Chapter 18
Additive Mixed Modelling Applied on Phytoplankton Time Series Data

A.F. Zuur, M.J Latuhihin, E.N. Ieno, J.G. Baretta-Bekker, G.M. Smith, and N.J. Walker

18.1 Introduction

This chapter looks at a data set where our first reaction was: 'How in heavens name are we going to analyse these data?' The data consist of a large number of phytoplankton species measured at 31 stations in Dutch estuarine and marine waters. Measurements took place 0–4 times per month from 1990 until present (2005). Environmental data (e.g. temperature, salinity, etc.) were also measured, albeit sometimes at different sampling times! The statistical analysis of these data is complicated for several reasons:

1. Environmental variables and phytoplankton variables were not always measured at the same time.
2. There may be temporal correlation, there may be spatial correlation, and both correlation structures may be complicated.
3. The data contain a large number of species.
4. The data are irregularly spaced.
5. There may be heterogeneity over time (e.g. more variation in summer than in winter).
6. Trends over time and in space may be non-linear.
7. The phytoplankton data were counted by different laboratories.

This chapter is a spin-off from a technical report produced by the first two authors of this book for Rijkswaterstaat – Centre for Water Management, a Dutch governmental department. In that report, univariate methods were applied on aggregated phytoplankton series. The motivation to use aggregated data was to reduce the large number of zeros in the original data. An alternative statistical analysis is to apply multivariate methods like the Mantel test, BIOENV and ANOSIM; see Clarke and Warwick (1994), Legendre and Legendre (1998), and Zuur et al. (2007) for details.

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_18,

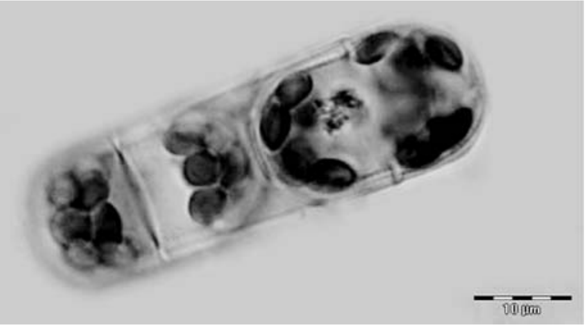


Fig. 18.1 *Melosira nummuloides* under the microscope; one of the small diatom species of the DIAT1-group (Photo C. Brochard – Koeman en Bijkerk bv)

The problem with these multivariate methods is that the permutation methods used to assess statistical significance ignore the temporal and spatial correlation structures in the data. Here, we follow the technical report approach and focus on a group of aggregated phytoplankton species. To save space, we only use one group: small diatoms (between 0 and 1,000 μm^3). These will be denoted by DIAT1 (Fig. 18.1). Other groups are not considered in this chapter.

As from Section 18.3, we describe an analysis that, in theory, can cope with some of the problems. It should be noted, however, that different analysis strategies are possible, but may give different results and that our chosen approach can be improved on and should be considered as a first attempt. However, given the complexity of the data, any statistical method will have serious difficulties with these data.

18.1.1 Biological Background of the Project

Marine biodiversity is under significant anthropogenic pressures such as physical engineering, physical and chemical pollution, eutrophication (enrichment with nutrients), and the introduction of invasive species. Eutrophication due to anthropogenic nutrient loading has greatly impacted ecological processes in marine waters, and therefore, a lot of effort has been put into reducing nitrogen and phosphorus discharges. To detect the effectiveness of such policy actions in the Netherlands, the Dutch national monitoring programme aims to provide the required information. In addition to a physical and chemical monitoring programme that had been running for several decades, Rijkswaterstaat began a biological monitoring programme for surface waters in the early 1990s. The primary goal of this programme is to provide biological information, especially in relation to long-term changes. The marine biological monitoring programme has been designed to assess ecosystem functioning and food-web relationships determine the structure of this system. Phytoplankton, the free-living, drifting, and mainly photosynthetic organisms in aquatic systems, is the major producer and forms the basis of the marine food web. Higher organisms such as benthic fauna, fish, and sea birds are all indirectly dependent on phytoplankton. Hence, information about the status of the phytoplankton community is essential to assess ecosystem functioning. In general, the growth of phytoplankton is regulated by underwater light and nutrient availability. Species composition and abundance of phytoplankton vary from season to season. The growth season usually starts with a bloom of diatoms in (early) spring followed by the blooming of *Phaeocystis* sp and in summer, blooms of (dino-)flagellates. Moreover, distinct differences can be detected between the various water bodies (in this chapter we call them areas), both in terms of species composition and abundance. The nutrient regime of the Dutch estuaries, the Delta in the south and the Wadden Sea and Ems estuary in the north, is mainly influenced by freshwater discharges with strongly elevated levels of nitrogen (N) and phosphorus (P) originating from farmland. For the North Sea ecosystem on the contrary, the much more oligotrophic Atlantic Ocean is the main source of nutrients. It seems reasonable to assume there is still some influence of riverine water in the coastal zone, but this rapidly decreases when going to the open sea. In general, increasing salinity goes hand in hand with decreasing nutrient concentrations. Nutrient enrichment usually results in an increase of phytoplankton biomass and often coincides with shifts in phytoplankton species composition. This latter phenomenon is due to different characteristics between individual algal species which have different storage capacities, nutrient uptake kinetics, etc. For example, silicon is an essential nutrient for diatoms, which is a major group of algae. But concentrations of this element seem unaffected by human activities. This implies that, due to eutrophication with N and P, it is likely that the species composition will change in the direction of increased abundance of algal species not dependent on silicon for their growth.

The mechanisms and implications of eutrophication for freshwater systems are reasonably well understood, but this is not the case for marine ecosystems, and the response of marine ecosystems to eutrophication is less predictable. It is suggested (Cloern, 2001) that the interaction between all the parameters characterizing a marine ecosystem – e.g. tidal regime, turbidity, depth, and biomass of benthic suspension feeders – play an important role. More precisely, the complex interaction of all physical and biological attributes operating together seems to act as a filter to modulate the response of an ecosystem to nutrient enrichment. As a result, some estuarine-coastal ecosystems appear to be highly sensitive to change in nutrient inputs, while others appear to be more resistant.

The main underlying question in this study is whether there are trends visible in the phytoplankton community, and if any, what trend, and whether there is a relationship with environmental variables. The rest of this chapter now follows the structure of the original technical report. Since April 1990, the species composition of phytoplankton has been monitored at 31 stations, which have been aggregated into ten different areas. Figure 18.2 presents the locations of the stations and defines the areas.

Water samples were collected from each sampling station and preserved with Lugol's solution, while at a limited number of stations a duplicate series was also counted live to improve identification. Samples were counted using an inverted microscope, and densities were subsequently calculated as number per liter. The

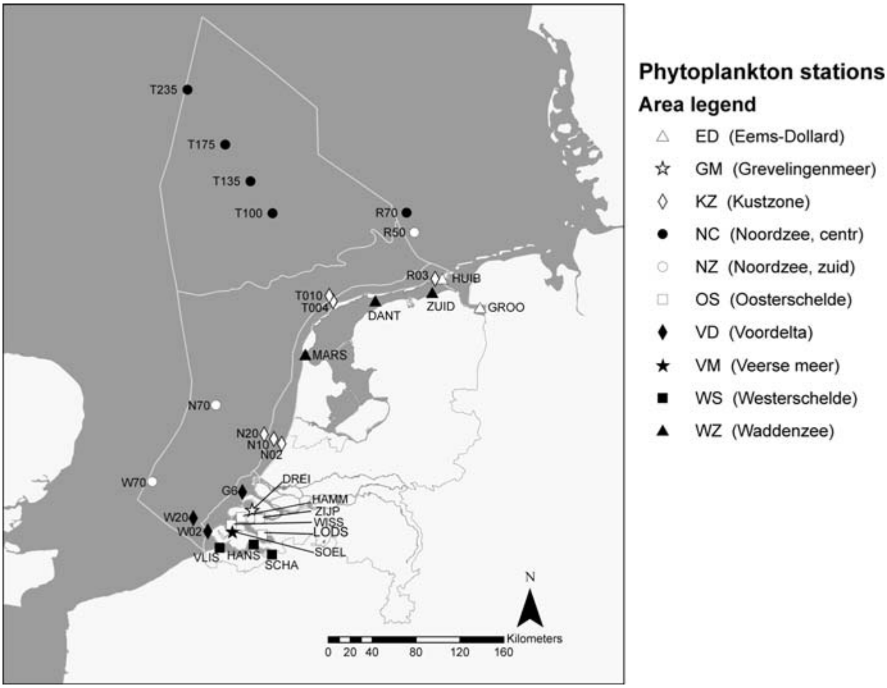


Fig. 18.2 Station locations and the area boundaries

sampling frequency depended on the season: monthly during winter and fortnightly during summer. All stations were sampled just below the water surface. When the water column is stratified, and this usually occurs on some, mainly offshore, stations in the summer, then samples were also taken at the thermocline and a few metres above the bottom with a Rosette sampler. This study is based on the Lugol-preserved samples taken close to the water surface.

