CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Introduction

Quantitative reconstruction of vegetation has been one of the primary goals in Quaternary palaeoecology and palynology (Sugita S., 2006). Generations of palynologists have observed that fossil pollen from differently sized sites represents vegetation at different spatial scales (Berglund, 1973; Jacobson and Bradshaw, 1981). Due to the growing need of past vegetation or land-cover reconstruction, there have been many different approaches and mechanisms by which this could be effectively handled. Among these publications are: Trajectories of land use change in Europe: a model-based exploration of rural features (Verburg, P.H, Berkel, D.v., Eupen, M.v., & Heiligenberg, H.A.R.v., 2010), Modeling of land cover and agricultural change in Europe: Combining the CLUE and CAPRI-Spat approaches.(Britz W., Verburg P.H., Leip A, 2010).

The notable one, upon which this thesis takes its first step, was proposed by M.-J. Gaillard et al.,(2010) and Sugita S.,(2006). Regional Estimates of VEgetation Abundance from Large Sites (**REVEALS**) was introduced by M.-J. et al., (2010) as a new method to discussing issues related to pollen-based reconstruction of the past land-cover. The REVEALS model estimates the percentage cover of species or taxa (group of species, genera, group of genera, or family). The species and taxa correspond to the pollen types that can be identified using pollen-morphological characteristics. REVEALS requires raw pollen counts, site radius, pollen productivity estimates (PPEs), and fall speed of pollen (FS) to estimate vegetation cover in percentages, (M.-J. et al., (2010)). The **REVEALS** model-based land-cover reconstruction has been demonstrated to provide better estimates of regional vegetation or land-cover changes than the traditional use of pollen percentages. For instance, the effectiveness of REVEALS has been empirically tested and shown to be satisfactory in southern Sweden (Hellman et al., 2008a, b).

In this thesis, we aim at using bio-climate variables to model land-cover of the past. Plants grow under different climatic conditions and on different soil types. The different plants that we have at any point in time give us the different land-cover types. Such land-cover types may include; Open-land, Summer-green and Ever-green among others. The data on the different plants were measured by Plant Functional Types (PFTs). The PFTs give us the group of plants that grow at any point on the globe. These PFTs were proposed by M.-J. Gaillard et al.,(2010). Among others, we attempt to build a regression model to verify the relationship between these PFTs and the bio-climate variables, the combinations of the PFTs and bio-climates that make up the land-cover types. We will use the models to predict PFT values for Europe using the REVEALS data.

The **LPJ** (Lund Potsdam Jena) **– GUESS** (General Ecosystem Simulator) model (**LPJ-GUESS**, Smith et al., 2001) is a dynamic, process-based vegetation model optimized for application across a regional grid that simulates vegetation dynamics based on climate data input. It represents landscape and stand-scale heterogeneity and, by resolving horizontal and vertical vegetation structure at these scales, more adequately accounts for the biophysical properties that influence regional climate variability.

Finally, we will compare our predicted PFTs using the simplified regression models to those of REVEALS and LPJ-GUESS.

1.2 OUTLINE

In Chapter one, we give a brief statement of the problem that we attempt to deal with. The main focus is on land-cover modeling using bio-climate variables. The approach is based on multiple linear regressions. We again give a concise description of the data and the background to their collection. Among the goals of this work is to verify how important bio-climate variables are in determining the PFTs.

Chapter two deals with models of the individual PFTs. This is where we construct multiple linear regressions for the ten PFTs using the bio-climate variables as covariates. We estimate the model parameters using the Least Squares principles and then assess the model assumptions using graphical residual analysis.

The third chapter deals with composite models. Here the ten PFTs will be grouped into three land-cover types: Evergreen canopy, Summer-green canopy and Open-land. As in chapter two, we fit regression models for these three land surface types. Then again, we apply the Least Squares Principles to estimate the model parameters. The plots of the residuals will be produced to assess how the models fit the data.

In the final chapter, we apply the models to the REVEALS and LPJ-GUESS datasets. Here we make predictions with the models for the individual PFTs as well as using the composite models. Again we compare our predicted PFTs using the simplified regression models and the REVEALS data to the original PFTs from REVEALS and LPJ-GUESS. Our conclusions will be deduced from these comparisons at the end.

1.3 Data description and Background

This thesis uses data from the methods proposed by M.-J. Gaillard et al., (2010) in discussing issues related to pollen-based reconstruction of the past land-cover. We have data on bio-climate variables from the project; LAND cover-CLIMate interactions (**LANDCLIM)** in NW Europe during the Holocene. These bio-climate variables date back to **1901-1950**. The **LANDCLIM** results were expected to provide crucial data to assess Anthropogenic Land Cover Change (**ALCC**) estimates for a better understanding of the land surface-atmosphere interactions (M.-J. Gaillard et al., 2010). The Plant Functional Types (PFTs) were also proposed by M.-J. Gaillard et al., (2010).

1.4 The REVEALS and LPJ-GUESS

Regional Estimates of VEgetation Abundance from Large Sites (**REVEALS**) was introduced by M.-J. et al., (2010) as a new method to discussing issues related to pollen-based reconstruction of the past land-cover. The REVEALS model estimates the percentage cover of species or taxa (group of species, genera, group of genera, or family). The species and taxa correspond to the pollen types that can be identified using pollen-morphological characteristics. REVEALS requires raw pollen counts, site radius, pollen productivity estimates (PPEs), and fall speed of pollen (FS) to estimate vegetation cover in percentages.

The **REVEALS** model-based land-cover reconstruction has been demonstrated to provide better estimates of regional vegetation/land-cover changes than the traditional use of pollen percentages. For instance, the effectiveness of REVEALS has been empirically tested and shown to be satisfactory in southern Sweden (Hell-man et al., 2008a, b). Based on this, the REVEALS would be used as a comparing standard for the predictions and estimations from our approach in modeling land-cover of the past.

The **LPJ** (Lund Potsdam Jena) **– GUESS** (General Ecosystem Simulator) model (**LPJ-GUESS**, Smith et al., 2001) is a dynamic, process-based vegetation model optimized for application across a regional grid that simulates vegetation dynamics based on climate data input. It represents landscape and stand-scale heterogeneity and, by resolving horizontal and vertical vegetation structure at these scales, more adequately accounts for the biophysical properties that influence regional climate variability. LPJ-GUESS has been interactively coupled to the Rossby Centre Regional Atmospheric model version 3 (**RCA3**), (Wramneby et al., 2009) and is being used to study the feedbacks of climate-driven vegetation changes on climate, via changes in albedo, roughness, hydrological cycling and surface energy fluxes. Preliminary results suggest that changes in treelines, phenology of conifer versus broadleaved trees, and LAI may modify the future climate development, particularly in areas close to treelines and in semi-arid areas of Europe (Wramneby et al., 2009).

1.5 The Bio-climate Variables

The table below shows the bio-climate variables. The first column shows the Bio-climates which would be used as the covariates in the models that will be constructed later on. In the second column, we have the descriptions of the bio-climates.

Table 1: **The Bio-climate Variables**

All values are calculated for period 1901 - 1950

Bio-climate variable	**Description**
MAXgdd5	maximum annual amount of growing degree days over 5 degrees
MINgdd5	minimum annual amount of growing degree days over 5 degrees
MEANgdd5	mean annual amount of growing degree days over 5 degrees
MAXmtmin	maximum monthly mean temperature of coldest months
MINmtmin	minimum monthly mean temperature of coldest months
MEANmtmin	mean monthly mean temperature of coldest months
MAXmtmax	maximum monthly mean temperature of warmest months
MINmtmax	minimum monthly mean temperature of warmest months
MEANmtmax	mean monthly mean temperature of warmest months
MAXamean	maximum annual mean temperature
MINamean	minimum annual mean temperature
MEANamean	mean annual mean temperature
MAXaprec	maximum annual mean precipitation
MINaprec	minimum annual mean precipitation for period
MEANaprec	mean annual mean precipitation
MAXsawcont_upper50	maximum annual soil water content of upper soil layer (50 cm)
MINsawcont_upper50	minimum annual soil water content of upper soil layer (50 cm)
MEANsawcont_upper50	mean annual soil water content of upper soil layer (50 cm)
MAXsawcont	maximum annual soil water content of soil layer (200 cm)
MINsawcont	minimum annual soil water content of soil layer (200 cm)
MEANsawcont	mean annual soil water content of soil layer (200 cm)

1.6 REVEALS and Bio-climate locations

In figure1 we have a map showing the points on the globe where data from REVEALS and LANDCLIM for bio-climate variables were available. In the figure, (o) shows that reveals data were available,(o) shows unavailability of reveals data and (·) indicates bio-climates points.

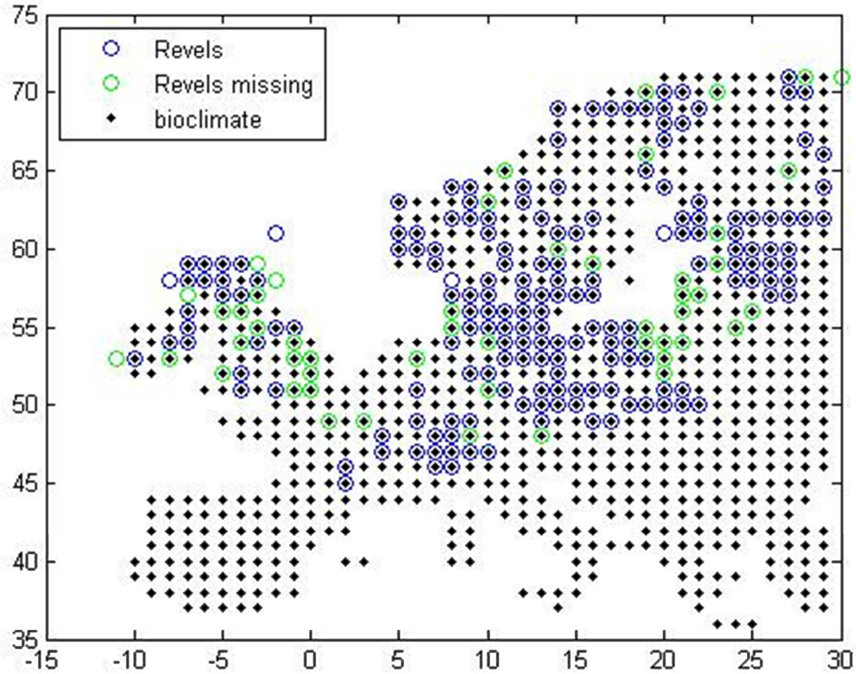


Figure 1: Reveals and Bio-climate availability points. In the figure, (o) shows that reveals data were available,(o) shows unavailability of reveals data and (·) indicates bio-climates points.

1.7 Project Goals

In this documentation, we aim at using bio-climate variables to model land-cover of the past. To this end, we will attempt to:

a) Figure out how important bio-climate variables are in explaining PFTs. This will be done using regression analysis.
b) Produce PFT estimates over Europe based on the simplified regression model and the REVEALS data.
c) Predict and compare PFTs estimates from the model to those from REVEALS and LPJ-GUESS datasets.

1.8 Statistical and Analytical Methods

Regression analysis is a statistical tool for the investigation of relationships between variables. It is this technique that will be used in modeling the PFTs. The relationship between the PFTs and the bio-climate variables have been modeled using multiple linear regressions.

1.8.1 Regression Analysis

In statistics, regression analysis includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

General linear model

In the more general multiple regression model, there are *p* independent variables:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

where x_{ij} is the i^{th} observation on the j^{th} independent variable, and where the first independent variable takes the value 1 for all *i* (so β_1 is on the line). $\mathbf{y}_i$ is the dependent variable and ε_i is the noise of the i^{th} observation.

The least squares parameter estimates $\hat{\beta}$, are obtained from ***p*** normal equations. The residuals can be written as

$$e_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}.$$

The normal equations are

$$\sum_{i=1}^{n} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i, \ \ j = 1, \ldots, p.$$

In matrix notation, the normal equations are written as

$$(\mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y},$$

Where the *ij* element of *X* is x_{ij}, the *i* element of the column vector *Y* is y_i, and the *j* element of $\hat{\beta}$ is $\hat{\beta}_j$. Thus *X* is $n \times p$, *Y* is $n \times 1$, and $\hat{\beta}$ is $p \times 1$. The solution is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error.
- The predictors are linearly independent.
- The errors are uncorrelated, that is, the variance-covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations.

1.8.2 Model selection using AIC

Variable selection

Which variables should be in the model?

If we have a limited number (***p***) of independent variables we can perform all possible linear regressions using the $\mathbf{2^p - 1}$ combinations and choose the best. This procedure quickly gets impossible when ***p*** is large.

Stepwise methods
Add (Forward selection) or remove (Backward elimination) or both (Stepwise regression) variables until we get a sufficiently good model. It is not guaranteed to find the best model. A model that explains as much of the variability as is practical is the best model (*not* as is possible).

The AIC
The Akaike Information Criterion (an information criterion-**AIC**) has been used by many people in statistical model selection. It is this methodology that we used in arriving at our final models. Below is a brief description of the AIC as published by Akaike H.(1974).

