

Alouette-1 – Ionogram Data Extraction Methodology

I. An Evergreen Document

The Canadian Space Agency (CSA) has developed an approach to extract data from scanned images of Alouette-1 ionograms, which were originally recorded on 35 mm negative films. However, the data extraction methodology presented is continually improving, and is being applied more generally to Alouette-2, ISIS-1 and ISIS-2 images of scanned ionograms. Therefore, it is expected that this document will be updated as new improvements are implemented.

This document will present in detail some of the errors in data processing and extraction. And it will provide instructions on how to re-process the data, so that one may take the approach presented and try to improve upon it.

It is helpful to gain an overview of the history of Alouette-1's data, and the data restoration efforts that occurred, which has led to the current set of Alouette/ISIS (Alouette-1, Alouette-2, ISIS-1, ISIS-2) scanned images:

- **Alouette-1 – A History of The Data from Canada's First Satellite Over 60 Years (2023) – Ravendra Naidoo**

There is more detailed information about the Alouette-1 scanned images and how to read them, which was written when scanning of the 35 mm film rolls started in 2017:

- **Alouette-1 Data Restoration (2017) – David Lessard**

Furthermore, much work was done from 2017 – 2022 that are included in the archive.

II. Data Acquisition

The images (.../Alouette-1/data/) are organized into Directories and Subdirectories. Directories correspond to a box of 35 mm negative film rolls; and Directories tend to start with the letter 'R'. Then, a particular Subdirectory within a Directory corresponds to a particular 35 mm film roll. Within that Subdirectory is a set of images, which are the scanned ionograms on that film roll (see Figure 1).

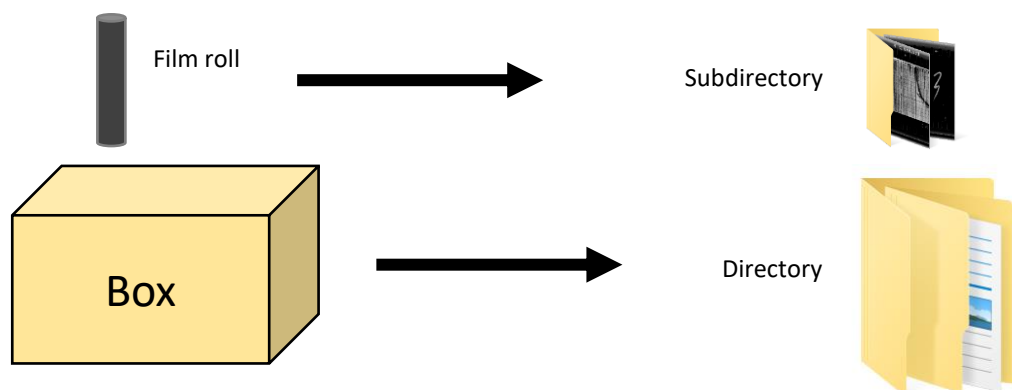


Figure 1: Subdirectories are folders of scanned images of ionograms from a film roll. A Directory corresponds to a box of film rolls.

2.1 FTP Entire Download

The simplest way to acquire the raw scanned images and the extracted ionograms is to download it entirely from the CSA FTP, using the following details:

FTP server: data.asc-csa.gc.ca

Username: anonymous

Password: [users must enter their email address]

Path to Alouette-1 data: users/OpenData_DonneesOuvertes/pub/Alouette-ISIS/Alouette-1/

Then in the directory for your project, rename the three folders:

- data → 02_downloaded
- result → 05_result
- log → 06_log

2.2 FTP Download by Subdirectories

However, it is an extremely large amount of data approximating ~ **2.6 TB**. An alternative approach was developed to download subdirectories of data one-at-a-time.

To download subdirectories one-at-a-time, first run the following Jupyter notebook:

- **00_Download Subdirectories from FTP.ipynb**

Follow the instructions in the notebook to set-up folders correctly in your project directory. Then test downloading of subdirectories.

Once subdirectories are downloading correctly, the process can be continued on a loop, by running the following Python script from the command line with 3 arguments:

```
Alouette_downloader.py [the file path to the root directory for your
project] [user's email address for FTP connection] [username for
download log]
```

For example:

```
python Alouette_downloader.py "C:/Projects/Alouette-1/data/"
"x@xmail.co" "UserX"
```

For the program to work, you will need to download the 'log' folder from the FTP directly and save it in the root directory for your project. Then rename the 'log' folder to '06_log'.

III. Inspection of Extracted Ionograms

Once you have acquired data, it is a good idea to inspect a sample of ionograms along with its corresponding extracted ionogram points. Run this Jupyter notebook:

- **01_Inspect Extracted Ionogram Points.ipynb**

If you had only downloaded a few Subdirectories of images following section 2.2, you will also need to download the corresponding mapped coordinates. Go to the FTP source for Alouette-1 mapped coordinates:

- `/users/OpenData_DonneesOuvertes/pub/Alouette-ISIS/Alouette-1/result/mapped_coords/`

, and download the corresponding mapped coordinates for the subdirectories you wish to inspect. Save these in '05_result/mapped_coords/' in your project root directory.

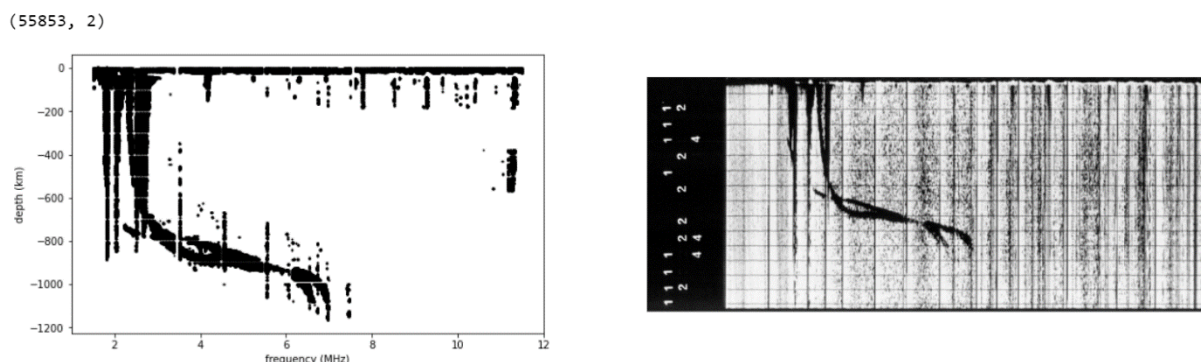


Figure 2: (left) Extracted datapoints mapped to a depth-frequency space; (right) the corresponding scanned image of the ionogram.

