distributions (Section 5.6.3, p.103), human errors when labelling groundtruth test sets, and with computer vision technologies, duplicated individuals and heterogeneous fields of views which bias class size estimates (Section 4.4, p.67). Beyond developing end-user-oriented uncertainty assessment methods, **explaining the basic concepts of classification errors to end-users remains challenging**. For instance, we identified crucial issues with the technical terminology (Section 6.7, p.134).

Regarding the need to develop end-users' classification literacy, **we come to the conclusion that the most valuable investments in future research should be placed in developing tutorials and guidelines** for explaining classification uncertainty. Developers of classification systems should be provided with guidelines to ensure that the uncertainty of their classification system is explained understandably and comprehensively, and that their choices when tuning classifiers address end-users' concerns. End-users should be provided with tutorials and guidelines that enable them to identify which uncertainty assessments are necessary, and require them from the developers of classification systems.

Investigating tutorials' understandability and completeness is also required for enabling valid user studies. When conducting user studies of tools to deal with classification results and their uncertainty, participants need to be provided with explanations of the tools and the classification results and uncertainty. These explanations can have critical impact on user studies. If explanations are not understandable or incomplete, users may perform poorly. Hence, **user performance may not reflect the quality of the tools, but the quality of the explanations**.

8.4 Epilogue

The issues with machine learning uncertainty extend beyond those covered in this thesis. For instance, besides the technical and educational implications discussed in this conclusion, Informatics Europe and ACM Europe[2] identify 4 other aspects of machine learning error and bias: the ethical, legal, economical, and societal implications. Different communities need to work together to address the different aspects of uncertainty issues. The public sector and civil society are crucial actors to drive the efforts required to address the implications of machine learning uncertainty. Public institutions and non-profit organizations have the leverage and interests for requiring transparent and accountable machine learning systems, while there is little benefit for commercial organizations to invest in providing them. *"While more fairness and justice would of course benefit society as a whole, individual companies are not positioned to reap the rewards. For most of them, in fact, [harmful machine learning systems] appear to be highly effective. Entire business models, such as for-profit universities and payday loans, are built upon them."*[3] As individual citizens and public servants are key stakeholders in machine learning uncertainty problems, this reinforces my personal opinion that essential and urgent future work is to invest in providing machine learning literacy to the general public.

[2]Report from Informatics Europe and ACM Europe Technology Policy Committee (EUACM), *When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making*, 2018. URL: *http://www.informatics-europe.org/component/phocadownload/category/10-reports.html?download=74*

[3]Book by Cathy O'Neil, *Weapons of Math Destruction*, 2016 (p.202).

Appendix A

Study of User Trust and Acceptance

This appendix details participants' answers in the user study introduced in Chapter 3.

- A summary of the questionnaire and related visualizations (Section A.1).
- The responses of each participant to the multiple choice questions (Tables A.1-A.3) and the free-form text questions (Tables A.5-A.7).
- Our interpretation of each participant's responses (Section A.2).

A.1 Questionnaire

We briefly recap the four types of questions and the concepts they intend to assess. We then provide visualizations that were displayed together with the questionnaire (Figures A.3 to A.2, complementing Figures 3.1 to 3.2 p.42).

Acceptance - Questions Q6-A-i to -iv:

- Question Q6-A-i evaluates the overall acceptance of the computer vision software.
- Question Q6-A-ii evaluates the acceptance of computer vision uncertainty compared to existing techniques in marine biology.
- Question Q6-A-iii evaluates the acceptance of computer vision uncertainty for scientific research in particular.
- Question Q6-A-iv evaluates the *personal attachment* to the computer vision software.

Trust - Questions Q1 and Q6-T-i to -iii:

- Question Q1 evaluates the overall trust in the computer vision results. It presented an example of fish counts provided by a fish detection software (Figure A.3) and asked how the trends observed in the fish counts are likely to be representative of real trends. The data used was artificially generated, and trends were simulated with various intensity (e.g., important increase or decrease, to stagnating fish counts), but presented as genuine to users. The same trends were presented across the steps of the experiment, so as to measure the impact that the additional information introduced at each step had on user trust.
- Question Q6-T-i evaluates the *perceived technical competence* of the fish detection software, and is adapted from (Madsen and Gregor 2000).
- Question Q6-T-ii evaluates the *perceived technical competence* of the method used for measuring fish detection errors.
- Question Q6-T-iii evaluates trust in the data produced by the video analysis software.

Actual Understanding - Questions Q2 to Q4:

- Question Q2 evaluates the effective user understanding of the technical concepts presented at each explanation step.
- Question Q3 evaluates user understanding of the scope of uncertainty issues involved in the computer vision system.
- Question Q4 evaluates if user understanding through the practical use of the technical information. It asked to compare two software and identify the one yielding the fewest errors (Figure A.1-A.2). This question was omitted at Step 3 because informal feedback collected prior to the experiment indicated that our visualization of classification errors for different thresholds were hard to understand. Furthermore the concepts explained at the last step were likely to overwhelm non-expert users. Thus we decided to avoid measuring user understanding of complex concepts using our poor data visualization support.

Perceived Understanding - Questions Q6-U-i to -iv:

- Question Q6-U-i evaluates if users think they understand the technical information provided in the tutorials.
- Question Q6-U-ii evaluates if users think they fully understand the video analysis processes, beyond the groundtruth evaluation process presented in the tutorials.
- Question Q6-U-iii evaluates if users think they understand the implications of using uncertain video analysis data containing classification errors for scientific purposes.
- Question Q6-U-iv evaluates if users think they understand how to handle the classification errors when performing scientific research based on the uncertain video analysis data (e.g., by applying statistical methods).

Information Needs - Questions Q3, Q5, and Q6-I-i to -v:

- Question Q3 investigate information needs regarding uncertainty issues beyond the technical concepts discussed in the explanations.
- Question Q5 investigates marine ecologists' need for estimating the number of

errors in classification end-results (i.e., using the methods introduced in Chapter 5).

- Question Q6-I-i evaluates if the explanations fulfil user information needs on the video analysis errors.
- Question Q6-I-ii evaluates if the user thinks the explanations generally fulfils the information needs on the video analysis method, while question Q6-I-iii evaluates if the user needs more information for her/his particular interests.
- Question Q6-I-iv evaluates if the information was easy to understand. If the information is difficult to understand, users may need more details in the explanations, and a different formulation of the information.
- Question Q6-I-v evaluates if the information is relevant, i.e., if some information is too detailed or superfluous, and does not address real user needs.

Figure A.1: The software to compare in question Q4 of Step 1 (Table 3.1).

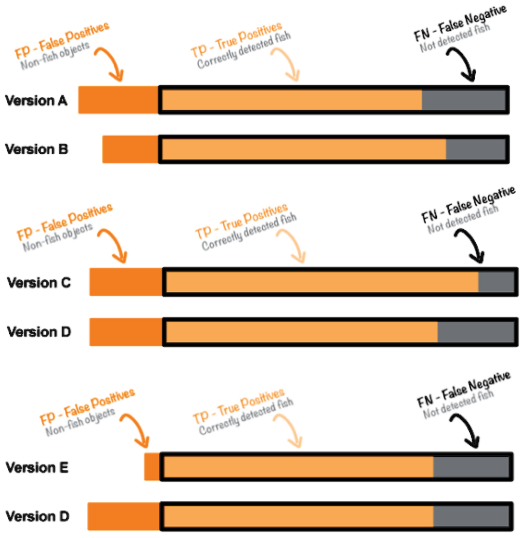

Figure A.2: The software to compare in question Q4 of Step 2 (Table 3.1).

