MRD 121, 183a, 116- Global and Site-specific Hazard/Particle Map

The "Hazard Maps" refer to two separate products the classify the surface terrain of Bennu and map out the spatial extent of hazards (objects > 21 cm in longest dimension), e.g. boulders, cobble, areas of extreme surface roughness, craters, linear features, and other geological features that considered "unsafe" for the OSRIS-REx TAGSAM. At the global scale, these features are mapped out in a thematic map of hazards and regions of interest (ROIs). At the site-scale, boulders and cobbles are mapped out in the cobble/particle map products.

The Hazard Maps consist of a PostGIS spatial database with a corresponding color-coded global mosaic which can be thresholded and translated into hazards-only binary mask. This three-part data product satisfies MRD-121 and 183a and will use the panchromatic base map generated by the IPWG during Detailed Survey as its input. The main hazards to be identified include boulders and other large objects (>21 cm), craters, linear features and other areas of exclusion. Areas of exclusion are defined as areas which are unfit for sample acquisition including, rocky terrain, overhangs, or areas in close proximity to multiple large objects. Areas will be classified as regions of interest if they satisfy the requirements that they are relatively smooth, accessible, and a safe distance from any large hazards. Hazards and regions of interest will be identified using both manual and automated methods. Manual methods include in-house 'experts' (members of the OSIRIS-REx mission and, specifically, the IPWG) who will manually identify surface features using the QGIS interface. Additionally, the citizen science website CosmoQuest is developing an online tool for educators and amateurs to contribute to boulder classification. CosmoQuest will deliver a relational database to the IPWG that will then be incorporated into the global hazard map database. Finally, hazards will be detected using a suite of machine learning algorithms found in the MATLAB Machine Learning and Statistical Toolbox. This process requires a small amount of manual training before automatically processing a large amount of data very quickly with a high rate of accuracy.

A spatial database is a database management system (DBMS) that uses special software (PostGIS) to extend a traditional relational database management system to store and query data in two and three dimensional space. The IPWG will be using shapefiles that store polygon geometry to identify hazards such as, boulders, craters, and linear features to classify areas on the asteroid's surface as either safe or unsafe for sampling. The shapefiles will be created using the open source geographic information system QGIS and an IPWG developed program that converts the output from the MATLAB Machine Learning and Statistical Toolbox to shapefiles using the convex hull function in the OpenCV library. By generating shapefiles tied to a Bennu specific datum, we will be able to load our data into PostGIS. The color-coded global mosaic and the hazards-only binary mask will be produced by running Structured Query Language (SQL) queries on the database.

Overview

Relevant ICD Data Products:

- Global Hazard Map (IP-2)
- Global Hazard Map, Ancillary Format (RD-15)
- Global Hazard Geodatabase (IP-18)
- Local Particle Maps (IP-5)
- Local Particle Geodatabase (IP-1)

What is the Data Type?

PostGIS spatial database

Color-coded global and site specific maps

Global global and site specific binary masks

What MRD does this data product satisfy or contribute to satisfying?

MRD 121: OSIRIS-REx shall image > 80% of the surface of Bennu with < 21cm spatial resolution (4-pixel criterion), once at 10am local time and once at 2pm local time, to produce a global mosaic, stereo images, mosaics of hazards and regions of interest, and image sequences of the asteroid surface.

MRD 183a: The Ground System shall produce the following data products on a global scale and for each candidate sample site in support of site selection during the encounter with Bennu: a. Safety Maps, b. Deliverability Maps, c. Sample-ability Maps, d. Science Value Maps

MRD 116: OSIRIS-REx shall, for > 80% of a 2-sigma TAG delivery error ellipse around at least 2 candidate sampling sites map the areal distribution and determine the particle size-frequency distribution of regolith grains < 2-cm in longest dimension.

What observations are required to provide the input data needed to make the data product?

