22 Crop pollination by honeybees in Argentina using additive mixed modelling

Basualdo, M., Ieno, E.N., Zuur, A.F. and Smith, G.M.

22.1 Introduction

Throughout the world, honeybees (*Apis mellifera L.*) are used as a pollinator in more than 40 types of commercial crops. This practise not only increases the crop yield, but also the quality of fruited varieties of commercial interest. The use of honeybees may improve crop production as a consequence of cross- pollination or due to the physical contact by foraging behaviour on the flowers.

Darwin formerly described the effect of cross-pollination in 1877. Further work was done by Waite in 1895, with studies carried out on pears showing the value of inter-plantation of cultivars and the roll of honeybees in transferring the pollen among them (Waite 1895).

In seed production systems, the pollination process also has a direct impact on crop production, and consequently on foraging legume or oilseed production such as the sunflower (*Helianthus annuus L.*). The sunflower is the second most important oilseed crop in the world, after soybean, as it is cholesterol-free and also has anti-cholesterol properties. In Argentina, the extension of sunflower cultivated areas for oilseed production has increased in recent years.

Seed from commercial hybrid sunflowers is produced using cytoplasmic Male Sterility. These lines, known as male-sterile (MS) or female, are fertilised with male lines denoted as male-fertile (MF) or restorer lines, allowing the recovery of fertility in the hybrid F1 generation. In the seed fields, the MS–MF lines are planted in separate rows in ratios ranging from 2:1 to 10:1 in order to optimise the number of seed-producing MS plants (Dedio and Putt 1980). To obtain adequate seed production, pollen has to be transferred from the MF lines to the MS lines. The honeybee is considered the most important pollinator of sunflowers (McGregor 1976), and at present, honeybee colonies are placed in sunflower fields to ensure adequate pollination. When honeybees move from MF to MS rows, cross-pollination should occur (Ribbands 1964; Delaude et al. 1978; Radford and Rhodes 1978; Drane et al. 1982). To ensure an adequate seed production, honeybees should visit both parental lines and be located uniformly within the MS. There is a direct linear relationship between the amount of seed produced and the number of visits that the sunflower head (capitulum) receives.

The aim of this case study chapter is to determine how honeybee foraging activity changes in response to days, state of flowering, time of day, temperature and visitation between the MF and the MS lines.

22.2 Experimental setup

The sunflowers used in this study were grown at the Universidad Nacional del Centro de la Provincia de Buenos Aires Campus in Tandil, Argentina. The Tandil district lies in the centre and southeast area of Buenos Aires province, at an altitude of 178 m (see Chapter 28 for a map reference). The area consists of a flat pampas phytogeography region surrounded by rounded slopes. The climatic conditions and soil development make this a particularly warm temperate pampas system and is considered one of the most productive areas in the world for agriculture and cattle rearing (Burkart et al. 1999).

For the experiment, 2 ha were sown with MF and MS sunflower transects at a density of 5 seeds per linear metre. The study plots contained six rows of MS plants followed by four rows of MF, in accordance with the seed company instructions, with a distance of 0.7 m between rows. Flowering extended from February 7 (15% of flowering) to February 16. Honeybee colonies composed of nine standard Langstroth frames were used. The colonies were placed at the edge of each experimental plot with their entrances facing the MS and MF rows, when 15% of the flowers were open (Basualdo et al. 2000).

Before flowering started, the state of flowering on MS and MF plants was estimated by tagging five capitula in five MS rows and five capitula in two MF rows. Each day, the total area of the tagged sunflower head with opened florets was estimated by measuring the maximum diameter of the total capitulum and the minimum diameter of unopened florets. The area of the capitulum with opened florets (OPFL) was estimated (De Grandi-Hoffman and Martin 1993). The number of honeybees collecting nectar on the tagged capitulum was counted daily throughout flowering period, twice in the morning (09.00–11.00 h) and twice in the afternoon (14.00–17.00 h). Daily air temperature was also recorded.

22.3 Abstracting the information

We now discuss how we can summarise the information above in terms of response and explanatory variables.

