ABSTRACT

FARJAT, ALFREDO EZEQUIEL. Optimal Seed Deployment under Climate Change using Spatial Models and Prediction of Genetic Merit in Loblolly Pine. (Under the direction of Brian J. Reich and Fikret Isik.)

Climate change is likely to affect agriculture and forestry globally. In this dissertation, we present methodologies to predict the effects of climate change on the dominant commercial pine species in the southeastern US making use of provenance tests data. The first chapter develops height growth prediction models using classical and penalized regression methods. The second chapter introduces a Bayesian spatial approach for modeling the relative performance of seed sources, which is used to develop a seed deployment tool. The third chapter presents methodology for the analysis of cloned progeny tests using multi-environmental trials data.

First, we present a statistical model to predict the effects of climate change on the height growth of loblolly pine (*Pinus taeda* L.) families using provenance test data. Provenance tests are a common tool in forestry designed to identify superior genotypes for planting at specific locations. The trials are usually replicated experiments established with seed from parent trees collected from different regions and grown at several locations. The geographic and climatic differences between the seed source and test site locations can be used to make predictions about the performance of provenances under different environmental conditions. Ordinary least squares, ridge regression, and LASSO regression were used to develop height growth prediction models. The models were tested using a hypothetical future climate scenario with 5% decrease in precipitation and 0.5°C increase in maximum and minimum temperatures, relative to historical average values. Under this scenario, local families from the coastal plains of Georgia, Florida, and South Carolina showed the highest performance relative to the current climate in their native environments. As these seed sources were moved to colder northern
and inland regions from their origin we observed decline in their height growth. Similarly, the climatic change scenario suggested that performance of northern seed sources decline significantly when they were moved to more southern warmer regions.

The second chapter presents a Bayesian spatial approach for modeling the expected relative performance of seed sources in terms of climate variables associated with the location of the origin of seed and the planting site. The proposed modeling technique accounts for the spatial dependence in the data and provides a flexible means to estimate effects associated with the origin and planting site locations. The statistical model was used to develop a quantitative tool for seed deployment aimed to identify the location of superior performing seed sources that could be suitable for a specific planting site under a specific climate scenario. Cross-validation results indicate that the proposed spatial models provide superior predictive ability compared to multiple linear regressions.

In the last chapter, a cloned progeny test of loblolly pine is analyzed to identify superior genotypes using multi-environmental trials data. The genetic material consisted of 51 crosses from 21 parents. Each cross had about 45 full-sib progeny resulting in a total of 2362 progeny that were cloned and tested in seven sites. Height, diameter, stem straightness, and fusiform rust incidence were assessed four years after planting. Genetic merits of clones were predicted for tree height using linear mixed models. Various covariance structures were employed to account for the heterogeneity in the data. Genotype-by-environment (G×E) interactions were assessed, and clusters of models were identified for making rankings based on genetic merit. The factor analytic formulation was parsimonious, informative, and provided a good approximation to the unstructured variance model. The environments exhibited relatively high pair-wise genetic correlation values, suggesting that G×E should not be a concern for the population under study. The clone means were reasonably highly repeatable suggesting that selection from tests of cloned progeny could be more efficient than from traditional seedling tests.
Optimal Seed Deployment under Climate Change using Spatial Models
and Prediction of Genetic Merit in Loblolly Pine

by

Alfredo Ezequiel Farjat

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2015

APPROVED BY:

Brian J. Reich
Co-chair of Advisory Committee

Fikret Isik
Co-chair of Advisory Committee

Joseph Guinness

Ross Whetten

Steven McKeand
DEDICATION

To Romina, whose love, support, and encouragement made this possible.
BIOGRAPHY

The author was born in Buenos Aires, Argentina. He graduated from the Universidad de Buenos Aires in 2005 with a Bachelor degree in Physics. He carried out his Physics’s thesis work in Tenaris’s Center for Industrial Research (CINI) in Campana, Buenos Aires, where he continued working after graduation for four years. In 2008, he was awarded a Fulbright scholarship to further his education in the United States. In 2009 he enrolled in North Carolina State University where he received a Masters degree in Financial Mathematics in 2011. After finishing his Masters degree, he continued at North Carolina State University as a doctoral student where he worked as Research Assistant and obtained a PhD in Statistics with co-major in Forestry & Environmental Resources in 2015.
ACKNOWLEDGEMENTS

I would like to thank and express my appreciation to my advisors Dr. Brian Reich and Dr. Fikret Isik for mentoring me throughout this research journey. I want to thank you both for your trust, patience, support, and guidance along the way. I am very grateful to Dr. Steve McKeand for encouraging my research and for allowing me to grow as a research scientist in his group. I would like to thank Dr. Ross Whetten for his many contributions, great ideas and suggestions for improving the research work. I would like to thank Dr. Joe Guinness for serving as my committee member at last moment and for his valuable comments and suggestions.

I would also like to thank my funding sources. The North Carolina State University Cooperative Tree Improvement Program, and its members, who established the field trials analyzed in this work. The Pine Integrated Network: Education, Mitigation and Adaptation Project (PINEMAP), a Coordinated Agricultural Project (CAP) funded by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (Grant 2011-68002-30185). The Southeastern Partnership for Integrated Biomass Supply Systems (IBSS) project supported by Agriculture and Food Research Initiative Competitive (Grant 2011-68005-30410) from the USDA National Institute of Food and Agriculture. The Department of Forestry and Environmental Resources at North Carolina State University. The USDA NIFA McIntire-Stennis (Project NCZ04214).
# TABLE OF CONTENTS

LIST OF TABLES ................................................................. viii

LIST OF FIGURES .............................................................. xi

Chapter 1  Modeling Climate Change Effects on Loblolly Pine  .................. 1

1.1 Introduction ................................................................ 1

1.2 Materials and Methods ................................................ 5

1.2.1 Genetic Material .................................................. 5

1.2.2 Experimental Design ............................................ 7

1.2.3 Data .................................................................. 8

1.2.4 Statistical Analysis .............................................. 10

1.2.5 Response surface of families ................................... 12

1.2.6 Relative response model for seed source families .......... 13

1.3 Regularization methods ................................................. 14

1.4 Results and Discussions ............................................... 17

1.4.1 Relative response of families .................................. 17

1.4.2 Response surface of families .................................. 18

1.4.3 Prediction model and cross validation ....................... 19

1.4.4 Climate change scenario ....................................... 21

1.5 Conclusions .............................................................. 24

1.6 Tables ................................................................... 26

1.7 Figures ................................................................... 29

Chapter 2  Optimal Seed Deployment under Climate Change using Spatial Models: Application to Loblolly Pine in the southeastern US ......... 37

2.1 Introduction .............................................................. 37

2.2 Description of the data ............................................... 41

2.2.1 Genetic Material .................................................. 41

2.2.2 Experimental Design ............................................ 43

2.2.3 Standardized Height ............................................ 43

2.2.4 Climate Variables ............................................... 44

2.3 Model formulation ...................................................... 45

2.3.1 Spatial seed source transport model ....................... 45

2.3.2 Spatial and Climate Predictions .............................. 49

2.4 Bayesian Formulation ................................................. 49

2.4.1 Bayesian Computation .......................................... 51

2.4.2 Missing Data ..................................................... 52

2.4.3 Prediction .......................................................... 54
Chapter 3  Prediction of Genetic Merit in a Clonal Population of Loblolly Pine  
3.1 Introduction .......................................................... 74
3.2 Materials and Methods .................................................. 77
 3.2.1 Genetic material and field design ................................. 77
 3.2.2 Statistical methods ................................................. 78
 3.2.3 Individual trial analysis ......................................... 80
 3.2.4 Multi-environment trial analysis ................................ 83
 3.2.5 Genetic correlation structures .................................. 86
 3.2.6 Loadings rotation and interpretation .......................... 88
 3.2.7 Exploration of genotype by environment interactions ...... 89
 3.2.8 Model fitting and comparison ................................... 90
 3.2.9 Clone mean repeatability ........................................ 92
 3.2.10 Rankings from clonal and seedling tests ...................... 92
3.3 Results and Discussions ................................................ 93
 3.3.1 Individual trial Analysis ....................................... 93
 3.3.2 Multi-environmental trial analysis ............................. 95
 3.3.3 Genotype by environment interactions ........................ 97
 3.3.4 Models comparison ............................................... 100
 3.3.5 Ranking within families ......................................... 102
 3.3.6 Model Fitting ..................................................... 103
 3.3.7 Repeatability of clone means .................................. 104
 3.3.8 Clonal versus seedling tests ................................... 105
3.4 Conclusions .............................................................. 105
3.5 Tables ................................................................. 106
3.6 Figures ................................................................. 116

REFERENCES ............................................................... 124

APPENDIX ................................................................. 137
  Appendix A  Appendix .................................................. 138
    A.1 Computational details ............................................ 138
      A.1.1 Data File ..................................................... 138
A.1.2 Fitting linear mixed models using ASReml-R . . . . . . . . . . . . 139
A.2 Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 146
A.3 Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 150
LIST OF TABLES

Table 1.1 Climatic and geographical variables used to develop the statistical model. Precipitation is measured in millimeters (mm) and temperature in Celsius degrees (°C).

Table 1.2 Regression model coefficients and standard errors for 8-year height of loblolly pine using ordinary least squares (OLS), Ridge, and LASSO regression models. The nine variables were chosen out of the initial set of 34 climatic variables for test site, family, and geographical variables for family. The models explained about 22% of the total variation in the 8-year height.

Table 1.3 Estimated mean and standard deviation of the mean squared errors (MSE) obtained from random splits of the data (training and test sets) for each regression method using cross-validation. The three regression methods gave essentially identical results.

Table 2.1 Posterior summaries for the parameters of CLIM model. The mean coefficients (β) are represented by capital letters for their corresponding covariates, and the covariance parameters (θ) are denoted with greek letters.

Table 3.1 Test site identification number, location, number of rows and columns, total number of observations, average height in meters and standard deviation for each trial. The sites were established in 2009 and 2010 (site 7) and the measurements were taken in 2013.

Table 3.2 Variance-covariance structures used in the linear mixed models for a) the residuals (R), and b) for the genetic effects (G). The mathematical form is expressed in matrix notation.

Table 3.3 Parameter estimates followed by standard errors in parenthesis from individual site analysis using tree height (m) as response variable. At each site, a pedigree based linear mixed model with spatially correlated residuals and independent measurement error was fit. $\sigma^2_g$ is the variance for the clone genetics effect; $\sigma^2$ represents the measurement error variance (nugget effect); $\tau^2$ is the spatial variance parameter (partial sill). Parameters $\rho_r$ and $\rho_c$ are the associated autocorrelation coefficients for the rows and columns, respectively. The structure of the spatial correlation matrix corresponds to a separable autoregressive process of first order in the row and column directions.
Table 3.4 Model-fit statistics. The models are named based on the structure of the genetic by environment ($G_g$) and residual ($R$) variance-covariance matrices, e.g. IID-BD for $G_g$=IID and $R$=BD. $N$ represents the number of variance components in the model to be estimated. $\log(L)$ is the logarithm of the restricted likelihood function evaluated at the parameter estimates. AIC and BIC are the Akaike and Bayesian information criteria, respectively. SEP and SED are the standard error of prediction and difference, respectively.

Table 3.5 Rotated loadings ($\hat{\lambda}_1^*$ and $\hat{\lambda}_2^*$) using the principal component representation and percentage of total variance accounted by the factors ($\hat{\omega}_1^2$, $\hat{\omega}_1^2 + \hat{\omega}_2^2$) at each environment for FA2-AR1 model.

Table 3.6 Genetic correlations between pairs of environments from FA2-AR1 model. The high correlations among sites 2 through 7 suggest a lack of genotype by environment interactions effects ($G \times E$) among these sites. Conversely, the relatively low correlation between site 1 and the others indicate the presence of $G \times E$.

Table 3.7 Rotated loadings ($\hat{\lambda}_1^*$ and $\hat{\lambda}_2^*$) using the principal component representation and percentage of total variance accounted by the factors ($\hat{\omega}_1^2$, $\hat{\omega}_1^2 + \hat{\omega}_2^2$) at each environment for FA2-BD model.

Table 3.8 Genetic correlations between pairs of environments from FA2-BD model. All the correlation pairs are greater than 0.80 suggesting the absence of important genotype by environment interactions.

Table 3.9 Pearson’s correlations (below diagonal) and rank correlations of clone breeding values (above diagonal) between the models.

Table 3.10 Pearson’s, Spearman’s rank, and within family rank correlation between the genetic merit of the clones at each site versus the genetic merit based on the multi-environmental trials analysis using the best model (FA1-AR1) and height as the response variable.

Table A.1 Overall summary statistics for the Atlantic Coastal Elite population. Sample mean, number of observations, minimum, maximum, and standard deviation of the measured traits at age 4 years.

Table A.2 Fixed effects from individual site analysis using height as response variable. At each site a linear row and column effect in addition to an intercept term were fit. The table shows the denominator degrees of freedom, the value for the F statistics and the associated p-values.
Table A.3  Parameter estimates followed by standard errors in parenthesis from individual site analysis using tree volume (dm$^3$) as response variable. At each site, a pedigree based linear mixed model with spatially correlated residuals and independent measurement error was fit. $\sigma^2_g$ is the variance for the clone genetics effect; $\sigma^2$ represents the measurement error variance (nugget effect); $\tau^2$ is the spatial variance parameter (partial sill). Parameters $\rho_r$ and $\rho_c$ are the associated autocorrelation coefficients for the rows and columns, respectively. The structure of the spatial correlation matrix corresponds to a separable autoregressive process of first order in the row and column directions.

Table A.4  Average within family rank correlations (range in parenthesis) of clone breeding values between all model pairs using height as the response variable.
LIST OF FIGURES

Figure 1.1 Geographic regions and county of origin of the selected mother trees. The seven regions were drawn based on the observed variation within the species as well as easily determined political boundaries in the southeastern United States. .................................................... 29

Figure 1.2 Test site locations measured for this study. These 16 test locations cover a substantial area of the natural range of loblolly pine (shaded area) in the east of the Mississippi river. .................................................... 30

Figure 1.3 Mean height deviation of seven seed sources (average of families) from site mean. Each curve can be interpreted as the relative performance of families from a particular geographic region (seed source) across test sites. The test sites are ordered in a southwest to northeast direction. Southern sources from SC, GF, and LG regions were not planted in VA1 site due to their expected poor cold tolerance (adaptability). .......................... 31

Figure 1.4 Observed heights (Height_{tf}) of pine families (f) at various sites (t) as function of family (TMIN_{f}) and site (TMIN_{t}) minimum temperature, and estimated response surface using ordinary least squares. The black dots represent the observed 8-year heights of loblolly pine across the southeastern United States. Planting families from warm regions (TMIN_{f} > 3) at colder sites (TMIN_{t} < −3) greatly reduces their mean height as seen by the red colors in the corner of the curve (left side of plot). Conversely, planting families from cold regions (TMIN_{f} < −3) at warmer sites (TMIN_{t} > 3) also greatly reduces their height as seen by the red colors in the corner of the curve (right side of plot). .......................... 32

Figure 1.5 Predicted 8-year height of loblolly pine local families across the southeastern U.S. using ordinary least squares (OLS) regression model for a hypothetical future scenario with 5% decrease in precipitation and an increase of 0.5 °C in maximum and minimum temperatures, relative to historical average values (reference period 1970-2010). The dark blue regions indicate where the local families are predicted to perform well. Predictions for the Gulf Coast of Florida, and north of Virginia are considered to lie outside the observed range defined by the spatial distribution of test sites. ............................................................ 33
Figure 1.6  Difference between the predicted 8-year height under the hypothetical climate change scenario and the predicted height using current climate conditions for local pine families. The chart illustrates the effect of the simulated climate change scenario on the growth of local families across the southeastern United States, relative to current climate conditions. Blue and light blue areas on the map indicate a positive difference or increase in mean height, whereas yellow and dark orange areas denote a negative difference or decrease in height. The Georgia-Florida coastal plain and part of South Carolina coastal plain regions exhibit the highest differences (gain in height) in growth relative to the current climate. Predictions for the Gulf Coast of Florida and Alabama, and north of Virginia are considered to lie outside the observed range defined by the spatial distribution of test sites.

Figure 1.7  Predicted 8-year height growth of coastal South Carolina loblolly pine families planted across the southeastern United States for a hypothetical climate change scenario (5% decrease in precipitation and 0.5°C increase in maximum and minimum temperatures), relative to historical average values (reference period 1970-2010). Coastal South Carolina families are expected to perform well across the Southeast, but highest growth (dark blue) is expected to be in southern Georgia, under the modeled climate change scenario. Predictions for the gulf coast of Florida, and north of Virginia are considered to lie outside the observed range defined by the spatial distribution of test sites.

Figure 1.8  Ordinary least squares (OLS) regression model predictions for 8-year height deviations from site mean of families from Virginia seed source for a hypothetical climate change scenario (5% decrease in precipitation and 0.5°C increase in minimum and maximum temperatures), relative to historical average values (reference period 1970-2010). The contour plot indicates that this seed source will only be superior in the most northern region (dark blue area) and perform poorly relative to other sources in the most southern region (red area), under the modeled climate change scenario.

Figure 2.1  Geographic regions of loblolly pine grouped by counties, location from where the seed sources were collected (black points), and test site locations (red triangles). The geographic regions are Virginia (VA), North Carolina Coastal Plain (NC), South Carolina Coastal Plain (SC), Georgia-Florida Coastal Plain (GF), Lower Gulf Coastal Plain (LG), Upper Gulf Coastal Plain (UG), and Piedmont (PD).
Figure 2.2  Plot to illustrate the underlying concept of the proposed model. The lines represent different climate scenarios, and the shading of dots denotes the expected response. The expected response of the underlying process can be characterized for future climate scenarios from current climate.

Figure 2.3  Boxplot of mean square errors (logarithmic scale) from cross-validation analyses for the models under consideration: a) leave-site-out, b) leave-source-out, and c) random split. The thick horizontal line inside each box represents the median and the black dot the mean. The bottom and top horizontal lines that define the box represent the first (25%) and third quantile (75%), respectively. The bottom and top whiskers extend to the minimum and maximum data value, respectively. The IID model is a non-spatial Bayesian multiple regression, GEO is a spatial model that uses solely geographical variables to define correlations; CLIM is a spatial model that uses climate variables only to define correlations, and GEO/CLIM model uses both geographical and climate variables to define correlations among observed values.

Figure 2.4  Trace plots, posterior histogram with approximated density superimposed, and auto-correlation function for the MCMC chains of $\tau^2$ (spatial variance), $\varphi_{T_i}$ (temp-site scale), and $\kappa_s$ (seed source smoothness) for CLIM model. There are 120,000 samples with a burn-in period of 20,000.

Figure 2.5  Predicted performance (standardized height) for seed sources from a) Nassau, FL (black point) and b) Appomattox, VA assuming that future climate conditions would be the same as the average of the last 20 years. The first colormap indicates that seeds from Nassau county will have a superior performance in the southern region. In contrast, the predicted standardized heights of seeds from Appomattox suggest a relative superior performance in the northern and colder regions of the Southeast. The standard deviation is plotted in panel c) and is the same for both predictions.

Figure 2.6  Predicted performance in Wayne County, NC as a function of seed source for two climate scenarios based on historical values: a) future climate conditions would be the same as the average of the last 20 years b) 1 $^\circ$C increase in minimum and maximum temperatures, and a 10% decrease in precipitations. The blue color indicates the location of top performing seeds. The arrows show the direction of the optimal seed source movement based on the top 1% standardized height.
Figure 2.7 The arrows show the direction of the optimal seed source movement to Southern Virginia (top) and Central South Carolina (bottom) based on the top 1% locations (dots) assuming three scenarios: 1) future climate conditions would be the same as the average of the last 20 years, 2) a 1°C increase in minimum and maximum temperatures, and a 10% decrease in precipitations relative to historical values, and 3) a 2°C increase in minimum and maximum temperatures, and a 20% decrease in precipitations relative to historical values. The colors correspond to the state of origin. 73

Figure 3.1 Test site locations across the southeastern United States. 116

Figure 3.2 Empirical semi-variogram from spatial analysis of test sites using height (m) as response variable. At each site, a pedigree based linear mixed model with spatially correlated residuals and independent measurement error was fit. The plots quantify the spatial correlation at each site. Most of the sites exhibit a smooth surface that increases in the row and column directions reaching a plateau of little variation, which indicates the absence of both global and extraneous trend. 117

Figure 3.3 Rotated and normalized loadings for the genetic effects from FA2-AR1 model using height as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments. 118

Figure 3.4 Rotated and normalized loadings for the genetic effects from FA2-BD model using height as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments. 119

Figure 3.5 Spaghetti plots of the empirical BLUP values for the clones as a function of environment for FA1-AR1 (top) and FA1-BD (bottom) models. The pattern of the spatial FA1-AR1 model indicates that there are important rank changes from site 1 to site 2, as evidenced by the number of non-parallel lines. On the other hand, the pattern of the FA1-BD model exhibits small rank changes throughout the environments since the lines are essentially parallel. 120

Figure 3.6 Scatter-plot of predicted clone genetic merit between factor analytic spatial (FA1-AR1) and standard non-spatial (IID-IID) model. The dashed lines indicate the 90% quantile for each model. Both rank and Pearson’s correlations between the predictions are about 0.90. 121

Figure 3.7 Histogram of within family rank correlations based on the clone genetic merit between factor analytic spatial (FA1-AR1) and standard non-spatial (IID-IID) models. The within family rank correlations ranged from 0.69 to 0.93 with an average of approximately 0.85. 122
Figure 3.8  Heat-map with dendrogram to illustrate clusters of models based on the average within family rank correlations of clone breeding values between model pairs. Similar performing model pairs are indicated with values close to one (red color).

Figure A.1  Rotated and normalized loadings for the genetic effects from FA2-AR1 model using volume as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments.

Figure A.2  Rotated and normalized loadings for the genetic effects from FA2-BD model using volume as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments.
Chapter 1

Modeling Climate Change Effects on Loblolly Pine

1.1 Introduction

Forest trees face challenges in adapting to changing environmental conditions because, unlike many animal species, trees may not be able to migrate or adapt quickly enough if climatic conditions rapidly become adverse. Forest trees typically have long life cycles, and a single individual can experience significant temporal climatic variation. In general, plants must respond to changes in environmental conditions either by adapting to new conditions, or by migration via pollen, seeds, or propagules to new locations where conditions are more favorable. Individual trees typically survive over long periods, and phenotypic plasticity may play an important role in their ability to survive. The challenge is to identify the genetic adaptations within species that will best allow trees to survive and adapt to future climates.

Planting seedlings predicted to be adapted to future climate conditions has been proposed as a forest management strategy to mitigate the potential negative impacts of climate change
Current climate change is altering environmental conditions rapidly relative to historical trends, so that, locally-adapted tree species may well experience extreme climatic conditions to which they are not well adapted (Skelly et al., 2007; Parry, 2007). Strategies for development of adaptive seed sources via genetic improvement are needed to mitigate the risks of adaptation of future forest tree plantations. New guidelines for deployment of seed sources are needed to ensure that plantations are adapted to future climates (Aitken et al., 2008).

In order to predict the impact of future climate change on tree phenotypes, the environmental and genetic effects need to be considered and quantified. A quantitative model could be used to accurately describe the intricate relationship between trees and environmental conditions and make predictions about the potential effects of climate change on growth and survival.

The establishment and analysis of provenance tests for investigating the genetic variation in forest tree species has a long tradition in forestry (Langlet, 1971). Provenance tests are common garden experiments, and their primary objectives are typically to identify fast growing and adaptable seed sources for reforestation. Provenance-progeny tests, if designed and managed properly, can also provide valuable information for assessing the response of tree populations to environmental change.

The effect of climatic variables on seed sources (genetics) has been previously studied to develop population response and transfer functions using provenance test data to assess future climate change scenarios (Schmidtling, 1994; Matyas, 1994). A population response function describes the performance of a single population across different planting locations as a function of the climate at those locations. The response function accounts for the environmental effects of planting location climate on the performance of a specific population. This approach has been used by several researchers to predict the impact of climate change on a particular population (Kapeller et al., 2012; O’Neill et al., 2008; Wang et al., 2006; Rehfeldt et al., 1999).
In general, the objective of studies involving population response functions is to determine the optimal climatic conditions for a given trait, that is when the trait of interest is expressed in its most favorable shape (Kapeller et al., 2013). The main drawback of the population response function approach is that it is population specific. Results obtained from a response function are only valid for the specific population for which it was developed; hence extrapolation of the results to other populations is not reliable.

In contrast to population response functions, a transfer function gives the performance of multiple populations as a function of what is called climatic transfer distance, which is the difference between the climate of the provenance and the climate of the planting site (Matyas, 1994; Rehfeldt et al., 1999; Andalo et al., 2005). The term provenance is used to describe the original location and climate of individuals. For instance, we can assess the transfer effect using the yearly average minimum temperature (TMIN) of the provenance and of the planting site (Schmidtling, 1994, 2001). It is generally hypothesized that tree populations are best adapted to their local climate conditions. Therefore, the best growth performance is expected at small climate transfer distances, whereas at larger climatic distances, trees are expected to perform poorly due to lack of adaptation (Kapeller et al., 2013). It is reasonable to derive population transfer functions for individual sites if data from populations from diverse climatic origins are available. Transfer functions account for the phenotypic effect of moving or transferring populations to different locations. However, the assumption that transfer effects are the same across different environments is often not valid. For example, the climatic transfer effect of $-4\,^\circ C$ of yearly average minimum temperature (TMIN) on phenotypes could be radically different, if a population is moved from $5\,^\circ C$ to $1\,^\circ C$ compared to a transfer of another population from $1\,^\circ C$ to $-3\,^\circ C$. Although the climatic transfer distance is the same in both cases, the effects on phenotype could be drastically different. Individual transfer functions have been developed using provenance climate, instead of provenance transfer distance as the independent variable (Wang
et al., 2006). This way, transfer functions can be used to identify optimal climatic conditions relative to each site. The main limitation of individual transfer functions is that they are test site specific; that is, an individual transfer function is valid only for the planting site for which it was developed.