PolyCam images acquired during the Baseball Diamond of Detailed Survey for the global panchromatic mosaic.

PolyCam images acquired during Orbital B for the 12 site-specific mosaics at 5 cm resolution.

PolyCam and MapCam (panchromatic) images acquired during Reconnaissance for the 2 site-specific mosaics at 2 cm resolution.

What is the spectral and/or spatial resolution of this data product?

The required spatial resolution is ≤ 21 cm (4 pixel resolution) for the global panchromatic mosaic from Detailed Survey (MRD 121) for > 80% of the surface of Bennu.

The required spatial resolution is ≤ 5 cm (4 pixel resolution) for the 12 site-specific panchromatic mosaics from Orbital B (MRD 576) for 100% of a 3σ TAG error ellipse.

The required spatial resolution < 2 cm (4 pixel resolution) for the 2 site-specific panchromatic mosaics from Reconnaissance for > 80% of a 2-sigma TAG delivery error ellipse.

When in the DRM are the observations that make the data product scheduled to be taken?

- Detailed Survey during Baseball Diamond
- Orbital B
- Reconnaissance

How long does it take to produce the data product?

Once the mosaicking is completed much of the processes to create the thematic hazard map will be automated. A significant portion of the identification will be done using a suite of algorithms found in the Matlab Machine Learning Toolbox. These machine learning based functions will quickly identify safe and unsafe areas at a measurable and high rate of accuracy. The algorithms will output masks which can be automatically read and converted to shapefiles using an IPWG produced program that uses OpenCV and the Shapefile C library. Hand classification will be the most time consuming process. It will be achieved using both in-house "experts" with additional input from the Citizen Science driven website CosmoQuest.

Of the data products the global binary mask will be produced the quickest. Since the detail of area identification is simply safe vs. unsafe, almost all of the classifying can be done with the automated machine learning classification algorithms. With training and processing times it should take one person 4-5 days to produce a usable binary hazard mask. This mask will be used to inform the Sampleability Map and therefore is the most time sensitive product. The output of the machine learning algorithms can be combined with the manually identified features to supplement the database and will include more nuanced classification including boulders, linear features, and craters. This data can be expected to take a longer time to collect, at least a month to acquire data from CosmoQuest and in-house identification. This data will be used to produce the color-coded global mosaic.

Is this product used for sample site selection, science value, or long term science?

This product will be used for sample site selection, science value and long term science.

Observation Requirements

For observation requirements for the input mosaics see the Data Product Description for the Global and Site Specific Image Mosaics.