IV. Working with the Processed Data

The CSA has already processed the entire set of all available Alouette-1 scanned ionograms. The results are extracted mapped coordinates (i.e: digital ionograms), along with their read metadata (i.e: the ground station that telemetry was received, date and time of topside sounding, etc).

Metadata for each fully processed and read image are found on the FTP server, here:

- `/users/OpenData_DonneesOuvertes/pub/Alouette-ISIS/Alouette-1/result/result_master.csv`

Information about images that were not fully processed or read, can be found in the intermediate results tables ('result_postprocess_raw.csv', 'result_stage2_raw.csv', 'result_stage1_raw.csv').

The extracted mapped coordinates, as described in section III, are found here:

- `/users/OpenData_DonneesOuvertes/pub/Alouette-ISIS/Alouette-1/result/mapped_coords/`

For details about the processing event, the telemetry stations, and other information about the scanned images, can be found here:

- `/users/OpenData_DonneesOuvertes/pub/Alouette-ISIS/Alouette-1/log/`

Alouette-1 ionograms have three different metadata types, as described in Naidoo (2023) and Lessard (2017), and which are depicted in Figures 3 and 4.

The digitization of Alouette-1 film rolls occurred in two batches. 'BATCH I' scanning started in 2017, while 'BATCH II' started in 2021.

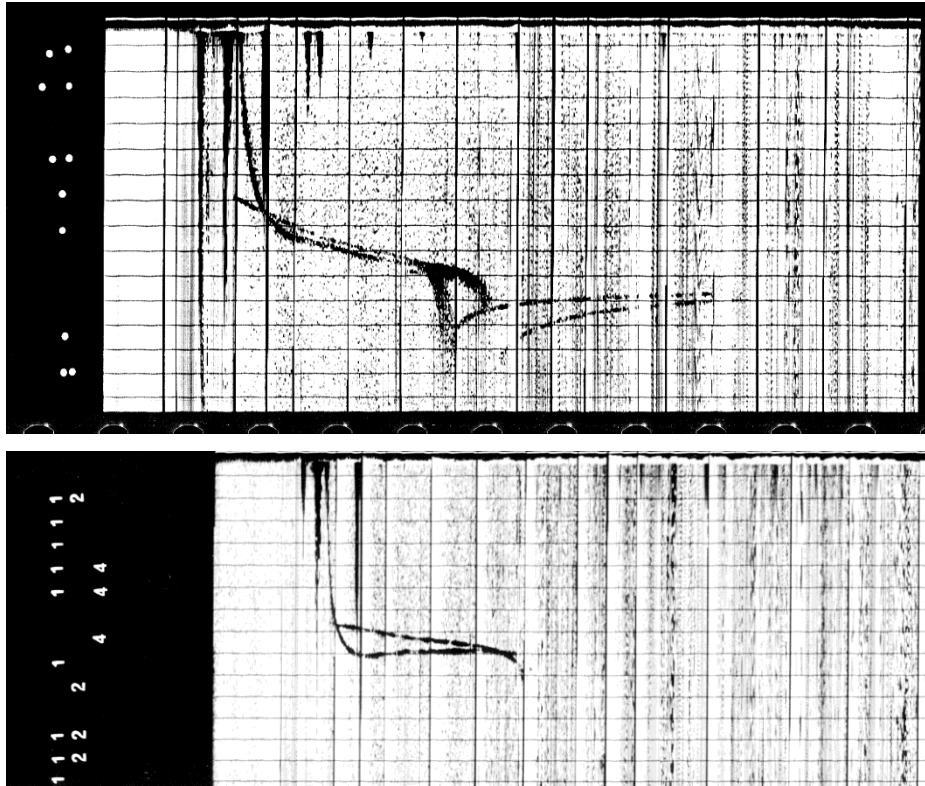


Figure 3: (top) 'dot' - Ionogram with dot metadata type. Early ionograms produced by DRTE encoded metadata into horizontal lanes, where each lane would represent a particular digit. The horizontal positions of dots within each horizontal lane would encode a number in binary format; (bottom) 'num' - Due to the ambiguity of the relative positions of each dot in each metadata lane, dots were replaced with numbers, but they were still arranged in the binary format. Therefore, one lane with a 1, 2, and 4 would be summed and represent 7. The new metadata format also included a lane to represent the year.

