Scanning Maps: Quantifying Errors to Inform Future Image Capture Efforts

Stanford University Libraries |G. Salim Mohammed | Digital and Rare Map Librarian | Luminary Class, ARL-LCDP, 2011-12 | June 23, 2012

ABSTRACT

In 2009, Stanford began an effort to scan its maps. Scanning large format items such as maps create a multitude of challenges. One of these challenges is to capture the map with specifications that meet all known repurposing needs. A prominent repurposing need is to ensure that the map can be consumed in a Geographic Information System (GIS). A team of Stanford University Library staff consisting of Patricia Carbajales, G. Salim Mohammed, Matt Pearson and Renzo Sanchez-Silva (noted here in alpha order) along with student assistants, conducted a detailed study of a Russian Topographic scanned map where details were visually inspected and checked for scanning errors. Data that was gathered helped us make recommendations for future scanning. The study was conducted earlier this year (2012). While the team continues to monitor scanning of maps at Stanford and this research is a work in progress, this poster focuses on the rationale of the initial identification and study of the problem, the study and results and recommendations. The work is a collective between all members of the team and much of the content domes from the study.

Rationale and Problem Identification

Why do we want to scan maps?

- Often, especially with historical data, the information is only available on the map.
- A scan allows us to archive the map digitally.
- Maps are increasingly scanned so that they can be accessed digitally, including on the web.
- Maps scanned with certain specifications can be, after further preparation, can be made GIS-ready.
- GIS-ready scans allow the user to layer one map on top of another.
- GIS also allows us to manipulate and analyze the data spatially.
- A process called *vectorization* allows us to extract features and create polygon layers. This allows us to also create three dimensional versions of features such as mountains and oceans; it also allows us to rasterize the vector lines for computer aided printing and processing. Rasterization converts vector layers such as contours showing altitude into continuous surfaces. Each layer can be manipulated differently.

Having scans that allow us to *vectorize* the information on the map, turned out to an important specification.

Scanning Specifications

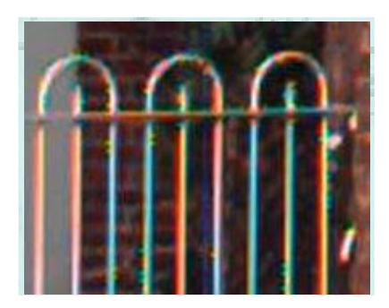
Therefore there were three primary specifications related to scanning:

- In general, items needs to be scanned keeping in mind all possible purposes.
- Scanning in color for image manipulation programs to distinguish between different data on the same map, i.e. to be able to ask the program to focus on the blue lines (rivers) and to create a vector file of rivers alone; besides scanning in color means the best possible representation of the object, without loss of information.
- Scanning also needs to be conducted in a high enough resolution so that the details are visible and distinguishable by image manipulation software.

Scanning Defects

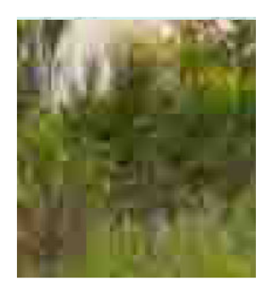
There are several defects that can appear while scanning. A few possibilities are illustrated below:

Banding


Bayer Pattern

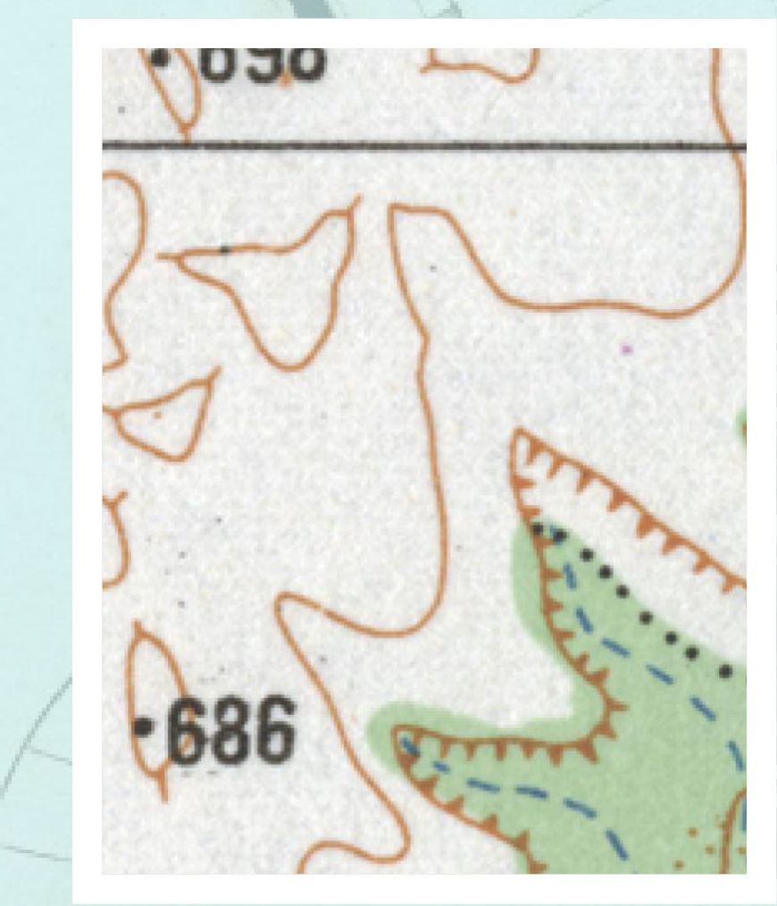
Moire

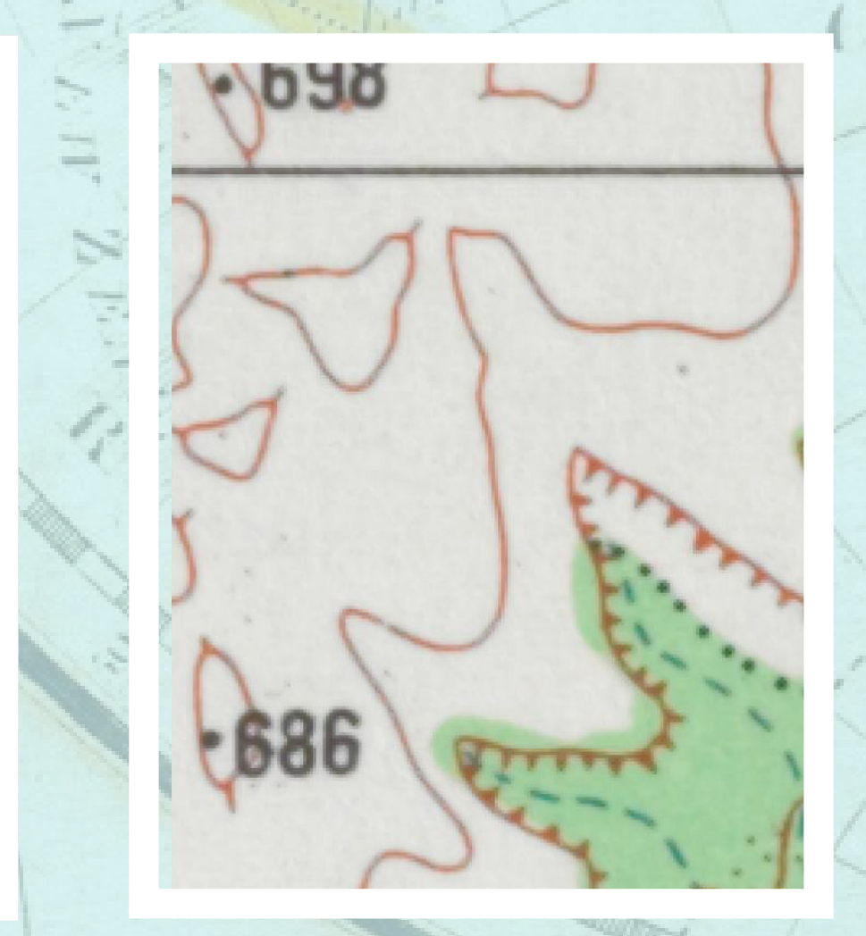
Posterization


Compression Artifacts

Discovering the defect

In October 2011, during normal quality control processes, it was discovered that topographic maps printed in the early 1900s that we were scanning had a defect—red lines were turning to grey or had missing links. This defect, collectively called the *Bayer-Moire* defect, affected approximately 4,000 maps and a determination had to be made to quantify the extent of the defect when compared to non-defective scans. We also had to determine how to scan in the future so that the defect disappears given the limits of our resources.


Non Defective Scan


Defective Scan. Note the orange lines have turned into red and grey.