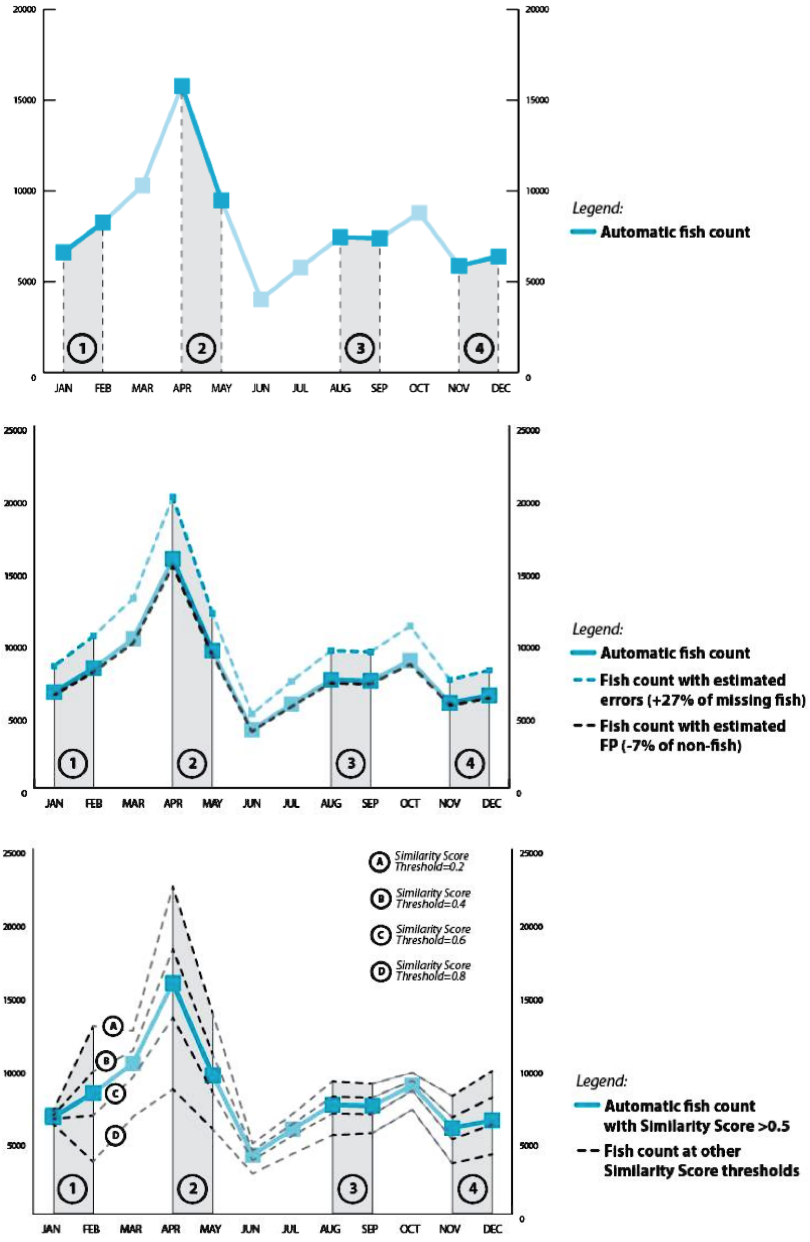

Figure A.3: The trends to assess in question Q1 (Table 3.1), at Step 1 to 3 (top to bottom).

			Participants														
Question	**Construct**	**Scale**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**	**P15**
Q1-i	TRUST	-2 to 2	1	1	0	0	0	-1	0	1	1	1	0	2	1	1	0
Q1-ii	TRUST	-2 to 2	1	-1	0	0	-1	-1	0	1	2	-1	0	2	1	1	0
Q1-iii	TRUST	-2 to 2	1	-1	0	0	0	1	0	1	0	2	0	0	1	1	0
Q1-iv	TRUST	-2 to 2	1	1	0	0	0	1	0	0	0	2	0	0	1	1	0
Q2-i	UNDERST.-act.	0 or 1	1	0	0	0	1	0	0	0	0	1	1	1	0	0	1
Q2-ii	UNDERST.-act.	0 or 1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Q2-iii	UNDERST.-act.	0 or 1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1
Q3-U-i	UNDERST.-act.	Y or N	Y	Y	Y	N	Y	Y	Y	N	N	Y	Y	N	Y	Y	Y
Q3-I-i	INFO NEED	Y or N	N	N	Y	Y	N	Y	Y	N	N	N	N	Y	N	N	Y
Q3-U-ii	UNDERST.-act.	Y or N	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y	N	Y	Y	Y
Q3-I-ii	INFO NEED	Y or N	N	N	Y	N	N	N	Y	N	N	N	N	N	N	N	N
Q3-U-iii	UNDERST.-act.	Y or N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y
Q3-I-iii	INFO NEED	Y or N	N	N	N	N	N	N	Y	N	N	Y	N	N	N	N	N
Q3-U-iv	UNDERST.-act.	Y or N	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y	N	Y	Y	Y
Q3-I-iv	INFO NEED	Y or N	N	N	Y	N	N	Y	Y	N	N	Y	N	Y	N	N	Y
Q3-U-v	UNDERST.-act.	Y or N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y
Q3-i-v	INFO NEED	Y or N	N	Y	N	Y	N	Y	Y	N	N	Y	Y	N	N	N	N
Q3-U-vi	UNDERST.-act.	Y or N	Y	Y	Y	Y	Y	N	Y	Y	N	Y	N	Y	Y	Y	Y
Q3-I-vi	INFO NEED	Y or N	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N
Q3-U-vii	UNDERST.-act.	Y or N	N	Y	Y	Y	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y
Q3-I-vii	INFO NEED	Y or N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Q4-i	UNDERST.-act.	0 or 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Q4-ii	UNDERST.-act.	0 or 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Q5	INFO NEED	Y, N, ?	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	?	Y	Y	?	?
Q6-A-i	ACCEPT.	-3 to 3	2	3	2	0	2	0	2	2	2	1	0	1	0	1	2
Q6-A-ii	ACCEPT.	-3 to 3	2	2	-2	1	0	-1	-2	-1	0	-1	-2	-2	-2	1	0
Q6-A-iii	ACCEPT.	-3 to 3	2	3	2	-2	1	-1	2	2	1	-2	-1	1	0	1	0
Q6-A-iv	ACCEPT.	-3 to 3	2	3	2	-2	0	-1	2	2	2	2	-2	0	2	2	2
Q6-T-i	TRUST	-3 to 3	2	2	0	0	0	0	1	2	0	0	1	-1	1	2	0
Q6-T-ii	TRUST	-3 to 3	2	2	0	0	-1	1	0	1	0	1	0	1	0	1	0
Q6-T-iii	TRUST	-3 to 3	2	2	1	1	0	-1	2	1	1	-1	-1	1	1	1	0
Q6-U-i	UNDERST.-perc.	-3 to 3	1	2	2	-2	2	-1	1	-2	1	2	-2	-1	2	2	2
Q6-U-ii	UNDERST.-perc.	-3 to 3	1	2	0	-2	-2	-1	1	-2	-2	-1	-1	-1	1	-2	-3
Q6-U-iii	UNDERST.-perc.	-3 to 3	1	3	2	-1	1	1	0	-2	1	-2	1	1	2	1	2
Q6-U-iv	UNDERST.-perc.	-3 to 3	1	2	0	-1	-1	1	0	-2	1	-2	1	-1	-1	-2	-3
Q6-I-i	INFO NEED	-3 to 3	2	3	2	0	-1	1	2	2	2	-1	1	1	-1	-1	-2
Q6-I-ii	INFO NEED	-3 to 3	2	1	-2	-2	-2	1	0	1	-2	-2	1	1	2	-2	-3
Q6-I-iii	INFO NEED	-3 to 3	-2	-3	-2	-3	-2	-1	2	-2	-2	-2	-2	0	-2	-3	-3
Q6-I-iv	INFO NEED	-3 to 3	1	-3	1	-2	0	1	1	-1	0	-1	-1	2	1	-2	-2
Q6-I-v	INFO NEED	-3 to 3	2	3	2	2	2	0	0	2	2	1	-2	1	2	2	3

Table A.1: Answers to multiple choice questions at Step 1. *Question Q2-ii was discarded because text feedback showed that users often misunderstood the term "manual fish count" as counts from diving observations, instead of counts from manual image labelling. Question Q5 evaluated the need for information on the potential errors to expect in the classification results (Y when needed, N when not needed, ? when user was not sure).*

			Participants														
Question	**Construct**	**Scale**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**	**P15**
Q1-i	TRUST	-2 to 2	1	1	0	-2	0	1	0	1	1	1	0	2	1	1	0
Q1-ii	TRUST	-2 to 2	1	-2	0	-2	0	-1	0	1	2	-1	0	2	1	1	0
Q1-iii	TRUST	-2 to 2	1	-1	0	-2	0	1	0	1	0	1	0	1	1	1	0
Q1-iv	TRUST	-2 to 2	1	1	0	-2	0	1	0	0	0	1	0	1	1	1	0
Q2-i	UNDERST.-act.	0 or 1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1
Q2-ii	UNDERST.-act.	0 or 1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Q2-iii	UNDERST.-act.	0 or 1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1
Q2-iv	UNDERST.-act.	0 or 1	0	0	1	0	1	0	0	1	1	0	0	0	1	1	1
Q2-v	UNDERST.-act.	0 or 1	0	0	0	1	0	1	0	1	0	0	0	1	0	1	1
Q2-vi	UNDERST.-act.	0 or 1	0	0	0	0	1	1	1	1	0	1	0	1	0	1	1
Q4-i	UNDERST.-act.	0 or 1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1
Q4-ii	UNDERST.-act.	0 or 1	1	0	1	0	1	1	1	0	1	0	0	1	1	1	1
Q4-iii	UNDERST.-act.	0 or 1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1
Q5	INFO NEED	Y, N, ?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	?	Y	Y	Y	Y
Q6-A-i	ACCEPT.	-3 to 3	2	3	2	0	2	1	2	2	1	1	0	2	0	1	2
Q6-A-ii	ACCEPT.	-3 to 3	2	2	-2	0	1	-1	-2	0	-2	-1	-3	-1	-1	0	-2
Q6-A-iii	ACCEPT.	-3 to 3	2	2	2	1	1	0	2	2	2	-1	0	2	1	1	0
Q6-A-iv	ACCEPT.	-3 to 3	2	3	2	1	0	0	2	1	2	0	-1	-1	1	2	2
Q6-T-i	TRUST	-3 to 3	2	3	2	0	0	-1	2	2	-2	-1	0	1	1	0	0
Q6-T-ii	TRUST	-3 to 3	2	3	2	0	-1	1	1	1	1	-2	-1	1	0	1	-2
Q6-T-iii	TRUST	-3 to 3	2	2	1	0	1	1	1	0	-1	0	-2	1	1	1	1
Q6-U-i	UNDERST.-perc.	-3 to 3	1	3	2	-1	2	0	1	-2	1	-1	-2	1	1	1	2
Q6-U-ii	UNDERST.-perc.	-3 to 3	1	3	0	2	-2	0	0	-1	0	-1	-2	1	1	-2	-3
Q6-U-iii	UNDERST.-perc.	-3 to 3	1	2	2	1	1	1	1	-1	1	-2	-1	1	1	-1	2
Q6-U-iv	UNDERST.-perc.	-3 to 3	1	2	1	1	-1	1	1	-1	1	-2	-1	-1	1	-1	2
Q6-I-i	INFO NEED	-3 to 3	2	3	2	0	-1	1	2	1	1	1	-1	1	0	-1	1
Q6-I-ii	INFO NEED	-3 to 3	2	2	0	0	-2	0	1	-1	-1	-2	-1	1	1	-1	-2
Q6-I-iii	INFO NEED	-3 to 3	-2	-3	-1	0	-2	1	2	-2	0	-2	-2	-1	-1	-2	-3
Q6-I-iv	INFO NEED	-3 to 3	1	2	0	-2	-1	0	2	1	1	-1	-1	1	0	-1	-2
Q6-I-v	INFO NEED	-3 to 3	2	3	1	1	2	0	1	2	2	1	0	-1	2	2	2

