University of the Witwatersrand
Johannesburg, South Africa

Doctoral Thesis

A Bayesian Framework for Forensic Investigations Involving Lightning Location System Data

Author:
Hugh Gordon Patrick Hunt

Supervisor:
Prof. Kenneth John Nixon

April 23, 2018

A thesis submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in fulfillment of the requirements for the degree of Doctor of Philosophy.
in the

Lightning and EMC Research Group
School of Electrical and Information Engineering
Faculty of Engineering and the Built Environment

Declaration of Authorship

I declare that this thesis is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

....
(Signature of Candidate)

.... day of .... year ....

Abstract

The work presented in this thesis extends and contributes to research into the forensic use of Lightning Location Systems and the uncertainty present due to location errors. While previous work in this area has produced some approaches that can be used in forensic investigations, there has not been a consistent, standardised approach to presenting Lightning Location System stroke reports as evidence in a legal environment. In the research presented, a Bayesian framework for representing Lightning Location System data as likelihood ratios and posterior probabilities is developed. The statistical models necessary for use of the framework are discussed and verified through ground-truth events and bivariate statistical analysis. Photographs of multiple lightning events to the Brixton tower in South Africa and current measurements at the Gaisberg tower in Austria are used as ground-truth data and bivariate statistical techniques are used to fit and evaluate different statistical models. It is shown that the bivariate Students' t-distribution is the best fit for Lightning Location System location errors, rather that the commonly assumed bivariate Gaussian distribution and that a bivariate Gaussian Mixture Model can be used to describe the prior probability of lightning occurrence in a region. It is demonstrated how the Bayesian framework can be used to present Lightning Location System data as evidence in a court of law. This represents a unique and valuable contribution to those working in the field of lightning location and, in particular, in forensic situations.

Acknowledgements

I would like to thank Professor Kenneth J. Nixon for his enthusiasm, guidance and support. I would also like to thank Professor Ian R. Jandrell, Dr. Michael D. Grant, Mr Jacques Naudé and Mr Ian McKechnie for their support and inspiration.

Thank you to my colleagues at the School of Electrical and Information Engineering, University of Witwatersrand as well as the staff of the Genmin and High Voltage Laboratories. A special thanks to Professor Paul Fatti for his knowledge and assistance with Bayesian probability.

The following organisations are gratefully acknowledged for funding and supporting this research and Lightning/EMC Research Group:

- The National Research Foundation (NRF) of South Africa (UNIQUE GRANT NO: 98244).
- Eskom through the Tertiary Education Support Programme (TESP).
- The Department of Trade and Industry (DTI) for Technology and Human Resources for Industry Programme (THRIP) funding.

Thanks are extended to the South African Weather Service for the use/assistance with the South African Lightning Detection Network (SALDN) data. In particular, Andrew van der Merwe, Morné Gijben, Nico Kroese, Michelle Hartslief and Mr Richard Evert of Eskom.

Thank you to Dr Gerhard Diendorfer, Head of the Austrian Lightning Detection and Information System (ALDIS) for providing the Gaisberg tower cases and the correlated ALDIS data. A special thanks to Dr Wolfgang Schulz of ALDIS for his guidance and assistance with lightning location system operations.

A special thanks to Dr Kleber Naccarato and Dr Marcelo Saba of the Brazilian Institute for Space Research for the unique opportunity to work on the RINDAT system and the *Natural Lightning to Common Buildings in São Paulo* Project. Thanks are extended to the ELAT group - Amanda, Diogo, Diovane, Luis and Raphael.

Finally, thank you to my friends and family. Your support, encouragement and much needed moments of distraction made this possible. To my wife Carina, for her invaluable assistance, unending patience and love and support - *Obrigado, eu te amo.*

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

Contents iv

List of Figures ix

List of Tables xiv

List of Symbols and Operators xvi

Nomenclature xvii

1 Introduction 1

1.1 Structure of Thesis 2

2 Approach Taken 5

2.1 Problem Description 5

2.2 Proposed Solution 7

2.3 Research Questions 8

2.4 Approach Taken 8

2.5 Scope of Thesis 9

2.6 Contribution of Thesis 9

2.7 Chapter Summary 10

3 Lightning Location Systems **11**

3.1 Overview 11

3.2 Detecting and Locating Lightning Events 12

3.2.1 The Lightning Flash 12

3.2.2 Lightning Localization 14

3.3 Lightning Location Systems 18

3.4 Lightning Location System Performance 19

3.4.1 Detection Efficiency 20

3.4.2 Location Accuracy 20

3.4.3 Ground-Truth Validation 22

3.5 Applications of LLS Data 23

3.5.1 Ground-Flash Density Maps 23

3.5.2 Forensic Investigations Involving LLS Data 25

3.6 Chapter Summary 26

4 Probability and Bayes' in a Forensic Context **28**

4.1 Overview 28

4.2 Bayesian Probability Theory 29

4.2.1 Sample Spaces and Events 29

4.2.2 Probability 29

4.2.3 Bayes' Theorem 30

4.2.4 Bayesian Inference 30

4.2.5 Bayesianism vs Frequentism 33

4.2.6 Decision Theory and Bayesian Networks 34

4.3 Bayes' in a Forensic Context 35

4.3.1 Bayes' Theorem in a Forensic context 36

4.3.2 Prior Probability in a Forensic context 38

4.3.3 Forensic Investigations with Bayesian Framework 38

4.4 Multivariate Statistics 38

4.4.1 Some Multivariate Distributions 39
4.4.2 Multivariate Maximum Likelihood Estimation (MLE) 43
4.4.3 Multivariate Goodness-of-fit Measures 45
4.5 Chapter Summary 45

5 Proposed Bayesian Framework **46**
5.1 Overview 46
5.2 Proposed Bayesian Framework 47
5.2.1 Assumptions and Scope 49
5.3 Likelihood Function $\boldsymbol{f(X_{lls_j}|L_{i_S})}$ 50
5.4 Prior Probability Function $\boldsymbol{p(L_{i_S})}$ 52
5.4.1 Non-informative (flat) prior 53
5.4.2 Informative prior 54
5.5 *NOT* Likelihood Function $\boldsymbol{f(X_{lls_j}|\overline{L_{i_S}})}$ 55
5.6 Applying the Framework 56
5.7 Chapter Summary 58

6 Bivariate Statistical Modelling **59**
6.1 Overview 59
6.2 LLS Data set $\boldsymbol{X}$ 60
6.3 Bivariate Visualization 60
6.4 Maximum Likelihood Estimation (MLE) 61
6.5 Goodness of Fit - Mahalanobis test 65
6.6 Marginal Visual Analysis (optional) 68
6.7 Chapter Summary 69

7 Ground-Truth Verification of Statistical Models **71**
7.1 Overview 71
7.2 South Africa 72
7.2.1 Brixton Tower Photographic Study 72

7.2.2 SALDN Data . . . 74

7.2.3 Bivariate Statistical Modelling: $\boldsymbol{SA_{Br}}$ . . . 75

7.2.4 Bivariate Statistical Modelling: $\boldsymbol{SA_{Jhb}}$ . . . 82

7.3 Austria . . . 88

7.3.1 Gaisberg Current Measurement . . . 88

7.3.2 ALDIS/EUCLID Data . . . 90

7.3.3 Bivariate Statistical Modelling: $\boldsymbol{Aus_{GBT}}$ . . . 90

7.4 Conclusions . . . 95

7.5 Chapter Summary . . . 97

8 Application of Bayesian Framework 99

8.1 Overview . . . 99

8.2 Forensic Investigation - Johannesburg, South Africa . . . 100

8.2.1 Applying the Bayesian Framework . . . 101

8.3 Ground-truth Comparison . . . 104

8.3.1 Log-likelihood Ratio Results . . . 104

8.3.2 Posterior Probability Results . . . 107

8.3.3 Conclusions . . . 108

8.4 Chapter Summary . . . 108

9 Conclusion 110

9.1 Recommendations for Future Research . . . 111

Appendix A *Sufficiently Small*, Discrete vs Continuous 113

A.1 Introduction . . . 113

A.2 Defining *Sufficiently Small* . . . 114

A.3 Univariate Bayesian Framework Example . . . 114

A.4 Conclusions . . . 118

Appendix B Lightning Events to Brixton Tower 2015-2016 **120**

B.1 Introduction 120

B.2 Photographed Flashes 120

B.3 Time-Correlated SALDN Data 122

B.4 Conclusion 123

Appendix C Algorithms and Source Code Implementation **125**

C.1 Maximum Likelihood Estimation (MLE) - Source Code 125

C.2 Mahalanobis Test - Source Code 127

C.3 Bayesian Framework - Source Code 130

References **134**

Bibliography **141**

List of Figures

2.1 Cases in which lightning location data indicates a location was (a) very likely struck (Building 1) and that locations were (b) very likely *NOT* struck. 6

2.2 Case in which lightning location data does not give a clear indication as to whether a location was struck (a), even with self-evaluation confidence ellipses (b). 6

2.3 Clustering of stroke reports making up a single flash. The number of ground termination points is unknown. 7

3.1 Lightning process for a downward negative flash showing the stepped leader (a-b), attachment and return stroke (c), continuing current (d), charge dissipation (e) and subsequent stroke (f-i). Adapted from Uman (1987). 13

3.2 Fast electric field measurement of downward negative lightning flash measured in Johannesburg, South Africa on 5 February 2017. Individual current pulses or "strokes" are visible. 14

3.3 Lightning process for an upward flash showing the upward propagating leader (a), initial continuing current (b-c), charge dissipation (d) and subsequent stroke (e-h). Adapted from Uman (1987). 15

3.4 Schematic demonstrating the principle of Magnetic Direction Finding. Adapted from Cummins and Murphy (1998). 16

3.5 Schematic demonstrating the principle of Time-of-Arrival (TOA). Adapted from Cummins and Murphy (1998). 17

3.6 Location of sensors in the EUCLID network as of 2015. Austria and the ALDIS network sensor locations indicated in red (Diendorfer, 2016). Image used with permission, courtesy of Dr W. Schulz. 18

3.7 Location of sensors for the South African Lightning Detection Network (SALDN). Image used with permission, courtesy of Mr Morné Gijben. . . 19

3.8 Gaussian distribution from where confidence ellipses are derived. Note the location error between the true stroke location and the computed stroke location. 21

3.9 Resolved LLS stroke location with 50% confidence ellipse shown. The stroke was reported in the region of the University of the Witwatersrand, Johannesburg, South Africa. 21

3.10 Ground-flash density maps of South Africa (a) CSIR map created with lightning counter data (b) SAWS map created with SALDN data (Gijben, 2012). 23

3.11 Gaussian smoothed map over the University of the Witwatersrand, Johannesburg, South Africa utilizing ten years of SALDN data. 25

4.1 Probability density $f(y)$ of random variable Y. 32

4.2 Example of a Bayesian network where multiple conditional dependencies are show, $P(A)$, $P(B|A)$, $P(D|B)$, $P(C|A,E)$, $P(E|B)$ and $P(F|C,D,E)$. 34

4.3 Likelihood ratio between two claims H_1 and H_2 with measured evidence, ϵ_1. Adapted from Aitken et al. (2010). 37

4.4 Example of a bivariate (k=2) Gaussian probability density function with $\boldsymbol{\mu} = (0,0), \sigma_{x_1} = \sigma_{x_2} = 1$ and $\rho_{x_1x_2} = 0$. 40

4.5 Univariate Gaussian, Students' t- ($\nu = 2$) and Cauchy distribution, $\mu = 0, \sigma = 1$. 42

5.1 Defined geographic area of investigation with parameters shown. The location/s of the event/s under investigation are to be found within this area. 47

5.2 Example of bivariate Students' t-distribution as the likelihood function - this describes the distribution of LLS reported stroke locations when location $\boldsymbol{L_i}$ is struck by lightning. 52

5.3 Bivariate uniform distribution for prior probability function as (a) probability density function $f(\boldsymbol{L_{i_S}})$, (b) probability mass function $p(\boldsymbol{L_{i_S}})$. . . 54

5.4 Presenting the results of the Bayesian framework (a) example of the log-likelihood ratio $\ln(LR_{jL_1-L_2})$ vs. t_j and (b) example of the posterior probability $p(\boldsymbol{L_{i_S}}|\boldsymbol{X_{lls_j}})$ vs. t_j. 57

6.1 Block diagram describing the methodology used to determine the statistical model of a LLS data set. 59

6.2 Modified EM algorithm to determine best fitting Students' t-distribution parameters for data set $\boldsymbol{X}$. 64

6.3 Block diagram showing the Mahalanobis goodness-of-fit test procedure. . 66

6.4 Example of Cumulative probability vs. Mahalanobis distances for different statistical models with parameters fitted to data set $\boldsymbol{X}$. $\boldsymbol{X}$ is data from a bivariate Students' t-distribution with $\boldsymbol{\mu} = \binom{0}{0}$, $\boldsymbol{\Sigma} = I$ and $\nu = 2$. 68

6.5 Example of marginal visual analysis with different statistical models fitted to (a) probability density estimate of $\boldsymbol{x}$ (b) Empirical cumulative density plot of $\boldsymbol{x}$. $\boldsymbol{x}$ is data from a univariate Students' t-distribution with $\mu = 0$, $\sigma = 1$ and $\nu = 2$. 69

7.1 The Johannesburg skyline, as viewed from a northern location looking south. Both the Hillbrow and the Brixton tower are visible. 72

7.2 Motion-triggered camera surveillance of Brixton tower (a) view of Brixton tower as seen by the camera system (b) location of camera relative to the location of Brixton tower. 73

7.3 Examples of upward lightning events from the Brixton tower captured by the surveillance camera system on 20-11-2015 at (a) 09:43:24 SAST (b) 10:24:48 SAST (c) 10:54:37 SAST 74

7.4 Aerial view of latitude vs. longitude for SALDN strokes time-correlated with events to the Brixton tower $\boldsymbol{SA_{Br}}$.The Brixton tower is located at $(28.0068°, -26.1925°)$. The marginal histograms in the latitudinal and longitudinal directions are also shown. 76

7.5 Bivariate visualization of data set $\boldsymbol{SA_{Br}}$: (a-b) Histogram count. (c-d) Probability density estimate. (e-f) Empirical cumulative density. All plots with bin size $0.0014° \times 0.0014°$. 78

7.6 Negative log-likelihood vs. degrees of freedom parameter for bivariate Students' t-distribution, ν_{Br}. When $\nu_{Br} = 1$, the fitted parameters are equivalent to a bivariate Cauchy distribution and as $\nu_{Br} \to inf$, the fitted parameters are equivalent to those of the bivariate Gaussian distribution. 79

7.7 Cumulative probability vs. Mahalanobis distance for the different statistical models compared with the empirical Mahalanobis distance of $\boldsymbol{SA_{Br}}$. 80

7.8 The **four** statistical models fitted to the marginal distributions of $\boldsymbol{SA_{Br}}$. 81

7.9 Aerial view of latitude vs. longitude for SALDN strokes in the Johannesburg region, $\boldsymbol{SA_{Jhb}}$. The marginal histograms in the latitudinal and longitudinal directions are also shown. 82

7.10 Bivariate visualization of data set $\boldsymbol{SA_{Jhb}}$: (a-b) Histogram count. (c-d) Probability density estimate. (e-f) Empirical cumulative density. All plots with bin size $0.0081° \times 0.0081°$. 84

7.11 Probability density function of the Gaussian Mixture Model (GMM) fitted to $\boldsymbol{SA_{Jhb}}$. The high density locations of Brixton tower and Hillbrow tower are clear in the plot. 85

7.12 Cumulative Mahalanobis distances for the different statistical models compared with the empirical Mahalanobis distance of $\boldsymbol{SA_{Jhb}}$ a) full curve b) enhanced section. 86

7.13 The **four** statistical models fitted to the marginal distributions of $\boldsymbol{SA_{Jhb}}$. 87

7.14 The Gaisberg radio tower, (a) on top of Gaisberg mountain, overlooking the town of Salzburg in the North of Austria and (b) its location, 47.8053° North and 13.1121° East. Images used with permission, courtesy Dr. Wolfgang Schulz 88

7.15 The current measurement equipment at Gaisberg tower: (a) the resistive shunt installation at the top of Gaisberg tower and (b) example of current wave-shape that is measured during lightning attachment. Images used with permission, courtesy Dr. Wolfgang Schulz 89

7.16 Aerial view of latitude vs. longitude for ALDIS strokes time-correlated with events to the Gaisberg tower, $\boldsymbol{Aus_{GBT}}$. The Gaisberg tower is located at $(13.1121°, 47.8052°)$. The marginal histograms in the latitudinal and longitudinal directions are also shown. 91

7.17 Bivariate visualization of data set $\boldsymbol{Aus_{GBT}}$: (a-b) Histogram count. (c-d) Probability density estimate. (e-f) Empirical cumulative density. All plots with bin size 0.0068° × 0.0068°. 93

7.18 Negative log-likelihood vs. degrees of freedom parameter for the bivariate Students' t-distribution, ν_{GBT}. When $\nu_{GBT} = 1$, the fitted parameters are equivalent to a Cauchy distribution and as $\nu_{GBT} \to \infty$, the fitted parameters are equivalent to those of the Gaussian distribution. 94

7.19 Cumulative probability vs. Mahalanobis distance for the different statistical models compared with the empirical Mahalanobis distance of $\boldsymbol{Aus_{GBT}}$. 94

7.20 The **four** statistical models fitted to the marginal distributions of $\boldsymbol{Aus_{GBT}}$. 96

8.1 GMM prior probability function density plot based on $\boldsymbol{X_{llsPr_q}}$ SALDN data from 2009-2013 thunderstorm seasons in Johannesburg, South Africa. The two high density points correspond with the locations of the Brixton and Hillbrow towers. 103

8.2 Log-likelihood Ratio $\ln(LR_{jBr-\overline{Br}})$ vs. t_j (Year-Month-Day HH:MM:SS UTC) with likelihood function set as (a) bivariate Gaussian distribution and (b) bivariate Students' t-distribution. 105

8.3 Log-likelihood Ratio $\ln(LR_{jBr-Hlbrw})$ vs. t_j (Year-Month-Day HH:MM:SS UTC) with likelihood function set as (a) bivariate Gaussian distribution and (b) bivariate Students' t-distribution. 106

8.4 Posterior probability $p(\boldsymbol{L_{Br_S}}|\boldsymbol{X_{lls}})$ vs. t_j (Year-Month-Day HH:MM:SS UTC) with (a) with Non-informative prior and likelihood function set as bivariate Gaussian distribution and (b) bivariate Students' t-distribution and (c) Informative prior with likelihood function set as bivariate Gaussian distribution and (d) bivariate Students' t-distribution. 107

A.1 Probability density $f(y)$ of random variable Y. 114

A.2 Probability of a random variable Y being in the range $\Delta y = [c, d]$, versus Δy. The two curves show the calculation using the integral $\int_c^d f(y)dy$ and the approximation $f(y)\Delta y$ 115

A.3 Uniform prior probability of all possible locations that may be struck by lightning. 115

A.4 Likelihood functions describing the distribution of LLS stroke reports when (a) location $\boldsymbol{L_5}$ is struck by lightning (b) locations $\boldsymbol{L_{1-10}}$ are struck by lightning. The stroke report $X_{lls} = 5.7$ is indicated by the blue, vertical line. 116

A.5 Posterior probability $p(\boldsymbol{L_{i_S}}|X_{lls} = 5.7)$ for each location $\boldsymbol{L_{1-10}}$. 117

A.6 Probability of being struck by lightning against locations $\boldsymbol{L}$ (a) non-uniform (informative) prior probability $p(\boldsymbol{L_{i_S}})$ (b) posterior probability $p(\boldsymbol{L_{i_S}}|X_{lls} = 5.7)$ for each location $\boldsymbol{L_{1-10}}$ with non-uniform prior probability function. 117

A.7 Posterior probability $p(\boldsymbol{L_{i_S}}|X_{lls} = 5.7)$ for each location $\boldsymbol{L_{1-40}}$ with (a) uniform prior probability function (b) with non-uniform prior probability function. 118

List of Tables

4.1 Confusion matrix for classification process assessment (Fawcett, 2005). . 35

4.2 Qualitative interpretation of LR as suggested by Kaye (2012) 37

5.1 Parameters for example bivariate Students' t-distribution as likelihood function. 51

7.1 Photographed Lightning Flashes to Brixton Tower 74

7.2 SALDN stroke reports within 1600 km^2 area around Johannesburg: $\boldsymbol{SA_{Jhb}}$ 75

7.3 Time-correlated SALDN data with photographed flashes that attached Brixton tower: $\boldsymbol{SA_{Br}}$ 75

7.4 Parameters for the statistical models fitted to $\boldsymbol{SA_{Br}}$ using MLE. 77

7.5 Mahalanobis test results for $\boldsymbol{SA_{Br}}$. Lowest values indicate best fit. . . . 80

7.6 Bivariate MLE fitted parameters for SALDN data in Johannesburg region $\boldsymbol{SA_{Jhb}}$ 83

7.7 Mahalanobis test results for $\boldsymbol{SA_{Jhb}}$. Lowest values indicate best quality of fit. 85

7.8 Number of lightning current measurements at the Gaisberg tower, Austria per year (Diendorfer et al., 2011a,b). 89

7.9 ALDIS data time-correlated with lightning attachments to the Gaisberg tower, $\boldsymbol{Aus_{GBT}}$ 90

7.10 Bivariate MLE fitted parameters fitted to $\boldsymbol{Aus_{GBT}}$ 92

7.11 Mahalanobis test results for $\boldsymbol{Aus_{GBT}}$. Lowest values indicate best fit. . 95

7.12 Mahalanobis test results for $\boldsymbol{SA_{Br}}$, $\boldsymbol{SA_{Jhb}}$ and $\boldsymbol{Aus_{GBT}}$. Lowest values indicate best fit. 95

8.1 Parameters of defined geographic area for the Brixton and Hillbrow tower investigation. 100

8.2 Parameters of defined time-frame for the Brixton and Hillbrow tower investigation . . . 101

8.3 Parameters of bivariate Gaussian likelihood function for forensic investigation, Brixton and Hillbrow tower. . . . 102

8.4 Parameters of bivariate Students' t-likelihood function for forensic investigation, Brixton and Hillbrow tower. . . . 103

8.5 Parameters of bivariate uniform (non-informative) prior probability function for forensic investigation, Brixton and Hillbrow tower. . . . 103

8.6 Parameters of bivariate GMM (informative) prior probability function for forensic investigation, Brixton and Hillbrow tower. . . . 103

8.7 Confusion matrix for Brixton tower. . . . 104

8.8 Ground-truth comparison of bivariate Gaussian and Students' t-models as likelihood functions. . . . 105

B.1 Date and time of photographed flashes 2015 - 2016. . . . 120

B.2 SALDN data time-correlated with photographed flashes 2015 - 2016. . . . 122

List of Symbols and Operators

The principal symbols and operators used in this thesis are summarised below. The first equation in which the symbol appears is given. Bold symbols are used to denote vectors and matrices throughout the thesis.

α	-	Azimuthal angle (equation 3.1).
σ	-	Standard deviation of statistical distribution (equation 3.1).
t	-	Time-stamp (YY-MM-DD HH:MM:SS) (equation 3.2).
φ	-	Longitudinal geographical coordinates (equation 3.4).
λ	-	Latitudinal geographical coordinates (equation 3.4).
θ	-	Parameter or parameters to be inferred (equation 4.6).
x	-	Measured data from an experiment (equation 4.6).
$p(x)$	-	Probability mass function (equation 4.7).
$f(x)$	-	Probability density function (equation 4.10).
H	-	Claim or proposition (equation 4.15).
ϵ	-	Obtained or measured evidence (equation 4.15).
$\boldsymbol{\mu}$	-	Mean of multivariate distribution (equation 4.18).
$\boldsymbol{\Sigma}$	-	Co-variance matrix of multivariate distribution (equation 4.18).
$\boldsymbol{\Delta}$	-	The Mahalanobis distance of multivariate distribution (equation 4.18).
w_q	-	Weighting of Gaussian mixture component (equation 4.20).
$g(x)$	-	Gaussian probability density function (equation 4.20).
ν	-	The degrees of freedom (equation 4.22).
$\boldsymbol{\hat{\theta}_{mle}}$	-	Maximum likelihood estimators (equation 4.28).
$\mathcal{L}$	-	Likelihood (equation 4.26).
ρ	-	The correlation coefficient (equation 5.10).
Mhl_T	-	Mahalanobis test statistic (equation 6.16).

Nomenclature

ALDIS	Austrian Lightning Detection and Information System
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BrasilDAT	Detecção de Descargas Atmosféricas do Brasil
DE	Detection Efficiency
ECM	Expectation-Conditionalization-Maximization
EM	Expectation-Maximization
EUCLID	European Cooperation for Lightning Detection
FN	False Negative
FP	False Positive
GBT	The Gaisberg Tower
GMM	Gaussian Mixture Model
GPS	Global Positioning System
GSP	Ground-Strike Point
IEC	International Electrotechnical Commission
IMPACT	Improved Accuracy from Combined Technology
LA	Location Accuracy
LDN	Lightning Detection Network
LLS	Lightning Location System
LR	Likelihood Ratio
MDF	Magnetic Direction Finding
MLE	Maximum Likelihood Estimation
NTP	Network Time Protocol
RINDAT	Rede Integrada Nacional de Detecção de Descargas Atmosféricas
SALDN	South African Lightning Detection Network
SAST	South African Standard Time
TN	True Negative
TOA	Time-of-Arrival
TP	True Positive
U.S.NLDN	U.S. National Lightning Detection Network
UTC	Coordinated Universal Time

Chapter 1

Introduction

Lightning can cause significant damage to property and assets. In the United States of America, the Insurance Information Institute estimated approximately $825 million paid out in more than 100,000 lightning claims in 2016, with an average claim of $7,500.00 (Insurance Information Institute, 2017). In Germany, insurers reported 300,000 claims due to lightning in 2016, resulting in an estimated EURO 210 million in paid claims according to the German association of insurers (Der Gesamtverband der Deutschen Versicherungswirtschaft, 2017). In South Africa, it has been estimated that insurance claims due to lightning mount to more than 500 million Rand per year, with the South African Weather Service (SAWS) noting that almost all insurance queries are lightning related (Gijben, 2012). These claim numbers include electrical surges and damages to components that could be due to lightning but it is not always clear. Lightning events can also result in loss of life (both human and livestock). In South Africa, it is estimated that about 50 people are killed each year (Blumenthal et al., 2009). In the United States, 38 people were killed by lightning in 2016 (Insurance Information Institute, 2017). Determining whether lightning was the true cause of these events is important, both for the impact on insurance claims and subsequent legal issues (was the damage caused by negligence etc.), as well as for accurate statistics on lightning injury and deaths.

Modern Lightning Location Systems (LLSs), also known as Lightning Detection Networks (LDNs), have become an invaluable tool in this respect, providing evidence of lightning for forensic investigations into such claims. These networks are able to detect individual lightning strokes, and report the location at which they occurred with great accuracy (Cummins and Murphy, 1998, Nag et al., 2015). The STRIKEnet® system utilizes data from the U.S. National Lightning Detection Network (U.S.NLDN) to provide forensic reports for clients in the United States (*STRIKEnet®*, 2017). In Europe, insurers are able to contact the operators of the European Cooperation for Lightning Detection (EUCLID) network to obtain lightning occurrence reports and in South Africa, reports can be obtained from SAWS, who operate the South African Lightning Detection Network (SALDN).

LLSs triangulate the location of lightning events based on reports from multiple sensors. These sensors are antennas and detect changes in electromagnetic fields, reporting directional and timing information, or both (Cummins and Murphy, 1998). Any errors in these reports (due to a number of reasons) will translate into an error in the reported location of a lightning event (Schulz, 1997). While these errors tend to be relatively small, they can lead to uncertainty in forensic investigations which may have a bearing on legal outcomes as to whether lightning, or some other reason, was the cause of property/asset damage or loss of life. Work has been done to evaluate this uncertainty by calculating probabilities but these calculations are based on the assumed bivariate Gaussian distribution of location errors and do not allow for the inclusion of prior (or regional) information in the calculation of the probability (Grant et al., 2011, Huddleston et al., 2012).

This thesis introduces a framework, based on a Bayesian approach to presenting evidence in a legal forum, that can be applied to LLS location reports for forensic investigations where lightning is a possible cause of damage to property, assets or loss of life. The framework allows for LLS reports within the relevant time-frame, as well as previous or prior reports within a region, to be considered in the investigation. Using ground-truth lightning data events to tall towers in South Africa and Austria, and bivariate statistical techniques, the necessary statistical models for the Bayesian framework are verified. The conclusions of the investigations also add value to the concepts of location accuracy and reliability of LLS data, specifically in regards to the use of the bivariate Gaussian distribution for LLS location errors.

1.1 Structure of Thesis

The thesis begins with a detailed description of the problem and the research questions to be answered, along with the approach taken to answering these questions and the conclusions and contributions that arise. From here, the thesis is divided into three main components: the first component consists of Chapters 3 and 4, and provides background information and a summary of the literature on LLSs and Bayes' in a forensic context. The next component is covered by Chapters 5 and 6, where the proposed Bayesian framework and the methodology for fitting the statistical distributions to bivariate data are explained in detail. The final component consists of Chapters 7 and 8, in which the statistical fitting techniques are applied to the ground-truth LLS data sets obtained from tall tower studies and the Bayesian framework is applied to the Johannesburg region in South Africa, as a case study, to demonstrate its use and the effect of different statistical models. The final conclusions of the thesis are then summarised in Chapter 9. After the conclusion there are four appendices, providing additional detail on the Bayesian framework, Brixton tower photographs and source code for the algorithms used.

The individual chapters and appendices are as follows:

Chapter 2 - Approach Taken: This chapter describes and defines the problem addressed by the thesis in detail. The choice of the Bayesian approach to forensic investigation as a proposed solution is introduced and the research questions and scope are defined. The approach taken to answer the research questions is described and the unique contributions of the thesis are presented.

Chapter 3 - Lightning Location Systems: This chapter provides a brief overview of the operation of Lightning Location Systems, with a specific focus on the methods of lightning localization and location errors. The South African Lightning Detection Network and Austrian Lightning Detection and Information System (from which data is used in this thesis) are discussed. The current research on the application of LLS data in ground-flash density maps and forensic investigations is presented.

Chapter 4 - Probability and Bayes' in a Forensic Context: This chapter discusses the use of Bayesian probability concepts in forensic investigation. The origins of Bayesian probability are presented with an explanation of Bayesian inference. The interpretation and application of this in forensic situations and the presenting of evidence in a legal forum is discussed. Some concepts of multivariate statistics are also included.

Chapter 5 - Proposed Bayesian Framework: This chapter presents the proposed Bayesian framework for forensic investigations utilizing Lightning Location System (LLS) stroke reports. The chapter describes the requirements for applying the framework in an investigation along with the relevant equations. The statistical models necessary for use of the framework are discussed in detail and a recommendation for presenting the results of an investigation is provided.

Chapter 6 - Bivariate Statistical Modelling: This chapter describes a methodology for fitting statistical models to bivariate data (eg. Lightning Location System (LLS) stroke locations) and evaluating the quality of fit of the model. Fitting of models is done using bivariate Maximum Likelihood Estimation (MLE) and Expectation-Maximization (EM) techniques. The quality of each fit is then assessed by a difference of squares comparison of the Mahalanobis distances, given the bivariate nature of the data set.

Chapter 7 - Ground-Truth Verification of Statistical Models: This chapter discusses the ground-truth data sets used to verify the proposed likelihood function and prior probability function of the Bayesian framework. LLS stroke reports are time-correlated with ground-truth lightning events to a tall tower in South Africa - the Brixton tower and in Austria - the Gaisberg tower. The methodology described in Chapter 6 is applied to these LLS data sets and the results of which statistical models best fit the LLS data are presented.

Chapter 8 - Application of Bayesian Framework: In this chapter, the proposed Bayesian framework described in Chapter 5 is applied to a case study using the statistical models found in Chapter 7 as the likelihood and prior probability functions. The case study is framed as a forensic investigation around the Brixton and Hillbrow towers in Johannesburg, South Africa, with the investigation aiming to determine whether the Brixton tower was struck between October 2015 - March 2016. Utilizing the ground-truth data sets from Chapter 7, it is possible to evaluate the Bayesian framework and the different statistical models.

Chapter 9 - Conclusion: The findings of the thesis are summarised and the contributions are presented. Recommendations for further research are made.

Appendix A - *Sufficiently Small*, Discrete vs Continuous: This appendix discusses the definition of "sufficiently small" in regards to Bayesian inference and the distinction between continuous and discrete variables used in the Bayesian framework. A simple hypothetical, univariate example of the Bayesian framework is also presented.

Appendix B - Lightning Events to Brixton Tower 2015-2016: This appendix provides a detailed breakdown of all photographed flashes to the Brixton tower and the exact date and time they occurred. during the 2015-2016 season. The number of time-correlated SALDN strokes for each event are also provided.

Appendix C - Algorithms and Source Code Implementation: This appendix provides the source code for the algorithms utilized in the bivariate statistical modelling methodology discussed in Chapter 6 and the proposed Bayesian framework in Chapter 5 and 8.

Chapter 2

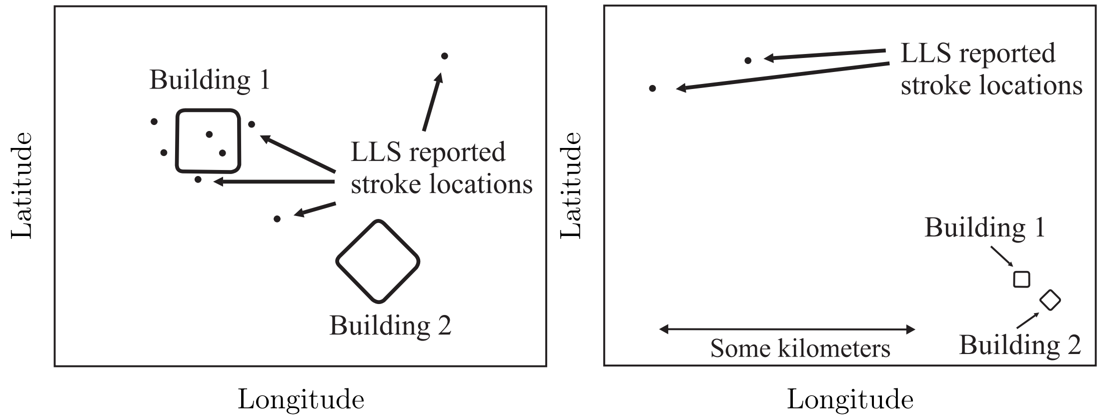
Approach Taken

This chapter describes and defines the problem addressed by the thesis in detail. The choice of the Bayesian approach to forensic investigation as a proposed solution is introduced and the research questions and scope are defined. The approach taken to answer the research questions is described and the unique contributions of the thesis are presented.

2.1 Problem Description

Figure 2.1 shows two examples of what a forensic report from a Lightning Location System (LLS) operator would look like. A geographic plot, with the reported locations of detected strokes shown by their reported latitude and longitude. In this hypothetical scenario, two buildings are also indicated on the figure and, one can imagine an investigation is underway to determine which of the buildings was struck by lightning. Figure 2.1a represents a report where it is intuitively clear what has happened - the LLS has detected a flash to Building 1, consisting of 8 strokes. The LLS has detected each of these strokes and reported their locations but, due to the slight errors present, they are not all reported at exactly the same location, even though the real event must have attached to a single location. The occurrence of these errors is discussed in more depth in Chapter 3. Nonetheless, the clustering makes it intuitively clear that Building 1 was struck. Similarly, the situation in figure 2.1b is intuitively clear that neither Building 1 nor Building 2 was struck, since the detected strokes were reported kilometers away.

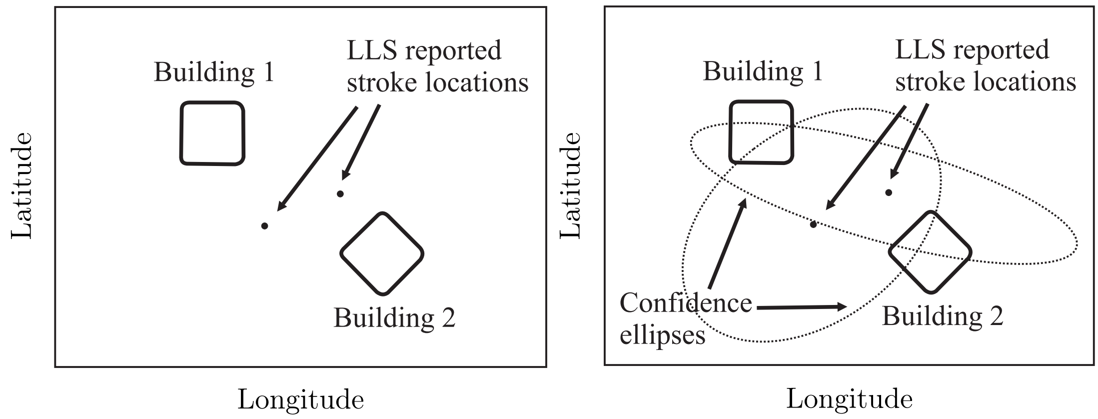
Both of these cases do have uncertainty and it cannot be said with absolute certainty as to what happened (due to the errors that occur in LLS reports). The problem becomes more pronounced (and unintuitive) in the scenario shown in figure 2.2. Here, two strokes are reported almost equally distant from both buildings. It is not possible to intuitively make a decision on which building was struck by lightning or if either building was struck. Even with the self-evaluation location accuracy techniques (confidence ellipses, figure 2.2b), it is still not possible to make an intuitive inference on what events actually occurred.



(a) Very likely Building 1 struck

(b) Very likely *NOT* struck

Figure 2.1: Cases in which lightning location data indicates a location was (a) very likely struck (Building 1) and that locations were (b) very likely *NOT* struck.



(a) Unintuitive case

(b) Unintuitive case with confidence ellipses

Figure 2.2: Case in which lightning location data does not give a clear indication as to whether a location was struck (a), even with self-evaluation confidence ellipses (b).

The problem can also be described in the context of ground-strike points. A lightning flash, consisting of multiple strokes, may have multiple terminations to the earth's surface (Bouquegneau, 2014, Cummins, 2014). As mentioned before, due to the errors present, a reported flash will look like a cluster of strokes, like in figure 2.3. Knowing whether a stroke, reported slightly further from the cluster, is a new termination point or if it is just a location error is not clear from the figure.

While an intuitive interpretation of these reports may be effective in the more extreme cases, many scenarios are not as easily interpreted and quantifying the degree of uncertainty is needed for the forensic reporting of evidence and decision making ie. the outcome of an insurance investigation. While there are proposed methods for calculating probabilities, they rely heavily on the self-evaluation confidence ellipses and the assumption that the location errors follow a bivariate Gaussian distribution (due to the random occurrence of measurement errors) (Grant et al., 2011, Huddleston et al., 2012,

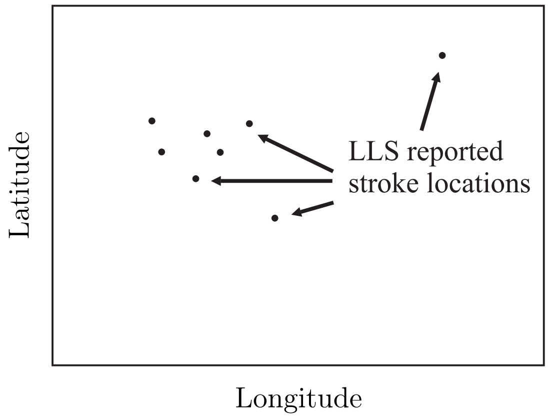


Figure 2.3: Clustering of stroke reports making up a single flash. The number of ground termination points is unknown.

Hunt et al., 2017). Also, they provide no consistent approach to incorporating prior knowledge of the situation/region or how to report the probabilities as evidence.

2.2 Proposed Solution

The proposed solution is to use a Bayesian probabilistic framework to assess forensic investigations involving LLS stroke reports. A Bayesian approach to assessing and reporting evidence in forensic investigations and legal forums is becoming increasingly preferred today (Aitken et al., 2010). The details of this approach are discussed later in Chapter 4 and Chapter 5 and are essentially based on Bayes' formula. Applying the Bayesian approach to a lightning investigation, Bayes' formula becomes:

$$P(\boldsymbol{L_S}|\boldsymbol{X_{lls}}) = \frac{P(\boldsymbol{L_S})P(\boldsymbol{X_{lls}}|\boldsymbol{L_s})}{P(\boldsymbol{X_{lls}})} \tag{2.1}$$

where $\boldsymbol{L_S}$ is the proposition that a specific geographic location was struck by lightning and $\boldsymbol{X_{lls}}$ is the LLS data (or evidence) detected in the time-frame of investigation. Bayes' formula allows us to calculate the probability that the specific geographic location was struck, given the evidence that was obtained, using the likelihood function $P(\boldsymbol{X_{lls}}|\boldsymbol{L_S})$, which is the probability of obtaining the evidence if the location was truly struck, and the prior probability function $P(\boldsymbol{L_S})$, which is the probability of the location having been struck in the past. The results can then be reported as odds or a posterior likelihood ratio:

$$\frac{P(\boldsymbol{L_S}|\boldsymbol{X_{lls}})}{P(\boldsymbol{L_{NS}}|\boldsymbol{X_{lls}})} = \frac{P(\boldsymbol{L_S})}{P(\boldsymbol{L_{NS}})} \times \underbrace{\frac{P(\boldsymbol{X_{lls}}|\boldsymbol{L_S})}{P(\boldsymbol{X_{lls}}|\boldsymbol{L_{NS}})}}_{LR} \tag{2.2}$$

where $\boldsymbol{LR}$ is the likelihood ratio - the ratio of the likelihood of different propositions ($\boldsymbol{L_S}$ or $\boldsymbol{L_{NS}}$) given the LLS data $\boldsymbol{X_{lls}}$, without any prior knowledge of the location under investigation. $\boldsymbol{L_{NS}}$ is the proposition that a specific geographic location was NOT struck by lightning. Theses two equations represent the main calculations of the Bayesian framework. The complete description of all terms and statistical models is discussed in detail in Chapter 5.

2.3 Research Questions

In order to design and apply such a framework, we need to understand the statistical models that describe the likelihood function, $P(\boldsymbol{X_{lls}}|\boldsymbol{L_S})$ and the prior probability function $P(\boldsymbol{L_S})$. The likelihood function is a description of how the LLS data would be distributed if the location was truly struck, and is therefore the distribution of location errors, usually assumed to follow a bivariate Gaussian distribution. The prior probability function would best be described by a historical data set of LLS strokes over an area that indicate which locations are struck more often than others - similar to a stroke or flash density map.

This leads to the following research questions:

1. What is the statistical model that best represents the distribution of LLS location errors when a specific geographic location is struck by lightning, $P(\boldsymbol{X_{lls}}|\boldsymbol{L_S})$? Can a better statistical model than the bivariate Gaussian distribution be found? (eg. bivariate Students' t-distribution).
2. What is the statistical model that best represents the distribution of locations struck by lightning in a area or region, $P(\boldsymbol{L_S})$?
3. What is the effect of different likelihood and prior probability functions on forensic investigations involving LLS data utilizing the Bayesian framework?

This thesis details the methodologies used to determine these statistical models and apply the Bayesian framework to lightning forensic investigations.

2.4 Approach Taken

The following approach is taken to answer the research questions:

1. The proposed Bayesian framework for utilizing LLS data in forensic investigations is explained, with the necessary definitions and assumptions being outlined. The proposed statistical models for the likelihood function - bivariate Gaussian, Students' t- and Cauchy distributions - are explained in terms of the framework and what their parameters mean as well as statistical models for the prior probability function - the bivariate uniform distribution (flat prior) and a Gaussian Mixture Model (GMM) (informative prior).

2. A methodology for modelling bivariate statistical distributions is developed in order to confirm if the distributions proposed for the likelihood function and prior probability function are good descriptions of real LLS data. The methodology uses Maximum Likelihood Estimation (MLE) and Expectation-Maximization (EM) techniques to fit the statistical models to LLS data, and a goodness-of-fit test based on the cumulative distribution of the Mahalanobis distance is used to determine the quality of fit.

3. The bivariate statistical modelling methodology is applied to LLS data from two countries - South Africa and Austria. To evaluate the statistical models proposed for the likelihood function, LLS strokes from the South African lightning Detection Network (SALDN) and the Austrian Lightning Detection and Information System (ALDIS) are time-correlated with ground-truth events to the Brixton tower, South Africa and the Gaisberg tower, Austria respectively, and are used to develop data sets of LLS location errors when a known location is repeatedly struck. For the statistical models of the prior probability functions, a data set of SALDN stroke reports for **five** thunderstorm seasons (2009-2013 and 2015-2016) in the Johannesburg region of South Africa is used.

4. The Bayesian framework is applied to a case study, the Brixton and Hillbrow tower in the Johannesburg region in South Africa, to demonstrate its use. Given the strokes time-correlated with events to the Brixton tower, it is possible to see the how the different proposed likelihood and prior probability functions effect the framework results.

2.5 Scope of Thesis

The scope of the thesis is limited to developing a probabilistic Bayesian framework that can be consistently applied in forensic investigations when LLS stroke reports are consulted. The statistical models that best represent the distribution of these reported stroke locations are key to formulating a Bayesian framework. The thesis does not address the framework needed when strokes are not detected. This would require understanding the statistical models of individual sensor measurements rather than resolved stroke reports and is a separate study. Similarly, the thesis does not address the framework needed for missclassification of cloud-to-ground events as cloud-to-cloud or vice-versa. The Bayesian framework described may be applied to any LLS data set of reported stroke locations, in any area for any period of investigation.

2.6 Contribution of Thesis

This thesis provides the following unique and significant contributions to the field of lightning location:

- A probabilistic Bayesian framework for utilizing LLS stroke reports in forensic investigations where lightning is a possible cause. The framework provides a method for reporting LLS data that is consistent with the reporting of evidence in legal forums - in the form of likelihood ratios and posterior probabilities, including prior information;

- Verification of the statistical models employed in the framework against ground-truth data from two different LLSs in two different countries (South Africa and Austria), and a demonstration of its application in a real-world case study;
- A methodology for fitting bivariate statistical models to LLS data and evaluating the quality of the fit of the models;
- The bivariate Students' t-distribution is a better fitting model for the location errors present in LLS reports than the more commonly assumed bivariate Gaussian distribution due to its heavier tails;
- Gaussian Mixture Models (GMMs) can be used to model the distribution of lightning over an area for prior probability calculations.

The contribution has a direct application in the use of LLS reports as evidence in forensic investigations. This is of critical importance for insurance claims, but also key in correctly determining statistics around lightning caused loss of life and damages.

The contribution can also play an important role in improving self-evaluation methods of LLSs and in clustering methods for ground-strike point algorithms.

2.7 Chapter Summary

This chapter has described the problem addressed by the thesis - the location errors present in LLS stroke reports can lead to uncertainty when LLS data is consulted in the context of a forensic investigation. The choice of a Bayesian approach to reporting this uncertainty has been made given its preference within legal forums as a method for reporting the strength of evidence. Given such a framework, the question as to which statistical models are best for the likelihood and prior probability functions is the focus of this research. The approach of using bivariate statistical modelling techniques applied to LLS strokes time-correlated with ground-truth lightning events is taken. The unique contributions of the thesis have been presented.

The following chapter provides a brief overview of LLSs and their operation.

Chapter 3

Lightning Location Systems

This chapter provides a brief overview of the operation of Lightning Location Systems, with a specific focus on the methods of lightning localization and location errors. The South African Lightning Detection Network and Austrian Lightning Detection and Information System (from which data is used in this thesis) are discussed. The current research on the application of LLS data in ground-flash density maps and forensic investigations is presented.

3.1 Overview

Lightning Location Systems (LLSs) or Lightning Detection Networks (LDNs) are networks consisting of a number antennas (referred to as sensors) placed over a region (eg. a country) that detect lightning events and report the time and location of lightning events (Cummins and Murphy, 1998). LLSs have greatly improved the ability of scientists and researchers to study the meteorological nature of thunderstorms and lightning occurrence as well as providing useful data to power utilities and the private sector such as the insurance industry. The chapter is divided into four sections:

3.2 Detecting and locating lightning events.

3.3 Lightning location systems.

3.4 Lightning location system performance.

3.5 Applications of LLS data.

The first section provides some brief background on lightning physics to understand detection and a description of the triangulation methods used to locate lightning events. Some examples of LLSs in different countries (specifically South Africa and Austria) are presented in the second section and evaluating the performance of LLSs is discussed in the next section. Finally, the use of LLS data in ground-flash density maps and forensic investigations is discussed.

