Chapter 16
Negative Binomial GAM and GAMM to Analyse Amphibian Roadkills

A.F. Zuur, A. Mira, F. Carvalho, E.N. Ieno, A.A. Saveliev, G.M. Smith, and N.J. Walker

16.1 Introduction

This chapter analyses amphibian fatalities along a road in Portugal. The data are counts of kills making a Gaussian distribution unlikely; restricting our choice of techniques. We began with generalised linear models (GLM) and generalised additive models (GAM) with a Poisson distribution, but these models were overdispersed. To solve this, you can either apply a quasi-Poisson GLM or GAM, or use the negative binomial distribution (Chapter 9). In this particular example, either approaches can be applied as the overdispersion was fairly small (around 5), but with many ecological data sets it can be considerably larger, in which case the negative binomial GLM (or GAM) is the natural choice. As many textbooks give examples using quasi-Poisson GAMs and GLMs and only a few using the negative binomial, we decided to use the negative binomial distribution.

We chose GAM because the relationships between roadkills and explanatory variables were non-linear. We address issues like collinearity, residual patterns, and spatial correlations.

16.1.1 Roadkills

Since the second part of the twentieth century, roads have become a common feature in contemporary landscapes. For example, in North America alone, the road network has reached eight million kilometres and road construction is still increasing. Roads provide people and goods mobility, and are a central element in society (Forman et al., 2002). However, their impact on wildlife can be harmful as they (i) fragment populations, (ii) present barriers to dispersal as well as access to food and mates, and (iii) restrict gene flow. Also a large numbers of fatalities can occur as a result of animal–vehicle collisions.

A.F. Zuur (✉)
Highland Statistics Ltd., Newburgh, AB41 6FN, United Kingdom

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_16,

The life cycle of most amphibians has an aquatic phase, corresponding to reproduction and to tadpole development and metamorphosis; and a terrestrial phase, when individuals use adjacent territory for foraging, shelter, periods of dormancy or overwintering (Semlitsch and Bodie, 2003). High levels of roadkills occur when roads cross amphibian migration routes to and from spawning sites or during juvenile dispersal (Langton, 2002).

The data presented in this chapter come from a two-year study on vertebrate roadkills in a National Road of southern Portugal (IP2, stretch Portalegre-Monforte, 27 km long). The surveyed road has paved verges with two lanes and a moderate amount of traffic (less than 10,000 vehicles per day). Road surroundings are dominated by cork *Quercus suber* and holm oak *Q. rotundifolia* tree stands, named 'montado' and open land, including pastures, meadows, and fallows.

The road was inspected for amphibian roadkills every two weeks between March 1995 and March 1997. Surveys were made by a car slowly (10–20 km per hour) driving along the road on the hard-shoulder. Each animal found dead was identified to species level, whenever possible, and its geographic location, on UTM coordinates, was determined with help of detailed cartography (1:2000) of horizontal and vertical road profiles and aerial photographs. All carcasses were removed from the road to avoid double counting.

For data analysis purposes, the road was divided in 500 m segments. The response variable is the total number of amphibian fatalities per segment. All animals found dead on each segment were allocated to the coordinates of its middle point. Figure 16.1 shows an example of one of the species recorded.

Detailed digital maps of land use were made through interpretation of aerial photographs corrected with field observations. Explanatory variables were identified from these maps using a Geographic Information System. A list with all available explanatory variables and the abbreviations used is given in Table 16.1.

Fig. 16.1 *Pelobates cultripes*, one of the species that was used in our data. The photograph was taken by Marco Caetano

Table 16.1 List of explanatory variables and the abbreviation used in this chapter

Variable	Abbreviation
Open lands (ha)	OPEN.L
Olive grooves (ha)	OLIVE
Montado with shrubs (ha)	MONT.S
Montado without shrubs (ha)	MONT
Policulture (ha)	POLIC
Shrubs (ha)	SHRUB
Urban (ha)	URBAN
Water reservoirs (ha)	WAT.RES
Length of water courses (km)	L.WAT.C
Dirty road length (m)	L.D.ROAD
Paved road length (km)	L.P.ROAD
Distance to water reservoirs	D.WAT.RES
Distance to water courses	D.WAT.COUR
Distance to Natural Park (m)	D.PARK
Number of habitat Patches	N.PATCH
Edges perimeter	P.EDGE
Landscape Shannon diversity index	L.SDI

They include areas occupied by each land cover class, total length of roads and water courses on a 2,000 m strip centred on each road segment; landscape indexes (total number of patches; total perimeter of edges between different land cover classes; and landscape Shannon diversity index which relates to landscape heterogeneity); and distances from the segment centre to water and to the southwest limit of S. Mamede Natural Park (a mountain range NE-SW oriented that is known for its high levels of humidity and rainfall, where landscapes are particularly well preserved and are good examples of harmonious interactions between man and nature).