Quantifying the information

A visual representation of the experiment in shown in Figure 22.1, although it must be emphasised that the real setup contained more transects (see above) that were not sampled. The five male-sterile transects are represented by open triangles

and two male-fertile transects by filled triangles. Each transect contained five capitulum. The number of honeybees on each of the 35 capitulum were counted twice in the morning and twice in the afternoon. In this chapter, we will use averages of the two morning and of the two afternoon counts. The reason for this will be explained later. Counts were started on day 7, and further measurements were taken on day 8, 9, 10, 11, 12, 13, 15 and 16. No measurements were taken on day 14. Therefore, the number of honeybees will be our response variable.

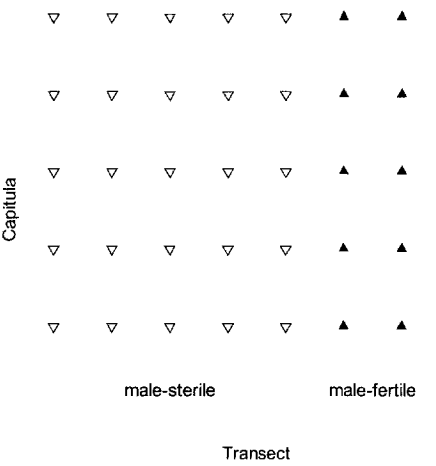

Figure 22.1. Sampling scheme. The graph is slightly misleading in the sense that the capitulum is not in a row. Male-sterile capitulum are labelled as Sex = 1 and male-fertile as Sex = 0.

We have an explanatory variable 'Transect' with values 1 to 7 denoting to which transect an observation belongs. This is a nominal variable and will allow us to check whether there are differences among the seven transects. It is also possible to quantify the difference between MS and MF more directly by using a nominal variable 'Sex' with a value of 0 if a sample was taken from an MF transect and 1 from a MS transect. Note that transects are nested within Sex, and we want to treat both nominal variables as fixed. Similarly, we created a nominal variable 'AMPM' with values 0 (AM) and 1 (PM) indicating at what time of the day the observation was made. All these nominal variables are used as explanatory variables. The day of sampling can also be used as a nominal variable. However, we coded the morning on day 7 as '1', the afternoon on day 7 as '2', the morning of day 8 as '3', etc. and named this explanatory variable 'Time'.

There are also a number of continuous explanatory variables. Air temperature was measured once per day. Another explanatory variable is the area of open florets (cm^2), denoted by 'PercFlower'. A pairplot (Chapter 4) of honeybees, temperature, PercFlower and time is shown in Figure 22.2. From the plot it can be

seen that at time 7–10 (days 10 and 11), the temperature was considerably lower. PercFlower had the highest values around day 11, which coincides with the lowest temperature. This might be an indication of collinearity (Chapter 4) between these two variables. However, the correlation was only 0.27, which does not provide enough justification to omit one of them. Figure 22.2 also shows that on the first day of the experiment, not all capitulum had the same area of open florets.

A summary of all explanatory variables is given in Table 22.1. As stated above, we wish to know the relationship between honeybee numbers and the explanatory variables. Therefore, the model we are after is of the form:

honeybees = F(Temperature, PercFlower, Transect, Sex, AMPM)

Where $F()$ stands for 'function of'. Using Transect and Sex in the same model caused numerical instability, and therefore, we used Transect in first instance, as it is more informative.