To improve consistency in the phytoplankton data over time, the first year of the time series, 1990, has been skipped because the phytoplankton monitoring only started in April that year. Moreover, some taxa were left out because they were not consistently counted over time, and many taxa that can be individually identified microscopically today, were lumped together in the early years. Thus, the initial number of about 600 different taxa was reduced to a dataset that contains 175 taxa from 1991 on. As explained above, in this chapter we only focus on an even more aggregated group of small diatoms.

The environmental variables that were used in the technical report are dissolved inorganic nitrogen (DIN = ammonium, nitrite plus nitrate), dissolved inorganic phosphorus (DIP), silicon, total nitrogen, total phosphorus (all in μmol/l), salinity, temperature (Celsius), Secchi depth (dm), and suspended matter (mg/l). Here, we only use DIN and temperature for illustrative purposes.

18.2 Data Exploration

Instead of starting with a discussion of the statistical modelling approach, we first apply a data exploration as it spreads some light on the type of data we are working with. We arbitrarily chose DIN for this. Standard data exploration tools for multiple time series are a `xyplot` and `bwplot` (both from the `lattice` package), boxplots, and Cleveland dotplots. Pairplots are less useful for this particular example, because the DIAT1, DIN and temperature data were not sampled at the same time. Figure 18.3 shows a graph of log-transformed DIN values versus time, for each station. The following code was used to make the graph.

```
> library(AED); data(RIKZDATAEnv); library(lattice)
> RIKZ2 <- RIKZDATAEnv #Saves space
> RIKZ1 <- RIKZ2[RIKZ2$Year > 1990, ]
> I <- !is.na(RIKZ1$DIN)
> RIKZ <- RIKZ1[I, ]
> RIKZ$LDIN <- log(RIKZ$DIN)
> RIKZ$fStation <- factor(RIKZ$Station)
> RIKZ$MyTime <- RIKZ$Year + RIKZ$dDay3 / 365
> xyplot(LDIN ~ MyTime | Station, data = RIKZ,
         xlab = "Time", col = 1, type = "h",
         strip = function (bg = 'white', ...)
         strip.default(bg = 'white', ...))
```

The first series of commands are used to access the data, discard data from 1990, remove rows with missing values (it makes the model validation easier), and it applies a logarithmic transformation. The variable `MyTime` is used to provide sensible axis labels. We used the option `type = "h"` to ensure that observations

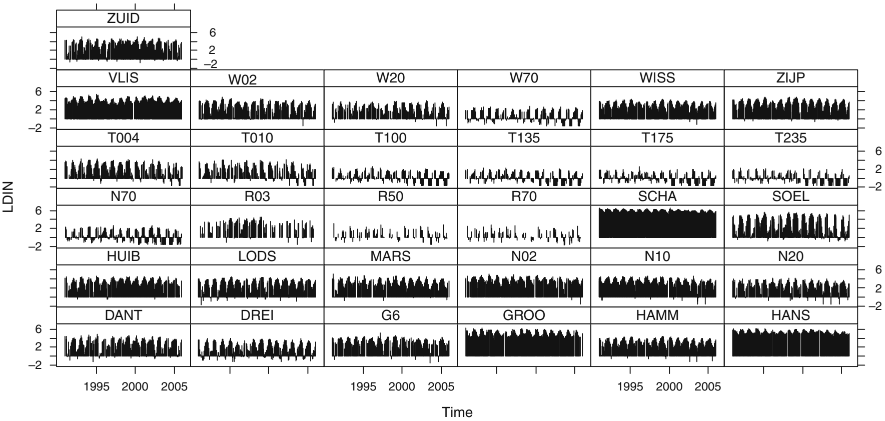


Fig. 18.3 Graph of log-transformed DIN (vertical axes) versus time (horizontal axes) per station. *Vertical lines* are used to show values

are not presented as points (in which case the graph becomes one big cloud of observations) or lines (missing values are not shown properly).

The graph shows there are differences per station in terms of absolute values, variation and number of missing values. The same graph for the untransformed data showed even more differences in absolute values per station. The question is now, what are we going to do with this? One option is to use log-transformed data, and another option is to standardise each time series. The latter option means that the data at stations like R50, with low values, become equally important as stations like ZUID with much higher values. Because DIN is a measure of available nutrients, we prefer *not* to make all series equally important, hence our choice for the log-transformation. The fact that we use a logarithmic transformation, and not a square root, is based on the range of the data.

For the same reason, a logarithmic transformation was applied on DIAT1. There was no need to transform temperature.

Figure 18.4 shows a conditional boxplot of the log transformed DIN values conditional on station. It shows that there are considerable differences between the stations, indicating that whatever model we apply, the term station has to be included.

We can either do this as a fixed term and pay the price of 30 regression parameters or use it as a random effect. The latter option makes more sense as it will allow us to make a general statement. If station is used as a fixed term, our statements and conclusions only hold for these particular 31 stations. Other advantages of using station as a random effect is that is saves a large number of parameters (one variance term versus 30 regression parameters) and it introduces a correlation structure between the observations at the same station (albeit it is the rather basic, compound symmetrical correlation structure; see Chapter 6).

Another aspect we need to take into account is seasonality. Figure 18.5 shows boxplots of log-transformed DIN values for each area conditional on month. Note

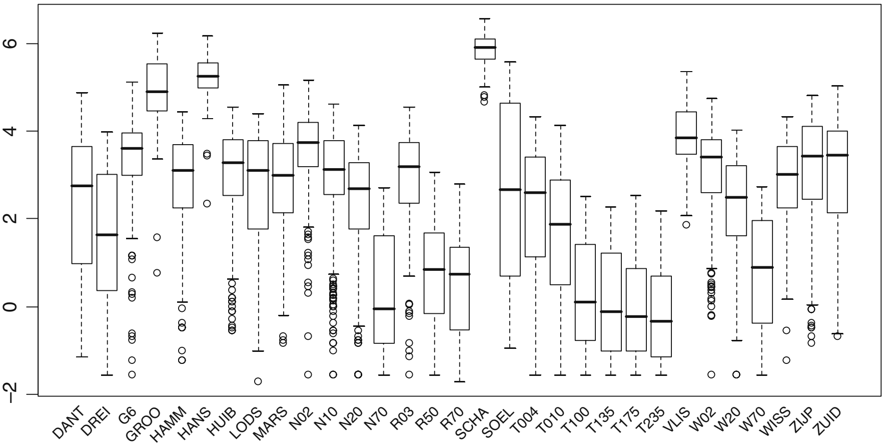


Fig. 18.4 Boxplot of log-transformed DIN conditional on Station. Mean values differ considerably per station, indicating that the term station has to be used in the model

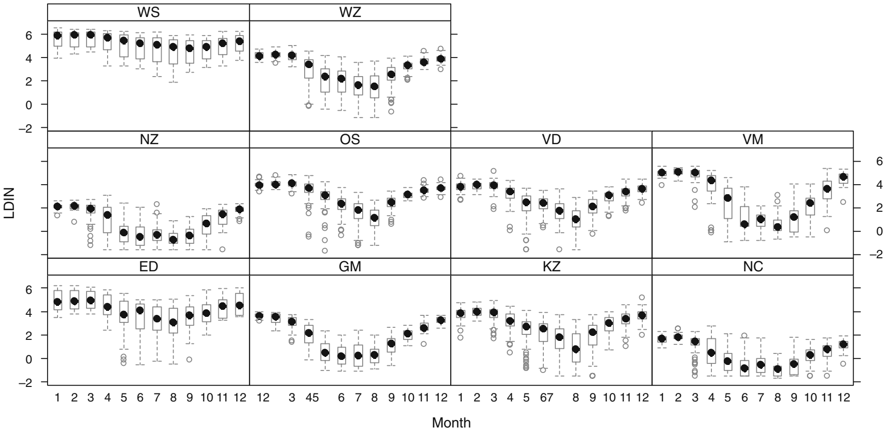


Fig. 18.5 *Boxplots* of log-transformed DIN versus month per area. Each area consists of various stations. There is a clear seasonal pattern that differs per area

there is a clear seasonal pattern at all areas, but not all patterns are identical. This means that we may have to include an interaction term between seasonal effects and area in the model. Note that for some months, in some areas, there is a group of observations outside the boxplot. For untransformed data, we saw similar patterns. We will discuss these points later.

