IMPROVING FOREST GROWTH ESTIMATION
BAYESIAN NETWORKS FOR INTEGRATING
SATELLITE IMAGES AND PROCESS-BASED
FOREST GROWTH MODELS

Yaseen Taha Mustafa
PhD dissertation committee

Chair
Prof. dr. ir. A. (Tom) Veldkamp University of Twente, ITC

Promoter
Prof. dr. ir. A. (Alfred) Stein University of Twente, ITC

Assistant promoter
Dr. V. A. (Valentyn) Tolpekin University of Twente, ITC

Members
Prof. dr. A.K. Skidmore University of Twente, ITC
Prof. dr. W. Verhoef University of Twente, ITC
Prof. dr. S.M. de Jong Utrecht University
Prof. dr.ir. G. Jongbloed Delft University of Technology

ITC dissertation number 202
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

Printed by: ITC Printing Department

© Yaseen Taha Mustafa, Enschede, The Netherlands
All rights reserved. No part of this publication may be reproduced without the prior written permission of the author.
IMPROVING FOREST GROWTH ESTIMATION
BAYESIAN NETWORKS FOR INTEGRATING
SATELLITE IMAGES AND PROCESS-BASED
FOREST GROWTH MODELS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, March 28, 2012 at 16:45 hrs

by

Yaseen Taha Mustafa
born on January 18, 1979
in Mosul, Iraq
This dissertation is approved by:

Prof. dr. ir. A. (Alfred) Stein (promoter)
Dr. V. A. (Valentyn) Tolpekin (assistant promoter)
To my family ...

Father,

Mother,

Wife, &

lovely son.
Summary

Carbon sequestration through forestry has the potential to play a significant role in ameliorating global environmental problems such as atmospheric accumulation of greenhouse gases (GHG) and climate change. Estimating the contribution of forests to carbon sequestration is commonly done by applying process-based forest growth models, such as the Physiological Principles Predicting Growth (3-PG) model. Output of such a model, however, suffers from uncertainties and requires a close attention of their inputs. Moreover, with the development of modern satellite remote sensing imagery, Moderate Resolution Imaging Spectroradiometer (MODIS) sensor also monitors the forest and its growth. Satellite images provide extensive data of forests with a large spatial coverage. These data, however, are not free of uncertainties and are often not available. To address the reduced uncertainty and limited data availability, this thesis presents an approach which has been constructed using graphical statistical models, in particular Bayesian networks. This approach was applied to the Speulderbos forest in The Netherlands, where detailed data were available. In order to explore and solve the issues and requirements of this approach, five studies were carried out in this thesis.

First, a Gaussian Bayesian network (GBN) was used to address the bias in a process-based forest growth model and the noise within satellite images. A novel inference strategy within the GBN was developed to take care of the different structures of the inaccuracies in the two data sources. The obtained outputs with the GBN were more accurate than either the 3-PG model or the MODIS estimate. Hence, the GBN improved the estimation of the forest growth estimates values by integrating a 3-PG model with MODIS imagery.

Second, the performance of GBN modeling for improving forest growth estimates was evaluated. This was done by considering the quality of GBN output and relationships among GBN variables, in order to determine the reliability, the applicability and the robustness of the GBN. For that purpose, both evidence propagation, and a sensitivity analysis of the input sources, i.e. the 3-PG model output and MODIS observations were addressed. In the context of this study, a thorough assessment of the sensitivity of the proposed GBN aimed to achieve a better understanding of the potential of GBNs in modeling forest growth,
and in particular to estimate the relevance of the GBN input parameters and the required level of precision needed to provide accurate estimates of forest growth estimates. Evidence propagation by means of the 3-PG model improved the GBN output, showing that its relative error with respect to the field data decreased by 2.0%. This improvement is stronger than propagating the same evidence through MODIS images as the relative error of GBN output increased by 4.5%. It was also found that a GBN is more sensitive to the variation of satellite estimates than to variation in forest growth model output.

Third, the performance of GBN modeling for forest growth estimates was investigated and assessed with missing input sources, i.e. satellite data. The satellite time series, however, contained gaps caused by persistent clouds, and cloud contamination. The EM-algorithm was formulated and integrated within a GBN to estimate these missing values. During a period of 26 successive months, the EM-algorithm was applied by making synthetic gaps at four different cases: successively and not successively missing values during two different winter seasons, successively and not successively missing values during one spring season, and not successively missing values during the full study time. The estimated values well represented the original values where the maximum values of the averaged absolute error between the original values and those estimated was equal 0.16. Moreover, the root mean square error of the GBN output reduced from 1.57 to 1.49.

Fourth, GBN modeling for forest growth estimates with the EM-algorithm was applied to real gaps of satellite data. The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) sensor was used as input source into the GBN. Moreover, the GBN was modified to improve forest growth estimates that show variation in space. The GBN output was more accurate than the 3-PG estimate, as the root mean square error reduced to 0.46, and the relative error to 5.86%. Moreover, the deviation of the averaged output of spatial GBN and field data was less than the deviation between the averaged output of the 3-PG and field data.

Finally, the spatio-temporal estimation of growth estimates of a heterogeneous forest using GBN was carried out. GBN combined the spatial version of the 3-PG model output with decomposed MODIS images. The Linear Mixture Model (LMM) was used to decompose MODIS pixels using class fraction derived from an aerial image and an ASTER image. In this way spatially heterogeneous output was produced. Results showed that the spatial output obtained with the GBN was more than 40% accurate than both the spatial 3-PG model output and the satellite estimate, with a root mean square error less than 0.53.

The study focused on improving forest growth estimates by using GBN to integrate two sources providing forest growth estimates, i.e. 3-PG model and satellite images. Such integration can be applied to different types of forests by giving more attention to remote sensing images. The study concluded that forest growth estimates could be improved using GBNs. Moreover, formulating the EM-algorithm within a GBN
can adequately handle missing satellite data and improve estimation of forest growth estimates. Ultimately, this study contributed to provide the means for an accurate and reliable estimation of forest growth estimates.
Samenvatting

Koolstof sekwstratie door middel van bosbouw biedt de mogelijkheid om een belangrijke rol te spelen bij het verlichten van wereldwijde milieuproblemen, zoals de opeenhoping van broeikasgassen in de atmosfeer en klimaatverandering. De bijdrage van bossen hieraan wordt als regel in beschouwing genomen bij het toepassen van proces-afhankelijke groeimodellen voor de groei ervan, zoals bij het 3-PG model voor de voorspelling van groei van een bos, dat op fysiologische principes is gebaseerd. De uitvoer van een dergelijk model heeft te lijden onder onzekerheden en vereist een zorgvuldige controle op de invoergegevens. Daarnaast is het ook mogelijk om met moderne satellietwaarnemingen, zoals verkregen met de Moderate Resolution Imaging Spectroradiometer (MODIS) sensor, het bos en zijn groei te monitoren. Satellietbeelden bieden uitgebreide hoeveelheden gegevens met een grote ruimtelijke dekking. Deze gegevens zijn echter niet vrij van fouten en zijn vaak ook niet beschikbaar. Teneinde om te gaan met de verminderte zekerheid en de beperkte beschikbaarheid van de gegevens presenteert dit proefschrift een benadering gebaseerd op grafische statistische methoden, in het bijzonder Bayesiaanse netwerken. Deze benadering is gebruikt in het Speulderbos in Nederland, waar gedetailleerde gegevens beschikbaar waren. Om verschillende aspecten en problemen te verkennen en op te lossen zijn in dit proefschrift vijf studies beschreven.

In de eerste plaats is een Gaussisch Bayesiaans Netwerk gebruikt om de onzuiverheid in een procesmodel en de ruis in satellietbeelden te analyseren. Een nieuwe schattingsmethode is ontwikkeld binnen een GBN om zorg te dragen voor de verschillende structuren in de onnauwkeurigheden in de twee gegevensbronnen. De uitvoer verkregen met het GBN is nauwkeuriger dan zowel de uitvoer van het 3-PG model als de schattingen die met MODIS worden gemaakt. In die zin heeft het GBN de schatting van de waarde van de groei van het bos verbeterd door het 3-PG model te integreren met het MODIS beeld.

In de tweede plaats is de prestatie van het modelleren van de schattingen met een GBN voor de groei van het bos geëvalueerd. Dit is gedaan door zowel de kwaliteit van de uitvoer van het GBN te bekijken als de relaties met andere GBN variabelen. Aandacht is geschonken aan de betrouwbaarheid, de toepasbaarheid en de robuustheid van een GBN. Voor dat doel zijn zowel de voortschrijding van bewijs als een gevoe-
Samenvatting

ligheidsanalyse van de invoergegevens uitgevoerd. Hierbij zijn zowel de uitvoer van het 3-PG model als de MODIS waarnemingen geanalyseerd. In de context van deze studie was het doel dat het voorgestelde GBN een beter begrip zou geven van de gebruiksmogelijkheden van een GBN in het modelleren van de groei van het bos en in het bijzonder om de relevantie van de invoergegevens van een GBN te schatten en de precisie vast te stellen die nodig is om nauwkeurige schattingen te maken voor de groei van het bos. Voortschrijding van bewijs via het 3-PG model verbeterde de uitvoer van het GBN zodanig dat de relatieve fout met betrekking tot gebruik van veldgegevens afnam met 2.0%. Een dergelijke verbetering is zelfs nog sterker dan bij het gebruik van MODIS beelden, waar de relatieve fout van de GBN uitvoer toenam met 4.5%. Ook bleek dat een GBN gevoeliger is voor de variaties in satellietsschattingen dan het groeimodel voor het bos zelf.

In de derde plaats is de prestatie van het modelleren met een GBN onderzocht en geschat bij het ontbreken van gegevensbronnen, in het bijzonder satellietgegevens. Satellietgegevens bevatten gaten die veroorzaakt worden door de langdurige aanwezigheid van wolken en verontreiniging in de vorm van wolken. Het EM-algoritme is geformuleerd en geïntegreerd binnen een GBN om dergelijke missende waarden te schatten. Gedurende een periode van 26 opeenvolgende maanden is dit algoritme eerst gebruikt om kunstmatig gaten in de tijdserie te maken: opeenvolgende en niet-opeenvolgende gaten in de waarnemingen zijn gemaakt gedurende twee winters en één voorjaar en niet-opeenvolgende gaten zijn gemaakt gedurende de gehele studieperiode. De geschatte waarden waren prima in overeenstemming met de oorspronkelijke gegevens, waar de maximum waarde van de gemiddelde absolute fout tussen de oorspronkelijke en de geschatte waarden gelijk was 0.16. Bovendien liep de wortel van de gemiddelde gekwadrateerde fout terug van 1.57 naar 1.49.

Vervolgens is het GBN modelleren voor de groei van het bos, met daarin geïntegreerd het EM-algoritme, toegepast om daadwerkelijke gaten in de satellietwaarnemingen op te vullen. De Advanced Spaceborne Thermal Emission en Reflection Radiometer (ASTER) sensor was gebruikt als een invoerbron voor gegevens voor het GBN. Bovendien was het GBN zelf aangepast om gegevens voor de groei van het bos te schatten die ruimtelijke variatie laten zien. De uitvoer van het GBN was nauwkeuriger dan de uitvoer van het 3-PG model, omdat de wortel uit de gemiddelde gekwadrateerde fout terugliep tot 0.46 en de relatieve fout tot 5.86%. Ook is de afwijking tussen de gemiddelde uitvoer van het ruimtelijke GBN en de veldgegevens kleiner dan die tussen de gemiddelde uitvoer van het 3-PG model en de veldgegevens.

Tenslotte is een ruimtelijk temporele schatting uitgevoerd voor een heterogene bos met behulp van een GBN. Het GBN combineerde de ruimtelijke versie van het 3-PG model met een decompositie van een MODIS beeld. Het lineaire gemengde model (LMM) is gebruikt om MODIS pixels te ontdoen, waarbij gebruik gemaakt werd van de fracties in de klassen die waren verkregen via een luchtfoto en een ASTER beeld. Op deze manier is ruimtelijk heterogene uitvoer verkregen. De resultaten
lieten zien dat de ruimtelijke uitvoer die verkregen was met een GBN meer dan 40% nauwkeuriger was dan zowel het ruimtelijke 3-PG model en de schatting die verkregen was met een satellietbeeld, met een wortel uit de gemiddelde gekwadrateerde fout die kleiner was dan 0.53.

Het proefschrift heeft zich gericht op het verbeteren van schattingen voor de groei van een bos door een GBN te gebruiken dat twee bronnen van informatie integreert: het 3-PG model en satellietbeelden. Een dergelijke integratie kan worden toegepast op verschillende typen bossen door meer aandacht te geven aan beschikbare satellietbeelden. De studie komt tot de conclusie dat schattingen voor de groei van een bos verbeterd konden worden door Gaussische Bayesiaanse Netwerken te gebruiken. Bovendien was het mogelijk om met een aangepast en geïntegreerd EM-algoritme ontbrekende satellietwaarden te schatten en de schattingen voor de groei van het bos te verbeteren. Ultimo heeft deze studie eraan bijgedragen om de middelen ter beschikking te stellen voor een nauwkeurige en betrouwbare schattingen van parameters voor de groei van het bos.
Acknowledgments

This research would not have been possible without the help, encouragement and appreciation of several individuals and organizations. However, I realized that it is impossible to name them all here. The help of those whose names are not mentioned is as greatly appreciated as the help of those whose names are.

First of all, I convey my sincere thanks and gratitude to my promoter Prof. Stein for his guidance and support. It has always been an honor to work with him. Heartily gratitude is extended to my co-promoter Dr. Tolpekin for his support, and sharing in depth scientific knowledge. Without his valuable comments, it would have been difficult to complete this research. I enjoyed working with my promoter and co-promoter, and I wish to continue in the future. I would also like to send my sincere thanks to Dr. Van Laake for his assistance, especially at the beginning of this research. I have built the idea of this research with him. Thank you very much.

Profound thanks are due to the faculty of the EOS Department for their continuous support and encouragement, especially to Teresa Brefeld and Muditha Heenkenda. Thanks to Dr. Paul van dijk, Loes Colenbrander, Bettine Geerdink, and Rébanna Spaan for their assistance and providing me with relevant information from the beginning to the end of my PhD study. My thanks go to Job Duim for his help in editing the cover of this dissertation. I would also like to thank Dr. Christiaan van der Tol and Murat Ucer for their help and providing me with information and data of the study area. Sincere appreciation to Dr. Rolf de By for formulating and modifying the LaTeX template that used for this thesis.

I am immensely grateful to my office mates, especially to Juan Pablo Ardila for his fruitful help and continuous support during my PhD study. I have learned a lot from you, thank you. Sincere appreciation to all PhD colleagues at EOS Department for their support, company, and sharing scientific knowledge. Great thanks to my country mates Bashar AlSadik and Ahmed AlSammarai and to their families. We as a family had a pleasure time together, thanks for your moral support.

My PhD has been obtained and funded by Erasmus Mundus for the first three years. Thank you Erasmus Mundus to give me this opportunity to do a PhD at ITC. It is a good opportunity to thank Angelique Holtkamp for her encouragement and assistance, especially in the beginning of my
Acknowledgments

The last year of my PhD has been funded by the Ministry of Higher Education and Scientific Research (MHESR) of Kurdistan region-Iraq. My heartfelt thanks to the MHESR of Kurdistan region-Iraq, headed by the Minister Prof. Dlawer A. Ala’Aldeen. I would also like to express my thanks to the president of the University of Dohuk Dr. Asmat M. Khalid and to the president of the University of Zakho Dr. Lazgin A. Jamil for their assistance and support in getting permission to continue my study and their efforts to get the fund from the MHESR of Kurdistan region. My special thanks also go to Prof. Nazar M.S. Numan for his support and efforts in finding this study. I would use this opportunity to thank my friends Ramadhan Abed, Razwan Mohammed, and Rebar Tahseen for their efforts and assistance in following up my documents and my status in Kurdistan region during my study.

Most importantly, I will never be deeply indebted to my parents for their love and prayers that enable me to achieve this goal. I am always thankful to them. Last and not least, I cannot find a suitable word to express my acknowledgments to my wife for her love, encouragement and support at every moment of my PhD study. Thanks a lot to almighty for blessing us with a handsome and keen son Yousif, who fill our life with joy and fun. We love you so much.
# Contents

Summary iii  
Samenvatting vii  
Acknowledgments xi  
Contents xiii  
List of Nomenclatures xix  

1 General introduction 1  
1.1 Background 2  
1.2 Modelling forest growth 3  
1.3 Remote sensing imagery 5  
1.4 Probabilistic graphical models 7  
1.5 Problem statement 8  
1.6 Research objectives 9  
1.7 Application 9  
1.8 Outline of the thesis 12  

2 Bayesian network modeling for improving forest growth estimates 15  
2.1 Introduction 17  
2.2 Model description 21  
2.3 Implementation 27  
2.4 Discussion 34  
2.5 Conclusions 36  

3 Performance evaluation of GBN modeling for forest growth estimates 37  
3.1 Introduction 39  
3.2 Materials and methods 41  
3.3 Implementation 47  
3.4 Experimental results 47  
3.5 Discussion 51  
3.6 Conclusion 54  

xiii
## Contents

4 Application of the EM-algorithm to estimate missing values in GBN 55  
4.1 Introduction ........................................ 57  
4.2 Materials and methods ............................ 58  
4.3 Implementation .................................. 64  
4.4 Experimental results ............................... 66  
4.5 Discussion ......................................... 71  
4.6 Conclusion ......................................... 72  

5 Improving forest growth estimates using a Bayesian network approach 75  
5.1 Introduction ......................................... 77  
5.2 Materials and methods ............................ 77  
5.3 Study area .......................................... 82  
5.4 Data description .................................... 83  
5.5 Implementation .................................... 84  
5.6 Results ............................................. 85  
5.7 Discussion and conclusions ........................ 86  

6 GBN to improve spatio-temporal growth estimates of heterogeneous forests 89  
6.1 Introduction ......................................... 91  
6.2 Materials and data .................................. 92  
6.3 Methods ............................................. 96  
6.4 Implementation .................................... 100  
6.5 Results ............................................. 101  
6.6 Discussion ......................................... 106  
6.7 Conclusion ......................................... 107  

7 Synthesis .................................................. 109  
7.1 General discussion .................................. 110  
7.2 Research findings ................................... 112  
7.3 Reflections .......................................... 115  
7.4 Limitations and further scope of research ........ 117  
7.5 General suggestions ............................... 119  

Bibliography ............................................... 121
List of Figures

1.1 (a) Map region of The Netherlands. (b) Location of Garderen village. (c) Location of study area of $1 \times 1$ km. 

1.2 A network, showing the connection and the relationship between thesis chapters.

2.1 Basic structure part of the 3-PG model.

2.2 The BN for $i^{th}$ iterations, $i \geq 1$, every triple nodes $LAI_{M_i}, LAI_{BN_i}$ and $LAI_{3PG_i}$ indicates the network at time $i$.

2.3 Workflow process of improving the LAI value.

2.4 Two normal Q-Q plots and histograms with the normal curve of $LAI_{M}$.

2.5 Normal Q-Q plots and histograms with the normal curve of the $LAI_{3PG}$.

2.6 LAI distribution of the Speulderbos forest obtained from four sources; $LAI_{FD}$, $LAI_{3PG}$, $LAI_{M}$, and $LAI_{BN}$ during the period July 2007–September 2009.

2.7 Sensitivity test of the BN after independently varying the $LAI_{3PG}$ and $LAI_{M}$ values.

2.8 Sensitivity test of the BN after simultaneously varying the $LAI_{3PG}$ and $LAI_{M}$ values.

3.1 The GBN for $i^{th}$ iterations, $i \geq 1$, $LAI_i$ indicates the network at time $i$. It consists of triple nodes $LAI_{3PG_i}, LAI_{BN_i}$ and $LAI_{M_i}$.

3.2 LAI_{BN} before and after evidence propagation in GBN through $LAI_{3PG}$ during (a) the winter seasons, (b) the summer seasons.

3.3 Sensitivity measure values ($S^p$) after $LAI_{3PG}$ has received evidence at two winter and two summer seasons.

3.4 $LAI_{BN}$ before and after evidence propagation in GBN through $LAI_{M}$ during (a) the winter seasons, (b) the summer seasons.

3.5 Sensitivity measure values ($S^p$) after $LAI_{M}$ has received evidence at two winter and two summer seasons.

4.1 The BN for the first, second and $i^{th}$ time step. Each time step consists of three nodes $LAI_{3PG_i}, LAI_{BN_i}$ and $LAI_{M_i}$.

4.2 BN with missing satellite observations.
List of Figures

4.3 Estimated parameters ($\mu^{\text{new}}_\text{x}, \sigma^{\text{new}}_\text{x}$) and the differences of two consecutive iterations in the EM-algorithm. .......................... 65
4.4 $\text{LAI}_{\text{BN}}$ and $\text{LAI}_{\text{FM}}$ values of the Speulderbos forest obtained before and after performing the EM-algorithm during the winter of 2008. ................................................................. 67
4.5 $\text{LAI}_{\text{BN}}$ and $\text{LAI}_{\text{FM}}$ values of the Speulderbos forest obtained before and after performing the EM-algorithm during the winter of 2009. ................................................................. 69
4.6 $\text{LAI}_{\text{BN}}$ and $\text{LAI}_{\text{FM}}$ values of the Speulderbos forest obtained before and after performing the EM-algorithm during the spring of 2009. ................................................................. 70
4.7 $\text{LAI}_{\text{BN}}$ and $\text{LAI}_{\text{FM}}$ values of the Speulderbos forest obtained before and after performing the EM-algorithm for not successfully missing $\text{LAI}_{\text{FM}}$ values during the period from July 2007 to September 2009. ................................................................. 70

5.1 (a) The BN for $i^{th}$ moments. Each moment consists of three nodes $\text{LAI}_{\text{SPG}}, \text{LAI}_{\text{BN}}$, and $\text{LAI}_{\text{SAT}}$. (b) Network with missing satellite observations. ................................................................. 80
5.2 Location of the study area. ................................................................. 83
5.3 The relationship between NDVI$_{\text{SAT}}$ and $e\text{LAI}_{\text{FD}}$. ......................... 84
5.4 LAI distribution of the Douglas fir trees obtained from four sources, namely, the field data, the 3-PG model, the ASTER images, and the GBN, during the period from June 26, 2010 to October 16, 2010. ................................................................. 85
5.5 Spatial LAI values of the Speulderbos forest obtained from one observation at June 26, 2010. ................................................................. 86

6.1 (a) Map region of The Netherlands, (b) Location of Garderen village (c) Location of study area, and (d) sampling design of the $1\times1$ km with five tree species. ................................................................. 94
6.2 (a) RGB image of the 1-km of the study area from ASTER bands 3, 2. (b) Map of the 1-km region using manual segmentation. ................................................................. 97
6.3 An example of a GBN output of one pixel with spatial scale of 250 m consists of three types. ................................................................. 98
6.4 Decomposed LAI MODIS into five LAI values of five species, Beech, Douglas fir, Hemlock, Japanese Larch, and Scotch pine. ................................................................. 101
6.5 LAI distribution of the Speulderbos forest of five species: (a) Beech, (b) Douglas fir, (c) Hemlock, (d) Japanese Larch, and (e) Scotch Pine. ................................................................. 102
6.6 Resulting relative errors of the spatial GBN output with respect to the field data of Douglas fir trees for the period of 14 months from June 26, 2010 to August 29, 2011. ................................................................. 104
6.7 The spatio-temporal distribution of the estimated LAI from four sources: $\text{LAI}_{\text{FD}}, \text{LAI}_{\text{SPG}}, \text{LAI}_{\text{SAT}}$, and $\text{LAI}_{\text{BN}}$. ................................................................. 105
6.8 LAI distribution of Douglas fir for the study area during the period June 2010 to August 2011. ................................................................. 105

xvi
List of Tables

2.1 Two normality tests. (a) Normality tests for the LAIM. (b) Normality tests for the LAIPG. 28

2.2 Mean values (Mean), standard deviation (σ), Root Mean Square Errors (RMSE) and Relative Errors (RE) for the various ways to estimate the LAI. 32

3.1 The minimum data required to run 3-PG model. 39

3.2 (a) Expectation and (b) variance values of average (Avg), maximum (Max), minimum (Min), and LAI value at middle time period (AtM*) of the season of all of LAIBN, LAIPG, and LAIFD for each season. 48

3.3 The δj value, LAIBN distribution after add the perturbation values to GBN, and sensitivity measure (Spj) value after LAIPG has received evidence in the 1st & the 2nd winter season. 49

3.4 The δj value, LAIBN distribution after add the perturbation values to GBN, and sensitivity measure (Spj) value after LAIPG has received evidence in the 1st & the 2nd summer season. 50

4.1 The RMSE and the RE of LAIM and LAIBN, and the AAE of LAIM. They are obtained before and after applying the EM-algorithm during the winter of 2008. 68

4.2 The RMSE and the RE of LAIM and LAIBN, and the AAE of LAIM. They are obtained before and after applying the EM-algorithm during the winter of 2009. 68

4.3 The RMSE and the RE of LAIM and LAIBN, and the AAE of LAIM. They are obtained before and after applying the EM-algorithm during the spring of 2009. 69

5.1 Mean values (Mean), standard deviation (σ), Root Mean Square Errors (RMSE) and Relative Errors (RE) for the various ways to estimate the LAI. 86

6.1 The 3-PG parameters of each species were determined from the following sources. 93

6.2 The area proportion (A), and the number of referenced pixels (N) for each species in the composition of MODIS pixels. 98
List of Tables

6.3 The $\mu$, the $\sigma$, tRMSE and tRE for the various ways to estimate the LAI: $LAI_{FD}$, $LAI_{3PG}$, $LAI_{SAT}$, and $LAI_{BN}$ with composed and decomposed LAI MODIS of five tree species. . . . . . . . . . . . . 103

6.4 The $\mu$, the $\sigma$, sRMSE and sRE for the various ways to estimate the LAI: $LAI_{FD}$, $LAI_{3PG}$, $LAI_{SAT}$, and $LAI_{BN}$; at 250 m spatial resolution of mixed species. . . . . . . . . . . . . . . . 104
## List of Nomenclatures

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-PG</td>
<td>Physiological Principles Predicting Growth model.</td>
</tr>
<tr>
<td>AAE</td>
<td>Averaged absolute error.</td>
</tr>
<tr>
<td>ASTER</td>
<td>Advanced Spaceborne Thermal Emission and Reflection Radiometer.</td>
</tr>
<tr>
<td>AST_07</td>
<td>ASTER surface reflectance product.</td>
</tr>
<tr>
<td>BN</td>
<td>Bayesian network.</td>
</tr>
<tr>
<td>DAG</td>
<td>Directed acyclic graph.</td>
</tr>
<tr>
<td>DHP</td>
<td>Digital hemispherical photograph.</td>
</tr>
<tr>
<td>EM-algorithm</td>
<td>Expectation Maximization-algorithm.</td>
</tr>
<tr>
<td>FPAR</td>
<td>Fraction of absorbed photosynthetically active radiation.</td>
</tr>
<tr>
<td>GBN</td>
<td>Gaussian BN.</td>
</tr>
<tr>
<td>GLA</td>
<td>Gap Light Analyzer software.</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System.</td>
</tr>
<tr>
<td>GSD</td>
<td>Ground sampling distance.</td>
</tr>
<tr>
<td>KLD</td>
<td>Kullback-Leibler divergence.</td>
</tr>
<tr>
<td>LAI</td>
<td>Leaf area index.</td>
</tr>
<tr>
<td>LMM</td>
<td>Linear Mixture Model.</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood.</td>
</tr>
<tr>
<td>MODIS</td>
<td>Moderate Resolution Imaging Spectroradiometer.</td>
</tr>
<tr>
<td>MOD13</td>
<td>Vegetation Indices MODIS product.</td>
</tr>
<tr>
<td>MOD15</td>
<td>LAI/FPAR MODIS product.</td>
</tr>
<tr>
<td>NDVI</td>
<td>Normalized difference vegetation index.</td>
</tr>
<tr>
<td>PAR</td>
<td>Photosynthetically active radiation.</td>
</tr>
<tr>
<td>PSF</td>
<td>Point spread function.</td>
</tr>
<tr>
<td>RE</td>
<td>Relative error.</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean square error.</td>
</tr>
<tr>
<td>SVI</td>
<td>Spectral vegetation indices.</td>
</tr>
</tbody>
</table>
### List of Nomenclatures

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>Extinction coefficient.</td>
</tr>
<tr>
<td>$LAI_{BN}$</td>
<td>LAI output of BN.</td>
</tr>
<tr>
<td>$LAI_{3PG}$</td>
<td>LAI output of 3-PG model.</td>
</tr>
<tr>
<td>$LAI_{M}$</td>
<td>Derived LAI from NDVI MODIS product.</td>
</tr>
<tr>
<td>$eLAI_{FD}$</td>
<td>Effective LAI measured on the ground.</td>
</tr>
<tr>
<td>$LAI_{FD}$</td>
<td>LAI field data after $eLAI_{FD}$ adjusted by clumping factor.</td>
</tr>
<tr>
<td>$LAI_{AST}$</td>
<td>Derived LAI from empirical relationship between $eLAI_{FD}$ and SVI ASTER data.</td>
</tr>
<tr>
<td>$LAI_{SAT}$</td>
<td>Derived LAI from Satellite data.</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean.</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>Covariance matrix.</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Standard deviation.</td>
</tr>
<tr>
<td>$\beta_{ij}$</td>
<td>Regression coefficients of BN output on its parents.</td>
</tr>
<tr>
<td>$pa_i$</td>
<td>Set of parents of node $i$.</td>
</tr>
<tr>
<td>$#pa_i$</td>
<td>Number of the set of parents of node $i$.</td>
</tr>
<tr>
<td>$\nu_i$</td>
<td>Conditional variance of node $i$ ($LAI_{BN}$) given its parents.</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Set of parameters ($\mu, \Sigma$).</td>
</tr>
<tr>
<td>$\bar{e}$</td>
<td>Evidence value, from field data, used in GBN.</td>
</tr>
<tr>
<td>$\delta_i$</td>
<td>Perturbation values at iteration $i$ to check the sensitivity of GBN.</td>
</tr>
<tr>
<td>$S_{i}^{pj}$</td>
<td>Sensitivity measure for the parameter $j$ at iteration $i$.</td>
</tr>
</tbody>
</table>
# General introduction

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Background</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Modelling forest growth</td>
<td>3</td>
</tr>
<tr>
<td>1.3</td>
<td>Remote sensing imagery</td>
<td>5</td>
</tr>
<tr>
<td>1.4</td>
<td>Probabilistic graphical models</td>
<td>7</td>
</tr>
<tr>
<td>1.5</td>
<td>Problem statement</td>
<td>8</td>
</tr>
<tr>
<td>1.6</td>
<td>Research objectives</td>
<td>9</td>
</tr>
<tr>
<td>1.7</td>
<td>Application</td>
<td>9</td>
</tr>
<tr>
<td>1.7.1</td>
<td>Study site description</td>
<td>9</td>
</tr>
<tr>
<td>1.7.2</td>
<td>Data description</td>
<td>10</td>
</tr>
<tr>
<td>1.8</td>
<td>Outline of the thesis</td>
<td>12</td>
</tr>
</tbody>
</table>
1. General introduction

1.1 Background

One of the major impacts on changing global climate is releasing a large amount of heat-trapping gases—in particular carbon dioxide (CO$_2$)—into the atmosphere. These gases contribute to higher global temperatures that could in turn increase the frequency of extreme weather events and have a profound impact on human health, sea levels, natural habitats, and agriculture. Among the greenhouse gases (GHGs), CO$_2$ is the most important due to its amount of approximately are 55% of the total proportion of GHG emissions (IPCC, 2007). With the increasing atmospheric concentration of CO$_2$, an extensive international literature has evolved concerning the understanding of the balance of carbon pools and fluxes in forests and forest products (Sedjo, 1993; Goodale et al., 2002; Piao et al., 2005; Woodbury et al., 2007).

Forests are important components of the global carbon cycle: they cover about $40 \times 10^6$ km$^2$, of the Earth’s land surfaces (FAO, 2010). Forests exchange large amounts of CO$_2$ and other gases with the atmosphere and store carbon, in various forms, in trees and soils. Storage of carbon in plants or soils is called “sequestered carbon”, whereas in the opposite way it is called “respired carbon” by returning it to the atmosphere when it has been used by trees or other organisms as energy for life. Carbon sequestration, thus, is the process by which atmospheric CO$_2$ is absorbed by trees through photosynthesis and stored as carbon in biomass (trunks, branches, foliage, and roots) and soils. Carbon sequestered in forests and wood products helps to offset fossil fuel emissions, and is a key driver of human-induced climate change.

Accordingly, forests play a major role in carbon sequestration (Kauppi et al., 1992; Nabuurs et al., 1997), thus affecting the speed of climate change. As a tree grows, it absorbs CO$_2$ from the atmosphere through photosynthesis, and uses solar energy to store carbon in its roots, stems, branches, and foliage. A growing forest is a carbon sink; in other words, it fixes more carbon through photosynthesis than what it releases via respiration. When the forest reaches maturity, equilibrium is created between the quantity of carbon fixed and the amount released. Forest growth has increased in recent decades, as shown by studies of temperate forests in North America (Sedjo, 1992) and Europe (Spiecker, 1999; Hunter and Schuck, 2002). Forest growth can be assessed through measuring and assessing its biophysical parameters as a leaf area index (LAI). LAI was first defined by Watson (1947) as the total one-sided area of photosynthetic tissue per unit ground surface area. LAI has an important influence on the exchange of energy, water vapor and carbon dioxide between terrestrial ecosystems and the atmosphere. Consequently, most distributed ecosystem process models that simulate carbon cycles require LAI as an input variable.

Forest biophysical parameters can be determined by means of measurements in the forest, by means of applying a forest growth model, and by means of a remote sensing image. Each approach has its own strategy.
of estimating biophysical parameters. However, those approaches are share the same rules in their estimation, i.e. to properly representing the forest canopy in terms of space and time. Sampling used to estimate biophysical parameter as LAI is probably as crucial as the approach used for estimation. Common to all methods and instruments is the question of spatial and temporal relevance, especially for ground-based LAI estimates. Parameters like canopy height and vertical stratification, plot dimensions, site topography, spatial integration of sensors, and canopy continuity are of importance for the procedure used and for the reliability of the results. The timing of sampling is related to the natural or incidental seasonal time-course of LAI. Recent research has improved LAI estimates through a better description and sampling of canopy heterogeneity (Jonckheere et al., 2004; Weiss et al., 2004; Martinez et al., 2009). It is difficult, however, to make any generalization, as each research requires its own strategy to fit its demands, bearing in mind the physiological process of interest (Weiss et al., 2004). Subsequently, sampling is often crucial as spatial variability in canopies is large, and replicates at several locations may be used to determine LAI.

1.2 Modelling forest growth

Forestry science combines experimentation and data analysis with planning and management. This dual task requires a means of quantitatively predicting future trends from current and historic measurements. Besides the forest importance and its role in the carbon sequestration, a better quantitative understanding of the growth processes in tree stands is of importance for the management and optimization of wood production and quality. For those reasons, several models have been established based on mathematical models.

Models are abstractions to encapsulate essential features of the system to be modeled. All models are approximations and therefore they should be testable; i.e. it should be possible to design experiments, or make measurements which, at least in principle, could falsify the hypotheses and show the model to be wrong. Models can be verbal, diagrammatic, graphical or mathematical. To make quantitative predictions about the behavior and responses of systems, models usually are written in mathematical terms. Most models are deterministic, which implies that in relationships of the form \( y = f(x) \) each particular value of \( x \) will lead to a unique value of the dependent variable \( y \). There is, however, uncertainty associated with all knowledge we have. This is particularly the case in relationships between biological variables, where there is often significant uncertainty in understanding of the fundamental processes, and natural biological variables. Most models are developed as research tools, and they are seldomly used by other scientists and even less frequently by non-scientists such as managers and policy makers (Sands, 1988).
1. General introduction

Forest growth models have a broader role in forest management and in the formulation of forest policy. Used to the advantage and in conjunction with other resource and environmental data, forest growth models can be used to make predictions, formulate prescriptions and guide forest policy. In the context of growth modelling, sufficient feedback should ensure an adequate inventory and reliable model predictions across the range of resource conditions and management prescriptions. These various roles draw on different qualities in growth models and provide a place for different kinds of models. There are three main types of forest growth models: empirical, process-based or mechanistic, and hybrid. Empirical forest growth models usually use the site index as a productivity rating to predict future forest growth from historical measurements of forest growth. Their application is therefore limited to the temporal and spatial scales under which they were developed (Battaglia and Sands, 1998; Landsberg, 2003b). Presently, industrial and governmental forest managers demand decision-support tools that overcome these limitations to prediction and extrapolation. Forest and plantation management questions have become more complex, including considerations of risk minimization from extreme events, mitigation of the effects of climate change, insect pests and environmental pollution, as well as sustainable forest management in terms of water yield (Battaglia and Sands, 1998; Coops and Waring, 2001b; Landsberg, 2003a; Nole et al., 2009). Process-based and hybrid approaches to forest growth modeling seek to overcome the limitations of empirical forest growth models to help answer these complex policy, management and scientific questions (Constable and Friend, 2000; Makela et al., 2000; Sands et al., 2000; Landsberg, 2003a). Process-based models may be applicable beyond the spatial and temporal scope of empirical models, if they can capture the fundamentals of forest system functions (Constable and Friend, 2000; Van Oijen, 2002).