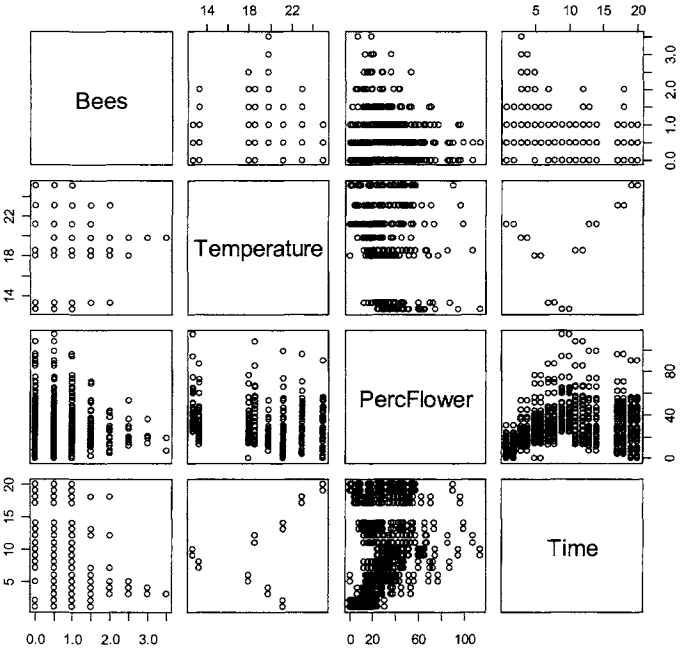

Figure 22.2. Pairplot showing the relationship among number of honeybees, temperature, PercFlower (the area of open florets) and Time.

Table 22.1. A summary of available explanatory variables.

Explanatory Variable	Remarks
Temperature	Continuous variable, one value per day
Days	Nominal variable which identifies day, with values 7–13, 15 and 16.
Transect	Nominal variable with values 1–7 to identify the seven transects.
Sex	Nominal variable with values 1–0 to identify male-sterile (transects 1–5) and male-fertile (transects 6 and 7).
PercFlower	Continuous variable measuring the area of the capitulum with open florets (cm^2).
AMPM	Nominal variable with values 0 (AM) and 1 (PM) identifying the time of the day that sampling took place.

22.4 First steps of the analyses: Data exploration

The first step in any analysis is to explore the data. Boxplots and Cleveland dotplots (Chapter 4) were made and confirmed that none of the continuous variables had outliers or extreme observations. A graph of the number of honeybees, temperature and PercFlower versus time has already been given in Figure 22.2. The number of honeybees observed is between 0 and 5. Previous studies have indicated that temperature is an important explanatory variable. A coplot of the number of honeybees versus temperature conditional on 'Transect' is given in Figure 22.3. To aid visual interpretation, a LOESS smoothing curve with a span of 0.5 was added. The graph suggests that in nearly all transects, the highest numbers of honeybees were observed at a temperature of approximately 20°C. Transect 7 (and also 6) seems to have a slightly different pattern, suggesting a possible interaction effect between temperature and Sex (or Transect). There seems to be only minor differences between AM and PM samples. The patterns in the graph suggest that if we want to include temperature in the models, we have three main options:

1. Create temperature classes and use this as a nominal explanatory variable in linear models.
2. Allow for an interaction between temperature (as a linear term) and any of the other variables.
3. Allow for a non-linear temperature effect.

The first option, converting temperature into classes and using this new variable as a nominal variable, is difficult as it will require arbitrary choices of which temperature values to combine. Including an interaction between transects and temperature as a linear term did not solve the problem. Therefore, we will model the honeybee data as a non-linear temperature effect using smoothing techniques (Chapter 7) together with a smoother-Transect or smoother-Sex interaction.

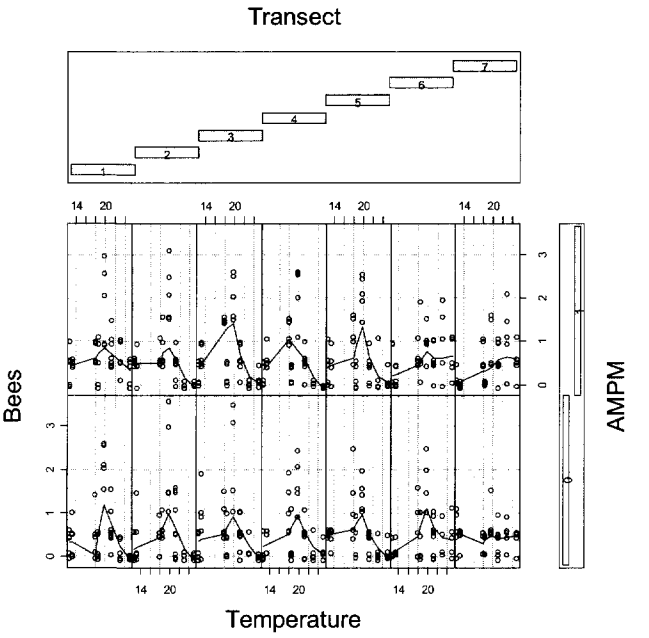

Figure 22.3. Coplot of the number of honeybees (vertical axis) and temperature (horizontal axis) conditional on Transect (blocks at the top) and time of day (blocks at the right hand side). The lower panels correspond to the AM data, and the upper row to the PM data. The left column contains the data from transect 1, and the rightmost column from transect 7. A LOESS smoothing line with a span of 0.5 was added to aid visual interpretation.

