

Uw kenmerk

Ons kenmerk

Doorkiesnummer

Bezoekadres

EP/IE/01/1

3051

Willem C. Van Unnikgebouw

Heidelberglaan 2

3584 CS, De Uithof, Utrecht

Datum

Onderwerp

1-2-01

Telefoon (030) 253 20 44

Fax (030) 254 06 04

Postbus 80115, 3508 TC Utrecht



Faculteit Ruimtelijke Wetenschappen

Universiteit Utrecht

Geachte geadresseerde,

Hierbij bied ik u het rapport "Spatial interpolation of sea bird densities on the Dutch part of the North Sea" aan. In dit rapport wordt een methode uitgelegd en toegepast voor het schatten van dichtheden van zeevogels op het Nederlands Continentaal Plat (NCP), aan de hand van striptransect-tellingen van zeevogels en informatie over waterdiepte en afstand tot de kust.

Zeevogeltellingen worden tweemaandelijks uitgevoerd door het Rijksinstituut voor Kust en Zee (RIKZ). Hierbij wordt drie dagen lang vanuit een vliegtuig volgens een bepaalde route de zeevogelstand opgenomen. In dit rapport zijn telgegevens van 1995 gebruikt. Als voorbeeld vindt u op pagina 72 van het rapport een kaart van waargenomen dichtheden van de Noordse Stormvogel in de periode augustus/september.

De gebruikte methode combineert gegeneraliseerde lineaire regressie voor trendmatige patronen als functie van eenvoudige parameters als de waterdiepte en de afstand tot de kust met geostatistische interpolatie van residuele variatie. Schattingen van dichtheden voor gebieden waar niet is waargenomen zijn onderhevig aan onzekerheid: de schattingsfout. Om deze fout tot uiting te brengen in de kaarten, is ervoor gekozen schattingen te presenteren in de vorm van 95% betrouwbaarheidsintervallen (zie figuur 5.2 op pagina 39).

Het rapport is de weerslag van onderzoek, uitgevoerd aan het Utrecht Centre for Landscape Dynamics (UCEL) van de Universiteit Utrecht, in opdracht van het Rijksinstituut voor Kust en Zee.

Mede namens de co-auteurs, hoogachtend,

Edzer J. Pebesma

01:55541



Centre for Geo-ecological Research

**Spatial interpolation of
sea bird densities on the
Dutch part of the North Sea**

E.J. Pebesma, R.N.M. Duin, A.M.F. Bio



Rijkswaterstaat

Rijksinstituut voor Kust en Zee/RIKZ
Bibliotheek (Den Haag)

C-5442 870

Spatial interpolation of sea bird densities on the Dutch part of the North Sea

Edzer J. Pebesma¹ Richard N.M. Duin²
Ana M.F. Bio¹

NOVEMBER 2000

¹Utrecht Centre for Environment and Landscape Dynamics (UCEL), Faculty of Geographical Sciences, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht; e.pebesma@geog.uu.nl

²Dutch National Institute for Coastal and Marine Management (RIKZ)

This research was supported by the Dutch National Institute for Coastal and Marine Management (RIKZ), and was supervised by R.N.M. Duin, J.G. Hartholt, C.M. Berrevoets and R.H. Witte.

Dit onderzoek werd uitgevoerd in opdracht van het Rijksinstituut voor Kust en Zee (RIKZ), voor de projecten HABIMAP, GEOMOD en TWINFOMEET

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 7 |
| 2 | Overview of existing methods | 9 |
| 2.1 | Spatial statistics | 9 |
| 2.2 | Linear geostatistical methods | 10 |
| 2.2.1 | Spatial dependence | 11 |
| 2.2.2 | Trend functions | 11 |
| 2.2.3 | Which trend model should we choose? | 12 |
| 2.2.4 | Conditional and unconditional predictions | 14 |
| 2.2.5 | Change of support and the global mean | 15 |
| 2.3 | Generalized linear and related models | 15 |
| 2.3.1 | Generalized linear models | 16 |
| 2.3.2 | Generalized additive models | 16 |
| 2.3.3 | Generalized estimating equations | 18 |
| 3 | Generalized linear geostatistics | 21 |
| 3.1 | Proposed method – A | 21 |
| 3.1.1 | Trend | 21 |
| 3.1.2 | Residual variation and spatial prediction | 22 |
| 3.1.3 | Predicting block means | 24 |
| 3.1.4 | Prediction maps | 25 |
| 3.2 | Proposed method – B | 25 |
| 3.2.1 | Spatial generalized estimating equations | 25 |
| 3.2.2 | Implementation | 26 |
| 3.3 | Handling known measurement errors | 26 |
| 3.3.1 | Systematic error (bias) | 28 |
| 3.3.2 | Random error (noise) | 29 |
| 4 | Case study: 1995 data | 31 |
| 4.1 | Exploratory data analysis | 31 |
| 4.2 | Explanatory variables | 32 |

| | | |
|----------|--|------------|
| 4.3 | Trend models | 33 |
| 4.4 | Residual models | 35 |
| 5 | Results | 37 |
| 5.1 | <i>Fulmarus glacialis</i> | 37 |
| 5.2 | <i>Sterna sandvicensis</i> | 38 |
| 5.2.1 | Period 3 | 38 |
| 5.2.2 | Period 4 | 42 |
| 5.2.3 | Period 5 | 45 |
| 5.3 | <i>Uria aalge/Alca torda</i> | 47 |
| 5.4 | <i>Melanitta nigra</i> | 49 |
| 5.5 | Method - B: <i>Fulmarus glacialis</i> | 49 |
| 6 | Discussion and conclusions | 55 |
| 7 | References | 59 |
| A | GLM output | 63 |
| A.1 | <i>Uria aalge/Alca torda</i> , period 2 | 63 |
| A.2 | <i>Fulmarus glacialis</i> , period 5 | 64 |
| A.3 | <i>Sterna sandvicensis</i> , period 3 | 65 |
| A.4 | <i>Sterna sandvicensis</i> , period 4 | 65 |
| A.5 | <i>Sterna sandvicensis</i> , period 5 | 66 |
| A.6 | <i>Uria aalge/Alca torda</i> , period 2, alternative model | 67 |
| B | Maps and Figures | 71 |
| C | the <i>Uria aalge/Alca torda</i> alternative | 89 |
| D | Block averages | 95 |
| D.1 | Theory | 95 |
| D.2 | Implementation | 95 |
| E | GEE: S-Plus output | 97 |
| F | S-Plus source code | 105 |
| G | Problems with estimating α | 107 |
| H | Mean-dependent covariances in gstat | 115 |
| H.1 | Simple kriging with known, non-constant mean | 115 |
| H.2 | Mean-dependent covariances | 116 |
| H.3 | Differences from previous approach | 117 |

| | | |
|----------|---------------------------------------|------------|
| I | Geostats 2000 conference paper | 119 |
| I.1 | Introduction | 119 |
| I.2 | The Monitoring Data | 120 |
| I.3 | Spatial Interpolation | 122 |
| I.3.1 | Generalised Linear Models | 122 |
| I.3.2 | Spatial correlation | 123 |
| I.3.3 | Parameter estimation | 123 |
| I.3.4 | Predicting $Z(s_0)$ | 124 |
| I.3.5 | Interval estimation | 124 |
| I.4 | Results | 125 |
| I.5 | Discussion | 126 |

1 Introduction

Since 1984, the Dutch National Institute for Coastal and Marine Management (RIKZ) monitors sea birds on the Dutch part of the North Sea (NCP) using an airborne observation technique (Baptist and Wolf, 1993). Since 1989 this monitoring is carried out systematically on a bi-monthly basis. Each monitoring round consists of three days of flying, following a fixed flight schedule. The goal of this monitoring is to get insight (i) in the spatial distribution of sea birds (and, to a lesser extent, marine mammals) over the NCP and (ii) in temporal changes in the spatial distribution of sea birds.

Up till now, reports have mainly shown descriptive statistics ('the data') of the monitoring program (e.g. Baptist and Wolf, 1993; Witte, 1995a-f) but present hardly any statistical inference based on these data regarding spatial distribution or temporal changes. This report is a first attempt to fill this gap: a method for the spatial interpolation of sea bird densities (counts) is proposed, based on both spatial (geostatistical) modelling and ecological (generalized linear) modelling of data. This method is applied to a selection of "typical" sea bird species and monitoring periods.

The selection of species and interpolation times was made such that both typical and extreme cases were present. Data from 1995 were used in this study. The species and observation period (month) chosen for this pilot study are:

- *Fulmarus glacialis* (GB-Fulmar; NL-Noordse Stormvogel), Aug/Sep (abundant on open sea)
- *Sterna sandvicensis* (Sandwich Tern; Grote Stern), all 6 periods (rare during winter; mostly restricted to coastal zone during the breeding season)
- *Uria aalge/Alca torda* (Guillemot/Razorbill; Zeekoet/Alk), Feb/Mar (many small clusters, mostly on open sea)
- *Melanitta nigra* (Common Scooter; Zwarte Zeeëend), Apr/May (highly clustered data: rare, but when present occurring in large groups, as

they are related to local abundance of their main food source, *Spisula subtruncata*)

To avoid the usually large prediction errors associated with predicting individual point observations, the size of blocks for which average bird densities were predicted was set at $5 \text{ km} \times 5 \text{ km}$, much larger than the size of individual observations (approximately 1 km^2). This block size is a trade-off: on the one hand statistical resolution (accuracy) calls for larger blocks (averages for larger blocks are estimated with smaller prediction errors) on the other hand geographical (spatial) resolution calls for small blocks (to show spatial differentiation). The choice for $5 \text{ km} \times 5 \text{ km}$ blocks seems to balance both aspects of resolution nicely.

Chapter 2 presents an overview of geostatistical and ecological approaches to statistical modelling of spatial data. Chapter 3 contains two proposed methods: one fairly straightforward and a second more elaborate approach. Chapter 4 describes the data and the application of the proposed method to these data and chapter 5 presents the main results. Chapter 6 briefly discusses the advantages and shortcomings of the method with respect to the current application and other relevant approaches.

2 Overview of existing methods

2.1 Spatial statistics

Spatial interpolation (or extrapolation) of observations implies that we estimate unobserved, unknown quantities from observed data, and for this we need methods that allow us to assess the accuracy of the estimate by providing a measure of the estimation error. We will therefore only address statistical models for the spatial interpolation here.

In statistics, observed data are usually treated as having a structural component and a noise (or error) term. Suppose we measure a quantity y , a statistical model for n observations $y_i, i = 1 \dots n$ could then be

$$y_i = \mu + e_i$$

where μ is the expected (or mean) value of the y_i , and e_i is the random error term or *residual*. In practice, there is nothing intrinsically random about e_i , we only treat it as random because we cannot explain (or understand) its variation. In the following, we will say that we *estimate* unknown, non-random quantities (μ) and that we *predict* random variables (e, y) when we want to assess their value.

In classic statistics, observations y_i are treated as being independent, which implies that there is no structure in the error, no variation in expected similarities between pairs y_i and $y_{j \neq i}$. In practice, observations that are taken closer together in space or in time are more alike than more distant observations, and this leads to spatial or temporal dependence between observed data. In presence of spatial dependence, classical statistical inference, based on the assumption of independent observations, is valid only when the data *locations* were chosen randomly from the (sub)population for which estimates or predictions are required (Hansen et al., 1983; De Gruijter and Ter Braak, 1990). The advantage of this is robustness, since no model for spatial variation (spatial dependence) is required. The disadvantage however is that information can be obtained only on a low spatial resolution, e.g. values

(such as means) for the complete area or for (relatively large) sub-areas. In practice, data are often not collected at randomly chosen locations.

When the goal is spatial interpolation, information is typically required on a high spatial resolution. In this case, spatial dependence of the residual can better be exploited in order to *let* an observation be more informative about its direct surrounding than more distant observations. Spatial statistics (Cressie, 1993) is the field of statistics that explicitly addresses spatial locations of observations, and that tries to use spatial structure in order to get optimal, location specific predictions (interpolated values). Again, these predictions should ideally come with an indication of prediction accuracy, the prediction error.

The challenge in applying statistics is to find a suitable model for the data. Mathematically speaking, any model may be applied to some data and will yield predictions and prediction errors, but results from different models may be contradicting. The question is then, are the predictions *good*, and are the prediction errors *realistic* measures of accuracy. Since all models at best approximate reality, the question arises what a good model is. In general, each model makes assumptions about the data, and these assumptions should agree (to a reasonable degree) with the observed data the model is applied to, and also with the prior information we have about the observed and modelled phenomenon.