Several forest growth models have been established to understand, estimate and predict forest growth. Examples are the Carbon Balance (CABALA) model (Mummery and Battaglia, 2004), the Forest Growth (FORGRO) model (Mohren and Van de Veen, 1995), and the physiological principles predicting growth (3-PG) model (Landsberg and Waring, 1997). In this thesis we focused on the 3-PG model. It is a flexible process-based forest growth model, with modest requirements for parameter values (Nightingale et al., 2008). The 3-PG model has been applied to single-species plantations and even-aged relatively homogeneous forests under different climatic and edaphic regions across the globe. Moreover, it has become more widely available within the forestry research community and it has been widely used around the world (Landsberg et al., 2001; Tickle et al., 2001; Coops et al., 2004; Xenakis et al., 2008; Nightingale et al., 2008; Feikema et al., 2010).

The 3-PG model has been developed as the spatial 3-PG model, and it has been applied to larger area by using either spatial databases, geographic information systems (GIS) or remote sensing (White et al., 2000; Coops and Waring, 2001, b; Tickle et al., 2001). It produces spatially...
and temporally explicit output at the scale of the input surfaces. It can be run for a period of several years in monthly timesteps, and it has been adjusted to run in 16 days timesteps. The model requires climatic data, site factors data and species specific data. The climate data includes extreme air temperature (°C), mean solar radiation (MJ m⁻²), total precipitation (mm), and frost days (days month⁻¹). Site factor data include latitude, maximum available water (mm), and soil fertility rate. The species data are rather different because those data characterize canopy properties. Most parameter values can be obtained from the literature or from field measurements. As output, it produces LAI, pool biomass (stem, foliage and root biomass), gross primary productivity (GPP), net primary productivity (NPP), stand water use, and available soil water as well. The output of interest in this thesis is the LAI, being an important indicator of vegetation status and a key parameter in process-based models to quantify the exchange of matter and energy flow between vegetation and the atmosphere.

1.3. Remote sensing imagery

Remote sensing deals with the detection and measurement of phenomena with devices sensitive to electromagnetic energy, such as cameras and scanners for light, thermal scanners for heat, and radar for radio waves. The basic components to measure remotely sensed data include the energy source, the transmission path, the target and the sensor. Remote sensing makes it possible to collect data at inaccessible areas with information available in various spectral, spatial, angular and temporal resolutions and polarization domains. Remote sensing in Earth resource analysis can be applied for physical, natural, and social sciences, such as geography, soil, biogeography, geology, hydrology, urban planning, agriculture, forestry, and marine sciences (Skidmore et al., 1997; Verhoef and Bach, 2007; Jensen, 2007; Wei et al., 2012). Remote sensing applications in Earth resource management include monitoring deforestation in areas as big as the Amazon Basin (Reddy et al., 2009), the effects of climate change (Hunt et al., 2004; Miehle et al., 2006), ecosystem productivity (Huemmrich et al., 2010), hydrology (Tang et al., 2009), and related environmental topics. Remote sensing also replaces costly and slow data collection on the ground, ensuring the process that there is no interference with areas or objects. Among several application areas, remote sensing images have been used to assess sequestered carbon and observe a forest growth (Hunt et al., 2004; Lopatin et al., 2006; Miehle et al., 2006). Major developments in remote sensing allow to estimate forest growth. Remote sensing of vegetation may be important, being a basic setting for living beings. Small alterations can have many consequences on other living organisms and the biosphere.

Remote sensing of vegetation provides valuable information about the vegetation type, biophysical properties (e.g. LAI and biomass) and biochemical properties (e.g. chlorophyll and leaf nutrient concentration).
1. General introduction

which are used to understand ecosystem functions, vegetation growth, and nutrient cycling (Myneni et al., 1995, 2002; Yang et al., 2006a; le Maire et al., 2011). Therefore, vegetation plays a major role in global physical and biogeochemical processes and strongly regulates regional and global climate. This role is based on a simple structural unit, the leaf. The photosynthetic capacity of leaves in a forest controls primary productivity, climate, water and carbon gas exchange, and radiation extinction and is, therefore, a key component of physiological, climatological and biogeochemical processes in ecosystems (Nemani et al., 1993; Patenaude et al., 2008; Qin et al., 2008). In this sense, leaves, quantitatively determined as a LAI, have been extensively studied using ground based and remotely sensed optical Earth observations. LAI addresses the confounding role of vegetation as biophysical, biochemical, and radiation regime determinant parameter in our biosphere.

Several studies have been carried out to estimate forest growth parameters as the LAI from different satellite sensors. LAI estimation values can either be provided by a spaceborne product, in particular the Moderate Resolution Imaging Spectroradiometer (MODIS), or they can be derived using empirical relationships between ground measurements of LAI and spectral vegetation indices (SVIs) from satellite images, for instance the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER). The ASTER and MODIS sensors were launched in December 1999 on board the NASA EOS Terra satellite platform (Yamaguchi et al., 1998; Myneni et al., 2002).

MODIS is designed to provide long-term global observations every 1 to 2 days at moderate spatial resolutions (250 m-1 km). It has 36 channels from visible to thermal-infrared sensors with a field of view (FOV) of ±55°, and a scene width of 2230 km. MODIS LAI product (MOD15), is composited every 8 days at a 1 km resolution (Yang et al., 2006a). This is widely used due to good validation results (Yang et al., 2006a,b; Yuan et al., 2011). The low spatial resolution of the MODIS LAI product, however, limits the utility at local scales, which may require higher spatial resolution data.

ASTER provides a higher spatial resolution surface spectral reflectance, temperature, and emissivity data (Yamaguchi et al., 1998). It has a wide spectral region with 14 bands from the visible to the thermal infrared with high spatial, spectral and radiometric resolution. The spatial resolution varies with wavelength: 15 m in the visible and near infrared (VNIR, 0.52–0.86 μm), 30 m in the short wave infrared (SWIR, 1.6–2.43 μm), and 90 m in the thermal infrared (TIR, 8.1–11.6 μm). Each ASTER scene captures a 60 × 60 km area (Abrams, 2000). The estimation of LAI by ASTER sensor has been investigated in several studies at various spatial scales and environments (Heiskanen, 2006; Ito et al., 2007; Zheng and Moskal, 2009). This estimation has been based on empirical relationships formulated between the forest variables measured in the field and satellite data, often expressed in the form of SVIs. Current techniques for estimating LAI, however, cannot be uniform for all sites (Wang et al., 2004; Heiskanen, 2006).
1.4 Probabilistic graphical models

Statistics, the science of handling uncertainty, attempts to model order into disorder. It is usually recognized as an extremely powerful research tool (Cressie, 1993). Moreover, statistics can be defined as a mathematical analysis involving the use of quantified representations, models and summaries for a given set of empirical data or real world observations. Statistical methods and approaches have been used in several applications in the real world. A major advantage of statistics is the modeling of uncertainty. Uncertainty about the outcome of any process is, in this view of nature, solely a function of our own ignorance. The more we know, the less uncertain we are.

In recent years, graph theory has established itself as an important mathematical tool in a wide variety of subjects, ranging from operational research and chemistry to genetics and linguistics, and from electrical engineering and geography to sociology and architecture. At the same time it has emerged as a worthwhile mathematical discipline in its own right (Lauritzen, 1996). A basic advantage of graph theory is that it helps to simplify a complex environment which can be expressed in a straightforward and understandable frame by discovering the relationship between variables.

A natural way of dealing with two major problems, uncertainty and complexity, can be provided by graphical models. Graphical models draw upon probability theory and graph theory. They provide intuitive ways in which both humans and machines can model a highly interactive set of random variables, as well as complex data structures, to enable them to be logical and useful, and make valid inference from data. Graphical models have become popular in recent years in several applications (Kalacska et al., 2005; Taroni et al., 2006; Werhli and Husmeier, 2007; Pourret et al., 2008). In particular, graphical models that allow encoding joint probability distributions, so-called probabilistic graphical models, attract attention.

From a historical perspective, the earliest traces of using graphical representations of probabilistic information occurred in statistical physics (Gibbs, 1902) and genetics (Wright, 1921). Probabilistic graphical models can be divided into directed and undirected graphs (Lauritzen, 1996). Pearl (1988) introduced probabilistic directed graphical models, called originally belief networks, whereas later the term Bayesian networks (BNs) became predominant. Over the last decades, Bayesian networks have emerged as a popular model for dealing with uncertainty. A Bayesian network is a mathematical methodology of combining the graphics and probabilities to express the mutual relationship between variables.

The syntax and semantics of Bayesian networks will be covered through the coming chapters. Here I restrict myself to an informal exposition that is sufficient to further outline the subjects covered in this thesis. A Bayesian network is a directed, acyclic graph that repres-
1. General introduction

\textbf{1.5 Problem statement}

Biophysical parameter values such as LAI have proved useful in a number of environmental applications, including estimating carbon assimilation by plants. Long-time series of accurate and precise LAI estimation values are essential for climate change studies to improve the parameterization of the surface-atmosphere interaction processes in a range of models. Accurate estimates of LAI are important for functional-structural plant models, since the leaf area influences rates of evapotranspiration and photosynthesis of trees (Nemani et al., 1993; Villalobos et al., 1995).

LAI is one of the parameter outputs of the 3-PG forest growth model. As for all models the 3-PG model suffers from model inadequacy, the condition that the model does not represent the behavior of the real system in all details. In a similar vein, 3-PG model parameters suffer from parameter uncertainty, which can have many causes, including those that are intimately linked to the model itself, e.g. “light use efficiency” as a proximate measure of the photosynthetic process. These lead to uncertainty in the 3-PG model output, of an uncertain magnitude. In iterative models, where the model output from one iteration functions as input to the next iteration, the uncertainty is compounded as an error over time. Such an error in the prediction may become large and therefore the 3-PG model needs to adjust and correct its output among iterations. Satellite remote sensing is particularly useful in this context. The reflectance from the forest canopy is easily observed and it is related to the central biophysical processes in forest growth models, namely photosynthesis, and as such it is closely related to important output parameters from those models. The radiance received by the satellite
sensor is composed of different sources of radiation, which may be intractable (e.g. atmospheric effects, scattering), difficult to separate (e.g. mixed pixels), or inherent in the properties of the forest or its environment (e.g. slope, internal shading). Combined these introduce errors such that a direct correlation between at-sensor radiance and forest canopy properties is at best weak and often not adequate for model correction purposes. Moreover, satellite images often contain gaps (missing data) due to atmospheric circumstances, such as the presence of clouds, or to incomplete track spatial coverage.

To address these problems, statistical methods based on graphical models, in particular Bayesian networks have been used in this thesis. The integration is performed by combining satellite data with the 3-PG model, thus attaining a more accurate LAI output. The innovation of this research is two-fold. The first is theoretically framed, as this research focused on exploring a Bayesian network in order to integrate the output of the forest growth model and satellite data. The second is from the practical application as this research intended to improve forest growth estimates derived from the 3-PG model. In this sense, it will help scientists to use this model for further applications in forestry.

1.6 Research objectives

The main objective of the research is to construct a method to optimally integrate the output of the 3-PG model with satellite data. To achieve this aim, the following specific objectives have been addressed:

- To develop a methodology by which a Gaussian Bayesian network (GBN) can be used to adjust the output of the 3-PG model using a series of satellite images.
- To evaluate the performance of Gaussian Bayesian network modeling for forest growth estimates.
- To investigate how the Gaussian Bayesian network performs in updating the 3-PG model when there are gaps in the series of satellite images.
- To implement Gaussian Bayesian network with different satellite images and to modify Gaussian Bayesian network to infer the spatial estimation of forest growth estimates.
- To improve spatio-temporal of growth estimates of heterogeneous forests using Gaussian Bayesian network.

1.7 Application

1.7.1 Study site description

To pursue the objectives of this study and considering the required data, the study was applied to a Speulderbos forest. The Speulderbos forest
is located between 51° 96′ 06″ to 52° 38′ 00″ N and 05° 61′ 69″ to 06° 08′ 90″ E (Figure 1.1), planted in 1962, near the village of Garderen, The Netherlands. A 46 m high tower with a climate station, operated by the National Institute for Public Health and the Environment (RIVM), is situated at 52° 15′ 08″ N, 05° 41′ 25″ E, within a dense 2.5 ha Douglas fir (Pseudotsuga menziesii) stand. The dominant species in the neighborhood of the Douglas fir stand are the Japanese Larch (Larix kaempferi), Beech (Fagus sylvatica), Scotch Pine (Pinus sylvestris) and Hemlock (Tsuga spp). The tree density is about 785 trees ha$^{-1}$ and the average tree height was about 32 m in 2006 (Lemoine et al., 1991; Su et al., 2009).

The forest floor is covered with leaves on which little vegetation is present. The single-sided LAI varies between 8 and 11 throughout the year (Steingröver and Jans, 1995). The nearest edge is at a distance of 1.5 km southeast from the site. A small clearing of 1 ha is situated to the north of the climate station (Van Wijk et al., 2001).

1.7.2 Data description

The study is primarily based on: the fieldwork measurements, MODIS TERRA Land satellite collection 5 data (Chapters 2, 3, 4, and 6), and ASTER satellite (Chapters 5 and 6).
1.7.2.1 Ground data

Field data are collected on the ground at the observation tower of the Speulderbos forest, which is equipped with a weather station and various scientific instruments. The following data were captured as needed by the 3-PG model:

- **Climate data**: 16-days mean temperature, solar radiation, rainfall, and frost days.
- **Site factors**: site-latitude, maximum available water stored in the soil, soil fertility rating.
- **Initial conditions**: stem, root and foliage biomass, stocking and soil water at some time.
- **3-PG parameters**: parameters characterizing the modeled species.

In addition, the LAI was measured on the ground to validate the estimated LAI by our research. The LAI is indirectly measured from the canopy transmission by the inversion of the measurements of the photosynthetically active radiation (PAR) above and below the canopy. The PAR data are acquired from the Speulderbos forest using four sensors placed at the tower. Two are located at the top of the tower to record the PAR outside the canopy forest in two opposite directions: an upward looking sensor to measure the incoming sun radiation and a downward looking sensor to measure the reflected radiation from the canopy. The other two sensors are located below the forest canopy to record the PAR inside the forest. These sensors are also located in two opposite directions: an upward looking sensor to measure the radiation penetration through the canopy coming from the sun and a downward looking sensor to record the reflected radiation from the soil and the forest components. The PAR data are recorded every 10 minutes during daytime. The calculation of the LAI from the PAR data is based on the relationship between leaf area and the light transmittance, described by the Beer–Lambert model (Maass et al., 1995). This model utilizes the light canopy transmittance in the Beer-Lambert equation to assess the monthly LAI of a tropical deciduous forest. The daily average LAI ground data has been computed from the recorded PAR data after LAI calculated at solar noon time. Furthermore, to verify the spatial variation of the LAI, measurements were randomly taken around the tower using LAI-2000 Plant Canopy Analyzer (Li-COR) between the days 274 and 287 of 2008, since they were significantly correlated ($R^2=0.74$). As shown by the studies in the area (Steingröver and Jans, 1995), the ground LAI observations of Douglas fir trees in Speulderbos is adjusted by a clumping factor. The adjusted LAI ground data ($LAI_{FD}$) used through the coming chapters as a validation data of the LAI output by this research.

Based on the agreement of the temporal resolution of the MODIS images and 3-PG model, the $LAI_{FD}$ measurements were computed by averaging the LAI daily average every 16 days, which consolidated a set of 50 $LAI_{FD}$ values.
1. General introduction

LAI measurements by ground techniques, however, are affected by random errors and bias (Whitford et al., 1995; Coops et al., 2004; Jonckheere et al., 2004). The measurements of the PAR are usually concentrated around solar noon to avoid high solar angles relative to the zenith. Hence, an uncertainty of the PAR measurement arises as to whether the solar angles should be expressed relative to a normal to the inclined surface or to the solar zenith angle.

1.7.2.2 MODIS data

The low spatial resolution of LAI MODIS product limits the utility at local scales as in this study. To increase the resolution to 250 m, LAI derived from the 250 m the normalized difference vegetation index (NDVI) MODIS product (MOD13Q1), by establishing the relation between the NDVI and the FPAR. Hence, I referred to this as a modified MODIS LAI to distinguish it from the LAI MODIS product provided by NASA. The modified MODIS LAI will be used throughout the coming chapters and labeled as LAIM. A stronger relation exists between the FPAR and the NDVI than between the LAI and the NDVI (Myneni et al., 1995, 2002). A linear relationship between the NDVI and the FPAR is present, which is approximately linear for green vegetation (Prince and Goward, 1995; Steinberg and Goetz, 2009). This relationship is sensitive to soil background, irradiance quality and canopy structure.

I used the FPAR values from the LAI/FPAR 8-day L4 global 1-km (MOD15A2), and the NDVI values from the Vegetation Indices 16-day L3 global 250 m (MOD13Q1) as well. A linear relationship was found between the NDVI and the FPAR with $R^2 = 0.71$ and RMSE = 0.29. The linear relationship was defined by comparing 16-day FPAR composites to the 16-day MODIS NDVI products during period of the study time, and the 250 m LAI was calculated as (Norman et al., 1996)

$$LAI = -\frac{\ln(1 - FPAR)}{c}, \quad (1.1)$$

here FPAR has a 250 m spatial resolution derived from the 250 m NDVI based on the relationship equation between them and $c$ is the extinction coefficient.

1.8 Outline of the thesis

This thesis is a compilation of seven chapters. Besides the introduction and synthesis, five remaining chapters are based on manuscripts published in or submitted to scientific journals. The structure and content used in submitting the manuscript have been largely retained in the thesis. Therefore, overlaps and repetitions may occur between individual chapters. A relation between chapters (2-6) is shown in Figure 1.2. Following is the brief summary of each chapter.
Chapter 1 provides a general introduction to the thesis. The role of the forest and its growth level in carbon sequestration is presented briefly, which is most likely having an influence on the speed of climate change. How forest growth could be estimated using process-based forest growth model and satellite remote sensing is introduced. Research problems of accurate forest growth estimates are presented, and to address these problems the objectives are mentioned. The subsequent chapters are based on these objectives to achieve the required results.

Chapter 2 focuses on construction of Gaussian Bayesian network to integrate 3-PG model output and satellite data. Network-building steps have been explained in detail covering the network updating process. The Gaussian Bayesian network is implemented with available data almost of two years.

Chapter 3 examines the performance of Gaussian Bayesian network that has been built in Chapter 2. It shows the reliability, the applicability and the robustness of the Gaussian Bayesian network. The work in this chapter is subjected to an evidence propagation and sensitivity analysis of the input sources, i.e. the 3-PG output and satellite data, into the Gaussian Bayesian network.

Chapter 4 investigates the application of Gaussian Bayesian network with presence a gap in the input sources, i.e. satellite data. The EM-algorithm is formulated and integrated with Gaussian Bayesian network to estimate missing satellite data. For the purpose of this chapter, synthetic gaps were created in the satellite data time series of two years. Moreover, the performance of Gaussian Bayesian network is assessed before and after performing EM-algorithm.

Chapter 5 considers the applicability of the presented method in Chapter 4 in terms of using different satellite sensor and the presence of real missing satellite data. Moreover, the Gaussian Bayesian network is modified to improve spatial estimation of forest growth estimates. This is performed by integrating the spatial version of the 3-PG model with the satellite data within Gaussian Bayesian network.

Chapter 6 presents the application of Gaussian Bayesian network to improve estimation of growth estimates for a heterogeneous forest that show variation in space and time. An effort has been considered in this chapter regarding to the satellite images. Linear Mixture Models (LMMs) were used to decompose satellite pixels using class fraction derived from finer resolution.

Chapter 7 provides a synthesis of the obtained results, reflection of the current research, and untimely followed by recommendatory remarks on the direction and scope for further research.
Figure 1.2: A network, showing the connection and the relationship between thesis chapters. The link direction points to the reliance of child on its parents, for example the directed link from chapter 2 to chapter 3 indicates that chapter 3 relies on chapter 2 and that it evaluates the performance of chapter 2.
Bayesian network modeling for improving forest growth estimates

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>17</td>
</tr>
<tr>
<td>2.1.1</td>
<td>The 3-PG model</td>
<td>18</td>
</tr>
<tr>
<td>2.1.2</td>
<td>LAI estimation using MODIS satellite imagery</td>
<td>20</td>
</tr>
<tr>
<td>2.1.3</td>
<td>Bayesian network</td>
<td>20</td>
</tr>
<tr>
<td>2.2</td>
<td>Model description</td>
<td>21</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Initialization of BN</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Updating of BN</td>
<td>23</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Sensitivity of BN</td>
<td>26</td>
</tr>
<tr>
<td>2.3</td>
<td>Implementation</td>
<td>27</td>
</tr>
<tr>
<td>2.3.1</td>
<td>LAI frequency distribution</td>
<td>27</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Example application</td>
<td>28</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Results</td>
<td>31</td>
</tr>
<tr>
<td>2.4</td>
<td>Discussion</td>
<td>34</td>
</tr>
<tr>
<td>2.5</td>
<td>Conclusions</td>
<td>36</td>
</tr>
</tbody>
</table>

2. Bayesian network modeling for improving forest growth estimates

Abstract

Estimating the contribution of the forests to carbon sequestration is commonly done by applying forest growth models. Such models inherently use field observations such as leaf area index (LAI), whereas a relevant information is also available from remotely sensed images. This paper aims to improve the LAI estimated from the forest growth model (Physiological Principles Predicting growth (3-PG)) by combining these values with the LAI derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite imagery. A Bayesian networks (BNs) approach addresses the bias in the 3-PG model and the noise of the MODIS images. A novel inference strategy within the BN has been developed in this paper to take care of the different structures of the inaccuracies in the two data sources. The BN is applied to the Speulderbos forest in The Netherlands, where the detailed data were available. This paper shows that the outputs obtained with the BN were more accurate than either the 3-PG or the MODIS estimate. It was also found that the BN is more sensitive to the variation of the LAI derived from MODIS than to the variation of the LAI 3-PG values. In this paper, we conclude that the BNs can improve the estimation of the LAI values by combining a forest growth model with satellite data.

Keywords: Bayesian Networks (BNs), Leaf Area Index (LAI), Moderate Resolution Imaging Spectroradiometer (MODIS), Physiological Principles Predicting Growth (3-PG) model.
2.1 Introduction

Forests are traditionally important as sources of fuel, building materials, paper, fiber, and timber. More recently, their importance as major storehouses of carbon has been realized, as well as their capacity to exchange carbon dioxide ($\text{CO}_2$) between the vegetation and the atmosphere, which can affect the rate of climate change. Forest growth leads to carbon fixation, and thus, it may turn out to be an important way to diminish the rate of global warming. Modern silviculture may increasingly be able to regulate the rate of forest growth. The application of forest growth models could be helpful in managing the forests by quantifying the storage amounts of $\text{CO}_2$ and by comparing different management scenarios (Mohren and Van de Veen, 1995; Landsberg and Waring, 1997).

A commonly measured variable in the forests is the leaf area index (LAI). For broadleaf canopies, the LAI is defined as the one-sided green leaf area per unit of ground area, whereas for coniferous canopies, it is defined as the hemisurface needle leaf area. In both instances, it is equal to the ratio of the total upper leaf surface of the vegetation divided by the surface area of the land on which the vegetation grows. The LAI values typically range from 3 to 15, whereas in some cases, the LAI is greater than 15 (Schulze, 1982). The LAI can be used to scale the measurements of photosynthesis, transpiration, and light interception from the leaf level to the canopy level. This can ultimately be used as an indicator of forest growth. A major challenge is to make a representative estimate of the LAI for large forest covered areas.

Traditionally, three independent ways are distinguished to determine the LAI: by means of the measurements in the forest, by means of applying a forest growth model, and by means of a remote sensing image. The LAI is directly or indirectly measured in the field. Direct methods normally involve destructive sampling of canopy elements, and in large complex forest, it may be impossible to collect sufficient samples to accurately characterize the structure. Indirect methods include both light interception instrumentation and hemispherical photography (Jonckheere et al., 2004). Such measurements, however, are time consuming, in particular, when obtaining a long time series. Process-based models have been used to estimate the LAI. Several of these models have been established recently, such as the physiological principles predicting growth (3-PG) model developed by Landsberg and Waring (1997), the CABALA model (Mummery and Battaglia, 2004) and the ForGro model (Mohren and Van de Veen, 1995). The 3-PG model, which we focus in this study, is a process-based, stand-level model of forest growth. It requires site and climatic data as input, and it outputs the time-course of stand development in a form that is familiar to the forest manager, as well as the LAI, biomass pools, stand water use, and available soil water. Moreover, the model output and several 3-PG model parameters (e.g., the average monthly root turnover rate, the fraction of the net
2. Bayesian network modeling for improving forest growth estimates

primary production (NPP) to the roots, the fertility rating, the canopy conductance, and the ratio of soil organism biomass) contain uncertainties. Such uncertainty is compounded as an error over time in the model iterative, and therefore, error in one parameter affects the determination of the next parameter in the modeling chain (Esprey et al., 2004; Fontes et al., 2006). Similarly, modern remote sensing can also provide the LAI estimates. For instance, the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor provides eight-day global data sets of the LAI. Such images, however, have a relatively coarse resolution as compared to the size of the forest stand. MODIS imagery also has uncertainties as instrument noise exists during image acquisition, and atmospheric characteristics are constantly changing (Yang et al., 2006a,b).

Statistical methods combining images with forest models and observations may help to reduce uncertainties, thus attaining a more accurate LAI estimated value. Graphical models, particularly Bayesian networks (BNs), provide a promising way of dealing with this due to their ability to integrate different sources and to deal with uncertainty in probabilistic way.

A BN is a probabilistic methodology for combining graphs and probabilities to express the relationships between variables (Pearl, 1988). In a BN, each node represents a variable that has a value, while the relations between the nodes are expressed in terms of conditional probabilities. The input data may influence any of the nodes. Once the data are obtained and assigned to the node, a Bayesian mechanism propagates their values to the subsequent nodes of the network. A BN can be used for the combination of information coming from separate sources with varying degrees of reliability. The BNs have been successfully applied to molecular biology, particularly in genetics and biotechnology (Imoto et al., 2003; Needham et al., 2007; Werhli and Husmeier, 2007), forensic sciences (Taroni et al., 2006), computer science, image processing, and artificial intelligence studies (Pourret et al., 2008).

The objective of this study is to explore the possibilities of the BNs to improve LAI estimations by combining the 3-PG model output with MODIS images. Ultimately, this may be used as a predicting model of the LAI values. The study is applied to a forest area in The Netherlands (Speulderbos forest), where the detailed data are available for the calibration of the 3-PG model and for the validation of the BN.

2.1.1 The 3-PG model

The 3-PG model is a process-based stand-level model of forest growth. A full description of the 3-PG has been provided by Landsberg and Waring (1997), and Sands and Landsberg (2002), and a brief illustration of the model is shown in Figure 2.1. It requires basic silvicultural and site and climatic data as inputs, and predicts as outputs the time-course of the stand development in a form that is familiar to the forest manager. It further produces the LAI \( (LAI_{3PG}) \), biomass pools, stand water use and available soil water. The 3-PG model bridges the gap between con-
2.1. Introduction

Conventional empirical models on the one hand and process-based carbon balance models on the other hand. It consists of five simple submodels: assimilation of carbohydrates; distribution of biomass between foliage, roots, and stems; determination of stem number; soil water balance; and conversion of biomass values into variables of interest to forest managers. It can be applied to plantations, i.e., even-aged relatively homogeneous forests. Of interest in this study is the LAI$_{3PG}$, being an important indicator of the vegetation status and a key parameter in process-based models to quantify the exchange of matter and energy flow between the vegetation and the atmosphere. The 3-PG model has been used to estimate and predict the LAI in different areas (Sands and Landsberg, 2002; Fontes et al., 2006).

![Diagram of 3-PG model](image)

**Figure 2.1:** Basic structure part of the 3-PG and the causal influences of its variables and processes. The symbols used stand for gross primary production (GPP), net primary production (NPP), air temperature (T), vapour pressure deficit (VPD), water (H$_2$O), carbon dioxide (CO$_2$), and leaf area index (LAI).

The 3-PG model can be run for any number of years using monthly weather data for each year or monthly averages for the year. In this paper, we modified it to run in 16-day timesteps, matching with the temporal resolution of MODIS images. A major concern, however, is the quality of the output. Uncertainty in the LAI output of the 3-PG model exists as the LAI values may be modified by conditions or events that are difficult to model or to predict, such as droughts and flooding or pests and diseases. Several studies considered the uncertainty and variation of 3-PG model parameters (Sands and Landsberg, 2002; Almeida et al., 2004a). Esprey et al. (2004) reported results from a sensitivity analysis of the 3-PG model, in particular, concerning the LAI.
2. Bayesian network modeling for improving forest growth estimates

2.1.2 LAI estimation using MODIS satellite imagery

The estimation of the LAI from satellite imagery may serve as a proxy for field measurements at regional and global scales (Kalacska et al., 2005). Many studies have been carried out to estimate the LAI values from different satellite sensors, in particular, using the MODIS sensors onboard the Earth Observing System Terra/Aqua platforms (Yang et al., 2006a,b, 2007). The MODIS sensor produces a standard suite of global products characterizing the vegetation cover, the LAI, and the fraction of absorbed photosynthetically active radiation (FPAR) at the 1-km spatial resolution based on observations and composited over an eight-day period. Morissette et al. (2006) reviewed techniques employed in various countries to produce and evaluate the LAI products derived from satellite measurements, and Yang et al. (2006b) summarized the experience of several collaborating investigators on the validation of the MODIS LAI products. In this study, we derived LAI MODIS \( \text{LAI}_{\text{MODIS}} \) from the normalized difference vegetation index (NDVI) MODIS product, every 16 days at a 250 m spatial resolution (Section 1.7.2.2), thus matching the temporal frequency of the 3-PG model and ground data. Hence, we refer to this as a modified MODIS LAI to distinguish from the MODIS LAI product provided by NASA. Current techniques for estimating LAI often failed to provide consistent values. Estimating LAI from one satellite instrument MODIS is an ill-posed inversion problem because the number of unknowns is always larger than the available bands due to the nature of the Earth’s complex environment. Furthermore, most LAI\(_{\text{MODIS}}\) data products are not continuous in space and time because of a cloud contamination, and an insufficient number of data points for retrieval. As a result, LAI\(_{\text{MODIS}}\) products need significant improvements. For this reason, some methods for reducing noise and constructing high-quality MODIS time series for further analysis have been formulated (He, 2007). Qin et al. (2008) estimated the LAI from remote sensing data by taking advantage of the physical-model inversion. Xiao et al. (2009) designed a temporally integrated inversion method to produce spatially and temporally continuous LAI products with relatively higher quality. These methods, however, have their own strengths and limitations. Uncertainty of the LAI remains, whereas the BNs provide a framework to combine the model output with the satellite data to estimate the LAI values more precisely.

2.1.3 Bayesian network

A Bayesian network is a network consisting of nodes linked with directed arcs (arrows) that allow us to carry out probabilistic reasoning. It is a mathematical methodology of combining graphs and probabilities to express relationships between variables identified by the nodes. Links in the network are configured as a directed acyclic graph (DAG), i.e. a graph without feedback (Jensen and Nielsen, 2007). Each node in a BN indicates a random variable, which is either discrete or continuous. The directed arcs linking the nodes are usually expressed in kinship terms...
such as the parents of a node (nodes with arrows pointing directly to this node), the children of a node (nodes with arrows pointing from this node), and the ancestors (parents at a higher level). These arcs are quantified by conditional probability tables (CPTs). Such networks can be designed using conventional scientific notions of cause and effect. For example, a node \( A \) with parents \( B_1, ..., B_n \) indicate the existence of an arrow from each of the \( B_i \)'s, \( i = 1, ..., n \), to \( A \). There does exist a CPT of \( \Pr(A \mid B_1, ..., B_n) \) (Pearl, 1988). Conditional dependence and independence of nodes in a graphical model allows for a substantial simplification of joint probabilities. For instance, two nodes \( A_1 \) and \( A_2 \) are conditionally independent, given a third set \( B \), if all paths between \( A_1 \) and \( A_2 \) are separated by a node \( B \) (Pearl, 1988; Jensen and Nielsen, 2007)

\[
\Pr(A_1 \mid A_2, B) = \Pr(A_1 \mid B).
\] (2.1)

### 2.2 Model description

#### 2.2.1 Initialization of BN

The main purpose of the BN, as used in this study, is to enhance the correspondence between the LAI values from the 3-PG model and MODIS estimate. An important assumption underlying the inference within the BN is that biophysical models aim to describe a physical process characterized by competition for resources and seasonality in as much detail as possible. Complications in this process, however, may cause a lack of understanding of the process or its parameters to support such a detailed description. This results in an approximation of the physical process only. Forest biophysical models invariably iterate for a long time (typically 40–120 years) at a weekly or monthly resolution, resulting in hundreds of iterations over the lifetime of the trees in the forest. Small errors in a single iteration quickly accumulate to a large bias even if the biophysical process is well described. Satellite imagery, on the other hand, has an unknown error associated with it, for instance, due to a suboptimal parametrization of the numerical analysis routine or to atmospheric conditions. The BN takes the different structures of the data quality into consideration when updating the LAI values to input them into the next iteration of the 3-PG model.

We start by building a simple network consisting of two nodes, first one refer to the LAI\(_{3\text{PG}}\) and second one refer to the LAI\(_M\). To serve the objective of this study, we introduce an intermediate node to combine LAI\(_M\) and LAI\(_{3\text{PG}}\), called LAI\(_{\text{BN}}\), where the subscript indicates the time step 1, taking values from both LAI\(_M\) and LAI\(_{3\text{PG}}\) (Figure 2.2(a)). This network contains two cliques of a child node with its parents, namely \{LAI\(_M\), LAI\(_{\text{BN}}\}\} and \{LAI\(_{3\text{PG}}\), LAI\(_{\text{BN}}\}\}. The connection type of Figure 2.2 is a converging connection type, defined as a BN node with two or more
2. Bayesian network modeling for improving forest growth estimates

The double ellipses, nodes, indicate that LAI is a continuous stochastic variable and follow the Gaussian distribution as illustrated in Section 2.3.1.

The quantitative part of the BN is a joint probability distribution written as a product of conditional probabilities that describes the dependences between the variables of the network. Given a set of variables $\text{LAI} = \{\text{LAI}_1, ..., \text{LAI}_n\}$, each with its parents, the joint probability distribution (JPD) of LAI is given by (Figure 2.2)

$$\Pr(\text{LAI}) = \prod_{i=1}^{n} \Pr(\text{LAI}_i | \text{pa}_i), \quad (2.2)$$

where $\text{pa}_i$ corresponds to the parent variables of $\text{LAI}_i$ (i.e., $\{\text{LAI}_{M_i}, \text{LAI}_{3PG_i}, \text{LAI}_{BN_i-1}\}$). A common type of a BN containing continuous variables is the Gaussian Bayesian network (GBN) (Shachter and Kenley, 1989; Geiger and Heckerman, 1994; Sullivant, 2008). It is a BN where the joint probability distribution of LAI is a multivariate Gaussian distribution $N(\mu, \Sigma)$.

The conditional probability distribution of $\text{LAI}_i$ that verifies the Equation (2.2), is a univariate Gaussian distribution with density

$$f(\text{LAI}_{BN_i} | \text{pa}_i) \sim N \left( \mu_i + \sum_{j=1}^{#\text{pa}_i} \beta_{ij}(\text{pa}_{ij} - \mu_{\text{pa}_ij}), \nu_i \right), \quad (2.4)$$

where $\mu_i$ is the expectation of $\text{LAI}_i$ at time $i$, the $\beta_{ij}$ are the regression coefficients of $\text{LAI}_{BN_i}$ on its parents, $#\text{pa}_i$ is the number of parents of $\text{LAI}_{BN_i}$, and $\nu_i$ is the conditional variance of $\text{LAI}_{BN_i}$ given its parents. Here $\Sigma_i$ is the unconditional variance of $\text{LAI}_{BN_i}$, $\Sigma_{i\text{pa}_i}$ are the covariances between $\text{LAI}_{BN_i}$ and its parents $\text{pa}_i$, and $\Sigma_{\text{pa}_i}$ is the covariance matrix of $\text{pa}_i$. For more details of the GBN concept, we refer to (Shachter and Kenley, 1989; Geiger and Heckerman, 1994), and for an illustrated example, we refer to (Castillo and Kjaerulff, 2003).