3.2 Detecting and Locating Lightning Events

In order to detect lightning events, and to then locate them, electromagnetic models of lightning are used where a lightning event is described as a current pulse flowing through a vertical channel from the cloud to ground. For such an occurrence, there will a variation in the electromagnetic field, radiated out from this vertical conducting channel. It is possible to then measure either the electric field component or the magnetic field component of this radiated electromagnetic pulse and *"detect lightning"*. If multiple sensors are used, it then becomes possible to triangulate the location of this conducting vertical channel. An actual lightning event however, consists of a number of physical processes before and after current flows to ground and it is important to distinguish exactly what process is detected by this approach.

3.2.1 The Lightning Flash

A lightning event is referred to as a lightning flash and may consist of multiple, subsequent pulses of current flowing through the same channel to the same ground location. Lightning begins with charge separation occurring within a cloud (Cooray, 2014, Rakov and Uman, 2003, Uman, 1987). This leads to concentrated charge centers forming within the cloud. The concentration of charge leads to intensified electric fields between the charge centers which in turn, can lead to a bidirectional leader beginning within the cloud. One end of the leader propagates back into the cloud while the other end, of opposite polarity, propagates downward from the cloud. Figure 3.1 shows the (a) downward propagating negative leader as it leaves the cloud base. The branching effect that is represented is known as the stepped leader process and occurs as the progressing leader enhances the electric field at its tip. Once the downward propagating leader is close enough to the ground (b), an upward connecting leader begins moving upwards from the ground. Note that the positive end of the bidirectional leader is still propagating within the cloud.

The next step in the process (c) is known as attachment and is when the downward propagating leader meets the upward connecting leader and the lightning channel is formed. Here we have charge reconstitution, where all the free charge in the channel and leader branches flows back through the channel and is transferred to ground. This is the return stroke and is the first current transient or pulse to flow to ground. Some small current may flow for a short time after the return stroke (d) known as continuing current. The charge is quickly dissipated (e) but the air in the channel remains ionized. Another leader may then find this ionized channel and use it. When it does (f), this charge now has a path to ground and begins to flow. Another current transient or pulse (g) occurs in the channel and this is known as a subsequent stroke. This charge once again dissipates (h-i). This can occur many times with some cases having more than 25 subsequent strokes. This process is what is known as a downward negative lightning flash.

Downward positive lightning flashes also occur and share a similar process. However, the leader process is usually not stepped but rather the presence of recoil leaders is observed. Downward positive flashes also tend to have longer continuing currents but fewer subsequent strokes.

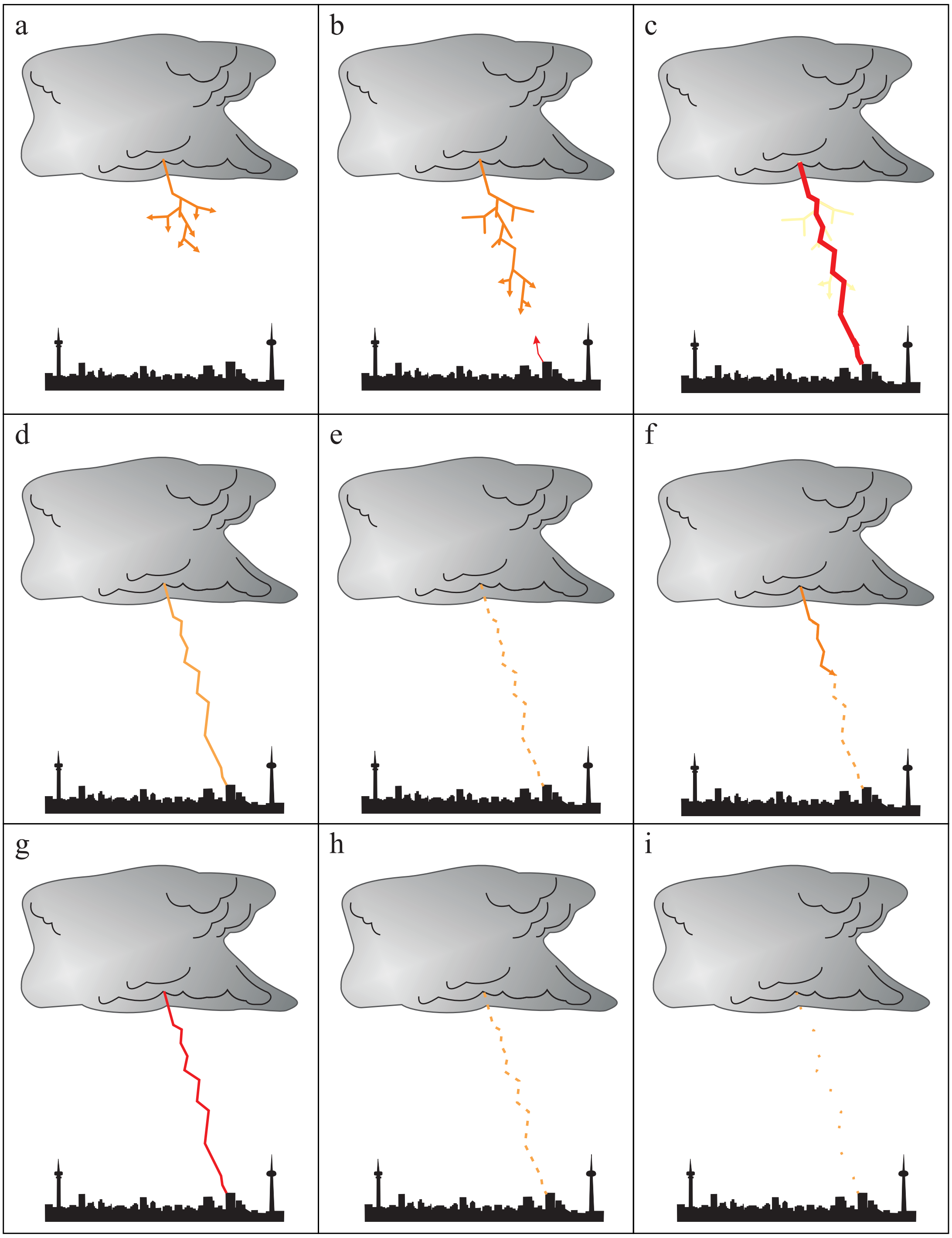


Figure 3.1: Lightning process for a downward negative flash showing the stepped leader (a-b), attachment and return stroke (c), continuing current (d), charge dissipation (e) and subsequent stroke (f-i). Adapted from Uman (1987).

If we make fast electric field or magnetic field measurements, we can see these current pulses or strokes as transients in the radiated fields. Figure 3.2 shows an example of a fast electric field waveform measured using a capacitive flat-plate antenna. The waveform was measured from a downward negative lightning event in Johannesburg, South Africa on 5 February 2017. The pulses (or strokes) can clearly be seen. It is this type of measurement that a LLS sensor will make and, with some threshold point set, detections of lightning strokes will be made.

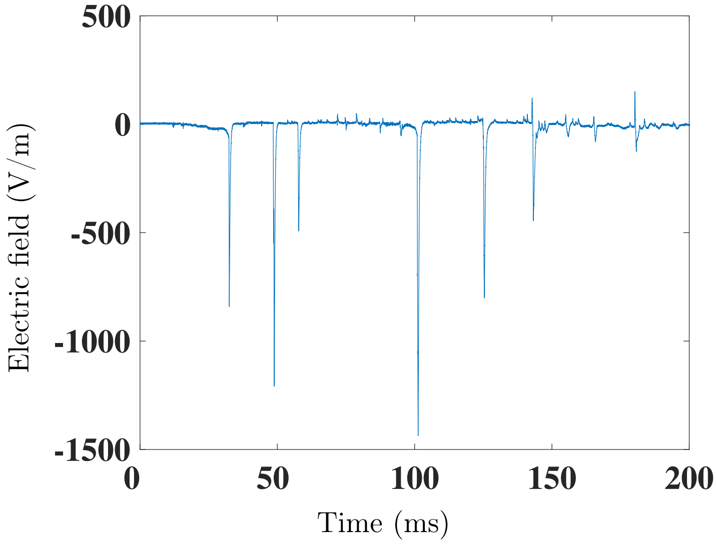


Figure 3.2: Fast electric field measurement of downward negative lightning flash measured in Johannesburg, South Africa on 5 February 2017. Individual current pulses or "strokes" are visible.

It is also possible to have an upward lightning flash, usually initiating from tall structures. Figure 3.3 shows the process for an upward flash. Here (a) the leader initiates from the point of high electric field intensity (usually the top of a tall structure) and propagates upward until it reaches the cloud (b). Current flows the whole time during this process but does not have the characteristics of defined transient like the downward flash - in other words, there is no return stroke but rather an initial continuing current. This current may last until (c) and then similar to the downward case, dissipate (d) but still leave an ionized channel. Once again, it is possible for charge centers to find this path (e) and for subsequent current to flow (f) ie. a subsequent stroke. The process (g-h) is then the same as in the downward case.

Upward lightning events can pose a problem for LLSs due to the slow rise of the initial continuing current rather than the quick transient of a return stroke. Given that some upward flashes only have initial continuing current, it is very likely that they will not be detected (Diendorfer et al., 2015).

3.2.2 Lightning Localization

Two main methods are used for locating lightning strokes (Cummins and Murphy, 1998, Diendorfer, 2007, Schulz, 1997):

- Magnetic direction finding (MDF)

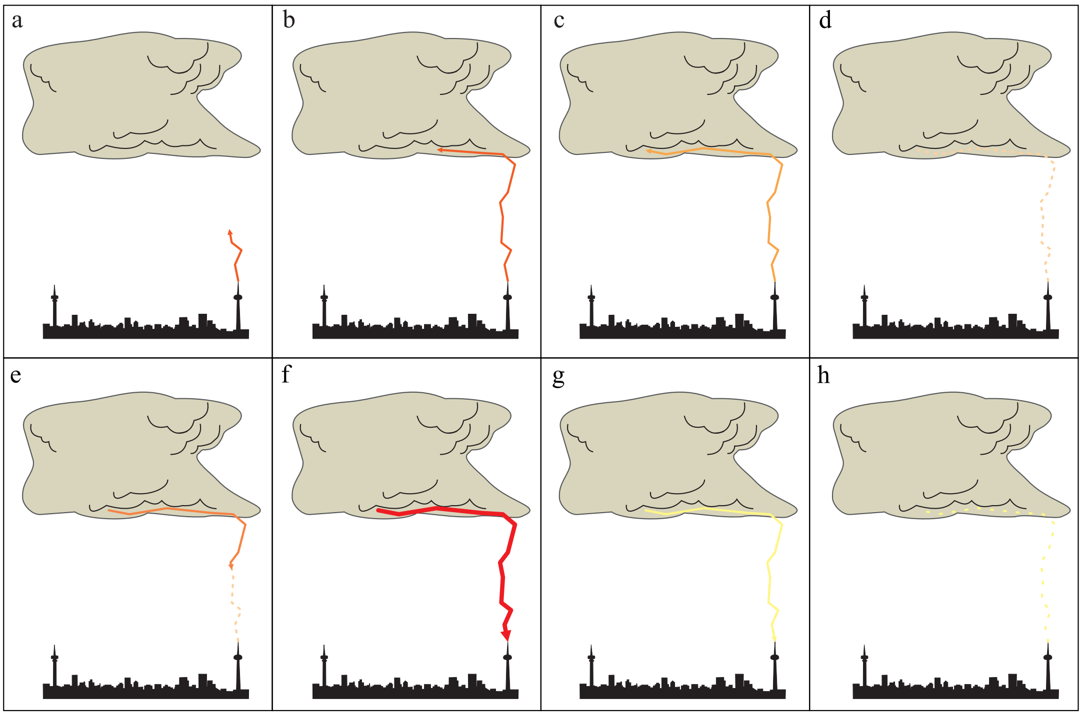


Figure 3.3: Lightning process for an upward flash showing the upward propagating leader (a), initial continuing current (b-c), charge dissipation (d) and subsequent stroke (e-h). Adapted from Uman (1987).

- Time-of-Arrival (TOA)

Many LLSs implement both of these techniques simultaneously in a method known as Improved Accuracy from Combined Technology (IMPACT) and information from both techniques is used in locating strokes.

Magnetic Direction Finding (MDF):

LLSs utilise two orthogonal loop antennas to detect the magnetic component of the radiated field from a lightning stroke. These antenna systems are wide-bandwidth so as to preserve the shape and polarity of the radiated field. A voltage proportional to the stroke magnetic field, multiplied by the cosine of the angle between the loop and the direction of the incoming field, is produced in each loop. The azimuth angle describing the direction of the lightning stroke discharge can then be determined by comparing the ratio of the two antenna outputs. When multiple sensors are used, the position of the lightning stroke can be inferred from the intersection of the reported directions (Schulz, 1997). Figure 3.4 shows a version of the commonly used schematic to demonstrate the concept.

Also indicated in the figure are the effects of errors in the measured azimuth angle from each sensor. Errors in the reported angle may arise due to a number of reasons: the non-vertical lightning channel, local noise in the area of a sensor and systematic

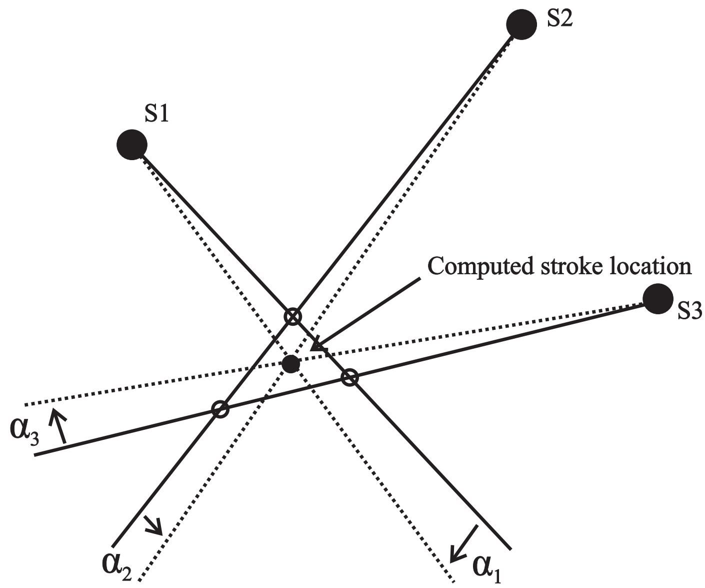


Figure 3.4: Schematic demonstrating the principle of Magnetic Direction Finding. Adapted from Cummins and Murphy (1998).

errors such as the calibration of a sensor or site errors. Systematic errors such as the site error can be corrected for but other errors fall into the category of random errors. The consequence of this is that different sensor reports do not all intersect at the same location as shown in figure 3.4.

In order to overcome this disagreement between sensor reports and compute a single location based on the angle information provided by multiple disagreeing sensors, the most probable or optimal position is determined by minimising the sum of squares of the reported angle measurements to find a "best fit" (equation 3.1) (Rakov and Uman, 2003).

$$\chi^2_{mdf} = \sum_{i=1}^{N_{sens}} \left(\frac{\alpha_i - \alpha_{ci}}{\sigma_{\alpha_{ci}}} \right)^2 \tag{3.1}$$

For a total number of sensors N_{sens}, α_i is the measured angle from the i-th sensor and α_{ci} is the unknown angle between each sensor and the computed stroke location. $\sigma_{\alpha_{ci}}$ is the measurement error estimate. This results in an optimal, resolved location and it is this that is reported as the stroke location by the LLS. The more sensors that partake in the solution, the more accurate it becomes.

Time-of-Arrival (TOA):

The Time-of-Arrival (TOA) technique determines the location of a lightning stroke based on the difference in time of detection at each sensor. The time of detection is the time the pulse crossed the threshold in figure 3.5. When multiple sensors report their detection times, there will be differences due to the different distances of each sensor from the location at which the stroke attached and therefore, the propagated fields arrive at each sensor at different times. If the difference in distance between the sensors is known, and the difference in the time of arrival is known, the location of the stroke can be inferred. Given two sensors and their reported times, there is only a certain range of locations that would allow the two sensors to report the times they did. If a third sensor is added, this reduces the number of possible locations enough to resolve a location. There is still some ambiguity but a fourth sensor removes this (Schulz, 1997). Figure 3.5 shows the commonly used representation of the TOA technique.

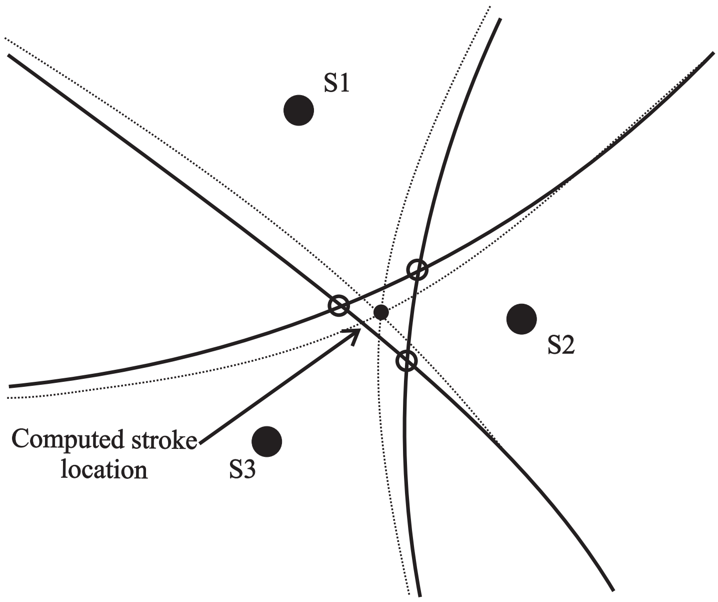


Figure 3.5: Schematic demonstrating the principle of Time-of-Arrival (TOA). Adapted from Cummins and Murphy (1998).

As with MDF, errors in arrival times may occur. In the case of TOA, any error in the time of detection of a stroke will correspond to an error in the time difference between the sensors and lead to disagreement in sensor reports. Similarly to MDF, a stroke position is resolved by minimising the sum of squares (equation 3.2) (Rakov and Uman, 2003).

$$\chi^2_{toa} = \sum_{i=1}^{N_{sens}} \left(\frac{t_i - t_{ci}}{\sigma_{t_{ci}}} \right)^2 \tag{3.2}$$

Where t_i is the measured time at the i-th sensor, t_{ci} is the unknown time difference between each sensor and the computed stroke location and $\sigma_{t_{ci}}$ is the measurement error estimate. As with MDF, the more sensors that participate in the solution, the more accurate it will become. The combination of both TOA and MDF techniques (equation 3.1 plus equation 3.2) allows for solutions even if only two sensors detect a stroke.

3.3 Lightning Location Systems

Many countries around the world operate LLSs to monitor lightning activity over their regions. In the United States, the U.S. National Lightning Detection Network (U.S.NLDN) consists of Vaisala inc. sensor technologies and has coverage over the entirety of North America, including part of Canada. In South America, the National Institute for Space Research operate two networks - RINDAT and BrasilDAT. One utilizing Vaisala Inc. technology (RINDAT) and the other operating with EarthNetworks® sensors. The European Cooperation for Lightning Detection (EUCLID) network operates throughout Europe and the South African Lightning Detection Network (SALDN) is the largest network operating in Africa. LLSs can also be found in China and Japan, amongst other countries.

The EUCLID network is a cooperation between a number of European countries with the goal of providing lightning data for all of Europe. Each member country of the EUCLID network runs their own independent LLS in their respective country. As a result, the EUCLID network consists of a range of different sensor types and more than 140 sensors covering all of Europe. While there is variety in the different sensor types, they are all Vaisala Inc. based technologies, operating in the same frequency range. Figure 3.6 shows the locations of all sensors operating in the EUCLID network as of 2015. Data from all sensors is processed centrally in Austria by the Austrian Lightning Detection and Information System (ALDIS) team to ensure consistency throughout Europe (Schulz et al., 2016). The ALDIS network consists of 8 Vaisala LS7002 sensors and is indicated in the figure in red (Diendorfer, 2016).

Figure 3.6: Location of sensors in the EUCLID network as of 2015. Austria and the ALDIS network sensor locations indicated in red (Diendorfer, 2016). Image used with permission, courtesy of Dr W. Schulz.

The SALDN consists of 24 Vaisala LS7000 sensors (Gijben, 2012). The location of these sensors are shown in figure 3.7. The network was originally installed in 2005 and consisted of 19 sensors. In late 2009 - early 2010, three more sensors were added. A second upgrade was performed mid-year 2011 in which two more sensors were added to the network and three of them relocated for better coverage (Gill, 2008). The network has been greatly instrumental in developing up-to-date ground flash density maps for South Africa.

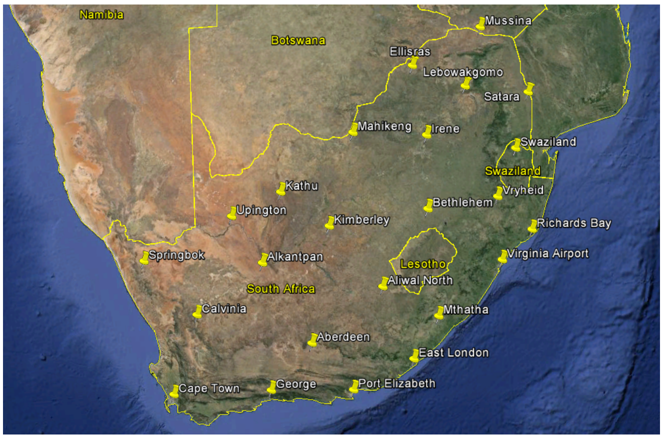


Figure 3.7: Location of sensors for the South African Lightning Detection Network (SALDN). Image used with permission, courtesy of Mr Morné Gijben.

3.4 Lightning Location System Performance

It is necessary to establish performance criteria for LLSs such that they may be evaluated and the reliability of the thunderstorm data ensured. Given that LLSs detect and report the occurrence of lightning events, the two main performance criteria are (Nag et al., 2015, Schulz, 1997):

- Detection efficiency (DE)
- Location accuracy (LA)

Detection efficiency is a measure of how many true lightning events are recorded and location accuracy is how accurate the reported locations of these events are. Two other performance criteria that are also relevant to the performance of LLSs are the estimation of the peak currents and classification of cloud-to-cloud/cloud-to-ground strokes. We will look at just the two main performance metrics.

3.4.1 Detection Efficiency

The detection efficiency of a LLS is defined as the ratio of the number of detected flashes to the number of actual flashes and is usually given as a percentage:

$$DE_{Flsh} = \frac{no.\ reported\ flashes}{no.\ actual\ flashes} \times 100\% \tag{3.3}$$

Given that LLSs detect individual strokes, and flashes are grouped subsequently, stroke detection efficiency (DE_{Str}) can be defined similarly. Detection efficiency is a metric used to describe whether a lightning location system is actually measuring real lightning events that occur. If the detection efficiency of a network is low, the flash density in the region of coverage will be underestimated. It is also useful for determining whether sections of a network, or sensors within the network, are functioning correctly.

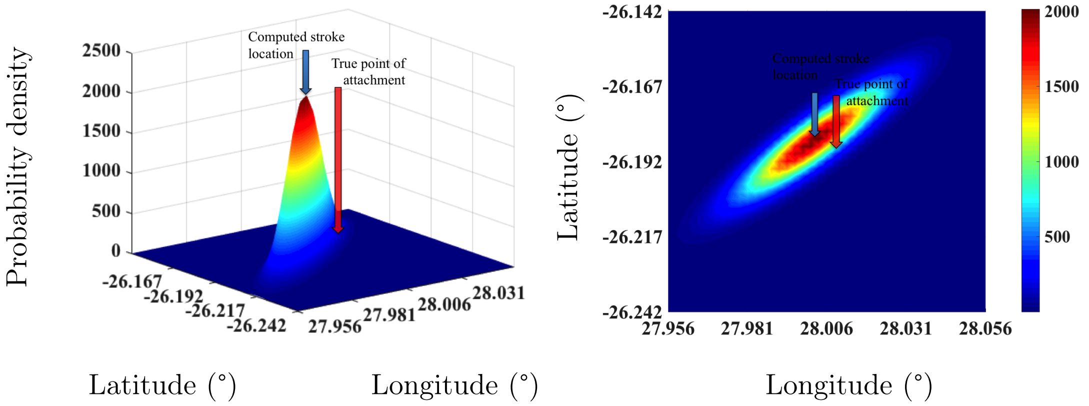
Estimating the detection efficiency of a LLS requires comparison with some other source of lightning event data - or example, photographic records, satellite data etc. Each of these has its own limitations and these limitations need to be considered in any comparison. LLSs operators regularly use self-evaluation techniques to estimate the detection efficiency of their networks in different regions of coverage. These techniques usually rely on self-comparison between sensors where one sensor in a network is compared with other sensors (Naccarato and Pinto, 2009, Schulz and Cummins, 2008).

3.4.2 Location Accuracy

Location accuracy is measure of how accurately a LLS reports the location at which a lightning stroke attached to ground (Cummins and Murphy, 1998, Schulz, 1997). It is measured as the difference between two geographic locations - the reported location and the true location. These differences are referred to as location errors and are usually given in kilometers. Since each stroke detection has its own reported location, each stroke has its own location error. Hence, location accuracy is a statistical measure and is usually given by the median or mode of a set of location errors. Similarly to detection efficiency, this can used to represent the performance of a LLS within a region.

The range in values for location errors is great, with poorly commissioned networks sometimes reporting strokes kilometers from their true location while well calibrated networks tend to have location accuracy values within 50-200 metres (Poelman et al., 2017, Schulz, 1997). If no systematic factors play a role in the location errors of reported strokes from a network, it is generally assumed that the location errors are distributed normally (Gaussian distribution). As with detection efficiency, estimating the location accuracy of a network in a region can be done using self-evaluation techniques but comparison with ground-truth data is necessary. Ground-truth data constitutes records of not only the time of lightning events but the location at which attachment occurred.

Self-evaluation of location errors is done by comparing individual sensor timing and angle reports (t_i and α_i from equations 3.2 and 3.1) of a historic data set of strokes with the individual sensor reports of resolved stroke locations (t_{ci} and α_{ci}). The variance of



(a) (b)

Figure 3.8: Gaussian distribution from where confidence ellipses are derived. Note the location error between the true stroke location and the computed stroke location.

the differences between i and ci is then found for each sensor ($\sigma_{\alpha_{ci}}$) and a Gaussian distribution is fitted. When a new stroke is detected, the distributions of each of the sensors that detected the stroke are combined, resulting in a bivariate Gaussian distribution in longitude and latitude where the optimal resolved location is set as the mean (Schulz, 1997, Stansfield, 1947). Figure 3.8 shows an example of a LLS stroke report where there is a location error (the true point of attachment is indicated) and the bivariate Gaussian description of the stroke is shown as well.

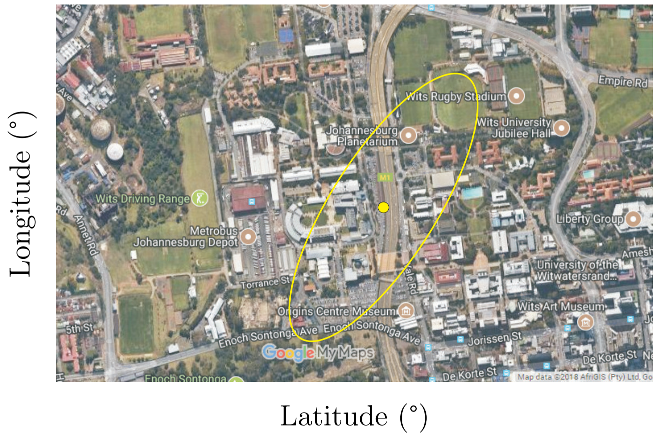


Figure 3.9: Resolved LLS stroke location with 50% confidence ellipse shown. The stroke was reported in the region of the University of the Witwatersrand, Johannesburg, South Africa.

From this, the major and minor axes of the 50^{th} percentile are determined which allow the distribution to be represented as a 50% confidence or error ellipse around the optimized stroke location. Figure 3.9 shows an example of a resolved stroke location its 50% confidence ellipse.

3.4.3 Ground-Truth Validation

While self-evaluation techniques are useful tools and aid LLS operators in monitoring their networks, validation against ground-truth lightning events is also important for understanding the performance of an LLS (Schulz, 1997, Schulz et al., 2016). Some of the types of ground-truth comparison are:

- Rocket-triggered lightning;
- Instrumented tall towers;
- Video studies.

Each of these have certain advantages and disadvantages. Rocket-triggered lightning studies are useful for assessing both detection efficiency and location accuracy since the exact location of ground attachment is known. However, there is concern that rocket-triggered lightning events do not represent natural downward lightning first return strokes. Tall tower observations and measurements are also useful given that the location is known but usually initiate upward lightning events, making detection by LLSs difficult. Video studies allow for detection efficiency to be evaluated but if location of attachment is not visible, it is not possible to assess location accuracy.

Rocket-triggered lightning events at Camp Blanding were used to validate the U.S.NLDN on a number of occasions. A study by Jerauld et al. (2005) in 2005 showed that U.S.NLDN detected 84% of 37 flashes rocket-triggered flashes and 60% of 159 strokes within these flashes. The location accuracy was found to have a median value of 600 m. A similar study performed in 2011 found flash and stroke detection efficiencies of 92% and 76% respectively, and a median absolute location error of 308 m (Nag et al., 2011).

An example of an instrumented tall towers is the Gaisberg tower in Austria. The top of the tower has a resistive shunt installed which enables the current waveform of all lightning events to the tower to be recorded. The ALDIS network has been validated against the measurements a number of times (Diendorfer, 2010). A study to evaluate the self-validation error ellipses was also performed at the tower, where it was shown that the confidence ellipse percentile is a good reflection of the number of strokes that truly attached within the ellipse area (Diendorfer et al., 2014). In a more recent evaluation of the EUCLID network, the instrumented Gaisberg tower played a role as one of the sources of ground-truth data, along with video and E-field recordings (Schulz et al., 2016).

The SALDN has only been compared with ground-truth data once, in 2014, in a study which involved photographs of lightning to the Brixton tower, a tall tower in Johannesburg, South Africa (Hunt et al., 2014). The tower has an approximate height of 250 m and is located at 26.1925° south and 28.0068° east. It was found that the network had a detection efficiency of 52% for upward lightning events to the tower and a median location accuracy of 250 m. It was not possible to assess the detection efficiency of downward events since only the tower was observed only only upward events occurred from it.

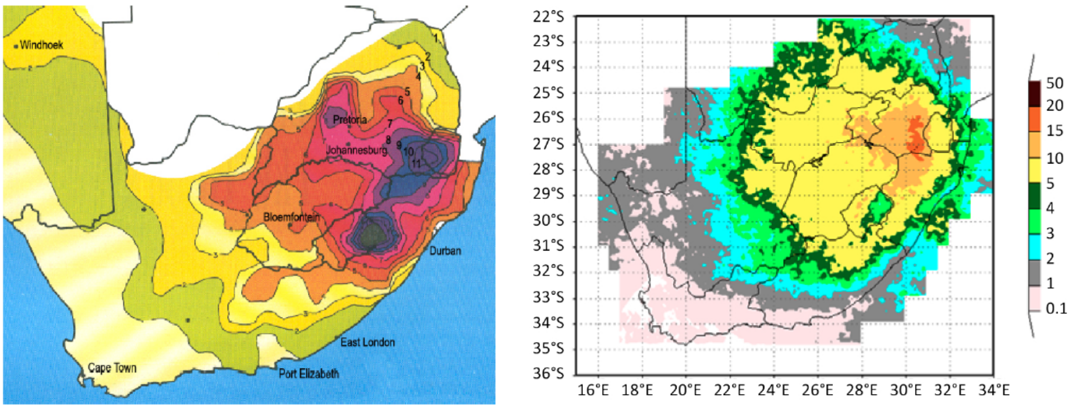
3.5 Applications of LLS Data

LLS data has applications in many fields. Here we look at two: ground-flash density maps and forensic investigations involving lightning.

3.5.1 Ground-Flash Density Maps

Ground-flash density maps are a means of quantifying lightning incidence in a region and report values as number of flashes per square kilometer per year ($flashes \times km^{-2} \times year^{-1}$). These type of maps are useful in understanding the meteorological nature of lightning and effects of topography, as well as being useful to power utilities and industry in helping them understand the risk exposure of assets.

LLSs have greatly enhanced the accuracy of these maps where lightning counters and thunderstorm day data sets were previously used. LLS stroke detections are grouped into flashes by reported time and location and then the number of flashes per square kilometer per year is determined. Figure 3.10 shows two ground-flash density maps of South Africa. The first map (figure 3.10a) was created in 1994 using data from lightning counters and was commissioned by the Council for Scientific and Industrial Research (CSIR), South Africa. From the map, maximum ground-flash densities of 11-12 $flashes \times km^{-2} \times year^{-1}$ were estimated.



(a) CSIR map

(b) SALDN map

Figure 3.10: Ground-flash density maps of South Africa (a) CSIR map created with lightning counter data (b) SAWS map created with SALDN data (Gijben, 2012).

The second map (figure 3.10b) was created in 2012 by Gijben (2012) at SAWS and utilizes data from the SALDN from 2006-2011. Here we see that the map based on LLS data indicates maximum ground-flash densities of 15-20 $flashes \times km^{-2} \times year^{-1}$, indicating that the previous map was underestimating the lightning incidence in the country. The IEC 62858 (2015) provides a clear set of recommendations as to the performance requirements of LLS data for use in ground-flash density maps.

Ground-Strike Points (GSPs):

More recently there has been a call for a third distinction to made between flashes and strokes when considering LLS data in ground-flash density maps and this is the ground-strike point (GSP) (Bouquegneau et al., 2013). Most stepped leaders attach to a single ground termination point to which current flows (and all subsequent strokes). However, it is possible for a second branched leader to attach to some other location and create a second channel with a ground termination point that is still part of the flash. Some of the charge in the subsequent strokes may then flow down this second channel (Cummins, 2014). Since LLSs detect individual strokes, the distinction as to whether a flash constituted one or more ground termination points may not be considered in the usual flash grouping algorithms and will simply be reported as a flash to a single location.

This has prompted the definition of the GSP when discussing the ground-flash density maps developed from LLS data as flash grouping algorithms may lead to underestimating the true number of ground terminations. There has been recommendations to increase all calculated densities by a factor of two (Bouquegneau, 2014, IEC 62858, 2015). This has also prompted the need for GSP grouping algorithms that can distinguish multiple termination points in a group of reported stroke locations. Stall et al. (2009) propose a method utilizing measured parameters such as rise-time and the chi-square value. Pédeboy and Schulz (2014) implemented a k-means clustering approach to distinguish GSPs with some success. Campos et al. (2015) further developed this, including the confidence ellipse error estimation in the k-means clustering approach with improved results.

Gaussian Smoothed Maps:

Another addition to ground-flash density maps is the use of confidence ellipses to create high-resolution (or resolution independent) density maps. Considering ground-flash density maps are flash count per area, the overall characteristics become lost visually in high-resolution maps (ie. grid squares less than a kilometer) since many grid squares will have zero flash counts within them. An approach by Bourscheidt et al. (2012) is to use the confidence ellipse of each flash as a Gaussian kernel and "smooth" each detection over the grid squares. The grids from each flash are all added together producing a "Gaussian count" that is independent of grid size Bourscheidt et al. (2013).

This allows for maps to be created for fairly small regions and still have relevant detail. Figure 3.11 shows an example of a Gaussian smoothed map over the University of the Witwatersrand in Johannesburg, South Africa which utilizes ten years of SALDN data. It should be noted that the values associated with this type of map are no longer $flashes \times km^{-2} \times year^{-1}$ and while Gaussian kernel smoothing appears as a mixture of bivariate Gaussian distributions, it is not a true Gaussian Mixture Model (GMM) as each component is not normalised.

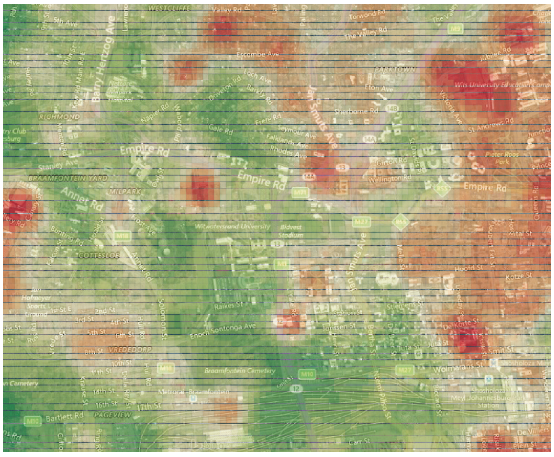


Figure 3.11: Gaussian smoothed map over the University of the Witwatersrand, Johannesburg, South Africa utilizing ten years of SALDN data.

3.5.2 Forensic Investigations Involving LLS Data

Another important application of LLS data is in forensic investigation - confirming the presence of lightning given damage to equipment or structures and even loss of life. Before such systems, forensic evidence from the event itself would the only available tool to establish lightning as a cause, unless there were eyewitnesses. Even then, this would usually only confirm the presence of a thunderstorm in the vicinity. With LLSs, it is possible to consult the data and find the detected strokes that may corroborate the event. Nonetheless, there is still uncertainty present due to the possibility of location errors.

Power utilities are often interested in determining the lightning performance of their power lines and, more specifically, if lightning was or was not the cause of a fault on the line. Fault detection equipment can trace the location of a fault on a line and if this can be correlated with LLS data in the same location and at the same time, utility operators are able to make a case for lightning as the cause of a fault. Of course, the possibility of location errors means that the reported strokes may not be reported exactly “on top” of the power line, allowing for some uncertainty as to whether they truly are the cause of the fault. One of the ways this is dealt with is by using the confidence ellipses to see if the lines fall within the ellipse areas (Nucci, 2007). Of course, this method has its flaws as not all strokes’ true location occurs within the confidence ellipse area. Hunt et al. (2014) further showed that using this method has a high possibility of false positives.

An example of a case involving loss of life can be found in the work by Anderson (2001) and Carte et al. (2002), where a number of children camping in a tent in the Northern Province of South Africa were killed on the 11th November 1994. The investigation consulted the Lightning Positioning and Tracking System (LPATS) – the system operated in South Africa at the time – and was able to confirm four lightning strokes in the vicinity at the time of the event (Anderson, 2001). However, in these types of cases, very rarely is LLS provided as evidence of the event but rather just to confirm lightning in the area. Grant et al. (2012, 2011) proposed a method for calculating the probability that a stroke was in an area by using the parameters of the bivariate Gaussian function described by the confidence ellipse. By integrating the bivariate Gaussian function over the area, as shown in equation 3.4, the probability that the stroke was in that area is calculated:

$$P(A_{str}) = \iint_A f(\varphi, \lambda)\mathrm{d}A \tag{3.4}$$

Where $P(A_{str})$ is the probability of a reported stroke having attached in an area, A is the area, φ is longitude and λ is latitude. Huddleston et al. (2010, 2012) used a similar method when attempting to determine whether U.S. NLDN stroke reports had attached inside a Naval satellite base. Hunt et al. (2017) used this to develop a further method, in which the probability that the area was struck by any reported strokes in a time-frame of inspection could be calculated. In all these cases, the values of probability calculated are highly dependent on the size of the area chosen as well as relying on the bivariate Gaussian description of location errors which, as Grant et al. (2011) note, may not be the best description of LLS location errors.

3.6 Chapter Summary

This chapter has provided a background summary of Lightning Location Systems (LLSs) and their operations along with some applications of LLS data.

LLSs detect and report the time and location of individual lightning strokes, which then are grouped into flashes by and time and location. Magnetic Direction Finding (MDF) and Time-of-Arrival (TOA) techniques are used to triangulate the location where a lightning stroke attached by utilizing detections from multiple sensors. The nature of these techniques and errors in measurement at individual sensors can lead to errors in the reported location of detected strokes. The Detection Efficiency (DE) - how many true strokes are detected - and Location Accuracy (LA) - the distribution of location errors - are key performance criteria for LLSs.

The distribution of LLS location errors is often assumed to follow a bivariate Gaussian distribution and this is used in the self-evaluation or location error estimation techniques communicated in the form of confidence ellipses. Validations of LLS performance are also often performed by comparison with ground-truth data - lightning to tall instrumented towers, rocket triggered lightning events and video studies.

LLS data is used extensively in creating flash density maps for meteorological and risk assessment purposes. The grouping of strokes into flashes is a key factor in creating such maps, and the introduction of Ground-Strike Points (GSPs) introduces another needed for accurate grouping algorithms. LLS data is also used extensively in forensic investigations - in powerline fault tracking, asset damage and even loss of life cases. Usually the proximity of reported strokes is all that is considered, with no regard given to the uncertainty due to location errors. In some cases, the confidence ellipse is used as a simple, visual representation of the uncertainty. Other methods that calculate the probability utilize the confidence ellipse parameters and therefore, the assumption of a Gaussian distribution of the location errors.

Many countries around the world operate LLSs, not least the Austrian Lightning Detection and Information System (ALDIS) in Austria and the South African Lightning Detection Network (SALDN) in South Africa.

The next chapter provides some background on Bayesian probability and and its use in a forensic context.

Chapter 4

Probability and Bayes' in a Forensic Context

This chapter discusses the use of Bayesian probability concepts in forensic investigation. The origins of Bayesian probability are presented with an explanation of Bayesian inference. The interpretation and application of this in forensic situations and the presenting of evidence in a legal forum is discussed. Some concepts of multivariate statistics are also included.

4.1 Overview

Forensic investigation requires the weighing of evidence against all possible causes. There is always uncertainty in the nature of the true events that transpired making it necessary to accurately and consistently quantify this uncertainty. The use of probability, especially a Bayesian approach, and likelihood ratios is a powerful tool for presenting evidence and any uncertainties in the claims that can be made. This chapter consists of three sections:

4.2 Bayesian probability theory

4.3 Bayes' in a forensic context

4.4 Multivariate statistics

The first section discusses some of the fundamental concepts of probability theory and how this leads to Bayes' theorem. Bayesian inference is explained and its role in decision theory and investigation is discussed. Next, the use of Bayesian approaches and inference in forensic situations is presented along with some examples. Finally, some multivariate statistical concepts are discussed, necessary for understanding the following chapters of the thesis.

4.2 Bayesian Probability Theory

While Bayesian probability is often referred to as a subjective probability or a *degree of belief* rather than an objective probability or a *relative frequency*, both interpretations share common definitions and are based on the fundamental logic of events and sample spaces (Jaynes, 2003).

4.2.1 Sample Spaces and Events

The set of all events, or all possible outcomes, from a random experiment is the sample space. The sample space can be discrete - where there is a finite, countable set of outcomes - or continuous if outcomes occur on an interval or infinite outcomes within an interval. Events are subsets of a sample space that result from a random experiment and can be a collection of related outcomes from a sample space (Montgomery and Runger, 2014, Murphy, 2012). New events can also be described in terms of:

- the **union** of two events, all outcomes that are contained in either of the events. Denoted $A \cup B$.
- the **intersection** of two events, all outcomes that are contained in both events. Denoted $A \cap B$.
- the **complement** of two events, the set of outcomes in the sample space that are not contained in the event. Denoted $\overline{\mathrm{A}}$.

Two events that have no outcomes in common ie. $A \cap B = \oslash$ are mutually exclusive.

4.2.2 Probability

Probability is a concept generally used to measure or quantify the uncertainty of whether an event is true or not. The probability of an event A being true is $P(A)$ and $0 \leq P(A) \leq 1$ where $P(A) = 1$ means A is certainly true and $P(A) = 0$ means A is certainly false. Any value between 1 and 0 indicates uncertainty in whether A is true or false.

From this definition, it follows that $P(S)$ (or the probability of all outcomes in the sample space) is equal to 1 and that the probability of mutually exclusive events is zero. The probability of the complement of an event, $P(\overline{\mathrm{A}})$, then becomes $1 - P(A)$. The probability of the union of two events is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.1}$$

If the two events are mutually exclusive, $P(A \cap B)$ is zero and the union is simply the sum of the probability of each event.

Since non-mutually exclusive events may have some of the same outcomes contained, it is possible to determine the probability of an event based on another event. This is known as conditional probability and is defined as such:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.2}$$

for $P(A) > 0$.

This describes the probability of A being true given that B is already known to be true.

4.2.3 Bayes' Theorem

Bayes' Theorem follows from the definition of conditional probability. Consider events A and B in equation 4.2. A similar equation could be written as follows:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{4.3}$$

Since the intersection of A and B ($A \cap B$) is equivalent to the intersection of B and A ($B \cap A$), then:

$$P(B)P(A|B) = P(A)P(B|A) \tag{4.4}$$

From this, Bayes' theorem is derived:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{4.5}$$

Now, the conditional probability of A being true given that B is true, is related to the probability of A being true, the probability of B being true and the probability that B would be true if A were true.

4.2.4 Bayesian Inference

The conditional nature of the probabilities in Bayes' theorem make it particularly useful for making inferences about outcomes based on a set of measured data (Gelman et al., 2013, Jaynes, 2003, Tsitsiklis, 2010). The symbols are usually changed in equation 4.5 to denote this particular use of Bayes' theorem where A becomes θ and is referred to as the parameter - this distinction is made to indicate that the outcome is whether the parameter θ takes on a specific value, rather than just whether the event A has happened or not. The event B is often replaced with x which is used to represent a set of measured data from an experiment, rather than simply whether the event B has happened or not. Bayes' theorem then becomes:

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)} \tag{4.6}$$

In this interpretation of Bayes' theorem, each of the terms in the formula take on a specific significance and are defined as follows:

- $P(\theta|x)$, the posterior probability function - the probability that θ has a specific value given the data x that was measured.
- $P(x|\theta)$, the likelihood function - the likelihood that the data x would be measured if θ truly has this specific value.
- $P(\theta)$, the prior probability function - the probability that θ has this specific value before any data was measured.
- $P(x)$, the probability that the data x is measured for any of the possible values that θ may have ie. $P(x) = \sum_{\theta} P(\theta)P(x|\theta)$.

The expression has the property of updating the probability - any prior knowledge about the probability of the parameter is updated by the measured data through the likelihood function. We can see that knowing the prior probability function and the likelihood function are essential to applying Bayesian inference. These functions are therefore both probability distributions/statistical models (either discrete or continuous).

The first and most apparent use of Bayesian inference is to determine the parameter (or parameters) of a statistical model or distribution given a set of data points. The likelihood function becomes the statistical model and theta becomes the parameters of the model. The parameters of the model which best fit the data can then be determined from which parameters have the highest probability (Bernardo, 1979, Congdon, 2003, Gelman et al., 2013). Bayesian inference is also at the heart of machine learning and *soft clustering* algorithms and techniques such as Expectation-Maximization (EM) (Murphy, 2012). An example is signal decoding where noise over a telecommunications line may affect the received data and it is necessary to infer the true sent data from the contaminated received data.

There are some subtle variations in the formula depending on the nature of θ and x and whether the values they can take on fall in discrete or continuous distributions (Tsitsiklis, 2010). Firstly, if both θ and x are discrete random variables, then the formula is equation 4.7:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\sum_{\theta} p(\theta)p(x|\theta)} \tag{4.7}$$

where $p(\theta)$ and $p(x|\theta)$ are the probability mass functions (usually denoted by the p) of both the prior probability distribution and the likelihood function. This is often referred to as a discrete prior and a discrete likelihood (David and Stone, 1973).

However, if both θ and x are continuous random variables, their distributions can no longer be described by probability mass functions but rather by probability density functions (denoted by f). Figure 4.1 shows the relationship between the probability density (probability/unit) and the probability of a random variable Y.

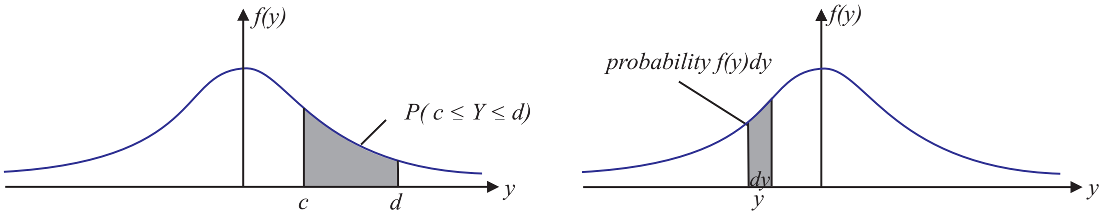


Figure 4.1: Probability density $f(y)$ of random variable Y.

The probability of Y being between two values $[c, d]$ is given by:

$$P(c \leq Y \leq d) = \int_c^d f(y)dy \tag{4.8}$$

If $[c, d]$ becomes an infinitesimal range ie. dy, then the probability that Y is in this infinitesimally range is $f(y)dy$.