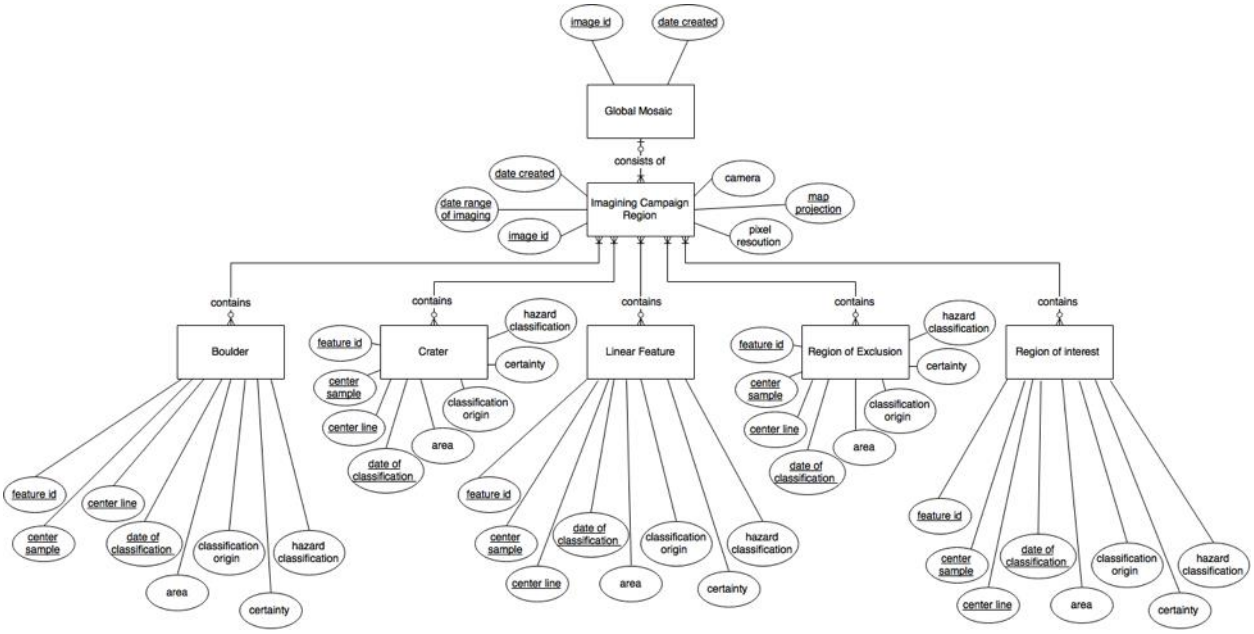
Data Product Structure and Organization

What is the structure of the data product (e.g. FITS file with 4 extensions)?

Binary mask: FITS file with x, y, z coordinate backplanes.

Color-coded global mosaic: FITS file

PostGIS spatial database E-R diagram:



Data Product Generation

By whom is the product generated?

IPWG members, primarily Carina Johnson.
What are the input products needed to produce the product?

Global and site-specific PolyCam panchromatic mosaics

Spice kernels

Asteroid shape model
Are there format expectations for the input products?

All input products to create the input mosaic will need to conform to the OCAMS SIS.

The input mosaic will need to be in ISIS3 cube format.

The asteroid shape model will need to be in Digital Shape Kernel (DSK) format.
What algorithms are used to generate products?

The Matlab Machine Learning and Statistical Toolbox is collection of classification, regression and clustering algorithms that will allow for a high level of automation in the feature detection process. The programs can output image masks that can then be inputted into makeShapefile and converted into postGIS compatible shapefiles.

The MATLAB Machine Learning and Statistical Toolbox is employed to explore/combine a sequence of learners that can be trained to classify the pixels nature (i.e. safe, unsafe, rock) and to rapidly prototype the algorithm to process data (i.e. global mosaic) and generate the thematic map in a timely fashion according to the OSIRIS REx data production schedule. The development of an accurate automatic classifier is articulated along the following phases:

1) Data Preprocessing and Feature Extraction phase

2) Training Set Generation phase

3) Classifier selection Phase

4) Training and Validation Phase.

After phase 1 through 4 are executed, the classifier is ready for deployment on the global mosaic. Visualization algorithm may display the maps according to the proper color code that identify the regions. Each phase is described in more detail below.

1) Data Preprocessing and Feature Extraction phase

The automatic classifier generates a class value per pixel of the global mosaic. The data preprocessing and feature extraction phase consist of preprocessing the data with a set of image processing filters and extract the features employed as input to the classifier. Generally, the global mosaic provides a single "predictor" (one single pixel value) which is not sufficient to provide adequate discrimination between the different thematic classes. The feature extraction process generates a set of "predictors" by running a set of sliding windows at different scales over the global mosaic and implementing a set of filters to extract the desired predictor. Example of filters are:

1. Median Filter

2. Standard Deviation Filter

3. Laplacian of Gaussian Filter

4. Prewitt Filter

5. Sobel Filter

Figure 1 shows an example of data preprocessing and feature extraction on an image of the Itokawa asteroid. A set of 12 predictors per pixels have been produced using the MATLAB Image Processing Toolbox, i.e. median filter, standard deviation, Laplacian of Gaussian (LoG),

Prewitt (Gradient Magnitude) filters at three different scales (3x3, 9x9 and 15x15 window centered on the pixel of interest).