The **Akaike information criterion** is a measure of the relative goodness of fit of a statistical model. It was developed by **Hirotsugu Akaike**, under the name of "an information criterion" (AIC), and was first published by Akaike in 1974. It is grounded in the concept of information entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality. It can be said to describe the tradeoff between bias and variance in model construction, or loosely speaking between accuracy and complexity of the model.

In the general case, the AIC is

$$\mathbf{AIC = 2p - 2}\boldsymbol{ln}\mathbf{(L)}$$

where ***p*** is the number of parameters in the statistical model, and ***L*** is the maximized value of the likelihood function for the estimated model.

An alternative explanation of the AIC is given as

Information per parameter:

$$AIC(p+1) = n \ln(SS(\text{Error})_p) + 2(p+1) - n \ln n$$

Tradeoff between small residual error and large number of parameters $(SS(\text{Error})_p$ decreases and $p+1$ increases with p. The best model is the one with the smallest AIC. Tends to give too large models. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value, Akaike (1974).

The first step was to build a regression model for the PFTs. All the bio-climate variables were included in the model as covariates. Then the AIC was used to arrive at the final models.

CHAPTER 2

MODELING THE PLANT FUNCTIONAL TYPES (PFTs)

2.1 The PFTs

In this section, we give a brief description of the PFTs that would be modeled as a function of the bio-climate variables. The table below shows the PFTs. In columns one to three, we have PFTs, Defintion of the PFTs and Plant taxa or Pollen morphological types respectively.

PFT	**PFT Definition**	**Plant taxa/Pollen-morphological types**
TBE1	Shade-tolerant ever-green trees	*Picea*
TBE2	Shade-tolerant ever-green trees	*Abies*
IBE	Shade-intolerant ever-green trees	*Pinus*
TSE	Tall shrub evergreen trees	*Juniperus*
IBS	Shade-intolerant summer-green trees	*Alnus, Betula, Corylus, Fraxinus, Quercus*
TBS	Shade-tolerant summer-green trees	*Carpinus, Fagus, Tilia, Ulmus*
TSD	Tall shrub summer-green trees	*Salix*
LSE	Low ever-green shrub	*Calluna*
GL	Grassland-All herbs	*Cyperaceae, Filipendula, Plantago lanceolata, Plantago Montana, Plantago media, Poaceae, Rumex p.p. mainly R.acetosa and R. acetosella/Rumex acetosa-t*
AL	Agricultural land	*Cereals(Secale excluded)/ Cerealia-t, Secale*

Some images of the plants that belong to each category of the PFTs

2.1.1 TBE1

Norway spruce

The plant functional type (PFT) in this category covers the plants called ***PICEA ABIES.*** This plant is called by others as ***Norway spruce.*** The picture to the left is one example of the plants that belong to the PFT called **TBE1.** The TBE1 therefore comes from shade tolerant, boreal, evergreen plants (***Picea abies***)

2.1.2TBE2

Fagus sylvatica

The plant functional type (PFT) in this category covers the plants called ***ABIES ALBA.*** The picture to the left is one example of the plants that belong to the PFT called **TBE2.**

2.1.3 IBE

Skots pine

PINUS SYLVESTRIS(Skots pine) with common names such as Tall(Swedish), Furu(Norwegian), Pino Silvestre(Italian, Spanish),Skovfyr(Danish), and Manty(Finnish) among others, is the main plant in this category of the PFTs. It is shade intolerant evergreen coniferous tree that releases pollen in mid to late spring (Farjon A, 2005); (Steven, H.M, & Carlisle, A, 1959, facsimile reprint 1996)

2.1.4 IBS

Manna ash

Fraxinus Ornus(**Manna ash** or **South European Flowering Ash**) is the main tree in this group of our PFT classification which we called ***IBE***. It is a species of Fraxinus native to Southern Europe and southwestern Asia (Rushforth, K., 1999)

2.1.5 TBS

Fagus sylvatica

Shade tolerant, temperate, summergreen plants .Examples : *Acer* spp, *Carpinus betulus*, *Fagus sylvatica*, *Tilia cordata*, *Ulmus glabra*

2.1.6 TSE

Juniperus communis

Tall shrub, evergreen (*Juniperus communis*)

2.1.7 TSD

Tall shrub, summer-green (*Salix* spp.)

salix spp leaves

2.1.8 LSE

Low shrubs, evergreen (*Empetrum nigrum*, *Calluna vulgaris*). These are ever-green plants.

empetrum nigrum

2.1.9 GL

Here we have grasses (herbs, C3 grasses).

Examples are Kentucky blue grass , Tall fescue grass and Ryegrass

kentucky blue grass

2.1.10AL (Agricultural land)

The PFTs in this group come from agricultural land (AL). So they are mainly from food crops, for instance wheat grass.

Wheat grass

2.2 Plant Functional Types (PFTs) Models

We assumed that there exists a linear relationship between the PFTs and the bio-climate variables. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Thus, i= number of observations (n= 174) and p= independent variables (21 bio-climate variables)

where $^{\mathrm{T}}$ denotes the transpose, so that $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is the inner product between vectors $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$.

These *n* equations can be written in vector form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$\mathbf{y}$ is a vectors of dependent variables, $\mathbf{X}$ is design matrix.

In matrix notation, the normal equations are written as

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

The solution is given as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

2.2.1 Estimated model parameters

After fitting a linear regression to the PFTs using the bio-climate variables as covariates, and using Least Squares to estimate the model parameters, the results are presented in the table below. The figures in the brackets are the standard errors of the resulting estimates in each cell. The rows show the bio-climate variables and the columns show the PFTs. The cells containing zeros means that those Bio-climate variables are not important to the corresponding PFTs. In other words, the zeros show that those covariates were not present in the models.

	TBE1	TBE2	IBE	IBS	TBS
Intercept	0.62(0.19)	-0.098(0.16)	-0.13(0.16)	-0.4(0.024)	-0.087(0.13)
MAXgdd5	0.0004(0.00012)	0.00025(0.000097)	0	-0.00024(0.00013)	-0.00017(0.00010)
MINgdd5	0.0008(0.00023)	0.00029(0.00016)	-0.00013(0.000072)	0	-0.00032(0.00017)
MEANgdd5	-0.0016(0.00031)	-0.00042(0.00022)	0	0.00035(0.00018)	0.00035(0.00010)
MAXmtmin	0	-0.026(0.010)	0	0.032(0.016)	-0.025(0.011)
MINmtmin	-0.019(0.011)	-0.029(0.0081)	0.011(0.0065)	0	0
MEANmtmin	-0.046(0.027)	0.027(0.019)	0	0.086(0.019)	-0.039(0.014)
MAXmtmax	0	-0.029(0.012)	0	0	0
MINmtmax	-0.14(0.030)	0	0	0.066(0.038)	-0.057(0.022)
MEANmtmax	0.13(0.0032)	0	0.04(0.018)	0.073(0.038)	0
MAXamean	0	0	-0.12(0.043)	0	0.16(0.044)
MINamean	0	0	-0.084(0.037)	0	0.15(0.034)
MEANamean	0.16(0.041)	0.046(0.024)	0.16(0.060)	-0.26(0.043)	-0.17(0.054)
MAXaprec	0	0.00011(0.000070)	-0.00025(0.000087)	0	0
MINaprec	0	0	-0.00049(0.00014)	0	0
MEANaprec	0	-0.00016(0.00011)	0.00068(0.00019)	0	-0.000059(0.000029)
MAXsawcont_upper50	-0.34(0.24)	0	0	1.1(0.34)	0
MINsawcont_upper50	0	-0.35(0.16)	0.80(0.29)	-0.71(0.39)	-1.0(0.27)
MEANsawcont_upper50	0	0	-1.1(0.27)	-0.84(0.40)	2.3(0.39)
MAXsawcont	0	-0.66(0.22)	0.89(0.23)	0	0
MINsawcont	0	0	-0.47(0.28)	0.83(0.37)	1.2(0.30)
MEANsawcont	0.33(0.20)	1.1(0.25)	0	0	-2.3(0.41)

	TSE	TSD	LSE	GL	AL
Intercept	0.023(0.031)	0.0077(0.026)	0.56(0.23)	1.5(0.35)	0.23(0.093)
MAXgdd5	0	-0.000028(0.000016)	0	0	0
MINgdd5	-0.000078(0.000038)	0	0	0.00074(0.00039)	-0.00048(0.00013)
MEANgdd5	0.00012(0.000048)	0.000036(0.000021)	0	-0.00082(0.00038)	0.00063(0.00017)
MAXmtmin	0	0	0	0	0.018(0.0071)
MINmtmin	0	0	0.019(0.0079)	0	0
MEANmtmin	0.0092(0.0020)	0	0	0	0
MAXmtmax	0.0064(0.0025)	0.0049(0.0026)	0	0	0
MINmtmax	0	0	0	0	0.091(0.017)
MEANmtmax	0	-0.0063(0.0034)	-0.028(0.019	0	-0.084(0.014)
MAXamean	-0.024(0.0053)	0	0	0	-0.058(0.016)
MINamean	0	0	0	0	0
MEANamean	0	0	0.1(0.062)	0	0
MAXaprec	0.000027(0.0000072)	0.0000091(0.0000060)	-0.00014(0.000060)	0.000068(0.000043)	0
MINaprec	-0.000051(0.000016)	-0.000021(0.000012)	0.00029(0.00013)	0	0
MEANaprec	0	0	0	0	0
MAXsawcont_upper50	0.15(0.11)	0.15(0.058)	0	-1.4(0.53)	0
MINsawcont_upper50	0	0	0	0	0
MEANsawcont_upper50	-0.27(0.12)	-0.092(0.044)	-1.5(0.49	3.8(0.82)	0
MAXsawcont	-0.2(0.10)	-0.086(0.053)	0	0	0.34(0.15)
MINsawcont	0	0.045(0.027)	0	0	0
MEANsawcont	0.33(0.11)	0	1.3(0.46)	-3(0.72)	-0.39(0.12)

Table 2: **Estimated model parameters.** The first column shows the covariates whiles the rest of the columns show the PFTs. In each cell we have the parameter estimate and its standard error (in parenthesis). The cells containing zero indicate that those covariates were not significant in the model for that PFT.

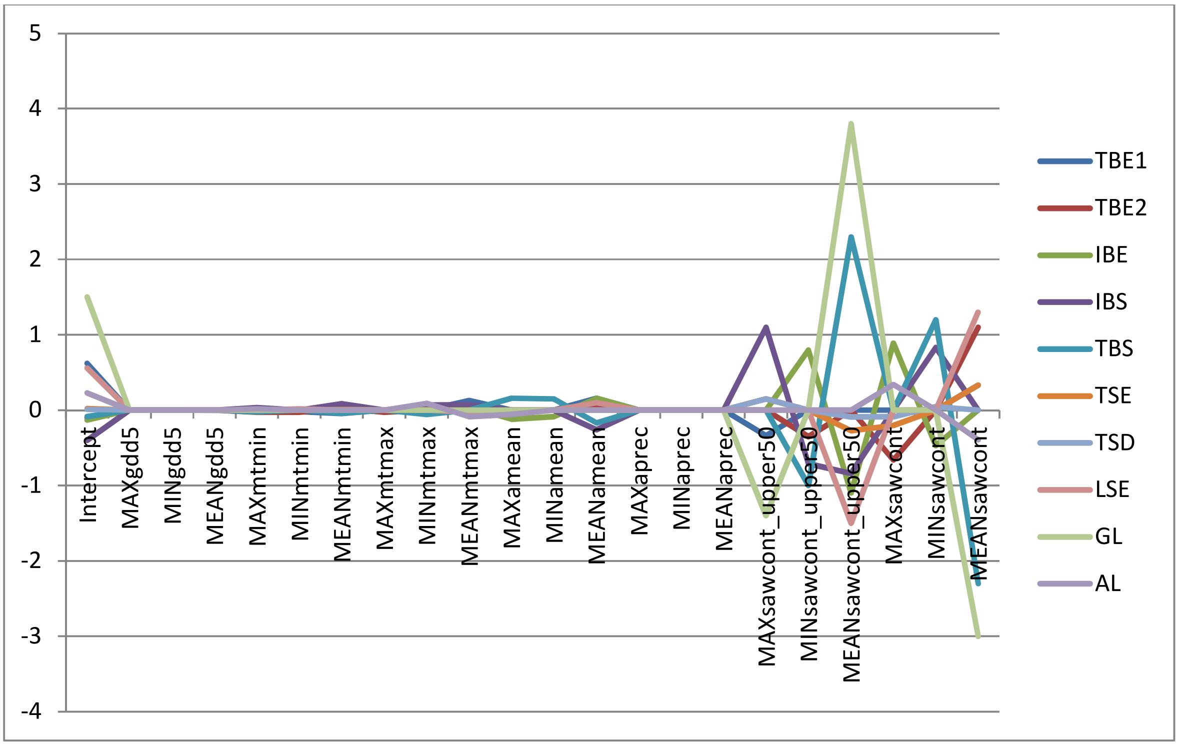


Figure 2: Plot of the estimated parameters: *The plot shows the bio-climate variables on the horizontal axis and the estimated model parameters on the vertical axis for the various PFT models. From this plot, we see a similar relationship between the PFTs and the bio-climate variables for all the models among the annual soil water content of the upper soil layers (50cm and 200cm).*

From table 2, we can write the models for each of the 10 PFTs as a function of the bio-climate variables. For instance, the regression equation for say TBE1 could be stated as

TBE1= $0.62 + 4.0\text{x}10^{-4}$Maxgdd5 + $8.0\text{x}10^{-4}$Mingdd5 – $1.6\text{x}10^{-3}$Meangdd5-$1.9\text{x}10^{-2}$MINmtmin – $4.6\text{x}10^{-2}$MEANmtmin – $1.4\text{x}10^{-1}$MINmtmax + $1.3\text{x}10^{-1}$MEANmtmax + $1.6\text{x}10^{-1}$MEANamean – $3.4\text{x}10^{-1}$ MAXsawcont_upper50 + $3.3\text{x}10^{-1}$MEANsawcont.

