



An algorithmic interpretation of historic cosmic ray recordings

VWF Mattana

 **orcid.org 0000-0001-8551-1157**

Dissertation submitted in partial fulfilment of the requirements
for the degree *Masters of Science in Computer Science* at the
North-West University

Supervisor:	Prof GR Drevin
Co-supervisor:	Dr RD Strauss

Graduation May 2018
21128707

PO Box 141
MTUNZINI 3867
15 November 2017

TO WHOM IT MAY CONCERN: DISSERTATION BY VWF MATTANA

This is to declare that I have done the editorial corrections to the dissertation :

"AN ALGORITHMIC RETRIEVAL OF HISTORIC COSMIC RAY RECORDS"

by VWF Mattana.

I am satisfied that when the corrections are made, the dissertation will be in an editorial state of adequate standard to be submitted for the Masters degree.

A handwritten signature in black ink, appearing to be 'B. Spoelstra', with a stylized, cursive script.

(Prof. Em.) B. Spoelstra

Abstract

The aim of this study is to make additional GLE (ground level enhancement) data recordings from the 1940s and 1950s available, as well as generating of a verified, accurate synthetic ground truth image of historic cosmic ray ionization data recordings. These images can subsequently be used to test data extraction algorithms. The ground truth image is a reference point from which measurements can be made, concerning the fitness of an algorithm. In this study, the ground truth image will be constructed from properties of the original image. This ground truth image will represent the data captured in the original image, and as such can be used to test data extraction algorithms on an image representative of the original data set, but with one major difference; the ground truth image is known, and reproducible. It is not a perfect replica of the original but rather a similar image where only the properties of interest are accurately described. A synthetic version of the ground truth image must to some degree, reproduce the distortions and optical artefacts present in the original image. The creation of such an image is possible when using image processing techniques, such as binarization, correlation, morphology, segmentation, amongst other techniques. The detection algorithm is tested on the synthetic ground truth image to evaluate the detection capabilities of the algorithm, using measures such as MSE, precision, accuracy and recall.

The ionization data of GLE #5 is extracted from the recordings of three stations, viz; Godhavn, Cheltenham, and Christchurch. The ionization data is then converted to percentage increase above background cosmic ray levels, for comparison to existing neutron monitor data. The neutron monitor data is sourced from a GLE database, as well as using Forbush's interpretation of the event.

Keywords: Historic cosmic ray recordings, image processing, ground level enhancement event, ground truth image.

Acknowledgements

I wish to acknowledge the following people for their contribution to this work:

Prof Harm Moraal,

Prof Ken McCracken,

Prof Günther Drevin,

Dr Du Toit Strauss.

Dedication

To:

Mom and Dad,

Quintin Robbetze,

Prof Günther Drevin,

Prof Harm Moraal.

Thank you all for all the love, support, patience, inspiration, and coffee. I would not have been able to do it without you.

Contents

1	Introduction	1
1.1	Aim	1
1.2	Background	2
1.3	Research question	3
1.4	Research methodology	4
1.5	Outline	4
1.6	Scope	5
2	Cosmic Rays	6
2.1	Cosmic rays	6
2.2	Ground level enhancement events	11
2.3	Neutron monitors	13
2.4	Model-C recorder details	16
2.5	Photographic film details	20
2.6	Interpretation of ionization data	24
2.7	Conclusion	24
3	Image Processing	25
3.1	Document image processing	26

3.1.1	Textual processing	26
3.1.2	Graphical processing	27
3.1.3	Chart image processing	27
3.2	Binarization	28
3.2.1	Evaluations of binarization techniques	31
3.3	Morphology	32
3.3.1	Erosion	32
3.3.2	Dilation	33
3.3.3	Opening	34
3.3.4	Closing	35
3.3.5	Watershed	35
3.4	Segmentation	36
3.5	Labelling	38
3.6	Hough transform	38
3.7	Correlation	39
3.8	Skew detection	40
3.9	Evaluation	40
3.9.1	Ground truth	40
3.9.2	Mean square error (MSE)	41
3.9.3	Accuracy and Recall in Image Processing	41
3.10	Conclusion	43
4	Method	44
4.1	Development of the algorithms	44

4.1.1	Process	45
4.1.2	Measure of success	46
4.1.3	GLE #5 data	46
4.1.4	Prior knowledge	47
4.2	Image processing	49
4.2.1	Sprockets	50
4.2.2	Pre-processing	51
4.2.3	Binarization	51
4.2.4	Hour markers	52
4.2.5	Scale lines	54
4.2.6	Data lines	54
4.2.7	Ground truth image	55
4.2.8	Inpainting	56
4.2.9	Synthetic image	56
4.2.10	Accuracy, precision, recall and MSE	56
4.3	Extraction of data from original image	57
4.3.1	Binning decisions	57
4.3.2	Conversion, sampling and quantization	57
4.3.3	Comparison	58
4.4	Conclusion	61
5	Results	62
5.1	Image processing	63
5.1.1	Ground truth	64

5.1.2	Synthetic image	65
5.1.3	Ionization data lines	66
5.1.4	Accuracy, precision, recall and MSE	67
5.2	Conversion	68
5.2.1	Ionization data binning	68
5.2.2	Percentage increase	69
5.3	Comparison	70
5.3.1	GLE #5	70
5.3.2	Forbush	71
6	Discussion	72
6.1	Input images and user edits	72
6.2	Ground truth and synthetic image	72
6.3	Accuracy, precision, recall and MSE	73
6.4	GLE data	74
6.4.1	Calibration constants	76
6.4.2	Ionization chamber inaccuracies	76
6.4.3	Cosmic ray anisotropy	76
7	Conclusion	77
7.1	Conclusion	77
7.2	Future work	78
8	Glossary	79
A	Appendix	81

Chapter 1

Introduction

1.1 Aim

The aim of this study is to test algorithms for the extraction of data relating to a short lived event, for which there is comparable scientifically approved data available. The creation of a synthetic ground truth image is a secondary aim, which is required to fulfill the primary aim of the study. This study is part of a larger study to extract and recover over 25 years of historic cosmic ray data recorded from 1935 until 1960. This study is an extension of the work done by Drevin (2008), Du Plessis (2010), Mattana *et al.* (2014), (Mattana and Drevin, 2017) and Mattana *et al.* (2017).

The short lived event of which the data is to be extracted is specifically the Ground Level Enhancement (GLE) event which occurred in 1954 (GLE #5), and was recorded by the Model-C recorder, for which we have the recordings in a digital format. A digital image processing algorithm was created to automatically extract the data contained in each scan of the photographic paper originally used to record the cosmic ray data. The extracted data will be compared with existing data for GLE #5, to test the feasibility of recovering as much of the cosmic ray data as possible. This data can be used in both long term space weather research, with the focus on weekly and monthly averages, as well as solar activity research, which is more focused on high resolution event data.

1.2 Background

This section will provide the reader with an orientational overview of the topics to be discussed as well as the outline of the study itself.

Cosmic rays

Cosmic rays are charged particles originating mainly from outside of our solar system, as well as particles from the Sun. Today we know that cosmic rays are high energy particles, specifically, 89% protons (hydrogen nuclei), 10% alpha particles (helium nuclei) and 1% nuclei of heavier elements (Simpson, 1983). These particles are constantly entering Earth's geomagnetic field, and interacting with our atmosphere.

Ground level enhancements

Sometimes, the solar particles will have enough energy to penetrate the atmosphere, and interfere with the readings of cosmic rays, as well as electronics on ground level. These occurrences are rare, and are named Ground Level Enhancement (GLE) events.

Neutron Monitors

The neutron monitor is a device for detecting and recording the neutrons which are secondary particles released from collisions between cosmic rays and atmospheric particles.

Model-C recorder

The recorder used by Compton *et al.* was named the Model-C recorder, and was implemented at stations around the world, in 1934. These stations recorded for over 20 years, and produced many rolls of photographic paper, with ionization data, temperature, and pressure data recorded. The recorder used an ionization chamber to detect the incoming cosmic rays, specifically recording the changes in the ionization rate. Seven of these recorders were placed around the world, and recorded for over 25 years.

Recordings

The photographic paper that the image is recorded on can contribute to the graininess of the image, and any imperfections on the material will be carried over into the digital format. The process of scanning these photographic papers (Figure 1.1) was undertaken by a private contractor for the National Geophysical Data Center (NGDC) in Boulder, Colorado. The images measure approximately 6000 by 905 pixels, have a pixel depth of 24 bits, and document a 19 hour period on average.

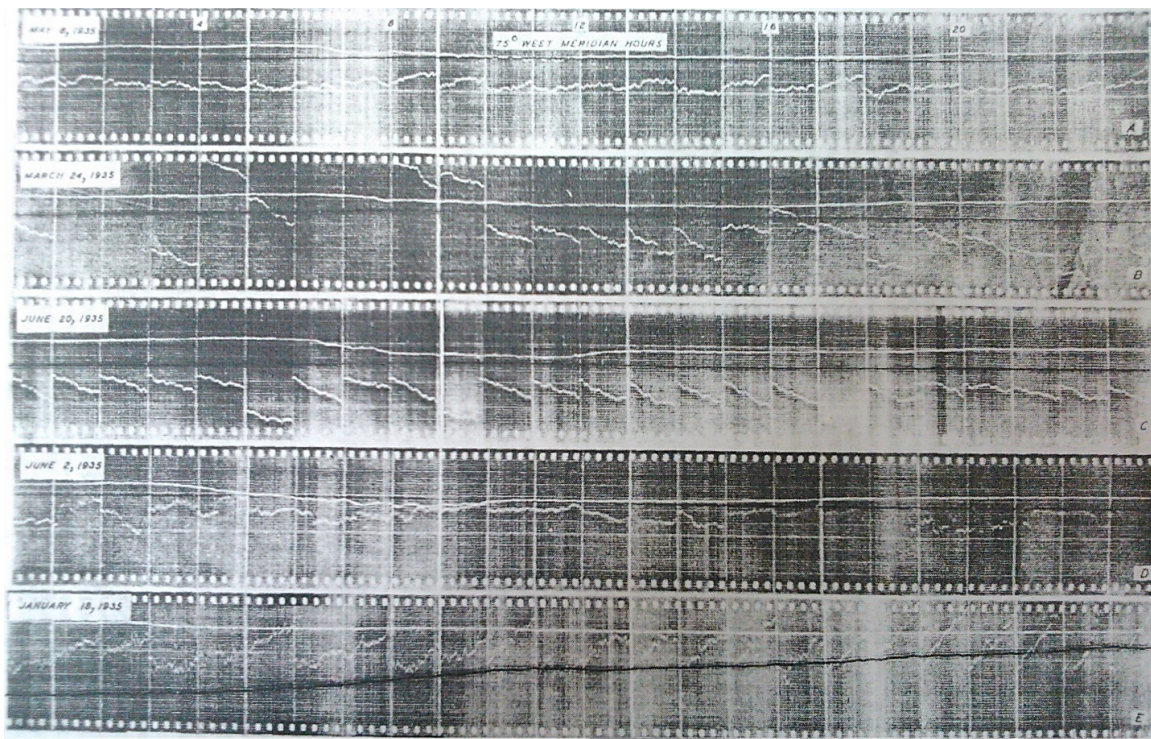


Figure 1.1: Examples of the photographic paper used by the Model-C recorder.

1.3 Research question

The research question can be formulated as follows: “With what level of accuracy can an automatic extraction/conversion algorithm extract the data from the photographic records produced by the Model-C recorder?”. This question describes the goals of the study, which are to accurately extract the ionization data, and convert it to cosmic ray counts. These values will be compared to the existing values for GLE #5 recorded by neutron monitors and available in a GLE database (Usoskin, 2016), as well as the hand extracted data from Scott Forbush (Forbush,

1958).

1.4 Research methodology

The research was approached from a quantitative, algorithmic, school of thought. The qualitative aspect of this research is minimal, as the data is already collected, and reflects a physical phenomenon, which is best described in the language of mathematics. The production of an artifact is a secondary result of automating the extraction and conversion process. The following artifacts are produced in the course of this study:

1. Algorithms to create ground truth and synthetic images.
2. Algorithms for the extraction of cosmic ray intensity readings made on the Model-C recorder.
3. The data extracted by using the algorithms on GLE #5 recordings.

The method used for this study can be described as the hypothetico-deductive method (Dodig-Crnkovic, 2002), which involves the creation and testing of a hypothesis to further the goal of making a firm deduction.

1.5 Outline

This study will be divided into the following chapters:

1. Introduction - The current chapter; provides context and overview of the subject matter.
2. Cosmic rays - A chapter devoted to the discussion of the literature on the technical details of cosmic rays and the research regarding them.
3. Image processing - A chapter focused on the discussion of literature on image processing, and the techniques and relevant research.
4. Method - Provides the technical details on the workings of the algorithm(s) created to automate the extraction process.

5. Results - Provides an overview of the results obtained from the algorithm, as well as a comparison to neutron monitor data.
6. Discussion - A discussion on the results, and their broader scientific context.
7. Conclusion - Closing statements, and an overview of the study, including recommended future work.

1.6 Scope

The goals of this study are clearly defined; find the percentage increase of the observed cosmic ray counting rate as accurately as possible for GLE #5 on the 23rd of February 1956, using an automated algorithm, as well as producing a synthetic ground truth image to test other data extraction algorithms. The automated algorithm will be written in Matlab, and will run as efficiently as possible, although this is not a primary concern. The primary concern is the accuracy of the observed cosmic ray counting rate extracted. There are other data lines on the images, but these are not important when investigating short term events such as GLEs and it is thus beyond the scope of this study to extract the temperature and barometric pressure data. User input is a valuable tool, although the algorithm will detect all aspects automatically if they are consistent in size and intensity. Cases exist where, due to the degradation of the photographic paper, or scanning process, the properties to be extracted cannot be detected. The inclusion of a user edited duplicate of the input image ensures accuracy even when there is damage to the film. The user input image is not required for the algorithm to function, and on good quality images, no user input will be necessary.

This study will be referencing image processing techniques, ranging from basic, to more advanced algorithms. More images of the photographic paper exist, which almost continuously document the period of 1934 to 1960. The other images, which do not contain details of the GLE #5 are not relevant to the aims of this study, as there is little to no way to verify the data contained in these images. Long term cosmic ray studies might be interested in these films, and with a bit of adjustment, the algorithm detailed in this study can convert these images to cosmic ray counts as well. However, the algorithm has not been optimized for such a process and the temperature and pressure data which might be relevant to such studies is not extracted by this algorithm.

Chapter 2

Cosmic Rays

This chapter deals with the necessary background, context, and research relevant to this study. cosmic rays will be discussed, followed by GLEs. Neutron monitors, and ionization chambers are then discussed, focusing on the Model-C recorder. The photographic paper used in the Model-C recorder is then discussed, as well as how to interpret the data on this photographic paper.

2.1 Cosmic rays

Cosmic rays are charged particles that are continuously entering Earth's geomagnetic field, and interacting with our atmosphere. As they interact with the atmosphere they produce a cascade of secondary particles, which are labelled as nuclear fragments in Figure 2.1. The cosmic rays have been shown to be partially responsible for lightning, as well as disrupting electronics on Earth. The extra-terrestrial nature of cosmic rays was discovered by Victor Hess in 1912, when he used hot air balloons to reach altitudes high enough to confirm the existence of ionized particles entering the Earth's atmosphere. It was assumed that these charged particles were beta-rays. The existence of cosmic rays were confirmed by W Kolhörster in 1913. These results were contested and ultimately refuted by an European board of scientific authorities. Millikan confirmed their existence in 1926, but claimed to have discovered cosmic rays himself, leading to strong reactions from V. Hess and W. Kolhörster (Carlson, 2012).

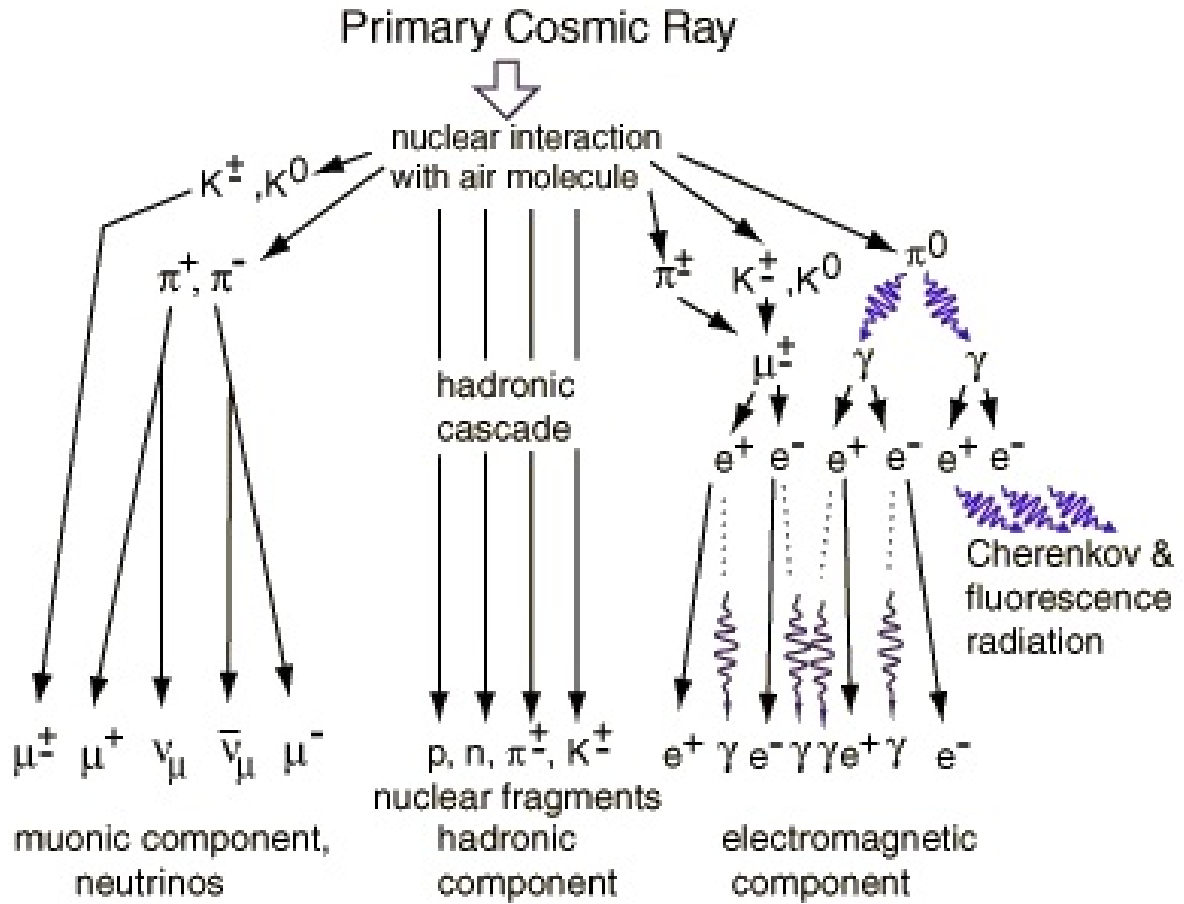


Figure 2.1: The process which leads to the production of secondary particles (Nave, 2006).

In 1963 the space age began with the launch of Sputnik 1, and a whole new era of cosmic ray detection became available. The availability of satellites allowed detectors to directly measure cosmic rays instead of measuring the secondary particles (Figure 2.1) created by cosmic rays. With an increased understanding of particle physics, neutron monitors were developed in the 1950s, and put into widespread use around the world. This coincided with the development and use of computers, and as such, the data recorded by neutron monitors is mostly in a digital format. However, in the flurry of technological advancement, it seems that the data recorded by the Model-C recorder (100+ station years) was stored in cardboard boxes, tucked away with other research materials of the late Professor Forbush, and forgotten. Today we have even more ways of measuring cosmic rays, from the IceCube particle detector at the South Pole, to satellite based detectors. Thus, in the past 100 years, our knowledge of cosmic rays has grown from discovery to a fertile field of study, with many implications. Due to the nature of the charged particles that make up cosmic rays, it is very difficult to pinpoint their original source.

However, cosmic rays are hypothesised to originate in pulsars, supernova, and radio galaxies. The origin of cosmic rays is an important aspect of contemporary research.

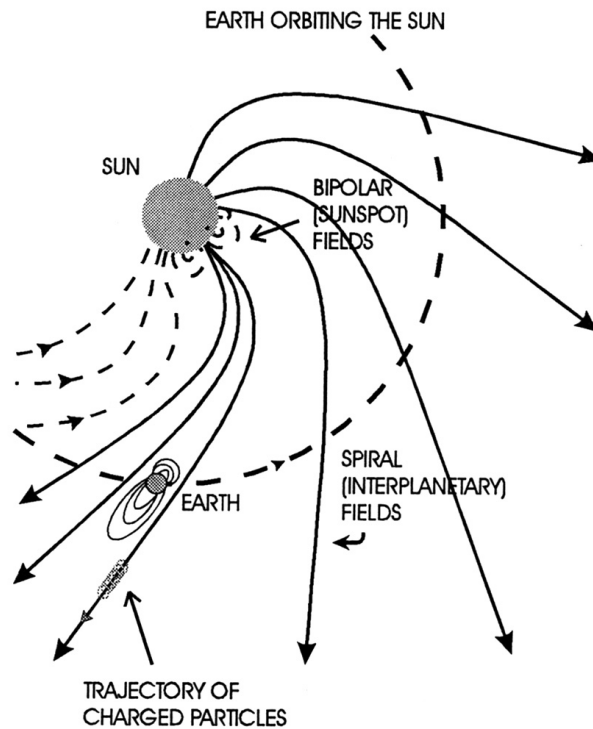


Figure 2.2: The interplanetary magnetic field, indicated by the spiral around the Sun.

Cosmic rays are influenced by magnetic fields, prominently the interplanetary magnetic field (IMF) as they enter our solar system. The IMF is caused by the Sun's magnetic field which is propagated outward by the solar wind (Figure 2.2). The solar wind is an extension of the solar corona, comprised of fast ionized gas (plasma), constantly emitted from the Sun's surface at 400 - 800 km/s (Potgieter, 2008). The plasma is a collection of charged particles which can 'store' a magnetic field. These magnetic fields can extend up to 120 AU away from the Sun, taking on a spiral form. An astronomical unit (AU) is defined as the distance between the Earth and the Sun. The IMF has an effect on charged particles, including cosmic rays, changing their direction, and trajectory, thus obscuring their origin. This causes the intensity of the cosmic rays to vary within our solar system as solar activity varies, and is known as the heliospheric modulation of cosmic rays.

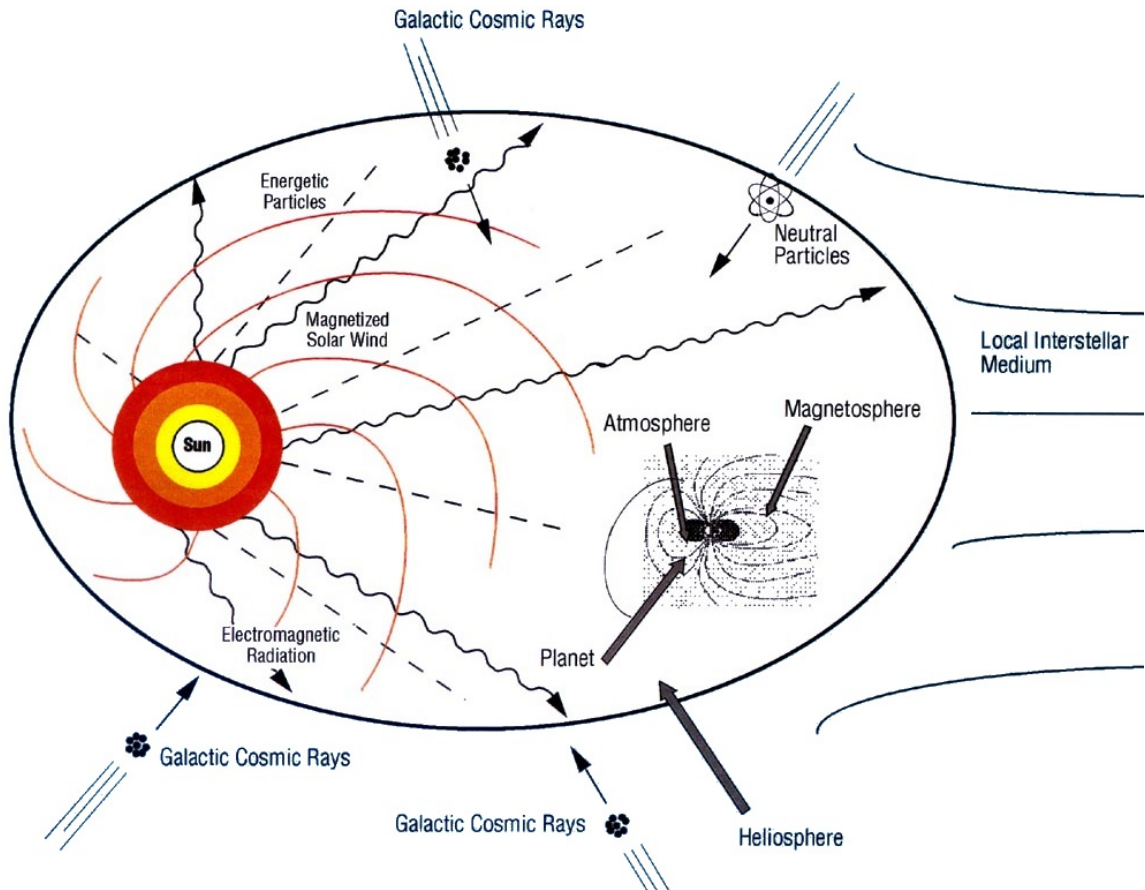


Figure 2.3: Diagram showing the context of cosmic rays entering the heliosphere.

The diagram in Figure 2.3 shows our solar system and the heliosphere, giving some context to the nature of cosmic rays. Due to the IMF, the direction from which the cosmic rays enter our atmosphere is not evenly distributed. This directional property is called anisotropy, and is due to the complex interaction of charged particles with a magnetic field. As a result of the interaction between our atmosphere and the cosmic rays, we can only record them first-hand while outside the atmosphere. When cosmic rays enter the atmosphere, they invariably collide with air molecules. This results in the discharge of secondary particles, which can be detected at ground level by detectors (neutron monitors or otherwise).

However, it is fortunate that the ratio of secondary particles to high energy particles remains constant (Anchordoqui *et al.*, 2004). Early cosmic ray researchers were under the impression that the Sun also released cosmic rays, as there is a correlation between counts made by ground based cosmic ray detectors and the occurrence of sunspots. It appeared as if the presence of a sunspot caused a sharp increase in the detection of cosmic rays over a short time period (2-3 hours). Today we know that there is a more complex interaction based on the solar wind,

IMF, and Earth's own magnetic field. cosmic rays have been linked to the seeding of clouds (Svensmark, 2007), as well as setting off the chain reaction which leads to lightning (Gurevich *et al.*, 1999). Cosmic rays are also important in carbon dating, as the formation of Carbon-14 is a result of incoming cosmic rays (Lingenfelter, 1963). Solar minima occur when the Sun is least active, with few observed sunspots. It was shown that cosmic ray levels peaked in the years; 1956, 1965, 1976, 1987 and 2009. These dates occur within solar activity minima, (Figure 2.4).

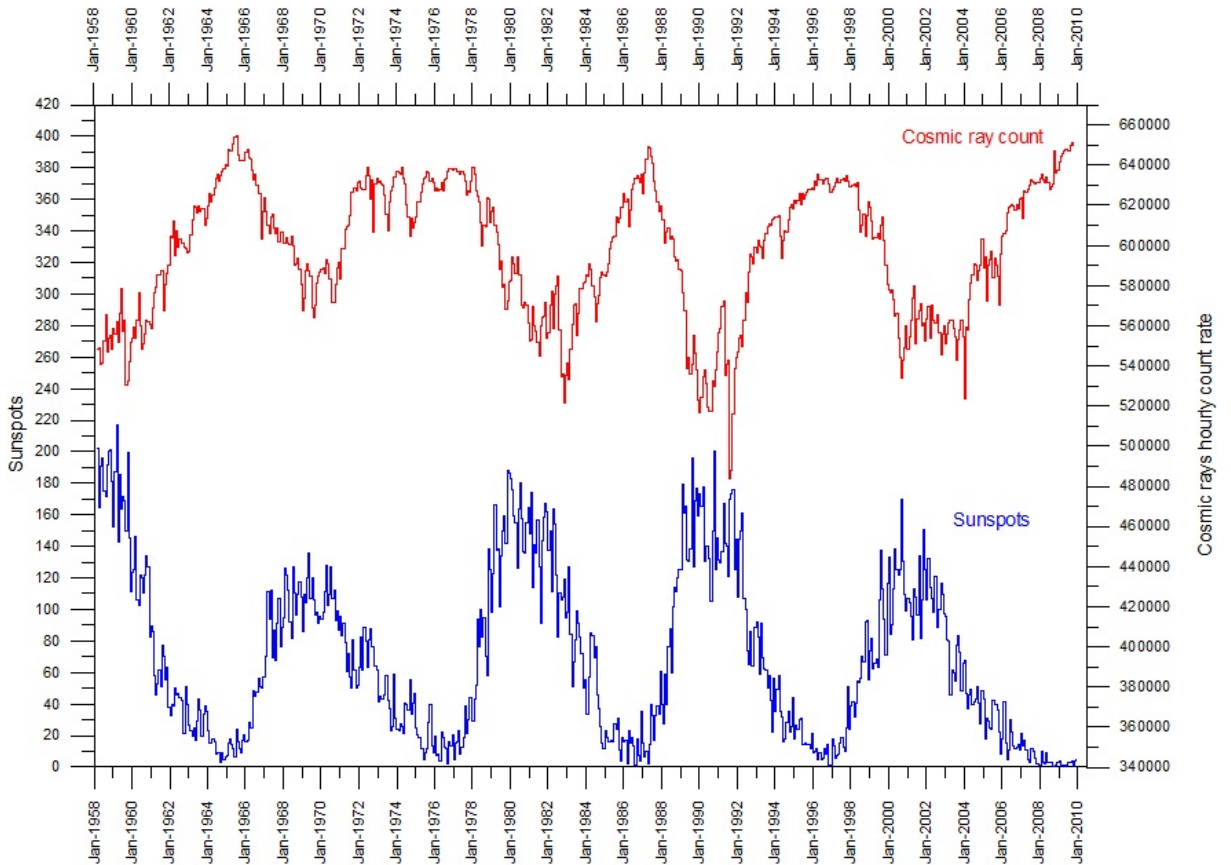


Figure 2.4: Diagram of solar activity vs. cosmic ray recordings.