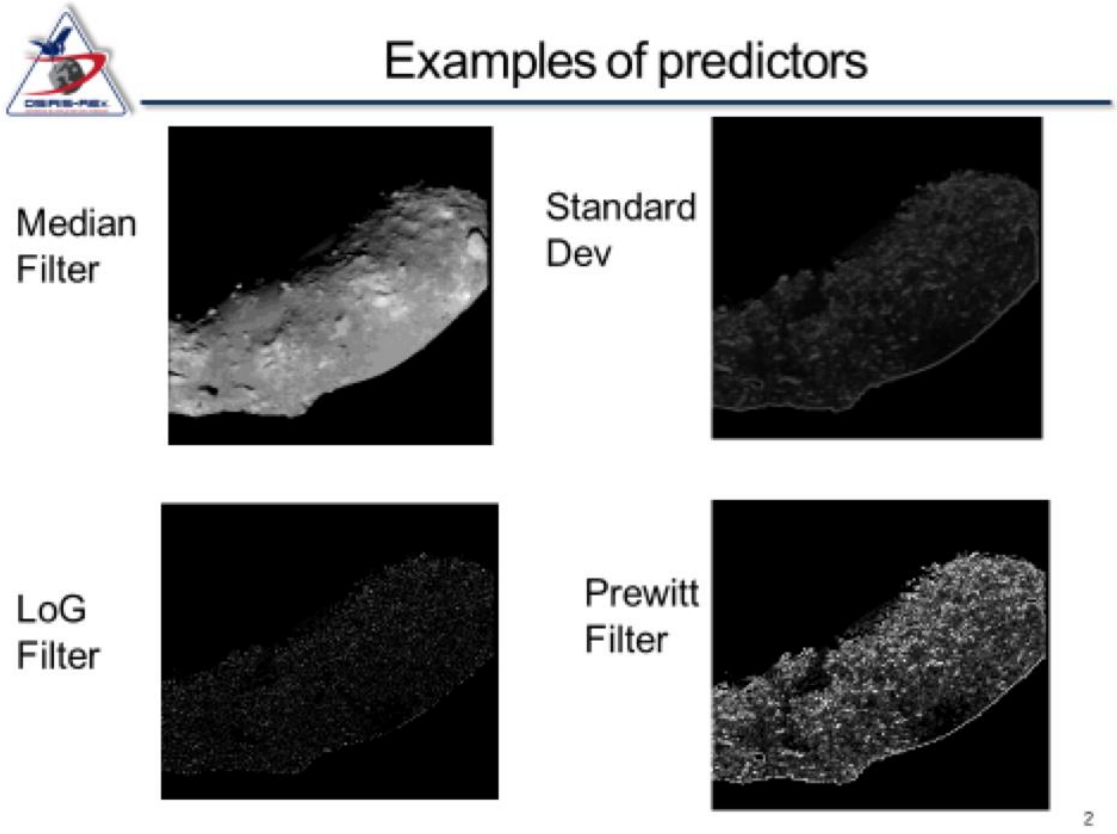


Figure 1: Examples of 4 predictors using a sliding window 3x3

2) Training Set Generation phase

The training set generation phase consists in generating training examples for each selected class. The process comprises loading the global mosaic and manually selecting areas on Bennu surface that correspond to the desired thematic class. Selection occurs by manually masking the different classes out of the images and then generating a table containing pixels labeled with the corresponding class. Figure 2. shows an example of the masks generated for the different training class. Masks have been generated using functions of the MATLAB Image Processing Toolbox. A set of total 19,319 training examples have been generated for the three considered classes.