The R code to produce Figs. 18.4 and 18.5 is given below. The code is self-explanatory, and in case of any doubts, consult the help files.

```
> #Figure 18.4:
> RIKZ$fMonth <- factor(RIKZ$Month)
> bwplot(LDIN ~ fMonth | Area, data=RIKZ, xlab="Month",
   strip = function(bg = 'white', ...)
   strip.default(bg = 'white', ...), col = 1,
   scales = list(rot = 45, cex = .6))
```

```
> #Figure 18.5
> boxplot(LDIN~fStation, data = RIKZ, xaxt = "n")
> text(1:31, par("usr")[3] - 0.25, srt = 45, adj = 1,
      labels = levels(RIKZ$fStation), xpd = TRUE,
      cex = 0.75)
```

18.3 A Statistical Data Analysis Strategy for DIN

If environmental and phytoplankton data had been measured at the same time, the following model could be our starting point.

$$\text{Phytoplankton data}_s = f(\text{environmental data}_s) + \text{noise}_s$$

The notation *f*() stands for '*is a function of*', well at least for the moment. The index *s* represents the sampling time (e.g. day). However, most statistical software routines will drop each observation where at least one of the variables is missing. This means that if the response and environmental variables are not sampled at the same time, you may end up with no data at all. Hence, conventional methods like linear regression, generalised linear modelling (GLM), or generalised additive modelling (GAM) cannot easily be used to model the function *f*. So, the first item from our list of problems, given in Section 18.1, is already causing a major headache. Our solution is to use a different model with the form:

$$\text{Variable} = f(\text{Time}) + \text{noise}$$

This means that each variable, either environmental or phytoplankton, is modelled as a function *f* of time, which represents the trend. We will apply this model on each variable, and compare the estimated trends. The advantage of this approach is that we do not have to compare the environmental and phytoplankton data directly, but just their temporal trends. These trends are smoothing functions over time and have values at the same time points. This allows us to compare the trends of different variables. We should note that our prime aim is to compare long-term trends and not the short-term (or, within-year) variation.

However, the bad news is this model is still complex. We still need to be able to deal with heterogeneity (more variation in summer months than in winter months), spatial and temporal correlation, non-linear trends, etc. A method that can potentially cope with this complexity is mixed modelling or if we allow for non-linear (or better: non-parametric) trends: additive mixed modelling.

In linear mixed modelling and additive mixed modelling, the model selection approach should follow a protocol that roughly contains the following steps (Fitzmaurice et al., 2004; Diggle et al., 2002; Chapters 5 and 6):

1. Start with a model that is as good as you can get it in terms of the fixed explanatory variables.
2. Using the fixed terms from step 1, find the optimal random structure. This means that for the noise component, we have to try different options (e.g. random effects, temporal correlation, different variances, etc).
3. For the optimal random structure found in step 2, find the optimal fixed structure.

This is a scheme that works well for linear mixed models applied on relatively small data sets, but for a large and complicated data set like ours, we have to be a bit more creative. In the remaining part of this section, we show how we sequentially develop our models and finally end up with something that seems to do the job. The starting point is again a model of the form

$$\text{LDIN} = \text{intercept} + f(\text{Time}) + \varepsilon$$

The ε represents the noise or unexplained bit and LDIN the log-transformed DIN data. First we need to add some indices. We have 31 stations, and measurements which are taken over time. This gives

$$\text{LDIN}_{is} = \text{intercept}_i + f_i(\text{Time}_s) + \varepsilon_{is}$$

where i is the station index from 1 to 31 and s represents the time units. The noise term ε_{is} is assumed to be normally and independently distributed with mean 0 and variance σ^2. The model above allows for a different trend at each station (the function f has a subscript i). If smoothing techniques are applied to model the function f, then it is almost impossible to fit a model with 31 trends on a data set of this size. If we are lucky, only one trend will be needed for all stations, and the subscript i can be dropped from the function f. From the data exploration section, we know that the model needs a long-term trend, station effect, and a seasonal component; so a possible starting point is

$$\text{LDIN}_{is} = \text{intercept} + \text{factor}(\text{Station}_i) + f(\text{Time}_s) + \text{factor}(\text{Month}_s) + \varepsilon_{is} \quad (18.1)$$

The function f is now a smoothing function over time and is typically modelled with a spline. We have seen the notation *factor* in various other chapters; it is used to tell R that the corresponding variable is categorical. Used here, it indicates the variables Station and Month are considered as categorical variables in Equation (18.1). The costs are 11 parameters for Month and 30 for Station. We will return to the 30 parameters for Station in a moment. For the seasonal component, we have multiple options. Instead of using a categorical variable Month, you can also use sinus or cosines functions (Pinheiro and Bates, 2000) or a smoother $f(\text{DayInTheYear}_s)$, where the variable DayInTheYear_s takes values between 1 and 365 (Wood, 2006). Based on initial runs, the latter option performs the best as judged by the AIC. Note that we are not too fussy about leap years.

Regarding the argument Time_s in the function $f(\text{Time}_s)$, we have two options. We can use the day of sampling expressed as the number of days since the first sampling day of the experiment (or since 1 January 1991). But you then need to ensure that sampling day for all variables is expressed relative to the same starting date! The second option is to use $f(\text{Year}_s)$, where Year_s has integer values between 1991 and 2005. It takes less computing time and is slightly easier for comparing trends of different variables, and this is the approach we use here.

Up to now, the model contains components for trends over time and trends within a year (the seasonal pattern). However, sampling took place at 31 stations and we also have the spatial coordinates for each station (denoted by X_i and Y_i). In the same way as temporal trends were added, we can include a spatial trend $f(\text{X}_i,\text{Y}_i)$, which is a 2-dimensional smoother. This gives the following model:

$$\begin{aligned}\text{LDIN}_{is} = {} & \text{intercept} + \text{factor}(\text{Station}_i) + f(\text{Year}_s) + f(\text{DayInTheYear}_s) \\ & + f(\text{X}_i, \text{Y}_i) + \varepsilon_{is}\end{aligned} \quad (18.2)$$

The problem with this model is that we are paying the penalty of 30 regression parameters for the station effect. This is in fact the same discussion that we had when random effects were introduced in Chapter 5. Are we really interested in knowing which stations have higher values than others? Do we want to make statements for only these 30 stations? In this case, the answer to both questions is no, and this is a typical example of using a random intercept for station. It allows us to make statements for all similar stations along the Dutch coast and saves several degrees of freedom. Therefore, model (18.2) becomes

$$\text{LDIN}_{is} = \text{intercept} + f(\text{Year}_s) + f(\text{DayInTheYear}_s) + f(\text{X}_i, \text{Y}_i) + a_i + \varepsilon_{is} \tag{18.3}$$

The random intercept a_i is assumed to be normally distributed with mean 0 and variance $\sigma_{\text{station}}{}^2$. So far, adding terms was based on common sense and some initial analyses. At this stage, it is perhaps useful to apply the model and see where it fails. This will guide further improvements, if needed.