Given that Bayes' formula (equation 4.6) is in terms of probabilities, if θ and x are continuous, it becomes:

$$f(\theta|x)d\theta = \frac{f(x|\theta)f(\theta)d\theta dx}{(\int_\theta f(x|\theta)f(\theta)d\theta)dx} \tag{4.9}$$

Since x is not varying (only θ), dx can be factored out of the integral in the denominator and the equation becomes:

$$f(\theta|x)d\theta = \frac{f(x|\theta)f(\theta)d\theta}{\int_\theta f(x|\theta)f(\theta)d\theta} \tag{4.10}$$

While $d\theta$ occurs on both sides of the equation, it is usually kept to denote a calculated probability since in the continuous case, the posterior probability is a density function $f(\theta|x)$. It is also possible to have a mixture of discrete and continuous priors and likelihoods, when either θ or x is discrete or continuous. If θ is continuous but x discrete, Bayes' formula becomes:

$$f(\theta|x)d\theta = \frac{p(x|\theta)f(\theta)d\theta}{\int_\theta p(x|\theta)f(\theta)d\theta} \tag{4.11}$$

Here, the posterior is a probability density function. If θ is discrete but x continuous, Bayes' formula becomes:

$$p(\theta|x) = \frac{p(\theta)f(x|\theta)}{\sum_\theta p(\theta)f(x|\theta)} \tag{4.12}$$

And the posterior is a probability mass function. This form of Bayes' theorem can be used to approximate equation 4.10, where $d\theta$ is approximated by a sufficiently small $\Delta\theta$ and is therefore discretized. This is further discussed in Appendix A.

4.2.5 Bayesianism vs Frequentism

As mentioned, there are two interpretations of probability - one being the frequentist perspective and the other being the Bayesian perspective. The differences are more philosophical, with both interpretations still seeing probability as a measure of uncertainty, but can still have an impact on a calculated result (Jaynes, 2003, Murphy, 2012).

From the frequentist point of view, there exists fundamental "true" probabilities of any outcomes in a given set having occurred. These "true" probabilities could be determined by an experiment run infinite times. Hence the term "objective" probability. Of course, this is not possible practically and we can only attempt to approximate this probability by running experiments.

The Bayesian perspective differs in that it does not consider probabilities to be "true" but rather that probabilities are a measure of the state of one's knowledge - or that all probabilities are actually conditional probabilities $P(A|B)$ and they are conditional on whatever information one may have. This is why the Bayesian perspective is often referred to as "subjective" probability - different people with different sets of knowledge will calculate different probabilities, even when assessing the same event.

The Bayesian perspective comes directly from Bayes' formula (equation 4.5), because it allows prior knowledge $P(A)$ to be updated by an experiment or likelihood $P(B|A)$. So, we may have some prior ideas about the probability of A occurring, we then run an experiment in which we measure B to infer A and update our prior beliefs with a new experiment (given that the experiment can never be this "true" probability as it is not infinite). This result is the posterior probability which can then become our prior knowledge should we run another experiment.

It is these characteristics of the Bayesian approach that make it appealing in practical inference or decision type problems - the concept of the likelihood, "when the outcome is true, what do I measure?" and how prior knowledge can be included in the calculation. Prior knowledge can be non-informative (sometimes referred to as a flat prior) where $P(A)$ is the same for all outcomes. For example, if I have a set of outcomes $A_1, A_2, ..., A_n$, a flat prior would be $P(A_1) = P(A_2) = ... = P(A_n) = \frac{1}{n}$. In other words, "I have no prior reason to believe any one outcome has happened more often then any of the others". An informative prior would change this, with each probability $P(A_{1,2,...,n})$ being different fractions of one.

4.2.6 Decision Theory and Bayesian Networks

Decision theory deals with the challenge of making decisions under conditions of uncertainty and, consequently, probability is an important concept (Jaynes, 2003). A distinction can be made between the idea of inference and decision making. The Bayesian framework is an incredibly useful inferential tool for example, and provides a consistent approach in which one can measure uncertainty and infer outcomes based how likely they may be or the probability of the outcome occurring/having occurred. This is not the same however, as making a decision. Taking a course of action based on what is known and what has been inferred about certain outcomes (Gittelson, 2013, Hansson, 1994). For example, it may be false that lightning was the cause of damage to a piece of equipment, but the decision of the insurance company on whether to be pay out to the client or not still needs to be made. Decision theory deals with the principles and algorithms necessary to make the correct or best decisions - often involving a combination of Bayesian probabilistic inference and loss or risk assessment functions (Turpin and Marais, 2004). By calculating probabilities and inferring outcomes, the expected loss or risk can be determined and this can be used to inform what is the best decision to make (Fenton and Neil, 2011).

Decisions of course, are not always dependent on the inference about a single outcome but rather multiple combined events or outcomes that effect each other, the certainty of which is not clear for any of them. This type of problem has lead to the concept of the Bayesian network, in which Bayesian inferences are “cascaded” to allow for the probability or uncertainty of a single outcome to be calculated (Murphy, 2012). Figure 4.2 gives an example of a Bayesian network.

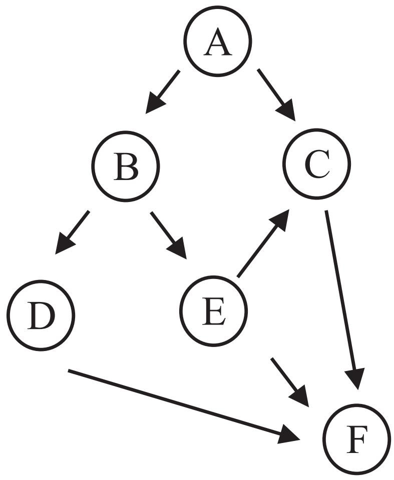


Figure 4.2: Example of a Bayesian network where multiple conditional dependencies are show, $P(A)$, $P(B|A)$, $P(D|B)$, $P(C|A, E)$, $P(E|B)$ and $P(F|C, D, E)$.

The network represents the conditional dependencies of each outcome on the other outcomes and how that leads to a final outcome F. F is dependent on the outcome of C, D and E which are in turn dependent on A, B and E. By following Bayes', the probability of F occurring, $P(F|C, D, E)$ can be determined.

One form of decision making is classification. For example, we infer the outcome of A, which may be "lightning struck the building." This is now classified as a lightning strike to the building. We may not be certain about the outcome A, but we attempt to make the classification based on our measure of that uncertainty.

Fawcett (2005) describes the evaluation of a classification process in terms of True Positives (TPs) and negatives (TNs) - when the classification is made correctly - and False Positives (FPs) and Negatives (FNs) - when the classification is made incorrectly. Table 4.1 describes this in what is termed a confusion matrix:

Table 4.1: Confusion matrix for classification process assessment (Fawcett, 2005).

	Real Pos.	Real Neg.
Inferred Pos.	True Positive	False Positive
Inferred Neg.	False Negative	True Negative

It is then possible to make assessments of a classification process:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4.13}$$

and

$$Precision = \frac{FP}{FP+TN} \tag{4.14}$$

Accuracy is therefore a measure of how many true classifications out of all classifications are made and *Precision* is a measure of how many negatives are correctly classified. The use of TPs, TNs, FPs and FNs has direct application in forensic investigation.

4.3 Bayes' in a Forensic Context

The Journal of the National Academy of Forensic Engineers defines forensic investigation and forensic engineering as *"the application of science and engineering in legal matters or matters which may relate to such"* (Specter, 2011). While forensic investigation is based on empirical scientific analysis, it differs from a scientific investigation in that the event has only happened once and cannot be repeated (Noon, 2001). The conclusion about the cause of the event needs to be drawn from a single run of an "experiment", and any prior knowledge that can inform the investigation. It is for these reasons that the Bayesian approach and tools of inference are a useful framework for assessing what occurred in a forensic investigation (Taroni et al., 2014).

In a legal situation, the results of a forensic investigation may be presented to a judge or panel of jurors and a decision needs to be made regarding the case at hand based the presentation of results by expert witnesses (Blair and Rossmo, 2010, Dawid, 2001). While there is resistance to presenting evidence to jurors and judges in the form of probability theory and statistical reasoning (they may not be educated in the field to interpret results correctly), it is becoming more and more common to have expert witnesses report their assessments of a forensic investigation in terms of a Bayesian framework involving posterior probabilities and likelihood ratios (Fenton et al., 2014a,b, Roberts and Aitken, 2014). This is especially true in the United Kingdom, where there is strong initiative by Aitken et al. (2010) to educate all parties involved in the legal proceedings in the fundamentals of statistics and Bayesian reasoning.

4.3.1 Bayes' Theorem in a Forensic context

The Bayesian perspective of viewing probabilities as *"strength in belief of a proposition"* rather than the frequentist interpretation as *"number of outcomes from a set of repeated events"* and its reliance on the likelihood function updating prior probabilities makes it ideally suited to evaluating uncertainty in forensic investigation, since the *"event"* is unrepeatable (Taroni et al., 2014).

In a legal interpretation, Bayes' formula becomes (Blair and Rossmo, 2010, Kruse, 2013, Roberts and Aitken, 2014, Taroni et al., 2004):

$$P(H|\epsilon) = \frac{P(H)P(\epsilon|H)}{P(\epsilon)} \tag{4.15}$$

Where H is the hypothesis or proposition and ϵ is the evidence that is available. H can take on any proposition. For example, it could be the prosecutor's claim that the defendant is guilty (H_1). Similarly, the defendant can make the claim that he is innocent and someone else is guilty (H_2). Bayes' allows the probability of both these claims to be assessed against the available evidence (equation 4.15) where:

- $P(H|\epsilon)$, the posterior probability - the probability that H is true given the evidence ϵ that was found.
- $P(\epsilon|H)$, the likelihood function - the function that gives the likelihood that the evidence ϵ would be obtained if H is actually true (a statistical model, probability density or mass function).
- $P(H)$, the prior probability of the proposition - the number of times that H was true in all other cases of this nature out of the total number of cases of this nature.
- $P(\epsilon)$, the probability of obtaining the evidence - $P(H)P(\epsilon|H) + P(\overline{H})P(\epsilon|\overline{H})$ ie. if H is true, ϵ would be obtained or if H is not true, other propositions may be true but ϵ was still obtained.

Competing propositions can then be assessed against each other in the form of posterior odds or the posterior likelihood ratio (Dawid, 2001):

$$\frac{P(H_1|\epsilon)}{P(H_2|\epsilon)} = \frac{P(H_1)}{P(H_2)} \times \underbrace{\frac{P(\epsilon|H_1)}{P(\epsilon|H_2)}}_{LR} \tag{4.16}$$

Since $P(\epsilon)$ is the same for both propositions, it cancels out when they are divided and the posterior odds become the product of the ratio of prior probabilities and the likelihood ratio, LR. Note that there is a distinction between the likelihood ratio and the posterior odds which is the prior probability ratio.

Figure 4.3 demonstrates the concept of the likelihood ratio LR where you have a likelihood function (probability density or probability mass function) for proposition H_1 given the evidence ϵ and a likelihood function for proposition H_2 given the evidence ϵ. Depending on what evidence was measured or found in the investigation, the likelihood of either H_1 or H_2 will vary accordingly (Royall, 1997).

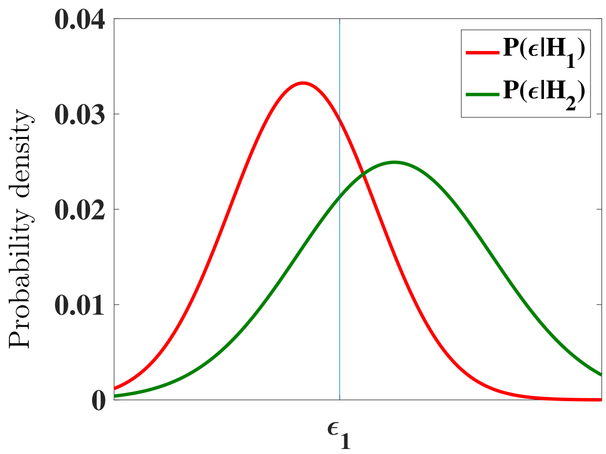


Figure 4.3: Likelihood ratio between two claims H_1 and H_2 with measured evidence, ϵ_1. Adapted from Aitken et al. (2010).

The figure shows a case in which the evidence ϵ_1 is more likely to have come from H_1 than from H_2. Hence, the likelihood ratio $\frac{P(\epsilon_1|H_1)}{P(\epsilon_1|H_2)} > 1$.

The value of the likelihood ratio being greater than or less than 1 has a clear interpretation (the numerator is more or less likely than the denominator) but the magnitude of how much more or less likely does not have a quantitative interpretation. There have been attempts to interpret likelihood ratio values in more qualitative terms for presentation to jurors and judges (Kaye, 2012, Martire et al., 2014, Royall, 1997). Table 4.2 gives an idea of this:

Table 4.2: Qualitative interpretation of LR as suggested by Kaye (2012)

Value of LR	Verbal equivalent
$> 1 - 10$	Weak or limited support
$10 - 100$	Moderate support
$100 - 1000$	Moderately strong support
$1000 - 10000$	Strong support
$10000 - 1000000$	Very strong support
> 1000000	Extremely strong support

It is based on this type of presentation, that judges and jurors are able to interpret the testimony of expert witnesses.

4.3.2 Prior Probability in a Forensic context

While the Bayesian framework is recognised as a powerful method for presenting evidence in legal cases due to its evaluation of the veracity of claims against the evidence in the form of the likelihood ratio, there is some contention over the role that the prior probability plays. The concern being the subjective nature in which prior probabilities can be calculated and that different expert witnesses may present different probabilities as their calculations come from different prior knowledge.

There are two ways this is addressed: one is to require all expert witnesses to use a flat or non-informative prior probability function when calculating probabilities but this has undergone criticism from those who support the Bayesian approach (Biedermann et al., 2007). Another approach is to require expert witnesses to only present their findings in the form of a likelihood ratio, LR and then leave the judge to apply his own prior probability using the product (equation 4.16) as suggested by Thompson et al. (2013).

4.3.3 Forensic Investigations with Bayesian Framework

The application of the Bayesian framework in forensic investigations ranges over a large number of subjects - from footwear identification for footprints found at crime scenes to quantifying likelihoods for digital investigations as done by Overill (2014). Often, there is more than one type of evidence available and this can be dealt with by the use of Bayesian networks, where different measurements that all have a bearing on the outcome can be connected given their dependencies.

The work of Biedermann has ranged from analysis of fire investigations to gunshot residue analysis, all involving the use of a Bayesian framework and Bayesian networks (Biedermann et al., 2011, Biedermann and Taroni, 2006, Biedermann et al., 2005). Facial identification type investigations is also often attempted with a Bayesian framework as done by Hepler (2005) and Allen (2008). The use of Bayesian networks for modelling crime situations and investigations is also undertaken using Bayesian methods (Baumgartner, 2005, Zoete et al., 2015). In all these cases, statistical models of the type of evidence collected, along with type of claims being investigated, are needed for the likelihood and prior probability functions.

4.4 Multivariate Statistics

Multivariate statistics deals with the distribution of data that has more than one random variable. For example, a univariate data set may be represented by a vector $\boldsymbol{x}$

which contains a range of values: $x_1, x_2, ..x_k$. A multivariate data set would instead be represented by a $k \times k$ matrix $\boldsymbol{X}$, where each column is a marginal random variable:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kk} \end{bmatrix} \tag{4.17}$$

For univariate data, we can describe the data set by set of statistics such as the mean μ and variance σ^2. Many of these same metrics can be described for multivariate data as well, with some adjustment. The mean instead becomes a $k \times 1$ vector $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)$. Variance is instead described by the co-variance matrix $\boldsymbol{\Sigma}$, a $k \times k$ matrix describing the variance (σ_{1-k}) of each marginal as well as the correlation (ρ) between marginals. Correlation describes whether the different marginal variables are independent or dependent of each other. If $\boldsymbol{\Sigma} = I$, where I is the identity matrix, the correlation is zero.

A descriptive statistic of multivariate distributions is the Mahalanobis distance:

$$\boldsymbol{\Delta} = \sqrt{(\boldsymbol{X} - \boldsymbol{u})^T \boldsymbol{\Sigma^{-1}} (\boldsymbol{X} - \boldsymbol{u})} \tag{4.18}$$

The Mahalanobis distance can be likened to the Euclidean distance of each data point from the mean, but weighted by the variance and correlation in the respective directions ie. if the co-variance matrix is the identity matrix, then the Mahalanobis distance will be the Euclidean distance.

4.4.1 Some Multivariate Distributions

Many common univariate statistical distributions have multivariate descriptions. Some of these are discussed here:

1. Multivariate Gaussian (Normal) distribution:

If we consider a $k \times k$ random column vector $\boldsymbol{X}$ that is normally distributed, then the probability density function will be:

$$f(\boldsymbol{X}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left[\frac{-(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})}{2}\right] \tag{4.19}$$

where:

$$\begin{aligned} \boldsymbol{\mu} &: \text{The mean matrix.} \\ \boldsymbol{\Sigma} &: \text{The co-variance matrix.} \\ f(\boldsymbol{X}) &: \text{The probability density.} \end{aligned}$$

Note the Mahalanobis distance term in equation 4.19. Figure 4.4 shows a plot of the probability density for a bivariate ($k = 2$) Gaussian distribution. The mean is set to $\boldsymbol{\mu} = (0, 0)$ and $\boldsymbol{\Sigma} = I$, where I is the identity matrix.

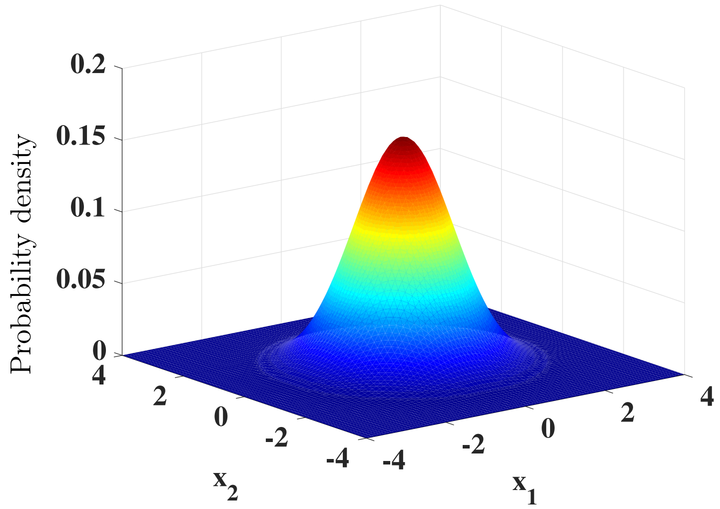


Figure 4.4: Example of a bivariate (k=2) Gaussian probability density function with $\boldsymbol{\mu} = (0, 0), \sigma_{x_1} = \sigma_{x_2} = 1$ and $\rho_{x_1 x_2} = 0$.

2. Gaussian Mixture Models (GMMs):

Multivariate Gaussian distributions are often used in mixture models. Mixture models are a mixture of different probability density functions used to describe clustering or sub-populations in univariate and multivariate data sets. A multivariate Gaussian Mixture Model (GMM) would be described by equation 4.20:

$$f(\boldsymbol{X}) = \sum_{q=1}^{Q} w_q g_q(\boldsymbol{X}) \tag{4.20}$$

where:

$$\begin{aligned} Q &: \text{The number of mixture components.} \\ w_q &: \text{The weight of each mixture component.} \\ g_q(\boldsymbol{X}) &: \text{The probability density of the multivariate Gaussian distribution.} \end{aligned}$$

Since the resulting function is a probability density, the integral over all space needs to be 1. But the integral over all space for each Gaussian mixture component is already 1, hence the need for each component to be weighted by w_q. If each component is weighted equally, $w_q = \frac{1}{Q}$. Each Gaussian mixture component will may have a different set of parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ - although it is common for Σ to be the same for each component. This is similar to the kernel in the Gaussian smoothing techniques although, in that case, the mixture components are not weighted and therefore do not represent a true GMM probability density function.

3. Multivariate Students' t-distribution:

Another univariate distribution that has a multivariate counterpart is the Students' t-distribution. The multivariate version of this distribution is obtained in the same way as the univariate version, and can be considered as a GMM where the variances of the Gaussians follow a chi-square distribution. So, for a normally distributed random variable, $\boldsymbol{Y} \sim \mathcal{N}(0, I)$ and the random variable $\boldsymbol{q} \sim \chi^2_\nu/\nu$. The variable t then follows a Student's t-distribution (Kotz and Nadarajah, 2004, Roth, 2013):

$$t \equiv t(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \boldsymbol{\mu} + \sqrt{\frac{\boldsymbol{\Sigma}}{\boldsymbol{q}}}\boldsymbol{Y}, \qquad \boldsymbol{Y} \sim \mathcal{N}(0, I), \qquad \boldsymbol{q} \sim \chi^2_\nu/\nu. \tag{4.21}$$

The probability density function of the multivariate Students' t-distribution has similar parameters to the multivariate Gaussian distribution which can also be described by the mean vector and co-variance matrix, but with an added parameter, the degrees of freedom parameter ν:

$$f(\boldsymbol{X}) = \frac{\Gamma[(\nu + k)/2]}{\Gamma(\nu/2)\nu^{k/2}\pi^{k/2}\sqrt{|\boldsymbol{\Sigma}|}} \left[1 + \frac{1}{\nu}(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right]^{-\frac{(\nu+k)}{2}} \tag{4.22}$$

where:

$\boldsymbol{\mu}$: The mean matrix.
$\boldsymbol{\Sigma}$: The co-variance matrix.
ν : The degrees of freedom.
k : number of dimensions of multivariate data.
$f(\boldsymbol{X})$: The probability density at the point $\boldsymbol{X}$.

Once again, note the Mahalanobis distance term in equation 4.22. The degrees of freedom parameter essentially acts as a scaling factor and is what leads to the heavier tails of the Students' t-distribution. If $\nu \to \infty$, equation 4.22 simply becomes the multivariate Gaussian probability density function. When $\nu = 1$, we have a special case of the Students' t-distribution - the Cauchy distribution, also known for its heavier tails. Figure 4.5 shows a comparison between a univariate Gaussian distribution, Students' t-distribution ($\nu = 2$) and a Cauchy distribution all with the same parameters (except for ν).

We can see the heavier tails in the Students' t-distribution and the Cauchy distribution and this same characteristic is carried into the multivariate versions of the distributions.

4. Multivariate uniform distribution:

The uniform distribution also carries into the multivariate domain, and is found when each of the marginals of $\boldsymbol{X}$ are uniformly distributed. There is no generalized probability density function for $k \times k$ dimensions of the uniform distribution, instead we can think

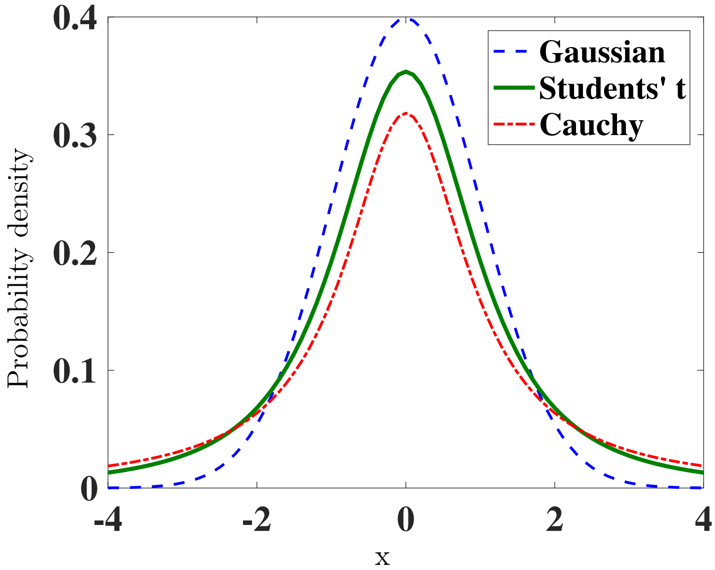


Figure 4.5: Univariate Gaussian, Students' t- ($\nu = 2$) and Cauchy distribution, $\mu = 0, \sigma = 1$.

of the multivariate uniform distribution as the product of each marginal distribution, where the univariate probability density function is:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for} \quad a \leq x \leq b, \\ 0 & \text{for} \quad x < a \quad \text{or} \quad x > b. \end{cases} \tag{4.23}$$

where:

a, b : Upper and lower parameters.
$f(x)$: The probability density at x.

The parameters of the uniform distribution are defined as the upper and lower limits but can also be written in terms of the mean and variance. Equation 4.23 then becomes:

$$f(x) = \begin{cases} \frac{1}{2\sigma\sqrt{3}} & \text{for} \quad -\sigma\sqrt{3} \leq x - \mu \leq \sigma\sqrt{3}, \\ 0 & \text{otherwise} \end{cases} \tag{4.24}$$

where:

μ : The mean.
σ : Standard deviation.
$f(x)$: The probability density at x.

4.4.2 Multivariate Maximum Likelihood Estimation (MLE)

These multivariate distributions can provide useful statistical models for the distribution of known measurements (for example, the reported locations of LLS strokes). However, for a given multivariate data set, the specific parameters of the distribution that best describe the data set are unknown. Maximum likelihood Estimation (MLE) is a means of determining the values of these parameters for a given statistical model/distribution which best fit the data set. MLE is essentially Bayesian inference (as discussed in Section 4.2.4), but without the prior probability function and therefore no posterior distribution. Instead, the inference is done by using the likelihood function alone and finding which parameters maximise it. The principles of MLE are the same in the multivariate case as they are in the univariate case - except there are more parameters to fit.

To apply MLE, we first have to decide which statistical model/distribution we are trying to fit the parameters to the data. We then specify the joint probability density function for all observations - shown in equation 4.25:

$$f(x_1, x_2, ..., x_N|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta})\cdot f(x_2|\boldsymbol{\theta})\cdots f(x_N|\boldsymbol{\theta}) \tag{4.25}$$

where

$$\begin{aligned} \boldsymbol{\theta} &: \text{the parameters } (\mu,\ \sigma \text{ etc.}) \text{ of the statistical model.} \\ f(x|\boldsymbol{\theta}) &: \text{the probability density function of the statistical model.} \end{aligned}$$

Essentially, the joint probability density function of all observations is the product of each observation substituted into the probability density function of the chosen statistical model to be fitted. Usually, when considering a probability density function, the parameter values $\boldsymbol{\theta}$ are fixed but here, it is the observation values ($\boldsymbol{x}$) that are fixed. We then define the likelihood function by looking at this from a different perspective - the likelihood of the parameter values being $\boldsymbol{\theta}$, given the data observations $\boldsymbol{x}$ that have been made (the same definition of the likelihood function in Bayesian inference). This is shown in equation 4.26:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{j=1}^{N} f(x_j|\boldsymbol{\theta}) \tag{4.26}$$

This can also be expressed as the log-likelihood function (equation 4.27), in which the product becomes a sum:

$$\ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{j=1}^{N} \ln f(x_j|\boldsymbol{\theta}) \tag{4.27}$$

Given that this function is dependent on the fixed values of the observed data, and gives the likelihood of specific parameter values given these fixed observations, the best fitting parameters (ie. statistical model) can be determined by finding the parameter values that maximise the likelihood function. In other words, the likelihood value that is calculated from the observed data set will vary for different parameter values. Hence, MLE is the maximization of the likelihood function, and the parameters that maximise the likelihood function are known as the maximum likelihood estimators, $\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}}$, of a statistical model, and are given by equation 4.28:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}} = \arg\max_{\boldsymbol{\theta}} \prod_{j=1}^{N} f(x|\boldsymbol{\theta}) \tag{4.28}$$

The minimum negative log-likelihood is used as a measure of the quality of the fit. The negative log-likelihood is simply the negative of equation 4.27 with the maximum likelihood estimated parameters:

$$\text{Neg.log-lik.} = -\ln \mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}}|\boldsymbol{x}) \tag{4.29}$$

Other statistics based on the negative log-likelihood are the Akaike Information Criterion (AIC):

$$\text{AIC} = 2p - 2\ln \mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}}|\boldsymbol{x}) \tag{4.30}$$

and the Bayesian Information Criterion (BIC):

$$\text{BIC} = \ln(k)p - 2\ln \mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}}|\boldsymbol{x}) \tag{4.31}$$

where p is the number of fitted parameters and k is the number of observations in data set $\boldsymbol{x}$. Both of these aim to penalise the negative likelihood value by including the number of fitted parameters (AIC) and the number of data points (BIC) in the statistic.

MLE is a key technique in many other approaches to fitting statistical models, for example EM. EM algorithms are conventionally used in clustering type mixture models or missing data problems. They are iterative algorithms that randomly assign parameter values to the distributions to be fitted and determine the likelihood that each of the data points belong to these distributions. Based on these likelihoods, the algorithms then re-assign weighted (weighted by the likelihoods) parameter values and iterate again. The algorithms converge when the likelihood converges to a maximum. The algorithms get their name from the two main iterated steps, the **E-Step** (Expectation step), where the likelihoods are determined, and the **M-Step** (Maximization step), where the new parameter values are assigned and the likelihood is calculated (Meng and van Dyk, 1997).

4.4.3 Multivariate Goodness-of-fit Measures

While the negative log-likelihood and the other criterion can provide some indication of the quality of the fit, they are highly dependent on the number of data observations used and therefore are only relative measure not allowing for comparison between different data sets. Hence, goodness of fit type tests such as Pearson's chi-square test or the Kolmogorov-Smirnoff test are preferred. While these tests are commonly used in univariate distribution fitting, they have been shown to have trouble when extended to multivariate distributions, especially Pearson's chi-squared test. This test relies on binning the data set and requires that no bin value is zero - simple to achieve in the univariate case but almost impossible with multivariate data. Other tests work well for certain multivariate distributions but fail when used with others, making comparison between statistical models/distributions difficult (Justel et al., 1997, McAssey, 2013, Palombo, 2011). There are many proposed goodness-of-fit tests for multivariate distribution fitting, many based on variations of the Kolmogorov-Smirnoff test as described by Justel et al. (1997), some based on quantile-quantile plots (Dhar et al., 2014) and one even based on the Mahalanobis distance, proposed by McAssey (2013).

4.5 Chapter Summary

This chapter has discussed Bayesian probability and the use of Bayesian inference in a forensic context.

The derivation of Bayes' theorem from the fundamentals of probability theory was presented and its interpretation in Bayesian inference discussed, with specific reference to the property of updating the prior probability through the likelihood function. It was also shown how Bayesian inference can have a mixture of continuous and discrete variables leading to a mixture of probability mass functions and probability densities.

The Bayesian approach to probability lends itself well to forensic investigation for a number of reasons - the ability to make inferences about a single "run" of the experiment and the updating of prior knowledge through the collection of evidence (the likelihood function) meaning the likelihood of different claims, H can be assessed against the same evidence, ϵ. Presenting this through the likelihood ratio and posterior probabilities is a powerful tool for communicating forensic evidence.

Many univariate distributions of multivariate versions. The multivariate Students' t-distribution is similar to the multivariate Gaussian distribution (both contain the Mahalanobis distance) but has slightly heavier tails in the distribution. Maximum Likelihood Estimation (MLE) techniques can be used to fit parameters of multivariate distributions to multivariate data sets however, assessing the quality of fit requires specialised test outside of the normal univariate goodness-of-fit tests.

The next chapter presents the proposed Bayesian framework for forensic investigations utilizing Lightning Location System (LLS) stroke reports.

Chapter 5

Proposed Bayesian Framework

This chapter presents the proposed Bayesian framework for forensic investigations utilizing Lightning Location System (LLS) stroke reports. The chapter describes the requirements for applying the framework in an investigation along with the relevant equations. The statistical models necessary for use of the framework are discussed in detail and a recommendation for presenting the results of an investigation is provided.

5.1 Overview

The proposed solution to presenting Lightning Location System (LLS) data as evidence for forensic investigations where lightning is a possible cause, is to use a probabilistic Bayesian framework based on the Bayesian principles discussed in Chapter 4. This chapter describes the framework and is divided into the following sections:

5.2 Proposed Bayesian framework.

5.3 Likelihood function.

5.4 Prior probability function.

5.5 *NOT* likelihood function.

5.6 Applying the framework.

In the first section, the requirements and definition of terms and variables of the proposed Bayesian framework are presented. From these, the two main equations of the framework, posterior probability and likelihood ratio (briefly shown in Chapter 2 - equation 2.1 and 2.2), are fully described. The assumptions the framework is based on are outlined and the proposed statistical models for the likelihood function and prior probability function - essential to utilizing the two main equations of the framework - are described in terms of the defined variables in the following sections. A *NOT* likelihood function is also derived and example plots for the posterior probability and likelihood ratio are recommended in the final sections.

5.2 Proposed Bayesian Framework

The Bayesian framework for forensic investigation involving LLS stroke reports begins on the premise that an event/s to be investigated (ie. damage etc.) has/have occurred. The framework requires that:

- there is a defined geographic area of investigation, in which the location/s at which the event/s occurred can be found, such that.

$$\varphi_{min} \leq \varphi_{event} \leq \varphi_{max} \qquad \text{and} \qquad \lambda_{min} \leq \lambda_{event} \leq \lambda_{max} \tag{5.1}$$

  where:

  φ : longitudinal geographical coordinates.
  λ : latitudinal geographical coordinates.

- there is a defined time-frame in which the event/s occurred, such that:

$$t_{str} < t_{event} < t_{fin} \tag{5.2}$$

- there is a LLS with coverage over the defined area.

The geographic area of investigation is divided into a grid with square size $\Delta\varphi \times \Delta\lambda$ as shown in figure 5.1.

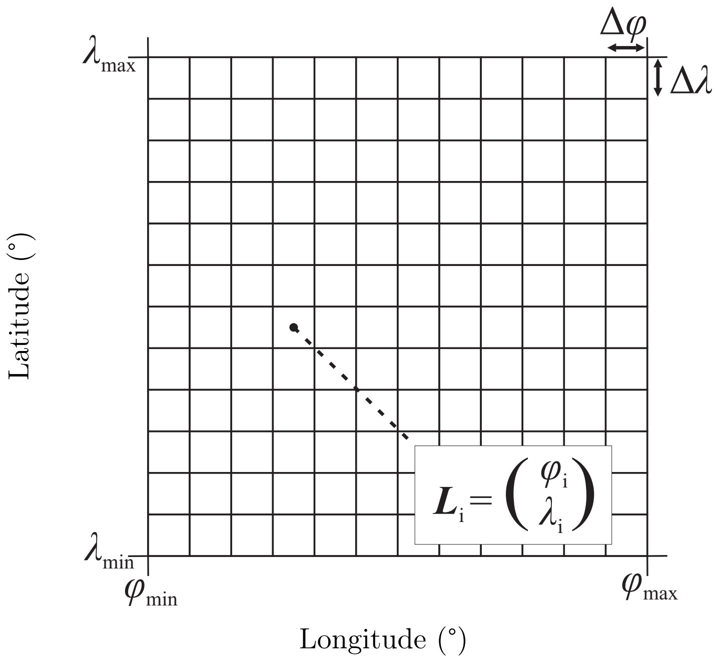


Figure 5.1: Defined geographic area of investigation with parameters shown. The location/s of the event/s under investigation are to be found within this area.

The set of all possible locations that may have been struck by lightning within the defined area of investigation will be finite and is then:

$$\boldsymbol{L_i} = \begin{pmatrix} \varphi_i \\ \lambda_i \end{pmatrix} \qquad i = 1, ..., N. \tag{5.3}$$

where:

φ_i : the longitudinal geographical coordinate of location $\boldsymbol{i}$.
λ_i : the latitudinal geographical coordinate of location $\boldsymbol{i}$.

And the reported locations of all strokes detected within the time-frame (and reported in the area) is then:

$$\boldsymbol{X_{lls_j}} = \begin{pmatrix} \varphi_j \\ \lambda_j \end{pmatrix} \qquad j = 1, ..., M. \tag{5.4}$$

where:

φ_j : The longitudinal geographical coordinate of stroke $\boldsymbol{j}$ detected at time t_j.
λ_j : The latitudinal geographical coordinate of stroke $\boldsymbol{j}$ detected at time t_j.
t_j : The reported time of detected stroke $\boldsymbol{j}$.

Given $\boldsymbol{L_i}$ is the coordinates of a location within the area of investigation, then $\boldsymbol{L_{i_S}}$ is a proposition or claim that *"location $\boldsymbol{L_i}$ was struck by lightning"*. Similarly, $\boldsymbol{L_{i_{NS}}}$ is the proposition or event that *"location $\boldsymbol{L_i}$ was NOT struck by lightning"* and $\boldsymbol{L_{i_{NS}}} = \overline{\boldsymbol{L_{i_S}}}$. $\boldsymbol{L_{i_S}}$ can therefore be likened to the claim, H and $\boldsymbol{X_{lls_j}}$ to the evidence, ϵ in the forensic interpretation of Bayes' theorem.

The posterior probability that a location $\boldsymbol{L_i}$ was struck by lightning given a single reported stroke at location $\boldsymbol{X_{lls_j}}$ is given by equation 5.5:

$$p(\boldsymbol{L_{i_S}}|\boldsymbol{X_{lls_j}}) = \frac{p(\boldsymbol{L_{i_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}})}{\sum_{i=1}^{N} p(\boldsymbol{L_{i_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}})} \tag{5.5}$$

(from the mixed form of Bayes' theorem, equation 4.12, Chapter 4, Section 4.2.4)

where:

- $p(\boldsymbol{L_{i_S}}|\boldsymbol{X_{lls_j}})$, the posterior probability function - the probability that location $\boldsymbol{L_i}$ was struck by lightning given that the LLS reported a stroke at $\boldsymbol{X_{lls_j}}$.
- $f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}})$, the likelihood function - the likelihood that the LLS would report a stroke at $\boldsymbol{X_{lls_j}}$ if location $\boldsymbol{L_i}$ was definitely struck by lightning (Section 5.3).

- $p(\boldsymbol{L_{i_S}})$, the prior probability function - the probability that $\boldsymbol{L_i}$ has been struck by lightning before the time-frame of investigation ie. before t_{str} (Section 5.4).
- $\sum_{i=1}^{N} p(\boldsymbol{L_{i_S}}) f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}})$, the probability that the the LLS would report a stroke at $\boldsymbol{X_{lls_j}}$ if any location in the area of investigation was struck by lightning.

The likelihood function $f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}})$ and the prior probability function $p(\boldsymbol{L_{i_S}})$ are statistical models that describe the distribution of $\boldsymbol{X_{lls_j}}$ and $\boldsymbol{L_i}$. The proposed models, the bivariate Students' t-distribution and the bivariate uniform distribution/Gaussian Mixture Model (GMM) are described in Section 5.3 and 5.4 respectively.

As described by the forensic interpretation of Bayes' theorem, competing claims can be assessed against each other (equation 4.16, Chapter 4, Section 4.2.4). In this case, the posterior odds between two claims/locations (for example, $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$) within the area of investigation can be expressed as:

$$\frac{p(\boldsymbol{L_{1_S}}|\boldsymbol{X_{lls_j}})}{p(\boldsymbol{L_{2_S}}|\boldsymbol{X_{lls_j}})} = \frac{p(\boldsymbol{L_{1_S}})}{p(\boldsymbol{L_{2_S}})} \times \underbrace{\frac{f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{1_S}})}{f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{2_S}})}}_{LR} \tag{5.6}$$

where LR is the likelihood ratio - the ratio of the likelihood of different claims given the evidence, without any prior probabilities considered:

$$LR_{L_1-L_2} = \frac{f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{1_S}})}{f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{2_S}})} \tag{5.7}$$

where $LR_{L_1-L_2}$ is the likelihood ratio between $\boldsymbol{L_{1_S}}$ and $\boldsymbol{L_{2_S}}$.

5.2.1 Assumptions and Scope

The proposed Bayesian framework presented above is based on the following assumptions:

1. There is a limit on the maximum possible location error that can be reported by an LLS such that it is possible to define the geographic area of investigation with large enough area that, if a location inside the area is struck by lightning, the detected stroke/s will not be reported with location/s ($\boldsymbol{X_{lls}}$) outside of the area.
2. The location errors of an LLS within a particular region are statistically consistent. In other words, while the location errors of individual detected strokes may vary within a particular region, the variance is consistent ie. the likelihood function for location $\boldsymbol{L_1}$ is the same as the likelihood function for location $\boldsymbol{L_N}$, just at different coordinates.
3. While $\boldsymbol{L_i}$ is truly a continuous set of locations (ie. $N = \infty$) within the defined geographic area (with prior probability density $f(\boldsymbol{L_{i_S}})$ of each location having

been struck by lightning), for a sufficiently small $\Delta\varphi$ and $\Delta\lambda$, each location $\boldsymbol{i}$ can be approximated by the grid square area $(\Delta\varphi \times \Delta\lambda)$ it is centred on (figure 5.1) since:

$$\text{as} \quad \Delta\varphi \to 0, \quad \varphi_i \pm \frac{\Delta\varphi}{2} \to \varphi_i \quad \text{and} \quad \Delta\lambda \to 0, \quad \lambda_i \pm \frac{\Delta\lambda}{2} \to \lambda_i \tag{5.8}$$

Therefore, $\boldsymbol{L_i}$ becomes a finite set of locations with prior probability function $p(\boldsymbol{L_{iS}})$, that can be given in terms of the density function $f(\boldsymbol{L_{iS}})$:

$$p(\boldsymbol{L_{iS}}) = \int_{\varphi_i-\frac{\Delta\varphi}{2}}^{\varphi_i+\frac{\Delta\varphi}{2}} \int_{\lambda_i-\frac{\Delta\lambda}{2}}^{\lambda_i+\frac{\Delta\lambda}{2}} f(\boldsymbol{L_{iS}}) d\varphi d\lambda \approx f(\boldsymbol{L_{iS}}) \Delta\varphi\Delta\lambda \tag{5.9}$$

More details can be found in Appendix A.

It should also be noted where the framework scope limitations are, as per Chapter 2, Section 2.5:

- Equation 5.5 does not account for the uncertainty of cloud-to-ground or intra-cloud stroke missclassifications, and will determine the probability based on reported location alone.
- Equation 5.5 does not account for undetected strokes and calculates a probability based only on reported strokes.

Next, the proposed statistical models for the likelihood function and the prior probability function are described and the *NOT* likelihood function is derived.

5.3 Likelihood Function $\boldsymbol{f(X_{lls_j}|L_{iS})}$

The likelihood function describes how the measured data is distributed if a claim is true. Here, this describes how the LLS data would be distributed if the proposition $\boldsymbol{L_{iS}}$ is true. In other words, the distribution of the location errors of reported strokes for a ground-truth location at which lightning has repeatedly occurred. As discussed in Chapter 3, these location errors are generally assumed to follow a bivariate Gaussian distribution. As noted however, outliers with large location errors do occur (Poelman et al., 2017) and it is proposed that a different distribution may better describe the location errors and prove to be a better statistical model for the likelihood function of the Bayesian framework.

The proposed statistical model for the likelihood function is the bivariate Students' t-distribution (described in Chapter 4). The bivariate Students' t-distribution is essentially a bivariate Gaussian distribution, but with heavier tails. In the context of the likelihood function for the Bayesian framework, the bivariate Students' t-distribution takes the form:

$$f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{is}}) = \frac{\Gamma[(\nu_i+2)/2]}{\Gamma(\nu_i/2)\nu_i\pi\sqrt{|\boldsymbol{\Sigma_i}|}}\left[1+\frac{1}{\nu_i}(\boldsymbol{X_{lls_j}}-\boldsymbol{\mu_i})^T\boldsymbol{\Sigma_i}^{-1}(\boldsymbol{X_{lls_j}}-\boldsymbol{\mu_i})\right]^{-\frac{(\nu_i+2)}{2}} \tag{5.10}$$

where:

$$\begin{aligned}
\boldsymbol{X_{lls_j}} &: \text{The reported location of stroke } j \text{ detected by the LLS, } \begin{pmatrix}\varphi_j & \lambda_j\end{pmatrix}.\\
\boldsymbol{\mu_i} &: \text{The mean matrix, } \begin{pmatrix}\mu_{i_\varphi} & \mu_{i_\lambda}\end{pmatrix}.\\
\boldsymbol{\Sigma_i} &: \text{The co-variance matrix, } \begin{pmatrix}\sigma_{i_\varphi}^2 & \rho_{i_{\varphi\lambda}}\sigma_{i_\varphi}\sigma_{i_\lambda}\\ \rho_{i_{\varphi\lambda}}\sigma_{i_\varphi}\sigma_{i_\lambda} & \sigma_{i_\lambda}^2\end{pmatrix}.\\
\nu_i &: \text{The degrees of freedom.}\\
f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{is}}) &: \text{The probability density at the point } \varphi_j,\ \lambda_j.
\end{aligned}$$

When $\nu_i \to \infty$, the bivariate Students' t-distribution becomes the bivariate Gaussian distribution:

$$f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{is}}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma_i}|}}\exp\left[\frac{-(\boldsymbol{X_{lls_j}}-\boldsymbol{\mu_i})^T\boldsymbol{\Sigma_i}^{-1}(\boldsymbol{X_{lls_j}}-\boldsymbol{\mu_i})}{2}\right] \tag{5.11}$$

Where the symbols have the same meaning as equation 5.10. When $\nu_i = 1$, the bivariate Students' t-distribution becomes the bivariate Cauchy distribution. Figure 5.2 shows an example of the bivariate Students' t-distribution as the likelihood function. The parameters for the example model are shown in table 5.1:

Table 5.1: Parameters for example bivariate Students' t-distribution as likelihood function.

μ_{i_φ}	μ_{i_λ}	σ_{i_φ}	σ_{i_λ}	$\rho_{i_{\varphi\lambda}}$	ν_i
28.0068°	−26.1925°	0.0025	0.0025	0.09	2

The mean $\boldsymbol{\mu_i}$:

The likelihood function is specific to a particular location $\boldsymbol{L_i}$, and shows the distribution of LLS reported data when that location is struck by lightning. It is therefore possible that $\varphi_i = \mu_{i_\varphi}$ and $\lambda_i = \mu_{i_\lambda}$. However, this is true only if the LLS location accuracy performance is subject to random errors alone but if any systematic error is present, there will be a shift in $\boldsymbol{\mu_i}$ from $\boldsymbol{L_i}$ - which will describe the effect of the systematic error.

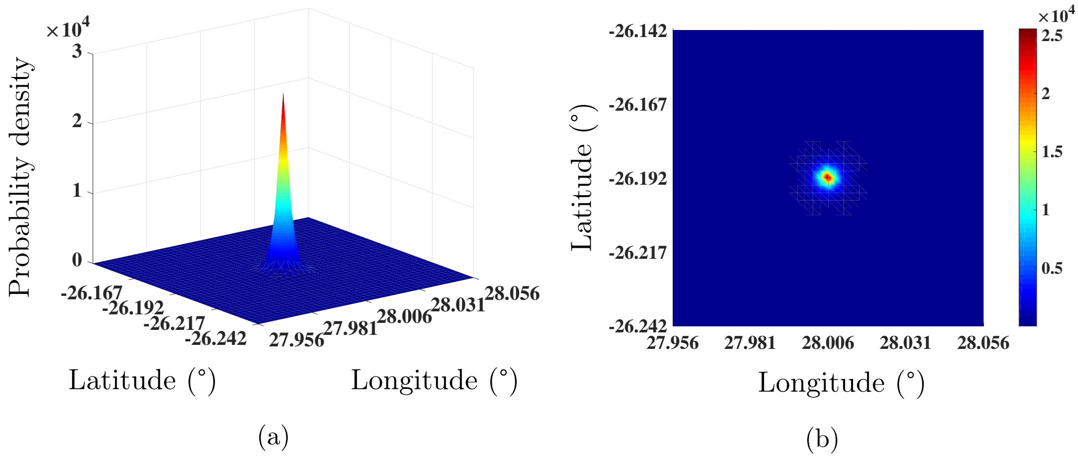


Figure 5.2: Example of bivariate Students' t-distribution as the likelihood function - this describes the distribution of LLS reported stroke locations when location $\boldsymbol{L_i}$ is struck by lightning.

The co-variance $\boldsymbol{\Sigma}_i$:

The co-variance matrix $\boldsymbol{\Sigma_i}$ describes the spread of the LLS data when location $\boldsymbol{L_i}$ is repeatedly struck by lightning. Both random errors and systematic errors in the reported locations of LLS strokes are described by these parameters and any baseline-type effects will be described by the co-variance matrix. Assumption 2 of the Bayesian framework manifests itself here - while the values of $\boldsymbol{\mu_i}$ will vary for each location $\boldsymbol{i}$, it is assumed the co-variance matrix will be the same for all locations within the area of investigation.