The underlying ecological question in this chapter is simple: is there a relationship between amphibian roadkills and any of the explanatory variables?

16.2 Data Exploration

The data were measured along the road, and the sampling positions are marked as dots in Fig. 16.2. The R code we used for this is as follows.

```
> library(AED); data(RoadKills)
> RK <- RoadKills
> library(lattice)
> xyplot(Y ~ X, aspect = "iso", col = 1, pch = 16,
         data = RK)
```

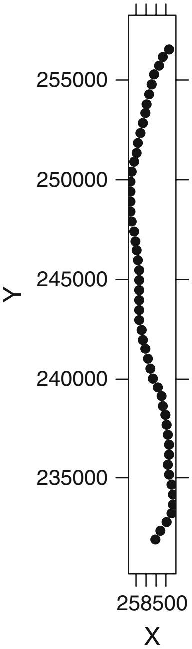

Fig. 16.2 Positions of the sampling points along the road

The first two commands access the data. We renamed the object `RoadKills` to `RK` as it is shorter. The `xyplot` command produces Fig. 16.2. The variable `D.PARK` is the distance (along the road) to the Natural Park, north of the sampling area. It therefore represents the distance (along the road) between each sampling point and the most northerly sampling site. If `D.PARK` had not been quantified, you would need to calculate the distance between each observation and the most northerly point yourself using the Pythagoras rule.

Using Cleveland dotplots (not shown here), pairplots, and initial GAM analyses, we decided to square root transform the explanatory variables POLIC, WAT.RES, URBAN, OLIVE, L.P.ROAD, SHRUB, and D.WAT.COUR.

There are 17 explanatory variables and only 52 observations. With such a low number of observations, we prefer not to use more than 5 or 6 explanatory variables, especially if we intend to use smoothing techniques. Furthermore, correlation coefficients between some of the explanatory variables are high. Because correlation coefficients only show pairwise correlations, we used variance inflation factors (VIF) to assess which explanatory variables are collinear and should be dropped before starting the analyses. VIFs were also used in Appendix A. We wrote our own R functions to calculate VIF values and these are part of our `AED` package. They are calculated with the following commands.

```
> RK$SQ.POLIC <- sqrt(RK$POLIC)
> RK$SQ.WATRES <- sqrt(RK$WAT.RES)
> RK$SQ.URBAN <- sqrt(RK$URBAN)
> RK$SQ.OLIVE <- sqrt(RK$OLIVE)
> RK$SQ.LPROAD <- sqrt(RK$L.P.ROAD)
> RK$SQ.SHRUB <- sqrt(RK$SHRUB)
> RK$SQ.DWATCOUR <- sqrt(RK$D.WAT.COUR)
> Z<-cbind(RK$OPEN.L, RK$SQ.OLIVE, RK$MONT.S,RK$MONT,
           RK$SQ.POLIC, RK$SQ.SHRUB, RK$SQ.URBAN,
           RK$SQ.WATRES, RK$L.WAT.C, RK$L.D.ROAD,
           RK$SQ.LPROAD, RK$D.WAT.RES, RK$SQ.DWATCOUR,
           RK$D.PARK, RK$N.PATCH, RK$P.EDGE, RK$L.SDI)
> corvif(Z)
```

The resulting VIF values are given in Table 16.2. As explained in Appendix A, a cut-off value of 5 or even 3 can be used to remove collinear variables; we used 3. To find a set of explanatory variables that does not contain collinearity, we removed one variable at a time, recalculated the VIF values, and repeated this process until all VIF values were smaller than 3. As a result, MONT, P.EDGE, N.PATCH, L.SDI, and SQ.URBAN were dropped. This means that we have 12 remaining explanatory variables. This is still a large number of variables!

We also present a scatterplot of all 12 selected explanatory versus the number of amphibian roadkills, see Fig. 16.3. We added a LOESS smoothing curve to help interpretation. The shape of the curves and the spread of the data around the smoothing curves do not look promising for good analysis. The only variables that seem to have a clear relationship with the roadkills are D.PARK and D.WAT.RES.

The R code for this graph is a bit a pain, but it is worth the effort. We basically need three columns of data. In the first column (`Killing12`), we copy and paste the variable containing the roadkills 12 times.