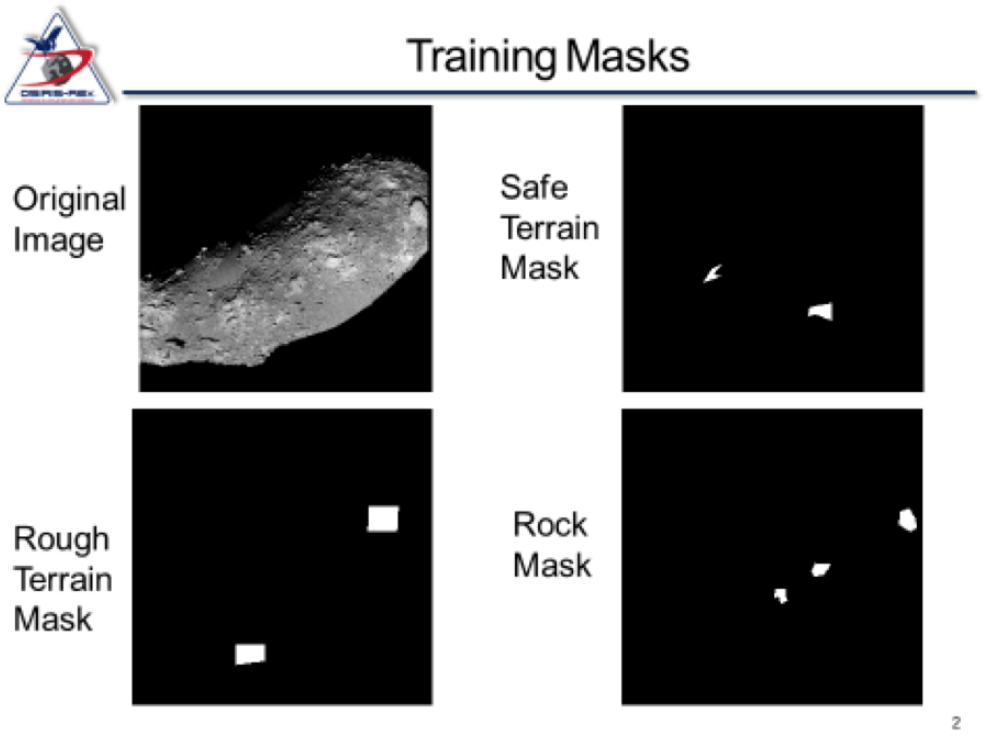


Figure 2. Training Masks employed for Training on Itokawa Images

3) Classifier selection and Training phase

Once the training set is generated, a proper classifier can be designed and analyzed. A variety of techniques to design the classifier are generally available. Classifiers performance are generally function of the available training set, predictors/features and selected techniques. We employ the MATLAB Classification Learner App to explore the design space for classifier selection and rapid prototyping of the classification algorithm. A large variety of algorithms are supported including:

1. Decision Trees
a. Simple/Medium/Complex Tree
2. Discriminant Analysis
a. Linear/ Quadratic Discriminant
3. Support Vector Machines (SVM)
a. Linear/Quadratic/Cubic SVM
b. Coarse/Medium/Fine Gaussian Kernel
4. Nearest Neighbor Classifier (KNN)
a. Coarse/Medium/Fine KNN
b. Ensemble Classifiers
i. Bagged Trees
ii. Boosted Trees

For the Itokawa Case Example we designed and implemented Bagged Decision Treee comprising 100 weak learners.

4) Training and Validation Phase

The training process occurs interactively within the Testing and Validation Phase. For a selected classifier, the MATLAB Classification Learner App is employed to import the data, train classifier on a set of training points and evaluate results. The process occurs iteratively where classifiers are selected and compared for accuracy before the final trained classifier is prototyped. Training set is generally partitioned in training and validation. Validation is necessary to avoid overfitting and evaluate the classifier performance on points that has not been trained. Generally cross-validation is implemented where the data set is partitioned into folds and accuracy of each fold is evaluated.