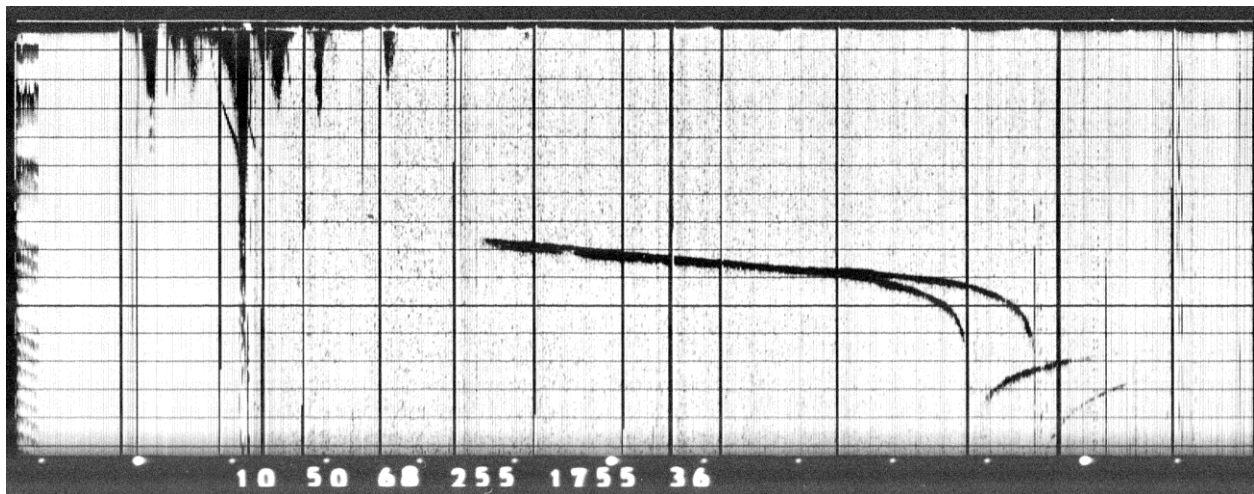


Figure 4: 'num2' - An easier to read numeric metadata format was introduced in 1965. A machine-learned text recognition model was used to read the metadata from ionograms with this format.

The total availability of fully extracted Alouette-1 data across time, and by ground station, is depicted in Figure 5.

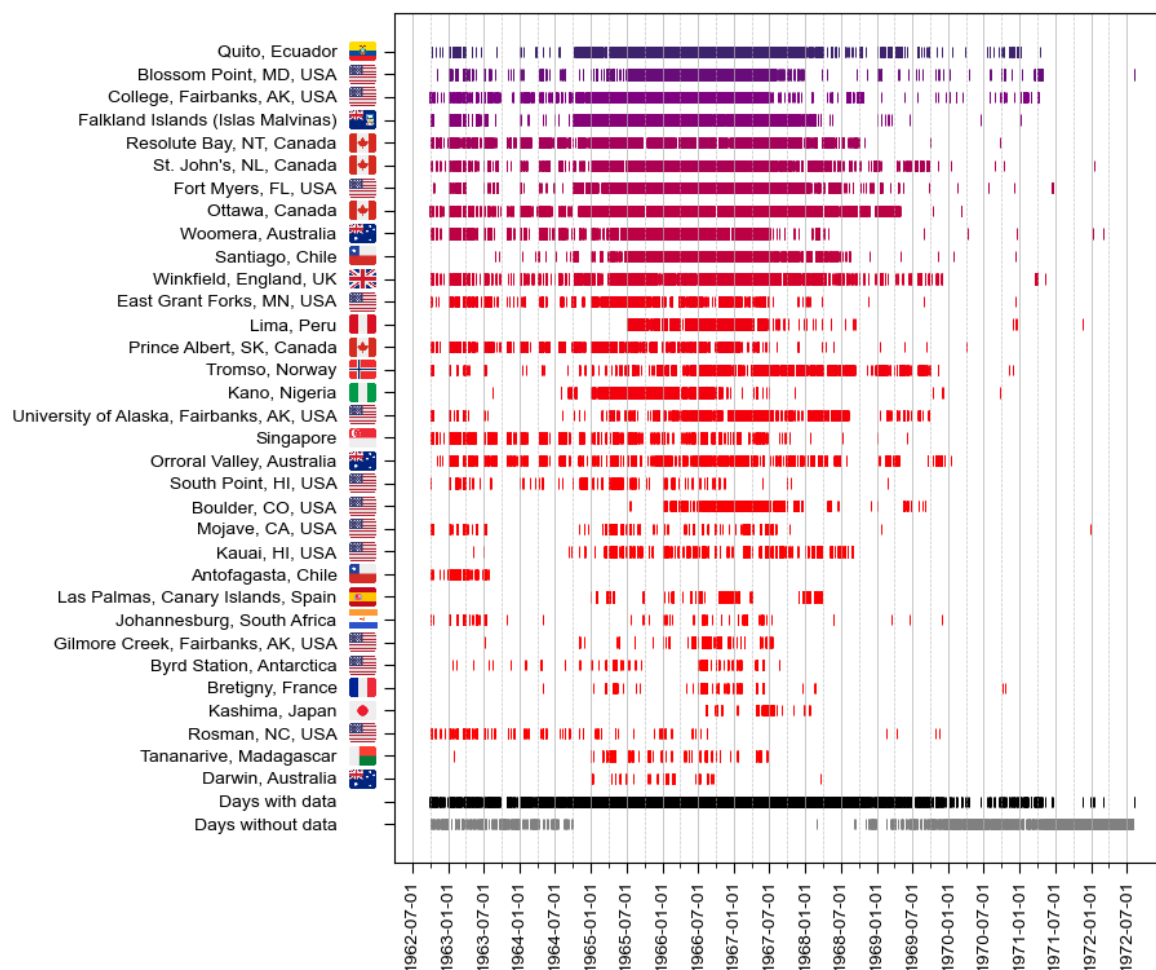


Figure 5: Current availability in days of Alouette-1 extracted data, by ground station.

V. Quality Analysis

Ultimately, **693,677 scanned images were fully processed and read** – representing a massive new dataset of Alouette-1 digitized topside ionograms.

However, out of 1,612,104 available Alouette-1 scanned ionogram images, **43.03%** were thought to be fully processed and read (i.e: the yield). Though as this section will discuss, there is some degree of error in data extraction and metadata reading within the set of ‘fully processed and read’ images.

As detailed in section VI, subdirectories of scanned images are processed through two stages (Stage 1, Stage 2), and then undergo a post-processing step to produce the ‘result_master’ dataset. After Stage 1 processing, 75.64% of images were fully processed, with 24.36% lost for various reasons (see Table 1). After Stage 2, the proportion of images that were considered fully processed reduced from 75.64% (after Stage 1) to 58.44%, for various reasons related to metadata misreading (see Table 2). After Stage 2 processing, 44.81% of images were classified as ‘num’, 9.06% were classified as ‘num2’ and 4.57% as ‘dot’ metadata type.