Wang et al. (2010) proposed a model called Universal Response Function (URF) in order to overcome the critical limitations of the population response function and the transfer function methods which are population specific and site specific, respectively. The URF model integrates both environmental effects of planting sites on phenotype (accounted for by population response functions) and genetic effects of climate on phenotype (accounted for by transfer functions). The approach uses all available information from provenances and test sites to estimate a joined response for a given environmental condition. Hence, the model allows predictions of the response of any population growing in any climate and can be used to construct deployment guidelines for reforestation. The URF was originally developed from the Illingworth lodgepole pine (Pinus contorta Douglas ex Loudon) provenance test established in British Columbia, Canada. The test is considered as one of the most comprehensive in the world comprising 140 populations and 62 test sites widely scattered across a large climate range.

It is very likely that climate change will have significant impacts on pine plantations across the southeastern United States. A report from the U.S. Global Change Research program (Karl et al., 2009) indicates that the southeastern United States will experience continued increases in the rate of warming through the end of the century, with a rise in average temperature of 2.5 to 5 °C by the 2080s, a magnitude of temperature increase that is predicted to decrease loblolly pine productivity by at least 10% (Schmidtling, 1994). Precipitation predictions for the region are less certain, but generally indicate that summertime precipitation will decline by 10% to 30% (Solomon, 2007). The expected precipitation decrease combined with elevated temper-
atures will likely increase vapor pressure deficits and soil water deficits. The physiological response to soil water and vapor pressure deficits shows genetic variation (Gonzalez-Benecke and Martin, 2010), suggesting negative impacts may be mitigated by genetic improvement of deployment populations.

In this study we investigated the climate response of loblolly pine families using the URF approach. Loblolly pine is the major commercial species in the southeastern United States (McKeand et al., 2003). It is subjected to an intensive breeding program and is planted widely for both pulpwood and solid wood products. Therefore, continued productivity of this species is vital to the region. Adverse effects of climatic variables on growth and adaptation could have serious economic implications.

The main objective of this work was to develop a statistical model to predict the responses of loblolly pine populations to climate change using the URF approach. The model takes into account both genetic and environmental effects through climatic variables. To demonstrate the use and extent of the proposed model, application examples are presented concerning the predicted performance of different loblolly pine families under a hypothetical future climatic scenario across the southeastern United States.

1.2 Materials and Methods

1.2.1 Genetic Material

The North Carolina State University Cooperative Tree Improvement Program (Cooperative) conducted an intensive selection effort in unimproved loblolly pine plantations between 1975 and 1981 to increase the size of breeding populations. The criteria for selecting trees for the breeding program were good growth, straight stems, and absence of fusiform rust galls caused
by the fungus *Cronartium quercuum f. sp. fusiforme*. In 1994, the Cooperative established the Plantation Selection Seed Source Study (PSSSS), a large replicated series of progeny tests in the southeastern United States, using plantation selections as parents. The experiment was designed to determine the patterns of genetic variation present in plantation selections from different geographic regions and to better understand genotype by environment interactions (Chamblee, 2011).

Since the selections used in the PSSSS were from plantations, and not from natural stands, the exact origin of each selection is not known. Movement of seeds up to 500 km from the original location was not uncommon in early plantations of loblolly pine in the Southeast, but based on the knowledge of foresters who established these plantations, we are confident that most of the sources of the seed are local or within a 100 to 200 km distance from the origin. Throughout this work, we will refer to seed sources instead of the provenances of individuals to denote location and climate, to more accurately describe the degree of knowledge of the true provenance of origin for each selected tree.

The selected trees used in the study were randomly chosen from a pool of plantation selections in seven geographic regions within the continuous natural range of loblolly pine in the southeastern United States. Figure 1.1 shows the location of the pine plantations from where the female parents were selected (mother tree origin). In the same figure, seven geographic regions of the natural range of loblolly pine are sketched: Virginia (VA), North Carolina Coastal Plain (NC), South Carolina Coastal Plain (SC), Georgia-Florida Coastal Plain (GF), Lower Gulf Coastal Plain (LG), Upper Gulf Coastal Plain (UG), and Piedmont (PD). These seven regions were somewhat arbitrarily defined by the Cooperative based on the observed geographic variation within the species as well as easily distinguishable state and physiographic boundaries. Geographic differences within loblolly pine often are not easy to define, and boundaries are not clear-cut. Thus, the determination of what constitutes a geographic region is often arbi-
trary (Zobel and Talbert, 1984). In each of seven regions, 20 selections were randomly chosen as females. These selections were then mated with a pollen mix of 40 pollen parents from the same region. As a result, a total of 140 maternal families were employed for the study, and a total of 280 male parents were used in the pollen mixes. The female and pollen parents were unrelated in order to avoid inbreeding and increase the genetic sample size.

1.2.2 Experimental Design

Field tests were planted from 1994 to 1997 using a randomized complete block design. The experiment was replicated at 25 sites (locations). Each test site had 24 blocks, with a single tree from each family and four seedlings from the unimproved local checklot randomly distributed throughout each block.

The test sites of the PSSS were managed in the same way following standard progeny test protocols that included site preparation before planting and weed and vegetation control during the experiment. No fertilization treatment was applied at any site. Some test sites were abandoned because of low survival, and in others the corresponding measurements were never taken. Only 16 test sites were available for this study. Nonetheless, the remaining planting sites cover a substantial area of the eastern part of the natural range of loblolly pine (Figure 1.2). In four northern test sites, southern sources from South Carolina Coastal, Georgia and Florida Coastal, and the Lower Gulf were not used due to space limitations and the expected poor adaptability of the southern sources in the northern sites (Schmidtling, 2001; Chamblee, 2011).
1.2.3 Data

Growth measurements such as height, diameter at breast height, stem straightness (1 to 6 categorical scale), fusiform rust disease incidence (presence or absence of galls), forking (presence or absence of forked stem), and survival were taken at age 8 years. In total, over 43,000 trees were measured. We used the height growth of trees at age 8 years as a response variable for the proposed model. Height can be considered an adaptive trait because it is heritable and likely to be subject to natural selection more than many other traits (McKeand et al., 2008). Taller trees compete better for light; hence natural selection for adaptation is likely to have acted on genes that affect height growth.

The climatic variables were estimated using the PRISM (Parameter-elevation Regressions on Independent Slopes Model) climate mapping system (Daly et al., 1994). PRISM is a climate analysis system that uses point data, a digital elevation model, and other spatial data sets to generate gridded estimates of annual, monthly and event-based climatic parameters. The PRISM system is basically an advanced spatial interpolation that is used to convert point-based historical climate into a gridded raster surface for climatic variables. It was created mostly to help estimate historical climate variables in regions of complex topography, and it has been used extensively to map precipitation and minimum and maximum temperatures over the United States. The time resolution of the PRISM mapping system was limited to monthly average estimates. The downside of the time constraint is that extreme or unusual climatic events that could have a big effect on height growth would be overlooked.

Yearly average total precipitation (PPT), yearly average minimum temperature (TMIN), and yearly average maximum temperature (TMAX) of both test site and pine families were used as explanatory variables to develop prediction models (Table 1.1). Climatic variables are described by their subscript. For instance, the subscript $t$ denotes test site (e.g. $PPT_t$), subscript
denotes the origin of the mother tree or female parent (e.g., PPTₘ), subscript \( p \) denotes pollen parent (e.g., PPTₚ), and subscript \( f \) denotes family. Each family climatic variable (e.g. PPTₖ) was constructed by taking the average between the climatic variable of the county of origin of the mother tree (e.g., PPTₘ) and the average of the same climatic variable of the 40 pollen parents (e.g., PPTₚ) from the same region. In this way, the location of the family is taken into account through the location of the female, and the climatic variables of both parents are considered and equally weighted.

In order to estimate the climatic variables for the mother trees and pollen parents, the twenty-year time reference period from 1970 to 1990 was used for the calculations. The centroids of the counties from where the selections were made were used as the reference location for female and male parents, because the exact origins of the trees in those plantations are unknown.

Climatic variables of 16 test sites were obtained by averaging the monthly observations of counties where test sites were established for the period from 1994 to 2006. The distinction between climate variables associated to families or test sites is crucial to develop a model for the response of pine families to different climatic variables. Specifically, these two components are intended to integrate both genetic factors, by considering the genetics of both female and pollen parents of the trees growing in the field tests, and the effects of climate at the test site location interacting with the genetic variation among the progeny of the selected parents.

The statistical model was developed using geographical variables for the families as predictors defined in terms of the latitude (LATₖ) and longitude (LONGₖ) coordinates that indicate the geographical location of the pine families. The hypothesis underlying the development of the model is that variation in phenotypes occurs geographically, and therefore, elements of this variation can be predicted from geographic descriptors.
1.2.4 Statistical Analysis

Multiple linear regression models were used to explain variation in height of trees. The climatic variables for the test sites (environmental effects), and climatic and geographic variables of pine families (genetic effects) were used as explanatory variables (Table 1.1). The general form of the model is given by:

\[ Y_{tf} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{1t}^2 + \beta_3 X_{2f} + \beta_4 X_{2f}^2 + \beta_5 X_{1t} X_{2f} + \beta_6 X_{3f} + \epsilon_{tf} \]  

where \( Y_{tf} \) is the height observation of family \( f \) at site \( t \); \( \beta_k \) for \( k = 1, \ldots, 6 \) are the regression coefficients associated to the explanatory variables of the model; \( X_{1t} \) is the collection of climate variables for test site \( t \); \( X_{2f} \) is the collection of climate variables for the family \( f \); \( X_{3f} \) is the collection of geographic variables for the family \( f \); and \( \epsilon_{tf} \iid \sim N(0, \sigma^2) \) is a random error term that follows a normal distribution with zero mean and constant variance \( \sigma^2 \). The model allows for a quadratic relationship between each predictor and the response with \( X_{1t}^2 \) and \( X_{2f}^2 \) referring to the set of squared climate variables for the test site and family, respectively, and for interactions between climate variables for the test site and family. Geographic variables for test sites are not included in the model because the presented model is aimed to identify optimal sites based on climatic conditions rather than geographical locations.

The model was constructed using the family and test site climatic variables listed in Table 1.1 in addition to seven geographical variables for the families (LAT\(_f\), LAT\(_f^2\), LONG\(_f\), LONG\(_f^2\), LAT\(_f \times\) LONG\(_f\), LONG\(_f^2\)/LAT\(_f\), and LONG\(_f^2\)/LAT\(_f\)). The climatic variables included both linear and quadratic terms, which gave 12 predictors, and 15 interaction terms. Hence, the number of potential independent variables was 34.

In all regressions, the explanatory variables were centered and scaled to have unit variance by subtracting the mean and dividing by the standard deviation. In addition, the response
variable, (8-year height) was centered. Centering and scaling does not affect the statistical inference in the regression model but makes the model more robust in terms of numerical stability, and forces the units of the regression coefficients to be the same.

Because of the large number of potential predictors (34), we pursued variable selection to identify a subset of predictors to be used in the prediction model. Reducing the number of predictors makes the model more interpretable, and can improve out-of-sample prediction by preventing over-fitting. In order to find a parsimonious model based on a subset of the 34 potential predictors, different approaches for model selection and parameter estimation were explored, namely the stepwise method (Hocking, 1976) combined with ordinary least squares (Rao, 2009), ridge regression (Hoerl and Kennard, 1970), and LASSO regression (Tibshirani, 1996). These methods are described in detail in the section Regularization Methods.

The model development and validation followed several steps: First, the stepwise selection method in the GLMSELECT procedure of SAS software Inc (2011) was used with the Schwarz Bayesian Information Criterion (Schwarz, 1978) for variable selection. This option balances the trade-off between maximizing the likelihood by adding predictors and penalizing the number of parameters in the model to avoid over-fitting. This first step is straightforward to implement and is of key importance to narrow down the total number of variables.

The second step involved the application of regression techniques for estimating the parameters of the model. In this stage, three approaches were explored; ordinary least squares, and two shrinkage methods, LASSO and ridge regression.

In the final step, a cross-validation of the explored models was conducted (Shao, 1993). Cross-validation measures how accurately the predictive model will perform in practice; that is, when predictions based on new or unseen data are carried out. Cross validation was implemented in R software (R Core Team, 2014) through an algorithm that randomly splits the data into training and validation sets. The statistical model was fit using the training data set,
and its predictive accuracy was assessed with the validation data set through the mean squared error (MSE). The cross-validation was repeated using 50,000 different random training and validation data sets. For all cases, the size of the training and validation data set was set to be half of the sample size (N/2=21,540).

The model offers great opportunity for further analysis. For instance, to predict the performance of any family based on some specific trait at a given planting site, the model becomes a transfer function since variables associated with the planting site are constant. In the same way, to predict the performance of a given seed source (families selected from a geographic region) at any planting site, the model reduces to a population response function since the variables associated to the family are constant (weighted average of the selected families). It is also possible to model the performance of local families given specific climatic conditions. In this case, the planting site climate variables are set equal to the family climate variables. In order to predict growth of any given family under a future climate scenario, historical climatic values associated to the test site are replaced by site future climate values.

1.2.5 Response surface of families

To graphically illustrate a potential application of the presented approach, a simple example was created using just test site (TMIN\textsubscript{t}) and family (TMIN\textsubscript{f}) yearly average minimum temperatures as predictors, and the observed 8-year height of family \( f \) at site \( t \) (HT\textsubscript{tf}) as the response variable. In this particular case, the statistical model takes the following form:

\[
HT_{tf} = \beta_0 + \beta_1 TMIN_t + \beta_2 TMIN_f + \beta_3 TMIN_t \times TMIN_f + \beta_4 TMIN_t^2 + \beta_5 TMIN_f^2 + e_{tf} \tag{1.2}
\]

In order to obtain a response surface for the 8-year height as function of family and site minimum temperature, the coefficients \( \beta_i \), for \( i = 0, \ldots, 5 \) need to be estimated. The coefficient
associated with $TMIN_t \times TMIN_f$ is the response of a particular family to different climates and can be interpreted as the family’s stability.

### 1.2.6 Relative response model for seed source families

In order to graphically illustrate the differences in the genetic expression of seed sources at each test site, the site mean was subtracted from the mean height of the pine families from each geographical region (Figure 1.1). These deviations were used to visualize the response trends of seed sources across the landscape (Figure 1.3). At a given test site, the deviation of a specific seed source from site mean can be interpreted as a measure of the performance of that seed source relative to the others present at the same site.

A prediction model for the relative performance of individual families from the seven geographic regions was developed using the same set of climatic and geographical previously described (section Statistical Analysis). As before, the climatic variables included linear, quadratic, and interaction terms, so variable selection was carried out to identify a subset of predictors. However, for this model, the response variable was the deviation of the observation from its corresponding site mean. This model was developed to visualize the response pattern of the pine families from each geographic region across the southeastern United States. For variable selection, we used the stepwise selection method in the GLMSELECT procedure of SAS software (Inc, 2011) under the cross-validation predicted residual sum of squares criterion.
1.3 Regularization methods

Multiple linear regressions are often used to investigate the relationship between the response variable $y$ and a set of $p$ predictors $x_1, \ldots, x_p$. Regression models are also used for predicting future responses. For example, given the predictors, the response is estimated by:

$$
\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \cdots + x_p\hat{\beta}_p
$$

(1.3)

where $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ is the vector of estimated coefficients that depend on the fitting procedure. For instance, Ordinary Least Squares (OLS) is based on finding the $p \times 1$ vector of regression coefficients $\hat{\beta}$ that minimizes the residual sum of squares. Hence, the OLS estimates can be written as follows:

$$
\hat{\beta}_{OLS} = \arg\min_{\beta} RSS = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 = \arg\min_{\beta} |y - X\beta|^2
$$

(1.4)

where $RSS$ is the residual sum of squares, $y_i$ the $i$th response, $x_i = (x_{i1}, \ldots, x_{ip})^T$ is the $i$th vector of predictors, $y$ is the $n \times 1$ response vector, and $X$ is the $n \times p$ design matrix of predictors. The solution to this problem is given by:

$$
\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty
$$

(1.5)

There are two criteria that are generally used to evaluate the quality of a model: prediction accuracy and model complexity. Prediction accuracy refers to the capacity of the model to make accurate predictions on future or unseen data; while complexity refers to the interpretation of the model. Simple and parsimonious models are preferred over more complicated models because they shed light on the relationship between the response variable and the predictors.
OLS is known to perform poorly in both prediction accuracy and model complexity when the number of predictors \((p)\) is large relative to the number of observations \((n)\), and also in the presence of multi-collinearity (high correlation among predictors). For these cases, OLS estimates are unstable. As a result, several regularization methods using penalization techniques have been proposed to improve OLS (Zou and Hastie, 2005). For instance, ridge regression (Hoerl and Kennard, 1970) is a shrinkage regression method that was originally intended to deal with multi-collinearity. The idea behind the method is to minimize the residual sum of squares subject to the constraint \(\sum_{i=1}^{p} \beta_i^2 \leq \alpha_R\), where \(\alpha_R\) is the ridge tuning parameter. Ridge regression estimates are obtained by minimizing the penalized residual sum of squares (Hoerl and Kennard, 1970):

\[
\hat{\beta}_R = \arg\min_{\beta} \left\{ |y - X\beta|^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \right\}
\]  

(1.6)

where \(\lambda\) is called the shrinkage parameter and needs to be tuned. As with OLS, the ridge regression estimate has closed form:

\[
\hat{\beta}_R = (X^T X + \lambda I_p)^{-1} X^T y
\]

(1.7)

The inclusion of \(\lambda\) makes the solution non-singular even if the Gramian matrix \(X^T X\) does not have an inverse. The shrinkage parameter \(\lambda\) controls the magnitude of the coefficients; so that the regression parameters from ridge regression converge to the parameters estimated by OLS \((\hat{\beta}_R \rightarrow \hat{\beta}_{OLS})\) as \(\lambda\) gets closer to zero \((\lambda \rightarrow 0)\), and the regression parameters tend to zero \((\hat{\beta}_R \rightarrow 0)\) as the shrinkage parameter gets infinitely large \((\lambda \rightarrow \infty)\). Ridge regression reduces the variability of the regression coefficient estimates by shrinking them towards zero. This results in an increase of prediction accuracy at the cost of a small bias. Although the
coefficients are shrunken towards zero in ridge regression, they never become exactly zero and all predictors are kept in the model. Therefore, ridge regression increases prediction accuracy in some cases but does not reduce the complexity of the model.

The LASSO method (Tibshirani, 1996) is another form of constrained least squares developed to improve both prediction accuracy and model interpretability. The solution minimizes the usual residual sum of squares subject to a constraint on the absolute values of the regression coefficients, \( \sum_{i=1}^{p} |\beta_i| < \alpha_L \), with \( \alpha_L \) the LASSO tuning parameter. The LASSO estimates are defined by:

\[
\hat{\beta}_{\text{LASSO}} = \text{argmin}_{\beta} \left\{ |y - X\beta|^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right\}
\]

Unlike OLS and ridge regression, the estimated coefficients cannot be expressed in a closed form solution. Therefore, numerical algorithms must be used to find a solution. The computation of the LASSO solution is a quadratic programming problem and can be tackled by standard numerical analysis algorithms. The R software package lars (R Core Team, 2014) and GLM-SELECT procedure of SAS software (Inc, 2011) implement LASSO regression. The LASSO reduces the variability of the regression coefficient estimates by shrinking them towards zero, and also reduces the complexity of the model by shrinking some of the coefficients to exactly zero. Therefore, unlike ridge regression, the LASSO also serves as a variable selection tool.

The OLS estimator is unbiased and has minimum variance among all the linear unbiased estimators (BLUE). While the ridge and LASSO estimators are biased, introducing a small bias often leads to a substantial decrease in variance, and hence to a substantial decrease in the mean square error of \( \hat{\beta} \). The effect of penalizing regression coefficients is to shrink estimates towards zero compared to OLS, which provides numerical and statistical stability.

We decided to explore regularization methods for constructing the model in order to avoid over-fitting and to deal with the presence of multi-collinearity among predictors. An over-fit
model will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data (Hawkins, 2004). The attention was centered on ridge regression and LASSO because they are most common type of regularization techniques used in regression analysis and are implemented in statistical packages (Endelman, 2011; de los Campos et al., 2013).

The performance of shrinkage methods depends heavily on the appropriate selection of the shrinkage coefficient \( \lambda \). Typical approaches for estimating parameter \( \lambda \) are the Bayesian information criterion (BIC) (Schwarz, 1978), the Akaike's information criterion (AIC) (Akaike, 1974), the Mallows Cp statistic (Mallows, 1973, 1995), the generalized cross-validation (Brown, 1993), the Hoerl-Kennard-Baldwin estimator (Hoerl et al., 1975), and the cross-validation method (Picard and Cook, 1984). The shrinkage parameter \( \lambda \) for the ridge regression was estimated using the Hoerl-Kennard-Baldwin approach implemented in \texttt{lm.ridge} function from the MASS package in R software (R Core Team, 2014). For the LASSO regression, the Mallows Cp statistic was used to estimate the optimal shrinkage parameter \( \lambda \) using the R package \texttt{lars}.

### 1.4 Results and Discussions

#### 1.4.1 Relative response of families

The mean height deviations of seven seed sources (difference between the average of families for a given geographic region and site mean) are presented in Figure 1.3. The line plots give an insight about the performance of the seed sources across the southeastern United States. Note that sites are ordered from southwest to northeast in the horizontal axis. Trees from the coastal South Carolina region (SC) showed superior performance not only in their native environment but also across all sites (line at the top). On the other hand, trees from Virginia
(VA) exhibited a poor performance relative to the others (line at the bottom); particularly in warmer climates. Families from the Georgia-Florida region (GF) displayed a fair performance across sites with a decreasing trend towards north. Other seed sources had similar average performance. These findings were not surprising, as better growth of coastal sources compared to inland or northern sources has been reported in several studies (Wells and Wakeley, 1966; Wells, 1983; Wells and Lambeth, 1983). One striking observation is that the difference between the relative performance of most northern seed sources (e.g. VA) and southern seed sources (e.g. SC) is large for southern sites, but the disparity gradually decreases at more northern sites and disappears at the NC2 site. This is expected since genetically superior SC seed source is not adapted to colder sites. On the other hand, the VA seed source performs better in its native environment than other seed sources (Chamblee, 2011).

1.4.2 Response surface of families

The coefficients $\beta_i$, for $i = 0, \ldots, 5$ of equation (1.2) that define the response surface for the mean height as function of family and site minimum temperature were estimated using ordinary least squares (OLS). The response surface (Figure 1.4) suggests that families from warm regions (with $T_{MIN_f} > 3^\circ C$) planted at colder sites (with $T_{MIN_t} < -3^\circ C$) tend to perform poorly as seen by the red colors in the corner of the curve (left side). Conversely, planting families selected from cold regions ($T_{MIN_f} < -3^\circ C$) at warmer sites ($T_{MIN_t} > 3^\circ C$) also greatly reduces their height as seen by the red colors in the corner of the curve on the right side of plot. The plot also suggests that families from colder regions ($T_{MIN_f} < 0$) tend to improve their performance when moved to warmer sites. Yet their performance declines after $T_{MIN_t} > 1$ (right side). We observe large variation between families at a given location as shown by the vertical distribution of black circles, particularly at sites with $T_{MIN}$ around 0.
degrees. Using this model as a framework, it is possible to find an optimal climate condition for a given family. This simplified and visual example can be extended to other climatic variables for the test sites, and climatic and geographic variables for the families.

It is important to emphasize that the analyzed data correspond to a short-term trial, and we are only analyzing the 8-year height response. Therefore, although an adaptation effect is likely playing a role in the observed results, we can only speculate about adaptability based on these data. Previous studies have shown the inherent variations in growth rate of loblolly pine seed sources and its association with physiographic and climatic effects (Wells, 1983).

1.4.3 Prediction model and cross validation

The selected statistical model for 8-year height of loblolly pine using OLS is shown in Table 1.2. The selected model included nine covariates chosen out of the initial set of 34. Ridge and LASSO regression coefficient estimates are also reported in the same table. For the LASSO regression, the estimated coefficient for the interaction between the total precipitation and maximum temperature for the test site (PPT_t × TMAX_t) is exactly zero; illustrating the variable selection aspect of LASSO regression. We did not report the standard errors of coefficient estimates from the ridge and LASSO regression because these procedures reduce the variance of estimators by introducing bias. The bias of each estimator could therefore be a major component of its mean squared error, whereas its variance may contribute only a small part. Reporting standard errors of a penalized estimate might thus be misleading, because the inaccuracy caused by the bias is completely ignored. Estimation of standard errors of penalized regression methods is a topic of current active research (Kyung et al., 2010).

The models given in Table 1.2 explained about 22% of the total variation in height growth. The estimated coefficients associated with the climate variables were highly significant. It is
interesting to note that only one climate variable associated to the pine family was selected as predictor of future growth; that is yearly average minimum temperature (TMIN). The other climate variables present in the model are related to the test site. Previous studies have drawn the attention to the importance of minimum temperature as predictor of performance (Schmidtling, 1994; Carter, 1996). Minimum temperature has been used by the United States Department of Agriculture (USDA) to define plant hardiness zones to determine which plants are most likely to thrive at a given location (USDA, 2015). Furthermore, the ability of tree populations to withstand the minimum temperatures of a defined geographic area was used to define the guidelines for seed movement for southern pines (Schmidtling, 2001, 2003).

The cross-validation results are shown in Table 1.3. The narrow range of mean squared error (MSE) values suggests that the models are highly consistent among the sub-samples. The three regression models gave very similar mean squared error (MSE). A formal hypothesis test reveals significant differences between OLS and LASSO, and between ridge and LASSO; but in practical terms, the three models exhibit the same performance. Therefore, despite its limitations we believe OLS is a reasonable choice for estimating the regression parameters, since it is the most popular regression approach and is implemented in most statistical packages.