Cressie (1993) classifies spatial data into three categories: geostatistical data, lattice data and point pattern data. Geostatistical data are data that can be observed (and predicted) at every location in the domain of interest, lattice data are observed only for larger, disjunct regions (polygons, e.g. administrative areas) and point pattern data are data on discrete objects that occur only at a limited number of point locations.

The category that applies to a certain variable does not only depend on the variable itself, but also on the way we observe it. Although the occurrence of individual birds is a discrete, point pattern process, in our case the data are collected as spatially aggregated counts (or densities) for given areas. Information about the spatial distribution of individual birds within the counted areas is lost, and the data should be treated as geostatistical data.

2.2 Linear geostatistical methods

In a geostatistical model, the observation y_i is treated as the value of the variable y at location s_i , denoted as $y(s_i)$. In turn, the observed value $y(s_i)$ is considered to be a realization of the random variable $Y(s_i)$, which is part

of the random field

$$\{Y(s), s \in \mathbf{D}\}$$

with \mathbf{D} the area studied. A general model for the variable is

$$Y(s) = \mu(s) + e(s)$$

with $Y(s)$ the variable that can be observed, $\mu(s)$ the expected value of Y or the *trend* at location s , and $e(s)$ a residual, zero-mean random variable.

2.2.1 Spatial dependence

In order to allow inference on the value of $Y(s_0)$ at an unsampled location s_0 from sample data, we need to know the spatial dependence between pairs $Y(s_i)$ and $Y(s_j)$. In general only a single realization of the random field $Y(s)$ is available, and this is not enough to tell something about the relation between $Y(s_i)$ and $Y(s_j)$. Therefore, we need to impose some structure on the random component of Y . The following assumptions are common:

Stationarity—If we assume that Y is a *stationary* process, the dependence between $Y(s_i)$ and $Y(s_j)$ is a function of the vector $h = s_i - s_j$ only. When multiple data pairs are available for a vector of (approximate) size h , these can be used to estimate the dependence between $Y(s_i)$ and $Y(s_j)$ – and all other pairs with this separation vector.

Isotropy—We can further assume that the process Y is isotropic, in which case the relation between $Y(s_i)$ and $Y(s_j)$ depends not on the vector h but only on $|h|$, the distance between s_i and s_j .

Multivariate distribution—When $Y(s)$ follows a multivariate Gaussian (normal) distribution, the complete multivariate distribution of $Y(s)$ is characterized by the mean and covariance of $Y(s)$. A wider class of distributions is obtained when a non-linear but *known* transform of $Y(s)$ follows a multivariate Gaussian distribution.

2.2.2 Trend functions

In the simplest model, the mean of $Y(s)$ is known, e.g.

$$Y(s_i) = \mu(s_i) + e(s_i)$$

with $\mu(s_i)$ a known function, and $e(s_i)$ a zero-mean, stationary random variable. When the value of $Y(s)$ at an arbitrary location s_0 , $Y(s_0)$ has to be

predicted, this model leads to *simple kriging*. Although the assumption that $\mu(s)$ is truly known is usually unrealistic, simple kriging is useful when for some other reason the estimation of the trend function cannot be integrated in the geostatistical analysis.

More realistic models for $Y(s)$ include an unknown mean function. The simplest form would involve an unknown but spatially constant mean m ,

$$Y(s_i) = m + e(s_i),$$

with $e(s_i)$ a zero-mean, stationary random function, and this leads to the use of *ordinary kriging* for the prediction of $Y(s_0)$.

A wider class of functions is obtained when the trend is modelled as a linear combination of unknown coefficients β_j and known base functions (or covariates, or explanatory variables) $X_j(s)$,

$$Y(s_i) = \sum_{j=0}^p \beta_j X_j(s_i) + e(s_i),$$

which would lead to using *universal kriging* for the prediction of $Y(s_0)$ (Cressie, 1993). Note that usually $X_0(s) \equiv 1$ and β_0 is an intercept, in which case the universal kriging model extends the ordinary kriging model. It should also be noted that the $X_j(s)$ should be *known* at all data locations, as well as at all locations s_0 where predictions for Y are required. In matrix notation, for n observations the last equation would read:

$$Y(s) = X(s)\beta + e(s)$$

and has expectation:

$$E(Y(s)) = \mu(s) = X(s)\beta \quad (2.1)$$

with $Y(s) = (Y(s_1), \dots, Y(s_n))'$, with $X_j(s) = (X_j(s_1), \dots, X_j(s_n))'$ the j -th column in the $n \times (p+1)$ matrix X , and with $\beta = (\beta_0, \dots, \beta_p)'$ the vector with unknown constants.

2.2.3 Which trend model should we choose?

When only observations and their spatial coordinates are available, the only base functions that can possibly be used for the trend are functions of the spatial coordinates, since only these functions are known at data locations and all other locations. Using (functions of) spatial coordinates as base functions may be effective in those cases where the presence of the trend is evident, e.g. because it is overly clear or when it should be there, according to prior knowledge of the (genesis of the) observed process. When in doubt, and

when there is little prior evidence for a complex coordinate trend function, assuming a constant mean to be *locally* constant may be sufficient, or even better, to accommodate local departures from a *globally* constant mean, than using coordinate polynomials (Journel and Rossi, 1989).

Examples of coordinate polynomial trend function in ecological applications are: Weseloh (1996), Augustin et al. (1996) and Buckland and Elston (1993). The latter state that greater biological knowledge is required to ensure inclusion of appropriate covariates. Use of spatial coordinates such as Easting and Northing as proxies for the relevant covariates will lead to poor model predictions beyond the calibration (data) region.

When, apart from spatial coordinates, other base functions (explanatory variables) are present that may be related to the observed process, these should be incorporated as base functions, whilst they explain part of the variation present in the data. Generally, by combining data and relevant base functions, more realistic predictions of Y will be obtained. Care should be taken when only linear models in the base functions are assumed: the base functions may have to be transformed non-linearly prior to their application in a linear function.

Usually, base functions will only succeed in explaining the slow, gradual variation in Y (Kitanidis, 1993). Using more (or better) base functions will therefore not only lead to a residual with less variance, but also to a residual with less spatial correlation. This argument can be reversed: when less (or no) base functions are present, observations (residuals) tend to be stronger spatially correlated than when these functions are present, because the residual will include the large frequency variation of the base functions.

Another argument in favour of using base functions to model the trend function is that we seem to understand (we can 'explain') the variation present in the trend function, but by lack of knowledge we will have to treat the remaining, residual variation as being random, since we cannot explain the latter. Understanding what is going on in the data is usually the preferred option in science. Finally, using information when it is available rather than ignoring it may always improve prediction.

Let us translate this to sea bird densities: suppose that we are studying *Fulmarus glacialis* on a moment that the individual birds are mostly occupied with collecting food, and we know the availability of one of its important sources of food. In this case, we can use this food variable to model the trend function in the data. However, this function may not explain the variance very well because

- the function used does not describe the true relation of *Fulmarus glacialis* to the food variable (lack of fit);

- the food variable is approximately known but prone to error, e.g. measurement or interpolation error (error in base functions);
- other, unknown food sources (or other unknown factors) have an important influence on the behaviour of *Fulmarus glacialis* as well;
- the observations of *Fulmarus glacialis* are subject to measurement error.

Increasing influence of either of the first three aspects will typically lead to increase of spatial correlation in the residual. In contrast to this, measurement error in the observations will usually add unstructured variation (noise).

2.2.4 Conditional and unconditional predictions

For the prediction of $Y(s_0)$, one can proceed in two ways: conditional to the data or unconditional¹ to the data.

When prediction is done unconditionally, the predicted value for $Y(s_0)$ is the trend function at s_0 :

$$\hat{Y}_u(s_0) = x(s_0)\hat{\beta}$$

with $x(s_0) = (X_0(s_0), \dots, X_p(s_0))$. At locations s_0 , $\hat{Y}_u(s_0)$ is the expected value of Y over the ensemble of *all realizations* of the random field Y . When a correlated value $Y(s_i)$ is nearby s_0 , or even when $Y(s_0)$ is known, it is not used to improve prediction because unconditional prediction involves all realizations together rather than one single realization, from which the data were obtained. Clearly, unconditional prediction can only capture the information present in the trend function.

For prediction of $Y(s_0)$, conditional to the observed data (conditional upon the realization at hand), the predicted value is

$$\hat{Y}_c(s_0) = x(s_0)\hat{\beta} + \hat{e}(s_0)$$

In addition to the trend function value at s_0 , $\hat{Y}_c(s_0)$ contains a non-zero predicted value for the residual. Although the residual has expectation zero over all realizations, conditional to the data the predicted value will be strongly varying, depending on the amount of correlated (nearby) data and the observed values.

When the goal is interpolation, we are clearly interested in the realization from which the data were obtained, and conditional prediction is the preferred option. In this case, the conditional prediction improves on unconditional prediction in the following ways:

¹this may seem a strange thing to do, but it tends to be the standard approach in most of the GLM or GAM applications to spatial ecological modelling.

- it shows more spatial differentiation, because local deviations from the trend function are revealed;
- it has smaller prediction variance (in case of predicting a single observation, i.e. the current realization);
- it reproduces the observed values (in case of absence of measurement error).

2.2.5 Change of support and the global mean

From the earliest stages on, geostatistics has been concerned with the *physical size* or *support* of measurements and objects for which predictions are required (Journel and Huijbregts, 1987). Often, measurements have a small support because the measurement device (or monitoring method) is constraint to physical limits. At this small support, the variable may show huge variations. Often however, the aim is to predict the observed variable at a much lower spatial resolution. At this lower resolution the variation of the variable is much smaller. Given point-measurements, it is possible that for unsampled point-locations the prediction error is huge, but that for larger areas (e.g. blocks the size of a few km²) the prediction error is reasonable.

Define the block average value of $Y(s)$ over B_0 as

$$Y(B_0) = \frac{1}{|B_0|} \int_{B_0} Y(u) du$$

with $|B_0|$ the area (or volume) of B_0 . The challenge in geostatistics is often to predict the block mean value $Y(B_0)$, conditional to the observed point-support data, $Y(s_i)$, along with a realistic prediction error. In linear geostatistics, this is obtained by (universal, ordinary or simple) block kriging. In non-linear geostatistics things may become much more complicated (Cressie, 1993). As a last resort, a simulation approach to approximate the conditional distribution of $Y(B_0)$ may be applied (Journel, 1992).

Conditional to the data, the *global mean* is $Y(B_0)$ with B_0 the complete area of interest. The best linear unbiased predictor of this quantity is obtained by the block kriging estimate.

2.3 Generalized linear and related models

2.3.1 Generalized linear models

Linear models for the trend are optimal when the data are normally distributed, and when there are no constraints to the possible values the data can take. In practice however, data may be severely constraint: they may be strictly positive, non-negative, or even take only the values 0 or 1. In that case, using a linear model for the trend may soon lead to predictions outside the range of possible values, and this should be avoided. For such data, the class of linear models is extended by the class of generalized linear models.

Generalized linear models (GLM, McCullagh and Nelder, 1989) extend the linear model (2.1) for the trend in regression (or universal kriging) models,

$$E(Y) = \mu(s) = X(s)\beta$$

to the more general form

$$g(\mu(s)) = X(s)\beta \quad (2.2)$$

allowing linear models on a different scale than the observation scale by choosing an appropriate *link* function $g(\cdot)$. The link function relates the predictor $\eta(s_0)$ to the expected value μ at location s_0 . The extension to normal linear models is the option of non-linear link functions. In addition to the link function, a *variance function* $V(\cdot)$ is defined. This function describes how the variance of observations depends on the mean value. The form is:

$$\text{Var}(Y(s)) = \phi(V(\mu(s))),$$

with ϕ a constant, the dispersion parameter. The usual set of link functions $g(\cdot)$ and variance functions $V(\cdot)$ provided by GLMs is found in McCullagh and Nelder (1989).

For application of GLMs, observations need to be independent, making these models unsuitable for spatial or temporal data (McCullagh and Nelder 1989, p. 21). GLM are mainly focused on model selection and estimation of systematic effects in data, not on the prediction of the value of unobserved data (interpolation).

