Exploration und Vorverarbeitung von MAGIC-Daten zur Gamma-Hadron-Separation

Diplomarbeit

von

Tobias Voigt

angefertigt unter Anleitung von Prof. Dr. Roland Fried,
vorgelegt der Fakultät Statistik
der Technischen Universität Dortmund
28. Juni 2010

Inhaltsverzeichnis

1 Einleitung

Die Erforschung und Beschreibung supermassiver schwarzer Löcher und anderer Quellen kosmischer Strahlung ist, was sich die Astroteilchenphysik zur Aufgabe gemacht hat. Dabei sind es von diesen ausgesendete Elementarteilchen -Protonen, Neutrinos und Photonen-, die als Botenteilchen dienen und die nötigen Informationen zur Beschreibung ihrer Quelle mit sich führen (vgl. z.B. Backes, 2008). Hier sind es vor allem hochenergetische Photonen, die wir im Folgenden als Gammas bezeichnen, die Information über ihren Herkunftsort beinhalten und daher in dieser Arbeit eingehender betrachtet werden. Die interessierende Information steckt dabei vor Allem in der Energie der Gammas. Die Messung dieser Gammas und die Bestimmung ihrer Energieverteilung ist daher von großem Interesse.
Cherenkov-Teleskope wie das MAGIC-Teleskop auf La Palma sind in der Lage, Gamma-Strahlung, die auf die Erdatmosphäre trifft, indirekt über Interaktion der Gammas mit der Erdatmosphäre zu messen. Das hauptsächliche Problem bei der Aufnahme solcher Gamma-Daten ist jedoch die Hintergrundstrahlung. Neben den eigentlich interessierenden Gammas werden nämlich noch Unmengen anderer Teilchen, zusammengefasst unter dem Begriff Hadronen, registriert, die mit der Quelle nichts zu tun haben. Bevor die eigentliche Auswertung der Gamma-Strahlung erfolgen kann, müssen also die interessierenden Gammas von den uninteressanten Hadronen bereinigt werden.
Diese Separation erweist sich jedoch als schwierig. Tatsächlich sehen nämlich Messdaten, die von Hadronen aufgenommen werden, ganz ähnlich aus wie die der Gammas. Ziel dieser Arbeit ist, durch Vorverarbeitung die Daten so aufzubereiten, dass die Klassifikation in Gammas und Hadronen einfacher erfolgen kann. Als Daten werden dabei aus den Rohdaten berechnete Kennzahlen, die sogenannten Hillas-Parameter, verwendet.
Im Folgenden wird zunächst in Kapitel 2 die Problemstellung dieser Arbeit formuliert und die vorliegenden Daten und deren Herkunft beschrieben. Dann werden in Kapitel 3 die verwendeten Methoden beschrieben. In der darauf folgenden Analyse werden die Daten zunächst in Kapitel 4 explorativ erforscht und dann in Kapitel 5 für die Separation von Gammas und Hadronen vorverarbeitet. Letztlich werden noch allgemeine Probleme bei der Klassifizierung in Kapitel 6 besprochen und in Kapitel 7 diskutiert.

2 Problemstellung

In diesem Kapitel formulieren wir die Problemstellungen dieser Arbeit. Dazu beschreiben wir zunächst das MAGIC-Experiment, aus dem die vorliegenden Daten stammen, da sich die Problemstellung erst bei Kenntnis des Experiments und der Analysekette formulieren lässt.

2.1 Luftschauer und das MAGIC-Teleskop

In dieser Arbeit unterscheiden wir im Wesentlichen zwischen zwei verschiedenen Formen kosmischer Strahlung: Hochenergetische Gamma-Strahlung, bestehend aus Photonen, nachfolgend Gammas genannt, und der Rest, genannt Hadronen, der im Wesentlichen aus Protonen besteht, aber auch beispielsweise Myonen beinhaltet. Dabei sind es die Gammas, für die man sich eigentlich interessiert. Dass Hadronen mit aufgezeichnet werden, ist ein unerwünschter Nebeneffekt der Art und Weise, wie Gammas detektiert werden. Zur Detektierung der Gammas nutzt man nämlich aus, dass sie beim Eintritt in die Atmosphäre mit dieser wechselwirken. Bei dieser Wechselwirkung wird eine Kettenreaktion ausgelöst, die einen Schauer aus Sekundärteilchen verursacht, im Folgenden Luftschauer oder einfach Schauer genannt (Grieder, 2010). Simulierte Luftschauer sind in Abbildung 1 zu sehen.

Die eigentliche Messung der Luftschauer erfolgt durch Cherenkov-Licht (vgl. Mazin, 2007). Dieses wird von hochenergetisch geladenen Teilchen im Luftschauer ausgesendet und kann von sogenannten Cherenkov-Teleskopen aufgezeichnet und auf einer Kamera abgebildet werden. Die Abbildung geschieht dabei so, dass die Form des Schauers erhalten bleibt.
Problematisch ist, dass nicht nur Gammas beim Eintritt in die Erdatmosphäre solche Luftschauer auslösen können. Auch Hadronen lösen Schauer aus, die wiederum Cherenkov-Licht erzeugen, das von Cherenkov-Teleskopen aufgezeichnet wird. Im Moment des Aufzeichnens des Cherenkov-Lichts sind hadronische Schauer nicht von Gamma-Schauern zu unterscheiden, weshalb die aufgenommenen Daten Gammas und Hadronen enthalten, die getrennt werden müssen.
Das zur Aufnahme der für diese Arbeit genutzten Daten genutzte Teleskop ist das MAGIC-Teleskop, das mit einem Durchmesser von 17 m das größte Cherenkov-

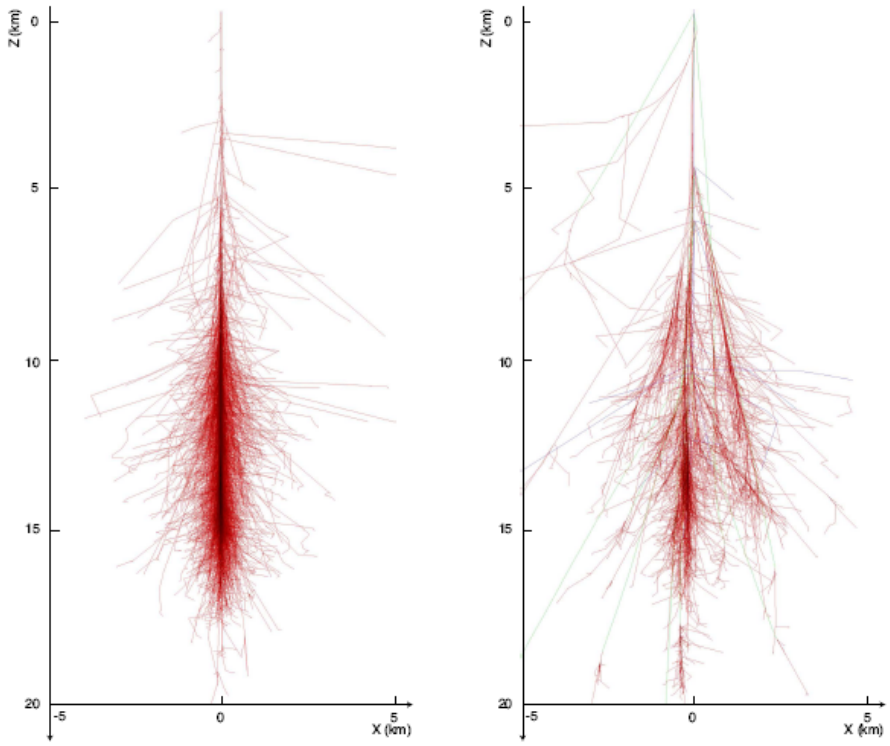

Abbildung 1: Simulierte Luftschauer eines Gammas (links) und eines Protons (rechts) (Sidro Martin, 2008)

Teleskop der Welt ist. Es steht auf dem Roque de los Muchachos auf der Kanareninsel La Palma. Trotz seiner Größe ist das Teleskop sehr beweglich und kann sich innerhalb von 25 Sekunden an jeden möglichen Punkt am Himmel ausrichten (Ferenc, 2005). Das hat es nicht zuletzt einer extremen Leichtbauweise zu verdanken: Das Gerüst des Teleskops besteht komplett aus Kohlefaserstäben.
Abbildung 2 zeigt das Teleskop und das im April 2009 in Betrieb genommene MAGIC II-Teleskop, dessen Daten hier noch nicht zur Verfügung standen, bei dem aber Verbesserungen an der Kamera (vgl. Hsu et al., 2007) und an den Spiegeln (vgl. Backes et al., 2007) vorgenommen wurden.

Abbildung 2: Die beiden MAGIC-Teleskope. Das ältere MAGIC-Teleskop (links), dessen Daten hier verwendet wurden und das neue MAGIC II-Teleskop (rechts) (MAGIC-Homepage, 2010).

Das Teleskop besteht im Wesentlichen aus parabolisch angeordneten Spiegeln, die das ankommende Licht fokussieren und einer Kamera, die das Licht in Photoelektronen umwandelt.
Die parabolisch geformten Spiegel des MAGIC-Teleskops sind zusätzlich isochron, das heißt, dass die Zeitstruktur der Luftschauer im Kamerabild erhalten bleibt. Dadurch können Zeitinformationen in die Analyse der Daten mit einbezogen werden

(Bastieri, 2005).
Die Kamera besteht aus 578 hexagonal angeordneten Photomultipliern (Photo Multiplier Tube, PMT, vgl. z.B. Errando, 2006). Diese wandeln Licht in Photoelektronen um. Dabei wird die Anzahl der Photoelektronen von der Intensität des Lichts bestimmt. Je heller das Licht ist, desto mehr Photoelektronen werden ausgelöst. Die Kamera besteht aus einem inneren und einem äußeren Bereich. Der innere besteht aus 396 PMTs mit einem Durchmesser von jeweils einem Zoll. Der äußere Bereich beinhaltet die restlichen 180 PMTs, die einen Durchmesser von 1,5 Zoll besitzen. Um die Zwischenräume zwischen den runden PMTs zu minimieren, sind ihnen hexagonale Lichttrichter vorgelagert, sogenannte *Winston Cones*. Die Kombination aus PMT und Winston Cone nennt man Pixel. Die schematische Darstellung der Kamerafläche ist in Abbildung 3 zu sehen.
Der äußere Bereich der Kamera dient zur Ergänzung des inneren. Schauer, die ausschließlich im äußeren Bereich der Kamera registriert werden, werden nicht aufgezeichnet. Die Pixel im äußeren Bereich werden nur gebraucht, wenn gleichzeitig auch Pixel im inneren Bereich einen Schauer registrieren. Dies hat mit den sogenannten Triggerzonen zu tun. Der innere Bereich der Kamera ist in Unterbereiche, die Triggerzonen, aufgeteilt. Ein Schauer wird nur dann aufgezeichnet, wenn mindestens vier benachbarte Pixel innerhalb einer Triggerzone die Schranke von sechs ausgelösten Photoelektronen überschreiten (Rissi, 2009). Die Triggerzonen im inneren Kamerabereich sind in Abbildung 4 dargestellt.

Der gesamte Prozess der Datenaufnahme ist in Abbildung 5 noch einmal veranschaulicht. Das eintreffende Teilchen löst einen Luftschauer aus. Dieser erzeugt Cherenkov-Licht, das vom MAGIC-Teleskop gebündelt wird, damit der Schauer auf einer Kamera abgebildet werden kann. Abgespeichert wird die Anzahl der ausgelösten Photoelektronen in jedem Pixel. Dies sind dann die Rohdaten.

2.2 Das bestehende Analyseverfahren

Die Analyse der MAGIC-Daten von der Datenaufnahme bis zur Gamma-Hadron-Separation gliedert sich in drei Schritte, die nach dem jeweils verwendeten Programm benannt werden. Zunächst erfolgt die Datenkalibration im Callisto-Schritt. Hiernach befinden sich die Daten immernoch in einem „Rohzustand“. Im Star-Schritt werden dann Ellipsen an die Schauer angepasst und Hillas-Parameter (s.u.) berechnet. Dann erfolgt schließlich im Ganymed-Schritt nach einigen so-

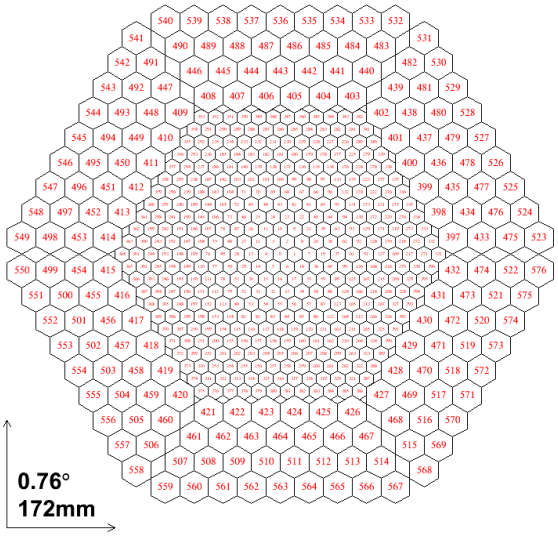

Abbildung 3: Die MAGIC-Kamera mit ihren durchnummerierten Pixeln (MARS-Software, 2010).

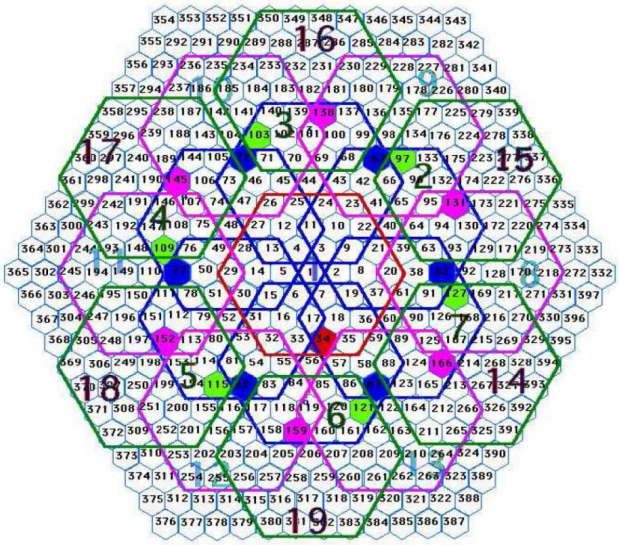

Abbildung 4: Die Triggerzonen im inneren Bereich der MAGIC-Kamera (Rissi, 2009).

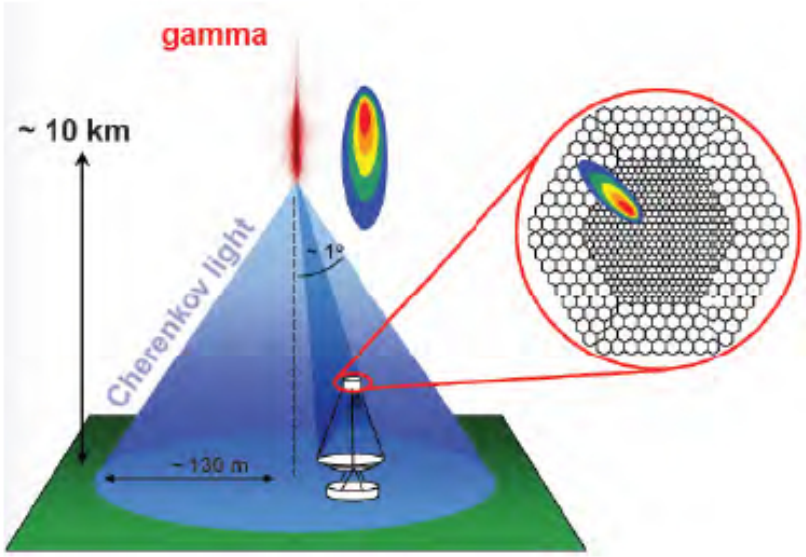

Abbildung 5: Der Prozess der Datenaufnahme mit Luftschauer, Cherenkov-Licht, dem Teleskop und der Kameraebene (Rügamer, 2006).

genannten Quality-Cuts die Separation von Gammas und Hadronen. Der nach Ganymed folgende Sponde-Schritt, in dem das Spektrum der Gammas bestimmt wird, wird hier nicht erläutert. Eine Erklärung dieses Schrittes findet sich zum Beispiel bei Thom (2009).
Im Folgenden wird erklärt, was in den einzelnen Schritten Callisto, Star und Ganymed passiert.

2.2.1 Callisto

Bei Callisto handelt es sich um den ersten Analyseschritt. Die Abkürzung steht für **Cal**ibrate **li**ght **s**ignals and **t**ime **o**ffsets. Wie der Name sagt, ist das Ziel dieses Analyseschritts die Kalibrierung der Daten.
Es gibt verschiedene Gründe, warum die Daten kalibriert werden müssen. Beispielsweise können manche Kamerapixel empfindlicher sein als andere. Außerdem kann es zu Zeitverzögerungen in manchen Pixeln kommen, zum Beispiel aufgrund unterschiedlich langer Kabelwege.
Zur Kalibrierung werden zunächst Kalibrationsdaten aufgenommen, indem einzelne Lichtpulse mit bekannter, für jeden Pixel gleicher, Intensität auf jeden der Pixel gesendet werden. Mit Hilfe dieser Daten kann das Umrechnungsverhältnis von

Licht in Photoelektronen und der zeitliche Versatz jedes Photomultipliers bestimmt und zur Kalibration genutzt werden.
In einem zweiten Kalibrierungsschritt werden Pedestal-Daten aufgenommen. Dabei handelt es sich um Daten des Nachthimmel-Hintergrunds, die im selben Himmelsausschnitt, aber an einer quellfreien Position aufgenommen wurden.
Mithilfe der Kalibrations- und Pedestaldaten werden die vorliegenden, interessierenden Daten um die aufgezeichneten Effekte bereinigt (Backes, 2005).

2.2.2 Star

Star steht für **St**andard **A**nalysis and **R**econstruction. In diesem Schritt werden aus den Originaldaten, also Pixeldaten, Hillas-Parameter (Hillas, 1985) gewonnen. Zunächst erfolgt aber vorher noch die Bildbereinigung. Hierbei werden solche Pixel entfernt, die offenbar nicht zum eigentlichen Schauerbild gehören.
Die Bildbereinigung erfolgt in mehreren Schritten (vgl. Thom, 2009):

1. Bestimme die CorePixel: Pixel, die eine Intensität oberhalb einer hohen Schranke aufweisen (mehr als 8,5 Photoelektronen).
2. Bestimme die Used Pixel: Pixel, die eine Intensität oberhalb einer niedrigeren Schranke haben (mehr als 4,5 Photoelektronen) und zu CorePixeln benachbart sind.
3. Entferne alle restlichen Pixel
4. Entferne isolierte CorePixel, also solche Pixel, deren Nachbarn alle entfernt wurden.

Abbildung 6 zeigt ein Beispiel eines Kamerabildes vor und nach der Bildbereinigung.

Es ist möglich, dass nach der Bildbereinigung mehrere, nicht zusammenhängende Flächen von nicht entfernten Pixeln entstehen. Diese werden im Folgenden Inseln genannt, wobei die größte als Hauptinsel (MainIsland) und die anderen als Neben-inseln (SubIslands) bezeichnet werden. In einem weiteren, verfeinerten Verfahren werden noch Zeitinformationen berücksichtigt. Hieraus werden dann zum Beispiel die *Slope*-Variablen gewonnen (s.u.).

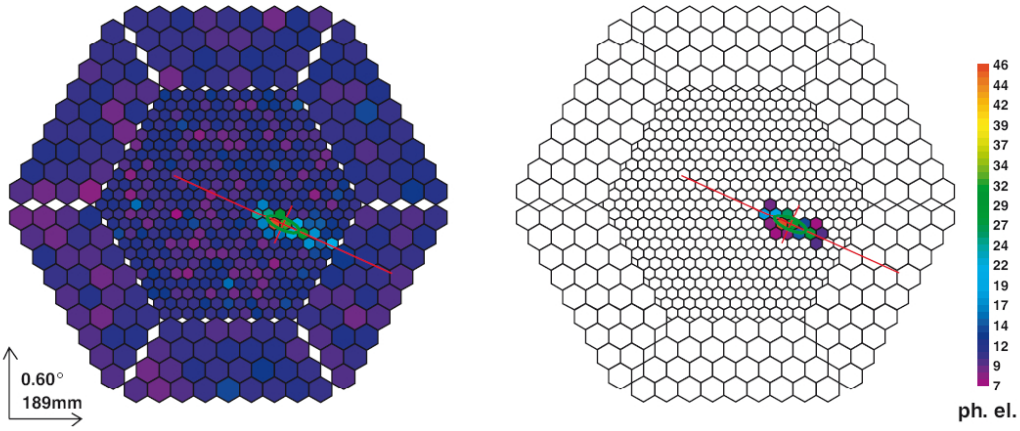

Abbildung 6: Das Kamerabild vor (links) und nach (rechts) der Bildbereinigung (Wagner, 2006).

Auf die Bildbereinigung folgt die Extraktion der Hillas-Parameter (Hillas, 1985). Das heißt, dass Ellipsen an die Schauer angepasst werden und dass die Parameter der Ellipsen (und andere Parameter) anstelle der Pixelinformation als Variablen verwendet werden. So reduziert man die Anzahl der Variablen von über 500 auf 33. Typische Hillas-Parameter sind *Length*, *Width* und *Area*, die die Abmessungen der angepassten Ellipse enthalten. Oder auch *Size*, die die Anzahl der Photoelektronen und damit die Lichtintensität des Schauers angibt und erfahrungsgemäß mit der Energie des eingetroffenen Teilchens zusammenhängt. Ein Überblick über alle Variablen, die in diesem Schritt konstruiert werden, ist in Tabelle 1 zu sehen. Genauere Informationen über die einzelnen Variablen finden sich in Abschnitt 4: Datenexploration.

2.2.3 Ganymed

Ganymed steht für **G**ammas **a**re **n**ow **y**our **m**ost **e**xciting **d**iscovery. Dieser Analyseschritt teilt sich in zwei Unterschritte:

- Quality-Cuts
- Gamma-Hadron-Separation

Variable	Beschreibung
Length	Länge der größeren Halbachse der Ellipse
Width	Länge der kürzeren Halbachse der Ellipse
Area	$\pi \cdot$ *Length* $\cdot$ *Width*
MeanX	x-Position des Schauerschwerpunkts (Center of Gravity, COG)
MeanY	y-Position des Schauerschwerpunkts (Center of Gravity, COG)
Delta	Winkel der längeren Halbachse zur x-Achse der Kamera
CosDelta	$\cos($*Delta*$)$
SinDelta	$\sin($*Delta*$)$
Asym	Differenz des Abstands des COG von der Quellposition und des Abstands des hellsten Pixels von der Quellposition
Size	Summe aller aufgenommener Photoelektronen
SizeSubIslands	Wie *Size*, nur für Nebeninseln
SizeMainIsland	Wie *Size*, nur für die Hauptinsel
NumSinglePixels	Anzahl nicht zu einer Insel gehörender Pixel
SizeSinglePixels	*Size* der nicht zu einer Insel gehörenden Pixel
Leakage1	Verhältnis der Anzahl Photoelektronen im äußersten Kameraring zu *Size*
Leakage2	Verhältnis der Anzahl Photoelektronen in den äußeren zwei Ringen zu *Size*
SlopeTrans	Transversaler Zeitgradient
SlopeLong	Longitudinaler Zeitgradient
NumIslands	Anzahl der Inseln inklusive Hauptinsel
Dist0	Entfernung des COG vom Mittelpunkt der Kamera
Conc	Verhältnis der Photoelektronen in den beiden hellsten Pixeln zu *Size*
Conc1	Verhältnis der Photoelektronen des hellsten Pixels zu *Size*
ConcCOG	Verhältnis der Photoelektronen im COG zu *Size*
ConcCore	Verhältnis der Photoelektronen in den CorePixels zu *Size*
M3Trans	Transversale Schiefe der Verteilung des Schauers
M3Long	Longitudinale Schiefe der Verteilung des Schauers
Borderline	Gesamtlänge des Randes aller Used Pixel
NumUsedPixels	Anzahl aller genutzten Pixel
NumCorePixels	Anzahl der CorePixel
UsedArea	Fläche aller genutzten Pixel
CoreArea	Fläche aller CorePixel
NumSatPixelsHG	Anzahl gesättigter Pixel (High Gain)
NumSatPixelsLG	Anzahl gesättigter Pixel (Low Gain)

Tabelle 1: Alle Variablen, die im Star-Schritt konstruiert werden.

Die Quality-Cuts (vgl. Bretz, 2006) dienen dazu, vor der eigentlichen Klassifikation solche Ereignisse auszusortieren, die entweder offensichtlich keine Gammas sein können, oder die unmöglich zu klassifizieren sind (zum Beispiel weil das Schauerbild zu klein ist).
Es werden nur solche Ereignisse behalten, die die folgenden Voraussetzungen erfüllen:

- Es ist bekannt, dass Gamma-Schauer niemals mehr als zwei Inseln bilden.
  NumIslands < 3
- Das Bild muss nach der Bildbereinigung aus mehr als fünf Pixeln bestehen.
  NumUsedPixels > 5
- Das Bild sollte möglichst komplett zu sehen sein, es sollte also möglichst wenig über den Rand der Kamera hinaus ragen.
  Leakage1 $< 0{,}3$

Außer diesen gibt es noch weitere Cuts, die aber weniger intuitiv und nicht interpretierbar sind. Siehe dazu Bretz (2006).

Für die darauf folgende Gamma-Hadron-Separation werden zur Zeit zwei verschiedene Verfahren verwendet. Eine der beiden Möglichkeiten ist das Anpassen eines Random-Forest zur Klassifizierung (vgl. Albert et al., 2008). Siehe dazu Abschnitt 3.4. Dieses Verfahren wird auch im Folgenden verwendet, um zum Beispiel die Klassifikationsgüte verschiedener Variablenselektionen zu testen. Dabei wird hier die ursprüngliche Implementierung von Breiman (2001) verwendet, die sich aber nur unwesentlich von der von Albert et al. (2008) unterscheidet.
Ein anderer Ansatz ist ein simpler Schnitt in der von den Variablen *Area* und *Size* aufgespannten Fläche. Bei diesem Area-Cut genannten Schnitt werden alle diejenigen Ereignisse als Gamma klassifiziert, die folgende Ungleichung erfüllen:

$$\textit{Area} < c_3 \cdot (1 - c_4 \cdot (\log_{10} \textit{Size}))^2$$

(Riegel, 2005).
Die Parameter c_3 und c_4 werden so bestimmt, dass die Signifikanz nach Li Ma (1983) maximal wird.

Es wird demnach eine Parabel durch die Daten gelegt, wobei alle unter der Parabel liegenden Beobachtungen als Gammas klassifiziert werden. Abbildung 7 veranschaulicht diesen Schnitt.