The meaning of this relationship is that, a per unit increase in the bio-climate variable with a positive coefficient will increase the plant functional type, in this case, TBE1 by the magnitude of the corresponding coefficient while a negative coefficient will reduce it as such when the other variables are held fixed. The equation also shows that, only ten of the twenty-one bio-climate variables were significant in the model for TBE1.

It is important also to note that, the magnitude of the coefficients of the covariates cannot be compared due to different scales of measurement. In fact, looking at the model as stated above for TBE1, we cannot say that the effect of MEANsawcont is higher than Maxgdd5 in the model even though the coefficients suggest that. The reason is due to differences in the measurement scale. However, we can compare the same bio-climate in different models. For instance, we can say TBE1 has stronger dependence on MEANamean than TBE2 as shown by the coefficients 0.16 and 0.046 respectively.

The rest of the models could be written in a similar manner.

2.2.2 Residual Analysis

The residuals of a fitted model are the differences between the responses observed at each combination of values of the explanatory variables and the corresponding prediction of the response computed using the regression function. Mathematically, the definition of the residual for the i^{th} observation in the data set is written

$$e_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}.$$

Model validation is possibly the most important step in the model building sequence. Often the validation of a model seems to consist of nothing more than quoting the $\boldsymbol{R}^2$ statistic from the fit (which measures the fraction of the total variability in the response that is accounted for by the model). Unfortunately, a high $\boldsymbol{R}^2$value does not guarantee that the model fits the data well. Use of a model that does not fit the data well cannot provide good answers to the underlying engineering or scientific questions under investigation.

There are many statistical tools for model validation, but the primary tool for most process modeling applications is graphical residual analysis. Different types of plots of the residuals from a fitted model provide information on the adequacy of different aspects of the model. Numerical methods for model validation, such as the $\boldsymbol{R}^2$ (coefficient of determination) statistic, are also useful, but usually to a lesser degree than graphical methods. Graphical methods have an advantage over numerical methods for model validation because they readily illustrate a broad range of complex aspects of the relationship between the model and the data. Numerical methods for model validation tend to be narrowly focused on a particular aspect of the relationship between the model and the data and often try to compress that information into a single descriptive number or test result.

In our models, we have assumed that the ε_i are normally distributed with mean zero and a constant variance. Thus $\varepsilon_i \approx N(\mathbf{0}, \mathbf{I}\sigma^2)$ and independent.

$\boldsymbol{e_i} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$, $\mathbf{I}\sigma^2 = \begin{pmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{pmatrix}$ and $\mathbf{0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$. Thus $\mathbf{I}$= (n × n identity matrix)

We assess how good our models fit the data by checking the assumption of normality of the residuals. We expect the residuals to be approximately normally distributed with mean zero and a constant variance. This is satisfied if the points of the Q-Q plot (Quantiles –Quantiles plot) lie in a straight line. The plot of the residuals should also show no patterns and scatter around zero and the density plot should be bell-shaped.

We take the model for **TBE1** as an example.

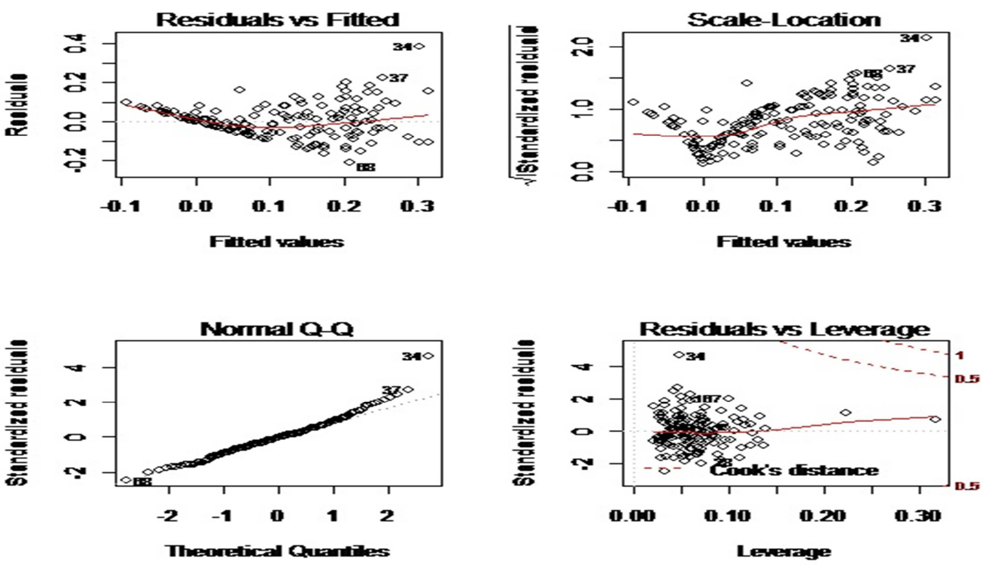


Figure 3: The Residuals and Q-Q plots of TBE1.*The figure shows the three different plots of the residuals and the Q-Q plot.*

The residuals (top left of figure3), residuals plotted against fitted values show some structural inadequacies and the presence of outliers. Also the residuals plot shows that the variance may not constant across all observations as assumed. However, the scatter plot of the residuals shown in figure 5 below show somewhat no pattern and the normality is also quite clear from the Q-Q plot. We conclude that the model is somehow good since the assumptions about the residuals are quite satisfied. The density plot as shown below is also bell-shaped as we expect it to be.

An important observation is the outliers. Outliers may have effect on the estimates from a given dataset. However, we did not take into account any significant effects on our models. Their inclusion or exclusion may not have much effect since they are just a few as compared to the size of the dataset. It would have been necessary to formally test using Cook’s Distance or other tests.

We therefore attribute the outliers to the dataset. This might have probably happened from faulty measurements or wrong coding of the data which are common errors in practice.

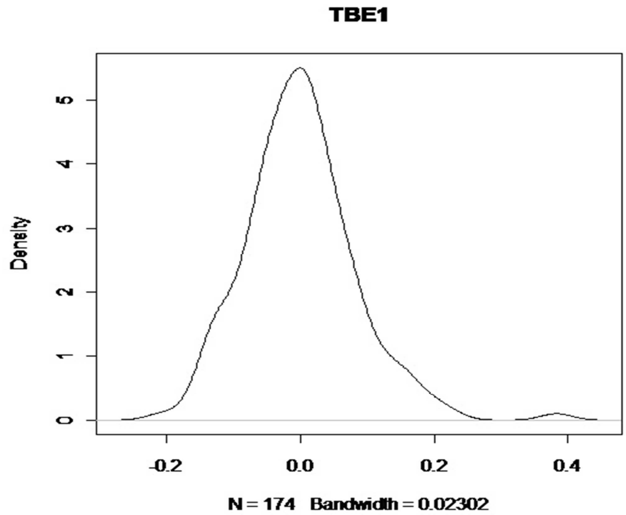


Figure 4: Density plot of TBE1.

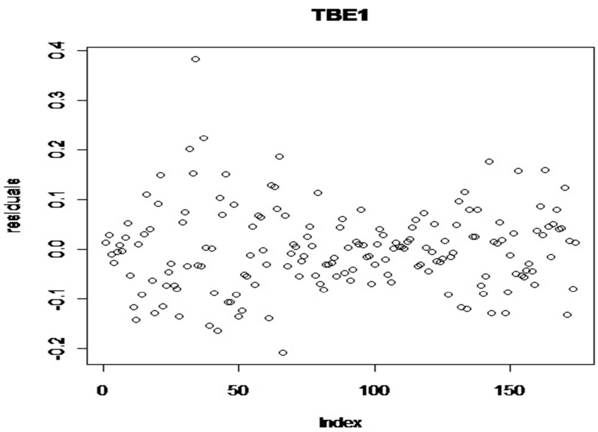


Figure5: Residuals plot of TBE1

Figure 6: Residuals and Q-Q plot of GL model. In the figure below we show the residuals and the QQ plots of the GL model.

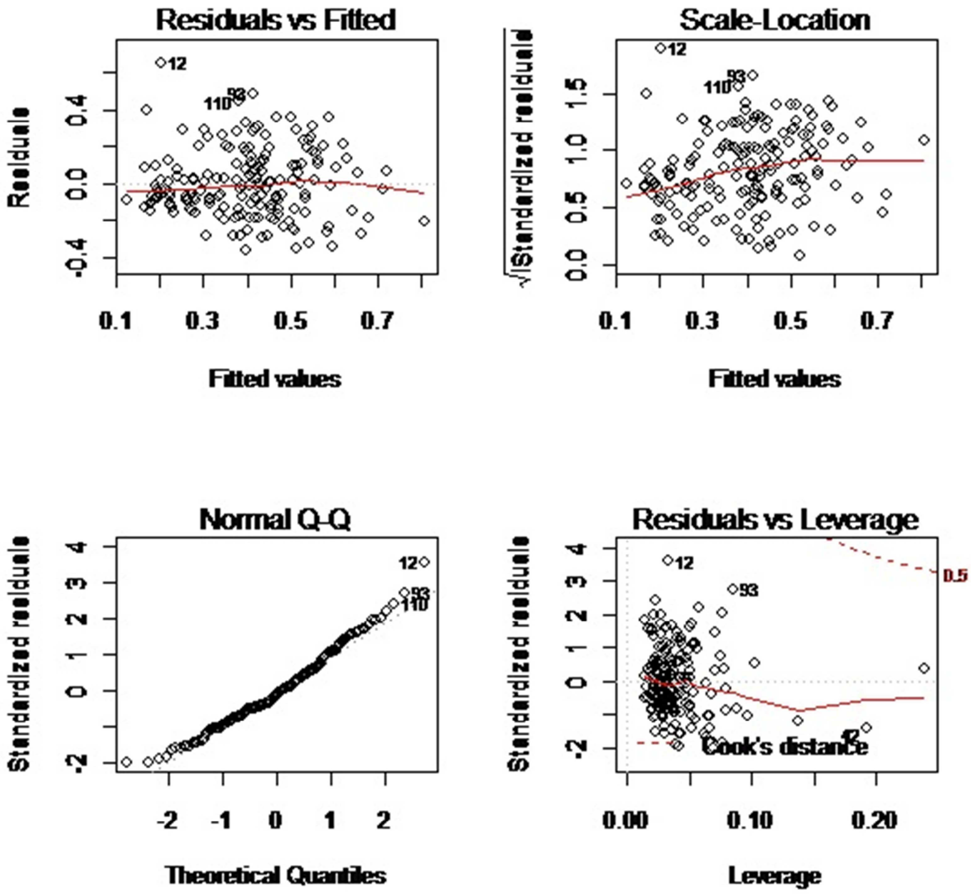
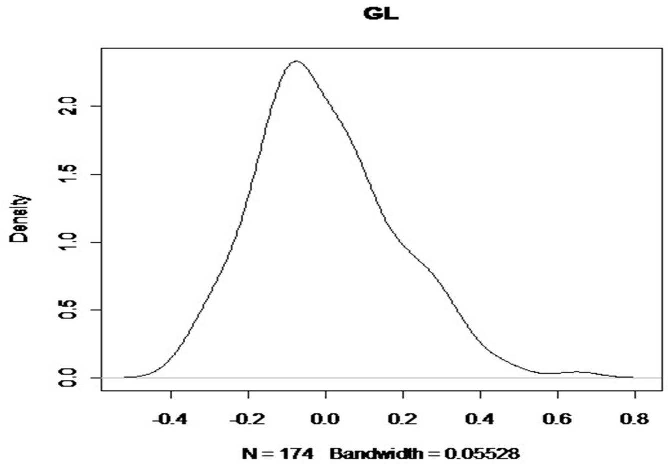


Figure 7: Density plot of GL model.