2.3.2 Generalized additive models

Generalized additive models (GAMs) extend generalized linear models (2.2) by allowing smooth, 'non-parametric' functions in the explanatory variables. the GAM replaces the linear predictor on the right side of (2.2) with an additive predictor:

$$g(\mu(s)) = \alpha + \sum_{j=1}^p f_j(X_j(s))$$

where the $f_j(\cdot)$ are arbitrary functions, e.g. *smooth functions* (Hastie and Tibshirani, 1990), each function containing a number of parameters that depends on the smoothness (smoother functions need more parameters). In a GAM, f_j can also be a linear predictor, and combinations of linear and smooth terms are allowed.

Just like in GLMs, GAMs assume observations to be independent, making these models unsuitable for spatial or temporal data. Augustin et al. (1996) extended the GLM by including an autoregression term, using spatial nearest neighbours.

The main disadvantages for using GAMs for spatial modelling in the present study are:

- The software used for modelling the trend (S-Plus) does not currently accommodate interactions with non-parametric smooth terms, although these are in principle possible. Methods for modelling interactions in GAM are proposed by Hastie and Tibshirani (1990, p. 266-268). Exclusion of interaction between Easting and Northing would make the approach unsuitable for the application of coordinates as exploratory variables, as the results are no longer rotation invariant (Augustin et al., 1996). (Note that we do not consider coordinates in our GLM)
- For non-parametric or combined models (containing parametric and non-parametric terms) prediction errors are only supplied for data locations. One example of a method to obtain approximate prediction errors for unobserved (interpolated) locations is using the Bootstrap (Augustin et al., 1998). An alternative, approximate, approach is the interpolation, for unobserved locations, of the point-wise prediction errors obtained for the data points.
- An important implication of smoothed response shapes is the impossibility of prediction beyond the modelled range of the independent variable. This is not necessarily a bad thing, since any statistical inference beyond the modelled variable range carries in it a certain danger (James and McCulloch 1990). If we were to use smooth functions of coordinates as predictors in our trend component, we could not extrapolate beyond the range of observation coordinates. The prediction area proposed in this study exceeds the modelled range to some extent.
- in the standard framework, prediction can not be done conditionally on observed (correlated) data (as is true for GLM).

Smoothed responses perform generally better than polynomials in fitting complex responses, with the same number of degrees of freedom. Third

and higher-order polynomials have serious drawbacks (Austin et al. 1990; Hastie and Tibshirani 1990; Huisman et al. 1993; Trexler and Travis 1993; Austin et al. 1994). They lack flexibility and tend to produce spurious and (ecologically) unrealistic response shapes (Bio et al. 1998). The fact that polynomials are bound to rigid shapes causes higher-order polynomials to display strange edge effects. To fit the centre of the gradient, they tend to wave their edges. When the trend prediction error is taken into account, this effect is somewhat less of a problem, as trend prediction errors tend to be large on the “waving” edges: this would only confirm that there is little information available in the area where we tend to extrapolate a fitted function.

Smoothed responses are also more sensitive to outliers than smoothers: because smoothers are local interpolators, outliers affect only the response to a part of the predictor gradient. They can be seen as a suitable tool for exploratory data analysis, for finding relations between dependent and independent variables.

2.3.3 Generalized estimating equations

Generalized estimating equations extend traditional generalized linear models to models suitable for correlated data. Most applications are available in the field of random effects models, such as longitudinal data, and are oriented towards model selection and hypothesis testing (Liang and Zeger, 1986; Zeger and Liang, 1986; Diggle et al., 1994)

Using an alternating iterative fitting of (i) the trend parameters β and (ii) the correlation (or correlation function) parameters α , so called “Generalized Estimation Equations” (GEE) are obtained. The GEE approach allows for general correlation structures, and has for instance been used for the estimation of the trend function (and testing hypotheses) from spatially correlated data (Albert and McShane, 1995; Gotway and Stroup 1997).

Generalized estimating equations (GEE, Liang and Zeger, 1986) extend the class of Generalized Linear Models (2.2) to models with correlated errors. Writing $\mu(s) = \mu$, $Y(s) = Y$ and $X(s) = X$:

- $\mu(\beta(\alpha)) = E(Y)$ (satisfying $g(\mu(\beta)) = X\beta$, with $g(\cdot)$ the link function)
- $V(\mu, \alpha) = \text{Cov}(Y)$ (satisfying $V(\mu, \alpha) = A(\mu)^{1/2}R(\alpha)A(\mu)^{1/2}/\phi$, with $A(\cdot)$ the variance function and $R(\alpha)$ the correlation matrix) and
- $D = \partial\mu(\beta)/\partial\beta$

generalized estimating equations solve $D^T V(\mu, \alpha)^{-1} \{Y - \mu(\beta(\alpha))\} = 0$ through the Gauss-Newton iterations

$$\alpha^{(k)} = f(\beta^{(k-1)}, Y - \mu(\beta^{(k-1)}))$$

and

$$\beta^{(k+1)} = \beta^{(k)}(\alpha^{(k)}) + \{D^T V^{-1} D\}^{-1} D^T V^{-1} [Y - \mu(\beta^{(k)})]$$

The iteration usually starts by taking α^0 such that V is a diagonal matrix (i.e., assuming uncorrelated observations). Then, after each estimate of β , the correlation parameters α are re-estimated using this newly obtained estimate of β .

3 Generalized linear geostatistics

Here, we will present two paths to obtain spatially interpolated sea bird densities. The first is based on combining a generalized linear model for the trend and a geostatistical model for the residual, based on known, non-constant variances and a stationary correlogram. The combination is a fairly crude construction, because their subsequent assumptions—uncorrelated residuals for estimating the trend, spatially correlated residuals for spatial prediction—are in conflict with each other. We will call this method A.

As an alternative, method B will follow a more elaborate estimating scheme based on the generalized estimating equations for combined estimation of trend function and correlation parameters, as laid out in section 2.3.3.

3.1 Proposed method – A

The general idea of the approach proposed here is to use GLM/GAM modelling for the trend and to use geostatistics for modelling and prediction of the residual. It is an implementation of, and, in some respects, an extension of Gotway and Stroup (1997).

3.1.1 Trend

At any location s , the trend can be modelled under a generalized linear model as

$$g(\mu(s)) = \eta(s) = x(s)\beta$$

Since the data we will be working with are count data (or more precisely: counts transformed to spatial densities), the obvious first attempt will be to model the data as a Poisson process. For Poisson data, the link function is

$$g(\mu(s)) = \log(\mu(s))$$

and the variance of observations is equal to the mean:

$$\text{Var}(Y(s)) = \mu(s)$$

Assuming independent observations, for this model any standard software package (S-Plus, SPSS, SAS, Minitab,...) can be used to estimate the trend parameters β .

In practice one will often encounter the situation that the variance of an observed process does not equal the variance of the assumed Poisson process (μ_i) but exceeds it systematically, in which case *over-dispersion* is present. Several references argue that over-dispersion is the rule rather than the exception. In our approach, over-dispersion will be modelled as being proportional to the mean (i.e., $\text{Var}(y(s_i)) = \phi\mu(s_i)$ for some constant dispersion parameter ϕ ; McCullagh and Nelder, 1989, p. 199). Then, the underestimation of prediction variance will be compensated by including this factor into the model for the residual.

An alternative model that we will consider for the trend is a conditional Poisson distribution for the non-zero counts, and the binomial model for the presence (non-zero) or absence (zero counts) of birds (similar to Welsh et al., 1996). This will be done for one species (*Uria aalge/Alca torda*), in appendix C.

3.1.2 Residual variation and spatial prediction

Clearly, on the observation scale the residual of a GLM model cannot be modelled as a stationary random variable, because it has a variance that depends on $\mu(s)$, and $\mu(s)$ varies among observations. For binomial variables, the variance is $\mu(s)(1 - \mu(s))$, for Poisson variables the variance is $\mu(s)$ (assuming no over-dispersion).

Knowledge of the variance of the process can be used to infer stationary spatial correlation, which will be modelled from standardized residuals. This is the approach taken here. It is based on Albert and McShane (1995) and Gotway and Stroup (1997).

In case of Poisson variables with no overdispersion, standardized or Pearson residuals are defined as

$$p(s_i) = \frac{Y(s_i) - \hat{\mu}(s_i)}{\sqrt{\hat{\mu}(s_i)}}. \quad (3.1)$$

Pearson residuals can be used to obtain estimates for the spatial correlation (Albert and McShane, 1995). Under correct model assumptions, these

residuals have zero mean and unit variance. The semivariogram of these residuals,

$$\gamma_p(s_i, s_j) = \frac{1}{2} E(p(s_i) - p(s_j))^2$$

can be estimated from Pearson residuals when the residuals are assumed to be second order stationary (i.e., $\gamma(s_i, s_j) = \gamma(s_i - s_j)$, meaning that the semivariance only depends on the spatial separation vector), which is equivalent to assuming that the raw, unstandardized residuals have a stationary correlogram¹. When a semivariogram model is fit to the sample residuals, correlation between any two location pairs can be obtained by using the correlogram $\rho(h) = 1 - \gamma_p(h)$ with h the distance between the locations.

When over-dispersion is present, the standardized residuals $p(s)$ will have a variance larger than 1, Eq. 3.1 assumes absence of over-dispersion. When the over-dispersion is modelled as a simple multiplication factor (under the Poisson mode, $\text{Var}(Y(s_i)) = \phi\mu(s_i)$), the variance of the $p(s_i)$ will approximate the over-dispersion factor ϕ when the number of parameters p is small compared to the number of observations n . Consequently, modelling a semivariogram from standardized residuals is equivalent to modelling a scaled inverse of the correlogram, or

$$\phi(1 - \rho(h)).$$

Spatial prediction for the observed process $Y(s)$ proceeds as follows (Gotway and Stroup, 1997): assuming the mean function for $Y(s) = (Y(s_1), \dots, Y(s_n))'$ and $Y(s_0)$ can be written as

$$E(Y(s)) = \mu(s)$$

$$E(Y(s_0)) = \mu(s_0)$$

with $\mu(s)$ an $n \times 1$ vector with the mean value at each data location and $\mu(s_0)$ the mean value at the (arbitrary) prediction location s_0 . Variances and covariances between all variables can be written as

$$\text{Var} \begin{pmatrix} Y(s) \\ Y(s_0) \end{pmatrix} = \begin{bmatrix} C & c_0 \\ c_0^T & \sigma^2(s_0) \end{bmatrix}$$

with C the $n \times n$ matrix with (known) covariances between observation locations, c_0 the $n \times 1$ vector with covariances between the s_i and s_0 , T denotes

¹ the correlogram gives the correlation between observations as a function of spatial separation vector

transpose and $\sigma^2(s_0) = \text{Var}(Y(s_0)) = \phi\mu(s_0)$. The best linear unbiased predictor can be expressed as

$$\hat{Y}(s_0) = \hat{\mu}(s_0) + c_0^T C^{-1}(Y(s) - \hat{\mu}(s)).$$

We can now proceed by taking $\hat{\mu}(s)$ and $\hat{\mu}(s_0)$ from the GLM for the trend. Given a semivariogram $\gamma_p(h) \approx \phi(1 - \rho(h))$ for the standardized residuals $p(s_i)$, variances and covariances for the response residuals

$$\hat{e}(s_i) = Y(s_i) - \hat{\mu}(s_i)$$

are obtained by

$$\text{Var}(Y(s_0)) = \phi\sigma^2(s_0)$$

$$\text{Cov}(Y(s_i), Y(s_j)) = \sigma(s_i)\sigma(s_j)\phi\rho(s_i - s_j)$$

with $\sigma^2(s_i) = \text{Var}(Y(s_i)) = \hat{\mu}(s_i)$. This procedure is equivalent to (i) estimating the trend $\mu(s)$ by using GLM, and (ii) predicting the residual by simple kriging:

$$\hat{e}(s_0) = c_0^T C^{-1} \hat{e}(s), \quad (3.2)$$

with $\hat{e}(s)$ the $n \times 1$ vector with (predicted) residuals at observation locations, $Y(s) - \hat{\mu}(s)$. The simple kriging variance is:

$$\text{Var}(\hat{e}(s_0) - e(s_0)) = \sigma^2(s_0) - c_0^T C^{-1} c_0. \quad (3.3)$$

“Standard” simple kriging equations as implemented in standard geostatistical packages need to be modified to work with a stationary correlogram, a dispersion parameter and non-stationary (mean dependent and therefore location specific) variances, semivariances and/or covariances.