The conditional distribution (Equation (2.4)) is recovered by considering that

$$f(\text{LAI}_{BN_i} | \text{pa}_i) \sim N \left( \mu_i + \sum_{j=1}^{#\text{pa}_i} \beta_{ij}\text{pa}_{ij}, \nu_i \right), \quad (2.5)$$

Under the assumption of Gaussianity, the variable $\text{LAI}_{BN_i}$ is completely specified by its expectation $\mu$ and covariance matrix $\Sigma$, obtained from the conditional distribution.
2.2. Model description

Following (Figure 2.2(a)), we notice that at the first time step the marginal distribution of $LAI_{BN1}$ given its parents $(LAI_{M1}, LAI_{3PG1})$ yields the expectation and covariance as

\[
E(LAI_{BN1}) = \beta_{LAI_{BN1}LAI_{M1}} \mu_{LAI_{M1}} + \beta_{LAI_{BN1}LAI_{3PG1}} \mu_{LAI_{3PG1}} + \mu_{LAI_{BN1}}
\]

(2.6)

and

\[
\Sigma_{LAI_{BN1}} = \begin{pmatrix}
\sigma_{LAI_{M1}LAI_{M1}} & 0 & \sigma_{LAI_{M1}LAI_{BN1}} \\
0 & \sigma_{LAI_{3PG1}LAI_{3PG1}} & \sigma_{LAI_{3PG1}LAI_{BN1}} \\
\sigma_{LAI_{BN1}LAI_{M1}} & \sigma_{LAI_{BN1}LAI_{3PG1}} & \sigma_{LAI_{BN1}LAI_{BN1}}
\end{pmatrix}.
\]

(2.7)

Here, $E(.)$ refers to the expectation, and the subscript $LAI_{BN1}$ of $\Sigma_{LAI_{BN1}}$ refers to the covariance matrix at the first time step. Both $\sigma_{LAI_{BN1}LAI_{M1}}$ and $E(LAI_{BN1})$ contribute to obtain LAI at the second time step in the network (Figure 2.2). Hence the marginal distribution of $LAI_{BN1}$ is

\[
LAI_{BN1} \sim N \left( \beta_{LAI_{BN1}LAI_{BN1}} \mu_{LAI_{BN1}} + \beta_{LAI_{BN1}LAI_{3PG1}} \mu_{LAI_{3PG1}} + \mu_{LAI_{BN1}}, \Sigma_{LAI_{BN1}} \right).
\]

(2.8)

The estimated LAI values for $LAI_{3PG1}$ are obtained from ground truth values. At this stage we assume that $LAI_{BN1}$ follows the same distribution as $LAI_{3PG1}$ in order to initialize and implement the network with truth values of LAI. This assumption will be generalized when the network extended as shown in the next section.

2.2.2 Updating of BN

The previous subsection formulated the basis of a BN. Now, we extend the BN over multiple time steps. To do so, the value of $LAI_{BN1}$, after being
Bayesian network modeling for improving forest growth estimates

derived from the BN, is inserted into the 3-PG model to update the LAI value at its next iteration (Figure 2.2(b)). This graphical representation is similar to Figure 2.2(a) but it refers to the second time step. The arrow from LAI_{BN_1} to LAI_{BN_2} links the first time step to the second time step, indicating that the new LAI value in LAI_{BN_2} is conditionally determined by the previous LAI value LAI_{BN_1}. This system can be further extended towards more than 2 time steps (Figure 2.2(c)). To deal with LAI_{BN_i}, for \( i \geq 2 \) and the prior probability distribution of the LAI, the following equation applies:

\[
LAI_{BN_i} = (1 - \tau) \rho LAI_{M_i} + \tau LAI_{3PG_i} + (1 - \rho)(1 - \tau)LAI_{BN_{i-1}} \tag{2.9}
\]

where \( \tau \) and \( \rho \) are weighing values, defined as \( \tau = \frac{|LAI_{M_i} - LAI_{M_{i-1}}|}{LAI_{M_{i-1}}} \) and \( \rho = \frac{|LAI_{3PG_i} - LAI_{3PG_{i-1}}|}{LAI_{3PG_{i-1}}} \). They are proportional to the change in LAI values obtained from MODIS images and 3-PG output.

Equation (2.9) includes BN output from the previous time step (LAI_{BN_{i-1}}) to ensure that the LAI values in the new node are consistent with LAI_M images, LAI_{3PG}, and BN output (LAI_{BN}) at the precedent iteration. This is based on the assumption that LAI values do not change sharply in a short period of time. In fact, these choice for \( \tau \) and \( \rho \) addresses the deviation between LAI field data (LAI_{FD}), LAI_{3PG} values and LAI_M images at two consecutive time steps. Weighing these values as in Equation (2.9) reduces the impact of large discrepancies between LAI values of 3-PG and MODIS, as shown in Section 2.3.3.

In Figure 2.2(b), the distribution at node LAI_{BN_2} can be expressed as

\[
f(LAI_{BN_2} | LAI_{M_2}, LAI_{BN_1}, LAI_{3PG_2}) \sim N \left( \mu_{LAI_{BN_2}} + \beta_{LAI_{BN_2}LAI_{M_2}} LAI_{M_2} + \beta_{LAI_{BN_2}LAI_{BN_1}} LAI_{BN_1} + \beta_{LAI_{BN_2}LAI_{3PG_2}} LAI_{3PG_2}, \nu_{LAI_{BN_2}} \right) \tag{2.10}
\]

Now, the expectation of LAI_{BN_2} can be found as

\[
E(LAI_{BN_2}) = \mu_{LAI_{BN_2}} + \beta_{LAI_{BN_2}LAI_{M_2}} \mu_{LAI_{M_2}} + \beta_{LAI_{BN_2}LAI_{BN_1}} \mu_{LAI_{BN_1}} + \beta_{LAI_{BN_2}LAI_{3PG_2}} \mu_{LAI_{3PG_2}} \tag{2.11}
\]

and covariance can be found as

\[
\Sigma_{LAI_{BN_2}} = \begin{pmatrix}
\sigma_{LAI_{M_2}LAI_{M_2}} & 0 & 0 & \sigma_{LAI_{BN_2}LAI_{BN_2}} \\
0 & \sigma_{LAI_{BN_2}LAI_{BN_1}} & 0 & \sigma_{LAI_{3PG_2}LAI_{BN_2}} \\
0 & 0 & \sigma_{LAI_{3PG_2}LAI_{3PG_2}} & \sigma_{LAI_{BN_2}LAI_{BN_2}} \\
\sigma_{LAI_{BN_2}LAI_{M_2}} & \sigma_{LAI_{BN_2}LAI_{BN_1}} & \sigma_{LAI_{BN_2}LAI_{3PG_2}} & \sigma_{LAI_{BN_2}LAI_{BN_2}}
\end{pmatrix} \tag{2.12}
\]

Thus, the marginal distribution of LAI_{BN_2} is equal to
2.2. Model description

\[ \text{LAI}_{BN_2} \sim N \left( \mu_{LAI_{BN_2}} + \beta_{LAI_{BN_2}} \mu_{LAI_{BN_2}} + \beta_{LAI_{BN_2}} \mu_{LAI_{BN_1}} + \beta_{LAI_{BN_2}} \mu_{LAI_{BN_2}} \mu_{LAI_{BN_1}} \right) \]

\[ + \mu_{LAI_{BN_2}} \mu_{LAI_{3PG_2}} \cdot \Sigma_{LAI_{BN_2}} \right). \quad (2.13) \]

The previous equations for the expectation and variance of LAI\(_{BN}\) are considering two iterations. For \( n \) iterations, the network is as follows (Algorithm 1; Figure 2.3)

- For \( i = 1 \), it is assumed that the LAI\(_{3PG_1}\) and LAI\(_{BN_1}\) values are the observed ground values. This assures that the BN starts with realistic values. The expectation and variance for LAI\(_{BN_1}\) are calculated from Equations (2.6) and (2.7). The new LAI value is inserted into the 3-PG model to update the LAI value at the second time step. Next, the 3-PG model produces a new LAI value, represented as the node LAI\(_{3PG_2}\) in the network.

- For \( i \geq 2 \), we define LAI\(_{BN_i}\) as in Equation (2.9), and the distribution of LAI\(_{BN_i}\) is equal to

\[ f \left( \text{LAI}_{BN_i} \mid \text{LAI}_{M_i}, \text{LAI}_{M_{i-1}}, \text{LAI}_{3PG_i} \right) \]

\[ \sim N \left( \mu_{\text{LAI}_{BN_i}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{M_i}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{BN_{i-1}}} \mu_{\text{LAI}_{M_{i-1}}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{3PG_i}} \right) \]

\[ + \mu_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{3PG_i}} \cdot \nu_{\text{LAI}_{BN_i}} \right) \quad (2.14) \]

whereas the expectation and the variance of LAI\(_{BN_i}\) after the network propagation are

\[ E \left( \text{LAI}_{BN_i} \right) = \mu_{\text{LAI}_{BN_i}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{M_i}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{BN_{i-1}}} \mu_{\text{LAI}_{M_{i-1}}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{3PG_i}} \]

\[ + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{3PG_i}} \mu_{\text{LAI}_{3PG_i}} \]

\[ \Sigma_{\text{LAI}_{BN_i}} = \left( \begin{array}{cccc}
\sigma_{\text{LAI}_{BN_i}, \text{LAI}_{BN_i}} & 0 & 0 & \sigma_{\text{LAI}_{BN_i}, \text{LAI}_{BN_{i-1}}} \\
0 & \sigma_{\text{LAI}_{BN_{i-1}, \text{LAI}_{BN_{i-1}}}} & 0 & \sigma_{\text{LAI}_{BN_{i-1}, \text{LAI}_{3PG_i}}} \\
0 & 0 & \sigma_{\text{LAI}_{3PG_i}, \text{LAI}_{3PG_i}} & \sigma_{\text{LAI}_{BN_i}, \text{LAI}_{3PG_i}} \\
\sigma_{\text{LAI}_{BN_i}, \text{LAI}_{BN_{i-1}}} & \sigma_{\text{LAI}_{BN_{i-1}, \text{LAI}_{BN_{i-1}}}} & \sigma_{\text{LAI}_{BN_{i-1}, \text{LAI}_{3PG_i}}} & \sigma_{\text{LAI}_{3PG_i}, \text{LAI}_{3PG_i}}
\end{array} \right) \].

\[ (2.16) \]

To obtain \( \Sigma_{\text{LAI}_{BN_i}} \), \( i \geq 1 \), the algorithm presented by Shachter and Kenley (1989) is used. Hence, the marginal distribution of LAI\(_{BN_i}\) is

\[ \text{LAI}_{BN_i} \sim N \left( \mu_{\text{LAI}_{BN_i}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{M_i}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{BN_{i-1}}} \mu_{\text{LAI}_{M_{i-1}}} + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{3PG_i}} \right) \]

\[ + \beta_{\text{LAI}_{BN_i}} \mu_{\text{LAI}_{3PG_i}} \mu_{\text{LAI}_{3PG_i}} \cdot \Sigma_{\text{LAI}_{BN_i}} \right). \quad (2.17) \]
2. Bayesian network modeling for improving forest growth estimates

New LAI values obtained as output from the BN, \( \text{LAI}_{BN} \), are compared with \( \text{LAI}_{FD} \) to assess the agreement between them. This new BN output is inputted into the 3-PG model to update the LAI value, whereas the 3-PG model produces new LAI values, represented as a node \( (\text{LAI}_{3PG_{i+1}}) \) in the network.

**Algorithm 1:** BN Implementation

**input**: Iteration number \( (N) \), \( \text{LAI}_{M} \), \( \text{LAI}_{3PG} \).

**output**: Updated values of forest growth parameter “intermediate BN node”.

**Initial step:** for \( i \leftarrow 1 \);

\( \text{LAI}_{BN} \) node \( \leftarrow f(\text{LAI}_{M} \) node, \( \text{LAI}_{3PG} \) node);

\( \text{LAI}_{BN} \) data \( \leftarrow \text{LAI}_{3PG} \) data \( \leftarrow \) field data;

after the BN propagates

\[
\text{LAI}_{BNi} \sim N \left( \mu_{\text{LAI}_{BNi}} + \beta_{\text{LAI}_{BNi}} \text{LAI}_{M} \mu_{\text{LAI}_{M}} + \beta_{\text{LAI}_{BNi}} \text{LAI}_{3PGi} \mu_{\text{LAI}_{3PGi}} \Sigma_{\text{LAI}_{BNi}} \right);
\]

Update the 3-PG model with \( \text{LAI}_{BNi} \) (which represents \( \text{LAI}_{3PG_{i+1}} \) in the network);

for \( i \leftarrow 2 \) to \( N \) do

\( \text{LAI}_{BNi} \) node \( \leftarrow f(\text{LAI}_{M} \) node, \( \text{LAI}_{BNi-1} \) node, \( \text{LAI}_{3PG} \) node);

calculate \( \text{LAI}_{BNi} \) data as Equation(2.9);

after the BN propagates

\[
\text{LAI}_{BNi} \sim N \left( \mu_{\text{LAI}_{BNi}} + \beta_{\text{LAI}_{BNi}} \text{LAI}_{BNi-1} \mu_{\text{LAI}_{BNi-1}} + \beta_{\text{LAI}_{BNi}} \text{LAI}_{3PGi} \mu_{\text{LAI}_{3PGi}} \Sigma_{\text{LAI}_{BNi}} \right);
\]

Statistical validation;

if \( \text{LAI}_{BNi} \approx \text{LAI}_{FD} \) then

the BN output is close to field observation;

else

Update 3-PG model with \( \text{LAI}_{BNi} \) (which represents \( \text{LAI}_{3PG_{i+1}} \) in the network);

end

end

2.2.3 Sensitivity of BN

Several methods have been proposed to measure the accuracy of the LAI derived from satellite imagery. Verger et al. (2008) evaluated the performance of the LAI derived from multisource images and their mutual agreement. Uncertainty in the LAI values derived from the MODIS images is introduced by factors like atmospheric variation, and functioning of the sensor. Similarly, the reliability of the 3-PG estimates depends on the accuracy of its input parameters. A full modeling of the uncertainty is
2.3 Implementation

Figure 2.3: Workflow process of improving the LAI value.

Beyond the frame of this study. However, we assess the sensitivity of the BN when the input values $LAI_{M}$ and $LAI_{3PG}$ are varied in steps of 0.25 units within the interval $[-1, 1]$. The sensitivity is also assessed when $LAI_{M}$ and $LAI_{3PG}$ values are varying simultaneously.

2.3 Implementation

The BN approach is applied to the Speulderbos forest in The Netherlands (Figure 1.1), where the ground and satellite data (Section 1.7.2) were available for a period of 26 months from July 2007 to September 2009. Moreover, the parameter values that used to implement the 3-PG model in this study were Douglas fir parameters, and were obtained from (Coops and Waring, 2001; Waring and McDowell, 2002).

2.3.1 LAI frequency distribution

Two tests of normality (Shapiro-Wilk and Lilliefors) were applied in order to verify the frequency distribution of the LAI values in each one of the MODIS images and the 3-PG model.

2.3.1.1 Modified MODIS LAI

The normality tests of the LAI values are taken in terms of spatial and temporal resolution. The tests were first applied to a single $LAI_{M}$ image for the whole study area (16 pixels with 250 m spatial resolution). Second, the tests were applied to the pixels at one location at successive $LAI_{M}$ images during the full study period (July 2007 until September 2009 at a 16-day temporal resolution). The Q-Q plots and the histograms of the $LAI_{M}$ values (Figure 2.4(a) and (b)) show some deviations from normality, but according to the tests of normality, the $LAI_{M}$ data does not deviate from a normal distribution (Table 2.1 (a)).
2. Bayesian network modeling for improving forest growth estimates

Table 2.1: Two normality tests. (a) Normality tests for the $\text{LAI}_M$. Column (A) is composed of 16 pixels covering the study area, and column (B) is a series of a single pixel followed from 50 $\text{LAI}_M$ images during 50 observation times. (b) Normality tests for the $\text{LAI}_3\text{PG}$. Column (A) is composed of 16 $\text{LAI}_3\text{PG}$ simulated values, and column (B) is a series of 50 output values during time series (50 iterations)

(a)

<table>
<thead>
<tr>
<th>Test</th>
<th>(A)</th>
<th>(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk P-value</td>
<td>0.09626</td>
<td>0.1714</td>
</tr>
<tr>
<td>Statistic (W)</td>
<td>0.9049</td>
<td>0.9684</td>
</tr>
<tr>
<td>Lilliefors P-value</td>
<td>0.1072</td>
<td>0.3825</td>
</tr>
<tr>
<td>Statistic (D)</td>
<td>0.1948</td>
<td>0.0883</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Test</th>
<th>(A)</th>
<th>(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk P-value</td>
<td>0.99</td>
<td>0.039</td>
</tr>
<tr>
<td>Statistic (W)</td>
<td>0.98</td>
<td>0.89</td>
</tr>
<tr>
<td>Lilliefors P-value</td>
<td>0.98</td>
<td>0.21</td>
</tr>
<tr>
<td>Statistic (D)</td>
<td>0.086</td>
<td>0.02</td>
</tr>
</tbody>
</table>

2.3.1.2 LAI output of 3-PG model

A test of normality has next been applied to the 16 LAI output values from the 3-PG model that was run at the same spatial and temporal resolution as the $\text{LAI}_M$ images. Although Figure 2.5(a) and (b) shows that these values deviate from normality, these values were not significantly different from normality (Table 2.1 (b)) according to the implemented tests.

2.3.2 Example application

We run the BN during 50 iterations, corresponding to a time period of 26 months, and we consider a homogeneous area of one km² around the climate station in the Speulderbos forest (Figure 1.1 (c)). Such an area is covered with 16 LAI values of the modified MODIS LAI at 250 m spatial resolution. We observe LAI output values from the 3-PG model, which was executed simultaneously in the BN for 50 iterations, also producing 16 values per iteration for 16-day periods. The MODIS and the 3-PG model were thus harmonized. The first two steps of the BN are presented here as explained in Algorithm 1, and the reminder iterations are similar to the second iteration. The results are shown in Figure 2.6 as follows

2.3.2.1 First iteration

The first iteration initializes the BN in its three nodes $\text{LAI}_M$, $\text{LAI}_3\text{PG}$, and $\text{LAI}_3\text{PG}$. The data of nodes $\text{LAI}_3\text{PG}$ and $\text{LAI}_3\text{PG}$ are the same and are ac-
2.3. Implementation

Figure 2.4: Two normal Q-Q plots and histograms with the normal curve of LAI$_m$. (a) Sixteen pixels covering the study area. The Q-Q plot and histogram of one LAI MODIS observation for 16 pixels. (b) Series of a single pixel followed from the 50 LAI$_m$ images during 50 observation times. The Q-Q plot and histogram of the 50 LAI MODIS images for one pixel.

quired from the LAI field data. Hence, LAI$_{3PG1}$ = LAI$_{BN1}$ $\sim$ N(7.47, 1.04), and LAI$_{M1}$ $\sim$ N(4.81, $7 \times 10^{-5}$).

The conditional distribution of LAI$_{BN1} |$ LAI$_{M1}$, LAI$_{3PG1}$ is equal to

$$\hat{\mu}_{LAI_{BN1}|LAI_{M1},LAI_{3PG1}} = \hat{\beta}_{LAI_{BN1},LAI_{M1}} LAI_{M1} + \hat{\beta}_{LAI_{BN1},LAI_{3PG1}} LAI_{3PG1} + \hat{\mu}_{LAI_{BN1}}$$

$$= 2.77 \times 10^{-15} LAI_{M1} + 1 \times LAI_{3PG1} + \hat{\mu}_{LAI_{BN1}} \quad (2.18)$$

where $\hat{\beta}_{LAI_{BN1},LAI_{M1}}$ and $\hat{\beta}_{LAI_{BN1},LAI_{3PG1}}$ are the regression coefficients relating LAI$_{BN1}$ to LAI$_{M1}$ and LAI$_{3PG1}$, respectively. The regression coefficients are close to zero or one at this first iteration step but will deviate from zero and one in subsequent iterations as the bias in the 3-PG model leads to an accumulation of error. The marginal distribution of LAI$_{BN1}$ is hence found by implementing the network with the data in Equations (2.6) and (2.7). We thus have
2. Bayesian network modeling for improving forest growth estimates

\[ E(LAI_{BN1}) = 7.44 \times 10^{-15} (4.81) + 7.47 + 7.47 = 14.94 \]  \hspace{1cm} (2.19)

\[ \Sigma_{LAI_{BN1}} = \begin{bmatrix} 0.007 & 0 & 1.7 \times 10^{-17} \\ 0 & 1.04 & 1.04 \\ 1.7 \times 10^{-17} & 1.04 & 2.08 \end{bmatrix}. \]  \hspace{1cm} (2.20)

Then, the distribution of node \( LAI_{BN1} \) after BN propagation is equal to \( LAI_{BN1} \sim N(14.94, 2.08) \)

### 2.3.2.2 Second iteration

The new LAI value \( LAI_{BN1} \) is inserted into the 3-PG model to calculate the development of the forest over the next time step. This produces the
LAI value at the next period. At this time step, the network consists of four nodes, namely, LAI\(_{M2}\), LAI\(_{BN1}\), LAI\(_{3PG2}\) and LAI\(_{BN2}\). The probability distribution of new LAI\(_{3PG2}\) after receiving an updated LAI and after running the 3-PG model is LAI\(_{3PG2}\) ∼ N(15.39, 0.0006), whereas LAI\(_{M2}\) ∼ N(4.51, 0.013), LAI\(_{BN1}\) ∼ N(14.94, 2.08), and LAI\(_{BN2}\) ∼ N(4.77, 0.016) according to Equation (2.9). Therefore, the expectation and variance of LAI\(_{BN2}\) | LAI\(_{M2}\), LAI\(_{BN1}\), LAI\(_{3PG2}\) are

\[
\hat{\mu}_{\text{LAI}_{BN2} | \text{LAI}_{M2}, \text{LAI}_{BN1}, \text{LAI}_{3PG2}} = \hat{\mu}_{\text{LAI}_{BN2}} + \hat{\beta}_{\text{LAI}_{BN2} \text{LAI}_{M2}} \text{LAI}_{M2} + \hat{\beta}_{\text{LAI}_{BN2} \text{LAI}_{BN1}} \text{LAI}_{BN1} + \hat{\beta}_{\text{LAI}_{BN2} \text{LAI}_{3PG2}} \text{LAI}_{3PG2}
\]

and the covariance can be found as

\[
\Sigma_{\text{LAI}_{BN2}} = \begin{pmatrix} 0.013 & 0 & 0 & 0.013 \\ 0 & 2.08 & 0 & 0.88 \\ 0 & 0 & 0.0001 & 1.2 \times 10^{-5} \\ 0.013 & 0.8 & 1.2 \times 10^{-15} & 0.029 \end{pmatrix}
\]

Therefore, the value of LAI output of the BN at this moment is distributed as LAI\(_{BN2}\) ∼ N(9.41, 0.029).

By observing the difference of this last value with the LAI field observation, we get | 7.49 – 9.41 | = 1.92, whereas the difference between the 3-PG LAI output, when 3-PG executed alone, and the field observation is equal to | 7.49 – 8.59 | = 1.1.

At this time step, the LAI\(_{BN}\) is still far from the validation data LAI\(_{FD}\), whereas after a few iterations, LAI\(_{BN}\) converges more rapidly than LAI\(_{3PG}\) (see Figure 2.6). Later iterations are similar to the second iteration.

The BN have been implemented using the C++ code. Initially, the 3-PG model produces 16 LAI values that are combined with the LAI\(_{M}\) values in the BN, which introduces the expectation and variance of the LAI.

### 2.3.3 Results

Figure 2.6 shows the LAI values estimated from the BN for a period of 26 months, along with the LAI derived from MODIS images, the 3-PG model, and LAI field data. The accuracy of the LAI 3-PG output is tested using the Root Mean Square Error (RMSE) and the Relative Error (RE) rate with respect to the LAI field data. We found an RMSE of 2.71 and an RE of 37.4%. On average, the LAI\(_{3PG}\) was 9.62, while the LAI\(_{FD}\) is 7.05 (Table 2.2). The 3-PG overestimated LAI values across the studied period.

The modified MODIS LAI shows a high deviation with respect to the LAI\(_{FD}\), with an RMSE and an RE of 3.26 and 44.1%, respectively (Table 2.2). Overall, for this study, the average LAI\(_{M}\) of 3.93 indicates an underestimation of the LAI values with respect to the LAI\(_{FD}\). Conversely, as shown in Section 2.3.2, after the first iteration, the BN yields more accurate LAI estimates than the 3-PG model and the MODIS images.
Figure 2.6: LAI distribution of the Speulderbos forest obtained from four sources, namely, the field data ($LAI_{FD}$), the 3-PG model ($LAI_{3PG}$), the MODIS images ($LAI_M$), and the BN ($LAI_{BN}$), during the period July 2007–September 2009.

Table 2.2: Mean values (Mean), standard deviation ($\sigma$), Root Mean Square Errors (RMSE) and Relative Errors (RE) for the various ways to estimate the LAI.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>$\sigma$</th>
<th>RMSE</th>
<th>RE rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LAI_{FD}$</td>
<td>7.05</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LAI_M$ data</td>
<td>3.93</td>
<td>0.92</td>
<td>3.26</td>
<td>44.1%</td>
</tr>
<tr>
<td>$LAI_{3PG}$ data</td>
<td>9.62</td>
<td>0.65</td>
<td>2.71</td>
<td>37.4%</td>
</tr>
<tr>
<td>$LAI_{BN}$ data</td>
<td>7.75</td>
<td>1.07</td>
<td>1.57</td>
<td>14.7%</td>
</tr>
</tbody>
</table>

(Table 2.2). For the 26 months period, we notice that the combination of the MODIS image and the 3-PG in a BN reduces the RMSE to 1.57 and the RE to 14.7%.

Figure 2.7 shows the output of the sensitivity test implemented by independently varying the $LAI_{3PG}$ and $LAI_M$ values. As shown in the Figure, the BN is more sensitive to the variations in the $LAI_M$ than to the variations of the $LAI_{3PG}$ values. Hence, for an added variation of the LAI to the MODIS LAI, the resulting error of the BN is larger than when adding the same variation to the 3-PG estimates. Furthermore, as shown in Figure 2.8, large errors of the LAI are obtained when the $LAI_M$ and the $LAI_{3PG}$ are simultaneously varied. We found a maximum error of 0.74 with $LAI_M + 1$ and $LAI_{3PG} - 1$. 

32
Figure 2.7: Sensitivity test of the BN after independently varying the $LAI_{3PG}$ and $LAI_M$ values in steps of 0.25 units within the interval $[-1, 1]$, across the studied period applied in the Speulderbos forest.

Figure 2.8: Sensitivity test of the BN after simultaneously varying the $LAI_{3PG}$ and $LAI_M$ values in steps of 0.25 units within the interval $[-1, 1]$, over the studied period applied in the Speulderbos forest.
2. Bayesian network modeling for improving forest growth estimates

2.4 Discussion

In this study, a methodology is presented to improve estimates of forest growth based on the manipulation of the LAI values using a BN that combines the LAI output of the biophysical 3-PG model with the LAI derived from satellite imagery. We used the BN to infer the updated LAI values from those predicted by the 3-PG model and the values reported by the modified MODIS LAI. As shown in Figure 2.6, the deviation of the BN output and ground measurement is lower than the deviation between the 3-PG model output and the ground measurement, indicating that the LAI output of the BN is more accurate than that of the 3-PG model output alone and is closer to the ground measurement. The high LAI values from the 3-PG (Figure 2.6) are due to an uncorrected simulation. Moreover, a sensitivity test of the 3-PG’s parameters could help the BN in improving the LAI estimates. The 3-PG is executed using a specific set of species and site parameters from 1962, when the forest was planted, to 2007, when we started implementing the BN, to generate the initial conditions of the foliage, root, and stem biomass. These initial conditions were used as inputs into the 3-PG model and were combined with the MODIS images when using the BN.

The strength of the presented methodology lies in the use of two sources of information and a combination of these in a BN to improve the estimate of the LAI values. It addresses two long-standing issues in biophysical modeling and satellite image analysis. In biophysical modeling, uncertainties in the model and its parameters lead to an accumulation of error over time (multiple iterations), which give the model output a bias. This bias can be reduced by refinement and proper parameterization of the model, but often, the available data of the biophysical system are not accurate enough to refine the model or to reduce the uncertainty in the parameters. In satellite remote sensing, there are varying instrument and environmental conditions that lead to errors, which are difficult to correct unless detailed local terrain and atmosphere conditions are available. The combination of these two sources of information provides a mechanism to reduce the bias in the biophysical model by integrating a data source that has no long-term bias but an error that is evaluated in the inference mechanism of the BN.

A major contribution of this study lies in the combination of satellite data with the 3-PG model within a BN. As a strategy for the consideration of these two products in a BN node, we resorted to the mathematical formulation in Equation (2.9). From this equation, we can identify the intermediate node of the BN based on the contribution of all of the MODIS images, 3-PG output, and precedent BN output. However, to account for the uncertainty in both MODIS images and 3-PG model, weighing factors are introduced. This new expression also includes the BN output of the precedent iteration due to the fact that the LAI values have no large changes within a relatively short period of time.

Regarding the influence of the input sources to the BN accuracy (Figure 2.6), we found out that the BN is more sensitive to the variation in
2.4. Discussion

The LAIM than to the variation in the LAI3PG (Figure 2.7). This indicates that LAIM has a significant influence on the BN. Hence, the BN proposed in this work gives more weight to LAIM (Equation (2.9)), but it is sensitive as well to LAI3PG, when they vary simultaneously (Figure 2.8). Prices LAIM as an input may help to get prices LAIBN output. Improving the accuracy of the method when the input sources are less sensitive is an area for future studies.

Also, some other issues require further work. For instance, the spatial resolution of the MODIS images (250 m) is insufficient in representing the spatial variability and distribution of trees within the forest. This brings additional uncertainty into the LAI estimation, which could be reduced by using finer resolution satellite imagery such as ASTER at 15–30 m resolution. Such spatial refinement would be particularly useful in assessing the variability in the reflectance and, thus, in the associated LAI values in smaller forest patches. Furthermore, the uncertainties in the example application (Section 2.3.2) that may result by the neighborhood of the 16 pixels have an impact on the resulting output of the BN. This may be addressed with a subpixel remote-sensing approach. Lastly, the multisource image fusion technique could improve the BN output so that the LAI is estimated more frequently.

In addition, the seasonal changes of the LAIBN (Figure 2.6) are not clear and smooth, whereas the modified LAIM shows a good agreement with the expected seasonal changes of the LAI through the year. As shown in Figure 2.6, the LAIBN increased rapidly in May 2008 and 2009, indicating that the LAIBN during the growing season follows the modified LAIM, due to the weighing Equation (2.9). The LAIBN after a few iterations, however, drops, showing unexpected seasonal growth. This inappropriate BN output is due to the fact that LAIBN is affected by the high LAI3PG values. Apparently, the LAIBN estimates reduce the uncertainties of the 3-PG output. Furthermore, the BN needs a long time series until the LAIBN gets close to the LAIFD and represents the changes in the seasonal growth. Figure 2.6 shows that the LAIBN started far from the LAIFD and converged to the LAIFD only after more than 7 months, corresponding to 13 iterations.

Regarding the applicability of our approach, the proposed BN requires satellite imageries and field data to estimate the LAI. The remote sensing data are often not available or of unknown quality due to atmospheric characteristics such as the presence of clouds and aerosols. Likewise, ground observations may be partly or entirely absent as some areas are difficult to reach or the instruments for field survey are too expensive for the national forest survey institutes. The use of the BNs, however, can be useful in dealing with this problem, which will be considered in further research.

In this work we selected a homogeneous forest area. This assumption has been verified after a significant correlation was found between the measured LAI at and around the tower. Homogeneity, however, rarely occurs within the forests. Nonhomogeneity of the forests is difficult to address when extracting the biophysical parameters from remote-
sensing images, particularly for relatively small areas. This may prohibit the extension of our method toward a more general applicability.

Spatial data are becoming increasingly important for forest vegetation management and decision making. Better estimations of the biophysical parameters provide forest managers information on forest growth, which may help in getting a better understanding of the forest and which ultimately can serve as an informative factor in climate change. The forest canopy data (e.g., vegetation parameters like LAI) play a major role in the simulation of the surface energy balance and, therefore, weather and climate prediction. It may also be useful for the forest scientists and the Intergovernmental Panel on Climate Change to apply it in directly or indirectly assessing the carbon stored in the forest, which is one of the important current issues in climate change mitigation and adaptation strategies.

2.5 Conclusions

This study presented a Bayesian network for improving the LAI estimation by combining the 3-PG output with that derived from satellite images. Some concepts of the BN are introduced and summarized. A new equation, Equation (2.9), is defined according to the relative error of the MODIS images and the 3-PG output. We have illustrated the framework by implementing the BN for the Speulderbos forest in The Netherlands. The study leads to the following conclusions.

- A BN is able to integrate the LAI values that come from different sources into a single reasoning framework.
- By using the satellite data, the output of the 3-PG forest growth model is improved. It closely matches the mean forest growth estimate, and it substantially reduces both the RMSE and the RE.
Performance evaluation of Gaussian Bayesian network modeling for forest growth estimates

3.1 Introduction 39
3.2 Materials and methods 41
  3.2.1 Source of uncertainties 41
  3.2.2 Gaussian Bayesian network 41
  3.2.3 Evidence propagation in a GBN 43
  3.2.4 Sensitivity analysis in a GBN 44
3.3 Implementation 47
3.4 Experimental results 47
  3.4.1 GBN performance with respect to the LAI_{3PG} 48
  3.4.2 GBN performance with respect to the LAI_{M} 50
3.5 Discussion 51
3.6 Conclusion 54

3. Performance evaluation of GBN modeling for forest growth estimates

Abstract

Leaf area index (LAI) is an important parameter to assess carbon sequestration. Several models have been established to estimate forest growth based on LAI. Recently, these models have been improved by including satellite image information. This study evaluates the performance of a Gaussian Bayesian network (GBN) for such integrated modeling. Evidence propagation and a sensitivity analysis are applied to the Physiological Principles Predicting Growth (3-PG) model combined with Moderate Resolution Imaging Spectroradiometer (MODIS) images. The study uses data collected during 26 successive months, split into two different seasons twice. Evidence propagation by means of the 3-PG model improved the GBN output, showing that the relative error of GBN output with respect to the field data decreased by 2.0%. This improvement is stronger than propagating the same evidence through MODIS images as the relative error of GBN output increased by 4.5%, and GBN is sensitive to the MODIS images as well. We conclude that the GBN is a robust network, being more robust with respect to the 3-PG model than with respect to the MODIS images.

Keywords: Gaussian Bayesian networks (GBNs), Evidence propagation, Leaf area index (LAI), Moderate Resolution Imaging Spectroradiometer (MODIS), Physiological Principles Predicting Growth (3-PG) model, Sensitivity analysis.
3.1 Introduction

Estimation and monitoring of forest growth has been on the agenda of forestry research in recent years. The reason is that trees play a major role in CO₂ sequestration and that they contribute to the reduction of carbon emissions to the atmosphere (Wamelink et al., 2009). The leaf area index (LAI) defined as one half the total leaf surface area per unit ground surface area projected on the local horizontal datum (m². m⁻²) (Black et al., 1991), is often used to monitor forest growth.

Three ways are distinguished to determine the LAI: (1) measurements in the field, (2) output of a forest growth model, and (3) observations from remote sensing products. LAI is measured in the field using direct or indirect methods. Direct methods are labor intensive and involve destructive sampling. Indirect methods include the use of ground-based optical instruments by measuring the transmission of radiation through the canopy, making use of the radiative transfer theory (Jonckheere et al., 2004). Ground measurements are time-consuming, in particular if long-term monitoring of spatial and temporal dynamics of leaf area development are required.

LAI is modeled using process-based model, such as the Physiological Principles Predicting Growth (3-PG) model (Landsberg and Waring, 1997). This requires the setting of specific parameters, site and climatic data as inputs (Table 3.1). Among several forest structural and physical variables, the 3-PG model also provides estimates of LAI based on foliage mass (tDM . ha⁻¹) and specific leaf area (SLA, m². kg⁻¹) (Landsberg and Waring, 1997). The 3-PG model has been used extensively to estimate and predict these variables in several areas (Nole et al., 2009; Coops et al., 2010).

Table 3.1: The minimum data required to run 3-PG model.

<table>
<thead>
<tr>
<th>Climate data</th>
<th>16-days mean of temperature, solar radiation, rainfall, and frost days.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site factor</td>
<td>site-latitude, maximum available water stored in the soil, soil fertility rating.</td>
</tr>
<tr>
<td>Initial conditions</td>
<td>stem, root and foliage biomass, stocking and soil water at some time.</td>
</tr>
<tr>
<td>3-PG Parameters</td>
<td>canopy quantum efficiency, temperature optimum and limits for photosynthesis, ratio of autotrophic respiration to gross photosynthesis, maximum leaf and canopy conductances, specific leaf area, maximum annual litterfall, litter and root turnover rates, soil fertility ranking, parameters for allometric equations with stem diameter.</td>
</tr>
</tbody>
</table>

Satellite remote sensing provides extensive spatial information of forest regions at different scales (Yang et al., 2006b; Hill et al., 2006). In particular, the LAI and the fraction of photosynthetically active radiation (FPAR) is supplied as an operational product of Moderate Resolution Imaging...
3. Performance evaluation of GBN modeling for forest growth estimates

ging Spectroradiometer (MODIS) onboard EOS (Earth Observation System) Terra and Aqua satellites (Barnes et al., 1998) with 1-km resolution and at 8-day intervals (Yang et al., 2006b). The operational MODIS algorithm ingests up to seven atmosphere-corrected surface spectral bi-directional reflectance factors (BRFs) and their uncertainties and provides pixel-based values of LAI, FPAR together with their respective dispersions. The algorithm, described in (Knyazikhin et al., 1999), uses look-up-tables (LUT) to achieve inversion of the three-dimensional radiative transfer problem. When this approach fails to provide a solution, a back-up method based on relations between the normalized difference vegetation index (NDVI), LAI/FPAR, and a biome map is used (Knyazikhin et al., 1999).