In the second column (`X12`), we concatenate the data of all 12 explanatory variables. The third column (`ID12`) needs to contain the name of the first explanatory

Table 16.2 Variance inflation factors for the full set of explanatory variables

Variable	GVIF	Variable	GVIF
OPEN.L	161.01	L.D.ROAD	4.41
SQ.OLIVE	34.44	SQ.LPROAD	3.38
MONT.S	3.96	D.WAT.RES	2.11
MONT	213.63	SQ.DWATCOUR	2.55
SQ.POLIC	3.89	D.PARK	2.91
SQ.SHRUB	3.32	N.PATCH	24.30
SQ.URBAN	14.03	P.EDGE	19.36
SQ.WATRES	1.98	L.SDI	10.02
L.WAT.C	3.64		

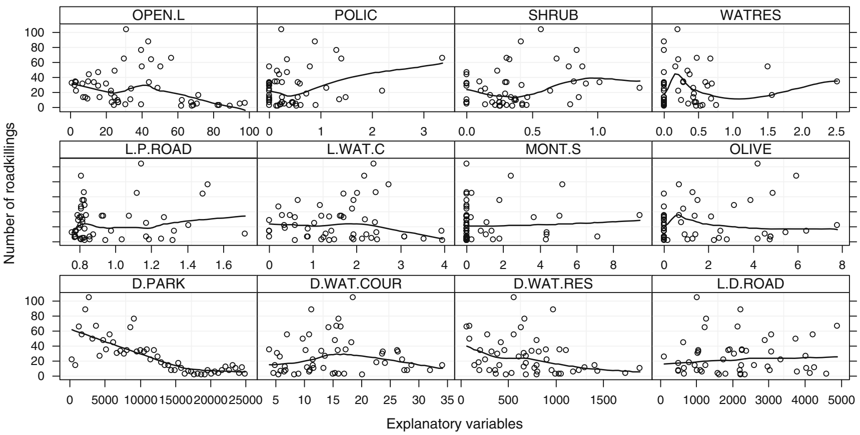

Fig. 16.3 Scatterplots of the number of amphibian roadkills (*y*-axis) against each of the 12 remaining explanatory variables. The heading in a panel indicates which explanatory variable is plotted along the *x*-axis. A smoothing (LOESS) curve was added in each panel

variable 52 times, the name of the second variable 52 times, etc. The rest is some fancy `xyplot` coding.

```
> X12 <- c(RK$OPEN.L, RK$SQ.OLIVE, RK$MONT.S,
            RK$SQ.POLIC, RK$SQ.SHRUB, RK$SQ.WATRES,
            RK$L.WAT.C, RK$L.D.ROAD, RK$SQ.LPROAD,
            RK$D.WAT.RES, RK$SQ.DWATCOUR, RK$D.PARK)
> Killings12 <- rep(RK$TOT.N, 12)
> I12 <- rep(c("OPEN.L", "OLIVE", "MONT.S", "POLIC",
             "SHRUB", "WATRES", "L.WAT.C", "L.D.ROAD",
             "L.P.ROAD", "D.WAT.RES", "D.WAT.COUR",
             "D.PARK"), each = 52)
> ID12 <- rep(I12, 12)
> library(lattice)
> xyplot(Killings12 ~ X12 | ID12, col = 1,
    strip = function(bg = 'white', ...)
    strip.default(bg = 'white', ...),
    scales = list(alternating = TRUE,
       x = list(relation = "free"),
       y = list(relation = "same")),
    xlab = "Explanatory variables",
    ylab = "Number of roadkillings",
    panel = function(x, y){
      panel.grid(h = -1, v = 2)
      panel.points(x, y, col = 1)
      panel.loess(x, y, col = 1, lwd = 2)})
```

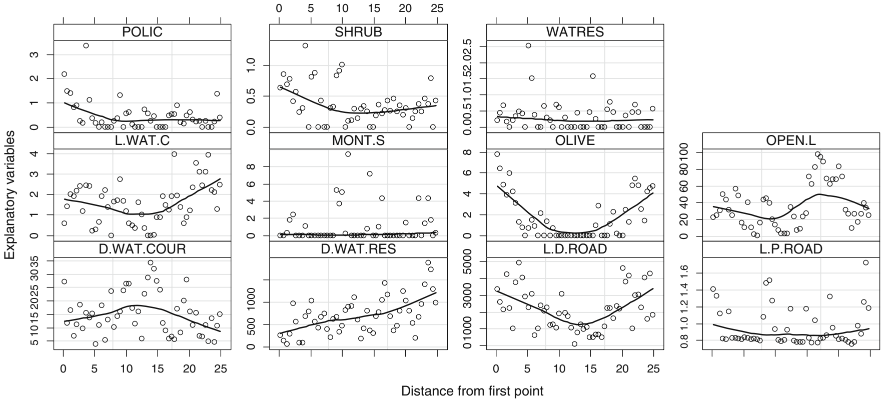

Fig. 16.4 Scatterplots of the explanatory variables (y-axis) versus the spatial variable D.PARK (distance from the first point expressed in km). The heading in a panel indicates which explanatory variable is plotted along the y-axis. A smoothing (LOESS) curve was added in each panel

The `strip` and `strip.default` options ensure that the boxes with the labels are white. The `scales` option allows for different ranges along the x-axes, but all y-axes have the same range. The `panel` function adds a grid, points, and a LOESS smoother; see also the bioluminescent case study in Chapter 17.