Table A.2: Answers to multiple choice questions at Step 2.

			Participants														
Question	**Construct**	**Scale**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**	**P15**
Q1-i	TRUST	-2 to 2	1	1	0	1	0	1	0	2	1	1	0	1	0	1	0
Q1-ii	TRUST	-2 to 2	1	-1	0	1	0	-1	0	1	2	1	0	2	1	1	0
Q1-iii	TRUST	-2 to 2	1	1	0	1	0	1	0	0	0	1	0	2	0	1	0
Q1-iv	TRUST	-2 to 2	1	1	0	1	0	1	0	0	0	1	0	2	0	1	0
Q2-i	UNDERST.-act.	0 or 1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0
Q2-ii	UNDERST.-act.	0 or 1	0	0	0	1	0	1	1	1	1	0	0	1	1	0	0
Q2-iii	UNDERST.-act.	0 or 1	0	0	1	0	1	0	0	1	0	1	0	0	1	1	0
Q2-iv	UNDERST.-act.	0 or 1	0	1	1	0	0	1	1	1	1	0	0	1	1	1	1
Q2-v	UNDERST.-act.	0 or 1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Q2-vi	UNDERST.-act.	0 or 1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1
Q6-A-i	ACCEPT.	-3 to 3	2	3	2	2	2	1	2	2	-1	1	-1	2	0	1	2
Q6-A-ii	ACCEPT.	-3 to 3	2	3	-1	1	1	0	-2	1	-2	-1	-2	1	-2	1	-2
Q6-A-iii	ACCEPT.	-3 to 3	2	3	2	1	2	-1	2	2	1	1	-2	2	1	1	1
Q6-A-iv	ACCEPT.	-3 to 3	1	3	2	0	0	0	2	2	2	1	-2	1	1	2	2
Q6-T-i	TRUST	-3 to 3	2	3	3	1	0	0	2	2	-1	1	0	2	1	1	0
Q6-T-ii	TRUST	-3 to 3	2	3	2	1	1	1	1	2	0	0	-1	3	1	-1	1
Q6-T-iii	TRUST	-3 to 3	2	3	1	1	1	0	1	2	1	0	-2	2	1	1	0
Q6-U-i	UNDERST.-perc.	-3 to 3	1	3	2	0	2	-1	0	1	0	-1	-2	1	1	-1	2
Q6-U-ii	UNDERST.-perc.	-3 to 3	1	3	1	0	-1	-1	0	-1	0	-1	-2	1	1	-2	-3
Q6-U-iii	UNDERST.-perc.	-3 to 3	1	3	2	0	2	1	1	2	1	1	-2	2	1	-2	2
Q6-U-iv	UNDERST.-perc.	-3 to 3	1	3	1	0	0	1	1	1	-2	-1	-2	1	1	-1	1
Q6-I-i	INFO NEED	-3 to 3	2	3	2	1	2	1	2	2	-1	1	-1	2	0	-1	1
Q6-I-ii	INFO NEED	-3 to 3	1	3	1	1	-1	1	1	1	-2	-2	-1	1	1	-2	-2
Q6-I-iii	INFO NEED	-3 to 3	-2	-3	-2	-1	-2	0	-1	-1	-1	-2	-2	-1	-1	-2	-3
Q6-I-iv	INFO NEED	-3 to 3	1	3	2	-1	2	0	2	-1	0	-1	-2	1	0	-1	-2
Q6-I-v	INFO NEED	-3 to 3	2	3	2	-1	2	1	1	2	1	1	-2	2	1	2	2

Table A.3: Answers to multiple choice questions at Step 3.

Question - Participant - Answer	
Q1	P5 Without any background information on the coral reef, location etc, it is very unlikely that any statement about how likely it is that this trends may occur in reality. So can a trend like the one showed here be really happening? Yes it may happen. Can I say that this likely to be what is happening there? No I can't without background information on location, species composition, etc.
	P6 A trend such as seen from April-May is possible, but only at certain dramatic circumstances. For instance a severe viral infection or something like that could decimated a population. Otherwise such a severe decrease is not likely.
	P7 Dear sir, madam, I am a terrestrial ecologist and therefore I am not very familiar with marine ecosystems. For this reason I find myself unqualified to give a thrustworthy judgement of the likelyness that the trends in as pictured in the above chart will occur in reality.
	P10 The abundance of fish in a certain area depends on complex biological and physical processes, interactions between behaviour, physiology, and habitat (e.g. depth, seabed) characteristics. In addition I think it is species dependend and therefore I would say it's difficult to argue whether these trends are likely to be observed in reality.
	P11 We need to have the real data or trend in reality. Otherwise, we will not be able to know whether the trend of software count is the same or different from the divers' census count. Is this what you said the "reality"?
	P12 For a small increase/decrease its hard to say if this is really what happened, due to some limitations of the method... like what you actually want to sample in terms of the overall habitat in a certain area, this because you only have a view of 8m for example. Depends on the sampling effort as well
Q2	P4 1.There are many schools of fish appear in the camera's field of view, different experts count will be different, but the machine will not. 2.Appeared several times in the field of view, the expert will not repeat count for the same fish, but the machine will repeat count.
	P5 I think that 3rd reason is the most probable, considering the consistently lower counts by the software. One question arises, are experts counting anything that they can identified as a fish, or only those that can they say what fish it is? If they do they are doing the former, advising them to do the later may take the counts closer together.
	P8 A single fish detected more then once compensates for the fishes that are being missed.
	P9 The first statement would lead to higher fish counts with the software than manually, this contradicts with the background information. I do believe however that this software may count rocks as fish. Then the error between manual and software counts would even be higher than 27% in reality.
	P10 I believe that especialy statement 2 is of importance!
	P11 1. I do not know whether the software is good enough to distinguish the "fish" and "rock". I can only believe the software can. 2. Diver can judge whether the fish swim out and in the camera field is the same or different individual. Certainly different divers may have different results. That is a bias by different observers. 3. some smaller body size fish or cryptic fish may not be detected by software. Some fishes if they swim too far away from the camera and could not be detected by software especially when the water visibility is not good. Divers should be able to see better than camera especially when camera lens has biofouling problem.
	P12 When doing the fish count manually it is more likely that the same fish wasnt been recorded several times. The reason for this is that you re better able to see differences in length and behavior between fish of the same species. You dont have this problem when using stereo cameras and using a relative abundance, so the maximum number seen in one frame.
Q3	P2 benthic fish can be miss count
	P3 video blocked by an object
	P11 If the camera field was changed. We should be able to detected by the monitor at lab. For the last two questions, I can not really answer. Because we should have both data, one from software count and another from divers count, in hand and then make a comparison study to find out is any other source of error.
	P12 The range of view.. especially if you want to compare the videos. For instance, when coral is blocking the view of the cameras. Also, the position of the cameras, because you can miss certain reef associated fish species when the cameras are pointing a bit upwards.
	P13 inter observer differences?
	P14 I don't know how you did the evalutation

Table A.4: Answers to free text questions at Step 1 (Part 1/2).