The degrees of freedom $\boldsymbol{\nu}_i$:

The degrees of freedom $\boldsymbol{\nu_i}$ can be thought of as *"steepness"*, or something approximate, and is measure of how much outliers are preferenced. For a lower $\boldsymbol{\nu_i}$, outliers will have a higher likelihood.

5.4 Prior Probability Function $p(\boldsymbol{L_{i_S}})$

The prior probability function describes what is known about the distribution of claims prior to measurements being made or evidence being collected. In this case, the prior probability function describes what is known about the set of locations $\boldsymbol{L_{i-N}}$ having been struck by lightning prior to the time-frame of investigation, t_{str}. in other words, which locations in the area have been struck more and which have been struck less. As mentioned in Chapter 4, the prior probability is the reason for the term *subjective* probability, associated with Bayesian inference, due to the many ways in which it can be chosen. Here, two types of priors are presented as options for the Bayesian framework:

- a non-informative (flat) prior - bivariate uniform distribution.

- an informative prior - bivariate Gaussian Mixture Model (GMM)

The statistical models chosen to achieve these are described below.

5.4.1 Non-informative (flat) prior

The non-informative or flat prior probability function describes the situation where nothing is known about which locations have been struck more often than others. The probability of any of the locations $\boldsymbol{L_{i-N}}$ having been struck is equal. Therefore, for a geographic area, the bivariate uniform distribution is chosen as it distributes the probability equally among all possible locations. For the framework, the bivariate uniform distribution takes the form:

$$f(\boldsymbol{L_{iS}}) = \begin{cases} \frac{1}{\varphi_{max}-\varphi_{min}} \times \frac{1}{\lambda_{max}-\lambda_{min}} & \varphi_{min} \leq \varphi_i \leq \varphi_{max} \text{ and } \lambda_{min} \leq \lambda_i \leq \lambda_{max}. \\ 0 & \text{for } \boldsymbol{L} \text{ everywhere else.} \end{cases} \tag{5.12}$$

where:

$\varphi_{max}, \varphi_{min}$: Upper and lower longitudinal limits of defined area.
$\lambda_{max}, \lambda_{min}$: Upper and lower latitudinal limits of defined area.
$f(\boldsymbol{L_{iS}})$: The probability density at the point φ_i, λ_i.

The probability density is the same for any location within the geographic area of investigation. However, since the framework chooses to describe the set of locations as a finite number N, the discrete form of the bivariate uniform distribution is:

$$p(\boldsymbol{L_{iS}}) = \begin{cases} \frac{1}{N} & \varphi_{min} \leq \varphi_i \leq \varphi_{max} \text{ and } \lambda_{min} \leq \lambda_i \leq \lambda_{max}. \\ 0 & \text{for } \boldsymbol{L} \text{ everywhere else.} \end{cases} \tag{5.13}$$

where:

N : The total number of locations $\boldsymbol{L_i}$ in the defined area.
$p(\boldsymbol{L_{iS}})$: The probability at the point φ_i, λ_i.

Given the parameters of the defined area and the grid square size, N can be determined as:

$$N = \frac{\varphi_{max} - \varphi_{min}}{\Delta\varphi} \times \frac{\lambda_{max} - \lambda_{min}}{\Delta\lambda} = \frac{1}{f(\boldsymbol{L_{iS}})\Delta\varphi\Delta\lambda} \tag{5.14}$$

Here we see that $p(\boldsymbol{L_{iS}}) = f(\boldsymbol{L_{iS}})\Delta\varphi\Delta\lambda$, rather than just the approximation given by equation 5.9. This is because, in the case of the uniform distribution, $f(\boldsymbol{L_{iS}})$ is a constant value and the integral in equation 5.9 simply becomes the limits integrated over 1.

Figure 5.3 shows $f(\boldsymbol{L_{is}})$ and $p(\boldsymbol{L_{is}})$ for a particular case where: $\varphi_{max} = 28.0318°$, $\varphi_{min} = 27.9818°$, $\lambda_{max} = -26.1675°$, $\lambda_{min} = -26.2175°$ and the grid square size is $\Delta\varphi = \Delta\lambda = 0.0025°$. Therefore, the number of locations $N = 400$ (from equation 5.14).

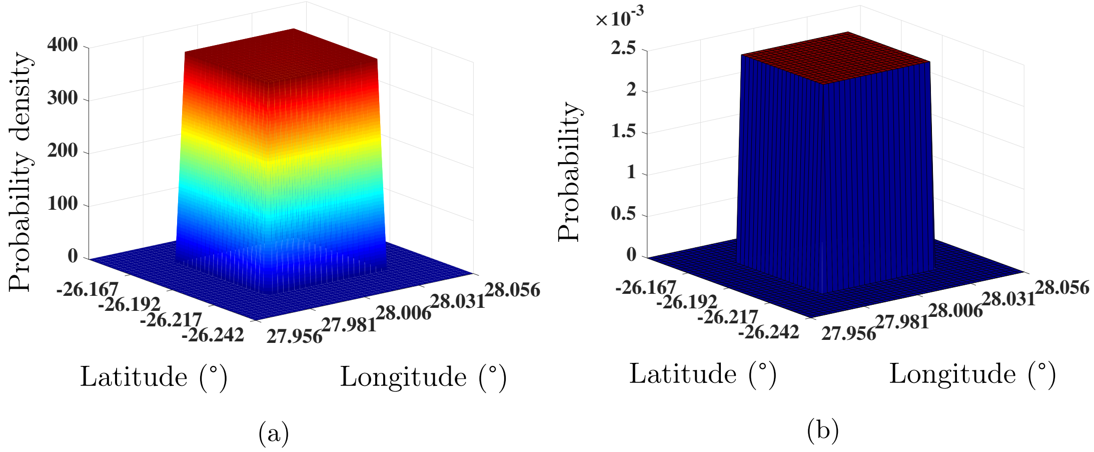


Figure 5.3: Bivariate uniform distribution for prior probability function as (a) probability density function $f(\boldsymbol{L_{is}})$, (b) probability mass function $p(\boldsymbol{L_{is}})$.

5.4.2 Informative prior

An informative prior is a prior probability function that described which claims have occurred more or less often previously. For example, tall towers are struck by lightning (or initiate upward flashes) more regularly than other locations. The non-informative prior described above would give us no indication of this - *"there is no reason to believe any location has been struck more often than any other location."*. This is of course, and we have LLS data to confirm so. It is proposed that an informative prior probability function for the Bayesian Framework be a bivariate GMM where:

$$f(\boldsymbol{L_{is}}) = \sum_{q=1}^{Q} w_q g_q(\boldsymbol{L_{is}}) \tag{5.15}$$

where:

Q : Number of mixture components.
w_q : The weight of each mixture component.
$g_q(\boldsymbol{L_{is}})$: Bivariate Gaussian distribution as a function of φ_i, λ_i.

Given that each defined geographic area of investigation may be in different regions and countries, the number of mixture components and the parameters of the bivariate

Gaussian distributions $g_q(\boldsymbol{L_{iS}})$ will differ for each investigation. Hence, a second data set of LLS reported stroke locations within the defined area of investigation is needed:

$$\boldsymbol{X_{llsPr_q}} = \begin{pmatrix} \varphi_q \\ \lambda_q \end{pmatrix} \qquad q = 1, ..., Q. \tag{5.16}$$

where:

φ_q : The longitudinal geographical coordinate of stroke $\boldsymbol{q}$ detected at time t_q.
λ_q : The latitudinal geographical coordinate of stroke $\boldsymbol{q}$ detected at time t_q.
t_q : The reported time detected stroke $\boldsymbol{q}$.

This data set of LLS strokes were all reported prior to the investigation ie. $t_Q < t_{str}$. The proposed GMM is then a mixture model of Q (the number of strokes in $\boldsymbol{X_{llsPr}}$) mixture components each taking the bivariate form shown in equation 5.11. For each mixture component, $\boldsymbol{\mu_i} = \boldsymbol{\mu_q}$, the reported location of each detected stroke. Given assumption 2, Section 5.2.1, we can assume that $\boldsymbol{\Sigma_i}$ will be the same for all locations and therefore can be used in each mixture component. Appendix A provides more detail on why this is valid within the framework.

5.5 *NOT* Likelihood Function $\boldsymbol{f(X_{lls_j}|\overline{L_{iS}})}$

As previously discussed, the likelihood ratio is an effective means of presenting how the evidence supports different claims. In this context, this would be used to determine which of two locations was struck by lightning (equation 5.6 and 5.7). However, when investigating a particular location $\boldsymbol{L_i}$, it can be useful to examine the likelihood ratio of $\boldsymbol{Li_S}$ against $\boldsymbol{Li_{NS}}$ (or $\overline{\boldsymbol{Li_S}}$):

$$LR_{L_i - \overline{L_i}} = \frac{f(\boldsymbol{X_{ldn_j}}|\boldsymbol{L_{iS}})}{f(\boldsymbol{X_{ldn_j}}|\overline{\boldsymbol{L_{iS}}})} \tag{5.17}$$

To do this, we need the likelihood function for $f(\boldsymbol{X_{ldn_j}}|\overline{\boldsymbol{L_{iS}}})$ or a *NOT* likelihood function.

Derivation of $\boldsymbol{f(X_{ldn_j}|\overline{Li_S})}$:

For a given stroke $\boldsymbol{X_{lls_j}}$, the probability that location $\boldsymbol{L_1}$ was *NOT* struck by lightning is:

$$p(\overline{\boldsymbol{L_{1S}}}|\boldsymbol{X_{lls_j}}) = 1 - p(\boldsymbol{L_{1S}}|\boldsymbol{X_{lls_j}}) \tag{5.18}$$

but, from equation 5.5, it can also be written as:

$$p(\overline{\boldsymbol{L_{1_S}}}|\boldsymbol{X_{lls_j}}) = \frac{p(\overline{\boldsymbol{L_{1_S}}})f(\boldsymbol{X_{lls_j}}|\overline{\boldsymbol{L_{1_S}}})}{\sum_{i=1}^{N} p(\boldsymbol{L_{i_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}})} \tag{5.19}$$

As equation 5.18 is equivalent to equation 5.19, then:

$$p(\overline{\boldsymbol{L_{1_S}}})f(\boldsymbol{X_{lls_j}}|\overline{\boldsymbol{L_{1_S}}}) = \sum_{i=1}^{N} p(\boldsymbol{L_{i_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}}) - p(\boldsymbol{L_{1_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{1_S}}) \tag{5.20}$$

Therefore:

$$f(\boldsymbol{X_{lls_j}}|\overline{\boldsymbol{L_{1_S}}}) = \frac{1}{p(\overline{\boldsymbol{L_{1_S}}})} \sum_{i=2}^{N} p(\boldsymbol{L_{i_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}}) \tag{5.21}$$

Which is essentially a weighted mixture of likelihood functions where the prior probability values determine the weighting. If the prior probability function is non-informative (ie. uniform distribution) then $p(\boldsymbol{L_{i_S}}) = \frac{1}{N}$ for all $\boldsymbol{i}$ and is a common factor. But $p(\overline{\boldsymbol{L_{1_S}}}) = 1 - p(\boldsymbol{L_{1_S}}) = 1 - \frac{1}{N}$. Therefore:

$$f(\boldsymbol{X_{lls_j}}|\overline{\boldsymbol{L_{1_S}}}) = \frac{1}{N-1} \sum_{i=2}^{N} f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{i_S}}) \tag{5.22}$$

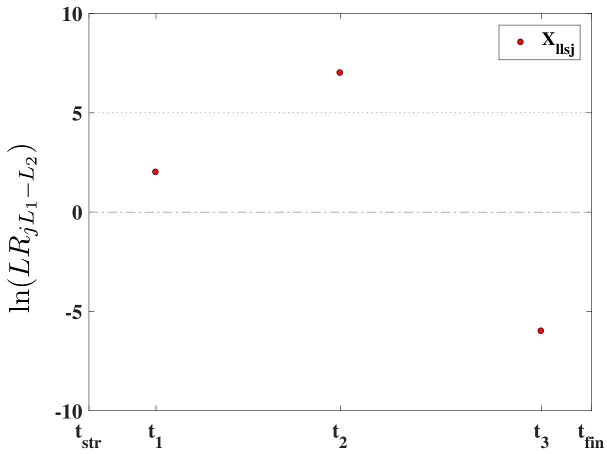
Equation 5.21 applies when an informative prior is used but reduces to equation 5.22 when a non-informative prior is used.

5.6 Applying the Framework

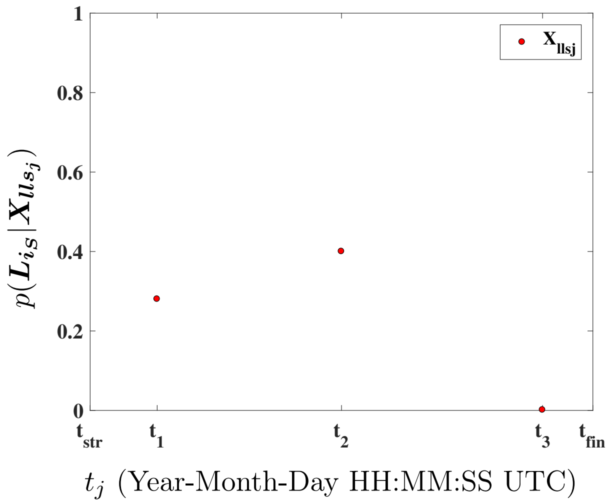
The application of the Bayesian framework to LLS stroke reports is in aid of determining whether certain claims are true. For example:

- $\boldsymbol{L_{i_S}}$: *"The location $\boldsymbol{L_i}$ was struck by lightning."*
- $\overline{\boldsymbol{L_{i_S}}}$: *"The location $\boldsymbol{L_i}$ was NOT struck by lightning."*

Figure 5.4 shows an example of presenting the results of applying the Bayesian framework. As discussed in Chapter 4, one form of presenting the evidence for these claims is the likelihood ratio LR. Figure 5.4a shows an example of how the likelihood ratio can be presented for LLS stroke reports - as a log-likelihood ratio $\ln(LR)$ against the time-stamp of each stroke t_j.



(a)



(b)

Figure 5.4: Presenting the results of the Bayesian framework (a) example of the log-likelihood ratio $\ln(LR_{jL_1-L_2})$ vs. t_j and (b) example of the posterior probability $p(\boldsymbol{L_{i_S}}|\boldsymbol{X_{lls_j}})$ vs. t_j.

The log-likelihood is chosen as the likelihood ratio values can range from $[-\infty; \infty]$ and this can lead to unclear plots. Critical values can then be established (along the lines of table 4.2, Chapter 4) and shown on the plot. For example, $\ln(LR_j) > 0$, which is the same as $LR > 1$. Since the likelihood ratio is a ratio of the support of the evidence for two claims, if the likelihood ratio is greater than 1, the claim of the numerator is more likely given the evidence. If not, the claim of the denominator is more likely. By showing this on the graph, it is easy to see which LLS stroke reports support the claim of the numerator or that of the denominator. Higher critical values can vary the strength of support as per the table 4.2. Similarly, the posterior probability can be plotted against the time-stamp of each LLS stroke report - $p(\boldsymbol{L_{i_S}}|\boldsymbol{X_{lls_j}})$ vs. t_j as shown in figure 5.4b. Here, the range is a probability and therefore between 0 and 1. This provides a measure of how certain we are that the location $\boldsymbol{i}$ was struck by lightning at time t_j, given the LLS measurements that were made.

5.7 Chapter Summary

This chapter has described the Bayesian framework proposed for use in forensic investigations utilizing Lightning Location System (LLS) stroke reports.

The premise of a forensic investigation is presented and the requirements - a defined geographic area of investigation, a time-frame of the event and LLS coverage - necessary to apply the Bayesian framework are presented along with the variables $\boldsymbol{L}$, the locations to be investigated, and $\boldsymbol{X_{lls}}$, the LLS data measured within the time-frame. The assumptions the framework is based on are described - there is a limit on the size of LLS location errors, the location accuracy of a LLS does not vary over small area and the continuous set of locations within the area can be approximated by the discrete variable $\boldsymbol{L}$, provide sufficiently small grid square sizes are used.

The two main equations of the Bayesian framework are then explained in terms of these variables: the posterior probability - the probability that a location $\boldsymbol{L_i}$ was struck by lightning given the reported LLS stroke $\boldsymbol{X_{lls_j}}$ - and the likelihood ratio - how much more likely it is that a location $\boldsymbol{L_1}$ was struck versus a location $\boldsymbol{L_2}$ given the reported LLS stroke $\boldsymbol{X_{lls_j}}$. Both these equations rely on the statistical models of the likelihood and prior probability functions.

The bivariate Students' t-distribution is proposed as a statistical model for the likelihood function, given the emphasis on heavier tails that it has in comparison to the bivariate Gaussian distribution. The bivariate uniform distribution is proposed as non-informative or flat prior probability function and a Gaussian Mixture Model (GMM) is proposed as a informative prior probability function. From these, a NOT likelihood function is derived - how LLS data is distributed when a location is NOT struck - so that the likelihood ratio of of a location being struck against it NOT being struck can be determined.

It is recommended to present the results of applying the framework as plots of the posterior probability and likelihood ratio against the time-stamp of each stroke reported in the area and within the time-frame of investigation.

The next chapter describes a methodology for fitting statistical models to bivariate data for the purpose of verifying the statistical models proposed for the likelihood and prior probability functions.

Chapter 6

Bivariate Statistical Modelling

This chapter describes a methodology for fitting statistical models to bivariate data (eg. Lightning Location System (LLS) stroke locations) and evaluating the quality of fit of the model. Fitting of models is done using bivariate Maximum Likelihood Estimation (MLE) and Expectation-Maximization (EM) techniques. The quality of each fit is then assessed by a difference of squares comparison of the Mahalanobis distances, given the bivariate nature of the data set.

6.1 Overview

Chapter 5 describes a number of statistical models proposed for the likelihood function and the prior probability function of the Bayesian framework. To confirm that these proposed models are indeed good choices, actual Lightning Location System (LLS) stroke data needs to be statistically modelled. Given that LLS strokes are reported as bivariate data (longitude and latitude), a methodology for fitting and assessing statistical models to bivariate data is needed.

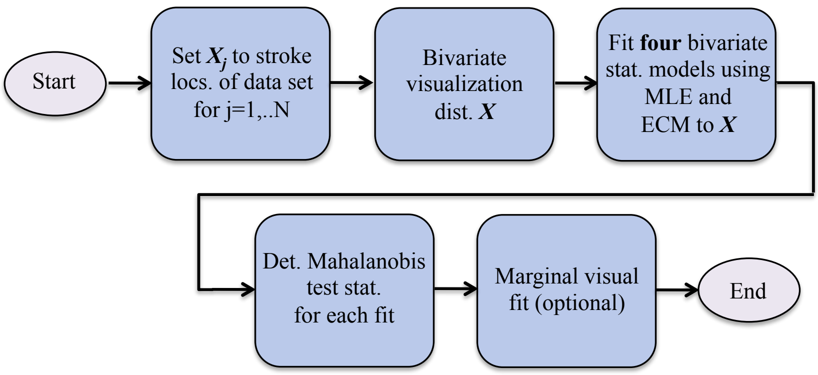


Figure 6.1: Block diagram describing the methodology used to determine the statistical model of a LLS data set.

Figure 6.1 shows a block diagram description of the bivariate statistical modelling methodology used to fit statistical models to bivariate LLS data. The chapter is organised into the four stages and fifth, optional stage of the methodology:

6.2 LLS data set $\boldsymbol{X}$.

6.3 Bivariate visualization.

6.4 Maximum Likelihood Estimation (MLE).

6.5 Mahalanobis goodness-of-fit test.

6.6 Marginal visual analysis (optional).

6.2 LLS Data set $\boldsymbol{X}$

Similarly to $\boldsymbol{X_{lls}}$ in Chapter 5, $\boldsymbol{X}$ represents a data set of M strokes detected by a LLS. While $\boldsymbol{X}$ can be any data set of LLS stroke locations, for the purposes of this investigation, it will be a data set of LLS strokes that were either time-correlated with ground-truth lightning events to tall towers (statistical model for the likelihood function) or LLS strokes reported in a specific area of investigation over a number of seasons (statistical model for the prior probability function).

$$\boldsymbol{X_j} = \begin{pmatrix} \varphi_j \\ \lambda_j \end{pmatrix} \qquad j = 1, ..., M. \tag{6.1}$$

where

φ_j : The longitudinal geographical coordinate of the detected stroke j.
λ_j : The latitudinal geographical coordinate of the detected stroke j.

6.3 Bivariate Visualization

To visualize the bivariate joint probability distribution of $\boldsymbol{X}$, the data first needs to be binned in both the longitudinal and latitudinal directions (the $\boldsymbol{\varphi}$ and $\boldsymbol{\lambda}$ marginal distributions). While it is possible for each marginal to have different bin widths, here $w_{\varphi i} = w_{\lambda i}$ where w_i is the width of a bin in degrees. The data can now be visualised in three ways:

- **Bivariate Histogram count:** This is the count or frequency of observations of the data set. Each bin value is the number of data points that fall within it (equation 6.2) and the sum of all bin values is M.

$$v_i = c_i \tag{6.2}$$

- **Probability density estimate:** This is an approximation to the bivariate probability density function of the data set. The value of each bin is the relative number of observations weighted by bin area (equation 6.3) and the sum of all bin values is 1.

$$v_i = \frac{c_i}{N \times A_i} \tag{6.3}$$

- **Empirical cumulative density:** This is an estimate of the cumulative density estimate. Each bin value is the cumulative relative number of observations (equation 6.4) and the last bin value is 1.

$$v_i = \sum_{j=1}^{i} \frac{c_j}{N} \tag{6.4}$$

where

$$\begin{aligned} v_i &= \text{the bin value.} \\ c_i &= \text{the number of data points in the bin.} \\ A_i &= w_{\varphi i} \times w_{\lambda i}, \text{ the area of each bin.} \\ N &= \text{the total number of data points in } \boldsymbol{X}. \end{aligned}$$

Each of these is a 3-dimensional plot with $w_{\varphi i}$ (longitude) by $w_{\lambda i}$ (latitude) by v_i for all bins. Given the surface nature of these plots, it is not possible to overlay fitted probability density functions of the statistical models for visual comparison.

6.4 Maximum Likelihood Estimation (MLE)

The **four** statistical models (and the number of parameters of each) that are fitted to the bivariate data $\boldsymbol{X}$ using MLE are:

- Bivariate Gaussian (normal) distribution, **five** parameters.
- Bivariate Students' t-distribution, **six** parameters.
- Bivariate Cauchy distribution, **five** parameters.
- Bivariate uniform distribution, **four** parameters.

A fifth statistical model is also fitted to $\boldsymbol{X}$ but not using MLE:

- Bivariate Gaussian Mixture Model.

1. Bivariate Gaussian distribution:

The joint probability density function of the bivariate Gaussian distribution (defined in Chapter 4, Section 4.4) has **five** parameters that need to be fitted to the data set $\boldsymbol{X}$. These are $\boldsymbol{\theta} : \mu_\varphi, \mu_\lambda, \sigma_\varphi, \sigma_\lambda, \rho_{\varphi\lambda}$ and can also be described as the mean vector and co-variance matrix shown in equation 6.5:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_\varphi \\ \mu_\lambda \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\varphi^2 & \rho_{\varphi\lambda}\sigma_\varphi\sigma_\lambda \\ \rho_{\varphi\lambda}\sigma_\varphi\sigma_\lambda & \sigma_\lambda^2 \end{pmatrix}. \tag{6.5}$$

When setting $f(\boldsymbol{X_j}|\boldsymbol{\theta})$ (equation 4.25)to the probability density function of the bivariate normal distribution, and maximising $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X})$ (equation 4.26), the maximum likelihood estimators become:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{X_j}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{j=1}^{N} (\boldsymbol{X_j} - \hat{\boldsymbol{\mu}})(\boldsymbol{X_j} - \hat{\boldsymbol{\mu}})^T \tag{6.6}$$

Which is simply the mean $\boldsymbol{\mu_X}$ and co-variance $\boldsymbol{\Sigma_X}$ of the bivariate data set $\boldsymbol{X}$.

2. Bivariate Students' t-distribution:

The common representation of the Students' t-distribution is used to determine the maximum likelihood estimators of the distribution (see Chapter 4, Section 4.4).

$$t \equiv t(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \boldsymbol{\mu} + \sqrt{\frac{\boldsymbol{\Sigma}}{\boldsymbol{q}}}\boldsymbol{Y}, \qquad \boldsymbol{Y} \sim \mathcal{N}(0, I), \qquad \boldsymbol{q} \sim \chi_\nu^2/\nu. \tag{6.7}$$

Given this representation, and using a weighted least squares procedure (Aeschliman et al., 2010, Lange et al., 1989), the maximum likelihood estimators become:

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{j=1}^{N} w_j \boldsymbol{X_j}}{\sum_{j=1}^{N} w_j}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{j=1}^{N} w_j(\boldsymbol{X_j} - \hat{\boldsymbol{\mu}})(\boldsymbol{X_j} - \hat{\boldsymbol{\mu}})^T \tag{6.8}$$

where

$$w_j = (2 + \nu)(\nu + (\boldsymbol{X_j} - \hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X_j} - \hat{\boldsymbol{\mu}}))^{-1} \tag{6.9}$$

Unlike the bivariate Gaussian distribution, these equations do not have a closed form solution and require iterative methods or approximations to solve (Lange et al., 1989). The method used here is a modified version of that described by Liu and Rubin (1995). Liu and Rubin describe how Expectation-Maximization (EM) and Expectation-Conditional-Maximization (ECM) algorithms can be used to fit multivariate Students' t-distributions and determine the parameters that maximise the likelihood function.

Equation 6.8 and 6.9 provide simple **E-** and **M-Steps** that allow for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ to be determined for a fixed ν. Beginning at iteration $r = 0$, with initial conditions set $\hat{\boldsymbol{\mu}}^{(0)} = \text{mean}(\boldsymbol{X})$ and $\hat{\boldsymbol{\Sigma}}^{(0)} = \text{cov}(\boldsymbol{X})$ and ν. Then:

- **E-Step**$^{(r+1)}$**:** $w_j^{(r+1)}$ is evaluated using $\hat{\boldsymbol{\mu}}^{(r)}$, $\hat{\boldsymbol{\Sigma}}^{(r)}$ and ν (equation. 6.9).
- **M-Step**$^{(r+1)}$**:** $\hat{\boldsymbol{\mu}}^{(r+1)}$ and $\hat{\boldsymbol{\Sigma}}^{(r+1)}$ are evaluated using $w_j^{(r+1)}$ (equation. 6.8) and the likelihood is calculated for $\hat{\boldsymbol{\mu}}^{(r+1)}$, $\hat{\boldsymbol{\Sigma}}^{(r+1)}$ and ν.

These steps continue to iterate until the likelihood converges to a maximum. However, this only determines the parameters given a fixed ν. To estimate ν as well, the **M-Step** needs to be conditionalized and the ECM and ECME algorithm are recommended by Liu and Rubin (1995). These algorithms are recommended as they are computationally efficient but since efficiency is not necessary for this methodology, a simpler method of iterating through values of ν is used.

Figure 6.2 shows the modified EM algorithm used to estimate ν. The algorithm begins with same initial conditions but iterator $p = 0$ as well and $\nu^{(0)}$ set very small. The **E-Step** and **M-Step** then iterate with $\nu^{(p)}$ fixed until the likelihood converges. The likelihood is then set to $\mathcal{L}^{(p)}$ and these parameters are compared with the initial conditions. If they are equal, this means the $\nu^{(p)}$ must be very large since the distribution is approximating a bivariate Gaussian function. Once this is reached, the p value for which $\mathcal{L}^{(p)}$ is the greatest ($pmax$) is determined and the maximum likelihood estimates are $\hat{\boldsymbol{\mu}}^{(pmax)}$, $\hat{\boldsymbol{\Sigma}}^{(pmax)}$ and $\nu^{(pmax)}$. The source code implementation of the algorithm can be found in Appendix C.

3. Bivariate Cauchy distribution:

The bivariate Cauchy distribution can be considered a special case of the the bivariate Students' t-distribution - they are equivalent when the degrees of freedom parameter of the bivariate Students' t-distribution is 1 ie. $\nu = 1$. Given this, the maximum likelihood estimation of the bivariate Cauchy distribution can be performed using the **E-** and **M-Step** discussed previously. Here, since $\nu = 1$ is fixed, the **E-** and **M-Step** can simply be iterated until the likelihood converges.

4. Bivariate uniform distribution:

As discussed in Chapter 4, the bivariate uniform distribution uses the maximum and minimum as parameters but they can be written as similar parameters to the bivariate Gaussian ($\boldsymbol{\theta} : \mu_\varphi, \mu_\lambda, \sigma_\varphi, \sigma_\lambda$) except that the bivariate uniform distribution has no correlation coefficient (or $\rho_{\varphi\lambda} = 0$) and therefore, only has **four** parameters that need to be

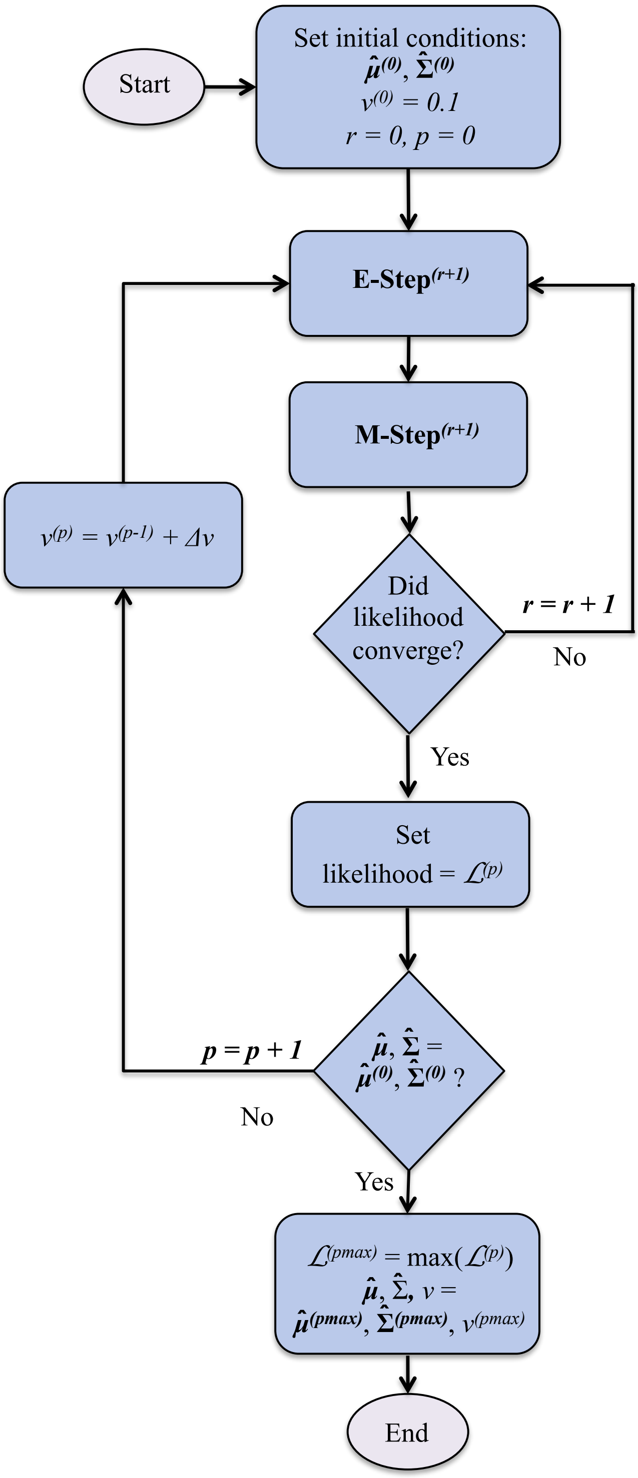


Figure 6.2: Modified EM algorithm to determine best fitting Students' t-distribution parameters for data set $\boldsymbol{X}$.

estimated by MLE. This is because, if the correlation coefficient was greater or less than zero, this would imply that one of the marginals was not uniform, meaning the model would not be a bivariate uniform distribution. The maximum likelihood estimators for the bivariate uniform distribution are:

$$\hat{\boldsymbol{\mu}} = \frac{1}{2}\begin{pmatrix} \max(\boldsymbol{\varphi}) - \min(\boldsymbol{\varphi}) \\ \max(\boldsymbol{\lambda}) - \min(\boldsymbol{\lambda}) \end{pmatrix} \tag{6.10}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{12}\begin{pmatrix} (\max(\boldsymbol{\varphi}) - \min(\boldsymbol{\varphi}))^2 & 0 \\ 0 & (\max(\boldsymbol{\lambda}) - \min(\boldsymbol{\lambda}))^2 \end{pmatrix}. \tag{6.11}$$

where $\boldsymbol{\varphi} = [\varphi_1, \varphi_2, ..., \varphi_M]$ and $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ..., \lambda_M]$.

5. Gaussian Mixture Model (GMM):

Unlike the other **four** statistical models, the GMM is not fitted to $\boldsymbol{X}$ using MLE. Instead, the fit is already defined where the number of mixture components is decided by the number of strokes in $\boldsymbol{X}$, M. The bivariate Gaussian parameters for the co-variance matrix $\hat{\boldsymbol{\Sigma}}$ fitted using MLE then become the co-variance matrix for each of the mixture components in the GMM. The mean of each mixture component simply becomes $\boldsymbol{\mu_j} = \boldsymbol{X_j}$ and the weighting factor is equal across all mixture components $w_j = \frac{1}{M}$. Even though this fit is not achieved using MLE, the quality of it can still be assessed as follows.

6.5 Goodness of Fit - Mahalanobis test

For the purposes of this methodology, a bivariate goodness-of-fit test proposed by McAssey (2013) is adopted. The test is empirically based (meaning it compares the observed data set to data sampled from the ideal model, rather than the ideal model itself) and uses the Mahalanobis distance to compare models with the observed data. Since the Mahalanobis distance features in the probability density function of the bivariate Gaussian distribution and the bivariate Students' t-distribution, and can be defined easily for the bivariate uniform distribution and a GMM, this test (referred to from now on as the Mahalanobis test) is ideal for this methodology.

From equation 4.18 in Chapter 4, Section 4.4, the Mahalanobis distance for $\boldsymbol{X}$ can be defined as:

$$\boldsymbol{\Delta_{jX}} = \sqrt{(\boldsymbol{X_j} - \boldsymbol{\mu_X})^T \boldsymbol{\Sigma_X^{-1}} (\boldsymbol{X_j} - \boldsymbol{\mu_X})} \quad \text{for} \quad j = 1, ..., M. \tag{6.12}$$

where

$\boldsymbol{X_j}$: the geographic location of stroke j.
$\boldsymbol{\mu_X}$: the mean of the bivariate data set $\boldsymbol{X}$.
$\boldsymbol{\Sigma_X}$: the co-variance matrix of the bivariate data set $\boldsymbol{X}$.
$\boldsymbol{\Delta_{jX}}$: the Mahalanobis distance of stroke $\boldsymbol{X_j}$.

One can consider the Mahalanobis distance of bivariate data as an elliptical description of distance from the mean. The test is based on the premise that, given a probability density, the Mahalanobis distances for data generated from this distribution will follow a distinct distribution itself - in short, allowing us to convert bivariate data to a univariate case. From this, more common test statistics can be applied.

Figure 6.3 shows a block diagram of the steps involved in the Mahalanobis test:

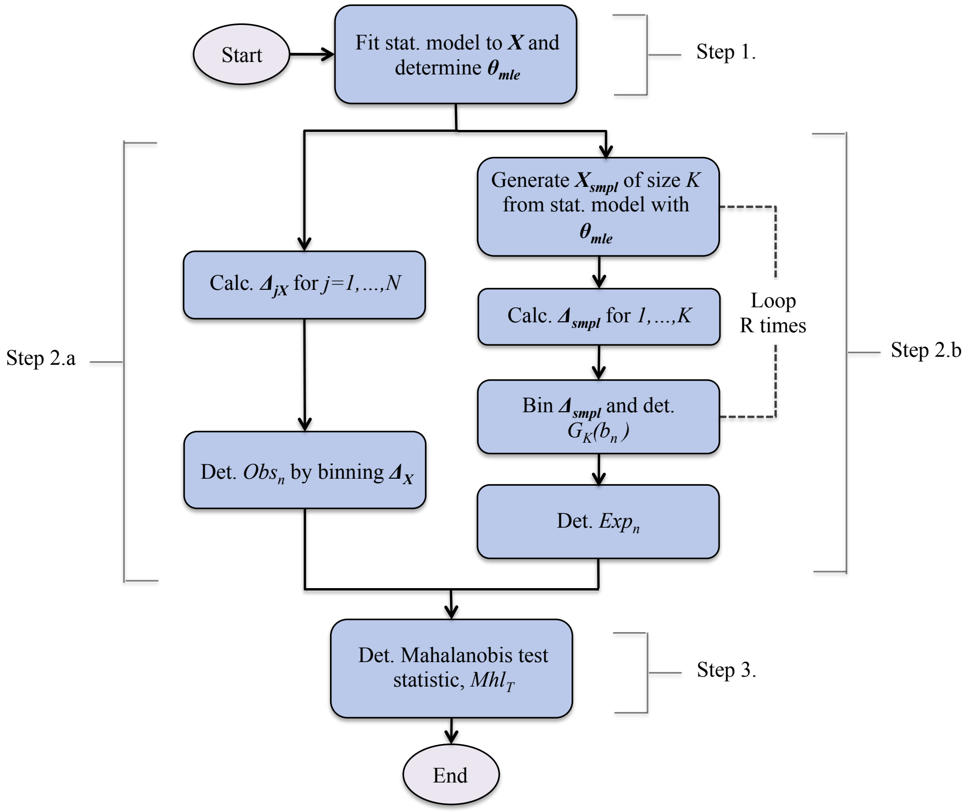


Figure 6.3: Block diagram showing the Mahalanobis goodness-of-fit test procedure.

The source code implementation of the Mahalanobis test can be found in Appendix C.

Step 1. MLE of $\boldsymbol{X}$

The first step of the test procedure is to fit the parameters ($\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}}$) of the statistical model to $\boldsymbol{X}$ using the MLE techniques described above. The full procedure needs to be applied for each of the **five** statistical models discussed above. *Step 2.a* and *Step 2.b* follow next and can occur in parallel.

Step 2.a Observed Counts

Calculate the Mahalanobis distance $\Delta_{1X}...\Delta_{N_X}$ for $\boldsymbol{X}$ using equation 6.12. Next, a set of bins (or bin edges) needs to be defined, $b_n = 0, ..., b_T$ where $b_T >> \max(\Delta_X)$ (we use $10,000$) and the bin width is b_w (we use 0.2). Therefore, the total number of bins, $T = \frac{b_T}{b_w}$ (50,000). We then determine the cumulative count of the observed ($\boldsymbol{X}$) Mahalanobis distances using:

$$Obs_n = \sum_{j=1}^{N} \mathbb{1}(\boldsymbol{\Delta_{jX}} \leq b_n), \qquad b_n > 0 \qquad \text{for} \quad n = 1, ..., T. \tag{6.13}$$

where $\mathbb{1}$ is the indicator function ie. 1 if the condition is met, 0 otherwise and n is the bin number.

Step 2.b Expected Counts

Next (or simultaneously) the expected counts (Exp_n) need to be determined. This is done by first generating a bivariate data sample ($\boldsymbol{X_{smpl}}$) of size K from the statistical model being tested with parameters set to the MLE parameters ($\boldsymbol{\hat{\theta}_{mle}}$) determined in step one. If K is large enough (we use 10,000), this bivariate sample will approximate the ideal model.

Using equation 6.12, but with $\boldsymbol{\mu_{X_{smpl}}}$ and $\boldsymbol{\Sigma_{X_{smpl}}}$, we calculate the Mahalanobis distances $\Delta_{1X_{smpl}} ... \Delta_{KX_{smpl}}$ for $\boldsymbol{X_{smpl}}$. Next, we determine the empirical cumulative distribution ($G_K(b_n)$) of $\boldsymbol{\Delta_{X_{smpl}}}$ using the same binning scheme as for the observed counts:

$$G_K(b_n) = \frac{1}{K} \sum_{s=1}^{K} \mathbb{1}(\boldsymbol{\Delta_{sX_{smpl}}} \leq b_n), \qquad b_n > 0 \qquad \text{for} \quad n = 1, ..., T. \tag{6.14}$$

Given that this step relies on generating random samples, it is advised to repeat the step R times and take the average of $G_K(b_n)$ to provide a consistent result. A typical value of 100 is used for R, although, the greater this value, the more accurate the test will become. The expected cumulative counts can now be calculated by multiplying by the sample size of $\boldsymbol{X}$:

$$Exp_n = M \times G_K(b_n) \qquad \text{for} \quad n = 1, ..., T. \tag{6.15}$$

Step 3. Test Statistic

Now that the observed and expected counts of the Mahalanobis distance have been determined, a test statistic may be applied. This could be any common statistic (Pearson's chi-square, Kolmogorov-Smirnoff etc.) but for the purposes here, we use the statistic given by McAssey (2013):

$$Mhl_T = \sum_{n=1}^{T} \frac{|Exp_n - Obs_n|}{Exp_n} \tag{6.16}$$

The statistical model which yields the smallest Mhl_T indicates the best fitting model.

Figure 6.4 shows an example of the empirical cumulative distributions (equation 6.14) of the Mahalanobis distance for the **five** statistical models fitted to the same data set. They are fitted to a bivariate Students' t-distribution with $\boldsymbol{\mu} = \binom{0}{0}$, $\boldsymbol{\Sigma} = I$ (the identity matrix) and $\nu = 2$. As can be seen, the empirical cumulative distribution of the Mahalanobis distance differs greatly for the **five** statistical distributions. The uniform distribution rises the slowest in the beginning, but reaches 1 first and the Cauchy distribution ($\nu = 1$) rises the fastest at first, but then takes the longest to reach 1 - indicating the acceptance of extreme outliers (or the heavier tails) of the distribution.

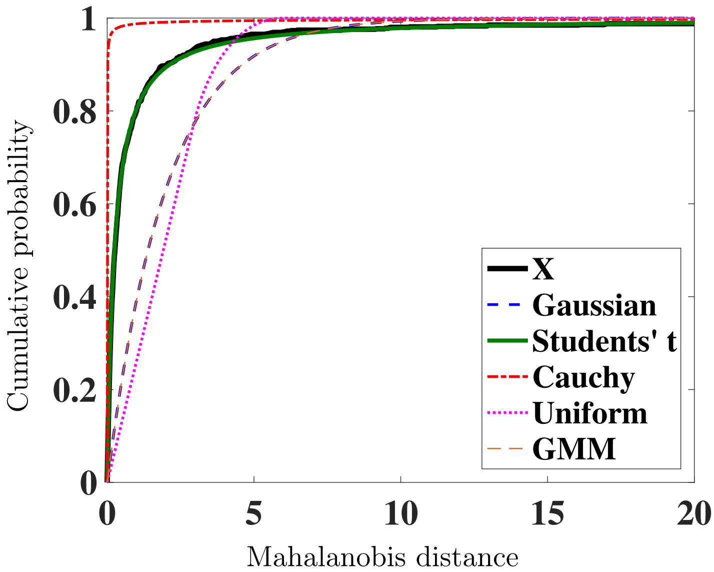


Figure 6.4: Example of Cumulative probability vs. Mahalanobis distances for different statistical models with parameters fitted to data set $\boldsymbol{X}$. $\boldsymbol{X}$ is data from a bivariate Students' t-distribution with $\boldsymbol{\mu} = \binom{0}{0}$, $\boldsymbol{\Sigma} = I$ and $\nu = 2$.

6.6 Marginal Visual Analysis (optional)

While the bivariate statistical models will have been fitted to $\boldsymbol{X}$ already, it is interesting to visualize the fits. This is not possible in the 3-Dimensional view provided by the bivariate visualization as plotting fitted probability density functions over the bivariate histograms would lead to visually confusing images.

It is possible instead, to fit univariate distributions to each of the the marginal distributions of $\boldsymbol{X}$ except for the GMM. This is because, for the bivariate Gaussian, the bivariate Students' t-distribution and the uniform distribution, the marginals all follow the respective distributions. In the case of the GMM though, this is not so. It should be noted though, that while the marginal distributions can be fitted with univariate statistical models, the fitted parameters will *not* be the same as the bivariate case since correlation between marginals is not accounted for. Therefore, this is purely for visual purposes.

Once the **four** univariate distributions have been fitted to both $\boldsymbol{\varphi}$ and $\boldsymbol{\lambda}$, the probability density function of each fit can be plotted against the estimated probability density function and the cumulative density function against the empirical cumulative density function respectively. The quality of the fit can then be visually assessed. Figure 6.5 shows an example of this. The data $\boldsymbol{x}$ is generated from a univariate Students' t-distribution with $\mu = 0$, $\sigma = 1$ and $\nu = 2$. The **four** univariate statistical models are fitted and the probability density estimate (figure 6.5a) and empirical cumulative density function (figure 6.5b) are shown.

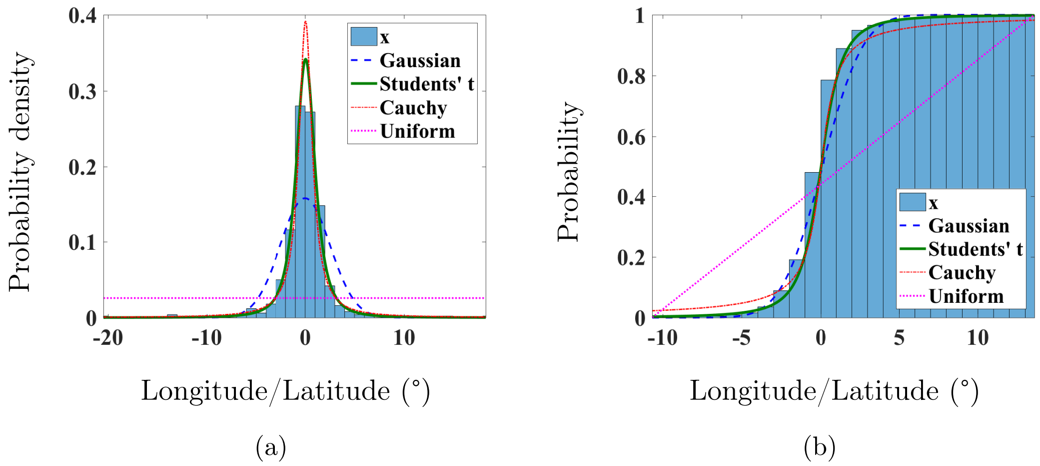


Figure 6.5: Example of marginal visual analysis with different statistical models fitted to (a) probability density estimate of $\boldsymbol{x}$ (b) Empirical cumulative density plot of $\boldsymbol{x}$. $\boldsymbol{x}$ is data from a univariate Students' t-distribution with $\mu = 0$, $\sigma = 1$ and $\nu = 2$.

6.7 Chapter Summary

This chapter has described the bivariate statistical modelling methodology used for fitting statistical models to bivariate Lightning Location System (LLS) stroke report data and assessing the quality of the fits.

The methodology uses Maximum Likelihood Estimation (MLE) to fit **four** bivariate statistical models (Gaussian, Students' t-, Cauchy and uniform) to LLS stroke report data and the maximum likelihood estimators for each of these statistical models is described. A modified version of the Expectation-Conditionalization-Maximization (ECM) algorithm employed by Liu and Rubin (1995) to solve the maximum likelihood estimators for the bivariate Students' t-distribution is described. A method for fitting a Gaussian Mixture Model (GMM) to the bivariate LLS data set is also presented.

Assessing the quality of the fitted statistical models is done by using the method proposed by McAssey (2013), in which the cumulative probability of the Mahalanobis distance of data sampled from each statistical model is compared with the cumulative probability of the Mahalanobis of the empirical LLS data. A test statistic is defined and the fitted model with the smallest value constitutes the best fit.

The next chapter describes ground-truth data time-correlated with LLS data in South Africa and Austria used to verify the proposed likelihood and prior probability functions by applying the bivariate statistical modelling methodology to the data sets.

Chapter 7

Ground-Truth Verification of Statistical Models

This chapter discusses the ground-truth data sets used to verify the proposed likelihood function and prior probability function of the Bayesian framework. LLS stroke reports are time-correlated with ground-truth lightning events to a tall tower in South Africa - the Brixton tower and in Austria - the Gaisberg tower. The methodology described in Chapter 6 is applied to these LLS data sets and the results of which statistical models best fit the LLS data are presented.