During transient solar phenomena, such as solar flares and coronal mass ejections (CMEs), the Sun can accelerate charged particles to cosmic ray like energies. These particles are referred to as solar energetic particles (SEPs). SEPs can interfere with cosmic ray recordings, if they are energetic enough, and are worth investigating due to the damage they could potentially cause, as well as leading to further insight in the study of space weather, and astrophysics.

A commonly used measure of a particle's 'penetration ability' is known as rigidity, and is an indication of how deeply it will penetrate the magnetic fields surrounding the Earth. Rigidity can also be described as the degree to which the trajectory of a charged particle will be deflected

in a magnetic field.

It has been observed that particles with rigidity less than the cutoff of 1 GV cannot penetrate the Earth's atmosphere. A particle that has at least 1 GV rigidity will penetrate through the atmosphere and create a cascade (Figure 2.1, p. 7) which reaches the Earth's surface. Not all SEPs will have enough rigidity to penetrate the atmosphere and magnetic field, and are best observed in orbit above the atmosphere.

An influx of SEPs is often attributed to solar flares, however this need not be the case, as CMEs and interplanetary shocks have also been shown to create SEP events (Reames, 1999). CMEs and interplanetary shock need not have a flare associated with them to produce SEP events, however these SEPs often do not have enough energy to penetrate our atmosphere and give rise to secondary particles. Solar flares have a more pronounced effect on cosmic ray detectors when they are on the western hemisphere of the Sun. This is due to the IMF, and solar winds which direct the SEPs to Earth following the curved magnetic field lines (Figure 2.2, p. 8). It is uncertain if solar flares are the cause of the SEP, or if the flares are a by-product of the process which accelerates particles (Duldig, 2001).

2.2 Ground level enhancement events

When these SEPs do have enough energy to penetrate our magnetic field and atmosphere, we see an increased number of secondary particles attributed to cosmic rays. When these SEPs reach the neutrally charged atmosphere of Earth they collide with the air molecules, and break down to produce secondary particles (Figure 2.1, p. 7). This additional source of secondary particles results in a sharp increase, or enhancement, in the number of particles counted by cosmic ray detectors at ground level (Belov *et al.*, 2005). These occurrences are named GLE events, and do not occur as often as SEP events, due to the higher energy requirements.

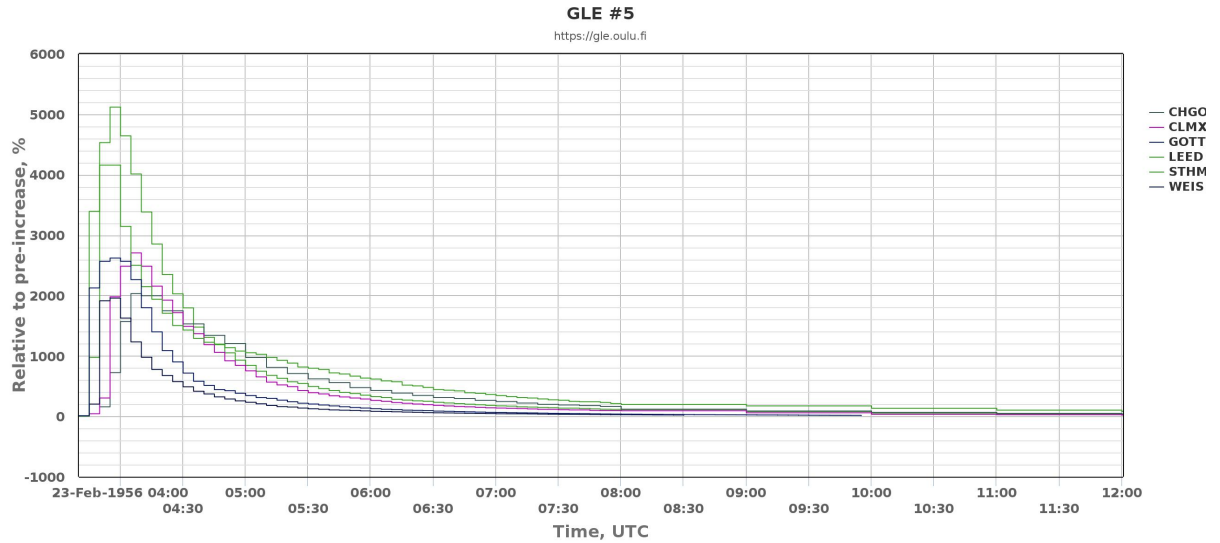


Figure 2.5: Increase of particle count relative to the the count before the increase of GLE 5, on 23/02/1956 (Usoskin, 2016).

During a GLE event, the SEPs will penetrate the Earth's atmosphere, and cause a cascade of secondary particles, which will be detected by neutron and cosmic ray monitors. GLE events are short lived and do not often last longer than a few hours of activity. The combination of these properties produces an impulse response on recordings of cosmic ray intensity (Figure 2.5). This impulse is easily recognizable, as normal cosmic ray activity changes over larger timescales. Forbush was the first to identify and name these events, and he did so using data recorded by the Model-C recorder (Forbush, 1938). He identified 5 GLEs, dated at; 28/2/1942, 7/3/1942, 25/07/1946, 19/11/1949, and 23/2/1956. The most prominent of the pre-space era GLEs occurred in 1956 (Forbush, 1958).

His calculations are said to be conservative, there may have been more GLEs present between 1942 and 1956. These events are not common, and in over 70 years of cosmic ray data, 72 GLE events have been identified, corresponding to roughly one per year.

GLEs sometimes cause a drastic change in the Earth's magnetic field, due to the mass ejection of magnetically charged plasma from the Sun. The largest events are associated with GLEs. This change in magnetic field can produce a geomagnetic storm, causing massive disruptions to the functioning of microelectronics on the surface of the Earth, potentially causing blackouts, or destroying micro-electronic circuits. The largest recorded geomagnetic storm caused by a CME occurred in 1859, and caused widespread disruptions to the then already established telegraph

network. This storm became known as the Carrington Event when RC Carrington observed and recorded the solar activity of the time (Cliver, 2006). It is reported that the storm lasted for up to 2 days, and that the telegraph stations could send messages without being connected to any power source.

2.3 Neutron monitors

The neutron monitor is a device for detecting and recording the neutrons released from collisions between cosmic rays and atmospheric particles. The neutron monitor was invented in 1948 (Simpson, 2000), and since then, there have been a number of different types of neutron monitors. All neutron monitors work on the same principles, which takes advantage of the way in which high energy and low energy neutrons interact with a variety of atoms. Neutrons will not noticeably interact with the electrons of an atom. High energy neutrons produce many low energy neutrons when they interact with a nucleus of an atom. This is due to the disruption they cause in the nuclei, especially heavy nuclei (Krüger, 2006). The interaction of high energy neutrons with atomic nuclei is rare. On the other hand, low energy neutrons interact with nuclei far more often, but these interactions tend to be elastic, not changing the structure of the nucleus. There are some anomalous nuclei (most notably ^{10}B and ^3He), which absorb low energy neutrons, and disintegrate. This disintegration releases very energetic charged particles (Simpson, 2000).

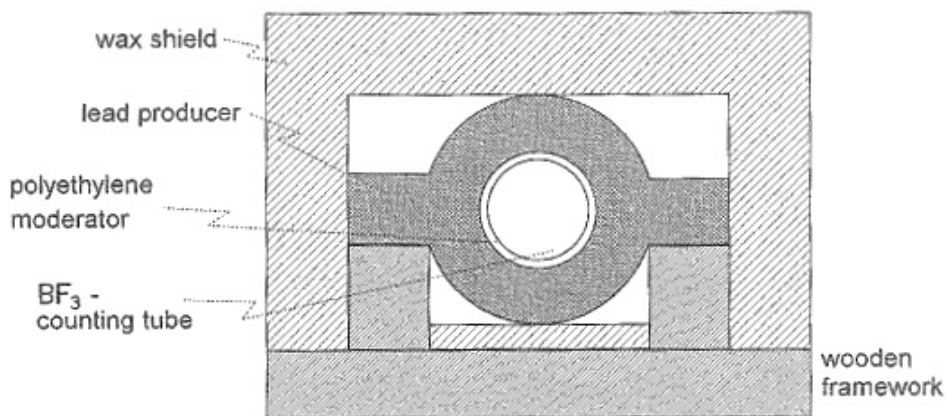


Figure 2.6: Cross section of a typical neutron monitor (Potgieter, 2008).

The major components in a neutron monitor are a reflector, producer, moderator, and proportional counter. The reflector, or shield, (wax shield in Figure 2.6) takes the form of a proton rich

shell, originally paraffin, but polyethylene is used in more recent models. This shield reflects environmental neutrons, which are not caused by cosmic rays and have lower energy, but offers little resistance to those high energy neutrons produced by cosmic ray cascades.

The producer acts to multiply the high energy particles, to make measuring them easier, as more particles will produce a greater response from the recorder. Lead acts as a producer of low energy neutrons and when high energy protons and neutrons interact with the lead producer, it produces an average of 10 lower energy neutrons, for each high energy neutron.

This lead producer serves two purposes, firstly to amplify cosmic ray readings, and secondly, it produces neutrons which will not easily escape the shielded reflector layer. The moderator is also a proton rich material which reduces the speed of the low energy neutrons, increasing the likelihood of them being detected (Simpson, 2000). Lastly, the proportional counter (counting tube in Figure 2.6) contains a Boron rich sample and an ionisable gas.

As the Boron sample disintegrates, the gas is ionized, producing an electrical signal. In Simpson's first monitors, ^{10}B was used as the active component, which produced a signal via the reaction ($n + ^{10}\text{B} \rightarrow \alpha + ^7\text{Li}$), producing an alpha particle and Lithium. More recent detectors use the reaction ($n + ^3\text{He} \rightarrow ^3\text{H} + p$), as the proportional counter. This reaction yields reaction products with total kinetic energy of 764 keV.

Many neutron monitors today have additional features which allow for the recording of the temperature, and barometric pressure. The temperature and barometric pressure are used to correct the count data, depending on the environment in which the monitor was deployed.

The NWU's Centre for Space Research presently has four neutron monitors. These are located at Potchefstroom, South Africa; Sanae, Antarctica; Hermanus, South Africa; and Tsumeb, Namibia (NWU, 2009) and data from them can be seen in Figure 2.7.

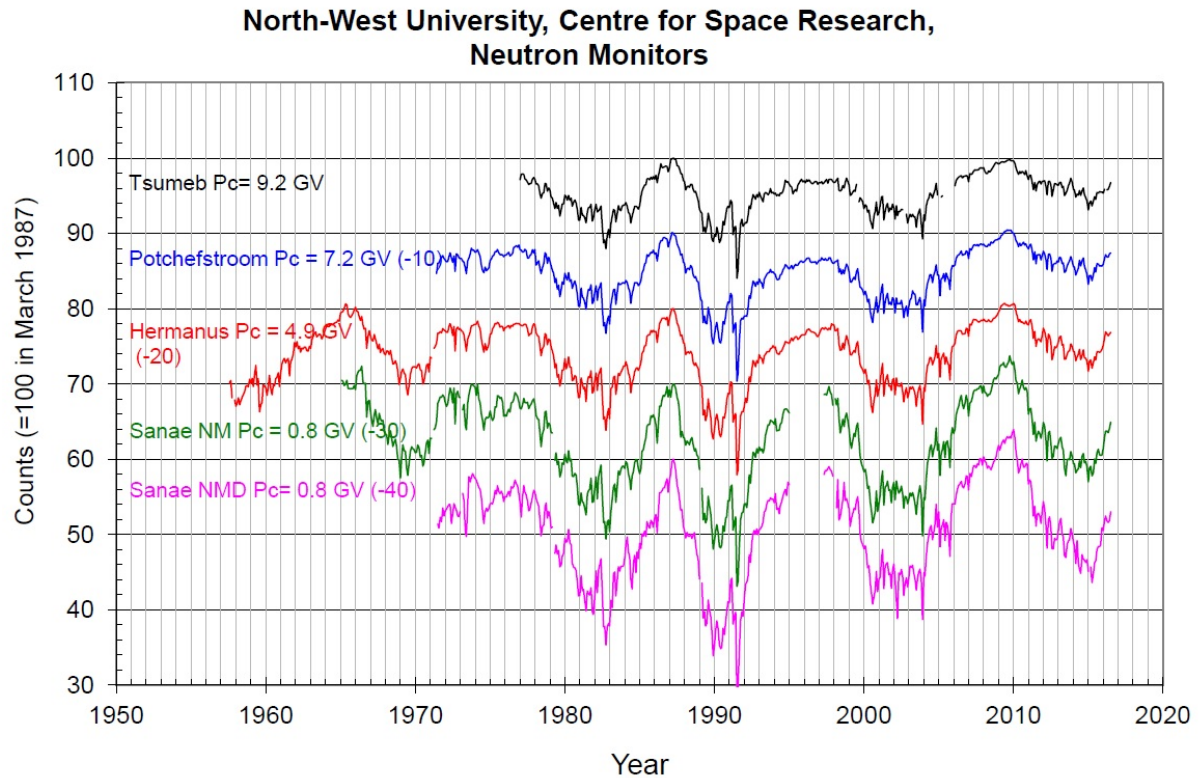


Figure 2.7: Neutron monitor data from stations maintained by the NWU.

Today cosmic rays are detected using neutron monitors, which are virtually maintenance free, and record readings digitally onto computer systems. However, this was not always the case. The newer detectors became the standard for measuring cosmic ray activity in the early 1950s, replacing the older ionization chamber detectors. These ionization chambers were typically large round metal casings, filled with an inert gas, and an electrometer to give readings when a charged particle entered the casing. The Model-C recorder was a prime example of an ionization chamber.

2.4 Model-C recorder details

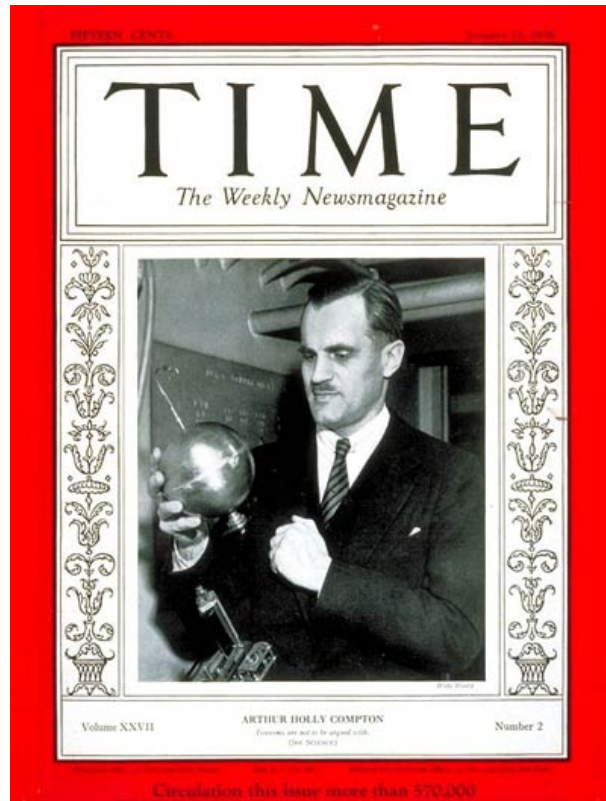


Figure 2.8: Compton on the cover of the Jan 13 1936 issue of TIME with the Model-C recorder.

In 1932, the Carnegie Institute in Washington DC assigned a team of researchers to study cosmic rays and their effects. The team of researchers endeavoured to construct an easily transportable cosmic ray recording device. This team was comprised of Compton, Wallace, and Bennet, amongst others. Arthur Holly Compton was a member of this team, who would later feature on the cover of TIME magazine (Figure 2.8) for his part in designing the Compton-Wollan-Bennett ionization chamber, also known as the Model-C cosmic ray ionization chamber, henceforth the Model-C recorder. These devices were placed around the world, and began recording data by 1933. They continued to produce data for 26 years, up till 1959. These records were used by Scott Forbush, an early cosmic ray researcher, to study the nature of sudden increases in cosmic ray readings.



Figure 2.9: A Lindeman electrometer.

This device was constructed from a metallic sphere which contains argon. This metallic sphere acted as a cathode, while an anode resides inside the device. The movement of charged particles through the chamber ionized the gas (created ion pairs). The electric field applied between the anode and cathode caused the electrons of each ion pair to move to the anode while the positively charged gas atom or molecule was drawn to the cathode. The movement of the ions to the collecting electrodes resulted in an electric charge arriving at the electrode. The ionization chamber was connected to a Lindeman electrometer, which recorded and stored this electric charge in a capacitor. As the capacitor became more charged, a needle casting a shadow on the photographic paper was lifted, recording the charge. The Lindeman electrometer (Figure 2.9) integrated the charge stored on the capacitor, therefore the recorded data represents a time integral of the ionization measured from the last grounding.

The electrometer is grounded for three minutes every hour to discharge the capacitor connected to the Lindeman electrometer. The data line also reset to its base position due to the grounding process.

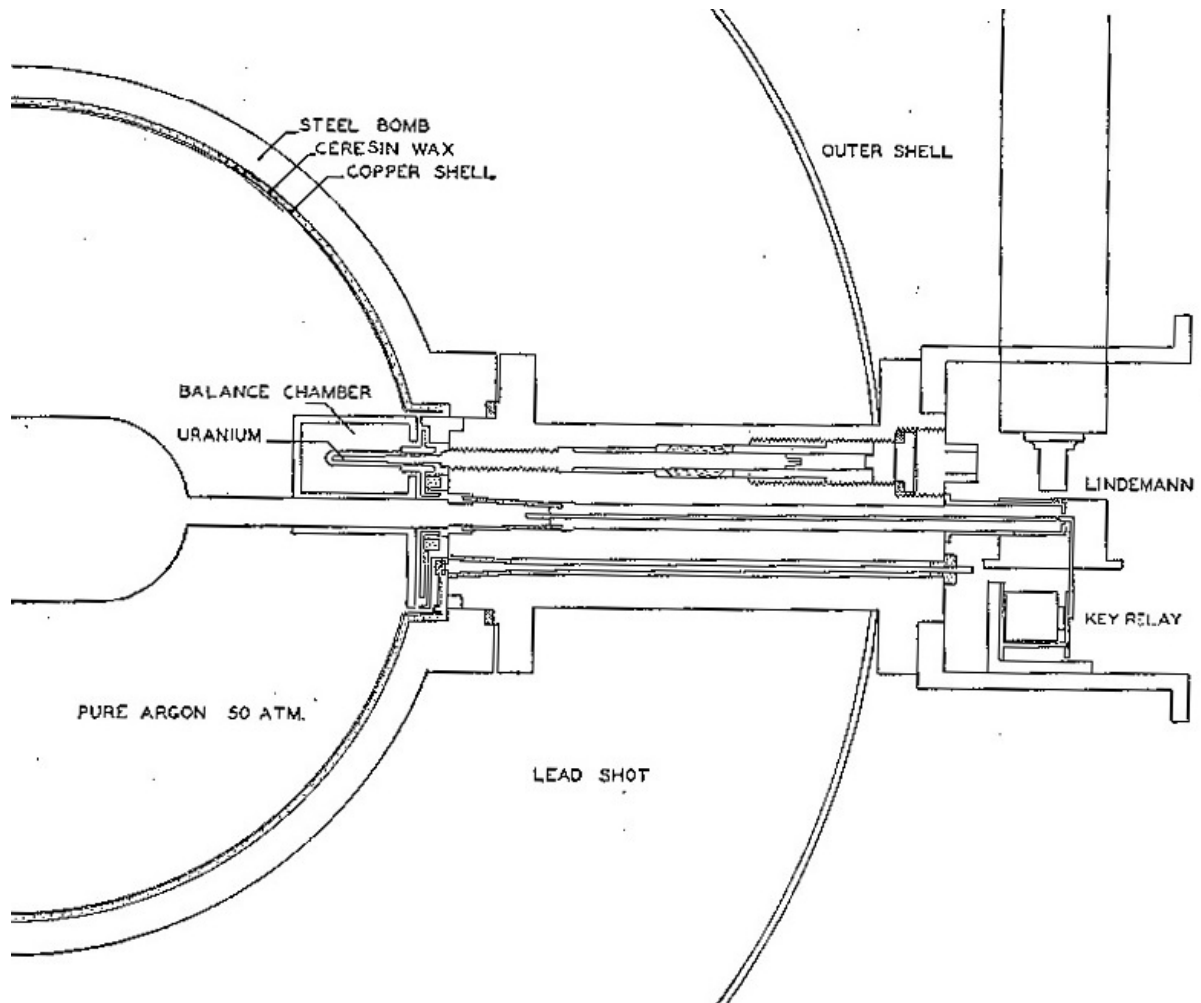


Figure 2.10: Detailed drawing of the ionization chamber (Compton *et al.*, 1934).

A separate smaller ionization chamber (Figure 2.10), contained an uranium sample and the Lindeman electrometer recorded the difference between the ionization currents of this chamber and the larger one which recorded the cosmic rays. As the uranium chamber normally had a larger ionization current than the cosmic ray chamber, the increase in cosmic ray ionization caused this difference to decrease. Therefore a drop in the electrometer reading represented an increase in the cosmic ray counts. The important aspect of the data is its gradient, and the way the gradient changes. A GLE is signified by a sharp drop in regards to the ionization data line on the photographic paper (photographic film). The apparatus was calibrated in a mine shaft to ensure accuracy, and was set so that a data line progressing at roughly 45 degrees to the horizontal can be regarded as the background level of the cosmic rays, with little change in the cosmic ray counts. Any deviation from this gradient is of interest and signifies a change in cosmic ray activity (Compton *et al.*, 1934). There are, additionally, recordings of the barometric

pressure and temperature on the photographic paper as well. The housing for these devices is shown in Figure 2.11.

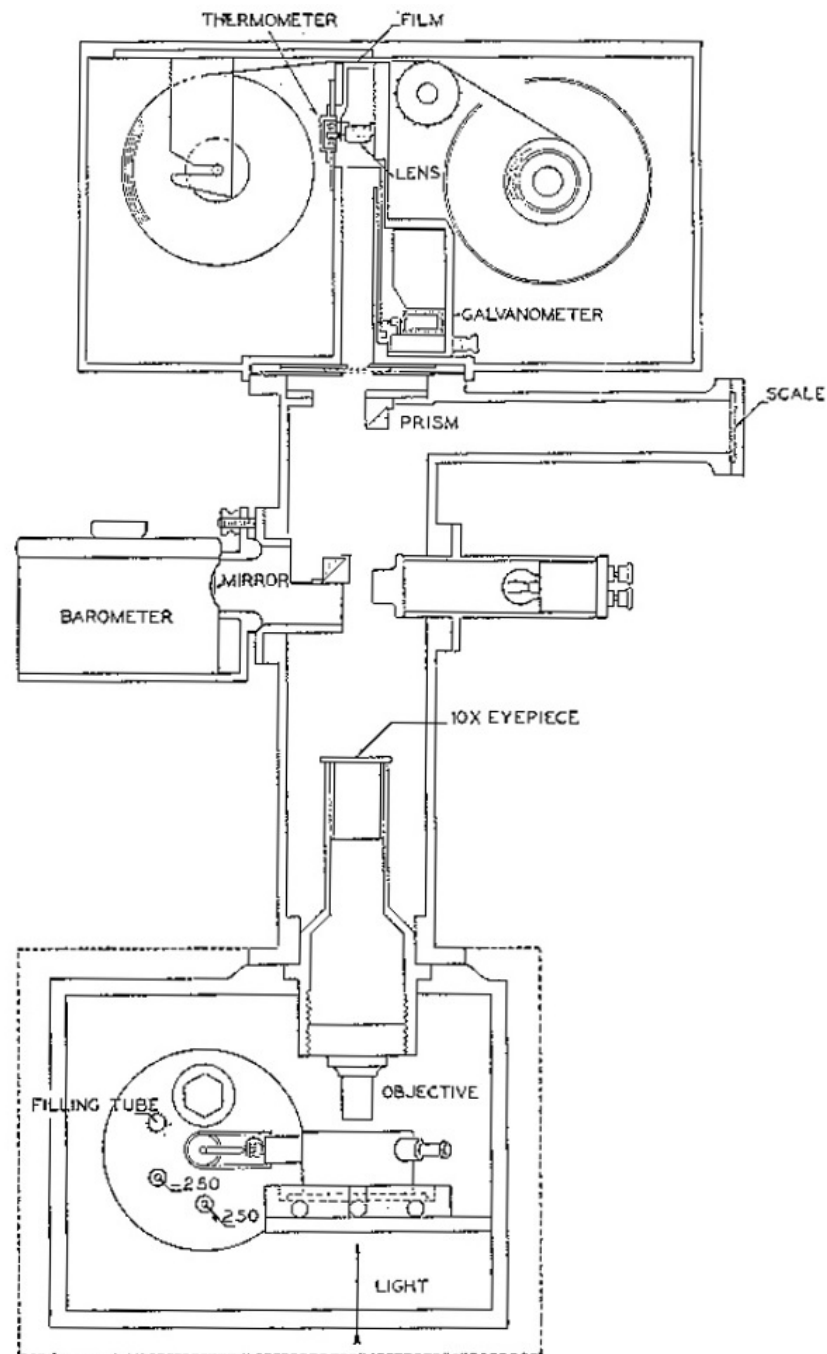


Figure 2.11: Detailed drawing of electrometer box, optical system, barometer, and recording camera (Compton *et al.*, 1934).

There were seven of these devices produced. Five of them were installed around the world at Cheltenham, Maryland (moved to Fredericksburg, Virginia); Huancayo, Peru; Mexico City, Mexico; Christchurch, New Zealand; and Godhavn, Greenland (Forbush, 1938).

The remaining two were used for specialized research projects in Virginia and Colorado in the USA. These stations recorded approximately 114 station years of cosmic ray ionization data on photographic paper. These records were used by Scott Forbush, who identified a number of early GLE events (Forbush, 1938).

These cosmic ray recordings have been well preserved, and their fidelity is exceptional, even by today's standards. This is due to the nature of the recording media, the photographic paper. It represents a continuous measurement, which is formed on a molecular layer, delivering a phenomenal resolution.

2.5 Photographic film details

The photographic paper that the data is recorded on can contribute to the graininess of the image, and any imperfections on the material will be carried over into the digital format. The process of scanning these photographic papers was undertaken by a private contractor hired by the National Geophysical Data Center (NGDC) in Boulder, Colorado. The images measure approximately 6000 by 905 pixels, have a pixel depth of 24 bits, and document a 19 hour period.



Figure 2.12: The locations of the stations which are relevant to this study.

The images used in this study document GLE #5, which occurred at about 4:30 am UTC, from stations located at (Figure 2.12):

1. Christchurch, New Zealand: 43.5° South and 172.6° East. 8 m above sea level.
2. Cheltenham, U.S.A: 83.7° North and 76.8° West. 72 m above sea level.

3. Godhavn, Greenland: 69.2° North and 53.5° West. 9 m above sea level.

The photographic paper is marked with linear horizontal and vertical lines, as well as nonlinear data lines recording the ionization, the barometric pressure, and the temperature. Figure 2.13 shows an annotated section of the photographic paper (about two and a half hours). The structure of the images on these photographic papers can be described by investigating the recording process which resulted in these structures. The recording process involves the mounting of the recording paper onto a sprocket drive, aligning the scale grid, as well as the recording needles.

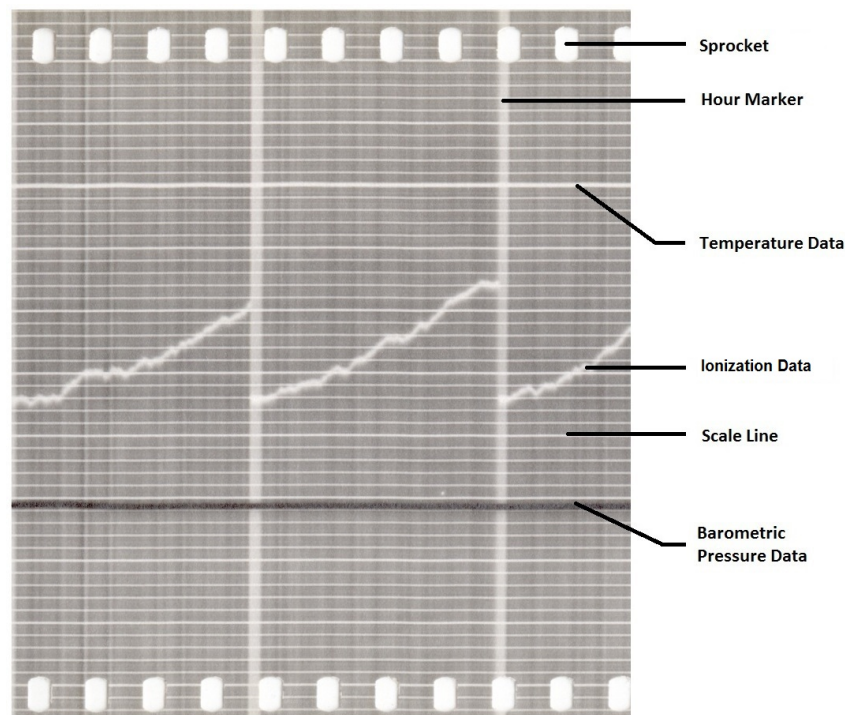


Figure 2.13: Diagram of a section of the photographic paper produced by the Model-C recorder.