22.5 Additive mixed modelling

Because the data exploration indicated that non-linear relationships might be expected, an additive or generalised additive model can be used. The question is whether a Gaussian or Poisson model should be used. In other words, is it additive modelling or generalised additive modelling with a Poisson distribution we want to work with? Theoretically, the Poisson distribution should be applied, as the data are counts. However, additive models are slightly easier to understand and work with. On top of this, we took averages for the morning and for the afternoon data, resulting in non-integer data. Therefore, we will start simple (Gaussian distribution) and do more complicated things (e.g., Poisson) if necessary. The first additive model we could apply is of the form:

$$\text{Honeybees} = \alpha + f_1(\text{Temperature}) + f_2(\text{PercFlower}) + \text{Transect} + \text{AMPM} + \varepsilon \quad (22.1)$$

The noise component ε in equation (22.1) is assumed to be normally distributed with expectation 0 and variance σ^2. Equation (22.1) models honeybees as a smoothing function of temperature, a smoothing function of PercFlower, and the nominal variables Transect and AMPM. The temperature and PercFlower effects are assumed to be the same for each transect, but Figure 22.3 indicated that this might be unrealistic. There are a couple of other potential problems. Basically we have a time series at each capitulum in each transect. This gives us 35 time series of length 20 with two missing values (day 14), and therefore we have to allow for auto-correlation. Another problem is that the variance might be different per transect, or per sex, or per day, or per morning-afternoon. It is even possible to allow for an increase in spread for increasing temperature (Pinheiro and Bates 2000), but we will not go that far. This brings us into the world of mixed modelling, or since smoothers are involved: additive mixed modelling (Chapter 8). First of all, we have to improve the notation of the model:

$$\begin{aligned}\text{Honeybees}_{ijs} = \alpha + f_1(\text{Temperature}_s) + f_2(\text{PercFlower}_{ijs}) \\ + \text{Transect}_j + \text{AMPM}_s + \varepsilon_{ijs}\end{aligned} \tag{22.2}$$

Honeybees_{ijs} is the value of honeybees at time s in capitulum i in transect j. In additive modelling we assume that the ε_{ijs} are independently normally distributed. We can allow for auto-correlation between honeybees at time s and t (in the same capitulum) by assuming that

$$\varepsilon_{ijs} = \rho\varepsilon_{ij,s-1} + \eta_{ijs} \qquad \text{or equivalently:} \quad cor(\varepsilon_{ijs}\varepsilon_{ijt}) = \rho^{|s-t|} \tag{22.3}$$

The term η_{ijs} is normally distributed. This is one of the most simple auto-correlations structures and is called the auto-regressive correlation of order 1. The further two time points are apart, the lower the correlation. If the difference between two time points is $|s - t| = 1$, then the correlation between the two observations is ρ, if $|s - t| = 2$, then it is ρ^2, etc. In order to implement such a model in a software package, sequential observations need to be identified by a unique number (Pinheiro and Bates 2000, Crawley 2002, mgcv helpfile in R). This was the motivation to take averages per morning and per afternoon. Other auto-correlation structures are discussed in Chapters 16, 26 and 35. We can also allow for different variances per transect by assuming

$$\varepsilon_{ijs} \sim N(0, \sigma_j^2) \tag{22.4}$$

Instead of different variances per transect, we can try different variances per sex; just replace the index j for σ by a k, where $k = 1, 2$.