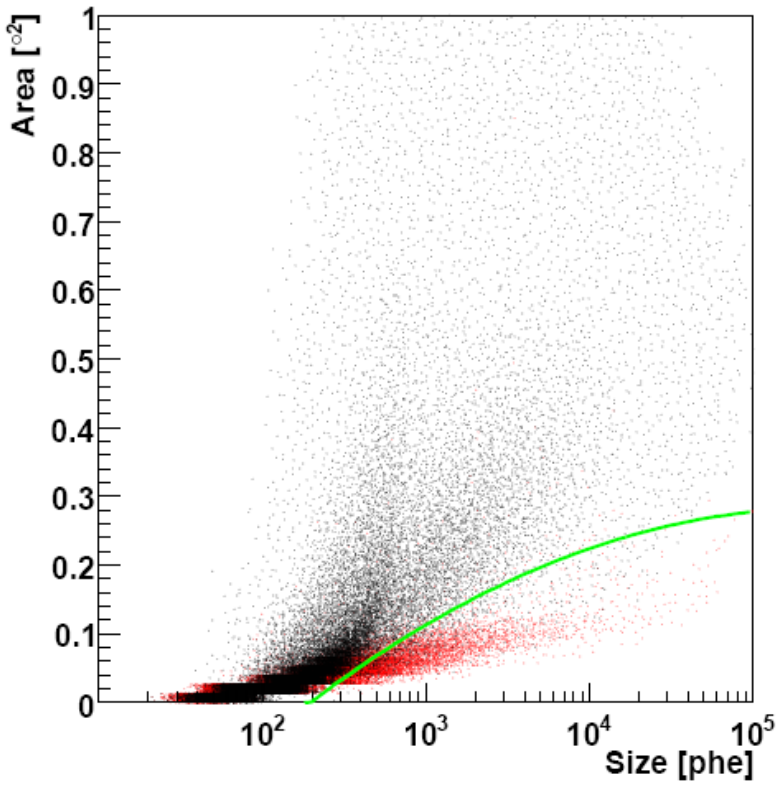

Abbildung 7: Der Area-Cut mit Gammas (rot) und Hadronen (schwarz). Alle Beobachtungen unter der Parabel werden als Gammas klassifiziert. (Riegel, 2006)

2.2.4 Off-Daten und Wobble-Modus

Benutzt man zum Anpassen eines Klassifizierers im Ganymed-Schritt simulierte Gammas, so steht man vor dem Problem, dass die bei der Simulation verwendeten Einstellungen unter Umständen nicht mit den in der echten Quelle vorkommenden übereinstimmen. Tatsächlich kennt man zum Beispiel das Energiespektrum einer Quelle nicht unbedingt, denn dies ist ja gerade, wofür man sich interessiert. Ist dies der Fall, kann es sein, dass ein Klassifizierer, der auf Testdatensätzen noch sehr gute Ergebnisse, also sehr reine Gamma-Stichproben, geliefert hat, auf den echten Daten sehr schlecht abschneidet. Anstatt also auf die Güte des Klassifizierers bei den echten Daten zu vertrauen, ist man daran interessiert, die Reinheit der Gamma-Stichprobe auch dann bestimmen zu können, wenn keine Informationen über die tatsächliche Klasse einer Beobachtung vorliegen (was bei echten Daten der Fall ist).
Um die Reinheit der nach der Klassifizierung entstehenden Gamma-Stichprobe zu bestimmen, kann man sich Off-Daten bedienen. Dabei handelt es sich um einen Datensatz, der komplett aus Hadronen besteht und der im selben Himmelsabschnitt über die gleiche Zeit aufgenommen wurde wie die Quelldaten, aber nachdem die Quelle weitergezogen ist. Klassifiziert man diese Daten auf die selbe Weise wie die Quelldaten, so hat man mit der Anzahl der Hadronen, die fälschlicherweise als Gamma klassifiziert wurden, eine Schätzung für die Anzahl der Hadronen in der Gamma-Stichprobe.
Diese Methode hat den Nachteil, dass nach der eigentlichen Datenaufnahme noch einmal die selbe Zeit aufgewendet werden muss um Off-Daten aufzunehmen. Bei diesem Verfahren wird vom On-Off-Modus des Teleskops gesprochen (Bretz, 2006).
Eine andere Möglichkeit, Off-Daten zu bekommen, ist, das Teleskop im Wobble-Modus zu betreiben (Fomin, 1994). Beim Wobble-Modus wird nicht, wie im On-Off-Modus, die Quelle im Kameramittelpunkt verfolgt. Stattdessen ist die Quellposition in der Kameraebene etwas verschoben. Symmetrisch zur Quellposition, also genau gegenüber des Kameramittelpunkts, wird dann eine sogenannte Off-Position definiert. Dies ist in Abbildung 8 zu sehen.
Diese Definition einer Off-Position erlaubt es, während der eigentlichen Datenaufnahme gleichzeitig auch Off-Daten zur Abschätzung der Reinheit der nach der Klassifikation erhaltenen Gamma-Stichprobe zu nehmen. Diese Abschätzung kann im Wobble-Modus jedoch nicht ohne den im Folgenden Abschnitt erläuterten

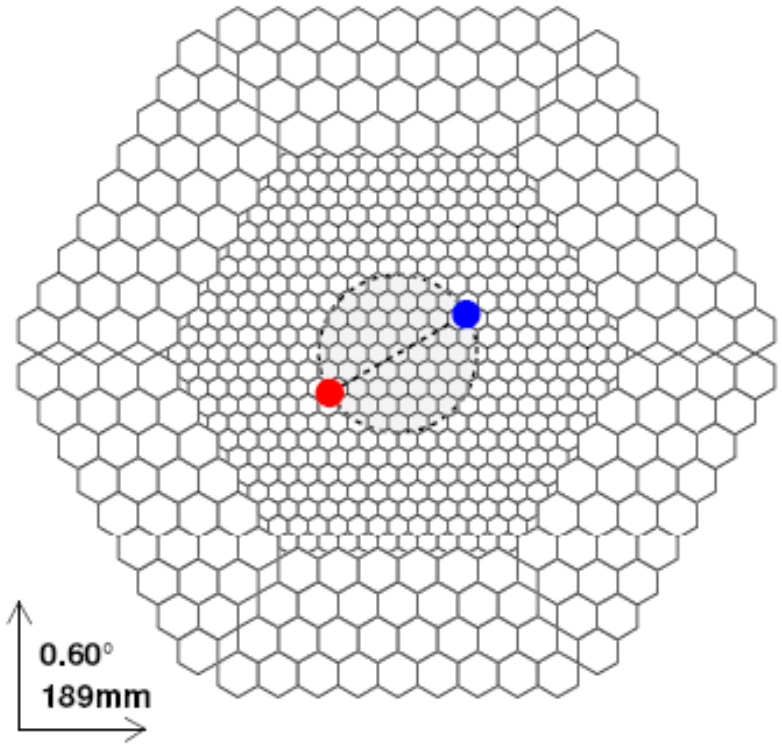

Abbildung 8: Der Wobble-Modus: Die echte Quellposition (blau) und die Off-Position (rot) (Doert, 2009)

θ^2-Cut erfolgen. Dort wird auch erläutert, wie mithilfe des Wobble-Modus und des θ^2-Cuts die Abschätzung der Reinheit funktioniert.

2.2.5 θ^2-Cut

Nach der Klassifikation wird immer noch ein weiterer Schnitt durchgeführt, egal, ob der Area-Cut oder ein Random Forest zur Klassifikation genutzt wurde. Bei dem Schnitt handelt es sich um den θ^2-Cut.
θ ist eine Variable, die sich aus den Hillas-Parametern berechnen lässt. Dazu berechnet man zunächst für jede Beobachtung eine geschätzte Quellposition, also die Stelle auf dem Teleskop, an der nach der Form der angepassten Ellipse zu urteilen die Quellposition liegen müsste, falls es sich bei der Beobachtung um ein Gamma handelt. Die Berechnung der geschätzten Quellposition erfolgt über eine weitere Variable *Disp* (Lessard, 2001). Dabei handelt es sich um eine Schätzung des Abstandes zwischen dem Schauerschwerpunkt (Center of Gravity, COG) einer Beobachtung und der geschätzten Quellposition. Es gibt verschiedene Ansätze zur Bestimmung von *Disp*. Einige werden in Rügamer (2006) ausführlich beschrieben und diskutiert.
Die eigentlich interessierende Variable θ ist dann gegeben durch

$$\theta^2 = \textit{Dist}^2 + \textit{Disp}^2 - 2 \cdot \cos(\textit{Alpha}) \cdot \textit{Dist} \cdot \textit{Disp},$$

wobei *Dist* der Abstand von der echten Quellposition zum COG des Schauers ist und *Alpha* der Winkel zwischen der großen Halbachse der Ellipse und der Strecke zwischen COG und echter Quellposition (Rügamer, 2006).
Bei θ handelt es sich also um den Abstand der geschätzten Quellposition von der echten. Für Gammas, die ja von genau dieser Quelle ausgehen, sollte dieser Abstand sehr häufig nahe bei Null sein, während er für Hadronen annähernd gleichverteilt ist.
Einen Überblick über die verschiedenen neu eingeführten Variablen gibt Abbildung 9.

Es wird dann der sogenannte θ^2-Cut durchgeführt. Dieser ist gegeben durch

$$\theta^2 < c_0,$$

wobei c_0 quellabhängig bestimmt wird und deutlich kleiner sein sollte als der Abstand zwischen Quell- und Off-Position (Bretz, 2006).

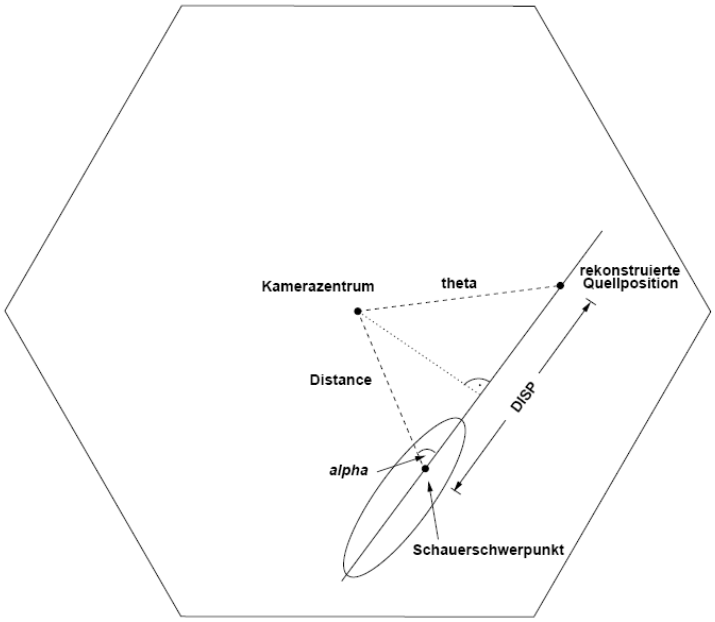

Abbildung 9: Die in diesem Abschnitt neu eingeführten Größen (Riegel, 2006).

Der θ^2-Cut hat den Vorteil, die Daten nach der eigentlichen Klassifikation noch weiter zu separieren, wobei - abhängig von der Wahl von c_0 - nur sehr wenige Gammas, aber viele Hadronen aussortiert werden.
Eine Häufigkeitsverteilung von θ^2 ist in Abbildung 10 dargestellt. Offenbar liegen bei Werten von θ^2 größer als 0,1 Signaldaten und Off-Daten auf einer Höhe, sodass davon ausgegangen werden kann, dass fast alle Beobachtungen dort in der Stichprobe verbliebene, also falsch klassifizierte, Hadronen sind. Es erscheint also sinnvoll, in den Signaldaten alle Werte mit einem großen θ^2 zu entfernen.

Die gute Klassifikationsgüte durch θ ließe sich auch im eigentlichen Klassifikationsverfahren nutzen, indem man es schon vor der Klassifikation berechnen und vom Klassifikationsverfahren nutzen ließe. Dadurch kann aber der θ^2-Cut nach der Klassifikation nicht mehr zur Abschätzung der Reinheit der entstehenden Gamma-Stichprobe verwenden. Hier ist es dem Anwender überlassen, ob eher eine bessere Klassifikation, oder die Fähigkeit der Abschätzung der Reinheit nach der Klassifikation gewünscht ist. In dieser Arbeit wird θ nicht während der eigentlichen Klassifikation genutzt.
Wie im vorhergehenden Abschnitt bereits beschrieben wurde, lässt sich mithilfe des θ^2-Cuts aus im Wobble-Modus aufgenommenen Daten die Reinheit der

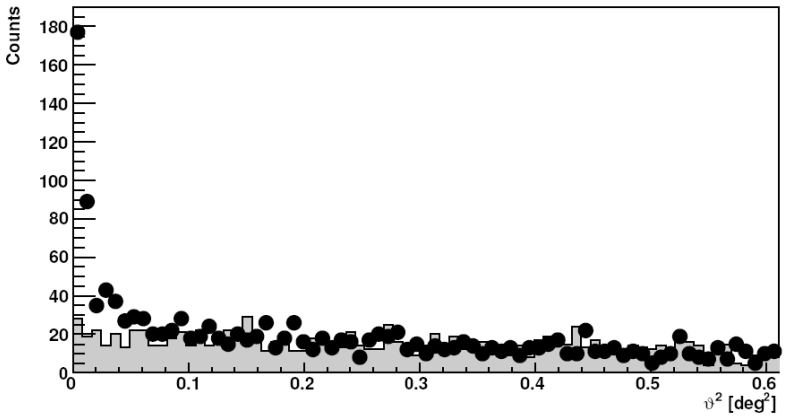

Abbildung 10: Eine Häufigkeitsverteilung von θ^2 für echte Daten nach der Klassifikation (schwarze Punkte) und Off-Daten (graue Fläche) (Bretz, 2006)

nach der Klassifikation erhaltenen Gamma-Stichprobe abschätzen. Das oben beschriebene θ wird dazu bezüglich beider Positionen, sowohl der echten Quellposition, als auch der Off-Position berechnet. Das zur Quellposition berechnete θ bezeichnen wir mit θ_Q, das zur Off-Position berechnete mit θ_{Off}. θ_Q ist also der Abstand der geschätzten Quellposition von der echten. Dieser Wert sollte für Gammas nahe bei Null sein, während er für Hadronen annähernd gleichverteilt ist, da sie nicht aus der Quelle stammen und Übereinstimmungen zwischen beobachteter und rekonstruierter Quelle rein zufälliger Natur sind. θ_{Off} ist der Abstand der geschätzten Quellposition von der Off-Position. Dieser Wert ist für Hadronen wiederum annähernd gleichverteilt. Für Gammas ist er annähernd gleich dem Abstand zwischen Quell- und Off-Position. Wendet man den θ^2-Cut auf Daten an, die Gammas und Hadronen enthalten, so ist das Ergebnis abhängig davon, ob θ_Q oder θ_{Off} in der Gleichung eingesetzt wurde. Verwendet man θ_Q, so werden bei einer geschickten Wahl von c_0 fast alle Gammas diese Gleichung erfüllen, sowie ein Anteil der Hadronen, der zufällig in diesem Bereich liegt. Verwendet man dagegen θ_{Off}, so werden fast keine Gammas die Gleichung erfüllen, aber wiederum ein Anteil der Hadronen, der zufällig in diesem Bereich liegt.
Da θ für Hadronen annähernd gleichverteilt ist, ist aber die Anzahl der Beobachtungen, die $\theta^2_{Off} < c_0$ erfüllen, annähernd gleich der Anzahl der Hadronen, die $\theta^2_Q < c_0$ erfüllen. Damit haben wir also eine Schätzung für die Anzahl der Hadronen, die nach dem θ^2-Cut in der Stichprobe verbleiben und somit für die Reinheit der Gamma-Stichprobe.

2.3 Die vorliegenden Daten

Die hier vorliegenden Daten stammen aus einem Zwischenschritt zwischen den Analyseschritten Star und Ganymed. Es handelt sich dabei um zwei Datensätze. Einer enthält simulierte Gammas, der andere echte, im Wobble-Modus aufgenommene Hadronen. Die Variablen des Datensatzes sind die in Tabelle 1 eingeführten. In beiden Datensätzen sind die in Abschnitt 2.2.3: Ganymed besprochenen Quality-Cuts bereits durchgeführt.
Der Gamma-Datensatz stammt aus einer Simulation (Doert, 2009). Die Einstellungen der Simulation wurden dabei so gewählt, dass die Beobachtungen solchen aus dem Crab-Nebel entsprechen. Für Informationen zum Crab-Nebel siehe zum Beispiel Albert et al. (2008). Der Datensatz besitzt 1.122.061 Beobachtungen in 33 Variablen.
Der Hadronen-Datensatz enthält echte, im Wobble-Modus aufgenommene Beobachtungen. Das bedeutet, dass das Teleskop bei der Aufnahme nicht auf ein quellenfreies Himmelsstück ausgerichtet wurde, sondern auf eine Quelle, in diesem Falle den Crab-Nebel. Da die Daten noch in keiner Weise klassifiziert wurden, ist davon auszugehen, dass dem Datensatz Gammas aus dieser Quelle beigemischt sind. Es ist bekannt, dass das Verhältnis von Gammas zu Hadronen in solchen Datensätzen etwa 1:1000 (Weekes, 2003) beträgt. Der Datensatz besteht aus 1.303.252 Beobachtungen, sodass davon auszugehen ist, dass etwa 13.000 Gammas in ihm enthalten sind. Die Anzahl der Variablen ist wie bei den Gammas 33. Eine Beschreibung der einzelnen Variablen erfolgt in Abschnitt 4: Datenexploration.

2.4 Zielsetzung

Ein wesentlicher Bestandteil der Analysekette der Gamma-Daten ist der Ganymed-Schritt. In diesem Schritt werden Gamma- von Hadronendaten getrennt. Das letztendliche Ziel dieses Schrittes ist es, einen möglichst reinen Gamma-Datensatz zu erzeugen, der im besten Fall überhaupt keine Hadronen mehr enthält. Wie im Folgenden gezeigt wird, ist das Ziel einer sehr reinen Gamma-Stichprobe sehr einfach zu erreichen, wenn man bereit ist, einen sehr großen Anteil der Gammas zu verlieren. Der Preis für eine sehr gute Reinheit ist also der systematische Verlust von Gamma-Ereignissen. Da wir aber so viele Gammas wie möglich für die wei-

teren Analyseschritte behalten wollen, ist ein weiteres Ziel, den Anteil der falsch, also als Hadron, klassifizierten Gammas so gering wie möglich zu halten. Oder andersherum den Anteil der richtig klassifizierten Gammas zu maximieren. Dieser Anteil richtig klassifizierter Gammas wird im Folgenden Recall genannt.
Das letztendliche Ziel ist also die gleichzeitige Maximierung der Reinheit und des Recalls. Es gilt:

$$\text{Reinheit} = \frac{\text{Anzahl richtig klassifizierter Gammas}}{\text{Anzahl aller als Gamma klassifizierter Beobachtungen}}$$

$$\text{Recall} = \frac{\text{Anzahl richtig klassifizierter Gammas}}{\text{Anzahl aller Gammas im Datensatz}}$$

Dieses Ziel entspricht im Wesentlichen dem der Minimierung der Fehlklassifikationsrate, also des Anteils falsch klassifizierter Beobachtungen, mit dem Unterschied, dass hier bei der eigentlichen Klassifizierung der Reinheit ein höheres Gewicht gegeben werden soll als dem Recall.
Außerdem ist die Betrachtung von Reinheit und Recall vergleichbar mit der Signifikanz nach Li und Ma (1983), die in der MAGIC-Analysekette als Gütemaß verwendet wird. Für diese ist jedoch eine Analyse der Daten bis hin zum und einschließlich des θ^2-Cuts notwendig. So weit wird die Analyse in dieser Arbeit aber nicht gehen, sodass wir hier das oben beschriebene andere, aber ähnliche Gütekriterium verwenden.
Um das Ziel der Maximierung von Reinheit und Recall zu erreichen, führen wir in dieser Arbeit einige Vorarbeiten durch. Um einen Einblick in die Struktur der Daten zu erhalten, werden wir zunächst die Variablen einzeln explorativ analysieren. Hierbei gehen wir explizit auf die Unterschiede zwischen Gammas und Hadronen ein, um schon hier Ansätze für eine Klassifikation zu finden.
Im Anschluss daran werden wir einige Vorverarbeitungen durchführen, die helfen werden, in der eigentlichen Klassifikation das oben genannte Ziel zu erreichen.
Eine wesentliche Fragestellung dabei ist, wie das Verhältnis von Gammas zu Hadronen in Trainings- und Testdatensätzen sein sollte. Hier gibt es das Problem, dass in der Realität der Anteil der Gammas an der Gesamtzahl der aufgezeichneten Ereignisse mit etwa 1:1000 sehr gering ist (Weekes, 2003). Die Frage ist, ob man dieses Verhältnis auch bei der Klassifikation in Trainings- und Testdatensätzen einhalten muss.
Weitere Aspekte bei der Vorverarbeitung werden Variablentransformationen und Variablenreduktion sein. Letztere bietet sich deshalb an, weil die Anzahl der aufgenommenen Beobachtungen immens ist. In einer Sekunde können mehrere Hun-

dert Luftschauer aufgezeichnet werden, sodass in einer einzigen Nacht mehrere Terrabyte an Daten anfallen können. Hier gibt es das Problem, dass mit steigender Datensatzgröße auch der benötigte Speicherplatz und vor allem die benötigte Rechenzeit der Auswertungen anwächst. Dem kann man abhelfen, indem man nicht benötigte Variablen aus dem Datensatz entfernt.
Letztlich werden wir noch untersuchen, wo es bei einer Klassifikation Probleme gibt. Es wird sich zeigen, dass Gammas, die schwierig als solche zu erkennen sind, gewisse Eigenschaften und Systematiken aufweisen.

Zusammengefasst lauten die Problemstellungen dieser Arbeit also:

- Explorative Analyse der Daten
- Untersuchung des Gamma-Hadron-Verhältnisses
- Transformationen einzelner Variablen
- Variablenreduktion
- Finden von Eigenschaften schwierig zu klassifizierender Gammas

3 Methoden

In diesem Kapitel beschreiben wir die in dieser Arbeit verwendeten Methoden. Zu beachten ist, dass für alle Berechnungen die Statistik-Programmiersprache R in der Version 2.10.1 durchgeführt wurden (R Development Core Team, 2009). Die Nutzung von zusätzlich zum Basisprogramm benutzter Pakete wird jeweils an entsprechender Stelle angegeben.

3.1 Box-Cox-Transformation

Die Box-Cox-Transformation (Box & Cox, 1964) ist ein Hilfsmittel, um Variablen mit einer schiefen Verteilung zur Symmetrie hin zu verschieben. Für eine Variable Y ist sie durch die Formel

$$Y_{BC} = \begin{cases} \frac{(Y+c)^p - 1}{p}, & p \neq 0 \\ \ln(Y+c), & p = 0 \end{cases}$$

gegeben.

Der Term $\ln(Y+c)$ ist dabei eine stetige Fortsetzung des anderen für $p = 0$. Die Parameter p und c müssen sinnvoll gewählt werden. c ist so zu wählen, dass $Y + c$ nur positive Werte aufweist, da die Transformation für negative Werte von $Y + c$ nicht notwendigerweise definiert ist. p ist der Parameter, der die Schiefe des Ergebnisses beeinflusst. Man kann versuchen, ihn so zu bestimmen, dass die transformierte Stichprobe möglichst symmetrisch wird.

3.2 Kullback-Leibler-Divergenz

Die Kullback-Leibler-Divergenz (Kullback & Leibler, 1951) ist ein Maß für den Abstand zwischen zwei Verteilungen mit Dichten f_1 und f_2 bezüglich des gleichen dominierenden Maßes λ. Sie ist gegeben durch

$$J(f_1, f_2) = \int (f_1(x) - f_2(x)) \log \frac{f_1(x)}{f_2(x)} \, \mathrm{d}\lambda(x).$$

Wir nutzen sie, um diejenigen Variablen zu bestimmen, die sich - ohne den Einfluss anderer Variablen - am besten zur Separation von Gammas und Hadronen eignen, indem wir die Divergenz zwischen der Verteilung von Gammas und Hadronen in jeder Variable berechnen. Eine große Kullback-Leibler-Divergenz bedeutet, dass sich diese Variable gut zur Separation eignet.
Da wir die Dichten f_1 und f_2 nicht kennen, schätzen wir sie mithilfe einer einfachen Kerndichteschätzung. Für eine aus einer Verteilung mit Dichte f stammenden Sichprobe $x_1, ..., x_n$ ist sie allgemein gegeben durch

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{j=1}^{n} k\left(\frac{t - x_j}{h}\right),$$

wobei k der zu wählende Kern und h die Bandbreite der Schätzung ist. Als Kern wählen wir den Gauss-Kern

$$k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right).$$

Die Bandbreite h lassen wir automatisch vom Programmpaket wählen. Dazu wird eine „rule of thumb“ von Silverman (1986) verwendet, die mit der Funktion `bw.rnd0()` im R-Basis-Paket `stats` implementiert ist.

3.3 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (vgl. Fried, 2009 und Hartung, 1999) ist ein Verfahren zur Dimensionsreduktion. Liegen Beobachtungen $\mathbf{X}_1, ..., \mathbf{X}_n$ von d quantitativen Merkmalen vor, so ist die Idee der Hauptkomponentenanalyse, dass aus den d Variablen wiederum d unkorrelierte Linearkombinationen (Hauptkomponenten, HKn) gebildet werden, die sukzessive einen sinkenden Anteil der Datenvariablilität erklären. Das heißt, die erste Hauptkomponente soll die meiste Variabilität erklären, die d-te Hauptkomponente die wenigste. Dabei ist darauf zu achten, dass die Gesamtvariabilität als Summe der einzelnen Varianzen der Variablen der Originaldaten gleich der Gesamtvarianz der Hauptkomponenten ist. Dies ist in der Hauptkomponentenanalyse der Fall, da die Spur einer Matrix invariant unter Basis-Transformationen ist.
Durch Auswahl der ersten $p < d$ Hauptkomponenten können die gesamten Daten in einem niedriger dimensionalen Raum ohne großen Informationsverlust re-

präsentiert werden. Die Idee der Hauptkomponentenanalyse ist auch in Abbildung 11 dargestellt.

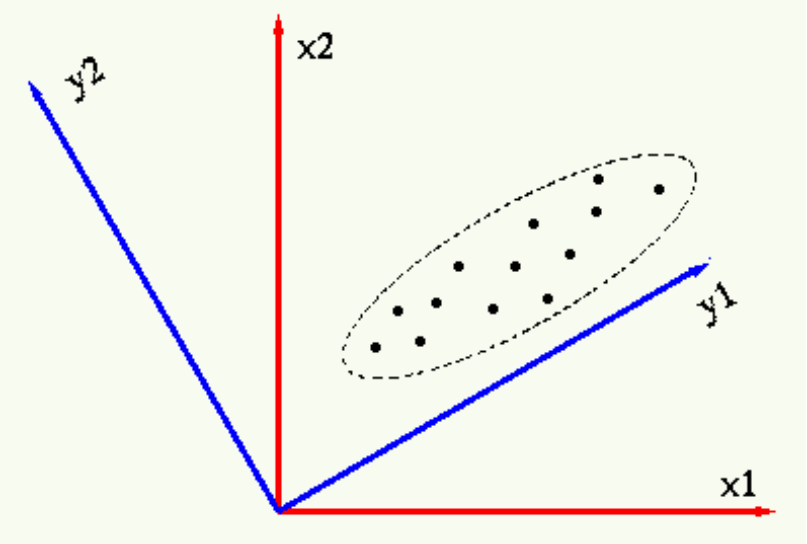

Abbildung 11: Die Hauptkomponentenanalyse: Die Variablen x_1 und x_2 sind stark korreliert. Es werden die Hauptkomponenten y_1 und y_2 gebildet, wobei y_1 maximale Varianz besitzt. y_2 enthält nur wenig Information über die Daten.
(http://fourier.eng.hmc.edu/e101/lectures/rotation.gif)

3.3.1 Konstruktion

Um die Hauptkomponenten zu berechnen, benötigt man zunächst die zentrierten Variablen. Seien die d Variablen $V_1, ..., V_d$ gegeben, sei also die Ausgangssituation wie folgt:

	V_1	V_2	$\cdots$	V_d
X_1	x_{11}	x_{12}	$\cdots$	x_{1d}
X_2	x_{21}	x_{22}	$\cdots$	x_{2d}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_n	x_{n1}	x_{n2}	$\cdots$	x_{nd}

Dann ist die j-te zentrierte Variable gegeben durch $Y_j = V_j - \overline{V}_j$, wobei $\overline{V}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ j = 1, \ldots, d$ und die Matrix der zentrierten Werte ist

$$\mathcal{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1d} \\ y_{21} & y_{22} & \cdots & y_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nd} \end{pmatrix} .$$

Mit dieser Ausgangssituation bestimmen wir nun zunächst die erste Hauptkomponente. Sei $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ ($'$ bezeichne die Transposition) der zentrierte Zufallsvektor mit empirischer Kovarianzmatrix $\mathbf{S}_Y = \frac{1}{n}\mathcal{Y}'\mathcal{Y}$. Dann ist die *1. Hauptkomponente* Z_1 definiert als eine normierte Linearkombination von $Y_1, \ldots, Y_d$ mit maximaler Varianz.
Sei $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1d})'$ der Koeffizientenvektor der 1. HK:

$$Z_1 = \mathbf{a}_1'\mathbf{Y} = a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1d}Y_d$$

mit empirischer Varianz

$$S_{Z_1}^2 = \mathbf{a}_1'\mathbf{S}_Y\mathbf{a}_1 .$$

$\mathbf{a}_1$ erhält man aus dem Maximierungsproblem

$$\text{Max!} \quad S_{Z_1}^2 = \mathbf{a}_1'\mathbf{S}_Y\mathbf{a}_1 \quad \text{unter} \quad \mathbf{a}_1'\mathbf{a}_1 = 1 .$$

Sei λ_1 der Maximalwert von $\mathbf{a}_1'\mathbf{S}_Y\mathbf{a}_1$.