Despite the simplicity of OLS, the approach is not always satisfactory because the variance of the estimator \( \text{var}(\hat{\beta}_{OLS}) = \sigma^2(X^TX)^{-1} \) is large when the inverse of the Gramian matrix is close to singular, or equivalently, when \( X \) is close to collinear. Therefore, in the presence of multi-collinearity, predictions based on OLS are overall not satisfactory. To achieve better predictions, regularization methods such as ridge and LASSO regression yield biased, yet numerically and statistically stable estimators. LASSO does both variable selection and shrinkage, while ridge regression is only a shrinkage estimator. The presented LASSO model is more parsimonious than the OLS and ridge models (Table 1.2), since it uses only eight predictors while keeping essentially the same predictive power.
Different geographical predictors that are associated with the origin of the pine families were tested (e.g. functions of LAT\(_f\) and LONG\(_f\)), and only the ratio LONG\(^2_f\)/LAT\(_f\) was found to be meaningful (\(\hat{\beta} = -0.06\)) as a predictor for height growth. This covariate expresses the estimated rate of change of the conditional mean height with respect to LONG\(^2_f\)/LAT\(_f\), while keeping all the other variables fixed. Therefore, conditional to all fixed climatic variables, it can be inferred that coastal pine families exhibit a better growth compared to inland families, and also, southern families show a better performance than northern ones. For instance, let a family originally from Nashville, NC with geographical coordinates LONG\(_f\) = 78 W, LAT\(_f\) = 36 N be the reference family from which relative measures are taken. Then the ratio gives LONG\(^2_f\)/LAT\(_f\) = 169. So, on average (while keeping all the other predictors fixed), a family originally from about 30 km west to the reference family (LONG\(_f\)=78.2304 W, LAT\(_f\)=36N) would be 0.06 m shorter, since the ratio LONG\(^2_f\)/LAT\(_f\) is increased by 1 unit. Following with the same reasoning, a family originally from approximately 30 km south of the reference (LONG\(_f\)=78 W, LAT\(_f\)=35.7882 N) would be 0.06 m shorter while holding other variables fixed. The same result was found when we averaged height of pine families from a given geographic region and deviated from site mean as illustrated in Figure 1.4.

1.4.4 Climate change scenario

We presented four examples to illustrate possible applications of the proposed regression model. A hypothetical climate scenario was created from historical data, assuming a 5% decrease in precipitation and 0.5 °C increase in maximum and minimum temperatures. The climate change is relative to historical average values from the reference period 1970-2010. The chosen climate scenario is in agreement with the direction of forecasted climate conditions (Karl et al., 2009). Pine families in the statistical model are identified by their minimum temperatures and
geographical locations. As a result, the use of the model comes with an important caveat; the predicted height for two genetically different families coming from the same location and planted in the same place will be identical.

The first example consists of predicting the 8-year height growth of loblolly pine growing in their native geographical range. That is planting only local pine families across the southeastern United States (Figure 1.5). The actual estimates were calculated on the centroids of most of the counties in the southeastern U.S. and then interpolated to obtain a surface. From the contour plot it can be seen that the higher predicted height values cover a large area of the Georgia-Florida (GF) and South Carolina (SC) coastal plain regions. The highest predicted response (height growth) was concentrated in the southeast of Georgia and along the coastal plains of South Carolina, and northern Florida, indicating the superior growth of these local families compared to other families in coastal plain regions under the considered climate change scenario. This scenario suggests that local pine families would perform poorly along the coastal plains of FL and AL because of water shortage and increased temperatures. However, the uncertainty of predictions is larger towards coastal AL and inland VA, compared to other regions, mainly because of the lack of experimental test sites in those areas (Figure 1.2). As a consequence, predicted values can be considered to lie outside the observed range of test sites, so the risk of obtaining misleading results is increased. The URF modeling approach requires that provenances or seed sources to be planted on a wide spectrum of test sites covering the natural range of climatic conditions of the species to enable statistically valid calibrations. Moreover, if possible, a few tests should be beyond the natural range of the species (Kapeller et al., 2013). Extrapolation of predicted values beyond the range of the observed data is not recommended for regression models, because they can produce meaningless results. Another issue with the spatial distribution of the test sites throughout the Southeast is that, the sites that are close to each other, such as AL1 and AL2, or AL4 and AL3, or SC1 and SC2, provide
essentially the same information as if each pair were just one site. This is because highly corre-
related responses are expected from families subjected to similar climatic conditions. Except
in the mountain region where altitude can considerably affect climatic conditions in short dis-
tances, the southeastern United States is characterized by a smooth spatial variation of climatic
variables. For instance, even though there are four test sites in Alabama, it can be argued that
effectively there are just two. One possible way to improve the statistical model is to relax the
assumption of independence of observations and assume a covariance structure for the residu-
als that takes into account the correlations between test sites based on the relative distance to
each other. Therefore, because of the higher degree of uncertainty relative to other areas, and
because of the spatial distribution of test sites, henceforth predictions about the gulf coast of
Florida and Alabama, and inland Virginia will be considered to lie outside the observed range
defined by the spatial distribution of test sites.

The second example illustrates the difference (gain or loss) between the hypothetical cli-
mate change scenario and current climate conditions for the predicted 8-year height for local
pine families (Figure 1.6). The chart shows the effect of the simulated climate change sce-
nario on the growth of local families across the southeastern United States, relative to current
climate conditions. It is interesting to note that local pine families from the Georgia-Florida
coastal plain and part of South Carolina coastal plain regions exhibit the highest differences
(gain) in growth. Therefore, these families are predicted to perform better in their natural
habitat under the modeled climate change scenario compared to their performance today. This
example demonstrates the usefulness of the model: the combined effect of the climate vari-
ables on the growth can be assessed and quantified to make deployment decisions and guide
breeding strategies to develop pine genetic stock for future plantations.

The third example shows the predicted response (height growth) of coastal South Carolina
pine families planted across the southeastern United States (Figure 1.7). This seed source was
chosen since it is considered the fastest growing provenance of loblolly pine in the southeastern United States (Chamblee, 2011). The plot indicates that South Carolina families exhibit a more robust performance relative to local families, specifically in central Georgia, coastal North and South Carolina.

The fourth application example shows the predicted 8-year height deviations from site mean of Virginia families throughout the southeastern United States for the same hypothetical climatic future scenario previously described (Figure 1.8). The blue colors in the predicted heights plot indicate that Virginia families will only be superior in the most northern regions. The relative performance of Virginia families decreases toward the south, and the performance is very poor in northern Florida; which is in agreement with previous studies (Schmidtling, 2001; Wells, 1983; Wells and Lambeth, 1983). This example can be considered extreme in the sense that it is very unlikely to observe movements of Virginia seed sources to southern regions. Yet, the example is useful to assess the model results by contrasting them against previous studies.

1.5 Conclusions

We presented a data-driven statistical framework to predict height growth of loblolly pine from seed source test data along with climatic and geographical variables. The statistical model based on regression analysis can be used to assess the environmental effects of planting sites (through population response functions) and genetic effects of climatic variables (through transfer functions) on the performance of loblolly pine seed sources. The model is designed to predict the growth of any given family under a specified climate scenario. Therefore, it can be used to develop forest management strategies to mitigate the negative impacts of climate change on plantation forestry. The response functions of seed sources and families can be used
to develop loblolly pine breeding and deployment strategies to meet the long-term challenges of climate change. In order to get meaningful results from the statistical model, close attention must be given to the data from provenance test, and to the validity range of predictions. Ideally, provenances should have been carefully planted in multiple test sites covering the natural range of climatic conditions of the species. Extrapolation of predicted values beyond the range of the observed data is not recommended for regression models because the risk of producing meaningless results is increased.
### 1.6 Tables

Table 1.1: Climatic and geographical variables used to develop the statistical model. Precipitation is measured in millimeters (mm) and temperature in Celsius degrees (°C).

<table>
<thead>
<tr>
<th>Climate variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPT&lt;sub&gt;t&lt;/sub&gt;</td>
<td>Test site yearly average total precipitation</td>
</tr>
<tr>
<td>TMIN&lt;sub&gt;t&lt;/sub&gt;</td>
<td>Test site yearly average minimum temperature</td>
</tr>
<tr>
<td>TMAX&lt;sub&gt;t&lt;/sub&gt;</td>
<td>Test site yearly average maximum temperature</td>
</tr>
<tr>
<td>PPT&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Female parent origin yearly average total precipitation</td>
</tr>
<tr>
<td>TMIN&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Female parent origin yearly average minimum temperature</td>
</tr>
<tr>
<td>TMAX&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Female parent origin yearly average maximum temperature</td>
</tr>
<tr>
<td>PPT&lt;sub&gt;p&lt;/sub&gt;</td>
<td>Pollen parents’ yearly average total precipitation</td>
</tr>
<tr>
<td>TMIN&lt;sub&gt;p&lt;/sub&gt;</td>
<td>Pollen parents’ yearly average minimum temperatures</td>
</tr>
<tr>
<td>TMAX&lt;sub&gt;p&lt;/sub&gt;</td>
<td>Pollen parents’ yearly average maximum temperatures</td>
</tr>
<tr>
<td>PPT&lt;sub&gt;f&lt;/sub&gt;</td>
<td>Family precipitation: (PPT&lt;sub&gt;p&lt;/sub&gt; + PPT&lt;sub&gt;m&lt;/sub&gt;) / 2</td>
</tr>
<tr>
<td>TMIN&lt;sub&gt;f&lt;/sub&gt;</td>
<td>Family minimum temperature: (TMIN&lt;sub&gt;p&lt;/sub&gt; + TMIN&lt;sub&gt;m&lt;/sub&gt;) / 2</td>
</tr>
<tr>
<td>TMAX&lt;sub&gt;f&lt;/sub&gt;</td>
<td>Family maximum temperature: (TMAX&lt;sub&gt;p&lt;/sub&gt; + TMAX&lt;sub&gt;m&lt;/sub&gt;) / 2</td>
</tr>
<tr>
<td>LAT&lt;sub&gt;f&lt;/sub&gt;</td>
<td>Latitude coordinate of female parent origin</td>
</tr>
<tr>
<td>LONG&lt;sub&gt;f&lt;/sub&gt;</td>
<td>Longitude coordinate of female parent origin</td>
</tr>
</tbody>
</table>
Table 1.2: Regression model coefficients and standard errors for 8-year height of loblolly pine using ordinary least squares (OLS), Ridge, and LASSO regression models. The nine variables were chosen out of the initial set of 34 climatic variables for test site, family, and geographical variables for family. The models explained about 22% of the total variation in the 8-year height.

<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS</th>
<th>Std. Error</th>
<th>Ridge</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPT&lt;sub&gt;t&lt;/sub&gt;</td>
<td>-13.84</td>
<td>0.17</td>
<td>-13.83</td>
<td>-12.39</td>
</tr>
<tr>
<td>PPT&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;t&lt;/sub&gt;</td>
<td>16.04</td>
<td>0.21</td>
<td>16.02</td>
<td>13.53</td>
</tr>
<tr>
<td>TMIN&lt;sub&gt;t&lt;/sub&gt;</td>
<td>15.39</td>
<td>0.33</td>
<td>15.36</td>
<td>10.79</td>
</tr>
<tr>
<td>TMIN&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;f&lt;/sub&gt;</td>
<td>-0.17</td>
<td>0.01</td>
<td>-0.17</td>
<td>-0.18</td>
</tr>
<tr>
<td>TMIN&lt;sub&gt;t&lt;/sub&gt; × TMIN&lt;sub&gt;f&lt;/sub&gt;</td>
<td>0.26</td>
<td>0.01</td>
<td>0.26</td>
<td>0.22</td>
</tr>
<tr>
<td>PPT&lt;sub&gt;t&lt;/sub&gt; × TMIN&lt;sub&gt;t&lt;/sub&gt;</td>
<td>-12.13</td>
<td>0.19</td>
<td>-12.12</td>
<td>-10.69</td>
</tr>
<tr>
<td>PPT&lt;sub&gt;t&lt;/sub&gt; × TMAX&lt;sub&gt;t&lt;/sub&gt;</td>
<td>-1.02</td>
<td>0.09</td>
<td>-1.02</td>
<td>0.00</td>
</tr>
<tr>
<td>TMIN&lt;sub&gt;t&lt;/sub&gt; × TMAX&lt;sub&gt;t&lt;/sub&gt;</td>
<td>-3.01</td>
<td>0.29</td>
<td>-2.99</td>
<td>-0.04</td>
</tr>
<tr>
<td>LONG&lt;sup&gt;2&lt;/sup&gt;&lt;sub&gt;f&lt;/sub&gt;/LAT&lt;sub&gt;f&lt;/sub&gt;</td>
<td>-0.06</td>
<td>0.01</td>
<td>-0.06</td>
<td>-0.05</td>
</tr>
</tbody>
</table>
Table 1.3: Estimated mean and standard deviation of the mean squared errors (MSE) obtained from random splits of the data (training and test sets) for each regression method using cross-validation. The three regression methods gave essentially identical results.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinary least squares (OLS)</td>
<td>2.6734</td>
<td>0.0172</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>2.6734</td>
<td>0.0172</td>
</tr>
<tr>
<td>LASSO regression</td>
<td>2.6865</td>
<td>0.0175</td>
</tr>
</tbody>
</table>
Figure 1.1: Geographic regions and county of origin of the selected mother trees. The seven regions were drawn based on the observed variation within the species as well as easily determined political boundaries in the southeastern United States.
Figure 1.2: Test site locations measured for this study. These 16 test locations cover a substantial area of the natural range of loblolly pine (shaded area) in the east of the Mississippi river.
Figure 1.3: Mean height deviation of seven seed sources (average of families) from site mean. Each curve can be interpreted as the relative performance of families from a particular geographic region (seed source) across test sites. The test sites are ordered in a southwest to northeast direction. Southern sources from SC, GF, and LG regions were not planted in VA1 site due to their expected poor cold tolerance (adaptability).
Figure 1.4: Observed heights (Height_{tf}) of pine families (f) at various sites (t) as function of family (TMIN_f) and site (TMIN_t) minimum temperature, and estimated response surface using ordinary least squares. The black dots represent the observed 8-year heights of loblolly pine across the southeastern United States. Planting families from warm regions (TMIN_f > 3) at colder sites (TMIN_t < -3) greatly reduces their mean height as seen by the red colors in the corner of the curve (left side of plot). Conversely, planting families from cold regions (TMIN_f < -3) at warmer sites (TMIN_t > 3) also greatly reduces their height as seen by the red colors in the corner of the curve (right side of plot).
Figure 1.5: Predicted 8-year height of loblolly pine local families across the southeastern U.S. using ordinary least squares (OLS) regression model for a hypothetical future scenario with 5% decrease in precipitation and an increase of 0.5°C in maximum and minimum temperatures, relative to historical average values (reference period 1970-2010). The dark blue regions indicate where the local families are predicted to perform well. Predictions for the Gulf Coast of Florida, and north of Virginia are considered to lie outside the observed range defined by the spatial distribution of test sites.
Figure 1.6: Difference between the predicted 8-year height under the hypothetical climate change scenario and the predicted height using current climate conditions for local pine families. The chart illustrates the effect of the simulated climate change scenario on the growth of local families across the southeastern United States, relative to current climate conditions. Blue and light blue areas on the map indicate a positive difference or increase in mean height, whereas yellow and dark orange areas denote a negative difference or decrease in height. The Georgia-Florida coastal plain and part of South Carolina coastal plain regions exhibit the highest differences (gain in height) in growth relative to the current climate. Predictions for the Gulf Coast of Florida and Alabama, and north of Virginia are considered to lie outside the observed range defined by the spatial distribution of test sites.
Figure 1.7: Predicted 8-year height growth of coastal South Carolina loblolly pine families planted across the southeastern United States for a hypothetical climate change scenario (5% decrease in precipitation and 0.5 °C increase in maximum and minimum temperatures), relative to historical average values (reference period 1970-2010). Coastal South Carolina families are expected to perform well across the Southeast, but highest growth (dark blue) is expected to be in southern Georgia, under the modeled climate change scenario. Predictions for the gulf coast of Florida, and north of Virginia are considered to lie outside the observed range defined by the spatial distribution of test sites.
Figure 1.8: Ordinary least squares (OLS) regression model predictions for 8-year height deviations from site mean of families from Virginia seed source for a hypothetical climate change scenario (5% decrease in precipitation and 0.5 °C increase in minimum and maximum temperatures), relative to historical average values (reference period 1970-2010). The contour plot indicates that this seed source will only be superior in the most northern region (dark blue area) and perform poorly relative to other sources in the most southern region (red area), under the modeled climate change scenario.
Chapter 2

Optimal Seed Deployment under Climate Change using Spatial Models: Application to Loblolly Pine in the southeastern US

2.1 Introduction

There is evidence that global climate is changing rapidly relative to historical trends (Stocker et al., 2014). As a result, there is considerable interest in predicting the biological response of forest trees to changes in the environment. Forest trees will likely experience climate changes and will have to either adapt or perish. To avoid extinction, tree populations must respond to adverse changes in environmental conditions either by migrating to new locations where conditions are more favorable, or adapting to the new conditions in the current location (Aitken et al., 2008). Unfortunately trees cannot migrate very rapidly. The changes in climate that took place during past ice ages were slow and some tree species were able to migrate as the glaciers advanced (Schmidtling et al., 2007). However, the changes in climate that the earth is currently
facing with global warming are likely to occur at a faster pace, and hence the survival of many tree species will be threatened.

In a climate change context, locally-adapted tree species will experience environmental conditions to which they are not well adapted (Skelly et al., 2007). As a result, planting seedlings predicted to be adapted to future climate conditions has been proposed as a forest management strategy aimed to mitigate the potential negative impacts of global climate change (Ledig and Kitzmiller, 1992). The risks of adaptation of future forest tree plantations can be mitigated with an assisted migration of suitable genetic material to ensure that plantations are adapted to future climate conditions (Aitken et al., 2008).

Loblolly pine (Pinus taeda L.) is a native pine species in the southeastern United States. Its natural range covers a diverse area that extends from the Atlantic and Gulf Coastal Plains to the Piedmont regions. Loblolly pine is the most hardy and versatile of all southern pines (Schultz, 1997) and is the major commercial species in the Southeast (McKeand et al., 2007). It is used in a broad range of products, such as pulpwood, plywood, and construction lumber. Hence it is planted widely for commercial timber production and subjected to an intense breeding program. In addition to supply resources for human use, loblolly forests provide watershed protection and vital organic infrastructure for countless species in the region. Therefore, the species has vital commercial and environmental value to the southeastern United States.

It is very likely that climate change will have significant impacts on pine plantations across the southeastern United States. The Southeast is expected to experience an increase in the rate of warming through the end of the century, with a rise in average temperature of 2.5 to 5°C by the 2080s (Karl et al., 2009). Precipitation predictions for the region are less certain, but the trend indicates a decline of 10% to 30% of summertime precipitation (Solomon, 2007). The expected precipitation decrease combined with elevated temperatures will likely increase vapor pressure and soil water deficits causing stress to the loblolly plantations and a decrease
in productivity of at least 10% (Schmidtling, 1994).

The establishment and analysis of provenance tests for investigating the genetic variation among forest trees has a long tradition in forestry (Langlet, 1971). Provenance tests are common garden experiments generally intended to identify superior genotypes for planting at specific environments or across environments. Such tests are usually replicated experiments established with seed from parent trees collected from different geographical regions within the species distribution and grown at several locations. In addition to its traditional use, provenance tests provide meaningful information for assessing the response of tree populations to environmental change (Schmidtling, 1994; Matyas, 1994). The geographic location from where the seed parents were collected, the location of the test sites where the progeny trees were grown, and the climate of these two sites can be used to make predictions about the relative performance of tree populations under different environmental conditions.

Predictive models to assess the effect of climate variables on tree phenotypes using provenance test data have been developed by different investigators. For instance, the population response function approach models the performance of a single population across different planting sites in terms of the climate at those locations. This approach was used by several researchers to predict the impact of climate change on a particular population (Schmidtling, 1994; Matyas, 1994; Rehfeldt et al., 1999; Wang et al., 2006; O’Neill et al., 2008). Transfer functions using provenance test data have been also developed. A transfer function gives the performance of multiple populations as a function of the climatic transfer distance, which is the difference between the climate of the original location of the individuals and the climate of the planting site (Matyas, 1994; Rehfeldt et al., 1999; Andalo et al., 2005). The Universal Response Function (Wang et al., 2010) approach integrates both the population response and transfer functions into a single model. It makes use of all available information from provenance tests to estimate a joint response of any population growing in any climate (McLane
et al., 2011; Kapeller et al., 2013; Farjat et al., 2014). All of these models are based on multiple linear regression analysis that ignore the spatial dependence structure of provenance trials and utilize climate and geographical variables as predictors. The other limitation of these approaches is that the joint response is constrained to a parametric linear function that may fail to characterize the underlying process.

In this work a spatial Bayesian approach is developed for modeling the relative performance of seed sources of loblolly pine using climate and geographical variables as predictors associated with the origin of the seed and the planting site. The underlying assumption is that both genetic and environmental effects are captured through these variables. The proposed modeling approach generalized previous universal response function methodology by accounting for complex non-linear climate relationships and spatial dependence. We model the response function as a Gaussian process over climate variables and the spatial coordinates of the origin and planting sites. This poses two types of climate/spatial dependence: for a given planting site we would expect similar relative performance for trees from nearby seed sources; and for a given seed source, we would expect similar relative performance at two nearby planting sites.

Also, since the goal of this study is to predict performance under different climate scenarios, ideally dependence between planting sites and seed sources would be explained by differences in climate variables. This motivates our covariance model which is a function of both spatial and climate variables from the planting site and seed sources. To account for all of these features in a computationally-feasible way, we employ a separable covariance function that is the product of the covariance between planting sites and the covariance between seed origins. The covariance is separable in the sense that the correlation between observations at different sites and from different origins is expressed as the product of two separate components. Under the proposed framework, we compared models based on geographic distances, climate distances, and a combination of both geographic and climate distances. Although Bayesian analysis often
requires higher computational cost compared to classical methods, especially in models with a large number of parameters, it provides a convenient setting for the proposed model and for the particular application that contains missing data. In addition, the formulation allows us to combine prior knowledge with the data and provides inferences that are exact and conditional on the data of any sample size, without relying on asymptotic approximations. The other advantage of the Bayesian approach includes the possibility of accounting for the uncertainty of the parameter estimates on any functional of the parameters.

The objectives of this work are to 1) develop a statistical model to predict the relative response of loblolly pine seed sources to climate change, 2) compare the predictive ability of the proposed model against multiple linear regression methods, and 3) use the model to create a quantitative tool for seed deployment aimed to identify the location of superior performing seed sources that are suitable for specific regions under a given climate scenario. To demonstrate the use and extent of the proposed model, application examples are presented and discussed concerning the predicted performance of loblolly pine seed sources for current and future climatic scenarios across the southeastern United States.

2.2 Description of the data

2.2.1 Genetic Material

The Cooperative Tree Improvement Program from North Carolina State University\(^1\) conducted an intensive selection effort in unimproved loblolly pine plantations between 1975 and 1981 aimed to increase the size of breeding populations. Trees were selected based on phenotypic attributes such as good growth, straight stems, and absence of fusiform rust galls (a pathological

\(^1\)http://www.treeimprovement.org
swelling caused by the fungus *Cronartium quercuum f. sp. fusiforme*). In 1994, some of these plantation selections were used as parents to establish the Plantation Selection Seed Source Study (PSSSS), a large replicated series of provenance-progeny tests in the southeastern United States. The experiment was designed to evaluate the actual patterns of geographic genetic variation of plantation selections and to assess genotype by environment interactions, so that informed decisions could be made about seed source deployment (Chamblee, 2011).

In early plantations of loblolly pine across the Southeast it was not uncommon to observe seed movements up to 500 km from their original location. As a result, the exact origin of the selections used as parents in the PSSSS is not known. Nevertheless, based on the knowledge of foresters who established these plantations, we are confident that most of the sources of the seed are local or within a 100 to 200 km distance from the origin. Therefore, throughout this work, instead of the original geographic source of seed (provenance) we will refer to seed sources to indicate the locality and climate where a source of seed was collected.

The trees selected for the PSSSS were chosen from a pool of plantation selections identified in seven geographic regions within the natural range of loblolly pine in the southeastern United States. Figure 2.1a shows the location of the pine plantations from where the female parents were selected. The plot displays the seven geographic regions of the natural range of loblolly pine defined by the Cooperative Tree Improvement Program of NCSU, namely: Virginia (VA), North Carolina Coastal Plain (NC), South Carolina Coastal Plain (SC), Georgia-Florida Coastal Plain (GF), Lower Gulf Coastal Plain (LG), Upper Gulf Coastal Plain (UG), and Piedmont (PD). These seven regions were delineated based on the observed geographic variation within the species as well as easily distinguishable state and physiographic boundaries. In each of seven regions, 20 selections were randomly chosen as female parents. These selections were then mated with a pollen mix of 40 pollen parents from the same region. As a result, a total of 140 maternal families were employed for the study, and a total of 280 male parents were used.
in the pollen mixes. The female and pollen parents were unrelated in order to avoid inbreeding and increase the genetic sample size.

2.2.2 Experimental Design

Field trials were established using a randomized complete block design. The experiment was originally replicated at 25 locations across the southeastern United States, however only 16 sites covering a substantial area of the natural range of loblolly pine were analyzed in this work (Figure 2.1b). Each site had 24 blocks comprising a single tree from each family and four genetically unimproved local seedlings used as control (mixed seedlots), randomly distributed within blocks. Test sites were managed following standard progeny test protocols that included site preparation before planting and weed and vegetation control during the experiment. No fertilization treatment was applied at any site. The experiment was intended to be balanced, yet southern seed sources from South Carolina Coastal, Georgia and Florida Coastal, and the Lower Gulf were not planted in four northern sites because of space limitations and expected poor adaptability (Schmidtling, 2001; Chamblee, 2011).

2.2.3 Standardized Height

Growth (height and diameter at breast height), stem straightness, fusiform rust disease incidence, forking, and survival of over 43,000 trees were assessed at age 8 years after planting. Although several traits were measured, the primary trait under consideration in this work is height because it is a heritable trait subjected to natural selection (McKeand et al., 2008) very likely to respond to changes in climate.