If you thought that this was complicated, well, there is more to come. We have now specified a model that contains fixed components (the smoothers, intercept and nominal variables) and random components (auto-correlation, different variances). The model selection process for such a model is similar as for mixed modelling (Chapter 8) and consists of three steps: (i) Start with a model that contains as many fixed terms as possible (a just beyond optimal model), (ii) using these fixed terms, find the optimal random structure, and (iii) for the optimal random structure, find the optimal fixed structure. A motivation for this approach was

given in Chapter 8. Starting with a 'just beyond optimal model' in terms of fixed components is simple for linear regression, just add as many main terms and interactions as possible, but for smoothing models this is slightly more complicated due to numerical instability. Based on the data exploration, it seems sensible to allow for a different non-linear temperature effect for each transect, and then we can determine in step 3 whether this is really necessary. We can do the same for PercFlower. Hence, our starting model for step 2 is

$$\begin{aligned}\text{Honeybees}_{ijs} = \alpha &+ f_1(\text{Temperature}_s) + f_1(\text{PercFlower}_{ijs}) + \\ &+ f_2(\text{Temperature}_s) + f_2(\text{PercFlower}_{ijs}) + \\ &+ f_3(\text{Temperature}_s) + f_3(\text{PercFlower}_{ijs}) + \\ &+ f_4(\text{Temperature}_s) + f_4(\text{PercFlower}_{ijs}) + \\ &+ f_5(\text{Temperature}_s) + f_5(\text{PercFlower}_{ijs}) + \\ &+ f_6(\text{Temperature}_s) + f_6(\text{PercFlower}_{ijs}) + \\ &+ f_7(\text{Temperature}_s) + f_7(\text{PercFlower}_{ijs}) + \\ &+ \text{Transect}_j \times \text{AMPM}_s + \varepsilon_{ijs} \\ \varepsilon_{ijs} \sim N(0,\sigma^2) \quad &\text{and} \quad cor(\varepsilon_{ijs}\varepsilon_{ijt}) = \rho^{|s-t|}\end{aligned} \tag{22.5}$$

The notation f_1(Temperature) means that a smoother of temperature is used for the data of transect 1. Technically, this is done using the 'by' command in the mgcv library in R; see also Chapter 35. Later, we will allow for different variances. A Transect × AMPM interaction term (plus main terms) was added.

To test whether we indeed need the auto-correlation coefficient, we applied a likelihood ratio test (Chapter 8) on a model with and without auto-correlation. It gave a test statistic of $L = 8.83$ with a p-value of 0.003, indicating that we need the auto-correlation. Note that we are not testing on the boundary (Chapter 8). Allowing for different variances per transect or per morning-afternoon caused numerical problems. As an alternative to check whether we need different variances in the model, we took the residuals from model (22.5) and plotted them against the transect. There were no clear differences in spread. The same holds for sex and AMPM. For this reason, we continue to step 3 with a random component that contains auto-correlation and one variance term σ^2.

In the third step, we search for the most optimal model in terms of fixed components. Most temperature smoothers in equation (22.5) were significantly different from 0 at the 5% level. The amount of smoothing was estimated by cross-validation (Chapter 7), and most temperature smoothers had 4 degrees of freedom (the maximum we allowed for due to the relatively small number of unique temperature values). However, all PercFlower smoothers had only one degree of freedom, and therefore we refitted the model using a linear PercFlower effect, and a PercFlower × Transect interaction term. Using a likelihood ratio test, the interaction term was not significant ($p = 0.19$), and the main term PercFlower was borderline significance ($p = 0.05$). The likelihood ratio test indicated that the AMPM × Transect interaction term was not significant ($p = 0.97$). We then dropped Transect (as a main term) from the model as it was the least significant term, followed by PercFlower and the smoother for transect 7. The remaining temperature

smoothers and the AMPM effect were all significantly different from 0 at the 5% level. Hence, the optimal model is given by

$$\begin{aligned}\text{Honeybees}_{ijs} = \alpha &+ f_1(\text{Temperature}_s) + f_2(\text{Temperature}_s) \\ &+ f_3(\text{Temperature}_s) + f_4(\text{Temperature}_s) + \\ &+ f_5(\text{Temperature}_s) + f_6(\text{Temperature}_s) + \text{AMPM}_s + \varepsilon_{ijs} \\ &\varepsilon_{ijs} \sim N(0,\sigma^2) \quad \text{and} \quad cor(\varepsilon_{ijs}\varepsilon_{ijt}) = \rho^{|s-t|}\end{aligned} \tag{22.6}$$