The advantage of the model in Equation (18.3) is that we have decomposed the time series into long-term trends and short-term trends. Each of these components can be extracted and compared with other environmental long-term and short-term trends, or with the phytoplankton short-term and long-term trends.

The following R code is used to implement the model in Equation (18.3).

```
> RIKZ$X <- RIKZ$X31UE_ED50  #spatial coordinates
> RIKZ$Y <- RIKZ$X31UN_ED50  #spatial coordinates
> library(mgcv)
> M1 <- gamm(LDIN ~ s(Year) + s(dDay3) + s(X, Y),
              random = list(fStation =~ 1), data = RIKZ)
```

The variable dDay3 contains the coding of the sampling day in a year, expressed as a number between 1 and 365. The results of this model are given in Fig. 18.6, which was produced with the following R code.

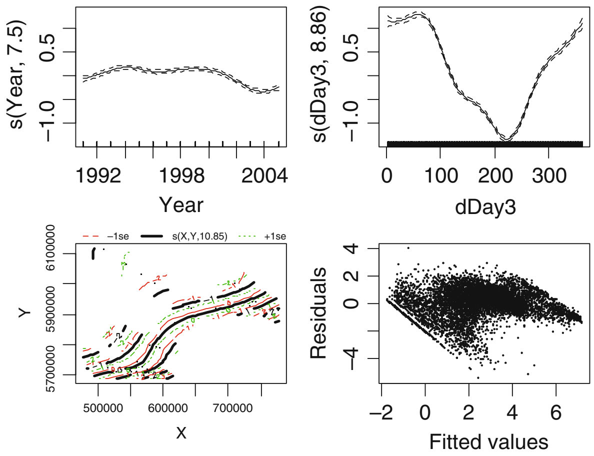


Fig. 18.6 Results for the model in Equation (18.3): *Upper left*: Smoothing component *f*(Year$_s$). *Upper right*: Smoothing component *f*(DayInTheYear$_s$). *Lower left*: Smoothing component *f*(X$_i$,Y$_i$). *Lower right*: Normalised residuals versus fitted values showing heterogeneity

```
> op <- par(mfrow = c(2, 2))
> plot(M1$gam, select = c(1))
> plot(M1$gam, select = c(2))
> plot(M1$gam, select = c(3))
> E <- resid(M1$lme, type = "normalized")
> F <- fitted(M1$lme)
> plot(x = F, y = E, xlab = "Fitted values",
       ylab = "Residuals", cex = 0.3)
> par(op)
```

There are various problems with the model in Equation (18.3) with both heterogeneity and patterns in the residuals. The latter problem is probably due to using only one smoother for long-term trends at all stations and using one seasonal component for all stations. The data exploration had already indicated that these patterns differ per station. Hence, a natural extension is to use multiple long-term trends and multiple seasonal smoothers. To find a balance between what is needed and what can be done with current software and the numerical capacity of computers, we introduce an interaction term between some of the smoothers and area. If we use one long-term smoother per area and one seasonal pattern per area, the model becomes

$$\text{LDIN}_{is} = \text{intercept} + f_{\text{area}}(\text{Year}_s) + f_{\text{area}}(\text{DayInTheYear}_s) + f(\text{X}_i, \text{Y}_i) + a_i + \varepsilon_{is} \qquad (18.4)$$

The term $f_{\text{area}}(\text{Year}_s)$ is the long-term smoother for a particular area (each area consists of multiple stations), and the same holds for the within-year pattern $f_{\text{area}}(\text{DayInTheYear}_s)$. Recall that there are 10 areas, meaning the model has 10 + 10 + 1 = 21 smoothers. Instead of the notation $f_{\text{area}}(\text{Year}_s)$, you can also use $f_a(\text{Year}_s)$ or even $f(\text{Year}_s)$:Area. The choice of notation depends on your own preference or the style of the journal you are aiming for. The R code to fit the model in Equation (18.4) is given by[1]

```
> M2 <- gamm(LDIN~
    s(Year, by = as.numeric(Area == "WZ"), bs = "cr") +
    s(Year, by = as.numeric(Area == "GM"), bs = "cr") +
    s(Year, by = as.numeric(Area == "VD"), bs = "cr") +
    s(Year, by = as.numeric(Area == "ED"), bs = "cr") +
    s(Year, by = as.numeric(Area == "OS"), bs = "cr") +
    s(Year, by = as.numeric(Area == "WS"), bs = "cr") +
    s(Year, by = as.numeric(Area == "KZ"), bs = "cr") +
    s(Year, by = as.numeric(Area == "NZ"), bs = "cr") +
    s(Year, by = as.numeric(Area == "NC"), bs = "cr") +
    s(Year, by = as.numeric(Area == "VM"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "WZ"), bs = "cr") +
```

[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.

```
    s(dDay3, by = as.numeric(Area == "GM"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "VD"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "ED"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "OS"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "WS"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "KZ"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "NZ"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "NC"), bs = "cr") +
    s(dDay3, by = as.numeric(Area == "VM"), bs = "cr") +
    s(X, Y), random = list(fStation =~ 1), data = RIKZ)
```

It looks intimidating, but it is only a simple extension of the model in Equation (18.3). The by option is used to ensure that the particular smoother is only applied on a subset of the data where the argument of the by option is equal to 1. The as.numeric() is used to convert the value TRUE to a 1 and FALSE to 0. The AIC of models (18.3) and (18.4) is 18366.07 and 16648.41, respectively. You can also try intermediate models with multiple long-term smoothers and a single seasonal smoother, or the other way around, but their AICs are all larger than 16648.41. We used cubic regression splines (bs = "cr") for the temporal trends because with large data sets these have shorter computing times than the default thin spline smoother.

The estimated long-term and seasonal smoothers obtained by this model are given in Figs. 18.7 and 18.8. The code to make these graphs is complex and given on the book website. Several long-term smoothers have similar patterns, e.g. WZ, ED, and GM. In fact, most trends have two peaks; one in the early 1990s and one towards the end of the 1990s. The trend for VM shows a strong decrease since 1998. As to the seasonal patterns per area, you can see that some areas (e.g. WS, ED) have a less strong seasonal pattern. The other areas have all slightly different shapes.

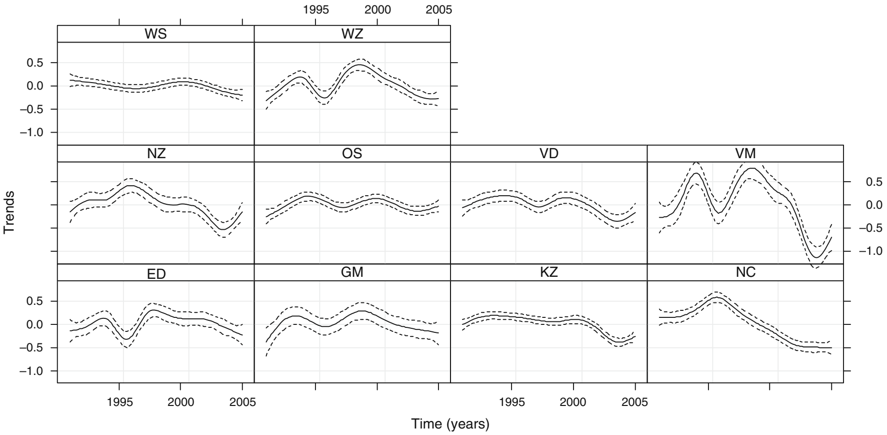


Fig. 18.7 Estimated long-term smoother for each area obtained by the model in Equation (18.4)

