# C H A P T E R 6

# FORECASTING EXCERISE APPLIED TO THE ECOLOGICAL DATA SET

## 6.1 Forecasting Methodology

Due to the experimental design, it was possible to split the data into two halves, such that plants from the same tub were separated into different data sets. Plants labeled as plant 1 were allocated to the training data set, and those labeled as plant 2 were allocated to the validation data set. The plant labels should have no significance. The forecasting exercise was carried out on both the linear mean model and the quadratic mean model. By fitting models to the training data set, model parameters were estimated, under both the linear mean model and under the quadratic mean model, for the no random effects model with $\boldsymbol{\omega}_i$ = UN, the random intercept and slope models with $\boldsymbol{\omega}_i$ = CSH and $\boldsymbol{\Sigma}$ = CSH, with $\boldsymbol{\omega}_i$ = CSH and $\boldsymbol{\Sigma}$ = UN, with $\boldsymbol{\omega}_i$ = ARH(1) and $\boldsymbol{\Sigma}$ = UN, $\boldsymbol{\omega}_i$ = AR(1) and $\boldsymbol{\Sigma}$ = UN, and with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN, the no random effects model with a TOEP covariance structure, the random intercept model with $\boldsymbol{\omega}_i$ = AR(1), and the OLS model. The first three models were chosen as they obtained good AIC and BIC values when fitted to the full data set under the linear mean model, with the first model performing well under both mean models. The random effects models with $\boldsymbol{\omega}_i$ = ARH(1) and $\boldsymbol{\Sigma}$ = UN and with $\boldsymbol{\omega}_i$ = AR(1) and $\boldsymbol{\Sigma}$ = UN were selected as they performed well under the quadratic model. The random effects model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN and the no random effects model with TOEP errors were selected as they performed well in the simulation study, and the OLS model was selected for the purposes of comparison.

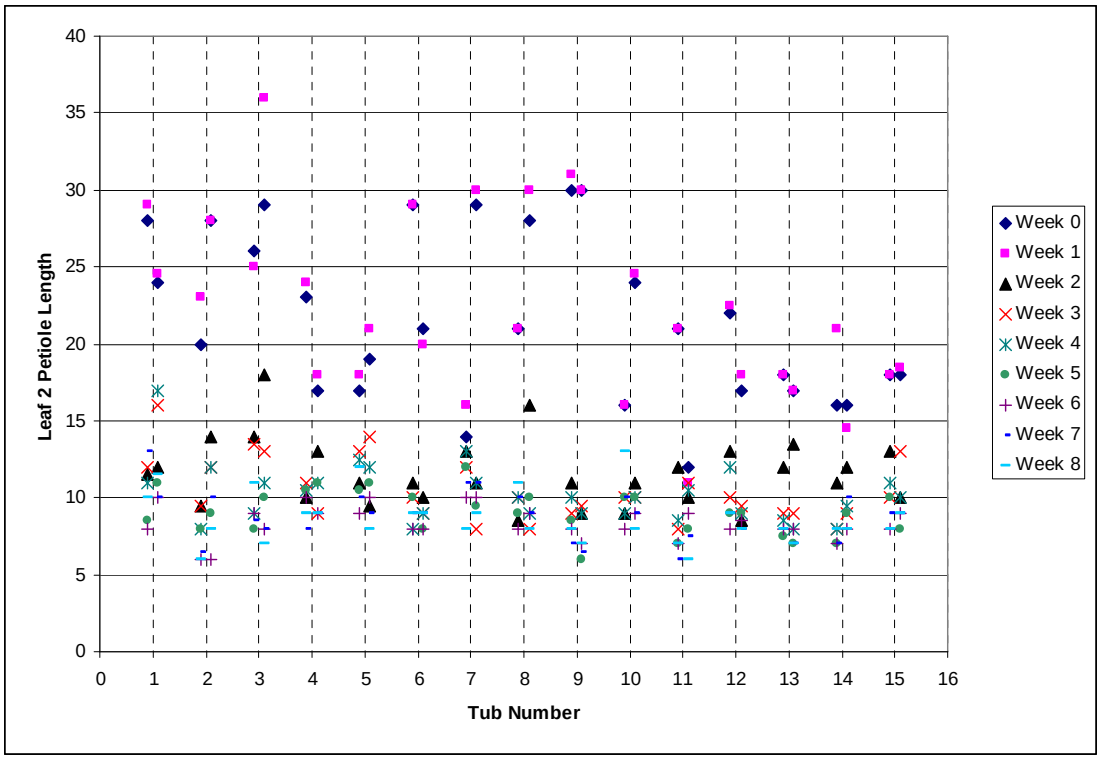


Fig. 6.1: Plot of observed responses for control plants according to each tub. The observations for plant 1 are offset to the left and those for plant 2 are offset to the right.