Now that we have this fancy R code, we would like to go one step back and focus on collinearity again. Figure 16.4 shows a similar plot as in Fig. 16.3, except that the explanatory variables are now plotted along the y-axis and the variable distance to the first point (D.PARK) along the x-axis.

We made this graph to get a feel for the spatial patterns of the explanatory variables. The variables OLIVE, D.WAT.RES, and L.D.ROAD show a clear pattern with D.PARK. Note that non-linear relationships are not picked up by the VIF. GAMs are rather sensitive to collinearity (Chapter 3), and we should not use D.PARK together with any of these three variables as they all represent the spatial position of the sampling locations. Because D.PARK has a clear ecological interpretation and it is easier to use in the independence verification later on, we decided to drop OLIVE, D.WAT.RES, and L.D.ROAD. The R code to produce Fig. 16.4 is similar to the code used for Fig. 16.3 and is not reproduced here.

16.3 GAM

The data exploration did not show any clear *linear* patterns between roadkills and the explanatory variables; so we need to move on to using a GAM. Furthermore, an initial GLM with a Poisson distribution and logarithmic link function gave an overdispersion of 5, and we therefore proceed with a GAM with a negative binomial distribution and logarithmic link function. The negative binomial distribution (Chapters 8 and 9) is useful if the variance is much larger than the mean. As it is for this data set, where the mean number of roadkills is 25.9 and the variance is 589.3. Recall from Chapter 9 that the negative binomial GAM is given by

$$RK_i \sim NB(\mu_i, k)$$

$$E(RK_i) = \mu_i \quad \text{and} \quad \text{Var}(RK_i) = \mu_i + \frac{\mu_i^2}{k}$$

$$\mu_i = e^{\alpha + f_1(\text{OLIVE}_i) + \ldots + f_{10}(\text{D.WAT.COUR}_i)}$$

RK_i is the number of amphibian roadkills at site i, where $i = 1, \ldots, 52$. The notation $f_j(X)$ stands for 'smoothing function of the explanatory variable X', and *NB* is a negative binomial distribution with mean μ_i and dispersion parameter k. The explanatory variables in the model are OPEN.L, MONT.S, SQ.POLIC, SQ.SHRUB, SQ.WATRES, L.WAT.C, SQ.LPROAD, SQ.DWATCOUR, and D.PARK. To fit the GAM, we can use the following R code.[1]

```
> library(mgcv)
> library(MASS)
> M1 <- gam(TOT.N ~ s(OPEN.L) + s(MONT.S) +
          s(SQ.POLIC) + s(SQ.SHRUB) + s(SQ.WATRES) +
          s(L.WAT.C) + s(SQ.LPROAD) + s(SQ.DWATCOUR) +
          s(D.PARK), family = negative.binomial(1),
          data = RK)
```

The package MASS is needed for the negative binomial distribution. The `(1)` in the code `negative.binomial(1)` means that the `gam` function will estimate the optimal dispersion parameter k.

The problem here is that this model gives an error message: `Model has more coefficients than data.` Because the model is applying cross-validation, some combinations of smoothers will use more than 52 degrees of freedom and we have only 52 observations. One option is to set an upper limit to the degrees of freedom; just extend the code with `s(OPEN.L, k = 4)` and do this for all terms in the GAM.

To find the optimal model, you can use shrinkage smoothers; these will also consider 0 degrees of freedom. If the model has multiple smoothers with 0 degrees of freedom, then you can drop them simultaneously. It is a faster alternative to a stepwise backward selection using, for example, the AIC or GCV. Shrinkage smoothers are obtained by using the *bs* option inside the *s* command, and specifying one of the shrinkage smoothers, for example, `s(OPEN.L, k = 4, bs = "ts")` or `s(OPEN.L, k = 4, bs = "cs")`. You can do this for all smoothers.

The `anova` command can be used to obtain *F*-statistics and approximate *p*-values for the smoothers. Results are not shown here, but most of the smoothers are not significant at the 5% level. Dropping the least significant smoother, and refitting the model until all terms are significant, is a highly confusing exercise for

[1] We used R version 2.6 and mgcv version 1.3–27. More recent versions of R and mgcv require a small modification to the code; see the book website (www.highstat.com) for updated code.

this data set. In one round, variables *x* and *y* are highly significant, and in another round, variable *z* is highly significant, but not *x* and *y*. This is clear evidence that there is still a certain degree of collinearity in the model. However, whichever model we applied, the variable D.PARK was always significant.