It should be noted here that deviding a non-zero residual through a *very* small value of $\sqrt{\hat{\mu}(s_i)}$ in Eq. 3.1 may leads to a highly extreme value for a Pearson residual. This occurs when possitive densities happen at very unexpected circumstances. Such outlying residuals may completely dominate the variogram and may have to be removed before a sensible analysis of spatial correlation can take place.

3.1.3 Predicting block means

Block means for a block B_0 (see section 2.2.5) are obtained by (i) replacing all point-point covariances in c_0 (i.e., the values $\text{Cov}(Y(s_i), Y(s_0))$) by the point-to-block covariances $\text{Cov}(Y(s_i), Y(B_0))$ in the simple kriging equations (3.2), and (ii) replacing the variance of the process $Y(s)$, $\sigma^2(s_0)$ in the simple kriging variance equation (3.3), by the variance of the block average process $Y(B)$, $\sigma^2(B_0)$. The latter variance is the average of pairwise covariances of all points in B_0 . Proofs of this and computational details are found in Journel and Huijbregts (1978) and Christensen (1991).

3.1.4 Prediction maps

Maps for the species densities can be obtained by adding the estimated trend to the predicted value of the residual:

$$\hat{Y}(s_0) = \hat{\mu}(s_0) + \hat{e}(s_0)$$

As noted by Gotway and Stroup (1997), when $\hat{\beta}$ is nonlinear (e.g. obtained by GLM) assessing the uncertainty associated with predictions based on this approach can be difficult.

In our, admittedly simple, approach, we approximate the prediction variance by

$$\hat{\sigma}_{\hat{Y}} = \text{Var}(\hat{Y}(s_0) - Y(s_0)) \approx \text{Var}(\hat{\mu}(s_0) - \mu(s_0)) + \text{Var}(\hat{e}(s_0) - e(s_0))$$

where $\text{Var}(\hat{\mu}(s_0) - \mu(s_0))$ is obtained from the GLM prediction, and $\text{Var}(\hat{e}(s_0) - e(s_0))$ is the simple kriging variance (associated with point kriging or block kriging).

Approximate 95% confidence intervals for the predicted bird density can be obtained by

$$[\hat{Y}(s_0) - 2\sqrt{\hat{\sigma}_{\hat{Y}}}, \hat{Y}(s_0) + 2\sqrt{\hat{\sigma}_{\hat{Y}}}] \quad (3.4)$$

These intervals can be shown on maps using one of the two ways, following the methods of Pebesma and De Kwaadsteniet (1997). Since we are constrained to black and white display here, only the second option is used.

3.2 Proposed method – B

3.2.1 Spatial generalized estimating equations

For spatial application of generalized estimating equations, α contains the variogram parameters, as the variogram completely determines the variance and correlation structure of a variable.

We can estimate α by (a) calculating the standard method-of-moment sample semivariogram for pre-defined distance classes from Pearson residuals obtained from the last β fitted, and (b) fitting a suitable parametric function to the sample residual variogram, using weighted least squares fitting (with the number of point pairs as weights). In addition, a known (pre-specified) measurement error can be given as an additive (variance) component to the Pearson residual variogram (Section 3.3). By ignoring unity constraints to the diagonal of $R(\alpha)$, the sill of the variogram becomes the dispersion parameter ϕ .

By using the variable ‘day’ as cluster id, we only included observation pairs for the variogram parameter estimation when they were observed at the same day, to avoid the cluttering of the spatial correlation estimates by temporal variation components.

3.2.2 Implementation

For the implementation of the GEE we used a modified and extended version of YAGS (Carey, 1998). YAGS was mainly developed for handling longitudinal (time series) data, but it was designed in such a way that extending it to work with arbitrary covariance functions (e.g. spatially correlated data) was allowed. The modifications and extensions to YAGS that were made for this purpose are listed in appendix F; they can be downloaded in source code form from the Internet.

3.3 Handling known measurement errors

Inevitably, observed sea bird density observations are subject to measurement errors. First of all, a systematic error will occur because it is likely that birds that are present will be missed, and it is unlikely that birds that are not present will be recorded. In addition to this mean or systematic error (bias), a random error will occur because

- counts for large groups or many small groups will be approximate
- the size of the observation strip fluctuates
- birds are easily missed (rather than observed while not being there), but the rate of missing fluctuates.

Here, we will address the consequences of measurement error for the proposed (generalized linear) modelling under fairly simplified circumstances (independently and identically distributed measurement error proportional to the true densities). Also, given estimates of the measurement error we would like to know what their influence is on the variogram of (Pearson) residuals.

Suppose the real (unobserved) density is Y and we observe densities \tilde{Y} , then, if measurement errors are taken proportional to the true densities we could assume the error model

$$\tilde{Y} = U \cdot Y$$

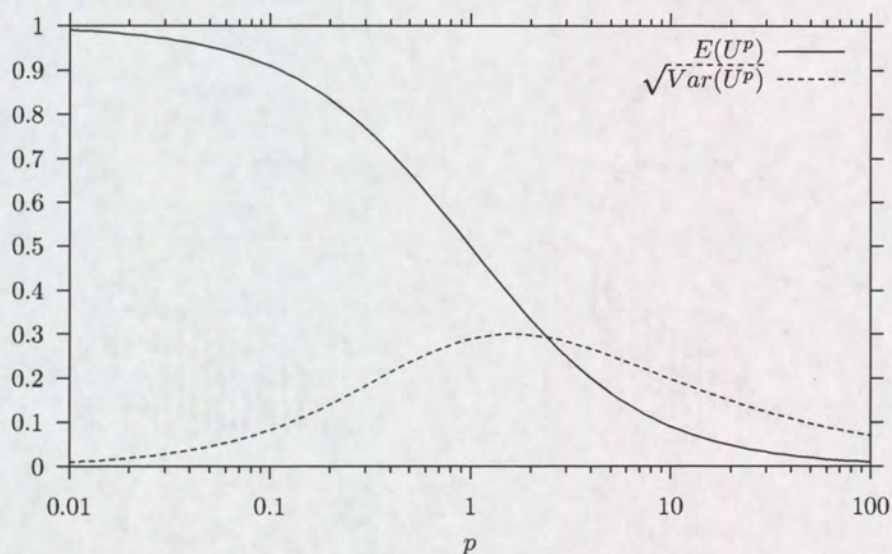
with

- U independent, identically distributed random variates
- U independent of Y
- $0 \leq U \leq 1$
- $E(U) = \mu_U$
- $\text{Var}(U) \equiv \sigma_U^2 > 0$

To provide an example of an error that complies to these restrictions, we can choose a suitable family of probability distributions. One possible family of distribution functions for U is that of powers of the uniform distribution on $[0, 1]$, $U^p \equiv (\text{Unif}[0, 1])^p$. It's first two moments are:

$$E(U^p) = \frac{1}{p+1}, \quad \text{Var}(U^p) = \frac{1}{2p+1} - \frac{1}{(p+1)^2}$$

Mean and standard deviation of U^p as a function of p are given in the following figure:



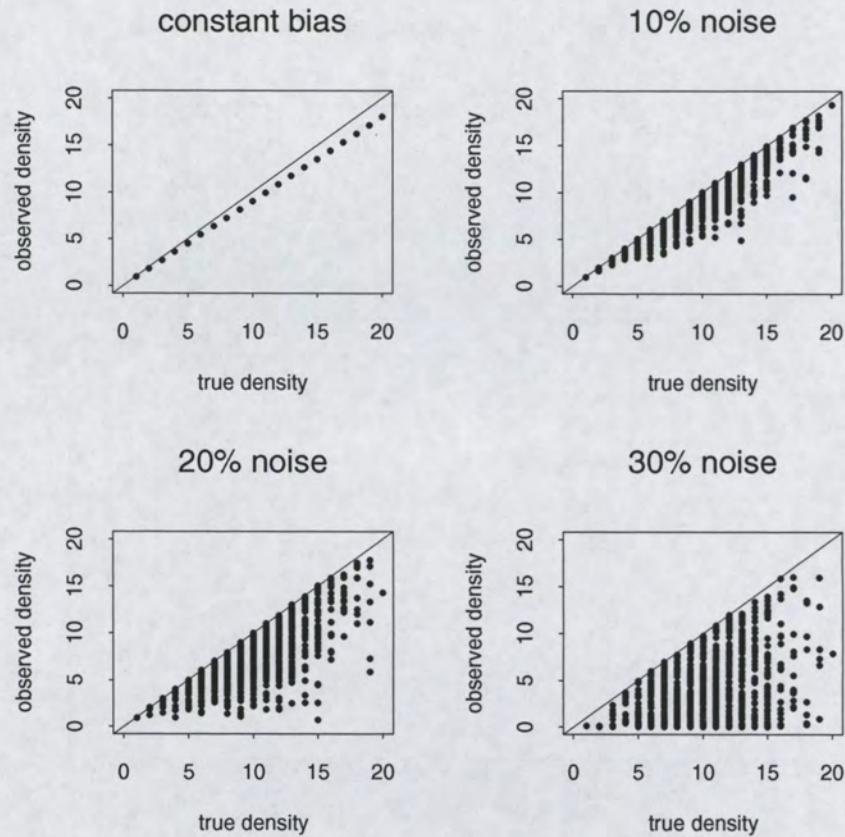
Proportional errors can be expressed as percentages. A proportional error of 20% means that given $Y = y_0$, the standard deviation of \tilde{Y} (as resulting from measurement error) is $0.2y_0$. Now when $\tilde{Y} = UY$, the conditional variance

$$\sigma_{\tilde{Y}|Y=y_0}^2 = \text{Var}(UY|Y = y_0) = \text{Var}(Uy_0) = y_0^2 \text{Var}(U) = y_0^2 \sigma_U^2$$

The corresponding standard deviation of \tilde{Y} is then $y_0\sigma_U$. For the family of $(\text{Unif}[0, 1])^p$ distributions, appropriate values for p corresponding to errors of 10%, 20% and 30% are:

| % error | $\sigma_{\tilde{Y} Y=y_0}$ | p |
|---------|----------------------------|-------|
| 10% | $0.1y_0$ | 0.126 |
| 20% | $0.2y_0$ | 0.354 |
| 30% | $0.3y_0$ | 1.5 |

The following plots give an example of constant measurement error and combinations of systematic and random measurement errors given the figures in this table.



3.3.1 Systematic error (bias)

Independence of U and Y implies that $E(\tilde{Y}) = E(UY) = E(U)E(Y)$. In generalized linear modelling with a log-link function (as in the methods pro-

posed here), the log of the mean response is modelled as a linear function of the covariates x :

$$\log(E(UY)) = \log(E(U)E(Y)) = \log(E(U)) + \log(E(Y)) = x^T \beta$$

Here, the intercept (a general mean level that is always present in β) absorbs the systematic part of the error, $E(U)$, and it will be biased (downward) by the amount of $\log(\mu_U)$ (a negative value).

3.3.2 Random error (noise)

When the systematic part of the error (bias, or mean error) has been taken care of, we may assume that the measurements are still distorted by a random error (noise) that is proportional to the true densities, as in

$$\tilde{Y} = U \cdot Y$$

but now with

- $E(U) = 1$ because the bias has been taken care of; this implies that $E(\tilde{Y}) = E(Y)$;
- U independent, identically distributed (IID)
- U independent from Y
- U has variance $\text{Var}(U) \equiv \sigma_U^2 > 0$.

The question now is: how do such measurement errors influence the variogram of Pearson residuals r_p , defined as

$$r_p = \frac{\tilde{Y}_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

The variogram of Pearson residuals, derived from \tilde{Y} , $\gamma_{\tilde{Y}}(h)$ will show larger values (semivariances) than that of Y when $\sigma_U^2 > 0$. Also, spatial correlation (the “structured” part of $\gamma_{\tilde{Y}}(h)$) can only be the result of spatial structure in Y , because U is IID and therefore contains no structure. The idea here is that when $U \equiv 1$ (i.e., $\sigma_U^2 = 0$) then $\gamma_{\tilde{Y}}(h) = \gamma_Y(h)$, and in any other case it will lead to an increase in variance that can be attributed to a nugget effect in the semivariogram.