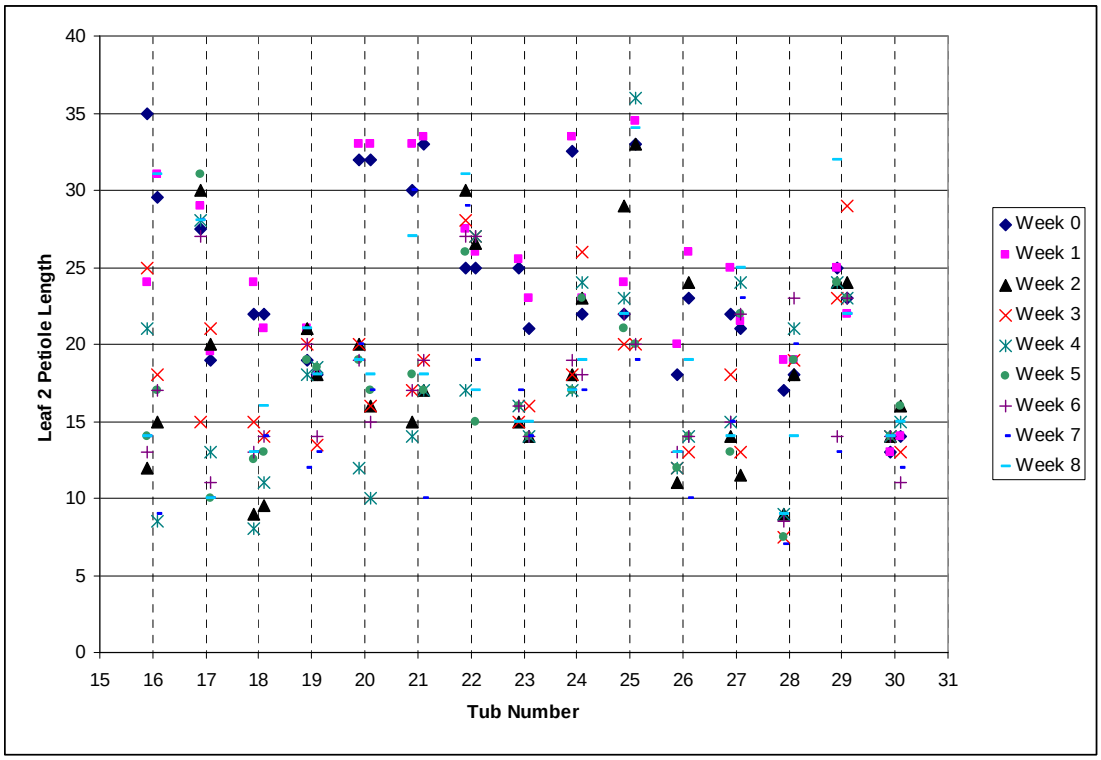


Fig. 6.2: Plot of observed responses for sprayed plants according to each tub. The observations for plant 1 are offset to the left and those for plant 2 are offset to the right.

The data set was split according to tub in order to determine how well the marginal estimates performed in comparison to the conditional estimates for the fixed effects. Plants grown in the same tub, and therefore in the same water, should be expected to have the same environmental variables. For example, nutrient levels through time, exposure to sunlight, weevil predation, etc. should be expected to be the same in each tub. Therefore it would be a reasonable assumption to think that the random effects for subjects from the same tub should be similar. Fig. 6.1 and Fig. 6.2 give plots of the observed data points of the control plants and sprayed plants respectively. These plots indicate that measurements taken at the same time from plants in the same tub are not necessarily more similar to each other than to measurements taken from others plants at that time. Tub 20, with sprayed plants, shows the greatest similarity between measurements taken in the same tub. The values for weeks 0 and 1 are identical, but then for later weeks, the measurements of the two plants begin to differ. Therefore plots of the data do not support the assumption that random effects of plants in the same tub would be similar. Nevertheless, splitting plants that were in the same tub into the training and validation data sets is still a reasonable approach.

