21 GAM for zero-inflated and spatial-temporal correlated sandeel data

Alain F Zuur, Elena N Ieno, Johan Craeymeersch,
Ingrid Tulp, Nicola SH Tien

In this chapter we analyse zero-inflated sandeel data using zero-altered Poisson (ZAP) models with spatial-temporal correlation, and we allow for non-linear covariate effects using generalised additive models (GAM). We also discuss how to extend the GAM by allowing for interactions between smoothers and categorical covariates using R-INLA.

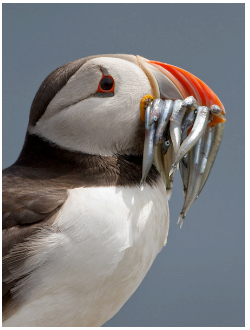
It is said that all roads lead to Rome, and that certainly seems to hold for the analysis of this sandeel data set. We had to make so many arbitrary decisions about how to continue the analysis after detecting yet another problem after fitting a new model. It may well be possible that different decisions would have given us a different final model, and therefore different biological conclusions. In addition to showing the reader how to execute all the models in R-INLA, we also try to explain our thinking patterns and how we eventually ended up in Rome.

Prerequisite for this chapter: Knowledge of ZAP models (Chapter 18), how to include spatial dependency (Chapter 19) and using smoothers in R-INLA (Chapter 20) is required.

21.1 Introduction

21.1.1 Sandeel

Sandeel are a small fish species, and an important source of food for many sea bird species. As the first part of the name suggests, they live in sandy sediments on the seabed and hide temporarily in small burrows to avoid predators. Because of this lifestyle, they are susceptible to changes in sediment characteristics, e.g. due to water velocity or man-induced disturbances via beam trawl fisheries. A beam trawl fishing boat uses a steel beam to keep the fishing net open, but the shoes and chains of the net also plough the sediment.



Tien et al. (2017) investigated the effect of various sediment-related covariates, water velocity and beam trawl-fishing indices on sandeel. The data are from the Voordelta, which is a protected nature area off the southern Dutch coast.

Around 400 sites were sampled in each of the years 2009, 2010, 2011 and 2012. Tien and co-authors focused on absence–presence data of three sandeel species. In this chapter we use sandeel (*Ammodytes tobianus*) abundance data and employ ZAP models with spatial-temporal correlation. We will also allow for non-linear covariate effects via smoothers. The data selection used in the chapter differs slightly from the data used in Tien et al. (2017), and therefore results may be slightly different.

21.1.2 Variables

Three sandeel species were sampled. In this chapter we will use the *A. tobianus* abundances as the response variable.

A large number of covariates were sampled, namely depth, water temperature, salinity, median grain size, percentage of coarse sand, percentage of medium sand, percentage of fine sand, percentage of very fine sand, percentage of silt, water velocity and the size of the sampled area. Many of these covariates are available as minimum, medium, average and maximum values. We used average values. There are also two fisheries disturbances indices, namely fisheries intensity by beam trawl fishery aiming at flatfish and aiming at shrimp. These indices are expressed in trawl minutes, and their effects on the sandeel abundances are of prime interest in the analyses.

For some of the covariates it may be a sensible option to use maximum values (e.g. for water velocity) instead of average values or to apply a time lag to allow for a delayed cause-effect relationship. We invite the interested reader to make changes to our set of selected covariates, but be forewarned; the road ahead of us is already complicated enough. Pottering around with time lags and minimum, maximum or median covariate values only adds extra problems.

This chapter contains a great deal of information. The Figure 21.1 mindmap shows the general layout of this chapter.

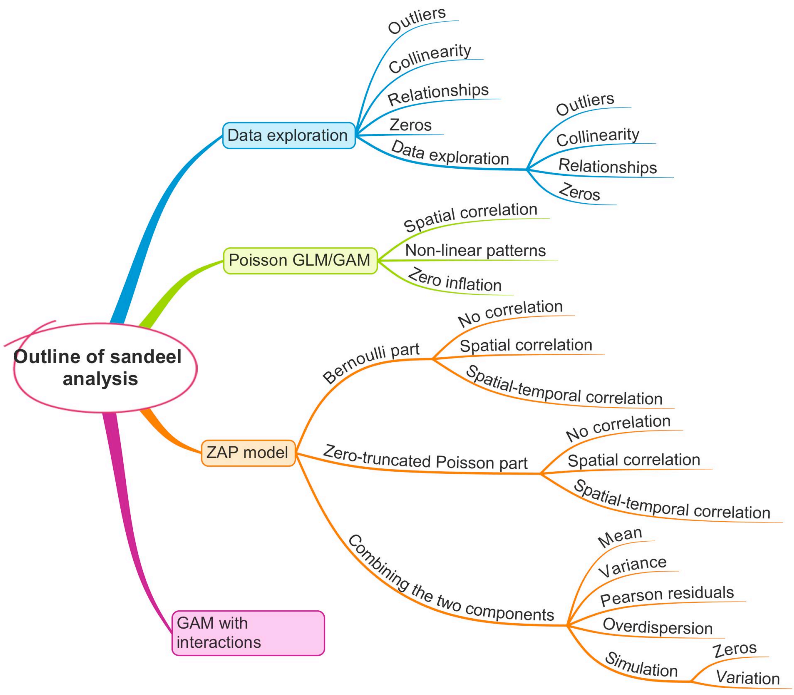


Figure 21.1. Outline of this chapter. We will start with data exploration, followed by an ordinary generalised linear model (GLM).

21.2 Data exploration

21.2.1 Spatial position of sampling locations

Figure 21.2 shows the spatial position of the sampling location per year. In each year around 400 locations were sampled; see Tien et al. (2017) for full details on sampling. Note that in 2015 the upper-right part of the study area was not sampled. It is an option to remove the data from this area also from the 2009–2012 data. Alternatively, we can include these locations, but when interpreting the results, we should take the unbalanced spatial-temporal nature of sampling design into account. Yet a third option is to analyse the data without these sampling locations, analyse the data with these sampling locations and compare the results. If results are similar (as is indeed the case here), then we can present the results based on all the data. If results are different, then we need to use the smaller data set.