Table 1: Top-5 Functions That Caused Loss – After Stage 1 Processing

Function	Images
image_segmentation.segment_images_in_subdir.segment_images: iono size outlier	81,777
metadata_translation.determine_leftside_metadata_grid_mapping.extract_centroids_and_determine_type	38,122
ionogram_content_extraction.extract_select_parameters.extract_fmin_and_max_depth	22,546
image_segmentation.trim_raw_metadata.trimming_metadata	10,550
image_segmentation.segment_images_in_subdir.segment_images: metadata size outlier	2,731

Table 2: Top Recorded Reasons for Loss – After Stage 2 Processing

Detail	Images
metadata could not be read by OCR	246,360
OCR read metadata contains letters	23,984
metadata was interpreted to be num type	4,087
metadata was interpreted to be dot type	4,073

In the post-processing step, further losses could be attributed to various reasons. For the 'num' and 'dot' metadata types, it is assumed that each metadata lane must have a range of 0 to 10. Therefore, in cases where a metadata lane was read to be > 10, the result was overwritten as 'NaN' (not a number). Then as metadata digits were combined to form the various metadata elements (i.e: year, day of year, hour, etc), if these exceeded logical bounds (such as day of year > 366), then the metadata element was also rendered as 'NaN'.

As a result, the timestamp associated with each ionogram could be rendered 'NaT' (not a time) if just one time element was 'NaN'. Therefore, a 'time quality' level was ascertained for each processed and read image in 'result_master'. A time quality of 1 means that the time is read to the nearest second (i.e: all time elements were read); time quality 2 – to the nearest minute; time quality 3 – to the nearest hour; time quality 4 – to the nearest day. In 'result_master' while the overall yield is 43.03%, the yield with at least time quality level 4 is 44.50%.

Finally, if a station number is read but it does not correspond with a known ground station, then this will also result in a loss. Further details about the general quality analysis can be found here:

- **04_Quality Analysis - General**

Processing yield is also limited by scan quality issues, film quality issues, and even telemetry tape quality issues. For instance, the NASA Goddard data restoration project documented that when reading the telemetry tape, there were sometimes more frequency markers detected than expected, and the ionogram frame sync pulse (that indicates the start of the ionogram) was sometimes not detected at all. These telemetry tape data quality issues would have likely impacted the analog ionograms produced in the 1960s, and such issues are observed in the scanned images today.

Therefore, it is often observed that the quality of extracted data from an image can be related to its subdirectory, where there are entire film rolls (subdirectories) that can have poor quality. As Figure 6 depicts, there are almost 700 subdirectories that have a very poor yield (< 10%).

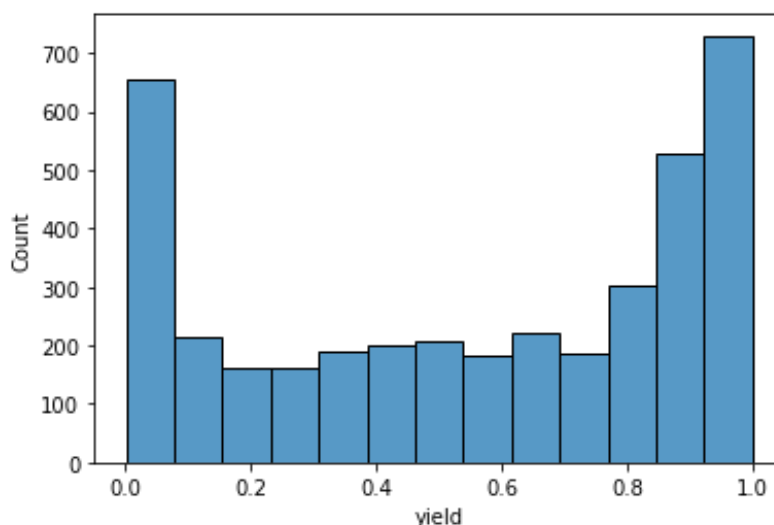


Figure 6: Histogram of yield in result_master dataset, by subdirectory.

5.1 Scan Error Detection

In the digitization of Alouette-1 35 mm film rolls, several issues arose that could be attributed specifically to the scanning of the film and the production of images. While it would require a great manual effort to check each subdirectory of images and make a qualitative assessment, two specific issues could be detected programmatically.

'Out-of-phase' images were those that spanned parts of two adjacent ionograms (see Figure 7). This type of scan error can be detected by checking the aspect ratio (height:width) of the image and detecting the number of digits in each image (in the case of 'num2' type images). A normal 'num2' image has an aspect ratio between 2 and 3, and has 15 digits on the bottom. Images with < 13 or > 17 digits are flagged as abnormal and potentially 'out-of-phase.'

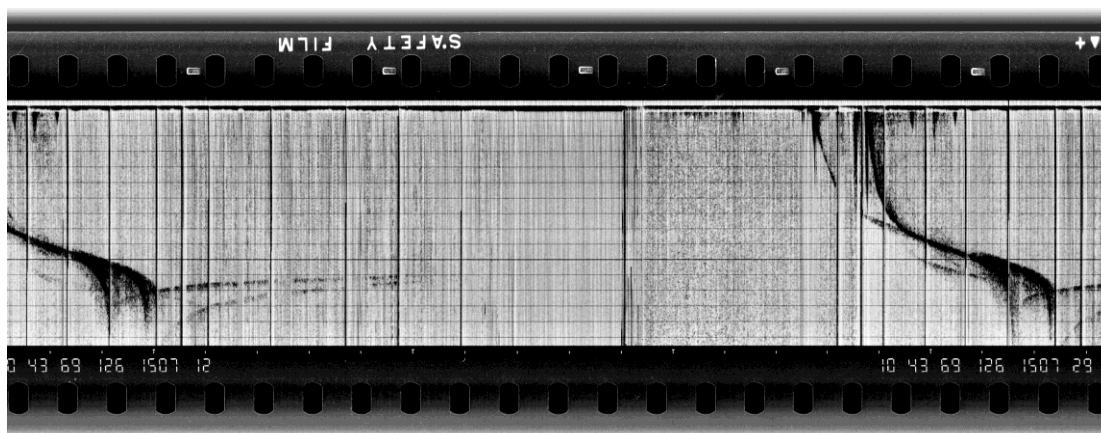


Figure 7: 'Out of phase' scanned image. The scan window should have been aligned to the start of the ionogram, in order to capture an entire ionogram, instead of parts of two adjacent ionograms.