The response $\mathbf{y}_i$, conditional on the random effect $\mathbf{b}_i$, is normally distributed with mean vector $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$. Therefore conditional estimates are subject-specific. The marginal density of $\mathbf{y}_i$ has mean structure $\mathbf{X}_i\boldsymbol{\beta}$, and therefore these estimates can be interpreted as population averaged (Verbeke & Molenberghs, 2000).

As in the model fitting exercise applied to the ecological data set, two different mean structures were considered for the forecasting exercise: the simplistic linear mean model, and the more complex quadratic mean model. This allowed the comparison

between the mean models as to which covariance structures resulted in the best predictions of the data. As in the model fitting exercise of the previous chapter, the quadratic model was applied to the logged lengths of the second petiole.

## 6.2 Forecasting Results for the Simplistic Linear Model

Table 6.1: Model parameters and information criteria for models fitted to the training data. The fixed effects estimates appear in the same order as indicated by the model equation for the simplistic mean model described in Chapter five.

| $\boldsymbol{\omega}_i$ | $\boldsymbol{\Sigma}$ | $\boldsymbol{B}_i$ | AIC | BIC | AICc |
|---|---|---|---|---|---|
| UN | None | $\begin{pmatrix} 18.97 \\ -3.66 \\ -0.17 \\ -0.63 \end{pmatrix}$ | 1428.8 | 1491.9 | 1447.6 |
| CSH | UN | $\begin{pmatrix} 16.41 \\ -5.96 \\ -0.11 \\ -0.60 \end{pmatrix}$ | 1527.9 | 1546.1 | 1529.3 |
| AR(1) | Intercept only | $\begin{pmatrix} 22.50 \\ -3.50 \\ -0.70 \\ -0.87 \end{pmatrix}$ | 1583.2 | 1587.4 | 1583.3 |
| TOEP | None | $\begin{pmatrix} 22.66 \\ -3.55 \\ -0.81 \\ -0.90 \end{pmatrix}$ | 1587.7 | 1600.3 | 1588.4 |
| VC | UN | $\begin{pmatrix} 22.26 \\ -3.61 \\ -0.75 \\ -0.86 \end{pmatrix}$ | 1619.6 | 1625.2 | 1619.7 |
| VC | None | $\begin{pmatrix} 22.26 \\ -3.61 \\ -0.75 \\ -0.86 \end{pmatrix}$ | 1677.6 | 1679.0 | 1677.6 |
| CSH | CSH | Non-convergence | | | |
| AR(1) | UN | $\boldsymbol{\Sigma}$ not positive definite | | | |
| ARH(1) | UN | $\boldsymbol{\Sigma}$ not positive definite | | | |

The model fitting results for the simplistic model are presented in Table 6.1. The model with $\boldsymbol{\omega}_i$ = CSH and $\boldsymbol{\Sigma}$ = CSH did not converge even though results were obtained under the full data set, therefore it does appear that sample size plays a big

role in the success of this covariance structure type in reaching convergence and getting good estimates. The random intercept and slope models with AR(1) or ARH(1) errors did not obtain valid estimates for the covariance, as for the full data set. Once again the UN covariance model obtained the smallest AIC, BIC and AICc values. The no random effects model with TOEP errors (hereafter referred to as the TOEP model) also obtained smaller information criteria compared to the model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN, as in the original analysis of the data. Comparing the fixed effects parameter estimates, those estimates from the TOEP model and the model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN are more similar to each other compared to the estimates for the UN model.

The estimates, both marginal and conditional, obtained from the models fitted to the training data set were compared to both the training data set and the validation data set and the forecasting statistics, mean square error (MSE) and mean absolute error (MAE), were obtained. The MSE is defined as $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and the MAE is defined as $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, where $y_i$ and $\hat{y}_i$ represent an observed and predicted value pair, respectively. Models with smaller values for these statistics predict the data better compared to models with larger values. Pan and Fang (2002) describe a very similar technique to selecting between models. They state that this method of model selection by means of using model prediction is extensively accepted in the literature.