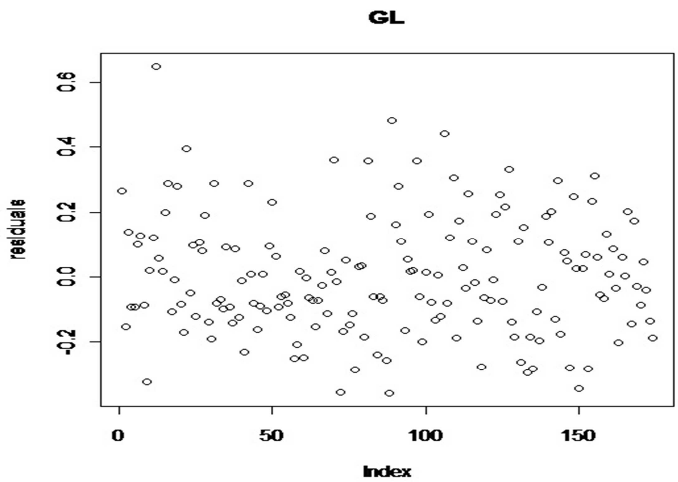


Figure 8: Residuals of GL model. *Shows no pattern and also scattered somewhat evenly around zero*

The residuals and density plots for the rest of the PFTs are shown in the APPENDIX A. Some of them look quite good.

Chapter 3

Composite Models

In Chapter two, the individual PFTs were modeled and the results presented in Table 2. It was seen that bio-climate variables are indeed important for PFTs.

In this Chapter, we attempt to model land-cover using the bio-climate variables. The ten PFTs were grouped into three land-surface types which we call land-cover types by M.-J. Gaillard et al.(2010).

The table below shows how the groupings were done. In columns one to four, we have PFTs, Definition of the PFTs, Plant taxa or Pollen morphological types and Land surface respectively. The Ever-green tree canopy is made up of TBE1, TBE2, IBE and TSE plant functional types. In terms of plant taxa, the Ever-green tree canopy is made of *Picea*, *Abies*, *Pinus* and *Juniperus* which are all Ever-green trees. Summer-green tree canopy is also made up of PFTs such as the IBS, TBS and TSD. The third land surface which is Open-land comprises of the PFTs LSE, GL (grass-land) and AL (agricultural land).

PFT	PFT Definition	Plant taxa/Pollen-morphological types	Land surface
TBE1	Shade-tolerant ever-green trees	*Picea*	Ever-green Tree Canopy
TBE2	Shade-tolerant ever-green trees	*Abies*	
IBE	Shade-intolerant ever-green trees	*Pinus*	
TSE	Tall shrub ever-green trees	*Juniperus*	
IBS	Shade-intolerant summer-green trees	*Alnus,Betula, Corylus, Fraxinus, Quercus*	Summer-green Tree Canopy
TBS	Shade-tolerant summer-green trees	*Carpinus, Fagus,Tilia, Ulmus*	
TSD	Tall shrub summer-green trees	*Salix*	
LSE	Low ever-green shrub	*Calluna*	Open-Land
GL	Grassland-All herbs	*Cyperaceae, Filipendula, Plantago lanceolata, Plantago Montana, Plantago media, Poaceae, Rumex p.p. mainly R.acetosa and R. acetosella/Rumex acetosa-t*	
AL	Agricultural land-cereals	*Cereals (Secale excluded)/ Cerealia-t, Secale*	

Table 3.1 Grouping of PFTs into Land-Cover types. *First column shows the PFTs, second shows PFT definitions, third shows Plant taxa or pollen-morphological types and the forth shows the Land-surface/land-cover.*

3.1 The Land covers

The PFTs were grouped into Evergreen Tree canopy, Summer-green Tree canopy and Open land as seen in Table 3.1. Using a similar approach as in chapter 2, a multiple linear regression model was then fitted to these land-cover types using the bio-climate variables as covariates.

Specifically,

Three vectors of land surface were formed by combining the PFTs as follows.

Ever-green = TBE1 + TBE2 + IBE + TSE

Summer-green = IBS + TBS +TSD

Open-Land = LSE + GL + AL

In matrix notation, the model could be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

Where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$\mathbf{y}$ is a vectors of the dependent variables and $\mathbf{X}$ is design matrix.

In matrix notation, the normal equations are written as

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

The solution is given as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

After fitting the model, the estimated parameters are shown in the table below.

Covariates	EVERGREEN	SUMMERGREEN	OPENLAND
Intercept	-1.1(0.032)	-0.32(0.24)	2.2(0.39)
MAXgdd5	0.00044(0.00018)	-0.00052(0.00019)	0
MINgdd5	0.0011(0.00040)	-0.00049(0.00033)	-0.00064(0.00043)
MEANgdd5	-0.0019(0.00056)	0.001(0.00039)	0.00079(0.00055)
MAXmtmin	-0.043(0.020)	0	0
MINmtmin	-0.051(0.0094)	0	0.0089(0.0059)
MEANmtmin	0	0	0
MAXmtmax	0	0	0
MINmtmax	-0.16(0.055)	0	0.14(0.056)
MEANmtmax	0.14(0.066)	0	-0.2(0.067)
MAXamean	0	0.24(0.068)	0
MINamean	0	0.17(0.057)	0
MEANamean	0.19(0.045)	-0.4(0.12)	0
MAXaprec	-0.00024(0.00016)	0	0
MINaprec	-0.00039(0.00028)	-0.00016(0.000081)	0
MEANaprec	0.00061(0.00037)	0	0
MAXsawcont_upper50	0	2.2(0.62)	-0.87(0.41)
MINsawcont_upper50	0.81(0.55)	-1.1(0.45)	1.6(0.62)
MEANsawcont_upper50	-2(0.83)	0	0
MAXsawcont	0	-1.1(0.71)	0
MINsawcont	-0.99(0.60)	1.5(0.54)	-1.2(0.66)
MEANsawcont	2.5(0.86)	-0.94(0.49)	0

Table 3.2 Estimated model parameters: Column1 shows covariates, columns 2, 3 and 4 show the estimated model parameters for Ever-green, Summer-green and Open-land respectively. The values in parenthesis show the standard error of corresponding estimate in each cell.

3.2 The Regression Equations

From the table 3.2, we state the equations as follows for the three land surface types.

Evergreen = $-1.1 + 4.4\times10^{-4}$ MAXgdd5 $+1.1\times10^{-3}$ MINgdd5- 1.9×10^{-3} MEANgdd5- 4.3×10^{-2} MAXmtmin – 5.1×10^{-2} MINmtmin - 1.6×10^{-1} MINmtmax + 1.4×10^{-1} MEANmtmax + 1.9×10^{-1} MEANamean – 2.4×10^{-4} MAXaprec – 3.9×10^{-4} MINaprec + 6.1×10^{-4} MEANaprec + 8.1×10^{-1} MINsawcont_upper50 – 2.0 MEANsawcont_upper50 – 9.9×10^{-1} MINsawcont + 2.5MEANsawcont

Summergreen = -3.2×10^{-1} - 5.2×10^{-4} MAXgdd5 – 4.9×10^{-4} MINgdd5 + 1.0×10^{-3} MEANgdd5 + 2.4×10^{-1} MAXamean + 1.7×10^{-1} MINamean – 4.0×10^{-1} MEANamean - 1.6×10^{-4} MINaprec + 2.2 MAXsawcont_upper50 – 1.1MINsawcont_upper50– 1.1MAXsawcont + 1.5MINsawcont - 9.4×10^{-1} MEANsawcont

Openland = 2.2 – 6.4×10^{-4} MINgdd5+7.9×10^{-4} MEANgdd5 + 8.9×10^{-3} MINmtmin + 1.4×10^{-1} MINmtmax – 2.0×10^{-1} MEANmtmax – 8.7×10^{-1} MAXsawcont_upper50 –1.6 MINsawcont_upper50 – 1.2 MINsawcont

From the models above, we find that evergreen land surface depends on 14 of the 21 bio-climate variables. These bio-climate variables include annual amount of growing degree days over 5 degrees , monthly mean temperature of coldest months, mean temperature of warmest months, annual mean temperature , annual mean precipitation and annual soil water content of soil layer (50 cm). This is what one should expect since these factors are very important for the plants that fall in this category.

Similarly, summer-green canopy depends on these factors as well. As per the model, summer-green depends on only 12 of our 21 bio-climate variables.

Open-land on the other hand depends on just 8 out of the covariates. An important bio-climate variable to all the land surface types is the annual soil water content of soil layer (50 cm). This affirms the fact that some plants mostly need water in the soil to grow while others do not depending on the sign of the estimated parameter of the bio-climate variable in question.

3.3 The Plot of Estimated model parameters.

In the following figures, we plot the parameters from the models against the covariates to compare the estimates from the models.

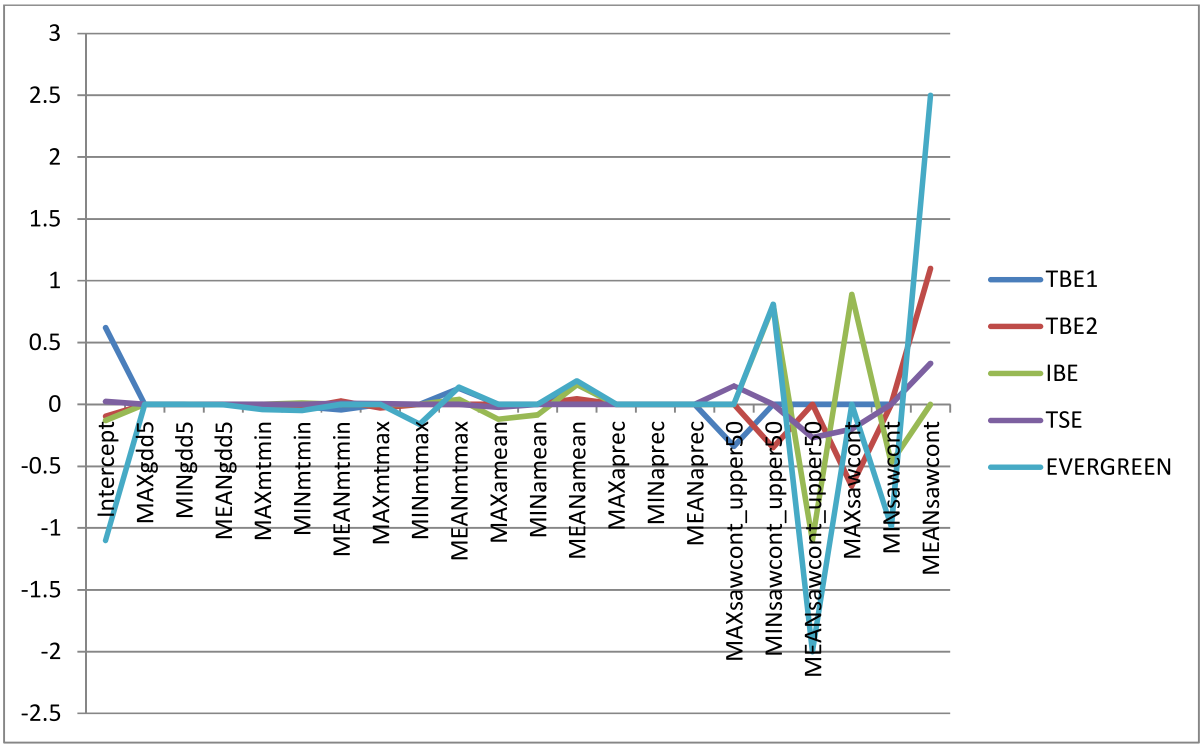


Figure 3.1 Plot of Estimated Parameters against Bio-climates (Ever-green canopy)

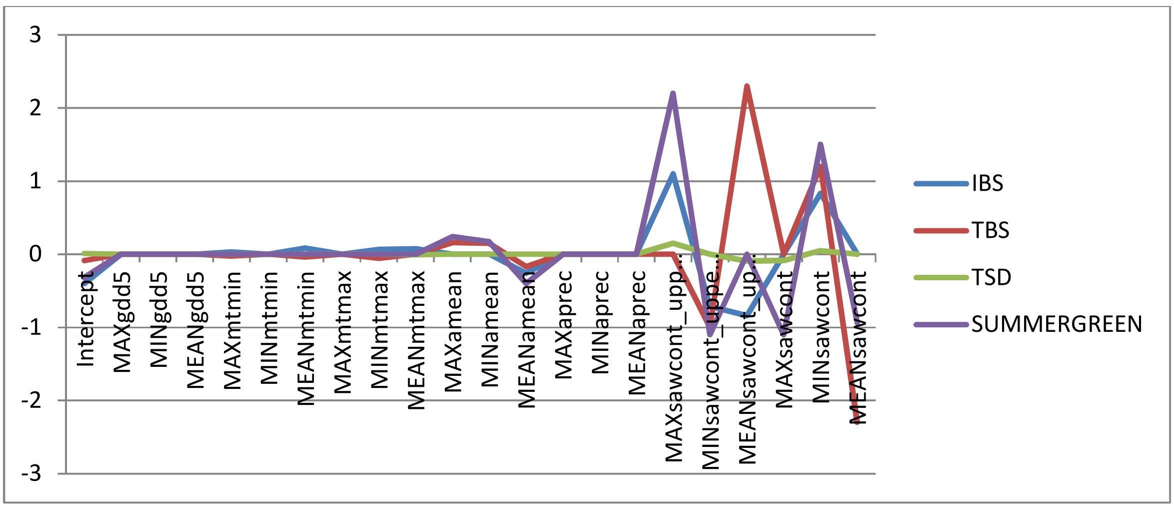


Figure 3.2 Plot of Estimated Parameters against Bio-climates (Summer-green canopy)

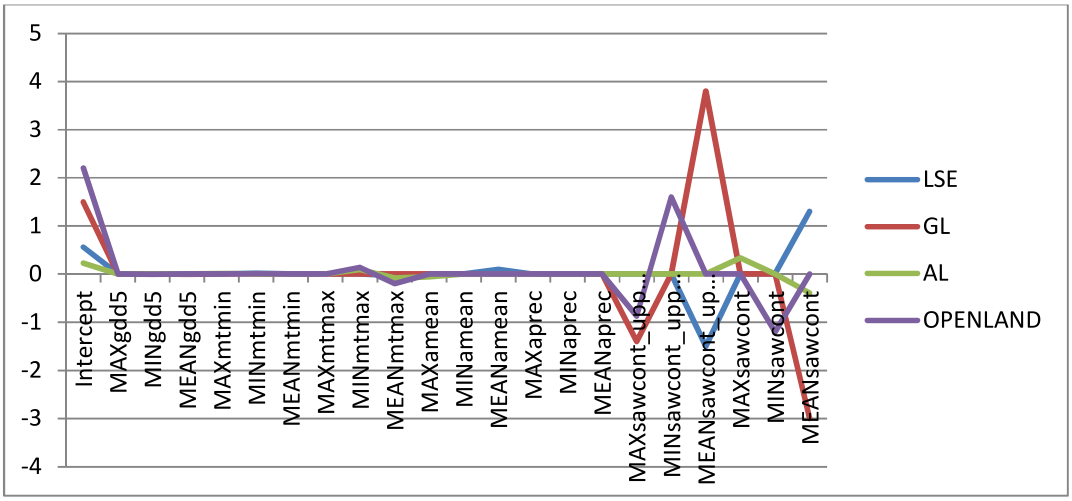


Figure 3.3 Plot of Estimated Parameters against Bio-climates (Open-land).