An abnormal aspect ratio, < 2 or > 3 could be indicative of an 'out-of-phase' image, or in many cases they are 'irregularly cropped.' It was found that some images were not cropped properly to display an entire ionogram (see Figure 8).

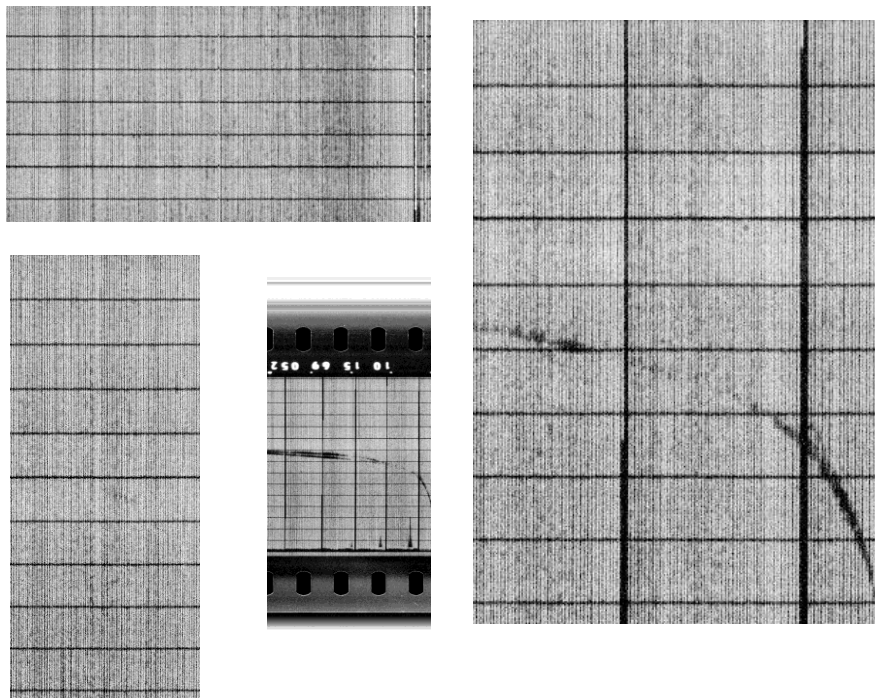


Figure 8: Examples of irregularly cropped images.

Further details about scan error detection can be found here:

- **A_Scan Error Detection Analysis.ipynb**
- **Aa_Scan Error Detection (BATCH II).ipynb**
- **Ab_Efficient Scan Error Detection (BATCH II).ipynb**
- **Ac_Scan Error Detection Results Analysis.ipynb**

By this method, for Alouette-1 images it was determined that about 4.7% of subdirectories may be out-of-phase and 0.1% were irregularly cropped.

5.2 Ionogram Mapping Accuracy

The data extraction algorithm attempts to read the metadata from a scanned ionogram image (see section 5.3) and extracts a set of datapoints that generally correspond to the ionogram trace (see Figure 2). The algorithm maps each datapoint from a (x, y) coordinate system of the image to a (frequency, depth) coordinate system, based on the reference lines on the ionogram (i.e: the grid system). The extraction process does pick up what are likely non-physical artifacts and noise that exist on the ionogram films, that were a result of scanning, or because of the data extraction itself (e.g: thick reference lines could be extracted).

Work was done to estimate the mapping accuracy of the data extraction algorithm, by drawing a random sample of 553 ionograms (399 images from BATCH I, 154 images from BATCH II) and manually reading the frequency and depth of the ionogram at the maximum-depth point. This is compared to the maximum-depth point from the corresponding set of extracted mapped coordinates (see Figure 9).

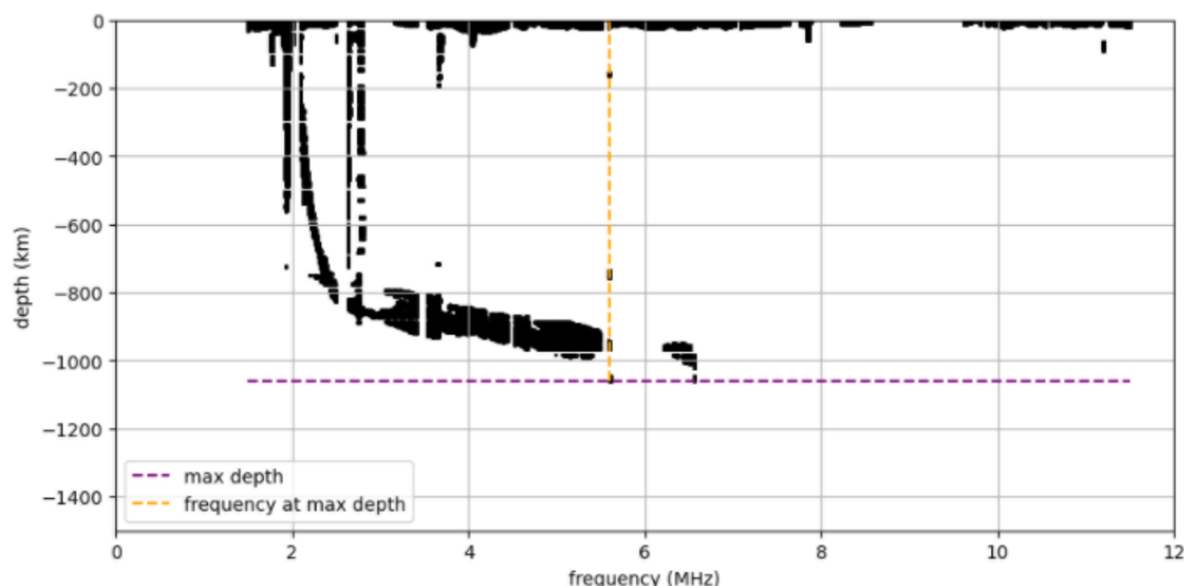


Figure 9: Maximum-depth point shown on a plot of extracted mapped coordinates.