Pearson residuals have zero expectation, and therefore $\text{Var}(r_p) = E(r_p^2)$:

$$E(r_p^2) = E\left(\frac{\tilde{Y}_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}\right)^2 = \frac{E(\tilde{Y}_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = \frac{\text{Var}(\tilde{Y}_i)}{\hat{\mu}_i} = \frac{\text{Var}(U_i Y_i)}{\hat{\mu}_i}$$

Now define $\text{Var}(Y_i) = \sigma_Y^2$, $\text{Var}(U_i) = \sigma_U^2$, $E(Y_i) = \mu_Y$ and $E(U_i) = \mu_U$. Then we can write $\text{Var}(U_i Y_i)$ as

$$\text{Var}(UY) = E(UY)^2 - (E(UY))^2$$

Because U and Y are independent, this can be written as

$$\text{Var}(UY) = (\sigma_U^2 + \mu_U^2)(\sigma_Y^2 + \mu_Y^2) - \mu_U^2 \mu_Y^2 = \sigma_U^2 \sigma_Y^2 + \mu_U^2 \sigma_Y^2 + \mu_Y^2 \sigma_U^2$$

When we take $\mu_U = 1$, this simplifies to

$$\text{Var}(UY) = \sigma_U^2 \sigma_Y^2 + \sigma_Y^2 + \mu_Y^2 \sigma_U^2 = \sigma_U^2 (\sigma_Y^2 + \mu_Y^2) + \sigma_Y^2$$

which, as expected, reduces to σ_Y^2 when $\sigma_U^2 = 0$.

In case of no measurement noise ($\sigma_U^2 = 0$), the variance (sill) of Pearson residuals is equal to

$$\text{Var}(r_p) = \frac{\sigma_Y^2}{\hat{\mu}_Y}$$

When a measurement error is present, this variance becomes

$$\text{Var}(r_p) = \frac{\sigma_U^2 (\sigma_Y^2 + \mu_Y^2) + \sigma_Y^2}{\hat{\mu}_Y}$$

The difference between these two is the variance “component” of the Pearson residual variogram that can be attributed to measurement error:

$$\frac{\sigma_U^2 (\sigma_Y^2 + \mu_Y^2) + \sigma_Y^2}{\hat{\mu}_Y} - \frac{\sigma_Y^2}{\hat{\mu}_Y} = \frac{\sigma_U^2 (\sigma_Y^2 + \mu_Y^2)}{\hat{\mu}_Y}$$

and this latter quantity can be used as an estimate of (or lower bound to) the nugget variance of the Pearson residual variogram.

4 Case study: 1995 data

4.1 Exploratory data analysis

Bird densities were recorded from a plane, flying at approximately 500 ft (150 m) above the sea surface (Baptist and Wolf, 1993). Birds were recorded non-stop on a 150 m wide strip on one side of the plane, or on both sides when observation (weather, light) conditions allowed this. Approximately, most counts were reported as totals of 2 minutes monitoring, resulting in strips of 6 km length, covering a surface of 1 km² each. The distribution of surface sizes of observations are shown in Fig. 4.1. A large number of counts registered for 1 minute were grouped (aggregated) to 2 minute counts, in order to get a more uniform monitoring strip size. Counts longer than 2 minutes were kept as such.

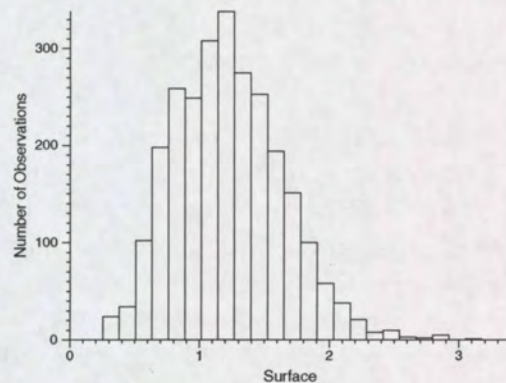


Figure 4.1: Surface (km²) of bird observations during 1995, with most 1-minute counts aggregated to 2-minute counts

To get rid of the (varying) strip sizes, all counts were divided by their strip surface. This results in a variable which will be called *observed bird density*, with units number of birds per km². Variation in strip size (apart

from the 1 minute counts that were aggregated) was ignored in the analysis. (As an alternative, one could work directly with the observed counts, and add the strip size as a covariate, cf. McCullagh and Nelder, 1989, p. 12, 206.)

In 1995, six monitoring rounds were done (Witte, 1995a-f). Each round, when completed, consists of three observation days. When possible, flying dates were kept close together, but obviously weather conditions always prevailed in finding days suitable for observation. The monitoring dates for each round are listed in table 4.1.

| period | round | dates |
|---------|-------|----------------------|
| Dec/Jan | 1 | Jan 13, Jan 30 |
| Feb/Mar | 2 | Mar 11, 12 and 13 |
| Apr/May | 3 | May 25, 26 and 28 |
| Jun/Jul | 4 | Jun 11, 12 and 13 |
| Aug/Sep | 5 | Aug 16, 17 and 21 |
| Oct/Nov | 6 | Oct 26 and 30, Nov 3 |

Table 4.1: Monitoring dates per monitoring round, 1995

In the following, we will mention e.g. *Sterna sandvicensis*, period 3, when we mean the Apr/May monitoring round. In fact, the results will only reflect the data collected on the monitoring days in table 4.1, but we will ignore this distinction. Table 4.2 presents a list of sample statistics for each species and period studied.

For *Sterna sandvicensis*, period 1, 2 and 6 provided too few positive counts (1, 2 and 1 respectively) to allow statistical modelling of the data. Also, for *Melanitta nigra*, period 3, having the largest counts, provided too few positive counts to allow statistical modelling. The only sensible information for these species/periods that can be provided here is the dot map showing the actual observations: occurrence of positive counts may be considered as 'rare' for these species/periods. Section 5.4 and chapter 6 further discuss the monitoring of *Melanitta nigra*, where very few but huge clusters occur.

4.2 Explanatory variables

Three variables that were considered to be 'possibly useful' for explaining the spatial variation in observed bird density were available at the time of this study: depth to sea bottom (depth, dep), distance to the Dutch coast (dis-

| species | period | # obs | # pos | mean | max | figure |
|------------------------------|--------|-------|-------|-------|------|----------|
| <i>Fulmarus glacialis</i> | 5 | 352 | 123 | 1.3 | 14.7 | p. 72-73 |
| <i>Sterna sandvicensis</i> | 1 | 351 | 1 | .0036 | 1.25 | p. 74-75 |
| | 2 | 364 | 2 | .0052 | 1 | p. 76-77 |
| | 3 | 462 | 18 | .073 | 3.64 | p. 78-79 |
| | 4 | 401 | 23 | .149 | 7.89 | p. 80-81 |
| | 5 | 352 | 41 | .296 | 13.4 | p. 82-83 |
| | 6 | 420 | 0 | 0 | 0 | |
| <i>Uria aalge/Alca torda</i> | 2 | 364 | 74 | .432 | 8.94 | p. 84-85 |
| <i>Melanitta nigra</i> | 3 | 462 | 3 | 8.32 | 2840 | p. 86-87 |

Table 4.2: Summary statistics. period—see table 4.1; # obs—number of observations; # pos—number of non-zero observations; mean—mean density; max—maximum observed density; All statistics are computed for observations within the study (mapping) area of Fig. 4.2

tance, dis; ldis denotes log-distance) and benthos biomass. Maps of distance to the coast, water depth and benthos are shown in Fig. 4.2.

In order to use maps as explanatory variables in regression modelling, the information in the map should not be subject to substantial errors. The benthos map was interpolated from relatively sparse point sample data, using inverse distance weighted interpolation. The spatial pattern in this map constitutes for large parts of interpolation artifacts. No information about the (interpolation) error for the benthos was available, and the map was considered to be not suitable to serve as an explanatory variable in the regression modelling.

4.3 Trend models

Bird densities were fitted to two continuous abiotic predictor variables, depth and distance to shore (Fig. 4.2), by multiple Poisson regression. We used the Generalized Linear Models (GLM) option of the statistical package used for this (S-Plus: Chambers and Hastie 1993; Venables and Ripley 1994).

Models were built using a forward stepwise selection of predictor variables with response shapes of increasing complexity. Starting with the null model (with the intercept only), each candidate predictor was added first as a first-order term, followed by terms of increasing order (quadratic, cubic, etc.). The drop in residual deviance caused by the addition of each term was compared to a χ^2 . At each step, the statistically most significant term was added to the

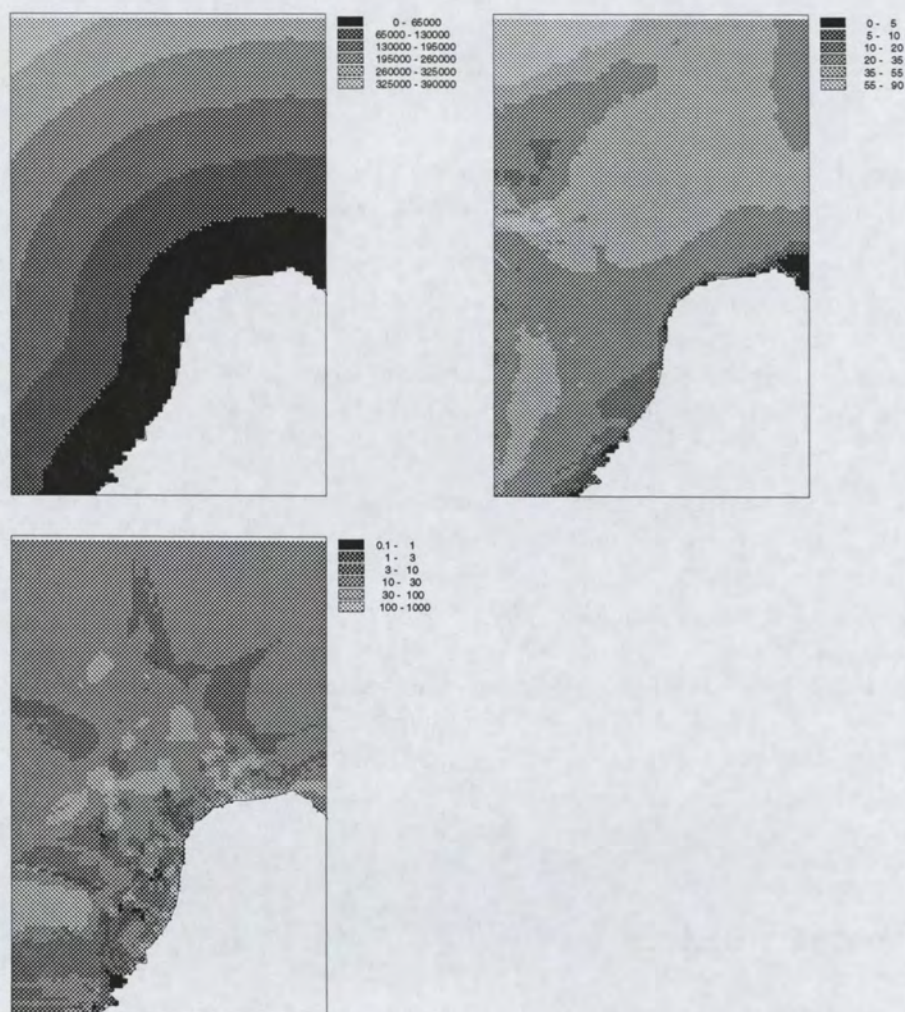


Figure 4.2: Maps of distance to the Dutch coast, m. (top, left), depth to sea bottom, m. (top, right) and benthos (bottom). The (x, y) coordinates of the map box extremes are (420646,5679620) and (775646,6224620), projection UTM31. The curved line is the Dutch North Sea coast line

null model, and the procedure repeated for the remaining terms or variable (Nicholls 1989; Yee and Mitchell 1991). Whenever not one of the single terms produced a significant decrease of deviance, the addition of two terms was tried. Selection stopped when, for neither of the variables, an increase in model order resulted in a change in deviance significant at the 0.01 level (Austin et al. 1990; Birks 1996).

As an indication of the variance explained by the regression model, the percentage of explained deviance (%D) is presented. The deviance depends on the proportion of positive observations in the data. Therefore, %D does not supply a direct measure of fit. It enables us, however, to compare models of different complexity applied to the same data. (Yee and Mitchell 1991; McCullagh and Nelder 1989; Smith 1994).