However, the shapes of the smoothers were all similar, and this raises the question of whether we should indeed use a smoother for each transect, or whether we can replace them by one overall temperature smoother, or two temperature smoothers for Sex (Table 22.1), or two smoothers for AMPM, or four smoothers for a Sex–AMPM combination. The model in which two temperature smoothers conditional of AMPM were used, was not better than the one in equation (22.6) as the likelihood ratio test gave a p-value of 0.53. The other comparisons gave:

Model	df	AIC	BIC	logLik	Test	L Ratio	p-value
1	16	1078.92	1150.05	−523.46			
2	6	1039.80	1066.48	−513.90	1 vs 2	19.11	0.04
3	12	1030.01	1083.36	−503.00	2 vs 3	21.79	<0.001
4	8	1010.83	1046.39	−497.41	3 vs 4	11.17	0.02

Model 1 is the model in equation (22.6). In model 2, we only use one temperature smoother for all transects. It is an improvement, but not by much. In model 3, we used four temperature smoothers, one for Sex = 0 and AMPM = 0, one for Sex = 1 and AMPM = 0, one for Sex = 0 and AMPM = 1 and one for Sex = 1 and AMPM = 1. Its AIC is lower than that of models 1 and 2. In model 4, we used two smoothers for temperature, one for Sex = 0 and one for Sex = 1. The likelihood ratio test shows that it is a significant improvement compared with model 3, and its AIC indicates that it is the most optimal model. It is given by

$$\begin{aligned}\text{Honeybees}_{ijs} = \alpha &+ f_{sex=0}(\text{Temperature}_s) + f_{sex=1}(\text{Temperature}_s) \\ &+ \text{AMPM}_s + \varepsilon_{ijs} \\ &\varepsilon_{ijs} \sim N(0,\sigma^2) \quad \text{and} \quad cor(\varepsilon_{ijs}\varepsilon_{ijt}) = \rho^{|s-t|}\end{aligned} \tag{22.7}$$

The numerical output for this model is as follows.

	Estimate	Std. Error	t-value	p-value
Intercept	0.530	0.030	17.391	<0.001
factor(AMPM)1	0.078	0.036	2.152	0.031

Approximate significance of smooth terms:

	edf	F	p-value
s(Temperature):Sex0	1.703	3.257	0.011
s(Temperature):Sex1	3.949	48.748	0.001

R-sq.(adj) = 0.292 Scale est. = 0.281 n = 630

The scale estimator is the variance of the noise. Because temperature has only a few unique values, we set the amount of smoothing for both smoothers equal to 4 degrees of freedom. Cross-validation (Chapter 7) was used to obtain the amount of smoothing, and it was estimated as 1.7 for the temperature smoother for the male fertile data (Sex = 0) and 3.9 for the male-sterile data (Sex = 1). The smoothing curves are given in Figure 22.4. For the male-sterile data, the largest numbers of honeybees are obtained for temperatures between 18°C and 21°C.

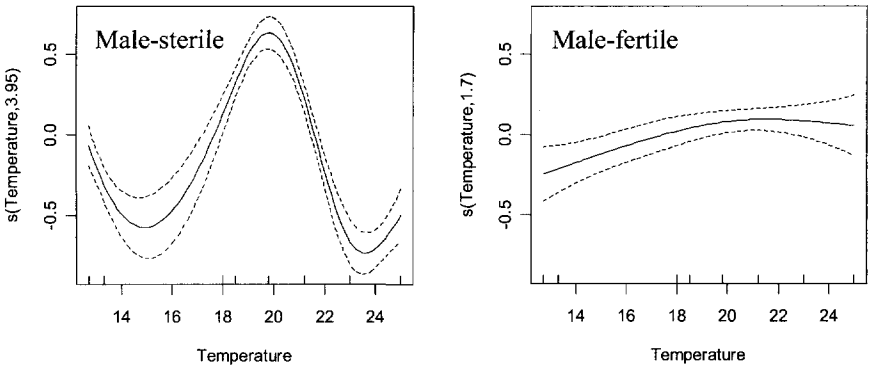

Figure 22.4. Smoothing curves for temperature for the male-sterile (Sex = 1) and male-fertile (Sex = 0) data. Dotted lines are 95% point-wise confidence bands.