LAI estimates derived from the 3-PG model and MODIS images overestimate or underestimate the true value, hence are not error-free (Hill et al., 2006; Xenakis et al., 2008). The potential use of LAI to determine forest structure and biomass has motivated numerous investigations in this area. An alternative approach recently proposed the use of statistical methods based on graphical models, in particular Gaussian Bayesian networks (GBNs). A GBN is able integrating 3-PG model output and MODIS images to achieve more accurate LAI estimation (Mustafa et al., 2011b).

The performance of a GBN has to be evaluated considering the quality of its outputs and the mathematic relationships between them, in order to determine the reliability, the applicability and the robustness of the GBN. To assess robustness, a GBN has to be subjected to a sensitivity analysis which investigates the effects of varying models parameters on its output (Kullback and Leibler, 1951). In such an analysis the sensitivity of output values has to be determined in the light of both the inaccuracies of input data and the uncertainty of the evidence fed to the network (Castillo et al., 1997; Gómez-Villegas et al., 2007). A sensitivity analysis of a GBN thus considers the propagation of evidence in the network and then evaluates the accuracy of its output. In this setup, evidence is considered as ground truth or true values, whereas the evidence propagation refers to the computation of the updated conditional distribution of the variables in a GBN, given a set of evidence to the network. As a consequence, this sensitivity analysis informs how modifications of GBN parameters influence the posterior marginal or conditional probabilities after introducing evidence or knowledge in the network (Gómez-Villegas et al., 2007).

The aim of this work is to extend the work of Mustafa et al. (2011b) towards evaluating the performance of a GBN for LAI modeling. This is achieved by addressing both an evidence propagation, and a sensitivity analysis of the input sources, i.e., the 3-PG output and satellite data. A similar approach to study the sensitivity analysis proposed (Gómez-Villegas et al., 2007) is used in this work, with a concrete application in the Speulderbos forest in The Netherlands where sufficient data are available. In the context of this study, a thorough assessment of the sensitivity of the proposed GBN aims to ultimately achieve a better understanding of the potential of GBNs in the modeling of forest growth,
3.2 Materials and methods

3.2.1 Source of uncertainties

Inaccurate LAI values may originate from inaccurate field data, uncertain model output, and approximate satellite observation. LAI ground measurements may provide the best estimation method. Most of these measurements, however, are prone to uncertainties due to measurement error, location and scale, weather characteristics, or to approximate method calculation. Calculations of LAI field measurement are achieved by using a mathematical formula like the Beer-Lambert formula (Maass et al., 1995). This formula requires an estimate of canopy light extinction efficiency ($c$), being one of the uncertainty sources in calculating LAI.

Inaccuracy in the 3-PG model output can be expected because the model invariably iterates for a long time, simulating a period of typically 40 to 120 years, at a weekly or monthly resolution resulting in hundreds of iterations over the lifetime of the trees in the forest. Small errors in a single iteration quickly accumulate, even if the biophysical process is well described. Therefore, many of the 3-PG model’s parameters such as the average monthly root turnover rate, the fertility rating, canopy conductance and the ratio of soil organism biomass require a close attention, and they essentially affect primary production and stem growth.

Similarly, inaccuracy in MODIS images can be expected due to atmospheric characteristics such as the presence of clouds or aerosols, the instrument sensor, mixed-pixel effect, instantaneous field of view (IFOV) (Cracknell, 1998) and the MODIS LAI algorithm (Yang et al., 2006b). In addition, inaccuracy is caused by the spatial resolution of MODIS images (Sprintsin et al., 2007). Moreover, other factors related to the spatial resolution could increase uncertainties of MODIS images, as averaging LAI MODIS pixels over many pixels.

Uncertainties occurring in any of the estimated LAI of the 3-PG model ($\text{LAI}_{3\text{PG}}$) or the derived LAI of MODIS images ($\text{LAI}_{\text{MODIS}}$) might influence the GBN output. Accordingly, we assessed the robustness of GBN estimates with occurred uncertainties of the estimated LAI by 3-PG and MODIS images.

3.2.2 Gaussian Bayesian network

GBNs are defined as Bayesian networks (BNs) where the joint probability density of $X = (X_1, X_2, \ldots, X_n)^T$ is a multivariate Gaussian distribution $N(\mu, \Sigma)$ with $\mu$ the $n$-dimensional mean vector and $\Sigma$ the $n \times n$ positive definite covariance matrix using a directed acyclic graph (DAG) to repres-
3. Performance evaluation of GBN modeling for forest growth estimates

![Diagram of GBN for LAI](image_url)

**Figure 3.1**: The GBN for \( i^{th} \) iterations, \( i \geq 1 \), \( LAI_i \) indicates the network at time \( i \). It consists of triple nodes \( LAI_{3PG} \), \( LAI_{BN} \), and \( LAI_M \) obtained from the 3-PG model, the GBN and a MODIS image, respectively. The link between intermediate nodes indicates that the new \( LAI_{BN} \) value at iteration \( i + 1 \) is conditionally determined by the \( LAI_{BN} \) value at iteration \( i \).

The dependence structure of the variables (Jensen and Nielsen, 2007; Gómez-Villegas et al., 2007).

As in BNs, the joint density can be factorized using the conditional probability densities of \( X_i \), \( i = 1, \ldots, n \) given its parents in the DAG, \( pa_i \subseteq \{X_1, \ldots, X_{i-1}\} \), such as:

\[
Pr(X) = \prod_{i=1}^{n} Pr(X_i|pa_i). \tag{3.1}
\]

Hereafter, we present a GBN for the variable \( X = LAI \). Hence, for a set of \( n^{th} \) consecutive LAI values, \( LAI = \{LAI_1, \ldots, LAI_n\} \) (Figure 3.1), Equation (3.1) can be applied as

\[
Pr(LAI) = \prod_{i=1}^{n} Pr(LAI_i|pa_i), \tag{3.2}
\]

where \( i \) refers to the time step and each \( LAI_i = \{LAI_{3PG}, LAI_{BN}, LAI_M\} \) represents a set of triple nodes of the LAI values estimated from MODIS, GBN, and 3-PG forest model, respectively.

In a GBN, the conditional probability distribution of the \( LAI_i \) represented by \( LAI_{BN} \) as the variable of interest given its parentage, is the univariate Gaussian distribution with density

\[
f(LAI_{BN}|pa_i) \sim N\left(\mu_i + \sum_{j=1}^{pa_i} \beta_{ij}(pa_{ij} - \mu_{pa_{ij}}) , \nu_i\right), \tag{3.3}
\]
3.2. Materials and methods

where \( \mu_i \) is the expectation of \( LAI_{BN_i} \) at time \( i \), the \( \beta_{ij} \) are the regression coefficients of \( LAI_{BN_i} \) on its parents, \( \#pa_i \) is the number of parents of \( LAI_{BN_i} \), and \( \nu_i = \Sigma_i - \Sigma_{pa_i} \Sigma_{pa_i}^{-1} \Sigma_i \) is the conditional variance of \( LAI_{BN_i} \) given its parents. Further, \( \Sigma_i \) is the unconditional variance of the \( LAI_{BN_i} \), \( \Sigma_{pa_i} \) are the covariances between \( LAI_{BN_i} \) and the variables \( pa_i \), and \( \Sigma_{pa_i} \) is the covariance matrix of \( pa_i \).

For \( i \geq 2 \) and based on the graphical part of GBN in Figuer 3.1, the conditional distribution of \( LAI_{BN_i} \) is equal to

\[
LAI_{BN_i} \sim N \left( \mu_{LAI_{BN_i}} + \beta_{LAI_{BN_i},LAI_{M_i}} (LAI_{M_i} - \mu_{LAI_{M_i}}) + \beta_{LAI_{BN_i},LAI_{BN_{i-1}}} \right),
\]

where \( \mu_{LAI_{M_i}}, \mu_{LAI_{3PG_i}}, \mu_{LAI_{BN_{i-1}}} \) and \( \mu_{LAI_{BN_i}} \) are the expectation values before network propagation of \( LAI_{M_i}, LAI_{3PG_i}, LAI_{BN_{i-1}}, \) and \( LAI_{BN_i} \) respectively. \( \Sigma_{LAI_{BN_i}} \) is the covariance matrix.

The posterior probability of \( LAI_{BN_i} \) in Equation (3.4) is conditionally determined by three parents \( LAI_{M_i}, LAI_{3PG_i}, \) and \( LAI_{BN_{i-1}} \). These parents are used as prior input information into the GBN, where every \( LAI_{M_i} \) consists of sixteen pixels of LAI value derived from MODIS images with 250 m resolution. Also every \( LAI_{3PG_i} \) consists of sixteen simulated LAI output from the 3-PG model, and every intermediate node \( LAI_{BN_i} \) consists of sixteen LAI values. Based on the contribution of all of the MODIS images, the 3-PG output, and the precedent GBN output, the prior information of the intermediate node is assumed and defined as

\[
LAI_{BN_i} = \rho (1 - \tau) LAI_{M_i} + \tau LAI_{3PG_i} + (1 - \rho)(1 - \tau) LAI_{BN_{i-1}},
\]

where \( \tau \) and \( \rho \) are the weighing values, defined as \( \tau = \frac{|(LAI_{M_i} - LAI_{M_i_{i-1}})|}{LAI_{M_{i-1}}} \) and \( \rho = \frac{|(LAI_{3PG_i} - LAI_{3PG_i_{i-1}})|}{LAI_{3PG_{i-1}}} \). They are proportional to the change in the LAI values obtained from MODIS images and 3-PG output. The values of \( \tau \) and \( \rho \) may vary from 0 to a large number (more than 1) if the difference between two consecutive moments of LAI are big value. In reality, however, is not true to find a big difference of LAI values for the same study area within a short period of time (16-days). Moreover, this prior information \( (LAI_{M_i} \) and \( LAI_{3PG_i} \) have been tested on Gaussianity. For more details, we refer to (Mustafa et al., 2011b).

3.2.3 Evidence propagation in a GBN

Evidence propagation is updating of the probability distribution of the network variables by introducing information about the state of one or more variables, known as evidence variables. Evidence is represented in the form of values for the variables, for example, based on an observation or an assumption.
In GBNs, evidence propagation on one variable updates the probability distributions of the other variables in the network. Evidence propagation calculates the posterior probability of a variable of interest after inserting evidence into the network through the variable (Jensen and Nielsen, 2007). Based on the evidence definition, the LAI field data \((LAI_{FD})\) will be used in this study as an evidence values in GBN.

Different algorithms have been proposed for evidence propagation in GBNs. We use the propagation model presented by Castillo et al. (1997). This is based on computing the conditional probability density of a multivariate Gaussian distribution given the set of evidential variables. To perform the evidence propagation in GBNs we consider a partition of the set of variables, as \(X = (E, X \setminus E)'\), where \(X = LAI_i, E \subset X\) being the set of evidential variables, and \(X \setminus E\) is the set of variables of interest. Let \(\bar{e}\) be the evidence about the variables \(E\). After performing evidence propagation into the network, the conditional probability distribution of the LAI variables given the evidence \(\bar{e}\) is assumed to follow a multivariate Gaussian distribution (Castillo et al., 1997; Mustafa et al., 2011b). This can be expressed mathematically as,

\[
LAI_{BN_i} | E = \bar{e} \sim N(\mu_{LAI|E=\bar{e}}, \sigma_{LAI|E=\bar{e}}),
\]

where \(\mu_i\) and \(\mu_e\) are the means of the \(LAI_{BN_i}\) and the variable that received the evidence \(\bar{e}\); respectively, before the evidence propagation in the GBN. Further, \(\sigma_{ii}\) and \(\sigma_{ee}\) are the variances of \(LAI_{BN_i}\) and the variable that received the evidence \(\bar{e}\); respectively, before propagating the evidence and \(\sigma_{ie}\) is the covariance between \(LAI_{BN_i}\) and the variable that received the evidence \(\bar{e}\) before the evidence propagation. Here \(\bar{e}\) in Equation (3.6) represents the average value of LAI evidence values. More details of \(\bar{e}\) values are considered in the coming section.

### 3.2.4 Sensitivity analysis in a GBN

A sensitivity analysis investigates the effects of inaccuracies in the parameters of a mathematical model. It aims to study how variation in model output can be appointed, qualitatively or quantitatively, to different sources of data (Morgan et al., 1990).

Knowledge of the forest growth output of a GBN, in particular its dependence on parameter values, gives information of the ways that parameters changes (evidence) affect forest growth estimates value. These parameters are associated with the 3-PG model, the MODIS images and field measurements. As LAI is a continuous variable, and we assumed a
Gaussianity distribution (Mustafa et al., 2011b), the associated parameters are the mean and the variance. The mean could change due to the way of deriving or estimating the LAI, or due to ways of calculating the mean. These are affecting the variance values as well. The question is whether the parameters may change and to which degree outcomes are sensitive to this change. It is also appropriate to ask whether LAI\textsubscript{3PG} or LAI\textsubscript{M} has the strongest influence on GBN output. This is addressed by a sensitivity analysis.

A key question with respect to a GBN is, what the effect is of changing a parameter \( p_j = (\mu_i, \mu_e, \sigma_{ii}, \sigma_{ee}, \sigma_{ie}) \). Here the subscripts \( i \) and \( j \) refer to the iteration number of the variable of interest (LAI\textsubscript{BN}), and the changed parameter \( j = 1, ..., 5 \), i.e. inaccurate parameter, respectively. Hence, the sensitivity analysis consists of five elements, which are considered in this study as follows

1. Change in \( \mu_i \), i.e. the mean value of the LAI\textsubscript{BN}. This variable contains uncertainty which is transmitted either through 3-PG model or through MODIS images. It may influence the GBN output during successive iterations. Perturbation equals \( \delta_1 = lv_i - \mu_i \), where \( lv_i \) is either the expected value of LAI based on expert knowledge or the evidence value at moment \( i \).

2. Change in \( \mu_e \) i.e. the mean value of the evidence variable. This change addresses an inaccurate evidence value due to the uncertainty factors associated with it, as illustrated in Section 3.2.1. The perturbation may equal

\[
\delta_2 = \begin{cases} 
\hat{\mu}_e - le_i & \text{if } |\hat{\mu}_e - le_i| \leq |Max_e - le_i| \\
Max_e - le_i & \text{otherwise}
\end{cases}
\]

where \( le_i \) is the LAI value at the moment \( i \) during a season period of the variable that received evidence before evidence propagates in the network, and \( \hat{\mu}_e \) and \( Max_e \) are the average and the maximum value during a season period, respectively, of the evidence variable before evidence propagation.

3. Change in the covariance of the variable of interest (\( \sigma_{ii} \)). Uncertainty in the LAI\textsubscript{BN} variance could exist due to accumulate uncertainties of input sources to GBN. In this case, a restriction condition \( \delta_3 > -\sigma_{ii} + \frac{\sigma_{ee}^2}{2} \) ensures to keep the covariance matrix positive definite (Gómez-Villegas et al., 2007). Here \( \sigma_{ii} \) and \( \sigma_{ee} \) are the covariance of the LAI\textsubscript{BN} and the evidence variable before evidence propagates in the network respectively, \( \sigma_{ie} \) is the covariance of LAI\textsubscript{BN} and the evidence variable before evidence propagation. Based on the restriction condition and to avoid a high deviation value of covariance matrix values, we set the perturbation factor equal to

\[
\delta_3 = \frac{(Max_{\sigma_{ii}} - \sigma_{ii}) + (Min_{\sigma_{ii}} + \sigma_{ii})}{2},
\]

where \( Max_{\sigma_{ii}} \) and \( Min_{\sigma_{ii}} \) are the maximum and the minimum of \( \sigma_{ii} \), respectively, observed from the GBN output before evidence propagation during a season period.
3. Performance evaluation of GBN modeling for forest growth estimates

4. Change in the covariance of the evidence variable ($\sigma_{ee}$). This change could happen due to uncertainty of the structural calculation or estimation of the evidence variable. Based on the positive sign of the covariance matrix, the restriction condition at this case is represented as $\delta_4 \geq -\sigma_{ee}(1 - \beta_{ie}^2)$ (Gómez-Villegas et al., 2007), where $\sigma_{ee}$ is the covariance of evidence variable before evidence propagation and $\beta_{ie}$ is the regression coefficient relating the variable $LAI_{BNi}$ to the evidence variable. We select the perturbation factor as $\delta_4 = (\text{Max } \sigma_{ee} - \sigma_{ee}) + (\text{Min } \sigma_{ee} + \sigma_{ee})^2$, where $\text{Max } \sigma_{ee}$ and $\text{Min } \sigma_{ee}$ are the maximum and the minimum value of $\sigma_{ee}$, respectively, observed from GBN output before evidence propagation during a season period.

5. Change in the covariance between the variable of interest and the evidence variable ($\sigma_{ie}$). Uncertainty could occur to $\sigma_{ie}$ due to the presence of uncertainty in the regression coefficient value between the variables of interest and the evidence variable. The uncertain values of the regression coefficient may occur due to the uncertain relationship of data of the variable of interest and the data of the evidence variable. Again to keep the positiveness definite sign of the covariance matrix, the restriction condition proposed as $(-\sigma_{ie} - \sqrt{\sigma_{ii} \sigma_{ee}}) \leq \delta_5 \leq (-\sigma_{ie} + \sqrt{\sigma_{ii} \sigma_{ee}})$ (Gómez-Villegas et al., 2007). Then the perturbation factor has been chosen as $\delta_5 = (\text{Max } \sigma_{ii} - \sqrt{\sigma_{ii} \sigma_{ee}})^2 + (\text{Min } \sigma_{ii} + \sqrt{\sigma_{ii} \sigma_{ee}})$, where $\sigma_{ii}$ and $\sigma_{ee}$ are the covariance of $LAI_{BNi}$ and the evidence variable respectively, and $\sigma_{ie}$ is the covariance between $LAI_{BNi}$ and the evidence variable before evidence propagation.

We carried out a sensitivity analysis to determine the degree to which each input variable into the GBN influenced the outcome variables. The proposed sensitivity analysis algorithm of Gómez-Villegas et al. (2007) is used, based on the Kullback-Leibler divergence (KLD) being a most common measures for comparing probability distributions (Kullback and Leibler, 1951):

$$\text{KLD}(f(w), f'(w)) = \int_{-\infty}^{\infty} f(w)\ln\frac{f(w)}{f'(w)} \, dw,$$

(3.7)

where $f(w)$ and $f'(w)$ are two probability densities with the same support. Such a sensitivity analysis for a GBN compares network output of two different models: the original model $N(\mu, \Sigma)$ and the perturbed model obtained after adding a perturbation $\delta_j$ to the $j^{\text{th}}$ parameter of the model. For a GBN, let $f(x_i | e)$ be the marginal density of interest after evidence propagation and $f(x_i | e, \delta_j)$ the density after adding the perturbation $\delta_j$ to the $j^{\text{th}}$ parameter of the initial model. Then, the sensitivity measure is equal to

$$S_{ij}^p\left(f(x_i | e), f(x_i | e, \delta_j)\right) = \int_{-\infty}^{\infty} f(x_i | e)\ln\frac{f(x_i | e)}{f(x_i | e, \delta_j)} \, dx_i, \quad (3.8)$$

46
where the subscript $i$ indicates the iteration number and $\delta_j$ the perturbation values. For the $j^{th}$ parameter $p_j$, the new value of the parameter defined as $p_j^{\delta_j} = p_j + \delta_j$.

For small $S_j^{\delta_j}$ values, this then may lead to the conclusion that the GBN is robust against a perturbation. For more details of the sensitivity analysis method in GBNs we refer to (Gómez-Villegas et al., 2007).

### 3.3 Implementation

This study is applied to the Speulderbos forest in The Netherlands where detailed field data and satellite data were available (Section 1.7.2). The study area has been assumed to be covered by a homogeneous forest of 1 km$^2$ around the climate station in the Speulderbos forest (Figure 1.1). A full description of the study area is given in Section 1.7.1.

The time period contains two winter seasons (October–March; 2008 & 2009), and two summer seasons (May–August; 2008 & 2009), Figure 2.6. We studied the performance of the GBN, based on its output, by inserting evidence into the GBN and observing a better matching of the GBN with the field data. The evidence value is represented as $\bar{e}$ value as in Equation (3.6), which refers to the average value over 16 values of the evidence value obtained from $LAI_{FD}$ (Mustafa et al., 2011b). The sensitivity analysis achieved after the evidence provided to the network.

The evidence propagation and the sensitivity analysis are considered for $LAI_{3PG}$, and $LAI_{M}$ during two seasons. We started carrying out evidence propagation and sensitivity analysis for both winter and summer season. These seasons have been chosen due to their influence to the accuracy of LAI values that estimated by 3-PG (Dye et al., 2004), and that derived from MODIS (Tian et al., 2004; Kalacska et al., 2005).

For each season, the sensitivity measure is checked at the central moment of the season. The methodology was implemented in C++ code. The accuracy of $LAI_{BN}$ is assessed using the root mean square error (RMSE) and the relative error (RE) with respect to the $LAI_{FD}$ after and before evidence propagates into the network as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (LAI_{BN_i} - LAI_{FD_i})^2}{n}},$$

and

$$\text{RE} = \left| \frac{LAI_{BN} - LAI_{FD}}{LAI_{FD}} \right| \times 100\%.$$

### 3.4 Experimental results

The averages, maximum and minimum expectation values of $LAI_{BN}$, $LAI_{3PG}$, and $LAI_{FD}$, and maximum and minimum variance values of $LAI_{BN}$, $LAI_{3PG}$, and $LAI_{FD}$ are found from Mustafa et al. (2011b), summarized in Tables 3.2a & b, respectively.
3. Performance evaluation of GBN modeling for forest growth estimates

Table 3.2: (a) Expectation and (b) variance values of average (Avg), maximum (Max), minimum (Min), and LAI value at middle time period (AtM*) of the season of all of LAI$_{BN}$, LAI$_{3PG}$, and LAI$_{FD}$ for each season.

<table>
<thead>
<tr>
<th>Season Period</th>
<th>Case</th>
<th>(a) Expectation values</th>
<th>(b) Variance values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LAI$_{BN}$</td>
<td>LAI$_{3PG}$</td>
</tr>
<tr>
<td>Winter 2008</td>
<td>AtM</td>
<td>8.20</td>
<td>8.35</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>8.02</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>9.42</td>
<td>9.59</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>6.47</td>
<td>6.64</td>
</tr>
<tr>
<td>Summer 2008</td>
<td>AtM</td>
<td>7.90</td>
<td>8.15</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>8.27</td>
<td>8.74</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>9.64</td>
<td>10.24</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>7.19</td>
<td>7.46</td>
</tr>
<tr>
<td>Winter 2009</td>
<td>AtM</td>
<td>7.73</td>
<td>7.90</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>6.99</td>
<td>7.06</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>7.78</td>
<td>7.75</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>6.05</td>
<td>6.44</td>
</tr>
<tr>
<td>Summer 2009</td>
<td>AtM</td>
<td>6.44</td>
<td>7.41</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>6.89</td>
<td>7.35</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>7.67</td>
<td>8.19</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>6.27</td>
<td>6.71</td>
</tr>
</tbody>
</table>

*The value of middle time period of winter and summer season is identified at day 1 and 192, respectively.

3.4.1 GBN Performance with respect to the LAI$_{3PG}$

3.4.1.1 Winter 2008 & 2009

The LAI$_{BN}$ after performing evidence propagation approximates LAI$_{FD}$ more rapidly than before evidence propagation. This can be observed in Figure 3.2a, where in the first winter the LAI$_{BN}$ approximates rapidly the LAI$_{FD}$. The deviation between LAI$_{BN}$ and LAI$_{FD}$ after evidence propagation becomes lower than before propagating the evidence as is shown in the second winter season. We found an RMSE of the LAI$_{BN}$ after and before evidence propagation to be equal to 1.49 and 1.57, respectively. Moreover, the RE of the LAI$_{BN}$ after and before evidence propagation is 12.4% against 14.7%. It shows that the LAI$_{BN}$ after evidence propagation is closer to the LAI$_{FD}$ than before evidence propagation. Therefore, getting evidence to LAI$_{3PG}$ from LAI$_{FD}$ during the winter seasons improves the GBN output.

Next, we apply a sensitivity analysis of the GBN at the middle time period of the season after inserting evidence to LAI$_{3PG}$. The sensitivity analysis is carried out for every winter season separately. The sensitivity measure values after LAI$_{3PG}$ has received evidence from field data ($\bar{e}=6.76, \bar{\bar{e}}=7.49$) at the first (day 1 of 2008) and the second (day 1 of 2009) winter season, respectively are shown in Table 3.3.

Table 3.3 shows the sensitivity measure values $S_{Pj}$ after inserting evidence to LAI$_{3PG}$ at the middle time period of the first and the second winter season are small (maximum of 0.88 and minimum of $8.7 \times 10^{-3}$).
3.4. Experimental results

Figure 3.2: LAI values obtained from various source. LAI\textsubscript{BN} before and after evidence propagation in GBN through LAI\textsubscript{3PG} during (a) the winter seasons, (b) the summer seasons.

Table 3.3: The $\delta_j$ value, LAI\textsubscript{BN} distribution after add the perturbation values to GBN, and sensitivity measure ($S^p_j$) value after LAI\textsubscript{3PG} has received evidence in the 1\textsuperscript{st} & the 2\textsuperscript{nd} winter season.

<table>
<thead>
<tr>
<th>$\delta_j$ value</th>
<th>LAI\textsubscript{BN} distribution</th>
<th>$S^p_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st}</td>
<td>2\textsuperscript{nd}</td>
<td>1\textsuperscript{st}</td>
</tr>
<tr>
<td>$p_1$</td>
<td>-0.059</td>
<td>0.12</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.26</td>
<td>-0.15</td>
</tr>
<tr>
<td>$p_3$</td>
<td>2.8 \times 10^{-3}</td>
<td>2.9 \times 10^{-3}</td>
</tr>
<tr>
<td>$p_4$</td>
<td>2.9 \times 10^{-4}</td>
<td>2.8 \times 10^{-4}</td>
</tr>
<tr>
<td>$p_5$</td>
<td>-2.3 \times 10^{-7}</td>
<td>-1.7 \times 10^{-8}</td>
</tr>
</tbody>
</table>

Thus leads to conclude that the GBN is not sensitive to the perturbations.

3.4.1.2 Summer 2008 & 2009

After inserting evidence to LAI\textsubscript{3PG} from the field data and propagating the network, we notice that the LAI\textsubscript{BN} comes closer to the LAI\textsubscript{FD} than without evidence. The deviation between LAI\textsubscript{BN} and LAI\textsubscript{FD} reduces after evidence propagation as shown in Figure 3.2b. The RMSE and the RE of LAI\textsubscript{BN} after and before evidence propagates equals 1.48 against 1.57 and 12.6% against 14.7%, respectively. It emphasizes that the LAI\textsubscript{BN} with evidence propagates better matches with LAI\textsubscript{FD} than without evidence. Hence, getting evidence to LAI\textsubscript{3PG} during the summer season serves as an indicator to improve the GBN output.

Next, a sensitivity analysis of the GBN after inserting evidence to LAI\textsubscript{3PG} is applied at the middle time period of the season period. The sensitivity measure values of the GBN after LAI\textsubscript{3PG} has received evidence from the field data ($\bar{e} = 7.95$, $\bar{e} = 6.39$) at the first (day 192 of 2008) and the second (day 192 of 2009) summer season, respectively, are shown in Table 3.4.
3. Performance evaluation of GBN modeling for forest growth estimates

Table 3.4: The $\delta_j$ value, $LAI_{BN}$ distribution after add the perturbation values to GBN, and sensitivity measure ($S^{p_j}$) value after $LAI_{3PG}$ has received evidence in the 1st & the 2nd summer season.

<table>
<thead>
<tr>
<th>$\delta_j$</th>
<th>$LAI_{BN}$ distribution</th>
<th>$S^{p_j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^1_1$</td>
<td>$N(7.92, 1.6 \times 10^{-3})$</td>
<td>$N(6.27, 0.01)$</td>
</tr>
<tr>
<td>$p^2_2$</td>
<td>$N(7.86, 1.6 \times 10^{-3})$</td>
<td>$N(6.32, 0.01)$</td>
</tr>
<tr>
<td>$p^1_3$</td>
<td>$2.8 \times 10^{-3}$</td>
<td>0.03</td>
</tr>
<tr>
<td>$p^1_4$</td>
<td>$9.2 \times 10^{-4}$</td>
<td>$5.9 \times 10^{-3}$</td>
</tr>
<tr>
<td>$p^1_5$</td>
<td>$-6.6 \times 10^{-5}$</td>
<td>$-1 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 3.4 indicates that the sensitivity measure values $S^{p_j}$ after inserting evidence to $LAI_{3PG}$ at the middle time period of first and second summer season are small (maximum of 0.76 and minimum of 8.9 $\times 10^{-4}$). Therefore, the GBN is not sensitive to perturbations made at this season. Sensitivity measure values $S^{p_j}$ of the five parameters perturbed at the two different seasons winter and summer after $LAI_{3PG}$ has received evidence are shown in Figure 3.3.

Figure 3.3: Sensitivity measure values ($S^{p_j}$) obtained from perturbing five parameters of GBN after $LAI_{3PG}$ has received evidence at two winter and two summer seasons.

From the results obtained of the evidence propagation and the sensitivity analysis in GBN with respect to the $LAI_{3PG}$ during the winter and summer seasons, we noticed that the $LAI_{BN}$ improves and that the GBN is a robust network.

3.4.2 GBN Performance with respect to the $LAI_M$

3.4.2.1 Winter 2008 & 2009

Evidence propagation is performed in the network through $LAI_M$ during two winter seasons. Figure 3.4a shows that the $LAI_{BN}$ with evidence
3.5. Discussion

Propagating deviates from the LAI<sub>FD</sub> more than without evidence. We found the RMSE and the RE of the LAI<sub>BN</sub> after and before evidence propagates are 1.76 against 1.57 and 19.0% against 14.7%, respectively. It shows low accuracy of LAI<sub>BN</sub> and indicate a poor matching between LAI<sub>BN</sub> and LAI<sub>FD</sub> after evidence propagation.

![Figure 3.4: LAI values obtained from various sources. LAI<sub>BN</sub> before and after evidence propagation in GBN through LAI<sub>M</sub> during (a) the winter seasons, (b) the summer seasons.](image)

3.4.2.2 Summer 2008 & 2009

Evidence propagation is carried out through inserting evidence to the LAI<sub>M</sub> from field data during two summer seasons. The LAI<sub>BN</sub> with evidence propagation deviates significantly from the LAI<sub>FD</sub> than without evidence, Figure 3.4b. The RMSE and the RE of the LAI<sub>BN</sub> with and without evidence propagation equals 1.63 against 1.57 and 17.5% against 14.7%, respectively. This shows that the LAI<sub>BN</sub> after evidence propagation is worse than without evidence and confirms that the LAI<sub>BN</sub> without evidence was better than with evidence. Therefore getting evidence to LAI<sub>M</sub> does not serve as an indicator to improve GBN output.

Furthermore, sensitivity measure values $S^p_j$ are shown in Figure 3.5, showing that the GBN output does not improve and the network is sensitive (not robust), as the sensitivity measure values far from zero (about 9.4).

3.5 Discussion

This study evaluated the performance and the robustness of a GBN for improving forest growth estimates using evidence propagation and sensitivity analysis. Our results showed that the GBN output improved after LAI<sub>3PG</sub> has received evidence from LAI<sub>FD</sub>. Moreover, the GBN was robust and little sensitive to LAI<sub>3PG</sub> variation. We observed a decrease
3. Performance evaluation of GBN modeling for forest growth estimates

Figure 3.5: Sensitivity measure values ($S_p^j$) obtained from perturbing five parameters of GBN after $LAI_M$ has received evidence at two winter and two summer seasons.

of the deviation between $LAI_{BN}$ and $LAI_{FD}$. In more detail, we observed that the $LAI_{BN}$ better matched with $LAI_{FD}$ during summer seasons than during winter seasons. The reason was that the uncertainty values that are part of the $LAI_{3PG}$ input into the GBN are largest during the winter seasons.

A major contribution of this work was to identify which of the $LAI_{3PG}$ and $LAI_M$ had the largest contribution and influence to the GBN output, thus providing information for users of the GBN. The study showed that the $LAI_{BN}$ approximates the $LAI_{FD}$ after inserting evidence into $LAI_{3PG}$. This was considered as an indicator of improving $LAI_{BN}$. The GBN failed to improve, however, after $LAI_M$ has received evidence from the $LAI_{FD}$, as shown that the RMSE value of $LAI_{BN}$ increased. The reason was that a GBN assigns more weight to $LAI_M$ (Equation (3.5)) as proposed in (Mustafa et al., 2011b). A GBN needs prior information (data) to propagate the network and get posterior probability (GBN output). Therefore, Equation (3.5) is resorted to have prior information of $LAI_{BN}$ variable at each moment of the GBN implementation. Equation (3.5) has a weighing parameters, $\rho$ and $\tau$, which rely on available information of $LAI_{3PG}$ and $LAI_M$, respectively.

The study showed that GBN output was sensitive with respect to the $LAI_M$ variation. The reason was first that observations from a MODIS image had a strong influence on GBN output (Mustafa et al., 2011b). Second, each pixel corresponds to a finite area which is generally elliptical and not square on the ground. Therefore, pixel’s neighborhoods has influence on the $LAI_M$. Third, we have made the assumption that the forest is homogeneous. However, even an apparent homogeneous forest shows variation. These three sources combined lead to a relatively large sensitivity of the GBN output to MODIS image data.

Our finding that the GBN was more sensitive to the $LAI_M$ than to the
3.5. Discussion

LAI\textsubscript{3PG} corresponds with earlier findings on testing for significance of the GBN. The additional component of this work was that we studied the sensitivity of the GBN in more detail, using a common GBN sensitivity algorithm. In this algorithm we used perturbation values based on the expected uncertainties of the contribution sources in the GBN. In particular, we assessed the sensitivity on the basis of a set of five parameters. The most sensitive parameters were the $p_{1}^{\delta_{1}}$ and $p_{5}^{\delta_{5}}$, being new parameters resulting from perturbing the original parameters as described in Section 3.2.4. The change affects both the mean and the covariance values of the LAI\textsubscript{BN} before and during evidence propagation in the GBN, thus these parameters combine the two input values; the 3-PG model and MODIS images, with their uncertainties. Other parameters, however, can contribute as well for example, the parameter obtained by adding a perturbation to the mean value of the evidence variable, $p_{2}^{\delta_{2}}$, contributes to a change in value of GBN output during evidence propagation.

Evidence propagation and the sensitivity analysis were carried out for LAI\textsubscript{3PG}, and LAI\textsubscript{M} during two winter and two summer seasons. These seasons have been selected due to their influence to the accuracy of estimated LAI by 3-PG (Dye et al., 2004), and that derived from MODIS (Tian et al., 2004). The evidence propagation of GBN was carried out first by providing evidence to LAI\textsubscript{3PG} and LAI\textsubscript{M} from LAI\textsubscript{FD} during the season period. This process helps to identify the indicator that improved GBN output. Sensitivity was checked after the evidence propagated in the network at the middle time period of each season. The reason of choosing the middle time period is that the other time steps within the selected season have the same impact on the GBN.

The proposed restriction conditions in the sensitivity analysis limit the application of this method to some degree. Such restrictions, however, are required to maintain positive definiteness of the covariance matrix, which is the case in a GBN. Nevertheless, other methods to carry out a sensitivity analysis may be considered, with less severe restrictive conditions. These methods can be used to study the sensitivity analysis in discrete BNs (Laskey, 1995). For our study, based on continuous data, this would require either an expansion of the methodology, or a discretization of the input variables (Neil et al., 2007). This could be considered for the further work in this field.

This study focused on using and evaluating the performance of a GBN for integrating remote sensing data and model data for LAI estimation. Other methods could also have been used as well, like for example the Kalman Filter or the particle filter that also allow for estimation and prediction of LAI values. We consider these as valuable alternatives, and their main properties are to be investigated. That could be done in a future study. Also, the current study considered the use of the Gaussian distribution. This distribution has the well-known advantages of being practically convenient as well as being the limiting distribution of other distributions. Other, non-Gaussian, distributions could be further explored within BNs, and one may consider them in the context of state
3. Performance evaluation of GBN modeling for forest growth estimates

We identify at this stage the possibility of the extension of our research, and hope to be able to further explore it in the future. The presented methodology is useful to perform a sensitivity analysis with a GBN for improving forest growth estimates, and it is useful to evaluate the effects of inaccurate parameters on the output. Such a method can be applied within a GBN to other forest areas as well.