Every hour the chamber was grounded for three minutes, resulting in the capacitor in the Lindeman electrometer resetting its charge, returning the data line to its default position, as well as brightening or dimming a lamp, which produces a vertical line to mark the passage of an hour on the photographic paper (Compton *et al.*, 1934). These markers will henceforth be referred to as hour markers, and vary in intensity from station to station, with some recordings having brightened hour markers, and others having darkened hour markers. It was established that the hour markers are 15 pixels wide on average. The gap separating the hour markers is approximately 300 pixels long, from end to beginning of a pair of hour markers. On some of the Model-C recorders the hour markers were recorded by deactivating the background lamp,

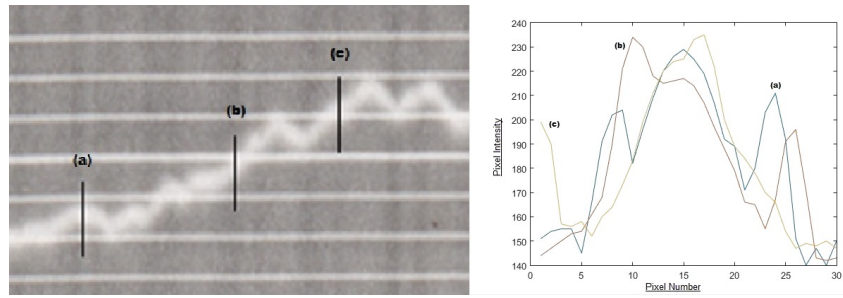
causing a dark hour marker instead of the bright one (Figure 2.14c).

Additionally, there are data recordings for cosmic rays intensity, barometric pressure, and temperature. The cosmic ray intensity was recorded continuously by recording the electric charge accumulating in the ionization chamber (Figure 2.14a). The gradient of the ionization line is related to the cosmic ray counting rate.

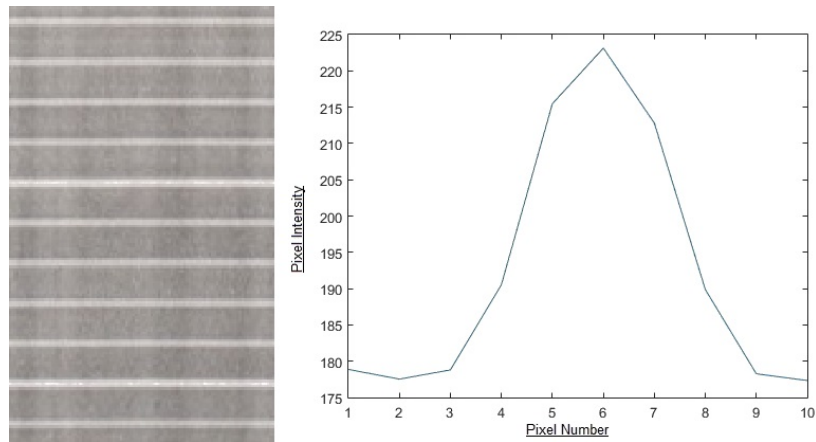
The horizontal lines that mark the image at about 2mm intervals are there to provide scale, and ease the reading of the recorded data. They will henceforth be referred to as scale lines. These scale lines overlap with both the hour markers and the data lines, as they are produced by a prism scored with parallel lines, projecting the lines onto the photographic paper. The scale lines are separated by 12 (± 2) pixels, and run parallel to each other, with only minor discrepancies as the photographic paper proceeds. Figure 2.14b shows a cross section of a sample of the scale lines.

The sprockets on the top and bottom of the film are remnants of the photographic paper, which was propelled using an apparatus that has gears matching the sprockets. These sprockets are identical for each roll of photographic film, and as such we can assume that each image has identical sprockets. However, some sprockets have damage around their edges, and this changes their shape. Sprockets are restricted to the top and bottom edges (80 pixels) of each image, and have a predictable separation of 24 (± 2) pixels. Figure 2.14c shows an example of a sprocket as well as a cross section of the sprocket.

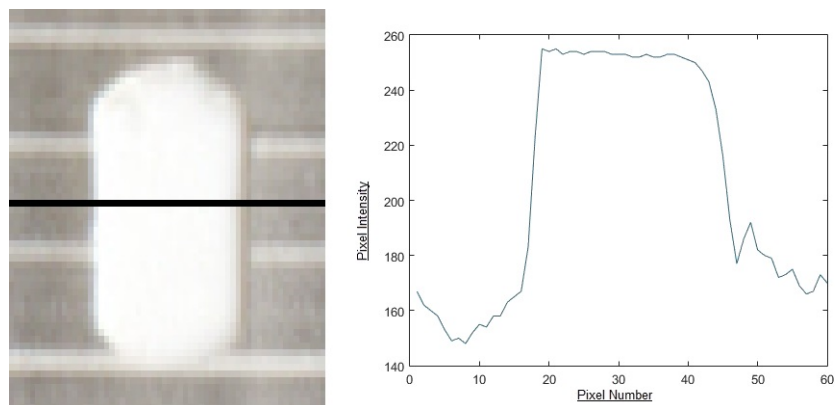
There are a number of artefacts and distortions to contend with on the photographic paper, some of which obscure the data, while others complicate the document analysis process. These artefacts include the sprockets, bends, bright spots, and damage to the photographic paper. The contrast and quality of the images also vary due to inconsistencies in the chemical processing of the photographic paper. Other examples of distortions include annotations that have been made on the photographic paper, in hand written script, with stickers, holes punched in the film, and even the bleed-through of annotations on the back of the paper (Compton *et al.*, 1934).



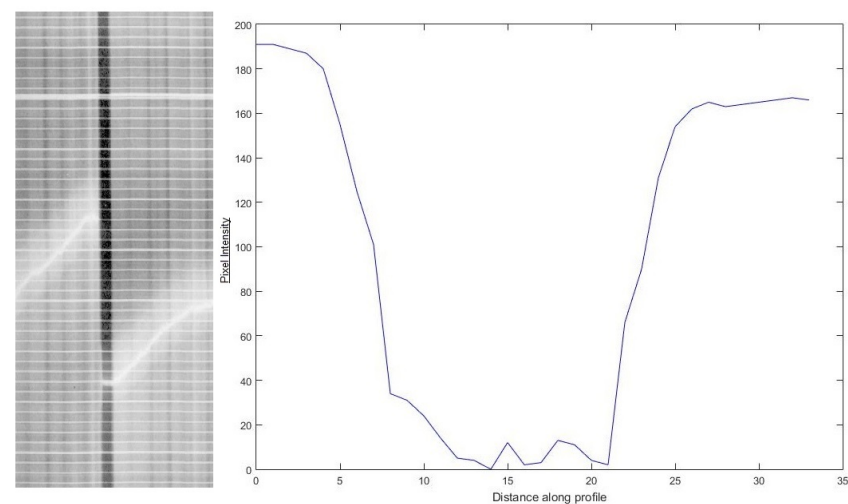
(a) Cross sections of the data line at different points.



(b) The average cross section of a scale line.



(c) Cross section of a typical sprocket.



(d) Cross section of an hour marker.

Figure 2.14: Some cross sections of important features in the cosmic ray records.

2.6 Interpretation of ionization data

As the data recording represents the time integral of the charge related to the ionization levels, we can derive the cosmic ray count from this information, using the derivative, or rate of change of the ionization recording, as well as a background value before the GLE occurred. Using these values we can determine the percentage increase above the background level over a period of time. Intensity-time profiles are the standard method for recording GLE data (Duldig, 1998). The percentage increase for a time interval can be obtained using the following formula:

$$Increase = \left(\left[1 - \frac{I}{I_0} \right] \right) \times 100\%. \quad (2.1)$$

where I represents the current segment's gradient, and is defined as:

$$I = \frac{\Delta c}{\Delta t}, \quad (2.2)$$

with Δc being the change in the ionization level over a period of time Δt . The background cosmic ray level is given by I_0 ,

$$I_0 = \frac{\Delta c_0}{\Delta t}, \quad (2.3)$$

with Δc_0 being the change in the background ionization level over a period of time Δt .

2.7 Conclusion

This concludes the discussion on the research context of cosmic rays, having addressed both the physical properties of cosmic rays, and the nature of the recordings which will be investigated. It is important to note that the ionization data to be extracted will be sourced directly from the original scanned images, instead of from the results of the image processing steps described in the Method (Chapter 4). These recordings are the focus of this study, and the image processing techniques and research context will be discussed in the following chapter.

Chapter 3

Image Processing

Image processing is the field concerned with the manipulation and enhancement of digital images. This can be approached mathematically by regarding an image as a two-dimensional function, where x and y are coordinates, with the intensity of the image at any point given by $f(x, y)$. The coordinates and intensity values are all finite, discrete values. Each coordinate in the image represents a picture element, also called a pixel. Digital image processing can be defined as the field of study concerned with the manipulation of pixels and the application of algorithms on the image function (O’Gorman and Kasturi, 1997). The field of digital image processing is often concerned with the restoration, correction and adjustment of images, taking an input image, and producing a resultant image as output. Many processes are related to the clarification of the image, such as correcting gamma, brightness, distortions, and noise.

There are numerous subfields in image processing, as varied as the number of images available. Examples in the greater field of image processing include facial recognition (Jain and Li, 2011), machine vision (Andersen *et al.*, 2015), inpainting of images (Bertalmio *et al.*, 2000), and even detecting the heart rate of an individual from cctv camera footage (Takano and Ohta, 2007).

One of the main goals of the field is the extraction of information from an image. An example of information extraction is document image processing which will be discussed in the following section, followed by a discussion of binarization, and other image processing techniques. Some relevant basic operations are described below, and a list of significant contributors to the current state of the technique is given.

3.1 Document image processing

Document image processing is designed to recognize the text and graphics components in images of documents, and to extract the composite information from the image. A hierarchy of document image processing (O’Gorman and Kasturi, 1997) is shown in Figure 3.1. In it we see that document image processing can be divided into textual processing and graphical processing. This is due to the different techniques existing to serve a variety of purposes, such as detecting text, locating images, or interpreting charts. Textual processing is divided into optical character recognition (OCR) and page layout analysis. Graphical processing is divided into line processing as well as region and symbol processing. Examples of the applications of these fields appear in circles in Figure 3.1.

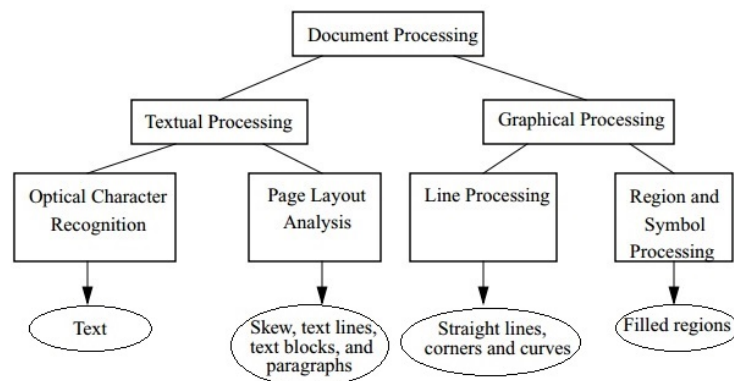


Figure 3.1: A hierarchy of document image processing (O’Gorman and Kasturi, 1997).

3.1.1 Textual processing

Textual processing is concerned with the textual aspects of an image, and the information stored as symbols. This procedure presents some challenges and some of the most commonly encountered methods are: optical character recognition (OCR), establishing the angle of skew, locating paragraphs, columns, lines of text, and individual words. Many advancements have been made in the field, and today OCR is among the most researched fields in pattern recognition (Antonacopoulos *et al.*, 2007), (Clawson *et al.*, 2013). Page layout analysis is concerned with identifying the locations of text blocks, images, headers, and other document elements. It is also concerned with skew correction, which is discussed later. There was once a division in the

field between textual processing and graphical processing, which is shrinking today as the field develops, and new techniques are developed. A change in perspective about user input also contributes to the closing divide. Where in the past autonomous systems were considered to be the ideal, today we see an acceptance of user input as smarter, and more natural than entirely autonomous systems (Chaudhuri, 2007), (Deng *et al.*, 2010).

3.1.2 Graphical processing

Graphical processing deals with line and symbol components that make up diagrams, charts and tables, demarcating divisions between sections of text, and other graphics. A number of these graphical components are built up of lines, and as such, processing procedures includes line fitting, line thinning, and detection of curves and corners (O’Gorman and Kasturi, 1997). Region and symbol processing is comprised of numerous image processing and machine vision techniques. These applications of image processing are highly domain specific, and often are custom built to solve a very specific problem (Kasturi *et al.*, 2002).

There are many different areas of interest within the field of graphical processing, such as digitization of musical scores (Calvo-Zaragoza *et al.*, 2016), historic document recognition and preservation (Lavrenko *et al.*, 2004) and graph recognition and interpretation (Huang, 2008). The techniques used within these fields are often very similar, as they make use of basic image processing techniques, or creating them from scratch (Dee and French, 2015). The difference in application and order of operation can have varying output and applications. As such many developments in the field are algorithmic, and represent a novel order of execution of often familiar functions. Other than improvements to basic functions, most applications of graphical image processing are tailored to their purpose, and would be ill suited to fulfilling other tasks (Chaudhuri, 2007).

3.1.3 Chart image processing

The graphical image processing subfield of chart recognition deals with the recognition of the chart type (bar, pie, line), chart components (text blocks), data, and the intended message of the chart (Huang, 2008). Chart image recognition is focused on identifying and extracting low-level symbols and text using either model based techniques, or by using learning based methods.

The classification of chart components is a core part of the process, and is assisted by high level association of text and symbols, to describe the chart images. The procedure proposed by Huang (Huang, 2008) makes use of four levels of processing, viz: the preparation level, the recognition level, the interpretation level, and the application level.

The preparation level is concerned with classifying components in the document into two categories, text and graphics. The graphical components are located by using processes such as vectorization, graphical symbol construction and chart model matching, while text recognition is accomplished by using text segmentation and OCR. Combining the results from the recognition level, the interpretation level builds chart descriptions by using the text/graphics association, chart data calculation and a generated description of these components. The final level of processing is the application level, which addressed subjects such as question answering, web based searches and supplementing traditional OCR techniques (Huang, 2008).

3.2 Binarization

We can consider binarization in document image processing as the process of segmenting pixels of an image into foreground, and background elements (Su *et al.*, 2013), (Kasturi *et al.*, 2002). The aim of binarization in an image with separate and distinct foreground and background is the tagging of pixels in the foreground as ON, and setting those in the background to OFF.

In the past higher quality results were obtained during the process of binarization if the original image was in a 0-255 grey-scale format (O’Gorman and Kasturi, 1997), as opposed to binarizing a colour image which has multiple levels contributing to the intensity of each pixel. Today, with the aid of hyperspectral imaging, colour images can reveal overwritten text, including text which has mixed ink (George and Hardeberg, 2015). This technique does require specialized capturing apparatus capable of recording both within and beyond the visible electromagnetic spectrum which is not always available, and as such the images referred to in this study are assumed to be in grayscale format. If the grey-level values for the foreground and background are relatively uniform over the range of the image, a global threshold can be used. This implies a single threshold value to be applied to every pixel to separate the foreground and background (Kasturi *et al.*, 2002). One threshold can be selected for the entire image, but this technique rarely works well, unless the document is almost perfectly recorded, and has no differing background gray

levels. If the background intensity varies too much, some sections will be either entirely black or entirely white. This is not regarded as a successful binarization. A sample image and a number of binarized versions, including a failed global threshold, are shown in Figure 3.2.

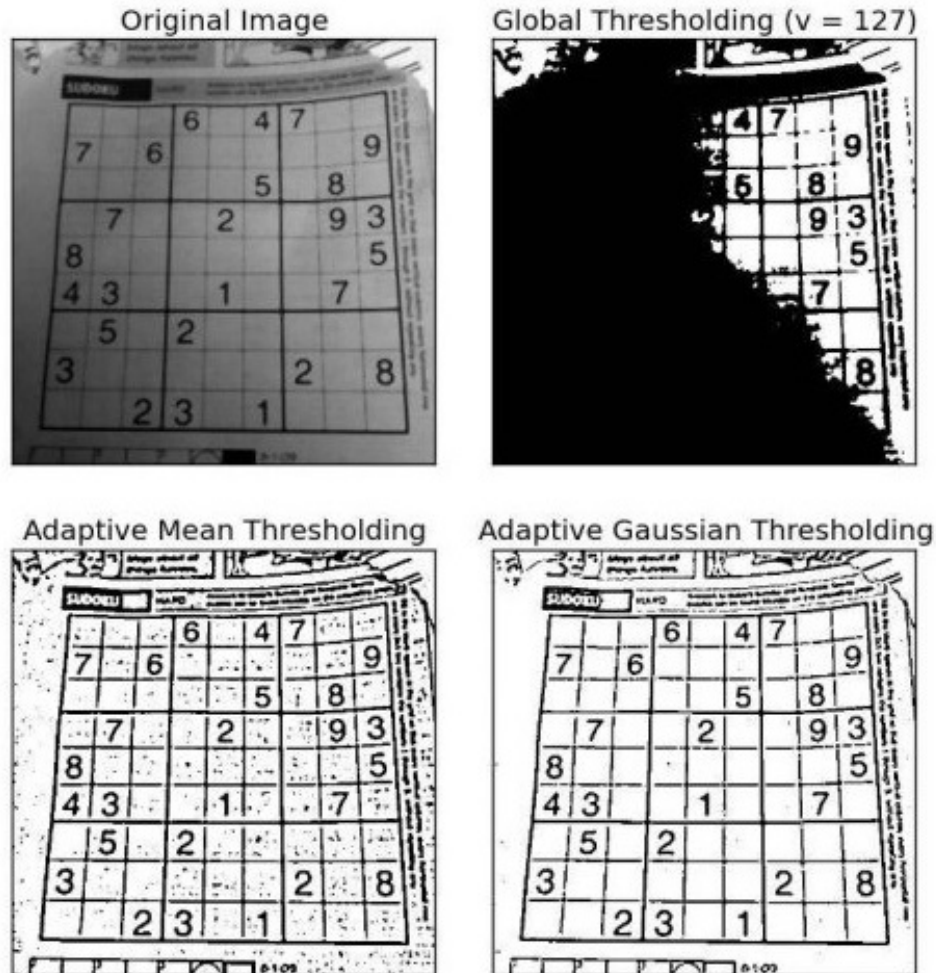


Figure 3.2: Binarization using a variety of methods, showing the weakness of global thresholding.

One of the easiest ways of selecting a global threshold, when the image is composed of distinct foreground and background elements, is to use the histogram of pixel intensities. The one peak will be significantly smaller than the other if one intensity is sparse compared to the other (e.g. text on a blank page). There will however be some section of the histogram which represents both background and foreground, and if the threshold is in this region there will be loss of foreground or background detail in the resulting binary image.

In situations like these there is a need for a local thresholding, which selects a different threshold value for different regions of the image based on a filter or mask around each pixel. An example of this problem is the binarization of a magazine page with plain text, a number of boxes

with text in them, a background of a different shade, and images present on the page. Multi-thresholding would be utilised in this instance, with a separate threshold for each region. Many multi-thresholding techniques require the number of levels beforehand (O’Gorman and Kasturi, 1997). Otsu’s algorithm makes use of grey scale histograms, and finds a threshold at some point between the peaks, depending on the number of levels (Otsu, 1975), as it provides a good separation of foreground and background on high quality images. Adaptive thresholding is a technique which selects different threshold values for different regions, by inspecting the grey-level intensities in a mask across the image. A number of different adaptive thresholding algorithms exist, such as Sauvola (Sauvola and Pietikäinen, 2000), and Gatos (Gatos *et al.*, 2006). Recently Sauvola’s algorithm has been improved (Lazzara and Géraud, 2014), rectifying some of its shortcomings, such as the static window parameters, failing at low contrast levels, and its unpredictability when processing contrast inverted images.

Attempts to judge the degradation of the image before binarization have been made with sufficient success to warrant interest in the subject. This approach is based on building prediction models to assist in the selection of thresholds from the image (Rabeux *et al.*, 2014). In this way, the thresholds for binarization are selected per image, and not for the entire data set, which improves the quality of the binarization process.

Histogram analysis is often central to determining a good threshold as demonstrated by Rabeux *et al.* (2014). An approach using histogram analysis to determine an appropriate threshold by using overlapping windows to histogram equalize a grey scale image was proposed by Singh *et al.* (2014). This method uses the histogram equalized sub-images to select a threshold between the two major peaks in the histogram by splitting the histogram in half, and finding the local peak for each half. If more than one peak is found, the right most peak is selected for the black half, and the leftmost peak is selected for the white half. Once these peaks have been located a threshold is calculated and the image is binarized. Some additional noise removal steps are taken after the binarization process (Singh *et al.*, 2014).

The correct thresholding is critical to the success of the binarization process. Automatic thresholding algorithms have been an area of much development for some time, with the process of automatically determining a good threshold improving frequently (Kasturi *et al.*, 2002). Edge noise adds to the complexity of the thresholding problem, and researchers have recently developed an effective technique using scarcity based algorithms (Hoang *et al.*, 2014). These

algorithms uses dictionary learning, based on the input image, to train the algorithm to detect noise in the image. This algorithm produces good results even with very noisy images (Figure 3.3).

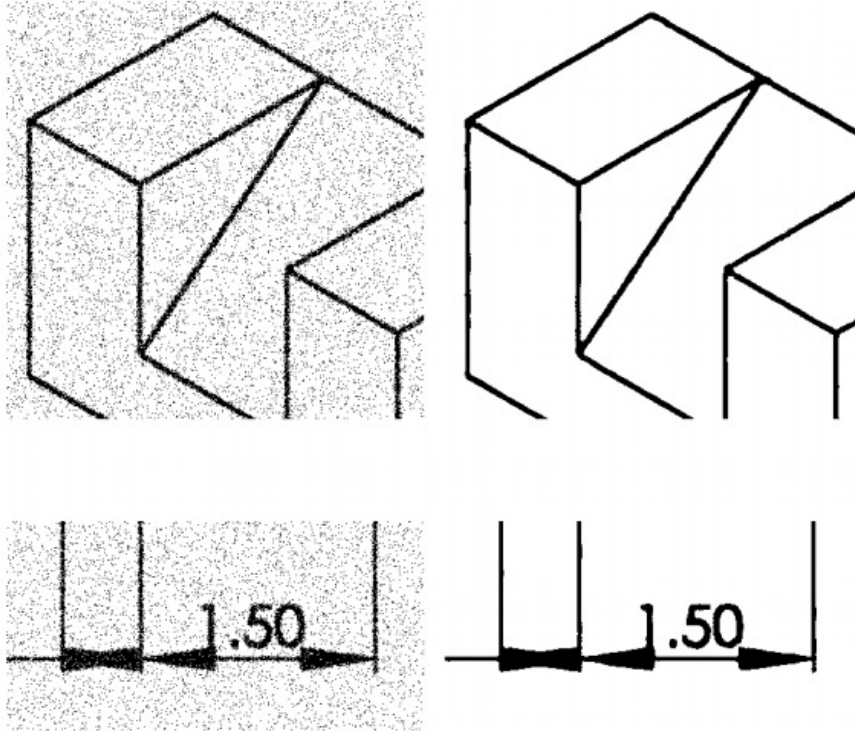


Figure 3.3: The results of the sparsity based algorithm of Hoang *et al.* (2014).

Artefacts such as holes, tears or ink blotches are detrimental to binarization and can often bias the threshold. Significant efforts have been made to automate the removal of noise clutter from binarized images, and high removal rates have been reported by Agrawal and Doermann (2013).

3.2.1 Evaluations of binarization techniques

Although the methods outlined above each have advantages and disadvantages, the choice of which procedure to use for an application is not a trivial one, and many studies have been done to evaluate the efficiency of the different techniques for a specific problem. For example, the extraction of text elements from historical documents requires extra attention to distortions and damage to the document, while commercial OCR applications assume that the document is in pristine or near pristine condition. Du Plessis and Drevin (2009) have shown that the method proposed by Gatos *et al.* (2006) is the superior method for binarizing cosmic ray records, although

it has some failings. Lamiroy and Sun (2013) compared Otsu (Otsu, 1975), Sauvola (Sauvola and Pietikäinen, 2000), and Wolf (Wolf and Doermann, 2002), showing that Sauvola and Wolf's techniques perform roughly equally well, while Otsu performs far worse (Lamiroy and Sun, 2013), due to its global threshold. Sauvola and Wolf both use local thresholding techniques. There are a number of measures used to evaluate the success of the binarization process such as precision, accuracy, recall, and F-measure. These techniques are discussed in a later section. These measures are used for a variety of document types, including handwritten documents (Ntirogiannis *et al.*, 2008b) as well as printed modern documents (Ntirogiannis *et al.*, 2008a).

3.3 Morphology

Morphology is the field in image processing devoted to extracting components in an image that can be used to represent a region or shape, such as a boundary, or skeleton (Gonzalez and Woods, 2002). Morphological processes can be low-level (using only basic operations), or they can be more advanced, building operations upon each other, or nesting operations. Morphology is used in segmentation, as demonstrated by Wu *et al.* (2015), who used morphological reconstruction to successfully segment images. Some low-level morphological processes, such as erosion, dilation, opening and closing, are described below.

Let $\mathbb{E}$ be an Euclidean space on an integer grid ($\mathbb{Z}^2$),

$$\mathbb{E} = \mathbb{Z}^2, \quad (3.1)$$

Let A be a binary image in $\mathbb{E}$, and let B be a set of 9 elements, that is, $B = \{(-1,-1), (-1,0), (-1,1), (0,-1), (0,0), (0,1), (1,-1), (1,0), (1,1)\}$. This set, B , can be visualized as a 3 x 3 square.

3.3.1 Erosion

The first basic morphological process that will be discussed is erosion. This process can be defined as follows: The erosion of the binary image A by the structuring element B is defined by:

$$A \ominus B = \{z | (B_z) \subseteq A\}, \quad (3.2)$$

where B_z is the translation of B by the vector z , i.e.

$$B_z = \{b + z | b \in B\}, \forall z \in \mathbb{E}. \quad (3.3)$$

In other words, this process checks a window, or structuring element, B around each pixel in the image, A . If the area being inspected contains an ‘OFF’ value, the centre pixel of B will be set to ‘OFF’. Otherwise if the entire region under the structuring element is ‘ON’, the centre pixel will remain ‘ON’. For example, the process of erosion is useful to thin bold text (Figure 3.4).

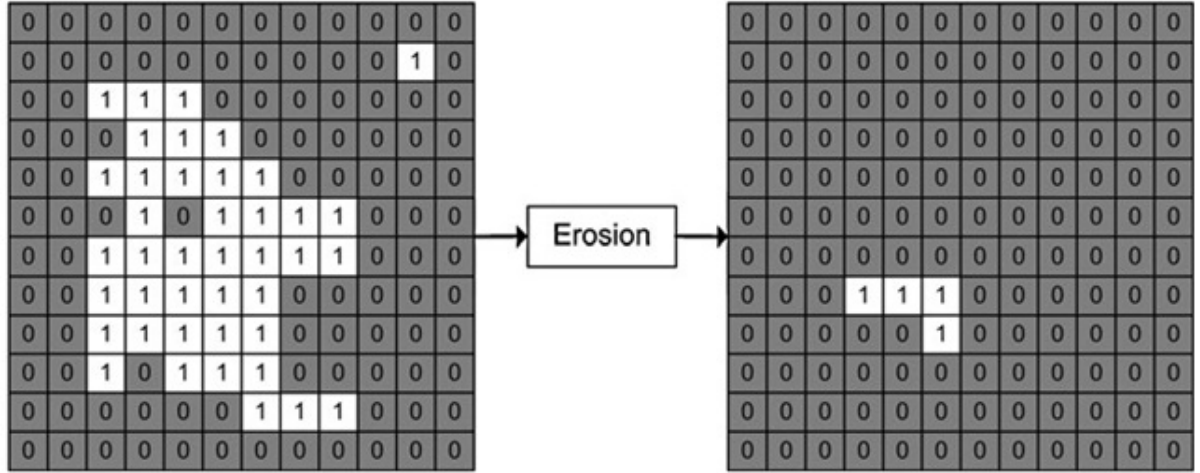


Figure 3.4: An example of erosion, removing the outer pixels.

3.3.2 Dilation

The second basic morphological operation is dilation. It is the dual process to erosion, thickening any ‘ON’ regions in the binary image. The dilation of A by the structuring element B is defined by:

$$A \oplus B = \{z | (\hat{B})_z \cap A \neq \emptyset\}, \quad (3.4)$$

where $\hat{B}$ is the reflection of B , given by:

$$\hat{B} = \{w | w = -b, \text{ for } b \in B\}. \quad (3.5)$$

The dilation is commutative, also given by:

$$A \oplus B = B \oplus A. \quad (3.6)$$

So, if the structuring element encounters any pixels within its borders which are marked ‘ON’, the centre pixel will be set to ‘ON’. This results in a thickening of any lines, or pixels present in the image (Figure 3.5). Dilation might be used to connect broken lines in a diagram, or cause text to appear bold.