Figure 3. shows the confusion matrix generated during the training and validation of the selected bagged decision tree (100 weak learners) on the training set for safe, rough and rock classification (Itokawa). Cross validation has been implemented by partitioning the training data in 5 folds. Accuracy is established to be of the order of ~98%.

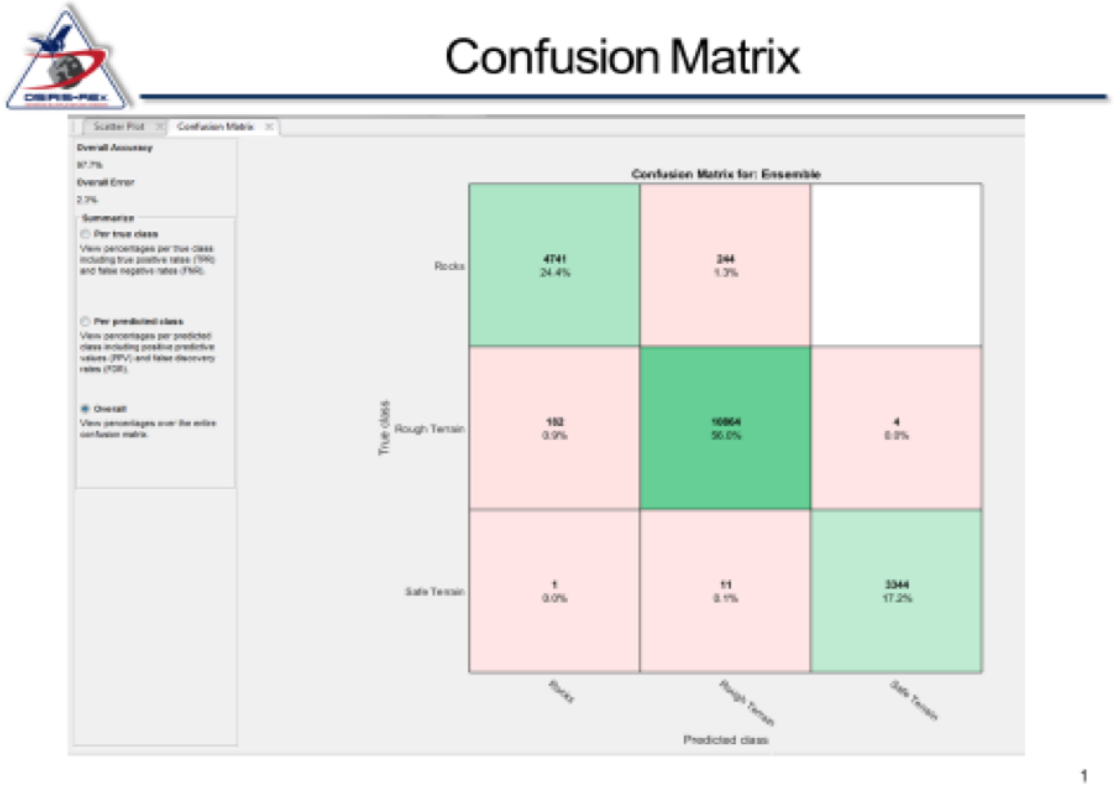


Figure 3. Confusion Matrix for the Bagged Decision Tree. Accuracy is 97.7%

After phase 1 through 4 are executed, the classifier is ready for deployment on the global mosaic. Visualization algorithm displays the maps according to the proper color code that identify the regions. Figure 4. Report the classification map (automated Thematic Hazard Map) generated by running the classifier over the Itokawa Image.