3.6 Conclusion

In this work, the performance of Gaussian Bayesian network is assessed using evidence propagation and a sensitivity analysis. Perturbation factors were proposed based on the uncertainties reflected by the parameters. Evidence propagation and sensitivity analysis are considered for two winter and two summer seasons due to their influence on the output. We conclude that providing evidence to the $LAI_{3PG}$ from $LAI_{FD}$ improves the GBN output. Moreover, the robustness of this GBN takes place after $LAI_{3PG}$ evidence propagates into the GBN. The explanation is that a GBN is a robust network and therefore evidence has an immediate effect on the output. Lastly, propagating evidence from $LAI_{FD}$ into the $LAI_{M}$ does not serve equally well to improve the GBN output, and that a GBN is not robust with respect to $LAI_{M}$.
Application of the Expectation Maximization-algorithm to estimate missing values in Gaussian Bayesian network modeling for forest growth

4.1 Introduction 57
4.2 Materials and methods 58
  4.2.1 Bayesian network 58
  4.2.2 The EM-algorithm for estimating missing values in a GBN 60
  4.2.3 Derivation and algorithmic steps of the EM-algorithm 61
4.3 Implementation 64
4.4 Experimental results 66
  4.4.1 Estimating $LAI_M$ during the winter of 2008 66
  4.4.2 Estimating $LAI_M$ during the winter of 2009 67
  4.4.3 Estimating $LAI_M$ during the spring of 2009 68
  4.4.4 Estimating $LAI_M$ during the full period 70
4.5 Discussion 71
4.6 Conclusion 72

This chapter has originally been published as: Y. T. Mustafa, V. A. Tolpekin, and A. Stein, 2012. Application of the Expectation Maximization Algorithm to Estimate Missing Values in Gaussian Bayesian Network Modeling for Forest Growth. IEEE Transactions on Geoscience and Remote Sensing 50(5).
4. Application of the EM-algorithm to estimate missing values in GBN

Abstract

The leaf area index (LAI) is a biophysical variable related to atmosphere-biosphere exchange of CO$_2$. One way to obtain LAI value is by the Moderate Resolution Imaging Spectroradiometer (MODIS) biophysical products. In this study we use this product to improve the physiological principles predicting growth (3-PG) model within a Gaussian Bayesian network (GBN) set-up. The MODIS time series, however, contains gaps caused by persistent clouds, cloud contamination, and other technique problems. We used the Expectation Maximization (EM)-algorithm to estimate these missing values. During a period of 26 successive months, the EM-algorithm is applied to four different cases: successively and not successively missing values during two different winter seasons, successively and not successively missing values during one spring season, and not successively missing values during the full study. Results show that the maximum value of the averaged absolute error between the original values and those estimated equals 0.16. This low value indicates that the estimated values well represent the original values. Moreover, the root mean square error of the GBN output reduces from 1.57 to 1.49 when performing the EM-algorithm to estimate the not successively missing values. We conclude that the EM-algorithm within a GBN can adequately handle missing MODIS LAI values and improves estimation of the LAI.

Keywords: Expectation Maximization (EM)-algorithm, Gaussian Bayesian networks (GBNs), leaf area index (LAI), Moderate Resolution Imaging Spectroradiometer (MODIS).
4.1 Introduction

Forests play a major role in the carbon cycle on Earth, by means of CO$_2$ absorption during photosynthesis. This leads to sequestration of a fraction of CO$_2$ in the atmosphere (Wamelink et al., 2009), thus serving to reduce the CO$_2$ ratio in the atmosphere and most likely influencing the speed of climate change. For these and other reasons, observing and analyzing forest growth have received wide attention (Bonan, 1993).

The leaf area index (LAI) is a vegetation biophysical parameter that is widely used to observe forest growth. It is defined as the one-sided area of leaf tissue per unit ground surface area ($m^2 \cdot m^{-2}$). The LAI is strongly related to gas-vegetation exchange, modeled as photosynthesis, evaporation and transpiration, rainfall interception, and carbon flux (Leuning et al., 2005; Dietz et al., 2006; Cleugh et al., 2007). Long-term monitoring of LAI can provide an understanding of dynamic changes in forest productivity and climate impacts on forest ecosystems.

Forest growth is commonly estimated using empirical models, for which LAI is an output parameter. Such a model, for example, is the Physiological Principles Predicting Growth (3-PG) model. It is a process-based forest growth model, developed by Landsberg and Waring (1997) that we will use in this study. It requires parameters, site and climate data as input and predicts the time-course of stand development on a monthly basis. A full detailed description of this model is provided in (Landsberg and Waring, 1997).

Remote sensing of vegetation provides valuable information about the LAI. In particular, the Moderate Resolution Imaging Spectroradiometer (MODIS) on-board Earth Observing System (EOS) Terra/Aqua platforms, provides LAI as a standard product (MOD15) at a 1 km resolution, every eight days (Yang et al., 2006a,b, 2007).

Statistical methods have been used in the past to estimate and improve forest growth estimates. Patenaude et al. (2008) calibrated the 3-PG model using remotely sensed data from a small forested region to simulate forest production. They used Bayesian calibration to reduce the uncertainties input. Also, graphical models have been developed in statistics to provide a flexible framework for specification and computation in complex systems (Lauritzen, 1992). A particular example of such models for estimating forest growth is the use of Bayesian networks (Kalacska et al., 2005; Mustafa et al., 2011b). A Bayesian network (BN) combines graphics and probabilities to express mutual relationships between variables. It uses a directed acyclic graph (DAG) to describe this relationship. Each node in the network represents a random variable, and the arc linking the nodes represents the relationship between variables. BNs have been widely applied in medical systems, computer science, image processing, artificial intelligence studies, and spatial data analysis (Kalacska et al., 2005; Pourret et al., 2008; Chan, 2009). Mustafa et al. (2011b) used a Gaussian Bayesian network (GBN) to improve LAI estimates by combining the 3-PG model output with MODIS images. The
applicability of their approach depends on availability of satellite images and field data. Satellite images, however, often contain gaps (missing values) due to atmospheric circumstance, such as the presence of clouds or to incomplete track spatial coverage.

For decades, researchers have relied on a variety of techniques to complete data by filling the missing values. These techniques are based on statistical or deterministic, and spatial or temporal approaches. Gao et al. (2008) presented an algorithm to produce spatially and temporally continuous time series of global LAI fields based on MODIS data. They applied their method in North America, and over specific locations containing deciduous broadleaf forest. A major breakthrough came in the 1970s with the advent of maximum likelihood (ML) estimation (Beale and Little, 1975; Dempster et al., 1977; Little and Rubin, 2002). A common method to find maximum likelihood for missing data is the expectation maximization (EM)-algorithm (Dempster et al., 1977). Handling missing data using the EM-algorithm within a GBN has not been fully explored and applied. In this study the EM-algorithm is formulated and applied to handle missing MODIS LAI values by estimating the missing parameters which are needed to execute the GBN approach in (Mustafa et al., 2011b).

The objective of this work is two-fold. First, we design an EM-algorithm to be used within a GBN. Second, we estimate missing MODIS LAI values and check the performance of the GBN. The methodology is applied to the Speulderbos study area in The Netherlands.

4.2 Materials and methods

4.2.1 Bayesian network

A BN is a graphical model that denotes joint probabilistic distribution among variables of interest based on their probabilistic relationships. The BN is a DAG, where each node represents a variable for which the label is either given or has to be assessed. The relationships between the nodes are expressed in terms of conditional probabilities and shown as arrows between nodes. The arrow starts at the condition event and expresses the probability of the event at its head (Pearl, 2000; Jensen and Nielsen, 2007).

BNs have been widely applied in different research domains. Mustafa et al. (2011b) designed a network to improve LAI estimation by combining LAI values derived from MODIS images and from the 3-PG model. Figure 4.1 shows the graphical part of the BN. The intermediate node \( (LAI_{BN}) \) represents the estimated LAI values of the BN as the variable of interest.

Based on the continuous variation of LAI over time, we showed in (Mustafa et al., 2011b) that we may assume that the LAI follows the Gaussian distribution. The most common type of a BN containing continuous variables is the GBN (Shachter and Kenley, 1989; Geiger and Heckerman, 1994; Heckerman et al., 1995). A GBN is a BN where the joint
4.2. Materials and methods

![Bayesian Network](image)

**Figure 4.1:** The BN for the first, second and $i^{th}$ time step. Each time step consists of three nodes $LAI_{3PG}$, $LAI_{BN}$, and $LAI_M$ obtained from the 3-PG model, the Gaussian Bayesian network and a MODIS image, respectively.

The probability distribution associated with its variables $LAI = \{LAI_1, ..., LAI_n\}$ is the multivariate Gaussian distribution $N(\mu, \Sigma)$, given by

$$f(LAI) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(LAI - \mu)^T\Sigma^{-1}(LAI - \mu)\right\}. \quad (4.1)$$

Here $\mu$ is the $n$-dimensional mean vector, and $\Sigma$ is the $n \times n$ positive definite covariance matrix with determinant $| \Sigma |$.

The conditional probability distribution of the $LAI_i$ represented by the $LAI_{BN_i}$ as the variable of interest given its parentage, is the univariate Gaussian distribution with density

$$f(LAI_{BN_i}|pa_i) \sim N\left(\mu_i + \sum_{j=1}^{#pa_i} \beta_{ij}(pa_{ij} - \mu_{pa_{ij}}), \nu_i \right), \quad (4.2)$$

where $\mu_i$ is the expectation of $LAI_{BN_i}$ at time $i$, the $\beta_{ij}$ represents the regression coefficients of $LAI_{BN_i}$ on its parents, $\#pa_i$ is the number of parents of $LAI_{BN_i}$, and $\nu_i = \Sigma_i - \Sigma_{ipa_i} \Sigma^{-1}_{pa_i} \Sigma_{ipa_i}$ is the conditional variance of $LAI_{BN_i}$ given its parents. Further, $\Sigma_i$ is the unconditional variance of the $LAI_{BN_i}$, $\Sigma_{ipa_i}$ are the covariances between $LAI_{BN_i}$ and the variables $pa_i$, and $\Sigma_{pa_i}$ is the covariance matrix of $pa_i$.

The regression coefficients relating $LAI_{BN_i}$ to MODIS LAI ($LAI_{M_i}$), precedent GBN output ($LAI_{BN_{i-1}}$), and LAI output of 3-PG model ($LAI_{3PG_i}$) can be expressed as $\beta_{LAI_{BN_i}LAI_{M_i}}$, $\beta_{LAI_{BN_i}LAI_{BN_{i-1}}}$, and $\beta_{LAI_{BN_i}LAI_{3PG_i}}$, respectively. Moreover, the expectation of $LAI_{M_i}$, $LAI_{BN_{i-1}}$, and $LAI_{3PG_i}$ are also expressed as $\mu_{LAI_{M_i}}$, $\mu_{LAI_{BN_{i-1}}}$, and $\mu_{LAI_{3PG_i}}$, respectively. Hence, for $i \geq 2$ and based on the graphical part of GBN in Figure 4.1, the conditional distribution of $LAI_{BN_i}$ equals...
4. Application of the EM-algorithm to estimate missing values in GBN

\[
LAI_{\text{BN}_i} \sim N \left( \mu_{LAI_{\text{BN}_i}} + \beta_{LAI_{\text{BN}_i}LAI_{\text{M}_i}} (LAI_{\text{M}_i} - \mu_{LAI_{\text{M}_i}}) + \beta_{LAI_{\text{BN}_i}LAI_{\text{BN}_{i-1}}} (LAI_{\text{BN}_{i-1}} - \mu_{LAI_{\text{BN}_{i-1}}}) + \beta_{LAI_{\text{BN}_i}LAI_{3PG_i}} (LAI_{3PG_i} - \mu_{LAI_{3PG_i}}), \Sigma_{LAI_{\text{BN}_i}} \right). \tag{4.3}
\]

The posterior probability of \(LAI_{\text{BN}_i}\) in Equation (4.3) is conditionally determined by three parents \(LAI_{\text{M}_i}\), \(LAI_{3PG_i}\), and \(LAI_{\text{BN}_{i-1}}\). These parents are used as prior input information into the GBN, where every \(LAI_{\text{M}_i}\) consists of sixteen pixels of LAI value derived from MODIS images with 250 m spatial resolution. Also every \(LAI_{3PG_i}\) consists of sixteen simulated LAI output from the 3-PG model, and every intermediate node \(LAI_{\text{BN}_i}\) consists of sixteen LAI values. Based on the contribution of all of the MODIS images, the 3-PG output, and the precedent GBN output, the prior information of the intermediate node is assumed and defined in (Mustafa et al., 2011b) as

\[
LAI_{\text{BN}_i} = \rho (1 - \tau) LAI_{\text{M}_i} + \tau LAI_{3PG_i} + (1 - \rho)(1 - \tau) LAI_{\text{BN}_{i-1}}, \tag{4.4}
\]

where \(\tau\) and \(\rho\) are the weighing values, defined as \(\tau = \left| \frac{LAI_{\text{M}_i} - LAI_{\text{M}_{i-1}}}{LAI_{\text{M}_{i-1}}} \right|\) and \(\rho = \left| \frac{LAI_{3PG_i} - LAI_{3PG_{i-1}}}{LAI_{3PG_{i-1}}} \right|\). They are proportional to the change in the LAI values obtained from the MODIS images and the 3-PG output. Moreover, this prior information has been tested on Gaussianity by Mustafa et al. (2011b), in terms of the temporal and the spatial resolution of the input sources into the GBN. They used the Shapiro-Wilk and Lilliefors tests for that purpose. For more details about a GBN of improving forest growth estimates and its mathematical explanation, we refer to (Mustafa et al., 2011b).

4.2.2 The EM-algorithm for estimating missing values in a GBN

The study aims to analyze a time series of satellite imageries. A common problem in such a series is the occurrence of incomplete observations (missing data). This is due to factors as persistent cloud cover, elevated aerosol loading, or sparse observations owing to insufficient repeat frequency of the satellite sensor. Mustafa et al. (2011b) used a time series of satellite data of two years to update estimated LAI values of the 3-PG model within a GBN. Their method requires a complete time series of satellite data as an input source into a GBN.

Several methods have been applied and proposed to estimate missing values (Beale and Little, 1975; Dempster et al., 1977; Little and Rubin, 2002; McLachlan and Krishnan, 2008). A well-established technique for estimating parameters of statistical models from incomplete data is the EM-algorithm. The origins of the EM-algorithm date back to the 1970s (Dempster et al., 1977). Early EM applications primarily focused
4.2. Materials and methods

on estimating a mean vector and a covariance matrix with missing data. Since then it has been extended to address a variety of incomplete-data estimation problems (McLachlan and Krishnan, 2008).

The EM-algorithm is broadly applicable for maximizing likelihoods and handling incomplete data problems. The idea of the algorithm is to alternate between computing the expectation of the sample log-likelihood conditional on the previous parameter estimates (the E-step) and maximizing this expectation with respect to the desired parameters to obtain parameter estimates for the next recursion (the M-step). It was shown in (Dempster et al., 1977) that this procedure is guaranteed to produce a monotonically increasing sequence of expected sample log-likelihoods which converges to a local maximum of the likelihood function. Thus, the EM-algorithm is formulated and applied to handle the missing satellite data. To do so, it estimates parameters in a GBN implementation on LAI.

4.2.3 Derivation and algorithmic steps of the EM-algorithm

Consider missing data of satellite images at the \(i^{th}\) moment \((i > 1)\) within a GBN as shown in Figure 4.2. The GBN output, \(\text{LAI}_{\text{BN}}^i\), conditionally depends on three nodes (variables), i.e., \(\text{LAI}_{\text{M}}^i\), \(\text{LAI}_{\text{BN}}^{i-1}\), and \(\text{LAI}_{\text{3PG}}^i\). In this work, we consider \(\text{LAI}_{\text{M}}^i\) as a missing value.

Let \((Y, X)\) be the complete data set at the \(i^{th}\) moment of GBN, with observed (complete) data \(Y = \{\text{LAI}_{\text{BN}}^i, \text{LAI}_{\text{3PG}}^i, \text{LAI}_{\text{BN}}^{i-1}\}\) and missing data \(X = \text{LAI}_{\text{M}}^i\) (Figure 4.2). For clarity, we abbreviate the GBN variables as \(y = \text{LAI}_{\text{BN}}^i, x = \text{LAI}_{\text{M}}^i, z = \text{LAI}_{\text{BN}}^{i-1}, w = \text{LAI}_{\text{3PG}}^i\); such that the model parameters of the GBN are named \(\mu_j, \sigma_j\) and \(\beta_{y_j}\) for \(j \in \{x, y, z, w\}\).

Hence Equation (4.3) can be reformulated as

\[
\gamma \sim N \left( \mu_y + \beta_{yx}(x - \mu_x) + \beta_{yz}(z - \mu_z) + \beta_{yw}(w - \mu_w), \sigma_y^2 \right). \tag{4.5}
\]
4. Application of the EM-algorithm to estimate missing values in GBN

The EM-steps to find new ML estimates for the parameters $\theta = (\mu_x, \Sigma_x)$ are as follows.

1. Choose an initial setting for the parameters $\theta$ and name it as $\theta^{old} = (\mu^{old}, \sigma^{old})$. Initial parameter values are guessed based on seasonal changes of LAI values obtained from MODIS observations

$$\theta^{old} = \begin{cases} 
(\mu - \frac{\mu_{x_i-2} - \mu_{x_i-1}}{\mu_{x_i-1}}, \sigma - \frac{\sigma_{x_i-2} - \sigma_{x_i-1}}{\sigma_{x_i-1}}) & \text{if } \mu_{x_i-2} \leq \mu_{x_i-1} \\
(\mu + \frac{\mu_{x_i-2} - \mu_{x_i-1}}{\mu_{x_i-1}}, \sigma + \frac{\sigma_{x_i-2} - \sigma_{x_i-1}}{\sigma_{x_i-1}}) & \text{Otherwise}
\end{cases} \quad (4.6)$$

where $\mu_x$ and $\sigma_x$ are the mean and the standard deviation values of the MODIS LAI, respectively. They are obtained either for the period from September to February (the non growing season), or for the period from March to August (the growing season). The determination of the period for which we need to obtain the $\mu_x$ and $\sigma_x$ is based on the occurrence of missing observations in that period.

In Equation (4.6), the $\frac{\mu_{x_i-2} - \mu_{x_i-1}}{\mu_{x_i-1}}$ and $\frac{\sigma_{x_i-2} - \sigma_{x_i-1}}{\sigma_{x_i-1}}$ are the relative changes of the mean and the standard deviation of the previous two MODIS LAI. Adding and subtracting these relative change values are based on the condition of increase or decrease the MODIS LAI values during the period of non growing or growing season.

2. E-step: compute the expectation (with respect to the $X$ data) of the likelihood function of the model parameters by including the missing variables as they were observed,

$$Q(\theta, \theta^{old}) = E_X[\log f(Y, X|\theta)|Y, \theta^{old}]$$

$$= \int \log f(Y, X|\theta)f(X|Y, \theta^{old})dX$$

$$= \int \log f(x, y, z, w|\theta)f(x|y, z, w, \theta^{old})dx, \quad (4.7)$$

where $\log f(x, y, z, w|\theta) = \log f(y|x, z, w, \theta)f(x|\theta)f(z|\theta)f(w|\theta)$ and $f(y|x, z, w, \theta)$ is the conditional distribution of $y$ given its parents $x, z$, and $w$, such that $f(y|x, z, w, \theta) \sim N(\mu_y + \beta_{yx}(x - \mu_x) + \beta_{yz}(z - \mu_z) + \beta_{yw}(w - \mu_w), \sigma_y^2)$, $f(x|\theta) \sim N(\mu_x, \sigma_x^2)$, $f(z|\theta) \sim N(\mu_z, \sigma_z^2)$, and $f(w|\theta) \sim N(\mu_w, \sigma_w^2)$.

Therefore, the $\log f(x, y, z, w|\theta)$ can be expressed as in the Equation (4.8).
\[ \log f(x, y, z, w | \theta) = -\frac{1}{2} \left( \frac{1}{\sigma_x^2} + \frac{\beta_{yx}^2}{\sigma_y^2} \right) x^2 \]

\[ + \left( \frac{y - \mu_y + \beta_{yx} \mu_x - \beta_{yz} (z - \mu_z) - \beta_{yw} (w - \mu_w)}{\sigma_y^2} \right) \beta_{yx} \]

\[ - \frac{1}{2} \left( \frac{y - \mu_y + \beta_{yx} \mu_x - \beta_{yz} (z - \mu_z) - \beta_{yw} (w - \mu_w)}{\sigma_y^2} \right)^2 \]

\[ + \frac{(z - \mu_z)^2}{\sigma_z^2} + \frac{(w - \mu_w)^2}{\sigma_w^2} + \frac{\mu_x^2}{\sigma_x^2} \] \[ \text{log} \left( 4 \pi^2 \sigma_y \sigma_x \sigma_z \sigma_w \right). \quad (4.8) \]

Based on the graphical representation of the GBN model, the second factor of the Equation (4.7) can be found as;

\[ f(x | y, z, w, \theta^{\text{old}}) = \frac{f(x, y, z, w | \theta^{\text{old}})}{\int f(x, y, z, w | \theta^{\text{old}}) \, dx}, \quad (4.9) \]

which can be written as \( Q = \int_{-\infty}^{\infty} V \left( -fx^2 + gx - h \right) e^{-ax^2 + bx - c} \, dx, \)

where \( V = \frac{\sqrt{\sigma_y^2 + \beta_{yx}^2} \sigma^{old}_x}{\sqrt{2\pi} \left( \sigma^{old}_y \right)^2 \sigma^{old}_x} \)

\[ f = \frac{1}{2} \left( \frac{1}{\sigma_x^2} + \frac{\beta_{yx}^2}{\sigma_y^2} \right), \]

\[ g = \left( \frac{y - \mu_y + \beta_{yx} \mu_x - \beta_{yz} (z - \mu_z) - \beta_{yw} (w - \mu_w)}{\sigma_y^2} \right) \beta_{yx}, \]

\[ h = \frac{1}{2} \left( \frac{y - \mu_y + \beta_{yx} \mu_x - \beta_{yz} (z - \mu_z) - \beta_{yw} (w - \mu_w)}{\sigma_y^2} \right)^2 + \frac{(z - \mu_z)^2}{\sigma_z^2} + \frac{(w - \mu_w)^2}{\sigma_w^2} + \frac{\mu_x^2}{\sigma_x^2} \]

\[ - \log \left( 4 \pi^2 \sigma_y \sigma_x \sigma_z \sigma_w \right), \]

\[ a = \frac{1}{2} \left( \frac{1}{\sigma^{old}_x} + \frac{\beta_{yx}^2}{\sigma_y^2} \right), \]

\[ b = \left( \frac{y - \mu_y + \beta_{yx} \mu_x - \beta_{yz} (z - \mu_z) - \beta_{yw} (w - \mu_w)}{\sigma_y^2} \right) \beta_{yx} + \frac{\mu^{old}_x}{\sigma^{old}_x}, \]

and \( c = \frac{b^2}{4a}. \)

Here \( \mu^{old} \) and \( \sigma^{old} \) refer to the guessed mean and standard deviation value of \( x \) variable (LAIx) obtained using the Equation (4.6). It can further be shown that the \( Q(\theta, \theta^{\text{old}}) \) equals

\[ Q(\theta, \theta^{\text{old}}) = \Omega \left( -\frac{f}{2a} + \frac{b}{2a} \frac{h}{(2a)^2} - h \right), \quad (4.10) \]

where \( \Omega = V \sqrt{\frac{\pi}{a}} e^{-c + \frac{b^2}{4a}}. \)

3. M-step: compute the ML estimates of the parameters \( \theta \) by maximizing the expected likelihood obtained during the E-step, i.e., \( \theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}}). \)
4. Application of the EM-algorithm to estimate missing values in GBN

Hence, by differentiating $Q(\theta, \theta_{\text{old}})$ with respect to $\theta$ and solving the equations for $\theta_{\text{new}} = (\mu_{\text{new}}^x, \sigma_{\text{new}}^x)$, the maximum values are found as

$$
\mu_{\text{new}}^x = \frac{\sqrt{\phi + \Psi}}{6 \lambda} + \frac{2 (-3 \delta \lambda + \alpha^2)}{3 \lambda \sqrt{\phi + \Psi}} + \frac{\alpha}{3 \lambda},
$$

(4.11)

$$
\sigma_{\text{new}}^x = \sqrt{\left(\mu_{\text{new}}^x\right)^2 - \frac{2 C}{B} \mu_{\text{new}}^x + \frac{E}{B}}.
$$

(4.12)

Here $\phi = -36 \delta \alpha \lambda + 108 \eta \lambda^2 + 8 \alpha^3$,

$$
\Psi = 12 \sqrt{3} \lambda^3 - \delta^2 \alpha^2 - 18 \delta \alpha \lambda \eta + 27 \eta^2 \lambda^2 + 4 \eta \alpha^3 \lambda,
$$

$\lambda = AB$,

$\alpha = 2 AC + DB$,

$\delta = AE + B^2 + 2 DC$,

$\eta = CB + dE$,

where $A = 4 \frac{\alpha^2 \Omega \beta_{yx}^2}{\sigma_y^2}$,

$B = 4 \alpha^2$, $C = 2 \alpha \beta \Omega$,

$D = -4 \alpha^2 \Omega (y - \mu_y - \beta_{yx} (z - \mu_z) - \beta_{yw} (w - \mu_w)) \beta_{yx}^2 + 2 \frac{\alpha \beta \Omega \beta_{yx}^2}{\sigma_y^2}$,

and $E = 2 \alpha \Omega + b^2 \Omega$.

4. Using the iterative EM-algorithm requires to check the convergence of the $\theta_{\text{new}}$ values. If $|\theta_{\text{new}} - \theta_{\text{old}}| > \varepsilon$ for a small value $\varepsilon$, then $\theta_{\text{old}} \leftarrow \theta_{\text{new}}$, and the algorithm returns to step 2 (Algorithm 2).

The convergence criterion $\varepsilon$ has been set in this study equal to $10^{-5}$ after examining the EM-algorithm for several cases of missing LAI$_m$ values. It ensures that the estimated $\theta$ differs by less than 0.1%. Figure 4.3 shows the changes of parameter estimates and the differences in those during consecutive iterations for one run of the EM-algorithm.

4.3 Implementation

The EM-algorithm, as described in Section 4.2.3, is applied to the Speulderbos forest in The Netherlands where the LAI is observed and is available as a time series from July 2007 to September 2009. The study area of this work has been assumed to be covered by a homogeneous forest of 1 km$^2$ around the climate station in the Speulderbos forest as it has been considered in (Mustafa et al., 2011b). Description of the study area and data used were given in Sections 1.7.1 and 1.7.2, respectively.

The time period contains two winter seasons (October-March; 2008 & 2009), two summer seasons (May-August; 2008 & 2009), and two spring seasons (end of March-mid of May; 2008 & 2009) (Figure 2.6). We have a
full set of observed MODIS LAI images during the monitoring period. To implement the approach of this work, synthetic gaps are introduced. This is done by removing some of MODIS LAI observations successively and not successively, following the common occurrence of missing satellite observations as explained below. The missing values are estimated using the EM-algorithm, and saved as a test set to compare with the original MODIS LAI.

Missing satellite imageries mainly occur during the winter season, due to the atmospheric conditions such as the presence of clouds and aerosols. Moreover, satellite images may not be available in other seasons due to the incomplete track spatial coverage. The EM-algorithm is applied to estimate missing MODIS LAI in four cases. The first case considers successively and not successively missing MODIS LAI values during the winter of 2008. The second case similarly considers the winter of 2009. The third case considers successively and not successively missing MODIS LAI values during the spring of 2009. The fourth case concerns not successively missing MODIS LAI values during the period from July 2007 to September 2009.

In each winter season, the EM-algorithm is carried out to estimate successively missing LAI$_M$ values during the first half (first two and a half months), the second half (second two and a half months), and the whole time period of each season. Moreover, the EM-algorithm is applied to estimate not successively missing LAI$_M$ values during each season. The performance of the GBN is assessed when performing the EM-algorithm. The method was implemented in C++ code. For clarity,
Algorithm 2: EM-algorithm in GBN

**Data:** LAI$_{BN}$, LAI$_{BN+1}$, LAI$_{3PG}$, and observed LAI$_M$ values

**Result:** Estimated parameters $\theta_{new} = (\mu_{new}, \sigma_{new})$ of the missing LAI$_M$

**Initial set**
Choose an initial $\theta_{old}$ values as in the Equation (4.6);

while $|\theta_{new} - \theta_{old}| > \varepsilon$

**E-Step**
Compute $\log f(Y, X|\theta)$ using Equation (4.8);
Compute $f(x|Y, \theta_{old})$ using Equation (4.9);
Compute $Q(\theta, \theta_{old})$ using Equation (4.7);

**M-Step**
Formulate $Q(\theta, \theta_{old})$ with respect to the $\theta$;
Solve $Q(\theta, \theta_{old})$ to find $\theta_{new}$;
Find a solution which maximizes $Q(\theta, \theta_{old})$ to get Equations (4.11), and (4.12);

the figures hereinafter only include the LAI$_M$, LAI$_{BN}$ and LAI$_{FD}$ as the values of interest.

The accuracy of LAI$_{BN}$ is assessed using the root mean square error (RMSE$_{LAI_{BN}}$) and the relative error (RE$_{LAI_{BN}}$) with respect to the LAI$_{FD}$ before and after performing EM-algorithm. The accuracy of LAI$_M$ is assessed as well using the RMSE$_{LAI_{M}}$ and RE$_{LAI_{M}}$ with respect to the LAI$_{FD}$ before and after performing EM-algorithm. The averaged absolute error (AAE$_{LAI_{M}}$) of the estimating LAI$_M$ with respect to the original LAI$_M$ is also calculated. The RMSE, RE and AAE are computed as follows:

$$\text{RMSE}_{LAI_{BN}} = \sqrt{\frac{\sum_{i=1}^{n}(\text{LAI}_{BN} - \text{LAI}_{FD})^2}{n}}, \quad \text{RMSE}_{LAI_{M}} = \sqrt{\frac{\sum_{i=1}^{n}(\text{LAI}_{M} - \text{LAI}_{FD})^2}{n}},$$

$$\text{RE}_{LAI_{BN}} = \left| \frac{\text{LAI}_{BN} - \text{LAI}_{FD}}{\text{LAI}_{FD}} \right| 100\% , \quad \text{RE}_{LAI_{M}} = \left| \frac{\text{LAI}_{M} - \text{LAI}_{FD}}{\text{LAI}_{FD}} \right| 100\% ,$$

and $\text{AAE}_{LAI_{M}} = \frac{1}{n} \sum_{i=1}^{n} |\text{Origin LAI}_{M} - \text{Estimated LAI}_{M} |$

4.4 Experimental results

4.4.1 Estimating LAI$_M$ during the winter of 2008

Estimation of missing LAI$_M$ values is carried out during the first half, the second half, and the whole time period of the season (Figure 4.4(a)-(d)). Estimated values are close to the original LAI$_M$ with AAE$_{LAI_{M}}$ values less than 0.1. The deviation between the LAI$_{BN}$ values after estimating the missing LAI$_{BN}$ values and the LAI$_{FD}$ becomes significantly lower than the deviation between the LAI$_{BN}$ values with the original LAI$_M$ and the
4.4. Experimental results

$LAI_{FD}$. Figure 4.4(a) shows LAI values after performing the EM-algorithm of estimating five successively missing $LAI_{M}$ values during the first half of the season. We found an RMSE$_{LAI_{BN}}$ of 1.53 and an RE$_{LAI_{BN}}$ of 13.2%. Figure 4.4(b) shows five successively missing $LAI_{M}$ values estimated during the second half of the season. The RMSE$_{LAI_{BN}}$ and the RE$_{LAI_{BN}}$ in this case are 1.55 and 14.2%, respectively, whereas in Figure 4.4(d) five not successively missing $LAI_{M}$ values during the whole time period of the season are estimated with an RMSE$_{LAI_{BN}}$ of 1.51 and an RE$_{LAI_{BN}}$ of 13.3%. The deviation between $LAI_{BN}$ and $LAI_{FD}$ increases, however, after performing the EM-algorithm (Figure 4.4(c)).

The eight successively missing $LAI_{M}$ values during the whole time period of the season are estimated using the EM-algorithm. The RMSE$_{LAI_{BN}}$ and the RE$_{LAI_{BN}}$ after and before performing the EM-algorithm equal 1.68 against 1.57 and 17.6% against 14.7%, respectively.

Estimating $LAI_{M}$ using the EM-algorithm represents the original $LAI_{M}$ particularly in the not successively missing case with an AAE$_{LAI_{M}}$ of 0.02 (Table 4.1). This indicates that the missing $LAI_{M}$ are estimated successfully.

![Figure 4.4: $LAI_{BN}$ and $LAI_{M}$ values of the Speulderbos forest obtained before and after performing the EM-algorithm during the winter of 2008; (a) five successively missing $LAI_{M}$ values during the first half of the season, (b) five successively missing $LAI_{M}$ values during the second half of the season, (c) eight successively missing $LAI_{M}$ values during the full season, (d) five not successively missing $LAI_{M}$ values during the full season.](image)

4.4.2 Estimating $LAI_{M}$ during the winter of 2009

After applying the EM-algorithm to estimate missing MODIS LAI at the same time steps of Section 4.4.1, we notice that the $LAI_{BN}$ is close to the $LAI_{FD}$ (Figure 4.5(a), (b), (d)). Figure 4.5(a) shows $LAI_{BN}$ and $LAI_{M}$ values
4. Application of the EM-algorithm to estimate missing values in GBN

Table 4.1: The root mean square error (RMSE) and the relative error (RE) of $L_{AM}$ and $L_{BN}$, and the averaged absolute error (AAE) of $L_{AM}$. They are obtained before and after applying the EM-algorithm during the winter of 2008.

<table>
<thead>
<tr>
<th>Cases</th>
<th>RMSE$<em>{L</em>{AM}}$</th>
<th>RMSE$<em>{L</em>{BN}}$</th>
<th>RE$<em>{L</em>{AM}}$</th>
<th>RE$<em>{L</em>{BN}}$</th>
<th>AAE$<em>{L</em>{AM}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>without missing (original $L_{AM}$)</td>
<td>3.26</td>
<td>1.57</td>
<td>44.1%</td>
<td>14.7%</td>
<td>0.05</td>
</tr>
<tr>
<td>Five successively missing $L_{AM}$ estimated during the first half</td>
<td>3.27</td>
<td>1.53</td>
<td>44.0%</td>
<td>13.2%</td>
<td>0.05</td>
</tr>
<tr>
<td>Five successively missing $L_{AM}$ estimated during the second half</td>
<td>3.26</td>
<td>1.55</td>
<td>44.0%</td>
<td>14.2%</td>
<td>0.02</td>
</tr>
<tr>
<td>Eight successively missing $L_{AM}$ estimated during the whole season period</td>
<td>3.23</td>
<td>1.68</td>
<td>43.6%</td>
<td>17.6%</td>
<td>0.16</td>
</tr>
<tr>
<td>Five not successively missing $L_{AM}$ estimated during the whole season period</td>
<td>3.27</td>
<td>1.51</td>
<td>44.1%</td>
<td>13.3%</td>
<td>0.02</td>
</tr>
</tbody>
</table>

after estimating five successively missing $L_{AM}$ values during the first half of the season. The RMSE$_{L_{BN}}$ equals 1.59 and the RE$_{L_{BN}}$ equals 14.7%. Furthermore, the five successively missing $L_{AM}$ values during the second half of the season are shown in Figuer 4.5(b). The RMSE$_{L_{AM}}$ and the RE$_{L_{BN}}$ in this case are 1.58 and 14.6%, respectively. Five not successively missing $L_{AM}$ values during the whole time period of the season are estimated with an RMSE$_{L_{BN}}$ of 1.51 and an RE$_{L_{BN}}$ of 14.4% (Figuer 4.5(d)), respectively. Moreover, the estimated missing $L_{AM}$ values are close to the original $L_{AM}$ values with AAE$_{L_{BN}}$ values below 0.08.

The strong deviation between $L_{BN}$ and $L_{FD}$ has occurred after applying the EM-algorithm to estimate eight successively missing $L_{AM}$ values during the whole time period of the season (Figuer 4.5(c)). The RMSE$_{L_{BN}}$ and RE$_{L_{BN}}$ after and before performing the EM-algorithm equal 1.69 against 1.57 and 17.0% against 14.4%, respectively. The quantitative results are shown in Table 4.2.

Table 4.2: The root mean square error (RMSE) and the relative error (RE) of $L_{AM}$ and $L_{BN}$, and the averaged absolute error (AAE) of $L_{AM}$. They are obtained before and after applying the EM-algorithm during the winter of 2009.

<table>
<thead>
<tr>
<th>Cases</th>
<th>RMSE$<em>{L</em>{AM}}$</th>
<th>RMSE$<em>{L</em>{BN}}$</th>
<th>RE$<em>{L</em>{AM}}$</th>
<th>RE$<em>{L</em>{BN}}$</th>
<th>AAE$<em>{L</em>{AM}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>without missing (original $L_{AM}$)</td>
<td>3.26</td>
<td>1.57</td>
<td>44.1%</td>
<td>14.7%</td>
<td>0.04</td>
</tr>
<tr>
<td>Five successively missing $L_{AM}$ estimated during the first half</td>
<td>3.25</td>
<td>1.59</td>
<td>44.0%</td>
<td>14.7%</td>
<td>0.04</td>
</tr>
<tr>
<td>Five successively missing $L_{AM}$ estimated during the second half</td>
<td>3.23</td>
<td>1.58</td>
<td>43.6%</td>
<td>14.6%</td>
<td>0.03</td>
</tr>
<tr>
<td>Eight successively missing $L_{AM}$ estimated during the whole season period</td>
<td>3.24</td>
<td>1.69</td>
<td>43.8%</td>
<td>17.0%</td>
<td>0.08</td>
</tr>
<tr>
<td>Five not successively missing $L_{AM}$ estimated during the whole season period</td>
<td>3.22</td>
<td>1.51</td>
<td>43.6%</td>
<td>14.4%</td>
<td>0.04</td>
</tr>
</tbody>
</table>

4.4.3 Estimating $L_{AM}$ during the spring of 2009

Missing $L_{AM}$ values are estimated during the second spring season only (the spring of 2009) as an extra test of the EM-algorithm. Figuer 4.6(a)
4.4. Experimental results

Figure 4.5: LAI_{BN} and LAI_{M} values of the Speulderbos forest obtained before and after performing the EM-algorithm during the winter of 2009; (a) five successively missing LAI_{M} values during the first half of the season, (b) five successively missing LAI_{M} values during second half of the season, (c) eight successively missing LAI_{M} values during the whole time period of the season, (d) five not successively missing LAI_{M} values during the whole time period of the season.

shows LAI values of estimating five successively missing LAI_{M} values during the season. There is a noticeable deviation with AAE_LAI_{M} of 0.09 between the estimated missing LAI_{M} and original LAI_{M}. Moreover, the LAI_{BN} deviates little from LAI_{FD} after performing the EM-algorithm than before applying EM-algorithm but it remains close to the LAI_{FD}. We found an RMSE_{LAI_{BN}} of 1.62 and an RE_{LAI_{BN}} of 15.8%.