Studying the Problem

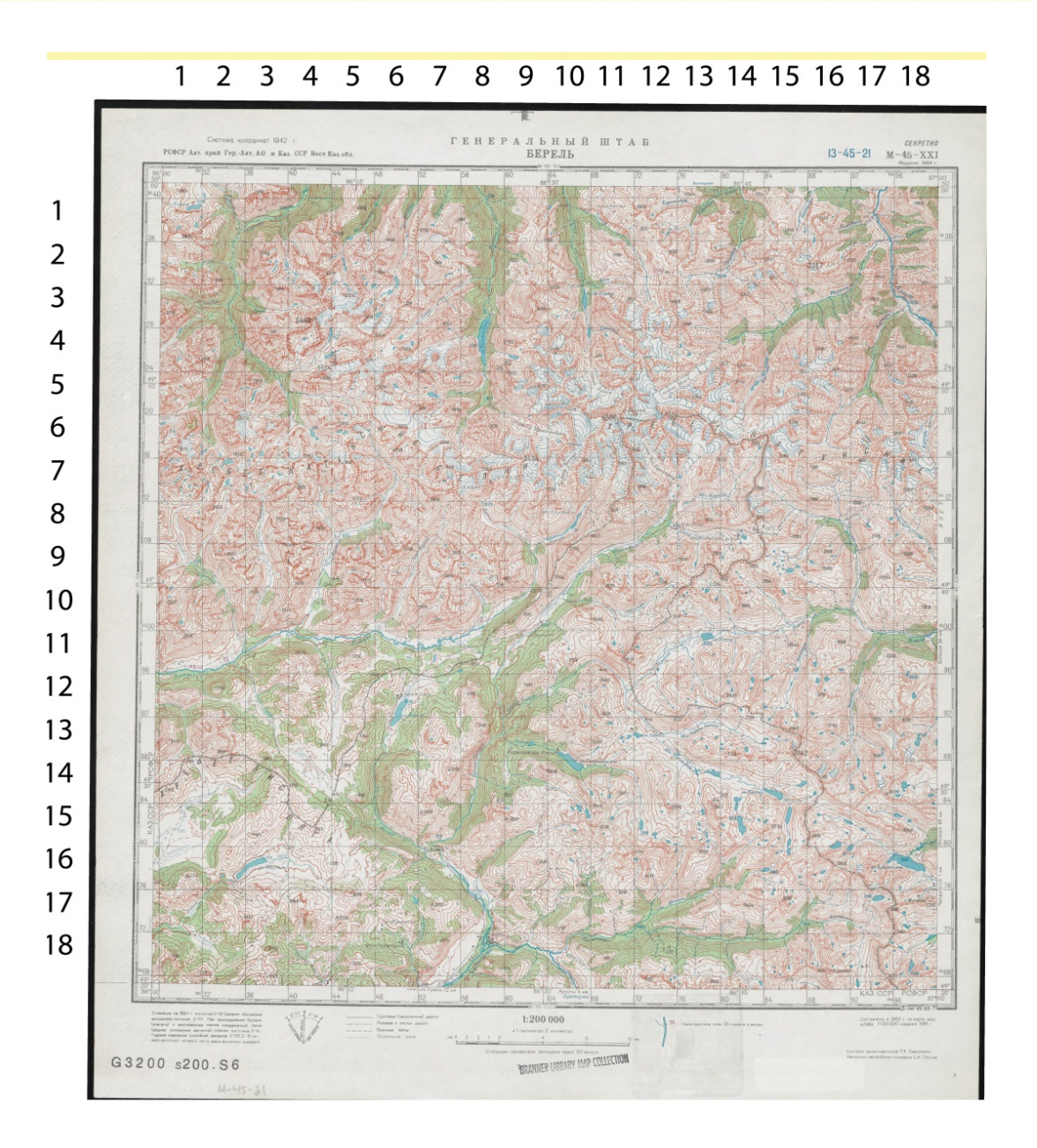
Selecting Versions to study

In order to know how defective a scan was , one had to measure the number of defects on the scan. The quality of defect was considered, but was harder to quantify, so we chose to focus on the sheer number. Also we needed a couple of other scans to compare the defective scan to. The defective scans were being scanned at 330 pixels per inch (PPI) , so we decided to make the following derivatives of the *same* map scan.

1. A 330 ppi version with the defect, using a P65 scanner.
2. A 330 ppi version without the defect, using a flat bed scanner.
3. A 400 ppi version without the defect, also using a flatbed scanner.

A defective map was chosen as is illustrated in the next column.