It was determined that for BATCH I, the extracted maximum-depth point was the same (within ± 1 MHz, and ± 100 km) as what was manually read just over 50% of the time. However, BATCH II showed significantly worse results with mapping being the same as what was manually read only $\sim 10\%$ of the time. This may be because ionogram quality was generally worse for BATCH II, as determined by the scan error analysis (see section 5.1) and qualitative observations. From these results one may conclude that the performance of the ionogram mapping aspect of the data extraction algorithm is strongly dependant on the quality of the ionogram images.

Details about mapping accuracy determination can be found here:

- **B_Determine Data Mapping Accuracy.ipynb**
- **Alouette-1 Data Mapping Update-AFerreira_20230321**

The usability of the extracted ionogram points will be limited by the type and rigour of scientific analysis required. However, taken in aggregate a large volume of extracted ionogram traces as these could still serve valuable purposes for scientific research and the Canadian public, as illustrated in Naidoo (2023).

Further work can be done to improve the extraction of the ionogram trace(s) while discarding line-based noise and other artifacts.

5.3 Metadata Reading Accuracy

As shown earlier, Alouette-1 ionograms have metadata encoded in one of three metadata types: 'dot', 'num', and 'num2'. A random sample of images for each metadata type was drawn. The metadata for each

image was then manually read and checked against what was read by the data extraction algorithm (see Table 3).

Table 3: Metadata Reading Accuracy of Random Sample of Alouette-1 Ionograms, by Metadata Type

Metadata Type	Images, Stage 2 processed (BATCH I, BATCH II)	Images, manually read from random sample	Images, entirely correctly read by algorithm from random sample	% of images correctly read by algorithm from random sample
num	579,054 (465,009, 114,045)	99	94	94.9%
num2	126,073 (38,149, 87,924)	49	42	85.7%
dot	12,265 (7,784, 4,481)	25	0	0.0%

The metadata reading accuracy for each metadata type random sample is reported as the proportion of images *entirely* read correctly by the data extraction algorithm. In the num type random sample, some images were misread because they were low quality num2 images; the algorithm was unable to correctly determine that the image was of num2 type, and so it was misclassified as num type. In the num2 random sample, some errors were due to the text recognition model incorrectly reading the number 1 as 7, and vice-versa. Or it could misread 6 as 8, and vice-versa. Finally, dot type metadata reading by the data extraction algorithm is clearly inadequate and should be improved in future work. This is understandable as it is also difficult to manually read dot metadata.

Looking more specifically at the metadata reading accuracy for each metadata element, generally there does not appear to be a systematic error in the misreading of a certain metadata element – in either the num or num2 type images. See Figures 10 and 11 on page 11 and 12 respectively.

As for the dot type images, the reading accuracy for each metadata element was zero for the random sample taken.

Further details regarding how the metadata reading accuracy was determined, can be found here:

- **C_Overall Metadata Reading Accuracy for Alouette-1.ipynb**
- **Ca_Metadata Reading Accuracy (BATCH I).ipynb**
- **Cb_Metadata Reading Accuracy (BATCH II).ipynb**