The forecasting error statistics (Table 6.2) reveal that models with simpler covariance structures predicted both the test data set and the validation data set better than the models which obtained the best goodness-of-fit statistics. The random effects model

with $\omega_i$ = CSH and $\Sigma$ = UN in particular has relatively large errors. This is in the case of both the validation and test data sets.

Table 6.2: Forecasting statistics for linear model estimates, both marginal and conditional, obtained using the training data set (fitted data) and the validation data set (paired data).

| **Marginal Estimates** | | | | | |
|---|---|---|---|---|---|
| **Covariance** | | **Fitted Data** | | **Paired Data** | |
| $\omega_i$ | $\Sigma$ | MSE | MAE | MSE | MAE |
| VC | None | 28.74 | 4.16 | 26.85 | 3.87 |
| VC | UN | 28.74 | 4.16 | 26.85 | 3.87 |
| TOEP | None | 28.80 | 4.19 | 26.78 | 3.92 |
| AR(1) | Intercept only | 28.97 | 4.25 | 27.02 | 3.94 |
| UN | None | 32.52 | 4.35 | 31.63 | 4.04 |
| CSH | UN | 48.87 | 5.20 | 49.13 | 5.07 |
| | | | | | |
| **Conditional Estimates** | | | | | |
| **Covariance** | | **Fitted Data** | | **Paired Data** | |
| $\omega_i$ | $\Sigma$ | MSE | MAE | MSE | MAE |
| VC | UN | 15.63 | 3.03 | 28.61 | 3.85 |
| AR(1) | Intercept only | 19.77 | 3.55 | 25.80 | 3.80 |
| CSH | UN | 42.82 | 4.41 | 58.95 | 5.60 |

Comparing the performance of the conditional estimates to those from their marginal counter parts, it seems that with regards to the test data set, the conditional estimates produced less error, as the MAE and MSE were smaller for the conditional estimates. When tested against the validation data set, the marginal estimates produced less error, except in the case of the random intercept model with AR(1) errors where the MAE and MSE were smaller for the conditional estimates but only slightly compared to the statistics for the marginal estimates. Therefore, from a "predictive power" point of view, it is better to use the marginal estimates to predict new data. Comparing the marginal estimates of the test data set to the marginal estimates of the validation data set, the models in fact performed better on the validation data set, except in the case of

random effects model with $\boldsymbol{\omega}_i$ = CSH and $\boldsymbol{\Sigma}$ = UN. Conversely, the conditional estimates fitted the test data set better compared to the validation data set, as expected due to the random effects in the mean model.

As in previous analyses, the model with TOEP error covariance structure performed well, obtaining error MSE and MAE values only marginally different from the minimum values. Interestingly, the OLS model, together with the random effects model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN, obtained the minimum MSE and MAE values. This is due to these two models having the same estimates for the mean structure.

## 6.3 Forecasting Results for the Quadratic Model

Table 6.3 presents the fitting results for the quadratic model fitted to the training data set. For the quadratic model, the no random effects model with unstructured covariance obtained the lowest AIC and AICc values, but obtained one of the highest BIC values. The random effects models with $\boldsymbol{\omega}_i$ = ARH(1) and $\boldsymbol{\Sigma}$ = UN and with $\boldsymbol{\omega}_i$ = AR(1) and $\boldsymbol{\Sigma}$ = UN obtained similar values for the information criteria, with the heterogeneous model obtain slightly lower AIC and AICc values, but higher BIC. The random effects model with $\boldsymbol{\omega}_i$ = AR(1) and $\boldsymbol{\Sigma}$ = UN obtained the lowest BIC value compared to all other model considered. The random intercept model with $\boldsymbol{\omega}_i$ = AR(1) obtained the next best set of AIC, BIC and AICc values, followed by the random intercept slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN and the TOEP model. The OLS model obtained the highest values for the information criteria. Comparing the estimates for the fixed effects, the values were very similar between all models, with the estimates of the adjustment parameters showing the most variability. As for the

linear model estimates, the no random effects model with $\boldsymbol{\omega}_i$ = VC (i.e. the OLS model) and the random intercept and slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN obtained the same fixed effects estimates. Under the quadratic mean model, the random effects models with CSH error structures resulted in non-covergence.