Alle weiteren Hauptkomponenten lassen sich ähnlich bestimmen, jedoch mit der Zusatzannahme der Unkorreliertheit zu allen vorangegangenen Hauptkomponenten. Die *j-te Hauptkomponente* Z_j, $j = 2, \ldots, d$, ist also eine normierte Linearkombination von $Y_1, \ldots, Y_d$, welche die Varianz maximiert unter allen zu Z_1, ..., Z_{j-1} unkorrelierten Linearkombinationen.

Sei $\mathbf{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jd})'$ der Koeffizientenvektor von Z_j,

$$Z_j = \mathbf{a}_j'\mathbf{Y} = a_{j1}Y_1 + a_{j2}Y_2 + \ldots + a_{jd}Y_d$$

mit empirischer Varianz

$$S_{Z_j}^2 = \mathbf{a}_j'\mathbf{S}_Y\mathbf{a}_j .$$

Dann ist die empirische Kovarianz zwischen Z_i und Z_j gleich

$$S_{Z_i,Z_j} = \mathbf{a}_i'\mathbf{S}_Y\mathbf{a}_j .$$

$\mathbf{a}_j$ ist also Lösung des Maximierungsproblems

$$\text{Max}_{\mathbf{a}_j}! \quad S^2_{Z_j} = \mathbf{a}_j' \mathbf{S}_Y \mathbf{a}_j \quad \text{unter} \quad \mathbf{a}_j' \mathbf{a}_j = 1\,, \quad \mathbf{a}_i' \mathbf{S}_Y \mathbf{a}_j = 0,\ i = 1, \ldots, j-1\,.$$

Sei λ_j dieses Maximum; es gilt nach Konstruktion $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{j-1} \geq \lambda_j$.

In der Praxis kann man sich bei der Berechnung der Hauptkomponenten zunutze machen, dass $\lambda_1, \lambda_2, ..., \lambda_d$ gleich den Eigenwerten und $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_d$ gleich den zugehörenden Eigenvektoren der empirischen Kovarianzmatrix $\mathbf{S}_Y$ von $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ sind.
Es gilt außerdem

$$S^2_{Z_i} = \mathbf{a}_i' \mathbf{S}_Y \mathbf{a}_i = \lambda_i,\ i = 1, \ldots, d\,,$$

und

$$S_{Z_i Z_j} = \mathbf{a}_i' \mathbf{S}_Y \mathbf{a}_j = 0,\ \forall i \neq j\,,$$

d.h. die λ_i sind gleich den empirischen Varianzen der Hauptkomponenten und die empirischen Kovarianzen zwischen den Hauptkomponenten sind (wie gefordert) Null.
In der Praxis kann es außerdem nützlich sein, statt den zentrierten Variablen, die standardisierten Variablen

$$Y_i = \frac{V_i - \overline{V}_i}{\sqrt{S_i^2}}, \quad S_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (V_{ij} - \overline{V}_i)^2,\ i = 1, \ldots, d\,.$$

zu benutzen. Dies ist insbesondere dann sinnvoll, wenn die Variablen stark unterschiedliche Varianzen, oder verschiedene Einheiten besitzen, da in solchen Fällen Variablen mit größerer empirischer Varianz einen stärkeren Einfluss auf die Kontruktion der Hauptkomponenten haben als solche mit kleinerer empirischer Varianz. Des Weiteren hängt sonst die Kontruktion der Hauptkomponenten von den Einheiten der Variablen ab.

3.3.2 Wahl der Hauptkomponenten

Bei der Konstruktion der Hauptkomponenten können immer genau so viele Hauptkomponenten berechnet werden, wie Variablen vorliegen. Da das Ziel der Hauptkomponentenanalyse aber eine Dimensionsreduktion sein soll, müssen aus diesen d Hauptkomponenten nach sinnvollen Kriterien die ersten p ausgewählt werden. Drei verschiedene Kriterien sind die Auswahl nach dem erklärten Prozent-

satz der Gesamtvarianz, nach der mittleren Variabilität, oder mithilfe eines Scree-Graphen. Bei der ersten Methode wählt man einen Prozentsatz der Gesamtvarianz, den die gewählten Hauptkomponenten erklären sollen. Wähle zum Beispiel p minimal, um mindestens 75% (90%, 95%) der totalen Variabilität zu erklären.
Der durch die ersten k Hauptkomponenten erklärte Prozentsatz ist gegeben durch:

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{i=1}^{d} \lambda_i}.$$

Um p nach der mittleren Variabilität der Daten zu bestimmen, wähle p so, dass die Varianz der p-ten Hauptkomponente gerade größer ist als die mittlere Varianz der Daten, also

$$p = \max \left\{ k : \lambda_k \geq \overline{\lambda} = \frac{1}{d} \sum_{j=1}^{d} \lambda_j \right\}$$

Die dritte Möglichkeit ist, p anhand eines Scree-Graphen zu bestimmen. Trage dazu die Varianzen gegen die Nummer der Hauptkomponenten auf und suche den „Knick“ im Graphen. Dies ist in Abbildung 12 veranschaulicht.

Wir benutzen in dieser Arbeit immer die erste Methode und bestimmen p anhand des erklärten Anteils an der Gesamtvarianz. Welcher Prozentsatz jeweils erklärt werden soll, ist an entsprechender Stelle angegeben.

3.4 Random Forest

Ein Random Forest (Breiman, 2001) ist ein Klassifikations- und Regressionsverfahren. Wir verwenden ihn hier als Klassifikator. Die Idee des Random Forest ist, eine Beobachtung von mehreren Entscheidungsbäumen klassifizieren zu lassen und sie so zu klassifizieren, wie es die Mehrzahl der Bäume getan hat.
Ein Entscheidungsbaum für sich genommen ist ein Klassifikationsverfahren, das den gesamten Variablenraum in M Regionen einteilt und jeder Region die Entscheidung für eine Klasse zuordnet. Dabei können auch mehrere Regionen zur gleichen Entscheidung führen.
Die Einteilung erfolgt, indem nacheinander Schnitte in einzelnen Variablen durchgeführt werden. Das Vorgehen wird in Abbildung 13 veranschaulicht. Dort ist dargestellt, wie ein einfacher Entscheidungsbaum mit zwei Variablen aussehen kann.

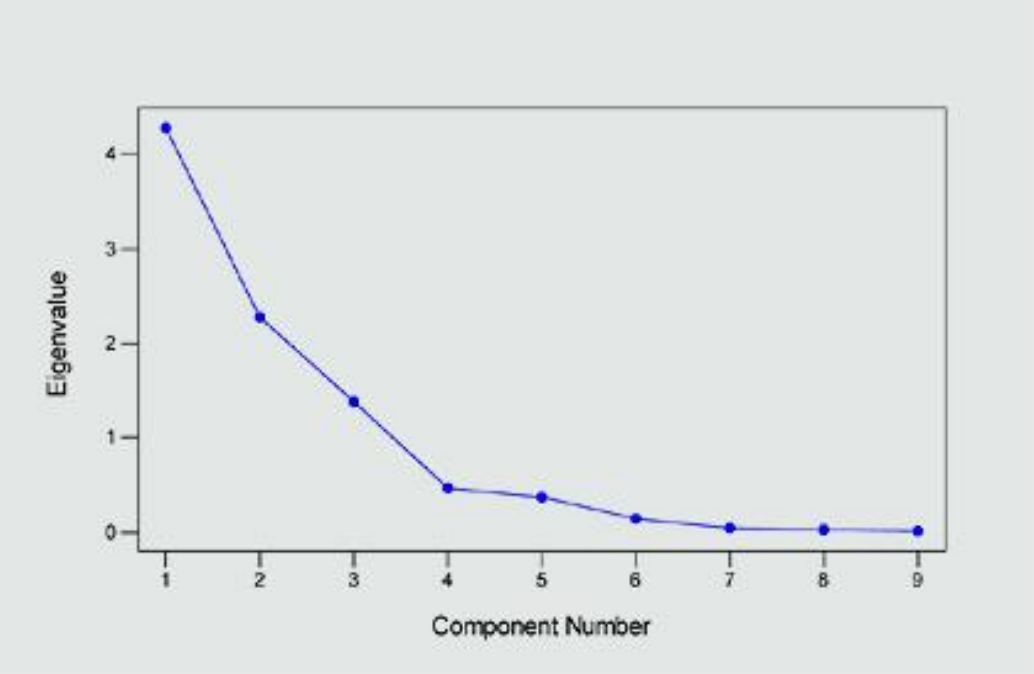

Abbildung 12: Beispiel eines Scree-Graphen. Der Knick ist hier bei der vierten Hauptkomponente zu sehen. Es sollten also die ersten vier HKn gewählt werden.
(http://neon.otago.ac.nz/chemlect/chem306/pca/Theory$_{PCA/images/screeplot.gif}$)

Die Schnitte werden so gewählt, dass die Beobachtungen innerhalb der beiden dadurch entstehenden Teilstichproben möglichst homogen und zwischen den Teilstichproben möglichst inhomogen sind. Als Maß kann dabei der Gini-Index oder die Entropie genutzt werden.

Um aus Entscheidungsbäumen einen Random Forest zu machen, nutze folgenden Algorithmus (Liaw Wiener, 2002):

1. Ziehe n_{tree} Stichproben mit Zurücklegen aus den vorliegenden Beobachtungen. n_{tree} ist dabei die Anzahl der Bäume, die der Random Forest enthalten soll.

2. Passe für jede der Stichproben einen Entscheidungsbaum an. Wende dabei KEIN pruning an und modifiziere den Baum wie folgt: Statt an jeder Verzweigung den besten Schnitt unter allen Variablen zu bestimmen, wähle von den Variablen zufällig m_{try} aus und bestimme den besten Schnitt nur unter diesen.

3. Um neue Daten zu klassifizieren, lasse die neuen Beobachtungen von jedem

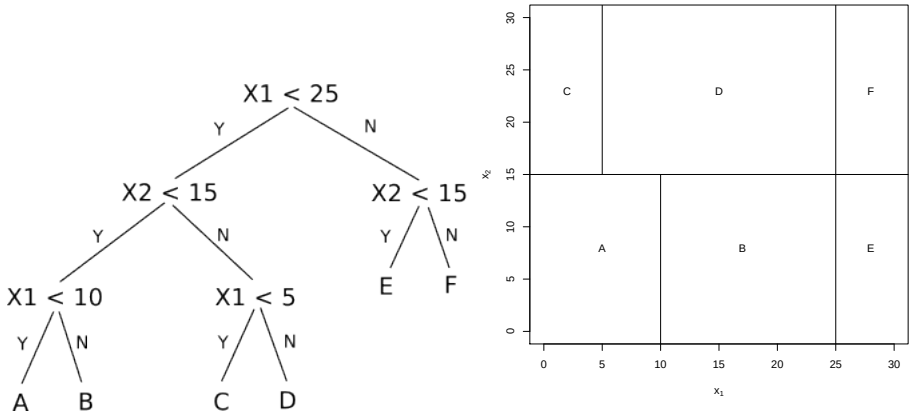

Abbildung 13: Links: Beispiel-Entscheidungsbaum, Rechts: Die Schnitte in der x_1-x_2-Ebene.
(http://www.ofai.at/ bernhard.jung/lva/SSME10/res/dt1.png)

der Bäume klassifizieren und bestimme die Klasse, die von der Mehrzahl der Bäume bestimmt wurde.

In dieser Arbeit verwenden wir immer Random Forests mit 500 Bäumen und nutzen an jeder Verzweigung 3 Variablen, um den besten Schnitt zu bestimmen.
Random Forests wurden in dieser Arbeit mit dem R-Paket `randomForest` von Liaw & Wiener (2002) erstellt.

3.5 Gewichtetes geometrisches Mittel

Um die in Kapitel 2 beschriebenen Größen Reinheit und Recall in einem einzigen Gütemaß zu vereinen, verwenden wir das gewichtete geometrische Mittel. Für Beobachtungen $x_1, ..., x_n$ ist es gegeben durch

$$\overline{x}_g = \prod_{i=1}^{n} x^{w_i}$$

mit Gewichten w_i, wobei $\sum_{i=1}^{n} w_i = 1$ gilt.
Im Falle von Reinheit und Recall bedeutet dies

$$Reinheit^{w_1} \cdot Recall^{w_2}.$$

Die Wahl der Gewichte ist dem Anwender überlassen. Da wir in dieser Arbeit mehr Gewicht auf die Reinheit als auf den Recall legen, werden die Gewichte auf 0,6 für Reinheit und 0,4 für Recall festgelegt.

3.6 Evolutionäre Algorithmen

Unter evolutionären Algorithmen (vgl. Bäck et al., 1997) versteht man eine Gruppe von Optimierungsalgorithmen, die sich - wie schon der Name impliziert - an die Evolution anlehnt. Die Idee ist, dass in einer Population jedes Individuum nach einem Optimalitätskriterium bewertet wird. Dann werden möglichst gute Individuen ausgewählt, aus denen Nachkommen erzeugt werden, die wiederum schlechtere Individuen ersetzen. So entwickelt sich die gesamte Population weiter in Richtung eines Optimums.
Die Struktur evolutionärer Algorithmen ist sehr einfach. Der Ablauf im Allgemeinen ist wie folgt:
Zunächst wird irgendwie - in der Regel zufällig gleichverteilt - eine Population von Individuen ausgewählt. Jedes dieser Individuen wird mithilfe der zu optimierenden Zielfunktion bewertet. Diese Bewertung nennt man Fitness.
Dann wird eine Gruppe von Individuen - in der Regel unter Einbeziehung der Fitness - zur Reproduktion ausgewählt. Die Auswahl dieser Eltern genannten Individuen kann auf vielfältige Weise erfolgen. Gängig sind die uniforme Selektion, bei der die Eltern zufällig gleichverteilt gewählt werden, die fitnessproportionale Selektion, die zwar auch zufällig Eltern auswählt, jedoch Individuen mit größerer Güte mit größerer Wahrscheinlichkeit auswählt als solche mit kleinerer Güte, oder auch die Turnier-Selektion, bei der zufällig eine Gruppe von Individuen ausgewählt wird, von denen dann diejenigen mit der besten Güte als Eltern dienen.
Sind die Eltern ausgewählt, werden im Crossover-Schritt daraus Nachkommen erzeugt, indem einzelne Eigenschaften der Eltern kombiniert werden. Wie viele Nachkommen erzeugt werden sollen und welche bzw. wieviele Eltern jeder Nachkomme besitzen soll, ist beliebig und dem Anwender überlassen. Häufig wird nur ein einziges von einem Elternpaar erzeugt. Die Art und Weise, wie aus den Eltern Nachkommen erzeugt werden, ist vielfältig variierbar und essenziell abhängig von der Struktur der Individuen, also des Suchraumes. Es gibt zahlreiche Crossover-Operatoren für Bitstrings oder reelle Zahlen. Hier sei auf Bäck et al. (1997) verwiesen.

Nach dem Crossover folgt die Mutation der Nachkommen. Die Mutation kann die einzige Möglichkeit sein, gewisse Teile des Suchraumes in die Suche mit einzubeziehen. So kann beispielsweise bei 0-1-Vektoren, die nur aus Nullen und Einsen bestehen, bei jedem Individuum der Population an einer bestimmten Stelle des Vektors eine 1 stehen, sodass auch alle Nachkommen immer die 1 an dieser Stelle erben werden. Der Algorithmus würde also möglicherweise nur in ein lokales Optimum laufen. Abhilfe bei diesem Problem schafft die Mutation. Hierbei werden Teile eines Nachkommen - bei Bitstrings üblicherweise einzelne Bits - zufällig variiert. Nach welchen Regeln die Variation geschieht, ist wiederum sehr vielfältig und dem Anwender überlassen. Hier sei wieder auf Bäck et al. (1997) verwiesen.
Nachdem die Nachkommen bewertet worden sind, erfolgt der letzte Schritt: Die Selektion zur Ersetzung. Hierbei werden diejenigen Individuen der Population ausgewählt, die durch die gerade erzeugten Nachkommen ersetzt werden. Die Auswahl dieser Individuen kann auf die gleiche Weise erfolgen wie die Auswahl zur Reproduktion, nur dass hierbei Individuen mit schlechter Fitness bevorzugt gewählt werden sollten.
Nach der Ersetzung wird überprüft, ob bestimmte Abbruchkriterien erfüllt sind. Ist dies nicht der Fall, wird die ganze Prozedur wiederholt, angefangen bei der Selektion zur Reproduktion. Für evolutionäre Algorithmen ist das gängigste Abbruchkriterium eine vorher gewählte Anzahl von Iterationen - bei evolutionären Algorithmen in der Regel Generationen genannt. Ist diese Anzahl Generationen erreicht, stoppt der Algorithmus. Im Folgenden ist der Ablauf des Algorithmus in einer Übersicht dargestellt:

1. Ziehe die Startpopulation.
2. Bewerte jedes Individuum in der Startpopulation.
3. Selektiere Individuen zur Reproduktion.
4. Erzeuge Nachkommen aus den selektierten Individuen.
5. Mutiere die Nachkommen.
6. Bewerte die Nachkommen.
7. Selektiere Individuen zur Ersetzung.
8. Überprüfe ein Abbruchkriterium. Ist es erfüllt, gehe weiter zu Punkt 9, ansonsten gehe zurück zu Punkt 3.

9. Ende.

In unserem Fall ist das Ziel des Algorithmus, aus den d vorliegenden Variablen eine Anzahl $k < d$ Variablen auszuwählen, die eine möglichst gute Separation von Gammas und Hadronen ermöglichen. Dazu nutzen wir einen Bitstring, dessen Länge der Anzahl der Variablen entspricht und in dem jeder Eintrag angibt, ob die entsprechende Variable bei der Separation mitgenutzt werden soll, oder nicht. Eine 1 an der Stelle i bedeutet dabei, dass die Variable verwendet wird. Liegen beispielsweise acht Variablen insgesamt vor und sollen von diesen die Variablen 1,2,5 und 7 zur Separation genutzt werden, so wird dies kodiert mit 11001010.
Wie bereits besprochen, gibt es für Bitstrings eine Vielzahl von Crossover- und Mutationsoperatoren.
Der gesamte Algorithmus wie er in dieser Arbeit verwendet wird, arbeitet wie folgt:

- **Initialisierung**: Erzeuge $m = 20$ Bitstrings deren Länge der Anzahl der Variablen im Datensatz entspricht. An jeder Stelle des Vektors soll dabei jeweils mit einer Wahrscheinlichkeit von 0,5 eine Eins oder eine Null stehen.

- **Bewertung der Startpopulation**: Wie schon beschrieben, soll das Ziel des Algorithmus sein, eine optimale Kombination aus Variablen auszuwählen, sodass die Separation von Gammas und Hadronen optimal ist. Die Zielfunktion muss also ein Gütemaß für die Separationsfähigkeit der einzelnen Variablenkombinationen sein. Dabei ist darauf zu achten, dass sich die Güte der Separation vor allem durch die Reinheit, also das Verhältnis von falsch klassierten Hadronen zu allen als Gammas klassifizierten Events, auszeichnet. Ein weiterer Faktor für die Güte der Trennung ist aber auch der Recall, also das Verhältnis von falsch klassierten Gammas zu der Anzahl aller tatsächlich vorliegender Gammas. Der Recall ist jedoch als weniger wichtig einzustufen als die Reinheit.
  Die Bewertung eines Individuums geschieht im Folgenden so, dass mit einem Trainingsdatensatz ein Random Forest mit den in dem Individuum vorliegenden Einsen entsprechenden Variablen angepasst wird. Mithilfe eines Testdatensatzes werden dann Reinheit und Recall bestimmt. Um diese beiden Werte in eine einzelne Zielgröße zu transformieren, benutzen wir das gewichtete geometrische Mittel aus beiden, wobei Reinheit ein Gewicht von 0.6 und Recall ein Gewicht von 0.4 bekommt. Die Fitness eines Individuums

S ist also:

$$f(S) = Reinheit^{0.6} \cdot Recall^{0.4}.$$

- **Selektion zur Reproduktion**: Als Selektion zur Reproduktion wird die Turnier-Selektion verwendet. Im Folgenden werden immer zwei Eltern zur Reproduktion verwendet. Es werden also zufällig $l = 5$ Individuen aus der Population ausgewählt, von denen die beiden mit der besten Fitness bestimmt werden, aus denen dann ein Nachkomme erzeugt wird.

- **Crossover**: Um aus den Eltern einen Nachkommen zu erzeugen, wird jede Stelle des Bitstrings einzeln betrachtet. Um die i-te Stelle des Nachkommen zu bestimmen, wird jeweils mit Wahrscheinlichkeit 0,5 die i-te Stelle eines der Elternteile für den Nachkommen übernommen (uniform crossover).

- **Mutation**: In dem so erzeugten Nachkommen wird jede Stelle des Bitstrings mit einer Wahrscheinlichkeit $p = 0,05$ mutiert. Das heißt, jede Eins wird mit einer Wahrcheinlichkeit von 0,05 zu einer Null und anders herum.

- **Bewertung des Nachkommen**: Die Bewertung des Nachkommen verläuft genauso wie die der initialen Population. Es wird mithilfe eines Trainingsdatensatzes ein Random Forest mit den in dem Nachkommen vorliegenden, den Einsen entsprechenden, Variablen angepasst. Mithilfe eines Testdatensatzes werden dann Reinheit und Recall bestimmt und durch das gewichtete geometrische Mittel zur Fitness vereint.

- **Selektion zur Ersetzung**: Zur Ersetzung wird stets deterministisch das Individuum mit der schlechtesten Fitness gewählt.

- **Abbruchkriterium**: Der Algorithmus stoppt nach einer vorgegebenen Anzahl von Iterationen. Das Individuum mit der zu diesem Zeitpunkt besten Fitness wird als Ergebnis ausgegeben.

  Ein ähnlicher Algorithmus wurde bereits von Helf (2009) auf MAGIC-Daten angewandt. Hier wurden jedoch die Problemstellung und einige Erkenntnisse dieser Arbeit genutzt, um einige Änderungen an dem Algorithmus vorzunehmen. Der wesentlichste Unterschied ist, dass hier kein Merkmalsgenerierungsschritt genutzt wurde und dass ein anderer Crossover-Operator und eine andere Fitness-Funktion verwendet wurden. Letztlich wurde der Algorithmus auch auf Daten mit teilweise unterschiedlichen Variablen angewendet.

3.7 t-Test

Der Student'sche t-Test ist ein statistischer Test, um Hypothesen über den Parameter μ einer normalverteilten Grundgesamtheit zu testen. Wir gehen also von einer Stichprobe $x_1, ..., x_n$ einer $N(\mu, \sigma^2)$-Verteilung aus, wobei σ^2 unbekannt ist. Wir können sowohl einseitige, als auch zweiseitige Hypothesen testen. Die einseitigen Hypothesen und Alternativen sind

H_{01}: $\mu \leq \mu_0$ gegen H_{11}: $\mu > \mu_0$ und
H_{02}: $\mu \geq \mu_0$ gegen H_{12}: $\mu < \mu_0$.

Die zweiseitige Hypothese und ihre Alternative lautet

H_{03}: $\mu = \mu_0$ $gegen$ H_{13}: $\mu \neq \mu_0$.

Um diese Hypothesen zum Niveau α zu testen, betrachten wir die Prüfgröße

$$t = \frac{\overline{x} - \mu_0}{s}\sqrt{n},$$

wobei $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ das arithmetische Mittel ist und $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ die empirische Standardabweichung der Stichprobe.

Bezeichnen wir das α-Quantil einer t-Verteilung mit $n-1$ Freiheitsgraden mit $t_{n-1;\alpha}$, so wird die Nulhypothese H_{01} verworfen, wenn

$$t > t_{n-1;1-\alpha};$$

Die Hypothese H_{02} wird verworfen, wenn

$$t < t_{n-1;\alpha};$$

Die zweiseitige Hypothese H_{03} müssen wir verwerfen, wenn

$$|t| > t_{n-1;1-\alpha/2}$$

(Hartung et al., 2005).

Alternativ kann auch jeweils der p-Wert bestimmt und abgelehnt werden, wenn dieser kleiner als α ist.

3.8 Wilcoxon-Tests

Mithilfe von Wilcoxon-Tests lassen sich ähnliche Hypothesen testen, wie mit dem t-Test, er kommt jedoch ohne Verteilungsannahmen aus. Wir nutzen in dieser Arbeit den Wilcoxon-Vorzeichen-Rangtest, um eine Stichprobe auf ihre Lage zu testen. Für den Vergleich von zwei Stichproben nutzen wir den Wilcoxon-Rangsummentest.

3.8.1 Wilcoxon-Vorzeichen-Rangtest

Der Wilcoxon-Vorzeichen-Rangtest ist ein Test auf die Lage einer Verteilung. Da er ohne jegliche Verteilungsannahmen auskommt, aber ähnliche Hypothesen testet, ist er eine Alternative zum t-Test, wenn nicht von Normalverteilung ausgegangen werden kann.
Die Annahmen bei diesem Test sind:

- Die Stichprobenvariablen $X_1, ..., X_n$ sind unabängig.
- $X_1, ...X_n$ haben eine stetige Verteilungsfunkion $F(x - \theta)$, wobei F symmetrisch um θ ist.

Die Testhypothesen lauten:

$H_{01}: \ \theta = \theta_0$ gegen $H_{11}: \ \theta > \theta_0$,
$H_{02}: \ \theta = \theta_0$ gegen $H_{12}: \ \theta < \theta_0$ und
$H_{03}: \ \theta = \theta_0$ gegen $H_{13}: \ \theta \neq \theta_0$.

Zur Überprüfung der Hypothesen definieren wir zunächst einige Hilfsgrößen. Sei $D_i = X_i - \theta_0$ mit Absolutbeträgen $|D_i| = |X_i - \theta_0|$, $i = 1, ..., n$, sowie $|D|_{(1)} < ... <$

$|D|_{(n)}$ die geordnete Statistik von $|D_1|, ..., |D_n|$. Sei weiterhin

$$V_i = \begin{cases} 1, & \text{wenn } |D|_{(i)} \text{ zu einer positiven Differenz} \\ 0, & \text{wenn } |D|_{(i)} \text{ zu einer negativen Differenz gehört.} \end{cases}$$

Dann ist die Teststatistik gegeben durch

$$V_n^+ = \sum_{i=1}^{n} V_i.$$

Zum Niveau α sind die jeweiligen Hypothesen abzulehnen, wenn

$H_{01}: \quad V_n^+ \geq v_{1-\alpha}^+,$
$H_{02}: \quad V_n^+ \leq v_{\alpha}^+,$
$H_{03}: \quad V_n^+ \geq v_{1-\alpha/2}^+ \quad \text{oder} \quad V_n^+ \leq v_{\alpha/2}^+.$

Kritische Werte v_α^+ bzw. $v_{-\alpha}^+$ sowie das Testverfahren in aller Ausführlichkeit sind in Büning & Trenkler (1994) zu finden.

3.8.2 Wilcoxon-Rangsummentest

Der Wilcoxon-Rangsummentest ist ein Verfahren, das zwei Stichproben mit Verteilungen F und G auf unterschiedliche Lage testet.
Für ihn müssen folgende Annahmen erfüllt sein:

- Die Stichprobenvariablen $X_1, ..., X_n, Y_1, ..., Y_m$ sind unabhängig.
- $X_1, ..., X_n$ und $Y_1, ..., Y_m$ haben stetige Verteilungsfunktionen F bzw. G.

Es gibt wiederum zwei einseitige und ein einseitiges Testproblem.

$H_{01}: \quad G(z) = F(z) \quad \text{gegen} \quad H_{11}: \quad G(z) = F(z-\theta) \quad \text{für alle } z \in \mathbb{R}, \theta < 0$
$H_{02}: \quad G(z) = F(z) \quad \text{gegen} \quad H_{12}: \quad G(z) = F(z-\theta) \quad \text{für alle } z \in \mathbb{R}, \theta > 0$
$H_{03}: \quad G(z) = F(z) \quad \text{gegen} \quad H_{13}: \quad G(z) = F(z-\theta) \quad \text{für alle } z \in \mathbb{R}, \theta \neq 0$

Wir bestimmen den Vektor $(Z_1, ..., Z_N)$ mit $n+m = N$, wobei $Z_i = 1$ ist, falls die i-te Variable in der kombinierten, geordneten Statistik eine X-Variable ist und $Z_i = 0$,

falls es sich um eine Y-Variable handelt.
Die Teststatistik ist dann gegeben durch

$$W_N = \sum_{i=1}^{N} iV_i.$$

Die jeweiligen Hypothesen werden abgelehnt, wenn

$$\begin{aligned} H_{01}: \quad & W_N \geq w_{1-\alpha}, \\ H_{02}: \quad & W_N \leq w_{\alpha}, \\ H_{03}: \quad & W_N \geq w_{1-\alpha/2} \quad \text{oder} \quad W_N \leq w_{\alpha/2}. \end{aligned}$$

Kritische Werte w_α und eine ausführliche Beschreibung des Tests sind wiederum in Büning Trenkler (1994) zu finden.