Question	Participant	Answer
Q4-i	P3	Because it is closer to the expert's count
	P4	Because of interference may not much difference of A and B counting the fish.
	P5	I don't think the difference is good enough. Maybe data on several runs and standard deviation of those runs will help to really see which one is better.
	P7	Version A gives the best estimate of the actual number of fish.
	P8	Version A is more precise, but that does not mean it is more accurate.
	P9	smallest difference
	P10	you need more information on the software... we have to take human as well as computer errors into account.
	P11	If the 5585 is correct count. Then, certainly Version A is better than B. Otherwise, B may be better than A.
	P12	Version A is closer to the number of fish counted by experts
	P13	difference between automatic and observer is smallest
Q4-ii	P3	Experts miight have missed fish too. But I would like to see if it were fish or other objects.
	P4	There will be differences in the analysis, I do not know which software to be believed.
	P5	Again, It doesn't matter what the difference sign is. The important is to see how good are the methods giving consistent counts.
	P7	Version C gives the best estimate of the actual number of fish.
	P8	But... see above
	P9	smallest difference
	P10	the numbers are closer together.
	P11	The same reason as my answer in above question
	P12	Its better to underestimate a certain result for further conclusions..
	P13	difference between automatic and observer is smallest
Q5	P3	it is more realistic. it changes the shape of the graph because it is relative
	P4	I think I will be to understand why we lost 27%, and then determine which data to be used. I am more inclined to choose "automatic fish count", but not absolute.
	P5	Because it gives an overview of the error.
	P7	The dashed corrects for any potential errors. In my opninion the dashed line gives the best estimate of the actual number of fish.
	P9	most real
	P10	it's relevant to know how much errors in the estimates you have, especially if you want to use the data for further analysis!
	P11	First of all, I should know how you estimate the missing fish and whether it is reasonable or not.
	P12	I think its a high percentage what you actual miss when you only focus on one method, so therefore I will include both versions of counting in the analysis
Q6	P4	1.Count the total number is not important, the important is the species, and number of individuals each species. The total number is no meaning in the ecological. 2.Species too much, it is recommended to count the number of dominant species or numbers of resident specie or numbers of semiresident.
	P11	If we do not have any evaluation study for the video analysis in advance. How could I know the video analysis is reliable or not.
	P12	You dont explain how the software is counting the fish. Does it react on movements or what is it?? It is also not fully understood how the sofware reacts on a fish that is in front of coral and well camouflaged for example or just very small fish..

Table A.5: Answers to free text questions at Step 1 (Part 2/2).

Question	Participant	Answer
Q1	P4	The trend is the focus, not the number. The trends of three methods are the same. So the results are the same. There were nothings to be compared.
	P5	The new information don't really solve the doubts expressed before. I don't think it add too much. It is expected by that type of software to have more or less constant errors, meaning that the trends are not going to change with more information as the observed trend is in general proportional to the real one.
	P11	MY answer or comments are still the same as my answer previously.
	P12	Im very convinced about the new line added to the graph with the fish count with estimated non-fish object. Now you include that high percentage what you miss with automatically running the software to identify the fish, i.e small fish.
Q4-i	P3	Higher percentage of TP
	P4	The trend is the focus, not the numbers.
	P5	not enough information. But of course a software that reduces both false negatives and false positives is obviously better.
	P6	'Cause it is the most accurate version and has the highest TP and lowest FP and FN.
	P7	Both the percentage of false positives and false negatives is smaller in version B.
	P9	smallest error
	P10	less FP and less FN = more accurate count...
	P11	We should do some evaluation on the accuracy of video analysis first before I can answer this question.
	P12	Version B has relatively more True Positive in the model, and with a decline of false parameters
	P13	difference with manual count is mallest
Q4-ii	P3	higher percentage of TP. Lower percentage of FN
	P4	Repeat, I take care the trend. Which methods I don't care.
	P5	Not enough information. Although the software able to reduce false negatives without compromising the false positives is better.
	P6	'Cause it is the most accurate version and has the highest TP and lowest FN.
	P7	The percentage of false positives is equal for both versions. However, the percentage of false negatives is smaller for version C.
	P9	smallest error
	P10	you could argue more TP are counted in version C but the overall count of both version are the same... so you can't differentiate between them.
	P11	Same as above
	P12	More TP than FN
	P13	difference with manual count is smallest
Q4-iii	P3	Higher percentage of TP. lower percentage of FP
	P4	Repeat, I take care the trend. Which methods I don't care.
	P5	Again the same. Of course a software that reduces the number of false positives without increasing false negatives is better.
	P6	'Cause it is the most accurate version and has the lowest FP.
	P7	The percentage of false negatives is equal for both versions. However, the percentage of false positives is smaller for version E.
	P9	smallest error
	P10	Less FP detected
	P11	Same as above
	P12	Total amount of counts which is True Positive relative to False Positive is higher. Less error
	P13	difference with manual count is smallest
Q5	P3	It should be +20 % ?
	P4	Repeat, I take care the trend. Which methods I don't care.
	P5	In this case the error seems constant all over the trend. But that may not be the case and I want to know when that happens
	P6	The automatic fish count clearly underestimates (by 27%) the true population, the fish count without estimated non-fish objects is good because it corrects for the non-fish objects but still misses the 27%. The fish count with estimated missing fish overestimates the true population because it doesn't correct for the non-fish objects. So a comparison of all three would be best, although it is the most laborious choice.
	P7	It is always better to have an estimate of possible error-margins
	P9	gives the best evaluation of the real situation
	P10	rather have the non-fish selections removed from my data-set then have more fish in my count of which a certain number are no-fish...
	P11	Same as above
	P12	Im not entirely convinced about estimating the missing fish only from a percentage of all videos that has to be analysed.
Q6	P12	Is it also possible to automatically identify the fish species?

Table A.6: Answers to free text questions at Step 2.

Question - Participant - Answer	
Q1	P11 I do not understand how these similarity were calculated. So I can not answer your questions appropriately.
Q6	P4 All the ways to fix it only in order to get the correct results. But the trend is the focus, not the numbers. Start trusted, it can only do so. If was "the garbage in, garbage out".
	P12 After reading and getting more information about errors that are produced with this method I get a better feeling in how the system works. So it might be that some given answers of the first two tasks or not entirely correct to my understanding

Table A.7: Answers to free text questions at Step 3.

A.2 Interpretation of participant responses

Our interpretation of user responses investigates the impact that the technical information introduced at each step had on user *understanding, trust, acceptance* and fulfilment of *information needs*. In particular, user *trust, acceptance* are compared with the *perceived* and *actual understanding*, so as to identify uninformed *trust* and *acceptance*. We considered both the quantitative measurements from the multiple choice questions, and the qualitative feedback from free-form text questions.

Participant P1 - The explanation steps had almost no impact on of the measurements, which were all relatively high, except *actual understanding* which varied from relatively high to moderate. We assume that trust and acceptance were uninformed.

Participant P2 - *Trust* increased over step from moderate (middle score) to relatively large while *acceptance* remained relatively high. *Actual Understanding* was moderate at Step 1 and 2 decreased to average at Steps 3. The text feedback indicates an accurate understanding, thus we assume trust and acceptance to be well-informed. *Perceived Understanding* was relatively high, and *Information Needs* increased from partly to fairly fulfilled. This is consistent with our assumption that P8 seeks well-informed trust and acceptance.

Participant P3 - The explanation steps had little impact on *trust* and *acceptance*, which remained relatively high and slightly increased at Step 3. *Actual Understanding* remained relatively high, which is consistent with the text feedback. Thus we assume trust and acceptance to be relatively well-informed. *Perceived Understanding* increased over steps from very low to average, and *information needs* remained perceived as partly fulfilled. This is consistent with our interpretation that P2 seeks well-informed trust and acceptance.

Participant P4 - The explanation steps had a significant impact on *trust* and *acceptance Trust* evolved from neutral (Step 1) to very low (Step 2) and relatively high (Step 3), while *acceptance* increased from relatively low to relatively high. *Actual Understanding* was moderate at Step 1, although the text feedback indicates an excellent understanding. It significantly lowered at Step 2 and 3, as the text feedback indicates a lost of interest for the materials. *Perceived Understanding*, however, increased after Step 1 as P4 understood crucial aspects of uncertainty: 1) the classification errors can impact the trends observed in the data; and 2) the variability of error rates can impact the extrapolation of classification errors in the data (Section 3.5.1). *The Information Needs* increased from largely to partly unfulfilled, which indicates that P4 seeks well-informed trust. The text and oral feedback also indicated that P4 seeks well-informed

trust, that the Information Needs on other uncertainty factors are largely unfulfilled, and that acceptance increased as P4 was willing to conduct experiments with the promising system (e.g., to assess the uncertainty issues).

Participant P5 - The explanation steps had little impact on *trust* and *acceptance*, although they slightly increased at each step. *Trust* remained moderate, and *acceptance* relatively high. *Actual Understanding* was very high at Step 1 and 2, with text feedback indicating an excellent understanding. Hence we assume that trust and acceptance were well-informed, although actual understanding was relatively low at Step 3. *Perceived Understanding* was moderate and *information needs* partly unfulfilled, although their scores improved at Step 3. This is consistent with our assumption that P5 seeks well-informed trust and acceptance.

Participant P6 - The explanation steps had little impact on *trust* and *acceptance*, which remained relatively low and slightly increased at Step 2. *Actual Understanding* was very high, but decreased at Step 3. Hence we assume that trust and acceptance are well-informed. *Perceived Understanding* remained moderate, and *information needs* partly fulfilled. This is consistent with our assumption that P6 seeks well-informed trust and acceptance.