7.1 Overview

To verify the statistical models proposed for the likelihood function and the prior probability function in Chapter 5, data sets of Lightning Location System (LLS) stroke reports are necessary. A data set of LLS stroke reports correlated with ground-truth lightning events to a known location is necessary to verify the statistical models proposed for the likelihood function and a data set of LLS stroke reports over a significantly large region (Chapter 5, Section 5.2.1) is necessary to verify the statistical models proposed for the prior probability function. In this chapter, three LLS data sets from two different countries are presented in the following sections:

7.2 South Africa, Brixton tower and the Johannesburg region

- $\boldsymbol{SA_{Br}}$: data set to verify the likelihood models.
- $\boldsymbol{SA_{Jhb}}$: data set to verify prior probability models.

7.3 Austria, Gaisberg tower

- $\boldsymbol{Aus_{GBT}}$: data set to verify the likelihood models.

The chapter is structured as follows: Firstly, the South African data sets $\boldsymbol{SA_{Br}}$ and $\boldsymbol{SA_{Jhb}}$ are presented and the ground-truth methodologies to obtain them are described. The methodology described in Chapter 6 is then applied to both of the data sets. Secondly, the Austrian data set $\boldsymbol{Aus_{GBT}}$ and the ground-truth investigation is presented and the methodology described in Chapter 6 is applied. The results of the quality of fit for each of the three data sets is then summarised.

7.2 South Africa

Johannesburg is the main economic city in South Africa, and is located in the North-East of the country in the Gauteng province. Flash densities of 10-15 $flashes \times km^{-2} \times year^{-1}$ have been recorded in the area, which makes Johannesburg an unusual city where the main industrial and economic center is in a relatively high flash density area. The city itself has two tall communications towers - the Hillbrow tower and the Brixton tower. Figure 7.1 shows the city skyline from a northern location looking south. Both towers are approximately 250 m tall with almost 5 km between them. The Hillbrow tower is located at 26.1868° south and 28.0494° east, and is situated in the city centre, surrounded by numerous other buildings. The Brixton tower is located at 26.1925° south and 28.0068° east, and is slightly outside the main city centre being surrounded by residential housing only. Both towers are located on a ridge that runs from West to East through the city.

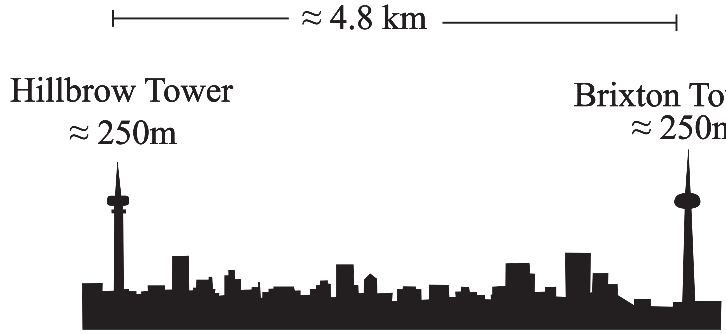
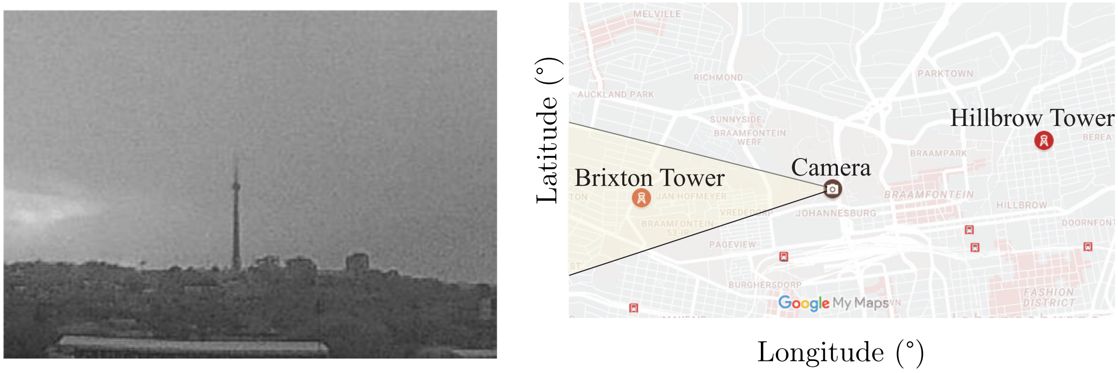


Figure 7.1: The Johannesburg skyline, as viewed from a northern location looking south. Both the Hillbrow and the Brixton tower are visible.

7.2.1 Brixton Tower Photographic Study

Since 2009, a surveillance system has been in operation monitoring the Brixton tower to photograph lightning attachments to the tower. The system initially consisted of a simple motion-triggered camera located east of the tower but was later upgraded to include multiple cameras. Figure 7.2a shows the westward looking view of the Brixton tower as seen by the camera system. It has an approximate view angle of 55° with 1 km either side of the tower in view. The camera system is located at the University of the Witwatersrand which is 1.88 km east of the Brixton tower (approximately half way between the Brixton tower and the Hillbrow tower). Figure 7.2b shows the location of the camera relative to the Brixton tower.



(a) (b)

Figure 7.2: Motion-triggered camera surveillance of Brixton tower (a) view of Brixton tower as seen by the camera system (b) location of camera relative to the location of Brixton tower.

The system operated from 2009 until 2013 but was shutdown after the 2012 - 2013 lightning season. It was re-initiated again in 2015 but with some modifications:

- **2009 - 2013 seasons:** A motion-triggered camera (a photograph is taken when there is a change in the tonality of the observed pixels, the sensitivity of which may be set) was operated from the University of Witwatersrand. The camera could operate at 30 frames-per-second but typically operated between 10-15 frames-per-second and therefore, had an approximate resolution of 60-100 ms and was able to capture numerous images of a lightning flash. However, the resolution was not enough to distinguish the number of individual strokes that constitute a flash. The camera was time-synchronised to the Network Time Protocol (NTP) server of the institution.
- **2015 - 2016 season:** For this season, the cameras used were simple web cameras and the motion-triggering was performed using using software (iSpy®) rather than the camera itself. This allowed for multiple cameras to be setup and improve the efficiency of the surveillance system. Similar frame rates were achieved, allowing for flashes to be photographed but not for individual strokes to be distinguished. This system was also time-synchronised to the NTP server of the institution. The setup was installed and became operational 1 October 2015 and was turned off at the end of the lightning season in March 2016.

Figure 7.3 shows some examples of the type of images captured by the surveillance camera system. Figures 7.3c, 7.3b and 7.3a show photographs of upward lightning events to the Brixton tower captured on the 20-11-2015 at 09:43:24, 10:24:48 and 10:54:37 South African Standard Time (SAST) (UTC + 02:00) respectively.

Table 7.1 shows a summary of all the photographed lightning attachments to the Brixton tower from the period of 2009 - 2013 and 2015 - 2016 with a total of 121 flashes to the Brixton photographed. Based on visual observation, the events are categorised into downward and upward lightning attachments. As expected of a tall tower, there are

(a) 20-11-2015 09:43:24 SAST

(b) 20-11-2015 10:24:48 SAST

(c) 20-11-2015 10:54:37 SAST

Figure 7.3: Examples of upward lightning events from the Brixton tower captured by the surveillance camera system on 20-11-2015 at (a) 09:43:24 SAST (b) 10:24:48 SAST (c) 10:54:37 SAST

significantly more upward events (104) than downward events (16). Also notable, is the much larger number (55) of photographed events in the 2015 - 2016 season. This is likely due to the improved surveillance system with multiple cameras allowing for a higher detection efficiency. For this period, the 55 flashes that attached to the tower occurred on 13 different days, meaning that the tower was often struck multiple times in one day. The full details (time, number, type) and photographs are included in table B.2 in Appendix B.

Table 7.1: Photographed Lightning Flashes to Brixton Tower

Season	No. Upward Events	No. Downward Events	Total No. Flashes
2009 - 2010	12	0	12
2010 - 2011	24	2	26
2011 - 2012	6	1	7
2012 - 2013	17	4	21
2015 - 2016	46	9	55
All	105	16	121

7.2.2 SALDN Data

The South African Lightning Detection Network (SALDN) was in operation over the seasons the photographic study was performed (2009-2013, 2015-2016) and has complete coverage over the Johannesburg area. For a 1600 km^2 area around the location of the

Brixton tower (sufficiently large enough to be greater than any location errors), Table 7.2 shows the number of strokes reported by the SALDN for each season. A total of 364705 strokes were detected in the region during the periods of the photographic study and this data set will be referred to as $\boldsymbol{SA_{Jhb}}$. It should be noted that while the quantities are large, these are detected strokes and have not been grouped into flashes. As such, these numbers are not an indication of ground-flash density.

Table 7.2: SALDN stroke reports within 1600 km^2 area around Johannesburg: $\boldsymbol{SA_{Jhb}}$

Season	Stroke Count
2009 - 2010	93077
2010 - 2011	75023
2011 - 2012	80545
2012 - 2013	49243
2015 - 2016	66817
All	364705

Given that each photograph of a lightning flash was time-stamped (synchronised with NTP server), these times were compared with the reported time-stamps of detected SALDN strokes. Strokes with reported time-stamps that correlated with the time-stamp of the photographs were selected as SALDN detections of events that attached to Brixton tower - henceforth referred to as $\boldsymbol{SA_{Br}}$. It should be noted that, given the camera resolution not allowing for individual stroke distinctions from the photographs, a group (flash) of reported strokes were time-correlated to each photograph. Table 7.3 shows the number of photographed flashes that were detected by the SALDN and how many reported strokes were time-correlated with the photographs for each of the lightning seasons. The complete, detailed table for each photograph from 2015-2016 can be found in Appendix B (table B.3).

Table 7.3: Time-correlated SALDN data with photographed flashes that attached Brixton tower: $\boldsymbol{SA_{Br}}$

Season	No. Photo Flashes	No. Detected by SALDN	Total No. Strokes
2009 - 2010	12	6	14
2010 - 2011	26	13	68
2011 - 2012	7	4	20
2012 - 2013	21	16	83
2015 - 2016	55	30	143
Total	121	69	328

This results in two data sets of LLS strokes ($\boldsymbol{SA_{Br}}$ and $\boldsymbol{SA_{Jhb}}$) that the bivariate statistical modelling methodology described in Chapter 6 can be applied to.

7.2.3 Bivariate Statistical Modelling: $\boldsymbol{SA_{Br}}$

$\boldsymbol{SA_{Br}}$ is the set of SALDN strokes that were detected and reported when the Brixton was struck by lightning with certainty (ground-truth photographs). Therefore, the distribution of the reported locations of $\boldsymbol{SA_{Br}}$ will be a description of how the location errors of the SALDN are distributed - the statistical model that best describes this data set will be the best choice of statistical model for the likelihood function described in Chapter 5. We now apply the bivariate statistical modelling methodology described in Chapter 6:

Stage 1. Data set $\boldsymbol{X} = \boldsymbol{SA_{Br}}$

Firstly, we let the data set $\boldsymbol{X}$ be the reported locations of all the strokes attaching to the Brixton tower that were detected by the SALDN for all 328 strokes in $\boldsymbol{SA_{Br}}$.

$$\boldsymbol{X_j} = \begin{pmatrix} \varphi_j \\ \lambda_j \end{pmatrix} = \boldsymbol{SA_{jBr}} \qquad j = 1, ..., 328. \tag{7.1}$$

Figure 7.4 shows an aerial view of latitude against longitude with the reported stroke locations indicated. The location of the Brixton tower can be found in the center at $(28.0068°, -26.1925°)$. As can be seen, the data is clustered around the location of the Brixton tower but a few strokes have fairly large location errors to the southwest of the tower.

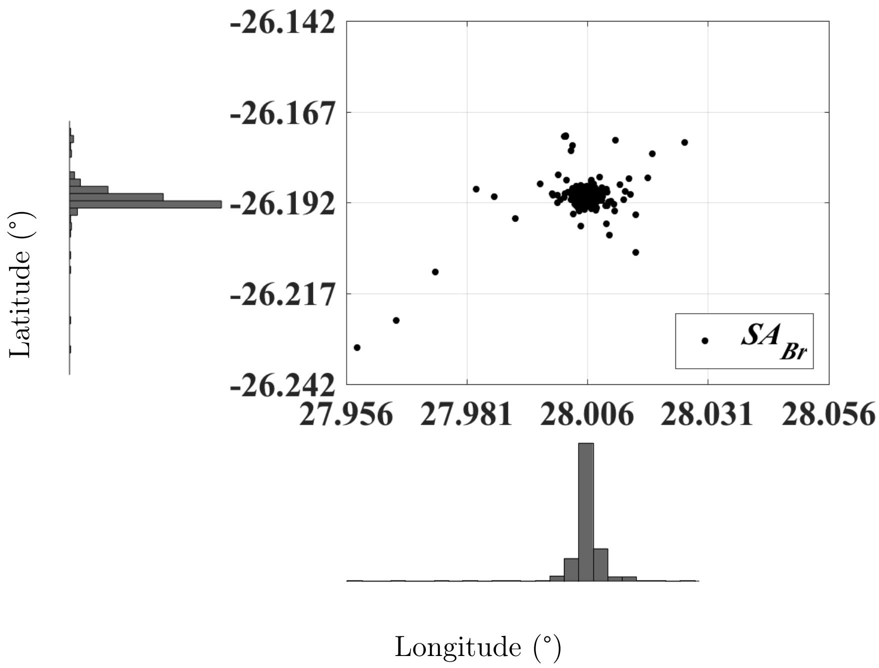


Figure 7.4: Aerial view of latitude vs. longitude for SALDN strokes time-correlated with events to the Brixton tower $\boldsymbol{SA_{Br}}$.The Brixton tower is located at $(28.0068°, -26.1925°)$. The marginal histograms in the latitudinal and longitudinal directions are also shown.

The marginal histograms are also shown in the figure and it appears that neither are symmetrical, implying some correlation between the longitudinal and latitudinal bivariate data. A proper visualization of the bivariate data is needed.

Stage 2. Bivariate visualization $\boldsymbol{SA_{Br}}$

Figures 7.5a-b, 7.5c-d and 7.5e-f show a bivariate histogram count, probability density estimate and empirical cumulative density plot with bin size 0.0014° × 0.0014° for the data set $\boldsymbol{SA_{Br}}$ respectively. The data is heavily clustered around the mode with the single bin at the mode containing the majority of the data. Nonetheless, some outliers are present giving the cumulative density function a steep gradient in both the longitudinal and latitudinal directions.

Stage 3. Maximum Likelihood Estimation $\boldsymbol{SA_{Br}}$

Table 7.4 shows the parameters for the **four** bivariate statistical models fitted to $\boldsymbol{SA_{Br}}$ using the Maximum Likelihood Estimation (MLE) techniques discussed in Chapter 6, Section 6.4.

Table 7.4: Parameters for the statistical models fitted to $\boldsymbol{SA_{Br}}$ using MLE.

Parameter	Gaussian	Students' t	Cauchy	Uniform	GMM
μ_{Br_φ}	29.0066°	29.0069°	29.0069°	27.993°	29.0066°
μ_{Br_λ}	−26.1915°	−26.1922°	−26.1922°	−26.2033°	−26.1915°
σ_{Br_φ}	0.0051	0.000607	0.000682	0.0196	0.0051
σ_{Br_λ}	0.0045	0.000568	0.000645	0.0169	0.0045
$\rho_{Br_{\varphi\lambda}}$	0.5395	0.0994	0.1113	-	0.5395
ν_{Br}	-	0.8318	1	-	-
loglik.	2632	3163	3161	-	-

Firstly, the parameters μ_{Br_φ}, μ_{Br_λ} are very close to the location of the Brixton tower with only slight differences, except for the uniform distribution. The fitted bivariate Gaussian distribution has a correlation coefficient of 0.5395 which was expected given the lack of symmetry in the marginal distributions. It is notable how much smaller the variances of the fitted bivariate Students' t-model are in comparison to those of the fitted bivariate Gaussian model - demonstrating that the bivariate Students' t-distribution is a "scaled" form of the bivariate Gaussian distribution. The parameters of the bivariate Cauchy model are very similar to those of the bivariate Students' t-model and, given that the degrees of freedom parameter $\nu_{Br} \approx 1$ for the bivariate Students' t-model, this follows.

Figure 7.6 shows a plot of the negative log-likelihood against the degrees of freedom parameter (ν_{Br}) of the bivariate Students' t-distribution. For each value of ν_{Br}, the rest of the model parameters are assessed to give the best fit. This demonstrates the modified Expectation-Maximization (EM) algorithm discussed in Chapter 6 where ν_{Br} is held constant, the other parameters of the bivariate Students' t-model are determined

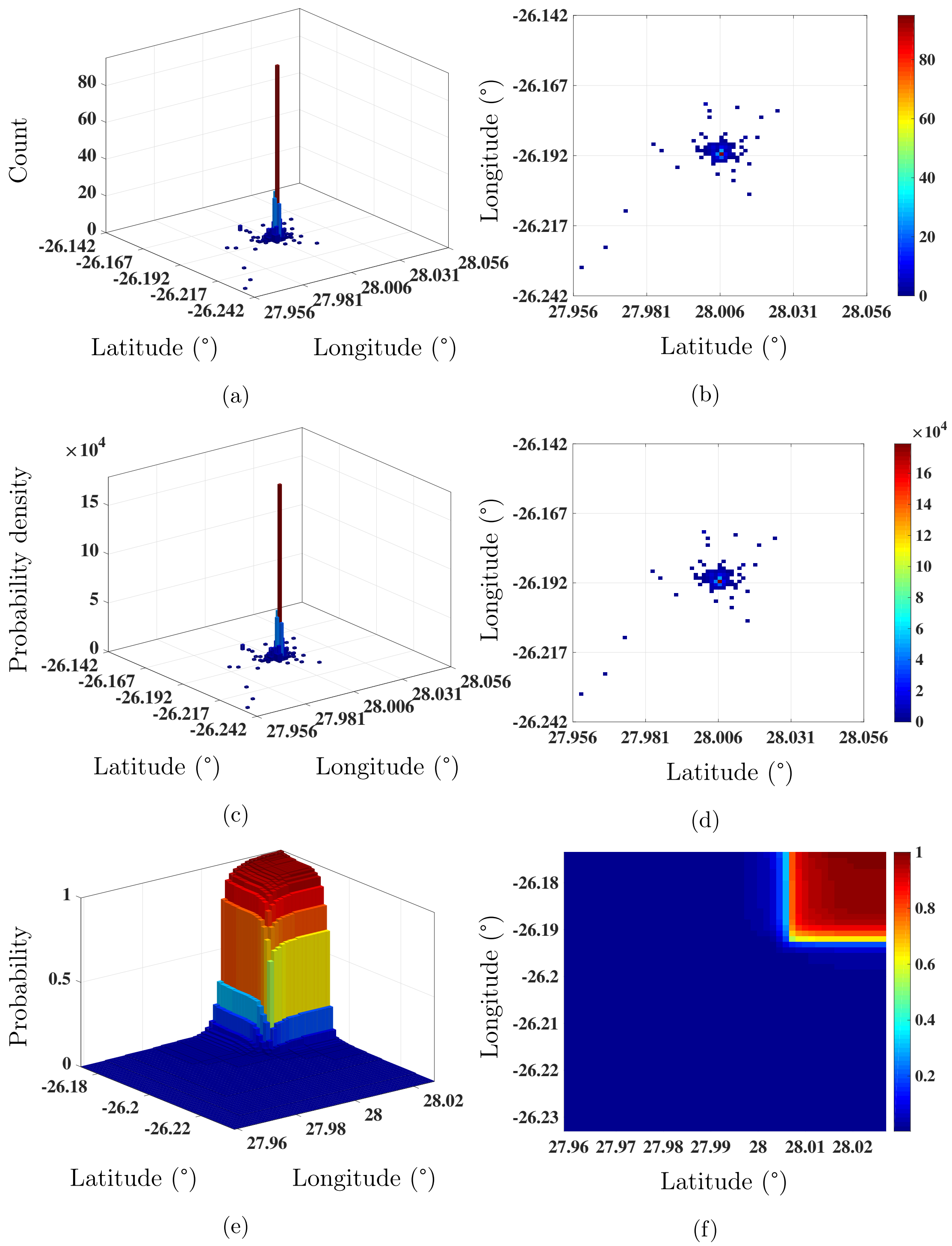


Figure 7.5: Bivariate visualization of data set $\boldsymbol{SA_{Br}}$: (a-b) Histogram count. (c-d) Probability density estimate. (e-f) Empirical cumulative density. All plots with bin size $0.0014^{\circ} \times 0.0014^{\circ}$.

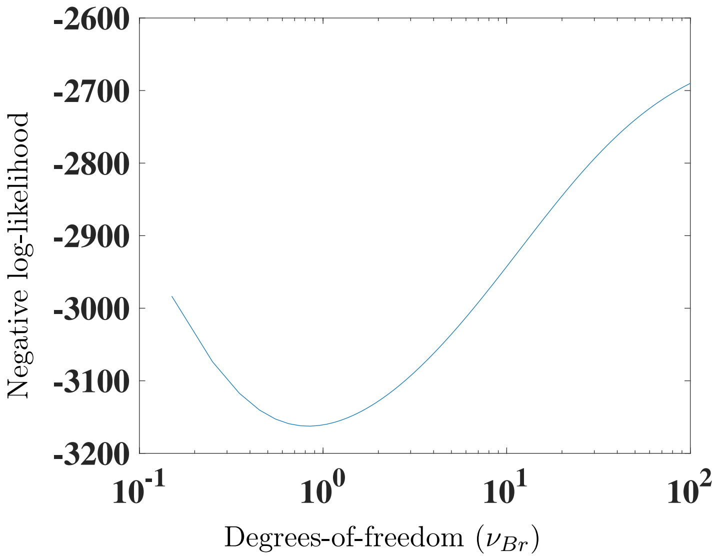


Figure 7.6: Negative log-likelihood vs. degrees of freedom parameter for bivariate Students' t-distribution, ν_{Br}. When $\nu_{Br} = 1$, the fitted parameters are equivalent to a bivariate Cauchy distribution and as $\nu_{Br} \rightarrow inf$, the fitted parameters are equivalent to those of the bivariate Gaussian distribution.

using EM and the negative-loglikelihood is calculated. ν_{Br} is then varied over a range of values. We can see the best fit for ν_{Br} from the lowest negative log-likelihood value.

Included in table 7.4, are the parameters of the Gaussian Mixture Model (GMM) although this is not fitted using MLE. Using the method described in Chapter 6, Section 6.4, we use the parameters of the fitted bivariate Gaussian model while varying the mean parameters $\mu_{Br_\varphi} = \varphi_{1-328}$, $\mu_{Br_\lambda} = \lambda_{1-328}$ (ie. the reported location of each stroke from $1 - 328$.). In this case, this simply reduces to the bivariate Gaussian fit equivalent to the MLE fitted bivariate Gaussian model.

Stage 4. Goodness-of-fit $\boldsymbol{SA_{Br}}$

Figure 7.7 shows the cumulative probability of the Mahalanobis distance for each of the fitted bivariate statistical models as well as for the empirical cumulative distribution of $\boldsymbol{SA_{Br}}$. We can see that the empirical data appears in between the bivariate Gaussian, GMM and uniform curve and the bivariate Students' t- and Cauchy curve. While the empirical curve does not rise as quickly as the bivariate Students' t- and the Cauchy model, it has a similar characteristic in that it does not reach 1 until much later (indicating the presence of extreme values/heavier-tails). In contrast, the bivariate Gaussian, GMM and uniform curves rise slower initially, but then reach 1 at much lower values of Mahalanobis distance.

Table 7.5 shows the results of Mahalanobis test for the **five** fitted statistical models. Also included are the negative log-likelihood values, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The bivariate Students' t- and

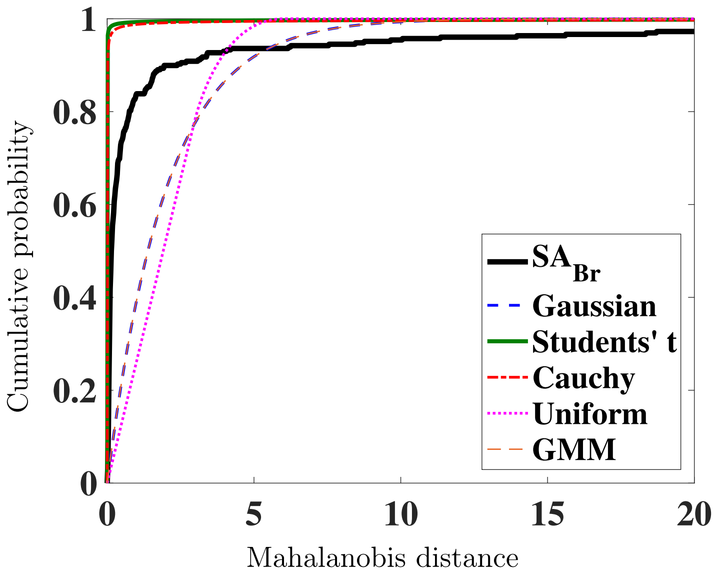


Figure 7.7: Cumulative probability vs. Mahalanobis distance for the different statistical models compared with the empirical Mahalanobis distance of $\boldsymbol{SA_{Br}}$.

Cauchy models have the lowest values across all metrics and are the best fit to the data with the uniform distribution being the worst description of the data.

Table 7.5: Mahalanobis test results for $\boldsymbol{SA_{Br}}$. Lowest values indicate best fit.

Bivariate Distribution	Mhl_T	Neg. Loglik.	AIC	BIC
Gaussian	23.0719	−2632	−5254	−5235
Students' t	17.8075	−3163	−6313	−6290
Cauchy	17.1153	−3161	−6312	−6293
Uniform	38.4571	-	-	-
GMM	23.0119	-	-	-

The bivariate Students' t-distribution therefore appears to be the best statistical model for the distribution of SALDN location errors, given the data set of reported strokes when the Brixton tower is stuck $\boldsymbol{SA_{Br}}$.

Stage 4.b Marginal visual fit $\boldsymbol{SA_{Br}}$

Figure 7.8 shows the fitted statistical models to the univariate marginal distributions for visual comparison. The GMM is not included since the marginal distributions will not be the same as the bivariate GMM.

By applying the bivariate statistical modelling methodology from Chapter 6 to $\boldsymbol{SA_{Br}}$ is has been possible to establish the bivariate Students' t-distribution (as the bivariate Cauchy distribution) as the best fit. The methodology can now be applied to $\boldsymbol{SA_{Jhb}}$.

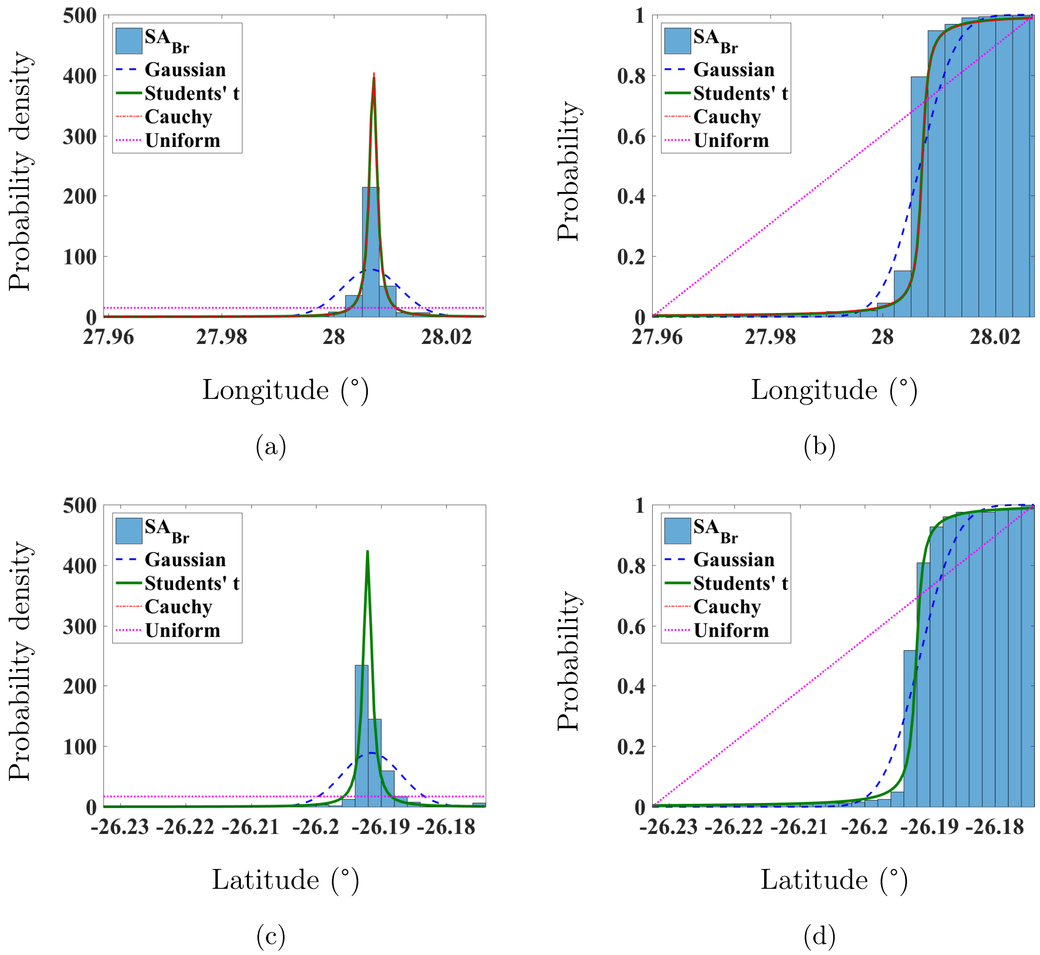


Figure 7.8: The **four** statistical models fitted to the marginal distributions of $\boldsymbol{SA_{Br}}$.

7.2.4 Bivariate Statistical Modelling: $\boldsymbol{SA_{Jhb}}$

$\boldsymbol{SA_{Jhb}}$ is the set of SALDN strokes that were detected and reported between 2009-2013 and 2015-2016 for a 1600 km^2 area around the Johannesburg region. $\boldsymbol{SA_{Jhb}}$ is a description of how SALDN strokes are distributed within the Johannesburg region - the statistical model that best describes this data set will be the best choice of statistical model for the prior probability function described in Chapter 5.

Stage 1. Data set $\boldsymbol{X} = \boldsymbol{SA_{Jhb}}$

We let the data set $\boldsymbol{X}$ be the reported locations of all 364705 strokes reported in the Johannesburg region, $\boldsymbol{SA_{Jhb}}$.

$$\boldsymbol{X_j} = \begin{pmatrix} \varphi_j \\ \lambda_j \end{pmatrix} = \boldsymbol{SA_{jJhb}} \qquad j = 1, ..., 364705. \tag{7.2}$$

Figure 7.9 shows an aerial view of latitude against longitude with the reported stroke locations indicated.

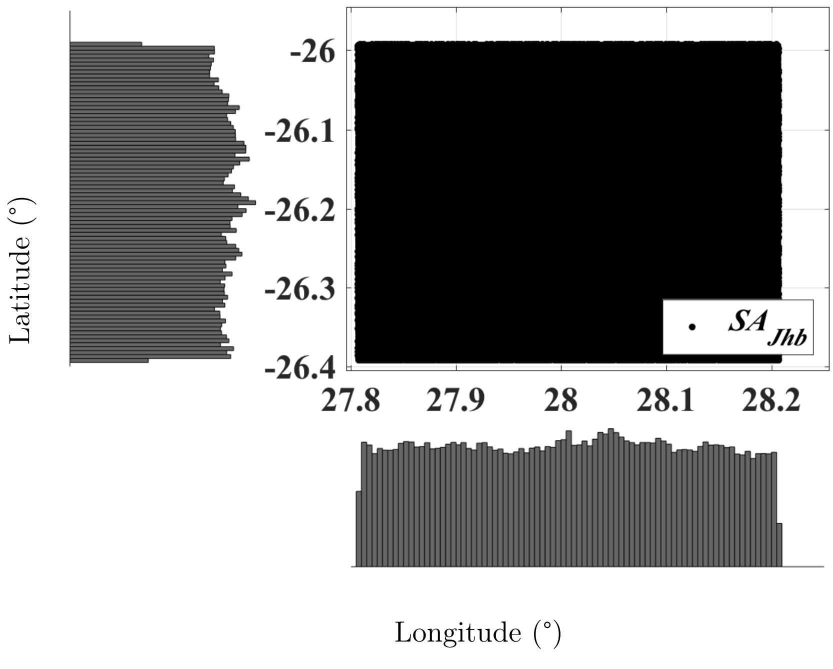


Figure 7.9: Aerial view of latitude vs. longitude for SALDN strokes in the Johannesburg region, $\boldsymbol{SA_{Jhb}}$. The marginal histograms in the latitudinal and longitudinal directions are also shown.

The figure only indicates that there is a large amount of lightning activity in the Johannesburg area, and that every location appears to have been struck at some point.

The marginal histograms show slight variation but it appears to be almost uniformly distributed.

Stage 2. Bivariate visualization $\boldsymbol{SA_{Jhb}}$

As described in Chapter 6, a proper bivariate visualization is needed. Figures 7.10a-b, 7.10c-d and 7.10e-f show a bivariate histogram count, probability density estimate and empirical cumulative density plot with bin size $0.0081° \times 0.0081°$ for the data set $\boldsymbol{SA_{Jhb}}$ respectively.

We now see a more accurate reflection of the distribution of SALDN strokes in the area - there are two locations where there are many more stroke reports than elsewhere. These coordinates correspond to the location of the Brixton tower and the Hillbrow tower, and it is clear that both towers are struck much more frequently than any other location in the area.

Stage 3. Maximum Likelihood Estimation $\boldsymbol{SA_{Jhb}}$

Table 7.4 shows the parameters for the **five** bivariate statistical models fitted to $\boldsymbol{SA_{Jhb}}$ using MLE.

Table 7.6: Bivariate MLE fitted parameters for SALDN data in Johannesburg region $\boldsymbol{SA_{Jhb}}$

Parameter	Gaussian	Students' t	Cauchy	Uniform	GMM
μ_{Jhb_φ}	29.0063°	29.0063°	29.0082°	28.0068°	$\varphi_{1-364705}$
μ_{Jhb_λ}	−26.1934°	−26.1933°	−26.1926°	−26.1925°	$\lambda_{1-364705}$
σ_{Jhb_φ}	0.1144	0.1140	0.0930	0.1154	0.0051
σ_{Jhb_λ}	0.1133	0.1129	0.0916	0.1154	0.0045
$\rho_{Jhb_{\lambda\varphi}}$	0.0100	0.0103	0.0248	-	0.5395
ν_{Jhb}	-	134.3271	1	-	-
loglik.	549980	548430	379060	-	-

Firstly, we notice that the bivariate Students' t-distribution model has the same fitted parameters as the bivariate Gaussian model and a ν_{Jhb} is greater than 100. This indicates that the best fitting bivariate Students' t-model is actual the bivariate Gaussian model, which is a bivariate Gaussian model with a wide variance, as close to the uniform model fitted parameters as possible. Once again, the bivariate GMM is included in the table, noting that the co-variance parameters are taken from the Gaussian fit to $\boldsymbol{SA_{Br}}$. Figure 7.11 shows a plot of the bivariate probability density function of the GMM with the parameters indicated in table 7.6. The high density locations of Brixton tower and Hillbrow tower are clear in the density plot.

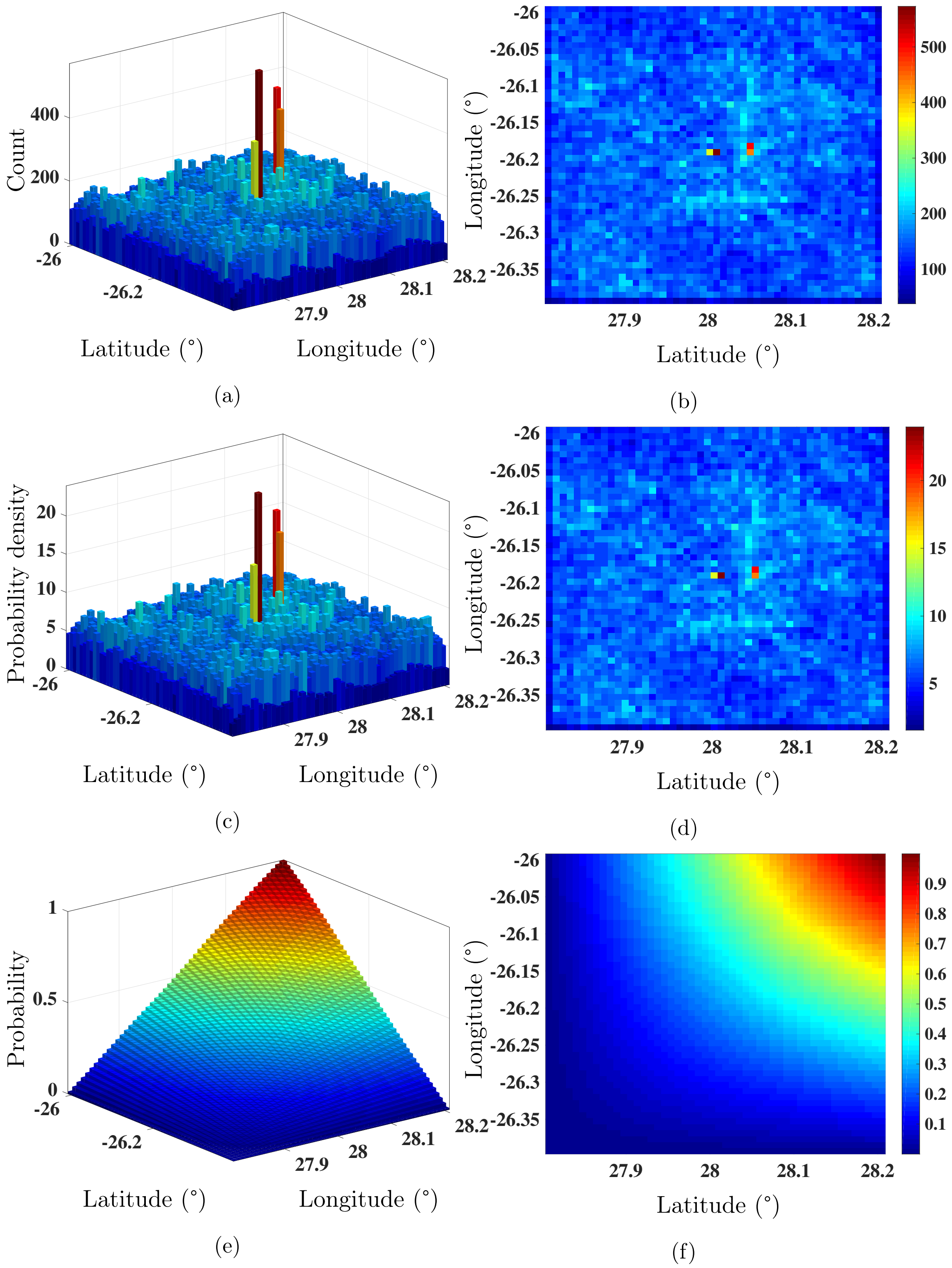
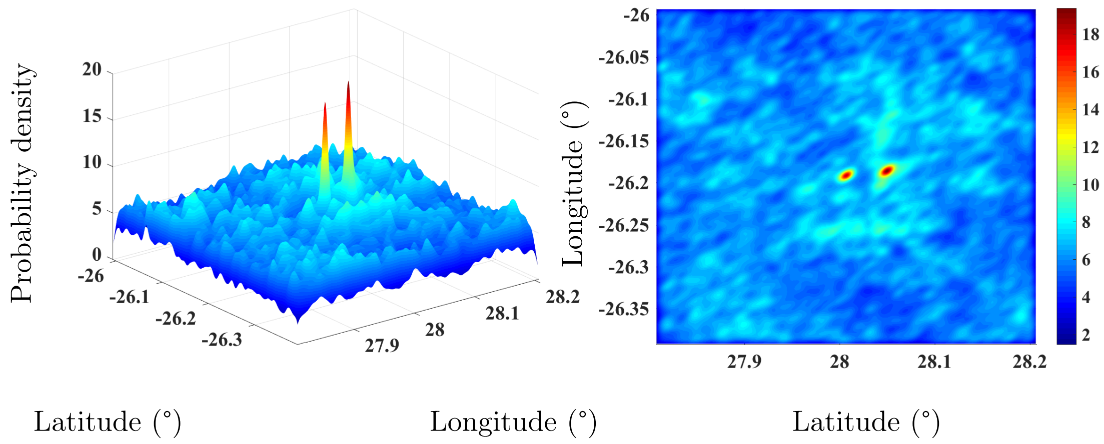


Figure 7.10: Bivariate visualization of data set $\boldsymbol{SA_{Jhb}}$: (a-b) Histogram count. (c-d) Probability density estimate. (e-f) Empirical cumulative density. All plots with bin size $0.0081^{\circ} \times 0.0081^{\circ}$.



(a) (b)

Figure 7.11: Probability density function of the Gaussian Mixture Model (GMM) fitted to $\boldsymbol{SA_{Jhb}}$. The high density locations of Brixton tower and Hillbrow tower are clear in the plot.

Stage 4. Goodness-of-fit $\boldsymbol{SA_{Jhb}}$

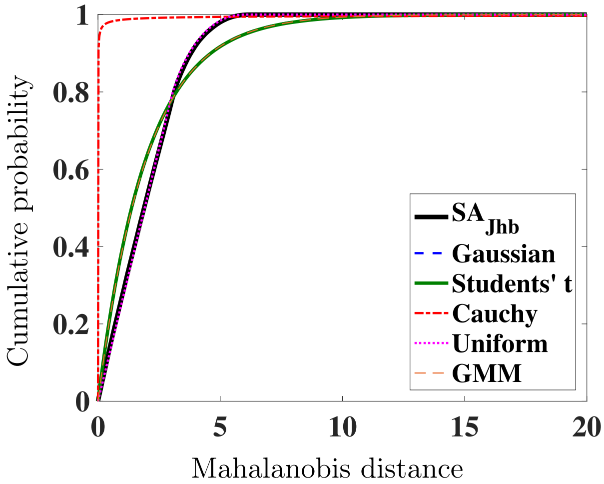
Figure 7.12a shows the cumulative probability vs. the Mahalanobis distance for each of the bivariate statistical models as well the empirical cumulative distribution of $\boldsymbol{SA_{Jhb}}$. Already we can see that the bivariate uniform model and the bivariate GMM match the empirical data the best, following the shape of the curve exactly. Figure 7.12b shows an enhanced section of the figure, allowing us to see that the uniform model fit deviates slightly from the empirical curve whereas the GMM fit continues to follow it precisely.

Table 7.7 shows the Mahalanobis test statistics (along with negative log-likelihood and the other criterion) for the statistical model fits to $\boldsymbol{SA_{Jhb}}$. The GMM has the lowest test statistic by far, followed by the uniform model fit. As expected the bivariate Cauchy model fit is the worst and the bivariate Gaussian and Students' t- are similar.

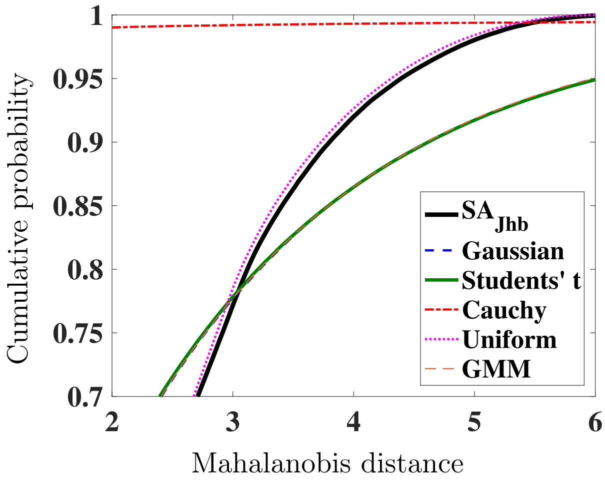
Table 7.7: Mahalanobis test results for $\boldsymbol{SA_{Jhb}}$. Lowest values indicate best quality of fit.

Bivariate Distribution	Mhl_T	Neg. Loglik.	AIC	BIC
Gaussian	4.5103	-549980	-1100000	-1099900
Students' t	4.6482	-548430	-1096900	-1096800
Cauchy	19.0805	-379060	-758110	-758050
Uniform	0.6803	-	-	-
GMM	0.0363	-	-	-

It appears that a GMM, based on the locations of strokes previously reported in an area is the best description for the data, although a uniform distribution is a good fit too.



(a)



(b)

Figure 7.12: Cumulative Mahalanobis distances for the different statistical models compared with the empirical Mahalanobis distance of $\boldsymbol{SA_{Jhb}}$ a) full curve b) enhanced section.

Stage 4.b Marginal visual fit $\boldsymbol{SA_{Jhb}}$

Figure 7.13 shows the statistical models fitted to the univariate marginal distributions for visual comparison. The GMM is not included as discussed previously. Here, the uniform model has the best visual fit. As noted earlier however, the high densities found at the location of the Brixton tower and the Hillbrow tower do not appear in the marginal distributions.

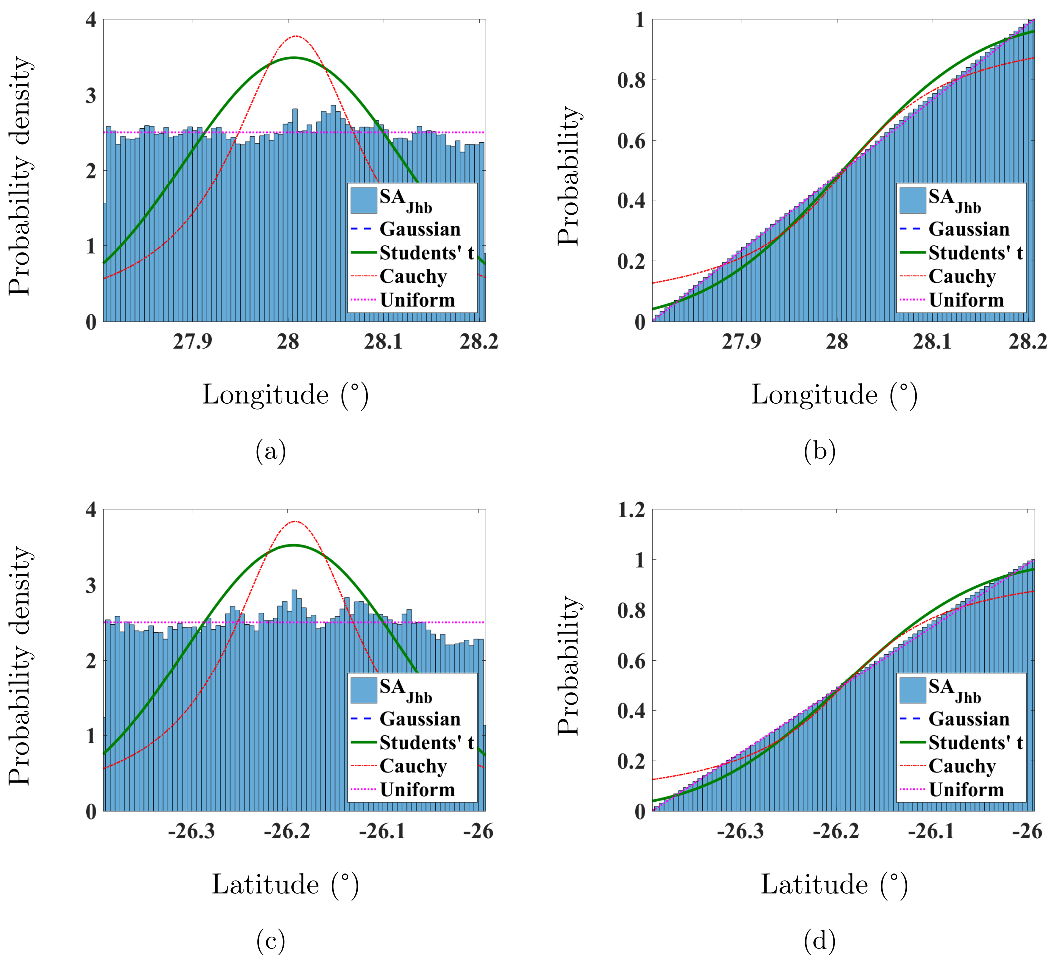


Figure 7.13: The **four** statistical models fitted to the marginal distributions of $\boldsymbol{SA_{Jhb}}$.

By applying the bivariate statistical modelling techniques described in Chapter 6, it appears that a bivariate GMM is the best statistical model for the prior probability function. Next, the methodology is applied to a LLS data set from Austria.