3.9 Hexagonale Klassierung

Aufgrund der sehr großen Anzahl von Beobachtungen können Diagramme, in denen zwei Variablen gegeneinander aufgetragen werden, sehr unübersichtlich werden, da viele Beobachtungen übereinander liegen. Aus diesem Grund klassieren wir die Daten vor der Darstellung in Sechsecke und färben jedes Sechseck gemäß der Anzahl der Beobachtungen in diesem Sechseck. Hier ist darauf zu achten, dass die so zur Klassierung verwendeten Sechsecke nichts mit den Pixeln der Kamera zu tun haben und unabhängig von diesen bestimmt werden.
Alle Diagramme mit so klassierten Daten in dieser Arbeit wurden mithilfe der R-Pakete `hexbin` von Carr et al. (2009) und `RColorBrewer` von Neuwirth (2007) erstellt.

4 Datenexploration

Wir werden nun die vorliegenden Daten explorativ analysieren. Dazu betrachten wir jede Variable einzeln. Zunächst untersuchen wir den Datensatz aber insgesamt auf Auffälligkeiten.

4.1 Erste Bereinigungen

Zunächst überprüfen wir den Datensatz auf fehlende und unsinnige Werte. Obwohl keine fehlenden Werte vorhanden sind, gibt es doch einige Beobachtungen, die Werte aufweisen, die eigentlich nicht auftreten dürften. So gibt es zum Beispiel Gamma-Events, bei denen *Width* und damit auch *Area* gleich Null sind. Genauso treten bei 20 der 1.122.061 Gamma-Events negative Werte in der Variable *SizeSubIslands* auf.
Unter den Hadronen gibt es dagegen ein Event, bei dem ein *Delta*-Wert außerhalb des zulässigen Bereichs zwischen $-\frac{\pi}{2}$ und $\frac{\pi}{2}$ auftritt.
Da die Anzahl dieser fehlerhaften Werte sehr gering ist im Vergleich zu der Gesamtzahl der Beobachtungen und da nicht nachvollziehbar ist, wie es zu den Fehlern gekommen ist und ob diese Fehler auch in anderen Variablen zu fehlerhaften Werten geführt hat, entfernen wir diese Beobachtungen aus dem Datensatz und schließen sie damit von weiteren Analysen aus.
Im reduzierten Datensatz befinden sich damit 1.122.040 Gamma- und 1.303.251 Hadronenevents. In den Folgenden Unterkapiteln werden nun die einzelnen Variablen analysiert.

4.2 *Length* / *Width* / *Area* / *Ellip*

Die Variablen *Length*, *Width* und *Area* beschreiben die Größe der angepassten Ellipse. *Length* ist dabei die Länge der größeren Halbachse, *Width* die der kleineren. *Area* beschreibt die Fläche der Ellipse, setzt sich also multiplikativ aus *Length* und *Width* zusammen. Es gilt:

$$Area = \pi \cdot Length \cdot Width.$$

Wir konstruieren zur Auswertung in dieser Arbeit außerdem noch eine vierte Variable *Ellip*. Wir definieren sie als

$$Ellip = \frac{Length - Width}{Length}$$

und gibt die „Elliptizität“ (engl. ellipticity) an, d.h. die Verformung des Ellipsoiden im Vergleich zu einem Kreis. Diese Variable kann nur Werte zwischen 0 und 1 annehmen, wobei 0 bedeutet, dass es sich bei dem Ellipsoiden um einen Kreis handelt und 1, dass er in Richtung der kleineren Halbachse keine Ausdehnung besitzt, dass es sich also um eine Linie handelt. Wir betrachten diese Variable, da bekannt ist, dass die von Gammas erzeugten Ellipsen bei gleicher Energie in der Regel „langgezogener“ sein sollten als die von Hadronen erzeugten. Dies ist in Abbildung 1 in Abschnitt 2.1 zu sehen. Einige Lage- und Streuungsmaße sind in Tabelle 2 dargestellt.

	Length		*Width*		*Area*		*Ellip*	
	Hadronen	Gammas	Hadronen	Gammas	Hadronen	Gammas	Hadronen	Gammas
Minimum	15,15	15.37	5,98	0,00	659	0	0,00	0,00
1. Quartil	27,99	27,75	15,77	15,64	1.446	1.403	0,33	0,38
Median	42,49	38,32	19,65	18,41	2.594	2.179	0,51	0,52
Arithm. Mittel	59,51	46,29	25,37	20,70	6.272	3.475	0,49	0,50
3. Quartil	78,49	58,10	28,17	22,99	6.634	3.937	0,66	0,63
Maximum	381,29	375,06	257,24	184,73	216.810	112.910	0,96	1,00
Standardabw.	42,96	24,50	16,20	8,55	9.331	3.729	0,20	0,17
Schiefe	1,53	1,33	2,62	3,25	3,75	4,26	-0,25	-0,51

Tabelle 2: Lage- und Streuungsmaße der vier Variablen *Length*, *Width*, *Area* und *Ellip*.

Zunächst ist auffällig, dass *Length*, *Width* und damit auch *Area* ein im Vergleich zu den drei Quartilen sehr hohes Maximum und im Zusammenhang damit auch eine recht starke Rechtsschiefe haben. Wie in den Histogrammen in Abbildung 14 zu sehen ist, sorgt diese Schiefe für eine große Unübersichtlichkeit der Daten. Eine geeignete Transformation könnte die Schiefe kompensieren und die Daten damit auch übersichtlicher gestalten. Siehe dazu den Abschnitt 5.1.
Des Weiteren ist schon an den Lagemaßen ein Unterschied zwischen Gammas und Hadronen zu erkennen. So sind die arithmetischen Mittel und alle drei Quartile von Length, *Width* und *Area* bei den Gammas kleiner als bei den Hadronen. Am deutlichsten ist der Unterschied in der Variable *Area*, sodass es so scheint, als wenn sich *Area* zur Separation von Gammas und Hadronen besser eignet als

Length und *Width* allein.
Dabei könnte aber die große Standardabweichung von *Area* in den Hadronen im Vergleich zu den Gammas problematisch sein, da sich die Verteilungen dadurch stark überdecken. Diese Überdeckung ist auch in den Histogrammen in Abbildung 14 zu sehen.
Die Verteilung der selbst konstruierten Variable *Ellip* scheint sich recht deutlich von den anderen zu unterscheiden, was den Unterschied von Hadronen und Gammas betrifft. Hier liegen die ersten beiden Quartile und das arithmetische Mittel bei den Gammas höher als bei den Hadronen. Der Unterschied ist allerdings klein.
Mehr Aufschluss über die Daten geben hier wieder die Histogramme in Abbildung 14. Hier ist der Unterschied zwischen Hadronen und Gammas, der schon mit bloßem Auge besser zu sehen ist als bei den anderen Variablen, sehr deutlich zu erkennen.
Zunächst ist offenbar der nicht in Tabelle 2 aufgeführte Modalwert der Daten, der sich durch die Schiefe recht deutlich vom arithmetischen Mittel und vom Median unterscheidet, bei den Gammas deutlich niedriger als bei den Hadronen. In diesem Zusammenhang ist die nächste Auffälligkeit, dass Gammas seltener als Hadronen hohe Werte von mehr als 0,7 annehmen, was der Behauptung, dass die von Gammas erzeugten Ellipsen „langgezogener“ sind, widerspricht. Andererseits ist auch noch die „bauchige“ Form der Verteilung in den Hadronen für Werte, die unter dem Modalwert liegen, auffällig.
Insgesamt scheint das Verhältnis von *Width* zu *Length* bei Hadronen gleichmäßiger über den gesamten möglichen Bereich zu streuen als bei Gammas, sodass sowohl ein stark nach oben, als auch ein nach unten vom Modalwert der Gammas abweichender Wert eher dafür spricht, dass es sich um ein Hadron handelt. Die Variable *Ellip* scheint sich also auch gut zur Separation zu eignen.

4.3 *MeanX* und *MeanY*

MeanX und *MeanY* geben die x- und y-Koordinate des Schwerpunkts (Center Of Gravity, COG) des Schauers auf der Kamera-Ebene an. Sie sind mit *MeanX* und *MeanY* gekennzeichnet, weil es sich um mit den Intensitäten der Pixel gewichtete arithmetische Mittel der x- bzw. y-Koordinaten der genutzten Pixel handelt. Das COG muss also nicht mit dem Schnittpunkt der Halbachsen oder einem Brennpunkt übereinstimmen.

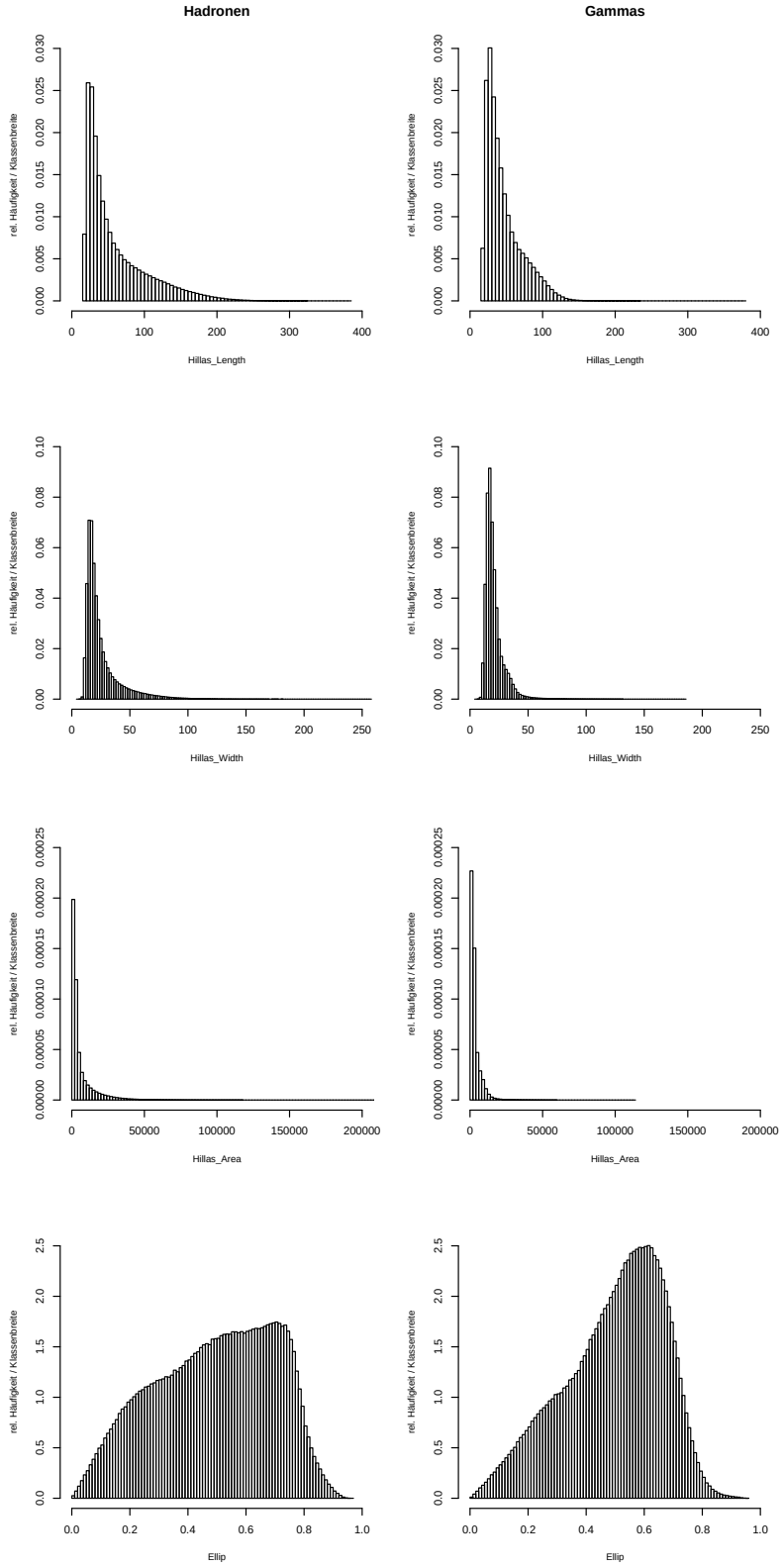

Abbildung 14: Histogramme von *Length*, *Width*, *Area* und *Ellip*

Einige Lage- und Streuungsmaße dieser beiden Variablen sind in Tabelle 3 zusammengestellt.

	MeanX		*MeanY*	
	Hadronen	Gammas	Hadronen	Gammas
Minimum	-456,68	-449,76	-429,43	-422,65
1. Quartil	-123,40	-41,18	-125,35	-124,58
Median	-6,55	58,89	-19,41	-3,55
Arithm. Mittel	-4,43	56,08	-13,31	-2,64
3. Quartil	114,90	167,68	102,70	118,77
Maximum	454,04	456,68	426,98	421,30
Standardabw.	151,43	146,24	145,96	153,44
Schiefe	0,04	-0,34	0,07	0,02

Tabelle 3: Lage- und Streuungsmaße der vier Variablen

Die deutlichste Auffälligkeit in Tabelle 3 ist, dass das COG bei Gammas offenbar im Mittel einen deutlich höheren x-Wert hat als bei Hadronen. Dies ist jedoch einfach zu erklären, da das Teleskop im Wobble-Modus arbeitet, d.h. die Quellposition, von der Gammas ausgesendet werden, liegt nicht im Kamera-Mittelpunkt, sondern ist in x-Richtung verschoben. Dementsprechend streuen die Gamma-Signale um diese Quellposition und nicht um den Kamera-Mittelpunkt.
Damit lässt sich aber nicht die große Abweichung des arithmetischen Mittels der y-Werte der Hadronen vom Nullpunkt erklären. Woher diese kommt, ist unklar. Die Abweichung ist nicht groß, doch scheint es so zu sein, dass Hadronen in y-Richtung eher negative Werte realisieren. Tatsächlich lehnt ein t-Test mit der Nullhypothese H_0: „Der wahre Erwartungswert der Variable *MeanY* ist größer oder gleich Null“ gegen die Alternative H_1: „Der wahre Erwartungswert ist kleiner als Null“ mit einem sehr kleinen p-Wert kleiner $2,2 \cdot 10^{-16}$ die Nullhypothese zum 5%-Niveau deutlich ab. Der sehr kleine p-Wert ist auf die sehr große Anzahl an Beobachtungen von über einer Million zurückzuführen, die trotz der hohen Varianz der Variable den Bereich, in dem die Nullhypothese nicht abgelehnt wird, sehr stark einschränkt. Ein 95% - Konfidenzintervall ist gegeben durch $[-\infty, -13{,}10]$. Dieses Phänomen, dass die Grenze(n) der Konfidenzintervalle nur wenig vom arithmetischen Mittel abweichen, wird uns noch öfter begegnen.
Eine weitere Auffälligkeit in Tabelle 3 ist die Linksschiefe der x-Werte der Gammas. Auch dieses Phänomen lässt sich mit dem Wobble-Modus erklären. Da sich die Quellposition auf der rechten Seite des Kamera-Mittelpunkts befindet und da die Kamera in beide Richtungen beschränkt ist, können die Werte nach unten weiter streuen als nach oben. Dies ist auch in Abbildung 16 zu sehen, auf die weiter unten noch genauer eingegangen wird.

Um weitere Einblicke in die Daten zu bekommen, sind in Abbildung 15 Histogramme dargestellt.
Teilweise bestätigen die Histogramme den durch die Lagemaße gewonnenen Eindruck. So ist zum Beispiel die Linksschiefe der Gamma-x-Werte deutlich zu sehen. Auch die Verschiebung der Hadronen-y-Werte in den negativen Bereich ist gut zu sehen, wobei hier noch deutlich wird, dass die Verteilung offenbar multimodal ist. Dies ist deshalb der Fall, weil Werte zwischen 0 und 100 offenbar seltener realisiert werden als man dies intuitiv vermuten würde.
Eine Erklärung für diese Verformung könnte sein, dass der Hadronendatensatz Gammas enthält. Da wir den genauen Anteil an Gammas nicht kennen, schließen wir dies als mögliche Erklärung nicht aus. Aufklärung könnten hier simulierte Hadronen-Daten bieten, bei denen bekannt ist, dass sie keine Gammas beinhalten.
Eine andere Erklärung dafür könnten defekte Kamerapixel in diesem Bereich sein. Dies ist ein weiterer Grund, statt Originaldaten simulierte Hadronen-Daten zu verwenden.
Was in Tabelle 3 nicht deutlich wurde, ist die symmetrische Multimodalität der Gamma-y-Werte. Es scheint also eine „bevorzugte“ Entfernung der Gammas von der Quellposition zu geben, die in den x-Werten (wohl wegen der Verschiebung der Quellposition) nicht zu sehen ist.

Da die beiden Variablen *MeanX* und *MeanY* zusammen genommen die genaue Lage des COG auf der Kameraebene angeben, ist es sinnvoll, sie gegeneinander aufzutragen, um sich ein Bild davon zu machen, wie die Verteilung der Schauer über die gesamte Kamera aussieht. Da wir hier sehr viele COGs auftragen, teilen wir den gesamten Bereich in Sechsecke ein und färben sie gemäß der Anzahl der Beobachtungen im jeweiligen Sechseck.
Die Plots sind in Abbildung 16 zu sehen.

Hier ist die sechseckige Form der Kamera sehr gut zu erkennen. Außerdem ist die Zweiteilung in einen inneren und einen äußeren Bereich zu sehen. Im äußeren Bereich scheinen sehr viel seltener Ellipsenschwerpunkte gemessen zu werden, als im inneren. Dies ist deshalb der Fall, da Events, die zu einem großen Teil im äußeren Kameraring auftreffen, gar nicht aufgezeichnet werden. Die Pixel im äußeren Ring werden nur dann gebraucht, wenn das Event eigentlich im inneren Bereich liegt, aber so groß ist, dass auch Pixel im äußeren benötigt werden, um es ganz abzubilden. Außerdem werden durch einen Quality-Cut alle Beobachtungen

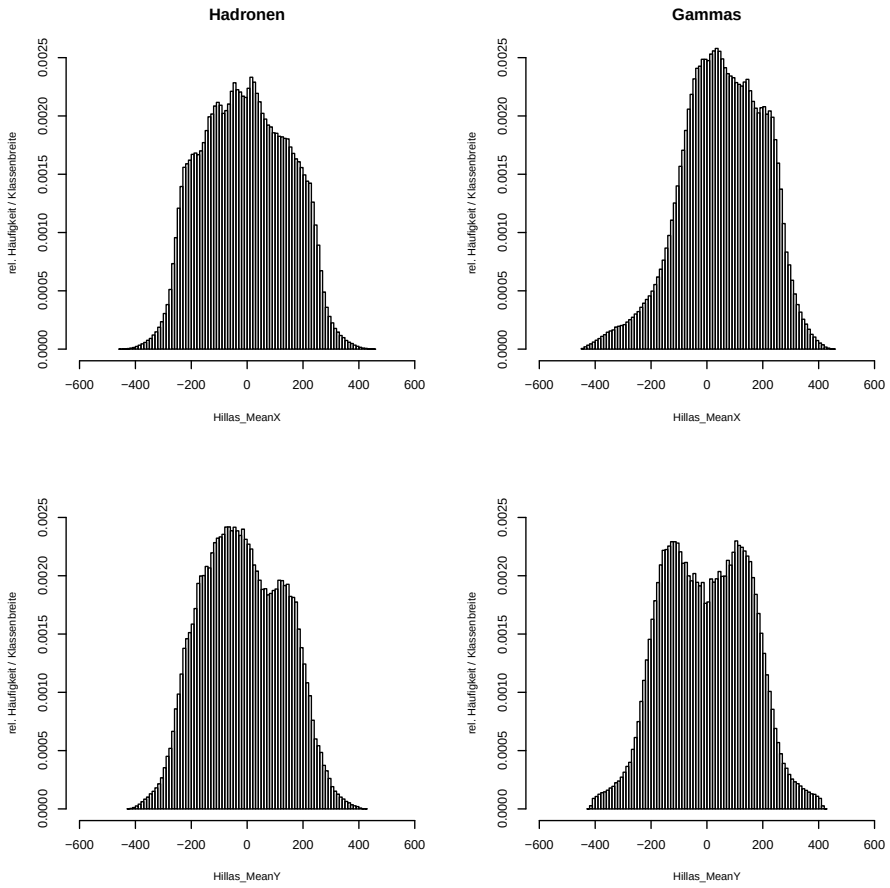

Abbildung 15: Histogramme der Variablen *MeanX* und *MeanY*

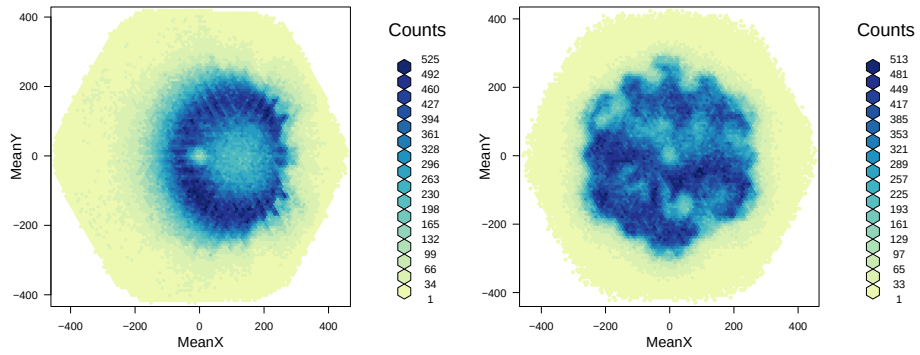

Abbildung 16: MeanX und MeanY gegeneinander aufgetragen und hexagonal klassiert. (links: Gammas, rechts: Hadronen)

entfernt, bei denen Leakage1 größer als $0{,}3$ ist. Vor Allem aber liegt dies an den in Abbildung 4 gezeigten Treiggerzonen, die nur im inneren Bereich der Kamera vorhanden sind.
Zudem ist hier nun ein ganz deutlicher Unterschied zwischen Hadronen und Gammas erkennbar. Während die Hadronen eher gleichmäßig über den gesamten Bereich streuen, formen die Gammas einen „Ring“ um ihre Quellposition. Dabei ist aber die Quellposition so weit verschoben, dass ein Teil des Rings auf den äußeren Kamerabereich fällt. Er wird dadurch abgeschnitten und erscheint unsymmetrisch. Dies ist wohl auch die Erklärung für die Asymmetrie der Variable *MeanX* und für die Mehrgipfligkeit der Variable *MeanY*. Die Vermutung liegt nahe, dass beide zumindest symmetrisch wären, läge die Quellposition im Kameramittelpunkt. Ferner liegt der Kameramittelpunkt, an dem sich kein Photomultiplier befindet und an dem entsprechend keine Messungen aufgezeichnet werden können, an einer Stelle großer Konzentration der Gammas, sodass auch hier die echte Verteilung verfälscht wird.
Bei den Hadronen gibt es ebenfalls mehrere Auffälligkeiten. Obwohl die Verteilung der Werte über die Kamera gleichmäßiger ist als bei den Gammas, so gibt es doch Bereiche, in denen eine deutlich höhere Konzentration festzustellen ist. Die Werte sind also offenbar nicht über den gesamten Kamerabereich gleichverteilt (auch dann nicht, wenn man den äußeren Bereich ausschließen würde). Eine Systematik, anhand derer man bevorzugte Bereiche der Hadronen ausmachen könnte, ist jedoch nicht auszumachen. Deutlich zu erkennen ist jedoch die „Lücke“ bei y-Werten zwischen 0 und 100. Wie schon im Histogramm ist hier zu sehen, dass in diesem Bereich offenbar die Konzentration von Hadronen deutlich geringer ist als in anderen Bereichen. Eine Erklärung für dieses Phänomen liefert aber auch diese Darstellung der Daten nicht.
Eine weitere Auffälligkeit ist, dass es offenbar „Inseln“ mit ähnlichen Konzentrationen gibt. Besonders auffällig ist dies bei den am weitesten oben liegenden Pixeln des inneren Bereichs, aber auch bei den meisten anderen. So ist der linke Rand des inneren Bereichs in drei Teile zu unterteilen, von denen die unteren beiden durch Pixel mit geringerer Intensität voneinander getrennt sind, wobei der obere generell niedrigere Intensitäten aufweist.
Weiterhin scheint das innere Sechseck im Vergleich zum äußeren gedreht zu sein. Es steht auf einer Spitze, statt wie das äußere auf einer Seite zu liegen. Auch dieses Phänomen ist mit den Triggerzonen erklärbar. Die Bereiche stärkerer Konzentration in Abbildung 16 stimmen genau mit den in Abbildung 4 gzeigten Triggerzonen des inneren Kamerabereichs überein.

4.4 *Delta*

Der Hillas-Parameter *Delta* gibt den Winkel der größeren Hauptachse der angepassten Ellipse zur x-Achse an. Ein Winkel von 0 bedeutet, dass die Hauptachse parallel zur x-Achse liegt. $\pm\frac{\pi}{2}$ bedeutet entsprechend eine Senkrechte. Demnach liegen die Extrema dieser Variablen auch etwa bei diesen Werten, wie in Tabelle 4 zu sehen ist.
Gleichzeitig liegen arithmetisches Mittel und Median sowohl bei den Gammas als auch bei den Hadronen etwa bei 0. Ein zweiseitiger Wilcoxon-Test auf Zentralwert 0 lehnt für Hadronen mit einem p-Wert von 0,29 zum 5%-Niveau nicht ab. Für Gammas tut er dies jedoch mit einem p-Wert von $2,2 \cdot 10^{-}16$. Trotz der augenscheinlich geringen Abweichung des arithmetischen Mittels von 0 ist die Abweichung nach unten offenbar hoch signifikant.

	Delta	
	Hadronen	Gammas
Minimum	-1,57	-1,57
1. Quartil	-0,80	-0,83
Median	0,00	-0,05
Arithm. Mittel	0,00	-0,02
3. Quartil	0,80	0,78
Maximum	1,57	1,57
Standardabw.	0,92	0,92
Schiefe	0,00	0,04

Tabelle 4: Lage- und Streuungsmaße der Variable *Delta*

Auffällig ist, dass die empirische Verteilung in den Hadronen offenbar symmetrisch ist, bei den Gammas jedoch leicht rechtsschief, was das Abweichen des arithmetischen Mittels von Null nach unten erklärt. Worin diese Rechtsschiefe begründet liegt, ist jedoch unklar.
Die Rechtsschiefe wird im Histogramm in Abbildung 17 nur schwach deutlich.
Außerdem ist hier abzulesen, dass offenbar bei Gammas Werte um 0 seltener vorkommen als andere. Dies hängt wiederum mit dem Wobble-Modus zusammen. Die Verschiebung der Quellposition und das damit verbundene geringe Vorkommen von Events, deren COG in y-Richtung nahe Null liegt, sorgt dafür, dass auch ein Delta in der Nähe von Null weniger häufig realisiert wird, da gerade solche Gamma-Events, die in y-Richtung bei Null liegen, sehr häufig auch ein Delta nahe Null realisieren. Dies ist in Abbildung 18 gut zu sehen. Das hat damit zu tun, dass

die größere Halbachse von Gamma-Schauern sich in der Regel auf die Quellposition ausrichtet. Bei Hadronen ist ein ähnliches Verhalten nicht zu erwarten. Abbildung 18 zeigt erwartungsgemäß keinen Zusammenhang zwischen Lage und Ausrichtung der Events. Daher ist unklar, weshalb Hadronen so selten Werte um Null realisieren.