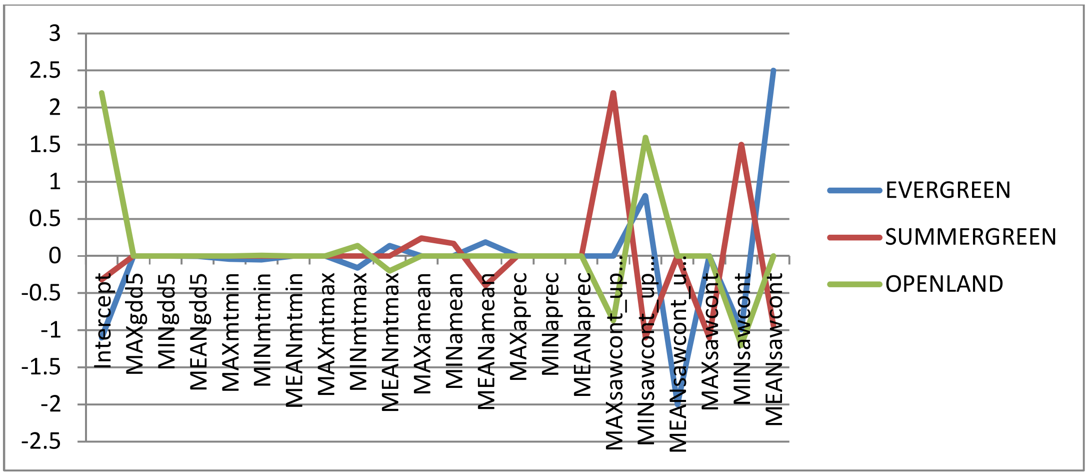


Figure 3.4 Plot of Estimated Parameters against Bio-climates. *The vertical axis shows the estimated model parameters of Ever-green (blue), summer-green (red) and open-land (green). On the horizontal axis we see the bio-climate variables.*

3.4 Residual Analysis

In the figure 3.5 below we present the residuals, QQ-plot and the density of the residuals for the three composite models.

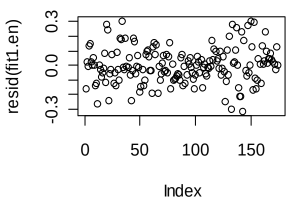
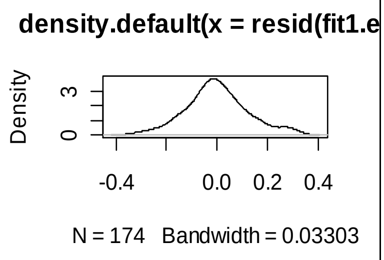
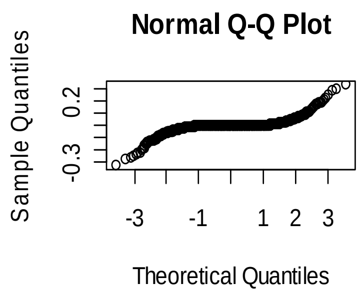
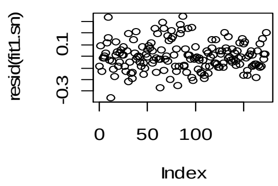
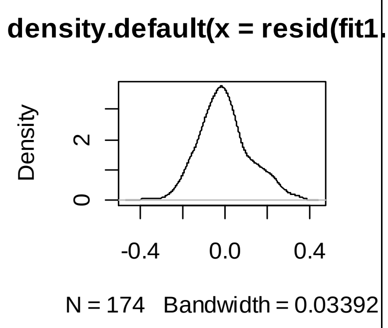
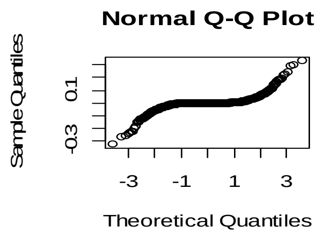
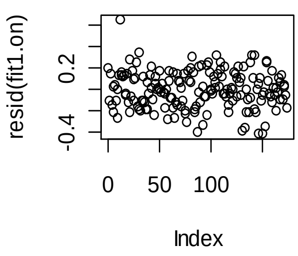
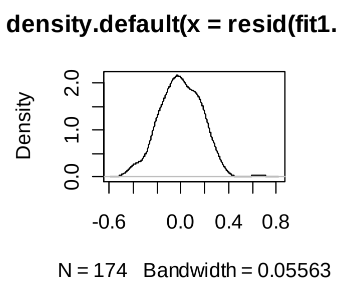
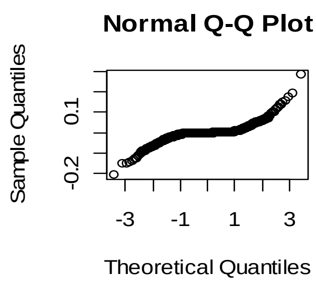
<table>
<tr><td>model</td><td>RESIDUALS</td><td>DENSITY PLOT</td><td>QUANTILES PLOT</td></tr>
<tr><td>EVERGREEN</td><td></td><td></td><td></td></tr>
<tr><td>SUMMERGREEN</td><td></td><td></td><td></td></tr>
<tr><td>OPENLAND</td><td></td><td></td><td></td></tr>
</table>

Figure 3.5: Column 1 shows the models, column 2 shows the residuals, column 3 shows the density plot and column 4 shows the Q-Q plots.

3.4.1 Ever-green Canopy

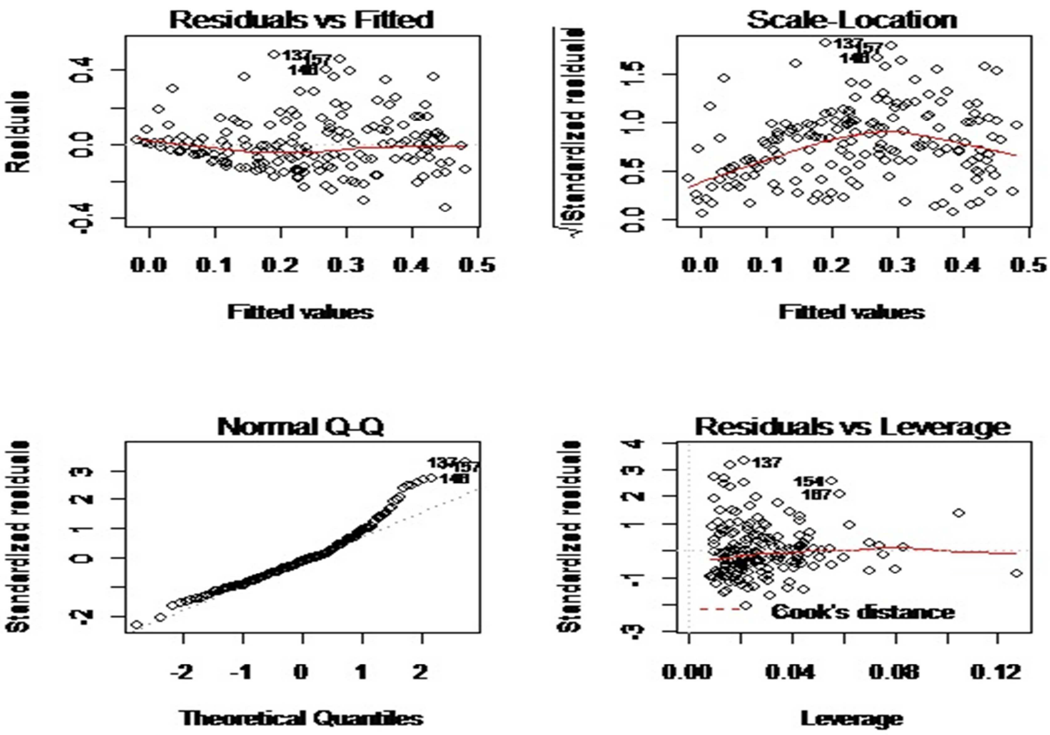


Fig. 3.6(a) Residuals (top left), Standardized residuals (top right), Q-Q plot (bottom left) and Cook’s distance of the Evergreen model.

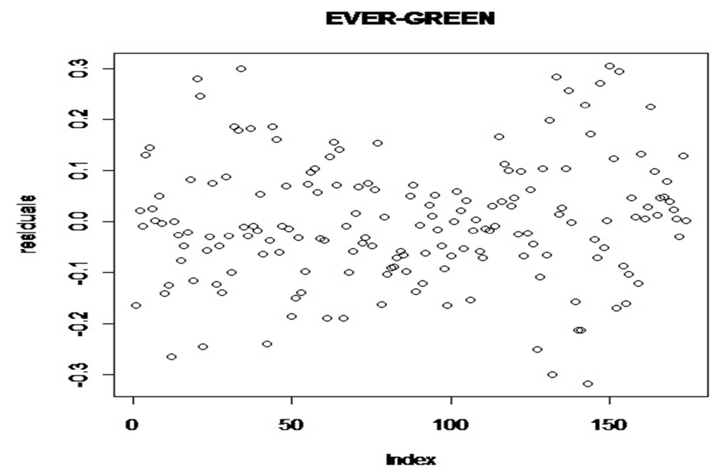


Figure 3.6(b) Residuals plot of Ever-green model.

3.4.2 Summer-green Canopy

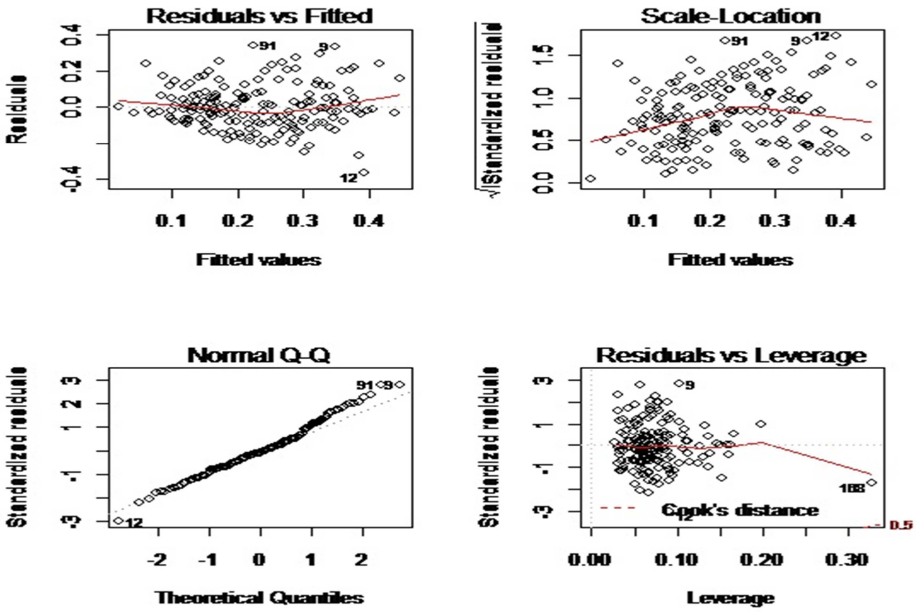


Fig. 3.7(a) Residuals (top left), Standardized residuals (top right), Q-Q plot (bottom left) and Cook's distance (bottom right) of the Summer-green model.