In order to quantify the performance of a specific seed source planted at a given site, the
8-year standardized height was obtained as follows:

\[ Y(s, t) = \frac{\bar{X}(s, t) - \hat{\mu}_t}{\hat{\sigma}_t} \]  

(2.1)

where \( \bar{X}(s, t) \) represents the mean height of trees from the seed source from geographic location \( s \) planted at test site \( t \), \( \hat{\mu}_t \) denotes the sample mean across seed sources obtained at site \( t \), and \( \hat{\sigma}_t \) is the sample standard deviation of seed source means from site \( t \). For a given test site, \( Y(s, t) \) measures the performance of a specific seed source relative to the others present in the same site. The standardized height was used as the response variable for the proposed model instead of the absolute height because the goal was to identify superior genotypes while accounting for environmental factors. Standardization removes site effects that may have influenced the growth and forces the response at different sites to have zero mean and variance equal to one.

### 2.2.4 Climate Variables

The climate variables associated with planting site and seed source were estimated using the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate mapping system (Daly et al., 1994, 2008). The PRISM system was created to estimate climate variables in regions of complex topography and it has been used extensively to map precipitation and temperature over the United States. The temporal resolution of the PRISM mapping system was limited to monthly average estimates.

Yearly average total precipitation (PPT), yearly average minimum temperature (TMIN), and yearly average maximum temperature (TMAX) of both test site and seed source locations were used as explanatory variables to develop prediction models. The hypothesis underlying the development of the model is that the relative variation in phenotypes is influenced
by climate and occurs geographically; therefore, this variation can be predicted from climate descriptors. Climate variables will be described by subscripts $s$ or $t$ throughout this work, denoting seed source and test site, respectively.

The climate variables of the 16 test sites were obtained by averaging the monthly estimates for the period 1994-2006, whereas the period 1970 to 1990 is used for the seed source climate variables. The centroids of the counties from where the selections were collected were used as the reference location for the seed sources.

2.3 Model formulation

2.3.1 Spatial seed source transport model

This section describes the statistical approach developed to model the relative performance of seed sources planted at different locations using geographical and climate variables as predictors. Let $Y(s, t, u_s, u_t)$ be the standardized height defined in (2.1) for seeds from geographical location $s \in \mathbb{R}^2$ and planted at site $t \in \mathbb{R}^2$, with associated climate variables from site $t$ and location $s$ denoted as $u_t = (u_{t1}, \ldots, u_{td})^T \in \mathbb{R}^d$ and $u_s = (u_{s1}, \ldots, u_{sd})^T \in \mathbb{R}^d$, respectively.

The standardized height is defined at every point over the region of interest $D \subset \mathbb{R}^{d+2} \times \mathbb{R}^{d+2}$, and modeled as a random process $Y(\cdot) = \{Y(s, t, u_s, u_t) : s, u_s \in \mathbb{R}^{d+2}; t, u_t \in \mathbb{R}^{d+2}\}$.

The proposed statistical model is inspired by classical geostatistics (Cressie, 1993) and takes the form:

$$Y(s, t, u_s, u_t) = \mu(u_s, u_t) + e(s, t, u_s, u_t)$$

(2.2)

where $\mu(u_s, u_t) = E[Y(s, t, u_s, u_t)]$ is the mean function assumed to be deterministic, and $e(s, t, u_s, u_t)$ is the mean-zero error. The mean term captures the general trend or large-scale spatial variations of the process under study using only climate variables. Geographical vari-
ables were not included in the mean function because the proposed model is aimed to identify optimal planting locations for a given seed source based on climatic conditions rather than geographical locations. We considered a linear parametric mean model:

$$\mu(u_s, u_t) = x(u_s, u_t)^T\beta$$

(2.3)

where $x(u_s, u_t)$ is a $p$-vector of covariates formed by the climate variables and $\beta$ is a $p$-vector of unknown regression parameters.

The second term in (2.2) accounts for the spatial dependence and is a function of both geographic locations and climate variables. The error process $e(\cdot)$ also accounts for the presence of an inherent measurement error in the observations. The measurement error is considered as an independent process without any spatial dependence. We decompose the error process into two additive contributions:

$$e(s, t, u_s, u_t) = \eta(s, t, u_s, u_t) + \varepsilon(s, t, u_s, u_t)$$

(2.4)

where $\eta(s, t, u_s, u_t)$ is the spatially-dependent process and $\varepsilon(s, t, u_s, u_t) \sim N(0, \sigma^2)$ is the measurement error process. The inclusion of the spatially-dependent process is based on the assumption that for a given planting site, performance of seed from neighboring origins would be similar; and conversely, performance of a given seed source at two neighboring sites would be approximately the same. Furthermore, dependence between planting sites and seed sources would be explained by differences in climate variables.

Figure 2.2 illustrates the underlying concept of the proposed model with a hypothetical and simplified example that uses minimum temperature and latitude as predictors. The lines in the plot represent different climate scenarios corresponding to current and future climate.
The idea is that given the current minimum temperature and latitude, the expected response of the underlying process can be characterized. In the same way, projections of future minimum temperature combined with latitude coordinates can be used to obtain the expected response of the process under consideration for the projected scenario.

The spatial covariance function of $\eta(s, t, u_s, u_t)$ is assumed to be separable, i.e., the product of the covariance for the planting sites, $C_t$, and the covariance for the seed sources, $C_s$.

The implications of the separability assumption are that the covariance across seed sources is the same for all planting sites, and the covariance across planting sites is the same for all seed sources. This separation is mathematically convenient and computationally advantageous because it allows the full covariance matrix to be written as the Kronecker product of two covariance matrices, which facilitates algebraic manipulations and the use of numerical methods to speed up the model fitting algorithm.

We considered a zero-mean Gaussian process for the spatially-dependent process $\eta(\cdot)$ with covariance function:

$$
\text{cov}[\eta(s, t, u_s, u_t), \eta(s', t', u'_s, u'_t)] = C(r_{ss'}, r_{tt'}, \theta) = \tau^2 C_s(r_{ss'}, \theta_s)C_t(r_{tt'}, \theta_t)
$$

(2.5)

where $\tau^2$ is the spatial variance, and $\theta_s$ and $\theta_t$ are the covariance parameters for seed sources and test sites, respectively; $r_{ss'}$ and $r_{tt'}$ denote generalized distances for seed sources ($s$ and $s'$) and test sites ($t$ and $t'$), respectively. These generalized distances were defined to include the climate variables for seed source and test site as follows:

$$
r_{ss'} = r_{ss'}(s, s', u_s, u'_s) = \sqrt{\left(\frac{||s - s'||}{\varphi_s}\right)^2 + \sum_{i=1}^{d} \left(\frac{u_{si} - u'_{si}}{\varphi_{u_{si}}}\right)^2}
$$

(2.6)
where \( \varphi_s \) and \( \varphi_t \) are unknown scale parameters associated with the seed source and test site geographical locations; \( \varphi_{u_{t_i}} \) and \( \varphi_{u_{t'i}} \) for \( i = 1, \ldots, d \) are unknown scale parameters for the climate variables observed at the location where seeds were collected and at the sites.

The spatially dependent error process \( \eta(s, t, u_s, u_t) \) is random but continuous in \( s, t, u_s, \) and \( u_t \). Valid isotropic covariance functions \( C_s \) and \( C_t \) were generated from the Matérn parametric family of correlation functions (Matérn, 1960, 1986). For instance, the correlation between observations from sites \( t \) and \( t' \) with generalized distance as in (2.7), is given by:

\[
C_t(r_{tt'}, \theta_t) = \frac{2^{1-\kappa_t}}{\Gamma(\kappa_t)} (r_{tt'})^{\kappa_t} K_{\kappa_t}(r_{tt'})
\]

(2.8)

where \( \theta_t = (\varphi_t, \varphi_{u_{t_1}}, \ldots, \varphi_{u_{t_d}}, \kappa_t) \) groups the scale and smoothness parameters associated with the geographic and climate variables for the test sites, \( K_{\kappa_t} \) is a modified Bessel function of the second kind of order \( \kappa_t \), \( \Gamma(\cdot) \) denotes the gamma function, and \( \kappa_t \) is the smoothness parameter for the sites, which directly controls the smoothness of the random field.

The scale parameters that control the range of the correlation are in the generalized distance \( r_{tt'} \). Similarly for the seed sources, the correlation function between observations from \( s \) and \( s' \) is denoted as \( C_s(r_{ss'}, \theta_s) \), where \( r_{ss'} \) is the generalized distance defined in (2.6), and \( \theta_s = (\varphi_s, \varphi_{u_{s_1}}, \ldots, \varphi_{u_{s_d}}, \kappa_s) \) is the vector of scale and smoothness parameters associated with the geographic and climate variables for the seed sources.

The Matérn class provides a flexible parametrization that includes the exponential correlation as a subclass with a smoothness parameter \( \kappa = \frac{1}{2} \). The square exponential correlation function is also a member that arises in the limit as \( \kappa \to \infty \), and represents the upper limit of...
smoothness in the class (Handcock and Stein, 1993). The smoothness parameter is important because its integer part determines the mean square differentiability of the underlying process, which affects the behavior of predictions. The historical development of this parametric family and its many areas of applications has been thoroughly documented by Guttorp and Gneiting (2006).

2.3.2 Spatial and Climate Predictions

The underlying concept of the proposed model was illustrated in Figure 2.2 using a simplified example comprising latitude and minimum temperature. The same concept can be extrapolated to higher dimensions by including multiple climatic and geographical variables to predict the response of the process under consideration. Hence, the proposed parametric model (2.2) can be used to provide spatially explicit estimates of seed source performance across the Southeast for current and future climates. For instance, to estimate the performance of a given seed source across the region of interest under current climate, we hold constant the variables associated with the seed, and we set the planting site climate variables equal to current climate values. Similarly, to predict the performance under hypothesized future climate values, the planting site climate variables are replaced by the site hypothesized future climate values. Following this approach, the performance of seed sources at a given planting site can be compared for current and hypothesized future climates, leading to optimal planting strategies.

2.4 Bayesian Formulation

Let $Y$ denote the $n$-vector of response values, that is, the vector of standardized heights ordered as sources within sites. Note that the dimension of the response vector is given by the product
of the number of sites \((n_t)\) by the number of seed sources \((n_s)\) in the study, that is, \(n = n_t n_s\).

Let \(X\) represent the \(n \times p\) design matrix of predictors, which for this study was a collection of covariates created from climate variables observed at the test sites and at the seed sources.

Let \(\Sigma = \Sigma(\theta)\) denote the variance-covariance matrix of the response vector, where \(\theta\) is a \(q\)-vector of unknown parameters that belong to a parameter space \(\Theta \subset \mathbb{R}^q\), within which the covariance function is valid. Since the error process is Gaussian, the joint distribution of \(Y\) is multivariate normal with mean \(X\beta\) and covariance matrix \(\Sigma(\theta)\), which will be denoted hereafter as follows: \(Y \sim N(X\beta, \Sigma(\theta))\). The structure of the spatial covariance matrix can be written in matrix notation as follows:

\[
\Sigma(\theta) = \tau^2 \Gamma(\theta_s, \theta_t) + \sigma^2 I
\]  

where \(\Gamma(\theta_s, \theta_t)\) is the spatial correlation matrix, \(\tau^2\) denotes the spatial variance or partial sill, \(\sigma^2\) represents the measurement error or nugget effect, and \(\theta_s\) and \(\theta_t\) are vectors that contain the scale and smoothness parameters from seed source and site, respectively. The separability assumption allows us to express the spatial correlation as a Kronecker product of two matrices. Hence, provided that the response vector is ordered as seed sources within test sites, we have:

\[
\Sigma(\theta) = \tau^2 \Gamma(\theta_t) \otimes \Gamma_s(\theta_s) + \sigma^2 I
\]  

where vector \(\theta\) contains all the unknown parameters that define the spatial covariance matrix, that is: \(\theta = (\tau^2, \sigma^2, \theta_s^T, \theta_t^T)^T\). Under the proposed framework, the likelihood of the data \(Y\) given the parameters of the model \(\beta\) and \(\theta\) is given by:

\[
Y|\beta, \theta \equiv Y|\beta, \tau^2, \sigma^2, \theta_s, \theta_t \sim N(X\beta, \tau^2 \Gamma(\theta_s, \theta_t) + \sigma^2 I)
\]
The Bayesian model is completed by assigning a prior distribution to all the model un-
knowns, that is $\beta$ and $\theta$. The vector of regression coefficients $\beta$ was assigned a relatively
non-informative prior given by: $\beta \sim N(0, \Sigma_\beta)$, where $\Sigma_\beta = \sigma^2_\beta I$ was considered diagonal. The hyper-parameter $\sigma^2_\beta$, that define the prior knowledge about the coefficients was chosen cor-
responding to a fairly vague prior. The parameters from the covariance matrix $\theta$ were assigned
a log-normal prior to enforce their positiveness, that is:

$$\log \theta_i \sim N(\mu_{\theta_i}, \sigma^2_{\theta_i})$$

(2.12)

where the location $\mu_{\theta_i}$ and scale $\sigma^2_{\theta_i}$ hyper-parameters were chosen to obtain fairly vague priors, except for the smoothness parameters $\kappa_s$ and $\kappa_t$ for which the priors were somewhat informative around the exponential covariance ($\kappa_s = \kappa_t = \frac{1}{2}$).

The posterior distribution of the the model parameters given the data $p(\beta, \theta | Y)$ is propor-
tional to the product of the likelihood of the data given the parameters $p(Y | \beta, \theta)$ and the prior
joint distribution of the parameters $p(\beta, \theta)$, that is:

$$p(\beta, \theta | Y) \propto p(Y | \beta, \theta)p(\beta, \theta) \propto p(Y | \beta, \theta) \prod_{i=1}^{p} p(\beta_i) \prod_{i=1}^{q} p(\theta_i)$$

(2.13)

where parameters $\beta$ and $\theta$ were assumed to be mutually independent among and within each other.

2.4.1 Bayesian Computation

The normalizing constant for the posterior distribution (2.13) is not available analytically. So,
Monte Carlo methods (Carlin and Louis, 2011; Robert and Casella, 2013) were used for es-
timation and inference of the model parameters. Specifically, a Gibbs sampler (Gelfand and
Smith, 1990) was developed to draw samples from the posterior distribution of the model parameters. The Metropolis algorithm (Metropolis et al., 1953) was used to sample from the full conditional distributions in cases where the prior and the likelihood did not form a conjugate pair, and hence, a closed form was not available. The regression coefficients \( \beta \) were sampled from the closed form full conditional \( p(\beta | \cdot) = N(\tilde{\mu}_{\beta}, \tilde{\Sigma}_{\beta}) \), where \( \tilde{\Sigma}_{\beta} = (X^T \Sigma^{-1} X)^{-1} \), and \( \tilde{\mu}_{\beta} = \tilde{\Sigma}_{\beta} X^T \Sigma^{-1} Y \). Samples from the full conditional distribution \( p(\theta | \cdot) \) were obtained from the Metropolis algorithm using a multivariate normal \( N(\theta^{(t-1)}, \tilde{\Sigma}_{\theta}) \) proposal density, where \( \theta^{(t-1)} \) denotes the previous sample and \( \tilde{\Sigma}_{\theta} \) is the proposal covariance matrix. The choice of \( \tilde{\Sigma}_{\theta} \) was thoroughly explored to accelerate the convergence of the MCMC chains. To avoid the issue of correlations among the elements of \( \theta \), the proposal covariance-matrix was set to be diagonal, that is \( \tilde{\Sigma}_{\theta} = \text{diag}(\tilde{\sigma}^2_{\theta_1}, \ldots, \tilde{\sigma}^2_{\theta_q}) \), and its elements adaptively chosen during the burn-in period. For each \( \theta_i \), an initial value of \( \tilde{\sigma}^2_{\theta_i} \) was picked based on preliminary runs and then modified based on the empirical proportion of accepted candidates so that the overall rate be close to the optimal acceptance rate of about 31% (Gelman et al., 1996). This approach is usually referred to as pilot adaptation (Gilks et al., 1998). The convergence of the model parameters was assessed from visual inspection of the trace plots for MCMC posterior draws, the estimated posterior density, and the auto-correlation function of the samples drawn. Finally, to assess convergence we conducted the Geweke’s diagnostic (Brooks and Roberts, 1998) with the first 10% and last 50% of each chain.

2.4.2 Missing Data

The two-way table formed by test sites by seed sources height means \( \bar{X}(s, t) \) contained around 5% of missing values. Most of the missing values occurred in the northern sites because some southern seed sources were not planted in the northern sites due to space limitations and
expected poor adaptability (Chamblee, 2011). As a result, the response vector of standardized heights $Y$ had missing values that impeded the necessary calculations of the assumed balanced configuration in the Bayesian framework. Therefore, in order to account for the missing values, an iterative data augmentation approach was used in the Gibbs sampler (Tanner and Wong, 1987; Li, 1988).

Let $Y_{com} = (Y_{mis}^T, Y_{obs}^T)^T$ represent the complete data set that contains the vector of missing imputed values $Y_{mis}$ and the vector of observed values $Y_{obs}$. The missing data $Y_{mis}$ were drawn iteratively from the conditional distribution of the missing data given the observed values and the model parameters, that is $p(Y_{mis}|Y_{obs}, \beta, \theta)$. Then, the model parameters were sampled from their full conditionals using the augmented data.

This approach can be understood as to generate samples from the joint posterior distribution $p(\beta, \theta, Y_{mis}|Y_{obs})$ from the full conditionals. After convergence of the Gibbs sampler, the sampled values of parameters $\beta$ and $\theta$ are treated as draws from the joint posterior distribution given the observed values $p(\beta, \theta|Y_{obs})$. Hence, inferences about the model parameters can be made based on the posterior samples.

The separation of the data set into missing and observed values, makes it possible to write the inverse of the covariance matrix $\Sigma_{com}$ in terms of its matrix block constituents:

$$\Sigma_{com}^{-1} = \begin{pmatrix}
\Omega_{mis} & \Omega_{mis,obs} \\
\Omega_{obs,mis} & \Omega_{obs}
\end{pmatrix}$$

(2.14)

where $\Omega_{mis}$ and $\Omega_{obs}$ are the blocks associated with the missing and observed values, respectively; and $\Omega_{mis,obs} = \Omega_{obs,mis}^T$ is the block for the conditional covariance matrix between missing and observed values. Since the data follow a multivariate Gaussian process $Y_{com} \sim N(\mu_{com}, \Sigma_{com})$, where $\mu_{com} = (\mu_{mis}^T, \mu_{obs}^T)^T$, the full conditional of the missing data
can be written in closed form:

\[ p(Y_{\text{mis}} | \cdot) = N \left( \mu_{\text{mis}} - \Omega_{\text{mis}}^{-1} \Omega_{\text{mis,obs}} (Y_{\text{obs}} - \mu_{\text{obs}}), \Omega_{\text{mis}}^{-1} \right) \]  

(2.15)

Note that \( \mu_{\text{mis}} = X_{\text{mis}} \beta \) and \( \mu_{\text{obs}} = X_{\text{obs}} \beta \) are obtained from partitioning the design matrix of predictors into missing \( X_{\text{mis}} \) and observed values \( X_{\text{obs}} \).

Updating the missing values from expression (2.15) in the Gibbs sampler requires reordering the data at each step to properly identify the matrix blocks, and computing the inverse of \( \Sigma_{\text{com}} \), which considerably slows down the sampler. For this reason, and to make the sampler more efficient the Metropolis algorithm (Metropolis et al., 1953; Carlin and Louis, 2011) was used to draw samples from the full conditional of the missing values.

### 2.4.3 Prediction

The value of the Gaussian process \( Y(\cdot) \) at the unobserved point \( \mathbf{x} = (s, t, u_s, u_t) \in D \) was predicted using the conditional expectation given the observations, that is:

\[ \hat{Y}(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x}) | Y(\mathbf{x}_1) = z_1, \ldots, Y(\mathbf{x}_n) = z_n] = \mathbb{E}[Y | Z] \]  

(2.16)

where \( Z = (z_1, \ldots, z_n)^T \) denotes the observed values of process \( Y \) at the points \( \mathbf{x}_1, \ldots, \mathbf{x}_n \in D \). This estimator is optimal in the sense that minimizes the expected squared prediction error (Ferguson, 1967). In general, it is difficult to obtain an analytical expression for predictor (2.16). However, it can be straightforwardly evaluated in closed form using the multivariate conditional Gaussian distribution:

\[ Y|Z \sim N \left( \mu_y + \Sigma_{yz} \Sigma_{zz}^{-1} (Z - \mu_z), \Sigma_{yy} - \Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zy} \right) \]  

(2.17)
where $\mathbf{Y}$ and $\mathbf{Z}$ are the vectors of unobserved and observed values, respectively, with joint multivariate normal distribution given by:

$$
\begin{pmatrix}
\mathbf{Y} \\
\mathbf{Z}
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
\mu_y \\
\mu_z
\end{bmatrix},
\begin{bmatrix}
\Sigma_{yy} & \Sigma_{yz} \\
\Sigma_{zy} & \Sigma_{zz}
\end{bmatrix}
\right)
$$

(2.18)

The model parameters $\beta$ and $\theta$ were estimated using the sample mean of the MCMC chains generated from their posterior distributions. Then, the posterior means $\hat{\beta}$ and $\hat{\theta}$ were evaluated in the quantities of expression (2.17) to obtain the predicted values $\hat{\mathbf{Y}}$ and their estimated covariance matrix $\hat{\Sigma}(\hat{\theta})$.

Predictions of the relative performance of seed sources under different climate scenarios were obtained using the approach above described. The centroids of the counties from where the selections were made were used as the geographic reference for the calculations. Current and future climate variables for the counties in the Southeast were used to obtain predictions for the relative performance (standardized height) of seed sources across the region. The actual predictions were obtained on the centroids of the counties, which were then interpolated to obtain a smooth surface to visualize the output.

### 2.5 Results

#### 2.5.1 Model Comparisons

The proposed formulation offers the flexibility to model the spatial dependence of the data in terms of combinations of geographic and climate variables related to test sites and seed sources. Four models were derived from the proposed framework and compared against each other through cross-validation. The explored models are all special cases of the model given
in Section 2.3 with identical predictors in their mean function but with different covariance functions.

1. The first model (IID) represents the benchmark from which the models were assessed. For this model the observations were assumed to be independent with the same measurement error variance. This non-spatial model is analogous to a Bayesian multiple linear regression (i.e., \( \tau^2 = 0 \)).

2. The second model (GEO) assumes that the correlation between observations is a function of solely geographical variables and does not consider any climate variable in the spatial covariance function (that is, \( \varphi_{u_{ij}} \to \infty \) and \( \varphi_{u_{ij}} \to \infty \) for \( i = 1, \ldots, d \), in equations (2.7) and (2.6)).

3. The third model (CLIM) ignores the geographical variables in the spatial covariance and uses climate variables only (that is, \( \varphi_s \to \infty \) and \( \varphi_t \to \infty \), in equations (2.7) and (2.6)). The climate variables used for this correlation model were minimum temperature and precipitation.

4. The fourth model (GEO/CLIM) is the full model (2.2) that combines both the geographical and the climate variables. So, the correlations between the observed values are expressed in terms of the generalized distances of equations (2.7) and (2.6). The climate variables utilized to define the generalized distances were minimum temperature and precipitation.

The prior distributions were the same for all four models. The vector of regression coefficients \( \beta \) was assigned a relatively non-informative prior, \( \beta \sim N(0, \sigma^2_{\beta} I) \), with \( \sigma^2_{\beta} = 5000 \). We considered different values for \( \sigma^2_{\beta} \) and posteriors were not sensitive to these choices. The parameters for the covariance matrix \( \theta \) were assigned a lognormal prior, \( \log(\theta_i) \sim N(\mu_{\theta_i}, \sigma^2_{\theta_i}) \),
with \( \mu_{\theta_i} = 0 \) and \( \sigma_{\theta_i}^2 = 3 \) except for the smoothness parameters \((\kappa_s, \kappa_l)\), in which case the location parameter was set to \( \log 0.5 \) and the scale parameter to 0.15. These hyper-parameters values were chosen relatively centered around the exponential covariance to simplify the analysis and to avoid identification issues associated with these parameters (Gelfand et al., 2010). The hyper-parameters for the other covariance parameters represent fairly vague priors and posteriors were not very sensitive to these choices.

The models were compared using \( K \)-fold cross-validation. The data were partitioned into \( K \) equally-sized subsamples. Each of the \( K \) subsamples was used exactly once as the validation set for testing the model, while the remaining \( K - 1 \) subsamples were treated as the training data for fitting the model. Then, the average of the mean square errors from the \( K \) folds was used as a measure of the predictive ability of the model. The sample was partitioned in three different ways for conducting the cross-validation analysis. To assess the predictive ability of the model based on unobserved sites, the data were partitioned using the test sites as the folds. This type of cross-validation will be referred as leave-site-out, and is arguably the most relevant given our goal of predicting performance under different climate scenarios for planting sites. In addition, to quantify the capacity of the models to predict the performance of unobserved seed sources, the data were randomly split by grouping the seed sources into 19 folds of 5 sources each. This type of cross-validation will be referred as leave-source-out. Finally, a typical 20-fold cross-validation was carried out by randomly splitting the sample into 20 folds disregarding the site or seed source grouping.

The cross-validation results are presented in Figure 2.3. The box-plots summarize the distribution of the mean square error (MSE) in the logarithmic scale for the explored models and for the three types of partitions of the data between training and validation sets. The leave-site-out cross-validation (Figure 2.3a) and the random split cross-validation (Figure 2.3c) show that the spatial models (GEO, CLIM, and GEO/CLIM) provide a superior predictive ability than
the non-spatial IID model in unobserved locations. On the other hand, the relative predictive powers of the three spatial models are essentially the same. The MSEs from the leave-source-out cross-validation (Figure 2.3b) are very similar to each other for all four models. Although the range of MSE values is smaller for the spatial models, overall these models provide a comparable predictive ability to the standard IID model in the leave-source-out cross-validation.

Based on the cross-validation analysis results, we decided to select the CLIM model for making predictions about the relative performance of seed sources under current and future climate conditions. Although the predictions are similar to the GEO model in the current climate, we believe that the CLIM model is more scientifically justifiable and will thus give more reliable predictions under different climate regimes.