7.3 Austria

This section discusses lightning observations made in Austria and how they are related to the Austrian Lightning Detection and Information System (ALDIS)/European Co-operation for Lightning Detection (EUCLID). The Gaisberg tower is a 100 m tall radio tower constructed on top of Gaisberg mountain (1287 m above sea level) nearby the town of Salzburg, in Austria. Figure 7.14a shows a photograph of the Gaisberg radio tower overlooking the town of Salzburg. The tower is located at 47.8052° North and 13.1121° East (shown in figure 7.14b) (Diendorfer et al., 2011a).

(a)

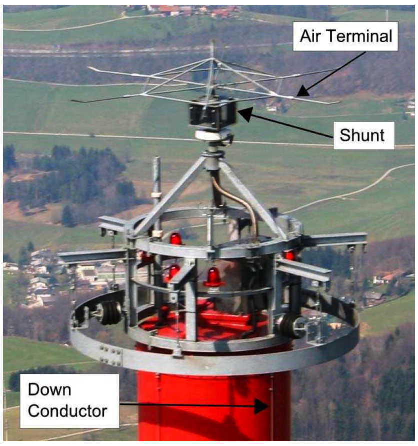
(b)

Figure 7.14: The Gaisberg radio tower, (a) on top of Gaisberg mountain, overlooking the town of Salzburg in the North of Austria and (b) its location, 47.8053° North and 13.1121° East. Images used with permission, courtesy Dr. Wolfgang Schulz

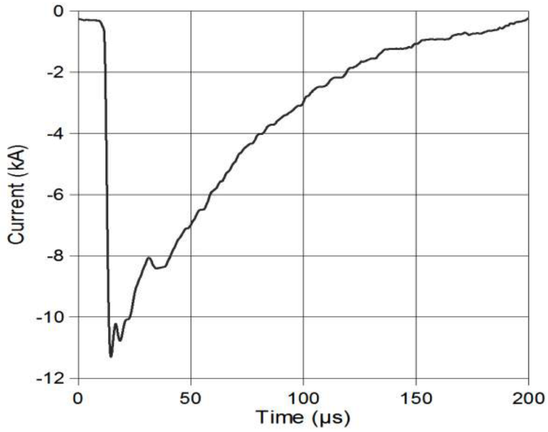
The flash density in the region is $2 - 3\ flashes \times km^{-2} \times year^{-1}$ but with a marked increase over Gaisberg mountain due to Gaisberg tower. Due to the large number of events, the Gaisberg tower presented itself as an ideal location to make lightning observations and measurements for the purposes of evaluating the Austrian Lightning Detection and Information System (ALDIS). Since 1998, a number of different measurements of ground-truth lightning events have been made, including: current measurements using a shunt installed on the tower, slow and fast electric field measurements and high-speed video recordings (Diendorfer, 2016, Diendorfer et al., 2009, 2011a).

7.3.1 Gaisberg Current Measurement

A shunt resistor is installed at the top of the Gaisberg tower, allowing for measurement of current wave-forms when the tower is struck by lightning or initiates upward flashes. A fibre optic connection carries the signal to the base of the tower where the data recording equipment is housed. All recordings are GPS time-stamped allowing for comparison with other measurements (Diendorfer et al., 2011b). Figure 7.15a indicates the shunt installed on the top of the Gaisberg tower and an example of the type of current wave-forms that are measured, figure 7.15b.



(a)



(b)

Figure 7.15: The current measurement equipment at Gaisberg tower: (a) the resistive shunt installation at the top of Gaisberg tower and (b) example of current wave-shape that is measured during lightning attachment. Images used with permission, courtesy Dr. Wolfgang Schulz

From 2000-2015, a total of 829 lightning events were recorded at the tower (Diendorfer, 2016). Table 7.8 shows a breakdown of the number of recorded events per year as published by Diendorfer et al. (2011a,b).

Table 7.8: Number of lightning current measurements at the Gaisberg tower, Austria per year (Diendorfer et al., 2011a,b).

Year	Total
2000	95
2001	61
2002	65
2003	32
2004	22
2005	59
2006	56
2007	99
2008	98
2009	65
2010	12
2011	28
2012	67
2013	10
2014	28
2015	32
Total	829

7.3.2 ALDIS/EUCLID Data

The Austrian Lightning Detection and Information System (ALDIS) consists of 8 Vaisala LS7002 sensors and has coverage over all of Austria. The network forms part of the European Cooperation for Lightning Detection (EUCLID) which has coverage over all of Europe. Table 7.9 shows the ALDIS stroke reports correlated with flashes to the Gaisberg tower from 2000 to 2015 (Diendorfer, 2010, Diendorfer et al., 2014, Schulz, 2015). This comes to a total of 1649 detected strokes and we let this be the data set $\boldsymbol{Aus_{Gais}}$.

Table 7.9: ALDIS data time-correlated with lightning attachments to the Gaisberg tower, $\boldsymbol{Aus_{GBT}}$

Year	No. Flashes	No. Detected by ALDIS	Total No. LLS Strokes
2000	95	33	161
2001	61	27	125
2002	65	33	118
2003	32	17	62
2004	22	8	11
2005	59	33	173
2006	56	12	50
2007	99	39	175
2008	98	60	373
2009	65	22	84
2010	12	2	4
2011	28	16	69
2012	67	15	42
2013	10	3	13
2014	28	15	57
2015	32	20	132
Total	829	355	1649

We can now apply the bivariate statistical model methodology described in Chapter 6 to $\boldsymbol{Aus_{GBT}}$.

7.3.3 Bivariate Statistical Modelling: $\boldsymbol{Aus_{GBT}}$

$\boldsymbol{Aus_{GBT}}$ is the set of strokes from the ALDIS network that were detected and reported when the Gaisberg tower was struck by lightning (ground-truth current measurements). The distribution of the reported locations of $\boldsymbol{Aus_{GBT}}$ are a description of how the location errors of the ALDIS network are distributed within the region of the Gaisberg tower.

Stage 1. Data set $\boldsymbol{X} = \boldsymbol{Aus_{GBT}}$

Firstly, we let the data set $\boldsymbol{X}$ be the reported locations of all the strokes attaching to the Gaisberg tower detected by the ALDIS network. This is a total of 1649 strokes.

$$\boldsymbol{X_j} = \begin{pmatrix} \varphi_j \\ \lambda_j \end{pmatrix} = \boldsymbol{Aus_{jGBT}} \qquad j = 1, ..., 1649. \tag{7.3}$$

Figure 7.16 shows an aerial view of latitude against longitude with the reported stroke locations indicated. The location of the Gaisberg tower can be found in the center at $(13.1121°, 47.8052°)$. Similarly to the Brixton tower in South Africa, the data is clustered around the location of the Gaisberg tower with a few outliers to the northeast of the tower. There seems to greater variance in the longitudinal directions.

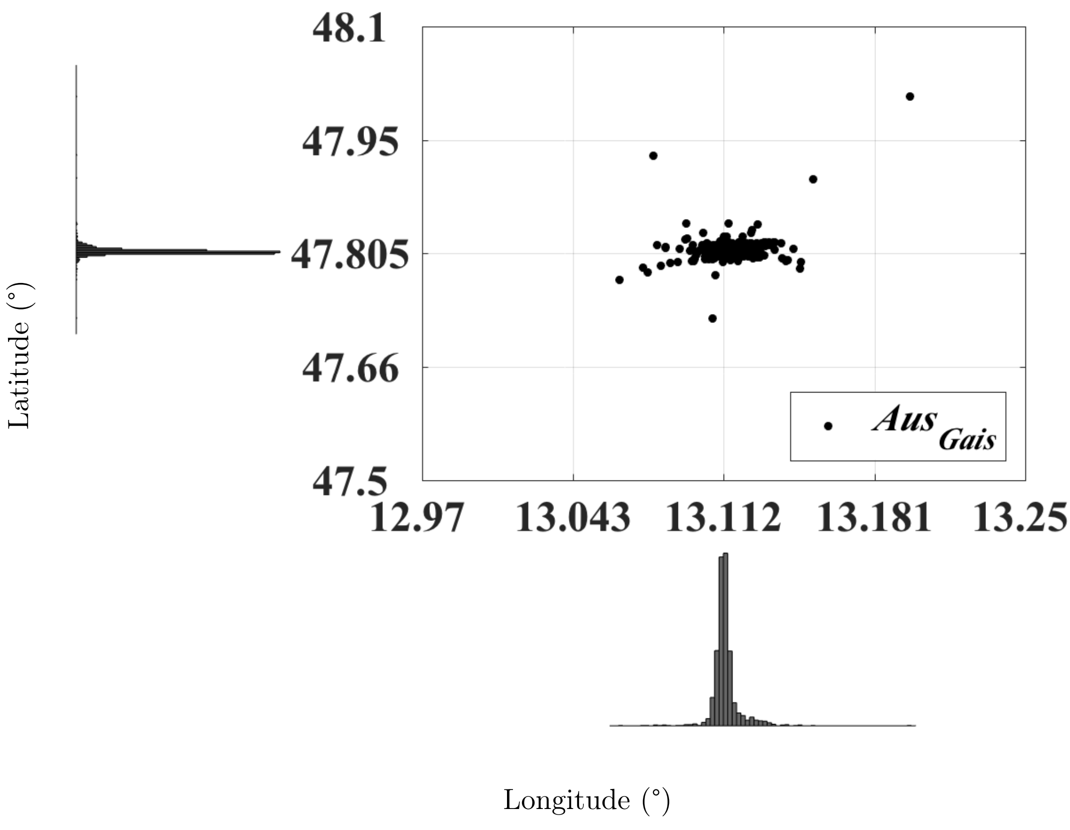


Figure 7.16: Aerial view of latitude vs. longitude for ALDIS strokes time-correlated with events to the Gaisberg tower, $\boldsymbol{Aus_{GBT}}$. The Gaisberg tower is located at $(13.1121°, 47.8052°)$. The marginal histograms in the latitudinal and longitudinal directions are also shown.

The marginal histograms are once again shown in the figure and there is correlation between the longitudinal and latitudinal marginal distributions as the histograms are not symmetrical.

Stage 2. Bivariate visualization $\boldsymbol{Aus_{GBT}}$

Figures 7.17a-b, 7.17c-d and 7.17e-f show the bivariate histogram count, probability density estimate and empirical cumulative density plot with bin size 0.0068° × 0.0068° for the data set $\boldsymbol{Aus_{GBT}}$ respectively. As with the Brixton tower, the data is heavily clustered around the mode with some outliers. The longitudinal variance is also noticeable here in the bivariate visualization.

Stage 3. Maximum Likelihood Estimation $\boldsymbol{Aus_{GBT}}$

Table 7.10 shows the parameters for the **four** bivariate statistical models fitted to $\boldsymbol{Aus_{GBT}}$.

Table 7.10: Bivariate MLE fitted parameters fitted to $\boldsymbol{Aus_{GBT}}$

Parameter	Gaussian	Students' t	Cauchy	Uniform	GMM
μ_{GBT_φ}	13.1128°	13.1120°	13.1120°	13.1311°	φ_{1-1649}
μ_{GBT_λ}	47.8075°	47.8066°	47.8065°	47.8649°	λ_{1-1649}
σ_{GBT_φ}	0.0062	0.0022	0.0018	0.0383	0.0062
σ_{GBT_λ}	0.0079	0.0018	0.0016	0.0831	0.0079
$\rho_{GBT_{\lambda\phi}}$	0.3652	0.2405	0.2229	-	0.3652
ν_{GBT}	-	1.8012	1	-	-
loglik.	11801	14090	14004	-	-

Similarly to the Brixton tower, the parameters μ_{GBT_φ}, μ_{GBT_λ} are all very close to the location of the Gaisberg tower. As expected for the bivariate visualization of the data, there is some correlation between the marginal distributions. For this data set, the scaling between the bivariate Gaussian parameters and the bivariate Students' t model parameters is not as pronounced as in the case of the Brixton tower, and the degrees of freedom of the (ν_{GBT}) is almost 2, rather than being close to the Cauchy model. Once again, the GMM model reduces to the fitted Gaussian model.

Figure 7.18 shows the negative log-likelihood against the degrees of freedom parameter (ν_{GBT}) of the bivariate Students' t-distribution. This demonstrates the results of the modified EM algorithm and shows how the bivariate Students' t model parameters are obtained for $\boldsymbol{Aus_{GBT}}$. The negative log-likelihood is lowest when $\nu_{GBT} \approx 2$

Stage 4. Goodness-of-fit $\boldsymbol{Aus_{GBT}}$

Figure 7.19 shows the cumulative probability of the Mahalanobis distances for each of the fitted bivariate statistical models as well as for the empirical cumulative distribution of $\boldsymbol{Aus_{GBT}}$. Here we see that the fitted bivariate Students' t-model almost matches the empirical data exactly. The empirical data exhibits the characteristic of the bivariate Students' t-distribution by rising quickly but then slowing and only reaching 1 much later.

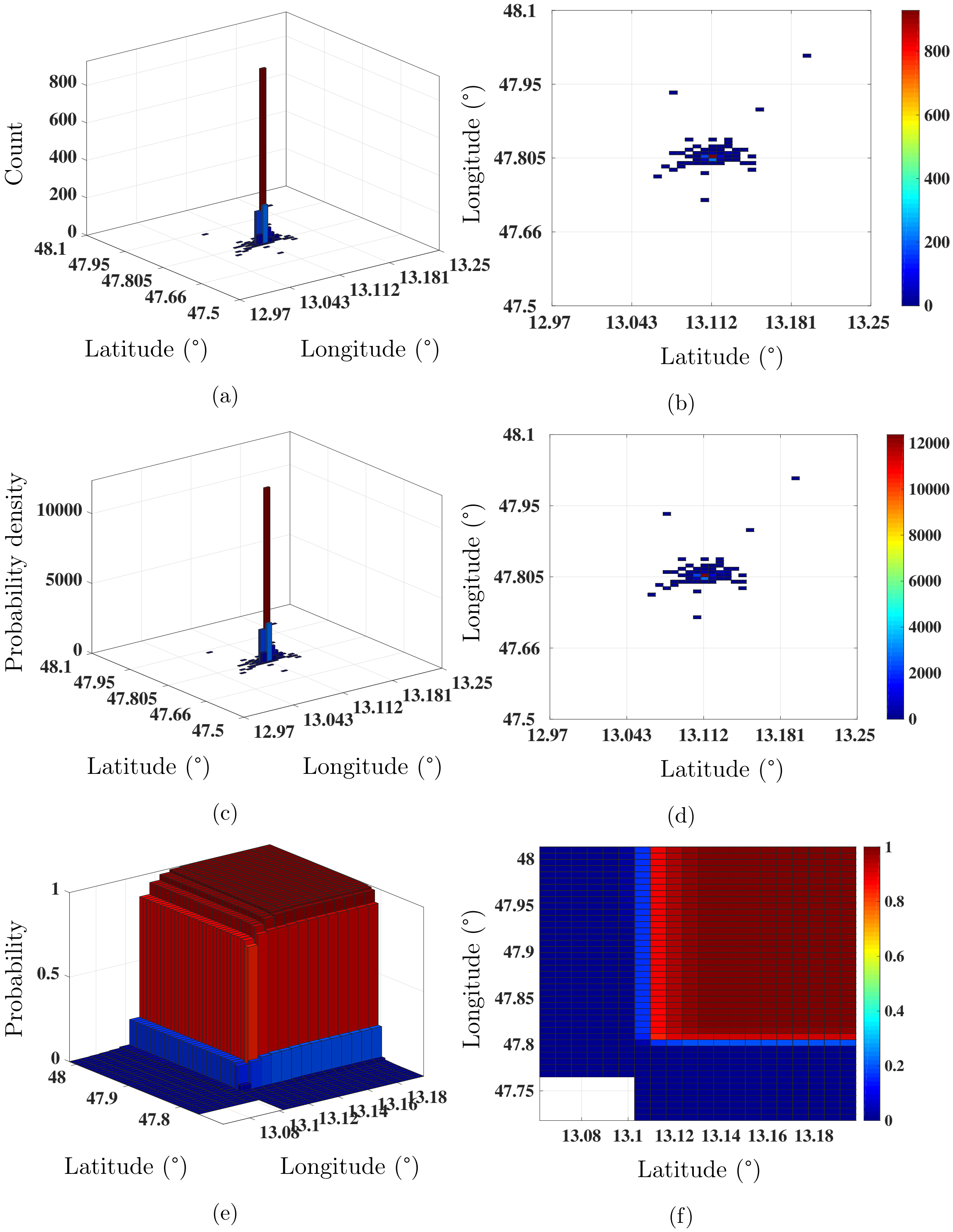


Figure 7.17: Bivariate visualization of data set $\boldsymbol{Aus_{GBT}}$: (a-b) Histogram count. (c-d) Probability density estimate. (e-f) Empirical cumulative density. All plots with bin size $0.0068° \times 0.0068°$.

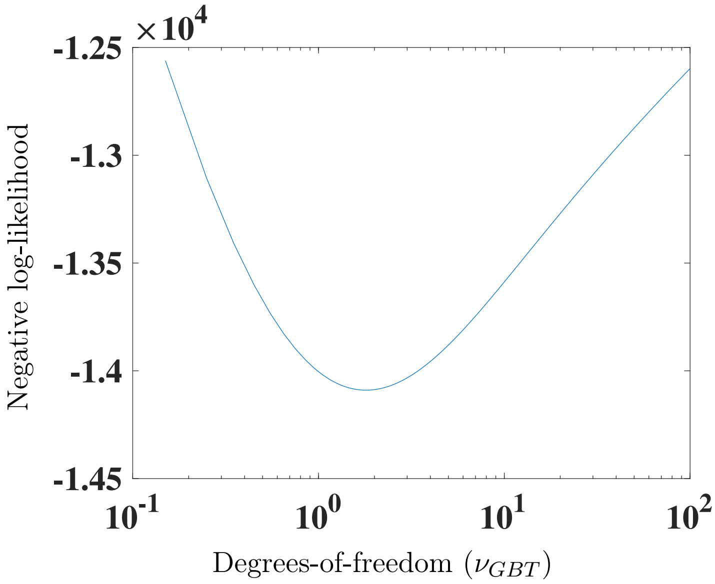


Figure 7.18: Negative log-likelihood vs. degrees of freedom parameter for the bivariate Students' t-distribution, ν_{GBT}. When $\nu_{GBT} = 1$, the fitted parameters are equivalent to a Cauchy distribution and as $\nu_{GBT} \to \infty$, the fitted parameters are equivalent to those of the Gaussian distribution.

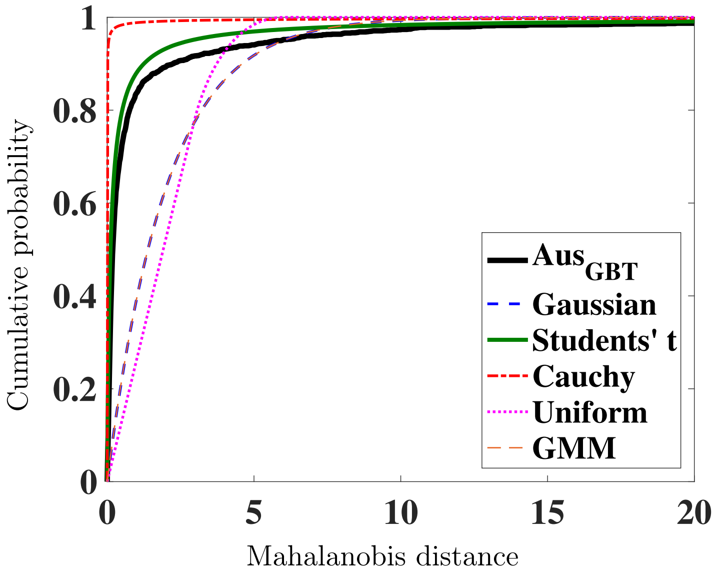


Figure 7.19: Cumulative probability vs. Mahalanobis distance for the different statistical models compared with the empirical Mahalanobis distance of $\boldsymbol{Aus_{GBT}}$.

Table 7.11 shows the results of Mahalanobis test for the **five** fitted statistical models along with the negative log-likelihood and the AIC and BIC. Here, it is clear that the bivariate Students' t-distribution is the best fit for the data, having a test statistic four times smaller than the bivariate Gaussian fit.

Table 7.11: Mahalanobis test results for $\boldsymbol{Aus_{GBT}}$. Lowest values indicate best fit.

Bivariate Distribution	Mhl_T	Neg. Loglik.	AIC	BIC
Gaussian	22.5194	-11801	-23592	-23565
Students' t	5.3849	-14090	-28168	-28135
Cauchy	13.1913	-14004	-27998	-27971
Uniform	37.4052	-	-	-
GMM	22.5425	-	-	-

As with the Brixton tower, the bivariate Students' t-distribution is the best fitting model for location errors in the ALDIS LLS.

Stage 4.b Marginal visual fit $\boldsymbol{Aus_{GBT}}$

Figure 7.20 shows the **four** statistical models fitted to the univariate marginal distributions for visual comparison.

As previously discussed, the marginals of the GMM do not follow the same distribution and can therefore not be shown on the figures.

7.4 Conclusions

Table 7.12 shows the Mahalanobis test statistic for each of the statistical model fits for all data sets.

Table 7.12: Mahalanobis test results for $\boldsymbol{SA_{Br}}$, $\boldsymbol{SA_{Jhb}}$ and $\boldsymbol{Aus_{GBT}}$. Lowest values indicate best fit.

	$\boldsymbol{SA_{Br}}$	$\boldsymbol{SA_{Jhb}}$	$\boldsymbol{Aus_{GBT}}$
Gaussian	23.0719	4.5103	22.5194
Students' t	17.8075	4.6482	5.3849
Cauchy	17.1153	19.0805	13.1913
Uniform	38.4571	0.6803	37.4052
GMM	23.0119	0.0363	22.5425

Firstly, considering $\boldsymbol{SA_{Br}}$ and $\boldsymbol{Aus_{GBT}}$, these data sets of LLS strokes allow us to see the distribution of LLS stroke reports when a known location is struck by lightning repeatedly - allowing us to see what statistical model is best for the likelihood function. For $\boldsymbol{SA_{Br}}$, we see that the bivariate Cauchy model is the best fit although the fitted Students' t-distribution has a degree of freedom parameter $\nu_{Br} \approx 1$ making it almost

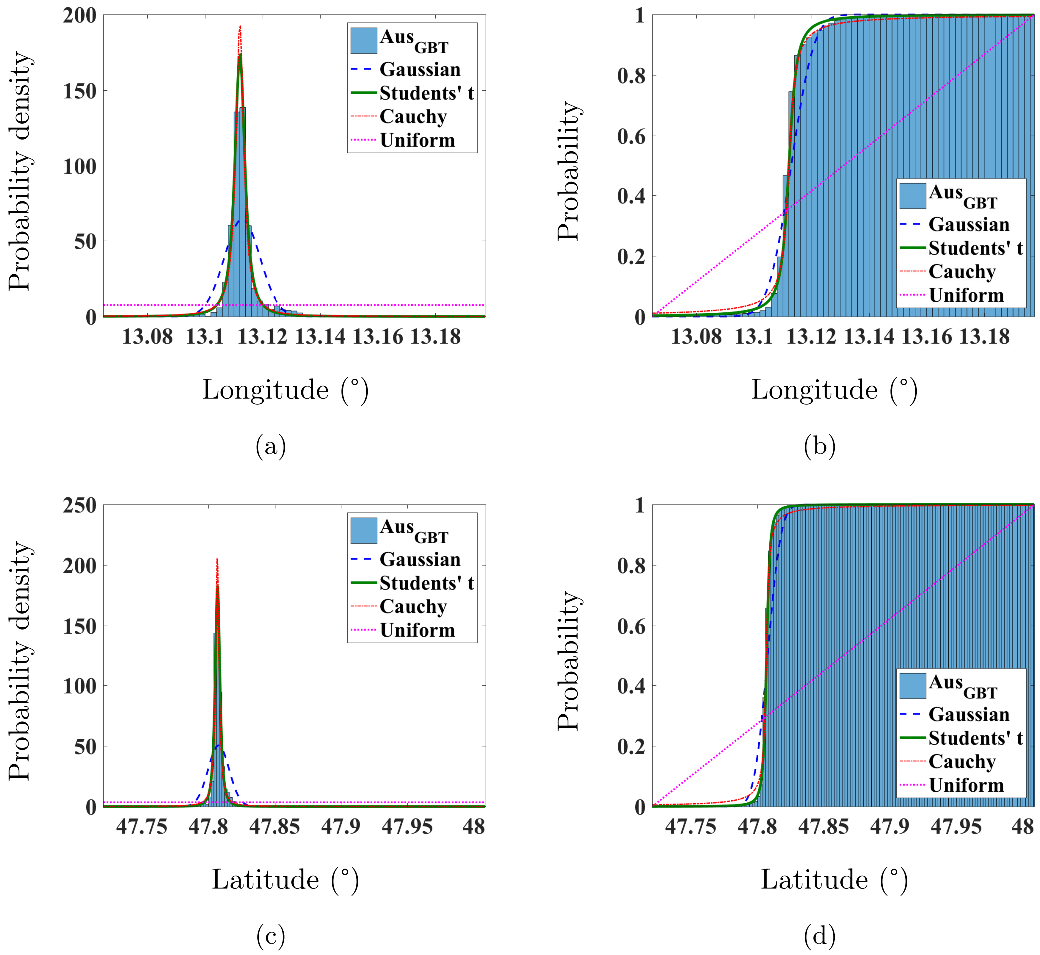


Figure 7.20: The **four** statistical models fitted to the marginal distributions of $\boldsymbol{Aus_{GBT}}$.

equivalent to the bivariate Cauchy model. For $\boldsymbol{Aus_{GBT}}$, the bivariate Students' t-model is clearly the best fit. The much greater number of strokes in $\boldsymbol{Aus_{GBT}}$ than $\boldsymbol{SA_{Br}}$ probably makes for the distinction between the bivariate Cauchy and Students' t-model. Overall, we see that a bivariate Students' t-distribution appears to be a better description for LLS errors than the bivariate Gaussian distribution, likely due to the presence of outliers in the data sets.

Considering $\boldsymbol{SA_{Jhb}}$, this allows us to see the distribution of strokes over a range of locations within an area of investigation and, while a bivariate uniform distribution is a better fit than a bivariate Gaussian or Students' t-distribution, the GMM gives the best fit overall, allowing the different frequency of events at different locations to be described.

7.5 Chapter Summary

This chapter has described the verification of the statistical models proposed for the likelihood function and the prior probability function of the Bayesian framework using ground-truth lightning event data time-correlated with Lightning Location System (LLS) stroke reports.

The city of Johannesburg, South Africa has been described with specific reference to the two tall communication towers - the Brixton tower and the Hillbrow tower. A study to photograph lightning events to the Brixton tower running from 2009-2013 and again in 2015-2016 has been described, with 121 flashes to the tower having been recorded. Data from the South African Lightning Detection Network (SALDN) time-correlated with these events lead to a data set, $\boldsymbol{SA_{Br}}$ of 328 stroke reports - the distribution of the location errors allowing for verification of the proposed likelihood function. The number of SALDN stroke reports within a 1600 km^2 around Johannesburg for the 2009-2013 and 2015-2016 period were also provided, making a data set of 364705 stroke reports, $\boldsymbol{SA_{Jhb}}$ - the distribution of these strokes allowing for the verification of the proposed prior probability function.

The Gaisberg tower - a tall communications tower - near Salzburg in Austria was described, along with the instrumentation installed at the tower to make ground-truth current measurements of lightning events from the tower. A total of 829 have been recorded between 2000 - 2015. The events were time-correlated with LLS data from the Austrian Lightning Detection and Information System (ALDIS) which resulted in a data set of 1649 stroke reports, $\boldsymbol{Aus_{GBT}}$ - the distribution of the location errors also allowing for verification of the proposed likelihood function.

The application of the bivariate statistical modelling methodology described in Chapter 6 to these LLS data sets - $\boldsymbol{SA_{Br}}$, $\boldsymbol{SA_{Jhb}}$ and $\boldsymbol{Aus_{GBT}}$ - has been described and it was found that the bivariate Students' t-distribution is the statistical model with the best fit for $\boldsymbol{SA_{Br}}$ and $\boldsymbol{Aus_{GBT}}$, indicating the distribution to be the best choice for the likelihood function, and the bivariate Gaussian Mixture Model (GMM) to be the best fit for $\boldsymbol{SA_{Jhb}}$ (with the bivariate uniform distribution being next best) indicating it to be the best choice for the prior probability function.

The next chapter demonstrates the use of the Proposed Bayesian Framework by applying it to a case study based around the Brixton and Hillbrow tower, South Africa.

Chapter 8

Application of Bayesian Framework

In this chapter, the proposed Bayesian framework described in Chapter 5 is applied to a case study using the statistical models found in Chapter 7 as the likelihood and prior probability functions. The case study is framed as a forensic investigation around the Brixton and Hillbrow towers in Johannesburg, South Africa, with the investigation aiming to determine whether the Brixton tower was struck between October 2015 - March 2016. Utilizing the ground-truth data sets from Chapter 7, it is possible to evaluate the Bayesian framework and the different statistical models.

8.1 Overview

To demonstrate the application of the Bayesian framework proposed in Chapter 7, the location of Johannesburg, South Africa and the two tall towers in the region (Brixton and Hillbrow tower), will again be examined. Here however, the situation will instead be framed as a forensic investigation. The following sections make up the chapter:

8.2 Forensic investigation - Johannesburg, South Africa.

8.3 Ground-truth comparison.

The forensic investigation is framed to answer the question as to whether the Brixton tower and/or the Hillbrow was struck by lightning during the 2015-2016 thunderstorm season. The requirements and definition of terms and variables described in Chapter 5 are performed for the application of the Bayesian framework to this investigation. Next, given the ground-truth data sets presented in Chapter 7, it is possible to compare the results of applying the Bayesian framework with ground-truth events to Brixton tower, allowing the framework (and the different statistical models for the likelihood and prior probability functions) to be evaluated.

8.2 Forensic Investigation - Johannesburg, South Africa

The forensic investigation is as follows - there are two locations of interest at which an event may have occurred, the Brixton tower $\boldsymbol{L_{Br}}$ and the Hillbrow tower $\boldsymbol{L_{Hlbrw}}$. Using LLS data, the investigation aims to answer the following types of questions:

- *"Was the Brixton tower struck by lightning between 2015-10-01 and 2016-03-01?"*
- *"When was the Brixton tower struck by lightning between 2015-10-01 and 2016-03-01 and when wasn't it?"*
- *"Is it more likely that the Brixton or the Hillbrow tower was struck by lightning at certain times between 2015-10-01 and 2016-03-01?"*

As discussed in Chapter 5, the first step of the Bayesian framework is to define the geographic area of investigation. This area needs to be such that both $\boldsymbol{L_{Br}}$ and $\boldsymbol{L_{Hlbrw}}$ are located within the area and is sufficiently large enough that reported LLS strokes with large location errors will still be found within the area. Therefore, an area of approximately 1600 km^2 around the location of Brixton tower is chosen, which includes the location of Hillbrow tower ($\approx$ 4 km east of Brixton tower.). Table 8.1 shows the relevant parameters of the defined geographic area:

Table 8.1: Parameters of defined geographic area for the Brixton and Hillbrow tower investigation.

	Longitude φ_i	Latitude λ_i
$\boldsymbol{L_{Br}}$	28.0068°	−26.1925°
$\boldsymbol{L_{Hlbrw}}$	28.0494°	−26.1870°
	φ_{lim}	λ_{lim}
Max.	28.2067°	−25.9926°
Min.	27.8069°	−26.3924°
	$\Delta\varphi$	$\Delta\lambda$
Δ	0.0014538°	0.0014538°

Also included in the table is the chosen grid size $\Delta\varphi \times \Delta\lambda$ (this is sufficiently small, as discussed in Appendix A. Therefore, the total number of locations $\boldsymbol{L_i}$ (Chapter 5, equation 5.3) within the defined geographic area of investigation is $i = 1, ..., N$ where $N = 75625$ (from equation 5.14).

The next step of the framework is to define a time-frame of investigation. The time-frame chosen is that described in Chapter 7, the lightning season in the Johannesburg region from 2015 to 2016 where the Brixton tower was monitored by a motion-triggered camera (October 2015 to March 2016). The time-frame parameters are shown in table 8.2.

Now that the time-frame has been defined, the LLS with coverage in the region - the South African Lightning Detection Network (SALDN) - can be queried for all strokes

Table 8.2: Parameters of defined time-frame for the Brixton and Hillbrow tower investigation

	Time (Year-Month-Day HH:MM:SS.SS UTC)
t_{str}	2015-10-01 00:00:00.00
t_{fin}	2016-03-01 00:00:00.00

detected within the time-frame and reported within the geographic area of investigation, $\boldsymbol{X_{lls_j}}$. For the parameters in table 8.1 and 8.2, a total of $M = 66780$ reported SALDN strokes are returned (equation 5.4).

Based on the ground-truth photographs and analysis conducted in Chapter 7, $\boldsymbol{X_{lls_j}}$ can be split into two separate data sets:

- $\boldsymbol{X_{lls_{jTo}}}$▼: SALDN strokes time-correlated with photographs of lightning attaching to the Brixton tower. A total of 143 strokes within the defined time-frame.
- $\boldsymbol{X_{lls_{jNot}}}$●: SALDN strokes *NOT* time-correlated with photographs of lightning attaching to the Brixton tower. A total of 66637 strokes within the defined time-frame and area.

Since this analysis was not performed for the Hillbrow tower (no motion-triggered camera captured ground-truth lightning events to the Hillbrow tower), it is not possible to distinguish SALDN strokes that may have attached to the Hillbrow tower. The SALDN can also be queried for all strokes reported in the area before the beginning of the time-frame t_{str}, to provide a prior data set $\boldsymbol{X_{llsPr_q}}$ (equation 5.16). This is similar to the data set of SALDN strokes reported in Johannesburg during the thunderstorm seasons where the photographic study was performed (referred to as $\boldsymbol{SA_{Jhb}}$ in Chapter 7). However, here we remove the 2015 - 2016 data as this is the time-frame of investigation. This yields a total of $Q = 297925$ reported SALDN strokes.

8.2.1 Applying the Bayesian Framework

Now that the geographic area of investigation and the time-frame have been defined, and the data sets $\boldsymbol{L}$ and $\boldsymbol{X_{lls}}$ are established, claims about the locations of events can be made. These claims can be stated as:

- $\boldsymbol{L_{Br_S}}$: *"The Brixton tower was struck by lightning."*
- $\overline{\boldsymbol{L_{Br_S}}}$: *"The Brixton tower was NOT struck by lightning."*
- $\boldsymbol{L_{Hlbrw_S}}$: *"The Hillbrow tower was struck by lightning."*

The Bayesian framework can be applied (Chapter 5, equation 5.5 and 5.6) to assess the probability of these claims being true given the evidence of the LLS data $\boldsymbol{X_{lls}}$ and attempt to answer the type of question a forensic investigation may pose, as described at the beginning.

Firstly, the likelihood ratios between the claims are assessed. The likelihood that Brixton tower was struck by lightning versus the likelihood that it was *NOT* struck by lightning is determined by applying equation 8.1 for each detected stroke $\boldsymbol{X_{lls_j}}$:

$$\ln(LR_{jBr-\overline{Br}}) = \ln(\frac{f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{Br_S}})}{f(\boldsymbol{X_{lls_j}}|\overline{\boldsymbol{L_{Br_S}}})}) \tag{8.1}$$

The likelihood that Brixton tower was struck by lightning versus the likelihood that Hillbrow tower was struck by lightning is then:

$$\ln(LR_{jBr-Hlbrw}) = \ln(\frac{f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{Br_S}})}{f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{Hlbrw_S}})}) \tag{8.2}$$

For each detected stroke $\boldsymbol{X_{lls_j}}$, the log-likelihood of whether that stroke supports each of the claims is calculated and, given that each stroke is detected at time t_j, the time of each claim is assessed. Next, the prior probability can be included and, from equation 5.5, the posterior probability of each of the claims being true for each detected stroke $\boldsymbol{X_{lls_j}}$ can be calculated:

$$p(\boldsymbol{L_{Br_S}}|\boldsymbol{X_{lls_j}}) = \frac{p(\boldsymbol{L_{Br_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{Br_S}})}{\sum_{\boldsymbol{L_{Br_S}}} p(\boldsymbol{L_{Br_S}})f(\boldsymbol{X_{lls_j}}|\boldsymbol{L_{Br_S}})} \tag{8.3}$$

This requires statistical models to be chosen for the likelihood functions $\boldsymbol{f(X_{lls_j}|L_{Br_S})}$, $\boldsymbol{f(X_{lls_j}|L_{Hlbrw_S})}$ and the prior probability function $\boldsymbol{p(L_{Br_S})}$.

Likelihood functions $\boldsymbol{f(X_{lls_j}|L_{Br_S})}$ and $\boldsymbol{f(X_{lls_j}|L_{Hlbrw_S})}$:

The likelihood functions chosen for the investigation are based on the results found in Chapter 7 (Section 7.2.3) and the parameters are shown in tables 8.3 and table 8.4:

Table 8.3: Parameters of bivariate Gaussian likelihood function for forensic investigation, Brixton and Hillbrow tower.

$\boldsymbol{L_i}$	μ_{i_ϕ}	μ_{i_λ}	σ_{i_ϕ}	σ_{i_λ}	$\rho_{i_{\lambda\phi}}$
Brixton	28.0066°	−26.1915°	0.0051	0.0045	0.5395
Hillbrow	28.0492°	−26.1858°	0.0051	0.0045	0.5395

Prior Probability function $\boldsymbol{p(L_{Br_S})}$:

The prior probability functions (non-informative and informative) chosen for the investigation are based on the parameters of the investigation (table 8.1) and $\boldsymbol{X_{llsPr_q}}$, the parameters of which are shown in tables 8.5 and 8.6:

Table 8.4: Parameters of bivariate Students' t-likelihood function for forensic investigation, Brixton and Hillbrow tower.

$\boldsymbol{L_i}$	μ_{i_ϕ}	μ_{i_λ}	σ_{i_ϕ}	σ_{i_λ}	$\rho_{i_{\lambda\phi}}$	ν_i
Brixton	28.0069°	−26.1922°	0.00061	0.00057	0.0994	0.8318
Hillbrow	28.0495°	−26.1865°	0.00061	0.00057	0.0994	0.8318

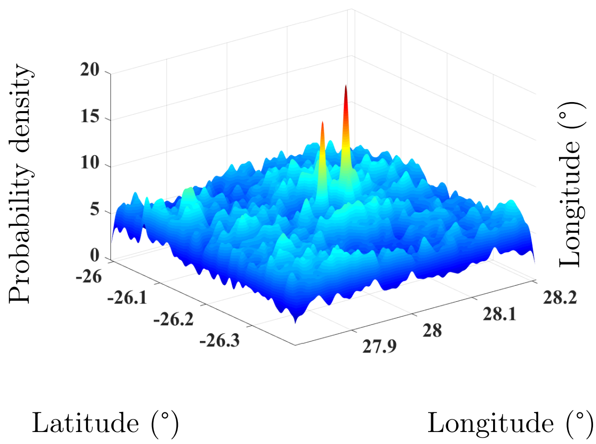
Table 8.5: Parameters of bivariate uniform (non-informative) prior probability function for forensic investigation, Brixton and Hillbrow tower.

φ_{max}	φ_{min}	λ_{max}	λ_{min}
28.2067°	27.8069°	−25.9926°	−26.3924°

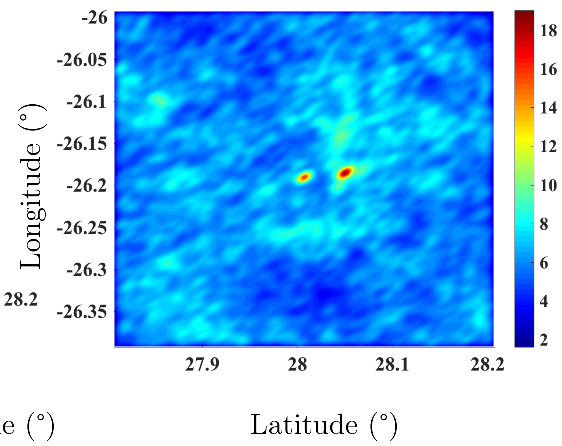
Table 8.6: Parameters of bivariate GMM (informative) prior probability function for forensic investigation, Brixton and Hillbrow tower.

Q	w_{1-Q}	μ_{Jhb_ϕ}	μ_{Jhb_λ}	σ_{Jhb_ϕ}	σ_{Jhb_λ}	$\rho_{Jhb_{\lambda\phi}}$
297925	3.36×10^{-6}	φ_{1-Q}	λ_{1-Q}	0.0051	0.0045	0.5395

Figure 8.1, shows the probability density function of the GMM - similar to that found in Chapter 7, Section 7.2.4 but with the data from the 2015 - 2016 thunderstorm season removed.



(a)



(b)

Figure 8.1: GMM prior probability function density plot based on $\boldsymbol{X_{llsPr_q}}$ SALDN data from 2009-2013 thunderstorm seasons in Johannesburg, South Africa. The two high density points correspond with the locations of the Brixton and Hillbrow towers.

8.3 Ground-truth Comparison

Given the ground-truth photographic study, and therefore the ability to separate $\boldsymbol{X_{lls}}$ into $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$●, it is possible to evaluate the performance of the two different likelihood functions (bivariate Gaussian and Students' t-models).

As discussed in Chapter 4, Section 4.3 and Chapter 4, table 4.2, two critical values V_{crit} can be established for the log-likelihood ratios $\ln(LR_{jBr-\overline{Br}})$ and $\ln(LR_{jBr-Hlbrw})$:

1. $\ln(LR_j) > 0$ (or $LR > 1$): There is weak-to-limited support that $\boldsymbol{L_{Br_S}}$ is true rather than $\overline{\boldsymbol{L_{Br_S}}}$ or $\boldsymbol{L_{Hlbrw_S}}$.
2. $\ln(LR_j) > 5$ (or $LR > 100$): There is moderately strong support that $\boldsymbol{L_{Br_S}}$ is true rather than $\overline{\boldsymbol{L_{Br_S}}}$ or $\boldsymbol{L_{Hlbrw_S}}$.

We can think of $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$● as positive and negative classifications of whether Brixton tower was struck by lightning and construct a confusion matrix as described in Chapter 4, table 4.1. Table 8.7 shows the distinctions between True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN):

Table 8.7: Confusion matrix for Brixton tower.

	$\boldsymbol{X_{lls_{To}}}$▼	$\boldsymbol{X_{lls_{Not}}}$●
$\ln(LR) > V_{crit}$	TP	FP
$\ln(LR) < V_{crit}$	FN	TN

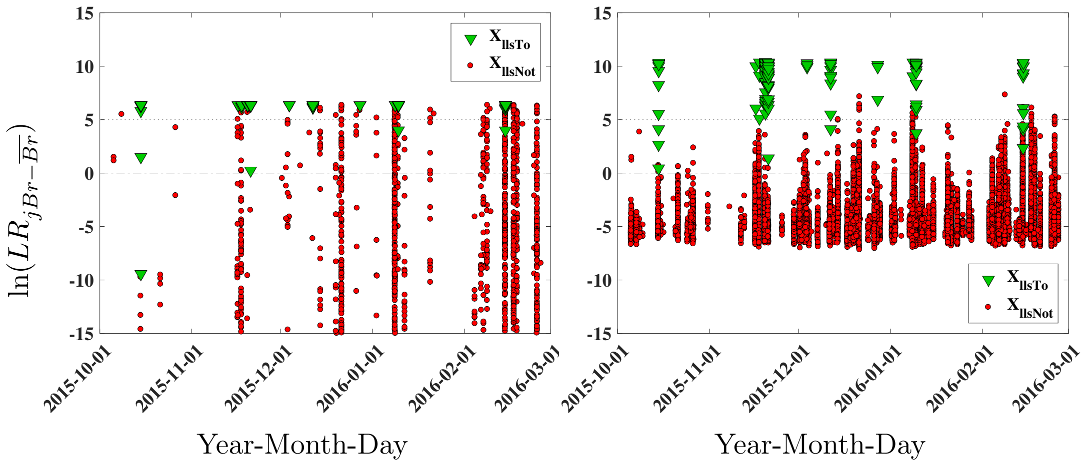
We can evaluate the different likelihood functions by comparing how many TPs and TNs the log-likelihood ratio critical values V_{crit} classify.

8.3.1 Log-likelihood Ratio Results

Figure 8.2 shows the log-likelihood ratio $\ln(LR_{jBr-\overline{Br}})$ (equation 8.1) vs. t_j, the time-stamp of each detected stroke $\boldsymbol{X_{lls_j}}$. Both $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$● are indicated on the figures. In figure 8.2a, the likelihood function is set to the bivariate Gaussian distribution and in figure 8.2b, it is set to the bivariate Students' t-distribution. The ground-truth comparison criteria are also indicated on the figures.

Looking at figure 8.2a, we can see that all true strikes to the Brixton tower ▼ have the highest likelihood ratio value. This is as expected given the claim that $\boldsymbol{L_{Br_S}}$ (Brixton tower was struck by lightning) is true. We can also see that there were at least two events to the tower every month of the season under investigation. In contrast, strikes that were *NOT* to Brixton tower ●, mostly have very low likelihood ratio values which is also to be expected. However, there is still a significant number of these strokes that have high likelihood ratio values in comparison with the true strokes, making it impossible to distinguish between high value likelihood ratio values.

If we consider figure 8.2b, where the bivariate Students' t-distribution is used as the likelihood function, we make a different observation. Firstly, the overall value of the



(a) Gaussian (b) Students' t

Figure 8.2: Log-likelihood Ratio $\ln(LR_{jBr-\overline{Br}})$ vs. t_j (Year-Month-Day HH:MM:SS UTC) with likelihood function set as (a) bivariate Gaussian distribution and (b) bivariate Students' t-distribution.

likelihood ratio for all strokes is much higher than in figure 8.2a, with no log-likelihood ratio value falling under -10. However, we can see that the true strokes ▼ dominate the higher range values and there is much more separation between strokes that truly attached to the Brixton tower and strokes that did not. This makes sense given the nature of the bivariate Students' t-distribution - it includes more outliers and therefore the likelihood value for each reported stroke will be higher. However, it has much steeper gradient closer to the mean, greatly preferencing strokes reported close to the true location.

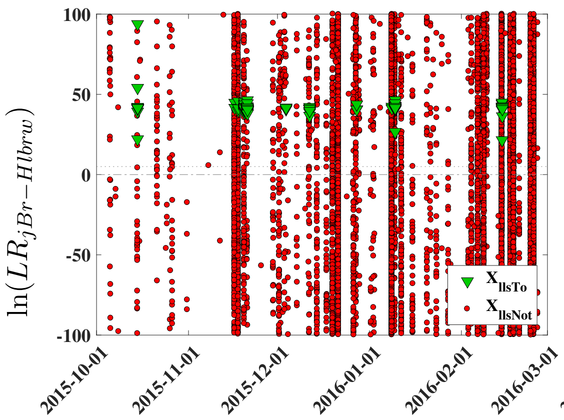
If we apply the critical values discussed in the previous section, we can categorise each stroke as either TP, FP, TN or FN and assess the effectiveness of the two different likelihood functions. Table 8.8 summarises this with a number of performance metrics shown as well. Accuracy (Acc.) and precision (Prec.) are calculated as per equation 4.13 and 4.14 in Chapter 4, Section 4.2.

Table 8.8: Ground-truth comparison of bivariate Gaussian and Students' t-models as likelihood functions.

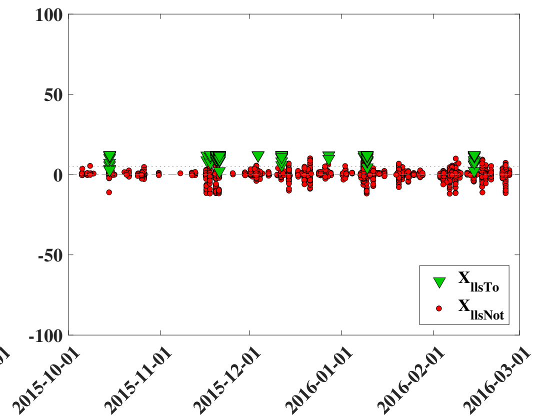
	$\ln(LR_j) > 0$		$\ln(LR_j) > 5$	
	Gaussian	Students' t	Gaussian	Students' t
TP	142	143	138	137
FP	451	1392	128	36
TN	66186	65254	66509	66601
FN	1	0	5	6
TP rate	0.993	1	0.965	0.958
FP rate	0.0068	0.0209	0.0019	0.00055
Acc.	0.9932	0.972	0.998	0.9994
Prec.	0.2395	0.0932	0.5188	0.7919

The bivariate Gaussian distribution as a likelihood function appears to perform better (lower FP rate, higher accuracy and precision) at the lower critical value threshold but at the higher threshold, the bivariate Students' t-distribution makes for a better likelihood function (lower FP rate, higher accuracy and precision). This emphasizes the observation in figure 8.2b, that there is better separation between strokes that truly attached to the Brixton tower and strokes that did not.