Participant P7 - The explanation steps had little impact on *trust* and *acceptance*, which remained relatively neutral, i.e., close to the middle score (average of minimum and maximum score). *Actual Understanding* was relatively high and decreased at Step 3, which is consistent with the text feedback. *Perceived Understanding* remained neutral, close to the middle score, and *information needs* were perceived as fairly fulfilled. Hence we assume that P7 seeks well-informed trust and acceptance.

Participant P8 - *Trust* was relatively high and slightly increased at each step. *Acceptance* remained very high, close to the maximum score. *Actual Understanding* decreased from high (Step 1) to low (Steps 2 and 3). Thus we assume trust and acceptance to be uninformed. *Perceived Understanding* was very high, increasing to reach the highest possible score, and *the Information Needs* increased from partly to fairly fulfilled. This is consistent with our assumption that trust and acceptance are uninformed.

Participant P9 - *Trust* remained moderate while *acceptance* decreased at each step, from relatively high to moderate. *Actual Understanding* was moderate to relatively high, and the text feedback indicates an accurate understanding. Hence we assume that trust and acceptance are well-informed, although understanding decreased at Step 3. *Perceived Understanding* and *information needs* increased and decreased together with *actual understanding*. This is consistent with our assumption that P9 seeks well-informed trust and acceptance.

Participant P10 - The explanation steps had little impact on *trust* and *acceptance*, which remained moderate although trust was lower at Step 2. *Actual Understanding* decreased at each step from the maximum to the average score, although the text feedback indicates an excellent understanding. Hence we assume that trust and acceptance are well-informed. *Perceived Understanding* remained relatively low, and *information needs* partly unfulfilled. This is consistent with our assumption that P10 seeks well-informed trust and acceptance.

Participant P11 - The explanation steps had little impact on *trust* and *acceptance*, although both showed a small decrease. *Trust* remained moderate to low, while *acceptance* remained very low. *Actual Understanding* was very high at Step 1, but low at Step 2 and 3 although the text feedback indicates an excellent understanding at all steps. *Perceived Understanding* decreased from relatively low to very low, and *information needs* were perceived as largely unfulfilled. Hence we assume that P6 seeks well-informed trust and acceptance.

Participant P12 - *Trust* and *acceptance* increased at each step, from moderate to relatively high. *Perceived Understanding* increased too, from relatively low to relative high. However, *actual understanding* decreased from the maximum score to a relatively high score, although the text feedback indicates an excellent understanding at all steps. Hence we assume that trust and acceptance are well-informed, which is consistent with the *information needs* being partly fulfilled.

Participant P13 - The explanation steps had almost no impact on any of the measurements. *Trust*, *actual understanding* and *perceived understanding* were relatively high, while *acceptance* was moderate and *information needs* partly fulfilled. We assume that trust and acceptance were well-informed.

Participant P14 - The information introduced at each step had little impact on *trust* and *acceptance*, which remained relatively high although although *trust* slightly decreased at each step. *Actual Understanding* was low at Step 1, which is consistent with the text feedback, good at Step 2, and low at Step 3. The text and multiple choice questions show that the key concepts of False Positives and False Negatives remained misunderstood at all steps. Thus we assume that trust and acceptance are uninformed. This interpretation is consistent with the *perceived understanding*, which was low and decreased over steps. *Information needs* remained perceived as largely unfulfilled (i.e., low score), which is consistent with the low user understanding and the uninformed trust and acceptance.

Participant P15 - The explanation steps had little impact on *trust* and *acceptance*. *Trust* remained relatively low, while *acceptance* remained relatively neutral, i.e., close to the middle score. *Actual Understanding* was very high at Step 1 and 2, but relatively low at Step 3. *Perceived Understanding* remained neutral, close to the middle score, and *information needs* were perceived as largely unfulfilled. Hence we assume that P5 seeks well-informed trust and acceptance.

Bibliography

Alsallakh, B., Hanbury, A., Hauser, H., Miksch, S., and Rauber, A. (2014). Visual methods for analyzing probabilistic classification datasets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12). (cited pp. 118, 121, 123, and 124)

Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *Symposium on Information Visualization (Infovis)*, pages 111–117. IEEE. (cited pp. 145, 146, and 148)

Artz, D. and Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71. (cited p. 45)

Beauxis-Aussalet, E. and Hardman, L. (2015). Multifactorial uncertainty assessment for monitoring population abundance using computer vision. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. (cited pp. 78, 79, and 105)

Beauxis-Aussalet, E. and Hardman, L. (2016). *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*, chapter Appendix 1 - User Interface and Usage Scenario. Spinger. (cited p. 160)

Beauxis-Aussalet, E., Palazzo, S., Nadarajan, G., Arlasnova, E., Spampinato, C., and Hardman, L. (2013). A video processing and data retrieval framework for fish population monitoring. In *ACM MultiMedia workshop on Multimedia Analysis for Ecological Data (MAED)*. (cited p. 59)

Bohrnstedt, G. and Goldberger, A. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association*, 64(328):1439–1442. (cited p. 110)

Boom, B. J., Beauxis-Aussalet, E., Hardman, L., and Fisher, R. B. (2016). Uncertainty-aware estimation of population abundance using machine learning. *Multimedia System Journal*. (cited pp. 106, 111, 112, 113, and 117)

Boom, B. J., Huang, P. X., Beyan, C., Spampinato, C., Palazzo, S., He, J., Beauxis-Aussalet, E., Lin, S.-I., Chou, H.-M., Nadarajan, G., et al. (2012). Long-term underwater camera surveillance for monitoring and analysis of fish populations. In

Workshop on Visual observation and analysis of Animal and Insect Behaviour (VAIB), held at the 21st International Conference on Pattern Recognition (ICPR). (cited p. 2)

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability evaluation in industry*, 189(194). (cited p. 128)

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods and Applications*. CRC Press, Taylor and Francis. (cited pp. 76, 77, 78, 79, 91, 92, 94, and 111)

Cappo, M., Speare, P., and De'ath, G. (2004). Comparison of baited remote underwater video stations (bruvs) and prawn (shrimp) trawls for assessments of fish biodiversity in inter-reefal areas of the great barrier reef marine park. *Journal of Experimental Marine Biology and Ecology*, 302(2):123–152. (cited pp. 27 and 29)

Card, D. H. (1982). Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, 48:431–439. (cited p. 76)

Cleveland, W. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387). (cited pp. 121, 123, 124, and 137)

Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons. (cited pp. 27 and 86)

Correa, C., Chan, Y.-H., and Ma, K.-L. (2009). A framework for uncertainty-aware visual analytics. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 51–58. (cited pp. 58 and 144)

Correll, M. and Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*. (cited p. 140)

Csurka, G., Zeller, C., Zhang, Z., and Faugeras, O. D. (1997). Characterizing the uncertainty of the fundamental matrix. *Computer vision and image understanding*, 68(1):18–36. (cited p. 61)

Drummond, C. and Holte, R. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1). (cited pp. 118 and 119)

Elias, M. and Bezerianos, A. (2011). Exploration views: understanding dashboard creation and customization for visualization novices. In *Human-Computer Interaction–INTERACT*, pages 274–291. Springer. (cited pp. 144, 145, and 148)

Elzen, S. v. d. and Wijk, J. J. v. (2011). Baobabview: Interactive construction and analysis of decision trees. In *IEEE Visual Analytics Science and Technology (VAST)*. (cited p. 118)

Endsley, M. R. (1988a). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 97–101. SAGE Publications. (cited p. 145)

Endsley, M. R. (1988b). Situation awareness global assessment technique (SAGAT). In *Proceedings of the National Aerospace and Electronics Conference (NAECON)*, pages 789–795. IEEE. (cited p. 164)

Endsley, M. R. (1995). Towards a theory of situation awareness in dynamic systems. *Human factors*, 37. (cited p. 127)

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8). (cited pp. 118 and 121)

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B*. (cited p. 110)

Fisher, R. B., Chen-Burger, Y.-H., Giordano, D., Hardman, L., and Lin, F.-P., editors (2016). *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*. Spinger. (cited p. 59)

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1):185–201. (cited p. 76)

Gill, T. and Hicks, R. (2006). Task complexity and informing science: A synthesis. *Information Science Journal*. (cited p. 132)

Grammel, L., Tory, M., and Storey, M. (2010). How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):943–952. (cited p. 145)

Grassia, A. and Sundberg, R. (1982). Statistical precision in the calibration and use of sorting machines and other classifiers. *Technometrics*, 24(2):117–121. (cited pp. 76, 77, and 79)

Griethe, H. and Schumann, H. (2006). The visualization of uncertain data: Methods and problems. In *SimVis*, pages 143–156. (cited pp. 144 and 145)

Harvey, E., Fletcher, D., and Shortis, M. (2001). A comparison of the precision and accuracy of estimates of reef-fish lengths determined visually by divers with estimates produced by a stereo-video system. *Fisheries Bulletin*, 99:63–71. (cited pp. 26 and 29)

Hay, A. M. (1988). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9(8). (cited pp. 76 and 77)

Hay, A. M. (1989). Global estimates from a confusion matrix, a reply to jupp. *International Journal of Remote Sensing*, 10(9). (cited p. 78)

He, J., van Ossenbruggen, J., and de Vries, A. P. (2013). Do you need experts in the crowd?: a case study in image annotation for marine biology. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 57–60. (cited p. 62)