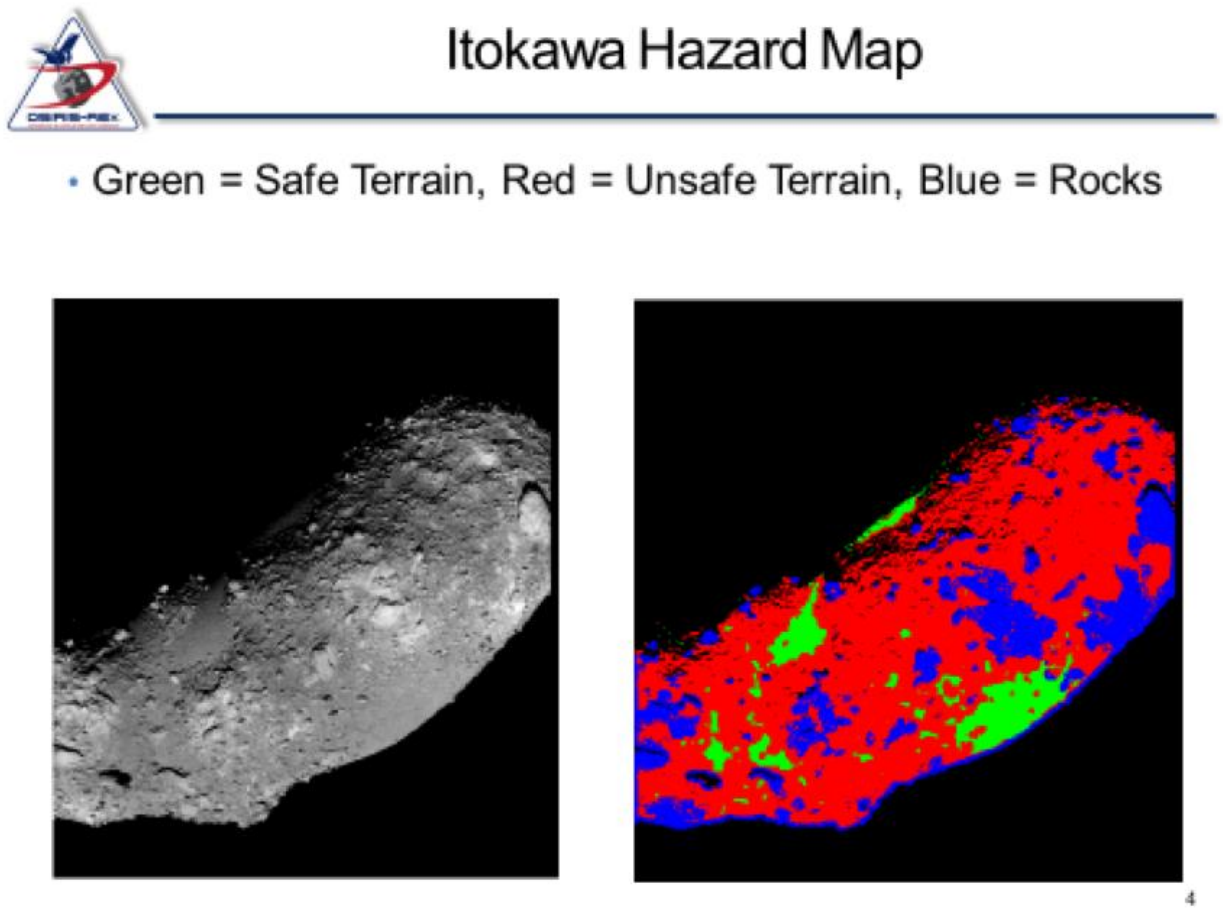


Figure 4. Automated Thematic Map of Hazards and Regions of Interest via the Trained Classifier

The output image masks of the MATLAB Machine Learning and Statistical Toolbox can then be inputted into makeShapefile and converted into postGIS compatible shapefiles. makeShapefile is an IPWG produced program to convert image masks to shapefiles. It uses a computer vision algorithm called convex hull found in the OpenCV library to identify blobs, trace their contours, find the minimal number of points to be able to recreate the shape, and output those points. These points are the sample/line coordinates of the blobs in the context of the original image file so mappt, an ISIS3 utility program is used to convert the sample/line coordinates to X/Y coordinates. From them shpcreate and shpadd, from the shapefile C library, are used to convert those X/Y coordinates into shapefiles that correspond to the geometry of the original mosaic file.