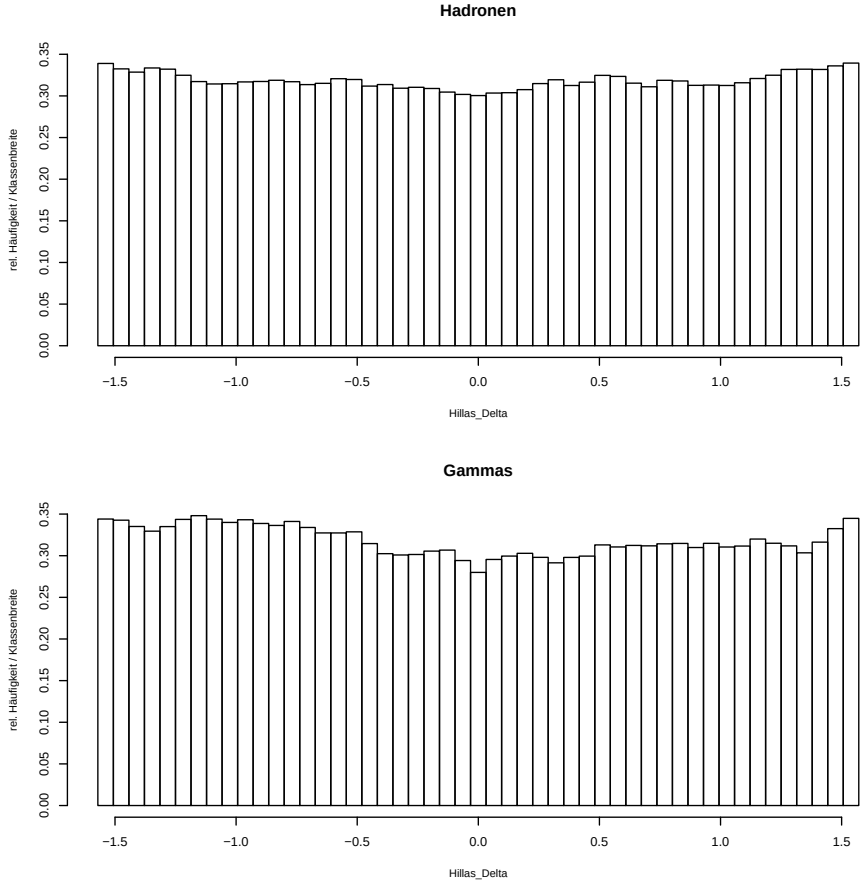

Abbildung 17: Histogramme der Variable *Delta*

Obwohl beide Verteilungen recht stark an eine Gleichverteilung erinnern, sprechen vor allem bei den Gammas die für diese Anzahl von Beobachtungen sehr starken Schwankungen dagegen.
Für sich genommen scheint die Variable *Delta* nicht sehr gut zur Separation von Gammas und Hadronen geeignet zu sein, da sich die Verteilungen sehr stark ähneln. Im Zusammenhang mit anderen Variablen sind jedoch sehr starke Unterschiede festzustellen, wie man zum Beispiel in Abbildung 18 sieht.

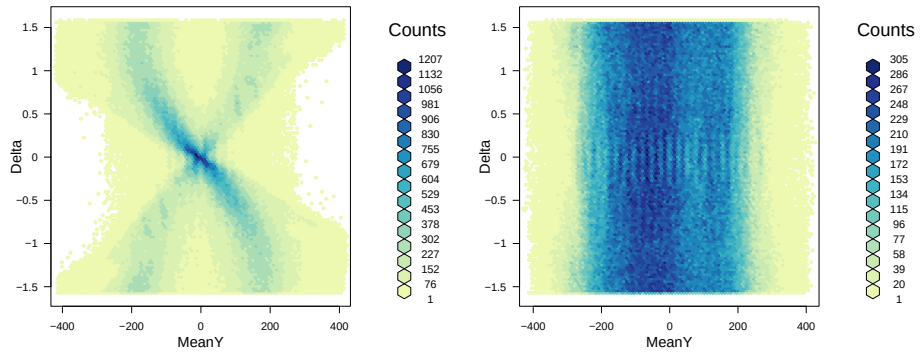

Abbildung 18: Delta gegen MeanY aufgetragen. Bei den Gammas ist eine klare Struktur zu erkennen, während die Variablen für Hadronen völlig unabhängig zu sein scheinen (links: Gammas, rechts: Hadronen)

4.5 *Size* / *SizeMainIsland* / *SizeSubIslands*

Es gibt insgesamt drei *Size*-Variablen: *Size*, *SizeMainIsland* und *SizeSubIslands*. Die *Size* eines Schauers ist die Summe der in allen benutzten Pixeln aufgenommenen Intensitäten. Da die Gesamtintensität und die Energie eines Schauers hoch korreliert sind, kann diese Variable als eine erste Abschätzung der Energie des aufgenommenen Schauers herangezogen werden.
SizeMainIsland und *SizeSubIslands* sind entsprechend die Summe der Intensitäten der Pixel der Haupt- bzw. Nebeninseln. Es gilt also

$$\mathit{Size} = \mathit{SizeMainIsland} + \mathit{SizeSubIslands}.$$

Tatsächlich ist dies im vorliegenden Datensatz nicht bei allen Beobachtungen der Fall. Die relative Abweichung ist dabei jedoch nicht größer als 10^{-5}. Die Abweichungen sind auf Rundungsfehler o.Ä. zurückzuführen.
In Tabelle 5 sind Lage- und Streuungsmaße dargestellt. Alle drei Variablen scheinen sehr stark rechtsschief zu sein. Dadurch ist sogar das arithmetische Mittel überall größer als das 3.Quartil.
SizeSubIslands ist bei mehr als 75% der Daten gleich 0. Dazu schauen wir uns noch die Variable *NumIslands* (Abschnitt 4.7) an, also die Anzahl der Neben-Inseln. Vermutlich ist *SizeSubIslands* deshalb gleich Null, weil es schlicht keine Neben-Inseln gibt.

Gammas scheinen tendenziell größere Werte bei der Size der Hauptinsel zu erzeugen, aber kleinere bei den Neben-Inseln. Dieser Unterschied schlägt sich aber fast nicht in den Quartilen wieder, sodass davon auszugehen ist, dass sich Gammas und Hadronen hier nur in extremen Werten, also einem kleinen Bruchteil der Daten unterscheiden.
Auffällig ist auch die deutlich höhere Schiefe der Hadronen gegenüber den Gammas bei *Size* und *SizeMainIsland*, während dies bei *SizeSubIslands* genau anders herum ist. Ebenso verhält es sich mit der Standardabweichung.

	Size		*SizeMainIsland*		*SizeSubIslands*	
	Hadronen	Gammas	Hadronen	Gammas	Hadronen	Gammas
Minimum	30,39	27,29	18,17	17,80	0,00	0,00
1. Quartil	90,30	75,44	87,05	74,28	0,00	0,00
Median	139,56	130,77	132,94	129,35	0,00	0,00
Arithm. Mittel	446,28	1.624,31	434,38	1.622,96	11,90	1,35
3. Quartil	265,69	332,55	249,44	331,16	0,00	0,00
Maximum	166.248,00	208.326,00	166.237,00	208.326,00	4.525,50	1.028,19
Standardabw.	2.068,21	6.950,30	2.064,54	6.949,60	40,08	9,90
Schiefe	21,49	8,17	21,59	8,17	8,48	23,73

Tabelle 5: Lage- und Streuungsmaße der *Size*-Variablen

Untransformiert wird es durch die starke Schiefe keinen Sinn haben sich Histogramme anzusehen. Außer einer Säule in der Nähe der 0 wird nichts zu sehen sein. Daher betrachten wir die Log-Transformierten Werte im Histogramm in Abbildung 19.
Aufgrund der vielen Nullen und der negativen Werte in *SizeSubIslands* ist dies hier jedoch nur in begrenztem Maße möglich, sodass wir hier zunächst eine Eins addieren und die Werte dann transformieren. So bleibt die Null als Minimum erhalten.

In den Histogrammen ist zu sehen, dass die Histogramme zu *Size* und *SizeMainIsland* fast identisch aussehen. Tatsächlich ist der Korrelationskoeffizient von Pearson zwischen den beiden Variablen größer als 0,99. Ebenso verhält es sich bei den Hadronen. Die Informationen in den Variablen sind also redundant. Es reicht scheinbar aus, eine von ihnen zu betrachten.
In den Histogrammen ist zu sehen, dass es bei den Gammas jeweils mehr hohe Werte als bei den Hadronen gibt. Gleichzeitig kommen aber auch sehr kleine Werte unter 4 häufiger vor als in den Hadronen. Hier scheinen also sehr große und sehr kleine Werte eher für ein Gamma als für ein Hadron zu sprechen. Im Mittelteil sind die Histogramme jedoch kaum voneinander zu unterscheiden.
In der Variable *SizeSubIslands* sind vor allem die Werte zu sehen, die gleich Null

sind. Dies scheint bei Gammas öfter der Fall zu sein, sodass wir davon ausgehen können, dass Gammas seltener Nebeninseln erzeugen.
Es ist aber in beiden Histogrammen noch zu sehen, dass es bei den positiven Werten von *SizeSubIslands* ein Maximum gibt. Bei den Gammas liegt dieses jedoch bei kleineren Werten als bei Hadronen. In dieser Variable ist also ein deutlicherer Unterschied zwischen Gammas und Hadronen zu sehen als in den anderen beiden. Es empfiehlt sich, aufgrund von Rechenzeit- und Speicherplatzersparnis, nur die Variablen *Size* und *SizeSubIslands* weiter zu betrachten.

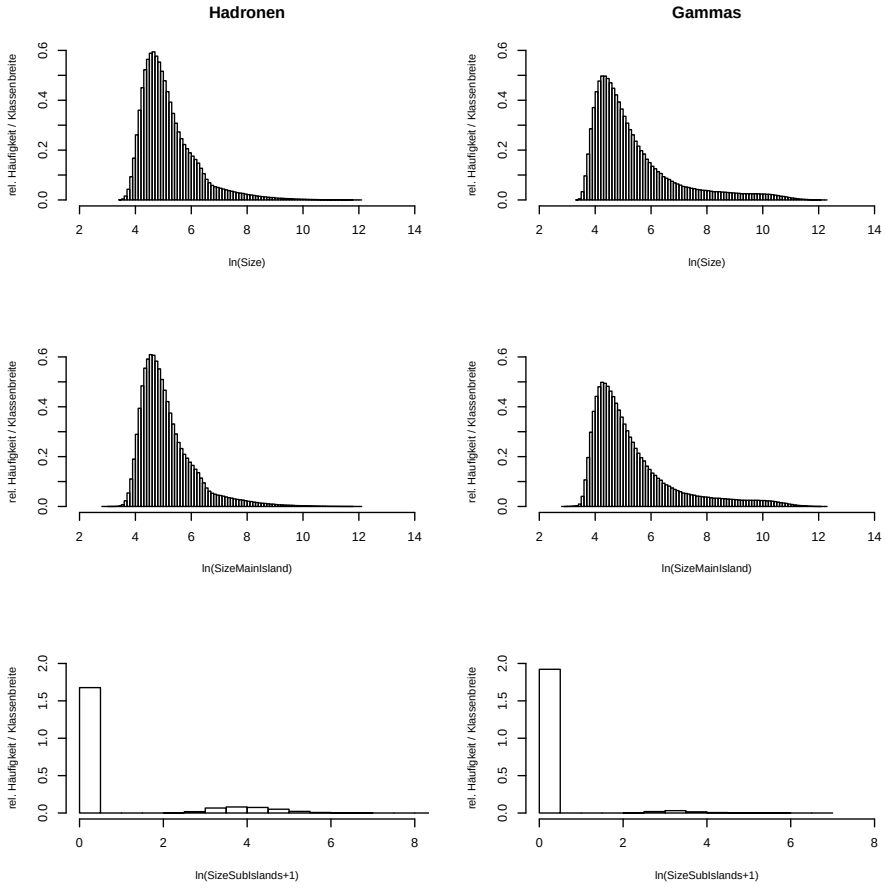

Abbildung 19: Histogramme von $\ln(\textit{Size})$, $\ln(\textit{SizeMainIsland})$ und $\ln(\textit{SizeSubIslands}+1)$

4.6 *SlopeLong* / *SlopeTrans*

Slope ist derjenige Hillas-Parameter, der die Zeitkomponente des Schauers enthält. Es ist der Zeitgradient des Schauers und gibt die „Geschwindigkeit“ an, mit der sich der Schauer über die Kamera bewegt hat. Dabei wird zwischen transversalem (*SlopeTrans*) und longitudinalem (*SlopeLong*) Slope unterschieden. Transversal ist hier in Richtung der kleineren Halbachse der Ellipse, longitudinal in Richtung der größeren. In Tabelle 6 sind wiederum einige Lage- und Streuungsmaße dargestellt.

	SlopeTrans		*SlopeLong*	
	Hadronen	Gammas	Hadronen	Gammas
Minimum	-0,19243	-0,22724	-0,22835	-0,21863
1. Quartil	-0,00448	-0,00492	-0,00299	-0,00592
Median	0,00007	0,00007	0,00006	-0,00118
Arithm. Mittel	0,00010	0,00009	0,00009	-0,00129
3. Quartil	0,00464	0,00510	0,00317	0,00318
Maximum	0,21544	0,22507	0,16731	0,22806
Standardabw.	0,00923	0,00973	0,00888	0,00806
Schiefe	0,07461	0,04711	0,00947	0,05733
Wölbung	11,68111	8,06560	17,57415	15,11860

Tabelle 6: Lage- und Streuungsmaße der beiden Slope-Variablen

Auf den ersten Blick sind Auffälligkeiten bei *SlopeLong* zu sehen. Das arithmetische Mittel und der Median liegen hier im Vergleich zu den anderen Variablen deutlich unter 0. Die Quartile und Extrema sind jedoch sehr ähnlich.
In *SlopeTrans* dagegen scheinen sich Hadronen und Gammas fast nicht zu unterscheiden. Ein Zweistichproben-Wilcoxon-Test zwischen zwischen Hadronen und Gammas ergibt für *SlopeTrans* einen p-Wert von 0,94. Für *SlopeLong* ist der p-Wert kleiner $2,2 \cdot 10^{-16}$. Es scheint also in *SlopeLong* einen signifikanten Unterschied zwischen den Mitteln von Gammas und Hadronen zu geben. In *SlopeTrans* dagegen nicht. Ein erster Hinweis darauf, dass sich SlopeLong besser zur Klassifikation eignet als *SlopeTrans*.
Beiden Variablen gemein ist die offenbar starke Konzentration der meisten Daten in der Nähe des Mittels. Aus diesem Grund ist in Tabelle 6 zusätzlich die Wölbung angegeben. Die sehr hohen Werte weisen auf eine sehr starke Konzentration der Werte am arithmetischen Mittel und damit auf eine starke „Spitzigkeit“ der Daten hin.
Durch die stark von 0 abweichende Wölbung ist hier auszuschließen, dass die Daten aus einer Normalverteilung stammen. Eine Annahme, die bei Betrachtung der

Histogramme in Abbildung 20 zunächst sinnvoll erscheint, obwohl die „Spitzigkeit“ auch dort zu sehen ist.
In den Histogrammen ist auch der Grund für die Verschiebung des Mittelwerts von *SlopeLong* in den Gammas zu sehen. Es gibt deutlich mehr negative Beobachtungen als positive. Das bedeutet, dass sich Gammaschauer häufiger von der Quellposition weg bewegen als darauf zu.
Es ist außerdem noch zu sehen, wie wenig sich die Histogramme von *SlopeTrans* für Gammas und Hadronen unterscheiden.

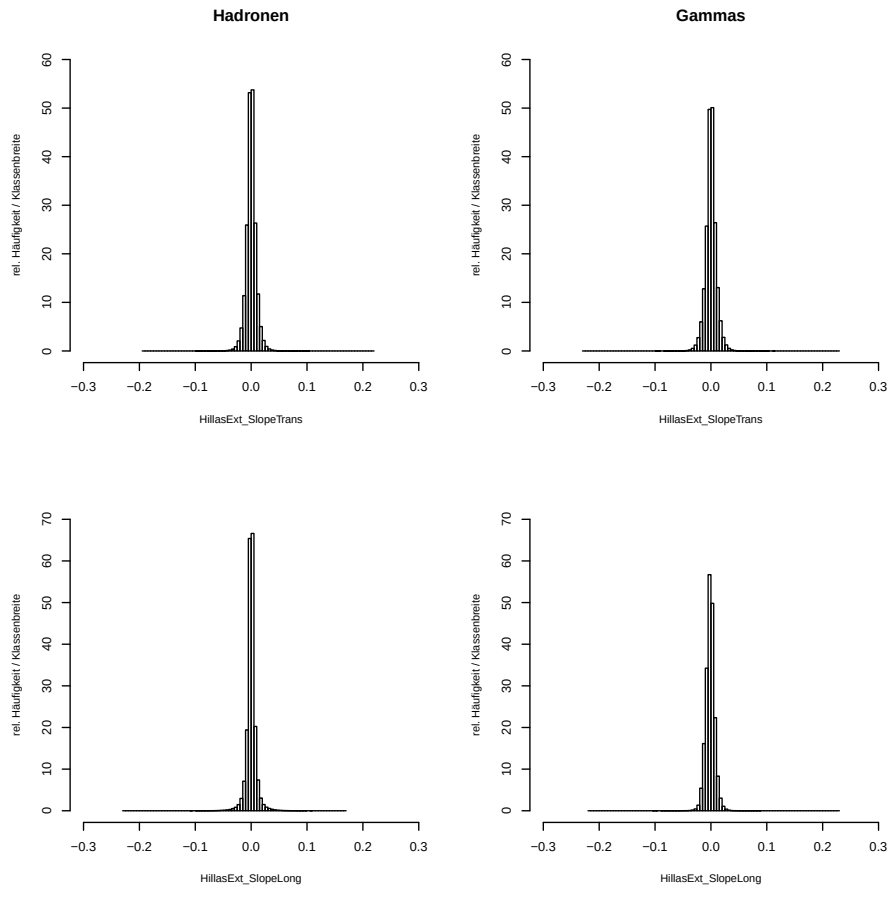

Abbildung 20: Histogramme der *Slope*-Variablen

4.7 *NumIslands*

NumIslands gibt die Anzahl der Inseln an, die bei dem Schauer entstanden sind. In dieser Variable wurde bereits ein Quality Cut durchgeführt, sodass jetzt nur noch Werte kleiner als drei im Datensatz vorkommen. Es sind also nur noch Beobachtungen vorhanden, die neben der Hauptinsel eine einzige Nebeninsel gebildet haben. Dies bedeutet auch, dass die in Abschnitt 4.5 betrachtete Variable *SizeSubIslands* nur die *Size* eben dieser einen Nebeninsel enthält. In Tabelle 7 ist eine Vierfeldertafel zu sehen, die Auskunft über die Häufigkeiten gibt, mit der die Werte 1 und 2 vorkommen. Offenbar ist es so, dass Hadronen deutlich öfter zwei Inseln realisieren als Gammas.

	NumIslands	
	Hadronen	Gammas
Eine Insel	1.092.852	1.078.511
Zwei Inseln	210.400	43.550

Tabelle 7: Lage- und Streuungsmaße der Variable *NumIslands*

Diese Vermutung wurde auch schon in Abschnitt 4.5 aufgestellt, wo zu sehen war, dass Hadronen scheinbar weniger oft eine Null in *SizeSubIslands* realisieren. Realisiert ein Schauer mehr als eine Insel, ist dies also offenbar ein Zeichen dafür, dass es sich eher um ein Hadron als um ein Gamma handelt.

4.8 *Leakage*

Die *Leakage*-Variablen beschreiben, wie weit die angepasste Ellipse über den Rand der Kameraebene hinaus ragt. Es wird dabei zwischen den Variablen *Leakage1* und *Leakage2* unterschieden. *Leakage1* ist der Anteil der Intensität in den Pixeln, die am äußersten Rand der Kamera liegen an der Gesamtintensität. *Leakage2* bezieht zusätzlich zu den Pixeln am Rand noch deren Nachbarn mit ein. *Leakage2* ist dementsprechend immer größer oder gleich *Leakage1*.
An den Lagemaßen in Tabelle 8 sehen wir, dass sich Gammas und Hadronen hier jeweils kaum unterscheiden. Es sind in beiden Variablen, sowohl bei Gammas, als auch bei Hadronen, mehr als 75% der Werte gleich Null. *Leakage1* unterscheidet sich bei zwei Nachkommastellen nur bei der Schiefe. Bei *Leakage2* sind zusätzlich noch das Maximum und die Standardabweichung verschieden.

Problematisch auf Leakage könnte sich der Wobble-Modus auswirken. Aufgrund der Verschiebung der Quellposition zum Rand hin, ist es wahrscheinlich, dass mehr Gammas einen großen Leakage-Wert haben, als wenn die Quellposition im Kameramittelpunkt wäre. Dadurch könnte ein größerer Unterschied zwischen Gammas und Hadronen vorhanden sein.
So, wie die Lage- und Streuungsmaße der beiden Variablen aber hier aussehen, ist der Unterschied zwischen ihnen so klein, dass vermutlich keine von beiden einen sinnvollen Beitrag zu einer Klassifizierung beitragen würde.

	Leakage 1		*Leakage 2*	
	Hadronen	Gammas	Hadronen	Gammas
Minimum	0,00	0,00	0,00	0,00
1. Quartil	0,00	0,00	0,00	0,00
Median	0,00	0,00	0,00	0,00
Arithm. Mittel	0,01	0,01	0,02	0,02
3. Quartil	0,00	0,00	0,00	0,00
Maximum	0,30	0,30	0,68	0,63
Standardabw.	0,04	0,04	0,05	0,07
Schiefe	4,99	3,73	4,07	3,58

Tabelle 8: Lage- und Streuungsmaße der beiden *Leakage*-Variablen

4.9 *Dist0*

Die Variable *Dist0* beschreibt den Abstand des COG zum Mittelpunkt der Kamera. Es gilt also

$$Dist0 = \sqrt{MeanX^2 + MeanY^2}.$$

Dabei ist zu beachten, dass der Mittelpunkt der Kamera nicht der Quellposition der Gammas entspricht. Den Absand des COG von der Gamma-Quellposition wird mit *Dist* bezeichnet. Diese Variable wird in dieser Arbeit jedoch nicht betrachtet.
In Tabelle 9 ist zu sehen, dass sich die Lage- und Streuungsmaße von *Dist0* bei Gammas und Hadronen kaum unterscheiden.
Hervorzuheben ist, dass die Schiefe jeweils nah an Null ist und dass die Werte generell sehr symmetrisch erscheinen. Dies werden wir anhand von Histogrammen weiter untersuchen. Diese sind in Abbildung 21 dargestellt.

	Dist0	
	Hadronen	Gammas
Minimum	0,49	0,29
1. Quartil	143,20	133,10
Median	201,00	199,00
Arithm. Mittel	196,60	200,40
3. Quartil	248,60	257,70
Maximum	462,50	457,10
Standardabw.	76,15	89,06
Schiefe	0,00	0,25

Tabelle 9: Lage- und Streuungsmaße der Variable *Dist0*

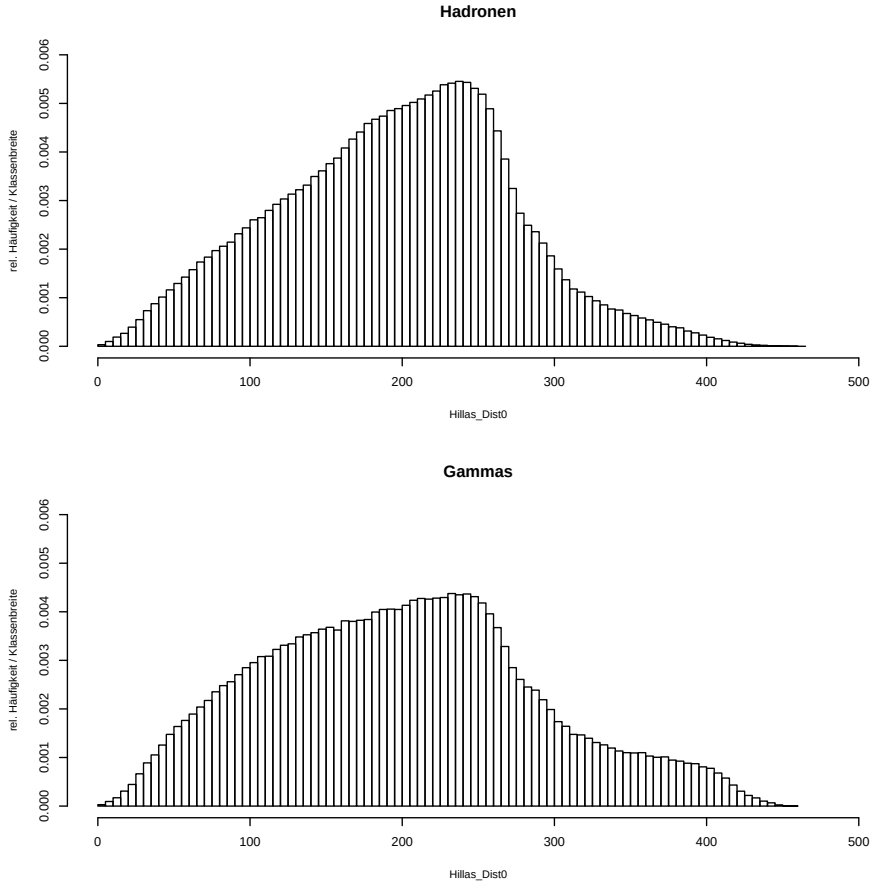

Abbildung 21: Histogramme der Variable *Dist0*

Alles in Allem erinnern die Verteilungen stark an die der Variable *Ellip* aus Abbildung 14.
Die Verteilungen sind offenbar nicht so symmetrisch, wie es den Anschein hatte. Beide Verteilungen flachen zu größeren Werten hin sehr stark ab. Dies liegt vermutlich an der sechseckigen Form der Kamera. Für kleinere Werte als den jeweiligen Modalwert erscheint der Anstieg bei Hadronen linear, die Gammas haben hier eine gewölbtere Form.

4.10 *Conc* / *Con1* / *ConcCOG* / *ConcCore*

Conc ist ein Maß für die Konzentration des Schauers. Es ist das Verhältnis der Intensitäten der beiden hellsten Pixel zur Gesamtintensität. *ConCore* ist entsprechend die Konzentration in den CorePixels, also das Verhältnis der Intensität der CorePixel zur Gesamtintensität. *ConcCOG* ist die Konzentration in dem Pixel, in dem das COG liegt. *Conc1* ist die Konzentration der Intensität im hellsten Pixel.
Einige Lage- und Streuungsmaße der vier Variablen sind in Tabelle 10 dargestellt. Auffällig ist hier, dass fast alle Werte der Lagemaße bei den Hadronen kleiner sind als bei den Gammas. Der Unterschied ist jedoch nicht sehr groß. Der bedeutendste Unterschied zwischen Gammas und Hadronen liegt in allen vier Variablen in der Schiefe. Es ist so, dass Hadronendaten rechtsschiefer sind als Gammadaten. Die Verteilungen der Variablen *Conc* und *ConcCore* in den Gammas sind sogar leicht linksschief.

	Conc		*Conc1*		*ConcCOG*		*ConcCore*	
	Hadronen	Gammas	Hadronen	Gammas	Hadronen	Gammas	Hadronen	Gammas
Minimum	0,007	0,011	0,003	0,006	0,000	0,000	0,000	0,000
1. Quartil	0,159	0,204	0,086	0,110	0,106	0,120	0,278	0,299
Median	0,272	0,298	0,149	0,164	0,262	0,329	0,344	0,372
Arithm. Mittel	0,276	0,296	0,155	0,166	0,289	0,336	0,346	0,367
3. Quartil	0,386	0,389	0,216	0,218	0,451	0,465	0,412	0,438
Maximum	0,713	0,708	0,485	0,490	0,823	0,895	0,825	0,806
Standardabw.	0,141	0,125	0,084	0,076	0,201	0,169	0,102	0,102
Schiefe	0,173	-0,015	0,409	0,245	0,378	0,195	0,108	-0,121

Tabelle 10: Lage- und Streuungsmaße der vier *Conc*-Variablen

Die Histogramme in Abbildung 22 geben weiteren Aufschluss über die Daten. Der deutlichste Unterschied in der Form der Histogramme ist bei *ConcCOG* zu sehen.

Hier ist der Modalwert der Gammas im Vergleich zu den Hadronen deutlich zu größeren Werten hin verschoben. Auch sonst unterscheiden sich die Histogramme stark. Während das Histogramm der Gammas recht symmetrisch erscheint (bis auf die Spitze bei kleinen Werten), fällt die Häufigkeit bei den Hadronen zu größeren Werten hin ab.
Am auffälligsten bei den ersten drei Variablen ist die Spitze, die bei einem Bereich kleiner Werte aus dem Histogramm hervorragt. Eine Erklärung für diese Spitze gibt es jedoch nicht. Hier wäre ein Ansatz für eine genauere Untersuchung gegeben.
Die Histogramme von *ConcCore* ähneln einander in der Form sehr stark. Die Hadronen sind im Vergleich zu den Gammas jedoch leicht zu kleineren Werten hin verschoben und im Histogramm der Gammas sieht man eine leichte Schiefe. Auffällig ist hier auch die in beiden Histogrammen vergleichsweise stark besetzte Klasse nahe der Null.
Insgesamt scheinen in allen Variablen recht deutliche Unterschiede zwischen Gammas und Hadronen zu bestehen. Jedoch sind auch hier einige Informationen redundant.