Heer, J., van Ham, F., Carpendale, S., Weaver, C., and Isenberg, P. (2008). Creation and collaboration: Engaging new audiences for information visualization. In *Information Visualization*, pages 92–133. Springer. (cited pp. 144 and 148)

Hetrick, N. J., Simms, K. M., Plumb, M. P., and Larson, J. P. (2004). *Feasibility of using video technology to estimate salmon escapement in the Ongivinuk River, a clear-water tributary of the Togiak River*. US Fish and Wildlife Service, King Salmon Fish and Wildlife Field Office. (cited p. 27)

Hoffrage, U., Krauss, S., Martignon, L., and Gigerenzer, G. (2015). Natural frequencies improve bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology*. (cited p. 136)

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining and Knowledge Management Process (IJDKP)*, 5(2). (cited pp. 119 and 121)

Huang, W., Eades, P., and Hong, S. (2009). Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*. (cited pp. 125, 127, and 132)

Jousselme, A.-L., Maupin, P., and Bossé, É. (2003). Uncertainty in a situation analysis perspective. In *Proceedings of the Sixth International Conference of Information Fusion*. IEEE. (cited p. 146)

Katila, M. (2006). *Forest Inverntory: Methodology and Applications*, chapter Correcting map errors in forest inventory estimates for small areas, pages 225–233. Number 13. Springer. (cited pp. 77 and 78)

Khan, A., Breslav, S., Glueck, M., and Hornbæk, K. (2015). Benefits of visualization in the mammography problem. *Int. J. Human-Computer Studies*. (cited pp. 119 and 136)

Kosinski, A. (2001). Cramer's rule is due to cramer. *Mathematics Magazine*, 74:310–312. (cited pp. 82, 94, and 95)

Kraan, M., Uhlmann, S., Steenbergen, J., Van Helmond, A., and Van Hoof, L. (2013). The optimal process of self-sampling in fisheries: lessons learned in the netherlands. *Journal of fish biology*, 83(4):963–973. (cited p. 26)

Krause, J., Dasgupta, A., Swartz, J., Aphinyanaphongs, Y., and Bertini, E. (2017). A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *IEEE Confrence on Visual Analytics Science and Technology*. (cited p. 118)

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9). (cited p. 125)

Langlois, T., Chabanet, P., Pelletier, D., and Harvey, E. (2006). Baited underwater video for assessing reef fish populations in marine reserves. *Fisheries Newsletter - South Pacific Commission*, 118. (cited p. 27)

Levina, E. and Bickel, P. (2001). The earth mover? s distance is the mallows distance: Some insights from statistics. In *null*, page 251. IEEE. (cited p. 103)

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons. (cited p. 72)

Liu, S., Wang, X., Liu, M., and Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*. (cited p. 118)

Lowry, M., Folpp, H., Gregson, M., and Suthers, I. (2012). Comparison of baited remote underwater video (bruv) and underwater visual census (uvc) for assessment of artificial reefs in estuaries. *Journal of Experimental Marine Biology and Ecology*, 416–417:243–253. (cited pp. 27 and 29)

Madsen, M. and Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of Eleventh Australasian Conference on Information Systems*, pages 6–8. Citeseer. (cited pp. 45 and 188)

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, pages 24–59. (cited p. 45)

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1):12–16. (cited p. 172)

McGuinness, B. (2004). Quantitative analysis of situational awareness (QUASA): Applying signal detection theory to true/false probes and self-ratings. DTIC Document. (cited p. 146)

Micallef, L., Dragicevic, P., and Fekete, J.-D. (2012). Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*. (cited p. 119)

Murch, G. M. (1984). Physiological principles for the effective use of color. *IEEE Computer Graphics and Applications*. (cited p. 124)

Pang, A. T., Wittenbrink, C. M., and Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390. (cited pp. 33, 58, and 144)

Ren, D., Amershi, S., Lee, B., Suh, J., and Williams, J. (2017). Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics*, 23(1). (cited pp. 118, 121, 123, and 124)

Saerens, M., Latinne, P., and Decaestecker, C. (2001). Adjusting the output of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14:21–44. (cited pp. 102 and 106)

Sebastiani, F. (2015). An axiomatically derived measure for the evaluation of classification algorithms. In *International Conference on The Theory of Information Retrieval*. (cited p. 119)

Senge, R., Del Coz, J. J., and Hüllermeier, E. (2014). On the problem of error propagation in classifier chains for multi-label classification. In *Data Analysis, Machine Learning and Knowledge Discovery*, pages 163–170. Springer. (cited p. 61)

Shafait, F., Mian, A., Shortis, M., Ghanem, B., Culverhouse, P., Edgington, D., Cline, D., Ravanbakhsh, M., Seager, J., and Harvey, E. (2016). Fish identification from videos captured in uncontrolled underwater environments. *ICES Journal of Marine Science*, 73(10):2737–2746. (cited p. 27)

Shieh, M. S. (2009). *Correction methods, approximate biases, and inference for misclassified data*. PhD thesis, Univ. of Massachusetts. (cited pp. 76, 77, 78, 79, 89, 91, 92, 100, and 111)

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Symposium on Visual Languages*, pages 336–343. IEEE. (cited pp. 145 and 146)

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4). (cited p. 119)

Spampinato, C., Beauxis-Aussalet, E., Palazzo, S., Beyan, C., van Ossenbruggen, J., He, J., Boom, B., and Huang, X. (2014). A rule-based event detection system for real-life underwater domain. *Machine vision and applications*, 25(1):99–117. (cited p. 31)

Spampinato, C., Palazzo, S., Giordano, D., Kavasidis, I., Lin, F.-P., and Lin, Y.-T. (2012). Covariance based fish tracking in real-life underwater environment. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 409–414. (cited pp. 61, 62, and 67)

Talbot, J., Setlur, V., and Anand, A. (2014). Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*. (cited pp. 121 and 124)

Tang, D., Stolte, C., and Bosch, R. (2004). Design choices when architecting visualizations. *Information Visualization*, 3(2):65–79. (cited p. 144)

Taylor, M., Baker, J., and Suthers, I. (2013). Tidal currents, sampling effort and baited remote underwater video (bruv) surveys: Are we drawing the right conclusions? *Fisheries Research*, 140(96–104). (cited p. 28)

Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics*, 14(1):187–202. (cited pp. 76, 77, and 79)

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., and Pavel, M. (2005). A typology for visualizing uncertainty. In *Visualization and Data Analysis (VDA)*, pages 146–157. (cited p. 144)

Tidwell, J. (2010). *Designing interfaces: Patterns for effective interaction design*. O'Reilly Media, Inc. (cited p. 124)

Trevor, J., Russell, B., and Russell, C. (2000). Detection of spatial variability in relative density of fishes: comparison of visual census, angling, and baited underwater video. *Marine Ecology Progress Series*, 198:249–260. (cited p. 29)

van der Aalst, W. M. P., Blichler, M., and Heinzl, A. (2017). Responsible data science. *Business and Information System Engineering*, 59(5):311–313. (cited pp. 1 and 36)

van Deusen, P. C. (1996). Unbiased estimates of class proportions from thematic maps. *Photogrammetric Engineering and Remote Sensing*, 62(4):409–412. (cited p. 76)

Viegas, F. B., Wattenberg, M., Van Ham, F., Kriss, J., and McKeon, M. (2007). Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128. (cited p. 149)

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B., Janssen, P., and Krayer von Krauss, M. P. (2003). Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated assessment*, 4(1):5–17. (cited p. 58)

Wang Baldonado, M. Q., Woodruff, A., and Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 110–119. ACM. (cited pp. 144, 145, 146, 148, and 149)

Wickens, C. D. and Carswell, C. M. (1997). Information processing. *Handbook of human factors and ergonomics*, pages 89–122. (cited p. 145)

Zhu, X. and Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210. (cited p. 61)

List of SIKS dissertations published since 2011

2011

01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
09 Tim de Jong (OU), Contextualised Mobile Media for Learning
10 Bart Bogaert (UvT), Cloud Content Contention
11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM), Strategic Decision-Making in complex games
19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach
21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT), Robust Brain-Computer Interfaces
45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013

01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing

03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36 Than Lam Hoang (TUe), Pattern Mining in Data Streams
37 Dirk Börner (OUN), Ambient Learning Displays
38 Eelco den Heijer (VU), Autonomous Evolutionary Art
39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

2014

01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
06 Damian Tamburri (VU), Supporting Networked Software Development
07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior
08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10 Ivan Salvador Razo Zapata (VU), Service Value Networks
11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

14 Yangyang Shi (TUD), Language Models With Meta-information
15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015

01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners

20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed colomn stores
28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31 Yakup Koç (TUD), On the robustness of Power Grids
32 Jerome Gard (UL), Corporate Venture Management in SMEs
33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34 Victor de Graaf (UT), Gesocial Recommender Systems
35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016

01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018

01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TUE), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TUE), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TUE), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis

2019

01 Rob van Eijk (UL), Web Privacy Measurement in Real-Time Bidding Systems. A Graph-Based Approach to RTB system classification.