Studying the Problem



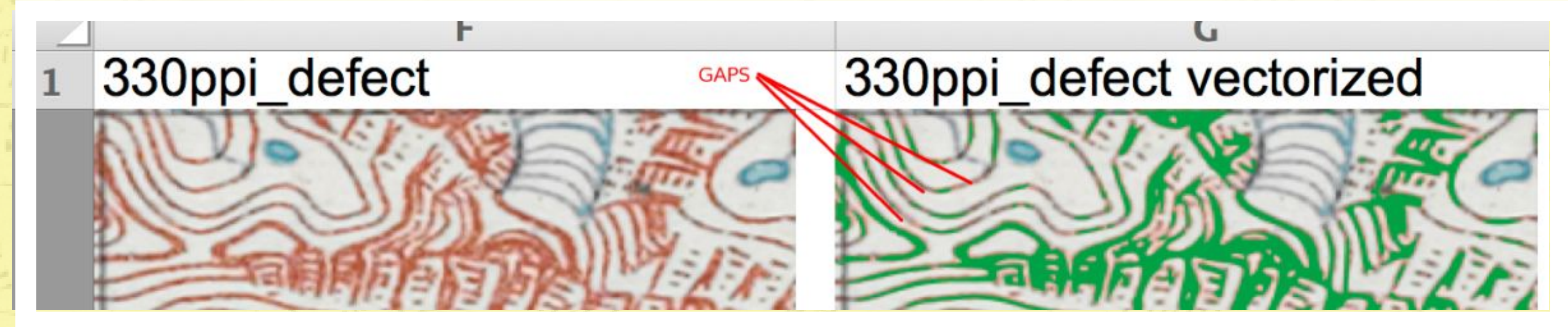
The chosen map (shown on the left) was then gridded, and then the three versions of the maps were vectorized.

Essentially, the red lines you see in the map in the next column were isolated and thickened. The red lines represent contours at different intervals depicting a certain height. These layers could help us create a three dimensional digital elevation model or DEM.

The tiles , both the original map and the vectorized version were then randomized using the below online tool at researchrandomizer.org

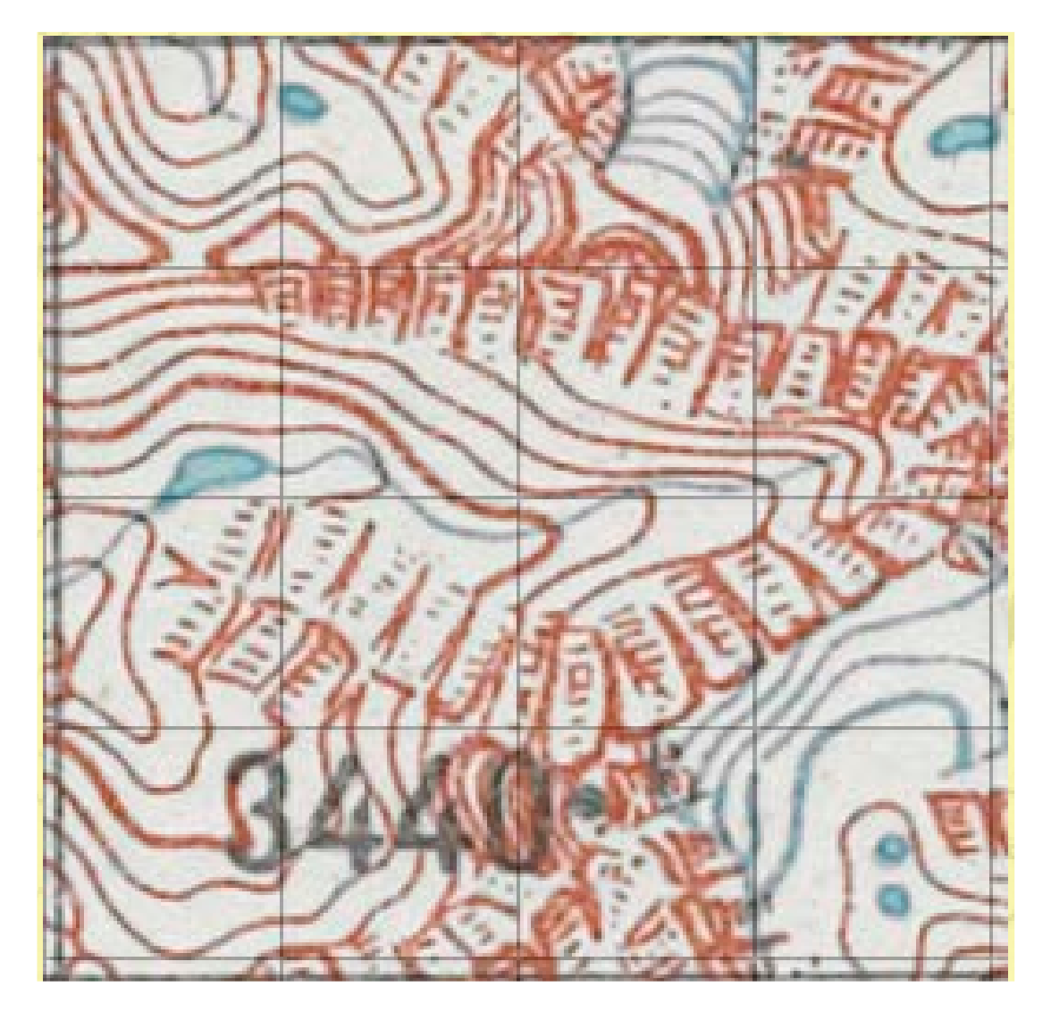
Six out of the 17 sets randomly generated above were used for the study. Details of what was studied and compared are highlighted below.