What calibration data are used to generate products?

Global and site-specific mosaics made from PolyCam L2 images is used to create this product. The images used to generate the input mosaics will be radiometrically calibrated by the OCAMS pipeline prior to use by the IPWG. ISIS3 will geometrically correct the images using the

PolyCam camera model.
Has a specific Science Team Member been assigned to produce this product?

Carina Johnson and Roberto Fufaro with assistance from Mathilde Westermann.
Will multiple versions of the product be generated?

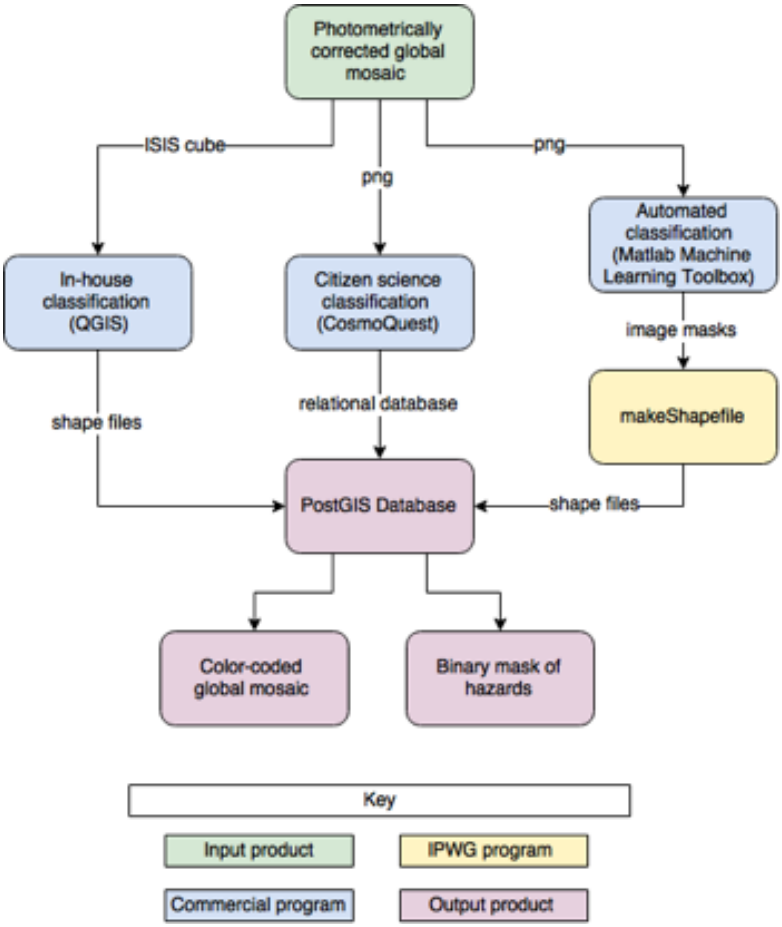
Yes, there will be a preliminary binary hazard map generated from the MATLAB Machine Learning Toolbox results as well as the in-house manual classifications. A new binary map and color-coded global mosaic will be generated when the CosmoQuest data becomes available. The database will constantly be updated as new data is generated. Additionally, these products will be regenerated every time a new shape model or a new photometric model is released.

Data Product Validation

How will the product be validated to ensure contents and formats are correct?

This data product will need to pass an IPWG acceptance test before being delivered to the SPOC. The IPWG acceptance test will ensure that format is correct.The IPWG lead will review the content of the data product and the data provenance before it is delivered to the SPOC.

Data Flow



Standards used to generate data product

Cartographic Standards

IPWG input mosaics will be generated in either equirectangular or polar stereographic cartographic projections. The equations describing these projections are included in the Image Processing Software Interface Specification (SIS) document.