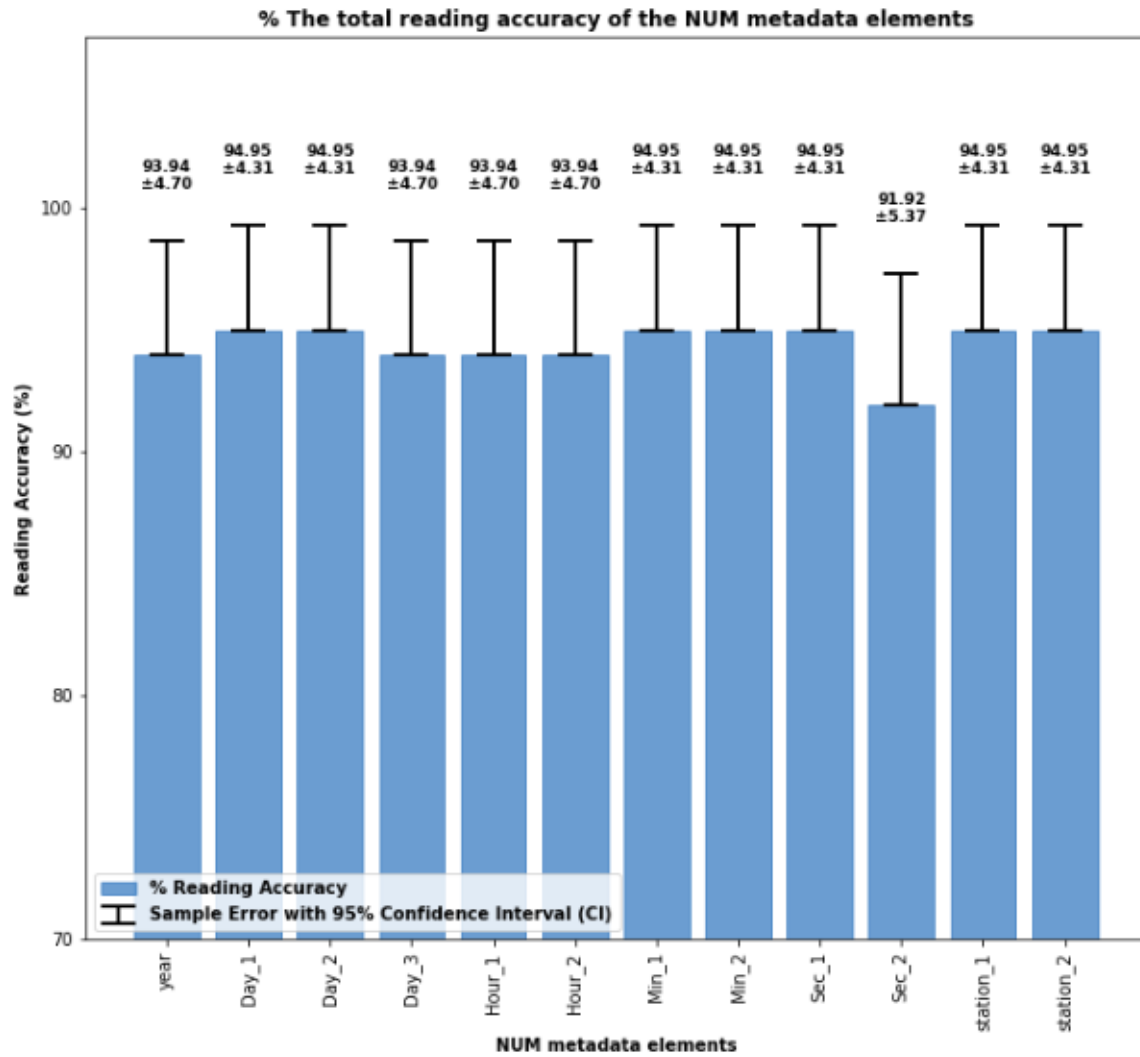


Figure 10: Metadata reading accuracy from random sample of num type ionograms, by metadata element.

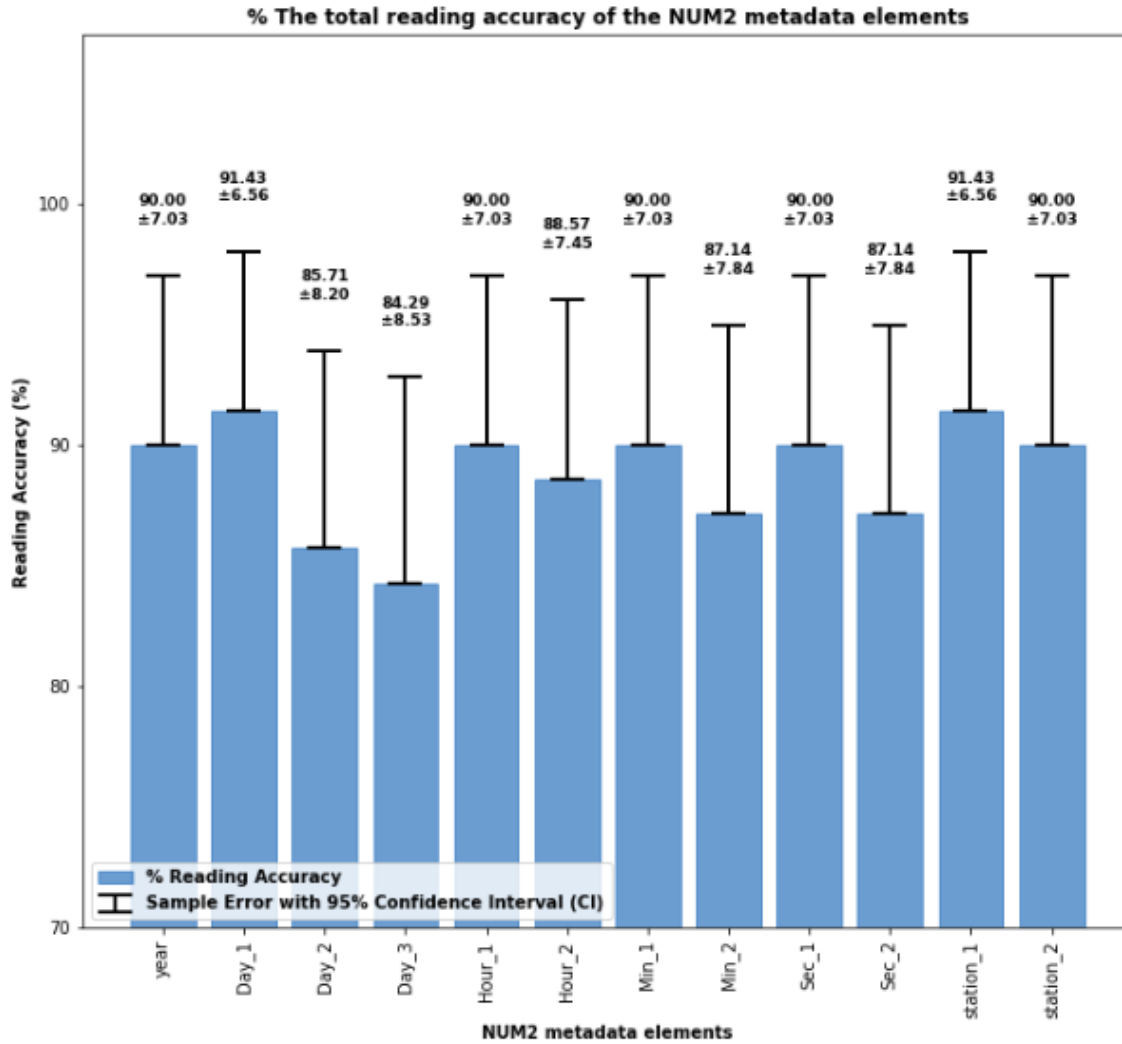


Figure 11: Metadata reading accuracy from random sample of num 2 type ionograms, by metadata element.

VI. Re-Processing

For many purposes the already processed data may be sufficient. However, one may wish to re-process the images, and even make improvements upon the current data extraction method.

5.4 Image Inventory

The image inventory is required for the processing method presented. It is already provided, here:

- `/users/OpenData_DonneesOuvertes/pub/Alouette-ISIS/Alouette-1/log/image_inventory.csv`

However, if somehow new scanned images were to become available, or for any other reason, the `image_inventory.csv` can be regenerated by running:

- `02_Image Inventory.ipynb`

5.5 Stage 1 Processing

Subdirectories of scanned images are processed through two stages: 'Stage 1' and 'Stage 2'. Stage 1 is the original, or legacy, data extraction algorithm that was developed from 2017 – 2020. A wrapper was developed around the original algorithm to manage and log performance of this processing step, along with other improvements.