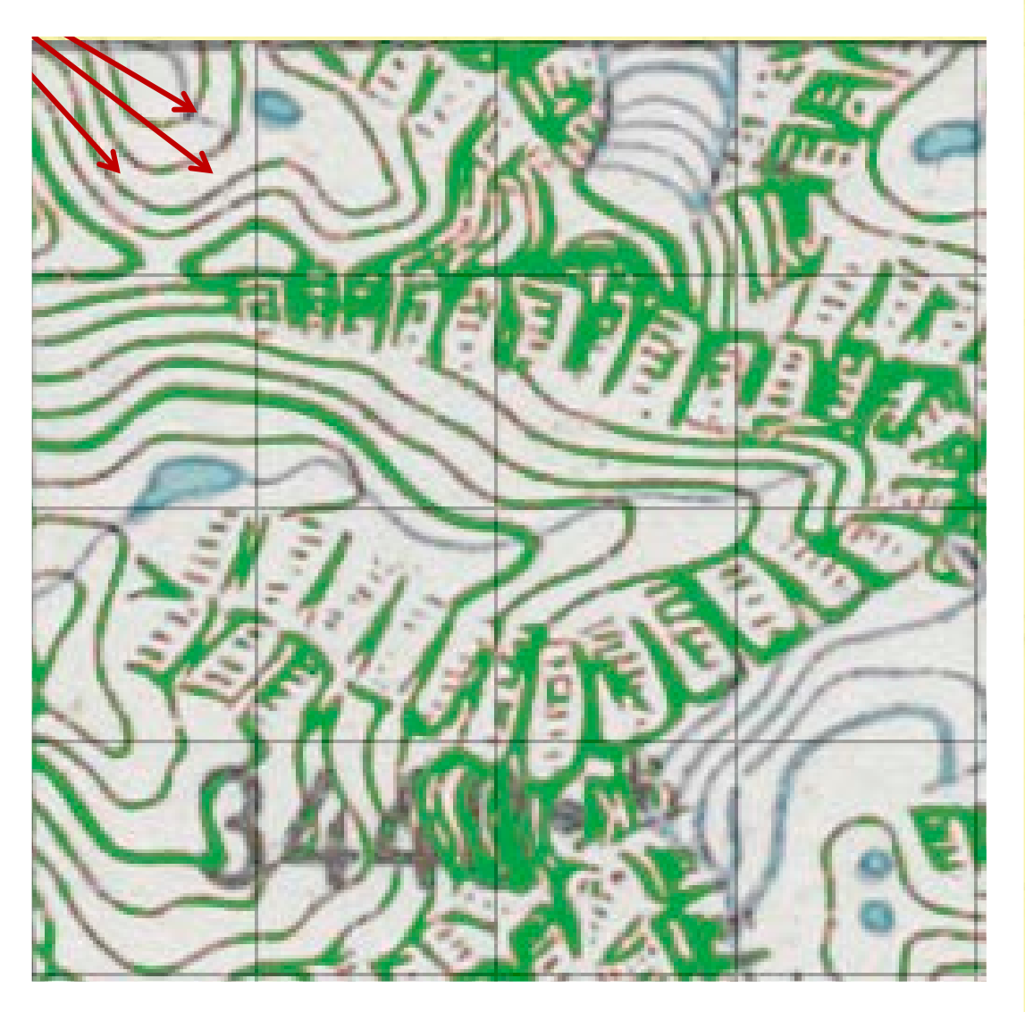
Vectorization results:.
Below you see a defective part of the map at 330 ppi vectorized. On your left is the map as scanned, on your right is you see the red lines exaggerated, making the "GAPS" clear. Another defect is called a bridge, where a line spills over an area that is it not supposed to.