2.5.2 Summarizing the posterior of the selected model

The posterior distributions for the CLIM model parameters are summarized in Table 2.1. The covariates that define the mean were chosen from a pool of variables using the Akaike Information Criterion (Akaike, 1974) in a stepwise algorithm. The selected variables included three climate variables for the planting sites (TMIN<sub>t</sub>, TMAX<sub>t</sub>, PPT<sub>t</sub>), three climate variables for the seed sources (TMIN<sub>s</sub>, TMAX<sub>s</sub>, PPT<sub>s</sub>), and two interactions (TMIN<sub>s</sub>TMIN<sub>t</sub>, TMIN<sub>s</sub>PPT<sub>s</sub>). This set of covariates explained about 30% of the variance in the standardized height using Ordinary Least Squares (Monahan, 2008). For the CLIM model, the error term includes eight parameters, namely the measurement error variance (σ²), the spatial variance (τ²), the site and seed source scale parameters for precipitation and minimum temperature (ϕ<sub>Pt</sub>, ϕ<sub>Ps</sub>, ϕ<sub>Tt</sub>, ϕ<sub>Ts</sub>), and the smoothness parameters (κ<sub>s</sub>, κ<sub>t</sub>). The results from Table 2.1 suggest that climate variables TMIN<sub>t</sub>, TMIN<sub>s</sub>, TMIN<sub>s</sub>TMIN<sub>t</sub>, PPT<sub>s</sub>, and TMIN<sub>s</sub>PPT<sub>s</sub> have a significant effect on the relative performance of loblolly pine
seed sources. Not surprisingly, the effect of yearly average minimum temperature for both site
and origin is fairly important. Previous studies have shown the importance of minimum temper-
ature as predictor of performance for southern pines (Schmidtling, 2001, 2003). Furthermore,
minimum temperature has been used by the United States Department of Agriculture to define
plant hardiness zones\(^2\) to determine which plants are most likely to thrive at a given location
(Farjat et al., 2014).

From Table 2.1 it can be inferred that the spatial process under consideration is important
given the significance and magnitude of the parameters that define the spatial dependence be-
tween observations. The scale parameter for the site precipitation \(\phi_P\) is the exception with
an extremely large posterior mean value that effectively eliminates site precipitation from the
spatial correlation function. The measurement error variance \((\sigma^2)\) is slightly larger than the
variance of the spatial process \((\tau^2)\). The tree measurements were taken in the field by differ-
ent crews, which might lead to measurement errors, however such errors are considered to be
small when considering averages. This supports the interpretation that \(\sigma^2\) might be reflecting
small-scale spatial variability. The posteriors of the scale parameters for the minimum tempera-
ture \((\phi_T, \phi_{Ts})\) indicate a stronger spatial dependence between observations based on minimum
temperature differences for the sites than based on minimum temperature differences for the
seed sources. Finally, the estimated posterior mean of the smoothness parameter for the seed
source \((\kappa_s)\) is very close to 0.5, i.e. the exponential correlation. On the other hand, the posterior
mean of the smoothness parameter for the site \((\kappa_t)\) is greater than 1, suggesting that the process
associated with the site is smoother than the seed source process.

The MCMC samples from the posterior distributions for some of the CLIM model parame-
ters are summarized by the trace plots and histograms in Figure 2.4. In addition, the plot shows
the auto-correlation function (ACF) for the MCMC chains to informally assess the similarity

\(^2\)http://planthardiness.ars.usda.gov/PHZMWeb/
between the samples as a function of the lag between them. From the histograms, it can be noticed that there is definitely Bayesian learning in the marginal posterior distributions relative to their respective priors. Visual inspection of the trace plots reveals that the chains are mixing well, that is successfully moving throughout the parameter space. The Geweke’s diagnostic suggested that convergence of the MCMC chains to the stationary distribution was reached after about 20,000 samples.

2.5.3 Projections and Seed Source Movements

In this section, examples to illustrate the capacity of the proposed model to make predictions about the performance of specific seed sources under current and future climate are presented. In addition, examples of optimal seed source movements to specific sites are presented.

The CLIM model was used to predict the relative performance of seed sources from the natural range of loblolly pine (Figure 2.1b) under different climate scenarios. For each scenario, a two-way table was created with the relative performance of seed sources as function of the county of deployment. The model is used to predict the performance of seed taken from each county in the Southeast and planted at every county in the Southeast. Thus, the location of top performing seed sources for planting at a given county can be identified.

Figure 2.5 shows the predicted relative performance (standardized height) of seed sources from Nassau, FL and Appomattox, VA under current climate. The climate scenario was created from historical data assuming that future climate conditions would be the same as the average of the last 20 years. The prediction locations correspond to the centroids of the counties and then interpolated to help with visualization. The predicted standardized heights indicate that seed sources from Nassau (Figure 2.5a) will have a superior performance in the southern region, specifically in Florida. The spatial trend is clearly evident indicating a decline in performance
in the northern regions. Not surprisingly, the predicted relative performance of seed sources from Appomattox (Figure 2.5b) exhibits a radically different pattern. In this case, the spatial trend indicates that seeds from Appomattox will be superior in the most northern regions with a decreasing performance towards the south. These results are in agreement with previous studies (Wells, 1983; Schmidtling, 2001). The standard deviations (Figure 2.5c) show a pattern of smaller uncertainty on a broad central region that include most of the observations and larger standard deviations in northern and southern areas.

Figure 2.6 shows the predicted relative performance (standardized height) of seed sources planted in Wayne County, NC for two climate scenarios created from historical values. The first scenario assumes that future climate conditions would be the same as the average of the last 20 years (Figure 2.6a). The color map displays the relative performance pattern for seed movement to Wayne, NC. Hence, the location of the top performing seed sources can be identified. The arrows show the direction of the optimal seed source movement based on the top 1% standardized heights and suggest a seed source movement from specific locations from South Carolina and central Georgia. The second scenario assumes a 1°C increase in minimum and maximum temperatures, and a 10% decrease in precipitations relative to historical values (Figure 2.6b). Under this scenario, the southern seed sources from the Atlantic Coastal Plain and central Georgia regions improve their relative performance in Wayne county. Furthermore, one of the arrows indicates an optimal seed source movement from Florida, in addition to movements of seed from South Carolina.

Figure 2.7 shows the optimal seed source movement to southern Virginia and Central South Carolina based on the top 1% performing locations for three future climate scenarios. The first scenario (1) is the most conservative and assumes that future climate conditions will be the same as the average of the last 20 years. The second scenario (2) was created assuming a 1°C increase in minimum and maximum temperatures, and a 10% decrease in precipitation relative
to historical values. The third scenario (3) is the most extreme and was created considering a 2°C increase in minimum and maximum temperatures, and a 20% decrease in precipitations relative to historical values. The arrows indicate the direction of the seed movement and the dots on the map show the location of the top performing seeds. To facilitate visualization, the color of the arrows were chosen depending on the state of origin.

The pattern of optimal seed source movement to Virginia (Figure 2.7a) under the most conservative scenario indicates that the Piedmont region of NC, SC, and GA as well as the VA region contain the top performing seed sources. The plot suggests that seed sources from the north of Georgia are suitable genetic material for deployment to a broad region in Virginia. Also, movements of seed from the SC Piedmont area to the Coastal Plains of Virginia are indicated. The intermediate climate scenario highlights nearly the same top performing locations but with differences in the deployment range. For this scenario, the seed sources from the Piedmont areas of Georgia and South Carolina exhibit great performance throughout Virginia. In the most extreme climate scenario, the seed movement is predominately from the Piedmont areas of South Carolina and Georgia. However, movements from the Coastal Plain of Georgia and Central Florida to the Coast of Virginia are indicated as optimal.

The arrows that define the optimal seed source movement to South Carolina (Figure 2.7b) indicates an important movement of seed from Florida to the Coastal Plains of South Carolina under the most conservative climate scenario. In addition, considerable seed movements from within the same region can be observed, which suggests the superior relative performance of SC seed sources in their local environment. The plot also shows optimal seed movements from central Georgia and the Gulf region to the Piedmont area of South Carolina. The pattern of the intermediate climate scenario suggests and increase in the range of the deployment of Florida seed sources to the Coastal Plains region of South Carolina and a reduction of seed movement from within the same region. In the most extreme climate scenario, the optimal seed movement
is predominately from Florida to both the Coastal Plain and the Piedmont regions of South Carolina. Although a considerable reduction of seed source movements from within the same region is observed, optimal movements from and to the Piedmont region of South Carolina are indicated.

2.6 Discussions

The proposed Bayesian spatial model exploits the spatial dependence between planting sites and seed sources from a provenance test to produce more accurate predictions than multiple linear regression analysis. The spatial dependence in the data is accounted for by fitting a non-parametric response surface and the separable Matérn covariance structure provides a flexible means to describe and estimate effects associated with the source of the seed and planting site locations.

The results indicate that the spatial dependence between observations cannot be ignored, and therefore predictions about the performance of different seed sources under future climate change scenarios can be considerably improved relative to traditional methods that constrain the response surface to a parametric linear function.

The proposed approach would be useful to help forest landowners to make an informed choice of seed sources based on their expectations and risk tolerance. Pine plantations can be seen as a portfolio made of different pine families or seed sources designed to meet some specific goals constrained to the available resources. Including seed sources adapted to future climate conditions in pine plantations would buffer the risks associated to climate change and its potential negative impacts.

The model uses Bayesian analysis because it provides a convenient and flexible setting to estimate the model parameters and to deal with the problem of the missing data. MCMC meth
ods make computation tractable for a wide range of parametric models and inferences about the model parameters are exact and conditional on the data, independently of the sample size and without using asymptotic approximations. So, the analysis provides more interpretable results about the uncertainty of the parameter estimates because Bayesian confidence intervals (credible sets) enable us to make probability statements about the likelihood of the parameter under consideration to fall into a given interval. We acknowledge that implementing a Bayesian analysis requires a higher computational cost compared to classical methods, however we believe that the flexibility and better predictive ability of the resulting models outweigh the computational burden.

The proposed model assumes that the environmental and genetic effects are accounted for by climate variables associated with the planting site and seed sources, respectively. However, among and within seed source variation could be affected by other factors related to the composition and evolution of the population under study. For instance, considerable genetic differences created from past events such as migration or gene flow can result in spatial patterns that are independent of climate.

The model was developed under the assumption that phenotypic variation of pine trees occurs geographically, which was documented by different studies of loblolly pine (Wells, 1983; Chamblee, 2011). The second important assumption is that geographic locations can be characterized by a set of climate variables, as a result, climate variables are used as predictors to identify optimal locations for planting. One of the caveats of identifying seed sources solely by a set of climate variables is that the predicted relative response of genetically different seed sources coming from the same location and planted in the same place will be identical. That is, the model gives the average relative performance of genetically diverse groups based on seed source similarities.

Future work could be aimed to extend the proposed model to account for the existing ge-
netic differences within seed sources from a given location. Specifically, modeling individual pine family performance is important, because most commercial plantations consist of only a few families. This extension will help to tackle the challenging task of identifying genetic adaptations within loblolly pine that will best allow trees to survive and adapt to future climates.

2.7 Conclusions

In this chapter we have developed a Bayesian spatial model to use data from a provenance test to predict the performance of different seed sources under climate change scenarios. We demonstrated that there is strong spatial dependence between both sites and seed sources, and that exploiting this fact results in better predictions than multiple linear regression analysis. We used the predictive model to map optimal seed source deployment, and develop graphical tools for visualization. The proposed statistical model can be used as a quantitative tool for designing forest management strategies oriented to mitigate the negative impacts of climate change. The model can be used as an interactive migration tool aimed to identify relatively superior seed sources within the species distribution that will best allow trees to thrive under future climate conditions.
### 2.8 Tables

Table 2.1: Posterior summaries for the parameters of CLIM model. The mean coefficients ($\beta$) are represented by capital letters for their corresponding covariates, and the covariance parameters ($\theta$) are denoted with greek letters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MEAN</th>
<th>SD</th>
<th>PERCENTILE</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>-0.007</td>
<td>0.056</td>
<td>-0.117</td>
<td>-0.007</td>
<td>0.102</td>
<td></td>
</tr>
<tr>
<td>TMIN$_t$</td>
<td>0.163</td>
<td>0.039</td>
<td>0.086</td>
<td>0.164</td>
<td>0.240</td>
<td></td>
</tr>
<tr>
<td>TMIN$_s$</td>
<td>2.413</td>
<td>0.563</td>
<td>1.307</td>
<td>2.415</td>
<td>3.514</td>
<td></td>
</tr>
<tr>
<td>TMIN$_s$ TMIN$_t$</td>
<td>0.260</td>
<td>0.033</td>
<td>0.195</td>
<td>0.260</td>
<td>0.325</td>
<td></td>
</tr>
<tr>
<td>TMAX$_t$</td>
<td>0.011</td>
<td>0.029</td>
<td>-0.046</td>
<td>0.011</td>
<td>0.069</td>
<td></td>
</tr>
<tr>
<td>TMAX$_s$</td>
<td>0.124</td>
<td>0.087</td>
<td>-0.048</td>
<td>0.124</td>
<td>0.294</td>
<td></td>
</tr>
<tr>
<td>PPT$_t$</td>
<td>0.000</td>
<td>0.020</td>
<td>-0.038</td>
<td>0.000</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td>PPT$_s$</td>
<td>-0.278</td>
<td>0.076</td>
<td>-0.427</td>
<td>-0.278</td>
<td>-0.130</td>
<td></td>
</tr>
<tr>
<td>TMIN$_s$ PPT$_s$</td>
<td>-2.056</td>
<td>0.542</td>
<td>-3.116</td>
<td>-2.058</td>
<td>-0.990</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.410</td>
<td>0.018</td>
<td>0.377</td>
<td>0.410</td>
<td>0.446</td>
<td></td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.296</td>
<td>0.044</td>
<td>0.221</td>
<td>0.293</td>
<td>0.392</td>
<td></td>
</tr>
<tr>
<td>$\varphi P_t$</td>
<td>2481</td>
<td>2122</td>
<td>679</td>
<td>1884</td>
<td>7845</td>
<td></td>
</tr>
<tr>
<td>$\varphi P_s$</td>
<td>0.690</td>
<td>2.270</td>
<td>0.017</td>
<td>0.254</td>
<td>3.341</td>
<td></td>
</tr>
<tr>
<td>$\varphi T_t$</td>
<td>6.508</td>
<td>2.774</td>
<td>3.071</td>
<td>5.862</td>
<td>13.741</td>
<td></td>
</tr>
<tr>
<td>$\varphi T_s$</td>
<td>2.147</td>
<td>12.403</td>
<td>0.020</td>
<td>0.324</td>
<td>14.916</td>
<td></td>
</tr>
<tr>
<td>$\kappa_t$</td>
<td>1.074</td>
<td>0.348</td>
<td>0.552</td>
<td>1.022</td>
<td>1.888</td>
<td></td>
</tr>
<tr>
<td>$\kappa_s$</td>
<td>0.499</td>
<td>0.197</td>
<td>0.215</td>
<td>0.467</td>
<td>0.973</td>
<td></td>
</tr>
</tbody>
</table>
2.9 Figures

(a) Geographic regions and seed sources

(b) Test sites and natural range of loblolly Pine

Figure 2.1: Geographic regions of loblolly pine grouped by counties, location from where the seed sources were collected (black points), and test site locations (red triangles). The geographic regions are Virginia (VA), North Carolina Coastal Plain (NC), South Carolina Coastal Plain (SC), Georgia-Florida Coastal Plain (GF), Lower Gulf Coastal Plain (LG), Upper Gulf Coastal Plain (UG), and Piedmont (PD).
Figure 2.2: Plot to illustrate the underlying concept of the proposed model. The lines represent different climate scenarios, and the shading of dots denotes the expected response. The expected response of the underlying process can be characterized for future climate scenarios from current climate.
Figure 2.3: Boxplot of mean square errors (logarithmic scale) from cross-validation analyses for the models under consideration: a) leave-site-out, b) leave-source-out, and c) random split. The thick horizontal line inside each box represents the median and the black dot the mean. The bottom and top horizontal lines that define the box represent the first (25%) and third quantile (75%), respectively. The bottom and top whiskers extend to the minimum and maximum data value, respectively. The IID model is a non-spatial Bayesian multiple regression, GEO is a spatial model that uses solely geographical variables to define correlations; CLIM is a spatial model that uses climate variables only to define correlations, and GEO/CLIM model uses both geographical and climate variables to define correlations among observed values.
Figure 2.4: Trace plots, posterior histogram with approximated density superimposed, and auto-correlation function for the MCMC chains of $\tau^2$ (spatial variance), $\varphi_{T_t}$ (temp-site scale), and $\kappa_s$ (seed source smoothness) for CLIM model. There are 120,000 samples with a burn-in period of 20,000.
Figure 2.5: Predicted performance (standardized height) for seed sources from a) Nassau, FL (black point) and b) Appomattox, VA assuming that future climate conditions would be the same as the average of the last 20 years. The first colormap indicates that seeds from Nassau county will have a superior performance in the southern region. In contrast, the predicted standardize heights of seeds from Appomattox suggest a relative superior performance in the northern and colder regions of the Southeast. The standard deviation is plotted in panel c) and is the same for both predictions.
Figure 2.6: Predicted performance in Wayne County, NC as a function of seed source for two climate scenarios based on historical values: a) future climate conditions would be the same as the average of the last 20 years b) 1 °C increase in minimum and maximum temperatures, and a 10% decrease in precipitations. The blue color indicates the location of top performing seeds. The arrows show the direction of the optimal seed source movement based on the top 1% standardized height.
Figure 2.7: The arrows show the direction of the optimal seed source movement to Southern Virginia (top) and Central South Carolina (bottom) based on the top 1% locations (dots) assuming three scenarios: 1) future climate conditions would be the same as the average of the last 20 years, 2) a 1°C increase in minimum and maximum temperatures, and a 10% decrease in precipitations relative to historical values, and 3) a 2°C increase in minimum and maximum temperatures, and a 20% decrease in precipitations relative to historical values. The colors correspond to the state of origin.
Chapter 3

Prediction of Genetic Merit in a Clonal Population of Loblolly Pine

3.1 Introduction

Progeny tests are commonly used in tree breeding programs to identify superior genotypes in specific environments or across environments. The tests are generally replicated in different environments to assess the magnitude of genotype-by-environment (G×E) interactions. These types of experiments are generally referred to as multi-environmental trials (METs). For forest trees, MET typically require large tracts of land in each environment, particularly for large number of genetic entries. Field trials are characterized by their environmental variability (Costa e Silva et al., 2001). The data collected from field trials are noisy and typically unbalanced due to missing observations (Ogut et al., 2014). Therefore, it is important to account for multiple sources of variation for predicting reliably genetic merit of individuals for ranking.

Linear mixed models have been intensively studied (McCulloch et al., 2008) and have become a popular technique for the analysis of MET data. Linear mixed models (Robinson,
1991) are particularly useful in settings where repeated measurements are made on the same statistical units, or where measurements are made on clusters of related statistical units. The advantages of linear mixed models over traditional approaches, such as ANOVA, include the simplicity to accommodate incomplete data, the flexibility to fit different error models, and the ability to consider random as well as fixed effects in the analysis. For a comprehensive overview of several linear mixed model approaches used in plant breeding refer to Smith et al. (2005).

Accounting for the heterogeneity in multi-environmental trials data is crucial to correctly estimate variance components, predict breeding values, and make sound selection decisions in forest tree breeding programs (Ogut et al., 2014). For the analysis of MET data, many covariance structures have been proposed to account for heterogeneity and to model G×E interactions. A simple variance-covariance structure may assume that site-specific variances are homogeneous and genotypes are independent across environments. These assumptions are not realistic because they ignore the relationship among genotypes as well as the spatial correlations between and within plots. For instance, Smith et al. (2001) and Kelly et al. (2009) used spatial analysis techniques to model the residuals between and within plots along with the factor analytic model to parsimoniously approximate a fully unstructured form of the genetic variance matrix in the context of METs.

Historically, tree breeding programs have tested full and half-sib progenies produced from different mating designs. However, the recent advances in vegetative propagation techniques such as rooted cuttings and somatic embryogenesis (Weng et al., 2010) have made possible the production of large number of clones for commercial deployment. Clonal forestry has been proposed as an alternative to conventional seed orchards to produce seeds for planting-stock production in the southeastern United States (Stelzer and Goldfarb, 1997). Clonal replication offers many advantages (Carson, 1986), together with the exceptional opportunity to propagate
tested desirable performers and the possibility to increase forest productivity over conventional strategies (Park, 2002). For instance, Isik et al. (2003) suggest that fusiform rust infection can be considerably reduced by deployment of selected clones. In addition, Isik et al. (2005) suggest that cloned progeny testing provides more reliable predictions of individual-tree breeding values than seedling progeny testing. Furthermore, the authors conclude that significant gains can be achieved with the implementation of clonal forestry in the southeastern United States.

The statistical analysis of METs data from clonal genetic tests considers repeated measures of the same genotype within the same trial and across different trials. In the presence of G×E interaction effects, clonally replicated genetic tests may considerably increase the precision of selections relative to seedling genetic tests since the performance of the clone is assessed across different environments. The same concept applies over time since measurements of the same ramet can be taken at different moments over its lifespan (Zamudio et al., 2008).

MET analyses usually require using heterogeneous variance-covariance structures for both the residuals and the genetic effects across environments for modeling G×E interactions. As a result, multiple predictions of genetic entries (one for each environment) are obtained. The challenge then is how to optimally weigh multiple predictions for the same genotype to obtain an aggregate genetic merit for each clone in the study. Prediction methods based on linear combination of effects and weighted averages across trials were suggested to tackle this problem (Gilmour et al., 2004; Welham et al., 2004; Kelly et al., 2007).

The objective of this study is to explore approaches that account for the multiple sources of variation present in the context of clonal tree breeding progeny test data to obtain accurate and reliable estimates of breeding values for selection rankings. The methods are developed and tested using clonal test data of the Atlantic Coastal Elite (ACE) breeding population of loblolly pine (Pinus taeda L.) from the North Carolina State University Cooperative Tree Improvement Program. The analyzed data set is very rare involving a large number of clones replicated
across various environments. We are not aware of any studies addressing the statistical analysis of such large clonal trials in forestry breeding. The method we introduce in this work may help others to analyze and interpret more efficiently this type of data.

3.2 Materials and Methods

3.2.1 Genetic material and field design

The initial genetic material of the Atlantic Coastal Elite population consisted of 76 crosses produced from 24 elite parents from the Atlantic Coastal Plain region of loblolly pine. Fusiform rust (caused by *Cronartium quercuum f.sp. fusiforme*) disease represents a threat to pine plantations in most of the Coastal Plain region. In order to make selection for disease resistance, seedlings were first screened at the USDA Resistance Screening Center (RSC) in Asheville, NC with a broad-base field inoculum that covered the expected deployment range (see Cowling and Young (2013) for details about the RSC). After six months, the seedlings’ response to the inoculation was evaluated through the presence or absence of rust galls caused by rust fungus. Seedlings with rust galls were deemed susceptible to rust and hence were eliminated from the study. The underlying assumption was that seedlings with a ”no gall” response would be more likely to be resistant in the field than seedlings with galls. Through this selection process, genotypes that may have been susceptible in the field to rust were eliminated, and the remaining seedlings were planted.

Based on the inoculation results, 51 out of the initial 76 crosses were selected for progeny testing in the field trials. From each full-sib family, an average of 46 progeny were cloned using rooted cutting techniques. A total of 2362 clones were tested in the field trials, and a single clone was represented at each test. Alpha-cyclic incomplete block row-column design
was used to accommodate large number of clones in the study (Williams and John, 1996). The
design allows controlling the trends at the location in two directions (rows and columns) to
better control the environmental noise (Williams et al., 2006). The study was established on
eight sites across the southeastern United States; data from seven of these sites were available
for this analysis (Figure 3.1, Table 3.1).

In order to compare the performance of the selected crosses, seedlings from seven coastal
common families were used as control or checks. These check families were randomly dis-
tributed throughout each test site. Growth (height and diameter at breast height), stem straight-
ness, fusiform rust disease incidence, forking, ramicorn branching, and survival were assessed
at age four years. In total, over 15,000 trees were measured. Although several traits were
measured, the primary trait under consideration in this work is height.

3.2.2 Statistical methods

Linear mixed models were used to estimate variance components and predict genetic merit
of cloned individuals. The general form of the linear mixed model can be written in matrix
notation as follows:

$$y = X\beta + Zu + e$$  \hspace{1cm} (3.1)

where $y$ denotes the $n \times 1$ vector of observations, $\beta$ is the $p \times 1$ vector of fixed effects, $X$ is a
$n \times p$ design matrix with covariates associated with the fixed effects, $u$ is the $q \times 1$ vector of
random effects, $Z$ is the $n \times q$ design matrix of covariates associated with the random effects,
and $e$ is the $n \times 1$ random vector of residuals. It is assumed that the random effects $u$ and
$e$ are mutually independent and jointly distributed multivariate normal with zero mean, and
covariance matrices given by: $\text{var}(u) = G(\gamma)$ and $\text{var}(e) = R(\phi)$, where it is emphasized that
these matrices are functions of the vectors of variance components $\gamma$ and $\phi$ associated with $u$. 

78
and \( e \), respectively. It follows that the response vector \( y \) has multivariate normal distribution, with mean \( \mathbb{E}[y] = X\beta \), and variance \( \text{var}(y) = V = ZG(\gamma)Z^T + R(\varphi) \). This will be denoted henceforth as: \( y \sim N(X\beta, V) \).

The Best Linear Unbiased Estimator (BLUE) for the fixed effects is given by:

\[
\hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}Y
\]  

which is also known as the generalized least squares estimator (Monahan, 2008). The Best Linear Unbiased Predictor (BLUP) for the vector of random effects \( u \) is defined as (Robinson, 1991):

\[
\tilde{u} = GZ^TV^{-1}(y - X\hat{\beta})
\]  

which is equivalent to the conditional expectation of \( u \) given \( y \) \( (\mathbb{E}[u|y]) \) under the normality assumption. Expressions 3.2 and 3.3 involve the computation of the inverse of the variance-covariance matrix \( V \), which could result in an intricate and time consuming process. Henderson (1975) derived the so-called mixed-model equations (MME) for jointly obtaining \( \hat{\beta} \) and \( \tilde{u} \) while avoiding the computation of \( V^{-1} \):

\[
\begin{bmatrix}
X^TR^{-1}X & X^TR^{-1}Z \\
Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
\tilde{u}
\end{bmatrix}
= \begin{bmatrix}
X^TR^{-1}y \\
Z^TR^{-1}y
\end{bmatrix}.
\]  

The advantage of this approach relies on the fact that the computation of the BLUE and BLUP is reduced to the solution of the system of equations 3.4. Furthermore, the mixed model equations depend on the inverses of matrices \( G \) and \( R \), which are easier to compute than \( V^{-1} \).