The EM-algorithm is applied to estimate three not successively missing LAI_{BN} values during the season as well (Figure 4.6(b)). At this case a small deviation between LAI_{BN} and LAI_{FD} is observed with an RMSE_{LAI_{BN}}, an RE_{LAI_{BN}}, and an AAE_{LAI_{BN}} equal to 1.56, 14.5%, and 0.06, respectively (Table 4.3).

Table 4.3: The root mean square error (RMSE) and the relative error (RE) of LAI_{M} and LAI_{BN}, and the averaged absolute error (AAE) of LAI_{M}. They are obtained before and after applying the EM-algorithm during the spring of 2009.

<table>
<thead>
<tr>
<th>Cases</th>
<th>RMSE_{LAI_{M}}</th>
<th>RMSE_{LAI_{BN}}</th>
<th>RE_{LAI_{M}}</th>
<th>RE_{LAI_{BN}}</th>
<th>AAE_{LAI_{M}}</th>
</tr>
</thead>
<tbody>
<tr>
<td>without missing (original LAI_{M})</td>
<td>3.26</td>
<td>1.57</td>
<td>44.1%</td>
<td>14.7%</td>
<td>0.09</td>
</tr>
<tr>
<td>Five successively missing LAI_{M}</td>
<td>3.25</td>
<td>1.62</td>
<td>44.2%</td>
<td>15.8%</td>
<td>0.09</td>
</tr>
<tr>
<td>estimated during the whole season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three not successively missing LAI_{M}</td>
<td>3.25</td>
<td>1.56</td>
<td>43.8%</td>
<td>14.5%</td>
<td>0.06</td>
</tr>
<tr>
<td>estimated during the whole season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Application of the EM-algorithm to estimate missing values in GBN

4.4.4 Estimating LAI_M during the full period

Finally, the EM-algorithm is carried out to estimate the missing LAI_M during the whole study period (16 not successively missing LAI_M values). Performance of the GBN is assessed during the test.

The LAI_BN after performing the EM-algorithm for estimating missing LAI_M values remains close to the LAI_FD. The deviation between LAI_BN and LAI_FD reduces after applying the EM-algorithm (Figure 4.7). We found an RMSE_{LAI_BN} value of 1.49 against 1.57 and an REL_{LAI_BN} value of 14.0% against 14.7%. The RMSE_{LAI_M} equals 3.27 and the RE_{LAI_M} equals 44.4%. Moreover, the AAE_{LAI_M} resulting after comparing the estimated values with the original values equals 0.16. This value is low with respect to the 16 missing LAI_M, thus confirming the success of formulating and implementing the EM-algorithm within a GBN to estimate the missing LAI_M and carrying out the GBN approach.

Figure 4.6: LAI_BN and LAI_M values of the Speulderbos forest obtained before and after performing the EM-algorithm during the spring of 2009; (a) five successively missing LAI_M values during the season, (b) three not successively missing LAI_M values during the full period of the season.

Figure 4.7: LAI_BN and LAI_M values at the Speulderbos forest obtained before and after performing the EM-algorithm for not successively missing LAI_M values during the period from July 2007 to September 2009.
4.5 Discussion

In this study the EM-algorithm is formulated within a Gaussian Bayesian network, and missing $LAI_M$ values are estimated.

Our results show that the missing $LAI_M$ values are estimated successfully, i.e. such that they well represent the original MODIS LAI trend. The GBN output with estimated missing $LAI_M$ values remains closer to the LAI field data than the LAI output of the 3-PG model. Overall the deviation between the GBN output and the LAI field data after estimating missing $LAI_M$ is lower than the deviation between the 3-PG model output and the LAI field data. Moreover, a smaller deviation occurred between GBN output and LAI field data in case of estimating not successively missing $LAI_M$ during the time study (Section 4.4.4). This emphasizes that the EM-algorithm within a GBN serves as a good approach to estimate missing MODIS LAI values.

The strength of the represented work lies in applying the EM-algorithm in a GBN to estimate the missing input source, $LAI_M$, of the GBN. Analytically, the most difficult portion of the EM-algorithm is the E-step. The reason for this is that the expectation must be computed over all LAI values. This may slow down the convergence of the EM-algorithm in some cases, requiring up to 800 iterations. Although other algorithms offer faster computation (Little and Rubin, 2002; McLachlan and Krishnan, 2008; Kuroda and Sakakihara, 2006), we selected the EM-algorithm because it guarantees convergence.

A well-known drawback of the EM-algorithm is that the convergence can be quite slow (McLachlan and Krishnan, 2008). In order to save computing time, it is essential to start with good initial parameters. For this reason we resorted Equation (4.6) such that we can identify the initial values to be close to the estimated $LAI_M$ values. The initial guess in Equation (4.6) has been formulated based on the seasonal change of LAI values that are derived from MODIS. However, the relative changes of $LAI_M$ are introduced in order to account the LAI change within a relatively short period of time which may show a change in the LAI trend.

From the results of performing the EM-algorithm to estimate the missing $LAI_M$, we observed that even a small difference between the estimated missing $LAI_M$ and the original $LAI_M$ has an impact on the resulting output of the GBN. This is due to the fact that a GBN is sensitive to $LAI_M$ variation and gives more weight to $LAI_M$ (Equation (4.4)), as proposed in (Mustafa et al., 2011b), since GBN needs prior information (data) to propagate the network and to get posterior probability (GBN output). The Equation (4.4) is resorted (Mustafa et al., 2011b) to include prior information of the $LAI_{BN}$ variable at each moment of the GBN implementation. Equation (4.4) has weighing parameters, $\tau$ and $\rho$, which rely on available information of $LAI_M$ and $LAI_{3PG}$, respectively. Hence, an increase of the RMSE$_{LAI_M}$ (Table 4.1) means that the $\tau$ value in the Equation (4.4) decreases and leads to a decrease of the prior value of the intermediate variable, $LAI_{BN}$, to be used in the next iteration (observation). This results in a decrease of the posterior probability of $LAI_{BN}$ and a decrease of the
4. Application of the EM-algorithm to estimate missing values in GBN

Differences between $LAI_{BN}$ and $LAI_{FD}$. This only happens in the case of successively missing $LAI_M$ for which the values of $\tau$ are calculated based on two or more estimations of successively missing $LAI_M$ values. This does not happen for not successively missing $LAI_M$; since then, the $\tau$ value decreases because of the decreasing $LAI_M$, resulting in a decrease of the prior value of $LAI_{BN}$ for the next iteration (Table 4.2 and 4.3). The reason is that values of $\tau$ are calculated based on the original $LAI_M$ rather than on those estimated by the EM-algorithm, hence resulting in a decrease GBN output.

The performance of the $LAI_{BN}$ is better in the case of estimating not successively missing $LAI_M$ than in the case of estimating successively missing $LAI_M$. The poorest $LAI_{BN}$ performance was in the case of estimating successively missing images. Moreover, a noticeable deviation of $LAI_{BN}$ is seen after applying the EM-algorithm during the spring season with successively and not successively missing $LAI_M$. The reason is that the spring season is the most critical time for forest growth, when the LAI values start changing and increasing. Overall, the results of this study show that the applicability of the EM-algorithm within a GBN is relatively limited for a long (more than 5 observations) series of successively missing $LAI_M$.

Some issues require further work. For instance, other methods could be included into the EM-algorithm to accelerate the speed of convergence (Kuroda and Sakakihara, 2006). Such improvements are e.g. based on the application to the EM-algorithm of optimization theory techniques such as Aitkin’s acceleration (Meilijson, 1989). These methods require some further preparatory analytical work and may increase the complexity of the implementation. Finally, ground observations, $LAI_{FD}$, may be partly or entirely absent as some areas are difficult to reach or the instruments for field-survey are too expensive for the national forest survey institutes. The EM-algorithm within a GBN, however, might be useful to deal with this kind of problem. This could be achieved by modifying the network such as include another node, variable, represents the ground data and applying EM-algorithm to the added variable.

4.6 Conclusion

This study formulated the EM-algorithm within a Gaussian Bayesian network to estimate missing MODIS LAI values. Performance of the network is evaluated by comparing its output with LAI field measurements before and after performing the EM-algorithm. The approach is examined for four cases: successively and not successively missing $LAI_M$ during various time periods of the winter of 2008, successively and not successively missing $LAI_M$ during various time periods of the winter of 2009, successively and not successively missing $LAI_M$ during the spring of 2009, and not successively missing $LAI_M$ during the full study period. We conclude that the missing $LAI_M$ values are estimated successfully using the EM-algorithm with 0.16 of the maximum value of the averaged
absolute error between the original values and those estimated. The presence of more than five successively missing $LAI_M$ influences the output of the network such that $LAI_{BN}$ does not match the $LAI_{FD}$, but that it approximates. We conclude, in particular, that the $LAI_{BN}$ improves after performing the EM-algorithm in the case of estimating not successively missing $LAI_M$ values during the full study period.
Improving forest growth estimates using a Bayesian network approach

5.1 Introduction 77
5.2 Materials and methods 77
  5.2.1 The 3-PG model 77
  5.2.2 Remote sensing imagery 78
  5.2.3 Bayesian networks 78
  5.2.4 EM-algorithm for estimating missing values in a GBN 80
5.3 Study area 82
5.4 Data description 83
  5.4.1 Ground data 83
  5.4.2 Satellite data 84
5.5 Implementation 84
5.6 Results 85
  5.6.1 GBN performance with point mode 3-PG model 85
  5.6.2 Spatial GBN performance with spatial 3-PG model 86
5.7 Discussion and conclusions 86

Improving forest growth estimates using a Bayesian network approach

Abstract

Estimating the contribution of forests to carbon sequestration is commonly done by applying forest growth models. Such models inherently use field observations, such as leaf area index (LAI), whereas relevant information is also available from satellite images. The purpose of this study is to improve the LAI estimated from the Physiological Principles Predicting Growth (3-PG) model by combining its output with LAI derived from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) sensor. A Bayesian network (BN) approach is proposed to take care of the different structure of the inaccuracies in the two data sources. It addresses the bias in the 3-PG model and the noise of the ASTER images. Moreover, the EM-algorithm is introduced into BN to estimate missing LAI ASTER data. This study shows that the outputs obtained with the BN were more accurate than the 3-PG estimate, as the root mean square error reduces to 0.46, and the relative error to 5.86%. We conclude that the EM-algorithm within a BN can adequately handle missing LAI ASTER values, and BNs can improve the estimation of LAI values. Ultimately, this method may be used as a predicting model of LAI values, and handling the missing data of ASTER images.
5.1 Introduction

Forests play a critical role in carbon sequestration (Wamelink et al., 2009), thus affecting the speed of climate change. Therefore, monitoring forest growth has received increasing attention (Bonan, 1993). An important parameter in observing forest growth is the leaf area index (LAI), defined as the ratio of leaf area to per unit ground surface area. It is a key biophysical variable influencing land surface photosynthesis, respiration, transpiration, leaf litterfall, and energy balance (Bonan, 1993). The LAI is estimated using process-based models, such as the Physiological Principles Predicting Growth (3-PG) model, being a stand-level model of forest growth. This model developed by Landsberg and Waring (1997) and has been used as a point mode. A grid mode (spatial) version of 3-PG model has been developed by Coops et al. (1998). Similarly, remote sensing imagery has been added to estimate LAI, for example from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) satellite (Heiskanen, 2006; Ito et al., 2007).

Statistical methods have been used to estimate LAI, in particular Bayesian networks (BNs) (Kalacska et al., 2005; Mustafa et al., 2011b). A BN is a directed acyclic graph (DAG) consisting of nodes and arcs, to represent variables and the dependencies between variables, respectively (Jensen and Nielsen, 2007). Gaussian Bayesian network (GBN) have improved LAI estimates by combining the 3-PG model output with MODIS images (Mustafa et al., 2011b). Their approach relies on availability of satellite images. Satellite images, however, often contain gaps (missing data) due to atmospheric characteristics. Mustafa et al. (2011a) integrated an EM-algorithm with GBN to estimate missing satellite date.

The objective of this work is two-fold. First, we use the GBN model to estimate LAI values using the finer resolution satellite ASTER imagery with the point mode 3-PG model. Second, we develop a spatial GBN model to improve estimation of LAI by considering a spatial version of 3-PG model bases on 15 m ASTER resolution.

ASTER imagery has been employed in this work because of its relatively high spatial resolution in the visible to near infrared bands, and high spectral resolution in the short-wave infrared bands. Provision of higher order data products, such as atmospherically-corrected surface reflectance data, is increasing the applicability of ASTER images (Abrams, 2000). Furthermore, ASTER imagery may help to reduce the uncertainty of the estimated LAI by GBN.

5.2 Materials and methods

5.2.1 The 3-PG model

The 3-PG model is a process-based stand-level model of forest growth. A full description of the 3-PG has been provided by Landsberg and Waring (1997). It requires few parameters, and site and climatic data as
5. Improving forest growth estimates using a Bayesian network approach

inputs. Its primary output variables are net primary production (NPP), the standing biomass in foliage, stem (i.e., all above-ground woody tissue) and roots, stem number, available soil water and evapotranspiration. It infers LAI (LAI_{3PG}), mean stem diameter at breast height, main stem volume, and mean annual increment. The 3-PG model has been modified to apply in a large forested areas. This is demonstrated by using spatial data bases, geographic information systems (GIS) or remote sensing and considered a spatial version of the 3-PG model (White et al., 2000; Coops and Waring, 2001b; Tickle et al., 2001). The spatial 3-PG produces spatially and temporally explicitly outputs at the scale of the input surfaces. Spatial outputs include variables such as above and below ground biomass, LAI, stem volume, and current annual increment. These models have been used in many areas like eastern Brazil, and British Columbia (Landsberg et al., 2001; Almeida et al., 2010; Coops et al., 2010). However, it is difficult to parameterize the models precisely, e.g., to minimize the uncertainty of the model output.

5.2.2 Remote sensing imagery

The estimation of LAI by satellite remote sensing, in particular ASTER sensor, has been investigated in several studies at various spatial scales and environments (Peng et al., 2003; Heiskanen, 2006; Ito et al., 2007; Zheng and Moskal, 2009). The ASTER instrument acquires surface data in the visible to near infrared (VNIR; three bands at 15 m/pixel), shortwave infrared (SWIR; six bands at 30 m/pixel), and thermal infrared (TIR; five bands at 90 m/pixel) wavelength regions of the electromagnetic spectrum (Abrams, 2000). Each ASTER scene captures a 60 km × 60 km area.

Estimation of the forest variables using satellite images has been based on empirical relationships formulated between the forest variables measured in the field and satellite data, often expressed in the form of spectral vegetation indices (SVI). Peng et al. (2003) compared twelve different vegetation indices (ranging from visible to shortwave infrared bands) with LAI and found that modified non-linear vegetation index (MNLI), simple ration (RS), and normalized vegetation index (NDVI) correlates best with LAI. Current techniques for estimating LAI often failed to provide consistent values. Furthermore, most LAI satellite data (LAI_{SAT}) products are not continuous in space and time because of a cloud contamination and an insufficient number of data points for retrieval.

5.2.3 Bayesian networks

A Bayesian network (BN) is a probabilistic graphical model that provides a graphical framework of complex domains with many of inter-related variables (Jensen and Nielsen, 2007). Mustafa et al. (2011b) designed a network to improve LAI estimation by combining LAI values derived from satellite images and estimated by the point mode 3-PG model. Figure 5.1 (a) shows the graphical part of BN. The intermediate node (LAI_{BN}) repres-
5.2. Materials and methods

ent the estimated LAI values of BN. Based on the continuous variation of LAI over time, it has shown in (Mustafa et al., 2011b) that LAI follows a Gaussian distribution. The common type of a BN containing continuous variables is the GBN (Shachter and Kenley, 1989).

A GBN is a BN where the joint probability distribution associated with its variables $LAI = \{LAI_1, ..., LAI_n\}$ is the multivariate Gaussian distribution $N(\mu, \Sigma)$, given by

$$f(LAI) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (LAI - \mu)^T \Sigma^{-1} (LAI - \mu) \right\}. \quad (5.1)$$

Here $\mu$ is the $n$-dimensional mean vector, and $\Sigma$ is the $n \times n$ positive definite covariance matrix with determinant $|\Sigma|$.

The conditional probability distribution of the $LAI_i$ represented by the $LAI_{BN_i}$ as the variable of interest given its parentage, is the univariate Gaussian distribution with density

$$f(LAI_{BN_i} | pa_i) \sim N(\mu_i + \sum_{j=1}^{#pa_i} \beta_{ij} (pa_{ij} - \mu_{pa_{ij}}), \nu_i), \quad (5.2)$$

where $\mu_i$ is the expectation of $LAI_{BN_i}$ at time $i$, the $\beta_{ij}$ are a regression coefficients of $LAI_{BN_i}$ on its parents, $#pa_i$ is the number of parents of $LAI_{BN_i}$, and $\nu_i = \Sigma_i - \Sigma_{ipa_i}^{-1} \Sigma_i$ is the conditional variance of $LAI_{BN_i}$ given its parents. Further, $\Sigma_i$ is the unconditional variance of the $LAI_{BN_i}$, $\Sigma_{ipa_i}$ are the covariances between $LAI_{BN_i}$ and the variables $pa_i$, and $\Sigma_{pa_i}$ is the covariance matrix of $pa_i$.

For $i \geq 2$, Mustafa et al. (2011b) defined an equation to identify the intermediate node of the GBN based on the contribution of all of the satellite data, 3-PG output, and precedent GBN output as

$$LAI_{BN_i} = \rho (1 - \tau) LAI_{SAT_i} + \tau LAI_{3PG_i} + (1 - \rho)(1 - \tau) LAI_{BN_{i-1}}, \quad (5.3)$$

where $\tau$ and $\rho$ are the weighing values, defined as $\tau = \frac{|LAI_{SAT_i} - LAI_{SAT_{i-1}}|}{LAI_{SAT_{i-1}}}$ and $\rho = \frac{|LAI_{3PG_i} - LAI_{3PG_{i-1}}|}{LAI_{3PG_{i-1}}}$. They are proportional to the change in the LAI values obtained from the satellite images and 3-PG output. Hence, the conditional distribution of $LAI_{BN_i}$ equals:

$$LAI_{BN_i} \sim N\left(\mu_{LAI_{BN_i}} + \beta_{LAI_{BN_i}LAI_{SAT_i}} (LAI_{SAT_i} - \mu_{LAI_{SAT_i}}) + \beta_{LAI_{BN_i}LAI_{BN_{i-1}}} (LAI_{BN_{i-1}} - \mu_{LAI_{BN_{i-1}}}) + \beta_{LAI_{BN_i}LAI_{3PG_i}} (LAI_{3PG_i} - \mu_{LAI_{3PG_i}}), \Sigma_{LAI_{BN_i}} \right). \quad (5.3)$$

For more details about a GBN of improving forest growth estimates, we refer to (Mustafa et al., 2011b).

In the current study, we modified the equation of the intermediate node in GBN model in order to model spatio-temporal $LAI_{BN_i}$ value by
contribution from all of the satellite data, 3-PG output, and precedent GBN output within two consecutive moments as follows:

\[
LAI_{Bn}^k = \begin{cases} 
\alpha - \lambda (1 - \alpha) LAI_{3PG}^k + \alpha LAI_{BN}^{k-1} + \lambda (1 - \alpha) LAI_{BN}^k, & \text{if } \left| \frac{LAI_{3PG}^k - LAI_{BN}^{k-1}}{LAI_{BN}^k} \right| \leq \left| \frac{LAI_{BN}^k - LAI_{BN}^{k-1}}{LAI_{BN}^k} \right|, \\
\lambda (1 - \alpha) LAI_{3PG}^k + \alpha LAI_{BN}^{k-1} + (1 - \lambda) (1 - \alpha) LAI_{BN}^k, & \text{Otherwise}
\end{cases}
\]  

(5.4)

Here \( i \) is the image number (the moment number), \( k = (n, m) \) is the pixel location, and \( \alpha = \gamma \delta, \lambda = \omega \eta \), where \( \gamma, \delta, \omega, \text{ and } \eta \) are the weights, defined as: 

\[
\gamma = \left| \frac{LAI_{3PG}^k - LAI_{BN}^{k-1}}{LAI_{BN}^k} \right|, \quad \delta = \left| \frac{LAI_{BN}^k - LAI_{BN}^{k-1}}{LAI_{BN}^{k-1}} \right|, \quad \omega = \left| \frac{LAI_{3PG}^k - LAI_{3PG}^{k-1}}{LAI_{3PG}^{k-1}} \right|, \quad \eta = \left| \frac{LAI_{BN}^k - LAI_{BN}^{k-1}}{LAI_{3PG}^{k-1}} \right|.
\]

They are proportional to the spatial and temporal change in the LAI values obtained from the satellite images and 3-PG output. Equation (5.4) includes the GBN output from the precedent iteration \( (LAI_{BN}^k) \) to ensure that the LAI values in the new node are consistent with the \( LAI_{3PG} \), \( LAI_{BN} \), and the \( LAI_{BN} \) at the precedent iteration. This is based on the assumption that the LAI values do not sharply change in a short period of time. In fact, this choice for \( \eta \) and \( \delta \) addresses the difference between \( LAI_{3PG} \) values, and \( LAI_{3PG} \) at two consecutive time steps. Moreover, \( \omega \) and \( \gamma \) consider the difference between the \( LAI_{3PG} \) and \( LAI_{3PG} \) at two neighbor pixels. Weighing these values (Equation (5.4)) reduces the impact of large discrepancies between the LAI values of 3-PG and satellite images.

### 5.2.4 EM-algorithm for estimating missing values in a GBN

The Expectation Maximization (EM)-algorithm is a technique for estimating parameters of statistical models from incomplete data. The EM-
algorithm is applicable for maximizing likelihoods. The EM-algorithm is formulated and applied to handle the problem of missing satellite data by estimating the missing parameters that are needed to implement a GBN approach (Mustafa et al., 2011a). In this section we give in brief the derivation of EM-algorithm with GBN model.

Consider missing data of satellite images at the $i$th moment ($i > 1$) of the GBN as shown in Figure 5.1 (a). The GBN output, $LAI_{BN}$, conditionally depends on three nodes (variables), i.e., $LAI_{SAT}$, $LAI_{BN, i-1}$, and $LAI_{3PG}$, where $LAI_{SAT}$ is considered as a missing value. Let $(X, Y)$ be the complete data set at the $i$th moment of GBN, with observed (complete) data $Y = \{LAI_{BN, i-1}, LAI_{3PG}, LAI_{BN, i}\}$ and missing data $X = LAI_{SAT, i}$ (Figure 5.1 (b)). For clarity, we rename the variables in the GBN model as $y = LAI_{BN, i}$, $x = LAI_{SAT, i}$, $z = LAI_{BN, i-1}$, and $w = LAI_{3PG, i}$, such that the model parameters of the GBN are named $\mu_j$, $\sigma_j$ and $\beta_{yj}$ for $j \in \{x, y, z, w\}$.

Hence Equation (5.3) can be reformulated as:

$$y \sim N \left( \mu_y + \beta_{yx} (x - \mu_x) + \beta_{yz} (z - \mu_z) + \beta_{yw} (w - \mu_w), \sigma_y^2 \right).$$

The EM-steps to find new ML estimates for the parameters $\theta = (\mu_x, \Sigma_x)$ are as follows.

1. Choose an initial setting for the parameters $\theta$, and name it as $\theta^{old} = (\mu^{old}, \Sigma^{old})$. These are guessed based on seasonal changes of LAI values that are obtained from satellite images as

$$\theta^{old} = \begin{cases} \left( \mu_x - \frac{\mu_x - \mu_{x, i-1}}{\mu_{x, i-1}}, \sigma_x - \frac{\sigma_x - \sigma_{x, i-1}}{\sigma_{x, i-1}} \right) & \text{if } \mu_{x, i-2} \leq \mu_{x, i-1} \\ \left( \mu_x + \frac{\mu_x - \mu_{x, i-1}}{\mu_{x, i-1}}, \sigma_x + \frac{\sigma_x - \sigma_{x, i-1}}{\sigma_{x, i-1}} \right) & \text{Otherwise} \end{cases}$$

where $\mu_x$ and $\sigma_x$ are the mean and the standard deviation values of the $LAI_{SAT}$, respectively. They are obtained either for the period from September to February (the non growing season), or for the period from March to August (the growing season). The determination of the period for which we need to obtain the $\mu_x$ and $\sigma_x$ is based on the occurrence of missing observations in that period.

In Equation (5.6), the $\frac{\mu_{x, i-2} - \mu_{x, i-1}}{\mu_{x, i-1}}$ and $\frac{\sigma_{x, i-2} - \sigma_{x, i-1}}{\sigma_{x, i-1}}$ are the relative changes of the mean and the standard deviation of the previous two $LAI_{SAT}$. Adding and subtracting these relative change values are based on the condition of increase or decrease the $LAI_{SAT}$ values during the period of non growing or growing season.

2. E-step: compute the expectation (with respect to the X data) of the likelihood function of the model parameters by including the missing variables as they were observed.
5. Improving forest growth estimates using a Bayesian network approach

\[ Q(\theta, \theta^{old}) = E_X[\log f(Y, X|\theta)|Y, \theta^{old}] \]
\[ = \int \log f(Y, X|\theta)f(X|Y, \theta^{old})dX \]
\[ = \int \log f(x, y, z, w|\theta)f(x|y, z, w, \theta^{old})dx \] (5.7)

3. M-step: compute the ML estimates of the parameters \( \theta \) by maximizing the expected likelihood obtained during the E-step, i.e., \( \theta^{new} = \arg \max_{\theta} Q(\theta, \theta^{old}) \).

Hence, by differentiation \( Q(\theta, \theta^{old}) \) with respect to \( \theta \), and solve the differentiation equations for \( \theta^{new} = (\mu^{new}_X, \sigma^{new}_X) \), the maximum values are found as

\[ \theta^{new} = \begin{cases} \mu^{new}_X = \frac{3 \phi + \Psi}{6 \lambda} + \frac{2(-3 \delta \lambda + \alpha^2)}{3 \lambda \sqrt{\phi + \Psi}} + \frac{\alpha}{3 \lambda} \\ \sigma^{new}_X = \frac{(\mu^{new}_X)^2 - \frac{2 C}{B} \mu^{new}_X + \frac{E}{B}}{\pi} \end{cases} \] (5.8)

where \( \phi = -36 \delta \alpha \lambda + 108 \eta \lambda^2 + 8 \alpha^3 \),
\( \Psi = 12 \sqrt{3} (4 \delta^3 \lambda - \delta^2 \alpha^2 - 18 \delta \alpha \lambda \eta + 27 \eta^2 \lambda^2 + 4 \eta \alpha^2 \lambda) \),
\( \lambda = AB \),
\( \alpha = 2 AC + DB \),
\( \delta = AE + B^2 + 2 DC \),
\( \eta = CB + dE \),
\( A = 4 a^2 \Omega \beta^2 \),
\( B = 4 a^2 \),
\( C = 2 a b \Omega \),
\( D = -4 a^2 \Omega (x - \mu_y - \beta_{yx} (z - \mu_z) - \beta_{yw} (w - \mu_w)) \beta_{yx} / \sigma^2 \),
\( E = 2 a \Omega + b^2 \Omega \).

Here \( \Omega = V \sqrt{\frac{\mu^2}{a} + \frac{d^2}{\pi a}} \), \( V = \frac{\sigma^2 + \beta^2_{yx} \sigma^{old} y^2}{\sqrt{2} \pi} \), \( a = \frac{1}{\gamma \sigma^{old} \beta^2_{yx}} \),
\( b = \frac{(y - \mu_y + \beta_{yx} \mu_z - \beta_{yw} (w - \mu_w)) \beta_{yx} \beta^{old}_{yx}}{\sigma^2} \),
\( \phi = \frac{\mu^{old}_{yx} \beta^{old}_{yx} \beta^{old}_{yx}}{\sigma^2} \), and \( c = b^2 / 4a^2 \).

4. Check for convergence of \( \theta^{new} \) values. If \( \left| \theta^{new} - \theta^{old} \right| \leq \varepsilon \) is not satisfied, then \( \theta^{old} \leftarrow \theta^{new} \) and the algorithm returns to M-step. \( \varepsilon \) is the stop criterion which has been selected to be \( 10^{-5} \).

For the calculation details of integration EM-algorithm with GBN approach, we refer to (Mustafa et al., 2011a).

5.3 Study area

The study area description was given in Section 1.7.1. In this study, we consider the area around the tower which consists of 2.5 ha Douglas fir trees only (Figure 5.2).
5.4 Data description

The study is primarily based on fieldwork measurements and the ASTER surface reflectance products (AST_07) data from 03 June 2010 to 16 October 2010.

5.4.1 Ground data

Ground data were collected at the observation tower of the Speulderbos forest, which is equipped with a weather station and various scientific instruments. The required data for 3-PG model are climate data (16 day mean temperature, solar radiation, rainfall, and frost days), site factors (site latitude, maximum available water stored in the soil, and soil fertility rating), initial conditions (stem, root, and foliage biomass; stocking; and soil water at specific time), and 3-PG parameters. The parameters values which are used in this study are obtained and used by Waring and McDowell (2002). The spatial datasets were required to implement the spatial 3-PG model in this work contained soil data as the stored water in the soil. These data were gridded, interpolated, converted into rasters and clipped to the study area at a 15 m resolution, using ESRI ArcGIS software version 10, same as ASTER resolution. In addition, the LAI was measured on the ground to validate the estimates of the $\text{LAI}_{\text{FN}}$ values. The LAI is indirectly measured from the canopy transmission by the inversion of the measurements of the photosynthetically active radiation (PAR) above and below the canopy. These LAI values are adjusted by a clumping factor. These adjusted LAI were considered as a true LAI ($\text{LAI}_{\text{FD}}$) and used for validation of GBN. For more details about calculating LAI at the Speulderbos forest, we refer to (Mustafa et al., 2011b). Furthermore, spatial measurements of the LAI were surveyed in June.
Improving forest growth estimates using a Bayesian network approach

A total 15 measurements were randomly taken at the study area using LAI-2000 Plant Canopy Analyzer (LI-COR). These measurements are considered as an effective LAI (eLAI<sub>FD</sub>). The ground points of the spatial LAI measurements were located using a handheld GPS device (Magellan Meridian Platinum). According to the manufacturer, the device should have an accuracy of 7 m for 95% of time. Furthermore, the GPS measurements were averaged over several minutes in order to enhance the accuracy. The spatial eLAI<sub>FD</sub> values used to establish the empirical relationship with ASTER data. We used effective LAI due to the fact that it is more closely associated with the nature of information gathered by above canopy remote sensors (Chen et al., 2004). Chen and Cihlar (1996) suggested using effective LAI for radiation interception consideration because eLAI<sub>FD</sub> is easier to measure than LAI<sub>FD</sub>, and because it has better correlation to the satellite vegetation indices than does LAI<sub>FD</sub>.

5.4.2 Satellite data

Three cloud free ASTER scenes of level 2 surface reflectance product (AST_07) were acquired from 03 June to 16 October 2010. Each scene was rectified using ten GCPs collected from 1:25 000 topographic map sheets with a final Root Mean Square Error (RMSE) of about 0.5 pixel (15 m). The VNIR bands of all images were resampled to a pixel size of 15 m<sup>2</sup> using nearest-neighbor resampling. A nonlinear relationship has been found between the calculated NDVI from ASTER images (NDVI<sub>SAT</sub>) and the spatial measurements of eLAI<sub>FD</sub> (eLAI<sub>FD</sub> = 0.74 exp(2.11NDVI<sub>SAT</sub>), R<sup>2</sup> = 0.80, Figure 5.3). The LAI is calculated based on the empirical relationship, and it has been considered as the LAI values from ASTER images (LAI<sub>AST</sub>). This LAI<sub>AST</sub> used in GBN as the input source of the satellite data.

5.5 Implementation

The GBN is applied to the Speulderbos forest in The Netherlands where the LAI<sub>FD</sub> is available as a time series from 03 June 2010 until 16 October 2010. Three ASTER images were available during the time study, whereas five images were missing due to the cloud coverage. The EM-algorithm is applied to estimate five missing LAI<sub>AST</sub> and used as an input into the GBN. The implementation of this study includes two cases. First is to implement the GBN with point mode 3-PG model for the full period.
of four months after missing LAI_{AST} estimated using the EM-algorithm. Second is to implement the GBN with spatial 3-PG model for one ASTER data.

5.6 Results

5.6.1 GBN performance with point mode 3-PG model

Figure 5.4 shows the LAI values estimated from the GBN for a period of four months, along with the LAI derived from the 3-PG model, LAI field data, and the LAI ASTER that were found from the empirical relationship between NDVI_{SAT} and eLAI_{FD}. The accuracy of the LAI 3-PG output is tested using the RMSE and the relative error (RE) rate with respect to the LAI field data. We found an RMSE of 1.40 and an RE of 18.7% (Table 5.1). The 3-PG overestimated the LAI values across the studied period. The LAI_{AST} shows a big difference with respect to the LAI_{FD}, with an RMSE and an RE of 3.08 and 45.8%, respectively (Table 5.1). The missing LAI_{AST} values are estimated with significant values such that no big deviation noticed with respect to the non-missing LAI_{AST} values. They are indicated in Figure 5.4 as a black square symbols. During the four months period, we notice that the combination of the LAI_{AST} values and LAI_{3PG} in a GBN reduces the RMSE to 0.46 and the RE to 5.8%.
5. Improving forest growth estimates using a Bayesian network approach

Table 5.1: Mean values (Mean), standard deviation ($\sigma$), Root Mean Square Errors (RMSE) and Relative Errors (RE) for the various ways to estimate the LAI.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>$\sigma$</th>
<th>RMSE</th>
<th>RE rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LAI_{FD}$</td>
<td>6.68</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$LAI_{AST}$</td>
<td>3.61</td>
<td>0.36</td>
<td>3.08</td>
<td>45.8%</td>
</tr>
<tr>
<td>$LAI_{3PG}$</td>
<td>7.88</td>
<td>0.40</td>
<td>1.40</td>
<td>18.7%</td>
</tr>
<tr>
<td>$LAI_{BN}$</td>
<td>6.51</td>
<td>0.51</td>
<td>0.46</td>
<td>5.8%</td>
</tr>
</tbody>
</table>

5.6.2 Spatial GBN performance with spatial 3-PG model

In this work, a full spatial LAI field dataset for the study area was not available at the ASTER resolution. The spatial output of the GBN is averaged over all pixels and compared with the $LAI_{FD}$ calculated from the PAR data at the tower. Therefore, the accuracy of the spatial GBN is assessed using the absolute error of the averaged spatial GBN with respect to the LAI field data at the tower. The absolute error between $LAI_{FD}$ and the $LAI_{AST}$, $LAI_{3PG}$, and $LAI_{BN}$ are 3.02, 1.65, and 0.64, respectively. The absolute error between $LAI_{BN}$ and $LAI_{FD}$ is lower than the absolute error between $LAI_{3PG}$ and $LAI_{FD}$. This indicates that the output of the GBN is more accurate than the 3-PG model output. Figure 5.5 (a), (b), and (c) show the spatial LAI estimated from ASTER images, spatial GBN output, and spatial 3-PG model of the study area.

5.7 Discussion and conclusions

In this work, the GBN with the point mode 3-PG model with the EM-algorithm is applied to the Douglas fir trees in the Speulderbos forest to estimate the missing LAI ASTER and improve LAI values during the period of four months (June till October 2010). Moreover, the spatial estimation of LAI with 15 m resolution is achieved for one ASTER image acquired in June 2010, by modifying the GBN to infer spatial LAI estimates using a
5.7. Discussion and conclusions

Our results show that the missing $LAI_{AST}$ is estimated with a small deviation among the non-missing $LAI_{AST}$ values. From the results we observed that the deviation of the GBN output and $LAI_{FD}$ is lower than the deviation between the 3-PG model output and $LAI_{FD}$, indicating that the LAI output of the GBN is more accurate than that of the 3-PG model output alone and is closer to the $LAI_{FD}$.