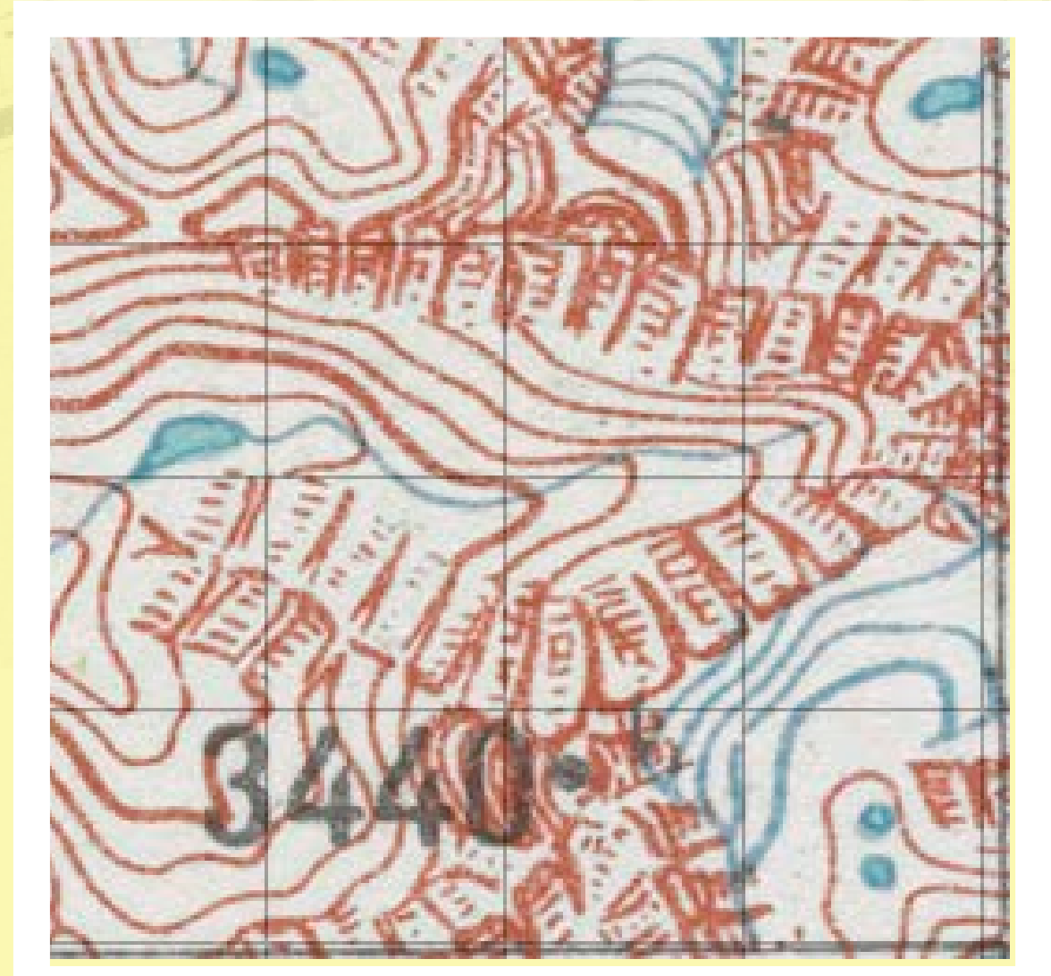
1	330ppi_defect	330ppi_defect vectorized