The mixed-model equations can be succinctly written as \( C\tilde{\xi} = W^TR^{-1}y \), where \( C \) is the coefficient matrix of the linear system of equations 3.4, \( \tilde{\xi} = (\hat{\beta}^T, \tilde{u}^T)^T \) is a vector that contains
the BLUE and the BLUP, and \( W = (X, Z) \) is the \( n \) by \( p + q \) matrix formed by the design matrices for the fixed and random effects, respectively. Confidence intervals are usually constructed using the prediction error variance, i.e. \( \text{var}(\hat{\xi} - \xi) \), instead of the variance of the estimator \( \hat{\xi} \) (Welham et al., 2004):

\[
\text{var}(\hat{\xi} - \xi) = C^{-1}
\]

(3.5)

Therefore, the prediction error variance of linear combinations of estimates, i.e. \( \lambda^T \hat{\xi} \) can be written as \( \lambda^T C^{-1} \lambda \).

The elements of the vector of variance components \( \theta = (\gamma^T, \varphi^T)^T \) that defines the variance-covariance matrix \( V \) are generally unknown. In practice they need to be estimated from the data. The most popular method for variance component estimation in linear mixed models is Restricted Maximum Likelihood (REML) (Patterson and Thompson, 1971).

The maximization of the restricted log-likelihood with respect to the vector of variance parameters \( \theta \), requires numerical solution through an iterative algorithm. The Average Information algorithm (Gilmour et al., 1995) is a second-order scheme that stands out in the context of variance components estimation for linear mixed models. This method can be considered as an extension of the Fisher-Scoring algorithm in which the expected information matrix is replaced in the updating formula for the variance parameters by an approximate average of the observed and expected information matrices (Gilmour et al., 2004).

### 3.2.3 Individual trial analysis

A separate analysis of individual trials was conducted to estimate site-specific variance components. The following linear mixed model was fit at each site:

\[
y = X\beta + Zu + e
\]

(3.6)
where \( \mathbf{y} \) denotes the \( n \times 1 \) response vector of measured heights, \( \mathbf{\beta} \) is the \( 3 \times 1 \) vector of fixed effects (overall site mean, row effect, column effect), \( \mathbf{X} \) represents the \( n \times 3 \) design matrix associated with the fixed effects, \( \mathbf{Z} \) is the incidence matrix for the observed clones, \( \mathbf{u} \) is the vector of random clone effects, and \( \mathbf{e} \) is the random error term. It is assumed that \( \mathbf{u} \) and \( \mathbf{e} \) are mutually independent and multivariate normally distributed with zero means. The variance matrix of clone effects is given by \( \mathbf{G} = \text{var}(\mathbf{u}) = \sigma_g^2 \mathbf{A} \), where \( \sigma_g^2 \) is the variance component associated with the clone genetics effect; matrix \( \mathbf{A} \) is the numerator relationship matrix (Falconer and Mackay, 1996) derived from the pedigree of analyzed individuals based on the probability of identical by descent gene pairs (Lynch and Walsh, 1998). The vector of residuals was decomposed into a spatially dependent process and an independent Gaussian process with zero mean (Smith et al., 2001). So, the variance of the error can be written in matrix notation as the sum of two contributions (Costa e Silva et al., 2001):

\[
\mathbf{R} = \text{var}(\mathbf{e}) = \tau^2 \mathbf{\Gamma} + \sigma^2 \mathbf{I}
\]

(3.7)

The first term (\( \tau^2 \mathbf{\Gamma} \)) in the above equation represents the variance of the spatial process, where \( \mathbf{\Gamma} \) is the spatial correlation matrix with associated spatial variance parameter \( \tau^2 \), usually referred as the partial sill in spatial statistics. The variance of the independent Gaussian process is captured by parameter \( \sigma^2 \) in the second term, which represents the measurement error, usually referred as the nugget effect (Gelfand et al., 2010).

The spatial analysis is focused on modeling the spatial patterns in the site. Thus, observations must be associated with some measure of location. In our data, the positions of the measured trees within a trial are tracked by the column and row number of a rectangular grid. The spatial process is generally assumed to depend only on the distance between the observations (spatial lag). For the analysis, it was assumed a separable two-dimensional autoregressive
process of the form (Cullis and Gleeson, 1991):

\[ \Gamma = \Gamma_r(\rho_r) \otimes \Gamma_c(\rho_c) \]  

(3.8)

where \( \Gamma_r(\rho_r) \) and \( \Gamma_c(\rho_c) \) are autoregressive correlation matrices of first order for rows and columns, respectively. Parameters \( \rho_r \) and \( \rho_c \) are the associated autocorrelation coefficients for rows and columns, respectively. There are many possible forms for the spatial correlation matrix \( \Gamma \) (Stefanova et al., 2009), however separable autoregressive processes of order one (AR1 \( \times \) AR1) are generally used because of their simplicity and adequate fit of spatial trends (Costa e Silva et al., 2001).

The errors of the fitted spatial models were assessed through their sample variograms (Gilmour et al., 1997). The sample variogram is a diagnostic tool commonly used in spatial statistics for describing the spatial correlation among observations.

The ratio of the clone genetic variance to the sum of the measurement error and genetic variance \( r = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e} \) was used to define the set of weights for the sites. Defined this way, the ratio of variance components provide a relative measure of the quality of each site to make inferences about genetic merit since it quantifies the strength of the genetic variation relative to the measurement error variation. The ratio defined above did not include the estimated spatial variance \( \tau^2 \) because the objective was to quantify the proportion of genetic variance embedded in the measurement error while explicitly excluding the variance associated with the spatial pattern.

The estimated site specific ratios \( r_i \) for \( i = 1, \ldots, t \) were used to define a weight factor to calculate the aggregate clone breeding values in the multi-environmental trial analysis. Let \( \hat{U}_i \) be the vector of predicted genetic merit for the clones in the \( i \)th environment, then the aggregate
vector of predictions $\tilde{U}$ was defined as a weighted average:

$$\tilde{U} = \sum_{i=1}^{t} w_i \hat{U}_i$$  \hspace{1cm} (3.9)

where $w_i = \frac{r_i}{\sum_{i=1}^{t} r_i}$ for $i = 1, \ldots, t$ are the weights derived from the ratio of variances at each site.

### 3.2.4 Multi-environment trial analysis

For the overall analysis, the data set consists of a total of $m$ genotypes grown at $t$ sites. However, typically the data are unbalanced because not all genotypes are present in all tests. Let $y = (y_1^T, \ldots, y_t^T)^T$ be the $n \times 1$ combined data vector of observations from $t$ trials, and assume that the $i$th trial comprises $n_i$ observational units. So, the total number of observation is $n = \sum_{i=1}^{t} n_i$. The linear mixed model used for the analysis can be written using matrix notation as follows:

$$y = X\beta + Zu + e$$ \hspace{1cm} (3.10)

where $\beta = (\beta_1^T, \ldots, \beta_t^T)^T$ is the vector of fixed effect with associated design matrix $X$; $u = (u_1^T, \ldots, u_t^T)$ is the $mt \times 1$ vector of random clone effects (ordered as clones within environments) with associated design matrix $Z$; and $e = (e_1^T, \ldots, e_t^T)^T$ denotes the vector of random residuals. This framework allows accounting for the uncertainty of peripheral sources of variation by including a vector of random non-genetic (peripheral) effects, but it was omitted from our analysis.

The vector of fixed effects included test site effects in addition to a row and column effect for each site. So, $\beta$ was a $3t \times 1$ vector where $\beta_i = (\mu_i, \text{row}_i, \text{col}_i)^T$ for $i = 1, \ldots, t$ denotes the vector of fixed effect for the $i$th site. The row and column effects at each site were included.
in the analysis to account for the presence of any linear spatial trend in the rectangular grid.

The vectors of random effects \( \mathbf{u} \) and \( \mathbf{e} \) are assumed to be mutually independent with joint multivariate normal distribution given by:

\[
\begin{bmatrix}
\mathbf{u} \\
\mathbf{e}
\end{bmatrix}
\sim N\left(
\begin{bmatrix}
0 \\
0
\end{bmatrix},
\begin{bmatrix}
\mathbf{G} & 0 \\
0 & \mathbf{R}
\end{bmatrix}
\right)
\] (3.11)

We fit various variance-covariance structures to account for heterogeneity in genetic and residual variances. The mathematical forms of the structures used to fit the linear mixed models are given in Table 3.2. The covariance matrix for the residuals vector \( \mathbf{e} \) can take many possible forms, though it is standard to assume a diagonal covariance of the form \( \mathbf{R} = \text{var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n \), where \( \sigma^2 \) represents the variance component associated with the residuals, and \( \mathbf{I}_n \) is the \( n \)-dimensional identity matrix. This form is generally utilized because of its simplicity and ease to fit but could fail to account for the presence of heteroskedasticity (non-constant variance) across trials. In such a case, a block diagonal variance-covariance matrix would provide a better fit at the expense of fitting one variance component for each site. The diagonally blocked covariance structure can be written in matrix notation as: \( \mathbf{R} = \bigoplus_{i=1}^{t} \sigma_i^2 \mathbf{I}_{n_i} \), where the direct sum operator \( \bigoplus \) has been used. Both the diagonal and block diagonal structures were considered in the fitted models. In addition, the residuals were modeled using a spatial approach where the spatial patterns at each site were accounted with a separable spatial correlation. Under this approach, the covariance matrix of the residuals was given by: \( \mathbf{R} = \bigoplus_{i=1}^{t} \tau_i^2 \mathbf{\Gamma}_i \), where \( \tau_i^2 \) and \( \mathbf{\Gamma}_i \) denote the spatial variance and spatial correlation matrix for the \( i \)th site, respectively. The spatial correlation associated with the \( i \)th site can be written as the Kronecker product of two
spatial correlations (Costa e Silva et al., 2001):

\[ \Gamma_i = \Gamma_{r_i}(\rho_{r_i}) \otimes \Gamma_{c_i}(\rho_{c_i}) \]  

(3.12)

where \( \Gamma_{r_i} \) and \( \Gamma_{c_i} \) are autoregressive correlation matrices for the rows and columns of the \( i \)th site, respectively. Parameters \( \rho_{r_i} \) and \( \rho_{c_i} \) are the associated autocorrelation coefficients for the rows and columns of the \( i \)th site, respectively.

The vector of random clone genetic effects was modeled as the sum of two contributions:

\[ u = u_g + u_\varepsilon \]  

(3.13)

where the first term in the right hand side represents the vector of clone by environment genetic effects, and the second term is the vector of residual genetic effects. It is assumed that both contributions are mutually independent, jointly distributed as multivariate normal with zero mean, and variances given by:

\[ \text{var}(u_g) = G_g \otimes A \]  

(3.14)

\[ \text{var}(u_\varepsilon) = \sigma_\varepsilon^2 I_n \]  

(3.15)

where \( G_g \) is a \( t \times t \) symmetric, non-negative definite matrix of genetic effects between environments, \( A \) is the numerator relationship matrix derived from the pedigree (Lynch and Walsh, 1998), and \( \sigma_\varepsilon^2 \) denotes the variance component associated with the residual genetic effects. This approach produces meaningful predictions for the genetic effects while accounting for genotype by environment interactions (Smith et al., 2001; Cullis et al., 2010; Beeck et al., 2010).
3.2.5 Genetic correlation structures

In order to account for the genetic correlations across and within environments, the elements of matrix $G_g$ need to be estimated. Therefore, it is important to consider several forms for $G_g$. The diagonal structure, that is $G_g = \bigoplus_{i=1}^{t} \sigma^2_{g_i}$, where $\sigma^2_{g_i}$ is the genetic variance for the $i$th environment, is a simple form that resembles a separate analysis of individual trials. However, the diagonal form does not consider genotype by environment interactions because clone effects are assumed to be independent between environments.

The simplest representation for modeling the variance structure that accommodates $G \times E$ interaction effects is the compound symmetric variance model. This structure can be written using matrix notation as: $G_g = \sigma^2_g I_t + \sigma^2_{ge} J_t$, where $\sigma^2_g$ and $\sigma^2_{ge}$ are the variance components for the genetic main effects and $G \times E$ interactions, respectively; $J_t$ is a $t \times t$ matrix in which all elements are one, and $I_t$ is the $t$-dimensional identity matrix. This model implies that all trials have the same genetic variance ($\sigma^2_g + \sigma^2_{ge}$) and all pair of trials have the same genetic covariance ($\sigma^2_{ge}$) and therefore the same genetic correlation. Although this model is easy to fit, it rarely provides a satisfactory fit, because it is very restrictive (Smith et al., 2015).

The most general variance-covariance structure for $G_g$ is the unstructured form, which accounts for the variance at each trial as well as for all the possible covariances between trials. This structure contains $t(t + 1)/2$ parameters and may be difficult to fit even for moderate number of trials because the number of estimates growths quadratically in $t$. As the number of trials increases, the capacity to reliably estimate the variance components and the ability to fit the model are compromised. Therefore, more parsimonious models are often preferred over the unstructured form for the analysis of MET data.

The factor analytic (FA) variance model (Smith et al., 2001) uses the multivariate technique of factor analysis (Johnson and Wichern, 2007) to approximate the unstructured variance.
model. Kelly et al. (2009), Cullis et al. (2010), Ogut et al. (2014), and Smith et al. (2015) among others, explored the FA model and concluded that it is the preferred model for analyzing METs data-sets. It is parsimonious, informative, and provides a good approximation to the unstructured variance model. The goal of the FA model is to account for the genetic covariances between environments in terms of a small number of factors.

The factor analytic model of order \( k \) (FA\( k \)) assumes that the vector of random genetic effect can be written in terms of a set of \( k \leq t \) random \( m \times 1 \) hypothetical factors \( f_r \) for \( r = 1, \ldots, k \) whose elements are usually referred as scores. Using matrix notation, this can be written as follows:

\[
\mathbf{u}_g = (\lambda_1 \otimes \mathbf{I}_m)f_1 + \cdots + (\lambda_k \otimes \mathbf{I}_m)f_k + \mathbf{\delta} \tag{3.16}
\]

where \( \lambda_r \) for \( r = 1, \ldots, k \) denotes the \( m \times 1 \) vector of the so-called loadings; and \( \mathbf{\delta} \) is a \( mt \times 1 \) vector of residuals for the model that represent the lack of fit of the regression. The key distinctive feature of the FA\( k \) is that both the covariates and the regression coefficients are unknown, so they have to be estimated from the data. This model can be written more compactly in matrix notation as follows:

\[
\mathbf{u}_g = (\mathbf{\Lambda} \otimes \mathbf{I}_m)\mathbf{f} + \mathbf{\delta} \tag{3.17}
\]

where \( \mathbf{\Lambda} \) is a \( t \times k \) matrix of environmental loadings, \( \mathbf{f} \) is a \( mk \times 1 \) vector of genetic scores, and \( \mathbf{\delta} \) represents the \( mt \times 1 \) vector of residuals. It is assumed that the joint distribution of \( \mathbf{f} \) and \( \mathbf{\delta} \) is the following multivariate normal:

\[
\begin{bmatrix}
\mathbf{f} \\
\mathbf{\delta}
\end{bmatrix} \sim \mathcal{N}
\left(
\begin{bmatrix}
0 \\
0
\end{bmatrix},
\begin{bmatrix}
\mathbf{I}_k \otimes \mathbf{A} & \mathbf{0} \\
\mathbf{0} & \mathbf{\Psi} \otimes \mathbf{A}
\end{bmatrix}
\right)
\tag{3.18}
\]
where $\Psi$ is a $t \times t$ diagonal matrix with the specific variances for each environment (trial). Hence, it can be shown that the variance-covariance matrix for the clone by environment genetic effects is given by:

$$\text{var}(u_g) = (\Lambda \Lambda^T + \Psi) \otimes A$$

(3.19)

Therefore, under the $\text{FA}_k$ variance model, the clone genetic variance between environments is approximated by:

$$G_g = \Lambda \Lambda^T + \Psi$$

(3.20)

The total number of parameters in the above variance model is $(k+1)t - k(k-1)/2$, which is linear in $t$. This represents the main advantage of the $\text{FA}_k$ variance model to approximate the unstructured variance model, in which the number of parameters is quadratic in $t$.

### 3.2.6 Loadings rotation and interpretation

For the $\text{FA}_k$ models, the matrix of environmental loadings $\Lambda = [\lambda_1, \ldots, \lambda_k]$ is not unique when $k > 1$, which makes the interpretation of both the loadings and scores challenging. The models were fit using ASReml-R (Butler et al., 2009) which uses an arbitrary constrained form of $\Lambda$ solely chosen for estimation and computational convenience (Smith et al., 2001). Therefore, a principal component representation was used to obtain a meaningful interpretation of the estimated loadings and scores. This representation corresponds to a rotation in which the rotated factors are orthogonal and capture the directions of maximum variance in the data.

In order to obtain the rotated matrix of loadings $\Lambda^*$, the singular value decomposition of $\Lambda$ was used:

$$\Lambda = UDV^T$$

(3.21)

where $U$ and $V$ are orthogonal matrices of dimension $t \times k$ and $k \times k$, respectively. $D$ is a
diagonal matrix with the $k$ singular values (square root of the eigenvalues of $\Lambda^T\Lambda$) arranged in decreasing order. So, the rotated matrix is given by:

$$\Lambda^* = \Lambda V = UD$$

where $\Lambda^* = [\lambda^*_1, \ldots, \lambda^*_k]$ contains the rotated loadings (Cullis et al., 2010). This principal component rotation means that the first vector accounts for the maximum amount of genetic covariance between environments, the second vector is orthogonal to the first one and accounts for the second largest covariance, and so on until the $k$th vector.

The rotated loadings from the factor analytic model with $k = 2$ were normalized to one and graphically displayed. The graphical display of loadings can be very informative to identify clusters of environments in terms of genetic correlations. The cosine of the angle subjected between the vectors for two environments represents the genetic correlation between these environments under the factor analytic approximation (Smith et al., 2001).

### 3.2.7 Exploration of genotype by environment interactions

The genetic by environment covariance matrix $G_g$ was obtained from the FA$k$ model. This matrix was then converted into a correlation matrix $C_g$ to further investigate the genetic correlations between pairs of environments. The correlation and covariance matrices are connected through $C_g = D_g G_g D_g$, where $D_g = \text{diag}(G_{g11}^{-1/2}, \ldots, G_{gtt}^{-1/2})$ is a diagonal matrix with elements given by the inverse of the square root of the diagonal elements of $G_g$. The information contained in the correlation matrix $C_g$ is potentially important for plant breeders, because it can reveal the clone by environment interaction between pairs of environments. For instance, in cases where the estimated genetic correlation is low, or possibly even negative, then the ranking of the clones between the environments will tend to differ considerably.
The genetic correlation matrix can be written in terms of the rotated loadings as follows:

\[ C_g = D_g (A^* A^{*T} + \Psi) D_g = \Omega \Omega^T + \Phi \] (3.23)

where \( \Omega = D_g A^* \) is the rotated matrix of loadings on the correlation scale and \( \Phi = D_g \Psi D_g \) is a diagonal matrix with \( t \) elements defined as follows:

\[ \phi_i = 1 - \sum_{r=1}^{k} \omega_{ri}^2 \] (3.24)

where \( \omega_{ri}^2 \) represents the proportion of total variance explained by the \( r \)th factor for the \( i \)th environment, and conversely \( \phi_i \) for \( i = 1, \ldots, t \) is the proportion of variance not explained by the underlying factors for the \( i \)th environment.

### 3.2.8 Model fitting and comparison

The explored models were fit using ASReml-R (Butler et al., 2009) which runs under R statistical package (R Core Team, 2014). The algorithm in ASReml-R estimates the variance components in the mixed models using the Restricted Maximum Likelihood (REML) approach (Patterson and Thompson, 1971) through the Average Information algorithm (Gilmour et al., 1995).

The fixed and random effects were estimated through their empirical values, that is the unknown variance parameters were replaced with their REML estimates (Patterson and Thompson, 1971) in the expressions for the best linear unbiased estimator 3.2 and predictor 3.3, respectively, which ignores any uncertainty in the estimation of the variance parameters.

The overall models were assessed relative to each other through the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978).
to take into account both the goodness of fit and model complexity. In addition, the restricted maximum likelihood ratio test (REMLRT) was used to assess nested models. Notice that factor analytic model of $k$th order (FA$k$) is nested within the FA model with $k+1$ factors. Also, the compound symmetric (CS) variance model is nested within the FA1 models, in the sense that the CS is an FA1 model in which the loadings are constrained to be equal, and all the elements of the diagonal matrix of specific variances are equal. Moreover, the diagonal structure (DIA) is also nested within a FA1 model in which the matrix of loadings is zero. The comparison of these nested models involves a null hypothesis in which the parameter of interest is on the boundary, thus the standard results regarding the asymptotic distribution of the REMLRT statistic are not valid. Stram and Lee (Stram and Lee, 1994) proposed an adjustment that can be used in these types of testing situations.

Model predictions for the genetic merit of the clones were compared using Pearson’s product-moment correlations, Spearman’s rank correlations (mathematically equivalent to the product-moment correlation coefficient after converting the raw scores to ranks), within family rank correlations (rank correlation among clones belonging to the same family), mean standard error of predictions, and mean standard error of the difference between all pairs of predictions. The standard error of the predictions for the clone breeding values as well as the standard error of the difference between predictions for all pairs of clones were estimated from expression 3.5. To obtain empirical values, the unknown variance parameters were replaced with their REML estimates (Patterson and Thompson, 1971), which we acknowledge can lead to underestimation of the prediction error variances since the uncertainty in the variance parameters is neglected. These two measures of the uncertainty in the predicted values were used to compare the accuracy of the explored models.

The correlations between models were displayed graphically using the heatmap function from R (R Core Team, 2014). The rows and columns of the resulting correlation matrices were
re-ordered to highlight the clustering among models based on the dendrogram obtained from the *agnes* package in R (Maechler et al., 2012) that uses an agglomerative hierarchical clustering algorithm (Kaufman and Rousseeuw, 2009). For the clustering analysis the dissimilarity matrix was given by $I - C_M$, where $C_M$ denotes the correlation matrix between models, and $I$ the identity matrix.

### 3.2.9 Clone mean repeatability

Each trial contains only one ramet per clone, therefore it is not possible to estimate means and variances of clone means at each site. In order to estimate the repeatability of clone means ($H^2_{cl}$) for each trait, the following expression was used:

$$H^2_{cl} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma^2 + \sigma^2_{\epsilon}}$$ \hspace{1cm} (3.25)

where $\sigma_g^2$ is the genetic variance explained by clone effect, $\sigma^2$ is the variance of the residual genetic effects, $\sigma^2_{\epsilon}$ is the residual variance, $n_c$ is the number of trees per clone per test ($n_c = 1$), and $t$ is the number of test sites ($t = 7$). The variance components in the above expression were estimated from a simple model assuming a single clone effect and a common variance to all sites. Standard errors for repeatabilities were estimated using delta method (Casella and Berger, 2001).

### 3.2.10 Rankings from clonal and seedling tests

In traditional seedling genetic tests, within-family selection is based on the performance of an individual tree in a specific environment, whereas in clonal genetic tests, within-family selection is based on the response of the same genotype across several environments. In order
to compare the rankings produced from clonal and seedling testing, correlation measures were computed between the genetic merit of the clones obtained from each site independently versus the genetic merit of the clones based on the multi-environmental trials analysis. For the comparison we used the Pearson’s, Spearman’s rank, and within family rank correlation values for each site. We assumed that predicted clone values from a single site analysis mimic seedling progeny tests, since each clone had only one copy.

### 3.3 Results and Discussions

#### 3.3.1 Individual trial Analysis

There are height growth differences across sites (Table 3.1). Site 4 (Pulaski, GA) exhibits the lowest growth (3.52 m), whereas site 3 (Greene, AL) has the highest growth (5.10 m). The lowest survival rate occurred in site 4. This site suffered from drought after planting. There is a big disparity in the number of trees measured ranging from 1294 in site 4 to 2410 in site 1. Summary statistics for the other measured traits (diameter at breast height, fusiform rust incidence, steam forking, and volume) are given in the appendix (Table A.1). The overall fusiform rust disease incidence was very low (2%) compared to typically observed values in field trials. This result suggests that the screening at the USDA Resistance Screening Center followed by planting only the disease-free cloned progeny was very effective. Several full-sib families were included in the tests with seedling progeny, and they averaged 17% rust incidence (galls in the steam). The frequency of the observed steam forking and ramicorn averaged 24% across sites with a wide range of 4% at site 1 and 47% at site 3. The overall mean stem volume was 10.2 dm$^3$ ranging from 3.7 dm$^3$ at site 4 to 16.2 dm$^3$ at site 3.

A pedigree based linear mixed model with spatially correlated residuals (autoregressive
of first order) and independent measurement error was fit at each site. Computational details about how the models were fit are given in Appendix A.1. The parameter estimates from the fit models for each trial are presented in Table 3.3. The wide range of the variance components estimates across sites can be attributed to scale effects. The ratios of the estimates to their respective standard errors suggest that the estimates are highly significant, with the exception of the genetic effect of site 4. Note that assuming a normal distribution of these ratios to compute p-values represents a large sample approximation of the actual distribution. The nugget effect (measurement error) is greater than the partial sill (spatial variance) across sites, reflecting the noisy character of forest tree data. The estimated autocorrelation parameters for rows and columns that define the spatial correlation structure are relatively high. These numbers suggest a strong spatial dependence for neighboring trees in both row and column directions. The ratio of genetic variance relative to the sum of measurement error and genetic variance ranged from 0.07 to 0.31. The wide range of this ratio indicates that some sites had more environmental noise than others. Therefore, it is reasonable to use the ratio of genetic variance to measurement error plus genetic variance as a weight factor to obtain aggregate breeding values in a multi-environmental data analysis.