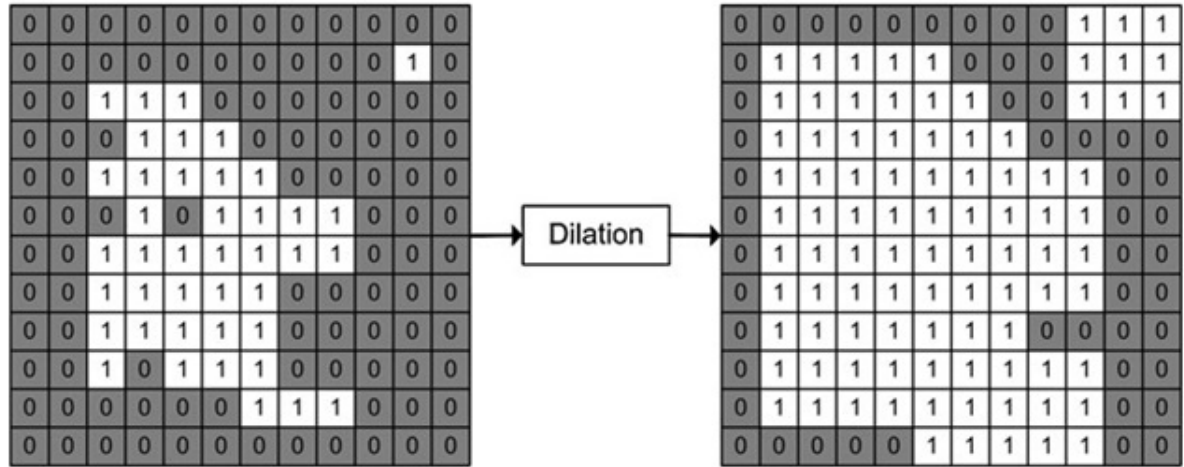


Figure 3.5: An example of dilation, adding pixels to the shape.

3.3.3 Opening

The first composite operation in morphology is opening, and consist of a process of erosion followed by dilation. The opening of A by B is obtained by the erosion of A by B , followed by dilation of the resulting image by B :

$$A \circ B = (A \ominus B) \oplus B. \quad (3.7)$$

This process results in the removal of any salt noise in the image, and thinning larger areas (Figure 3.6). This process can be nested to remove larger regions, for example to remove a 2x2 block of 'ON' pixels. The process would consist of two iterations of erosion, followed by two iterations of dilation. This technique is very useful in removing noise in an image.

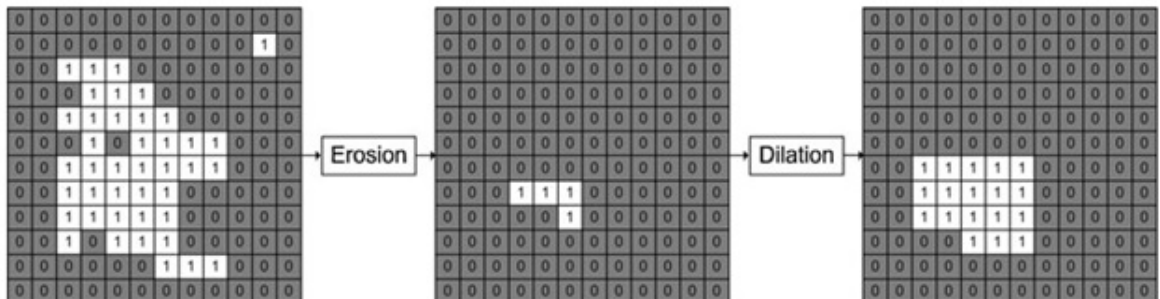


Figure 3.6: The effect of morphological opening.

3.3.4 Closing

The last morphological process that is discussed in this study is the process of closing. Closing is a composite process comprised of a dilation followed by a erosion. This process is dual to opening. The closing of A by B is obtained by the dilation of A by B , followed by erosion of the resulting structure by B :

$$A \bullet B = (A \oplus B) \ominus B. \quad (3.8)$$

This technique allows the closing of holes in a section of ‘ON’ pixels, while not removing noise (Figure 3.7). Much like the process of opening, closing can also be applied iteratively, to close holes of larger sizes.

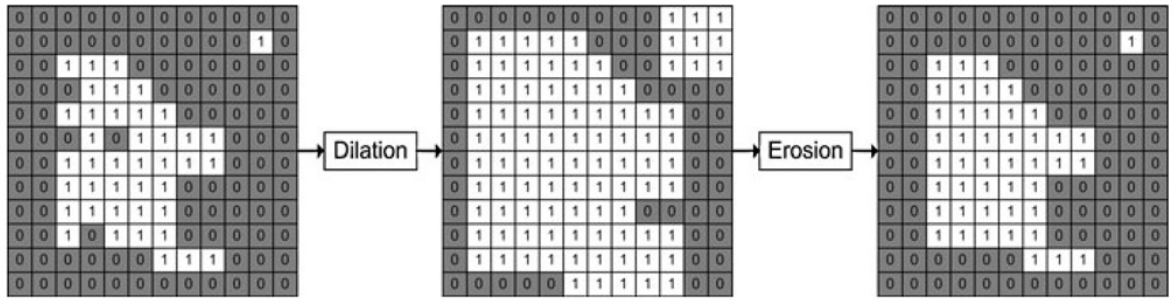


Figure 3.7: The effect of morphological closing.

3.3.5 Watershed

In terms of mathematical morphology, the concept of watershed transformation is adapted from the topographical concept of a watershed line. A watershed line is a region which separates two catchment basins. In terms of image processing, the grey-scale image can represent a topographical relief, where the pixel values represent the elevation at that point. Digabel and Lantuéjoul proposed the watershed transformation (Digabel and Lantuéjoul, 1978). There are a number of ways to define the watershed transformation, as explained in the literature (Beucher and Meyer, 1992), (Vincent and Soille, 1991), which are beyond the scope of this study.

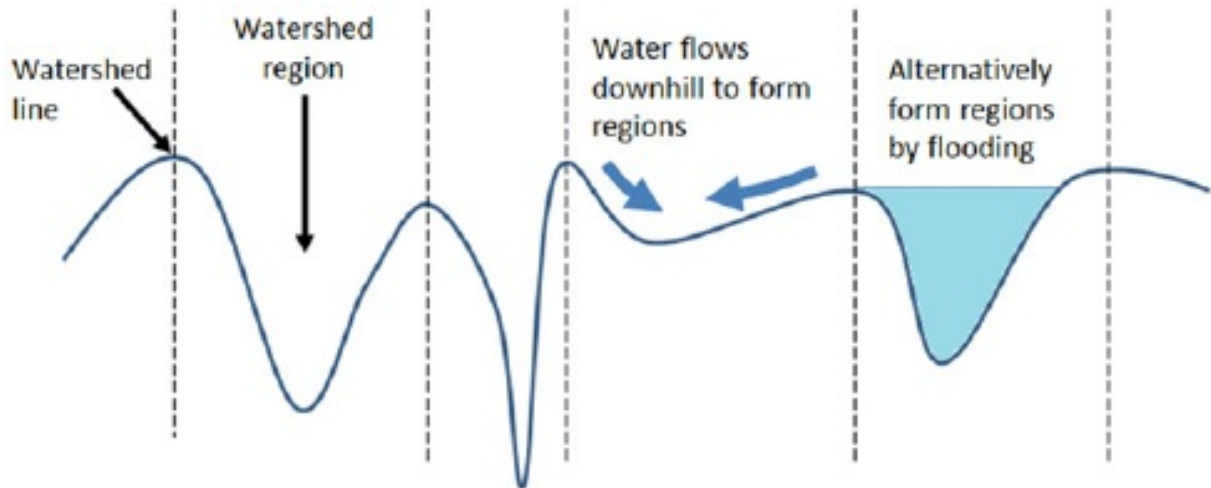


Figure 3.8: Illustration of watershed lines, and the regions which lie between these lines (MacAulay *et al.*, 2015).

Binarization based on adaptive water flow is suggested by Valizadeh and Kabir (2013), by considering an image as a landscape, with peaks and troughs. If one considers rain over such a terrain, the valleys would fill with water, while the peaks would not be submerged. The process is based on filling each valley halfway, and subjecting the filled regions to a classification process (Figure 3.8). This approach is used effectively on documents with differing intensities, such as historical documents.

3.4 Segmentation

Segmentation is the process of dividing an image into its component parts, such as foreground, and background. The foreground can consist of a number of objects, resulting in more than two levels (Gonzalez and Woods, 2002). Segmentation in document processing is mainly focused on extracting and separating text areas, graphical segments, as well as paragraphs and sections from the background (Li *et al.*, 2008). Segmenting an image into logical entities is nontrivial, and often requires some a-priori knowledge of typical documents. Typical a-priori assumptions about documents include that lines of text run horizontally (or vertically) parallel to each other, or that paragraphs are separated with whitespace, or an accent (such as an asterisk or a more ornate design). Depending on the granularity of the segmentation, there are a number of approaches that can be used. Some examples include the retrieval of historical documents (Fernández-Mota

et al., 2014), or the segmentation of musical scores from text (Pedersoli and Tzanetakis, 2016) or even the understanding of comic books (Rigaud *et al.*, 2015).



Figure 3.9: The effect of Canny edge detection on an x-ray of a pair of hands.

Due to the nature of image processing, applications in the broader field of image processing can often be applied to document image processing, or other subfields with relative ease, due to the similarities that all images share, such as pixel intensities, or edges between regions of differing intensities. Edge detection features heavily in segmentation applications, with the Canny edge detection algorithm being robust and reliable (Canny, 1986) (Figure 3.9).

Features such as lines, circles, squares, text, and roads (Sonka *et al.*, 2014) have been identified and extracted with varying degrees of accuracy. Lines, and simple geometric shapes have been identified with great success and minimal computational power, within the document image processing environment, such as identifying roads from raster maps (Chiang and Knoblock, 2013).

Segmentation can be described as the grouping together of perceptually significant regions of an image, reflecting global properties of the image (Rucklidge *et al.*, 2001). It is a broad field with applications in almost all the subfields of image processing, as well as OCR.

There is also work done on the segmentation of graphical elements such as charts, into component elements, such as axes, data points, markers, and dividers (Huang, 2008). The medical field is also interested in image processing, as many diagnostic techniques rely on images, such as x-rays, or even circulatory system markers, used to detect breaks or problems in the blood stream (Niczyporuk and Miller, 1991).

The problem that is to be solved in this study is mainly a segmentation problem, focused on identifying the different attributes in the historic cosmic ray recordings. There has been work done in this vein, and it is relevant to this study (Mattana *et al.*, 2014), (Mattana and Drevin, 2017), (Drevin *et al.*, 2007), (Drevin, 2008), (Du Plessis and Drevin, 2009).

3.5 Labelling

Feature level analysis is concerned with global image statistics, such as the skew of the text, the length of lines, spacing, and a number of other factors concerned with OCR. In graphical feature-level analysis, features include line width, range of curvature, rectangles, circles and other geometric forms (O’Gorman and Kasturi, 1997). Only a mention will be made of techniques in this field as most of the applications of these procedures are not relevant to the problem addressed in this study. The main processes in feature-level analysis are line and curve fitting, which attempts to describe, and attach meaning to the lines and curves found in the previous steps. Critical point detection is the area of interest concerned with finding junctions, loops, end points, and labelling them as such for use in recognition software (Gonzalez and Woods, 2002).

3.6 Hough transform

The Hough transform (Ballard, 1981) is a technique used to extract features in an image, such as lines at any angle, circles, or ellipses. The technique uses a voting system for building an accumulator space, which represents the likelihood of a match between the input shape, and the imperfect candidates in the image. The local maxima in this accumulator space represent object candidates in the image. This transform gives an image consisting of sinusoidal shaped curves (Fig 3.10). This image spans horizontally from 0° to 360° . For instance, if vertical lines need to be found, one needs only inspect the columns 85-95 of the Hough transform, as that is where any 90° line maxima will be found in the transform.

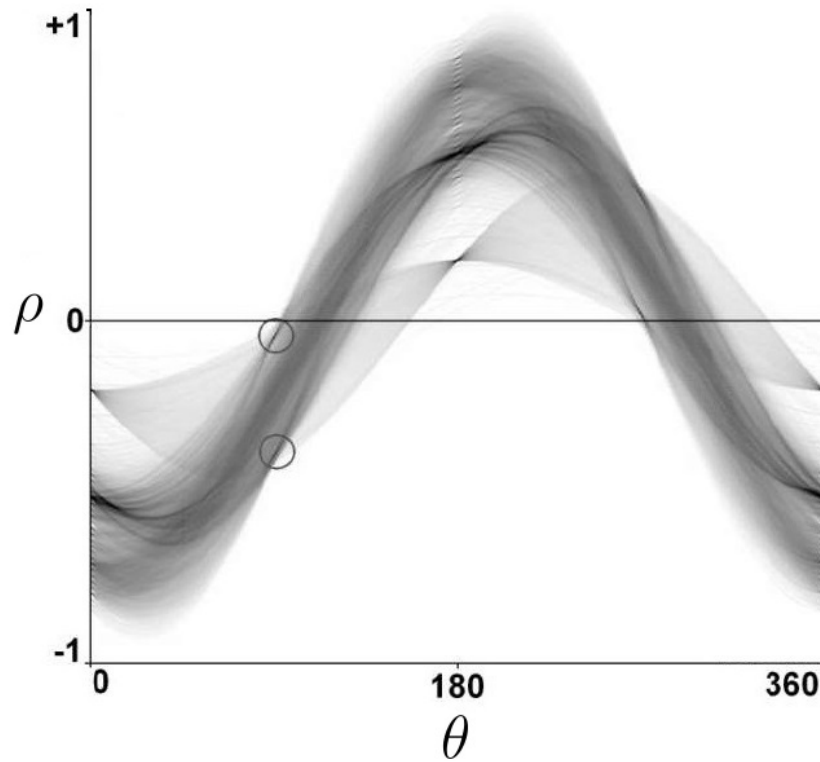


Figure 3.10: A Hough transform of an hour segment of cosmic ray recordings, with hour markers circled (Du Plessis and Drevin, 2009).

3.7 Correlation

Correlation is an image processing technique used to find subsections of an image, in a larger unknown image. As a function, it takes the sub-image, and the parent image, and returns a map of where the sub-image best fits into the parent image. Since the function is normalized, a direct match will result in a value of 1 at the relevant pixel. It uses a windowed averaging technique, which compares the average intensity of the sub-image to the average intensity of the current window (which is always the same size as the sub-image). This allows one to find the location of a sub image, within the parent image (Gonzalez and Woods, 2002). A correlation coefficient is found using the equation:

$$r_{ij} = \frac{\sum_m \sum_n [f(m+i, n+j) - \bar{f}][g(m, n) - \bar{g}]}{\sqrt{\sum_m \sum_n [f(m, n) - \bar{f}]^2 \sum_m \sum_n [g(m, n) - \bar{g}]^2}}, \quad (3.9)$$

where the maxima of r_{ij} is the point where the smaller image is most likely to lie. The symbols m and n represent the position of the current pixel being processed. The symbol $\bar{f}$ is the average of the windowed area f , and $\bar{g}$ is the average of the smaller image, g .

3.8 Skew detection

Often, documents are not scanned perfectly, and will be skew. This can be a major problem, since many assumptions are made based on the spatial characteristics of typical documents, such as lines of text that run horizontally or vertically, or that paragraphs are separated by a horizontal line in many languages. Correcting skew is usually far easier than adapting an algorithm to compensate for the skew. Papandreou *et al.* (2014) developed an efficient technique based on profiles of the documents, which relies on the assumption that lines of text are parallel to each other, either horizontally or vertically. Hough transforms can be used to locate the skew of segments, as demonstrated by Shafii and Sid-Ahmed (2015).

3.9 Evaluation

This section deals with the measures used to evaluate the success or failure of this study, and will discuss ground truth images, the mean square error, as well as accuracy, precision, and recall in image processing.

3.9.1 Ground truth

A ground truth image is a standard against which the success of the algorithm is measured. It can be considered as the memorandum, or ‘correct answer’ to the question which is posed during the process of segmentation or binarization. The use of ground truth images is widespread, and is seen in contests such as DIBCO (Pratikakis *et al.*, 2013), (Gatos *et al.*, 2009). It is sometimes handmade, with the author creating it carefully, or it can be automatically extracted from an input image. The drawback to handmade ground truth images is the amount of effort and time such a process demands. The drawback of automated ground truth creation is the possibility of misrepresenting the actual content of the image. A middle ground between these two approaches

seems to be the preferred approach, balancing accuracy of human detection, with the speed of automated detection.

There are problems with blindly accepting the ground truth image as perfect, though, as the process which is responsible for the creation of the ground truth may have errors itself. This would then compromise the accuracy of the results obtained from any algorithm tested against this ground truth. For example, if a ground truth image has an omission, and the algorithm to be tested accurately detects the element which was omitted from the ground truth, that algorithm would score worse than an algorithm which makes mistakes consistent with the ground truth image. This is possible prevented by multiple checks being put in place, and using independent ground truth images to obtain a statistically significant measure of the accuracy of an algorithm, based on more than one ground truth image. This is not always practical due to the effort this kind of testing would require. However, studies have been done to calculate the precision, accuracy, and recall using probabilistic methods, with incomplete or damaged ground truth images (Lamiroy and Sun, 2013).

3.9.2 Mean square error (MSE)

If we consider $\hat{Y}$ to be a vector (or matrix in our case) of n results from a function, and Y is the array of values observed to correspond to the inputs of the same function, then the MSE of the function can be given by:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (3.10)$$

Stated differently, the MSE is the ‘mean’ ($\frac{1}{n} \sum_{i=1}^n$) of the ‘square’ of the errors between data points $((\hat{Y}_i - Y_i)^2)$. This quantity is easily calculated for any sample. However, the MSE has the drawback of assigning larger weights to outliers (Bermejo and Cabestany, 2001).

3.9.3 Accuracy and Recall in Image Processing

An often used method of testing accuracy is to investigate the ratio of false positives, and false negatives. The formulae for the precision, P , and recall, R , are given below.

$$P = \frac{|\{\text{relevantelements}\} \cap \{\text{retrievedelements}\}|}{|\{\text{retrievedelements}\}|} \quad (3.11)$$

$$R = \frac{|\{relevantelements\} \cap \{retrievedelements\}|}{|\{relevantelements\}|} \quad (3.12)$$

If we consider true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn), we can rewrite the formulae for precision P , recall R and accuracy A , discussed below (Rijsbergen, 1979).

Precision can be considered to be the fraction of correctly classified elements (tp) over all the elements classified in that way ($tp + fp$). Another way of interpreting precision is to consider the grouping or consistency of the classifier. Figure 3.11 shows this as a curve, with the width of this curve labelled as precision. Precision can be expressed as:

$$P = \frac{tp}{tp + fp} \quad (3.13)$$

Recall can be described as the fraction of positives which are correctly labelled as positives. Recall is sometimes called sensitivity (Powers, 2011). Recall can be expressed as:

$$R = \frac{tp}{tp + fn} \quad (3.14)$$

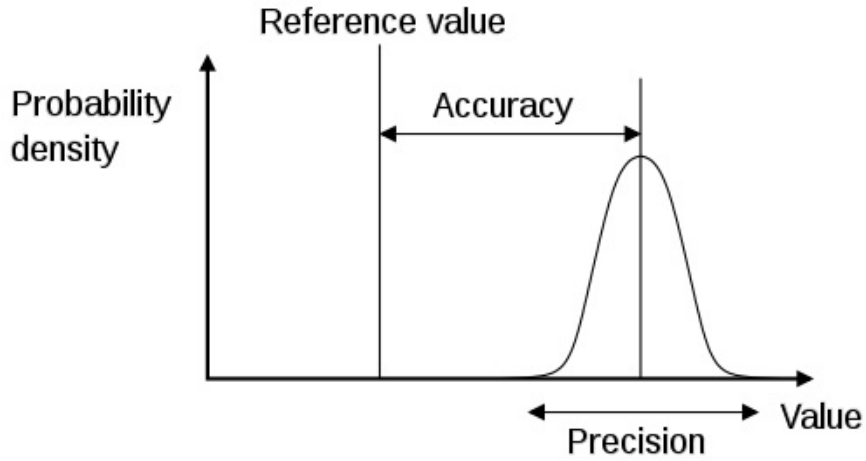


Figure 3.11: A diagram showing the difference between precision and accuracy.

Accuracy can be considered as a measure of how correct the positive results are. In terms of Figure 3.11 the accuracy is the distance to the reference value. Accuracy can be expressed as:

$$A = \frac{tp + tn}{tp + tn + fp + fn} \quad (3.15)$$

These measures can be used to quantify the accuracy, precision and recall of the extraction algorithms. The ground truth images are critical in evaluating the extraction algorithms, if these measures are to be used to quantify the quality of the extraction algorithms.

3.10 Conclusion

The focus of this Chapter was on the image processing techniques that are used in this study to create ground truth images and test the data extraction algorithms using the synthetic ground truth images. The ground truth image was re-used to locate areas of interest in the original image, as well as to enable the testing of varied data extraction algorithms. The following chapter will discuss the method, and implementation of the techniques and research presented in the previous chapters, focused on the studies' aim of extracting the recorded ionization levels, as well as creating a ground truth image, and a synthetic image, both of which are to be used for the testing of data extraction algorithms. For clarity, there are two image processing algorithms at work in this study. The first is related to the data extraction, where important attributes of the original images are isolated and combined to create a ground truth image. This has a dual purpose, whereby the extracted attributes are representative of a ground truth image, and reproducing the data bearing attributes in a controlled image. This first algorithm is used in the extraction of the ionization data, when applied to the original image, as well as being used as a step in the creation of a synthetic image. The second algorithm is used only in the creation of the synthetic image, as the second algorithm is focused on adding texture, optical artifacts and other distortions to the ground truth image produced by the first algorithm. The data extraction algorithm can be tested against other data extraction algorithms by using the synthetic image as a source image. However, the data extracted and compared to the existing cosmic ray database is extracted from the original image, and not the synthetic image.

Chapter 4

Method

In this chapter a description is given of the steps taken to answer the research question. To do so as accurately as possible the method was implemented using the records of the ground-level enhancement event of GLE #5. This GLE was chosen due to the availability of neutron monitor data of the same event. The historic photographic paper had already been scanned for the NGDC, and as such the extraction was to be done using digital image processing. This data had also been analysed by Forbush (1957) providing another reference against which the method could be verified.

4.1 Development of the algorithms

The decisions made in constructing the algorithms are described in the following sections, detailing the process followed, the measures used to decide the success or failure of the study, rationalization of the data used, against which to test the analysis of this work. The aim of this study was to accurately digitize (convert into values) the cosmic ray data. Priority was given to accuracy, and all the steps laid out were based on measurable properties of the images.

4.1.1 Process

The processes that were followed are described below:

1. Extract the ground truth image (extraction algorithm):
 - (a) Remove sprockets from the input image.
 - (b) Pre-process the desprocketed input image.
 - (c) Binarize the image produced in the previous step.
 - (d) Identify hour marker locations.
 - (e) Divide the input image up into hour segments, for individual analysis.
 - (f) Identify scale lines.
 - (g) Analyse the scale lines and reduction of scale line intensities.
 - (h) Identify data lines using the image obtained from the previous step.
 - (i) Produce a ground truth image, using the located hour markers, scale lines, and data lines.
2. Generation of the synthetic image (reproduction algorithm):
 - (a) Map distortions and optical artefacts found in the original images, as well as using an inpainting technique to produce the texture of the synthetic ground truth image.
3. Test and improve the extraction algorithm:
 - (a) Apply extraction algorithm to the synthetic image to extract features.
 - (b) Compare the extracted features to the ground truth image, using MSE, precision, recall, and accuracy.
 - (c) Iterate this phase until satisfactory results are obtained.
4. Use the extraction algorithm to extract ionisation data from historic (original) cosmic ray image.
5. Calculate the percentage increase values for the data line extracted from the original image.
6. Compare the percentage increase data to existing data for GLE #5, obtained from the GLE database (Usoskin, 2016), and Forbush (Forbush, 1957).

A detailed description of the operations used is presented in the following pages.

4.1.2 Measure of success

Ground truth images and synthetic images were created to test extraction algorithms. Although in this case the extraction algorithm was the same algorithm that built the ground truth data, the synthetic image did include some differences, and as such this comparison between the test data, and the original ground truth image was mostly for comparison to other extraction techniques such as the techniques proposed by Drevin (Drevin, 2008). The calculated percentage increase above the pre-event (background) cosmic ray level was compared to neutron monitor data of the same event, as well as to the data from Forbush (1957) and Usoskin (2016).

4.1.3 GLE #5 data

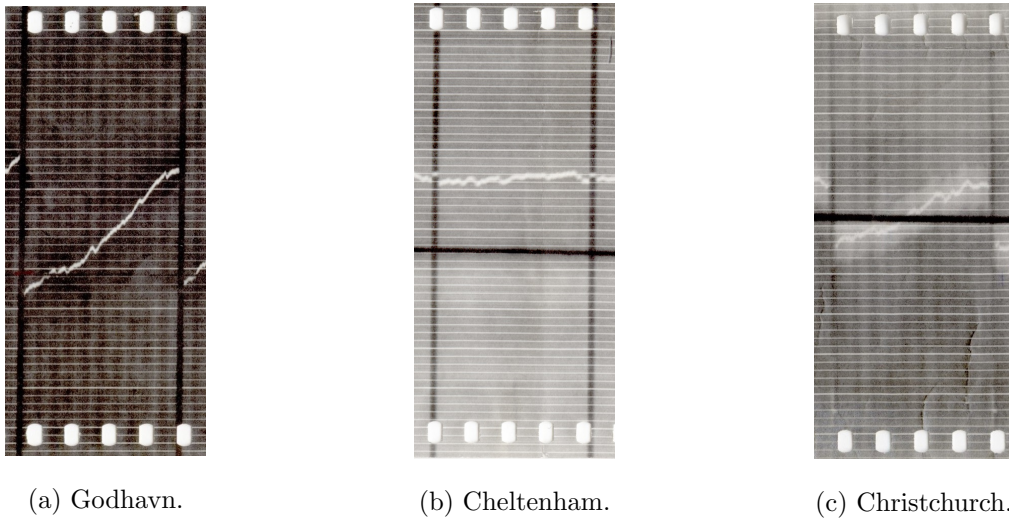


Figure 4.1: Typical hour sections for each of the available stations documenting GLE #5.

GLE #5 on 23/2/56 was recorded by the Model-C recorders at Cheltenham, Christchurch, and Godhavn.

The Godhavn station recording gave good results, due to the high contrast of the image (Figure 4.1a). A few scale lines were apparently missing, but this was due to either the pressure line (in black), overlapping the scale line in the input image, or areas of varying contrast obscuring the scale lines.

The Cheltenham station recording, with its brighter background, and lower contrast also provided good results due to the clarity of the ionization data line (Figure 4.1b). The damage to the sprockets also contributed to the slightly worse scores than the Godhavn station image.

The Christchurch station had low contrast, faint hour markers, damage to the sprockets, and film, having been cut, and repaired with adhesive tape. All of the stations were recording at regular speed, with black hour markers. These images were unedited and are referred to as the original input images (Appendix A).

4.1.4 Prior knowledge

There were a number of details about the photographic paper, and the Model-C recorders which could be used to make the extraction of features easier. This a-priori knowledge is discussed in the following sections, relating to each feature individually.

This prior knowledge, when programmed into the algorithm could be considered as an expert system, specialised in the extraction of these specific features (hour markers, scale lines, and data lines). The method implemented was similar in approach to the method used by Rigaud *et al.* (2015) in extracting content from comic strips. The first step was to identify the basic structure of the image, then make use of the a-priori knowledge to produce temporary results. These temporary results were then refined by the same process, yielding validated results. This process is illustrated in Figure 4.3 using a version adapted from Rigaud *et al.* (2015).

Figure 4.2a is the process used in this study revealed a number of similarities which were used by the inference-engine to interpret the image (finding hour segments and average scale line separations). These ‘deductions’ were made using a-priori knowledge built into the artefact as well as statistics from the image itself, and could be called the knowledge base. The combination of the inference-engine and the knowledge base made up the expert system which interpreted the Model-C recordings. From this process a ground truth image was created, as shown in Figure 4.2a, as described by Step 1. This image was then subject to some alterations to create the synthetic image, Step 2. The resulting synthetic image was then processed by the same extraction algorithm (the expert system) to obtain a result image which can be compared to the ground truth image, Step 3a. This result image was then compared to the ground truth to determine the accuracy, precision and MSE of the extraction algorithm, Step 3b. This comparison of the

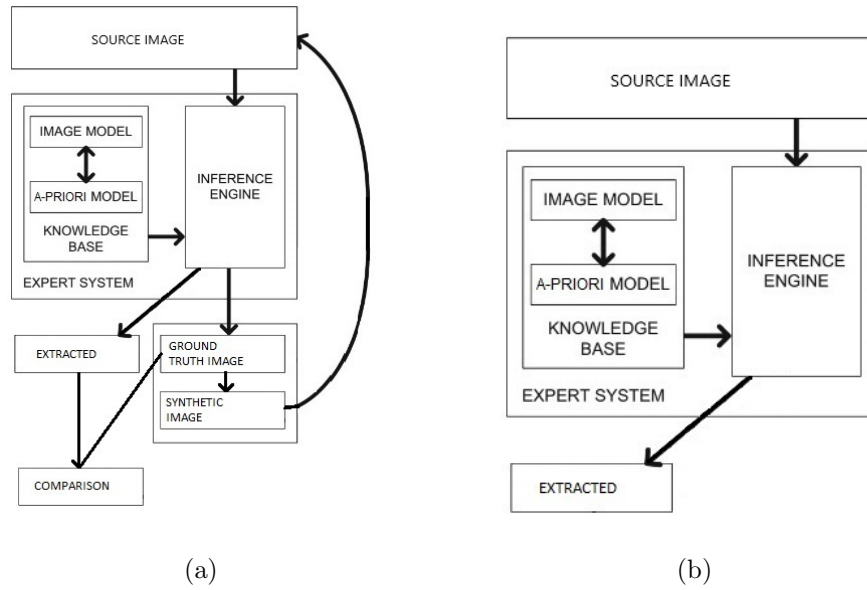


Figure 4.3: Diagrams showing the relation of the algorithm to the expert system in different steps in the process (Adapted from Rigaud *et al.* (2015)).

extraction algorithm's results can be iterated to tweak the algorithm as needed, Step 3c.