In addition, for *Uria aalge/Alca torda* an alternative approach was taken. The observed densities were first converted to a binary variable, assigning 0 to all zero densities and 1 to non-zero densities. Next, a (multiple) logistic regression model was obtained for this binary variable. The positive observations were finally modelled as a Poisson variable. Logistic and Poisson models were both obtained in the stepwise manner described above.

4.4 Residual models

As explained in subsection 3.1.2, the spatial correlation is modelled from Pearson residuals, $p(s_i)$. For 20 disjunct distance classes¹ $h_{[j]} = [(j-1) \times 5000, j \times 5000]$, $j = 1, \dots, 20$, the sample variogram of the Pearson residuals is obtained by

$$\hat{\gamma}_p(h_{[j]}) = \frac{1}{2N_j} \sum_{i=1}^{N_j} (p(s_i) - p(s_i + \tilde{h}))^2$$

with N_j the number of residual pairs considered for class h_j , and with $\tilde{h} \in h_{[j]}$.

To prevent temporal variability between consecutive observation days to interfere with the modelling of the spatial variation (correlation), a constraint was added to exclude residual pairs for the estimation of the spatial correlation when they were observed on different days. This constraint excluded about 25% of the point pairs, and improved the modelling of spatial correlation considerably.

No attempt was made to model the spatial correlation as a function of separation *direction*, we assumed the residuals to be isotropic. One of the reason for this is that the spatial distribution of residual pairs with different

¹distance classes of width 5000 were chosen to get both sufficient point pairs for all semivariance estimates and enough information on short distances at the same time

direction classes is very uneven over the area: most SW-NE pairs will be along the coast whereas most SE-NW pairs will be on the open sea.

For *Sterna sandvicensis* in period 3 and 4, the sample semivariogram showed surprisingly large values at the origin, and smaller values at larger distances. This indicates a negative correlation between residuals, which may result from fitting of many parameters to relatively few (positive) observations. For these species/periods², the sample covariogram was estimated by

$$\hat{C}(h_{[j]}) = \frac{1}{2N_j} \sum_{i=1}^{N_j} (p(s_i) - \hat{m})(p(s_i + \tilde{h}) - \hat{m})$$

with \hat{m} the sample mean of the Pearson residuals. A valid model fitted to the sample covariogram was used to model the spatial correlation. The semivariogram model $\gamma(h)$ was derived from the covariogram model $C(h)$ by

$$\gamma(h) = C(0) - C(h).$$

All residual semivariogram models found were of the exponential form, for $h > 0$:

$$\gamma(h) = a + b(1 - \exp(-h/c))$$

with a the nugget variance (i.e., zero for $h = 0$ and a for $h > 0$), b the (partial) sill and c the range parameter. In the semivariogram plots, such a semivariogram model is expressed as

$$a \text{ Nug}(0) + b \text{ Exp}(c).$$

Note that c is the range *parameter* and that the semivariogram *range*, the distance at which the exponential part of the semivariogram is at 95% of its (asymptotic) maximum value, approximates $3a$.

Semivariogram model parameters were estimated by weighted nonlinear regression, using weights $N_j/\gamma^2(\bar{h}_j)$ for semivariograms and N_j for covariograms, \bar{h}_j being the average separation distance of the N_j point pairs.

All geostatistical computations were done using gstat (Pebesma and Wesseling, 1998). For the purpose of simple kriging non-stationary residuals the source code was modified.

Sample semivariograms, fitted semivariogram models and their parameters are shown for each species in the corresponding sections in Chapter 5.

²we recommend a more thorough analysis of regression modelling and residual correlation modelling along the lines of proposed method B here

5 Results

For each species the semivariograms (or covariograms) are shown. In addition, four maps are shown with the position of the approximate prediction interval (Eq. 3.4) for the 5 km \times 5 km block mean bird density relative to four different levels. This relative position can be: *lower* when the complete interval falls below the level, *higher* when the interval falls above the level, or *not distinguishable* when the interval straddles the level. In the latter case, predicted density and level cannot be distinguished based on available information.

Appendix B shows the observed data for each species/period considered, and shows histograms of the observed densities and of the observed non-zero densities, along with a Poisson distribution having the same mean.

5.1 *Fulmarus glacialis*

The GLM resulting from the stepwise model selection has the following features:

| | |
|----------------------|---------------------------------|
| Terms in final model | dis + dep + dis \times dep |
| Null Deviance | 1354 on 351 degrees of freedom |
| Residual Deviance | 526.3 on 348 degrees of freedom |
| Explained Deviance | 61% |
| Dispersion parameter | 2.091 |

where 'dis' denotes distance, 'dep' denotes depth and dis \times dep indicates the first-order interaction (product) between distance and depth.

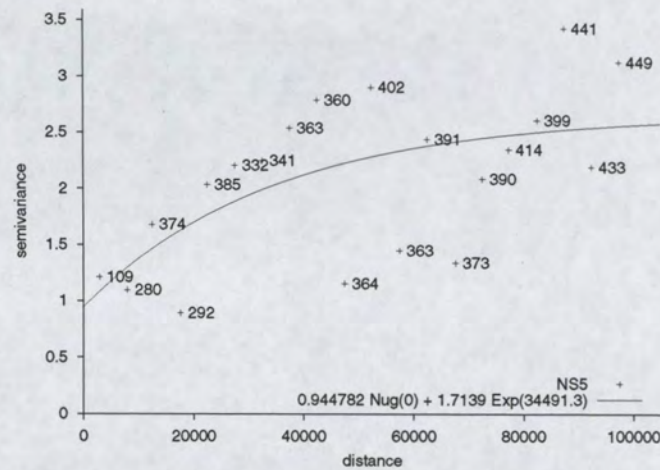


Figure 5.1: Sample semivariogram (+) and fitted semivariogram model (—) for *Fulmarus glacialis*, period 5. Numbers indicate N_j , the number of residual pairs

5.2 *Sterna sandvicensis*

For the first two and for the last period, no GLM could be fit for this species. This is probably due to the sparse number of positive observations.

Distance to coast was converted to $\log(\text{distance})$ before it was used as an explanatory variable, because all *Sterna sandvicensis* observations occurred at very small distances.

5.2.1 Period 3

For the period Apr/May 1995, the GLM resulting from the stepwise model selection has the following features:

| Terms in final model | $\log(\text{dis}) + \log(\text{dis})^2$ |
|----------------------|---|
| Null Deviance | 226.9 on 461 degrees of freedom |
| Residual Deviance | 187.7 on 459 degrees of freedom |
| Explained Deviance | 17% |
| Dispersion parameter | 1.269 |

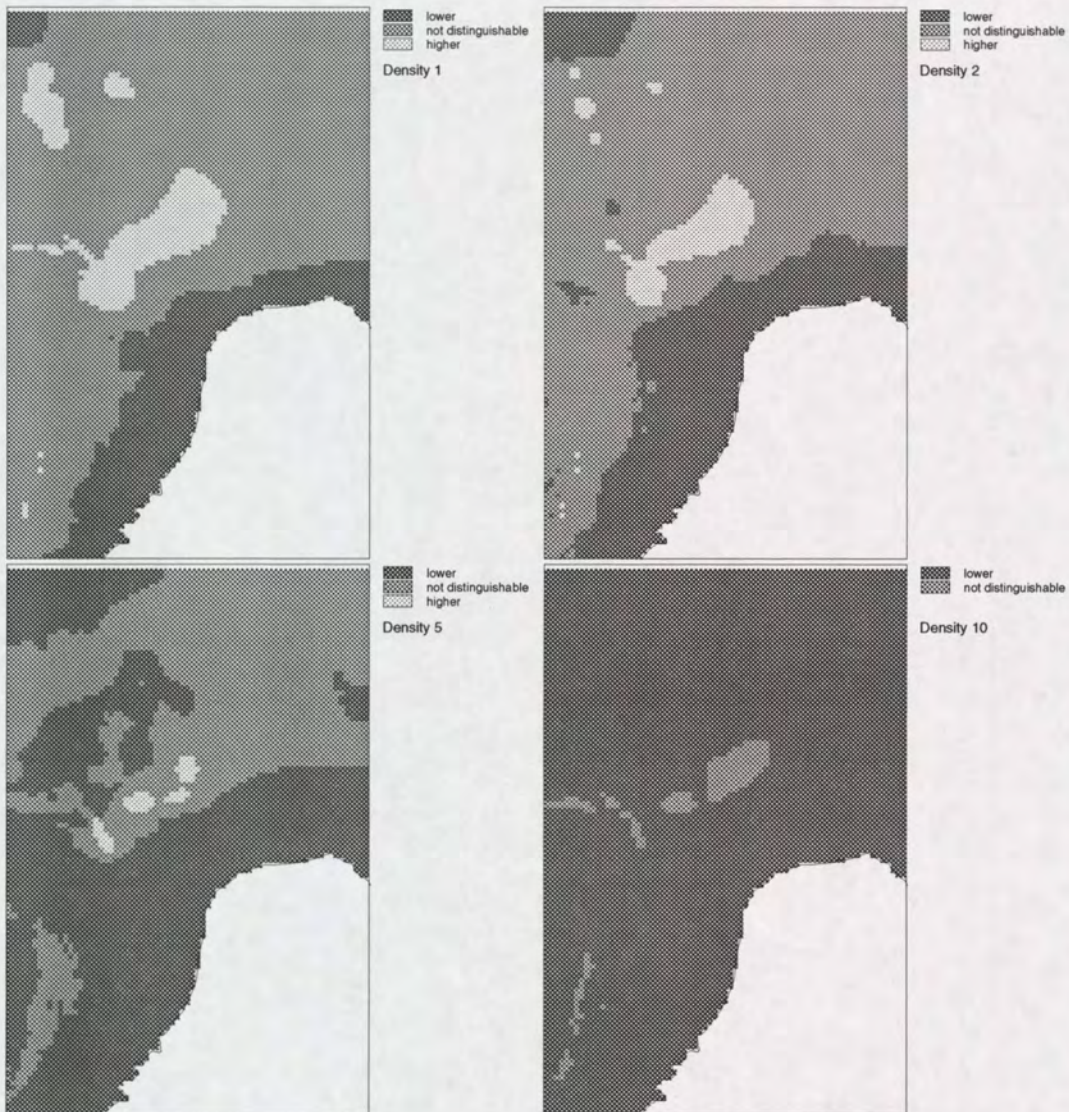


Figure 5.2: Map of *Fulmarus glacialis*, period 5. 95% Prediction intervals for 5 km \times 5 km block mean densities related to four density levels

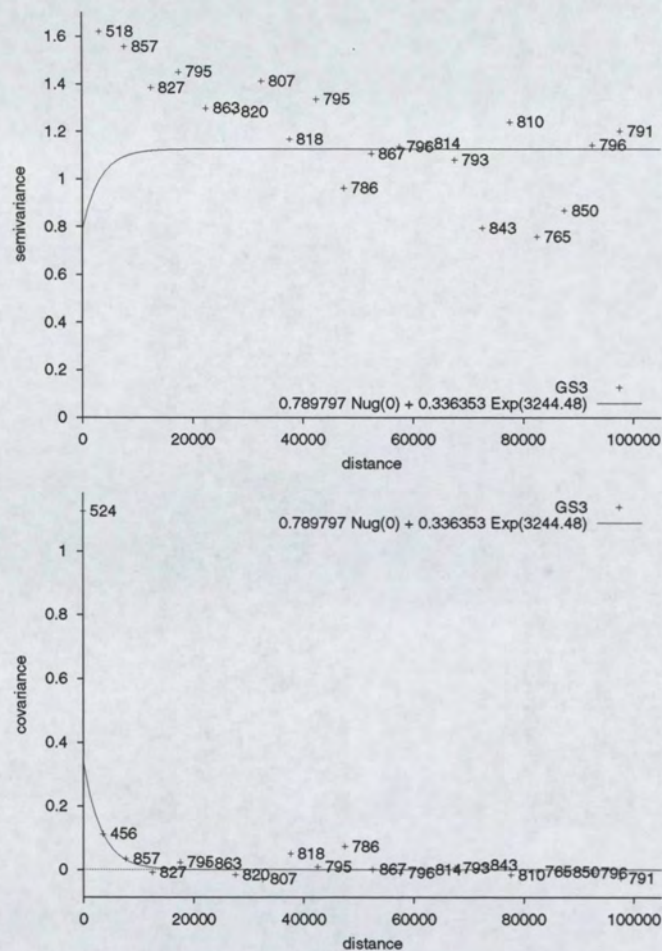


Figure 5.3: Sample semivariogram (+, top) and sample covariogram (+, bottom) with fitted model (—) for *Sterna sandvicensis*, period 3. Numbers indicate N_j , the number of residual pairs. Semivariogram model was fitted to the sample covariogram