The sample variograms from the individual site analysis (Figure 3.2) indicate the existence of a spatial dependence in the row and column directions for most of the analyzed trials. In most of the sites, the surface has a smooth appearance and increases in the row and column directions reaching a plateau of little variation. These results suggest the absence of both global and extraneous trends (Stringer and Cullis, 2002). Sites 1 through 5 are fairly balanced regarding the number of rows and columns. In contrast, the number of rows in sites 6 and 7 is much bigger than the number of columns (particularly in site 6), resulting in a long and narrow grid. Linear row and column fixed effects were fit at each site (Table A.2). The inclusion of these fixed effects was based on visual inspection of the sample variogram followed by a formal
assessment using conditional F-tests (Kenward and Roger, 1997). Higher order polynomials for the row and column direction were fit, but the shape of the resulting variograms (dictated by the nugget and partial sill) and overall results did not change considerably.

### 3.3.2 Multi-environmental trial analysis

The mathematical form of the structures used to fit the linear mixed models along with the corresponding abbreviations for later reference are listed in Table 3.2. The model abbreviations are used to uniquely identify the model. For instance, if \( G \) is factor analytic of order 1 (FA1) and \( R \) has a block diagonal structure (BD), then the model is identified by FA1-BD, that is the abbreviation of both structures connected by a dash. The model-fit statistics for the explored models are presented in Table 3.4 and the computational details are given in Appendix A.1. All the models were fit with the same mean structure consisting of a site mean and a row and column fixed effects. The difference between models is in their covariance structure of genetic and residual random effects. As a result, throughout this work, models are identified based on the structure of the genetic by environment matrix (\( G_g \)), which includes the clone effect and the clone by site interactions, and the residual (\( R \)) covariance matrix.

Overall the spatial models provided a better fit relative to the non-spatial, and the factor analytic structure showed good performance in approximating the most general form of the variance-covariance matrix with a reduced number of parameters. Among the non-spatial models, the factor analytic of first order (FA1-BD) attained a restricted log-likelihood value similar to the most general form, that is the unstructured (US-BD), with considerably less parameters (22 against 36). In addition, there was a small difference in the restricted log-likelihood values between the factor analytic models (\( \Delta \log(L) = 2 \)). As a result, the FA1-BD produced the lowest AIC, since the FA2 has seven extra parameters. The BIC criterion, which emphasizes
parsimony, favored the CS-BD model with the lowest value among the non-spatial models.

The spatial models produced a much better fit for the data relative to the non-spatial models. This is reflected in the much higher restricted log-likelihood and lower AIC and BIC values. In this context, the FA1-AR1 model produced the lowest AIC value and reached almost the same restricted log-likelihood value as FA2-AR1 ($\Delta \log(L) = 1$), which contains seven parameters more. The CS-AR1 model, the most parsimonious among the spatial models, produced the lowest BIC value. Notice that the parsimony emphasis of the BIC could lead to the choice of models that may under-fit the data. Although the CS models provided a good trade-off between goodness of fit and model complexity as evidenced by their lowest BIC values, the underlying structure implies that all sites have the same genetic variance and all pair of sites have the same covariance. So, if there is interest in assessing genotype by environment interactions, the compound symmetric structure will constrain and oversimplify the analysis in a way that important $G \times E$ effects may be overlooked. On the other hand, Smith et al. (2015) reported that AIC tend to select unnecessarily complex models. Therefore, it is worth emphasizing that information criteria provide guidelines to choose among statistical models, and they should not be taken as formal statistical tests.

FA1-BD model is nested in FA2-BD and produced a better fit based on the restricted maximum likelihood ratio test, since we failed to reject the null hypothesis in favor of the full model ($\text{REMLRT} = 4$ on 7 degrees of freedom, $p = 0.78$). CS-BD can be considered a special case of the FA1-BD model and produced an inferior fit ($\text{REMLRT} = 76$ on 12 degrees of freedom, $p < 0.001$). An analogous behavior was found for the spatial models, the FA1-AR1 model is nested in FA2-AR1 and we failed to reject the null hypothesis in favor of the full model ($\text{REMLRT} = 2$ on 7 degrees of freedom, $p = 0.96$). Also, FA1-AR1 produced a better fit than CS-AR1 ($\text{REMLRT} = 108$ on 12 degrees of freedom, $p < 0.001$).

The standard error of the predicted values (SEP) for the random clone effects exhibited
similar values across models ranging from 0.16 to 0.20 for the first (25%) and third (75%) quantiles, respectively. The mean of the SEP across models ranged from 0.17 to 0.19, and for the median, this range was narrower, from 0.17 to 0.18. The average of the standard error of the difference (SED) between all pairs of predicted values is another measure of uncertainty about the predictions listed in Table 3.4. The SED for the clones were comparable across models with most of the values ranging from 0.23 to 0.27, corresponding to the first and third quantiles, respectively. Therefore, we concluded that overall the explored models performed similarly based on the uncertainty of predictions.

### 3.3.3 Genotype by environment interactions

Genotype by environment interaction (G × E) takes place when the relative genetic merits of individuals vary across environments. Modeling G × E is critical for tree breeders, because it allows them to make informed decisions regarding defining the breeding zones and deployment of genetic material. We found that the FA formulation efficiently modeled the G × E effects for tree height. Specifically, the FA1 and FA2 models provided a good fit for both spatial and non-spatial models. Table 3.5 presents the REML estimates of the rotated loadings ($\hat{\lambda}_1^*$ and $\hat{\lambda}_2^*$) and the percentage of variance explained by the factors for each environment ($\hat{\omega}_1^2$ and $\hat{\omega}_1^2 + \hat{\omega}_2^2$) from model FA2-AR1. The FA2-AR1 model provides a satisfactory fit for most sites with the lowest value of 77% of accounted variance for site number 3. The first latent variable ($\hat{\lambda}_1^*$) explains 79% for the first site, and considerably less for the others, with the lowest value of 48% for site number 6. In addition, the second latent variable ($\hat{\lambda}_2^*$) is primarily a contrast between the first site and the other sites. These results suggest that the first site is not as highly correlated with the other sites. The genetics by environment correlation matrix estimated for the FA2-AR1 model (Table 3.6) confirms that the correlation between site 1 and the other sites is lower than
the correlations among the other sites. The high correlations among sites 2 through 7 suggest a lack of G×E among those environments. Conversely, the relatively lower correlations between site 1 and the others indicate that the rank between these environments considerably changed and hence G×E may be relevant.

The rotated and normalized loadings from the FA2-AR1 model were plotted in the unit circle (Figure 3.3) to graphically display clusters of environments based on genetic correlations. This plot illustrates the strong agreement between sites, except for site 1. The angle between the loading vector associated with site 1 and the other sites is considerably bigger than the angle subtended among the other vectors. The cosine of the angle subtended by the vectors for two environments represents the genetic correlation (Smith et al., 2001). Therefore, the normalized loadings suggest two clear and distinct clusters from the environments; that is site 1 and the other sites.

Table 3.7 presents the REML estimates of the rotated loadings (\(\hat{\lambda}_1^*\) and \(\hat{\lambda}_2^*\)) and the percentage of variance explained by the factors for each environment (\(\hat{\omega}_1^2\) and \(\hat{\omega}_1^2 + \hat{\omega}_2^2\)) from model FA2-BD. The non-spatial FA2-BD model provides an overall satisfactory fit for most sites with the lowest value of 82% of accounted variance for site 5. In contrast to the results obtained from the spatial FA2-AR1 model, these results suggest a lack of G×E interaction effects, which is noticeable for the relatively high proportion of genetic variance explained by the nested FA1 model. In this case, the first latent variable (\(\hat{\lambda}_1^*\)) explains around 90% of the variance for all sites. The second latent variable (\(\hat{\lambda}_2^*\)) basically contrasts site 1 with sites 2 and 4, and fully explain the variance of those sites. These results suggest that sites are overall well correlated among each other with differences between site 1 and sites 2 and 4. The genetics by environment correlation matrix estimated from FA2-BD model (Table 3.8) confirms the overall strong correlation among sites since all correlation pairs are greater than 0.8, which supports the assumption of absence of relevant genotype by environment interactions.
The rotated and normalized loadings from the FA2-BD model are graphically illustrated in Figure 3.4. The plot shows a strong overall agreement between sites in terms of their genetic correlations considering the small angle subtended among the vectors. However, these set of normalized rotated loadings suggest three clusters of environments. The first cluster is formed by site 1, which seems to be less correlated with the other sites. Another cluster can be identified by sites 2 and 4. The third cluster contains the remaining sites, which appear to be highly correlated among each other considering the very small angle between them.

The spaghetti plots of the empirical BLUP values for the clones as a function of the trials for models FA1-AR1 and FA1-BD are presented in Figure 3.5. The lines connect the estimated breeding values for the clones across the environments and allow to graphically illustrate rank changes. The pattern of the spatial FA1-AR1 model indicates that there are important rank changes from site 1 to the other sites, as evidenced by the number of non-parallel lines. Then from site 2 to site 7, the lines seem to be relatively flat suggesting small rank changes. On the other hand, the pattern of the FA1-BD model exhibits small rank changes throughout the environments since the lines are essentially parallel to one another.

The main difference between the underlying genetic correlation structures estimated from the best spatial (FA1-AR1, FA2-AR1) and non-spatial models (FA1-BD, FA2-BD) is about site 1. The spatial models suggest an overall genetic correlation between site 1 and the rest of around 0.4, while the non-spatial models indicate a larger value, of about 0.8. Although these findings may look suspicious considering the lack of agreement between models, we have reasons to be confident about the obtained results. The trials were managed in a similar manner following standard progeny test protocols, however site 1 has two distinctive features to stand out from the rest. First, posses the highest number of observations, and second, has the highest site quality ratio \( r = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e} \), measured by the ratio of the clone genetic variance to the sum of the measurement error and genetic variance for height. That is, this site exhibits the lowest
relative environmental noise along with the largest sample. Thus, we are confident with the data coming from this site. In addition, the spatial models provide a much better fit to the data relative to the non-spatial models based on the fit statistics shown in Table 3.4. Hence, we are also confident with the model performances. Perhaps other factors not considered in the presented analysis explain the differences in the observed relative genetic merit of individuals across environments, or possibly, the spatial analysis revealed a pattern and magnitude of genotype by environment interactions that could have been overlooked using a non-spatial approach.

Volume is a trait of economic importance that is typically used for selection. Parameter estimates for each trial from the spatial models using volume as the response variable are presented in the appendix (Table A.3). Although volume data are noisy, there is evidence of spatial dependence in the row and column directions. The sample variograms exhibited a similar pattern as for height. The proportion of genetic variance relative to measurement error and genetic variance was narrower than for height and ranged from 0.07 to 0.26. The rotated and normalized loadings from the FA2-AR1 (Figure A.1) and FA2-BD (Figure A.2) models using volume as the response variable suggest an overall agreement with height results regarding the genetic correlations among environments.

### 3.3.4 Models comparison

The change of relative rankings between the explored models was quantified through correlation measures and clusters of models were identified. Figure 3.6 presents a scatter-plot of the predicted clone breeding values of the FA1-AR1 model against the simple IID-IID. These predictions are in the same units as the height was measured (meters) and were obtained from ASReml-R package (Butler et al., 2009), which uses the methods described by Welham et al.
(2004) and Gilmour et al. (2004). The predicted genetic merit of the clones from both models exhibited a strong positive correlation. The Spearman and Pearson’s correlation between the predictions are 0.90 and 0.91, respectively. The dashed lines in the scatter-plot indicate the 90% quantile of clone breeding values for each model. Thus, the upper right quadrant groups the intersection of the top 10% clones from both models. The degree of association between the rankings from each model is surprising considering the difference in complexity between the models and the underlying assumptions regarding genotype by environment interaction effects. Notice that the IID-IID model does not account for $G \times E$ interactions effects and the residual structure assumes the same variance across sites. Similar scatter-plots to Figure 3.6 are obtained from the predictions of the spatial against the non-spatial models.

Table 3.9 presents the Pearson and Spearman (rank) correlations between the estimated clone breeding values for the explored models. The upper left block formed by the non-spatial models contains pair correlation values of 0.97 or more. These results suggest that for the ACE population, rankings based on breeding values obtained from non-spatial models are relatively insensitive to the choice of variance-covariance structure. For instance, the Pearson’s and rank correlations between the FA2-BD model (29 parameters) and the simple IID-IID model (3 parameters), are both around 0.98.

The block of spatial models in the bottom right of Table 3.9 indicates a strong association among the FA1-AR1, FA2-AR2, and DIA-AR1 models, with pairwise correlations very close to one. The CS-AR1 model is the exception in this group because the clone breeding values are less correlated with the other spatial models, with a value of around 0.91. Overall the correlation between the clone breeding values from the non-spatial and the spatial models are about 0.90, with the exception of the spatial compound symmetric CS-AR1 with higher correlation values of around 0.96. Hence, for rankings based on clone breeding values, the spatial models give very similar results, with the exception of the CS-AR1. The DIA-AR1
model, which does not account for G×E interaction effects, produced very similar rankings as the FA models, which are aimed to fully describe these interaction effects. Perhaps the relatively absence of G×E across most sites explains the agreement among models for similar rankings.

### 3.3.5 Ranking within families

The change of ranking within the 51 families was assessed through the within family rank correlations, that is the rank correlation among clones belonging to the same family. Figure 3.7 shows a histogram of the 51 family rank correlations between the clone breeding values obtained from the best model (FA1-AR1) and the simplest model (IID-IID). The histogram of the within family rank correlations is relatively symmetric with a range of values that goes from 0.69 to 0.93, and exhibits a peak at the average value of 0.85. These results indicate that some families are more sensitive than others regarding the chosen model for estimating the clone breeding values. Relatively low correlations observed from some families between FA1-AR1 versus IID-IID models suggest that choosing poor models may cause erroneous ranking within families and would reduce genetic gains from within-family selections.

Table A.4 presents the average within family rank correlations between the clone breeding values from all model pairs. Similarly to the overall correlations, the spatial and non-spatial models were roughly clustered based on their average correlation values. The spatial models were highly correlated among each other with the exception of CS-AR1 that exhibited a relative lower value. The within family correlations between the spatial and non-spatial models were around 0.85 with the exception of model CS-AR1, for which the correlations were higher, approximately 0.91.

Figure 3.8 presents a heat map of the average within family rank correlations. This plot was
obtained using the *agnes* R package (Maechler et al., 2012) from the dissimilarity matrix \( \mathbf{I} - \mathbf{C} \), where \( \mathbf{C} \) represents matrix of within family rank correlations, and \( \mathbf{I} \) the identity matrix. The dendrograms on the margins and the pattern exhibited by the heat map suggest the presence of clusters formed by the models. The spatial models formed one cluster. Another bigger cluster can be identified by the non-spatial models with perhaps two sub-clusters, specifically models FA2-BD, US-BD, FA1-BD are on one group, and models CS-BD, IID-BD, DIA-BD in another. Finally, models IID-IID and CS-AR1 form two separate clusters.

### 3.3.6 Model Fitting

The models that utilized the FA covariance structures to account for the heterogeneity in variances across multiple environments were superior based on the Akaike Information Criterion. Furthermore, the FA structure combined with a spatial modeling of the residuals greatly improved the fit as shown in Table 4. However, the fitting process for the FA1-AR1 and FA2-AR1 models was computationally demanding and time consuming. A thorough exploration of the parameter space was needed to obtain sensible starting values, otherwise the fitting algorithm failed to converge. Considering the computational burden for analyzing multi-environmental trials, our experience indicates that the spatial analysis may be limited to a moderate number of typically sized trials. On the other hand, the FA structure combined with a diagonally block covariance structure for the residuals was straightforward to fit. For instance, the FA1-BD and FA2-BD provided parsimonious and informative models that could be fit faster and without running into convergence problems since the computational burden of the mixed model analysis was greatly reduced. Heterogeneous residual variances are used in forest progeny trials because of the observed large growth differences among and within environments. Analytical procedures should account for these differences to be efficient in the prediction of genetic
merit.

3.3.7 Repeatability of clone means

Since each site contained only one ramet per clone, the simple IID-IDD model was used to estimate the repeatability of clone means for each trait. The clone means were moderately highly repeatable for height \( H^2_{HT} = 0.58 \pm 0.07 \), whereas clone means for diameter were less repeatable \( H^2_{DBH} = 0.40 \pm 0.08 \). For volume, the repeatability of clone means was moderately high \( H^2_{VOL} = 0.42 \pm 0.08 \).

A key parameter in the accuracy of the prediction of breeding values from forest tree progeny tests is the heritability of the population under study. This parameter quantifies the proportion of variation among individuals that can be attributed to genetic factors and gives a measure of the correlation between phenotypic values and breeding values (Falconer and Mackay, 1996). In general, the reported values of narrow sense individual-tree heritability from traditional seedling tests of loblolly pine are fairly low, particularly for growth traits (Balocchi et al., 1993; Isik et al., 2004; Ogut et al., 2014). Thus, the moderately high clone mean repeatability values from this study suggest that selection from tests of cloned progeny could be more efficient than traditional seedling tests. For instance, Isik et al. (2004) found higher within-family heritability from clonally replicated tests compared to seedling testing and concluded that clonally replicated progeny tests yielded greater genetic gain than seedling tests, even with their higher associated costs and a longer cycle, because of the greater genetic gains from within family selections.
3.3.8 Clonal versus seedling tests

Table 3.10 presents the correlation values between the genetic merit of the clones from each site independently versus the genetic merit of the clones based on the MET analysis using the best model (FA1-AR1). The correlations between the predicted breeding values of clones based on one site and breeding values based on MET analysis were high, around 89%. Not surprisingly, site 1 exhibits the lowest correlation values since FA1-AR1 model predicts smaller genetic correlations between this site and the others. In fact, these results are in agreement with the estimated genetic correlation structure between environments. In addition, the effect on selection was clear with a 50% agreement in the top 100 clones between the MET analysis and the ranking from the first site, compared to a 60 to 70% agreement between the MET and the ranking from the other sites.

3.4 Conclusions

Among the explored models for the genetic analysis of traits, the factor analytic approach performed particularly well in terms of providing a parsimonious and informative model for assessing genotype by environment interactions, and to make accurate predictions of genotypes. The spatial analysis greatly improved the fit relative to the traditional homogeneous residual structure. This approach accounts for the spatial variation within trials at the cost of increasing the computational requirements, which probably limits the method to a moderate number of typically sized trials. On the other hand, the non-spatial diagonally blocked structure for the residuals reduced the computational burden of the mixed model analysis while giving satisfactory results. Hence, this method would be more suitable for the joint analysis of large multi-environmental trials.
3.5 Tables

Table 3.1: Test site identification number, location, number of rows and columns, total number of observations, average height in meters and standard deviation for each trial. The sites were established in 2009 and 2010 (site 7) and the measurements were taken in 2013.

<table>
<thead>
<tr>
<th>Site ID</th>
<th>Location</th>
<th>Row</th>
<th>Columns</th>
<th>Observations</th>
<th>Avg. height (m) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wayne, GA</td>
<td>40</td>
<td>62</td>
<td>2410</td>
<td>4.58 (0.95)</td>
</tr>
<tr>
<td>2</td>
<td>Colleton, SC</td>
<td>40</td>
<td>62</td>
<td>2161</td>
<td>5.02 (0.91)</td>
</tr>
<tr>
<td>3</td>
<td>Greene, AL</td>
<td>40</td>
<td>62</td>
<td>2201</td>
<td>5.10 (0.62)</td>
</tr>
<tr>
<td>4</td>
<td>Pulaski, GA</td>
<td>40</td>
<td>62</td>
<td>1294</td>
<td>3.52 (0.56)</td>
</tr>
<tr>
<td>5</td>
<td>Pierce, GA</td>
<td>48</td>
<td>52</td>
<td>2302</td>
<td>4.69 (0.73)</td>
</tr>
<tr>
<td>6</td>
<td>Beaufort, NC</td>
<td>146</td>
<td>17</td>
<td>1948</td>
<td>4.54 (0.79)</td>
</tr>
<tr>
<td>7</td>
<td>Colleton, SC</td>
<td>80</td>
<td>23</td>
<td>1627</td>
<td>3.74 (0.53)</td>
</tr>
</tbody>
</table>
Table 3.2: Variance-covariance structures used in the linear mixed models for a) the residuals (R), and b) for the genetic effects (G). The mathematical form is expressed in matrix notation.

### a) Covariance structure for the residuals

<table>
<thead>
<tr>
<th>Structure (Abbreviation)</th>
<th>Mathematical form</th>
<th>Number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent (IID)</td>
<td>( \sigma^2 I_n )</td>
<td>1</td>
</tr>
<tr>
<td>Block diagonal (BD)</td>
<td>( \bigoplus_{i=1}^t \sigma^2_i I_{n_i} )</td>
<td>( t )</td>
</tr>
<tr>
<td>Autoregressive (AR1)</td>
<td>( \bigoplus_{i=1}^t \tau_i^2 \Gamma_{r_i}(\rho_{r_i}) \otimes \Gamma_{c_i}(\rho_{c_i}) )</td>
<td>( 3t )</td>
</tr>
</tbody>
</table>

### b) Covariance structure for the genetic effects

<table>
<thead>
<tr>
<th>Structure (Abbreviation)</th>
<th>Mathematical form</th>
<th>Number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent (IID)</td>
<td>( \sigma^2 )</td>
<td>1</td>
</tr>
<tr>
<td>Diagonal (DIA)</td>
<td>( \bigoplus_{i=1}^t \sigma^2_i )</td>
<td>( t )</td>
</tr>
<tr>
<td>Compound symmetric (CS)</td>
<td>( \sigma^2 I_t + \sigma^2 J_t )</td>
<td>2</td>
</tr>
<tr>
<td>Factor analytic ( k )th order (FA( k ))</td>
<td>( \Lambda \Lambda^T + \Psi )</td>
<td>( (k + 1) t - k(k - 1)/2 )</td>
</tr>
<tr>
<td>Unstructured (US)</td>
<td>( \Sigma(\sigma_{ij}); 1 \leq i, j \leq t )</td>
<td>( t(t + 1)/2 )</td>
</tr>
</tbody>
</table>

\( I_n \): \( n \)-by-\( n \) identity matrix  
\( J_n \): \( n \)-by-\( n \) matrix of ones  
\( \Lambda \): \( t \)-by-\( k \) matrix of loadings  
\( \Psi \): \( t \)-by-\( t \) diagonal matrix of specific variances  
\( \Gamma_r \): Rows correlation matrix  
\( \Gamma_c \): Columns correlation matrix  
\( \rho_r \): Row autocorrelation parameter  
\( \rho_c \): Column autocorrelation parameter
Table 3.3: Parameter estimates followed by standard errors in parenthesis from individual site analysis using tree height (m) as response variable. At each site, a pedigree based linear mixed model with spatially correlated residuals and independent measurement error was fit. \( \sigma^2_g \) is the variance for the clone genetics effect; \( \sigma^2 \) represents the measurement error variance (nugget effect); \( \tau^2 \) is the spatial variance parameter (partial sill). Parameters \( \rho_r \) and \( \rho_c \) are the associated autocorrelation coefficients for the rows and columns, respectively. The structure of the spatial correlation matrix corresponds to a separable autoregressive process of first order in the row and column directions.

<table>
<thead>
<tr>
<th>Site</th>
<th>Genetic (( \sigma^2_g ))</th>
<th>Nugget (( \sigma^2 ))</th>
<th>Partial sill (( \tau^2 ))</th>
<th>Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Row (( \rho_r ))</td>
</tr>
<tr>
<td>1</td>
<td>0.17 (0.06)</td>
<td>0.38 (0.04)</td>
<td>0.20 (0.03)</td>
<td>0.54 (0.06)</td>
</tr>
<tr>
<td>2</td>
<td>0.13 (0.05)</td>
<td>0.61 (0.03)</td>
<td>0.10 (0.02)</td>
<td>0.61 (0.10)</td>
</tr>
<tr>
<td>3</td>
<td>0.04 (0.01)</td>
<td>0.26 (0.01)</td>
<td>0.09 (0.02)</td>
<td>0.39 (0.11)</td>
</tr>
<tr>
<td>4</td>
<td>0.01 (0.01)</td>
<td>0.18 (0.01)</td>
<td>0.12 (0.02)</td>
<td>0.66 (0.07)</td>
</tr>
<tr>
<td>5</td>
<td>0.09 (0.03)</td>
<td>0.39 (0.02)</td>
<td>0.06 (0.01)</td>
<td>0.43 (0.13)</td>
</tr>
<tr>
<td>6</td>
<td>0.05 (0.02)</td>
<td>0.43 (0.02)</td>
<td>0.29 (0.13)</td>
<td>0.99 (0.01)</td>
</tr>
<tr>
<td>7</td>
<td>0.06 (0.02)</td>
<td>0.18 (0.01)</td>
<td>0.04 (0.01)</td>
<td>0.91 (0.03)</td>
</tr>
</tbody>
</table>
Table 3.4: Model-fit statistics. The models are named based on the structure of the genetic by environment \((G_g)\) and residual \((R)\) variance-covariance matrices, e.g. IID-BD for \(G_g=IID\) and \(R=BD\). \(N\) represents the number of variance components in the model to be estimated. \(\log(L)\) is the logarithm of the restricted likelihood function evaluated at the parameter estimates. AIC and BIC are the Akaike and Bayesian information criteria, respectively. SEP and SED are the standard error of prediction and difference, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>(N)</th>
<th>(\log(L))</th>
<th>AIC</th>
<th>BIC</th>
<th>Mean SEP</th>
<th>Mean SED</th>
</tr>
</thead>
<tbody>
<tr>
<td>IID-IID</td>
<td>3</td>
<td>-2302</td>
<td>4610</td>
<td>4633</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td>IID-BD</td>
<td>9</td>
<td>-1808</td>
<td>3635</td>
<td>3703</td>
<td>0.17</td>
<td>0.24</td>
</tr>
<tr>
<td>DIA-BD</td>
<td>15</td>
<td>-1804</td>
<td>3637</td>
<td>3750</td>
<td>0.18</td>
<td>0.26</td>
</tr>
<tr>
<td>CS-BD</td>
<td>10</td>
<td>-1758</td>
<td>3536</td>
<td>3611</td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td>FA1-BD</td>
<td>22</td>
<td>-1720</td>
<td>3484</td>
<td>3650</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>FA2-BD</td>
<td>29</td>
<td>-1718</td>
<td>3494</td>
<td>3712</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td>US-BD</td>
<td>36</td>
<td>-1712</td>
<td>3495</td>
<td>3767</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td>DIA-AR1</td>
<td>29</td>
<td>-1387</td>
<td>2833</td>
<td>3051</td>
<td>0.17</td>
<td>0.24</td>
</tr>
<tr>
<td>CS-AR1</td>
<td>24</td>
<td>-1367</td>
<td>2781</td>
<td>2962</td>
<td>0.17</td>
<td>0.24</td>
</tr>
<tr>
<td>FA1-AR1</td>
<td>36</td>
<td>-1313</td>
<td>2699</td>
<td>2970</td>
<td>0.18</td>
<td>0.25</td>
</tr>
<tr>
<td>FA2-AR1</td>
<td>43</td>
<td>-1312</td>
<td>2710</td>
<td>3035</td>
<td>0.18</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Table 3.5: Rotated loadings ($\hat{\lambda}_1^*$ and $\hat{\lambda}_2^*$) using the principal component representation and percentage of total variance accounted by the factors ($\hat{\omega}_1^2$, $\hat{\omega}_1^2 + \hat{\omega}_2^2$) at each environment for FA2-AR1 model.