4.11 *NumSinglePixels* / *SizeSinglePixels*

NumSinglePixels und *SizeSinglePixels* beinhalten Informationen zu isolierten Pixeln, also solchen Pixeln, die weder zur Hauptinsel, noch zu einer Nebeninsel gehören. *NumSinglePixels* gibt dabei die Anzahl der vorkommenden isolierten Pixel an. *SizeSinglePixels* ist die *Size*, also die Gesamtintensität aller isolierter Pixel.

Wie auch in Tabelle 11 zu sehen ist, sind die Werte der Variable *NumSinglePixels* diskret.
Die Unterschiede zwischen Hadronen und Gammas sind in beiden Variablen gravierend. Bis auf das Maximum, das bei Gammas jeweils deutlich unter dem der Hadronen liegt, liegen alle Lagemaße bei den Gamas sehr deutlich über denen der Hadronen. Besonders deutlich wird dies beim ersten Quartil von *SizeSinglePixels*. Dieses ist bei den Gammas sogar größer als das dritte Quartil bei den Hadronen. Auch die Schiefe ist zwischen Gammas und Hadronen sehr unterschiedlich. Lediglich die Standardabweichung ist bei beiden ähnlich.

Einen besseren Einblick in die Variablen bietet Abbildung 23, wo eine Häufigkeits-

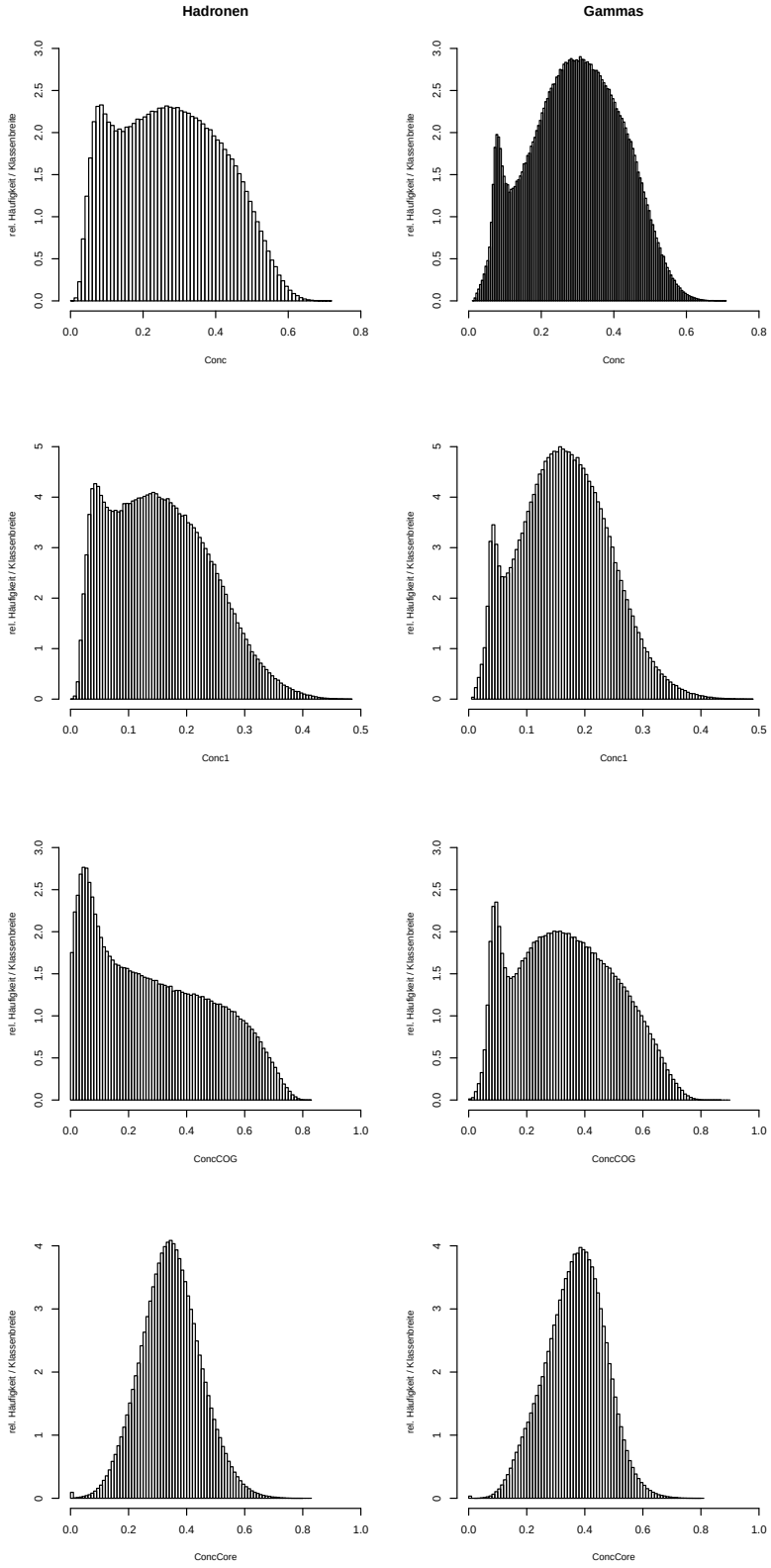

Abbildung 22: Histogramme der *Conc*- Variablen

	NumSinglePixels		*SizeSinglePixels*	
	Hadronen	Gammas	Hadronen	Gammas
Minimum	0	0	0,00	0,00
1. Quartil	1	5	6,34	43,06
Median	2	7	14,09	57,41
Arithm. Mittel	-	-	20,79	58,92
3. Quartil	3	9	29,61	73,05
Maximum	33	22	497,80	257,50
Standardabw.	1,61	2,51	22,32	22,30
Schiefe	1,46	0,32	2,03	0,43

Tabelle 11: Lage- und Streuungsmaße der beiden Variablen *NumSinglePixels* und *SizeSinglePixels*

verteilung von *NumSinglePixels*, bzw. ein Histogramm von *SizeSinglePixels* zu sehen ist.

Wie zu erwarten war, haben die Verteilungen von *NumSinglePixels* und *SizeSinglePixels* jeweils etwa die gleiche Form. Wenige isolierte Pixel führen auch zu einer kleinen Size der isolierten Pixel.
Dafür unterscheiden sich die Daten zwischen Hadronen und Gammas umso stärker. Der Eindruck aus Tabelle 11 bestätigt sich. Während sich beide Variablen bei den Hadronen bei Null häufen und große Werte selten angenommen werden, sehen die Verteilungen der Gamma-Daten fast schon symmetrisch um ihren Modalwert aus, der dabei deutlich von Null verschieden ist.
Durch die großen Unterschiede zwischen Gammas und Hadronen wären diese Variablen sehr gut geeignet für eine Gamma-Hadron-Separation. Dass solch große Unterschiede in den Daten vorhanden sind, erscheint aber ungewöhnlich.
Zwar ist es möglich, dass diese Werte tatsächlich so sind, wie sie sind, intuitiv würde man hier aber zunächst von einem Fehler - zum Beispiel in der Simulation - ausgehen. Zumindest bedarf dieser Sachverhalt weiterer investigativer Eingriffe.
An dieser Stelle entfernen wir beide Variablen aus den Datensätzen, da nicht mit Gewissheit ein Fehler in der Simulation ausgeschlossen werden kann, der die Ergebnisse dieser Arbeit verfälschen würde.

4.12 *M3Long* / *M3Trans*

M3 ist das dritte Moment der Verteilung der Intensität entlang der Halbachsen. Die Intensitäten werden also auf die Halbachse projiziert und so als eindimensiona-

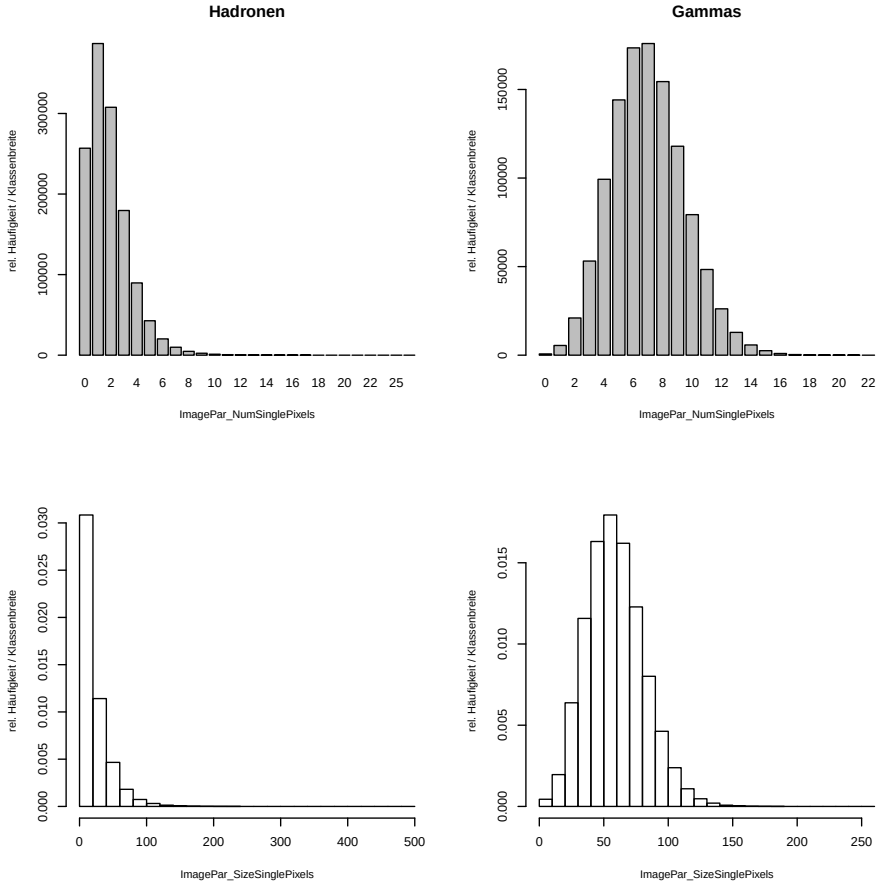

Abbildung 23: Histogramme der Variablen *NumSinglePixels* und *SizeSinglePixels*

le Verteilung entlang dieser gesehen. *M3Long* ist das dritte Moment entlang der größeren Halbachse, *M3Trans* entsprechend das dritte Moment entlang der kleineren.
Ein negativer Wert bei *M3Long* bedeutet dabei, dass das COG weiter von der Quellposition entfernt ist als der Mittelpunkt der Ellipse. Ein positiver entsprechend das Gegenteil.
Teilte man M3 noch durch die dritte Potenz der Standardabweichung, so ergäbe sich die Schiefe. Auch so ist ein stark von Null abweichender Wert des dritten Moments aber schon ein Indikator für eine große Schiefe.

Aufgrund der zu typischen Form der Gamma-Schauer ist auch zu erwarten, dass Gammas eher negative Werte aufweisen, während bei Hadronen negative und positive Werte gleich oft vorkommen sollten. Tabelle 12 zeigt, dass genau dies der Fall ist. Unter Berücksichtigung der Standardabweichung ist das arithmetische Mittel jedoch nur wenig zum Negativen hin verschoben. Die Abweichung von Null ist dennoch hoch signifikant. Ein Wilcoxon-Test ergibt einen p-Wert kleiner $2,2 \cdot 10^{-16}$.
Eher überraschend sind die Werte für *M3Trans*. Hier wäre zu erwarten gewesen, dass sowohl Gammas, als auch Hadronen symmetrisch verteilt sind. Stattdessen zeigt sich bei den Gammas das gleiche wie bei *M3Long*: Ein arithmetisches Mittel kleiner Null und leichte Linksschiefe. Diese Linksschiefe ist ohnehin überraschend, da bei einem arithmetischen Mittel kleiner Null eigentlich eine Rechtsschiefe zu erwarten wäre. Auch das 1. und 3. Quartil der Gammas bei *M3Long* lassen eher auf eine Rechtsschiefe schließen.
Es fällt auf, dass Hadronen jeweils eine größere Standardabweichung haben.

Die Histogramme in Abbildung 24 geben mehr Aufschluss über die Daten.

	M3Long		*M3Trans*	
	Hadronen	Gammas	Hadronen	Gammas
Minimum	-347,20	-254,40	-318,10	-216,60
1. Quartil	-22,76	-23,23	-15,91	-14,62
Median	4,60	-9,39	2,46	-4,44
Arithm. Mittel	0,44	-3,89	0,25	-1,21
3. Quartil	23,33	18,69	16,18	13,63
Maximum	314,60	255,70	272,90	224,80
Standardabw.	48,31	33,47	33,86	21,66
Schiefe	0,00	-0,25	0,03	-0,34

Tabelle 12: Lage- und Streuungsmaße der beiden M3-Variablen

Auffällig ist hier, dass alle Histogramme zwei Maxima in einem kleinen Abstand von Null aufweisen und dass *M3Long* offenbar öfter Werte kleiner Null als größer Null aufweist, während die anderen drei Histogramme recht symmetrisch aussehen.

Außerdem ist hier zu sehen, dass sich Gammas und Hadronen tatsächlich eher in *M3Long* unterscheiden. Die Widersprüche in den Lagemaßen könnten auf die hohe Standardabweichung zurückzuführen sein, zusammen mit dem seltenen Vorkommen hoher Werte, die zum Beispiel das arithmetische Mittel stark beeinflussen. Ein Unterschied zwischen Gammas und Hadronen ist vor allem im mittleren Teil von *M3Long* zu sehen.

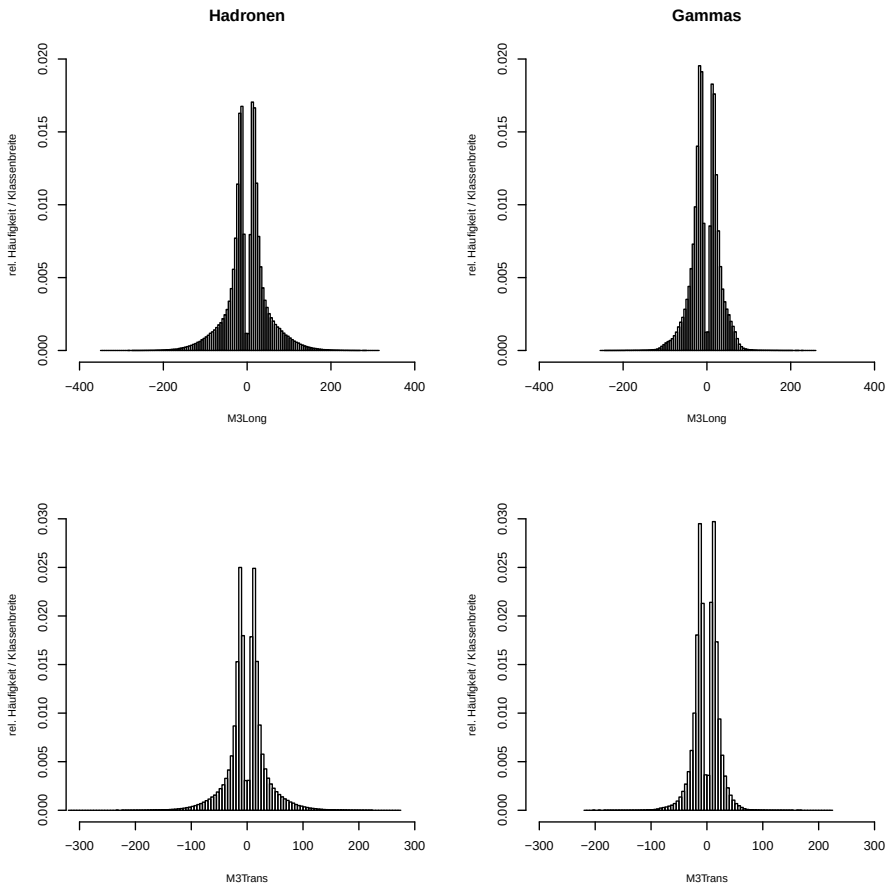

Abbildung 24: Histogramme der *M3*-Variablen

4.13 *Borderline*

Borderline gibt die Länge des Randes der Hauptinsel an. Es ist zu erwarten, dass Hadroneninseln „ausgefranster“ aussehen als Gammainseln und deshalb einen ausgedehnteren Rand haben, d.h. einen größeren Wert bei *Borderline*.

Lage- und Streuungsmaße der Variable sind in Abbildung 13 zu sehen.

Tatsächlich scheinen die Hadronen hier größere Werte zu realisieren. Das arithmetische Mittel ist bei Hadronen deutlich größer als bei den Gammas. Ein Wilcoxon-Test für zwei Stichproben ergibt einen p-Wert kleiner $2,2\ cdot 10^{-16}$ und bestätigt damit, dass der Unterschied signifikant ist.
Auffällig ist auch die deutlich größere Standardabweichung der Hadronen. Gleichzeitig sind aber Minimum und Maximum der Daten jeweils etwa gleich. Hadronen und Gammas streuen in der Variable *Borderline* also über den gleichen Bereich, wobei Gammas aber stärker konzentriert sind. Dies ist auch gut in den Histogrammen in Abbilung 25 zu sehen.

	BorderLine	
	Hadronen	Gammas
Minimum	91,98	92,50
1. Quartil	144,30	141,90
Median	205,00	184,70
Arithm. Mittel	281,10	219,90
3. Quartil	360,80	265,60
Maximum	1651,00	1571,00
Standardabw.	188,96	105,86
Schiefe	1,57	1,50

Tabelle 13: Lage- und Streuungsmaße der Variable *Borderline*

In beiden Histogrammen scheinen sich die Beobachtungen bei ihrem Minimum zu häufen. Die Häufigkeit nimmt zu großen Werten hin dann aber ab. Wobei bei Hadronen deutlich mehr große Beobachtungen vorliegen als bei Gammas. Hierdurch ist wohl auch der große Unterschied zwischen den Standardabweichungen zu erklären.
Sehr große Werte scheinen sowohl bei Gammas, als auch bei Hadronen nur sehr selten aufzutreten. Bei Hadronen liegen 99% der Daten unter 900, obwohl das Maximum bei 1.651 liegt. Bei Gammas sind 99% sogar schon bei einem Wert von 550 erreicht, bei einem Maximum von 1.571.

Ein ähnliches Verhalten zeigte sich auch in der Variable *Area*. Dies ist ein Zeichen dafür, dass große Ereignisse deutlich seltener vorkommen als kleine.

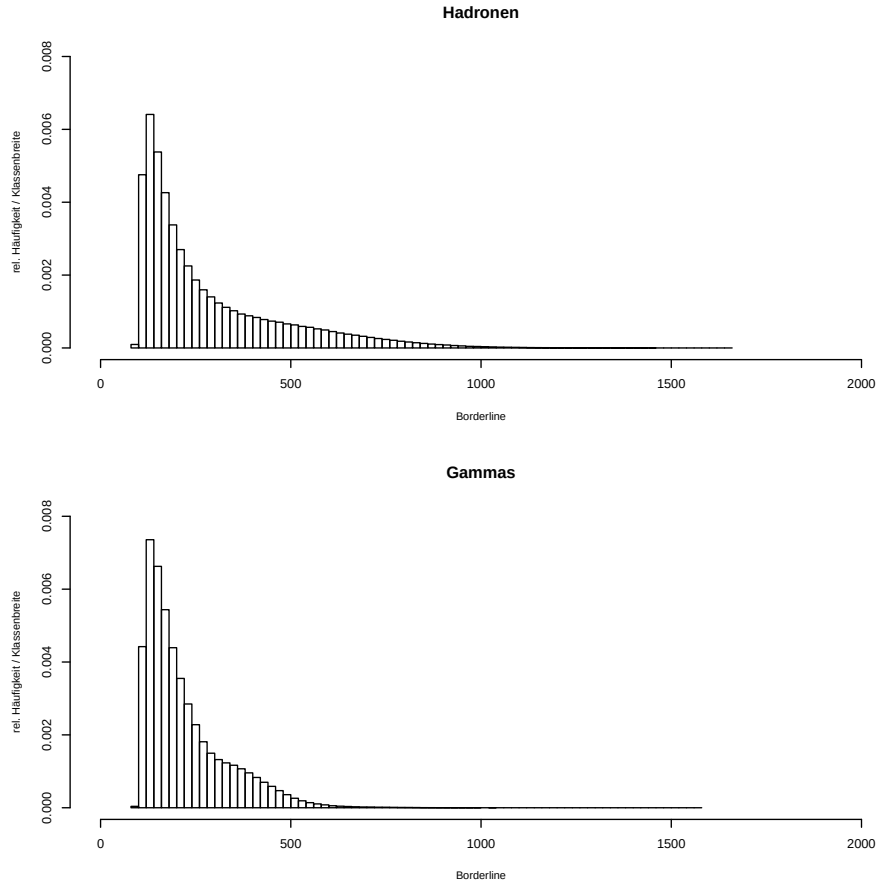

Abbildung 25: Histogramme der Variable *Borderline*

4.14 *UsedArea* / *NumUsedPixels*

NumUsedPixels und *UsedArea* geben die Anzahl der letztlich zur Ellipsenanpassung genutzten Pixel und deren Gesamtfläche an. *UsedArea* ist hierbei nicht zu verwechseln mit *Area*. *Area* gibt die Fläche der angepassten Ellipse an, während *UsedArea* die Summe der Flächen aller genutzer Pixel angibt. Da die Fläche der genutzten Pixel mit der Anzahl der genutzten Pixel direkt zusammenhängt, ist klar, dass diese Variablen einen sehr hohen Korrelationskoeffizienten untereinander haben. Dieser liegt hier bei 0,96. Tatsächlich wäre der Zusammenhang exakt linear,

wenn alle Pixel des Teleskops die gleiche Fläche hätten.
Die Lage- Und Streuungsmaße in Tabelle 14 weisen darauf hin, dass sich Gammas und Hadronen in diesen Variablen nicht sehr stark unterscheiden. Das Minimum und das erste Quartil stimmen jeweils exakt überein, Median, drittes Quartil und Maximum unterscheiden sich nur unwesentlich. Auffällig ist dabei, dass der Median und das dritte Quartil in den Hadronen größer sind, das Maximum aber kleiner ist als in den Gammas.
Durch das größere Maximum ist auch die Standardabweichung bei den Gammas etwas größer als bei den Hadronen.
Die Schiefe ist bei allen Variablen sehr groß, was auch gut in den Histogrammen in Abbildung 26 zu sehen ist. Hier sind zwar kleinere Unterschiede zu erkennen, diese sind aber so klein, dass die Histogramme fast nicht zu unterscheiden sind. Die Variablen *NumUsedPixels* und *UsedArea* scheinen sich also nicht sehr gut zur Trennung von Hadronen und Gammas zu eignen.

	UsedArea		*NumUsedPixels*	
	Hadronen	Gammas	Hadronen	Gammas
Minimum	4677,00	4677,00	6	6
1. Quartil	7015,00	7015,00	9	9
Median	11690,00	10130,00	14	13
Arithm. Mittel	22390,00	22710,00	-	-
3. Quartil	21820,00	20260,00	24	21
Maximum	692100,00	716300,00	494	514
Standardabw.	35358,28	39142,26	28,73	29,24
Schiefe	5,86	6,06	5,41	6,75

Tabelle 14: Lage- und Streuungsmaße der beiden Used-Variablen

4.15 *CoreArea* / *NumCorePixels*

Core-Pixel sind - wie in Abschnitt 2.2.2 beschrieben - solche Pixel, die eine Schranke von 8,5 Photoelektronen übersteigen. Für jedes Event gibt es eine Anzahl an Core Pixeln, für die auch eine Fläche berechnet werden kann. Die Lage- und Streuungsmaße der entsprechenden Variablen *CoreArea* und *NumCorePixels* sind in Tabelle 15 dargestellt.
Das Minimum von *CoreArea* ist bei Hadronen und Gammas gleich. Es entspricht der Fläche genau eines Pixels. Auch sonst sind die Lagemaße sehr gleich. Das erste Quartil stimmt in beiden Variablen bei Hadronen und Gammas überein. Bei *CoreArea* auch das dritte. Den einzigen wesentlichen Unterschied gibt es beim

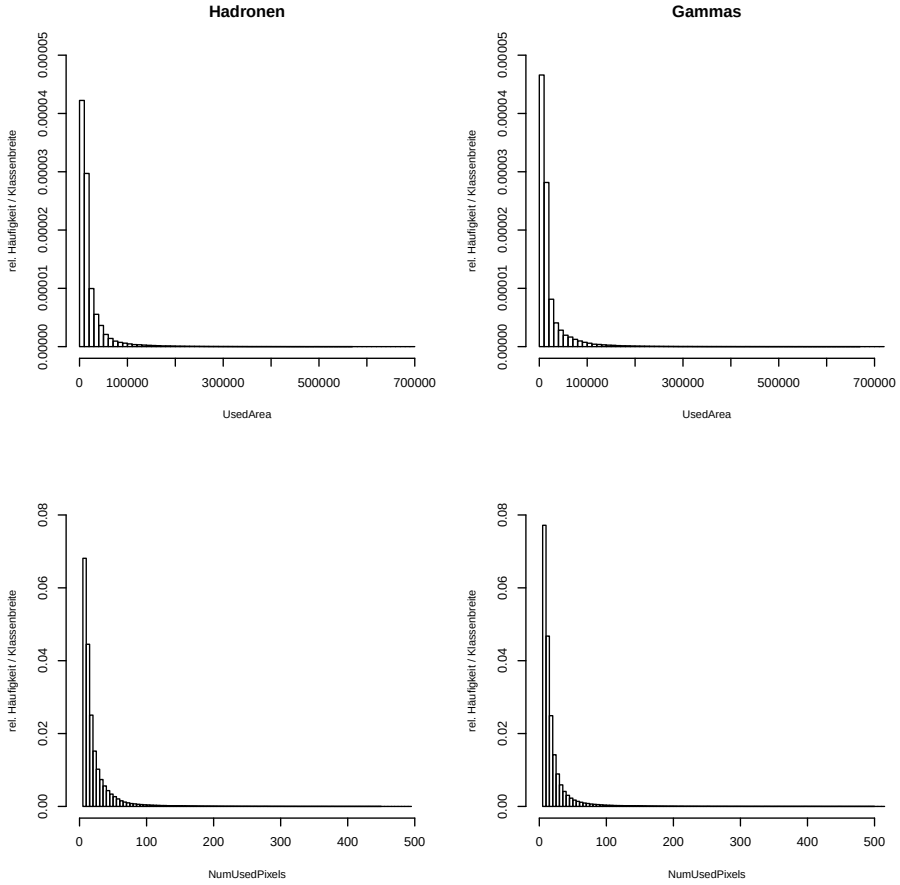

Abbildung 26: Histogramme der Variablen *UsedArea* und *NumUsedPixels*

Maximum. Hier liegen die Gammas recht stark über den Hadronen. Die Standardabweichung ist dementsprechend ebenfalls bei den Gammas höher.

	CoreArea		*NumCorePixels*	
	Hadronen	Gammas	Hadronen	Gammas
Minimum	779,40	779,40	1	1
1. Quartil	4677,00	4677,00	6	6
Median	7794,00	7015,00	10	9
Arithm. Mittel	16210,00	17800,00	-	-
3. Quartil	14810,00	14810,00	17	16
Maximum	641500,00	664100,00	466	492
Standardabw.	29186,56	34738,26	23,96	26,53
Schiefe	6,74	6,33	6,33	7,05

Tabelle 15: Lage- und Streuungsmaße der beiden Core-Variablen

Die sehr hohe Schiefe ist gut in den Histogrammen in Abbildung 27 zu sehen. Hier ist mit bloßem Auge auch fast kein Unterschied zwischen Hadronen und Gammas auszumachen.