Once a satisfactory extraction algorithm is decided upon, the ionisation data can confidently be extracted from the source image. This process is illustrated by Figure 4.2b. This step correlates to Step 4 in the process used. This ionization data can be converted into a more useful measure, describing the percentage increase above background cosmic ray levels as described in Step 5 of the process. To verify the extracted data, comparison to some form of ground truth is preferred, and in this case, the neutron monitor data acted as the ground truth values for GLE #5.

Of the three images used, the image from Godhavn was in the best condition (Figure A.1a, p. 82), with high contrast between the data lines, scale lines, and background. The last hour marker section was far lighter than the rest of the image's background. This image had some damage at about 2/3 of its length, which had been repaired with adhesive tape applied vertically.

The second image, from Cheltenham (Figure A.5a, p. 86), was of intermediate quality, had fair contrast, and an even tone in the background, but some of the data lines were very faint, if not impossible to differentiate from the light background. This image had some damage along the sprockets.

The third image, from Christchurch (Figure A.9a, p. 90), was in the poorest condition, with very little contrast between the background and the hour markers. The barometric pressure

also overlapped with the data, obscuring a good section of the recordings. This image also had some damage at about $2/3$ of its length, which had been repaired with adhesive tape applied diagonally.

The sprockets were located horizontally in the top and bottom 80 pixels of the image. They had an average separation of 45 pixels along the horizontal axis. They were approximated by an ideal sprocket mask, of size 36×52 pixels.

The scale lines were horizontal lines across the image, with a height of roughly 3 pixels. They were separated by 13 – 16 pixels. There were 60 scale lines per image. According to Compton *et al.* (1934) every 5th scale line was brighter than the others.

The hour markers were vertical lines across the image, with a width of 15 pixels on average for the regular recording mode. The average separation of the hour markers was 285-310 pixels, depending on the recording station, and hour marker alignment. There were 19 hour markers per image.

The cosmic ray percentage increase data was contained in the gradient of the ionization values. These ionization values were recorded as an irregular white line, beginning at the same position on each hour marker. The data line was the brightest part of the film, excluding the sprockets. Depending on the changes in ionization values the ionization data line might have been thinner than usual if a large change occurs.

The extractions of the temperature and pressure data lines were not in the scope of this study, and were left for further work.

4.2 Image processing

This section deals with the image processing steps involved in the algorithm, which is described below. The image processing was undertaken with a goal in each step. The goal was often to detect and locate a specific feature, such as sprockets, hour markers, scale lines, or ionization data lines.

4.2.1 Sprockets

One can safely assume that the most prominent artefact in the input image would be the holes punched in the paper, known as the sprockets. These holes were punched into the photographic paper by the manufacturers, as these holes were standard and necessary to drive the photographic paper, using a geared apparatus. In this step the aim was to find and remove the sprockets from the input image, so as to better process into binary, as well as a greyscale image. In the film industry these sprockets have undergone a number of developments and reforms and some example sprockets are shown in Figure 4.4. The sprockets seen in our data set resembled the BH shape in Figure 4.4.

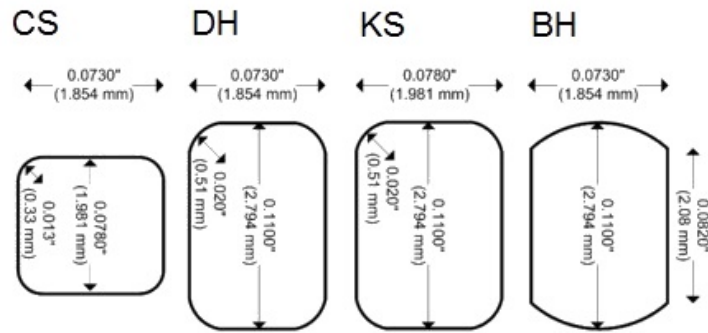


Figure 4.4: Different types of photographic paper sprockets. Smith (2017)

The sprockets produced areas of brightness, or high intensity, during the scanning process. These bright spots disrupted the histogram, and made accurate binarization very difficult. These sprockets had to be located and removed from our working image if any progress was to be made in extracting the other features. Firstly, a good sample sprocket had to be located on the image, as each image had the potential for different sprockets, as the recordings spanned a number of years, and it seems as if the suppliers of the photographic film changed at some point. Locating a good sample sprocket could be done automatically, making the process more user-friendly. This was accomplished by finding the maximum of the correlation function between an ideal sprocket profile, and the input RGB (Red, Green, Blue) image. Once this prime sample sprocket was found, another correlation function was applied to the input image, using the prime sample sprocket as a template. This second correlation function was then thresholded to retain the top 20% of the possible matches. This top percentile was a good indication of where the sprockets lie on the image. An ideal sprockets mask was applied to the RGB input image, and pixels which

corresponded to the mask were set to NaN (Not a Number). This resulted in the sprockets being removed from our image, which was then ready for further pre-processing.

4.2.2 Pre-processing

The input image was ready to be converted to greyscale, after the sprockets had been removed. This step focused on converting the colour image into a greyscale image with good contrast. Due to the size of the image, and the variance of intensity across the image, a block processing procedure was implemented. This block processing procedure divided the image into smaller sections, where the light intensity was more uniform, and the attributes of interest were of sufficient scale. The block size was set at 20x20 pixels, to ensure that no attribute occupied the entire block at once. The resulting greyscale image was then histogram-equalized. This histogram-equalised gray scale image was then ready to undergo binarization, separating the foreground from the background.

A user edited image could also be supplied to the algorithm, to assist in detecting areas of ambiguity. These difficult features included hour markers which were brighter or darker than expected, data lines which were too faint to accurately extract, or unexpected damage to the paper which could have complicated thresholding.

4.2.3 Binarization

The binarization step of the image processing procedure was focused on creating a low noise binary image, with the attributes of the recording in white, and the background in black. The process of binarization was approached by dividing the image into subsections, and regarding them independently of the other subsections (block processing). This was done to negate the effect of varying background intensity over the image. This block processing procedure made use of an iterative process to determine the best threshold. A starting threshold was found using Otsu's method (Otsu, 1975), and applied to the block. The total number of foreground pixels in the binarized block were counted, and this value was used as a measure of the block's brightness, referred to as the percentage of 'ON' pixels in a block. The threshold obtained from Otsu's method was then modified depending on the count of 'ON' pixels in the block.

Calculations by the author show, in the next paragraph, that the foreground elements occupied roughly 20% in an average block, and as such the ‘ON’ pixel percentage should have been similar.

Consider an hour segment 900 pixels in height, and 300 pixels in length. In this segment the hour marker was 15 pixels wide on average, and as such occupied 5% of the vertical space in the image. There were on average 60 scale lines, each about 2 pixels thick. These scale lines occupied at most 14% of the image. The data lines were less easily defined, but could be estimated to occupy about 2% of the image, due to overlap with the already counted scale lines and hour markers. The thickness of the data lines varied from hour segment to hour segment, as well as between different recordings. As such the ideal value of ‘ON’ pixel percentage in an hour marker segment laid between 19% and 21% ‘ON’ pixels per segment. The segments were not ideal however and this threshold was extended to 16 - 22% depending on the image being investigated.

The threshold was adjusted up or down to move the ‘ON’ percentage closer to the calculated values. This process was iterated until the ‘ON’ percentage was between 16% and 22%, with the upper bound being slightly tighter, to reduce the effect of noise in the background. Once the blockwise process was completed, the binarized image was subjected to morphological opening, removing any ‘ON’ noise in the image.

At this stage it is useful to note that the *width* of a line was measured perpendicular to the line in question. For example, the width of a scale line (horizontal) was the vertical dimension of the scale line, whereas the width of a vertical line (such as an hour marker) was its horizontal dimension. This usage of the term is followed throughout this study.

4.2.4 Hour markers

The next step was to locate the hour markers, or vertical darkened lines in the utilized recordings, using the results of the previous steps to assist this process. The gray scale image was divided into 20 segments, each spanning the height of the image. For each segment, the column totals were determined, and a moving average was found (5 pixels wide) for each column. The column with the lowest moving average was selected as a candidate hour marker, and was transcribed onto the blank output image. In this way, we found up to 21 potential hour markers, even though only 19 of them were true hour markers, due to the first or last one being cut off at

some length. Once the entire image had been processed as described, and the candidate hour markers had been found, the false hour markers were removed, using the a-priori knowledge (the average separation of the hour markers for the current image). Any hour marker within less than 285-310 pixels of the previous hour marker was removed. The exact pixel distance depends on the recording, as an average separation for the hour markers was determined as the algorithm proceeded through the image.



Figure 4.5: Diagram of the inherent skew of the hour markers.

Once the hour markers had been located and recorded on a dedicated image, it was possible to segment the image according to the hour markers. This segmentation was first used to determine the angle at which the hour marker laid, cropping a window 31 pixels wide around each hour marker location. The angle of the hour marker was found by using the Hough transform of the edges of this line. From the resulting lines, the most prominent line was drawn onto a blank image. Once the angle, θ , had been found by using the Hough transform, a new hour marker was drawn in at this angle, however, the hour marker's pivot point was not at the top of the image, but rather near the middle of the image's height. From Equation 4.1, it was possible to find the horizontal distance, $2d$, of a line drawn from the top of the image, to the bottom, at an angle θ . Once this distance was known, the new hour marker was drawn with the pivot position in the centre of the image's height using half this distance, d .

$$d = \frac{H * \tan(\theta)}{2}. \quad (4.1)$$

This new hour marker was then used as a mask, and shifted left and right, until a minimum

intensity was found. At this stage the new tilted hour marker was drawn onto the final hour marker image. These tilted hour markers were used as the ground truth hour markers from this point forward.

4.2.5 Scale lines

Once the hour markers had been found, the histogram equalized gray scale image was segmented along the middle of each hour marker line. This aim of this step was to locate scale lines, in each segment, as accurately as possible. In this way each hour segment was inspected individually and the continuous scale lines and data lines were located. Due to the varying intensity of the scale lines on the images, the decision was made to approach each hour marker segment as an individual set of scale lines. The scale line location process was also a block processing technique, using blocks that spanned the width of the image, and allowed 60 such blocks to fit into the image. This ensured that roughly 2 scale lines were visible in each block. The row sums of the block were then obtained, and a running average of 3 pixels was determined for each row total. The maximum row in the block was marked on a black output image as a candidate scale line. Once all candidate scale lines have been marked, the output image was checked against a-priori knowledge, such as the average separation, width, and count. False scale lines were removed, and estimated scale lines were inserted where scale lines were missing. These candidate scale lines were then individually adjusted to lie on the local maxima of column intensity. This resulted in a collection of horizontal lines which approximated the input image. These lines were then used as a guiding rail, for finding the real scale line per column. This technique used a candidate system spanning 5 pixels above and below each guide scale line, with the moving average of 3 pixels in width. This process made it possible to obtain an accurate description of the local maxima along the scale line approximations, which reflected the scale lines.

4.2.6 Data lines

This step was concerned with locating the ionization data line in each hour segment. The scale lines were the brightest regions on the image. Therefore, utilizing the regional maxima of the greyscale image. it was possible to locate the ionization data. However, the scale lines were also registered as regional maxima, and as such the scale lines needed to be eliminated. Care

was taken in removing scale lines, as the scale lines overlapped with the data. The solution was to subtract the average scale line intensity from a small neighbourhood around the scale lines that have been found previously. In this way the brighter region of the scale line where the data line crossed remained bright, whereas the rest of the scale line was far less intense. The regional maxima of the grey scale image were then de-noised using morphological opening, with a structuring element of 25 pixels in size. The remaining region was a good indication of where the data could be found. This region was then used as a mask, and the column-wise candidate process was undertaken again, this time with a moving average of 5 pixels in size. This returned a data line for regions detected by the regional maxima algorithm provided by Matlab.

4.2.7 Ground truth image

Here the attributes located in previous steps were combined to create a ground truth image. These extracted elements were then combined into a ground truth image, where the hour markers, scale lines, and the ionization data lines were represented. The hour markers were 5 pixels wide, and the scale lines were 1 pixel wide. The ionization data line was also a single pixel wide.

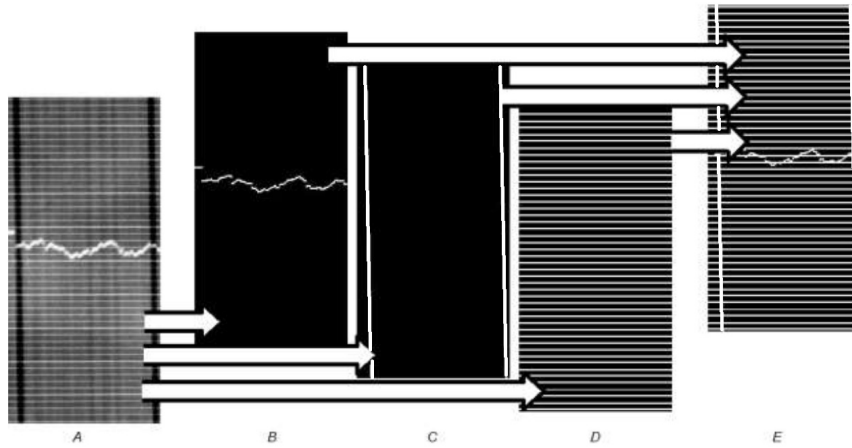


Figure 4.6: Process of creating the ground truth image from the original image (A). The data line (B), hour markers (C), and scale lines (D) were extracted and combined into the ground truth image (E). This image used an hour segment for illustrative purposes, the algorithm was applied to full size images. Adapted from Mattana *et al.* (2014).

4.2.8 Inpainting

This step focused on using the ground truth image created in the previous step as well as the input image to fill in the background texture. The technique, inpaintNaNs (D’Errico, 2012), was used to create an estimate of the input image, based on a seed image consisting of the ground truth image, and the input image. More specifically the intersection of the ground truth image and the input image produced the seed image. This seed image was then inpainted, filling any NaN areas on the seed image (between scale lines, hour markers, and data lines) with appropriate colour using a Δ^2 approximation (D’Errico, 2012).

4.2.9 Synthetic image

The aim of this step was to apply other distortions to the image from the previous step, creating a synthetic ground truth image. Once the texture had been inpainted, the data line was drawn onto the synthetic image. However, the data line was first distorted by adding a Gaussian function to approximate the absolute *sinc* distribution (Bracewell, 1986) that a band of interfering light would have. This was done to replicate the process used in the Model-C recorder where light would pass through a narrow slit, and be recorded by the photographic paper. The line would show interference bands on a small scale, due to the slit through which the light passed. This smoothed data line was then added to the synthetic image. The synthetic ground truth image was complete at this stage.

4.2.10 Accuracy, precision, recall and MSE

As a proof of concept, and for reference purposes, the resulting synthetic image was processed in the same way the original input image was, undergoing the entire extraction algorithm process again, (including a user edited guide image) to obtain the features of interest. These images were compared to the ground truth image. The features extracted from the synthetic image were then compared to the ground truth image using the MSE, as well as precision, recall, and accuracy. The results of the calculation of these measures for the three images, documenting GLE #5 were tabulated in Table 5.1 on p. 67.

4.3 Extraction of data from original image

This phase of the study focused on extracting the percentage increase above background cosmic ray levels from the ionization data lines. The extraction process involved the conversion of the ionization data line to a binned data line, from which the gradient could be calculated between each binned interval. These gradient values were used in the calculation of the percentage increase above background ionization levels.

4.3.1 Binning decisions

As the percentage increase above background cosmic ray levels was intrinsically a gradient measurement, some binning method was required. A number of binning alternatives were considered, such as, to obtain regression values for the bins, or to use the average value, or to use the median. The question was raised whether the binning should be done on the ionization data, or on the gradient data. The option of using the median before finding the gradient was selected, as this reflected the data well, and smoothed the data, rather than exacerbating outliers, as regression or averaging techniques would have done.

4.3.2 Conversion, sampling and quantization

The conversion step was directed at converting the binned ionization data to percentage increase data, and the procedure that was followed is detailed below. Once the data was located the ionization value coordinates were identified. These coordinates were saved as a comma separated file, containing the pixel location of the data point, with the origin in the bottom left of the original image. If there was no data point, the value would be saved as NaN. The median over 5 minute intervals, using approximately 24 pixels, was used in Equation 2.1 (p. 24). This median value was used for sampling and quantization, to discretize the data. The centre of the adjacent 5 minute interval was used to calculate the percentage increase above the background levels. The background level was determined using the first 3 hour sections on the recording. The percentage increase above the background cosmic ray level was calculated on an hourly scale using I_0 in Equation 2.2 (p. 24).

Five minute intervals were chosen because the data recorded on the GLE database was recorded in 5 minute intervals, whereas the hourly interval was chosen because the data published by Forbush were divided in hourly intervals. In some instances the data would be marked as NaN. This missing data could be attributed to data lines on the input image being broken or overwritten by temperature lines, or such a sudden decrease in the readings that the data line reached the bottom edge of the photographic paper before the end of the hour segment. Unfortunately in the latter case the data which was missing was of much interest in the study of GLEs. For the sake of accuracy, no attempts were made to interpolate or extrapolate this missing data, and the missing sections were labelled as NaN.

4.3.3 Comparison

Two sets of data were used in the comparison, the first being neutron monitor data from the GLE database (Usoskin, 2016) (Figure 4.7), and the second being ionization data which Forbush extracted by hand in 1958 (Forbush, 1958) (Figure 4.8, and 4.9 p. 60). The data from Forbush's papers was concerned with GLE #1 - #5 (only GLE #5 was of interest in this study), and was not binned in a clear way, leaving the binning open to interpretation. The neutron monitor data was sourced from the GLE database (Usoskin, 2016). The GLE data from the GLE database was in 5 minute intervals, and covered GLE #5 to GLE #72.

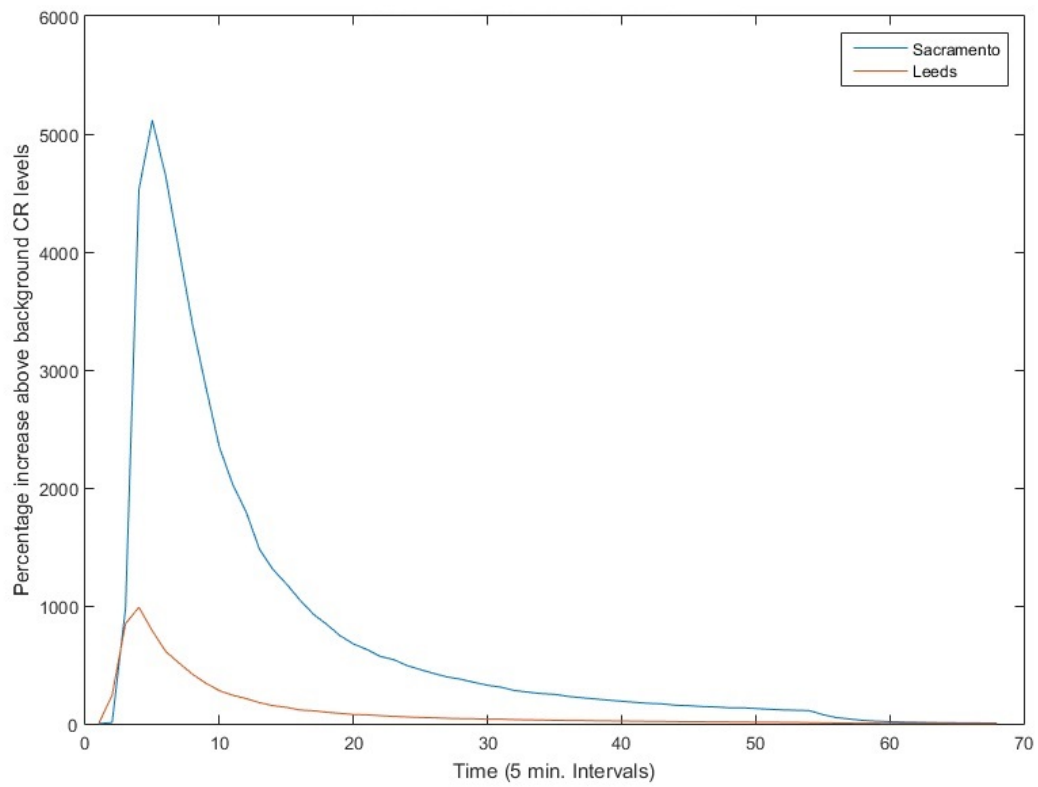


Figure 4.7: The data for GLE #5 from Sacramento and Leeds, sourced from the GLE database (Usoskin, 2016). The time axis begins at 3:50am UTC, with each interval representing 5 minutes.

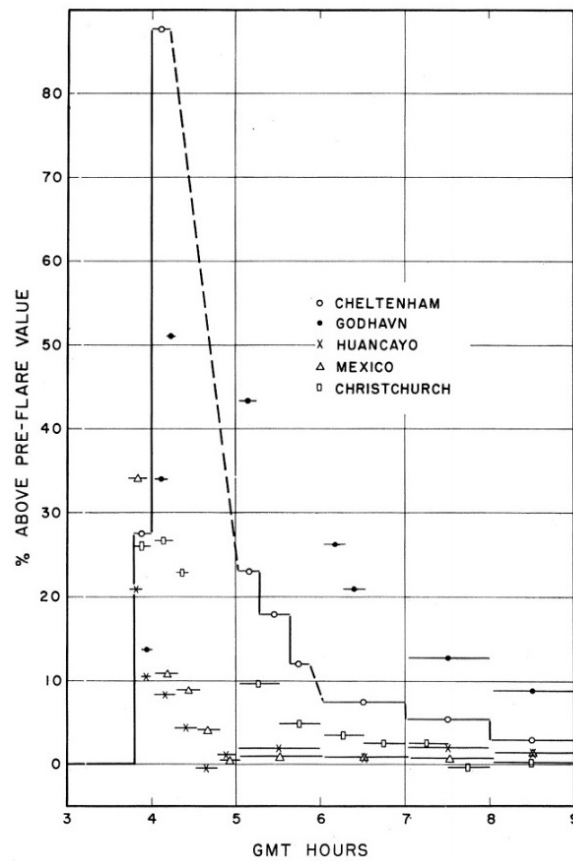


Figure 4.8: The results for GLE #5 as interpreted by Forbush (1957) from observations by various ionization chambers.

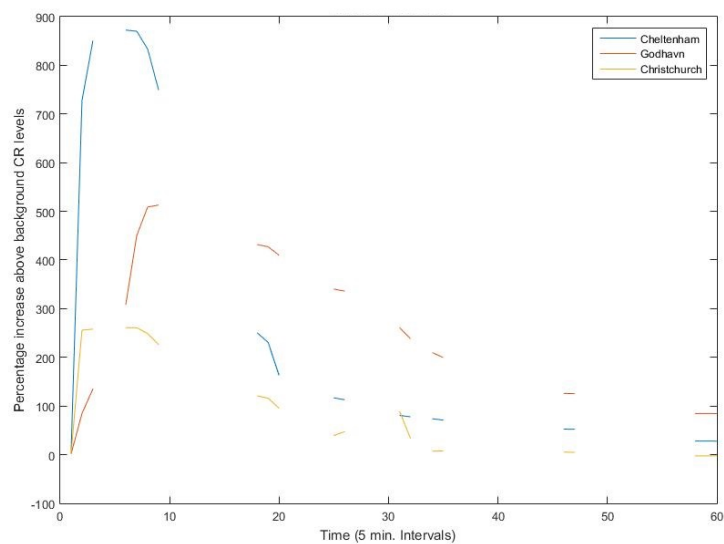


Figure 4.9: The results for GLE #5 recorded by the Model-C ionization chamber, with the time axis beginning at 3:50 am. Digitized by the author.

4.4 Conclusion

This concludes the discussion of the method used in this study to automatically create ground truth images as well as synthetic ground truth images. The methods used to extract the data were also discussed. The algorithms presented here can be separated into two distinct groups, the first being the data extraction algorithm (which also produces a ground truth image), and the second being the synthetic image construction algorithm (which is focused on adding texture, and optical artifacts to the ground truth image).

For the purpose of comparing the extracted ionization data to existing records for GLE #5, the data extraction algorithm was applied to the original image, and the data lines were converted to percentage increase above background level.

For the purpose of comparing forthcoming extraction algorithms, the synthetic image can be used to compare the results of data extraction algorithms with respect to accuracy, precision, recall. In the following chapter, the results of this method was examined and discussed.

Chapter 5

Results

This chapter presents the results of the method outlined and described in the previous chapter. The aim was to present the results in an accessible form. In order to give a clear illustration of the results, the analysis of a section of the photographic paper obtained from godhavn is presented. This was done because the original images were too large to make the required details clearly visible. The full size images are presented in Appendix A. In Figures 5.1 - 5.3 (p. 63 - 65), the period 2:30 am - 6:30 am is documented. In Figures 5.4 - 5.5 (p. 66 - 68) the entire original images range is presented graphically. In further images in this chapter, the ionization data extracted was compared to data collected by S Forbush, and to data obtained by neutron monitors.

5.1 Image processing

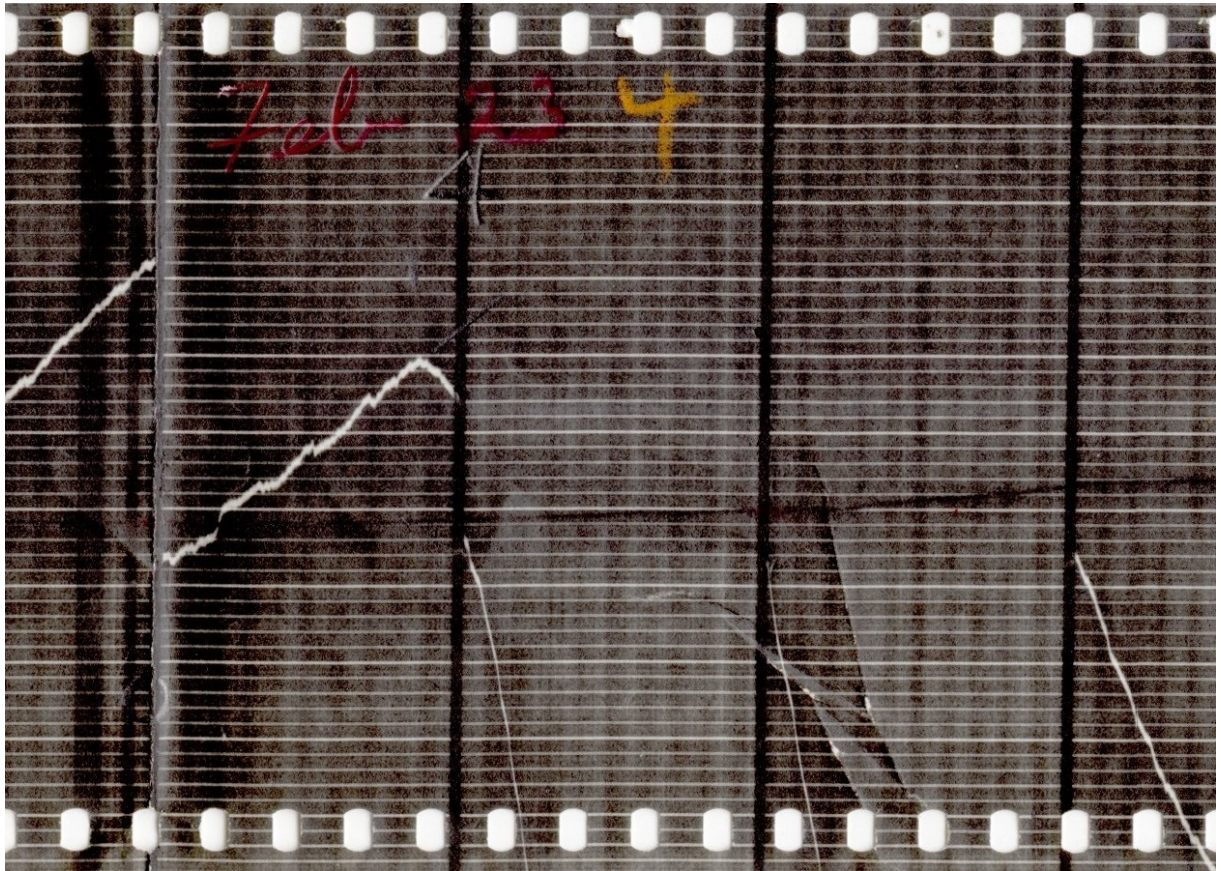


Figure 5.1: The period 2:30 am - 6:30 am on 23/02/1956, recorded at Godhavn, documenting the GLE, seen by a sharp decrease in the ionisation data line.

Only one section of the image was considered in the discussion of the results, so as to focus on the details of the image. The section used to discuss the initial processing steps spans the period from 2:30 am - 6:30 am on 23/02/1956 (Figure 5.1). This section was chosen as it details GLE #5 during its peak activity, and showcases some of the most challenging aspects of the historical data. Some of the data lines exceed the limits of the photographic paper, extending downwards from the 4:00 am, 5:00 am, and 6:00 am hour markers. Note the sharp drop. Also visible is damage to the paper in the form of folds and creases, which can be seen near the data line in the bottom half of the 5:00 am hour segment. There were also notations, such as the time and the date, seen in the upper half of the image, where the time was marked as ‘4’, (referring the 4:00 am hour marker). There was also damage to one of the top sprockets, as can be seen in the upper middle of the 4:00 am segment. The 3:00 am hour marker was also damaged.