<table>
<thead>
<tr>
<th>Site</th>
<th>$\hat{\lambda}_1^*$</th>
<th>$\hat{\lambda}_2^*$</th>
<th>$\hat{\omega}_1^2$</th>
<th>$\hat{\omega}_1^2 + \hat{\omega}_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.75</td>
<td>0.38</td>
<td>79</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>-0.36</td>
<td>-0.30</td>
<td>59</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>-0.15</td>
<td>-0.11</td>
<td>50</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>-0.15</td>
<td>-0.11</td>
<td>66</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>-0.29</td>
<td>-0.19</td>
<td>55</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>-0.23</td>
<td>-0.19</td>
<td>48</td>
<td>83</td>
</tr>
<tr>
<td>7</td>
<td>-0.25</td>
<td>-0.17</td>
<td>62</td>
<td>88</td>
</tr>
</tbody>
</table>
Table 3.6: Genetic correlations between pairs of environments from FA2-AR1 model. The high correlations among sites 2 through 7 suggest a lack of genotype by environment interactions effects ($G \times E$) among these sites. Conversely, the relatively low correlation between site 1 and the others indicate the presence of $G \times E$.

<table>
<thead>
<tr>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 4</th>
<th>Site 5</th>
<th>Site 6</th>
<th>Site 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.39</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.46</td>
<td>0.99</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.44</td>
<td>0.88</td>
<td>0.78</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>0.91</td>
<td>0.80</td>
<td>0.91</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.47</td>
<td>0.93</td>
<td>0.82</td>
<td>0.94</td>
<td>0.83</td>
<td>0.85</td>
<td></td>
</tr>
</tbody>
</table>

111
Table 3.7: Rotated loadings ($\hat{\lambda}_1^*$ and $\hat{\lambda}_2^*$) using the principal component representation and percentage of total variance accounted by the factors ($\hat{\omega}_1^2$, $\hat{\omega}_1^2 + \hat{\omega}_2^2$) at each environment for FA2-BD model.

<table>
<thead>
<tr>
<th>Site</th>
<th>$\hat{\lambda}_1^*$</th>
<th>$\hat{\lambda}_2^*$</th>
<th>$\hat{\omega}_1^2$</th>
<th>$\hat{\omega}_1^2 + \hat{\omega}_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.38</td>
<td>0.15</td>
<td>86</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>-0.46</td>
<td>-0.10</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>-0.23</td>
<td>-0.01</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>-0.21</td>
<td>-0.04</td>
<td>97</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>-0.39</td>
<td>-0.01</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>6</td>
<td>-0.31</td>
<td>-0.01</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>7</td>
<td>-0.34</td>
<td>0.01</td>
<td>90</td>
<td>90</td>
</tr>
</tbody>
</table>
Table 3.8: Genetic correlations between pairs of environments from FA2-BD model. All the correlation pairs are greater than 0.80 suggesting the absence of important genotype by environment interactions.

<table>
<thead>
<tr>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 4</th>
<th>Site 5</th>
<th>Site 6</th>
<th>Site 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.83</td>
<td>0.93</td>
<td>0.94</td>
<td>0.89</td>
<td>0.86</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>0.87</td>
<td>0.93</td>
<td>0.94</td>
<td>0.89</td>
<td>0.86</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>0.85</td>
<td>0.99</td>
<td>0.94</td>
<td>0.89</td>
<td>0.86</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>0.85</td>
<td>0.92</td>
<td>0.89</td>
<td>0.93</td>
<td>0.85</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>0.88</td>
<td>0.93</td>
<td>0.90</td>
<td>0.93</td>
<td>0.86</td>
<td>0.89</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Table 3.9: Pearson’s correlations (below diagonal) and rank correlations of clone breeding values (above diagonal) between the models.

<table>
<thead>
<tr>
<th></th>
<th>IID-HD</th>
<th>IID-BD</th>
<th>DIA-BD</th>
<th>CS-BD</th>
<th>FA1-BD</th>
<th>FA2-BD</th>
<th>US-BD</th>
<th>DIA-AR1</th>
<th>CS-AR1</th>
<th>FA1-AR1</th>
</tr>
</thead>
<tbody>
<tr>
<td>IID-HD</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.95</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>IID-BD</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>DIA-BD</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.91</td>
<td>0.97</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>CS-BD</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>FA1-BD</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>FA2-BD</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>US-BD</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>DIA-AR1</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>CS-AR1</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>FA1-AR1</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
</tbody>
</table>

114
Table 3.10: Pearson’s, Spearman’s rank, and within family rank correlation between the genetic merit of the clones at each site versus the genetic merit based on the multi-environmental trails analysis using the best model (FA1-AR1) and height as the response variable.

<table>
<thead>
<tr>
<th>Site</th>
<th>Pearson</th>
<th>Spearman’s rank</th>
<th>Within family rank (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.86</td>
<td>0.84</td>
<td>0.86 (0.71-0.95)</td>
</tr>
<tr>
<td>2</td>
<td>0.92</td>
<td>0.91</td>
<td>0.90 (0.90-0.96)</td>
</tr>
<tr>
<td>3</td>
<td>0.87</td>
<td>0.86</td>
<td>0.88 (0.79-0.96)</td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
<td>0.92</td>
<td>0.89 (0.80-0.95)</td>
</tr>
<tr>
<td>5</td>
<td>0.87</td>
<td>0.85</td>
<td>0.88 (0.79-0.96)</td>
</tr>
<tr>
<td>6</td>
<td>0.90</td>
<td>0.89</td>
<td>0.88 (0.79-0.95)</td>
</tr>
<tr>
<td>7</td>
<td>0.90</td>
<td>0.89</td>
<td>0.88 (0.79-0.95)</td>
</tr>
</tbody>
</table>
3.6 Figures

Figure 3.1: Test site locations across the southeastern United States.
Figure 3.2: Empirical semi-variogram from spatial analysis of test sites using height (m) as response variable. At each site, a pedigree based linear mixed model with spatially correlated residuals and independent measurement error was fit. The plots quantify the spatial correlation at each site. Most of the sites exhibit a smooth surface that increases in the row and column directions reaching a plateau of little variation, which indicates the absence of both global and extraneous trend.
Figure 3.3: Rotated and normalized loadings for the genetic effects from FA2-AR1 model using height as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments.
Figure 3.4: Rotated and normalized loadings for the genetic effects from FA2-BD model using height as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments.
Figure 3.5: Spaghetti plots of the empirical BLUP values for the clones as a function of environment for FA1-AR1 (top) and FA1-BD (bottom) models. The pattern of the spatial FA1-AR1 model indicates that there are important rank changes from site 1 to site 2, as evidenced by the number of non-parallel lines. On the other hand, the pattern of the FA1-BD model exhibits small rank changes throughout the environments since the lines are essentially parallel.
Figure 3.6: Scatter-plot of predicted clone genetic merit between factor analytic spatial (FA1-AR1) and standard non-spatial (IID-IID) model. The dashed lines indicate the 90% quantile for each model. Both rank and Pearson’s correlations between the predictions are about 0.90.
Figure 3.7: Histogram of within family rank correlations based on the clone genetic merit between factor analytic spatial (FA1-AR1) and standard non-spatial (IID-IID) models. The within family rank correlations ranged from 0.69 to 0.93 with an average of approximately 0.85.
Figure 3.8: Heat-map with dendrogram to illustrate clusters of models based on the average within family rank correlations of clone breeding values between model pairs. Similar performing model pairs are indicated with values close to one (red color).
REFERENCES


J. K. Stringer and B. R. Cullis. Application of spatial analysis techniques to adjust for fertility
trends and identify interplot competition in early stage sugarcane selection trials. *Crop and

Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statis-

USDA. United States Department of Agriculture, Agricultural Research Service.

in selecting lodgepole pine populations for future climates. *Global Change Biology*, 12:

Tongli Wang, Gregory O’Neill, and Sally Aitken. Integrating environmental and genetic effects
to predict responses of tree population to climate. *Ecological Applications*, 20(1):153–163,
2010.

Sue Welham, Brian Cullis, Beverley Gogel, Arthur Gilmour, and Robin Thompson. Prediction
2004.

O. O. Wells and C. C. Lambeth. Loblolly pine provenance test in southern arkansas 25th year

Osborn O. Wells. Southwide pine seed source study–loblolly pine at 25 years. *Southern


APPENDIX
Appendix A

Appendix

A.1 Computational details

This section presents some of the most relevant R (R Core Team, 2014) code required for fitting the models outlined in the paper. The goal is to provide sufficient information to implement the algorithms.

A.1.1 Data File

The data is read into R from a csv file that contains all the necessary information for fitting the models. The following shows the code to load the data into R object named dat and how to convert the loaded variables to numerical mode for later use. In addition, the pedigree file is loaded and the inverse of the numerator relationship matrix is computed.

```r
# Load ASreml-R library
library(asreml)

# Read in ACE data
dat=read.csv(file='ACE_data.csv',header=T,colClasses='character')

# Convert numerical variables
```

138
# Height (feet)
dat$HT=as.numeric(dat$HT)
# Diameter (inches)
dat$DBH=as.numeric(dat$DBH)
# Volume (cubic feet)
dat$VOL=as.numeric(dat$VOL)
# Height (meters)
dat$HTm=as.numeric(dat$HTm)
# Diameter (centimeters)
dat$DBHcm=as.numeric(dat$DBHcm)
# Volume (cubic meters)
dat$VOLdm=as.numeric(dat$VOLdm)
# Rust disease incidence (0-1 categorical variable)
dat$RUST=as.numeric(dat$RUST)
# Straightness (1 to 6 categorical variable)
dat$STRT=as.numeric(dat$STRT)
# Forking or Ramicorn (0-1 categorical variable)
dat$FORK_RAM=as.numeric(dat$FORK_RAM)
# ROW continuous (not categorical)
dat$ROWc=as.numeric(dat$ROW)
# COL continuous (not categorical)
dat$COLc=as.numeric(dat$COL)

# Load pedigree file
ace.ped=read.csv(file='ACE_ped_complete.csv',
                     colClasses='character',
                     header=T,
                     sep=',')

# Get inverse of numerator matrix A
ace.ainv = asreml.Ainverse(ace.ped)$ginv

### A.1.2 Fitting linear mixed models using ASReml-R

The linear mixed models described in the Materials and Methods section 3.2 were fit using ASReml-R (Butler et al., 2009), which is a package for R that can be obtained from VSNI International (http://www.vsni.co.uk/). Below are the code lines used to fit the models for the individual site analysis described in section 3.2.3 and the multi-environmental trial analysis described in section 3.2.4.
# Individual Site Analysis

# Fit a spatial model at each site independently

# number of sites
n.sites=7

# Initialize results table to store variance component estimates
results=matrix(NA,nrow=n.sites,ncol=7)
# name columns accordingly
colnames(results)=c('SITE','CLONE','NUG','SILL','ROW','COL','r')

# Initialize loop across sites
for(i in 1:n.sites)
{
  ith=which(dat$TEST==i) # index for the ith site
  Data.site.i=dat[ith,] # Data for the ith site
  fit = asreml(fixed = HTm ~ 1 + ROWc + COLc,
               random = ~ ped(CLON) + units,
               rcov= ~ ar1(ROW):ar1(COL),
               ginverse= list(CLON=ace.ainv),
               data=Data.site.i,
               workspace=128e6,
               pworkspace=128e6,
               na.method.Y='include',
               na.method.X='include',
               maxit = 100,
               family=asreml.gaussian(link='identity',dispersion=NA) )

  # Variance components
  vc <- summary(fit)$varcomp

  # Get variance components estimates
  CLON.var <- vc['ped(CLON)!ped', 'component']
  NUG.var <- vc['units!units.var', 'component']
  SILL.var <- vc['R!variance', 'component']
  ROW.cor <- vc['R!ROW.cor', 'component']
  COL.cor <- vc['R!COL.cor', 'component']

  # clone variance plus nugget variance
  Vp=(CLON.var + NUG.var)
  # ratio
  r = CLON.var/Vp
  # Store variance component estimates
  results[i,]=c(i, CLON.var, NUG.var, SILL.var, ROW.cor, COL.cor, r)
}

# Array with ratio of variance components for each site
R=results[,7]
# Standardized weights for each site
weights=R/sum(R)

###########################################################################
# Multi-environmental Trial Analysis
###########################################################################
# Standard Models (non-spatial)
###########################################################################
# Fit IID-IID model
fit1 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc, 
             random = ~ ped(CLON) + ide(CLON),
             rcov= ~ units,
             ginverse= list(CLON=ace.ainv),
             data=dat,
             workspace=128e6,
             pworkspace=128e6,
             na.method.Y='omit',
             na.method.X='omit',
             family=asreml.gaussian(link='identity',dispersion=NA) )

###########################################################################
# Fit IID-BD model
fit2 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc, 
             random = ~ ped(CLON) + ide(CLON),
             rcov= ~ at(TEST):units,
             ginverse= list(CLON=ace.ainv),
             data=dat,
             workspace=128e6,
             pworkspace=128e6,
             na.method.Y='omit',
             na.method.X='omit',
             family=asreml.gaussian(link='identity',dispersion=NA) )

###########################################################################
# Fit DIA-BD model
fit3 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc ,
             random = ~ diag(TEST):ped(CLON) + ide(CLON),
             rcov= ~ at(TEST):units,
             ginverse= list(CLON=ace.ainv),
             data=dat,
             workspace=128e6,
             pworkspace=128e6,
             na.method.Y='omit',
             na.method.X='omit',
             maxit=100,
             family=asreml.gaussian(link='identity',dispersion=NA) )
# Fit CS-BD model
fit4 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc,
random = ~ ped(CLON) + TEST:ped(CLON) + ide(CLON),
rhocov =~ at(TEST):units,
ginverse = list(CLON=ace.ainv),
data = dat,
workspace = 128e6,
pworkspace = 128e6,
na.method.Y = 'omit',
na.method.X = 'omit',
maxit = 100,
family = asreml.gaussian(link='identity',dispersion=NA) )

# Fit FA1-BD model
fit5 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc,
random = ~ fa(TEST):ped(CLON) + ide(CLON),
rhocov =~ at(TEST):units,
ginverse = list(CLON=ace.ainv),
data = dat,
workspace = 128e6,
pworkspace = 128e6,
na.method.Y = 'omit',
na.method.X = 'omit',
maxit = 1000,
family = asreml.gaussian(link='identity',dispersion=NA) )

# Fit FA2-BD model
fit6 = asreml(fixed = VOLm ~ TEST + at(TEST):ROWc + at(TEST):COLc,
random = ~ fa(TEST,2):ped(CLON) + ide(CLON),
rhocov =~ at(TEST):units,
ginverse = list(CLON=ace.ainv),
data = dat,
workspace = 128e6,
pworkspace = 128e6,
na.method.Y = 'include',
na.method.X = 'include',
maxit = 100,
stepsize = 0.0001,
R.param = 'initial.values.M6.txt',
G.param = 'initial.values.M6.txt',
family = asreml.gaussian(link='identity',dispersion=NA) )

# Fit US-BD model
fit7 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc,
random = `us(TEST):ped(CLON) + ide(CLON),
rcov= `at(TEST):units,
ginverse= list(CLON=ace.ainv),
data=dat,
workspace=128e6,
pworkspace=128e6,
na.method.Y='include',
na.method.X='include',
R.param='initial.values.M7.txt',
G.param='initial.values.M7.txt',
maxit=1000,
family=asreml.gaussian(link='identity',dispersion=NA) )

###########################################################################
# Spatial Models without nugget effecs (AR1 X AR1)
###########################################################################
# Fit DIA-AR1 model
fit8 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc,
random = ` diag(TEST):ped(CLON) + ide(CLON),
rcov= ` at(TEST):ar1(ROW):ar1(COL),
ginverse= list(CLON=ace.ainv),
data=dat,
workspace=256e6,
pworkspace=256e6,
na.method.Y='include',
na.method.X='include',
maxit=5000,
R.param='initial.values.M8.txt',
G.param='initial.values.M8.txt',
family=asreml.gaussian(link='identity',dispersion=NA) )

###########################################################################
# Fit CS-AR1 model
fit9 = asreml(fixed = HTm ~ TEST + at(TEST):ROWc + at(TEST):COLc,
random = ` ped(CLON) + TEST:ped(CLON) + ide(CLON),
rcov= ` at(TEST):ar1(ROW):ar1(COL),
ginverse= list(CLON=ace.ainv),
data=dat,
workspace=512e6,
pworkspace=512e6,
na.method.Y='include',
na.method.X='include',
maxit=5000,
R.param='initial.values.M9.txt',
G.param='initial.values.M9.txt',
family=asreml.gaussian(link='identity',dispersion=NA) )

###########################################################################
# Fit FA1-AR1 model
Some of the arguments used to fit models may need to be modified based on the particular analyzed data set and computing platform. For details and documentation for the ASReml-R package, refer to Butler et al. (2009). Some of the useful functions to extract information from asreml object fit1 (IID-IID model) are given below. The same functions can be used to extract information from the other models fit. The comments on the code (followed after #) describe the specific purpose of each function.

# Prediction of clone breeding values using model IID-IID from MET analays
p.fit1=predict(fit1,
  classify='CLON',
  present=list('TEST', prwts=weights),
  sed=T)
# Displays the estimated variance components, along with standard errors
summary(fit1)$varcomp

# Gives the number of residual degrees of freedom ( length(y) - rank(X) )
n=fit1$nedf

# Displays the REML log-likelihood
logLik=fit1$loglik

# Gives the number of covariance parameters in the model
npar=fit1$nwv

# AIC criterion
AIC=-2*logLik + 2*npar

# BIC criterion
BIC=-2*logLik + log(n)*npar

# Displays fixed effects estimates along with their standard errors.
summary(fit1,all=T)$coef.fixed

# Displays BLUPs and their standard errors.
summary(fit1,all=T)$coef.random

# Table of predicted clone breeding values with their standard errors.
idx=44:2405 # indices of the clones
BV=p.fit1$predictions$pvals[idx,

# Average of standard error of predictions
AVSEP=mean(p.fit1$predictions$pvals[idx,3])

# Matrix of standard error of prediction differences for the clones
p.fit$predictions$sed[idx,idx]

# Average of standard error of prediction differences
p.fit$predictions$avsed
### A.2 Tables

Table A.1: Overall summary statistics for the Atlantic Coastal Elite population. Sample mean, number of observations, minimum, maximum, and standard deviation of the measured traits at age 4 years.

<table>
<thead>
<tr>
<th>Trait</th>
<th>MEAN</th>
<th>N</th>
<th>MIN</th>
<th>MAX</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT (m)</td>
<td>4.55</td>
<td>13943</td>
<td>0.34</td>
<td>7.47</td>
<td>0.92</td>
</tr>
<tr>
<td>DBH (cm)</td>
<td>7.66</td>
<td>13874</td>
<td>0.25</td>
<td>17.02</td>
<td>2.60</td>
</tr>
<tr>
<td>RUST freq</td>
<td>0.02</td>
<td>14460</td>
<td>0</td>
<td>1</td>
<td>0.16</td>
</tr>
<tr>
<td>STRT score</td>
<td>3.30</td>
<td>13509</td>
<td>1</td>
<td>6</td>
<td>1.25</td>
</tr>
<tr>
<td>VOL (dm$^3$)</td>
<td>10.23</td>
<td>13874</td>
<td>0.85</td>
<td>48.14</td>
<td>6.79</td>
</tr>
<tr>
<td>FORK/RAM freq</td>
<td>0.24</td>
<td>13941</td>
<td>0</td>
<td>1</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Table A.2: Fixed effects from individual site analysis using height as response variable. At each site a linear row and column effect in addition to an intercept term were fit. The table shows the denominator degrees of freedom, the value for the F statistics and the associated p-values.

<table>
<thead>
<tr>
<th>Site</th>
<th>Parameter</th>
<th>Denominator d.f.</th>
<th>F statistic</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>47.1</td>
<td>643.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Row</td>
<td>29.5</td>
<td>56.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>36.1</td>
<td>16.03</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>Intercept</td>
<td>17</td>
<td>803.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Row</td>
<td>10.1</td>
<td>1.01</td>
<td>0.339</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>17.6</td>
<td>4.67</td>
<td>0.044</td>
</tr>
<tr>
<td>3</td>
<td>Intercept</td>
<td>17.5</td>
<td>1583</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Row</td>
<td>12.2</td>
<td>6.40</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>25.9</td>
<td>7.42</td>
<td>0.011</td>
</tr>
<tr>
<td>4</td>
<td>Intercept</td>
<td>14.4</td>
<td>376.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Row</td>
<td>12.2</td>
<td>3.39</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>19.2</td>
<td>4.76</td>
<td>0.043</td>
</tr>
<tr>
<td>5</td>
<td>Intercept</td>
<td>25.9</td>
<td>1619</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Row</td>
<td>11.6</td>
<td>4.38</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>25.4</td>
<td>0.54</td>
<td>0.466</td>
</tr>
<tr>
<td>6</td>
<td>Intercept</td>
<td>5.6</td>
<td>90.81</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Row</td>
<td>4.4</td>
<td>0.31</td>
<td>0.603</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>6.6</td>
<td>1.79</td>
<td>0.221</td>
</tr>
<tr>
<td>7</td>
<td>Intercept</td>
<td>18.7</td>
<td>1233</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Row</td>
<td>13.0</td>
<td>0.18</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>Column</td>
<td>10.3</td>
<td>1.80</td>
<td>0.208</td>
</tr>
</tbody>
</table>
Table A.3: Parameter estimates followed by standard errors in parenthesis from individual site analysis using tree volume (dm$^3$) as response variable. At each site, a pedigree based linear mixed model with spatially correlated residuals and independent measurement error was fit. $\sigma^2_g$ is the variance for the clone genetics effect; $\sigma^2$ represents the measurement error variance (nugget effect); $\tau^2$ is the spatial variance parameter (partial sill). Parameters $\rho_r$ and $\rho_c$ are the associated autocorrelation coefficients for the rows and columns, respectively. The structure of the spatial correlation matrix corresponds to a separable autoregressive process of first order in the row and column directions.

<table>
<thead>
<tr>
<th>Site</th>
<th>Genetic ($\sigma^2_g$)</th>
<th>Nugget ($\sigma^2$)</th>
<th>Partial sill ($\tau^2$)</th>
<th>Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Row ($\rho_r$)</td>
</tr>
<tr>
<td>1</td>
<td>2.29 (0.84)</td>
<td>9.49 (1.55)</td>
<td>4.34 (0.13)</td>
<td>0.42 (0.07)</td>
</tr>
<tr>
<td>2</td>
<td>7.34 (3.52)</td>
<td>42.16 (13.43)</td>
<td>6.35 (0.34)</td>
<td>0.75 (0.09)</td>
</tr>
<tr>
<td>3</td>
<td>4.96 (2.12)</td>
<td>35.64 (9.50)</td>
<td>5.25 (0.23)</td>
<td>0.58 (0.13)</td>
</tr>
<tr>
<td>4</td>
<td>0.20 (0.10)</td>
<td>1.79 (0.52)</td>
<td>0.92 (0.29)</td>
<td>0.80 (0.06)</td>
</tr>
<tr>
<td>5</td>
<td>4.28 (1.73)</td>
<td>15.13 (4.13)</td>
<td>1.60 (0.26)</td>
<td>0.50 (0.15)</td>
</tr>
<tr>
<td>6</td>
<td>1.18 (0.71)</td>
<td>14.65 (6.09)</td>
<td>26.67 (0.68)</td>
<td>0.99 (0.01)</td>
</tr>
<tr>
<td>7</td>
<td>0.65 (0.25)</td>
<td>1.81 (0.47)</td>
<td>0.35 (0.24)</td>
<td>0.94 (0.03)</td>
</tr>
</tbody>
</table>
Table A.4: Average within family rank correlations (range in parenthesis) of clone breeding values between all model pairs using height as the response variable.

|        | IID-IID | IID-BD     | DIA-BD     | CS-BD   | FA1-BD   | FA2-BD     | US-BD     | DIA-AR1   | CS-AR1   | FA1-AR1   | FA2-AR1   |
|--------|---------|------------|------------|---------|----------|------------|-----------|-----------|----------|-----------|-----------|-----------|
|        | 0.94    | 0.99       | 0.99       | 0.99    | 0.99     | 0.99       | 0.99      | 0.85      | 0.85     | 0.99      | 0.99      |
|        | (0.86-0.98) | (0.97-0.99) | (0.98-0.99) | (0.98-0.99) | (0.99-0.99) | (0.99-0.99) | (0.99-0.99) | (0.66-0.92) | (0.75-0.94) | (0.76-0.95) | (0.76-0.95) | (0.76-0.95) |
A.3 Figures

Figure A.1: Rotated and normalized loadings for the genetic effects from FA2-AR1 model using volume as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments.
Figure A.2: Rotated and normalized loadings for the genetic effects from FA2-BD model using volume as response variable. The cosine of the angle between the vectors of two environments represents the genetic correlation between these environments.