Stage 1 processing can occur by running multiple instances of `Alouette_processor.py`. You can run as many instances as your system will allow. Multiple systems can even be used to run instances, so long as the root directory for the project is in a shared location.

Details about setting up the root directory are in '00_Download a Subdirectory from FTP.ipynb'. Ensure that the root directory has the following folders:

- `04_processed`
- `04a_unprocessed`
- `05_result`
- `06_log`

Furthermore, set up an instance directory. If only running one instance, the instance directory can be the root directory. If running multiple instances, set up a separate instance directory for each instance (e.g: 'Stage1_instance1/', 'Stage1_instance2/', etc). In either case ensure that each instance directory has the following folders:

- `03_processing`
- `05a_result_local`

To start Stage 1 processing, run:

```
Alouette_processor.py [the file path to the scanned images, organized
into Directories and Subdirectories] [the file path to root directory
for your project] [The file path to the instance directory] [username
(for logging purposes)] [instance number]
```

For example:


```
python Alouette_processor.py "C:/Projects/Alouette-1/data/"
"C:/Projects/Alouette-1/root/" "C:/Projects/Alouette-
1/Stage1_instance_12/" "UserX" 12
```

A process_log.csv is produced and saved in the 06_log folder. This is helpful to keep track of processing, and what is accessed by multiple instances to ensure a subdirectory is not processed more than once. It is also useful to analyze the processing performance. If one is processing the entire Alouette-1 image set, be prepared to run multiple instances, and perhaps multiple systems, to bring down the projected processing time required. Processing with 12 instances of Alouette_processor.py on a high-performance machine with a 24-core CPU and 128 GB RAM still took two to three days to process.

After Stage 1 processing is complete, open:

- **03_Post-Processing.ipynb**

In this Jupyter notebook, only run the first section:

Generate processed_inventory, concatenate Stage 1 results - RUN BEFORE starting Stage 2 processing (OCR processing stage)

This will concatenate the Stage 1 results for each Subdirectory, and save it as ~/05_result/result_stage1_raw.csv

5.6 Stage 2 Processing

The Stage 2 processing step is relatively more straightforward. This stage was developed in 2023 to apply a machine-learned text recognition model (keras_ocr) to detect 'num2' type images, and then read that form of metadata.

The keras_ocr model used has a convolutional neural network (CNN) to first locate text on a given image. Then, the text is read with a text recognition classifier. The model's use of the CNN can be a computationally intensive process. Therefore, a GPU version of the code was developed for this Stage, in addition to a standard CPU version. Leveraging a system's GPU enabled this stage to progress significantly faster.

To start Stage 2 processing, run:

```
Alouette_processor2.py [the file path to root directory for your
project] [username (for logging purposes)] [instance number] [batch
size]
```

For example:

```
python Alouette_processor2.py "C:/Projects/Alouette-1/root/" "UserX" 1
8
```

And, to run the GPU version:

```
Alouette_processor2_gpu.py [the file path to root directory for your
project] [username (for logging purposes)] [instance number] [batch
size]
```

Because it is a computationally demanding process, it was found that only 1 CPU instance and 1 GPU instance could be run before system resources were fully utilized – though your system specifications may allow for more instances. The batch size can also be varied based on your system specifications. However, if the batch size is set too large, the program will crash.

5.7 Post-Processing

After Stage 2 processing is complete, open:

- **03_Post-Processing.ipynb**

Now run the remainder of the script, starting from:

Concatenate 'OCR pass' results - START HERE after Stage 2 processing (OCR processing stage)

This step will process the datetime, retrieve the ground station details, among other things, to save the ~/05_result/result_master.csv file.

5.8 Quality Analysis

General quality analysis and statistics can be produced after post-processing, to determine the yield, and other things.

Run:

- **04_Quality Analysis - General**

VII. Future Work

Many things could be done to improve the ionogram data extraction yield, which is currently **43.03%**. Nevertheless, there will always be a limit to the yield because of scan quality, film quality, or quality issues with the original telemetry tape.

At least two improvements to the data extraction method could significantly improve the yield. First, it appears that the Stage 1 processing step is not able to reliably read the 'dot' metadata type, if at all. In the least, the mapping of the relative position of dots to binary metadata in each metadata lane should be checked and recalibrated. Or the entire process ought to be rebuilt to be able to detect and read dot type scanned ionogram images. While dot type images may represent a relatively small proportion of the total set of Alouette-1 scanned images, they occur within the first year of operation of Alouette-1. The early years (1962-65) of Alouette-1 have the greatest scientific interest, as they are the earliest and only topside soundings of the ionosphere made by humanity, until the launch of Alouette-2 on November 29, 1965.

The default keras_ocr text recognition model that was used could be further trained to better read 'num2' metadata numbers, and to only read numbers and not letters. For instance, there were at least 23,984 num2 images that seemed to adhere to the metadata format (i.e: had 15 characters beginning with '10'), but then had certain numbers misread as letters. There are also several instances where an image that is actually of num2 type was not read, perhaps due to scan quality issues. Further training of the text recognition model could improve the reading accuracy.

As described, the data extraction method occurs in two stages – to first run the old data extraction algorithm, and then apply the new text recognition stage to read num2 metadata. The algorithm should be refactored such that the metadata type can be detected initially, and then read according to its type. This refactoring and rebuilding of the code base could resolve certain long-standing issues and may improve yield.

Furthermore, it can enable the algorithm to be more adaptable to be able to read Alouette-2, ISIS-1 and ISIS-2 ionograms.

Yet with some modifications, ISIS-1 and ISIS-2 ionograms seem to have been read successfully.

Prepared by: Ravendra Naidoo, Roksana Sheikholmolouki, Ashley Ferreira, Marianne Fortier, Jackson Cooper, Benjamin Cannings, Émilie Fillion

Date: August 23, 2023