A major contribution of this work is to apply the methods of (Mustafa et al., 2011a,b) to finer spatial resolution images (ASTER). It shows the applicability of their methods with different satellite images. Moreover, the GBN is modified to infer 15 m spatial estimation of LAI using the spatial 3-PG model. It addresses the spatial LAI values pixel by pixel of ASTER image and 3-PG model output. As a strategy for the consideration of these two products, ASTER and 3-PG model, in a spatial GBN node, we resorted to the mathematical formulation in Equation (5.4). From this equation, we can identify the intermediate node of the spatial GBN based on the contribution of the ASTER images, 3-PG output, and precedent spatial GBN output. However, to account for the uncertainty in both ASTER images and 3-PG model, weighing factors are introduced. This new expression also includes the spatial GBN output of a precedent moment due to the fact that the LAI values have no large changes within a relatively short period of space and time.

A drawback of this work is that the spatial values of the LAI ground measurements of the study area are not available. This was a restriction during validation with respect to the spatial GBN and that was also the reason of applying the spatial GBN for one ASTER observation. However, this study can be applied to a small or a big area whenever the appropriate field data are available.

Furthermore, some other issues require further work. For instance, the time period of this study needs to be extended to include at least one year. This may show a better explanation of the model in terms of LAI seasonal change. Also, ground measurements should be collected for the study area at the same spatial and temporal resolution of ASTER images to use as a validation of the spatial GBN model. Further, the modified version of GBN needs to be applied with more than one satellite observation. This may be addressed by developing the EM-algorithm with the GBN of (Mustafa et al., 2011a) by integrating the EM-algorithm with the spatial version of the GBN to estimate the spatial missing LAI satellite data.

From the present work, we conclude that the GBN model can be applied with different satellite images. Moreover, spatial estimation of LAI can be done with the modified version of the GBN, such that the deviation of the averaged LAI output of spatial GBN and LAI field data is less than the deviation between the averaged LAI output of 3-PG and LAI field data.
Using Gaussian Bayesian networks to improve spatio-temporal growth estimates of heterogeneous forests

6.1 Introduction 91
6.2 Materials and data 92
  6.2.1 The spatial 3-PG model and its input requirements 92
  6.2.2 Remote sensing data 92
  6.2.3 LAI ground data 93
6.3 Methods 96
  6.3.1 Decomposing MODIS LAI 96
  6.3.2 Gaussian Bayesian network 97
6.4 Implementation 100
6.5 Results 101
  6.5.1 Decomposed \( \text{LAI}_M \) 101
  6.5.2 GBN 101
6.6 Discussion 106
6.7 Conclusion 107

6. **GBN to improve spatio-temporal growth estimates of heterogeneous forests**

**Abstract**

Canopy leaf area index (LAI), defined as one half the total leaf surface area per unit ground surface area projected on the local horizontal datum. It is a quantitative measure of canopy foliar area. Spatial estimation of LAI can be used to generate spatial estimates of carbon exchanges. It can be indicative of canopy structure responses to competition, disease and climate change, making it useful for efficient forest management. In this study, Gaussian Bayesian networks (GBNs) were applied to improve the spatio-temporal estimation of LAI of a heterogeneous forest located in The Netherlands. It combined the spatial version of physiological principles predicting growth (3-PG) model output with decomposed Moderate Resolution Imaging Spectroradiometer (MODIS) images. The Linear Mixture Model (LMM) was used to decompose MODIS pixels using class fraction derived from an aerial and an Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) images. In this way spatially heterogeneous output was produced. Results showed that the spatial output obtained with the GBN was more than 40% accurate than both the spatial 3-PG model output and the satellite estimate, with a root mean square error less than 0.53. We conclude that the GBNs can improve the spatial estimation of the LAI values of a heterogeneous forest by combining a spatial forest growth model with satellite images.

**Keywords**: Gaussian Bayesian networks (GBNs), leaf area index (LAI), Linear Mixture Model (LMM), mixed pixels, Moderate Resolution Imaging Spectroradiometer (MODIS), Physiological Principles Predicting Growth (3-PG) model.
6.1 Introduction

There is an increasing demand for reliable estimates of forest growth due to the role of forest in carbon sequestration. Quantitative forest growth assessment can be achieved by measuring and estimating biophysical parameters, such as the leaf area index (LAI). The LAI is defined as one half the total leaf surface area per unit ground surface area projected on the local horizontal datum (m$^2$·m$^{-2}$) (Watson, 1947). Estimation of LAI is important as it is related to ecological processes, rates of photosynthesis, transpiration and evapotranspiration (Leuning et al., 2005), and is found to be the most influencing biophysical parameter to primary productivity. A good estimate for LAI is helpful to determine above-ground carbon stock, which needed for carbon trading under “Clean development Mechanism” of the Kyoto protocol. Spatial estimation of LAI can be used to generate spatial estimates of carbon exchanges. It can be indicative of canopy structure responses to competition, disease and climate change, making it useful for efficient forest management.

Several methods to estimate this index are in general use i) measurements in the field, ii) forest growth models, iii) observations from remote sensing images. Field measurements of LAI are achieved either by direct or by indirect methods (Breda, 2003; Weiss et al., 2004). Direct methods are labor intensive and involve destructive sampling. Indirect methods are rather complicated as they include both light interception instrumentation and hemispherical photography. Further, the representative uses of field data remains problematic because of large spatial and temporal variability (Breda, 2003).

As an alternative, process-based models such as the Physiological Principles Predicting Growth (3-PG) model (Landsberg and Waring, 1997) have been developed to estimate the LAI. A spatial version of the 3-PG model has been developed by Coops et al. (1998) using spatial data bases. It requires site and climate data as inputs and it can generate several types outputs that can be used to understand the processes involved in tree growth, among which is the LAI. The 3-PG model output is subject to uncertainty due to 1) uncertainty in the model parameters and in the input data, and 2) due to the design of the model for a homogeneous forest, whereas by means of forests are heterogeneous in space and time.

The third form of LAI data collection is remote sensing imagery. For instance, the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor provides eight day global data sets of the LAI at a 1 km spatial resolution (Myneni et al., 2002). Satellite imagery, however, also has uncertainties due to 1) presence of the instrument noise during image acquisition, and 2) due to mixed pixels caused by heterogeneity of the forest.

Statistical methods based on graphical models, in particular Gaussian Bayesian networks (GBNs) are helpful to reduce uncertainties, thus attaining a more accurate estimated LAI value. Mustafa et al. (2011b) used a GBN to improve LAI estimates by integrating the 3-PG model output.
6. GBN to improve spatio-temporal growth estimates of heterogeneous forests

with MODIS data. Their approach is extended in (Mustafa et al., 2012) to improve the spatial estimation of LAI by further integrating a spatial version of 3-PG model based on a 15 m resolution with ASTER data, called a spatial GBN. Their approach is applied to a small homogeneous forested area and considers solely the variation in space.

The aim of this work is to improve LAI estimation for heterogeneous forests that show variation in space and time. To do so, we decompose an MODIS image and thus produce separate data for individual tree species. Thereafter we combine these data with the spatial 3-PG output using GBNs. The study is applied to a heterogeneous forest, called the Speulderbos, in The Netherlands.

6.2 Materials and data

6.2.1 The spatial 3-PG model and its input requirements

Landsberg and Waring (1997) developed a deterministic process-based forest growth model, called the 3-PG model, that is based on a number of established biophysical relationships. Coops and Waring (2001,b), Tickle et al. (2001), and White et al. (2000) have demonstrated the application of the 3-PG model to large forested areas using spatial data bases, geographic information systems (GIS) or remote sensing and considered a spatial version of the 3-PG model. Used in this way, the model produces spatially and temporally explicit outputs at the scale of the input. It requires climate and site data. The climate data used in this study are 16 days precipitation, solar radiation, and temperature from June, 2010 to August, 2011. These data were derived from three weather stations (Haarweg, Loobos, and Speulderbos). The climate data were gridded, interpolated, converted into rasters and clipped to the study area at a 250 m resolution using ESRI ArcGIS software version 10.

Two types of soil data are used as site data. These data are required to initialize the 3-PG model. The first type of data are soil fertility, which take values between 0 (very low fertility) and 1 (nutrition non-limiting). The second type of data are a measure of the maximum amount of the available water stored in the soil. As the detailed soil data for the entire study area were not available, we assumed a fixed value of soil fertility with a rank of 0.6, and we used the maximum available soil water was derived from the available soil data around the Speulderbos’s climate station.

The 3-PG model requires knowledge of parameters values for each tree species. Parameter values for the tree species were derived from the literature and assigned to each tree species as shown in Table 6.1.

6.2.2 Remote sensing data

There are several advantages gained from remote sensing imageries, although, they have also been criticized in terms of geometric formulation
6.2. Materials and data

of pixel, ground sampling distance (GSD), point spread function (PSF), mixed pixels, and resampling (Cracknell, 1998). In this study and for the sake of simplicity, we will assume that pixels have a square footprint, we will ignore the effect of PSF, and we will assume the GSD equals the size of the footprint.

Several studies have been carried out to estimate the LAI values from different satellite sensors. LAI estimation values can be provided by a spaceborne product, in particular MODIS. The MODIS sensor was launched in December 1999 on board the NASA EOS Terra satellite platform (Myneni et al., 2002). MODIS is designed to provide long-term global observations every 1 to 2 days at moderate spatial resolutions (250 m–1 km) (Myneni et al., 2002).

Both MODIS and ASTER data were used in this study. The MOD15A2 LAI/FPAR (Fraction of Photosynthetically Active Radiation) product is recorded at a nominal 1 km resolution and composited over 8 days (Myneni et al., 2002). In (Mustafa et al., 2011b), we derived LAI from the normalized difference vegetation index (NDVI) MODIS product (MOD13Q1) at a nominal 250 m spatial resolution and composited over 16 days. The derived 250 m LAI was used in this study as the LAI value from MODIS in order to match with the temporal frequency of the 3-PG model and to produce LAI at a finer spatial resolution (Section 1.7.2).

A cloud free ASTER scene of level 2 surface reflectance product (AST_07) was acquired on 21 May 2011. This level 2 surface reflectance product contains atmospherically corrected data for both the Visible Near-Infrared (VNIR) and Shortwave Infrared (SWIR) sensors. The scene was rectified using 25 GCPs collected from 1:25000 topographic map sheets with a Root Mean Square Error (RMSE) of approximately 0.6 pixel (15 m). The use of ASTER image in this study is to identify the area proportion of each tree species within a MODIS pixel.

Table 6.1: The 3-PG parameters of each species were determined from the following sources.

<table>
<thead>
<tr>
<th>Tree species</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beech</td>
<td>Lebaube et al. (2000), Nole et al. (2009)</td>
</tr>
<tr>
<td>Douglas fir</td>
<td>Waring and McDowell (2002)</td>
</tr>
<tr>
<td>Hemlock</td>
<td>Mencuccini and Grace (1996); Breuer et al. (2003)</td>
</tr>
<tr>
<td>Japanese larch</td>
<td>Liang et al. (2004); Toda et al. (2007)</td>
</tr>
<tr>
<td>Scotch pine</td>
<td>Mencuccini and Grace (1996); Xenakis et al. (2008)</td>
</tr>
</tbody>
</table>

6.2.3 LAI ground data

6.2.3.1 Spatial sampling

Design of a spatial sampling strategy is a key issue when performing ground measurements that need to be representative for the whole canopy. It includes aspects that reflect the heterogeneous nature of the area, and the area of the each class (McCoy, 2005). The sampling strategy
is critical for getting an accurate characterization of the heterogeneity in the study area.

An area of 1 km² around the tower of the Speulderbos forest has been considered in this study, Figure 6.1. A full description of study area was given in Sections 1.7.1. Prior to the field work, a natural color aerial image with a 0.5 m resolution acquired in summer of 2011 was used. This image interpreted visually to explore the forest cover species, and then the sample points were distributed for each species, Figure 6.1(d). Data sampling in this study was designed following a stratified random sampling method in order to assure a reliable representation of the existent LAI variability, paying special attention to the possible loss in efficiency with inappropriate stratification (Fassnacht et al., 1997). Using the stratified random sampling method, we assigned a specific number of sample points to each tree species based on the proportion to the area that’s covered by that species. Thus, each sample point occurs within an area not greater than 1000 m² for a particular species. This is done to ensure that each cover species has a sufficient number of sample points (McCoy, 2005). In total 155 sample points were distributed over the study area, stratified by the tree species (Figure 6.1(d)).

LAI ground measurements were collected five times within a 16 days period, from the end of April to the end of June 2011. An initial survey of the forest area was carried out in March 2011 before the forest canopy
got dense. This was done to locate the sample points using two handheld GPS devices (Garmin eTrx and MobileMapper™). These locations were labeled by marking the trees in order to take consecutive LAI measurements within every 16 days at the same locations. The GPS devices had an accuracy within a range of ±3 m and 2–5 m for 95% of time as confidence interval, respectively. Furthermore, the GPS measurements were averaged over a period of several minutes in order to enhance the accuracy.

6.2.3.2 Methods and measurements

Optical techniques are commonly used indirect technique for in-situ LAI estimates. These are fast and easy ways of sampling over large spatial areas (Morisette et al., 2006). In this study, the Digital Hemispherical Photographs (DHPs) technique was used for LAI measurements. This optical technique has proven to be robust in terms of low sensitivity to illumination conditions, independence from ancillary information on canopy optical properties and potential to derive clumping index (Weiss et al., 2004).

The LAI is computed from the hemispherical photographs from gap fraction estimates in different zenithal and azimuthal ranges. We used a Canon EOS 5D camera (21 megapixel) with a 15 mm F2.8 EX fisheye lens and a 180° field-of-view (15 mm, f/2.8, Sigma, Tokyo, Japan). The camera with fisheye lens was held at 1.3 m above the ground and mounted on the tripod and pointed at 90° in the vertical direction. The camera was aligned such that the front face of the camera faced East, and the left side of the image was directed towards the North. Each sample point location was the center of three photo locations with a distance of 5 m. These three photos were averaged after the photos were processed. All measurements were made under diffuse light conditions to avoid introducing errors due to the presence of sunlit foliage.

The DHP data were processed using the Gap Light Analyzer (GLA) software developed by Institute of Ecosystem Studies in the USA. Canopy structure data (gap fraction, canopy openness, effective LAI) and gap light transmission indices could be extracted based on a user-specified day of interest. The GLA has been used in several similar applications with satisfactory results, see (Hardy et al., 2004) and the references therein. The LAI values derived from optical techniques are called the effective LAI. The true LAI can be adjusted from the effective LAI values by a clumping index. The clumping index quantifies the effect of nonrandom spatial distribution of foliage on LAI measurements (Weiss et al., 2004). More information concerning the performance of the camera, lens setup and GLA software for hemispherical photo analysis was presented in (Frazer et al., 2001).

Furthermore, the LAI is measured indirectly from the canopy transmission by the inversion of the measurements of the photosynthetically active radiation (PAR) above and below the canopy. The PAR data are acquired from the Speulderbos climate station using four sensors placed
6. GBN to improve spatio-temporal growth estimates of heterogeneous forests

at the tower. The tower is located within a dense 2.5 ha Douglas fir stand. Two sensors are positioned at the top of the tower to record the PAR outside the canopy forest. The other two sensors are positioned below the forest canopy to record the PAR inside the forest. PAR data are recorded every ten minutes during daytime. Calculation of LAI from PAR data is based on the relationship between leaf area and light transmittance, described by the Beer-Lambert model (Maass et al., 1995). The daily average LAI ground data has been computed from the recorded PAR data after LAI calculated at solar noon time. These LAI data were available for a period of 14 months from June 26, 2010 to August 29, 2011 and considered as a temporal LAI ground data of Douglas fir. For more details of calculating LAI from PAR data, we refer to (Mustafa et al., 2011b). For validation, the LAI ground measurements are adjusted by a clumping factor and were labeled as LAI field data ($LAI_{FD}$). The clumping factor for each tree species was derived from the literatures (Smith, 1993; Stenberg et al., 1994; Walter et al., 2003; Wagner and Hagemeier, 2006).

6.3 Methods

6.3.1 Decomposing MODIS LAI

To decompose the MODIS LAI values towards a finer grid, we used the Linear Mixture Model (LMM). This LMM assumes that the spectrum measured by a sensor is a linear combination of the spectra of all components within the pixel (Brown et al., 2000). Researchers have applied this model for land cover estimation and classification (Brown et al., 2000; Tolpekin and Stein, 2009). According to its definition, the LAI over a specific area can be interpreted as a linear composition of LAIs in subareas. This is achieved by considering the LAI MODIS pixel as a linear combination of LAI values of all species present in the MODIS pixel and assuming that the LAI of one species is the same for the entire study area. The LMM is thus expressed as

$$y_k = \sum_j A_{k,j} z_j + \epsilon_k,$$ (6.1)

where $y_k$ refers to the LAI MODIS for pixel $k$ ($k = 1, ..., 16$), $A_{k,j}$ is a matrix of the area proportion values of species $j$ ($j = 1, ..., 5$) in pixel $k$, $z_j$ is the LAI of tree species $j$, and $\epsilon_k$ is the residual for pixel $k$. The area proportion in the MODIS pixel is between 0 and 1, i.e., $\sum_j A_{k,j} = 1$.

A reference grid of 15 m pixel size was constructed by manual digitizing forest species. Digitization was supported on visual interpretation of the ASTER and aerial images. Area proportions of each species within MODIS pixels were derived from the reference grid, Figure 6.2. This was done by counting the number of reference pixels of each species within MODIS pixel. Table 6.2 shows the area proportions, and the number of ASTER pixels for each species in the composition of MODIS pixels.
### 6.3. Methods

In this way, the explanatory variables in Equation (6.1) are the LAI MODIS of each MODIS pixel and the area proportion of each species in the MODIS pixels. Hence, solving Equation (6.1) using least squares leads to retrieving LAI values of each species in the MODIS image. The resulting LAI values by the LMMs were considered the decomposed LAI MODIS and labeled as $LAI_{SAT}$. The resulting residual values were used to assess the fit of the model. Residual values were also used as a standard deviation in the spatial GBN for the resulting LAI values by LMM of each species.

We applied this procedure to all MODIS images that were used in this study between July 26, 2011 and August 13, 2011, by assuming that there is no major change in the forest composition as identified by the area proportion of each species during the time period of this study.

![Figure 6.2](image)

**Figure 6.2:** (a) RGB image of the 1-km of the study area from ASTER bands 3, 2 with the tree species map. (b) Map of the 1-km region using manual segmentation with an overlaid MODIS grid of 250 m resolution.

#### 6.3.2 Gaussian Bayesian network

A Bayesian networks (BNs) is a probabilistic graphical model useful to study a set of random variables with a specified dependence structure (Jensen and Nielsen, 2007). Formally, a BN is defined by a pair $(G, P)$ where $G$ is a directed acyclic graph (DAG), with nodes representing random variables $X = \{X_1, ..., X_n\}$, and arrows representing probabilistic dependencies between variables, and $P = \{P(X_1 | pa_1), ..., P(X_n | pa_n)\}$ is a set of conditional probability densities (one for each variable) with $pa_i$ the set of parents of node $X_i$ in $G$.

In the current work, the variable of interest is the LAI. Under the assumption as shown in (Mustafa et al., 2011b) that LAI follows the...
6. **GBN to improve spatio-temporal growth estimates of heterogeneous forests**

### Table 6.2: The area proportion (A), and the number of referenced pixels (N) for each species (Beech, Douglas fir, Hemlock, Japanese Larch, and Scotch Pine) in the composition of MODIS pixels.

<table>
<thead>
<tr>
<th>MODIS pixel (x,y)</th>
<th>Beech N</th>
<th>Beech A</th>
<th>Douglas fir N</th>
<th>Douglas fir A</th>
<th>Hemlock N</th>
<th>Hemlock A</th>
<th>Japanese Larch N</th>
<th>Japanese Larch A</th>
<th>Scotch Pine N</th>
<th>Scotch Pine A</th>
<th>No trees N</th>
<th>No trees A</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>84</td>
<td>0.29</td>
<td>141</td>
<td>0.49</td>
<td>0</td>
<td>0.00</td>
<td>18</td>
<td>0.06</td>
<td>33</td>
<td>0.11</td>
<td>13</td>
<td>0.04</td>
</tr>
<tr>
<td>(1,2)</td>
<td>4</td>
<td>0.01</td>
<td>163</td>
<td>0.56</td>
<td>0</td>
<td>0.00</td>
<td>117</td>
<td>0.40</td>
<td>0</td>
<td>0.00</td>
<td>5</td>
<td>0.02</td>
</tr>
<tr>
<td>(1,3)</td>
<td>50</td>
<td>0.17</td>
<td>22</td>
<td>0.08</td>
<td>44</td>
<td>0.15</td>
<td>151</td>
<td>0.52</td>
<td>0</td>
<td>0.00</td>
<td>22</td>
<td>0.08</td>
</tr>
<tr>
<td>(1,4)</td>
<td>256</td>
<td>0.89</td>
<td>0</td>
<td>0.00</td>
<td>19</td>
<td>0.07</td>
<td>8</td>
<td>0.03</td>
<td>6</td>
<td>0.02</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(2,1)</td>
<td>134</td>
<td>0.46</td>
<td>51</td>
<td>0.18</td>
<td>15</td>
<td>0.05</td>
<td>68</td>
<td>0.24</td>
<td>21</td>
<td>0.07</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(2,2)</td>
<td>0</td>
<td>0.00</td>
<td>21</td>
<td>0.07</td>
<td>123</td>
<td>0.43</td>
<td>143</td>
<td>0.49</td>
<td>0</td>
<td>0.00</td>
<td>2</td>
<td>0.01</td>
</tr>
<tr>
<td>(2,3)</td>
<td>8</td>
<td>0.03</td>
<td>0</td>
<td>0.00</td>
<td>110</td>
<td>0.38</td>
<td>136</td>
<td>0.47</td>
<td>18</td>
<td>0.06</td>
<td>17</td>
<td>0.06</td>
</tr>
<tr>
<td>(2,4)</td>
<td>163</td>
<td>0.56</td>
<td>0</td>
<td>0.00</td>
<td>46</td>
<td>0.16</td>
<td>18</td>
<td>0.06</td>
<td>62</td>
<td>0.21</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(3,1)</td>
<td>106</td>
<td>0.37</td>
<td>0</td>
<td>0.00</td>
<td>11</td>
<td>0.04</td>
<td>143</td>
<td>0.49</td>
<td>0</td>
<td>0.00</td>
<td>29</td>
<td>0.10</td>
</tr>
<tr>
<td>(3,2)</td>
<td>29</td>
<td>0.10</td>
<td>79</td>
<td>0.27</td>
<td>44</td>
<td>0.15</td>
<td>137</td>
<td>0.47</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(3,3)</td>
<td>145</td>
<td>0.50</td>
<td>24</td>
<td>0.08</td>
<td>0</td>
<td>0.00</td>
<td>15</td>
<td>0.05</td>
<td>85</td>
<td>0.29</td>
<td>20</td>
<td>0.07</td>
</tr>
<tr>
<td>(3,4)</td>
<td>77</td>
<td>0.27</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>15</td>
<td>0.05</td>
<td>174</td>
<td>0.60</td>
<td>23</td>
<td>0.08</td>
</tr>
<tr>
<td>(4,1)</td>
<td>201</td>
<td>0.70</td>
<td>71</td>
<td>0.25</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>17</td>
<td>0.06</td>
</tr>
<tr>
<td>(4,2)</td>
<td>72</td>
<td>0.25</td>
<td>52</td>
<td>0.18</td>
<td>0</td>
<td>0.00</td>
<td>140</td>
<td>0.48</td>
<td>0</td>
<td>0.00</td>
<td>25</td>
<td>0.09</td>
</tr>
<tr>
<td>(4,3)</td>
<td>153</td>
<td>0.53</td>
<td>0</td>
<td>0.00</td>
<td>99</td>
<td>0.34</td>
<td>3</td>
<td>0.01</td>
<td>34</td>
<td>0.12</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>(4,4)</td>
<td>154</td>
<td>0.53</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>19</td>
<td>0.07</td>
<td>42</td>
<td>0.15</td>
<td>74</td>
<td>0.26</td>
</tr>
</tbody>
</table>

**Figure 6.3:** An example of a GBN output of one pixel with spatial scale of 250 m consists of three types. \( \text{LAI}_{\text{SAT}} \) refers to the decomposed LAI MODIS for each tree species, \( \text{LAI}_{3\text{PG}} \) is the LAI output from spatial version of the 3-PG model, and \( \text{LAI}_{\text{BN}} \) is the spatial output of GBN.
Gaussian distribution, our attention focused on the use of a Gaussian Bayesian network (GBN). A GBN is a BN where the joint probability distribution associated with its variables \( \text{LAI} = \{ \text{LAI}_1, ..., \text{LAI}_n \} \) (Figure 6.3 (a)) is the multivariate Gaussian distribution \( N(\mu, \Sigma) \), given by
\[
 f(\text{LAI}) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\text{LAI} - \mu)^T \Sigma^{-1} (\text{LAI} - \mu) \right\}.
\]
Here \( \mu \) is the \( n \)-dimensional mean vector, and \( \Sigma \) is the \( n \times n \) positive definite covariance matrix with determinant \( |\Sigma| \).

The conditional probability distribution of the LAI_i represented by the LAI_{BN,i}, as the variable of interest given its parentage, is the univariate Gaussian distribution with density
\[
 f(\text{LAI}_{BN,i}|p_{ai}) \sim N \left( \mu_i + \sum_{j=1}^{#p_{ai}} \beta_{ij}(p_{aij} - \mu_{p_{aij}}), \nu_i \right),
\]
where \( \mu_i \) is the expectation of LAI_{BN,i} at time \( i \), \( \beta_{ij} \) represents the regression coefficients of LAI_{BN,i} on its parents, \( #p_{ai} \) is the number of parents of LAI_{BN,i}, and \( \nu_i = \Sigma_i - \sum_{j=1}^{#p_{ai}} \Sigma^{-1}_{p_{ai}} \Sigma_{p_{ai}} \) is the conditional variance of LAI_{BN,i} given its parents. Further, \( \Sigma_i \) is the unconditional variance of the LAI_{BN,i}, \( \Sigma_{p_{ai}} \) are the covariances between LAI_{BN,i} and the variables \( p_{ai} \), and \( \Sigma_{p_{ai}} \) is the covariance matrix of \( p_{ai} \).

The regression coefficients relating LAI_{BN,i} to LAI satellite data (LAI_{SAT,i}), precedent GBN output (LAI_{BN,i-1}), and LAI output of 3-PG model (LAI_{3PG,i}) can be expressed as \( \beta_{LAI_{BN,i},LAI_{SAT,i}}, \beta_{LAI_{BN,i},LAI_{BN,i-1}}, \) and \( \beta_{LAI_{BN,i},LAI_{3PG,i}} \), respectively. Moreover, the expectation of LAI_{SAT,i}, LAI_{BN,i-1}, and LAI_{3PG,i} are also expressed as \( \mu_{LAI_{SAT,i}}, \mu_{LAI_{BN,i-1}}, \) and \( \mu_{LAI_{3PG,i}} \), respectively. Hence, for \( i \geq 2 \) and based on the graphical part of GBN as in Figure 6.3 (a) or (b), or (c), the conditional distribution of LAI_{BN,i} equals
\[
 \text{LAI}_{BN,i} \sim N \left( \mu_{LAI_{BN,i}} + \beta_{LAI_{BN,i},LAI_{SAT,i}} (LAI_{SAT,i} - \mu_{LAI_{SAT,i}}) + \beta_{LAI_{BN,i},LAI_{BN,i-1}} (LAI_{BN,i-1} - \mu_{LAI_{BN,i-1}}) + \beta_{LAI_{BN,i},LAI_{3PG,i}} (LAI_{3PG,i} - \mu_{LAI_{3PG,i}}), \Sigma_{LAI_{BN,i}} \right).
\]

The posterior probability of LAI_{BN,i} in Equation (6.3) is conditionally determined by three parents LAI_{SAT,i}, LAI_{3PG,i}, and LAI_{BN,i-1}. These parents are used as prior input information into the GBN and has been tested on Gaussianity by Mustafa et al. (2011b), in terms of the temporal and the spatial resolution. They used the Shapiro-Wilk and Lilliefors tests for that purpose. Moreover, the prior information of the intermediate node in Figure 6.3 (a) or (b), or (c) has been defined to modeled the spatio-temporal LAI_{BN,i} value by contribution from all of the satellite data, 3-PG output and precedent GBN output (Mustafa et al., 2012) as in Equation (6.4),
\[
 \text{LAI}_{BN,i} = \begin{cases} 
 (1 - \lambda_1) \text{LAI}_{3PG,i} + \alpha \text{LAI}_{3PG,i} + \lambda_1 \text{LAI}_{BN,i-1}, & \text{if } |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i-1}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i-1}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i-1}| \\
 \lambda_1 \text{LAI}_{3PG,i} + \alpha \text{LAI}_{3PG,i} + (1 - \lambda_1) \text{LAI}_{BN,i-1}, & \text{if } |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i-1}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i-1}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i}| \leq |\text{LAI}_{3PG,i} - \text{LAI}_{BN,i-1}| \\
 \lambda_1 \text{LAI}_{3PG,i} + \alpha \text{LAI}_{3PG,i} + (1 - \lambda_1) \text{LAI}_{BN,i-1}, & \text{otherwise}
\end{cases}
\]
(6.4)
Here \( i \) is the image number (the moment number), \( k = (n, m) \) is the pixel location, and \( \alpha = \gamma\delta, \lambda = \omega\eta, \) where \( \gamma, \delta, \omega, \) and \( \eta \) are the weights, defined as \( \gamma = \frac{|LAI_{k_i}^{BN} - LAI_{k_i}^{SAT}|}{LAI_{k_i}^{BN}}, \delta = \frac{|LAI_{k_i}^{BN} - LAI_{k_i-1}^{BN}|}{LAI_{k_i}^{BN}}, \omega = \frac{|LAI_{k_i}^{3PG} - LAI_{k_i}^{3PG}|}{LAI_{k_i}^{3PG}}, \) and \( \eta = \frac{|LAI_{k_i}^{3PG} - LAI_{k_i-1}^{3PG}|}{LAI_{k_i}^{3PG}} \). Those weights (\( \gamma, \delta, \omega, \) and \( \eta \)) are proportional to the spatial and temporal change in the LAI values obtained from the satellite data and 3-PG output. The values of weighing factors may vary from 0 to a large number. It is unreasonable, however, to find a big difference (more than one) of LAI values within a short time period (e.g., 16 days). For more details about a GBN of improving forest growth estimates and its mathematical explanation, we refer to (Mustafa et al., 2012, 2011b).

The diagram in Figure 6.3 illustrates the procedure of implementing GBN, as called spatial GBN (Mustafa et al., 2012), with a spatial support of 250 m contains, for example, three species. The spatial GBN is implemented for each species individually, and the outputs of all implemented spatial GBN is summed after multiplied by its proportion to produce LAI value of a mixture pixel with support of 250 m.

### 6.4 Implementation

The above approach is applied to the Speulderbos forest in The Netherlands. The LAI ground measurements measured by DHP in space and time is for all tree species available from April 23, 2011 to June 26, 2011. Moreover, LAI ground measurements calculated from PAR data are for the Douglas fir available at single location from June 26, 2010 to August 29, 2011. In the current study, we used a spatial version of 3-PG model with a support of 250 m. This is the same resolution as that of the MODIS images.

Prior to implementing the spatial GBN, LAI MODIS is decomposed by applying the LMM to find the LAI values for each species individually for the entire study area. This approach is assumed to be valid throughout and is applied subsequently to all other LAI MODIS images.

The results are presented in terms of satellite image as well as by statistical data. The Root Mean Squared Error (RMSE) and the relative error (RE) were used to evaluate how well the estimated values obtained from all of the satellite data, spatial 3-PG model, and spatial GBN compare to the LAI_FD. The RMSE, and RE rate for spatial and temporal assessment are computed respectively as

\[
\text{sRMSE} = \sqrt{\frac{\sum_{p} (LAI_{Q}^{k} - LAI_{FD}^{k})^2}{p}}, \quad \text{sRE} = \frac{|LAI_{Q}^{k} - LAI_{FD}^{k}|}{LAI_{FD}^{k}} \times 100\%,
\]

\[
\text{tRMSE} = \sqrt{\frac{\sum_{s} (LAI_{Q}^{k} - LAI_{FD}^{k})^2}{s}}, \quad \text{tRE} = \frac{|LAI_{Q}^{k} - LAI_{FD}^{k}|}{LAI_{FD}^{k}} \times 100\%,
\]

where \( Q \in \{BN, SAT, 3PG\}, LAI_{FD}^{k} \) is the LAI field data, \( k \) is the pixel location, \( p \) is the number of pixels, and \( s \) is the number of samples. The
6.5 Results

6.5.1 Decomposed \( LAI_M \)

Figure 6.4 shows the LAI values of the five species resulting from decomposing the LAI MODIS pixel. The result is satisfactory with a residual of 0.14, 0.05, 0.13, 0.06, and 0.03 for Beech, Douglas fir, Hemlock, Japanese Larch, and Scotch Pine, respectively. The LAI resulting from this decomposing is used as input into the GBN as the \( \mu_{LAI_{SAT}} \) value. Moreover, the residual values are interpreted as the uncertainty of calculated LAI values. Hence, the residual is assigned to the standard deviation \( \sigma_{LAI_{SAT}} \).

6.5.2 GBN

Figure 6.5 (a)-(e) shows the temporal LAI values for each species estimated from the spatial GBN for a period of 14 months, from June 26, 2010 to August 13, 2011, along with the LAI derived from the spatial 3-PG model, and the LAI satellite data that were found from the decomposed LAI MODIS pixel. The spatial LAI ground measurements of all tree species were available for a period from April 23, 2011 to June 26, 2011 (Figure 6.5 (a), (c)-(e)), whereas, the LAI ground measurements of Douglas fir trees were available for a period of 14 months from June 26, 2010 to August 29, 2011 (Figure 6.5 (b)).

Table 6.3 shows the tRMSE and tRE values of estimated LAI from the four sources, \( LAI_{FD} \), \( LAI_{3PG} \), \( LAI_{SAT} \), and \( LAI_{BN} \), for all five species.

\( LAI_{FD} \) used to calculate sRMSE and sRE, is the average \( LAI_{FD} \) value of all species present within a pixel of 250 m resolution.

Furthermore, to see the site of the improvements with the current work, we compared the results of the current work with the results of (Mustafa et al., 2011b). This is carried out for the same study area and with the same data set (MODIS, 3-PG, and field data) of the Douglas fir.
6. GBN to improve spatio-temporal growth estimates of heterogeneous forests

Figure 6.5: LAI distribution of the Speulderbos forest of five species: (a) Beech, (b) Douglas fir, (c) Hemlock, (d) Japanese Larch, and (e) Scotch Pine. They are obtained from four sources, namely, the field data ($LAI_{FD}$), the spatial 3-PG model ($LAI_{3PG}$), the decomposed LAI satellite ($LAI_{SAT}$), and the spatial GBN ($LAI_{BN}$) with composed and decomposed LAI MODIS, during the period June 2010 to August 2011. The $LAI_{FD}$ are plotted whenever they were available.

The calculated tRMSE and tRE were applied for the averaged LAI over all nominal pixels and for all sources of each tree species across available $LAI_{FD}$ values. We found that the $LAI_{BN}$ is closer to the $LAI_{FD}$ than $LAI_{3PG}$ and $LAI_{SAT}$ having lower tRMSE and tRE values. Moreover, for the period of 14 months, we notice that the spatial GBN output with the decomposed LAI MODIS (bold black line in Figure 6.5) is closer to the LAI field data than the spatial GBN output with the composed LAI MODIS (bold black dished line in Figure 6.5). The RMSE and the RE of spatial GBN output with decomposed and composed LAI for all species are shown in Table 6.3. Overall, the improvement of the spatial GBN with decomposed LAI was between 47.62% and 66.98% than the spatial GBN output with composed LAI (Figure 6.5).

LAI spatial values were also assessed using sRMSE and sRE values. We found that the spatial GBN output at a spatial resolution of 250 m is more accurate than the output of the spatial 3-PG model and satellite data. The sRMSE and sRE values between $LAI_{BN}$ and $LAI_{FD}$ were lower than those between $LAI_{3PG}$ and $LAI_{FD}$ and those between $LAI_{SAT}$ and $LAI_{FD}$ (Table 6.4). Figure 6.7 shows the spatio-temporal distribution of...
6.5. Results

Table 6.3: The mean values ($\mu$), the standard deviation ($\sigma$), temporal root mean square errors (tRMSE) and temporal relative errors (tRE) for the various ways to estimate the LAI: field data ($LAI_{FD}$), the spatial 3-PG model ($LAI_{3PG}$), the satellite data ($LAI_{SAT}$), and the spatial GBN ($LAI_{BN}$) with composed and decomposed LAI MODIS of five tree species.