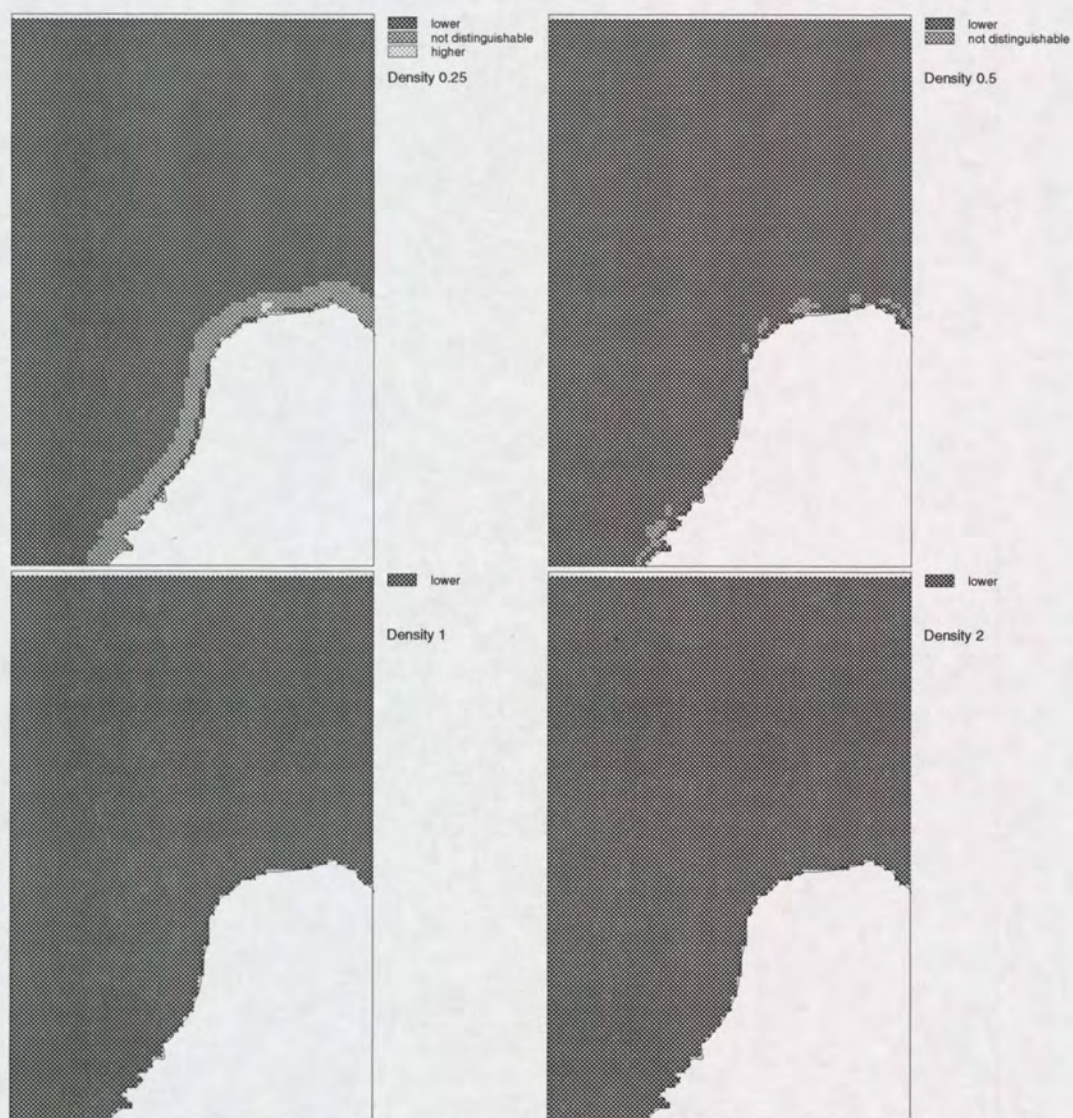


Figure 5.4: Map of *Sterna sandvicensis*, period 3. 95% Prediction intervals for $5 \text{ km} \times 5 \text{ km}$ block mean densities related to four density levels

5.2.2 Period 4

For the period Jun/Jul 1995, the final GLM is as follows:

| | |
|----------------------|---|
| Terms in final model | $\log(\text{dis}) + \log(\text{dis})^2$ |
| Null Deviance | 382 on 400 degrees of freedom |
| Residual Deviance | 251.4 on 398 degrees of freedom |
| Explained Deviance | 34% |
| Dispersion parameter | 1.449 |

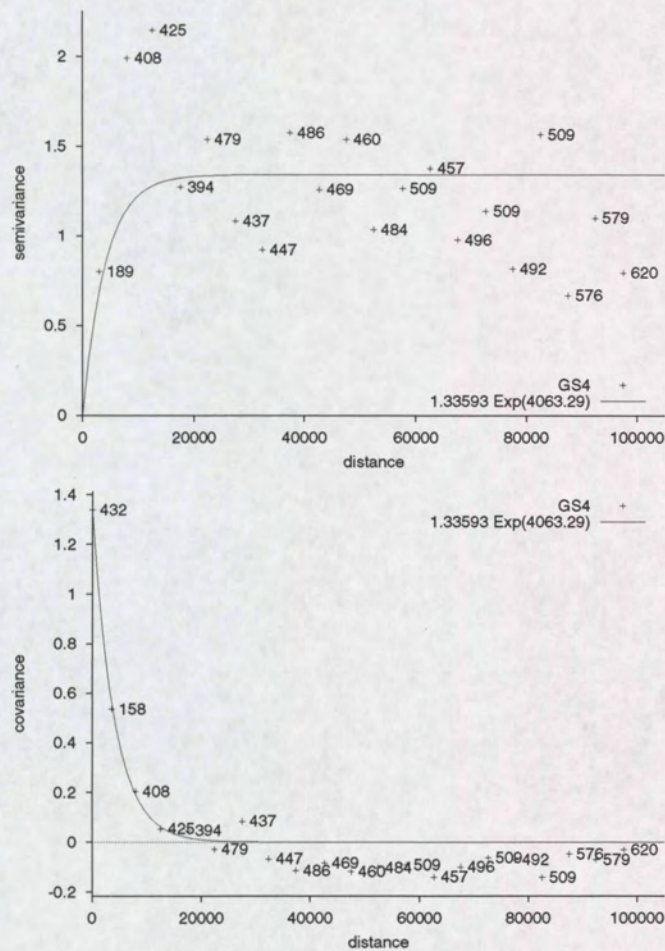


Figure 5.5: Sample semivariogram (+, top) and sample covariogram (+, bottom) with fitted model (—) for *Sterna sandvicensis*, period 4. Numbers indicate N_j , the number of residual pairs. Semivariogram model was fitted to the sample covariogram

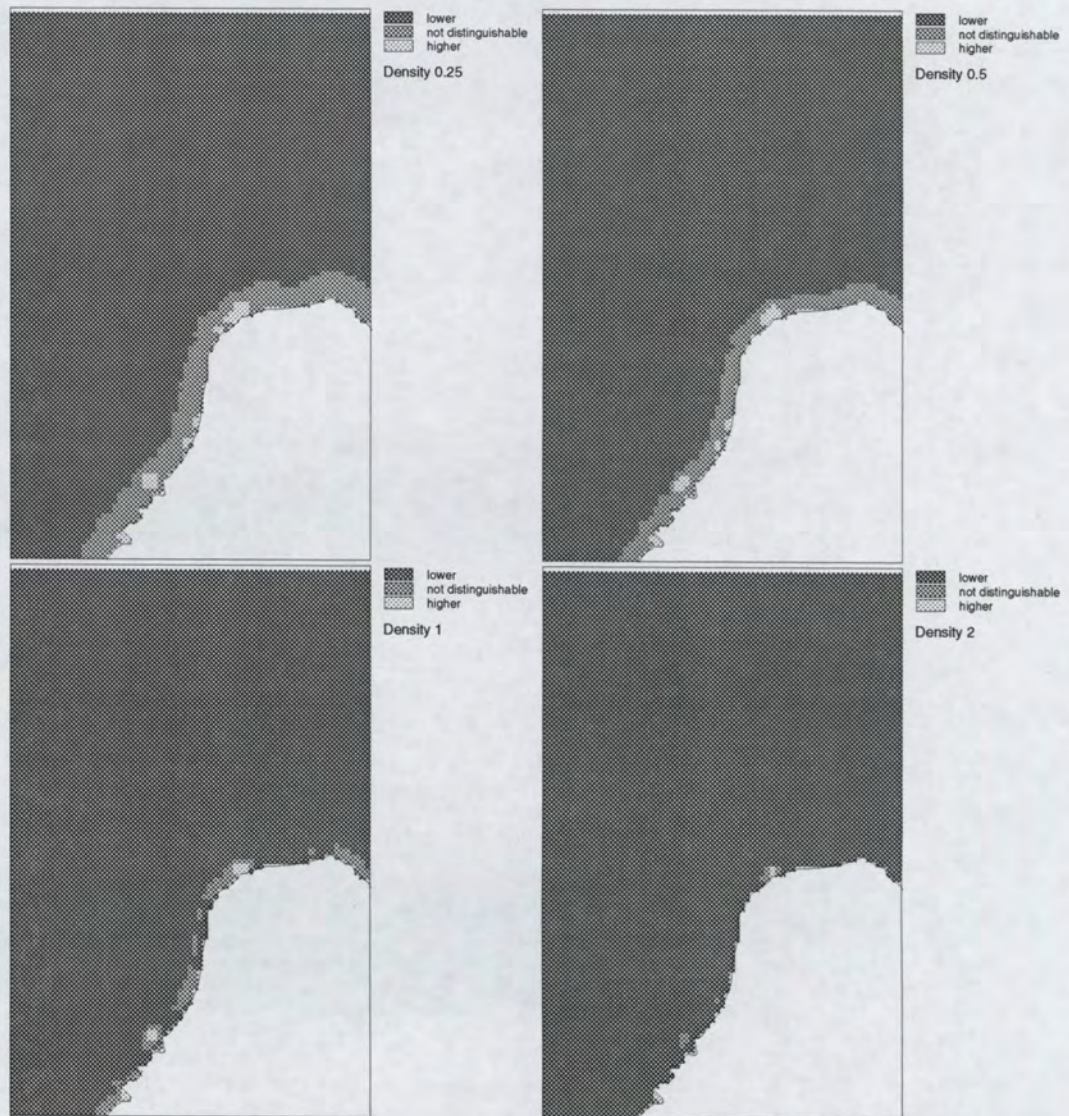


Figure 5.6: Map of *Sterna sandvicensis*, period 4. 95% Prediction intervals for 5 km \times 5 km block mean densities related to four density levels

5.2.3 Period 5

For the period Aug/Sep 1995, the final GLM is as follows:

| | |
|----------------------|--|
| Terms in final model | $\log(\text{dis}) + \log(\text{dis})^2 + \text{dep}$ |
| Null Deviance | 529.4 on 351 degrees of freedom |
| Residual Deviance | 321.5 on 348 degrees of freedom |
| Explained Deviance | 39% |
| Dispersion parameter | 2.854 |

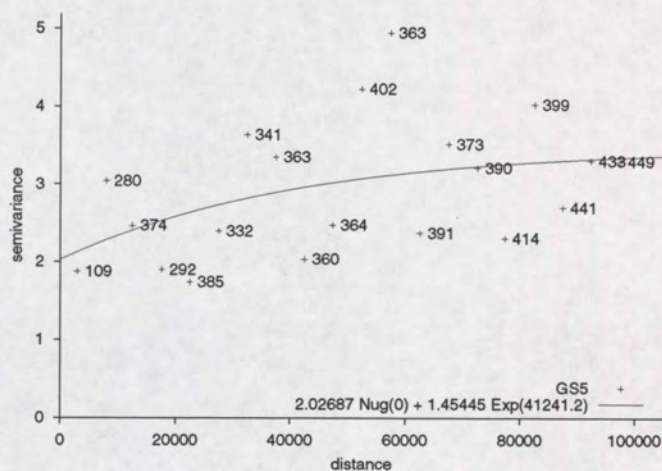


Figure 5.7: Sample semivariogram (+) and fitted semivariogram model (—) for *Sterna sandvicensis*, period 5. Numbers indicate N_j , the number of residual pairs