5.1.1 Ground truth

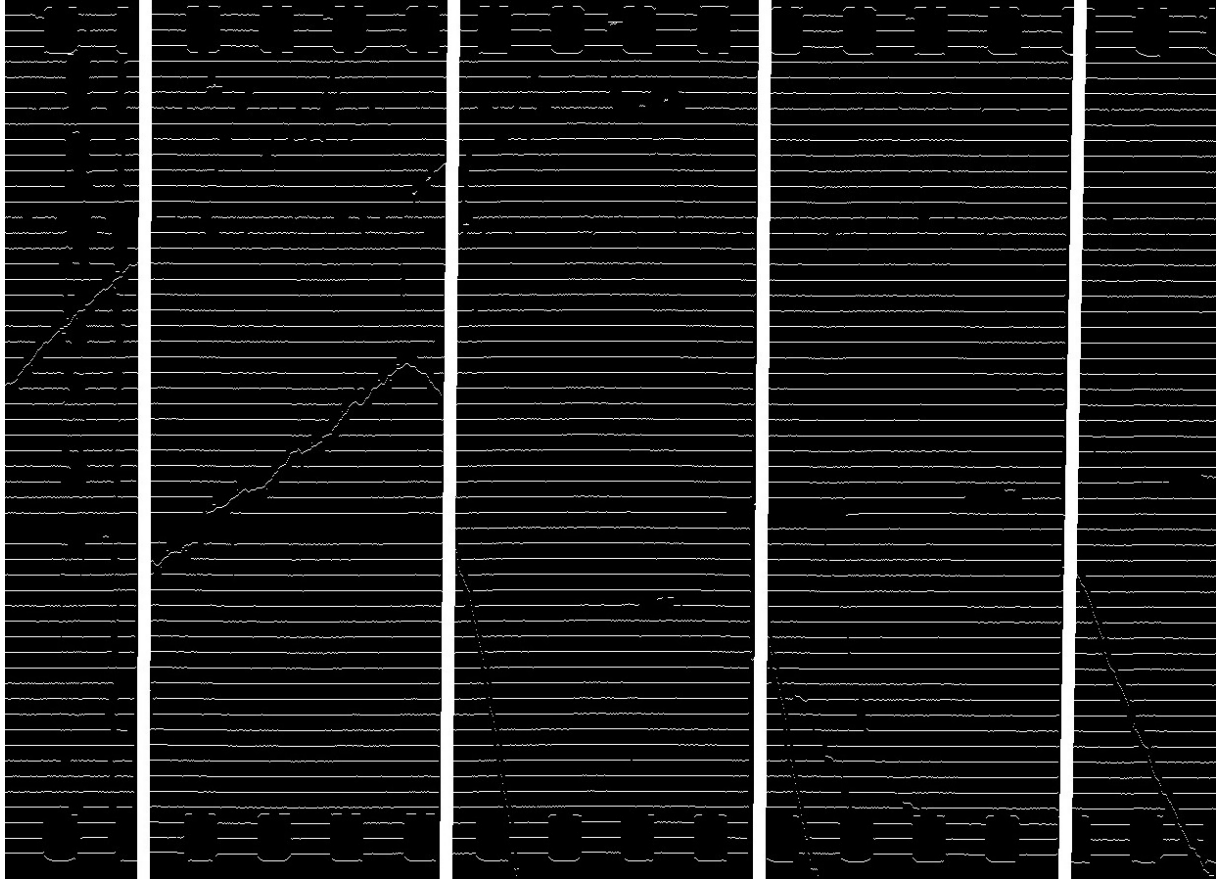


Figure 5.2: Ground truth image, with hour markers (vertical), scale lines (horizontal), and ionization data (nonlinear line).

The result of the image processing steps detailed in the previous chapter yielded a ground truth image (Figure 5.2). This ground truth image represents the structures in the original images, specifically, the hour markers, scale lines, and ionization data line. The hour markers were 5 pixels wide, while the rest of the elements were a single pixel wide. Upon first inspection one will notice missing sections of scale lines. However, if one inspects Figure 5.1 closely, it becomes apparent that there are sections where the dark pressure line obscured a scale line, or part thereof.

The ground truth image was used as a framework to identify the ionization data on a histogram equalized grayscale version of the original image. A sample of a ground truth image is shown in Figure 5.2, while the full images can be seen in Appendix A.

5.1.2 Synthetic image

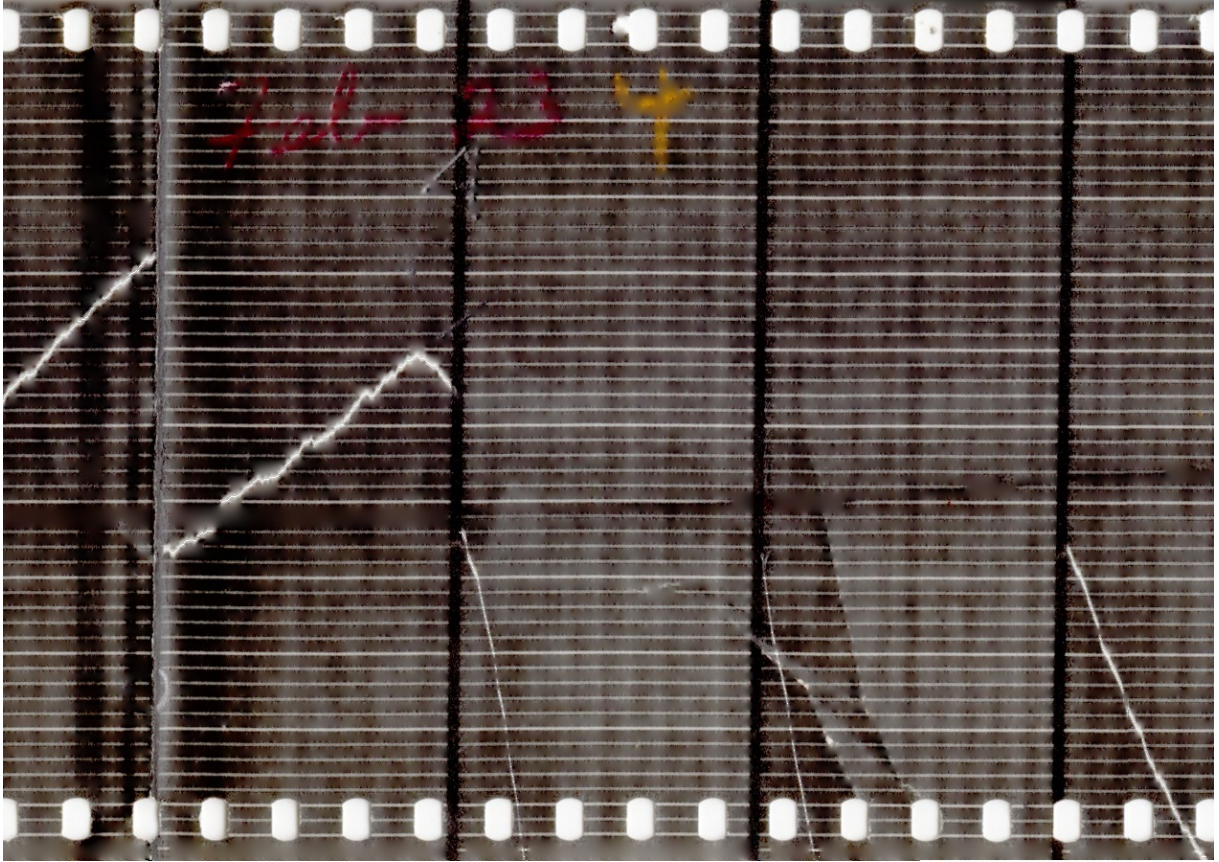


Figure 5.3: Synthetic ground truth image produced by the algorithm.

The ground truth images were used as a skeleton upon which in-painting could be done to reconstruct the texture of the original image. As such, the intersection of the ground truth and the grayscale image was used as a seed for inpainting. The resulting synthetic image could be used to test extraction algorithms, and compare different algorithms, as the source data for the synthetic image was known. The result (Figure 5.3) was visually similar to the original source image. It was also based directly on the ground truth image, which can be of use when testing alternative data extraction techniques.

5.1.3 Ionization data lines

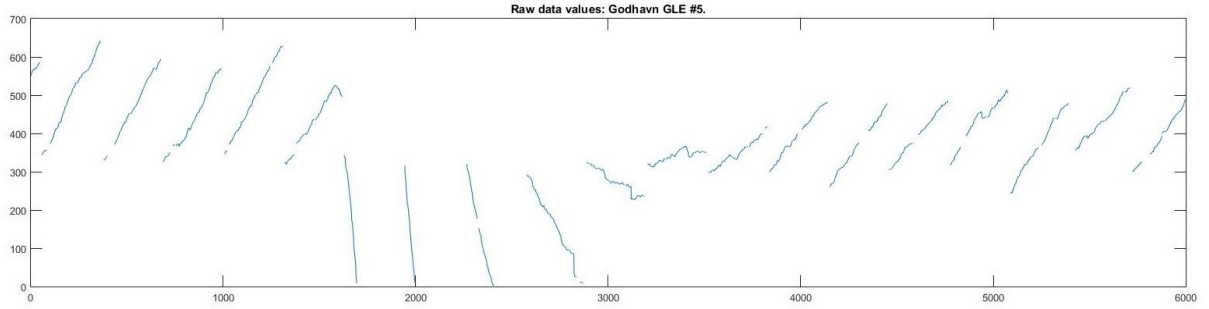


Figure 5.4: The ionization data for the Godhavn image, with pixel height plotted against pixel coordinates, where each pixel on the x-axis represents 12 seconds, and the axis begins at 11:00 pm.

Arguably the most important feature on these recordings was the ionization data lines. These lines represented the ionization recorded by the Model-C recorder, and to obtain the percentage increase which GLE #5 produced, it was necessary to retrieve this attribute of the image as accurately as possible. The raw extracted data lines for the Godhavn observation of GLE #5 is shown in Figure 5.4. These data have been converted to a numeric representation from the data line found in the ground truth, as can be seen in Appendix A. The horizontal axis represents the pixel number on which the data point occurs, however this axis also represented time, with each pixel measuring approximately 12 seconds. The vertical axis was measured in pixel height. There were cases where no data points were located on a line, and as such the corresponding height value was stored as Not a Number (NaN). This had an effect on the discretized interval, and if there were no values in the interval, data would appear to be missing in Figures 5.4 - 6.2 (p. 66 - 75).

5.1.4 Accuracy, precision, recall and MSE

The comparison of the results of the test image to that of the original ground truth, for each of the three recordings investigated, is presented in Table. 5.1. The MSE row shows the MSE (mean square error) between the extracted attributes and the ground truth image, as well as the MSE between the extracted ionization data line and the ground truth data line. The MSE was calculated for each recording. A lower value indicated a more accurate reproduction. The equations (Eq. 3.10 p. 41 and Eq. 3.13 - 3.15 p. 42) were used to calculate the values shown below. The MSE shows the square of the distance between pixels of the resulting data, and the pixels of the ground truth image. The precision accuracy and recall have been defined earlier, and they give a measure of how good the prediction was. The precision, recall, accuracy and F-measure are also calculated from the extracted attributes and the ground truth image, as well as the extracted ionization data line and the ground truth's data line. All of these measures were normalised to lie between 0-1, with a higher value being better.

Table 5.1: Precision of results

	Godhavn		Cheltenham		Christchurch	
Measure	Extract	Data	Extract	Data	Extract	Data
MSE	599.71	1.8203	345.76	20.3916	1446.9	34.3610
Precision	0.9446	0.9822	0.9630	0.8108	0.8758	0.6222
Recall	0.9604	0.9868	0.9824	0.8291	0.8967	0.4813
Accuracy	0.9908	1.0000	0.9947	0.9997	0.9777	0.9995
F-Measure	0.9524	0.9845	0.9726	0.8199	0.8861	0.5428

5.2 Conversion

5.2.1 Ionization data binning

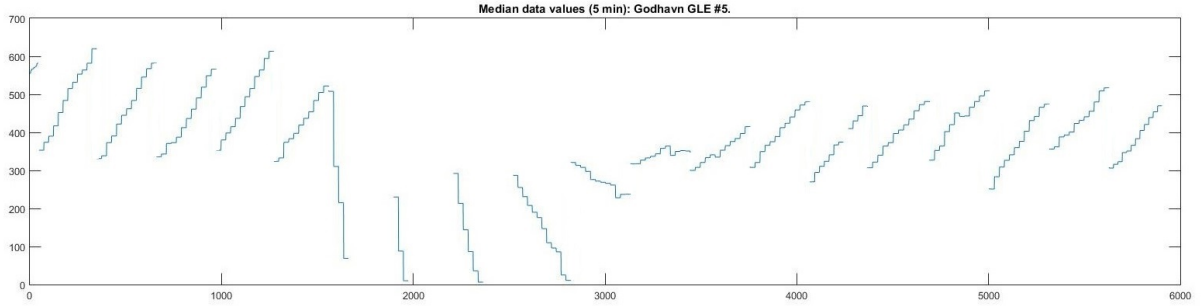


Figure 5.5: Five minute interval binning (28 pixels approx.), with pixel height plotted against pixel coordinates, where each pixel on the x-axis represents 12 seconds, with the axis beginning at 11:00 pm.

The raw location data was subjected to median binning, with each bin representing a five minute interval, and consisting of about 28 pixels (this number varied depending on the width of the hour section). The median was used instead of the average, due to the median's resistance to outliers and noise. The data sourced from the GLE database was recorded in 5 minute intervals, and as such the decision was made to use 5 minute interval binning so as to match the existing data. The extracted data is shown in Figure 5.5. The 5 minute intervals greatly simplified the comparison of the percentage increase data. The constant needed to correct the data for these settings was specific to each recorder, as each individual Model-C recorder was calibrated separately. This resulted in each recorder having a different scaling constant, which meant that the separation of the scale lines had different values attributed to them, based on the recording station's calibration constant. However, the gradient of the data line was where the percentage increase data could be found, using Equation 2.1, p. 24. Thus in this study only the relative increase above background levels was considered. The binned data for all the recordings processed can be found in Appendix A.

5.2.2 Percentage increase

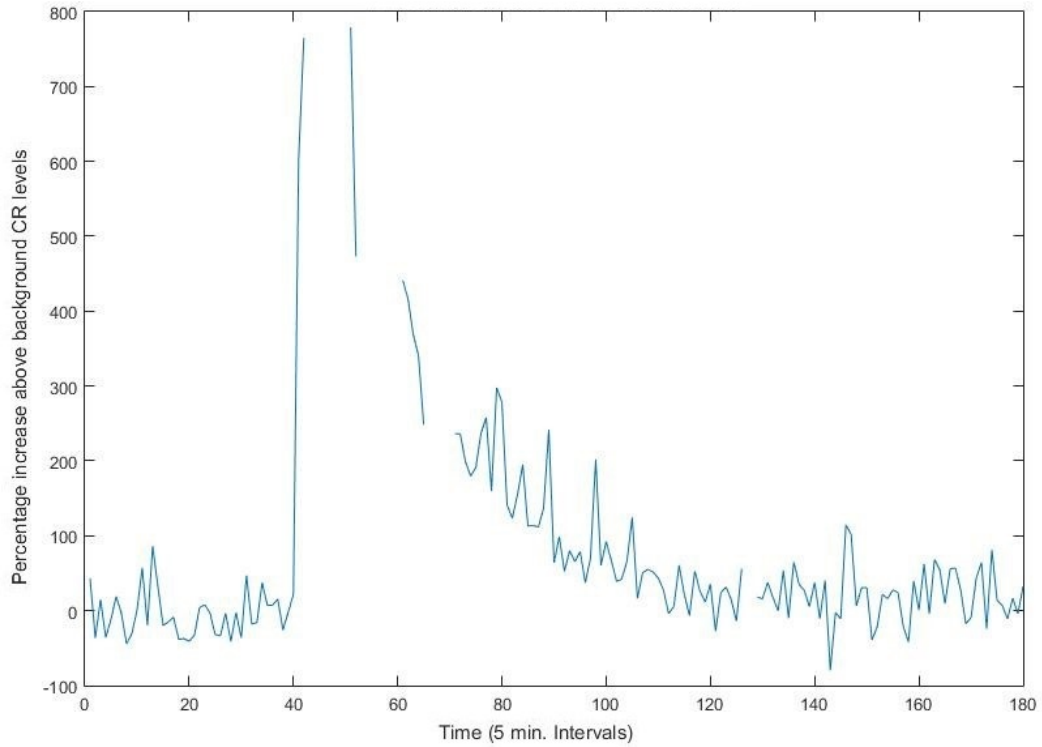


Figure 5.6: Percentage increase above background cosmic ray levels at Godhavn 23/02/1956. The data which shows a sharp increase and slower decrease are indicative of a GLE, with the x-axis representing data bins spanning 5 minutes or 25 pixels beginning at 11:00 pm, and the y-axis shows percentage increase above background cosmic ray levels.

The gradient between each bin (or step) in the data was of great interest. Due to Equation 2.1 (p. 24), the percentage increase over the bin length (five minutes) could be found (Figure 5.6). Figure A.4 shows the percentage increase in the cosmic ray intensity over the background level for the full span of the image recorded at Godhavn. The percentage increase plots for the other stations can be found in Appendix A. The missing segments were due to the data lines reaching the bottom edge of the photographic paper before the Model-C recorder could reset its position in the following hour.

5.3 Comparison

5.3.1 GLE #5

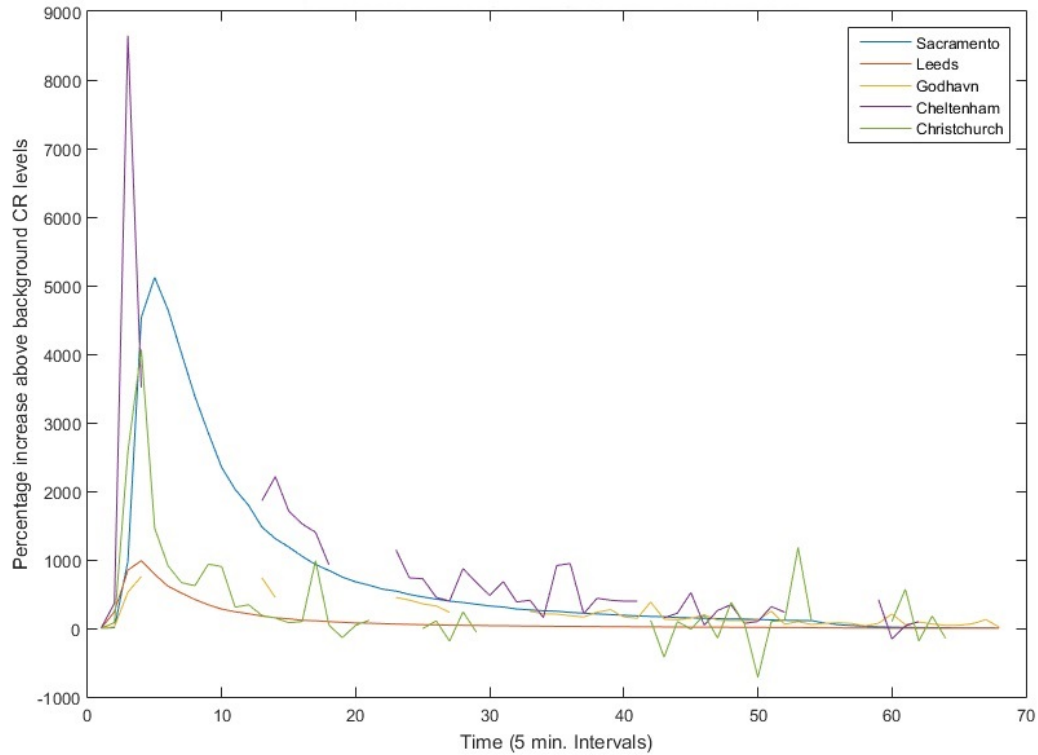


Figure 5.7: A comparison of the data lines (Godhavn, Cheltenham, and Christchurch) to existing neutron monitor records (Leeds and Sacramento) (Usoskin, 2016). The data begins at 3:50 am UTC, with each interval representing 5 minutes, and ends at 9:45 am. The comparison of the results considers the data from beginning of the event, so as to compare the height and shape of the recordings

There are documented results from Model-C recorders, as well as neutron monitors available for GLE #5, and these were the values that the results were compared to. Due to the nature of the data, and calibration constants which were unknown at the time of doing this study (due to fragmented log books), the shape of the percentage increase plot was the main feature of interest (Figure 6.2). The height of the peak could be adjusted by using the correct constant for the particular apparatus which recorded the data. There was potentially a way to calibrate the data using temperature and pressure, but that was outside the scope of this study.

5.3.2 Forbush

To compare the data Forbush extracted relating to GLE #5 the values were multiplied by 10, so as to make the data visible in comparison to the data extracted. Apart from the large difference in the scale of the data, there were some significant similarities. The data Forbush extracted shared a similar shape as it is extracted from the same recordings. The similarity was prominent when referring to the sharp increase and gradual decrease (Figure 4.8, 5.8, pp. 60, 71). The Cheltenham values were also much higher than the values from the other stations, as seen in the extracted data. It was possible that Forbush had access to complete logbooks, and could correct his data, but the values he used were not available at the time of doing this study.

We do see some differences, with the Christchurch values not reaching the same kind of heights we see in our extracted data or in the other records for GLE #5.

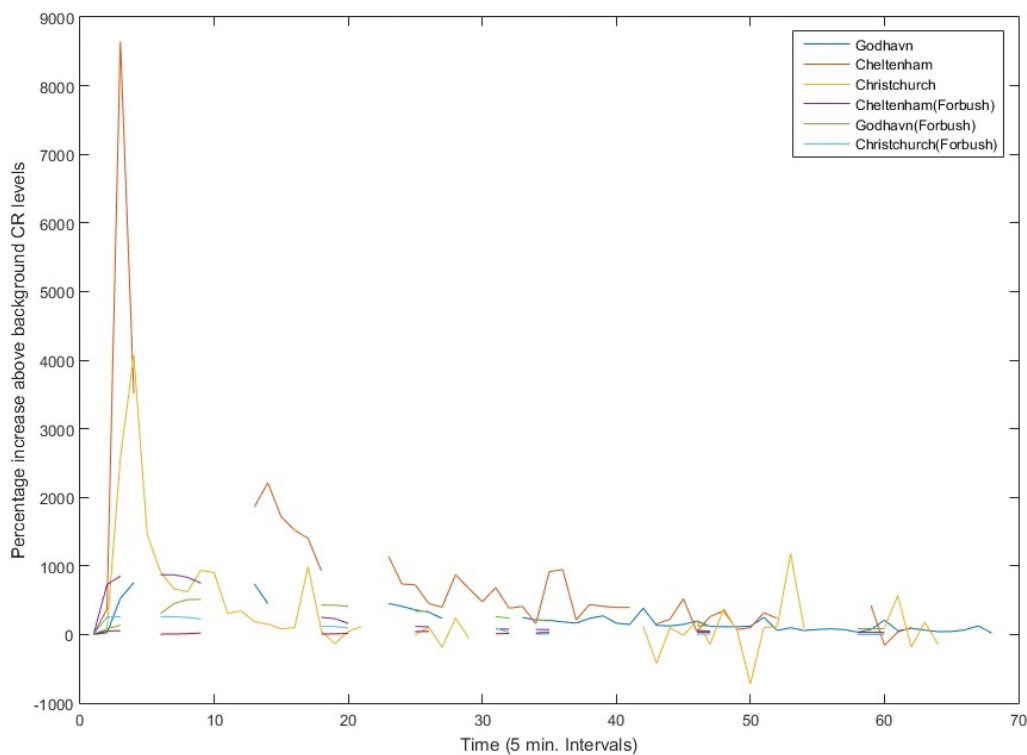


Figure 5.8: A plot of the results extracted from Figure 4.8 (p. 60), by (Forbush, 1957) compared to the data extracted in this study. The data begins at 3:50 am UTC, with each interval representing 5 minutes.

Chapter 6

Discussion

In this chapter the implications, interpretations, failings and successes of the results obtained, are investigated and discussed. A closer look is taken at the results, pointing out areas of interest, confusion, and success.

6.1 Input images and user edits

As the image was split into hour long segments, approximately 300 pixels long, these segments were approached individually. Most segments were segmented with very few errors. Other segments did not fare so well, and scale lines might have been missed, due to either lack of contrast, overlapping pressure line, or damage to the photographic paper. In cases where the data line was too thin to reliably detect using the unaided algorithm, a user edited image was employed to assist the location, which is mentioned again later in this chapter.

6.2 Ground truth and synthetic image

The ground truth image produced by the algorithm was the same size as the input image, and contained the hour markers, scale lines, and data line. The hour markers were 5 pixels wide, while the other elements were all a single pixel wide. The data lines were also saved, recording the height of the ionization data line for each column of the image.

The resulting ground truth image shows the attributes of interest as they appeared on the

input image. There may be inconsistencies between the abstract conception of the attributes and the ground truth image produced. The abstracted properties of the recorded data were accounted for by considering the a-priori knowledge that was available about these attributes. The algorithm used this higher level understanding of the properties, and returned what can be quantitatively located on the input image. As such, damage on the input image may have rendered certain properties unrecognizable; however, clarifying such damage is the purpose of the user input image. The user input image required some insight as to what the algorithm could detect, and what could be missed. This meant that the user would likely have to run the algorithm more than once to supply a satisfactory user input image. This was a section of work recommended for further study, with a responsive user interface for assisting detection.

The resulting ground truth image was then overlaid on the input image, and the dark sections of the ground truth image were set to NaN (excluding sprocket locations). This seed image was then used with the “inpaintNaNs technique”, to produce a textured ground truth image. The data line was then lightly distorted using a Gaussian function, to approximate the interference pattern of the light on the photographic paper. The placement of the distorted data line was the final step in the generation of the ground truth image.

6.3 Accuracy, precision, recall and MSE

The features extracted from the synthetic image was compared to the ground truth image, and metrics such as the MSE, precision, accuracy, and recall were determined. Due to the difficulties posed by the quality of the Christchurch recording, the scores for the Christchurch recording were worse than for the other recordings. The hour markers and scale lines were still extracted with a satisfactory level of accuracy, but the data lines were broken by the pressure line overlapping with much of the ionization data. This also resulted in this image having a significant amount of noise in the percentage increase data. Most of the images scanned for the NGDC were in better condition than the Christchurch recording of GLE #5.

6.4 GLE data

The percentage increase data which was used to compare GLE data was compared to that of neutron monitors from Leeds and Sacramento, as these locations offer insight into the range of GLE #5. The Leeds recording was one of the recordings with the smallest percentage increase for GLE #5, whereas the Sacramento recording had one of the highest recorded peaks for GLE #5, (Figure 6.1). The extracted percentage increases were also compared to the data Forbush collected regarding GLE #5, and can be seen in Figure 4.9 (p. 60).

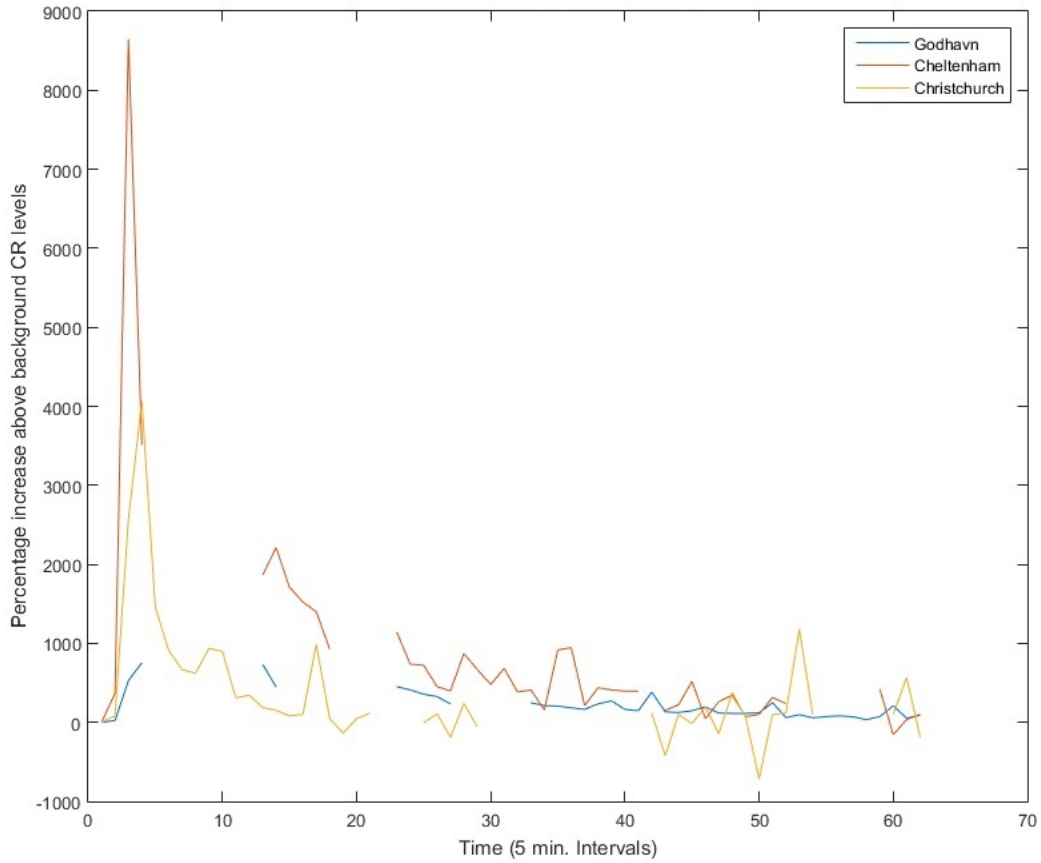


Figure 6.1: A close-up of the extracted data (Godhavn, Cheltenham, and Christchurch) beginning at 3:50 am UTC.

GLE database

The extracted percentage increase above the background level was compared to the neutron monitor data from the GLE database for Leeds and Sacramento as comparative data sources, spanning the range of GLE #5 (Figure 6.2). The important aspect of these graphs was their general shape, as the calibrations of the Model-C recorder were unknown to the author. Normalization constants should be applied to these curves, based on the pressure and temperature of the time and place where the recordings were taken. Additionally, parts of the data set were missing due to the limitations of the Model-C recorder.