Another problem with this whole approach is that by setting an upper limit to the degrees of freedom, we may miss important variables that have a highly nonlinear effect. We therefore follow a different model selection approach of forward selection. We started with a GAM that used only one explanatory variable, fitted 9 different models, and compared their AICs (obtained by the `AIC` command). The model with the (by far) lowest AIC was the one with D.PARK. Its AIC was 352.5, whereas the second best model (with only OPEN.L) had an AIC of 423.2. We then fitted 8 GAMS, each with two explanatory variables, one which was D.PARK. The combination D.PARK and OPEN.L had the lowest AIC (340.3). We then continued with GAMs containing 3 smoothers: D.PARK, OPEN.L, and a third variable, but no combination gave a model with a better AIC and significant smoothers. Hence, following a forward selection approach, we end up with D.PARK and OPEN.L. To run this model in R, use

```
> M2 <- gam(TOT.N ~ s(OPEN.L) + s(D.PARK),
           family = negative.binomial(1), data = RK)
> anova(M2)
```

```
Family: Negative Binomial(28.7654)
Link function: log
Formula:
TOT.N ~ s(OPEN.L) + s(D.PARK)

Approximate significance of smooth terms:
            edf Est.rank      F  p-value
s(OPEN.L) 8.107    9.000  3.641  0.00282
s(D.PARK) 8.727    9.000 26.509 6.78e-13
```

The smoother for D.PARK is highly significant, and the OPEN.L smoother has a *p*-value of 0.003. The estimated smoothers are given in Fig. 16.5. The pattern for D.PARK can be seen in all sub-optimal models and also in the data exploration graphs. It shows a clear decrease along the gradient up to 18 km from the park and a slight increase after that distance. It is important to note that D.PARK reflects the distance to a mountain range where rainfall and humidity are higher than in surrounding flat areas. A decreasing gradient in these environmental conditions is expected to take place along the road as we move south. Moreover, both edges of the sampled road are in the boundaries of two localities (Portalegre and Monforte) with an agriculture matrix of small orchards and vegetable gardens. These are places where water availability in small ponds and channels for irrigation proposes is usually high. So the pattern found for D.PARK smoother is consistent with amphibian needs, concerning water availability. As higher amphibian abundances are expected in moist and wetter areas you would expect higher numbers of roadkills in these environments. The highest fatalities occurring in the first few kilometres of the sampled road probably also reflects the cumulative effect of water availability on

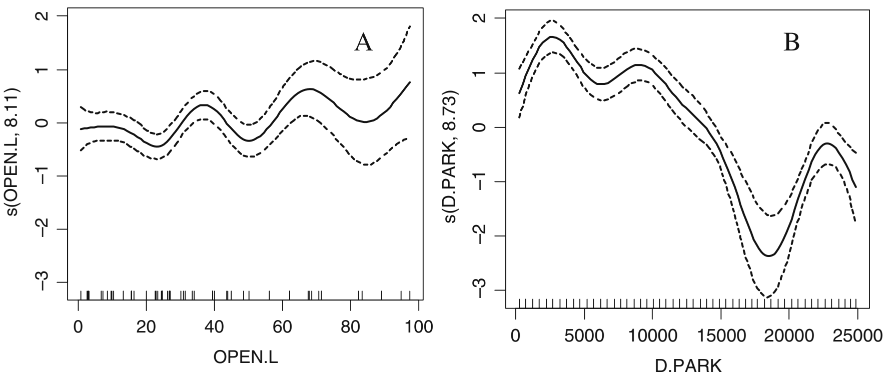

Fig. 16.5 **A**: Smoother for OPEN.L. **B**: Smoother for D.PARK. Both smoothers are from the optimal GAM model

mountain range and land use at this end of the road. At the other end of the road, only the land use is influencing the results.

The shape and interpretation of the smoother for OPEN.L is unclear. Based on the vertical ranges in both panels, the variable D.PARK contributes more to the fitted values than OPEN.L. We are rather tempted to drop OPEN.L from the model as we expect that the bumpy pattern may reflect some collinearity problems between D.PARK and OPEN.L. Figure 16.4 already indicated some sort of pattern between them, and perhaps, it was not a good idea to use them both.

The output of the `summary` command (not shown here) shows that this model explains 93.7% of the variation, and the dispersion parameter for the negative Binomial distribution is 28.7 (see also Chapter 9).