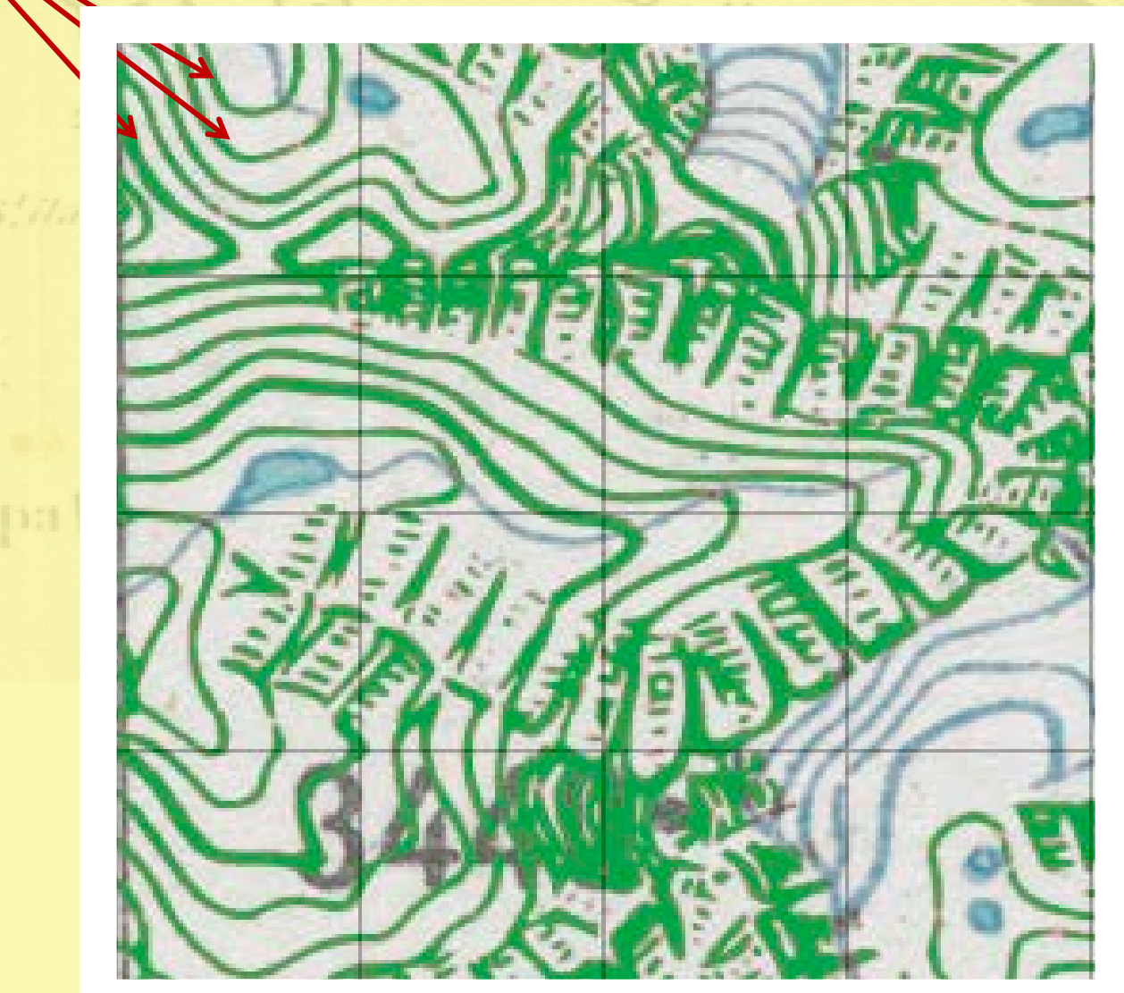
Sample tile that was visually examined and studied:


Original map scanned at 330 ppi, **with** defect.



Defective map, Vectorized, 330 ppi


Original map scanned at 330 ppi, **without** defect.



Defective map, Vectorized, 330 ppi **without** defect

Results and Recommendations

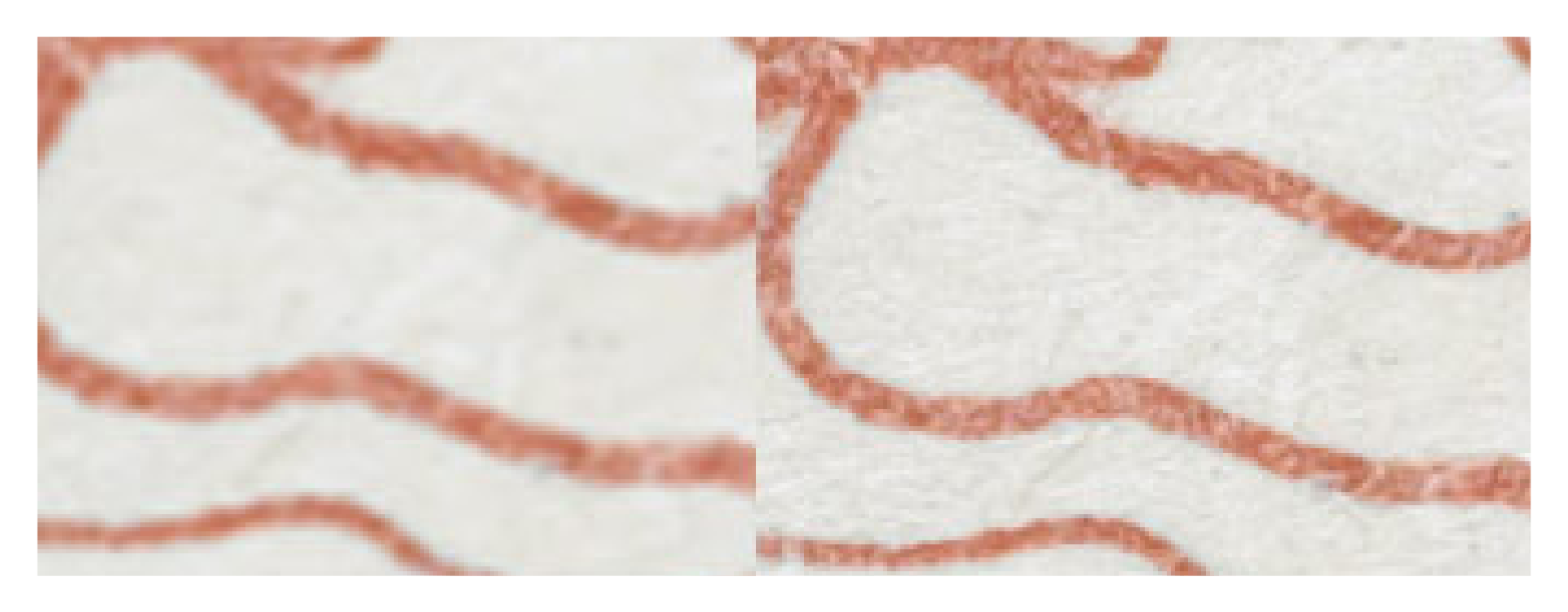
Visual examination and notation of each kinds of errors were tallied and summarized. The chief finding of the study was as follows:

1. Defective 330ppi files had 3.25 times as many vectorization errors as the defect-free 400ppi version.
2. The defective 330ppi file had 4.89 times as many vectorization errors as the defect-free330ppi version.
3. 400ppi image files had 1.64 times as many vectorization errors as defect-free 330ppi version

4. **Defective 330ppi file had 4.36 times as many defects as defect-free 330 ppi version.** This is an important finding as 330 ppi was the standard we were using prior to this study for affected topographic maps.

330 ppi versus 400 ppi, performance and noise

Interestingly the defect-free 330 ppi vectorized files produced fewer errors than the 400 ppi files, which is counter-intuitive. One would think that a file of higher resolution would have few errors. What is happening is that the 400 ppi file is recording more information which is competing with the contour like information. The data present in the contour line s is what we are interested in capturing, but there is easy way to distinguish it from "noise." The smaller file at 330 ppi (below, image on the left) has an ample number of pixels to record the contour lines, but the 400 ppi image (below, on the right) also has detail that records paper texture and nuances of the ink and the fiber of the paper. The below images illustrate



Legacy files Versus New Scans

Approximately 4,000 files were considered defective. We outlined the following options for these files:

1. Process on-demand. Should a student be interested and requests vectorization, existing map image files are vectorized using the Potrace method described below; requests will require 2-3 hours of students and the Geospatial Manager for preparing each individual file for automatic vectorization using Potrace-based documentation provided by the GIS Software Developer. The Potrace method is one way to digitally manipulate the image to reduce the impact of the error in scanning.

2. Re-Image. Re-imaging the entire map collections will require approximately 10 weeks of dedicated studio time with 1 full-time (40hr/wk) photographer and 2 part-time (19 hr/wk) staff.

3. Batch processing. It was approximated that this will require 3-4 days of developer time (24-32 hrs) and result in additional overhead for stewarding vector files that will need to accompany the map image files through accessioning.

Recommendation: Given the low demand for vectorization and the enormous costs associated with re-imaging and batch processing (which would result in some level of defective vectorization), the decision was made to notate that these files have an issue. We recommended that we looked into what it would take to streamline a process-on-demand model.

Future Scanning

Given current resources including limitations on the hardware (the camera), it was determined that the map, when necessary be split into small tiles or sections and scanned at 400 ppi and then digitally stitched. The study therefore recommended that we photograph the Topographic maps in 4-6 parts at 400ppi with the currently installed PhaseOne P65+ scan back (kind of camera) and then stitch images in Photoshop. The Bayer Moiré defect is minimized and its impact on successful automated vectorization of the topographical map contour lines is also minimal. Other options which involve hardware upgrades and corresponding funding needs are available and discussed in the report. We recommended that we will not continue to produce defective 330ppi image files.

Note that further study and recommendations regarding hardware options were extensively studied by Matt Pearson.

Many thanks to the team who worked on the project. I also want to thank Tom Cramer, Deardra Fuzzell, Jane Ingalls, Michael Olson, Julie Sweetkind-Singer, Stu Syndman and Wayne Vanderkull, who played other roles in this study. Special thanks to Julie Sweetkind-Singer and Bob Schwartzwalder for making it possible to complete my tenure at the Association of Research Libraries Career and Leadership Program and to Paula Mochida formerly of the University of Hawaii at Manoa for sponsoring me in the first place.

For the encouragement, support and coordination by Mark Puente – many many thanks.

And to Jon Cawthorne for helping me with my transition to Stanford, job offer to acceptance!

Last but not least, thanks go to the members of my Luminary Class – specially to Lisa Shiota and Nikhat Ghouse who helped with feedback on the poster and for general support. And for Alanna Aiko Moore for taking care of a few details!

Shine On Luminaries! You Make Me Proud.

Background Map: The World. G. Boynton Sc. Published by Lincoln & Edmands. (1832). View it at: http://bit.ly/BoyntonMap Courtesy davidrumsey.com