Next, the log-likelihood ratio of whether a reported LLS stroke struck Brixton tower or struck Hillbrow tower is examined. Similar to figure 8.2, figure 8.3 shows the log-likelihood ratio $\ln(LR_{jBr-Hlbrw})$ (equation 8.2) vs. t_j. Once again, $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$● are indicated on the figure. The distinction between when the bivariate Gaussian distribution and when the bivariate Students' t-distribution is used as the likelihood function is similarly shown in figures 8.3a and 8.3b respectively.



(a) Gaussian



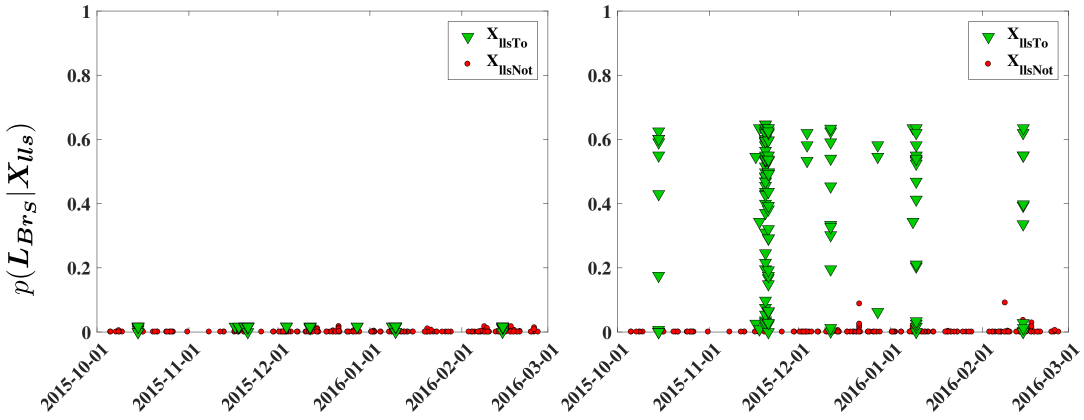
(b) Students' t

Figure 8.3: Log-likelihood Ratio $\ln(LR_{jBr-Hlbrw})$ vs. t_j (Year-Month-Day HH:MM:SS UTC) with likelihood function set as (a) bivariate Gaussian distribution and (b) bivariate Students' t-distribution.

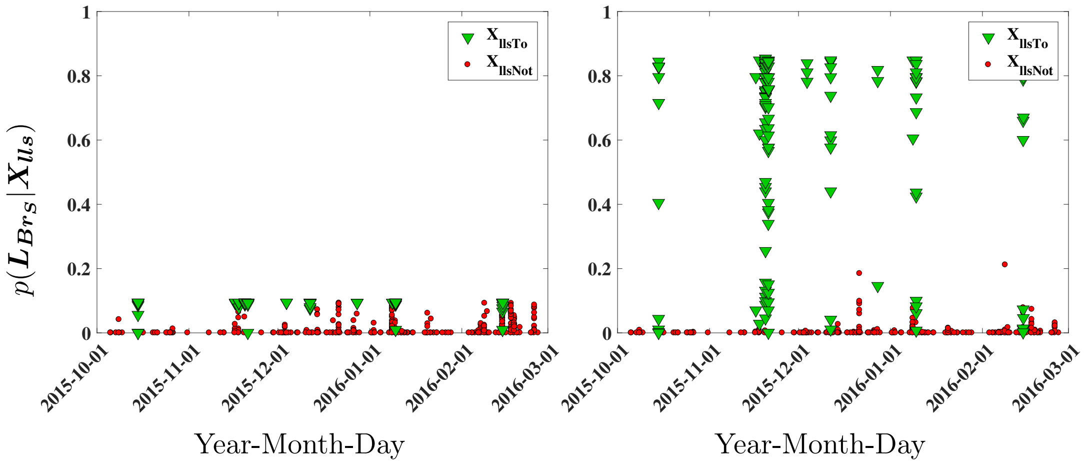
Considering figure 8.3a, the likelihood ratio values are high for the reported strokes that truly struck Brixton tower. In the previous case the location of the Brixton tower was being compared to all other locations whereas now, it is only being compared to the location of the Hillbrow tower so this is expected. All other strokes (strokes that did not attach to Brixton tower) have a very large range of values, some greater than the true strokes themselves. This is because, no matter how far a stroke was reported from either of the towers we are still looking at the likelihood ratio of it having attached to Brixton tower over the Hillbrow tower - even if it attached to neither. However, when we use the bivariate Students' t-distribution as the likelihood function, we obtain a different plot as seen in figure 8.3b. Here, the strokes that truly attached to the Brixton tower have the highest values and the range of log-likelihood values for all other strokes is much smaller, making the true cases clearer.

8.3.2 Posterior Probability Results

Here, we include the prior probability function and look at the posterior probability of the Brixton tower being struck by lightning for each reported SALDN stroke $\boldsymbol{X_{lls_j}}$. This is done firstly with the non-informative/flat prior probability function and secondly with the informative prior probability function. Figure 8.4 shows the posterior probability of Brixton tower having been struck by lightning $p(\boldsymbol{L_{Br_S}}|\boldsymbol{X_{lls}})$ against the time-stamp of each detected stroke t_j for both $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$●.



(a) Non-informative with Gaussian (b) Non-informative with Students' t



(c) Informative with Gaussian (d) Informative with Students' t

Figure 8.4: Posterior probability $p(\boldsymbol{L_{Br_S}}|\boldsymbol{X_{lls}})$ vs. t_j (Year-Month-Day HH:MM:SS UTC) with (a) with Non-informative prior and likelihood function set as bivariate Gaussian distribution and (b) bivariate Students' t-distribution and (c) Informative prior with likelihood function set as bivariate Gaussian distribution and (d) bivariate Students' t-distribution.

Figure 8.4a shows the posterior probability of Brixton tower being struck by lightning with the bivariate Gaussian model as the likelihood function and the uniform (flat) model as the prior probability function. All values of probability are low here (under 10%) making distinction between $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$● difficult. Figure 8.4b shows the same but with the bivariate Students' t-model as the likelihood function. Here we see that strokes from $\boldsymbol{X_{lls_{To}}}$▼ have much higher probabilities, many greater than 50% but strokes from $\boldsymbol{X_{lls_{Not}}}$● still have much lower posterior probability values. This is the same as what was seen with the log-likelihood ratios, where the bivariate Students' t-model makes a much better distinction between the TPs and the TNs.

Figures 8.4c and 8.4d show the same two figures - the posterior probability with the bivariate Gaussian model (figure 8.4c) and the bivariate Students' t-model (figure 8.4d) as likelihood functions - but with the GMM as an informative prior probability function. We can already see the difference in using the informative prior probability function with the Gaussian likelihood function, where strokes from $\boldsymbol{X_{lls_{To}}}$▼ have increased probability values as compared to the non-informative case. However, these values are still fairly low probabilities. Similarly, in the case of the bivariate Students' t-model (figure 8.4d), the probability values have increased, with many strokes from $\boldsymbol{X_{lls_{To}}}$▼ having probability values greater than 80%. While the use of the informative prior probability model clearly helps to distinguish strokes that did attach to the Brixton tower, the use of the bivariate Students' t-distribution as the likelihood function has a much greater impact on this distinction.

8.3.3 Conclusions

By applying the Bayesian Framework to a case study in which we aim to determine whether the Brixton tower was struck by lightning over the 2015-2016 lightning season period, we are also able to see the effect of using different likelihood and prior probability statistical models. Firstly, looking at the likelihood ratio LR of whether Brixton tower was struck or NOT struck, we see that the bivariate Students' t-distribution as a model for the likelihood function does much better at classifying the strokes that were time-correlated with events to Brixton tower $\boldsymbol{X_{lls_{To}}}$▼ from events that did not attach to the Brixton tower $\boldsymbol{X_{lls_{Not}}}$● than the bivariate Gaussian distribution as a likelihood model.

This can also be seen when we look at the likelihood ratio LR of whether Brixton or Hillbrow tower was struck by a stroke - with the likelihood function as a bivariate Students' t-distribution, the strokes $\boldsymbol{X_{lls_{To}}}$▼ are distinguished from all other strokes while it is impossible to make such a distinction with the bivariate Gaussian model. By looking at the posterior probabilities that each stroke attached to the Brixton tower, we are able to see the effect of the prior probability function. While the GMM does make differentiation between $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$● easier than a non-informative, uniform prior probability function, the difference is minimal and the effect of using the bivariate Students' t-distribution as a likelihood function has a much greater improvement on classification.

8.4 Chapter Summary

This chapter has demonstrated the application of the Bayesian framework in a real-world case study in which South African Lightning Detection Network (SALDN) stroke reports are utilized in a forensic investigation.

The region of Johannesburg, South Africa, where the Brixton tower and the Hillbrow tower are located, is framed as a forensic investigation. The investigation is to determine when the Brixton tower was struck by lightning during the 2015-2016 thunderstorm season. The requirements of applying the Bayesian framework are demonstrated - defining a geographic area of investigation, 1600 km^2 around Johannesburg and a time-frame of investigation, 1 October 2015 to 1 March 2016. The Johannesburg region falls within the coverage of the SALDN and the data within the region for the 2015-2016 time-frame is obtained for the investigation, $\boldsymbol{X_{lls}}$. Data from before the time-frame is also obtained to provide prior knowledge about lightning activity in the region, $\boldsymbol{X_{llsPr}}$.

Given the photographic study described in Chapter 7, it is known when the Brixton tower was struck by lightning and the SALDN strokes that were measurements of events to the tower are differentiated from events that did not by time-correlation - $\boldsymbol{X_{lls_{To}}}$▼ and $\boldsymbol{X_{lls_{Not}}}$●. Using the concepts of True Positives (TPs), True Negatives (TNs), False Positives (FPs) and False Negatives (FNs), the effectiveness of the Bayesian framework and the different statistical models is evaluated.

The likelihood function is set to both the bivariate Gaussian distribution and the Students' t-distribution and, through the use of log-likelihood ratio plots it is clear that the bivariate Students' t-distribution distinguishes SALDN strokes that attached to the tower (▼) from strokes that did not (●) much better than the bivariate Gaussian distribution when set as the likelihood function. The prior probability function is set to both he bivariate uniform distribution and a bivariate Gaussian Mixture Model (GMM) calculated from $\boldsymbol{X_{llsPr}}$. Using posterior probability plots it is seen that the bivariate GMM (informative prior) makes a clearer distinction between strokes that attached to Brixton tower (▼) and strokes that did not (●).

Finally, this chapter has shown that the Bayesian framework is effective at quantifying the uncertainty in Lightning Location System (LLS) stroke reports and, through the use of likelihood ratios and posterior probability plots, is an effective means of presenting LLS data as evidence in forensic investigations.

The next chapter presents the final Conclusions of the thesis.

Chapter 9

Conclusion

Modern Lightning Location Systems (LLSs) have become increasingly used in forensic investigations where lightning may be a possible cause of damage to equipment and assets, or even loss of life. This is of interest to a range of industries, from power utilities to insurance companies. While the detection of lightning strokes and the reporting of their time and location is valuable information that was previously unavailable, there is still uncertainty in these reports due to location errors. In legal situations, where decisions are on the basis of forensic investigation, this uncertainty can cast doubt on the outcome if not properly quantified and communicated. While there exist some approaches to calculating probabilities of LLS stroke reports attaching in specific areas, these rely on the assumption that LLS location errors follow a bivariate Gaussian distribution and do not include any prior knowledge about the region or case in the calculation. Also, there exists no consistent approach to communicating these probabilities.

Bayesian probability theory and inference is a powerful framework for assessing the uncertainty of events and outcomes. The manner in which prior information is updated by measured evidence in the form of the likelihood function makes it ideally suited to forensic investigation and is becoming widely used as means of communicating forensic evidence in legal forums, through the use of likelihood ratios and posterior probabilities.

This thesis, utilizing a Bayesian approach, has made the following contributions to the use of LLS data in forensic investigations:

- A probabilistic Bayesian framework for utilizing LLS stroke reports in forensic investigations where lightning is a possible cause. The framework provides a method for reporting LLS data that is consistent with the reporting of evidence in legal forums - in the form of likelihood ratios and posterior probabilities, including prior information: Investigators are able to define an area and time-frame of investigation, and determine the probability of any location in the area being struck by lightning given reported LLS strokes in the area/time-frame, including prior LLS reports in the area;
- Verification of the statistical models employed in the framework against ground-truth data from two different LLSs in two different countries (South Africa and

Austria), and a demonstration of its application in a real-world case study: Ground-truth lightning events to the Brixton tower in South Africa and the Gaisberg tower in Austria are time-correlated with the data from the respective LLSs (143 strokes and 16 respectively). These data sets allow for modelling of the distribution of LLS location errors, and in the South African case study, a comparison point to determine the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) events classified by the framework;

- A methodology for fitting bivariate statistical models to LLS data and evaluating the quality of the fit of the models: Given LLS stroke reports are a bivariate data set, Maximum Likelihood Estimation (MLE) and Expectation-Maximization (EM) approaches are utilized to fit parameters of different bivariate distributions to LLS data and the quality of the fit is assessed based on a bivariate goodness-of-fit test proposed by McAssey (2013), using the Mahalanobis distance;

- The bivariate Students' t-distribution is a better fitting model for the location errors present in LLS reports than the more commonly assumed bivariate Gaussian distribution due to its heavier tails: Applying the bivariate statistical modelling techniques to the ground-truth data sets from South Africa yielded a Mahalanobis test statistic of 23.0719 for the bivariate Gaussian model and 17.8075 for the bivariate Students' t-model, and in Austria, 22.5194 and 5.3849 respectively. In the case study, the bivariate Students' t-model had a lower FP rate, higher accuracy and precision than the bivariate Gaussian model;

- Gaussian Mixture Models (GMMs) can be used to model the distribution of lightning over an area for prior probability calculations: Applying the bivariate statistical modelling techniques to LLS stroke reports within the Johannesburg, South Africa region from multiple thunderstorm seasons, a Mahalanobis test statistic of 0.0363 is found for the GMM in comparison to the 0.6803 of the bivariate uniform distribution and the 4.5103 of the bivariate Gaussian distribution.

These results provide valuable contributions to the use of LLS stroke reports in forensic investigations as well as to the general field of LLS location accuracy and system performance.

9.1 Recommendations for Future Research

The conclusions of this thesis provides a base on which further research in two areas can be performed:

1. Confidence ellipses and the self-evaluation techniques to estimate location errors. The self-evaluation of location accuracy in LLSs relies on fitting univariate Gaussian functions to the timing errors of individual sensors. If instead, univariate Students' t-distributions were fitted, the confidence ellipses could based around the bivariate Students' t-distribution and may provide tighter ellipses for more accurately located strokes and distinguish large location errors better.

2. Ground-Strike Point (GSP) grouping algorithms. These algorithms determine which strokes in a cluster attached to one location and which attached to a different location. The Bayesian framework would be the perfect method to develop a *soft-clustering* type algorithm that assesses the probability of there being strokes to multiple ground-strike points rather than the *hard-clustering* of k-means type algorithms currently used.

Appendix A

Sufficiently Small, Discrete vs Continuous

This appendix discusses the definition of "sufficiently small" in regards to Bayesian inference and the distinction between continuous and discrete variables used in the Bayesian framework. A simple hypothetical, univariate example of the Bayesian framework is also presented.

A.1 Introduction

The concept of "sufficiently small" is alluded to a number of times in this thesis. Specifically, in Chapter 4, Section 4.2.4 on Bayesian inference and the continuous, discrete and mixed form of Bayes' theorem and again in Chapter 5, Section 5.2.1, on the assumptions of the Bayesian framework for Lightning Location System (LLS) stroke reports - in particular, assumption 3. In both these cases, the concept is a condition for approximating a continuous variable by a discrete one.

In Chapter 5, the concept is used to define the set of all possible locations $\boldsymbol{L}$ within the area of investigation as a finite, discrete variable of length N. Of course, we know this is not the case as the set of all possible locations within in an area must be a continuous variable with an infinite number of locations. This presents a problem in that, the prior probability of any one location within this set having been struck by lightning will be zero, making calculations of probabilities meaningless. The assumption discussed in the chapter is that each location can be approximated by a small area (making $\boldsymbol{L}$ discrete), and therefore making the calculation possible. This also then allows the mixed form of Bayes' theorem to be used, removing the need for integration to be performed. The question of how small each location's area should be is what is termed "sufficiently small", and this is defined here.

Also discussed here, is a simple hypothetical, univariate example of the Bayesian framework which helps to explain the Bayesian framework as well as demonstrate the approximation of $\boldsymbol{L}$ as a discrete variable and the definition of "sufficiently small". This section also shows why the Gaussian Mixture Model (GMM) in Chapter 5, Section 5.4

is a valid approach to the prior probability function within the premise of the Bayesian framework itself.

A.2 Defining *Sufficiently Small*

If we consider the random variable Y discussed in figure 4.1 from Chapter 4 here:

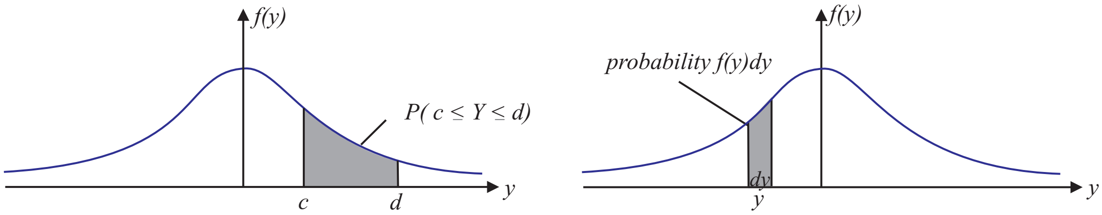


Figure A.1: Probability density $f(y)$ of random variable Y.

We recall that the probability of Y being between two values $[c, d]$ is given by:

$$P(c \leq Y \leq d) = \int_c^d f(y)dy \tag{A.1}$$

If $[c, d]$ becomes an infinitesimal range ie. dy a single point, then the probability that Y is in this infinitesimally range is $f(y)dy$. So, if $[c, d]$ is very small that it is almost a single point, but not quite (Δy rather than dy) the probability that Y is in this range is $f(y)\Delta y$, at least approximately. Therefore, we can approximate equation A.1 as such:

$$P(c \leq Y \leq d) = \int_c^d f(y)dy \approx f(y)\Delta y \tag{A.2}$$

Figure A.2 shows a plot of the probability of the random variable Y being within the range $\Delta y = [c, d]$ against Δy. $f(y)$ is a univariate Gaussian function with mean $\mu = 0$ and variance $\sigma = 1$. Two curves are shown: the integral between the limits $[c, d]$ and the approximation, $f(y)\Delta y$.

As the range $\Delta y = [c, d]$ grows, the approximation $f(y)\Delta y$ becomes a less accurate estimation of the probability. The approximation completely fails for values of $f(y)\Delta y$ greater than ≈ 2, where probabilities greater than 1 are calculated - which is meaningless. Nonetheless, we can see that for very small values of $f(y)\Delta y$ (less than ≈ 0.25) the probability is estimated correctly. It is this concept that defines "sufficiently small" for the purposes of this thesis.

A.3 Univariate Bayesian Framework Example

Consider the hypothetical situation in which there are 10 different locations $\boldsymbol{L_{1-10}}$ in a line and it is only possible for these locations to be struck by lightning and nowhere

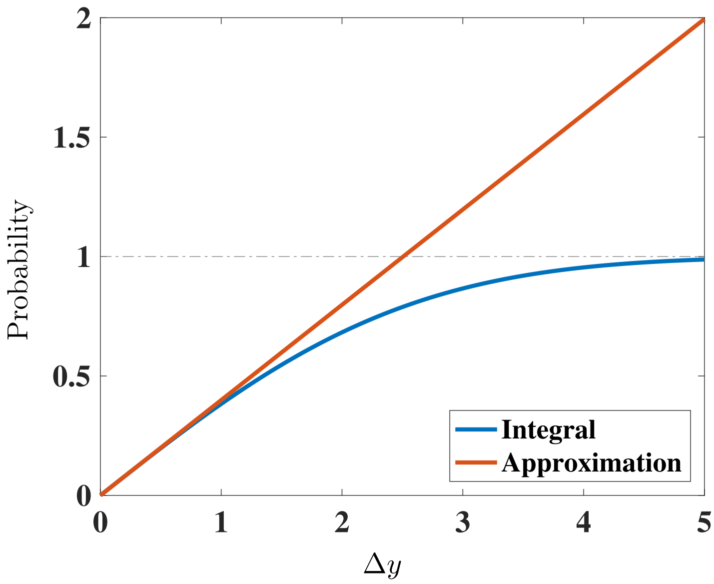


Figure A.2: Probability of a random variable Y being in the range $\Delta y = [c, d]$, versus Δy. The two curves show the calculation using the integral $\int_c^d f(y)dy$ and the approximation $f(y)\Delta y$

elsewhere (a real-world approximation may be 10 equally tall towers, evenly spaced and with protection areas that cover all possible locations between them). If we consider the probability of each of these locations being struck by lightning with no prior knowledge ($\frac{1}{10}$), we find ourselves with a uniform prior probability function as shown in figure A.3:

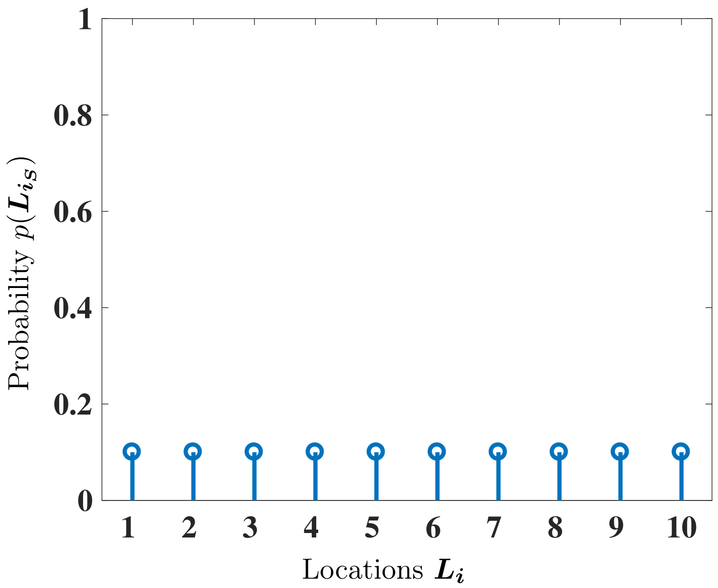


Figure A.3: Uniform prior probability of all possible locations that may be struck by lightning.

Now we consider that this region or "line" of locations falls within the coverage of a LLS which can detect any of the events to $\boldsymbol{L_{1-10}}$ and this will be reported as $\boldsymbol{X_{lls}}$. However, the LLS is prone to location errors and may not report a detected event at its exact location. Unlike $\boldsymbol{L}$, $\boldsymbol{X_{lls}}$ is continuous and can be anywhere between $\boldsymbol{L_1}$ and $\boldsymbol{L_{10}}$. Let us also assume that the likelihood function $f(\boldsymbol{X_{lls}}|\boldsymbol{L_{is}})$ is a Gaussian distribution with variance 0.5. Therefore, if for example, location $\boldsymbol{L_5}$ is repeatedly struck by lightning, the LLS will report stroke locations that follow a distribution as shown in figure A.4a. This likelihood function of course can be applied to each of the 10 locations and figure A.4b shows the same function for location $\boldsymbol{L_5}$ along with the likelihood function for each of the other locations.

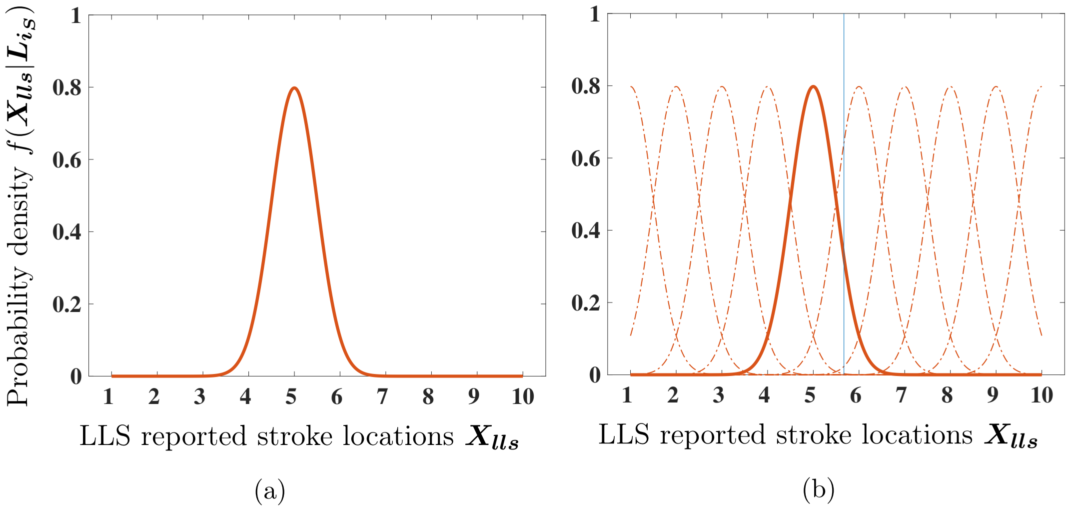


Figure A.4: Likelihood functions describing the distribution of LLS stroke reports when (a) location $\boldsymbol{L_5}$ is struck by lightning (b) locations $\boldsymbol{L_{1-10}}$ are struck by lightning. The stroke report $X_{lls} = 5.7$ is indicated by the blue, vertical line.

Now, consider a particular time-frame in which the LLS detects a stroke with reported location $X_{lls} = 5.7$ (indicated in the figure with a blue, vertical line). Due to the presence of location errors, the stroke report is in between two locations - there is a possibility that it attached to any one of the 10 different locations and we can see how these differ from where $X_{lls} = 5.7$ falls in each of the likelihood functions - $f(X_{lls} = 5.7|\boldsymbol{L_{is}})$. The question is now, which location was actually struck?

We can use our posterior probability equation from our Bayesian framework to calculate the probability of each location having been struck given the LLS stroke report:

$$p(\boldsymbol{L_{is}}|X_{lls} = 5.7) = \frac{p(\boldsymbol{L_{is}})f(X_{lls} = 5.7|\boldsymbol{L_{is}})}{\sum_{i=1}^{N} p(\boldsymbol{L_{is}})f(X_{lls} = 5.7|\boldsymbol{L_{is}})} \tag{A.3}$$

Figure A.5 shows the posterior probability (equation A.3) of each location having been struck given the LLS stroke report.

Given the Gaussian likelihood function, the uniform prior probability function and the reported stroke location $X_{lls} = 5.7$, the location with the highest probability of being struck is location $\boldsymbol{L_6}$, followed by $\boldsymbol{L_5}$. In this way, the probability of a location being struck by lightning given location errors present in LLS stroke reports can be assessed.

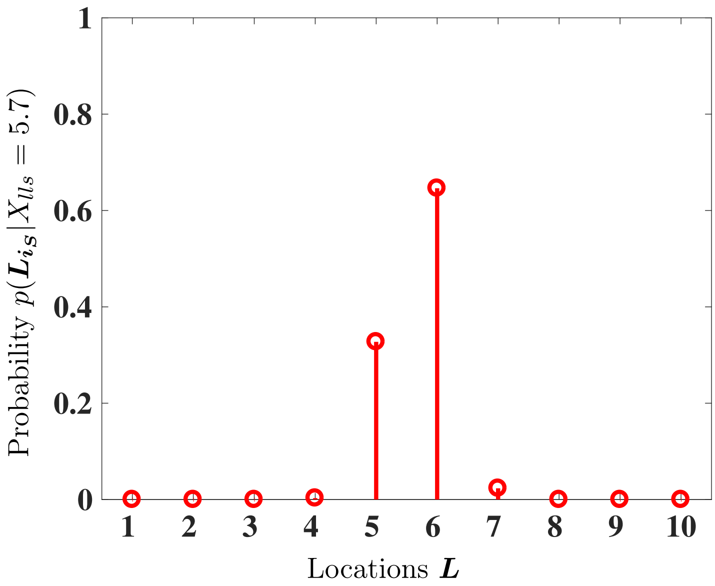


Figure A.5: Posterior probability $p(\boldsymbol{L_{is}}|X_{lls} = 5.7)$ for each location $\boldsymbol{L_{1-10}}$.

Of course, if prior information is available - previous LLS or ground-truth data - that suggest one of the locations is struck more often than the others, our prior probability function may look more like something in figure A.6a. Then, using equation A.3 the posterior probability becomes that shown in figure A.6b. The probability that location $\boldsymbol{L_6}$ was struck is even greater given what we know about how often it was struck by lightning previously.

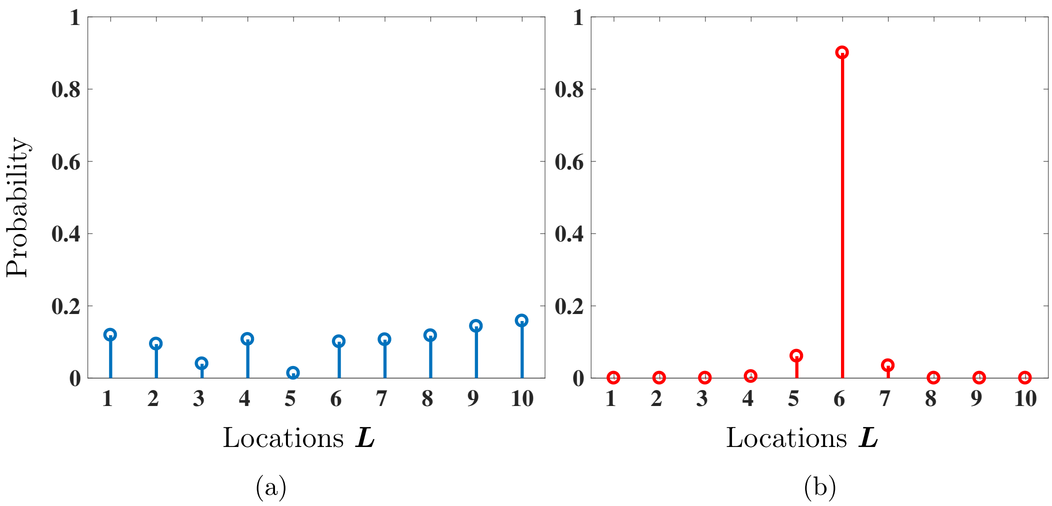


Figure A.6: Probability of being struck by lightning against locations $\boldsymbol{L}$ (a) non-uniform (informative) prior probability $p(\boldsymbol{L_{is}})$ (b) posterior probability $p(\boldsymbol{L_{is}}|X_{lls} = 5.7)$ for each location $\boldsymbol{L_{1-10}}$ with non-uniform prior probability function.

This example provides a good description of how the Bayesian framework functions although, only in one dimension. The complete framework utilizes the multivariate domain of both longitude and latitude. The other simplification of this example is of course the discrete nature of $\boldsymbol{L}$ and the large separation between each location. For the framework described in Chapter 5, we wish to approximate a continuous variable by making the difference between each location "sufficiently small". Figure A.7 shows the univariate case but with 40 different locations (rather than just 10), within the same distance - the distance between each location has become smaller, closer approximating a continuous set of locations. If the criteria for sufficiently small differences are met, the continuous case will be approximated.

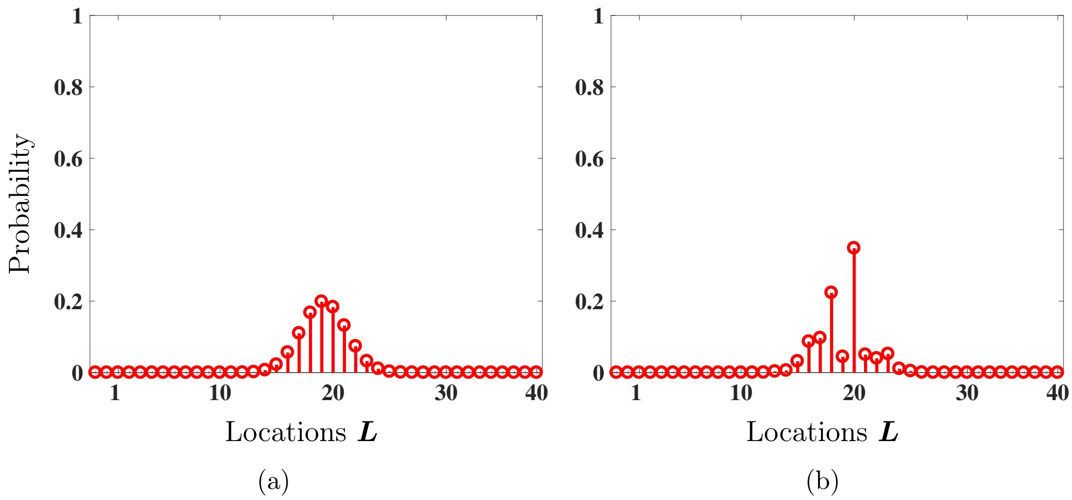


Figure A.7: Posterior probability $p(\boldsymbol{L_{is}}|X_{lls} = 5.7)$ for each location $\boldsymbol{L_{1-40}}$ with (a) uniform prior probability function (b) with non-uniform prior probability function.

Figure A.7a shows the posterior probability results for the 40 different locations with a uniform prior probability function. Here we can see that the Gaussian likelihood function and the posterior probability function are the same (except one is a function $\boldsymbol{L}$ the other of $\boldsymbol{X_{lls}}$). This observation is used for the GMM approach to the prior probability distribution - if each stroke report in the prior data set was the mean of the posterior distribution with a uniform prior, that can now become the prior for the next investigation. This is updating property of Bayes' theorem. Figure A.7b shows the posterior probability results for the 40 different locations with a non-uniform prior probability function and we can see how the likelihood function and the prior probability interact to produce the posterior distribution.

A.4 Conclusions

If the area of investigation for application of the Bayesian framework is divided into a grid with square sizes that are "sufficiently small", then each grid square will approximate a point location. "Sufficiently small" is the size of the grid square in which the product of the distribution and the gird square size is equivalent integral over the area of the grid square size. This means that, for the purposes of the Bayesian framework, the parameter (or claim) locations $\boldsymbol{L}$ can be considered discrete and the measurement (or evidence), LLS stroke reports $\boldsymbol{X_{lls}}$ can be considered as continuous. This means that the mixed form of Bayes' theorem can be used where the likelihood function is a probability density and the prior probability function is a probability mass function. Finally, the use of a Gaussian Mixture Model (GMM) for the prior probability function is valid because of the updating property of Bayesian probability.

Appendix B

Lightning Events to Brixton Tower 2015-2016

This appendix provides a detailed breakdown of all photographed flashes to the Brixton tower and the exact date and time they occurred. during the 2015-2016 season. The number of time-correlated SALDN strokes for each event are also provided.

B.1 Introduction

This appendix provides the complete list of all photographs of lightning attachments to the Brixton tower during the 2015-2016 thunderstorm season in Johannesburg, South Africa used in Chapters 7 and 8 as ground-truth evidence of lightning attachment to a known location.

B.2 Photographed Flashes

Photographs of lightning attachments to the Brixton tower in Johannesburg, South Africa were captured over the period of 1 October 2015 to 1 March 2016. A total of 55 photographs were obtained and these were categorised into upward and downward events from visual inspection of branching. Table B.2 shows lists each flash along with its type (upward or downward) and the the date and time the flash to the Brixton tower was captured in both South African Standard Time (SAST) and Universal Coordinated Time (UTC).

Table B.1: Date and time of photographed flashes 2015 - 2016.

Photograph (type)	Date and Time (SAST)	Date and Time (UTC)
1 (Upward)	14/10/2015 22:05:21	14/10/2015 20:05:21

Continued on next page

Table B.1 – *Continued from previous page*

Photograph (type)	Date and Time (SAST)	Date and Time (UTC)
2 (Upward)	14/10/2015 22:07:10	14/10/2015 20:07:10
3 (Upward)	14/10/2015 22:08:36	14/10/2015 20:08:36
4 (Upward)	14/10/2015 22:10:30	14/10/2015 20:10:30
5 (Downward)	14/10/2015 22:48:19	14/10/2015 20:48:19
6 (Upward)	14/10/2015 22:53:25	14/10/2015 20:53:25
7 (Downward)	14/10/2015 22:57:19	14/10/2015 20:57:19
8 (Upward)	16/11/2015 16:35:08	16/11/2015 14:35:08
9 (Upward)	16/11/2015 18:42:10	16/11/2015 16:42:10
10 (Upward)	16/11/2015 19:34:51	16/11/2015 17:34:51
11 (Upward)	17/11/2015 21:51:53	17/11/2015 19:51:53
12 (Upward)	17/11/2015 21:55:38	17/11/2015 19:55:38
13 (Upward)	17/11/2015 21:58:06	17/11/2015 19:58:06
14 (Downward)	19/11/2015 22:12:54	19/11/2015 20:12:54
15 (Upward)	19/11/2015 22:16:32	19/11/2015 20:16:32
16 (Upward)	19/11/2015 22:38:41	19/11/2015 20:38:41
17 (Upward)	19/11/2015 22:41:16	19/11/2015 20:41:16
18 (Upward)	19/11/2015 22:42:54	19/11/2015 20:42:54
19 (Upward)	19/11/2015 22:45:41	19/11/2015 20:45:41
20 (Upward)	19/11/2015 22:47:12	19/11/2015 20:47:12
21 (Upward)	19/11/2015 22:51:51	19/11/2015 20:51:51
22 (Upward)	19/11/2015 22:56:22	19/11/2015 20:56:22
23 (Upward)	19/11/2015 23:07:58	19/11/2015 21:07:58
24 (Upward)	20/11/2015 21:11:12	20/11/2015 19:11:12
25 (Downward)	20/11/2015 21:14:37	20/11/2015 19:14:37
26 (Upward)	20/11/2015 21:39:40	20/11/2015 19:39:40
27 (Upward)	20/11/2015 21:43:23	20/11/2015 19:43:23
28 (Upward)	20/11/2015 22:17:57	20/11/2015 20:17:57
29 (Downward)	20/11/2015 22:21:58	20/11/2015 20:21:58
30 (Upward)	20/11/2015 22:22:56	20/11/2015 20:22:56
31 (Upward)	20/11/2015 22:24:48	20/11/2015 20:24:48
32 (Upward)	20/11/2015 22:49:50	20/11/2015 20:49:50
33 (Upward)	20/11/2015 22:52:51	20/11/2015 20:52:51
34 (Upward)	20/11/2015 22:54:37	20/11/2015 20:54:37
35 (Upward)	20/11/2015 22:59:55	20/11/2015 20:59:55
36 (Upward)	20/11/2015 23:02:59	20/11/2015 21:02:59
37 (Upward)	20/11/2015 23:43:56	20/11/2015 21:43:56
38 (Upward)	20/11/2015 23:47:40	20/11/2015 21:47:40
39 (Upward)	21/11/2015 22:17:04	20/11/2015 22:17:04
40 (Upward)	02/12/2015 23:48:49	02/12/2015 21:48:49
41 (Upward)	03/12/2015 23:26:30	03/12/2015 21:26:30
42 (Upward)	04/12/2015 23:31:47	03/12/2015 23:31:47
43 (Upward)	04/12/2015 23:33:59	03/12/2015 23:33:59
44 (Upward)	11/12/2015 22:12:14	11/12/2015 20:12:14
45 (Upward)	11/12/2015 22:17:53	11/12/2015 20:17:53
46 (Upward)	27/12/2015 18:58:57	27/12/2015 16:58:57

Continued on next page

Table B.1 – *Continued from previous page*

Photograph (type)	Date and Time (SAST)	Date and Time (UTC)
47 (Upward)	08/01/2016 15:54:38	08/01/2016 13:54:38
48 (Upward)	09/01/2016 18:11:58	09/01/2016 16:11:58
49 (Upward)	09/01/2016 18:15:11	09/01/2016 16:15:11
50 (Upward)	09/01/2016 18:17:50	09/01/2016 16:17:50
51 (Downward)	14/02/2016 19:14:30	14/02/2016 17:14:30
52 (Downward)	14/02/2016 19:17:25	14/02/2016 17:17:25
53 (Downward)	14/02/2016 21:20:48	14/02/2016 19:20:48
54 (Downward)	26/02/2016 20:38:35	26/02/2016 18:38:35
55 (Upward)	26/02/2016 20:43:17	26/02/2016 18:43:17

B.3 Time-Correlated SALDN Data

Table B.3 shows the same photographed lightning flashes to the Brixton tower but now with the time-correlated data from the South African Lightning Detection Network (SALDN) provided as well. The date, time and type of each flash is shown, along with whether it was detected by the SALDN and if so, how many stroke reports were time-correlated with the event. A total of 143 SALDN strokes were time-correlated with the 55 flashes to Brixton tower during the 2015-2016 thunderstorm season.

Table B.2: SALDN data time-correlated with photographed flashes 2015 - 2016.

Photograph (type)	Date and Time (UTC)	Detected by SALDN	No. Strokes
1 (Upward)	14/10/2015 20:05:21	Yes	4
2 (Upward)	14/10/2015 20:07:10	No	-
3 (Upward)	14/10/2015 20:08:36	No	-
4 (Upward)	14/10/2015 20:10:30	Yes	2
5 (Downward)	14/10/2015 20:48:19	No	-
6 (Upward)	14/10/2015 20:53:25	Yes	4
7 (Downward)	14/10/2015 20:57:19	Yes	1
8 (Upward)	16/11/2015 14:35:08	-	2
9 (Upward)	16/11/2015 16:42:10	No	-
10 (Upward)	16/11/2015 17:34:51	No	-
11 (Upward)	17/11/2015 19:51:53	Yes	3
12 (Upward)	17/11/2015 19:55:38	No	-
13 (Upward)	17/11/2015 19:58:06	No	-
14 (Downward)	19/11/2015 20:12:54	No	-
15 (Upward)	19/11/2015 20:16:32	Yes	5
16 (Upward)	19/11/2015 20:38:41	Yes	1
17 (Upward)	19/11/2015 20:41:16	No	-
18 (Upward)	19/11/2015 20:42:54	Yes	4

Continued on next page

Table B.2 – *Continued from previous page*

Photograph (type)	Date and Time (UTC)	Detected by SALDN	No. Strokes
19 (Upward)	19/11/2015 20:45:41	Yes	8
20 (Upward)	19/11/2015 20:47:12	Yes	3
21 (Upward)	19/11/2015 20:51:51	Yes	11
22 (Upward)	19/11/2015 20:56:22	Yes	15
23 (Upward)	19/11/2015 21:07:58	No	-
24 (Upward)	20/11/2015 19:11:12	No	-
25 (Downward)	20/11/2015 19:14:37	No	-
26 (Upward)	20/11/2015 19:39:40	Yes	4
27 (Upward)	20/11/2015 19:43:23	Yes	1
28 (Upward)	20/11/2015 20:17:57	Yes	9
29 (Downward)	20/11/2015 20:21:58	No	-
30 (Upward)	20/11/2015 20:22:56	Yes	3
31 (Upward)	20/11/2015 20:24:48	Yes	9
32 (Upward)	20/11/2015 20:49:50	Yes	1
33 (Upward)	20/11/2015 20:52:51	No	-
34 (Upward)	20/11/2015 20:54:37	Yes	5
35 (Upward)	20/11/2015 20:59:55	No	-
36 (Upward)	20/11/2015 21:02:59	No	-
37 (Upward)	20/11/2015 21:43:56	Yes	1
38 (Upward)	20/11/2015 21:47:40	No	-
39 (Upward)	20/11/2015 22:17:04	No	-
40 (Upward)	02/12/2015 21:48:49	No	-
41 (Upward)	03/12/2015 21:26:30	Yes	3
42 (Upward)	03/12/2015 23:31:47	No	-
43 (Upward)	03/12/2015 23:33:59	No	-
44 (Upward)	11/12/2015 20:12:14	Yes	11
45 (Upward)	11/12/2015 20:17:53	No	-
46 (Upward)	27/12/2015 16:58:57	Yes	3
47 (Upward)	08/01/2016 13:54:38	Yes	2
48 (Upward)	09/01/2016 16:11:58	Yes	2
49 (Upward)	09/01/2016 16:15:11	No	-
50 (Upward)	09/01/2016 16:17:50	Yes	14
51 (Downward)	14/02/2016 17:14:30	Yes	1
52 (Downward)	14/02/2016 17:17:25	Yes	7
53 (Downward)	14/02/2016 19:20:48	Yes	4
54 (Downward)	26/02/2016 18:38:35	No	-
55 (Upward)	26/02/2016 18:43:17	No	-

B.4 Conclusion

A total of 55 flashes to the Brixton tower were recorded by the motion-triggered surveillance system between the 1 October 2015 and the 1 March 2016 in Johannesburg, South

Africa. 9 of these flashes were downward attachments to the tower while the other 46 were upward events that initiated from the tower. The events occurred on 10 different days with many days having multiple events throughout a storm. A total of 143 stroke detections made by the South African Lightning Detection Network (SALDN) could be time-correlated with the flashes to the Brixton tower with the most strokes (15) correlated with a upward initiated event on the 19/11/2015 at 20:56:22 UTC. In total, 29 of the 55 photographed flashes were detected by the SALDN with 4 of the downward events being detected and 25 of the upward events being detected.

Appendix C

Algorithms and Source Code Implementation

This appendix provides the source code for the algorithms utilized in the bivariate statistical modelling methodology discussed in Chapter 6 and the proposed Bayesian framework in Chapter 5 and 8.

C.1 Maximum Likelihood Estimation (MLE) - Source Code

Section 6.4 of Chapter 6 describes the maximum likelihood estimators for the statistical models for bivariate data. The following source code calculates these for a bivariate Gaussian distribution fitted to LLS stroke report data:

Listing C.1: *Matlab code MLEGaussian.m*

```
% Gaussian Maximum Likelihood Estimators - 29 January 2018
%
% This code determines the maximum likelihood parameter estimates for a bivariate
% Gaussian fit to Lightning Location System (LLS) reported stroke locations

LLSdata = importdata‘(LLSdata.csv', ',');                 % load LLS data set

lat = LLSdata.data(:,2);                                  % Select locations
long = LLSdata.data(:,3);

mydata = [long lat];

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function GausParams = MLEGaussian(mydata)

 options = statset('Display','final');
 GausFit = gmdistribution.fit(mydata,1,'Options',options);
 aNLLGauss = GausFit.NegativeLogLikelihood;
 AICg =  2*5 + 2*aNLLGauss;
 BICg = log(length(mydata))*5 + 2*aNLLGauss;

GausParams = [GausFit.mu,GausFit.Sigma,aNLLGauss,AICg,BICg];

```

```
end
```

The following source code calculates the maximum likelihood estimators for a bivariate uniform distribution fitted to LLS stroke report data:

Listing C.2: *Matlab code MLEuniform.m*

```
% Uniform Maximum Likelihood Estimators - 29 January 2018
%
% This code determines the maximum likelihood parameter estimates for a bivariate
% uniform fit to Lightning Location System (LLS) reported stroke locations

LLSdata = importdata‘(LLSdata.csv', ',');                 % load LLS data set

lat = LLSdata.data(:,2);                                  % Select locations
long = LLSdata.data(:,3);

mydata = [long lat];

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function UniParams = MLEuniform(mydata)

unif(1,:) = mle(mydata(:,1),'distribution','Uniform');
unif(2,:) = mle(mydata(:,2),'distribution','Uniform');

UniParams = [unif(1,:),unif(2,:)];

end
```

Figure 6.2 in Chapter 6, Section 6.4 shows a block diagram representation of modified EM algorithm to determine best the fitting bivariate Students' t-distribution parameters. The algorithm is based on the work by Liu and Rubin (1995) and the implementation of the function $fitt.m$ was done by Robin Ince and can be found at https://github.com/robince/tdistfit and http://www.robinince.net.