The numerical output of our model presented above also indicates that AMPM (time of sampling) is significant but with a p-value close to 0.05. Hence, we should not consider this explanatory variable as important. The relevant model validation graphs to assess the assumptions of normality and homogeneity are given in Figures 22.5 and 22.6. Normality for the residuals of the male-sterile (Sex = 1) seems to be OK, although one can argue about normality of the male-fertile (Sex = 0) residuals. As to homogeneity (Figure 22.6), one can see patterns in these residuals, but this is due to the observed honeybees values only taking on a certain number of unique values, and this results in typical banding of the residuals. However, one can see a small increase in spread for larger fitted values. To understand why this is the case, we plotted residuals versus each nominal explanatory variable. There was no indication that for some levels of transects, sex or AMPM, the spread of residuals was larger. Figure 22.7 gives two examples. Figure 22.8 shows residuals versus temperature conditional on Sex. The larger residuals for the male-sterile data are obtained for temperature around 20°C. This is also the range where higher honeybee values were measured. It may be an option to allow for different residual spread per temperature regime, or use a Poisson distribution.

The model in equation (22.7) gave $\rho = 0.16$. This means that the correlation between two sequential samples is $\rho = 0.16$. If the gap is two units, then the correlation is $\rho^2 = 0.16^2 = 0.03$. These values are rather low.

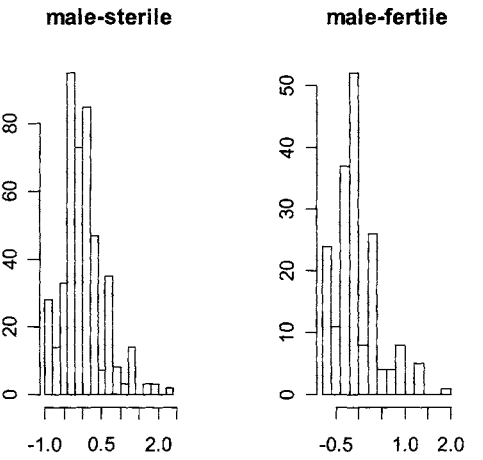

Figure 22.5. Histogram of residuals conditional on male-sterile (Sex = 1) and male-fertile (Sex = 0).

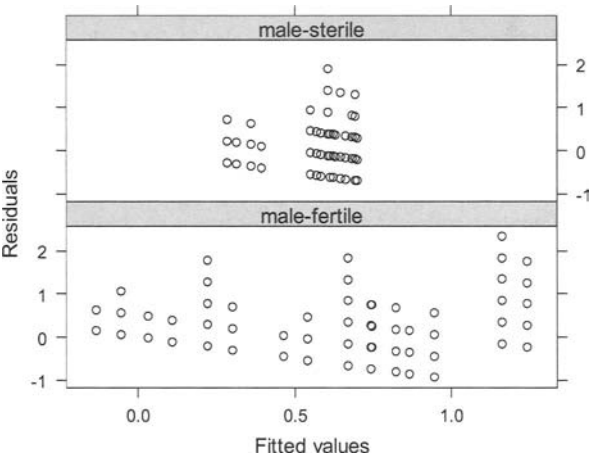

Figure 22.6. Residuals versus fitted values conditional on male-sterile (Sex = 1) and male-fertile (Sex = 0).

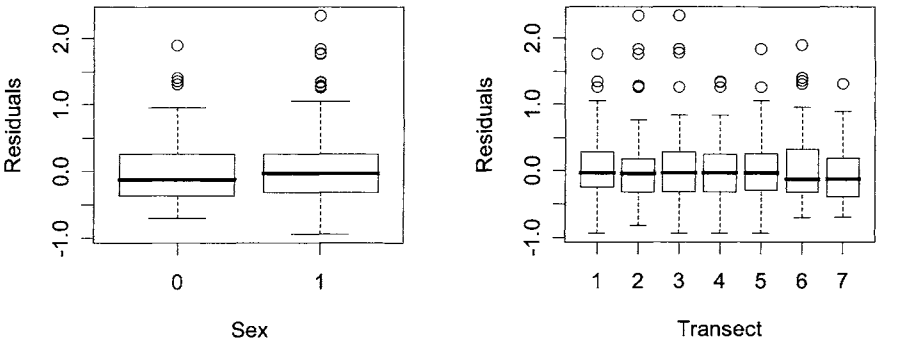

Figure 22.7. Model validation graphs showing that the spread in residuals is homogenous per level of Sex and Transect.