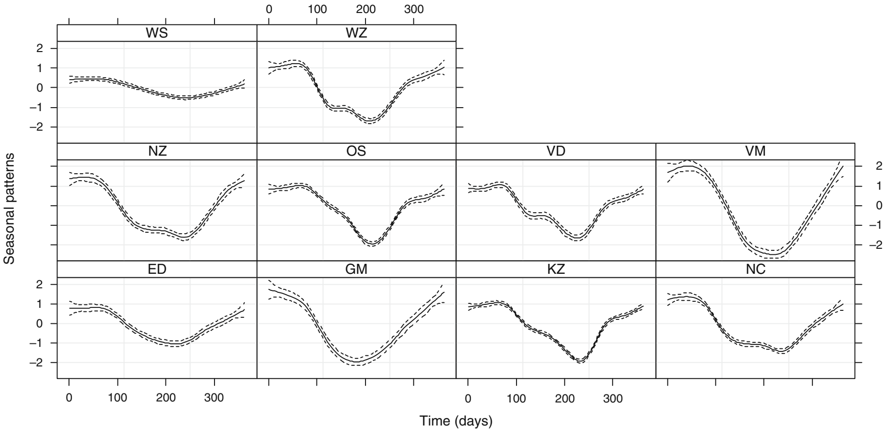


Fig. 18.8 Estimated seasonal pattern per area obtained by the model in Equation (18.4). The *x*-axis contains the days from 1 to 365

As part of the model validation process, we plotted normalised residuals versus month (for all stations), see Fig. 18.9. If you wonder why we did this, then the answer is 'common sense'. In most ecological systems, the spread in the data differs between months or seasons. You can get the same message by redrawing the lower right panel in Fig. 18.6 and use different colours per month or season. Figure 18.9 shows that there is more variation in spring and summer than in autumn and winter, which violates the homogeneity assumption. The figure was created with the following R code.

```
> E2 <- resid(M2$lme, type = "n")
> plot(E2 ~ RIKZ$fMonth, xlab = "Month",
        ylab = "Normalised residuals")
```

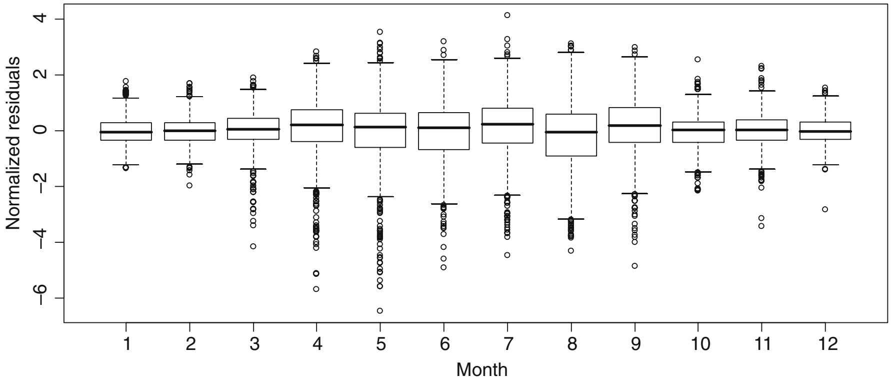


Fig. 18.9 Normalised residuals plotted versus month obtained by model (18.4)

A solution for the heterogeneity problem is to relax the assumption that the residuals ε_{is} are normally distributed with mean 0 and variance σ^2. Instead, we can use a Normal distribution with mean 0 and variance $\sigma_m{}^2$, where m stands for month. Hence, the residuals are allowed to have a different spread per month. The problem is that computing time for such a model for these data can be long (hours on a modern computer), and therefore, it may be a more realistic option to use a different variance per season (four variances) or per 6-month period (two variances). We decided to go for four variances and define the seasons as months 1–3, 4–6, 7–9, and 10–12. However, further fine-tuning of the model can still be achieved. The R code for the model with four variances is a simple extension of the previous R code and is not reproduced here. We only have to define a variable defining the four seasons:

```
> n <- length(RIKZ$Month)
> RIKZ$M14 <- vector(length = n)
> RIKZ$M14[1:n] <- 0
> RIKZ$M14[RIKZ$Month >= 1 & RIKZ$Month  <= 3] <- 1
> RIKZ$M14[RIKZ$Month >= 4 & RIKZ$Month  <= 6] <- 2
> RIKZ$M14[RIKZ$Month >= 7 & RIKZ$Month  <= 9] <- 3
> RIKZ$M14[RIKZ$Month >= 10 & RIKZ$Month <= 12] <- 4
> RIKZ$fM14 <- factor(RIKZ$M14)
```

Allowing for different variances is done with the `weights` option and the `varIdent` structure; see also Chapter 5. All we have to do is add the code `weights = varIdent(form = ~1 | fM14)` to the `gamm` function presented above.

Unfortunately, this model did not converge. Using the `varIdent` function with two variances (two seasons) neither converged. However, using 10 long-term smoothers and one seasonal pattern for all smoothers plus four variances for the seasons (and a random intercept) did not cause any numerical problems. If this sort of numerical trouble happens, it can be quite a challenge to sort out. One option is to increase the number of iterations in the `gamm` routine or reduce the convergence criteria, see the help file of `gamm` how to do this. Other options are to fix the degrees of freedom (and not use cross-validation) or set the four variances to fixed values (e.g. based on the residual variation of previous models). Unbalanced data, missing values, etc., can also cause convergence problems. We tried all of this, but without success. However, replacing the 10 seasonal smoothers by a `fMonth × Area` component in the model in Equation (18.4) and re-running the code did not give any converge problems. The estimated long-term smoothers obtained by this model had nearly identical shapes as those in Fig. 18.7. Furthermore, extending this model with four residual variances did not cause any numerical problems. Again, its estimated long-term trends are similar as to those in Fig. 18.7 and are therefore not presented again.

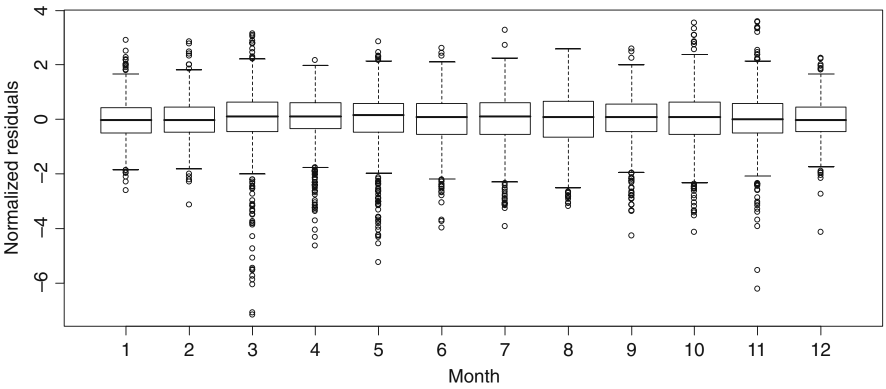


Fig. 18.10 Normalised residuals of the model with 10 long-term smoothers, seasonal components modelled by `fMonth` × `Area`, a spatial trend, a random intercept for stations, and four variances. Note that `Area` is automatically a factor due to its coding. Residuals are grouped per month

However, the model validation did show some problems. Although the residual spread is approximately the same in all months, we still have more negative residuals in the spring and summer than in the autumn and winter (Fig. 18.10). We had already seen this behaviour in the data exploration section. By plotting the residuals versus time for each station (Fig. 18.11), we can see that there is no clear pattern in these large values. One option to deal with this is to include a nested (within station) random intercept for month. This allows for random variation around the seasonal pattern, and this variation can be different per month. However, this would only

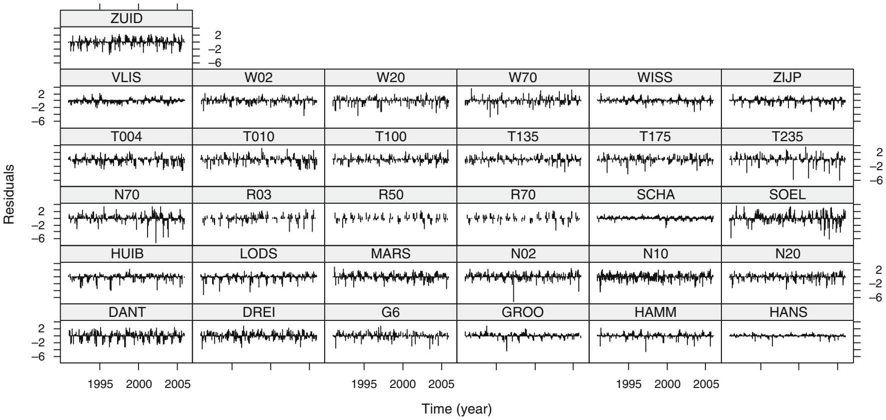


Fig. 18.11 Normalised residuals of the model with 10 long-term smoothers, seasonal components modelled by `fMonth` × `Area`, a spatial trend, a random intercept for stations, and four variances. Residuals plotted versus year for each station

hide the fact that we have large observations in some months and stations. This may well be a sampling issue; not all stations are sampled on the same day due to practicalities of travel arrangements (some stations are separated by 250 km). If DIN values are high (for a short period) in a certain region, then you may measure it at one station, but values may have dropped already by the time you reach the next station. So, instead of hiding it in random effects, we will leave it as it is. The R code to produce Figs. 18.10 and 18.11 is not reproduced here as it closely follows earlier code.