Table 6.3: Model parameters and information criteria for models fitted to the training data The fixed effects estimates appear in the same order as indicated by the model equation for the simplistic mean model described in Chapter five.

| $\boldsymbol{\omega}_i$ | $\sum$ | $\boldsymbol{B}_i$ | **AIC** | **BIC** | **AICc** |
|---|---|---|---|---|---|
| UN | None | $\begin{pmatrix} 2.9989 \\ -0.0943 \\ 0.0101 \\ 0.1632 \\ 0.2866 \\ -0.3558 \\ -0.0551 \\ 0.1999 \\ 0.2788 \end{pmatrix}$ | -68.7 | -5.6 | -49.4 |
| ARH(1) | UN | $\begin{pmatrix} 3.0857 \\ -0.1247 \\ 0.0127 \\ 0.0750 \\ 0.2319 \\ -0.3804 \\ -0.0459 \\ 0.2509 \\ 0.2981 \end{pmatrix}$ | -57.4 | -39.2 | -55.9 |
| AR(1) | UN | $\begin{pmatrix} 2.9995 \\ -0.1000 \\ 0.0108 \\ 0.1607 \\ 0.2922 \\ -0.3272 \\ -0.0501 \\ 0.1983 \\ 0.2524 \end{pmatrix}$ | -48.3 | -41.3 | -48.3 |

Table 6.3 (cont.): Model parameters and information criteria for models fitted to the training data The fixed effects estimates appear in the same order as indicated by the model equation for the simplistic mean model described in Chapter five.

| $\omega_i$ | $\Sigma$ | $B_i$ | AIC | BIC | AICc |
|---|---|---|---|---|---|
| AR(1) | Intercept only | $\begin{pmatrix} 2.9944 \\ -0.0991 \\ 0.0108 \\ 0.1674 \\ 0.2988 \\ -0.3255 \\ -0.0501 \\ 0.1947 \\ 0.2478 \end{pmatrix}$ | -44.0 | -39.8 | -43.9 |
| VC | UN | $\begin{pmatrix} 3.0068 \\ -0.1012 \\ 0.0108 \\ 0.1519 \\ 0.2800 \\ -0.3294 \\ -0.0501 \\ 0.2019 \\ 0.2607 \end{pmatrix}$ | -43.0 | -37.4 | -42.9 |
| TOEP | None | $\begin{pmatrix} 2.9917 \\ -0.0989 \\ 0.0108 \\ 0.1650 \\ 0.3018 \\ -0.3193 \\ -0.0507 \\ 0.1989 \\ 0.2381 \end{pmatrix}$ | -37.6 | -25.0 | -36.8 |
| VC | None | $\begin{pmatrix} 3.0068 \\ -0.1012 \\ 0.0108 \\ 0.1519 \\ 0.2800 \\ -0.3294 \\ -0.0501 \\ 0.2019 \\ 0.2607 \end{pmatrix}$ | 99.6 | 103.1 | 99.6 |
| CSH | UN | Non-convergence | | | |
| CSH | CSH | Non-convergence | | | |

The forecasting results appear in Table 6.4. The results show that for the validation data set, the marginal estimates for models with random effects perform better compared to the conditional counter parts. Therefore, as for the simplistic linear model, the marginal estimates are better at predicting a new data set. Comparing the training data set to the validation data set, marginal estimates predicted the validation data set better compared to the training data set, and conditional estimates predicted the training data set better compared to the validation data set. For the training data set, the conditional estimates from the random intercept and slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN gave the best predictions. For the validation data set, the random intercept and slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN and the OLS model obtained that lowest MAE values. The random intercept and slope model with $\boldsymbol{\omega}_i$ = ARH(1) and $\boldsymbol{\Sigma}$ = UN obtained the lowest MSE value. Several models obtained very similar forecasting statistics for both the training data set and for the validation data set. These include the OLS model, the random intercept and slope models with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN and with $\boldsymbol{\omega}_i$ = AR(1) and $\boldsymbol{\Sigma}$ = UN, the random intercept model with $\boldsymbol{\omega}_i$ = AR(1), and the TOEP model. The model which gave the worst estimates for training data set was the no random effects model with unstructured errors. This is surprising, as this model obtained the lowest AIC value.