As part of the model validation, we also need to look at independence. Sites close to each other may have similar roadkill levels. To verify this, we can plot the residuals against the spatial coordinates. The problem is that the ranges along the horizontal and vertical axes in Fig. 16.2 make it rather difficult to do this. However, we can distort the shape of the picture a little bit by omitting the `aspect` option in the `xyplot` command:

```
> E <- resid(M2, type = "pearson")
> I <- vector(length = length(E))
> I[E < 0] <- 1
> I[E >=0] <- 16
> library(lattice)
> xyplot(Y ~ X, cex = 2 * abs(E) / max(abs(E)),
         pch = I, col = 1, data = RK)
```

The resulting graph (Fig. 16.6) shows the distorted road, but it allows for a visual inspection of the residuals. There are no immediate clear clusters of negative or positive residuals, and there are no clear clusters of large (in absolute sense) values.

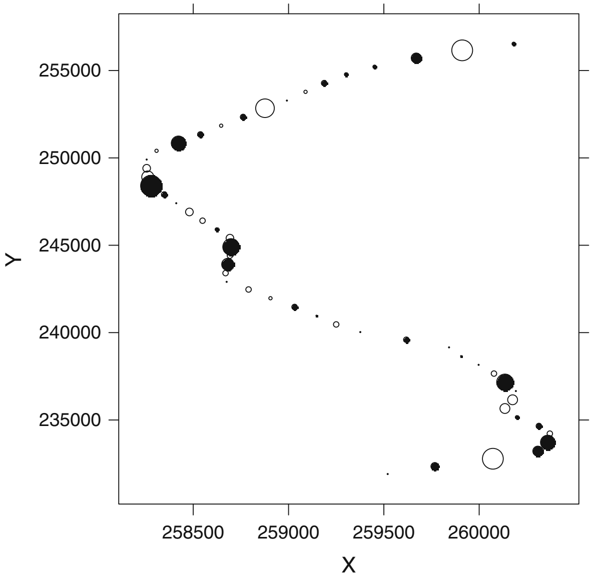

Fig. 16.6 Residuals of the optimal GAM model are plotted versus the spatial coordinated of the sites. The scales of the axes were distorted to make all the *dots* visible. The larger the *dot*, the larger the residual (in absolute sense). *Filled circles* have a negative sign for the residuals and the open circle positive values. There should be no clustering of positive or negative residuals or clustering of large values

Another part of the model validation process consists of plotting the residuals of the optimal GAM model versus all explanatory variables; see Fig. 16.7. You should not be able to see any patterns in these graphs. In this case, there are some patterns, but they are not strong enough as judged by the AIC (adding any of these terms as a smoother to the model results in higher AIC or non-significant smoothers) to be of concern.

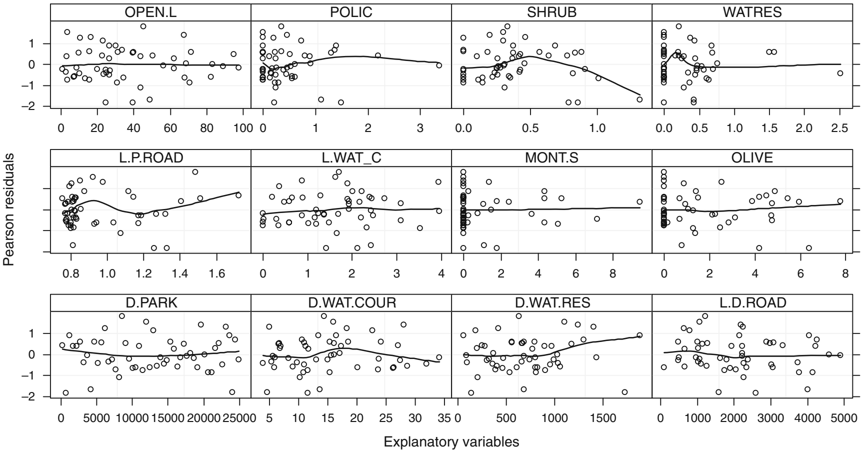

Fig. 16.7 Residuals of the optimal GAM model versus all explanatory variables. A LOESS *smoothing curve* was added

The R code to produce Fig. 16.7 is similar to the code used for Figs. 16.3 and 16.4 and is not presented here. Just replace the first column for the vertical axes by the residuals obtained by `residuals(M2, type = "pearson")`.