Plotting normalised residuals versus fitted values showed that there is still a certain degree of heterogeneity in the residuals. This is because some stations have less variation in DIN values. This can also be seen in the data exploration section and even in Fig. 18.11. Another way of spotting this is to plot fitted values against residuals and use a different colour per station. It is difficult to solve this. It is not practical to use the `varIdent` structure and 31 different variances as computing time would drastically increase. A better option is to scale each time series, for example, by using: $\text{LDIN}_{is}/\max(\text{LDIN}_{is})$. Such a standardisation ensures all the time series have similar *variation*, but the average values can still be different (this in contrary to centring and dividing by the standard deviation).

The last aspect we look at (as part of the model validation process) is spatial patterns in residuals. We made a bubble plot of averaged (per station) residuals (Fig. 18.12), and there seems to be no clear clustering of positive (or negative)

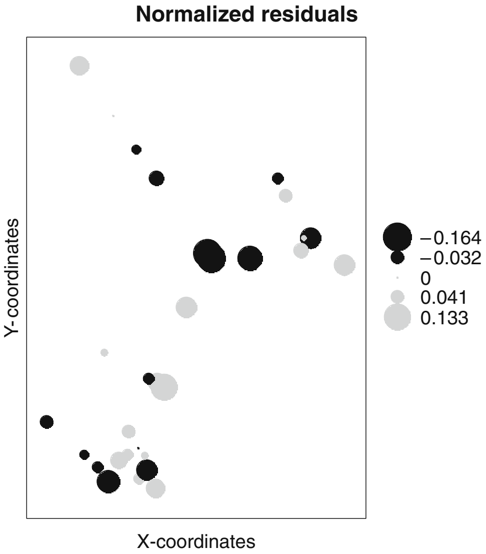


Fig. 18.12 *Bubble plot* for averaged residuals per station. *Large dots* represent large residuals with *black dots* for negative residuals and *grey dots* for positive residuals. The R code to produce this graph is available from the book website

residuals. It is also possible to make this graph for data of each year or each season. Alternatively, variograms can be made of residuals per station or per year. It would be a nice challenge to make an `xyplot` with multiple variograms in it, but we will leave this as an exercise for the reader.

To allow for spatial or temporal correlation in the residuals, you can attempt to add the correlation option to the `gamm` function, but our initial attempts resulted in convergence problems due to the large sample size. So, we are pushing things a little bit too far with current software and hardware.

As to the numerical output, all trends were highly significant. However, we advise being cautious with these *p*-values as there is considerable residual information left in the model. It may be an option to allow for more smoothers for the series or analyse these time series separately.

18.4 Results for Temperature

The same modelling strategy was applied on the (untransformed) temperature data. We started the model selection process from scratch. The data exploration showed that the patterns over time show less variation compared to the DIN data. Figure 18.13 shows the boxplots of temperature per month for each area. It will be interesting to test whether the seasonal pattern change per area or not.

The same strategy used for the DIN analysis was followed. The AIC showed that the model with 10 long-term smoothers, 10 seasonal smoothers, a spatial trend, and a random intercept was the best model. There was only minor (visual) evidence of heterogeneity and therefore no real need to use multiple variances per season. The estimated long-term smoothers are shown in Fig. 18.14. It may be an option

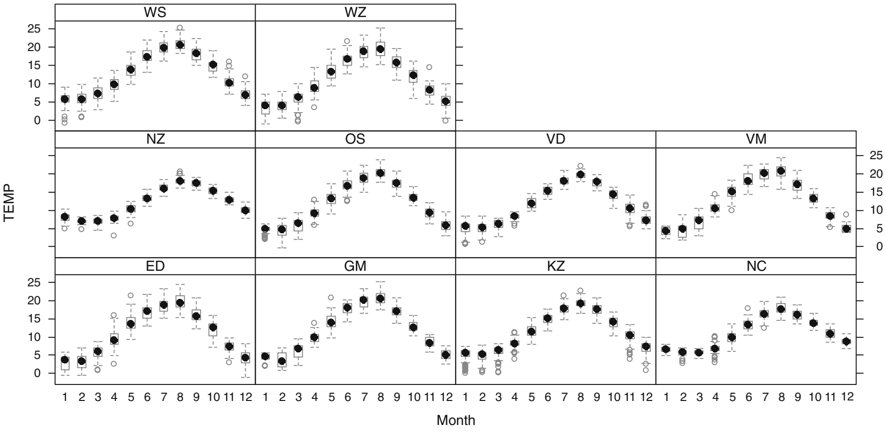


Fig. 18.13 Temperature per month for each area

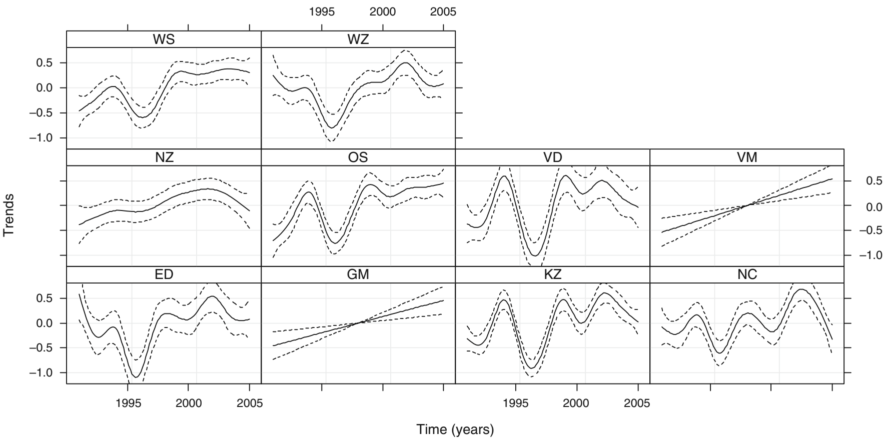


Fig. 18.14 Long-term trends for temperature by area. The *solid line* is the smoother and the *dotted lines* are 95% confidence bands

to group some of the areas and use only one smoother for them, but this makes the comparison with the phytoplankton data, presented later, more difficult. The 10 seasonal components are given in Fig. 18.15; note that NZ, NC, VD, and KZ trends are slightly different from the others. The shape of these curves also shows why a sinus function would not work; the patterns are not symmetrically shaped during the year. The spatial trend *f(X,Y)* is presented in Fig. 18.16.

The R code for the temperature data analysis is identical to the code used in the previous section and is not presented again.