Summary

Handling classification uncertainty is a crucial challenge for supporting efficient and ethical classification systems. This thesis addresses uncertainty issues from the perspective of end-users with limited expertise in machine learning. We focus on uncertainties that pertain to estimating class sizes, i.e., numbers of objects per class. We aim at enabling non-expert end-users to conduct uncertainty-aware and scientifically-valid analysis of class sizes.

We research the means to support end-users' understanding of class size uncertainty. After investigating the specific use case of in-situ video monitoring of animal populations, where classes represent animal species, we derive generalizable methods for:

- Assessing the uncertainty factors and the uncertainty propagation that result in high-level errors and biases in class size estimates.
- Estimating the magnitude of classification errors in class size estimates.
- Visualizing classification uncertainty when evaluating classification systems, and interpreting class size estimates.

We first study the high-level information needs that can or cannot be addressed by computer vision techniques for monitoring animal populations. The uncertainty issues inherent to each data collection technique, and high-level requirements for uncertainty assessment are identified. We further investigate the information that support end-users in developing informed uncertainty assessments. We explore how information about classification errors impacts users' understanding, trust and acceptance of the computer vision system. We highlight unfulfilled information needs requiring additional uncertainty assessments, and high-level user-oriented information that uncertainty assessments must provide.

From these insights, we identify key uncertainty factors to address for enabling scientifically valid analyses of classification results. Our scope includes uncertainty factors beyond the computer vision system, arising from the conditions in which the system is deployed. We identify the interactions between uncertainty factors, how uncertainties propagates to high-level information, and the uncertainty assessment methods that are applicable or missing.

We further investigate uncertainty assessment methods for estimating the numbers of errors in classification end-results, using error measurements performed with test sets. Class sizes can be corrected to account for the potential False Positives and False Negatives in each class. We identify existing methods from statistics and epidemiology, and highlight the unaddressed case of disjoint test and target sets, which impacts the variance of the error estimation results. We introduce 3 new methods:

- The Sample-to-Sample method estimates the variance of error estimation results for disjoint test and target sets.
- The Maximum Determinant method uses the determinant of error rate matrices as a predictor of the variance of error estimation results.
- The Ratio-to-TP method uses atypical error rates that have properties of interest for predicting the variance of error estimation results.

We then focus on the means to communicate uncertainty to end-users with limited expertise in machine learning. We introduce a simplified design for visualizing classification errors. Our design uses raw numbers of errors as a basic yet complete metric, and simple barcharts where several visual features distinguish the actual and assigned classes. We present a user study that compares our simplified visualization to well-established visualizations. We identify the main difficulties that users encountered with the visualizations and with understanding classification errors.

Finally, we introduce a visualization tool that enables end-users to explore class size estimate, and the uncertainties in specific subsets of the data. We present a user study that investigates how the interface design supports user awareness of uncertainty. We highlight the factors that facilitated or complicated the exploration of the data and its uncertainties.

Our research contributes to enabling the scientific study of animal populations based on computer vision. Our results contribute to a broader range of applications dealing with uncertain computer vision and classification data. They inform the design of comprehensive uncertainty assessment methods and tools.

Samenvatting

Het omgaan met onzekerheid in classificatietaken is een cruciale uitdaging voor het ondersteunen van efficiënte en ethisch verantwoorde classificatiesystemen. Dit proefschrift behandelt onzekerheidsvraagstukken vanuit het perspectief van eindgebruikers die beperkte expertise hebben op het gebied van machine learning en systemen voor beeldherkenning. We concentreren ons hierbij op onzekerheid ten aanzien van het schatten van klassengroottes, oftewel het aantal gevallen per klasse. Het doel is om eindgebruikers zonder expertkennis in staat te stellen academisch verantwoorde data analyses m.b.t. klassengroottes uit te voeren, en hen hierbij rekening te laten houden met de onzekerheid die hierbij komt kijken.

We onderzoeken de middelen die eindgebruikers inzicht moeten bieden in de onzekerheid t.a.v. klassengroottes. Na het bestuderen van in-situ videomonitoring van dierpopulaties hebben we algemeen toepasbare methoden ontwikkeld voor:

- Het beoordelen van de onzekerheidsfactoren en de daarmee gepaard gaande fouten en onzuiverheden bij het schatten van de klassengroottes.
- Het visualiseren van de classificatieonzekerheid bij het beoordelen van classificatiesystemen en het interpreteren van schattingen van klassengroottes.
- Het bepalen van de omvang van classificatiefouten bij het schatten van klassengroottes.

Onze gebruikersstudies vormen een belangrijke basis door het analyseren van de informatiebehoeften van de gebruikers (Hoofdstuk 2), waar aanvullend ook specifieke aandacht uitgaat naar de behoeften ten aanzien van de onzekerheid bij classificatietaken (Hoofdstuk 3). Uit dit onderzoek concluderen we wat de belangrijke onzekerheidsvraagstukken zijn en inventariseren we de methoden om de onzekerheid te bepalen (Hoofdstuk 4). Vervolgens introduceren we nieuwe methoden voor het schatten van het aantal fouten in de resultaten van classificatietaken en voor het corrigeren van de daaruit voortvloeiende vertekening bij het schatten van de klassengroottes (Hoofdstuk 5). Ten slotte onderzoeken we nieuwe visualisatietechnieken voor het bepalen van classificatiefouten (Hoofdstuk 6) en voor het analyseren van klassengroottes en bijbehorende onzekerheden (Hoofdstuk 7). We sluiten af met het bespreken van de implicaties van onze resultaten (Hoofdstuk 8).

We beschrijven de informatie die kan worden geboden door systemen die voor digitale beeldherkenning worden ingezet in wetenschappelijk onderzoek naar dierpopulaties. We bestuderen dit toepassingsdomein door marien ecologen te interviewen. Ook wordt een vergelijking gemaakt tussen de standaard technieken voor het verzamelen van gegevens, op basis waarvan we algemene informatiebehoeften afleiden. Na het interviewen van experts op het gebied van digitale beeldherkenning

identificeren we de behoeften die al dan niet middels video monitoringstechnieken kunnen worden ingevuld. Ten slotte bespreken we de onzekerheidsvraagstukken die inherent zijn aan elke techniek voor het verzamelen van gegevens, en identificeren we hoog-niveau eisen en wensen ten aanzien van het beoordelen van onzekerheid.

We onderzoeken de informatie die eindgebruikers ondersteunen bij de bepaling van onzekerheid. Onze tweede gebruikersstudie onderzoekt hoe informatie over classificatiefouten van invloed is op het begrip, het vertrouwen en de acceptatie van gebruikers met betrekking tot het systeem voor digitale beeldherkenning. We verzamelen hiertoe aanvullende feedback van gebruikers ten aanzien van onzekerheidsfactoren, en bespreken de relaties tussen het (on)begrip van onzekerheid, het vertrouwen en de acceptatie van de gebruikers. Onze conclusies werpen licht op onvervulde informatiebehoeften die aanvullende technieken voor het bepalen van onzekerheid vereisen, en tevens op de hoog-niveau informatie die met behulp van deze technieken aan gebruikers moet worden aangeboden.

We identificeren de belangrijkste factoren van onzekerheid waar rekening mee gehouden dient te worden bij wetenschappelijk valide analyses van beeldherkenningsresultaten. We richten ons op in- situ video monitoringstechnologieën, zoals die geïmplementeerd zijn binnen het Fish4Knowledge systeem om tellingen van individuele dieren per soort te uit te voeren met behulp van gefixeerde, niet-stereoscopische onderwatercamera's. We houden hierbij rekening met onzekerheidsfactoren die los staan van het beeldherkenningssysteem en voortkomen uit de omgeving waarin het systeem wordt ingezet (zoals het blikveld en de plaatsing van de camera). Na het specificeren van het typische systeem voor beeldherkenning en de bijbehorende randvoorwaarden voor implementatie, worden de onzekerheidsfactoren benoemd die uit interviews met marien ecologen en computerdeskundigen gedestilleerd zijn. Vervolgens identificeren we de interacties tussen onzekerheidsfactoren, en beschrijven we hoe onzekerheid doorwerkt tot het niveau van hoog-niveau informatie. Ten slotte identificeren we de bestaande en missende technieken voor het bepalen van onzekerheid.

We identificeren methoden voor het schatten van het aantal fouten in classificatieresultaten, waarbij we gebruik maken van de foutmetingen uit testsets. Deze methoden leveren zuivere schattingen op van de klassengroottes en richten zich niet primair op het identificeren van welke specifieke items verkeerd zijn geclassificeerd. De klassengroottes kunnen op zo'n wijze worden gecorrigeerd dat er voor elke klasse rekening wordt gehouden met potentiële foutpositieve en foutnegatieve gevallen (de zogenaamde false positives en false negatives). We reviewen de bestaande statistische en epidemiologische methoden om schattingsfouten te corrigeren, en onderzoeken hun geschiktheid voor classificatiemodellen in de context van beeldherkenning. Vervolgens breiden we de correctiemethoden uit met het schatten van het aantal fouten van de verschillende klassen. We identificeren de niet eerder verkende casus van disjuncte test- en doelsets, hetgeen implicaties heeft voor de variantie van de resultaten van foutcorrectie en -schatting. We introduceren vervolgens drie nieuwe methoden:

- De Sample-to-Sample methode schat de variantie van de resultaten van foutcorrectie en foutschatting in het geval van disjuncte test- en doelsets.
- De Ratio-to-TP methode gebruikt atypische foutratio's die eigenschappen hebben die relevant zijn voor het schatten van de variantie van de resultaten van foutschattingen.
- De Maximum-Determinant methode gebruikt de determinant van foutratio's, geformuleerd als een confusion matrix, als een voorspeller van de variantie van de resultaten van foutschattingen, voorafgaand aan het toepassen van het classificatiemodel op de doelsets.