Table 6.4: Forecasting statistics for quadratic model estimates, both marginal and conditional, obtained using the training data set (fitted data) and the validation data set (paired data).

| **Marginal Estimates** | | | | | |
|---|---|---|---|---|---|
| **Covariance** | | **Fitted Data** | | **Paired Data** | |
| $\omega_i$ | $\Sigma$ | MSE | MAE | MSE | MAE |
| VC | None | 0.0705 | 0.2048 | 0.0651 | 0.2001 |
| VC | UN | 0.0705 | 0.2048 | 0.0651 | 0.2001 |
| AR(1) | Intercept only | 0.0705 | 0.2048 | 0.0654 | 0.2003 |
| TOEP | None | 0.0706 | 0.2049 | 0.0654 | 0.2003 |
| AR(1) | UN | 0.0706 | 0.2048 | 0.0652 | 0.2003 |
| ARH(1) | UN | 0.0708 | 0.2043 | 0.0650 | 0.2009 |
| UN | None | 0.0715 | 0.2049 | 0.0661 | 0.2017 |
| | | | | | |
| **Conditional Estimates** | | | | | |
| **Covariance** | | **Fitted Data** | | **Paired Data** | |
| $\omega_i$ | $\Sigma$ | MSE | MAE | MSE | MAE |
| VC | UN | 0.0228 | 0.1171 | 0.0821 | 0.2017 |
| AR(1) | Intercept only | 0.0319 | 0.1432 | 0.0707 | 0.2007 |
| AR(1) | UN | 0.0247 | 0.1218 | 0.0783 | 0.2066 |
| ARH(1) | UN | 0.0272 | 0.1166 | 0.0825 | 0.2138 |

## 6.4 Discussion

This short forecasting exercise demonstrates that the information criteria may not necessarily be choosing the models that have the best mean structure estimates, as these measures include the estimates of the covariance parameters as well. The MSE and MAE measures on the other hand are only concerned with how well the model predicts the mean structure of the data. In this particular case, it appears that the models with simpler covariance structures gave better predictions compared to the models with more complicated covariance structures. This is the case under both the simplistic linear model and the more complicated quadratic model. In particular the

random intercept and slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN and the OLS model obtained the best predictions under both mean models.

The marginal estimates from the random effects models were better at predicting the validation data set compared to the conditional estimates from these models. Conversely, the conditional estimates predicted the training data set better. This can be attributed to "shrinkage" brought about by the random effects. A consequence of the individual-specific coefficients is to "shrink" the prediction for the $i^{th}$ individual towards the population-averaged response profile (Fitzmaurice *et al*., 2004). This demonstrates that the conditional estimates can be used when there is interest in the estimated response of a particular subject in the sample, but the marginal estimates should rather be used to estimate subjects outside of the sample. This data was originally modelled with the tub as the subject and the plant effect nested within tub. This approach did not perform any better compared to using the individual plants as subjects, and so was not used in order to keep the model as simple as possible. This approach is supported since the conditional estimates did not perform better at predicting the validation data set compared to the marginal estimates, or were very similar, indicating that the plants within the same tub were not more similar to each other compared to plants in other tubs. Plots of the observed responses for each plant against the tub in which the plant was located suggest the same conclusion (Fig. 6.1 and Fig. 6.2).

Of the models which were predicted to perform well under the simulation study, the random intercept and slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN obtained competitive information criteria compared to other models with relatively simple covariance

structures, and obtained the best predictions compared to models under all other covariance structures. The model with the TOEP covariance structure performed well relative the information criteria, obtaining close to the minimum values for these measures compared to other models with simple covariance structures. This model also obtained values for the MAE and MSE that were close to the minimum, indicating good estimates for the parameters in the mean structure. Therefore it can be concluded that both the random intercept and slope model with $\boldsymbol{\omega}_i$ = VC and $\boldsymbol{\Sigma}$ = UN and the no random effects model with TOEP covariance structure are good choices for this relatively complex data set. The third model which performed well under the simulation study, the random intercept and slope model with $\boldsymbol{\omega}_i$ = AR(1) and $\boldsymbol{\Sigma}$ = UN, obtained good forecasting statistics under the quadratic model, but was unusable for the simplistic linear model as the estimated random effects covariance matrix was not positive definite.

The OLS model obtained relatively poor information criteria, but obtained the best MAE and MSE values under both the linear and quadratic mean models. This indicates that the OLS model estimated the mean structure well, but estimated the covariance matrix poorly. This supports the fact that the OLS estimator for the mean is consistent and unbiased (Verbeke & Molenberghs, 2000; Demidenko, 2004).