<table>
<thead>
<tr>
<th></th>
<th>Statistics</th>
<th>$LAI_{FD}$</th>
<th>$LAI_{SAT}$</th>
<th>$LAI_{3PG}$</th>
<th>$LAI_{BN}$ with composed LAI</th>
<th>$LAI_{BN}$ with decomposed LAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beech</td>
<td>$\mu$</td>
<td>5.38</td>
<td>4.01</td>
<td>7.03</td>
<td>5.97</td>
<td>5.65</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>0.41</td>
<td>1.01</td>
<td>0.51</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>tRMSE</td>
<td>0.86</td>
<td>1.65</td>
<td>0.63</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRE</td>
<td>15.00%</td>
<td>30.90%</td>
<td>11.05%</td>
<td>3.72%</td>
<td></td>
</tr>
<tr>
<td>Douglas fir</td>
<td>$\mu$</td>
<td>6.08</td>
<td>4.42</td>
<td>6.81</td>
<td>6.04</td>
<td>6.20</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>0.31</td>
<td>0.82</td>
<td>0.48</td>
<td>0.82</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>tRMSE</td>
<td>1.64</td>
<td>0.67</td>
<td>1.06</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRE</td>
<td>26.16%</td>
<td>9.28%</td>
<td>6.32%</td>
<td>4.67%</td>
<td></td>
</tr>
<tr>
<td>Hemlock</td>
<td>$\mu$</td>
<td>5.74</td>
<td>3.32</td>
<td>7.22</td>
<td>6</td>
<td>6.03</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>0.42</td>
<td>0.7</td>
<td>0.91</td>
<td>0.49</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>tRMSE</td>
<td>2.51</td>
<td>0.95</td>
<td>0.46</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRE</td>
<td>43.32%</td>
<td>16.48%</td>
<td>5.86%</td>
<td>3.22%</td>
<td></td>
</tr>
<tr>
<td>Japanese Larch</td>
<td>$\mu$</td>
<td>6.02</td>
<td>4.12</td>
<td>6.47</td>
<td>6.05</td>
<td>6.09</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>0.52</td>
<td>0.99</td>
<td>0.68</td>
<td>0.47</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>tRMSE</td>
<td>1.58</td>
<td>0.73</td>
<td>0.42</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRE</td>
<td>24.90%</td>
<td>11.20%</td>
<td>6.79%</td>
<td>3.17%</td>
<td></td>
</tr>
<tr>
<td>Scotch Pine</td>
<td>$\mu$</td>
<td>6.29</td>
<td>3.83</td>
<td>6.84</td>
<td>6.1</td>
<td>6.21</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>0.45</td>
<td>0.76</td>
<td>0.94</td>
<td>0.42</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>tRMSE</td>
<td>2.44</td>
<td>1.07</td>
<td>0.5</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tRE</td>
<td>38.12%</td>
<td>15.42%</td>
<td>5.92%</td>
<td>2.81%</td>
<td></td>
</tr>
</tbody>
</table>

$^1$ The $\mu$ and $\sigma$ of $LAI_{FD}$ of all species (except Douglas fir) were calculated for the period from April 23, 2011 to June 26, 2011, whereas for Douglas fir were calculated for the period from June 26, 2010 to August 29, 2011.

the estimated LAI from the four sources, $LAI_{FD}$, $LAI_{3PG}$, $LAI_{SAT}$, and $LAI_{BN}$, at the spatial resolution 250 m of mixed species, after considering the proportion of each species in the composition of each pixel. It shows that the 3-PG is overestimating and that the LAI satellite data are underestimating the LAI values across the studied period. Moreover, it shows that the $LAI_{FD}$ values occurred within the error bands (the standard deviation) of the $LAI_{BN}$ and emphasize that $LAI_{BN}$ values, even with their errors, are close to the field data. Overall, for this study, the spatial GBN output matches the field data better than the spatial 3-PG output and the decomposed LAI MODIS pixel values.

Figure 6.6 shows the relative errors of implementing this approach for Douglas fir, where the field data were available. It indicates that the relative errors are rather large (> 0.10) and unstable during the first six months. Overtime, however, the relative errors decreased to approximately 0.02 and became more stable, especially after January 2011.
6. GBN to improve spatio-temporal growth estimates of heterogeneous forests

Table 6.4: The mean values ($\mu$), the standard deviation ($\sigma$), spatial root mean square error (sRMSE) and spatial relative error (sRE) for the various ways to estimate the LAI: field data ($LAI_{FD}$), the spatial 3-PG model ($LAI_{3PG}$), the satellite data ($LAI_{SAT}$), and the spatial GBN ($LAI_{BN}$); at 250 m spatial resolution of mixed species.

<table>
<thead>
<tr>
<th>Date</th>
<th>LAI</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>sRMSE</th>
<th>sRE rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>23-April</td>
<td>$LAI_{FD}$</td>
<td>5.04</td>
<td>0.39</td>
<td>0.98</td>
<td>17.97%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{SAT}$</td>
<td>4.11</td>
<td>0.10</td>
<td>1.21</td>
<td>26.81%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{3PG}$</td>
<td>6.35</td>
<td>0.30</td>
<td>1.40</td>
<td>26.81%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{BN}$</td>
<td>4.97</td>
<td>0.19</td>
<td>0.37</td>
<td>6.43%</td>
</tr>
<tr>
<td>9-May</td>
<td>$LAI_{FD}$</td>
<td>5.33</td>
<td>0.48</td>
<td>1.52</td>
<td>25.77%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{SAT}$</td>
<td>4.07</td>
<td>0.19</td>
<td>1.21</td>
<td>20.25%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{3PG}$</td>
<td>6.58</td>
<td>0.49</td>
<td>1.40</td>
<td>26.81%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{BN}$</td>
<td>5.36</td>
<td>0.32</td>
<td>0.57</td>
<td>7.73%</td>
</tr>
<tr>
<td>25-May</td>
<td>$LAI_{FD}$</td>
<td>5.51</td>
<td>0.35</td>
<td>1.38</td>
<td>24.19%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{SAT}$</td>
<td>4.16</td>
<td>0.21</td>
<td>1.26</td>
<td>21.98%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{3PG}$</td>
<td>6.69</td>
<td>0.26</td>
<td>1.26</td>
<td>21.98%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{BN}$</td>
<td>5.34</td>
<td>0.21</td>
<td>0.42</td>
<td>6.01%</td>
</tr>
<tr>
<td>10-June</td>
<td>$LAI_{FD}$</td>
<td>5.69</td>
<td>0.45</td>
<td>1.68</td>
<td>27.93%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{SAT}$</td>
<td>4.07</td>
<td>0.25</td>
<td>1.24</td>
<td>20.35%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{3PG}$</td>
<td>6.79</td>
<td>0.30</td>
<td>1.26</td>
<td>21.98%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{BN}$</td>
<td>5.35</td>
<td>0.53</td>
<td>0.76</td>
<td>10.88%</td>
</tr>
<tr>
<td>26-June</td>
<td>$LAI_{FD}$</td>
<td>6.24</td>
<td>0.40</td>
<td>1.73</td>
<td>26.85%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{SAT}$</td>
<td>4.55</td>
<td>0.15</td>
<td>0.81</td>
<td>11.38%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{3PG}$</td>
<td>6.91</td>
<td>0.57</td>
<td>0.81</td>
<td>11.38%</td>
</tr>
<tr>
<td></td>
<td>$LAI_{BN}$</td>
<td>5.98</td>
<td>0.26</td>
<td>0.59</td>
<td>8.01%</td>
</tr>
</tbody>
</table>

Figure 6.6: Resulting relative errors of the spatial GBN output with respect to the field data of Douglas fir trees for the period of 14 months from June 26, 2010 to August 29, 2011.

Figure 6.8 shows the $LAI_{BN}$ output of both, the spatial GBN and the GBN of the spatially homogeneous approach (Mustafa et al., 2011b) for Douglas fir. We found that the output of the spatial GBN (bold black line in Figure 6.8) is closer to the field data than the output of the GBN (bold black dished line in Figure 6.8). Moreover, the spatial GBN output represents the seasonal change behavior of LAI better than the GBN output. By using spatial information in the spatial GBN of the current
study the tRMSE reduced to 0.35 and the tRE to 4.67%, whereas the tRMSE and tRE of the GBN output were equal to 0.95 and 11.18%, respectively.

Figure 6.7: The spatio-temporal distribution of the estimated LAI from four sources, field data ($LAI_{FD}$), the spatial 3-PG model ($LAI_{3PG}$), the satellite data ($LAI_{SAT}$), and the spatial GBN ($LAI_{BN}$). (a) spatial distribution of the LAI at April 23, 2011, (b) the temporal distribution of the averaged LAI, (c) spatial distribution of the LAI at June 26, 2011.

Figure 6.8: LAI distribution of Douglas fir for the study area during the period June 2010 to August 2011. They are obtained from four sources, namely, field data ($LAI_{FD}$), the satellite data ($LAI_{SAT}$), the GBN ($LAI_{BN}$), and the spatial GBN (spatial $LAI_{BN}$).
6.6 Discussion

Improve estimation of LAI values that show variation in space and time is achieved for a heterogeneous Speulderbos forest in the Netherlands, by combining the LAI output of the spatial 3-PG with that derived from the satellite images using the spatial GBN.

Our results showed that the deviation between the spatial GBN output and LAI field data is lower than the deviation between the spatial 3-PG model output and the LAI field data. This is indicating that the spatial GBN output was more accurate than that of the spatial 3-PG model output alone and was closer to the LAI field data. The inaccurate LAI output of the spatial 3-PG model was due to the several factors related to the uncertainties of input data as the soil data, and specific parameter values of each tree species (Tickle et al., 2001; Xenakis et al., 2008). Therefore, the 3-PG model parameters and soil input data require a close attention to reduce their uncertainties. The spatial GBN output also has uncertainties as the resulting errors. These errors, however, decreased over time after six months of the spatial GBN implementation. Moreover the spatial GBN output is more accurate than the GBN output of the spatially homogeneous approach.

The strength of this approach is that the spatial LAI values can be estimated, with satisfactory level of accuracy, for a heterogeneous forest. This is, however, not the case in the 3-PG model and even in a mixed MODIS pixel. Given that the 3-PG model is designed for use in a homogeneous forest, however, the spatial GBN is used to estimate LAI values for a heterogeneous forest by combining the spatial information from two sources, i.e. the spatial 3-PG model and satellite data. For that purpose, remote sensing techniques were used to classify and determine the forest cover species, prior to the spatial GBN implementation. This technique required a fine resolution image and the use of Linear Mixture Model approach.

A major contribution of this work was to apply the method of (Mustafa et al., 2012) to infer the spatio-temporal estimation of LAI values for a heterogeneous forest. It showed the applicability of the method with different tree species. Moreover, the implementation of the spatial GBN was achieved for each tree species individually. The final output of the spatial GBN was calculated for all presence species in composition of mixed pixels with spatial support of 250 m. This was calculated by considering the area proportion of each species in the composition of the mixed pixels. In this way a reliable and accurate LAI values can be estimated for mixture pixels.

Identifying LAI estimates of different species in the composition of a mixed MODIS pixel was based on using LMMs. The LMMs are commonly used in remote sensing to decompose the spectral signal in pixels. They are used to assess the accuracy of the land cover type by identifying the area proportion (fraction) of each species in the image pixel. In this study, we used LMMs to decompose the biophysical variable LAI where the information of the area proportion of each species was available. This
information was gained from counting the number of reference pixels; which was constructed using the support from aerial and ASTER images, of each species within MODIS pixel. By using area proportion information, we identified the value of interest (LAI) of all presence species in the pixels. As outcome from using LMMs in this study, we obtained five decomposed LAI values (one for each species) from each LAI MODIS image. These decomposed LAI values were used to implement the spatial GBN for each species individually. Moreover, two assumptions were proposed to implement the LMM in this study. The first was that no major change in the forest composition in terms of area proportion of each species occurred during the study period. The second was that the LAI of a specific species had the same value during the entire study area. The latter assumption was made in order to solve the Equation (6.1). These assumptions, however, could be considered as sources of uncertainty that may influence the accuracy of the spatial GBN output. The spatial GBN output with decomposed LAI, however, was better than the spatial GBN output with composed LAI. Furthermore, there may uncertainties occur due to misregistration between referenced grid and MODIS image, however, the treatment of this issue is beyond of the scope of this study.

Several other issues require further work. For instance, the time period of this study needs to be extended to include longer time as well as larger spatial scale. This may show a better explanation of the model in terms of LAI seasonal changes. Also, ground measurements should be collected for the study area at the same spatio-temporal support of the satellite images to be used as a validation of the spatio-temporal GBN model.

Spatial data of the forest are becoming increasingly important for forest vegetation management and decision making. The reliable and accurate estimations of the biophysical variables as LAI are useful and have served as inputs to functional models of ecosystem biogeochemistry. It provides a better understanding of the forest, description of the spatial pattern of forest structure, and which ultimately can serve as an informative factor in climate change.

6.7 Conclusion

With this work, the spatial GBN has been shown to improve the spatio-temporal estimation of LAI values by combining the spatial output of the 3-PG model and satellite data in a forest with various tree species. Moreover, LAI MODIS is decomposed using LMM producing LAI value for each tree species individually. In this way the spatial GBN was implemented for individual strata that consisted of trees of a single species. The spatio-temporal accuracy of the spatial GBN is assessed from the end of April to the end of June 2011, where the spatial LAI field data were available. The temporal accuracy of the spatial GBN output for the Douglas fir species assessed for the period of fourteen months, from end of June, 2010 to mid of August, 2011. The study concludes
that the spatio-temporal LAI output of the spatial GBN is more accurate than the spatio-temporal LAI output of the spatial 3-PG model. The deviation of the LAI output of the spatial GBN is lower than that of the LAI output of the spatial 3-PG model from the LAI field data. Lastly, given the presented method and results in this study, our approach can be applied successfully to estimate LAI values for any heterogeneous forest.
7.1 General discussion
7.2 Research findings
  7.2.1 Objective 1
  7.2.2 Objective 2
  7.2.3 Objective 3
  7.2.4 Objective 4
  7.2.5 Objective 5
7.3 Reflections
  7.3.1 Integration
  7.3.2 Reliability
7.4 Limitations
7.5 General suggestions
7. Synthesis

7.1 General discussion

Forests form one of the world’s most important renewable resources. They play an important role in preserving biodiversity, protecting critical watersheds and providing livelihoods. As governments commit themselves to sustainable forest management and climate change initiatives, such as the United Nations Framework Convention on Climate Change (UNFCCC) and the Kyoto Protocol, there is an increasing demand for precise estimates of forest growth at local and global scales (UNFCCC, 1998). At the 5th of December 2010, the “Forest Day” at the United Nations climate change conference in Cancun, Mexico, two of the Nature Conservancy’s leading forest experts, Jeff Fiedler and Frank Lowenstein, sat down to brainstorm their list of “top 10 reasons why forests matter” (The-Nature-Conservancy, 2010). The list briefly is: absorbing and storing carbon, home to people, source of jobs and livelihoods, wood for furniture, lumber, firewood and other products, habitat for mammals, birds, insects, preventing flooding, conserving soil and water, natural beauty, places for recreation and tourism, and regulating regional climate. The latter is important as well in the Middle East. For example in Iraq, sandstorms have swept across its cities recently and this is due to several reasons, among which is deforestation caused by lack of forest information and lack of good management. Trees and forests act as a barrier to filter and reduce the amount of sand carried by the wind. Therefore, monitoring and observing forest and provide information on its growing are important both for researchers in the climate change domain and for forest managers.

Existing models for estimating forest growth, such as process-based ones, provide information of the forest status and assess the growth of the forest. Such models, however, contain uncertainties and suffer from model inadequacy resulting in model output which does not reflect the forest system in all details. During the last three decades, the use of process-based models by forest managers has been met with some resistance (Welch et al., 1981; Sands, 1988; Battaglia and Sands, 1998; Makela et al., 2000; Sands et al., 2000). The main cause is the need to calibrate and validate these models using real data to give them credibility, reliability and determine levels of uncertainty (Welch et al., 1981; Sands et al., 2000; Mummery and Battaglia, 2004). Remote sensing imagery provides information of several forest parameters. For example, biophysical parameters can be provided by a spaceborne product or can be derived using empirical relationships between ground measurements and spectral vegetation indices from satellite images. Although this approach is ideal for a quick assessment of forest growth, it still contains uncertainties, whereas the data are of a limited precision and are often not available.

A process-based forest growth model, such as the 3-PG model, is a deterministic model. It can be defined as a description of the behavior of a forest system in terms of a set of functional relationships (e.g. plant
growth processes), their interactions with each other and the system environment. It is a generic stand model, in the sense that its structure is not site or species-specific. The inaccurate output of such a model is due to several factors related to the uncertainties of input data, in particular soil data, and specific parameter values of each tree species (Esprey et al., 2004; Fontes et al., 2006). Furthermore, any change in the forest structure due to the unexpected disasters, as the cutting of trees, forest fire, or flooding cannot be estimated directly by the process-based forest growth model. Existence of such a disaster leads to uncertain output of the model. Therefore, the uncertainty is aggravated as an error over time and, thus, the error in the prediction may become large. Consequently, the process-based forest growth model needs to adjust and correct its output during successive iterations. One of the solutions that can help to update the model parameters and its output is the use of satellite images. Satellite images provide quick information, and the images can be used to observe changes in the forest. Satellite data, i.e. normalized difference vegetation index (NDVI), may be used directly to adapt the model by using it in mathematical model calculations. Coops et al. (1998) utilized satellite data as input to the 3-PG model and called it the 3-PGS model. Although, satellite data in that way were used to feed and update the process-based forest growth model, the uncertainty remains. Uncertainties in satellite data are due to atmospheric characteristics such as the presence of clouds or aerosols, the instrument sensor, mixed-pixel effect, instantaneous field of view (IFOV), and image processing (Cracknell, 1998). To reduce the uncertainties, statistical methods were applied in this study based on graphical models, in particular Bayesian networks. Such methods can reduce uncertainties, thus leading to more accurate and precise forest growth assessments.

Bayesian networks are a combination of graph theory and probability theory. They are increasingly valuable and popular for modeling uncertain and complex domains such as ecosystems, environmental management, computer science, and forensic sciences (Taroni et al., 2006; Pourret et al., 2008). At best, they provide a robust and mathematically coherent framework for the analysis of this kind of problems. A useful aspect of Bayesian networks is that technically, there is no such thing as “too few data”. There are no minimum sample sizes required to perform an analysis. Bayesian networks take into account all available data (Myllömäki et al., 2002). Moreover, the structure of Bayesian networks can be learned by defining the model structure based on subject matter knowledge and using the data to define conditional probability distributions. This is an area of active research, and although the statistical theory is well understood, the methods are still under development, since their computational requirements are hard (Myllömäki et al., 2002; Jensen and Nielsen, 2007). In addition, Bayesian network models have the advantage that they can easily and in a mathematically coherent manner incorporate knowledge of different accuracies and from different sources. Data measured at different levels of accuracy (e.g. absence/presence and quantity data) can be combined as well.
7. Synthesis

This study aimed to integrate and incorporate two different sources with different level of accuracy, i.e. the 3-PG model output and satellite data, within Gaussian Bayesian networks. Gaussian Bayesian networks are a common type of Bayesian network dealing with continuous variables.

In the current chapter I firstly synthesize the most important results of previous chapters, in order to gain a better understanding of the study. Secondly, it reflects the study. Lastly, it outlines the scope for future research.

7.2 Research findings

The research findings of specific research objectives were stated in Section 1.5, and are summarized as below:

7.2.1 Objective 1

Objective one was to develop a methodology by which a Gaussian Bayesian network can be used to adjust the output of the 3-PG forest growth model using a series of satellite data.

The capability of Gaussian Bayesian network was explored in integrating two different sources with different level of accuracy and improving the forest growth estimates. The framework was illustrated by implementing the Gaussian Bayesian network for a period of 26 months from July 2007 to September 2009. The findings were:

- A Gaussian Bayesian network was able to integrate the forest growth estimates that come from different sources into a single reasoning framework.
- For the period of 26 months, the combination of satellite data and the 3-PG model in a Gaussian Bayesian network reduced the RMSE from 2.71 to 1.5 and the RE from 37.4% to 14.5%.
- By using satellite data, the output of the 3-PG model closely matched the mean forest growth estimate.
- The Gaussian Bayesian network was more sensitive to the variation of the forest growth estimates derived from satellite image than to the variation of the estimated output of 3-PG model.

7.2.2 Objective 2

Objective two was to evaluate the performance of Gaussian Bayesian network modeling for forest growth estimates.

The work of objective 1 was extended towards evaluating the performance of a Gaussian Bayesian network for forest growth estimates modeling, by addressing both evidence propagation, and sensitivity analysis of the input sources, i.e., the 3-PG model output and satellite data. The specific findings related to objective two were:
7.2. Research findings

- Providing evidence to the 3-PG model output from field data improved the Gaussian Bayesian network output.
- The robustness of the Gaussian Bayesian network took place after the evidence of the 3-PG output propagated into the Gaussian Bayesian network, where the evidence had an immediate effect on the output.
- Propagating evidence from field data into the forest growth estimates that were derived from satellites did not serve equally well to improve the Gaussian Bayesian network output.

7.2.3 Objective 3

Objective three was to investigate how the Gaussian Bayesian network performs in updating the 3-PG model when there are gaps in the series of satellite images.

The EM-algorithm was formulated to be used within a Gaussian Bayesian network to estimate missing satellite data that used as an input into the Gaussian Bayesian network. Thereafter performance of the Gaussian Bayesian network was checked. The approach of this objective was implemented by introducing synthetic gaps in the satellite data successively and not successively during winter seasons and spring. The specific findings related to the objective three were:

- The missing satellite data were estimated successfully using the EM-algorithm. The maximum value of the averaged absolute error between the original data and those estimated was 0.16.
- The Gaussian Bayesian network performance was better in the case of estimating not successively missing satellite data than in the case of estimating successively missing satellite data. Whereas the poorest performance of the Gaussian Bayesian network was observed in the case of estimating successively missing satellite data.
- A noticeable deviation of Gaussian Bayesian network output was seen after applying the EM-algorithm during the spring. This was due to the fact that the spring season was the most critical time for forest growth, when the leaves started growing, became more abundant.
- The presence of more than five successively missing satellite data influenced the output of the network such that Gaussian Bayesian network output did not match the field data. However, the Gaussian Bayesian network output remained closer to the field data than the 3-PG model output.
- The Gaussian Bayesian network output, particularly, improved after the EM-algorithm performed in the case of estimating not successively missing satellite data during the full study period.
7. Synthesis

7.2.4 Objective 4

Objective four was to implement the Gaussian Bayesian network with different satellite images and to modify Gaussian Bayesian network to infer the spatial estimation of forest growth estimates.

Part of this objective was to apply the approach of objectives 1 & 3 with satellite images of a finer resolution and with real missing data. The second part of this objective was to modify the Gaussian Bayesian network to perform spatial estimation of the forest growth estimates using a spatial version of the 3-PG model. This study was implemented for a period from June 2010 to October 2010. The findings were:

- Gaussian Bayesian network modelling for forest growth estimates could be applied with different satellite images.
- The missing satellite data were estimated with significant values such that a small deviation was noticed with respect to the non-missing data during the study time.
- The spatial estimation of forest growth estimates could be done with the modified version of the Gaussian Bayesian network.
- The deviations between the output of the spatial Gaussian Bayesian network and field data were smaller than to the deviations between the spatial 3-PG model output and field data.

7.2.5 Objective 5

Objective five was to improve spatio-temporal estimation of growth estimates of heterogeneous forest using modified Gaussian Bayesian network.

This study covered the improvement of the forest growth estimates values that show variation in space and time of a heterogeneous forest containing various tree species. For such purposes, the satellite pixels were decomposed using a finer resolution satellite image. This study was implemented for a period of 14 months from April 2010 to June 2011. The highlights of this study were:

- The spatio-temporal output of the spatial Gaussian Bayesian network was more accurate than the spatio-temporal output of the spatial 3-PG model.
- The output of the spatial Gaussian Bayesian network of all tree species with the decomposed satellite pixels was more accurate (between 47.82% and 66.98%) than spatial Gaussian Bayesian network output with composed satellite pixels.
- The more reliable and stable output of the Gaussian Bayesian network with lower RE was observed after six months of start implementing Gaussian Bayesian network.
- Given the spatial information, the output of the spatial Gaussian Bayesian network of the current objective was closer to field data than to the output of the Gaussian Bayesian network of the spatially homogeneous approach (objective 1).
7.3 Reflections

The study focused on improving the accuracy of the forest growth estimates and found an application in the Speulderbos forest in The Netherlands. Two different data sources, i.e., a process-based forest growth model and satellite data, were integrated using statistical tools. As reflections from this study, two perspectives can be determined: integration and reliability.

7.3.1 Integration

Accurate information on forest growth estimates is important at the global, national and local scales. This is particularly the case nowadays, because of the role of forests in the climate change, the use of forests for recreation and other services, their functioning for biological diversity and their contribution to national economies. Such situations require a wide range of complete and accurate information on forest growth. This information, however, is often available at different levels of accuracy or may not be available for a certain space and time. Moreover, to make sound decisions, a solid estimate of forest parameters should be obtained. An important way to obtain reliable estimates is by integrating different approaches which are associated with the same forest parameter. The integration strategy is the key to combining data and provides a valuable information to assist decision makers to have actual evidence and help them come up with a solution. This is noticeable in the recent years. For example, agencies and organizations of environmental protection try to facilitate the use of integrated approaches to inform and improve local, regional and national policy decisions relevant to climate change adaptation and mitigation strategies (Jones, 2011).

Integration refers to incorporating, combining, assimilating, or mixing sources of information. There are several definitions for the integration. It is defined in medical science as anabolism, coordination, or assimilation. In psychology, it refers to the organization of the psychological or social traits and tendencies of a personality into a harmonious whole. In mathematics it means the process of computing an integral; the inverse of differentiation, and thus increasing the dimensionality.

In this study, integration referred to the combination of derived data from different approaches taking into consideration the consistency level of these approaches. By means of this definition, the study showed that the output of forest growth models integrating satellite data was able to take into consideration the level of accuracy. This turned into more reliable results. Therefore, integration of the two approaches enhances the values of both data sets, resulting in a significant improvement in consistency values. This was done in a statistical way using Bayesian networks. This study showed that Bayesian networks are an efficient tool for such integration as they improved estimated values. Bayesian networks addressed uncertainties in both the process-based forest growth
model and the satellite data. This opens possibilities to address other aspects in particular reliability of the estimated values.

7.3.2 Reliability

Process-based forest growth models may only approximate actual processes and changes in forests. Also, satellite data may be imprecise to provide reliable estimates of forest growth and they may often not available. Both result in a lack of output reliability. In this context the definition of reliability refers to the degree of consistency between two equivalent forms of a measure or estimate. Reliability differs from mean validity, both being not necessarily related. Validity concerns the extent to which a measure actually measures what it is supposed to do. Reliability puts a maximum limit on the correlation that can be observed between two measures. It is, therefore, possible for scores to be perfectly reliable on a measure that has no validity or for scores to have no reliability on a measure that is perfectly valid.

There are three aspects of reliability: equivalence, stability and internal consistency. Equivalence refers to the amount of agreement between two or more approaches that are administered at nearly the same moment and location. The output of the integration approach with the output of the sources that used in the integration were compared with the field data at the same moment and location. A better agreement between the output of the integration approach and field data was shown. Therefore, this study showed an equivalence in results relative to the result of the sources that used in the integration approach.

Stability occurs when the same or similar scores are obtained with repeated testing for the same group. In other words, the scores are consistent from one time to the next. In this study, it became apparent through chapters 2 to 6. This approach was implemented with the same group, satellite data, process-based forest growth model, and field data and the results of this approach gave a more accurate estimation (Chapters 2, 5, and 6).

Internal consistency refers to the degree of homogeneity of components in an approach—the extent to which responses to the various components (sources) of the integration approach correlate with one another or with a score on the approach as a whole. The sources for the integration represent the seasonal change behavior of forest growth. However, they either underestimate or overestimate growth estimates, unless integration is done in a statistical way as in this study.

Overall, integrating the two sources process-based forest growth model and the satellite data provided more reliable growth estimates than implementing each source independently.
7.4 Limitations and further scope of research

Two main drawbacks of this study could be notified and both were related to the data availability and study area. First, the study area is of $1 \times 1$ km. It is a rather small area for a domain of forest research. One of the important reasons to restrict the study to this small area is the availability of field data, whereas no such defined modeling research was done in this area so far. Furthermore, the study area is a heterogeneous forest, consisting of five tree species. Such a limitation was adequate to use the Bayesian network in the forest growth field. The focus of this study concerned data integration and modeling of forest growth. The study can be extended by applying this approach to a large forest area if the detailed field and satellite data are available.

Second, a reliable source of collecting data plays an important role in the precise estimation of forest growth estimates. Collection of reliable climate and soil data, and specific parameter values are required for an accurate output of the 3-PG model. In this study, most field data were acquired from the Speulderbos climate station. The spatial data for the spatial 3-PG model, however, were acquired from three climate stations. These climate data were gridded, interpolated, converted into rasters and clipped to the study area at a 250 m resolution using ESRI ArcGIS software version 10. This data may be prone to uncertainties. Uncertainties may arise from the instruments or the location or the climatic characteristics or the methods used to process the data. Also, satellite data are not free of uncertainties due to several factors like atmospheric variation, functioning of the sensor, or image processing techniques.

This study focused on integrating satellite data and process-based forest growth model data for estimation of the forest growth estimates using a Gaussian Bayesian networks. Other methods could also have been used as well, such as the Kalman filter (Czaplewski et al., 1988; Steve, 2001). It allows to estimate and predict forest growth estimates, however, it requires available historical time series data (Xiao et al., 2011). In addition, this study considered the use of the Gaussian distribution. This distribution has the well-known advantage of being practically convenient as well as being the limiting distribution of other distributions. Other, non-Gaussian, distributions could be further explored within Bayesian networks (Holmes, 2008; Kjaerulff and Madsen, 2008). This could be identified as the possibility of extending this research.

Specific limitations of each objective and their recommendation for the further research are summarized below.

Related to objective 1, the spatial resolution of the satellite images (250 m) is limited in representing the spatial variability and distribution of trees within the forest. This introduces additional uncertainty into forest growth estimates, which could be reduced by using finer resolution satellite images. Furthermore, the uncertainties that result from the neighborhood of the satellite pixels have an impact on the output of
7. Synthesis

A Gaussian Bayesian network. This may be addressed with a subpixel remote-sensing approach. Lastly, multisource image fusion technique could improve Gaussian Bayesian network output so that the forest growth estimates is estimated more frequently. At some part of the study time, particularly at the beginning of implementation, the Gaussian Bayesian network output was not well representing the seasonal changes of the forest growth estimates. The inappropriate Gaussian Bayesian network output is due to the fact that the Gaussian Bayesian network output is affected by the high values of the 3-PG output. The inaccurate output of the 3-PG model was caused by several factors related to the uncertainties of input data as the soil data, and specific parameter values of each tree species (Tickle et al., 2001; Xenakis et al., 2008). Therefore, the 3-PG model parameters and soil input data require a close attention to reduce their uncertainties. Likewise, the Gaussian Bayesian network needs at least six months until its output approximates the field data at a sufficient precision and represents the changes in the seasonal growth. Also, homogeneity of the forest study area has been assumed. This rarely occurs within the forests. Nonhomogeneity of the forests is difficult to address when extracting the biophysical parameters from satellite images, particularly for relatively small areas. This may prohibit the extension of this method toward a more general applicability by considering unmixing in remote sensing.

Related to objective 2, performance of the Gaussian Bayesian network was evaluated using evidence propagation and sensitivity analysis of Gaussian Bayesian networks. These methods have restrictions, however, as they were required to maintain positive definiteness of the covariance matrix. Other methods to carry out a sensitivity analysis may be considered, with less severe restrictive conditions, to study the sensitivity in discrete Bayesian networks (Laskey, 1995). For this study, based on continuous data, this would require either an expansion of the methodology, or a discretization of the input variables (Neil et al., 2007). This could be considered for further work in this field.

Related to objective 3, the EM-algorithm was formulated within a Gaussian Bayesian network, to handle the problem of missing satellite data and to estimate their value. A well-known drawback of the EM-algorithm is that the convergence can be slow (McLachlan and Krishnan, 2008). This may in turn slowdown the process of Gaussian Bayesian networks. The speed of convergence of the EM-algorithm, however, can be accelerated by including other methods into the EM-algorithm (Kuroda and Sakakihara, 2006). Such improvements are, e.g., based on the application to the EM-algorithm of optimization theory techniques such as Aitkin’s acceleration (Meilijson, 1989).

Related to objectives 4 and 5, the spatio-temporal estimation of forest growth estimates was achieved taking the heterogeneity of forest into account. Decomposition of the satellite pixels of 250 m resolution was achieved with Linear Mixture Models (LMM). An assumption was used in this study to implement the LMM. No major change in the forest composition has been assumed during the study period in terms of area.
proportion of each tree species. This assumption, however, could be considered as sources of uncertainty which may have an influence the accuracy of the spatial Gaussian Bayesian network output. For further exploration, frequent observation needs to be done for the study area to quantify a precise area of each tree species in the satellite pixels. Moreover, satellite images of a finer resolution could be used with less mixed pixels. Lastly, ground measurements should be collected at the same spatio-temporal support of the satellite images to be used as a validation of the spatio-temporal Gaussian Bayesian network.

7.5 General suggestions

Researchers in forestry may wish on the basis of the findings in this thesis to design a system for monitoring and managing forest. Such a system has as a target to serve application requirements to give a maximum benefit at a minimum effort. Such approaches generally restrictive due to several factors. It depends on the particular needs to be gained from such a system and the required degree of accuracy. The goal and the outcome of the system should be outlined. Based on this thesis, estimating forest growth requires inclusion of both a forest model, and satellite images. In the rest of this section, I provide an explanation on the ideal system that should be used to estimate forest growth, if requirements are available and can be set at the onset.

The best forest model to choose is the one that is most useful for an application, and the choice should be based on the application and on resources. It is the end use that finally determines the best approach for modeling forest stands. Applications requiring growth models include site evaluation, testing hypotheses of growth, estimating expected yields, and establishing optimal management regimes. Each of these applications may require a different approach to growth modeling. No single model can serve all these needs, and all put different requirements to a model, but some general guidelines can be given. A spatial version of the 3-PG model is considered to be an appropriate model if it is well parameterized for individual species. This model is chosen based on its advantages. For instance, it has a relatively low demand for data; it has several types of outputs, and it produces spatially and temporally explicit outputs at the scale of the input. The spatial 3-PG model requires spatial input of climate and soil data that need to be collected at least at three sites (stations). To have a reliable and accurate output, however, the model needs a frequent calibration at some moment of its implementation (Paul et al., 2007; Fontes et al., 2006). Therefore, if the required data for implementation and calibration are available, the spatial 3-PG model could be a good choice for estimating forest growth.

Information on forest growth could be gained from satellite images if the forest model cannot be used due to unavailability of the required data. This may not lead to a total ignoring of the use of forest models, because forest model can estimate tree parameters that cannot be
estimated easily by satellite images, such as the tree diameter. There are several advantages of using satellite images in forestry. The use of images requires specialized expertise in order to select those images from satellite sensors that use the appropriate spectral band for a forest application. For instance, a suitable sensor for classifying urban images will not necessarily be appropriate for forest mapping, because buildings reflect electromagnetic waves differently from trees and hence show a different spectrum with the same sensor. Moreover, the cost of images may also be a severe constraint in the selection. A high spatial resolution image usually gives better results, but it requires more money and perhaps more time to analyze it. On the other hand, a too low resolution image may not reveal enough details to detect changes. Finally, selection of images is based on the application requirement. For example, a suitable sensor for forest mapping will not necessarily be suitable for estimating forest parameters. Also, it depends on the application scale in terms of spatial or temporal resolution. In general, for forest growth estimates, a moderate image spatial resolution every 16 or 30 days would be a sufficient to use, such as those obtained by the ASTER satellite (Heiskanen, 2006).

Finally, statistical approaches that include integrating sources, as Bayesian networks, should be included. Bayesian networks may compensate for the limitations of the two above approaches. For example, if required data for calibrating forest model are not available, or the satellite data are not available. A Bayesian network may help to make probabilistic statements using prior knowledge and the various sources of data at the present moment as well as from historical records.


Coops, N. C., Waring, R. H., Landsberg, J. J., 1998. Assessing forest productivity in Australia and New Zealand using a physiologically-based model driven with averaged monthly weather data and satellite-


Bibliography


Bibliography

(MOD15). Algorithm, theoretical basis document. Greenbelt, MD 20771, USA.


Bibliography


Bibliography


Bibliography


131


Biography

Yaseen Taha Mustafa was born on 18 January 1979 in Mosul-Iraq. In 1996, he joined Department of Mathematics, College of Computer Science and Mathematics, The University of Mosul, Iraq, where he received his bachelor degree in Mathematics. From 2000 to 2003 he was appointed as an assistant researcher in the Department of Mathematics, College of Science, The University of Dohuk, Dohuk, Kurdistan region, Iraq. In 2005, he earned his MSc degree in Mathematics from the University of Duhok. From 2005 to 2008, he was appointed as a lecturer in the Department of Mathematics, College of Education, at the same University. In 2008, he began to pursue the present PhD research from the Department of Earth Observation Science, Faculty of Geo-Information Science and Earth Observation of the University of Twente (ITC), Enschede, The Netherlands. His research interests include the area of remote sensing, including mathematical and statistical tools, such as Bayesian networks. During his PhD, he published three ISI peer reviewed paper, and presented his research in three international conferences. He received the Best Student Paper Award at the ASPRS conference 2011 in Milwaukee, Wisconsin, USA for his paper. This thesis is the output of his PhD research.
Author’s Publications

Published ISI Journal Articles


Articles Under Review


Other Refereed Journal Article

Conference full paper

ITC dissertations

A complete list of ITC dissertations is online on the ITC website: www.itc.nl/research/phd/phd_graduates.aspx.

This dissertation has number 202.