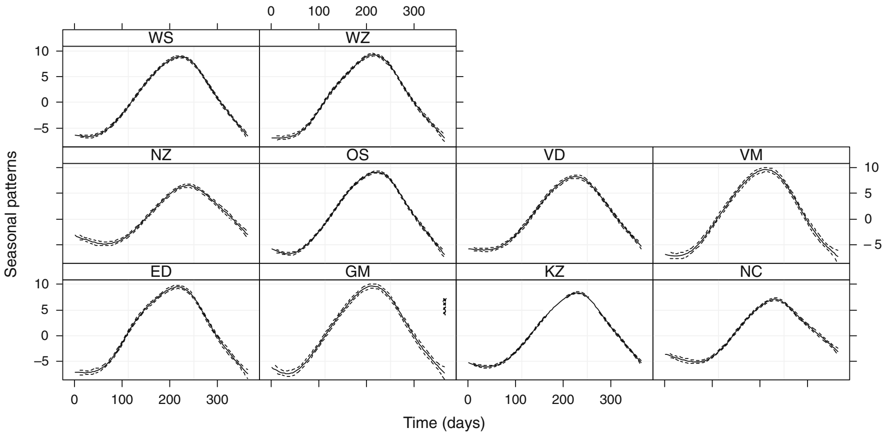


Fig. 18.15 Seasonal components for temperature by area. The *solid line* is the smoother and the *dotted lines* are 95% confidence bands

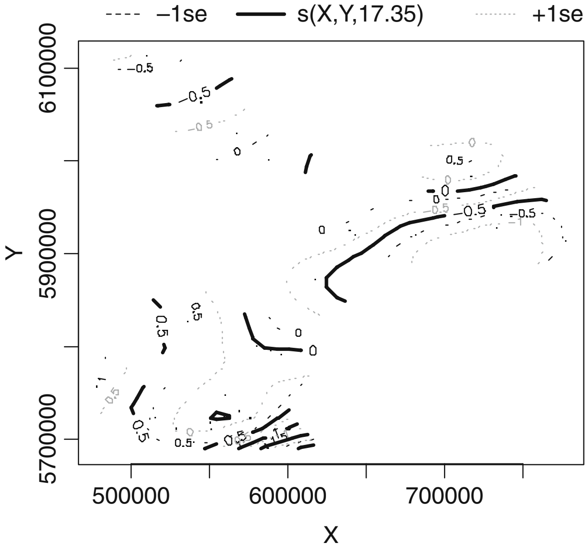


Fig. 18.16 Spatial trend for the temperature data. The *solid line* represents the *contour lines* and the *dotted lines* are confidence bands. It is also possible to plot this graph as a 3-dimensional picture

18.5 Results for DIAT1

In this section, the aggregated DIAT1 (diatoms between 0 and 1,000 μm^3) phytoplankton series are analysed. An initial data exploration was carried out, and this indicated that a log-transformation was needed.

The main difference with the environmental data and these data is that during the 15 years of sampling, four different laboratories were successively involved with the counting of the phytoplankton. There was no overlap between the laboratories and there is a clear ‘laboratory’ effect that can be seen using a simple boxplot or a more advanced boxplot produced by the bwplot function from the lattice package, see Fig. 18.17. At most areas, values taken by laboratory D are the highest. However, this was also the laboratory that took the most recent samples. If we include the term factor(Laboratory) in model (18.4) and replace LDIN by LDIAT1, the categorical variable laboratory is highly significant. In fact, the estimated trends for the model with and without the laboratory effect are only slightly different, especially during the period when laboratory D was in charge of counting. The question then rises, whether there is indeed a laboratory effect or whether abundances have increased during the period when laboratory D counted. Unfortunately, there is no way we can distinguish between the two. The only thing that we can say is that the estimated laboratory effect (as measured by estimated parameters for each level of the categorical variable) is larger than you would expect based on common sense approach to our existing ecological understanding. We have found similar changes of abundance between other years, not corresponding with

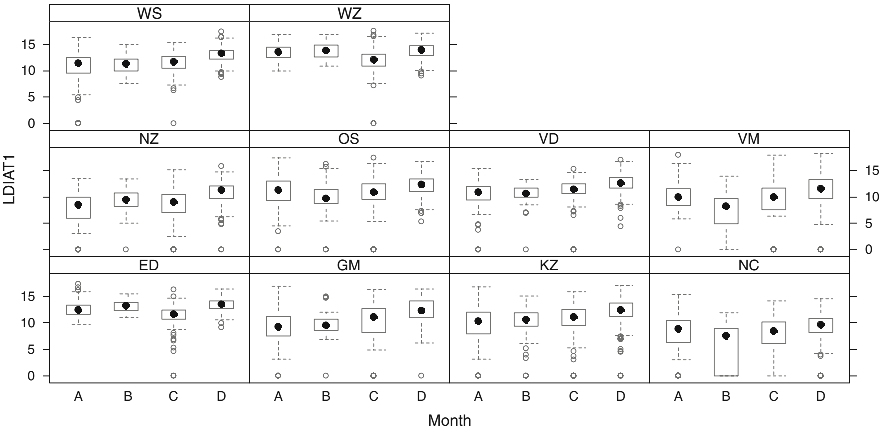


Fig. 18.17 *Boxplot* of log-transformed DIAT1 conditional on laboratory (represented by A, B, C, and D) per area

factor Lab. As well as looking further into this particular DIAT1-group either on the level of species or per station, the Lab-pattern appeared not to be of a structural kind. Summarising, we cannot say whether any changes over time in abundances are due to a laboratory effect or whether it represents a real change. We therefore concluded that the laboratory effect is small compared to observed changes and ignored it.

The modelling approach followed similar lines used for the environmental variables. Note that most long-term trends in Fig. 18.18 seem to increase up to about 2001. Seasonal patterns are rather different per area (Fig. 18.19). Some areas show a clear diatom blooming in early spring followed by a smaller bloom in late autumn. The spatial pattern is given in Fig. 18.20.

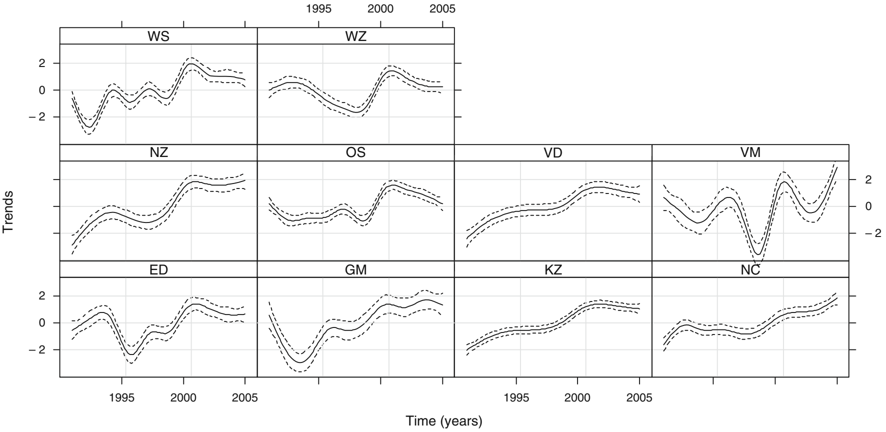


Fig. 18.18 Estimated trends for log-transformed diatoms (DIAT1). The model did not contain a laboratory effect

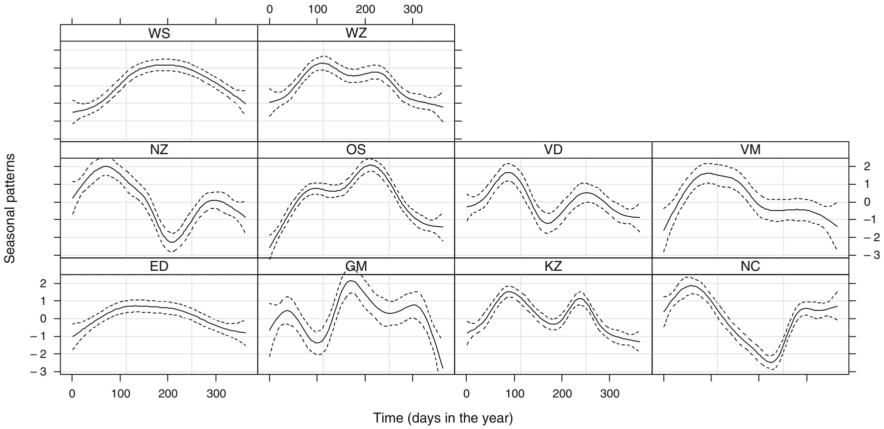
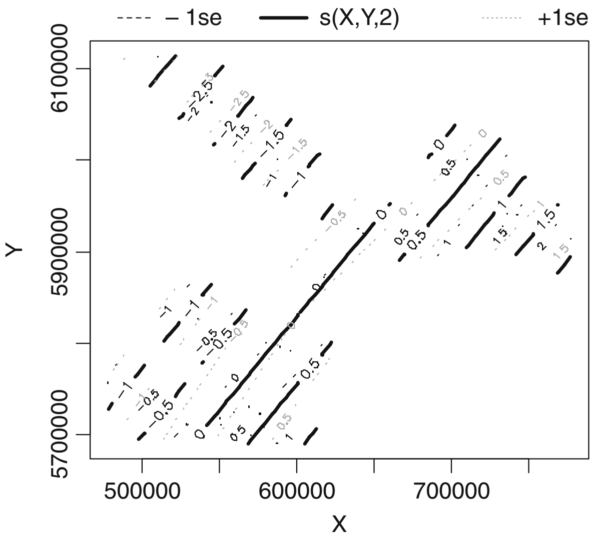