16.4 Understanding What the Negative Binomial is Doing

Before continuing with the GAMM section, we show what the negative binomial model is doing. In this discussion, it is easier to use a model that only contains D.PARK. Besides, we were not impressed with the role of OPEN.L anyway. The following code fits the negative binomial GAM with only D.PARK as a smoother.

```
> M3 <- gam(TOT.N ~ s(D.PARK), data = RK,
            family = negative.binomial(1))
```

The estimated smoother has a similar pattern as Fig. 16.5B. The estimated parameter *k* is 11.8. To better visualise what this model is doing, we will draw the fitted values on the real scale, add confidence bands around the fitted values, and superimpose values from a negative binomial distribution with the mean value given by the fitted GAM values and the dispersion parameter of 11.8. The following code achieves this, and the results are given in Fig. 16.8.

```
> M3Pred <- predict(M3, se = TRUE, type = "response")
> plot(RK$D.PARK, RK$TOT.N, cex = 1.1, pch = 16,
      main = "Negative binomial GAM",
      xlab = "Distance to park",
      ylab = "Number of road killings")
```

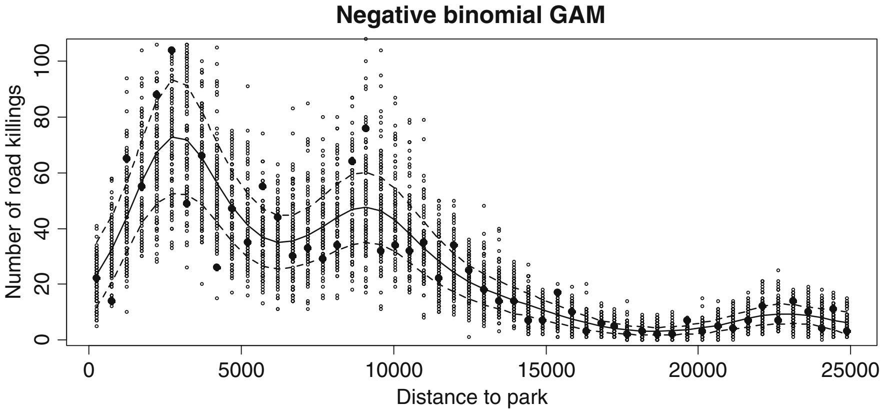

Fig. 16.8 Fitted values (*solid line*) and approximate 95% confidence bands (*dotted lines*) for the mean obtained by the negative binomial GAM. The large *filled dots* are observed values, and the *small dots* are 100 random samples per site taken from a negative binomial distribution

```
> I <- order(RK$D.PARK)
> lines(RK$D.PARK[I], M3Pred$fit[I], lwd = 2)
> lines(RK$D.PARK[I], M3Pred$fit[I] +
            2 * M3Pred$se.fit[I], lty = 2, lwd = 2)
> lines(RK$D.PARK[I], M3Pred$fit[I] -
            2 * M3Pred$se.fit[I], lty = 2, lwd = 2)
> for (i in 1:52){
   y <- rnbinom(100, size = 11.8, mu = M3Pred$fit[i])
   points(rep(RK$D.PARK[i], 100), y, cex = 0.5)}
```

The `predict` command takes the results from the GAM model and predicts fitted values on the response scale. The `plot` command sets up the graph with the observed values (`pch = 16` produces filled circles). The `lines` commands are used to draw the fitted values (solid thick line in the middle) and approximate pointwise 95% confidence bands (thick dotted lines) for the mean. So far, we have used no new R code; all this was used earlier in Chapter 3. The new bit comes now. The loop takes the fitted values at site *i* (given by μ = `M3Pred$fit[i]`), and using a parameter of $k = 11.8$, it draws 100 values from a negative binomial distribution, which are superimposed on the graph with the `points` command at the value of D.PARK for each the site. It gives an impression of the likely (road killing) values at any particular site. Unfortunately, it is rather difficult to draw the negative binomial density curves on top of this graph as we did in Chapter 2 for the Normal distribution. Figure 16.9 shows density curves for different values along the D.PARK gradient, and you can imagine these density curves on top of the fitted values in Fig. 16.8.

Density curves from a Poisson GAM are considerably less wide.