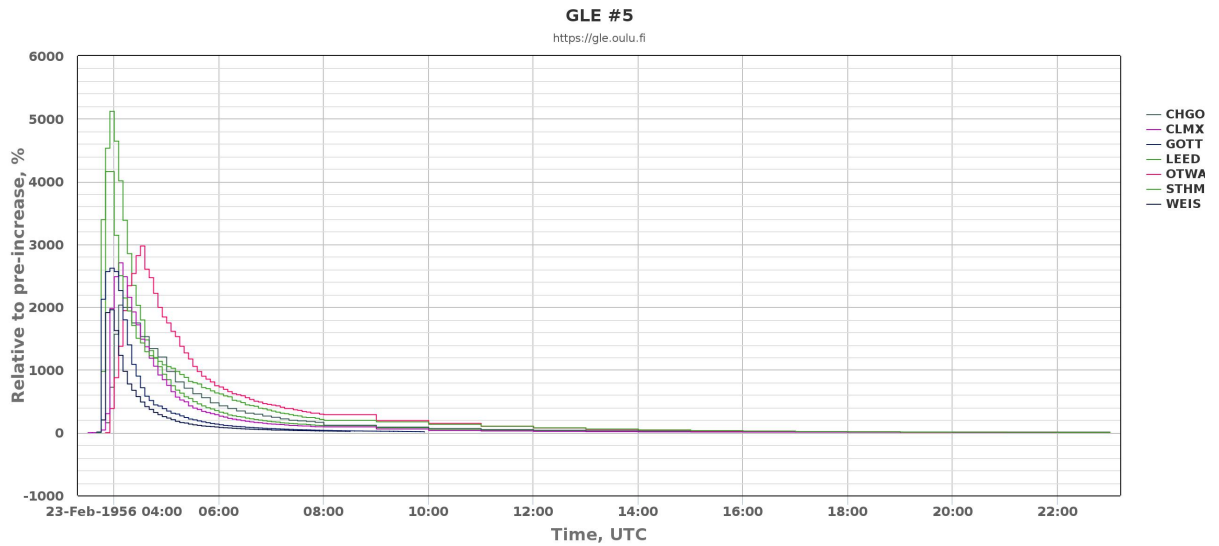


Figure 6.2: A collection of recordings stored by the GLE database.

The images do share common attributes, and these included a sharp rise to a peak which tapers off more slowly. This trend was seen in all three data sets, and can be considered to be consistent with the existing (more accurate) neutron monitor data. Additional discrepancies may be attributed to the inaccuracies of the Model-C recorder itself.

Forbush

The ionization data from Forbush's paper was different to the data we extracted in two major ways; firstly, our percentage increase values were much higher; secondly, our peaks were sharper, rising more dramatically. The high values, observed in the Cheltenham recording, were also represented in Forbush's data. Concerns were raised about the error bars used in Figure 4.8

(p. 60), as they left some uncertainty about the time scale he used to calculate the values. The data extracted from Figure 4.8 (p. 60) was multiplied by 10, so as to compare it to the much larger results (Figure 5.8). We see that the peaks correlate to the same time, but the height of the peaks are considerably different.

6.4.1 Calibration constants

In order to correctly normalize the extracted percentage increase values, certain details had to be known beforehand. These details included the pressure, temperature, altitude, time, and most importantly, any calibration constants used by the recording device. The majority of these details were available for the data in this study; however, the most important details (calibration constants) were missing, as the log books for the recorders are incomplete. This means that even with all the other details, the data will not be truly corrected for. This led to the inspection of the shape of the percentage increase plot, as this remained unchanged with the multiplication of a constant. The main features of a GLE was the sharp increase of readings, and the subsequent tapering off and return to background levels after a few hours.

6.4.2 Ionization chamber inaccuracies

The neutron monitor gained popularity for a reason, and that reason was that it recorded more accurately than the ionization chamber predecessors. Thus the ionization data that was extracted must be considered as less accurate than the neutron monitor to begin with, and as such a perfect match between the extracted data, and the comparison data was not expected, nor achieved. The correlation between the shapes of the data is nevertheless remarkable, and it is significant in validating the source data.

6.4.3 Cosmic ray anisotropy

The directionality of cosmic rays was significant in the detection of GLEs. Some GLE events are considered to be very anisotropic, and the location of recording can have a large effect on the shape of the data (McCracken, 2016). This may also play a role in the range of data obtained, and may also influence the normalization constants.

Chapter 7

Conclusion

This chapter provides an answer to the research question posed in the introduction. It also suggests directions in which this study can be improved upon, or taken further.

7.1 Conclusion

In conclusion, GLE #5 has been extracted from historical GLE records with an acceptable level of accuracy, using automated image processing techniques. Two algorithms are described, the first being a data extraction algorithm, which isolates attributes of interest, and the second which produces a synthetic image from a ground truth image for use in testing future algorithms.

This study serves as a proof of concept, and validates ground truth images for further extraction of historical cosmic ray records by producing data comparable with the existing neutron monitor data for GLE #5. The synthetic and ground truth images allow these algorithms to be used to test other extraction algorithms.

The study had two main aims, the first being the accurate extraction of historic cosmic ray ionization data. This was achieved by applying the data extraction algorithm to the original image, and converting the ionization data lines into percentage increase values, to be compared to existing data for GLE #5. The second aim was to lay the groundwork for testing more refined data extraction algorithms, by producing a synthetic image, upon which new data extraction algorithms can be tested. This was achieved by using the data extraction algorithm to extract data from the original image and create a ground truth image from these results. The repro-

duction algorithm was then applied to the ground truth image to produce a synthetic image. The data extraction algorithm was then applied to the synthetic image to produce extracted attributes which can be compared to the ground truth image. This process can be extended to accommodate future extraction algorithms to compare.

7.2 Future work

Future work that should be done includes long-term data extraction (daily, monthly, annual values), statistical rationale for binning decisions, user interface for user input on the source image, development and testing of better extraction algorithms, extraction of temperature and barometric pressure, normalization of the ionization data (taking into account temperature and pressure), and finding the correct calibration constants for the Model-C recorders used.

Chapter 8

Glossary

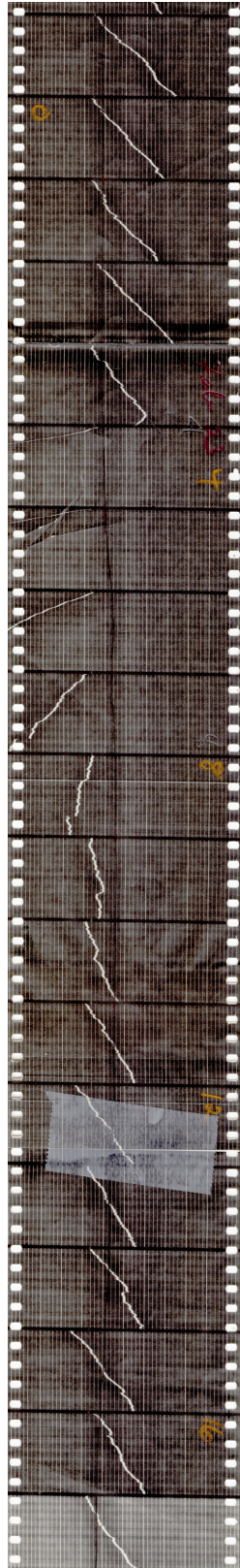
- CR - Cosmic rays, energetic charged particles which originate in stars, and are carried by solar winds. They enter the Earth's magnetic field almost constantly.
- GLE events - Ground level enhancement events, which occur when CRs penetrate the Earth's atmosphere and cause a shower of secondary particles as they collide with our atmosphere.
- SEP - Solar energetic particles, specifically originating from the Sun.
- IMF - Interplanetary magnetic field originating from the Sun, and spiralling outwards, carrying charged particles in the magnetized plasma.
- AU - Astronomical Unit, defined as the distance between the Earth and the Sun.
- NGDC - National Geophysical Data Center, a center for the collection of geophysical data relating to Earth, its oceans, and observations of Earth from space.
- NaN - Not a number, a tag used in Matlab to designate a data point which does not contain numeric data.
- MSE - Mean square error, in this context, a measure of pixel-wise distance between the result and the ground truth.
- GMT - Greenwich meridian time.
- UTC - Coordinated Universal Time, identical to GMT.

-
- RGB - Red green blue, a colour format used to represent colour images as the combination of the three primary optical colours.

Appendix A

Appendix

This chapter is a repository for all the results obtained in the course of this study, and is arranged according to recording station. Thus, all the results from Godhavn will be displayed together, then all the results from Cheltenham, and finally the results from Christchurch. Each station has images for the input data, the user edited data, the ground truth image, the synthetic image, the raw ionization location values graph, binned ionization location values graph, and percentage increase over background cosmic ray levels. The recordings are rotated clockwise at a right angle, resulting in the recordings beginning at the top of the page.



(a)



(b)

Figure A.1: Godhavn station recording of GLE #5, with the first hour marker at 11:00PM, UTC. (a) Original scanned image, (b) User edited version of the original image.

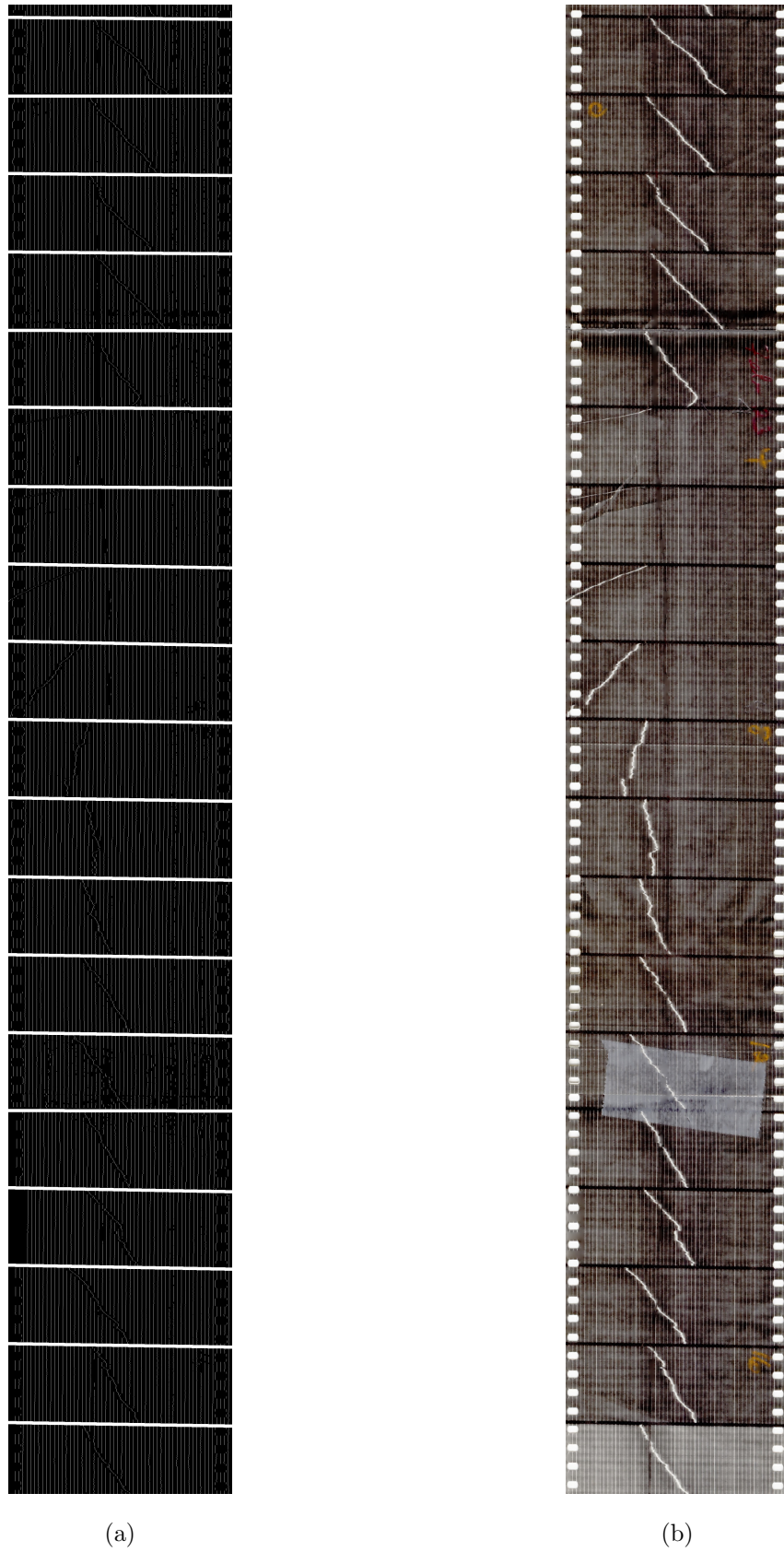


Figure A.2: Godhavn station results of GLE #5, with the first hour marker at 11:00PM, UTC. (a) Ground truth image of Godhavn station recordings, (b) Synthetic ground-truth image of Godhavn station recordings.

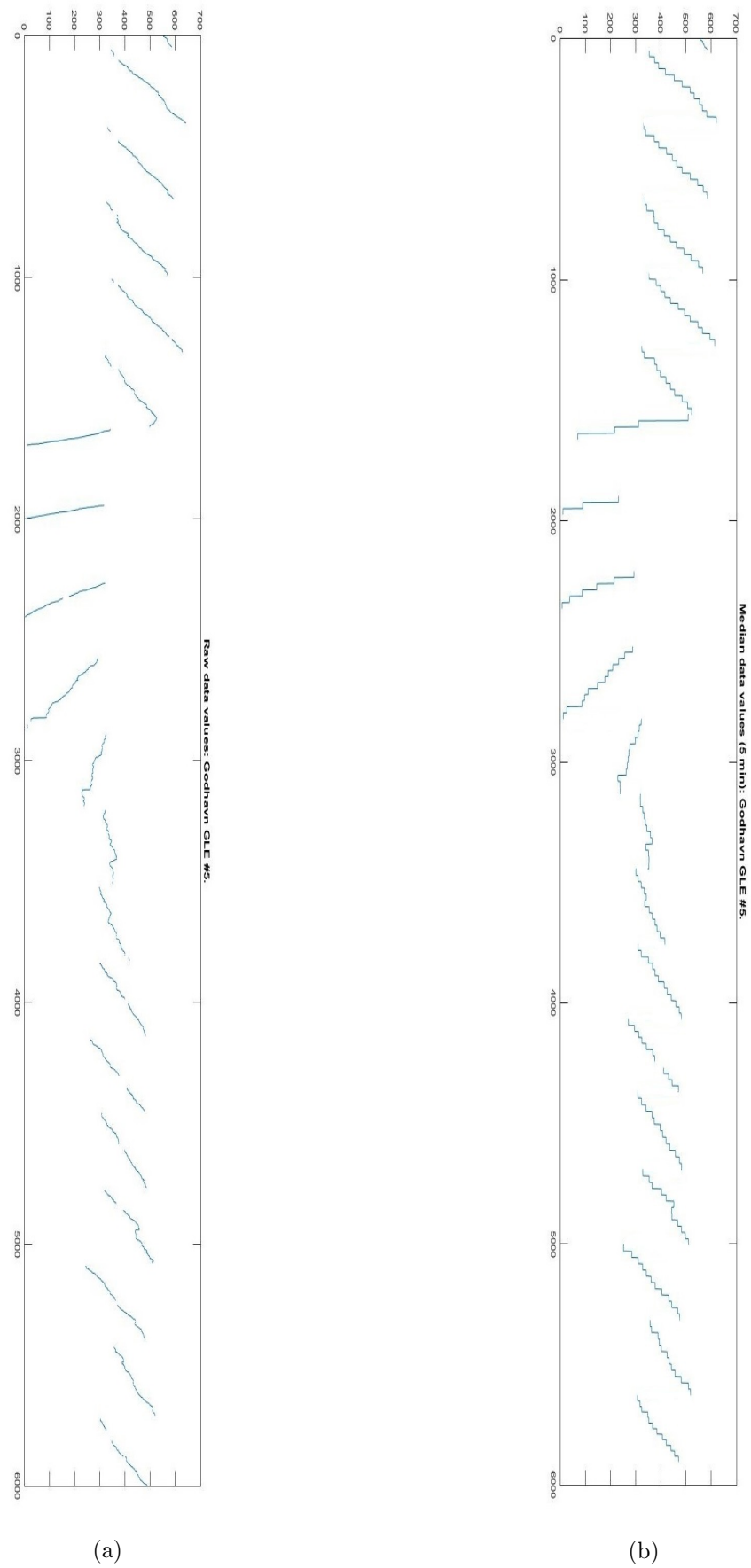


Figure A.3: Godhavn data extraction results. (a) Raw output data line, (b) Data averaged over 5 minute intervals.

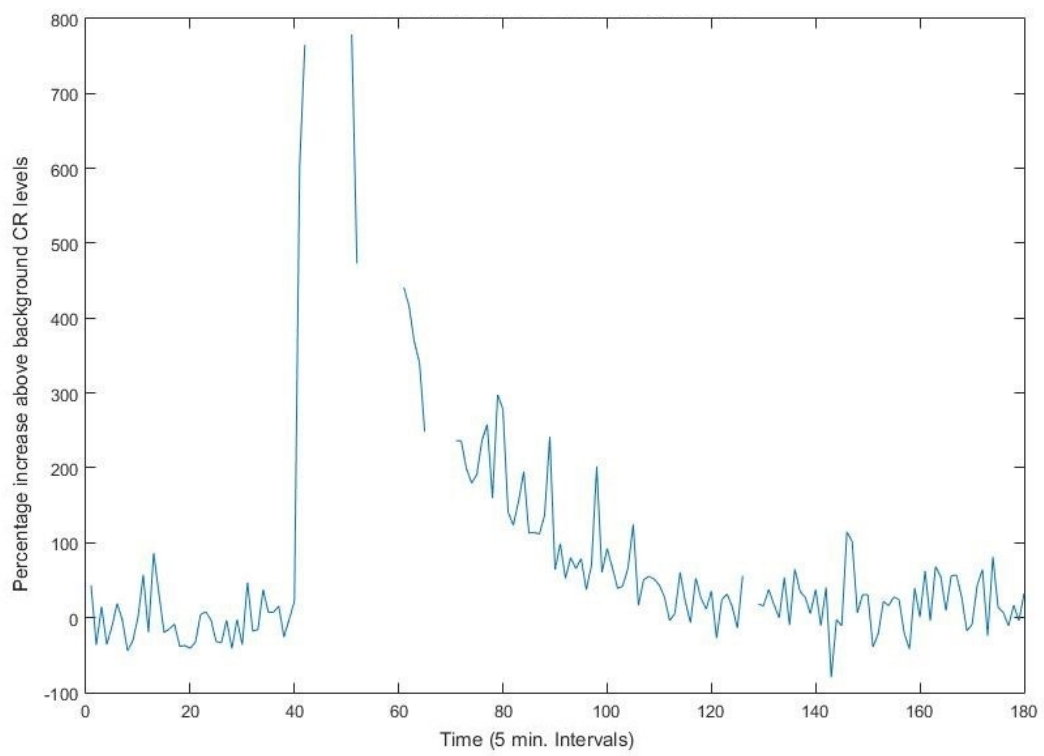
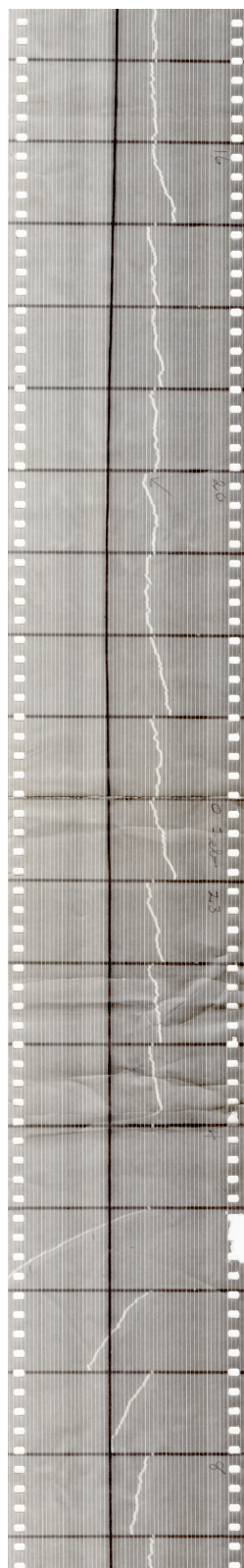
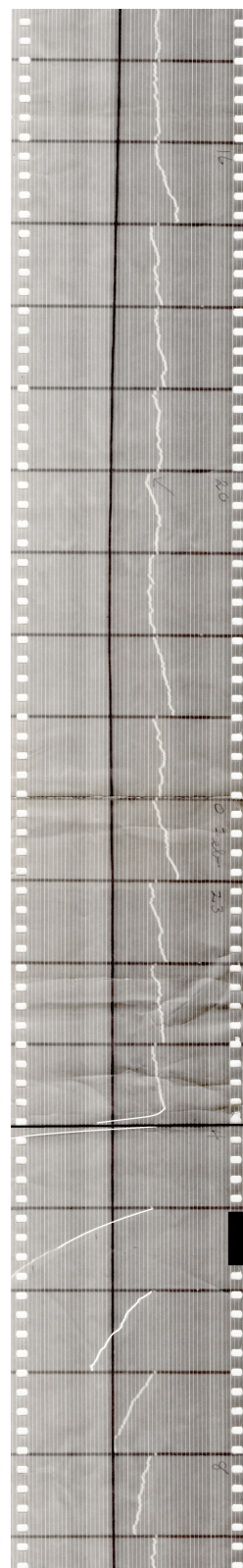


Figure A.4: Percentage increase above background level for Godhavn recording.



(a)



(b)

Figure A.5: Cheltenham station recording of GLE #5, with the first hour marker at 03:00PM, UTC. (a) Original scanned image, (b) User edited version of the original image.

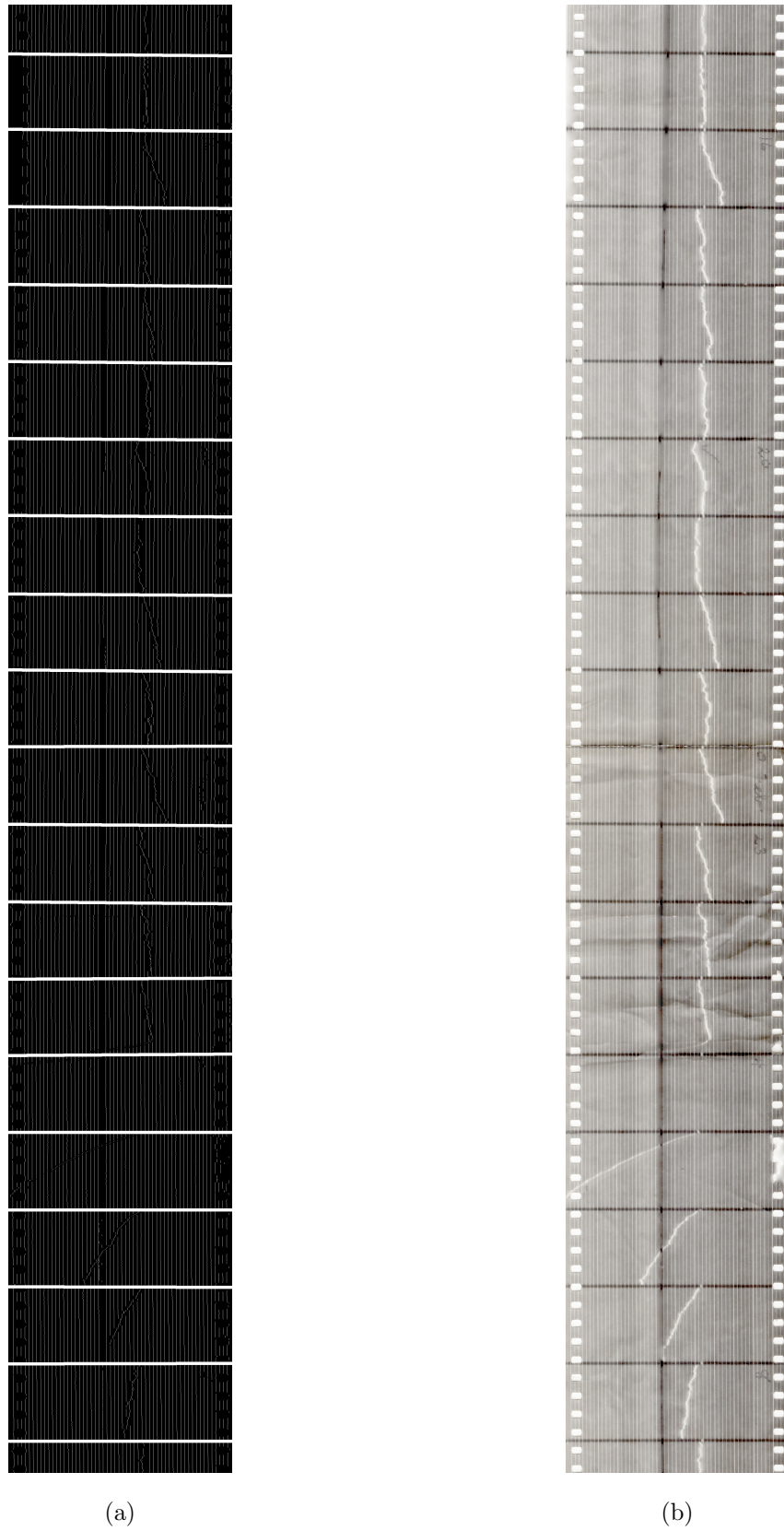


Figure A.6: Cheltenham image processing results. (a) Ground truth image, (b) Synthetic ground truth image.

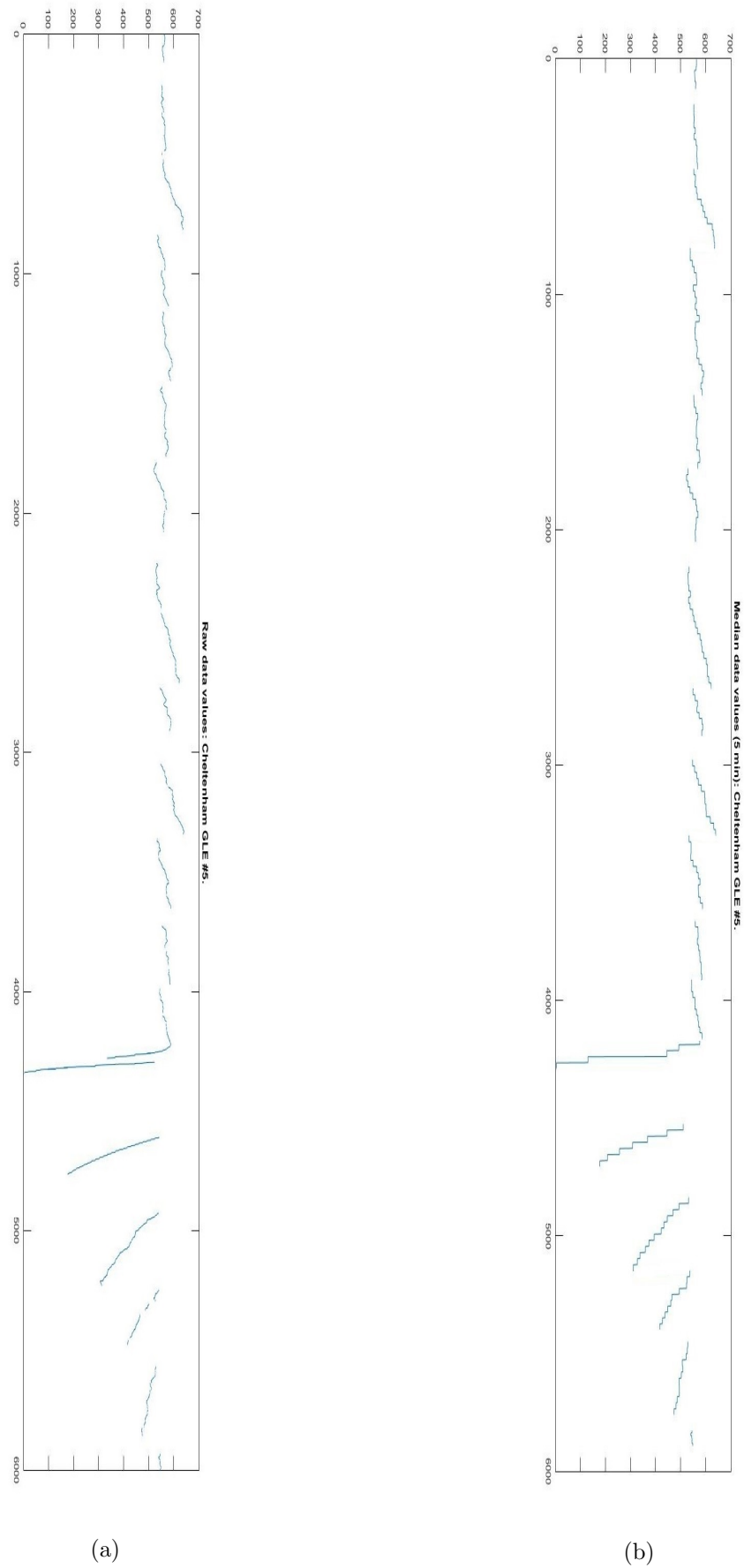


Figure A.7: Cheltenham data extraction results. (a) Raw output data line, (b) Data averaged over 5 minute intervals.