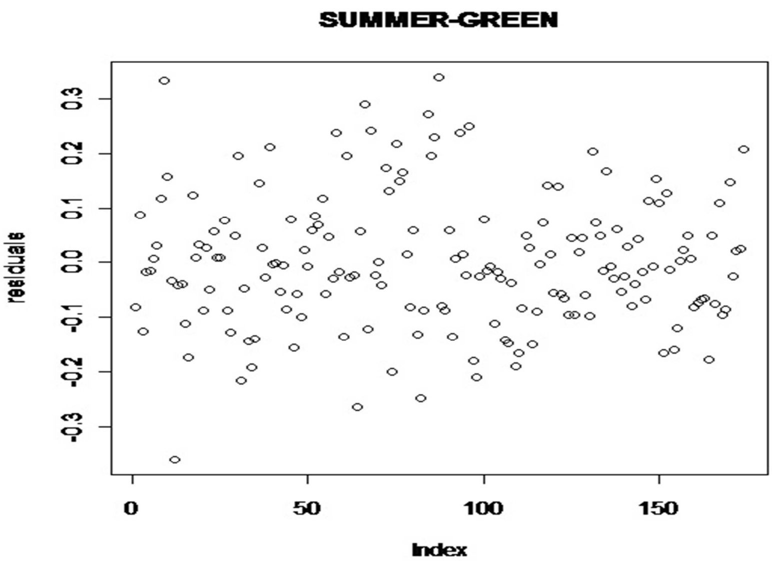


Figure 3.7(b) Residuals on index scale of Summer-green model.

3.4.3 Open-land

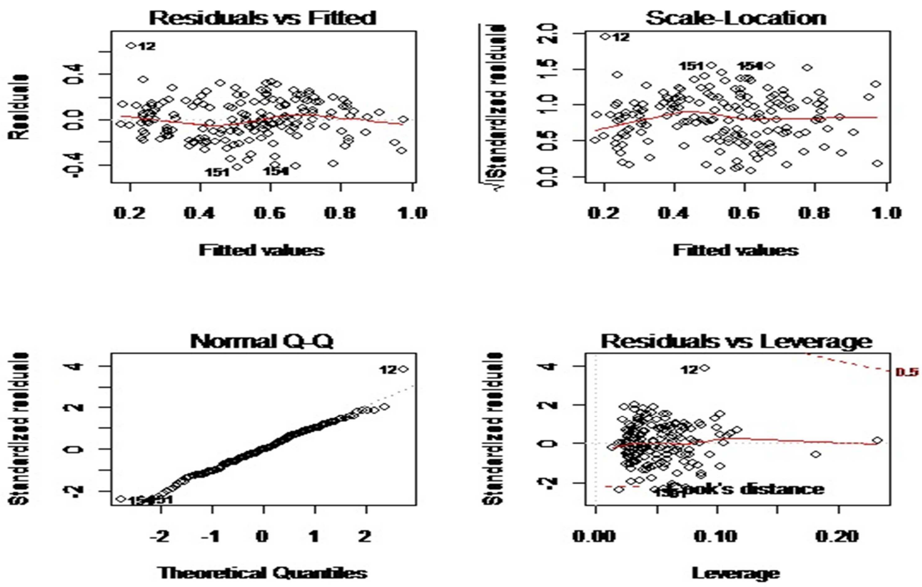


Fig. 3.8(a) Residuals (top left), Standardized residuals (top right), Q-Q plot (bottom left) and Cook's distance (bottom right) of the Ever-green model.

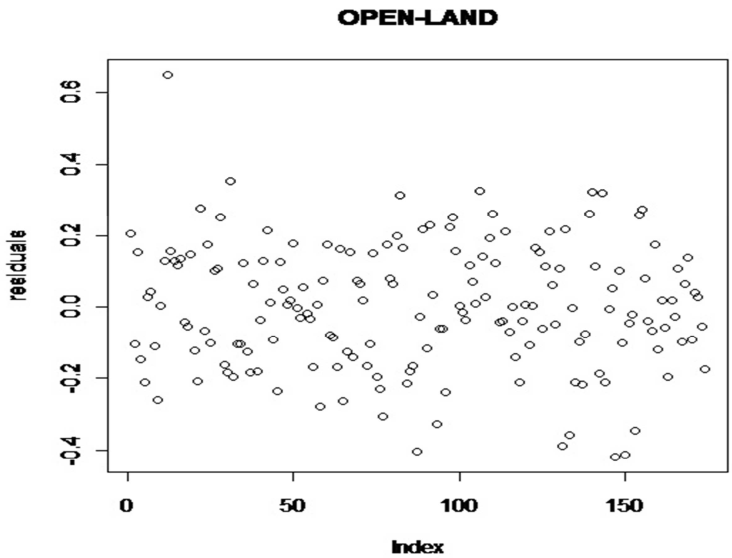


Figure 3.8(b) Residuals of Open-land model.

3.5 Summary and Discussions

Most significant covariates

It is important to note that most of the bio-climate variables were present in some of the models. So in conclusion, we say that the plant functional types (PFTs) depend on the bio-climate variables. Naturally, one should expect plant to grow well when they have their favorable climatic conditions. These include, soil water content, temperature, precipitation among others.

Interactions of covariates

In fitting the models, we first introduced the interaction terms of the covariates. However, those terms were seen to be insignificant.

Regression equations

Regression analysis is a statistical tool for the investigation of relationship between variables. It is this technique that has been used throughout in this thesis. In the end, the following regression equations were arrived at for the three different land surfaces which we called land-cover. Evergreen canopy depends on more bio-climate variables than the other two.

Ever-green = $-1.1 + 4.4\times10^{-4}$ MAXgdd5 $+1.1\times10^{-3}$ MINgdd5- 1.9×10^{-3} MEANgdd5- 4.3×10^{-2} MAXmtmin – 5.1×10^{-2} MINmtmin - 1.6×10^{-1} MINmtmax + 1.4×10^{-1} MEANmtmax + 1.9×10^{-1} MEANamean – 2.4×10^{-4} MAXaprec – 3.9×10^{-4} MINaprec + 6.1×10^{-4} MEANaprec + 8.1×10^{-1} MINsawcont_upper50 – 2.0 MEANsawcont_upper50 – 9.9×10^{-1} MINsawcont + 2.5MEANsawcont

Summer-green = -3.2×10^{-1} - 5.2×10^{-4} MAXgdd5 – 4.9×10^{-4} MINgdd5 + 1.0×10^{-3} MEANgdd5 + 2.4×10^{-1} MAXamean + 1.7×10^{-1} MINamean – 4.0×10^{-1} MEANamean - 1.6×10^{-4} MINaprec + 2.2 MAXsawcont_upper50 – 1.1MINsawcont_upper50– 1.1MAXsawcont + 1.5MINsawcont - 9.4×10^{-1} MEANsawcont

Open-land = 2.2 – 6.4×10^{-4} MINgdd5+7.9×10^{-4} MEANgdd5 + 8.9×10^{-3} MINmtmin + 1.4×10^{-1} MINmtmax – 2.0×10^{-1} MEANmtmax – 8.7×10^{-1} MAXsawcont_upper50 –1.6 MINsawcont_upper50 – 1.2 MINsawcont

The nature of the relationship between the PFTs and the bio-climates were similar to those between the Land-cover types and the bio-climate variables. That is to say, the individual PFTs models were somewhat consistent with the composite models.

The Residuals

The residuals plots as seen in the figures above (figure 3.6, 3.7 and 3.8) look quite good. The index plot of the residuals shows no pattern and are scattered around the zero line. This shows that they somewhat random in nature.

The residuals plotted against the fitted values also look somehow good. An important observation is the outliers. We see outliers in all the plots for the PFTs and the composite models. However we did not take into account their effect on the estimated model parameters. This is clearly not a good idea though. The density plots of the residuals do not deviate much from normality. This is seen in the Q-Q plots as well. It is however important to note the skewness which is seen in the density plots most especially in the individual PFTs models.

All these notwithstanding, we will proceed to use these models in the next chapter where we do the predictions for the PFTs for the entire Europe.

Chapter 4

Predictions and Comparisons

In chapter two, we fitted models for the individual PFTs. We extended our models by grouping the PFTs into three, namely: Evergreen, Summer-green and Open-land, and fitted models for them in Chapter three. The nature of the relationship was seen to be similar for the individual PFTs and the composite models. As a next step in our modeling sequence, we predict the PTFs based on our models and the REVEALS data.

Regression models predict a value of the *Y* (dependent) variable given known values of the *X* (independent) variables. Prediction within the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction outside this range of the data is known as extrapolation. Performing extrapolation relies strongly on the regression assumptions. The further the extrapolation goes outside the data, the more room there is for the model to fail due to differences between the assumptions and the sample data or the true values.

It is generally important that when performing extrapolation, one should accompany the estimated value of the dependent variable with a prediction interval that represents the uncertainty. Such intervals tend to expand rapidly as the values of the independent variable(s) moved outside the range covered by the observed data.

In figure 4.1 we have a map showing the points on the globe where data from REVEALS and LANDCLIM for bio-climate variables were available. In the figure, (o) shows that REVEALS data were available, (o) shows unavailability of reveals data and (·) indicates bio-climates points.

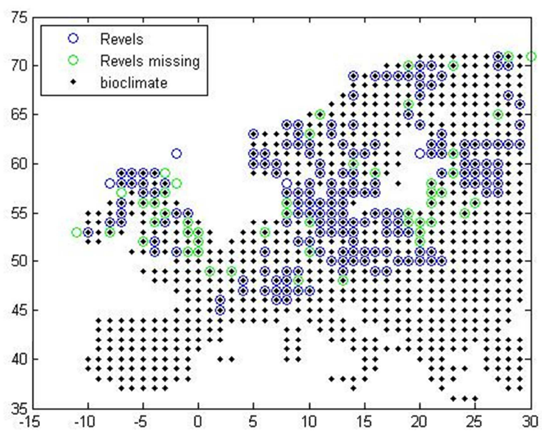


Figure 4.1: REVEALS and Bio-climate variables availability points

4.1 REVEALS
In constructing our regression models for the PFTs, we used the REVEALS dataset. The REVEALS dataset has 178 points as we see in the figure 4.1 above. For these locations, 4 did not have data on bio-climate variables and so were deleted from the dataset giving a total of 174 data points for our model building. These locations have data on PFTs and bio-climate variables.

4.1.1 Plot of PFTs from REVEALS
In the figure below, we see the plot of the PFTs from the REVEALS dataset. These plots are from the original measurements of PFTs from the REVEALS model-based land-cover reconstruction. The PFTs abundance increases from blue to red as shown in figure 4.2.

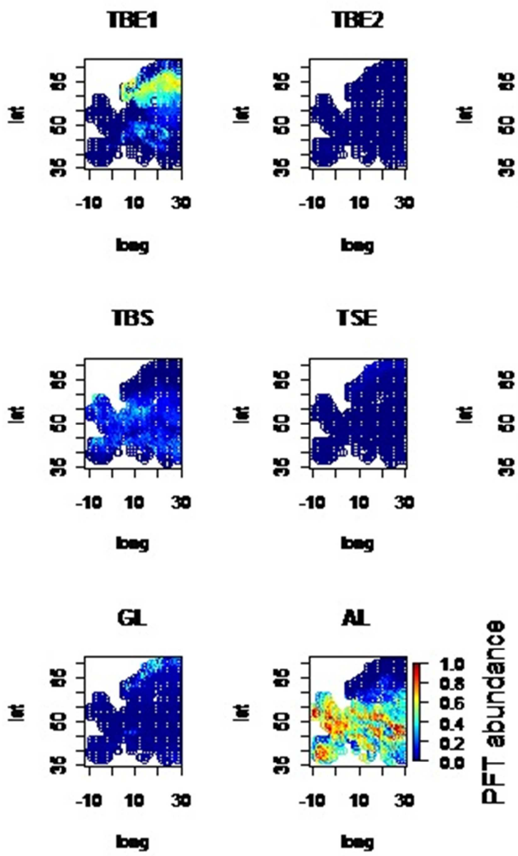


Figure 4.2 PFT plots from REVEALS

4.1.2 Composite plots from REVEALS
The figure 4.3 below shows the land-covers from the REVEALS data. As already stated in chapter 3, the PFTs were grouped into three land-cover types namely: Ever-green, Summer-green and Open-land. These plots were made using the original PFTs from REVEALS.

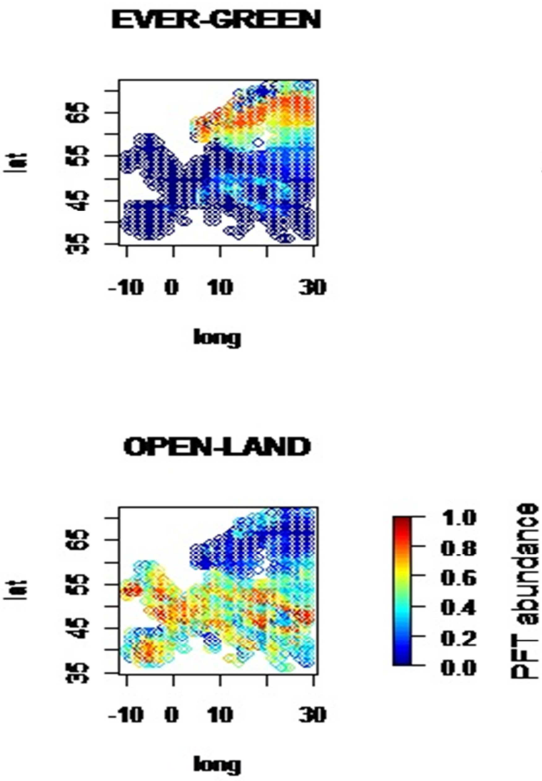


Figure 4.3 Land-covers from REVEALS. Top left shows the composite plot of Ever-green canopy, top right shows Summer-green canopy and lower shows Open-land.