We introduceren een vereenvoudigd ontwerp voor het visualiseren van classificatiefouten, d.w.z. van de fouten die gevonden zijn met behulp van een ground truth testset en die standaard in een confusion matrix worden gepresenteerd. We vermijden hierbij het weergeven van de foutratio, aangezien die verkeerd kunnen worden geïnterpreteerd. Vanuit onze ontwerpprincipes kiezen we voor absolute foutaantallen als een eenvoudige maar complete meting, evenals voor eenvoudige barcharts waar verschillende visuele kenmerken het onderscheid duiden tussen de feitelijke en geschatte klassen. We presenteren een gebruikersstudie die onze vereenvoudigde visualisatieaanpak vergelijkt met standaard visualisaties (ROC curve, confusion matrix en heatmap). We identificeren tenslotte de belangrijkste problemen die gebruikers tegenkomen bij het werken met de visualisaties en het begrijpen van classificatiefouten, hierbij rekening houdend met de basiskennis van de gebruiker.

We introduceren een uitgebreide visualisatietool waarmee eindgebruikers de klassengroottes kunnen monitoren en zij onzekerheden in specifieke subsets van de gegevens kunnen onderzoeken. We introduceren een interactieontwerp voor het onderzoeken van klassengroottes en de onderliggende onzekerheidsfactoren (zoals de kwaliteit van het videomateriaal en de tekortkomingen van het beeldherkenningsalgoritme). We onderzoeken middels een gebruikersstudie het interfaceontwerp en de wijze waarop deze gebruikers bewust maakt van de onzekerheden. We benoemen de factoren die de exploratie van data en bijbehorende onzekerheden faciliteren of bemoeilijken. Hierbij wordt er in het bijzonder aandacht besteed aan het feit dat gebruikers zich niet altijd bewust zijn van cruciale onzekerheidsfactoren. We sluiten af met aanbevelingen voor het verbeteren van het ontwerp van dergelijke interfaces.

Ons onderzoek draagt bij aan wetenschappelijke studies naar dierpopulaties op basis van digitale beeldherkenning. Onze resultaten dragen bij aan uiteenlopende toepassingen voor het omgaan met onzekerheid in systemen voor beeldherkenning en classificatie, in de zin dat zij de basis vormen voor het ontwerpen van een uitgebreide set van methoden en tools voor het bepalen van de onzekerheid.

Curriculum Vitae

Emmanuelle (Emma) Beauxis-Aussalet was born and raised in France, until she moved to The Netherlands for her PhD research. She obtained her first Bachelor of Graphic Design in 2004, and her second Bachelor of Webmaster-Webmarketer in 2006. She then obtained a Master of Digital Communication in 2007 (cum laude). Her interest in digital technologies led her to obtain a second Master of Computer Science in 2008, on the topic of Distributed Systems.

While studying for her Bachelors and her first Master, Emmanuelle worked part-time in communication agencies. She gained a 2-year professional experience as a webmaster and designer at CVB agency, working on internal communication and intranet websites for Renault. She gained a 1-year professional experience as a project manager at G2 agency (formely Grrrey), working on websites for major companies (SNCF, Varilux, Nokia). After finishing her second Master, she worked for 3 years at Thales as a R&D Engineer specialized in semantic technologies, system design and user interfaces. During her doctoral research, she worked for 2 years as a data specialist at LightHouse IP, providing the company with systems to collect, extract and analyze intellectual property data from over 100 data sources.

Emmanuelle conducted her doctoral research at the NWO institute Centrum Wiskunde e Informatica (CWI) from 2011 to 2018. As part of the Information Access group and the Fish4Knowledge project (http://www.fish4knowledge.eu), she added her design skills to the team and gained knowledge of information retrieval, computer vision, machine learning and statistics from her colleagues. She led the Classee project (http://classee.project.cwi.nl) in collaboration with the University of Amsterdam and Amsterdam Data Science. She started to work part-time in 2012, for personal reasons, and reduced her work hours in 2016 to take charge of her position at LightHouse IP.

Emmanuelle now works at the Digital Society School of the Amsterdam University of Data Science (also known as the Hogeschool van Amsterdam, HvA). Her role as the Senior Track Associate for the Data-Driven Transformation Track includes supervising learners and establishing research directions for innovative projects in collaboration with industrial partners. Her research interests include the transfer and application of data-driven technologies for the best interest of society, the development of machine learning literacy in the general public, and the development of explainable and accountable artificial intelligence.

Acknowledgements

The work presented in this thesis would have been impossible to carry out without the support of many colleagues, friends and family members. Nothing at all could have happened without Lynda Hardman, from whom I have learned so much about academic work and work-life balance. She has been able to see the best and the worst in me, and remained focused on bringing the best out of me. Her skills, rigor, patience and outstanding human qualities were essential to my progress, and brought much light into my tunnel.

My colleagues from the Information Access group at CWI were also essential to my progress. They kept my ignorance and morale in check, and their bright minds were great sources of ideas and joy. The CWI Personnel and IT departments were of tremendous help too. I could always count on them when needed, and they make it possible to work in one of the best working environments. Exchanging ideas with colleagues from other research groups at CWI has been most stimulating, and at times decisive for my research. Among the great people with whom I worked, directly or indirectly, I am especially grateful to Arjen (de Vries), Nishant, Desmond, Joost, Tiziano, Martin, Elya, Jiyin, Myriam, Astrid, Tessel, Thaer, Gebre, Jacco, Martine, Bikkie, Erik, Peter, Tom, Arjen (de Rijke), Max, Hannes and Léon.

My colleagues within the Fish4Knowledge project were also outstanding. Our adventures in Taiwan and Europe are unforgettable, especially when we ate blue jellyfish straight out of the sea at a Lanyu Island beach. For all the good moments we shared, intellectually and culturally, I would like to thank Robert Fisher, Fang-Pang Lin, Professor Shao, Hsiu-Mei, Karen, Bas, Concetto, Simone, Isaak, Daniela, Phoenix, Cigdem, Gaya, and Jessica.

Alongside my doctoral research, I had the chance to work with extremely kind colleagues at LightHouse IP. My special thanks go to Willem and Wiegert for their understanding and flexibility, and for welcoming me as a researcher with such an atypical schedule. My luck with outstanding working environments now continues within the Digital Society School at the Amsterdam University of Applied Science. Each of my new colleagues, with their creative and diverse mindsets, lets me appreciate larger perspectives at the confluence of science, design and society. Finding myself working among them gives a deeper meaning to my studies and work experiences. I am very much looking forward to the new adventures ahead of us.

Beyond my professional life, my personal life in Amsterdam has been rich and jolly thanks to the fantastic friends around me. There is much to stay about your friendship, and much to keep of the record. I would rather not mention any juicy

story or deep connection in particular, but let you remember them yourselves and know that our time together is most precious to me. Of those who have broadened my mind and my smile, much is owed to Sergio, Maarten, Steven, Tiago, Dome, Andreia, Igor, Romulo, Bea, Luis, Ana Sofia, Noortje, Roberto, Laura, Fleur, Wout, Henrique, Delyana, Luka, Oana, Victoria, and Lorraine. My special thanks to Deba, Teresa, Eleni, Pablo, Bram and Ays, my CWI companions who were most supportive and understanding.

My dear new friends from Amsterdam cannot overshadow my dear old friends from France, in particular Loriane, Laure, Magali, Guillaume and Lola. It was a sacrifice to leave you, but it certainly confirmed how important you are to me. A special thanks to my friend Gaëlle whose precious advice guided me into the academic world, before and during my PhD.

My deepest gratitude goes to my family for their unfailing support. No matter how far we may be from each other, knowing that you care for me was the most powerful source of strength. You were the invisible hand that carried me through the most difficult times, and that lifted my spirit at all times. What a blessing it is to be born among you, my parents Marielle and François, and my siblings Romain, Yann, Pierre, David and Anna. This blessing extends to our grandparents, and to our large family with its many branches of cousins and (great) uncles and aunts. My most tenderly geeky memories goes to my grandfather Lucien who taught me HTML and CSS, and fantastic stories about how he contributed to apply the first computers in the industry, with punch cards and palm-sized transistors.

Le meilleur pour la fin, Ralph gave me essential support and insights, about research, life or music. I could not accurately portray his wisdom and patience, or the joy and peace he brings to my life. He does not like me to go over the top, and although I am very tempted to do so here to pay him the right tribute, I'll keep it short: *you are the best*.


l' Acacia
la Lune
la Neige
le Plafond
l' Orage
le Désert
magritte