Listing C.3: *Matlab code MLEStudentst.m*

```
% ’Students t- Maximum Likelihood Estimators - 29 January 2018
%
% This code determines the maximum likelihood parameter estimates for a bivariate
% ’Students t-fit to Lightning Location System (LLS) reported stroke locations.
% It is based on the paper ``ML estimation of the t distribution using EM and its
% extensions, ECM and ”ECME by Liu and Rubin (1995). The implementation of the function
% fitt.m was done by Robin Ince and can be found at https://github.com/robince/tdistfit
% and http://www.robinince.net/blog/2013/04/18/fitting-multivariate-t-distributions-to-data/.

LLSdata = importdata‘(LLSdata.csv', ',');                 % load LLS data set

lat = LLSdata.data(:,2);                                  % Select locations
long = LLSdata.data(:,3);

mydata = [long lat];

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function StudentstParams = MLEStudentst(mydata)

 [mu_t S_t nu_t] = fitt(mydata);
 logt = mvtLogpdf(mydata,mu_t,S_t,nu_t);
 aNLLtdist = -sum(logt);
```

```
 AICt =  2*6 + 2*aNLLtdist;
 BICt = log(length(mydata))*6 + 2*aNLLtdist;

 StudentstParams = [mu_t,S_t,nu_t,aNLLtdist,AICt,BICt];

end
```

The following source code calculates the maximum likelihood estimators for a bivariate Cauchy distribution and utilizes the modified EM algorithm based on the work by Liu and Rubin (1995), the implementation of the function $fitt_{f}ixnu.m$ was done by Robin Ince and can be found at https://github.com/robince/tdistfit and http://www.robinince.net.

Listing C.4: *Matlab code MLECauchy.m*

```
% Cauchy Maximum Likelihood Estimators - 29 January 2018
%
% This code determines the maximum likelihood parameter estimates for a bivariate
% Cauchy fit to Lightning Location System (LLS) reported stroke locations.
% It is based on the paper ``ML estimation of the t distribution using EM and its
% extensions, ECM and ”ECME by Liu and Rubin (1995). The implementation of the function
% fitt_fixnu.m was done by Robin Ince and can be found at https://github.com/robince/tdistfit
% and http://www.robinince.net/blog/2013/04/18/fitting-multivariate-t-distributions-to-data/.

LLSdata = importdata‘(LLSdata.csv', ',');                 % load LLS data set

lat = LLSdata.data(:,2);                                  % Select locations
long = LLSdata.data(:,3);

mydata = [long lat];

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function CauchyParams = MLECauchy(mydata)

 nu_c = 1;
 [mu_t S_t nu_t] = fix_nu(mydata,nu_c);
 logc = mvtLogpdf(mydata,mu_c,S_c,nu_c);
 aNLLcdist = -sum(logc);
 AICc =  2*6 + 2*aNLLcdist;
 BICc = log(length(mydata))*6 + 2*aNLLcdist;

 CauchyParams = [mu_c,S_c,nu_c,aNLLcdist,AICc,BICc];

end
```

C.2 Mahalanobis Test - Source Code

Figure 6.3 in Chapter 6, Section 6.5 shows a block diagram representation of the Mahalanobis test algorithm based on the work by McAssey (2013). The code that calculates the Mahalanobis test statistic for each of the statistical models is as follows:

Listing C.5: *Matlab code MahalanobisTest.m*

```
% Mahalanobis Test - 29 January 2018
%
% This code performs a goodness of fit test on a bivariate data set of
% Lightning Location System (LLS) stroke location
```

```
% reports. It is based on the paper ``An empirical goodness-of-fit test
% for multivariate distributions??????
% by M. P. McAssey (2013).
%



LLSdata = importdata(LLSdata.csv', ','); % load LLS data set

lat = LLSdata.data(:,2);                                  % Select locations
long = LLSdata.data(:,3);

mydata = [long lat];

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Initial maximum likelihood fit - Step 1.
% Determine theta_mle for each statistical model.

[GausFitmu,GausFitSigma,aNLLGauss,AICg,BICg] = MLEGaussian(mydata); % Fit Gaussian

[unif(1,:),unif(2,:)] = MLEuniform(mydata);                        % Fit uniform

[mu_t,S_t,nu_t,aNLLtdist,AICt,BICt] = MLEStudentst(mydata);        % Fit Students??? t-

[mu_c,S_c,nu_c,aNLLcdist,AICc,BICc] = MLECauchy(mydata)            % Fit Cauchy


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Step - 2a.

Mhlnbismy = mahalanobis(mydata, cov(mydata), mean(mydata));% Calculate Mahalanobis distance
                   % of LLS stroke reports

BinWidth = 0.02;
BinLimits = [0 20];

 myd = histogram(Mhlnbismy,'BinWidth',BinWidth,'BinLimits',BinLimits);
 myd.Normalization = 'CDF';
 binedges = myd.BinEdges;
 binwidth = myd.BinWidth;
 binnums = myd.NumBins;
 bincountsmydata = myd.BinCounts;

 Obs_cnts = cumsum(bincountsmydata);                     % Det. Obs_n binning LLS
 Obs_prb = Obs_cnts./length(Mhlnbismy);                  % Mahalanobis distance


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Step - 2.b

for R = 1 : 1 : 100;

% Generate random data - X_smpl and Mahalanobis_smpl for each statistical model


mvnrnddata = mvnrnd(GausFitmu,GausFitSigma,10000);         % Gaussian
Mhlnbismvnrnd = mahalanobis(mvnrnddata, cov(mvnrnddata), mean(mvnrnddata));


mvtrnddata = diag(sqrt(diag(S_t))) * mvtrnd(S_t,nu_t,10000)'; % Students??? t-
mvtrnddata = mu_t + mvtrnddata';
Mhlnbismvtrnd = mahalanobis(mvtrnddata, cov(mvtrnddata), mean(mvtrnddata));

mvcrnddata = diag(sqrt(diag(S_c))) * mvtrnd(S_c,nu_c,10000)'; % Cauchy
mvcrnddata = mu_c + mvcrnddata';
```

```
Mhlnbismvcrnd = mahalanobis(mvcrnddata, cov(mvcrnddata), mean(mvcrnddata));

mu_u = [(unif(1,2)+unif(1,1))/2 (unif(2,2)+unif(2,1))/2];   % uniform
Difs_u = [(unif(1,2)-unif(1,1)) 0; 0 (unif(2,2)-unif(2,1))];
S_u = [((unif(1,2)-unif(1,1))^2)/12 0; 0 ((unif(2,2)-unif(2,1))^2)/12];

mvurnddata = mu_u + (Difs_u * (-0.5 + rand(10000,2))')';
Mhlnbismvurnd = mahalanobis(mvurnddata, cov(mvurnddata), mean(mvurnddata));


% Bin random generated data to determine G_k(b_n)


maxs = [max(Mhlnbismy); max(Mhlnbismvnrnd); max(Mhlnbismvtrnd); max(Mhlnbismvcrnd); max(Mhlnbismvurnd)];
maxMhlnbis = round(max(maxs));


  mvn = histogram(Mhlnbismvnrnd,'BinEdges',binedges);         % Gaussian
  mvn.Normalization = 'CDF';
  bincountsn = mvn.BinCounts;

    GNtaccentg = cumsum(bincountsn)./length(Mhlnbismvnrnd);
    GNtaccentavg(:,R) = GNtaccentg';

 mvt = histogram(Mhlnbismvtrnd,'BinEdges',binedges);          % Students??? t-
 mvt.Normalization = 'CDF';
 bincountst = mvt.BinCounts;

    GNtaccentt = cumsum(bincountst)./length(Mhlnbismvtrnd);
    GNtaccentavt(:,R) = GNtaccentt';

 mvc = histogram(Mhlnbismvcrnd,'BinEdges',binedges);          % Cauchy
 mvc.Normalization = 'CDF';
 bincountsc = mvc.BinCounts;

    GNtaccentc = cumsum(bincountsc)./length(Mhlnbismvcrnd);
    GNtaccentavc(:,R) = GNtaccentc';

 mvu = histogram(Mhlnbismvurnd,'BinEdges',binedges);         % uniform
 mvu.Normalization = 'CDF';
 bincountsu = mvu.BinCounts;

    GNtaccentu = cumsum(bincountsu)./length(Mhlnbismvurnd);
    GNtaccentavu(:,R) = GNtaccentu';


end

% Expected count average over R = 100 repetitions

Exp_cntsg = mean(GNtaccentavg').*length(Mhlnbismy);
Exp_cntst = mean(GNtaccentavt').*length(Mhlnbismy);
Exp_cntsc = mean(GNtaccentavc').*length(Mhlnbismy);
Exp_cntsu = mean(GNtaccentavu').*length(Mhlnbismy);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Step - 3. Mahalanobis test statistic for each fit

Mhlstatn = sum((abs(Exp_cntsg-Obs_cnts))./Exp_cntsg);
Mhlstatt = sum((abs(Exp_cntst-Obs_cnts))./Exp_cntst);
Mhlstatc = sum((abs(Exp_cntsc-Obs_cnts))./Exp_cntsc);
Mhlstatu = sum((abs(Exp_cntsu-Obs_cnts))./Exp_cntsu);
```

C.3 Bayesian Framework - Source Code

The following source code implements the proposed Bayesian framework presented in Chapter 5 and utilized in Chapter 8.

Listing C.6: *Matlab code BayesCode.m*

```
% Bayesian Framework - 29 January 2018
%
% This code performs the calculations used in the Bayesian
% framework.


LLSdata = importdata‘(LLSdata.csv', ',');          % load LLS data set

lat = LLSdata.data(:,2);                                   % Select locations
long = LLSdata.data(:,3);

Xlls = [long lat];

        DateString = LLSdata.textdata;                        % Load time-stamps of strokes
        formatIn = 'yyyy-mm-dd HH:MM:SS';
        DateVector = datevec(DateString,formatIn);
        year = DateVector(:,1);
        month = DateVector(:,2);
        day = DateVector(:,3);
        hour = DateVector(:,4);
        minute = DateVector(:,5);
        seconds = DateVector(:,6);

        nseconds = LLSdata.data(:,1);
        secondwnano = seconds + (nseconds./1000000000);

for iTo = 1 : 1 : length(year)
    t(iTo) = etime([year(iTo) month(iTo) day(iTo) hour(iTo) minute(iTo) secondwnano(iTo)↩
         ],[1970 1 1 0 0 0]);
end

 stringtime = datestr(719529+t/86400,'yyyy-mm-dd HH:MM:SS.FFF');

 TickDates = etime([2015 10 1 00 00 00; 2015 11 1 00 00 00; 2015 12 1 00 00 00; 2016 1 1 00 ↩
      00 00; 2016 2 1 00 00 00; 2016 3 1 00 00 00],[1970 1 1 0 0 0]);
 tickdate = datestr(719529+TickDates/86400,'yyyy-mm-dd');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Define location of events and investigation


    L_lon_br = 28.0068;   % Latitude and Longitude of the Brixton tower
    L_lat_br = -26.1925;


    L_event = [L_lon_br L_lat_br];    % This can be any location of interest

    maxLon =  L_lon_br + 0.1;   % Define area of investigation such that
    minLon =  L_lon_br - 0.1;   % the location of the event falls within
    maxLat =  L_lat_br + 0.1;   % the area.
    minLat =  L_lat_br - 0.1;


    n_lon = 275; n_lat = 275;         % Define grid over area of investigation.
    lonstep = (maxLon-minLon)/n_lon;
    latstep = (maxLat-minLat)/n_lat;
    gX = minLon+(lonstep/2):lonstep:maxLon-(lonstep/2);
```

```
    gY = minLat+(latstep/2):latstep:maxLat-(latstep/2);
    [L_lon,L_lat] = meshgrid(gX,gY);
    Grid_zeros = zeros(n_lon,n_lat);
    Grid_ones = ones(n_lat,n_lon);

    L = [L_lon(:) L_lat(:)];         % Define L, set of all possible locations within
               % the area of investigation.


    i_event = index(L, L_event, lonstep, latstep);


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


% Prior Calculation:


% Non-informative or flat prior - Uniform distribution

    LongUnfrmVal = 1/(maxLon - minLon);
    LatUnfrmVal = 1/(maxLat - minLat);
    UnfrmVal = LatUnfrmVal*LongUnfrmVal;
    UnfrmVal_PrVal = UnfrmVal*lonstep*latstep;

 f_L_grid = UnfrmVal.*Grid_ones;    % Uniform distribution
 f_L = f_L_grid(:);                 % prior probability density function


               % OR %

% Informative prior - GMM

% f_L_GMM_grid = importdata‘(GMM.csv', ',');
% f_L_GMM = f_L_GMM_grid(:);
% f_L = f_L_GMM;                    % prior probability density function


 p_L = f_L.*(lonstep*latstep);      % prior probability mass function




%  Now the measurement is made:



for j = 1 : 1 : length(LLSdata);

% Gaussian Likelihood Calculation:

  f_Xlls_L_event(j) = twodgauss(Xlls(j,1), Xlls(j,2), L_event(1)-0.0002, L_event(2)+0.001, ↩
      0.0051, 0.0045, 0.5395);

  % Parameters of Gaussian function from Chapter 7.

 % f_Xlls_L_event(j) = twod_t_dist(Xlls(j,1), Xlls(j,2), L_event(1)+0.0001, L_event(2)↩
      +0.0003, 0.00060742, 0.00056824, 0.0994, 0.8318);

  % Parameters of ’Students t- function from Chapter 7.

% Evidence Calculation:

     p_Xlls_temp = p_L.*twodgauss(Xlls(j,1), Xlls(j,2), L(:,1)-0.0002, L(:,2)+0.001, 0.0051, ↩
          0.0045, 0.5395);

     %p_Xlls_temp = p_L.*twod_t_dist(Xlls(j,1), Xlls(j,2), L(:,1)+0.0001, L(:,2)+0.0003, ↩
          0.00060742, 0.00056824, 0.0994, 0.8318);
```

```

    p_Xlls(j) = sum(p_Xlls_temp);

     % Integral solution for denominator in continuous case

% Posterior Calculation:

     p_L_Xlls_event(j) = (p_L(i_event).*f_Xlls_L_event(j))./p_Xlls(j);


% NOT likelihood function:

    f_Xlls_L_event_ns(j) = (p_Xlls(j) - p_L(i_event).*f_Xlls_L_event(j))./(1 - p_L(i_event));
    p_L_Xlls_event_ns(j) = (f_Xlls_L_event_ns(j).*(1 - p_L(i_event)))./p_Xlls(j);


end



   % log Likelihood ratios (no prior ie. not posterior odds)

  % Event Location vs NOT Event Location (ie. Brixton struck vs Brixton not struck.)

    LR_event_event_ns = f_Xlls_L_event./f_Xlls_L_event_ns;
    lg_LR_event_event_ns = log(LR_Br_Br_ns_To);


  % Plots

   Postplot = plot(t, p_L_Xlls_event, 'v','MarkerFaceColor',[0 0.5 0],'MarkerEdgeColor','k','MarkerSize',10);
   hold on;

 %  lnLRplot = plot(t, lg_LR_event_event_ns, 'v','MarkerFaceColor',[0 0.5 0],'MarkerEdgeColor','k','MarkerSize',10);
 %  hold on;
```

The following source code generates the Gaussian Mixture Model (*GMM.csv*) for use as the informative prior in the Bayesian framework source code above.

Listing C.7: *Matlab code GMM.m*

```
% Baussian Mixture Model - 29 January 2018
%
% This code performs the calculates the Gaussian Mixture Model (GMM)
% and outputs it as GMM.csv for use in BayesCode.m


LLSdata = importdata‘(LLSdata.csv', ',');          % load LLS data set

 lat = LLSdata.data(:,2);                                  % Select locations
 long = LLSdata.data(:,3);

 XllsPr = [long lat];

    maxLon = max(long);
    minLon = min(long);
    maxLat = max(lat);
    minLat = min(lat);

 nlat = 275; nlon = 275;                              % Define grid over area of investigation.
 latstep = (maxLat-minLat)/nlat;
 lonstep = (maxLon-minLon)/nlon;
```

```
gX = minLon+(lonstep/2):lonstep:maxLon-(lonstep/2);
gY = minLat+(latstep/2):latstep:maxLat-(latstep/2);
[X_e,Y_e] = meshgrid(gX,gY);
Gaus_grid_sum = zeros(nlat,nlon);
Grid_sum = zeros(nlat,nlon);
Gaus_grid_neg_prod = ones(nlat,nlon);

for q = 1 : 1 : size(XllsPr)

    g_q = twodgauss(X_e, Y_e, XllsPr(q,2), XllsPr(q,1), 0.0051, 0.0045, 0.5395);

 % Parameters of Gaussian function from Chapter 7.

    Gaus_grid = (g_q./size(XllsPr));
    Gaus_grid_sum = Gaus_grid_sum + Gaus_grid;

end;

dlmwrite‘(GMM.csv', Gaus_grid_sum);
```

References

Aeschliman, C., Park, J. and Kak, A. (2010). A novel parameter estimation algorithm for the multivariate t-distribution and its application to computer vision, *European Conference on Computer Vision* .

Aitken, C., Roberts, P. and Jackson, G. (2010). *Fundamentals of probability and statistical evidence in criminal proceedings: guidance for judges, lawyers, forensic scientists and expert witnesses*, London: Royal Statistical Society. Available at www.rss.org.uk/statsandlaw.

Allen, R. J. (2008). Exact solutions to Bayesian and maximum likelihood problems in facial identification when population and error distributions are known, *Forensic Science International* **179**(2): 211 – 218.

Anderson, R. B. (2001). Does a fifth mechanism exist to explain lightning injuries?, *IEEE Engineering in Medicine and Biology* .

Baumgartner, K. A. C. (2005). *Bayesian network modeling of offender behavior for criminal profiling*, Master's thesis, Department of Mechanical Engineering and Materials Science in the Graduate School of Duke University.

Bernardo, J. (1979). Reference posterior distributions for Bayesian inference, *J. R. Statist Soc.* **41**(2): 113–147.

Biedermann, A., Bozza, S. and Taroni, F. (2011). Probabilistic evidential assessment of gunshot residue particle evidence (part ii): Bayesian parameter estimation for experimental count data, *Forensic Science International* **206**: 103–110.

Biedermann, A. and Taroni, F. (2006). A probabilistic approach to the joint evaluation of firearm evidence and gunshot residues, *Forensic Science International* **163**(1): 18 – 33.

Biedermann, A., Taroni, F., Delemont, O., Semadeni, C. and Davison, A. (2005). The evaluation of evidence in the forensic investigation of fire incidents. part ii. practical examples of the use of bayesian networks, *Forensic Science International* **147**(1): 59 – 69.

Biedermann, A., Taroni, F. and Garbolino, P. (2007). Equal prior probabilities: Can one do any better?, *Forensic Science International* **172**(2): 85 – 93.

Blair, J. P. and Rossmo, D. K. (2010). Evidence in context: Bayes' theorem and investigations, *Police Quarterly* **13**(2): 123–135.

Blumenthal, R., Jandrell, I. R., Anderson, R. B. and Trengrove, E. (2009). Recent lightning research in South Africa with a special focus on keraunopathology, *16th International Symposium on High Voltage* **K-1**: 1–6.

Bouquegneau, C. (2014). The need for an international standard on Lightning Location Systems, *23rd International Lightning Detection Conference* .

Bouquegneau, C., Lecomte, P., Coquelet, F., Poelman, D. and Crabbé, M. (2013). Lightning flash and strike-point density in Belgium, *8th Asia-Pacific International Conference on Lightning* .

Bourscheidt, V., Cummins, K. L., Pinto, O., Jr. and Naccarato, K. P. (2012). Methods to overcome Lightning Location System performance limitations on spatial and temporal analysis: Brazilian case, *Journal of Atmospheric and Oceanic Technology* **29**.

Bourscheidt, V., Pinto, O., Jr. and Naccarato, K. P. (2013). Improvements on lightning density estimation based on analysis of Lightning Location System performance parameters: Brazilian case, *IEEE Transactions on Geoscience and Remote Sensing* .

Campos, L. Z., Cummins, K. L. and Jr., O. P. (2015). An algorithm for identifying ground strike points from return stroke data provided by Lightning Location Systems, *2015 Asia-Pacific International Conference on Lightning (APL), Nagoya, Japan* .

Carte, A. E., Anderson, R. B. and Cooper, M. A. (2002). A large group of children struck by lightning, *Annals of Emergency Medicine* .

Congdon, P. (2003). *Applied Bayesian Modelling*, John Wiley & Sons, Ltd.

Cooray, V. (2014). *The lightning flash: 2nd edition*, The Institution of Engineering and Technology, Michael Faraday House, Six Hills Way, Stevenage, Herts, SG1 2AY, United Kingdom.

Cummins, K. L. (2014). Mapping the impact of terrain on lightning incidence and multiple ground contacts in cloud-to-ground flashes., *International Conference on Atmospheric Electricity* **15**.

Cummins, K. and Murphy, M. (1998). An overview of Lightning Locating Systems: History, techniques and data uses with an in-depth look at the U.S. NLDN, *IEEE Transactions on Electromagnetic Compatability* **40**.

David, A. P. and Stone, M. (1973). Expectation consistency and generalized Bayes inference, *Annals of Statistics* **1**(3): 478–485.

Dawid, A. (2001). *Bayes's Theorem and Weighing Evidence by Juries.In: Bayes's Theorem.*, British Academy.

Der Gesamtverband der Deutschen Versicherungswirtschaft (2017). Last Accessed: 12 September 2017.
URL: *http://www.xprimm.com/GERMANY-Less-claims-paid-for-damages-caused-by-lightning-in-2016-but-more-expensive-costs-per-claim-articol-117,124-9882.htm*

Dhar, S., Chakraborty, B. and Chaudhuri, P. (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests, *Bernoulli* **20**(3): 1484–1506.

Diendorfer, G. (2007). Lightning Location Systems (LLS), *IX International Symposium on Lightning Protection* pp. 1–13.

Diendorfer, G. (2010). LLS performance validation using lightning to towers, *2010 International Lightning Detection Conference (ILDC)*.

Diendorfer, G. (2016). A review of 25 years of lightning research in Austria from 1991-2015, *World Meeting on Lightning*.

Diendorfer, G., Pichler, H. and Mair, M. (2009). Some parameters of negative upward-initiated lightning to the Gaisberg tower (2000–2007), *IEEE Transactions on Electromagnetic Compatibility* **51**(3): 1–12.

Diendorfer, G., Pichler, H. and Schulz, W. (2014). EUCLID located strokes to the Gaisberg tower – accuracy of location and its assigned confidence ellipse, *2014 International Lightning Detection Conference (ILDC)*.

Diendorfer, G., Pichler, H. and Schulz, W. (2015). LLS detection of upward initiated lightning flashes, *Asia-Pacific International Conference on Lightning (APL)*.

Diendorfer, G., Zhou, H. and Pichler, H. (2011a). A review of 10 years of lightning measurement at Gaisberg tower in Austria, *3rd International Symposium on Winter Lightning (ISWL2011)*.

Diendorfer, G., Zhou, H., Pichler, H. and Thottappillil, R. (2011b). Review of upward positive and bipolar lightning flashes at the Gaisberg tower, *Asia-Pacific International Conference on Lightning (APL)*.

Fawcett, T. (2005). An introduction to roc analysis, *Pattern Recognition Letters* **27**: 861–874.

Fenton, N., Berger, D., Lagnado, D., Neil, M. and Hsu, A. (2014a). When "neutral" evidence still has probative value (with implications from the barry george case), *Science and Justice* **54**(4): 274–287.

Fenton, N. and Neil, M. (2011). The use of Bayes and causal modelling in decision making, uncertainty and risk, *Cepis Upgrade* **12**(5): 10–21.

Fenton, N., Neil, M. and Hsu, A. (2014b). Calculating and understanding the value of any type of match evidence when there are potential testing errors, *Artificial Intelligence and Law* **22**: 1–28.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*, CRC Press.

Gijben, M. (2012). The lightning climatology of South Africa, *South African Journal of Science* .

Gill, T. (2008). A lightning climatology of South Africa for the first two years of operation of the South African Weather Service Lightning Detection Network: 2006-2007, *20th International Lightning Detection Conference* pp. 1–12.

Gittelson, S. N. (2013). *Evolving from Inferences to Decisions in the Interpretation of Scientific Evidence*, PhD thesis, Universite de Lausanne Faculte de Droit et des Sciences Criminelles.

Grant, M. D., Tordiffe, A. S. W. and Blumenthal, R. (2012). Case study of lightning-caused death of critically endangered tragelaphus sp., *2012 International Conference on Lightning Protection (ICLP).*

Grant, M., McKechnie, I., Jandrell, I. and Nixon, K. (2011). Probablistic interpretation of LDN confidence ellipses with reference to forensic applications, *7th Asia-Pacific International Conference on Lightning* .

Hansson, S. O. (1994). *Decision Theory: A Brief Introduction*, KTH Stockholm.

Hepler, A. B. (2005). *Improving Forensic Identification Using Bayesian Networks and Relatedness Estimation: Allowing for Population Substructure*, PhD thesis, faculty of North Carolina state university.

Huddleston, L. L., Roeder, W. P. and Merceret, F. J. (2010). A method to estimate the probability that any individual lightning stroke contacted the surface within any radius of any point, *2010 International Lightning Detection Conference (ILDC).*

Huddleston, L. L., Roeder, W. P. and Merceret, F. J. (2012). A probabilistic, facility-centric approach to lightning strike location, *in NASA Technical Reports Server (NTRS), BiblioGov* .

Hunt, H., Liu, Y. C. and Nixon, K. (2014). Evaluation of the South African Lightning Detection Network using photographed tall tower lightning events from 2009-2013, *32rd International Conference on Lightning Protection (ICLP), Shanghai, China)* .

Hunt, H., Nixon, K. and Naude, J. (2017). Using Lightning Location System stroke reports to evaluate the probability that an area of interest was struck by lightning, *Electric Power Systems Research Journal* .

IEC 62858 (2015). *IEC 62858:2015 Edition 1: Lightning density based on Lightning Location Systems (LLS) – General principles.*

Insurance Information Institute (2017). Last Accessed: 12 September 2017.
URL: *http://www.iii.org/fact-statistic/lightning*

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*, Cambridge University Press, New York.

Jerauld, J., Rakov, V. A., Uman, M. A., Rambo, K. J., Jordan, D. M., Cummins, K. L. and Cramer, J. A. (2005). An evaluation of the performance characteristics of the U.S. National Lightning Detection Network in Florida using rocket-triggered lightning, *Journal of Geophysical Research: Atmospheres* **110**(D19).

Justel, A., Peña, D. and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit, *Statistics & Probability Letters* **35**(3): 251 – 259.

Kaye, D. H. (2012). Likelihoodism, bayesianism, and a pair of shoes, *Jurimetrics* **53**(1).

Kotz, S. and Nadarajah, S. (2004). *Data Analysis in Forensic Science: A Bayesian Decision Perspective*, Cambridge University Press.

Kruse, C. (2013). The Bayesian approach to forensic evidence: Evaluating, communicating, and distributing responsibility, *Social Studies of Science* **43**(5): 657–680.

Lange, K., Little, R. and Taylor, J. (1989). Robust statistical modeling using the t-distribution, *Journal of the American Statistical Association* **84**(408): 881–896.

Liu, C. and Rubin, D. (1995). ML estimation of the t-distribution using EM and its extensions, ECM and ECME, *Statistica Sinica* **5**: 19–39.

Martire, K. A., Kemp, R. I., Sayle, M. and Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect, *Forensic Science International* **240**: 61–68.

McAssey, M. P. (2013). An empirical goodness-of-fit test for multivariate distributions, *Journal of Applied Statistics* **40**(5): 1120–1131.

Meng, X.-L. and van Dyk, D. (1997). The EM algorithm - an old folk-song sung to a fast new tune, *Journal of the Royal Statistical Society. Series B (Methodological)* **59**(3): 511–567.

Montgomery, D. C. and Runger, G. C. (2014). *Applied Statistics and Probability for Engineers*, John Wiley & Sons, Inc., New York, 6th Ed.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*, Massachusetts Institute of Technology.

Naccarato, K. and Pinto, O. (2009). Improvements in the detection efficiency model for the Brazilian Lightning Detection Network (BrasilDAT), *Atmospheric Research* **91**(2): 546 – 563.

Nag, A., Mallick, S., Rakov, V. A., Howard, J. S., Biagi, C. J., Hill, J. D., Uman, M. A., Jordan, D. M., Rambo, K. J., Jerauld, J. E., DeCarlo, B. A., Cummins, K. L. and Cramer, J. A. (2011). Evaluation of U.S. National Lightning Detection Network performance characteristics using rocket-triggered lightning data acquired in 2004–2009, *Journal of Geophysical Research: Atmospheres* **116**(D2).

Nag, A., Murphy, M., Schulz, W. and Cummins, K. (2015). Lightning Locating Systems: Insights on characteristics and validation techniques, *Earth and Space Science* **2**(4): 65–93.

Noon, R. (2001). *Forensic Engineering Investigation*, CRC Press LLC.

Nucci, C. (2007). Inferring the correlation between lightning events and voltage dips in distribution networks, *International Symposium on Lightning protection (SIPDA).*

Overill, R. E. (2014). Quantifying likelihood in digital forensic investigations, *Journal of Harbin Institute of Technology* **21**(6).

Palombo, G. (2011). Multivariate goodness of fit procedures for unbinned data: An annotated bibliography, *ArXiv e-prints* .

Pédeboy, S. and Schulz, W. (2014). Validation of a ground strike point identification algorithm based on ground truth data, *23rd International Lightning Detection Conference* .

Poelman, D., Schulz, W., R.Kaltenboeck and Delobbe, L. (2017). Analysis of lightning outliers in the EUCLID network, *Atmospheric Measurement Techniques* **10**.

Rakov, V. A. and Uman, M. A. (2003). *Lightning Physics and Effects*, Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, England.

Roberts, P. and Aitken, C. (2014). *The Logic of Forensic Proof: Inferential Reasoning in Criminal Evidence and Forensic Science.RSS Practitioner Guide No 3.*, London: Royal Statistical Society. Available at www.rss.org.uk/statsandlaw.

Roth, M. (2013). On the multivariate t-distribution, *Technical report*, Department of Electrical Engineering Linkopings universitet.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*, Chapman and Hall/CRC. ISBN-13: 978-0412044113.

Schulz, W. (1997). Performance evaluation of Lightning Location Systems, *Faculty of Electrical Engineering, Technical University of Vienna* .

Schulz, W. (2015). Location accuracy improvements of the Austrian Lightning Location System during the last 10 years, *Asia-Pacific International Conference on Lightning (APL)*.

Schulz, W. and Cummins, K. (2008). A method to determine relative stroke detection efficiencies from multiplicity distributions, *2008 International Lightning Detection Conference (ILDC)*.

Schulz, W., Diendorfer, G., Pedeboy, S. and Poelman, D. R. (2016). The European Lightning Location System EUCLID part 1: Performance analysis and validation, *Natural Hazards and Earth System Sciences* **16**(2): 595–605.

Specter, M. M. (2011). Forensic engineering electrical engineering edition, *NAFE Journal Papers* **2**.

Stall, C. A., Cummins, K. L., Krider, E. P. and Cramer, J. A. (2009). Detecting multiple ground contacts in cloud-to-ground lightning flashes, *Journal Atmospheric of Oceanic Technology* **26**.

Stansfield, R. G. (1947). Statistical theory of DF fixing, *Proceedings of Institute of Electrical Engineers* .

STRIKEnet® (2017). Last Accessed: 13 September 2017.
URL: *http://www.corelogic.com/products/lightning-verification.aspx*

Taroni, F., Biedermann, A., Garbolino, P. and Aitken, C. (2004). A general approach to Bayesian networks for the interpretation of evidence, *Forensic Science International* **139**(1): 5 – 16.

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P. and Aitken, C. (2014). *Data Analysis in Forensic Science: A Bayesian Decision Perspective*, John Wiley and Sons, Ltd.

Thompson, W., Vuille, J., Biedermann, A. and Taroni, F. (2013). The role of prior probability in forensic assessments, *Frontiers in Genetics* **4**: 220.

Tsitsiklis, J. (2010). 6.041 probabilistic systems analysis and applied probability. License: Creative Commons BY-NC-SA.

Turpin, S. M. and Marais, M. A. (2004). Decision-making: Theory and practice, *ORiON* **20**(2): 143–160.

Uman, M. A. (1987). *The lightning discharge*, Academic Press.

Zoete, J., Sjerps, M., Lagnado, D. and Fenton, N. (2015). Modelling crime linkage with Bayesian networks, *Science & Justice* **55**(3): 209 – 217.

Bibliography

Aitkin, M., Boys, R. J. and Chadwick, T. (2005). Bayesian point null hypothesis testing via the posterior likelihood ratio, *Statistics and Computing* **15**(3): 217–230.

Baran, I. and Costea, M. (2014). Fault analysis and Lightning Location Systems, **17**: 489 – 496.

Becerra, M. (2008). *On the Attachment of Lightning Flashes to Grounded Structures*, PhD thesis, Uppsala University - Faculty of Science and Technology.

Berger, G. (2012). Recent investigations on lightning accidents: The real ones ... and some other, *2012 International Conference on Grounding and Earthing and 5th International Conference on Lightning Physics and Effects*.

Blom, P. S., Marcillo, O. and Arrowsmith, S. J. (2015). Improved Bayesian infrasonic source localization for regional infrasound, *Geophysical Journal International* **203**(3): 1682.

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence, *Statistics in Medicine* **21**: 2563–2599.

Borghetti, A., Nucci, C. A., P., M. and Bernardi, M. (2006). A statistical approach for estimating the correlation between lightning and faults in power distribution systems, *9th International Conference on Probabilistic Methods Applied to Power Systems*.

Borghetti, A., Nucci, C. A. and Paolone, M. (2003). Effect of tall instrumented towers on the statistical distributions of lightning current parameters and its influence on the power system lightning performance assessment, *European Transactions on Electrical Power* **13**(6): 365–372.

Bourscheidt, V., Cummins, K. L., Jr., O. P. and Naccarato, K. P. (2012). Methods to overcome Lightning Location System performance limitations on spatial and temporal analysis: Brazilian case, *Journal of Atmospheric and Oceanic Technology* **29**(9): 1304–1311.

Bourscheidt, V., Jr., O. P. and Naccarato, K. P. (2010a). The effect of different configurations of the Brazilian Lightning Detection Network on the cloud-to-ground lightning spatial distribution, *2010 International Lightning Detection Conference (ILDC)*.

Bourscheidt, V., Jr., O. P. and Naccarato, K. P. (2010b). Using the Gaussian-based confidence ellipse provided by Lightning Location Systems to create more reliable lightning density maps, *2010 International Lightning Detection Conference (ILDC)*.

Carreira-Perpinan, M. A. (2000). Mode-finding for mixtures of Gaussian distributions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(11): 1318–1323.

Chen, X. and Goodwin, B. K. (2013). Spatio-Temporal Modeling of Lightning Fires on Forestland: A Compensation Scheme, *Technical Report 151285*.

Chronis, T., Cummins, K., Said, R., Koshak, W., McCaul, E., Williams, E. R., Stano, G. T. and Grant, M. (2015). Climatological diurnal variation of negative CG lightning peak current over the continental United States, *Journal of Geophysical Research: Atmospheres* **120**(2): 582–589.

Cramer, J. A. and Cummins, K. L. (2014). Evaluating location accuracy of Lightning Location Networks using tall towers, *2014 International Lightning Detection Conference (ILDC)*.

Cummins, K. L., Cramer, J. A., Biagi, C. J., Krider, E. P., Jerauld, J., Uman, M. A. and Rakov, V. A. (2006). The U.S. National Lightning Detection Network: Post-upgrade status, *2nd Conference on Meteorological Applications of Lightning Data*.

Cummins, K. L., Murphy, M., Cramer, J. A., Scheftic, W., Demetriades, N. and Nag, A. (2010). Location accuracy improvements using propagation corrections: A case study of the U.S. National Lightning Detection Network, *2010 International Lightning Detection Conference (ILDC)*.

Cummins, K. L., Murphy, M. J. and Tuel, J. V. (2000). Lightning detection methods and meteorological applications, *IV International Symposium on Military Meteorology Hydro-Meteorological Support of Allied Forces and PfP Members Tasks Realization*.

den Hout, A. V. and Alberink, I. (2016). Posterior distributions for likelihood ratios in forensic science, *Science & Justice* **56**(5): 397 – 401.

Dickson, A. S. (2014). Action volume ratio - a method to classify the danger of lightning in any given volume, *Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg* .

Djurica, V., Kosma, J. and Milev, G. (2008). A multiple power line corridor and lightning error-ellipse spatial processor for real-time correlator, *2008 International Lightning Detection Conference (ILDC)*.

Djurica, V. and Kosmac, J. (2006). LLS accuracy improvements by measurements colleted by RLDN, *2006 International Lightning Detection Conference (ILDC)*.

Djurica, V., Mandeljc, R. and Milev, G. (2010). A new analysis method could locate a faulty or noisy directional sensor in a Lightning Location Network, *2010 International Lightning Detection Conference (ILDC)*.

Elhor, A. and Pensky, M. (2000). Bayesian estimation of locations of lightning events, *Sankhya : The Indian Journal of Statistics* **62**: 202–216.

Erntell, F. (2013). *On modeling insurance claims using copulas*, Master's thesis, Lund Universtity.

Evert, R. and Schulze, G. (2005). Impact of a new Lightning Detection and Location System in South Africa, *Inaugural IEEE PES 2005 Conference and Exposition in Africa* pp. 356–363.

Flinn, F. C., Roeder, W. P., Buchanan, M. D., McNamara, T. M., McAleenan, M., Winters, K. A., Fitzpatrick, M. E. and Huddleston, L. L. (2010). Lightning reporting at 45th weather squadron: Recent improvements, *2010 International Lightning Detection Conference (ILDC)*.

Forbes, F. and Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering, *Statistics and Computing* **24**(6): 971–984.

Grant, M. D., Nixon, K. J. and Jandrell, I. R. (2012). Positive polarity: Misclassification between intracloud and cloud-to-ground discharges in the southern african lightning detection network, *2012 International Lightning Detection Conference (ILDC)*.

Hall, P. and York, M. (2001). On the calibration of Silverman's test for multimodality, *Statistica Sinica* **11**: 515–536.

Holle, R. L. (2008). Annual rates of lightning fatalities by country, *20th International Lightning Detection Conference (ILDC)* .

Holle, R. L. (2015). Some aspects of global lightning impacts, *2015 AMS Annual Meeting* .

Holle, R. L. and Lopez, R. E. (1996). Insured lightning-caused property damage in three western states, *Journal of Apllied Meteorology* .

Jayakumar, V. (2004). *Estimating power, energy, and action integral in rocket-triggered lightning*, Master's thesis, University of Florida.

Kappennman, J. (1999). Analysis of the French Lightning Locating System location accuracy: Impact of the recent technological upgrades on performances, *IEEE T&D Conference*.

Kern, A. and Dikta, G. (2008). Probability of damage of electrical and electronic systems due to indirect lightning flashes – investigation of data from German insurance companies, *2008 International Conference on Lightning Protection (ICLP)*.

Krider, E. P. and Noggle, R. C. (1975). Broadband antenna systems for lightning magnetic fields, *Journal of Applied Meteorology* **14**(2): 252–256.

Kunreuther, H. and Pauly, M. (2005). Insurance decision-making and market behavior, *Foundations and Trends in Microeconomics* **1**(2): 63–127.

Lafkovici, A., Hussein, A. M., Janischewskyj, W. and Cummins, K. L. (2006). Performance analysis of the North American Lightning Detection Network using CN tower lightning data, *2006 International Lightning Detection Conference (ILDC)*.

Lakota, G., Milev, G., Djurica, V. and Žitnik, B. (2014). Use of high resolution flash density data and GIS analysis methods in power utilities networks, *2014 International Conference on Lightning Protection (ICLP)*, pp. 1229–1233.

Li, X. and Hunter, D. K. (2008). Probabilistic model of triangulation, pp. 301–306.

Lopez, J., Murphy, M., Aranda, J., Gaztelumendi, S., Maruri, M. and de la Vega, D. (2010). Analysis of the geographical configuration of a Lightning Detection Network and its influence in the accuracy of lightning location, *2010 International Lightning Detection Conference (ILDC)*.

Luan, J., Wang, X., Wang, P. and Wang, Y. N. D. (2013). Error correction of lightning location data, *Meteorological and Environmental Research*, Vol. 4, pp. 1–5.

Mallick, M. (2002). Maximum likelihood geolocation using a ground moving target indicator (gmti) report, **4**: 1561–1570.

Marcek, M. (2005). An application of Bayesian approach for lightning current arrester sales modelling and forecasting, *Journal of Information, Control and Management Systems* **3**(2): 145–150.

Meyers, G. (2005). On predictive modeling for claim severity, *Casualty Actuarial Society Forum* pp. 215–253.

Michishita, K., Hongo, Y. and Yokoyama, S. (2009). Estimation of lightning striking points by time-of-arrival method with small network, *IX International Symposium on Lightning Protection*.

Miki, M., Rakov, V. A., Rambo, K. J., Schnetzer, G. H. and Uman, M. A. (2002). Electric fields near triggered lightning channels measured with pockels sensors, *Journal of Geophysical Research: Atmospheres* **107**(D16): ACL 2–1–ACL 2–11.

Mingli, C., Xingsheng, L., Changming, G. and Zhengmo, G. (1991). A parametric method of site error estimation for a Lightning Location System, *Journal of Meteorological Research* **5**(3): 370.

Muller, M., Formayer, H., Diendorfer, G., Arpaci, A., Gossow, H. and Vacik, H. (2011). Analysis of lightning induced forest fires in Austria, *5th International Wildland Fire Conference*.

Murphy, M. J. (2006). When flash algorithms go bad, *2006 International Lightning Detection Conference (ILDC)*.

Naccarato, K. P., Saraiva, A. C. V., Saba, M. M. F. and Schumann, C. (2012). First performance analysis of BrasilDAT total lightning network in southeastern Brazil, *2012 International Conference on Grounding and Earthing and 5th International Conference on Lightning Physics and Effects*.

Naccarato, K. and Pinto, O. (2009). Improvements in the detection efficiency model for the Brazilian Lightning Detection Network (BrasilDAT), *Atmospheric Research* **91**(2): 546 – 563.

Nauslar, N. J. (2014). Examining the lightning polarity of lightning caused wildfires, *2014 International Lightning Detection Conference (ILDC)*.

Nestares, O., Fleet, D. J. and Heeger, D. J. (2000). Likelihood functions and confidence bounds for total-least-squares problems, *IEEE Conference on Computer Vision and and Pattern Recognition* **1**: 523–530.

O. Pinto Jr. (2008). The main advances and the most common misleading applications related to Lightning Location System data, *2008 International Lightning Detection Conference (ILDC)*.

O'Fallon, M. J. and Butterfield, K. D. (2005). A review of the empirical ethical decision-making literature: 1996-2003, *Journal of Business Ethics* **59**(4): 375–413.

Panyukov, A. and Avramenko, A. (2006). Forming of the map of probable quantity of lightning discharges for increasing accuracy of the single-point thunderstorm location system, *2006 International Lightning Detection Conference (ILDC)*.

Pédeboy, S. (2015). Distribution lightning fault correlation and advanced applications of lightning data, *2015 International Symposium on Lightning Protection*.

Pinto, I., Jr., O. P. and Naccarato, K. (2006). How ground flash density obtained by Lightning Location Networks can be used in lightning protection standards: A case study in Brazil, *2006 International Lightning Detection Conference (ILDC)*.

Poser, S., Bornstein, B. H. and McGorty, E. K. (2003). Measuring damages for lost enjoyment of life: The view from the bench and the jury box, *Law and Human Behavior* **27**(1): 53–68.

Rai, J., Sharma, D. K., Chand, R. and ad M. Israil, K. K. (2006). Horizontally and vertically polarized components of lightning electric fields, *19th International Lightning Detection Conference*.

Renni, E., E, K., Antonioni, G., Bonvicini, S., Spadoni, G. and Cozzani, V. (2009). Risk assessment of major accidents triggered by lightning events, **9**.

Robertson, B., Vignaux, G. A. and Berger, C. E. H. (2011). Extending the confusion about Bayes, *The Modern Law Review* **74**(3): 444–455.

Rocchi, M. B. L., Sisti, D., Ditroilo, M., Calavalle, A. and Panebianco, R. (2005). The misuse of the confidence ellipse in evaluating statokinesigram, *Italian Journal of Sport Sciences* **12**: 169–172.

Saraiva, A. C. V., Saba, M. M. F., Jr., O. P., Cummins, K. L., Krider, E. P. and Holle, R. L. (2010). On the variability of lightning characteristics over thunderstorm life-cycles, *2010 International Lightning Detection Conference (ILDC)*.

Schumann, C. and Saba, M. M. F. (2012). Continuing current intensity in positive ground flashes, *International Conference on Lightning Protection (ICLP)*.

Schumann, C., Saba, M. M. F., Schulz, W. and da Silva, R. G. B. (2011). Electric field waveforms of positive return strokes, *XIV International Conference on Atmospheric Electricity*.

Severson, V., Murray, P. and Heeter, J. (2014). Establishing cause and effect relationships between lightning flash data and airplane lightning strike damage, *2014 International Lightning Detection Conference (ILDC)*.

Shaw, W. and Lee, K. (2008). Bivariate student t-distributions with variable marginal degrees of freedom and independence, *Journal of Multivariate Analysis* **99**(6): 1276 – 1287.

Simon, T., Umlauf, N., Zeileis, A., Mayr, G. J., Schulz, W. and Diendorfer, G. (2017). Spatio-temporal modelling of lightning climatologies for complex terrain, *Natural Hazards and Earth System Sciences* **17**(3): 305–314.

Simpson, L. A., Botai, J., De Koning, E. and Sivakumar, V. (2011). Comparing lightning polarity and cloud microphysical properties over regions of high ground flash density in South Africa, *27th Annual Conference of the South African Society for Atmospheric Sciences: the Interdependent Atmosphere, Land and Ocean, Hartbeespoort, 22-23 September 2011.*

Sjastad, K., Lucy, D. and Andersen, T. (2016). Lead isotope ratios for bullets, forensic evaluation in a bayesian paradigm, *Talanta* **146**: 62 – 70.

Solakiewicz, R. and Koshak, W. (2008). Potential use of a Bayesian network for discriminating flash type from future GOES–R Geostationary Lightning Mapper (GLM) data, *2008 International Lightning Detection Conference (ILDC).*

Srihari, S. N. (2010). *Analysis of Footwear Impression Evidence*, PhD thesis, University at Buffalo.

Starr, S. (1998). *Development of algorithms and error analyses for the short baseline lightning detection and ranging system*, National Aeronautics and Space Administration, John F. Kennedy Space Center.

Steiner, M., Deierling, W., Ikeda, K. and Bass, R. (2014). Implications of lightning data uncertainty on personnel safety and operational efficiency decisions, *2014 International Lightning Detection Conference (ILDC).*

Strauss, C., Rosa, M. B. and Stephany, S. (2013). Spatio-temporal clustering and density estimation of lightning data for the tracking of convective events, *Atmospheric Research* **134**: 87 – 99.

Stroupe, A. W., Martin, M. C. and Balch, T. (2001). *Merging Gaussian Distributions for Object Localization in Multi-Robot Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Suzuki, F., Tsuji, H., Shimizu, H. and Shimizu, M. (2014). Evaluation of the new LS7001 network in the chubu region of japan after site-error corrections, *2014 International Lightning Detection Conference (ILDC).*

Vieland, V. J. and Hodge, S. E. (1998). Statistical evidence: A likelihood paradigm, *The American Journal of Human Genetics* **63**: 283–289.

Vlek, C., Prakken, H., Renooij, S. and Verheij, B. (2013). Modeling crime scenarios in a Bayesian network, pp. 150–159.

Wankhede, A. G. and Sariya, D. R. (2013). Damage due to lightning when it strikes the face, *Forensic Science International* **224**(1): e1 – e3.

Warner, T. A., Cummins, K. L. and Orville, R. E. (2012). Upward lightning observations from towers in Rapid City, South Dakota and comparison with National Lightning Detection Network data, 2004–2010, *Journal of Geophysical Research: Atmospheres* **117**(D19).

Wells, J. K. (2006). *Investigation of factors affecting the region of origin estimate in bloodstain pattern analysis*, Master's thesis, University of Canterbury.

Wen-Rong, S., Jin-Xiu, Z., Xiao-Lu, C. and Hao, Z. (2016). Lightning characteristics based on date observed with Shanghai Lightning Location System, *2016 International Conference on Condition Monitoring and Diagnosis (CMD)*, pp. 775–778.

Woolford, D. G. and Braun, W. J. (2007). Convergent data sharpening for the identification and tracking of spatial temporal centers of lightning activity, *Environmetrics* **18**(5): 461–479.

Zhang, Y., Lu, W., Zhang, Y., Zheng, D., Chen, L., Chen, S. and Yan, X. (2014). Evaluation for the performance of the Guangdong-Hongkong-Macau Lightning Location System, *2014 International Lightning Detection Conference (ILDC).*