4.16 *Asym*

Asym ist, wie der Name angibt, neben den beiden *M3*-Variablen eine weitere Variable, die Information über die Asymmetrie des Schauers enthält. Es ist die Differenz von *Dist* und dem Abstand des Pixels mit höchster Intensität von der Quellposition. Dieser ist ein Repräsentant für den Schwerpunkt des Schauers. Ist der hellste Pixel weiter weg von der Quellposition als der Ellipsenmittelpunkt, so ist *Asym* negativ. Ist er näher, ist *Asym* positiv.
Wie bei *M3Long* ist auch hier in den Lagemaßen (Tabelle 16) zu sehen, dass Gammas kleinere Werte realisieren als Hadronen. Der Unterschied ist vor allem im 3. Quartil zu sehen.
Zusätzlich und anders als bei *M3Long* unterscheiden sich hier auch die Standardabweichung, die Schiefe und die Wölbung recht stark.
Die Standardabweichung ist bei den Gammas deutlich geringer als bei den Hadronen. Dies führt zu einer stärkeren Konzentration der Daten und damit zu der geringeren Wölbung.
Der Unterschied in der Schiefe ist nicht sehr bedeutend. Es ist jedoch anzumerken, dass Hadronen fast perfekt symmetrisch zu sein scheinen, während Gammas

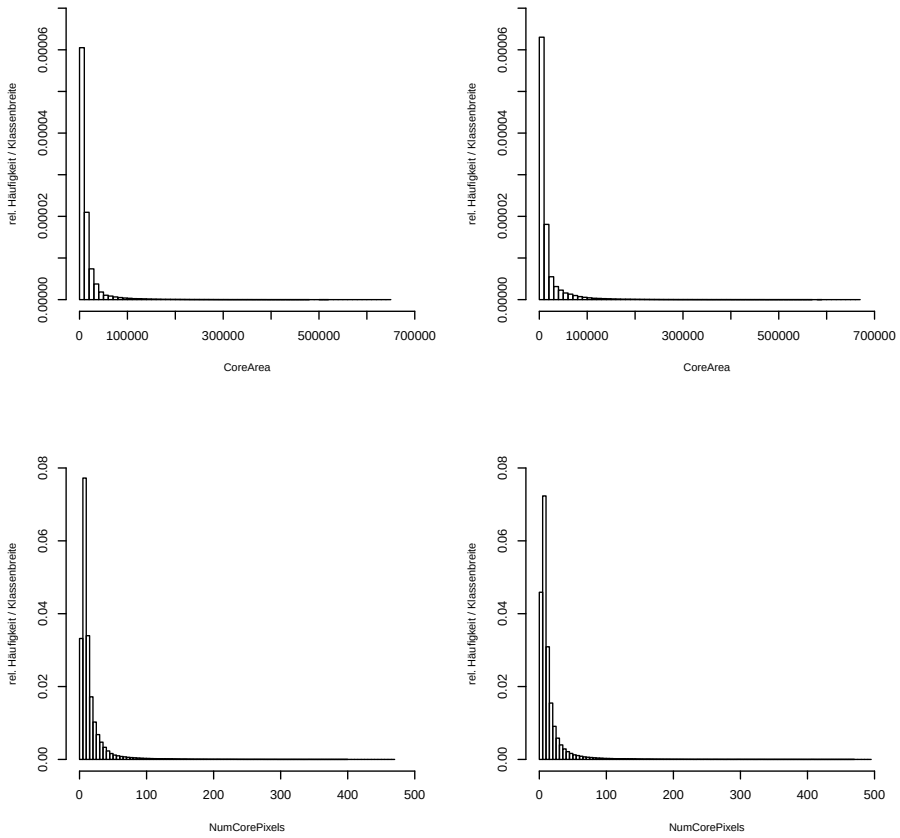

Abbildung 27: Histogramme der Variablen *CoreArea* und *NumCorePixels*

leicht rechtsschief sind.

	Asym	
	Hadronen	Gammas
Minimum	-563,20	-472.70
1. Quartil	-24,11	-23.99
Median	0,73	-3.20
Arithm. Mittel	0,57	-4.70
3. Quartil	25,10	16.28
Maximum	515,80	423.80
Standardabw.	58,37	37.31
Schiefe	0,02	-0.25
Wölbung	4.85	2.05

Tabelle 16: Lage- und Streuungsmaße der Variable *Asym*

Die Histogramme in Abbildung 28 liefern nicht viele neue Erkenntnisse. Gut zu sehen ist die geringere Standardabweichung der Gammas und dass mehr Werte kleiner Null als größer Null vorliegen.
Zusätzlich ist noch zu sehen, dass stark von Null abweichende Werte sowohl bei Gammas als auch bei Hadronen sehr selten sind.
Insgesamt ist die Interpretation dieser Variable die gleiche wie die von *M3Long*. Gamma-Schauer scheinen öfter als Hadronen eine Asymmetrie aufzuweisen, die für eine Verschiebung des Schauerschwerpunkts weg von der Quellposition verantwortlich ist.

4.17 *NumSatPixelsHG* / *NumSatPixelsLG*

Beide Variablen sind sowohl bei Hadronen, als auch bei Gammas konstant Null und werden daher vor weiteren Analysen aus den Datensätzen entfernt.

4.18 Übersicht

In diesem Kapitel haben wir festgestellt, dass es Auffälligkeiten in den Daten gibt, die uns dazu veranlassten, Veränderungen an den Datensätzen vorzunehmen. Einige Beobachtungen waren scheinbar fehlerhaft und werden daher entfernt und bei der weiteren Auswertung nicht berücksichtigt. Der Gamma-Datensatz besteht nach deren Entfernung aus 1.122.040 Beobachtungen, der Hadronen-Datensatz

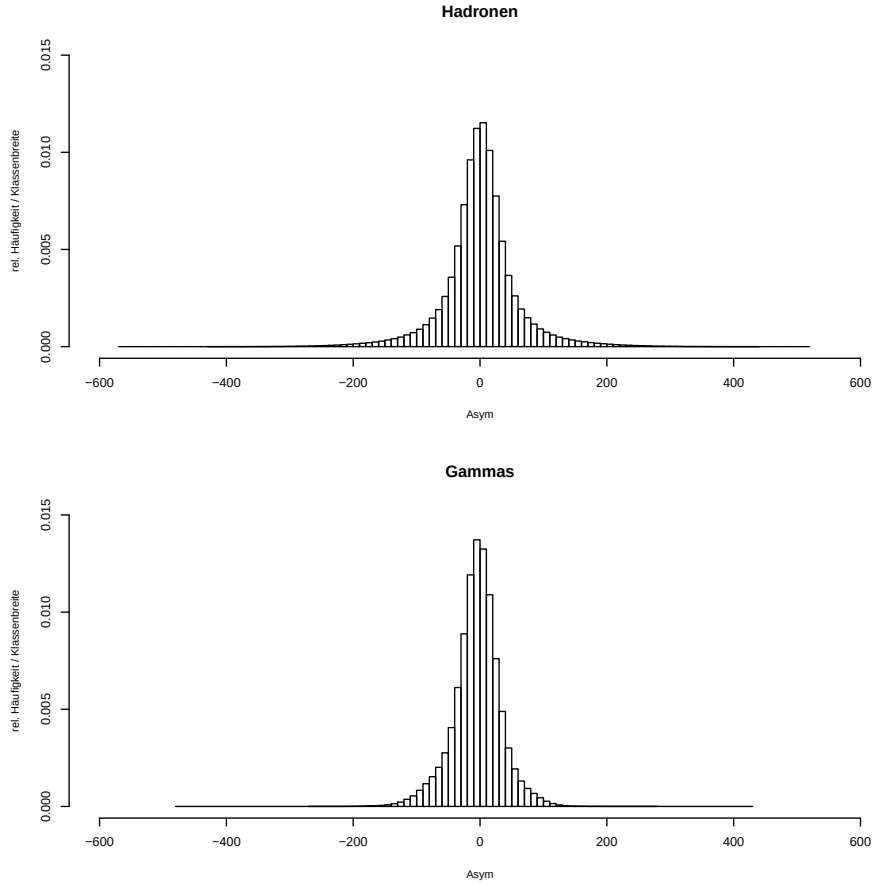

Abbildung 28: Histogramme der Variable *Asym*

aus 1.303.251. Außerdem wurden die Variablen *NumSinglePixels* und *SizeSinglePixels*, sowie *NumSatPixelsHG* und *NumSatPixelsLG* aus den Datensätzen entfernt. Dafür kam die in dieser Arbeit definierte Variable *Ellip* neu hinzu. Es werden also im Weiteren 30 Variablen betrachtet.

5 Vorverarbeitung

Wir führen nun Vorverarbeitungen durch, die die eigentliche Gamma-Hadron-Separation erleichtern sollen.

5.1 Transformationen

Symmetrische Verteilungen haben bei statistischen Auswertungen große Vorteile. Zunächst erfordern viele Verfahren eine solche Symmetrie, um überhaupt verwendet werden zu dürfen. Des Weiteren haben wir im vorherigen Abschnitt gesehen, dass stark schiefe Verteilungen nur schlecht mit bloßem Auge interpretierbar sind. Außerdem profitieren einige Verfahren, wie zum Beispiel die im Folgenden genutzte Hauptkomponentenanalyse, von symmetrischen Verteilungen, obwohl die Symmetrie nicht unbedingt vorausgesetzt werden muss. Tatsächlich ist die Hauptkomponentenanalyse bei nicht elliptischen Verteilungen nicht unbedingt sinnvoll.
Aus diesen Gründen werden wir in diesem Abschnitt durch Transformationen die Schiefe in sämtlichen Variablen, bei denen dies möglich ist, beseitigen. Dies geschieht mithilfe der in Abschnitt 3.1 eingeführten Box-Cox-Transformation.
Die Parameter c und p dieser Transformation werden für jede Variable individuell so bestimmt, dass die Schiefe nach der Transformation (annähernd) Null ist.
Der Parameter c dient dabei lediglich dazu, die Daten in den positiven Bereich zu verschieben. Er ist nur dann wichtig, wenn die Variable Werte kleiner oder gleich Null annehmen kann. Er wird für eine Variable x bestimmt als:

$$c = -\min\left(\min x, 0\right) + 1$$

Falls das Minimum von x also negativ oder Null ist, wird c so gewählt, dass das neue Minimum nach der Verschiebung um c bei 1 liegt. Falls das Minimum von x positiv ist, wird lediglich die 1 zu allen Werten addiert. Werte, die sehr nah bei Null liegen, können nach einer Box-Cox-Transformation ungewollt groß werden. Dies wird dadurch umgangen.
Der Parameter p wird nach der Festlegung von c iterativ optimiert. Die Zielgröße ist dabei der Absolutbetrag der Schiefe. Diese Zielgröße soll minimiert werden. Das Verfahren, mit dem optimiert wird, ist eine Kombination aus der Suche mithilfe des goldenen Schnitts und einer sukzessiven parabolischen Interpolation. Die

Optimierung wird mithilfe der Funktion `otimize` des R-Basispakets vorgenommen (R Development Core Team, 2009).

Die letztlich gewählten Parameter-Einstellungen für jede Variable sind in Tabelle 17 aufgeführt.

Die Variablen *NumIslands*, *NumUsedPixels* und *NumCorePixels*, sowie *SizeSubIslands* und *Leakage* werden nicht transformiert. *NumIslands*, *NumUsedPixels* und *NumCorePixels* sind ordinal skaliert, weshalb eine Box-Cox-Transformation nicht sinnvoll ist. *SizeSubIslands* und die beiden *Leakage*-Variablen beinhalten sehr viele Nullen, sodass ein sehr großer Wert für p nötig ist, um die Schiefe zu beseitigen. Diese Variablen werden so belassen, wie sie sind.

5.2 Gamma-Hadron-Verhältnis

Ein großes Problem bei der Gamma-Hadron-Separation ist das in der Realität vorkommende Verhältnis von Gamma- zu Hadronenschauern. Dieses liegt bei etwa 1:1000. Das bedeutet, dass auf jedes Gamma, das vom Teleskop registriert wird, 1000 Hadronen registriert werden.
Dies ist für die Gamma-Hadron-Separation deshalb problematisch, weil dementsprechend auch dort mit diesem realen Verhältnis gearbeitet werden sollte. Dadurch müssen aber unter Umständen die Trainings- und Testdatensätze sehr groß gewählt werden. Möchte man beispielsweise einen Testdatensatz, der nur 1.000 Gammas beinhalten soll, so muss man einen Datensatz mit 1 Mio Hadronen-Beobachtungen wählen. Bei entsprechend mehr Gammas kommt man schnell in Datensatzgrößen, die bei einer Auswertung kaum zu bewältigen sind. Und auch auf großen Rechenclustern, die solche Datenmengen bewältigen können, ist die Rechenzeit unter Umständen immens.
Ein Anliegen ist es also, alternative Lösungsansätze für dieses Problem zu finden.

5.2.1 Testdaten

Für Testdatensätze ist eine Problemlösung recht einfach. Hier ist davon auszugehen, dass die Fehlklassifikationsrate von Gammas und Hadronen innerhalb eines Random Forests konstant bleibt. Das heißt, passt man einen Random Forest an

Variable	c	p
Length	1	-0,65
Width	1	-1,20
Area	1	-0,56
Ellip	1	-0,54
MeanX	457,68	1,19
MeanY	430,43	0,94
Delta	2,57	0,96
SinDelta	2	0,92
CosDelta	1	3,58
Size	1	-0,60
SizeMainIsland	1	-0,61
SizeSubIslands	-	-
SlopeTrans	1,23	0,56
SlopeLong	1,23	0,74
NumIslands	-	-
Leakage1	-	-
Leakage2	-	-
Dist0	1	0,86
Conc	1	0,57
Conc1	1	-1,33
ConcCOG	1	-0,24
ConcCore	1	0,99
M3Trans	319,07	1,01
M3Long	348,25	1,01
Borderline	1	-0,92
UsedArea	1	-0,70
NumUsedPixels	-	-
CoreArea	1	-0,49
NumCorePixels	-	-
Asym	564,23	0,97

Tabelle 17: Parameter der Box-Cox-Transformation

und lässt zwei Testdatensätze von diesem Random Forest klassifizieren, wobei einer der Datensätze ein Gamma-Hadron-Verhältnis von 1:1 und der andere eines von 1:1000 hat, so wird der Anteil falsch klassifizierter Gammas an der Gesamtzahl der Gammas innerhalb der Testdatensätze gleich sein. Selbiges gilt für Hadronen. Dies ist in Tabelle 18 verdeutlicht. Dort wurde ein Random Forest angepasst und zwei verschiedene Testdatensätze damit klassifiziert. Der eine Datensatz beinhaltete 5000 Gammas und 5000 Hadronen. Der andere 500 Gammas und 50000 Hadronen. Wie zu sehen ist, sind die Fehler innerhalb der Klassen fast identisch.

		klassifiziert		
		g	h	class error
real	g	4229	771	0,1542
	h	1012	3988	0,2024

		klassifiziert		
		g	h	class error
real	g	424	76	0,1520
	h	10283	39717	0,2057

Tabelle 18: Vierfeldertafeln für eine Gamma-Hadron-Separation. Oben mit einem Verhältnis von 1:1 , unten mit einem Verhältnis von 1:100. In den Spalten stehen die echten Klassenzugehörigkeiten und in den Spalten, als was sie klassifiziert wurden. Außerdem angegeben ist die Fehlklassifikationsrate innerhalb der echten Klassen.

Das bedeutet, dass wir, um eine realistische Reinheit zu berechnen, einen Testdatensatz mit einem Gamma-Hadron-Verhältnis von 1:1 nutzen und die Anzahl falsch klassifizierter Hadronen (die wir zur Berechnung der Reinheit benötigen) mit 1000 multiplizieren können. Die Reinheit ist dann etwa so groß wie sie bei einem realistischen Verhältnis gewesen wäre.

5.2.2 Trainingsdaten

Bei Trainingsdatensätzen gestaltet sich die Problemlösung schwieriger. Zwar kann man hier zumindest bei Random Forests Gewichte eingeben, sodass ein beliebiges Verhältnis von Gammas zu Hadronen im Traingsdatensatz vorliegen darf, das Verhältnis, mit dem man trainiert, hat jedoch einen sehr starken Einfluss auf die Güte der Klassifikation und muss nicht unbedingt dem realen Verhälnis entsprechen.
In Tabelle 19 ist auch gut zu sehen, wie sich dieser Einfuss auswirkt. Es wurden

mit den aufgeführten Gamma-Hadron-Verhältnissen jeweils 10 Random Forests angepasst. Jeder wurde dazu genutzt, einen zufällig gezogenen Testdatensatz, in dem das Verhältnis immer gleich 1:10 war, zu klassifizieren. Es wurden jeweils Reinheit und Recall berechnet und über die 10 Random Forests gemittelt. Die Ergebnisse sind in Tabelle 19 zu sehen.

Verhältnis	Reinheit	Recall	geom.Mittel
1:1000	1	0,002	0,083
1:100	0,93	0,04	0,264
1:50	0,87	0,08	0,334
1:10	0,37	0,27	0,326
1:5	0,23	0,41	0,290
1:2	0,09	0,71	0,206
1:1	0,05	0,88	0,157

Tabelle 19: Reinheit und Recall für verschiedene Gamma-Hadron-Verhältnisse im Trainingsdatensatz. Der Testdatensatz hatte ein Verhältnis von 1:1 und wurde entsprechend hoch gewichtet.

Offenbar gibt es hier einen echten Trade-off zwischen Reinheit und Recall. Für ein Verhältnis von 1:1000 ist die Reinheit sehr gut, der Recall aber sehr schlecht. Bei einem Verhältnis von 1:1 verhält es sich genau anders herum.
Betrachten wir das geometrische Mittel aus beiden mit Gewichten 0,6 für Reinheit und 0,4 für Recall, dann sieht man ein Maximum bei einem Verhältnis von 1:50. Dieses scheint unter diesem Kriterium das beste Verhältnis für Trainingsdatensätze zu sein.
Es ist dabei zu beachten, dass die Gewichte im geometrischen Mittel anwenderabhängig sind und frei gewählt werden können. Es ergibt sich aber, dass auch bei einer noch höheren Gewichtung der Reinheit von zum Beispiel 0,7 das Verhältnis 1:50 erste Wahl bleibt. Lediglich bei einer höheren Gewichtung des Recalls sollte zu einem Verhältnis von 1:10 übergegangen werden.
Im Folgenden nutzen wir also für alle Random Forests ein Gamma-Hadron-Verhältnis von 1:1 für alle Datensätze. Wir gewichten sie jedoch so, dass das tatsächliche Verhältnis 1:50 für Trainingsdatensätze und 1:1000 für Testdatensätze entspricht.

5.3 Variablenreduktion

Wie in Abschnitt 4: Datenexploration beschrieben wurde, lässt sich bei manchen der 29 betrachteten Variablen mit bloßem Auge nur sehr schwierig ein Unterschied

zwischen Gammas und Hadronen ausmachen.
Des Weiteren ist die Korrelation zwischen manchen Variablen sehr hoch. Zwischen *Size* und *SizeMainIsland* beträgt sie zum Beispiel 0,98.
Es stellt sich also die Frage, ob zur Gamma-Hadron-Separation nicht auch deutlich weniger Variablen ausreichen als die gesamten hier vorliegenden.
Im Folgenden nehmen wir daher eine Variablenreduktion vor und überprüfen im nächsten Schritt, ob sich dadurch die Trennungsqualität im Vergleich zu den Originaldaten unter Benutzung eines Random Forests (dem Verfahren, das bisher zur Separation verwendet wird) verschlechtert. Bleibt die Trennungsqualität gleich, oder wird sogar besser, so spricht nichts gegen die Variablenreduktion.
Zur Reduktion verwenden wir drei verschiedene Methoden:
Zum einen wählen wir Variablen anhand der Kullback-Leibler-Divergenz aus. Dadurch schließen wir Variablen aus, die scheinbar wenig zur Trennungsqualität der Daten beitragen.
Außerdem verwenden wir eine Hauptkomponentenanalyse, die vor allem dazu geeignet ist, Korrelationen zwischen den Variablen zu beseitigen und gleichzeitig die Dimension der Daten zu reduzieren.
Im letzten Schritt verwenden wir einen evolutionären Algorithmus, um diejenigen Variablen zu selektieren, die die beste Trennung liefern.
Es ist bei allen Verfahren jeweils das Ziel, die Variablen wenigstens um die Hälfte, also auf etwa 15 zu reduzieren.

5.3.1 Kullback-Leibler-Divergenz

Zur ersten, einfachen Reduktion der Variablen nutzen wir in diesem Abschnitt die Kullback-Leibler-Divergenz. Dazu bestimmen wir für jede Variable die Kullback-Leibler-Divergenz zwischen Hadronen- und Gamma-Daten wie in Abschnitt 3.2 beschrieben. Anhand dieser Werte wählen wir dann die Variablen zur Reduktion aus. Dabei wählen wir die 15 Variablen mit den größten Kullback-Leibler-Divergenzen, da dies bedeutet, dass in diesen Variablen die größten Unterschiede zwischen den Verteilungen von Gammas und Hadronen existieren. Wir wählen deshalb 15, weil dies genau der Hälfte der 30 Originalvariablen entspricht.
Hierbei ist darauf zu achten, dass wir nur die eindimensionalen Dichten der Variablen schätzen und damit die Kullback-Leibler-Divergenz berechnen. Dadurch kann es zum Beispiel passieren, dass hoch korrelierte Variablen gewählt werden, die wechselseitig wenig Trennungsqualität beitragen, aber für sich genommen eine

gute Trennungsqualität liefern. Dies kann man umgehen, indem man die Kullback-Leibler-Divergenz von höherdimensionalen, gemeinsamen Dichten berechnet und diese als Kriterium zur Variablenauswahl herzieht. In dieser Arbeit beschränken wir uns jedoch auf den eindimensionalen Fall.
Die ausgewählten Variablen und deren Kullback-Leibler-Divergenzen sind in Tabelle 20 dargestellt. Eine Auflistung der Kullback-Leibler-Divergenzen der anderen Variablen ist in Tabelle A.1 im Anhang zu finden.

Variable	KL-Div
NumIslands	0,987
SizeSubIslands	0,699
ConcCOG	0,441
Length	0,340
BorderLine	0,322
Area	0,256
Size	0,223
SizeMainIsland	0,211
MeanX	0,209
Width	0,204
SlopeLong	0,181
Ellip	0,165
Leagake1	0,140
M3Trans	0,132
Leakage2	0,129

Tabelle 20: Ausgewählte Variablen zur Variablenreduktion und deren Kullback-Leibler-Divergenzen.

5.3.2 Hauptkomponentenanalyse

Wie in Abschnitt 3.3 beschrieben, konstruiert eine Hauptkomponentenanalyse aus den 29 vorhandenen Variablen 29 Linearkombinationen mit sukzessiv kleiner werdender Varianz. Diese Linearkombinationen nennt man Hauptkomponenten. Die Idee der Hauptkomponentenanalyse ist, dass dann die ersten Hauptkomponenten die meiste Varianz (= Information) enthalten und dass man die anderen weglassen kann, da sie nur wenig Information enthalten.
Wir betrachten in diesem Abschnitt eine Hauptkomponentenanalyse mit und eine ohne Skalierung. Wie viele Hauptkomponenten behalten werden, ist dem Anwender überlassen. Wir behalten hier so viele Hauptkomponenten, dass ihre kumulative Varianz gerade größer als 95% der Gesamtvarianz ist. Dies ist bei der HKA mit Skalierung der Fall, wenn wir die ersten 16 Hauptkomponenten betrachten. Nach

diesem Kriterium haben wir die Anzahl der Variablen also wie bei der Nutzung der Kullback-Leibler-Divergenz etwa halbiert, wie es das Ziel war. Wie sich die ersten 16 Hauptkomponenten aus den Variablen zusammensetzen ist in Tabelle 21 zu sehen.

In der Regel sind Hauptkomponenten schwierig oder gar nicht zu interpretieren. In diesem Fall ist aber jedenfalls die erste Hauptkomponente, die auch eine deutlich größere Varianz hat als die nachfolgenden, recht gut zu interpretieren. Hier liegt großes Gewicht auf *Length*, *Width*, *Area*, *NumUsedPixels* und anderen Variablen, die die Größe des Ereignisses angeben, sowie auf *Size*, das als eine grobe Annäherung der Energie gewertet werden kann, während kleines Gewicht auf *Conc* liegt. Kleine Werte dieser Variable entsprechen also kleinen Ereignissen, während große Werte großen Ereignissen entsprechen. Die erste Hauptkomponente erklärt also die Varianz, die dadurch entsteht, dass es kleine und große Ereignisse gibt.
Intuitiv interessant für die Klassifizierung scheint die siebte Hauptkomponente zu sein. Durch sie werden solche Ereignisse beschrieben, die einen großen positiven Wert in *MeanX*, *Dist0*, *Asym* und *SlopeLong*, aber einen großen negativen Wert in *M3Long* und *M3Trans* haben. In Abbildung 16 in Abschnitt 4.3 haben wir gesehen, dass die verwendeten Gamma-Simulationen eher auf der rechten Seite der Kamera liegen, also im Allgemeinen einen höheren Wert in *MeanX* haben als Hadronen. Gleichzeitig sind Gammas im Allgemeinen mit ihrer großen Halbachse auf die Quellposition ausgerichtet, wobei der Schwerpunkt des Schauers zur Quellposition hin verschoben ist, was eine Asymmetrie bzw. Schiefe bedeutet. Es ist also zu erwarten, dass Gammas in Hauptkomponente Sieben eher kleine Werte realisieren, Hadronen dagegen größere.
Als Ergänzung und zum Vergleich führen wir außerdem noch eine Hauptkomponentenanalyse ohne Skalierung durch. Die entsprechende Rotationsmatrix befindet sich in Tabelle A.2 im Anhang. Hier reichen schon die ersten zwei Variablen, um 95% der Varianz zu erklären. Für einen Vergleich mit der HKA mit Skalierung verwenden wir aber auch hier 16 Hauptkomponenten.

5.3.3 Evolutionärer Algorithmus

Als drittes Verfahren zur Variablenselektion verwenden wir einen evolutionären Algorithmus wie er in Abschnitt 3.6 vorgestellt wurde.

	HK1	HK2	HK3	HK4	HK5	HK6	HK7	HK8	HK9	HK10	HK11	HK12	HK13	HK14	HK15	HK16
Length	0,27		0,18	-0,09		-0,11					0,08	0,09	0,10			
Width	0,24		-0,08		-0,13	0,14				0,14	-0,15	-0,28	-0,29			
Delta		0,68	0,16					-0,06	0,10							
Size	0,27		-0,06	0,08	-0,08	-0,09					0,08		-0,10			
MeanX			-0,19	-0,14	-0,19	0,07	0,59	-0,09	0,11	0,13	0,16	-0,14	0,25	-0,56	0,08	0,30
MeanY		-0,16			-0,06			-0,39	0,80	0,22	-0,03	0,33	-0,08			
Dist0	0,12	0,09	-0,36	-0,20	0,28		0,17							-0,12	0,31	-0,76
BorderLine	0,28		0,15	-0,07		-0,09										
Area	0,28		0,10									-0,07	-0,07			
CosDelta			0,09	0,09					-0,07	0,67	0,29	-0,18	0,19	-0,29	0,13	-0,10
SinDelta		0,68	0,16						0,10				0,00			
Asym			-0,14	-0,35	-0,35		0,16		-0,07	0,29	0,20		0,35	0,55	-0,34	-0,18
M3Long			-0,18	-0,44	-0,32		-0,30			-0,14				0,15	0,67	0,28
M3Trans			-0,12	-0,35	-0,32	-0,08	-0,40		0,11	-0,31			-0,08	-0,45	-0,50	-0,16
SlopeLong			0,06		-0,33	0,12	0,11		-0,40	0,29	-0,15	0,67	-0,31	-0,18		-0,12
SlopeTrans							-0,07	-0,91	-0,36	-0,08	-0,06	-0,18				
NumIslands	0,10	-0,10	0,37	-0,25	0,07	0,46				-0,08	0,14					
SizeSubIslands	0,09	-0,07	0,27	-0,24	0,09	0,55					0,07	-0,10	-0,12			-0,07
SizeMainIsland	0,26		-0,11	0,12	-0,10	-0,12					0,07		-0,11			
Leagake1	0,15	0,08	-0,35	-0,15	0,37	0,17	-0,14		-0,06	0,07		0,20			-0,17	0,30
Leakage2	0,16	0,08	-0,34	-0,18	0,37	0,15	-0,09		-0,06	0,08		0,18			-0,14	0,23
Conc	-0,28					0,10				-0,06	0,08	0,06				
Conc1	-0,28					0,09				-0,06	0,10	0,06				
ConcCOG	-0,27		-0,15	0,10		0,09				-0,03	0,07					
ConcCore			-0,20	0,24	-0,15	0,12				-0,25	0,79	0,10	-0,27			
NumUsedPixels	0,20		-0,08	0,29	-0,18	0,30	-0,12			-0,15	-0,14	0,11	0,39		0,06	-0,06
NumCorePixels	0,19		-0,10	0,30	-0,17	0,30	-0,13			-0,16	-0,15	0,12	0,42		0,06	-0,07
UsedArea	0,28				-0,07	-0,09						-0,06	-0,09			
CoreArea	0,28			0,06	-0,07	-0,07										
Ellip	0,14	-0,08	0,30	-0,18	0,17	-0,32			-0,06	-0,19	0,25	0,36	0,35	-0,07		
Standardabweichung	3,41	1,42	1,41	1,33	1,23	1,18	1,11	1,00	0,99	0,97	0,97	0,96	0,93	0,85	0,72	0,69
Anteil an der Varianz	0,39	0,07	0,07	0,06	0,05	0,05	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,02	0,02	0,02
Kumulativer Anteil	0,39	0,45	0,52	0,58	0,63	0,68	0,72	0,75	0,78	0,82	0,85	0,88	0,91	0,93	0,95	0,96

Tabelle 21: Zusammensetzung der ersten 16 Hauptkomponenten bei einer HKA mit Skalierung. Werte mit einem Absolutbetrag kleiner oder gleich 0,05 wurden der Übersicht halber weggelassen.