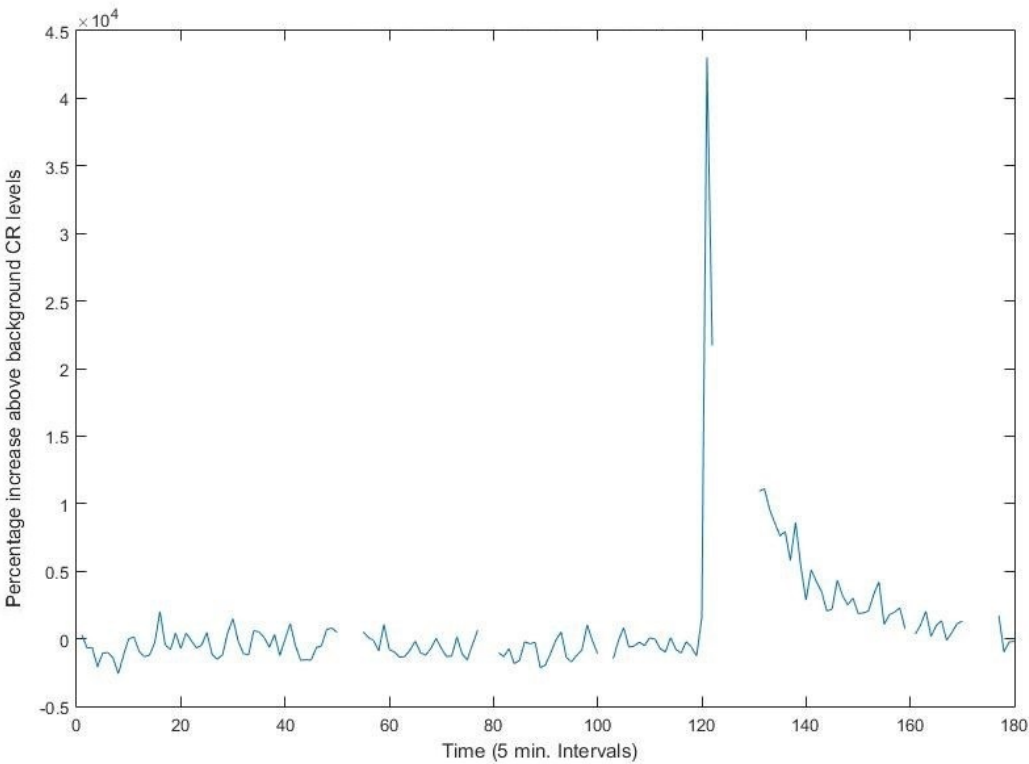


Figure A.8: Percentage increase above background level for Cheltenham recording.

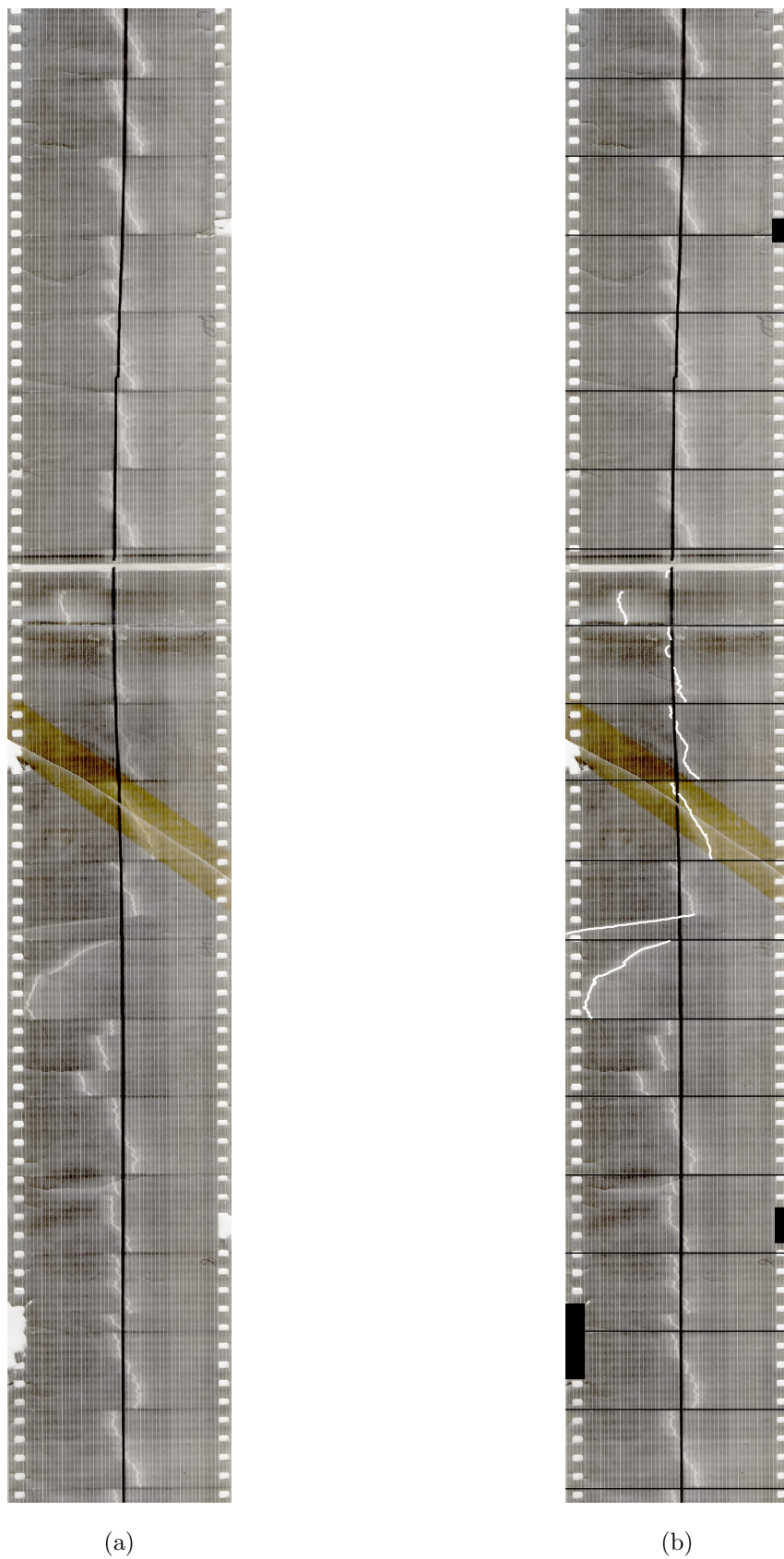


Figure A.9: Christchurch station recording of GLE #5, with the first hour marker at 05:00PM, UTC. (a) Original scanned image, (b) User edited version of the original image.

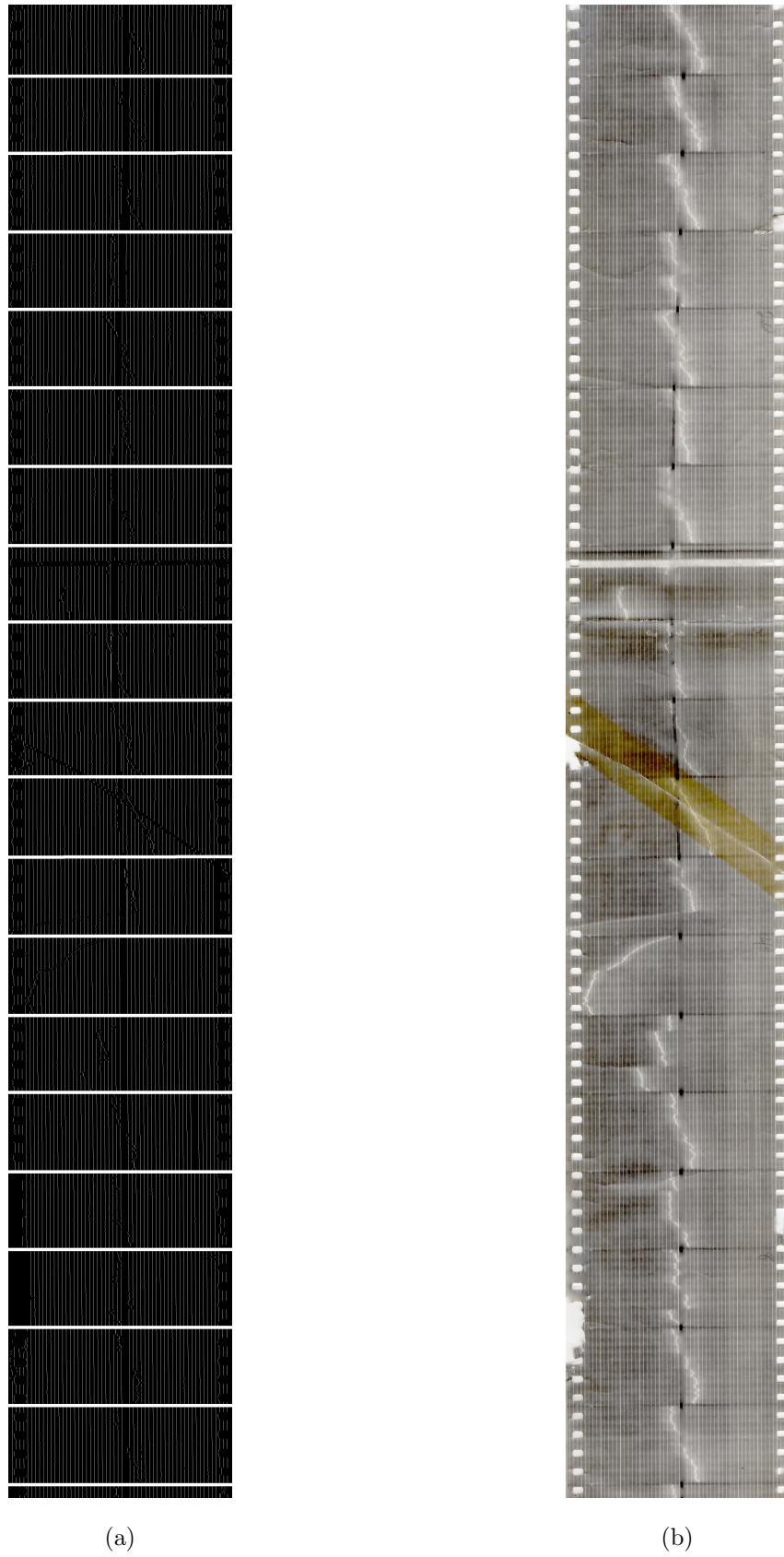


Figure A.10: Christchurch image processing results. (a) Ground truth image, (b) Synthetic ground truth image.

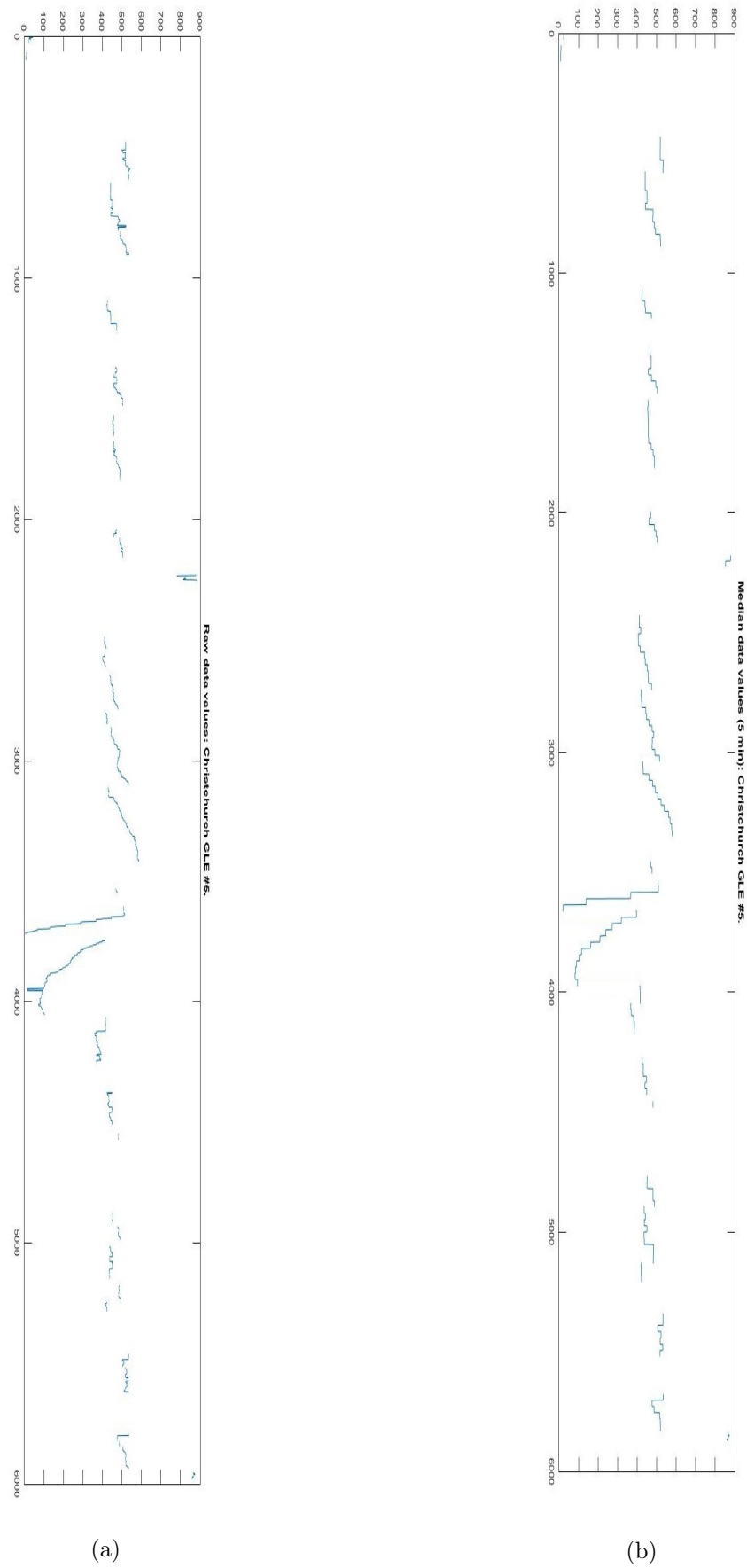


Figure A.11: Christchurch data extraction results. (a) Raw output data line, (b) Data averaged over 5 minute intervals.

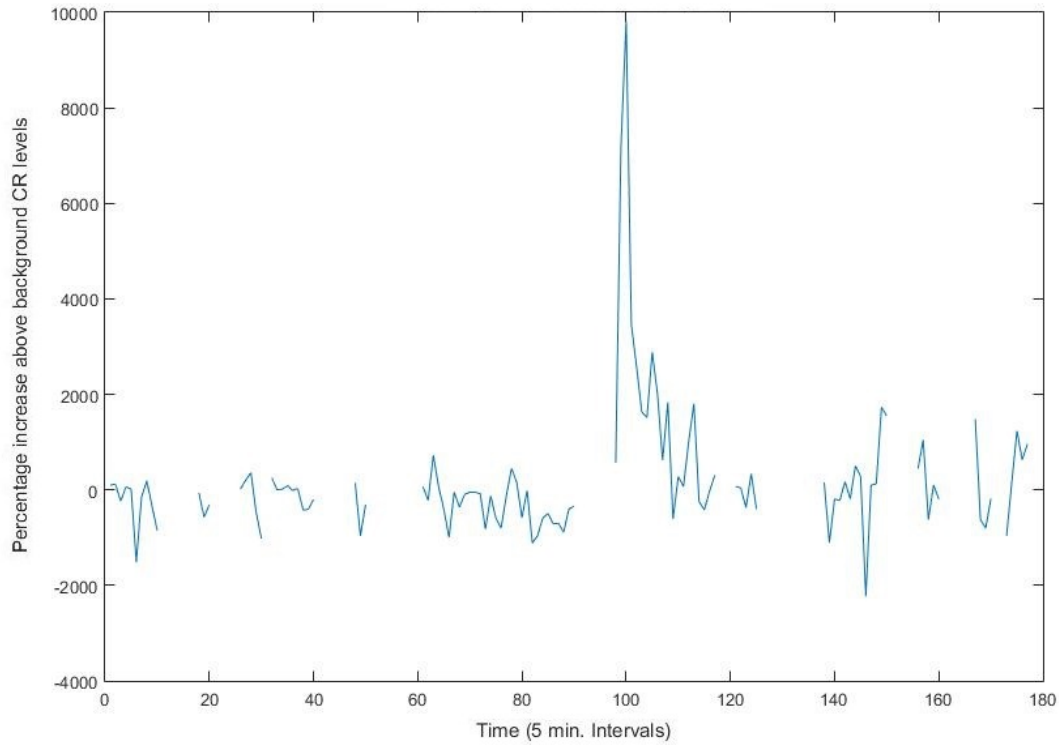


Figure A.12: Percentage increase above background level for Christchurch recording.

Appendix B

Bibliography

Bibliography

- Agrawal, M. and Doermann, D. (2013). Clutter noise removal in binary document images. *International Journal on Document Analysis and Recognition (IJ DAR)*, 16(4), 351–369.
- Anchordoqui, L., Dova, M.T., Mariazzi, A., McCauley, T., Paul, T., Reucroft, S. and Swain, J. (2004). *High energy physics in the atmosphere: Phenomenology of cosmic ray air showers*. vol. 314. Elsevier, pp. 145–207.
- Andersen, M.R., Jensen, T., Lisouski, P., Mortensen, A.K., Hansen, M.K., Gregersen, T. and Ahrendt, P. (2015). *Kinect depth sensor evaluation for computer vision applications*. vol. 1. pp. 1–35.
- Antonacopoulos, A., Gatos, B. and Bridson, D. (2007). Page segmentation competition. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2. IEEE, pp. 1279–1283.
- Ballard, D.H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2), 111–122.
- Belov, A., Eroshenko, E., Mavromichalaki, H., Plainaki, C. and Yanke, V. (2005). A study of the ground level enhancement of 23 February 1956. *Advances in Space Research*, 35(4), 697–701.
- Bermejo, S. and Cabestany, J. (2001). Oriented principal component analysis for large margin classifiers. *Neural Networks*, 14(10), 1447–1461.
- Bertalmio, M., Sapiro, G., Caselles, V. and Ballester, C. (2000). Image inpainting. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., pp. 417–424.
- Beucher, S. and Meyer, F. (1992). *The morphological approach to segmentation: the watershed transformation*. Optical engineering-New York-Marcel Dekker Incorporated, 34, 433–433.

- Bracewell, R.N. (1986). The Fourier transform and its applications, vol. 31999. McGraw-Hill New York.
- Calvo-Zaragoza, J., Micó, L. and Oncina, J. (2016). Music staff removal with supervised pixel classification. *International Journal on Document Analysis and Recognition (IJDAR)*, 1–9.
- Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-8(6), 679–698.
- Carlson, P. (2012). A century of cosmic rays. *Phys. Today*, 65(2), 30–36.
- Chaudhuri, B.B. (2007). Digital document processing: major directions and recent advances. Springer Science & Business Media.
- Chiang, Y.Y. and Knoblock, C.A. (2013). A general approach for extracting road vector data from raster maps. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(1), 55–81.
- Clawson, R., Bauer, K., Chidester, G., Pohontsch, M., Kennard, D., Ryu, J. and Barrett, W. (2013). Automated recognition and extraction of tabular fields for the indexing of census records. *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, p. 86580J86580J.
- Cliver, E. (2006). The 1859 space weather event: Then and now. *Advances in Space Research*, 38(2), 119–129.
- Compton, A., Wollan, E. and Bennett, R. (1934). A Precision Recording Cosmic-Ray Meter. *The Review of Scientific Instruments*, 5, 415–422.
- Dee, H. and French, A. (2015). From image processing to computer vision: Plant imaging grows up. *Functional Plant Biology*, 42(5), iii–v.
- Deng, F., Wu, Z., Lu, Z. and Brown, M.S. (2010). BinarizationShop: a user-assisted software suite for converting old documents to black-and-white. *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, pp. 255–258.
- D’Errico, J. (2012). “inpaint nans” MATLAB Central File Exchange. <https://www.mathworks.com/matlabcentral/fileexchange/4551-inpaint-nans/>. Accessed: 2016-10-06.

- Digabel, H. and Lantuéjoul, C. (1978). Iterative algorithms. Proc. 2nd European Symp. Quantitative Analysis of Microstructures in Material Science, Biology and Medicine, vol. 19. Stuttgart, West Germany: Riederer Verlag, pp. 85–99.
- Dodig-Crnkovic, G. (2002). Scientific methods in computer science. Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia. pp. 126–130.
- Drevin, G. (2008). Adaptive frequency domain filtering of legacy cosmic ray recordings. Proceedings of the 11th Joint Conference on Information Sciences. pp. 15–20.
- Drevin, G., Moraal, H. and McCracken, K. (2007). Determining the skew and scale in images of Compton-Wollan-Bennett ionization chamber recordings. Information Sciences 2007. World Scientific, pp. 888–894.
- Du Plessis, T. (2010). Die versyfering van historiese kosmiese straal-data - The digitization of historic cosmic-ray data.
- Du Plessis, T. and Drevin, G.R. (2009). Evaluating methods to binarize historic cosmic-ray recordings. IEEE EUROCON 2009.
- Duldig, M. (1998). An extended analysis of the September 1989 cosmic ray ground level enhancement. Journal of Geophysical Research, 103(A10), 23–733.
- Duldig, M. (2001). Australian cosmic ray modulation research. Publications of the Astronomical Society of Australia, 18(01), 12–40.
- Fernández-Mota, D., Lladós, J. and Fornés, A. (2014). A graph-based approach for segmenting touching lines in historical handwritten documents. International Journal on Document Analysis and Recognition (IJ DAR), 17(3), 293–312.
- Forbush, S.E. (1938). On world-wide changes in cosmic-ray intensity. Physical Review, 54(12), 975.
- Forbush, S.E. (1957). Solar influences on cosmic rays. Proceedings of the National Academy of Sciences, 43(1), 28–41.
- Forbush, S.E. (1958). Cosmic-ray intensity variations during two solar cycles. Journal of Geophysical Research, 63(4), 651–669.

- Gatos, B., Ntirogiannis, K. and Pratikakis, I. (2009). ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). ICDAR, vol. 9. pp. 1375–1382.
- Gatos, B., Pratikakis, I. and Perantonis, S. (2006). Adaptive Degraded Document Image Binarization. *Pattern Recognition*, 39, 317–327.
- George, S. and Hardeberg, J.Y. (2015). Ink classification and visualisation of historical manuscripts: Application of hyperspectral imaging. *Document Analysis and Recognition (ICDAR)*, 2015 13th International Conference on. IEEE, pp. 1131–1135.
- Gonzalez, R.C. and Woods, R.E. (2002). *Digital image processing*. Pearson Education.
- Gurevich, A., Zybin, K. and Roussel-Dupre, R. (1999). Lightning initiation by simultaneous effect of runaway breakdown and cosmic ray showers. *Physics Letters A*, 254(1), 79–87.
- Hoang, T.V., Smith, E.H.B. and Tabbone, S. (2014). Sparsity-based edge noise removal from bilevel graphical document images. *International Journal on Document Analysis and Recognition (IJДАР)*, 17(2), 161–179.
- Huang, W. (2008). *Scientific chart image recognition and interpretation*. Ph.D. thesis, National University Of Singapore.
- Jain, A.K. and Li, S.Z. (2011). *Handbook of face recognition*. Springer.
- Kasturi, R., Ogorman, L. and Govindaraju, V. (2002). Document image analysis: A primer. *Sadhana*, 27(1), 3–22.
- Krüger, H. (2006). *A calibration neutron monitor for long-term cosmic ray modulation studies*/H. Krüger. Ph.D. thesis, North-West University.
- Lamiroy, B. and Sun, T. (2013). Computing precision and recall with missing or uncertain ground truth. *Graphics Recognition. New Trends and Challenges*. Springer, pp. 149–162.
- Lavrenko, V., Rath, T.M. and Manmatha, R. (2004). Holistic word recognition for handwritten historical documents. *Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on*. IEEE, pp. 278–287.
- Lazzara, G. and Géraud, T. (2014). Efficient multiscale Sauvola's binarization. *International Journal on Document Analysis and Recognition (IJДАР)*, 17(2), 105–123.

- Li, C., Huang, R., Ding, Z., Gatenby, C., Metaxas, D. and Gore, J. (2008). A variational level set approach to segmentation and bias correction of images with intensity inhomogeneity. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 1083–1091.
- Lingenfelter, R.E. (1963). Production of carbon 14 by cosmic-ray neutrons. Reviews of Geophysics, 1(1), 35–55.
- MacAulay, G.D., Senin, N., Giusca, C.L., Leach, R.K. and Ivanov, A. (2015). Review of feature boundary identification techniques for the characterization of tessellated surfaces. Surface Topography: Metrology and Properties, 3(1), 013 002.
- Mattana, V.W. and Drevin, G.R. (2017). An algorithmic retrieval of historic cosmic ray records. Poster, ICRC 2017.
- Mattana, V.W., Drevin, G.R. and Roux, P. (2014). The creation of synthetic digital ground-truth images of historic cosmic ray data recordings. Honours Dissertation, North-West University Potchefstroom.
- Mattana, V.W., Drevin, G.R. and Roux, P. (2017). The creation of synthetic digital ground-truth images of historic cosmic ray data recordings. Lamiroy B., Dueire Lins R. (eds) Graphic Recognition. Current Trends and Challenges., vol. 9657. GREC, Springer, Cham.
- McCracken, K. (2016). Private communications.
- Nave, C.R. (2006). Hyperphysics, hosted by the Department of Physics and Astronomy. Georgia State University, URL=<http://hyperphysics.phy-astr.gsu.edu/hbase/hph.html>.
- Niczyporuk, M.A. and Miller, D.C. (1991). Automatic tracking and digitization of multiple radiopaque myocardial markers. Computers and biomedical research, 24(2), 129–142.
- Ntirogiannis, K., Gatos, B. and Pratikakis, I. (2008*a*). An objective evaluation methodology for document image binarization techniques. Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on. IEEE, pp. 217–224.
- Ntirogiannis, K., Gatos, B. and Pratikakis, I. (2008*b*). An Objective Evaluation Methodology for Handwritten Image Document Binarization Techniques.

- O’Gorman, L. and Kasturi, R. (1997). Document Image Analysis. New York: IEEE.
- Otsu, N. (1975). A Thresholding Selection Method from Gray-Level Histograms. *Automatica*, 11(285-296), 23–27.
- Papandreou, A., Gatos, B., Perantonis, S.J. and Gerardis, I. (2014). Efficient skew detection of printed document images based on novel combination of enhanced profiles. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(4), 433–454.
- Pedersoli, F. and Tzanetakis, G. (2016). Document segmentation and classification into musical scores and text. *International Journal on Document Analysis and Recognition (IJDAR)*, 1–16.
- Potgieter, M.S. (2008). Solar cycle variations and cosmic rays. *Journal of Atmospheric and Solar-Terrestrial Physics*, 70(2), 207–218.
- Powers, D.M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 1(2), 37–63.
- Pratikakis, I., Gatos, B. and Ntirogiannis, K. (2013). ICDAR 2013 document image binarization contest (DIBCO 2013). 2013 12th International Conference on Document Analysis and Recognition. IEEE, pp. 1471–1476.
- Rabeux, V., Journet, N., Vialard, A. and Domenger, J.P. (2014). Quality evaluation of degraded document images for binarization result prediction. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(2), 125–137.
- Reames, D.V. (1999). Particle acceleration at the Sun and in the heliosphere. *Space Science Reviews*, 90(3-4), 413–491.
- Rigaud, C., Guérin, C., Karatzas, D., Burie, J.C. and Ogier, J.M. (2015). Knowledge-driven understanding of images in comic books. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(3), 199–221.
- Rijsbergen, C.v. (1979). Information retrieval. The Information Retrieval Group, <http://dsc.glasgow.ac.uk/preface.html>.
- Rucklidge, W.J., Huttenlocher, D.P. and Felzenswalb, P. (2001). Method and apparatus for image processing employing image segmentation using tokenization. US Patent 6,295,371.

- Sauvola, J. and Pietikäinen, M. (2000). Adaptive Document Image Binarization. *Pattern Recognition*, 33, 225–236.
- Shafii, M. and Sid-Ahmed, M. (2015). Skew detection and correction based on an axes-parallel bounding box. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(1), 59–71.
- Simpson, J.A. (2000). The cosmic ray nucleonic component: The invention and scientific uses of the neutron monitor. *Cosmic Rays and Earth*. Springer, pp. 11–32.
- Singh, B.M., Sharma, R., Ghosh, D. and Mittal, A. (2014). Adaptive binarization of severely degraded and non-uniformly illuminated documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(4), 393–412.
- Smith, M. (2017). 35mm_sprockets.png. https://en.wikipedia.org/wiki/File:35mm_sprockets.png. Accessed: 2017-03-02.
- Sonka, M., Hlavac, V. and Boyle, R. (2014). *Image processing, analysis, and machine vision*. Cengage Learning.
- Su, B., Lu, S. and Tan, C.L. (2013). Robust document image binarization technique for degraded document images. *IEEE transactions on image processing*, 22(4), 1408–1417.
- Svensmark, H. (2007). Cosmoclimatology: a new theory emerges. *Astronomy & Geophysics*, 48(1), 1–18.
- Takano, C. and Ohta, Y. (2007). Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8), 853–857.
- Usoskin, I. (2016). GLE Database GLE #5. <https://gle oulu.fi>. Accessed: 2016-09-25.
- Valizadeh, M. and Kabir, E. (2013). An adaptive water flow model for binarization of degraded document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(2), 165–176.
- Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE transactions on pattern analysis and machine intelligence*, 13(6), 583–598.

- Wolf, C. and Doermann, D. (2002). Binarization of low quality text using a markov random field model. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3. IEEE, pp. 160–163.
- Wu, Y., Peng, X., Ruan, K. and Hu, Z. (2015). Improved image segmentation method based on morphological reconstruction. *Multimedia Tools and Applications*, 1–13.