Fig. 18.19 Estimated seasonal patterns for log-transformed diatoms (DIAT1). The model did not contain a laboratory effect

Fig. 18.20 Estimated spatial patterns for log-transformed diatoms (DIAT1). The *solid lines* are the *contour lines* and the *dotted lines* are 95% confidence bands. The model did not contain a laboratory effect



18.6 Comparing Phytoplankton and Environmental Trends

In the previous three sections, we applied a Gaussian GAMM on multiple times series for DIN, temperature, and DIAT1. For each variable, we have 10 long-term trends. The question now is whether there is any relationship between the DIAT1 trends and the DIN and temperature trends. One may be tempted to consider DIAT1 as a response variable and DIN and temperature as explanatory variables. However, the original data set had approximately 8–10 explanatory variables, and there was considerable collinearity between these variables. This makes it rather difficult to pinpoint any of the variables as *the* driving variable. An extra interpretation problem is caused by the seasonal patterns in the original data, as this may cause the high correlations between the explanatory variables. We therefore go for the simple approach comparing the long-term DIAT1 trends with the DIN and temperature trends using Pearson correlation coefficients. There is no point in comparing the ED DIAT1 trend with the WS DIN trend as these areas are 250 km apart. Hence, it makes more sense to compare the DIAT1 and environmental trends per area.

A word of caution is also needed. Long-term trends tend to be smooth functions by definition, and the Pearson correlation coefficient between two smooth functions tends to be high. Furthermore, we are going to calculate 20 correlation coefficients, which means that there are potential problems with multiple testing. Our view on this is to just calculate the correlations, present them graphically, see which combinations have the highest correlations, and refrain from interpreting *p*-values. The estimated Pearson correlations are given in Table 18.1 and a graphical presentation of these correlations in Fig. 18.21. The graphical presentation may look like overkill, but it is useful if more environmental variables are used. Another way to present the estimated correlations is presented in Fig. 18.22; the correlations between the DIAT1 and environmental variables are presented in two panels, the font size of the labels is proportional to the (absolute) estimated correlations. The advantage of this graph is that you have a better overview where (spatially) the areas with high correlations are.

The R code to calculate the correlation between the trends and to produce Figs. 18.21 and 18.22 is rather complicated and is given on the book website. The main problem in the R code is to access the estimated smoothers. By default, the `plot.gam` function is creating smoothers of length 100; hence, the smoothers in for example Fig. 18.18 are interpolated curves. Here we used long-term smoothers of length 15 (because there are 15 years).

Table 18.1 Estimated Pearson correlation coefficients between the DIAT1 trends and the corresponding (i.e. same area) DIN and temperature trends

	DIN	TEMP
DIAT1WZ	–0.29	0.62
DIAT1GM	–0.33	0.78
DIAT1VD	–0.36	0.48
DIAT1ED	–0.3	0.68
DIAT1OS	–0.2	0.37
DIAT1WS	–0.39	0.66
DIAT1KZ	–0.61	0.58
DIAT1NZ	–0.54	0.79
DIAT1NC	–0.85	0.3
DIAT1VM	–0.51	0.32

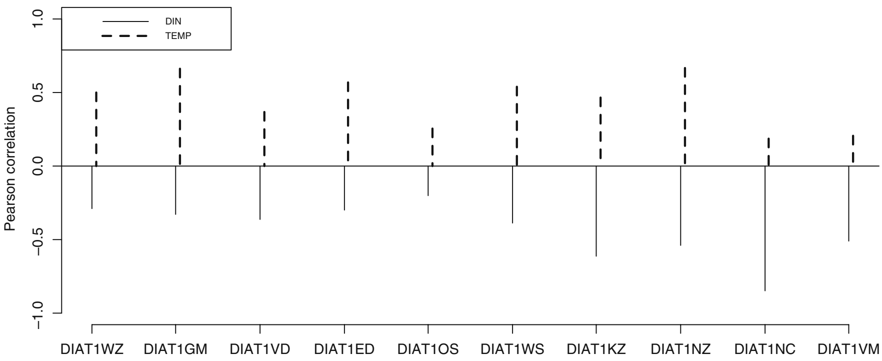
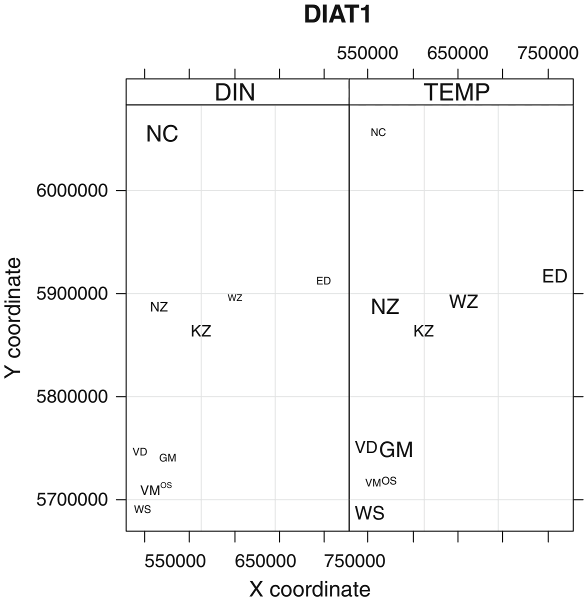


Fig. 18.21 Graphical presentation of the estimated Pearson correlation coefficients in Table 18.1. The endpoint of a line gives the value (along the *y*-axis) of a DIAT1 trend with the corresponding environmental variable (for the same area)

Fig. 18.22 Graphical presentation of the estimated Pearson correlation coefficients. The font size of the labels for an area is proportional to the absolute value of the estimated correlation



18.7 Conclusions

The analysis of the Rijkswaterstaat time series was one of the more challenging exercises in this book. However, it is a type of data set you are very likely to come across if you work with ecological or environmental monitoring data. By no means is this a finalised analysis. It would, for example, be good to add temporal residual correlation structures using, for example, the option `correlation = corAR1(form =~ dDay1 | fStation)` within the `gamm` function. At the time of writing, our (new) computer (with a Windows operating system) was not able to carry out such analyses for this data set (and due to the complex mathematical calculations, it is unlikely to run neither on a Mac, LINUX, or UNIX operating system). It would be even better to use a spatio-temporal residual correlation structure, which you would have to program yourself. Before making any attempts to include a correlation structure, you should make an experimental variogram of the normalised residuals per station and plot the experimental variograms in a lattice plot. If these suggest there is no temporal correlation, then there is no point trying to add a temporal correlation structure inside the model.

The data presented here is merely an illustration how to deal with data of this type and is a spin-off from a technical report. The original report used more environmental variables and more phytoplankton groups. Because we only used a small part of the data here, we will not go into a biological discussion of the results.

The technical aspects of the analysis of multiple phytoplankton species are simple; just apply the same methodology on the most important species and use good visualisation tools to present the results.

18.8 What to Write in a Paper

If this chapter was your work, you are faced with a dilemma. The residuals of the estimated models still show patterns. So you either have to present this as a paper with preliminary results and make it clear that further work is going on or you can argue that this is as much as can be done with current hardware and software, and because all terms in the model are highly significant, the results are reasonable robust. Whichever route you go, you have to be very careful with the interpretation of the results due to the remaining patterns. However, in the technical report we analysed the data slightly differently, but the estimated long-term trends were nearly identical to the ones presented here. Perhaps, some simulation studies to assess sensitivity would be a useful addition to convince the referees.