4.2 LPJ-GUESS

The **LPJ** (Lund Potsdam Jena) **– GUESS** (General Ecosystem Simulator) model (**LPJ-GUESS**, Smith et al., 2001) is a dynamic, process-based vegetation model optimized for application across a regional grid that simulates vegetation dynamics based on climate data input. It represents landscape and stand-scale heterogeneity and, by resolving horizontal and vertical vegetation structure at these scales, more adequately accounts for the biophysical properties that influence regional climate variability. LPJ-GUESS has been interactively coupled to the Rossby Centre Regional Atmospheric model version 3 (**RCA3**), (Wramneby et al., 2009) and is being used to study the feedbacks of climate-driven vegetation changes on climate, via changes in albedo, roughness, hydrological cycling and surface energy fluxes. Preliminary results suggest that changes in treelines, phenology of conifer versus broadleaved trees, and LAI may modify the future climate development, particularly in areas close to treelines and in semi-arid areas of Europe (Wramneby et al., 2009). The **LPJ-GUESS** dataset contains the data on PFTs values as well as bio-climate variables.

4.2.1 Plots of PFTs from the LPJ-GUESS

In the figure 4.4, we show the PFTs plots from the LPJ-GUESS dataset. The plots look similar to those from the REVEALS dataset. It is important to state that the plots were made using the raw PFT values from the LPJ-GUESS dataset.

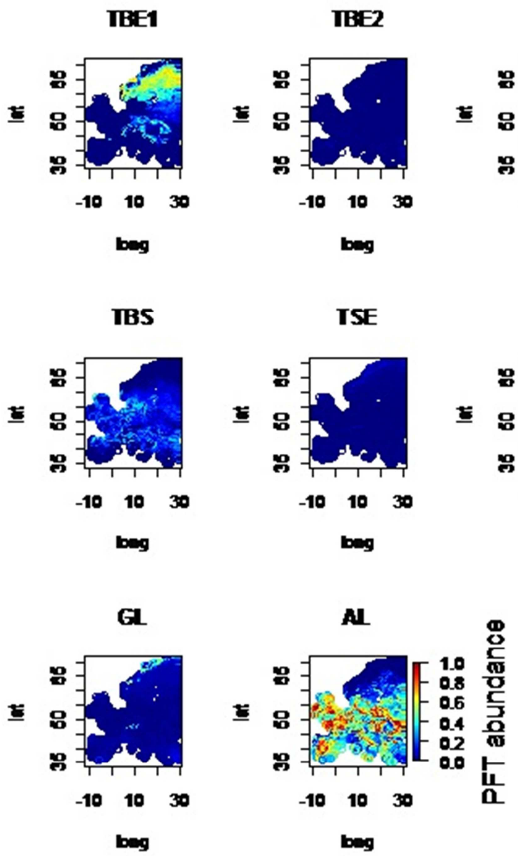


Figure 4.4 Plots of PFTs from LPJ-GUESS.

4.2.2 Composite plots from LPJ-GUESS

The images in figure 4.5 show the composite plots from the LPJ-GUESS dataset. These plots were made using the original values of PFTs in the LPJ-GUESS dataset. The PFT abundance increases from blue to red as shown in the legend of figure 4.5.

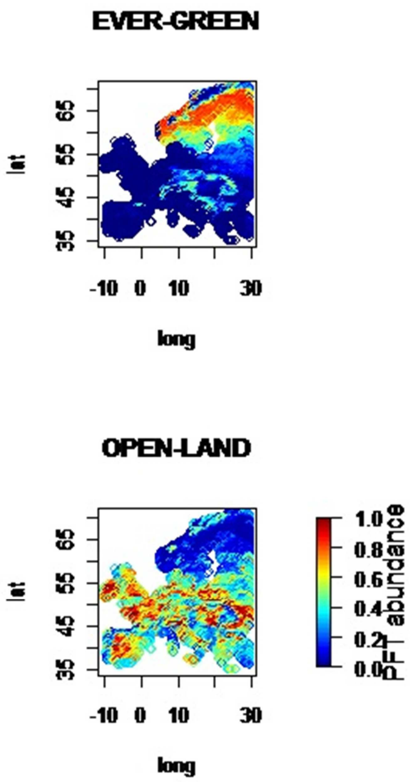


Figure 4.5 **Land-covers from LPJ-GUESS**. Top left shows the composite plot of Ever-green canopy, top right shows Summer-green canopy and the down one shows Open-land.

4.3 Predicted PFTs and Land-covers

Regression models predict a value of the *Y* (dependent) variable given known values of the *X* (independent) variables. Prediction within the range of values in the dataset used for model-fitting is known informally as interpolation. In this section we predict the PFTs using the simplified regression models. First, we make predictions for the individual PFTs and secondly, for the composite models. It is worth noting that these predictions are done within the dataset used for the modeling: REVEALS Dataset.

4.3.1 Predicted Plant Functional Types (PFTs)

The figure below shows the predicted PFTs using the simplified regression models for the individual PFTs and the REVEALS data. Using the regression models in Table 2 and the REVEALS dataset, we made predictions for the individual PFTs. Here we used the PFTs as the dependent variables and the bio-climate variables as independent. The predictions are therefore the conditional expectations of the PFTs given the bio-climate variables.

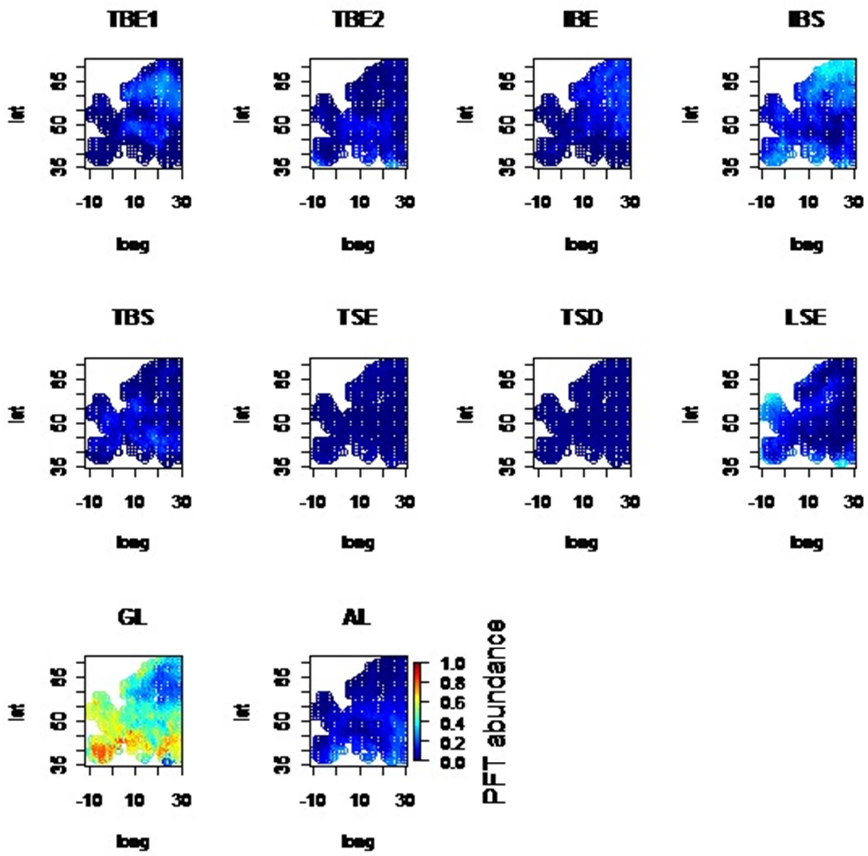


Figure 4.6 Predicted PFTs using the REVEALS dataset.

4.3.2 Estimated Land-covers

In this section, we attempt to estimate the land-covers: Ever-green, Summer-green and Open-land. To this end, we use the land-cover or composite regression models in Table 3.2 and the REVEALS dataset. So we compute the expected land-covers conditional on the bio-climate variables from the REVEALS dataset. Specifically, *E(Ever-green|Bio-climates), , E(Summer-green|Bio-climates), E(Open-land|Bio-climates)* were computed.

Figure 4.7 shows the estimated land-covers based on the simplified regression models in Section 3.1 and the REVEALS dataset.

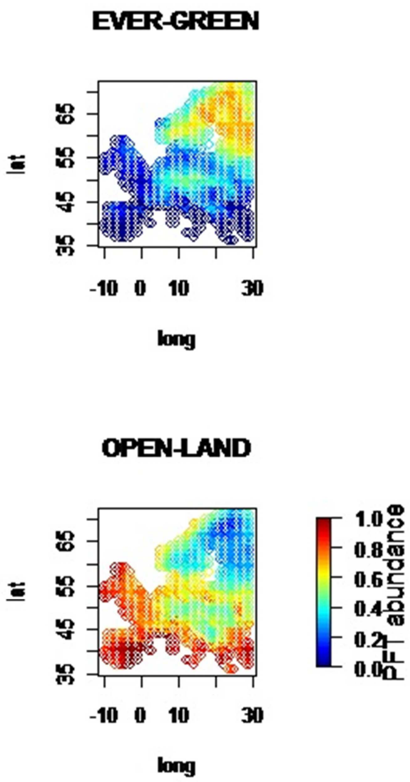


Figure 4.7 Estimated land-covers from REVEALS. Ever-green (top-left), Summer-green (top-right) and Open-land (down).

4.4 Comparisons

To validate our models, we now have to make comparative analysis using our estimates based on the Reveals dataset, the REVEALS and LPJ-GUESS. We will make the comparison using some of the individual PFT models and the composite models.

4.4.1 Criteria for Comparisons

In this section, we compare our PFT predictions using the regression models and the original PFT values from REVEALS. The idea employ here is to find the point-wise difference between our predicted PFT values and the original PFT values from REVEALS and normalized with the standard errors of the predicted PFT. That is at all locations where we have the REVEALS data, we compute the quantity:

$$\frac{Estimated_{pft} - REVEALS_{pft}}{SE(est)_{pft}}$$

where

$Estimated_{pft}$ = predicted PFT value using the model and the REVEALS data.

$REVEALS_{pft}$ = original PFT value from the REVEALS.

$SE(est)_{pft}$ = Standard error of the estimated PFT.

If this quantity is less than -1.96, then our predicted PFT at this location is considered too small compared to the original PFT value from the REVEALS and a value greater than 1.96 indicates too big estimate from our model. A value within the interval [-1.96, 1.96], indicates good estimate from the model.

Thus, our estimates compare well with REVEALS if the condition below is satisfied.

$$-1.96 \leq \frac{Estimated_{pft} - REVEALS_{pft}}{SE(est)_{pft}} \leq 1.96$$

4.4.2 Predicted PFTs and original PFTs from REVEALS.

Applying the criteria in section 4.4.1, we compared our predicted PFTs and REVEALS PFTs using the individual PFT models TBE1, TBE2 and GL.

In the figures shown below, the red(o) indicates that our estimate at that location was either too large or too small compared to the original PFT value in the REVEALS data whiles green(o) indicates good estimates.

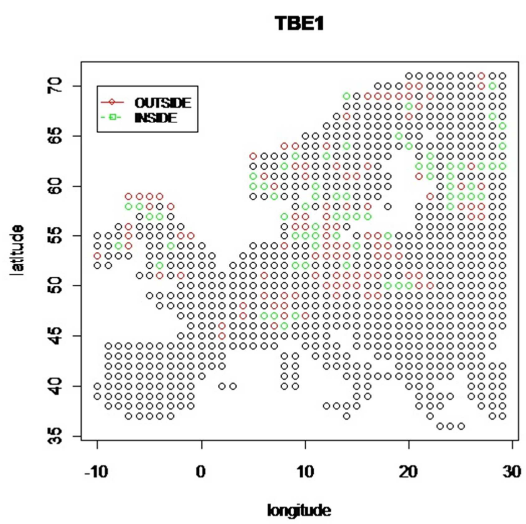


Figure 4.8 Comparing Reveals PFTs and Estimated PFTs using TBE1. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons.