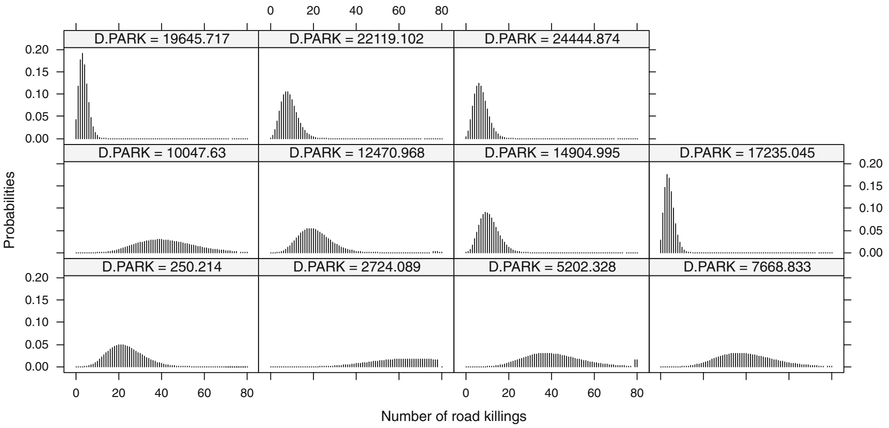

Fig. 16.9 Examples of negative binomial distributions. The density curves have a parameter of $k = 11.8$, and the mean value μ was taken from the fitted values at certain arbitrary chosen values along the D.PARK gradient

16.5 GAMM: Adding Spatial Correlation

In the previous section, we applied a GAM and found that the optimal model contains a smoother for D.PARK and OPEN.L. The residuals were plotted against the spatial coordinates, and we could not see any clear spatial patterns in these residuals. Instead of making this plot, we can also make a variogram of the residuals. The easiest option is to use the function `Variogram` from the `nlme` package, which is designed to work with the `gls`, `lme`, and `gamm` functions. All we need to do now is to rerun the GAM as a GAMM, just like we reran the linear regression with a GLS in Chapter 4 and use the `Variogram` function on its results. The code is given below and the resulting graph in Fig. 16.10, where there is a minor indication that points close to each other are more similar than points further separated along the road (this can be seen from a slightly increasing pattern in the variogram). However, one can equally well argue that the points form a horizontal band of points, indicating independence.

```
> library(nlme)
> RK$D.PARK.KM <- RK$D.PARK / 1000
> M4 <- gamm(TOT.N ~ s(OPEN.L) + s(D.PARK), data = RK,
             family = negative.binomial(theta = 11.8))
> M4Var <- Variogram(M4$lme, form =~ D.PARK.KM,
                     nugget = TRUE, data = RK)
> plot(M4Var, col = 1, smooth = FALSE)
```

It is also possible to add a spatial correlation structure to the model and see whether it improves anything. This can easily be done by using one of the available correlation structures `corExp`, `corSpher`, `corRatio`, or `corGaus`. According to the protocol defined in Chapters 4 and 5, we should start with a model containing smoothers of all explanatory variables. However, such a model did not converge. We therefore used the optimal model from the GAM with D.PARK and OPEN.L and

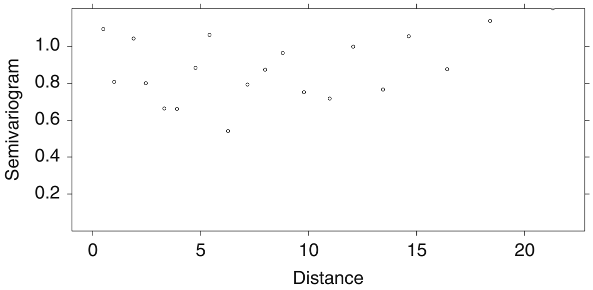

Fig. 16.10 Variogram of the residuals of the optimal GAM model with D.PARK and OPEN.L as smoothers. The variogram indicates independence as the points seem to form almost a cloud of horizontal points. Spatial correlation is present if we can see an increasing pattern up to a certain level

added a spatial correlation structure. Of all the spatial correlation structures, only the corGaus converged. This model is fitted by the following R code.

```
M5 <- gamm(TOT.N ~ s(OPEN.L) + s(D.PARK), data = RK,
           family = negative.binomial(theta = 11.8),
           correlation = corGaus(form =~ D.PARK.KM,
           nugget = TRUE))
```

However, the estimated range is close to 0, meaning that the chosen correlation structure makes no sense. When fitting these models without the smoother for OPEN.L, most convergence problems disappeared, but the spatial correlation functions gave rather different ranges and sills. It may be better to choose the range and sill interactively based on the residuals from the optimal GAM models in Fig. 16.10.

16.6 Discussion

This chapter provides an example of a data analysis that shows how important it is to do a good model validation and how confused you can get from a GAM if you ignore collinearity before starting the analysis. Having a large number of explanatory variables that are all linked to the spatial position of the sites (distance to water, distance to a park, etc.) does not help.

The results indicate that the variable D.PARK is the most important variable explaining amphibian roadkills. The optimal GAM model also contained OPEN.L, but the shape of the smoother is difficult to interpret and the model contained a small (tiny) amount of residual spatial correlation.

16.7 What to Write in a Paper

You need to emphasise the data exploration and problems with collinearity. You also need to discuss the interpretation of the variable D.PARK and why it was used as an explanatory variable. It is important to explain that there is no violation of independence in the model.