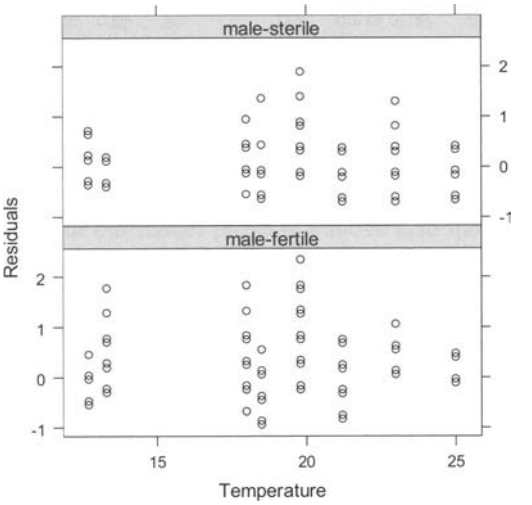

Figure 22.8. Residuals versus temperature conditional on Sex.

22.6 Discussion and conclusions

Initial exploration of the honeybee data indicated that non-linear relationships were evident for some explanatory variables and that the inclusion of linear interaction terms did not solve this problem. Therefore, we applied additive modelling. Indeed, this method showed a non-linear relationship of temperature.

The optimal model contained a temperature smoother for male-sterile (Sex = 1) and male-fertile (Sex = 0) data. However, the model showed a certain amount of heterogeneity, and this means that we have to be careful with interpreting p-values close to borderline significance, as is the case for AMPM. Further model improvement may be obtained by allowing for different variances per temperature regime. Alternatively, a Poisson model can be tried. One might wonder why we should make a lot of fuss about using different variances and correlation structures. The simple answer is, that by including these factors in our model, we can reduce the probability of obtaining a type I error.

The remainder of the discussion will concentrate on whether the additive model was the most optimal model or whether a generalised additive modelling approach should have been used. Interestingly, we did apply the GAM with a log link function and Poisson distribution, and we also converted the honeybee data to presence–absence data and applied a GAM with a binomial distribution and logistic link during the course of our data analysis. Both types of models can be extended with auto-correlation resulting in generalised additive mixed modelling. However, the conclusions resulting from these models were similar to the ones obtained by the additive mixed model.

The selected model indicated that the temperature effect on honeybees was different between MS and MF lines. Within the MS block, a uniform distribution of honeybees was found (different temperature patterns per sex were better compared to difference patterns per transect), which supports the finding of similar studies carried out by Skinner (1987). In this respect, planting schemes used in this study do not affect the seed production.

For the MS block, the maximum number of honeybees was estimated to occur at temperatures between 18°C and 21°C. Foraging activity declined outside this temperature range for the MS block, and temperature seems to be an important factor for honeybee activity. This observation has also been made in previous studies conducted both in this area and also in other areas where a decrease in foraging activity was observed when the temperature dropped by 2°C and 4°C respectively (Núñez 1982; Skinner 1987). At higher temperatures (> 22°C) honeybee activity decreased. However, we cannot say for sure that temperature is the driving factor for honeybee activity as a certain amount of higher dimensional collinearity between temperature, PercFlower, Sex and time exists. Repeating this experiment under different temperature regimes may give more information about the cause–effect relationship.

Acknowledgement

This work was part of the PhD project carried out by the first author who was particularly indebted to Enrique Bedascarrasbure and David de Jong for useful suggestions towards her guidance. Alfonso Lorenzo and Paul Ens kindly provided assistance in data gathering. This work was supported by SECYT-UNCPBA, and Don Atilio Company was acknowledged for supplying the seeds. We would like to thank Alex Douglas for valuable comments on an earlier draft.