Bei der Optimierung nutzen wir aufgrund der sehr hohen Rechenzeit als Trainingsdatensatz eine Stichprobe mit 10.000 Beobachtungen, bei dem das Verhältnis von Gammas zu Hadronen 1:50 beträgt und einen Testdatensatz mit 50.000 Beobachtungen, bei einem Gamma-Hadron-Verhältnis von 1:1, wobei wir die falsch klassifizierten Hadronen zur Berechnung der Reinheit mit dem Faktor 1.000 gewichten.

Die 11 zur Selektion empfohlenen Variablen sind in Tabelle 22 dargestellt.

Width	*MeanX*	*Delta*
Dist0	*M3Long*	*M3Trans*
SlopeLong	*SlopeTrans*	*NumIslands*
SizeMainIsland	*NumUsedPixels*	

Tabelle 22: Vom evolutionären Algorithmus ausgewählte Variablen zur Variablenreduktion

5.3.4 Testen der Trennungsqualität

Nachdem wir nun drei verschiedene Variablenreduktionen durchgeführt haben, werden wir in diesem Abschnitt überprüfen, ob die Daten dadurch Trennungsqualität gewonnen oder verloren haben. Wenn die Klassifizierung mithilfe der reduzierten Daten zumindest nicht schlechter ist als die mit den Originaldaten, dann spricht nichts gegen eine Variablenreduktion.
Zum Test verwenden wir eine Stichprobe der Daten. Die Stichprobe enthält 408.000 Beobachtungen, wobei 8.000 Beobachtungen Gammas und 400.000 Hadronen sind. Das Verhältnis von Gammas zu Hadronen beträgt also 1:50. Mit diesem Datensatz führen wir nacheinander die vorgestellten Verfahren zur Variablenreduktion durch: Die Reduktion mittels Kulback-Leibler-Divergenz (KL-Div), die Hauptkomponentenanalyse mit und ohne Skalierung (HKAmit/HKAohne) und die Reduktion auf die vom evolutionären Algorithmus vorgeschlagenen Variablen (Evol).
Mit dem Originaldatensatz und den reduzierten Datensätzen führen wir dann jeweils eine 10-fache Kreuzvalidierung durch. In jedem Durchlauf der Kreuzvalidierung werden Reinheit und Recall ermittelt und dann über die zehn Durchläufe gemittelt. Dabei multiplizieren wir die Anzahl der falsch klassifizierten Hadronen mit 20, um auf das realistischere Gamma-Hadron-Verhältnis von 1:1000 zu kommen, wie in Abschnitt 5.2 beschrieben.
Die Ergebnisse sind in Tabelle 23 aufgeführt. Zusätzlich zu mittlerer Reinheit und

Recall ist auch noch das geometrische Mittel mit Gewichten 0,6 für Reinheit und 0,4 für Recall angegeben, um beide Werte in einem Gütemaß zu vereinen.

	Reinheit	Recall	geom. Mittel
Original	0,42	0,14	0,27
KL-Div	0,27	0,15	0,21
HKAmit	0,49	0,03	0,15
HKAohne	0,53	0,04	0,19
Evol	0,55	0,11	0,29

Tabelle 23: Ergebnisse der Klassifizierung unter Nutzung der verschiedenen Variablenreduktionen.

Offenbar liefert die Hauptkomponentenanalyse eher schlechte Ergebnisse. Sowohl mit als auch ohne Skalierung ist vor allem der Recall sehr schlecht. Dabei verbessert sich allerdings die Reinheit recht stark. Die HKA ohne Skalierung erreicht bei der Reinheit sogar das zweitbeste Ergebnis. Durch den schlechten Recall schneiden beide HKAn im geom. Mittel, also insgesamt, am schlechtesten ab.
Die Auswahl mittels Kullback-Leibler-Divergenz liegt im Mittelfeld der Ergebnisse. Während die Reinheit sich stark verschlechtert hat, ist der Recall leicht besser geworden. Das geometrische Mittel liegt wegen der schlechteren Reinheit aber deutlich unter dem der Originaldaten.
Heraus sticht der evolutionäre Algorithmus, der vor allem bei der Reinheit sehr gut abschneidet. Der Recall ist hier zwar auch etwas schlechter als bei den Originaldaten, die deutlich verbesserte Reinheit führt hier jedoch zu einem verbesserten Gesamtergebnis. Dabei ist gleichzeitig die Reduzierung der Variablen beim evolutionären Algorithmus am stärksten. Hier wurden nur 11 Variablen vorgeschlagen.
Es spricht also nicht nur nichts gegen eine Variablenreduktion, sie ist auch sehr empfehlenswert, um eine bessere Trennungsqualität des Random Forest zu erzielen. Des Weiteren profitiert man von geringerem Rechenaufwand und weniger Speicherplatzverbrauch durch weniger Variablen.

6 Eigenschaften schwierig zu klassifizierender Gammas

Wie wir gesehen haben, ist, egal, ob wir einen reduzierten Datensatz, oder alle Variablen verwenden, der Recall grundsätzlich sehr schlecht. Das heißt, dass viele Gammas nicht als solche erkannt werden. Wir werden nun untersuchen, welche einfacher und welche schwieriger zu klassifizieren waren.
Intuitiv lässt sich vermuten, dass die Schwierigkeit, ein Gamma zu klassifizieren, mit der Größe des Schauers zusammenhängt. Je größer ein Schauer ist, desto einfacher lässt sich eine Struktur erkennen, die klar für oder gegen ein Gamma spricht. Dagegen ist es bei kleinen Ereignissen, die im Extremfall nur aus 5 Pixeln bestehen, sehr schwierig, einen Unterschied zwischen Gammas und Hadronen zu erkennen.
Wie in Abschnitt 5.3.2 beschrieben, beinhaltet die erste Hauptkomponente alle Information über die Größe eines Ereignisses. Wir betrachten im Folgenden also, welche Beobachtungen dieser HK im Allgemeinen richtig und welche falsch klassifiziert werden.
Dies ist in Abbildung 29 zu sehen. In der ersten Hauptkomponente stehen große Werte für große Ereignisse, während kleine Werte für kleine Ereignisse stehen. Auffällig ist in der Abbildung, dass bei großen Werten der Anteil richtig klassifizierter Gammas sehr hoch ist. Er liegt fast durchgehend bei über 88%. Allein zwei Klassen ganz rechts haben etwas kleinere Werte, die aber auch noch bei 67% oder mehr liegen. Aufgrund der dünn besetzten Klasse kann dies jedoch leicht eine zufällige Abweichung sein.
Zu kleineren Werten, also kleineren Ereignissen, hin wird der Anteil richtig klassifizierter Gammas kleiner. Das Minimum liegt bei 23%.
Offenbar sind also tatsächlich größere Ereignisse deutlich einfacher zu klassifizieren als kleine.
Gleichzeitig ist aber auch zu sehen, dass die Klassen bei größeren Werten dünner besetzt sind als kleinere. Wie sich bei *Area* und *Size* schon gezeigt hat, ist also auch hier zu sehen, dass kleine Ereignisse häufiger vorkommen als große.

Eine weitere Frage, die man sich neben der nach der Größe eines Schauer stellen kann, ist, wie die erkannten Gammas in der von *Area* und $\log_{10}(\textit{Size})$ aufgespannten Fläche liegen. Wie in Abschnitt 2.2.3 beschrieben, wird teilweise im Ganymed-Schritt zur Separation statt des Random Forest ein einfacher Schnitt in dieser Ebe-

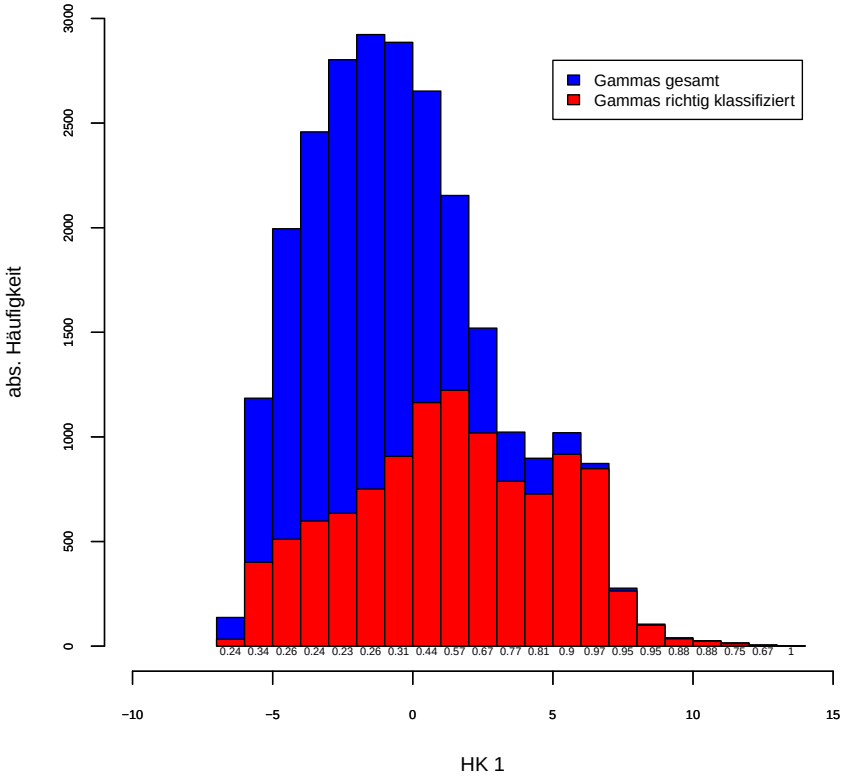

Abbildung 29: Der Anteil richtig klassifizierter Gammas an der Gesamtanzahl der Gammas in der ersten Hauptkomponente. Unter jeder Säule ist der Anteil der richtig klassifizierten Gammas in dieser Klasse angegeben.

ne genutzt, um Gammas und Hadronen zu trennen. Insbesondere werden dabei, wie in Abbildung 7 zu sehen war, niemals Beobachtungen mit einer Size kleiner als 10^2 als Gamma klassifiziert. In Abbildung 30 ist zu sehen, wie die Trennung mit dem Random Forest aussieht. Wie schon in Abbildung 29 zu sehen war, ist die Klassifizierung von Beobachtungen von Werten mit großer Size problemlos. Zu kleineren Werten von Size hin werden jedoch viele Gammas falsch klassifiziert, besonders dann, wenn sie eine etwas größere Area besitzen. Vor allem werden in diesem Bereich kleinerer Size auch viele Hadronen fälschlicherweise als Gamma klassifiziert. Positiv anzumerken ist jedoch, dass, im Gegensatz zum Area-Cut, auch viele Gammas mit kleiner Size richtig als solche erkannt werden. Ob der Vorteil der richtig erkannten Gammas bei kleiner Size den Nachteil der falsch als Gamma klassifizierter Hadronen überwiegt, ist noch zu untersuchen, was die Möglichkeiten dieser Arbeit jedoch übersteigt.

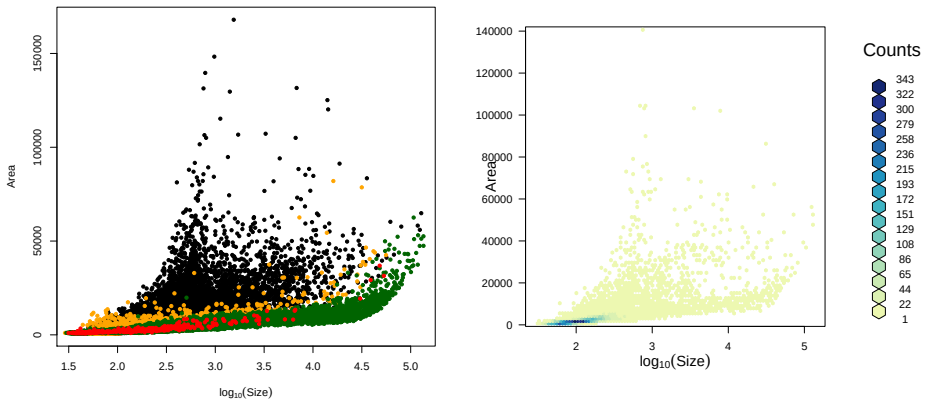

Abbildung 30: Links: Area gegen $log_{10}(\textit{Size})$. Zu sehen sind richtig klassifizierte Hadronen (schwarz), richtig klassifizierte Gammas (grün), als Gamma klassifizierte Hadronen (rot) und als Hadron klassifizierte Gammas (gelb). Rechts: Area gegen $log_{10}(\textit{Size})$, klassiert.

Dass es wichtig sein kann, auch für kleine Ereignisse eine ordentliche Klassifizierung zu erzielen, zeigt ebenfalls Abbildung 30. Wie man sehen kann, konzentriert sich ein großer Teil der Beobachtungen in einem kleinen Bereich, nämlich bei kleinen Size- und Area-Werten. Verzichtet man auf diese Daten (wie es beim Area-Cut immer der Fall ist), verzichtet man bewusst auf den größten Teil der Gamma-Daten.

7 Diskussion und Zusammenfassung

Wir konnten mithilfe von verschiedenen Vorverarbeitungen die Qualität der Separation von Gammas und Hadronen deutlich verbessern. Die Ergebnisse erscheinen jedoch zunächst immer noch unbefriedigend. So sind die Werte für die erzielte Reinheit mit einem Maximum von 0,56 nach der Vorverarbeitung sehr schlecht. Dies bedeutet nämlich, dass noch mehr als 40% Hadronen nach der Klassifikation in der Stichprobe vorhanden sind. Hier ist jedoch zu beachten, dass an dieser Stelle noch kein θ^2-Cut durchgeführt wurde. Durch diesen kann sich die Reinheit der Stichprobe noch deutlich verbessern. Dieser Sachverhalt ist noch zu untersuchen.
Bei der Interpretation der Ergebnisse dieser Arbeit sollten verschiedene Gesichtspunkte beachtet werden. Da wäre beispielsweise der Fakt, dass mit simulierten Gammas, aber echten Hadronendaten gearbeitet wurde. Hier können systematische Fehler erkennbar werden, die die Separation vereinfachen (oder auch erschweren) können, die aber in echten Daten nicht vorhanden sind. Inwieweit solche Fehler vorhanden sind, ist nicht bekannt. Die gravierenden Unterschiede zwischen Gammas und Hadronen in den Variablen *NumSinglePixels* und *SizeSinglePixels* könnten aber ein Hinweis auf deren Existenz sein.
An dieser Stelle gibt es ohnehin auch Klärungsbedarf: Kann ein solch großer Unterschied zwischen Gammas und Hadronen existieren? Wenn ja, ist in diesen beiden Variablen großes Potential erkennbar, die Qualität eines jeden Klassifikators zur Gamma-Hadron-Separation deutlich zu verbessern.
Ein weiteres Problem besteht in der Struktur der Hadronendaten. Diese wurden aufgenommen, während das Teleskop im Wobble-Modus arbeitete, also während es auf eine Quelle ausgerichtet war. Dies bedeutet zwangsläufig, dass die Hadronendaten mit Gammas durchsetzt sind, und zwar mit dem oben angesprochenen Verhältnis von etwa 1:1000. Bei einem Trainingsverhältnis von 1:50, das hier benutzt wurde, bedeutet das, dass auf 20 Gammas, von denen man weiß, dass es Gammas sind, ein Gamma kommt, das als Hadron gekennzeichnet ist. Dies kann entscheidenden Einfluss auf eine Klassifikation haben. Die Nutzung von simulierten oder echten, im On-Off-Modus aufgenommenen, Hadronendaten scheint für die Zukunft unerlässlich.
In Abschnitt 6 haben wir gesehen, dass große Events, also zum Beispiel mit einer großen *Area* und *Size*, einfacher zu klassifizieren sind als solche mit kleineren Werten in diesen Variablen. Dieser Fakt muss bei der weiteren Auswertung unbedingt berücksichtigt werden. Er bedeutet, dass in der Gamma-Stichprobe, die

man nach der Separation erhält, überproportional viele höherenergetische Ereignisse vorkommen als niederenergetische (denn vor allem *Size* ist ein Indikator für die Energie des Ereignisses). Wir haben zum Beispiel gezeigt, dass sich die Verteilungen der Gammas vor und nach der Separation in der ersten HK deutlich unterscheiden.
Optimalerweise würde man einen anderen Ansatz als den Random Forest finden, um an die schwierig zu klassifizierenden Beobachtungen zu kommen, denn offensichtlich eignet sich dieser so wie er hier verwendet wurde nicht dazu. Die Wahl fiel in dieser Arbeit deshalb auf dieses Verfahren, weil es eines der Verfahren ist, das bereits zur Klassifikation echter Daten eingesetzt wird. Eine Antwort auf die Frage, ob es bessere Verfahren für speziell dieses Problem gibt, steht jedoch noch aus. Nicht nur um kleine Ereignisse besser klassifizieren zu können, auch generell sollte überprüft werden, ob es bessere Alternativen zum Random Forest gibt. Eventuell kann man dabei schon vor dem Ganymed-Schritt in Callisto ansetzen, um beispielsweise andere (oder weitere) Variablen als die Hillas-Parameter zu extrahieren. Hier wäre die Anpassung einer zweidimensionalen Dichtefunktion an den Schauer denkbar. Statt oder zusätzlich zu den Hillas-Parametern könnten dann die Parameter der Verteilung zur Klassifikation verwendet werden. Weitere mögliche Variablen wären Gütemaße, die die Anpassung der Verteilung an den Schauer beschreiben oder verschiedene Momente der Verteilung.

Insgesamt kann als Ergebnis dieser Arbeit festgehalten werden, dass eine Reduktion der Hillas-Parameter von 32 auf 11 sinnvoll ist. Nicht nur, um Rechenzeit und Speicherplatz einzusparen, sondern auch, um die Güte der Trennung von Gammas und Hadronen zu verbessern.
Weiterhin war zu sehen, dass vor allem kleine Events schwierig zu klassifizieren sind, während bei großen Events keine Probleme bestehen. Es sollte ein Verfahren entwickelt werden, das explizit in kleinen Events nach Unterschieden zwischen Hadronen und Gammas sucht und diese zur Klassifikation nutzt.

Literaturverzeichnis

- Albert, J. (2007), *VHE Gamma-Ray Observation of the Crab Nebula and its Pulsar with the MAGIC Telescope*, The Astrophysical Journal 674
- Albert, J. et al. (2008) *Implementation of the Random Forest Method for the Imaging Atmospheric Cherenkov Telescope MAGIC*, Nuclear Instruments and Methods in Physics Research A588, S. 424
- Backes, M. et al. (2007), *Long term monitoring of bright TeV Blazars with the MAGIC Telescope*, Astronomische Nachrichten 328, S. 677
- Backes, M. (2008), *Langzeitbeobachtung von TeV-Blazaren - Quellenanalyse mit MAGIC und Konzeption eines dedizierten Teleskops*, Diplomarbeit, Technische Universität Dortmund
- Bäck, T. et al. (1997) *Handbook of Evolutionary Computation*, Taylor and Francis Group, New York
- Bastieri, D. et al. (2005), *The Mirrors of the MAGIC-Telescope*, Proceedings of the 29th International Cosmic Ray Conference, Bd. 5, S. 283, Pune
- Box, G.E.P., Cox, D.R. (1964), *An Analysis of Transformations*, Journal of the Royal Statistical Society B26, S. 211-252
- Breiman, L. (2001), *Random Forests*, Machine Learning 45, S. 5
- Bretz, T. (2006), *Observations of the Active Galactic Nucleus 1ES 1218+304 with the MAGIC-Telescope*, Dissertation, Bayerische Julius-Maximilians-Universität, Würzburg
- Büning, H., Trenkler, G. (1994), *Nichtparametrische statistische Methoden*, 2. Aufl., de Gruyter, Berlin/New York
- Carr, D. et al. (2009), *hexbin: Hexagonal Binning Routines*, R package version 1.20.0
- Doert, M. (2009) *Automated Production of Standardized and Tailor-made Monte Carlo Simulations for the MAGIC Telescopes*, Diplomarbeit, Technische Universität Dortmund

- Errando, M. (2006), *Study of optical properties of last generation photodetectors for Cherenkov astronomy applications*, Masterarbeit, Universitat Autonoma de Barcelona
- Ferenc, D. et al. (2005), *The MAGIC Gamma-Ray Observatory*, Nuclear Instuments and Methods in Physics Research A553, S. 274
- Fomin, V.P. (1994), *New methods of atmospheric Cherenkov imaging for gamma-ray astronomy. I. The false source method*, Astroparticle Physics, Volume 2, Issue 2 S. 137
- Fried, R. (2009) *Skript zur Vorlesung Multivariate statistische Verfahren*, Technische Universität Dortmund, unveröffentlicht
- Grieder, P.K.F. (2010) *Extensive Air Showers*, Springer, Heidelberg
- Hadasch, D. (2008), *Study of the MAGIC Performance at High Zenith Angles and Application of the Results on a Very High Energy Gamma Ray Flare of the Blazar PKS 2155-304*, Diplomarbeit, Technische Universität Dortmund
- Hartung, J., Elpelt, B. (1999), *Multivariate Statistik*, 7. Auflage, Oldenbourg, München/Wien
- Hartung, J. et al.(2005), *Statistik - Lehr- und Handbuch der angewandten Statistik*, 14. Aufl., Oldenbourg, München/Wien
- Helf, M. (2009), *Genetische Merkmalsselektion für die Gamma-Hadron-Separation im MAGIC-Experiment*, Studienarbeit, Technische Universität Dortmund
- Hillas, A.M. (1985), *Cherenkov Light Images of EAS Produced by Primary Gamma.*, Proceedings of the 19th International Cosmic Ray Conference ICRC, Bd. 3, S. 445, San Diego
- Hsu, C.C. et al. (2007), *The Camera of the MAGIC-II Telescope*, Proceedings of the 30th International Cosmic Ray Conference ICRC, Merida
- Kullback, S., Leibler, R.A. (1951), *On Information and Sufficiency*, Annals of Mathemtical Statistics 22, Nr. 1, S.79
- Lessard, R.W. et al. (2001), *A New Analysis Method for Reconstructing the Arrival Direction of TeV Gamma Rays Using a Single Imaging Atmospheric Cherenkov Telescope*, Astroparticle Physics 15, S. 1

- Li, T.P., Ma, Y.Q. (1983), *Analysis Methods for Results in Gamma-Ray Astronomy*, Astrophysical Journal 272, S. 317

- Liaw, A., Wiener, M. (2002), *Classification and Regression by randomForest*, RNews Vol. 2/3, S. 18-22

- MAGIC-Kollaboration (2010), *Offizielle Homepage*, http://magic.mppmu.mpg.de/

- MARS-Software (2010), Online erhältlich unter http://magic.astro.uni-wuerzburg.de/mars/

- Mazin, D. (2007), *A Study of Very High Energy γ-Ray Emission From AGNs and Constraints on the Extragalactic Background Light*, Dissertation, Technische Universität München

- Neuwirth, E. (2007), *RColorBrewer: ColorBrewer palettes*, R package version 1.0-2

- R Development Core Team (2009). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Wien, URL http://www.R-project.org.

- Riegel, B. (2005), *Systematische Untersuchung der Bildparameter zur Entwicklung einer Standardanalyse für das MAGIC-Teleskop*, Diplomarbeit, Julius-Maximilians-Universität, Würzburg

- Rissi, M.T. (2009) *Detection of Pulsed Very High Energy Gamma-Rays from the Crab Pulsar with the MAGIC telescope using an Analog Sum Trigger*, Dissertation, Swiss Federal Institute of Technology, Zürich

- Rügamer, S. (2006), *Systematische Studien der Verwendung der Zeitstruktur von Luftschauern zur Reduktion des Untergrundes in MAGIC-Daten*, Diplomarbeit, Julius-Maximilians- Universität, Würzburg

- Sidro Martin, N. (2008), *Discovery and Characterization of the Binary System LSI +61303 in Very High Energy Gamma-Rays with MAGIC*, Dissertation, Universitat Autonoma, Barcelona

- Silverman, B. W. (1986), *Density Estimation*, Chapman and Hall, London

- Wagner, R. (2006), *Very High Energy Blazar Astrophysics - Measurement of Very High Energy Gamma-Ray Emission from Four Blazars Using the MAGIC Telescope and a Comparative Blazar Study*, Dissertation, Technische Universität München
- Weekes, T.C. (2003), *Very High Energy Gamma-Ray Astronomy*, Institute of Physics Publishing, Bristol/Philadelphia

Anhang

Variable	KL-Div
NumIslands	0,987
SizeSubIslands	0,699
ConcCOG	0,441
Length	0,340
BorderLine	0,322
Area	0,256
Size	0,223
SizeMainIsland	0,211
MeanX	0.209
Width	0.204
SlopeLong	0.181
Ellip	0.165
Leagake1	0.140
M3Trans	0.132
Leakage2	0.129
M3Long	0.122
CoreArea	0.083
Conc	0.082
Asym	0.080
Conc1	0.073
Dist0	0.071
ConcCore	0.064
NumCorePixels	0.042
NumUsedPixels	0.030
MeanY	0.024
UsedArea	0.021
SlopeTrans	0.011
Delta	0.002
SinDelta	0.002
CosDelta	0.001

Tabelle A.1: Alle Variablen und deren Kullback-Leibler-Divergenzen.

	HK1	HK2	HK3	HK4	HK5	HK6	HK7	HK8	HK9	HK10	HK11	HK12	HK13	HK14	HK15	HK16
Length												0,08		-0,13		
Width																
Delta										-0,79	0,09					0,61
Size														-0,16	0,12	
MeanX	-1,00															
MeanY		-1,00														
Dist0			0,10	0,73	0,54	-0,40		0,06								
BorderLine																
Area																
CosDelta										-0,11	-0,99					
SinDelta										-0,61	0,07					-0,79
Asym			0,51	-0,07	-0,51	-0,68		0,11								
M3Long			0,77	0,06	0,12	0,41		-0,45								
M3Trans			0,34		0,08	0,33		0,88								
SlopeLong																
SlopeTrans																
NumIslands												0,24	0,97			
SizeSubIslands				0,17	-0,21	0,08	0,96									
SizeMainIsland														-0,15	0,13	
Leagake1														-0,06		
Leakage2														-0,10	0,08	
Conc												-0,11		0,59	-0,19	
Conc1														0,27	-0,08	
ConcCOG												-0,18		0,61	-0,09	
ConcCore														0,26	0,95	
NumUsedPixels			-0,10	0,49	-0,47	0,23	-0,19		-0,66							
NumCorePixels			-0,09	0,43	-0,40	0,20	-0,19		0,75							
UsedArea																
CoreArea																
Ellip												0,94	-0,25	0,23		
Standardabweichung	483,27	102,69	50,63	43,00	35,73	34,63	27,85	25,50	2,18	1,11	1,01	0,34	0,21	0,12	0,10	0,06
Anteil an der Varianz	0,93	0,04	0,01	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Kumulativer Anteil	0,93	0,97	0,98	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Tabelle A.2: Zusammensetzung der ersten 16 Hauptkomponenten bei einer HKA ohne Skalierung. Werte mit einem Absolutbetrag kleiner oder gleich 0,05 wurden der Übersicht halber weggelassen.

Eidesstattliche Erklärung des Urhebers

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht habe.

Dortmund, den 28. Juni 2010

Unterschrift

Einverständniserklärung des Urhebers

Ich erkläre mich hiermit einverstanden, dass meine Diplomarbeit nach §6 (1) des URG der Öffentlichkeitdurch die Übernahme in die Bereichsbibliotheken zugänglich gemacht wird. Damit können Leser der Bibliothek die Arbeit einsehen und zu persönlichen wissenschaftlichen Zwecken Kopien aus dieser Arbeit anfertigen. Weitere Urheberrechte werden nicht berührt.

Dortmund, den 28. Juni 2010

Unterschrift