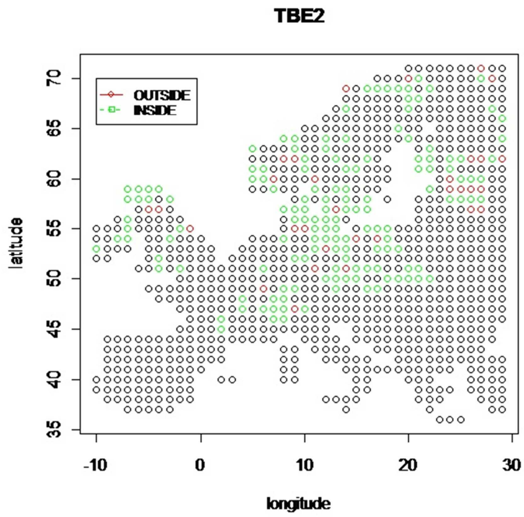


Figure 4.9 Comparing Reveals PFTs and Estimated PFTs using TBE2. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons.

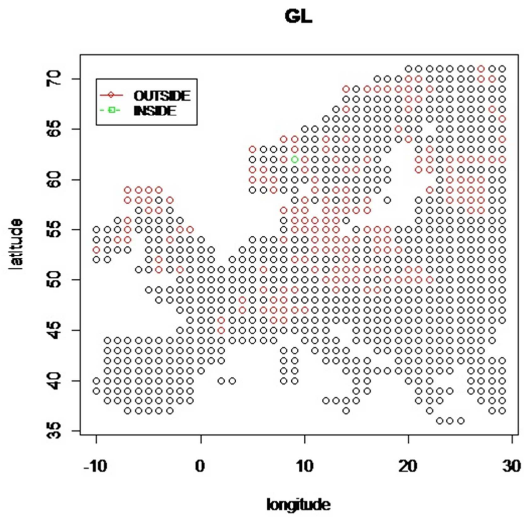


Figure 4.10 Comparing Reveals PFTs and Estimated PFTs using GL. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons.

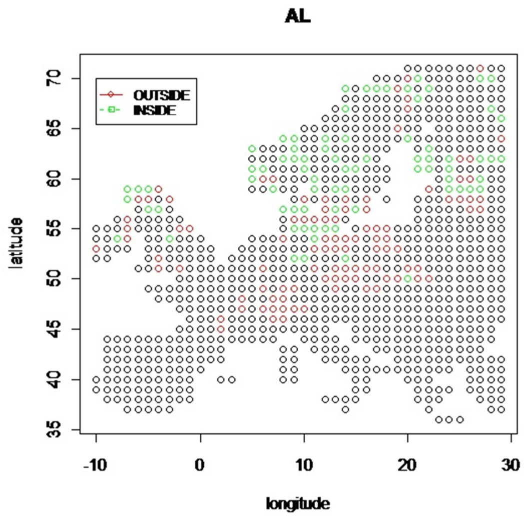


Figure 4.11 Comparing Reveals PFTs and Estimated PFTs using AL. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons.

From the figures (4.8, 4.9, 4.10 and 4.11), we find that predicted PFTs using TBE1 model compares quite well with the original PFTs from REVEALS. The comparison gets better with TBE2 where much more green points are seen from the figure 4.9. We however find a very strange results looking at the comparison with GL (Grass-land) model where we see only one location that gives good estimate of PFT. The case is completely different looking at the estimates from AL (Agricultural land).

The rest of the plots for this comparison are included in the Appendix. Most of them give good PFT estimates as compared to the Reveals PFTs.

4.4.3 Predicted and REVEALS PFTs by Land-cover models.

In section 4.4.2 we looked at how well our predicted PFTs compare with those from the REVEALS using the individual PFT models. In this section we want to do a similar comparative analysis using our composite models or the Land-cover models: Ever-green, Summer-green and Open-land. From the figures (4.12, 4.13 and 4.14), we find that Ever-green model gives much better PFT predictions compared to Summer-green and Open-land models.

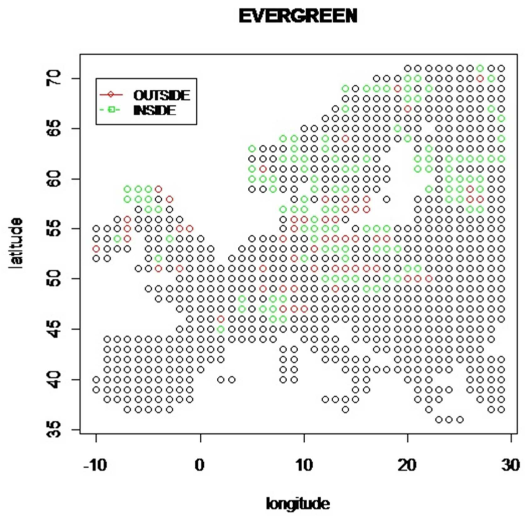


Figure 4.12 Comparing Reveals PFTs and Estimated PFTs using Ever-green model. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons. No REVEALS data at black locations (o).

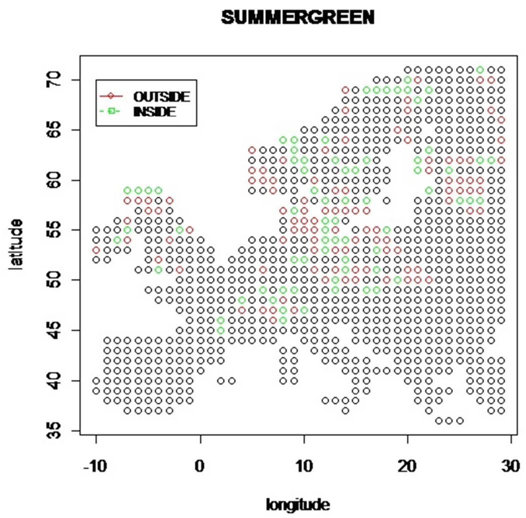


Figure 4.13 Comparing Reveals PFTs and Estimated PFTs using Summer-green model. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons. No REVEALS data at black locations (o).

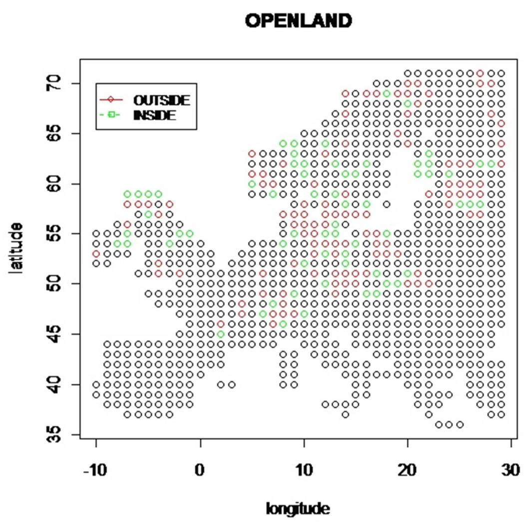


Figure 4.14 Comparing Reveals PFTs and Estimated PFTs using Open-land model. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons. No REVEALS data at black locations (o).

4.4.4 Predicted PFTs and LPJ-GUESS PFTs

In section 4.4.3, we predicted PFTs using the simplified Land cover models and the REVEALS dataset. It turned out that Ever-green model gives good PFTs predictions at most locations compared to Summer-green and Open-land models. It is important to state that, we made interpolation in section 4.4.3, which is making predictions within the dataset used for model fitting.

In this section however, we make extrapolation: making predictions outside the dataset used for model fitting. We used the REVEALS dataset to fit the models and now we want to apply the models to the LPJ-GUESS dataset.

After making the predictions, we used the criteria as explained in section 4.4.1. Thus as follows:

$Estimated_{pft}$ = predicted PFT value using the model and the LPJ-GUESS data.

$SE(est)_{pft}$ = Standard error of the predicted PFT.

$GUESS_{pft}$ = original PFT in GUESS dataset.

$$-1.96 \leq \frac{Estimated_{pft} - GUESS_{pft}}{SE(est)_{pft}} \leq 1.96$$

For all the locations, we deem our estimates as good compared to the original PFT if the above condition is satisfied.

Any value of $\frac{Estimated_{pft} - GUESS_{pft}}{SE(est)_{pft}}$ outside the interval [-1.96, 1.96] means that we either have very small or too big estimate compared to the original PFT at that location. That is 95% confidence interval around the estimated PFTs at all the locations on the map. In the figures below, the green points show that our estimates are good and the red points show otherwise.

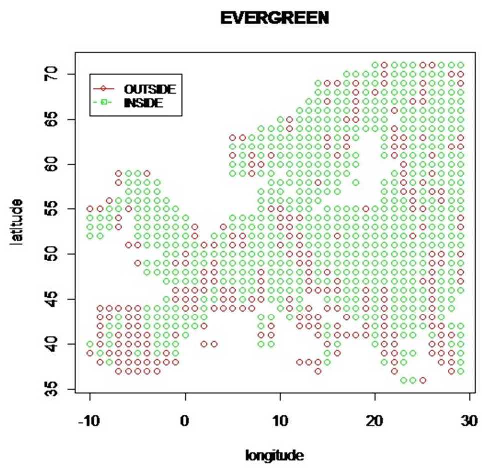


Figure 4.15 Comparing GUESS PFTs and Estimated PFTs using Ever-green model. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons.

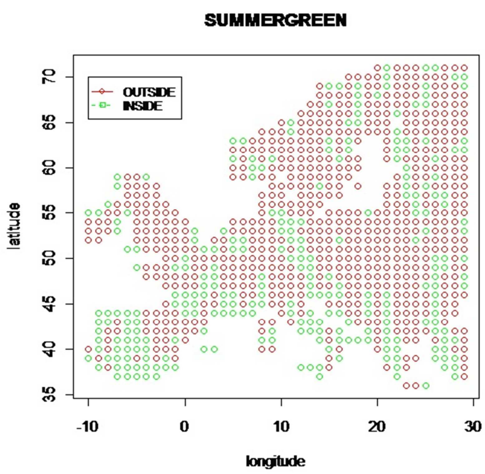


Figure 4.16 Comparing GUESS PFTs and Estimated PFTs using Summer-green model. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons.

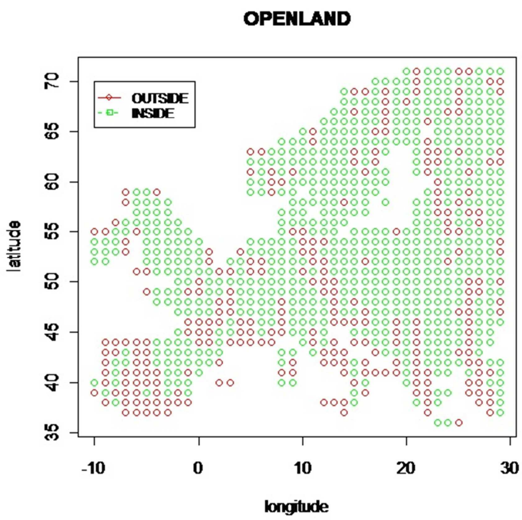


Figure 4.17 Comparing GUESS PFTs and Estimated PFTs using Open-land model. Green points (o) show locations where we have good comparisons and Red points (o) show locations where we have bad comparisons.

From the figures (4.15, 4.16 and 4.17), it is evident that Ever-green and Open land models make better predictions than Summer-green model.

Similar plots were made using the individual PFT models. Most of them do not give good predictions using the LPJ-GUESS dataset. See Appendix B. So the appropriate models to be used for predicting PFTs are the Land-cover or composite models because they fit the data better than the individual PFT models.

CHAPTER 5

CONCLUSIONS

It has been seen that bio-climate variables are important to the growth of plants thereby helping plants to produce pollens. Naturally, one should expect plants to grow well when they have their favorable climatic conditions. These include, soil water content, temperature, precipitation among others. Given a reliable and well-measured data of bio-climate variables and plant functional types, it is possible to use regression analysis to obtain a linear relationship between these plant functional types and the bio-climate variables. Consequently, it is feasible to model land-cover when we have bio-climate variables and plant functional types using multiple linear regression.

The land-cover models, gave better predictions when applied to the LPJ-GUESS dataset. For example, the Ever-green and open land models gave comparatively good estimates of PFTs.

Regression analysis is a statistical tool for the investigation of relationship between variables. It is this technique that was used in fitting the models. It has proven to be effective in explaining the relationship between the PFTs and the bio-climate variables. However, other statistical technique such as Logistic regression could have equally been used.

Future Work

Using a *Spatial Model* to evaluate the spatial structure that is left in the residuals.

PFT model output from LPJ-GUESS is 0.5 degree grid and PFT from REVEALS at the modern time window (the last 100 years) and at the spatial scale of 1 degree grid.

REFERENCES

M.-J. Gaillard et al. 2010; Holocene land-cover reconstructions for studies on land cover-climate feedbacks.

Farjon, A. (2005). Pines Drawings and Descriptions of the Genus Pinus 2nd ed. Brill ISBN 90-04-13916-8.

Steven, H. M., & Carlisle, A. (1959, facsimile reprint 1996). The Native Pinewoods of Scotland. Castlepoint Press.

Rushforth, K. (1999). Trees of Britain and Europe. Collins ISBN 0-00-220013-9.

Verburg, P.H.,Berkel, D.v., Eupen, M.v., & Heiligenberg, H.A.R.v., 2010. Trajectories of land use change in Europe: a model-based exploration of rural features. Landscape Ecology 25(2): 217-232.

Modeling of land cover and agricultural change in Europe: Combining the CLUE and CAPRI-Spat approaches.(Britz W, Verburg PH, Leip A, 2010).