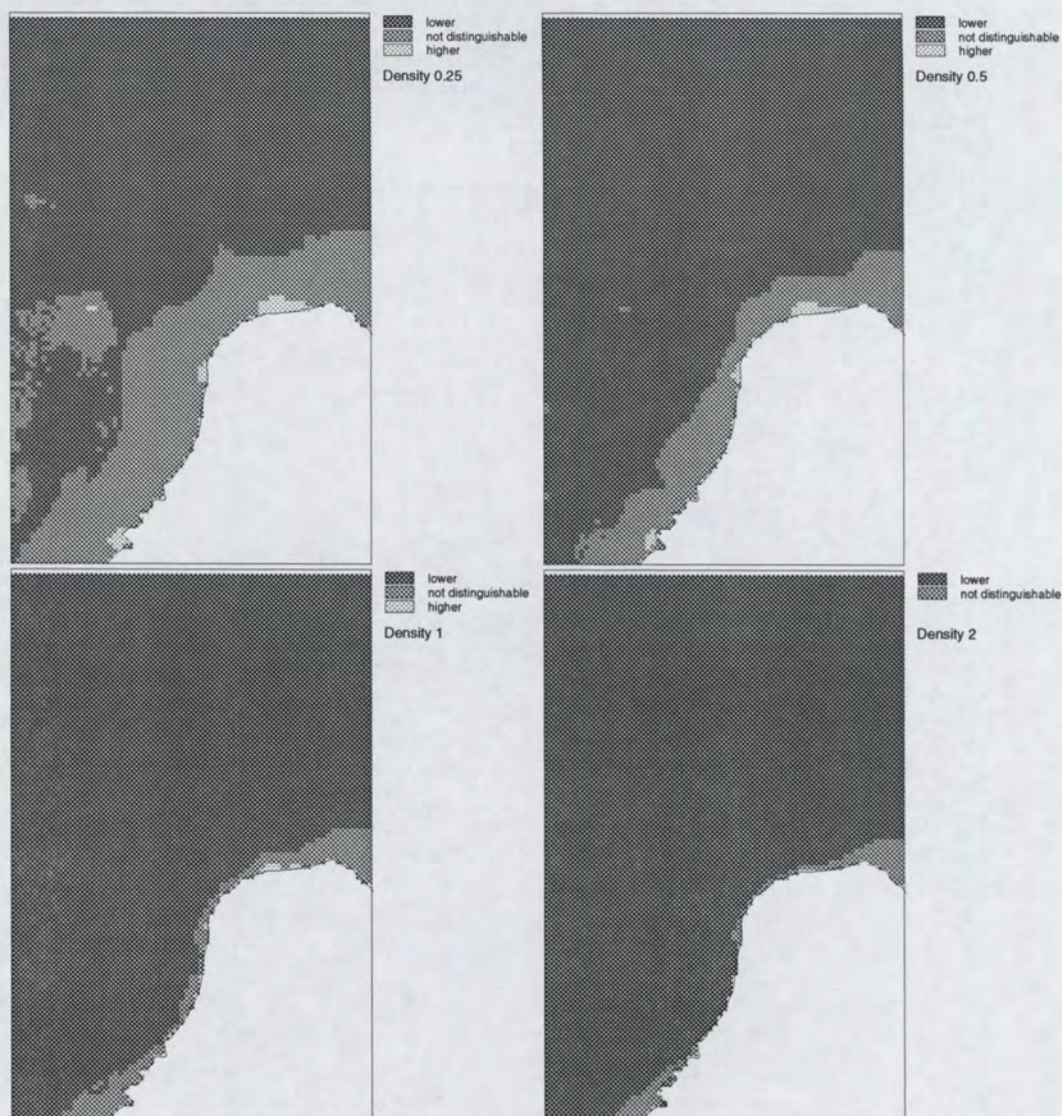


Figure 5.8: Map of *Sterna sandvicensis*, period 5. 95% Prediction intervals for $5 \text{ km} \times 5 \text{ km}$ block mean densities related to four density levels

5.3 *Uria aalge/Alca torda*

The GLM resulting from the stepwise model selection has the following features:

| | |
|----------------------|--|
| Terms in final model | $\text{dis} + \text{dis}^2 + \text{dis}^3 + \text{dep} + \text{dep}^2 + \text{dep}^3 + \text{dep}^4 + \text{dep}^5 + \text{dis} \times \text{dep}$ |
| Null Deviance | 614.6 on 363 degrees of freedom |
| Residual Deviance | 420.2 on 354 degrees of freedom |
| Explained Deviance | 32% |
| Dispersion parameter | 2.104 |

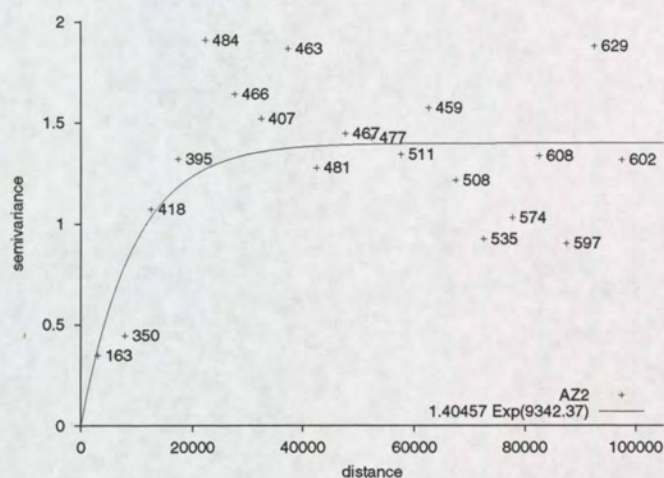


Figure 5.9: Sample semivariogram (+) and fitted semivariogram model (—) for *Uria aalge/Alca torda*, period 2. Numbers indicate N_j , the number of residual pairs

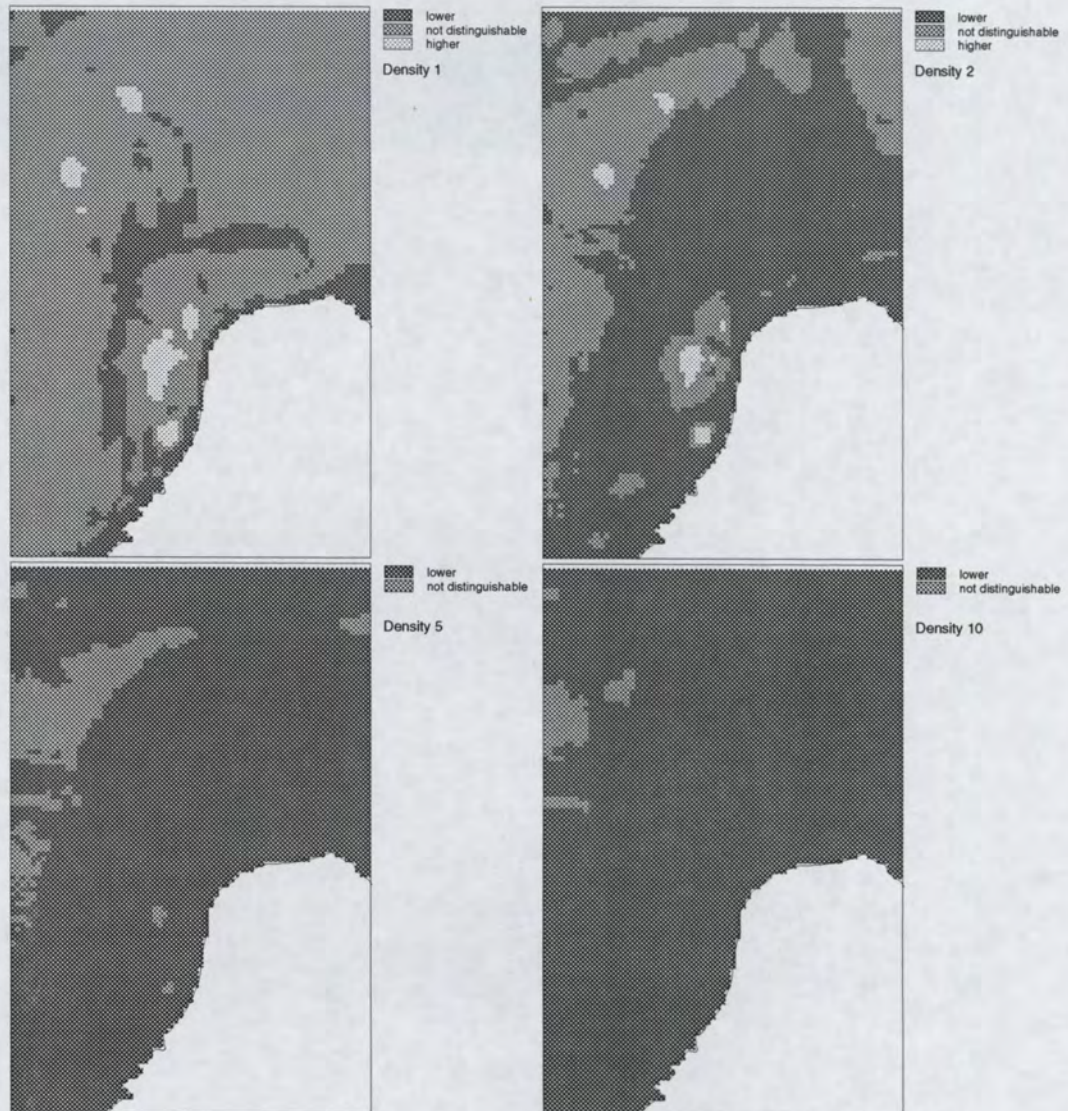


Figure 5.10: Map of *Uria aalge/Alca torda*, period 2. 95% Prediction intervals for 5 km \times 5 km block mean densities related to four density levels

5.4 *Melanitta nigra*

Very few large cluster of *Melanitta nigra* were observed during the period Apr/May 1995. As a consequently no GLM or residual model was obtained.

The best summary of the data seems to be the data themselves (observed bird densities: 18.52, 987.65 and 2839.51 birds/km²) accompanied by a map where these densities were observed, found on p. 86.

5.5 Method – B: *Fulmarus glacialis*

In order to compare results obtained from using generalized estimating equations (GEE) or the simpler general linear modelling (GLM) approach we applied both to a single model for *Fulmarus glacialis*, period 5. The model for the trend was:

$$\log(\mu(s)) = \beta_0 + \beta_1 \times \text{dis}(s) + \beta_2 \times \text{dep}(s)$$

with $\mu(s) = E(Z(s))$, the expected density at location s , with $\text{dis}(s)$ the distance to coast at location s (m), with $\text{dep}(s)$ the depth at location s (m). The variogram for the residual term was taken as an exponential model without a nugget:

$$\gamma_r(h) = c_1(1 - \exp(-h/a))$$

with sill c_1 and range a .

| | GEE | (s.e.) | GLM | (s.e.) |
|-----------|----------|--------|---------|---------|
| β_0 | -0.868 | 0.58 | -0.695 | 0.22 |
| β_1 | -2.49E-6 | 4.2e-6 | 1.78E-6 | 1.75e-6 |
| β_2 | 0.0406 | 0.016 | 0.0207 | 0.0086 |
| c_1 | 5.93 | | 33.0 | |
| a | 28500 | | 334000 | |

Table 5.1: Regression and variogram coefficients

Table 5.1 shows the trend and variogram coefficients for both approaches. Figure 5.11 shows the sample variograms and fitted models of both approaches, figure 5.12 shows the maps with confidence intervals for the GEE approach, and figure 5.13 shows the confidence intervals for the GLM approach. Note that in the resulting maps only the prediction (block kriging) variance $\text{Var}(\hat{e}(s_0))$ of the residual term was accounted for, because the GEE does not (as of yet) provide trend estimation variances $\text{Var}(\hat{\mu}(s_0))$.

Even though the regression coefficients are rather different, the maps from both approaches are surprisingly similar. Clearly, most of the information comes from the spatial correlation in the residual. Both approaches have a strong spatial correlation.

Larger differences may be expected when a more complex trend function is fitted, and more complex variogram models are fitted in the procedure. However, as noted in Appendix G, this may also result in complications (divergence) in the fitting procedure.

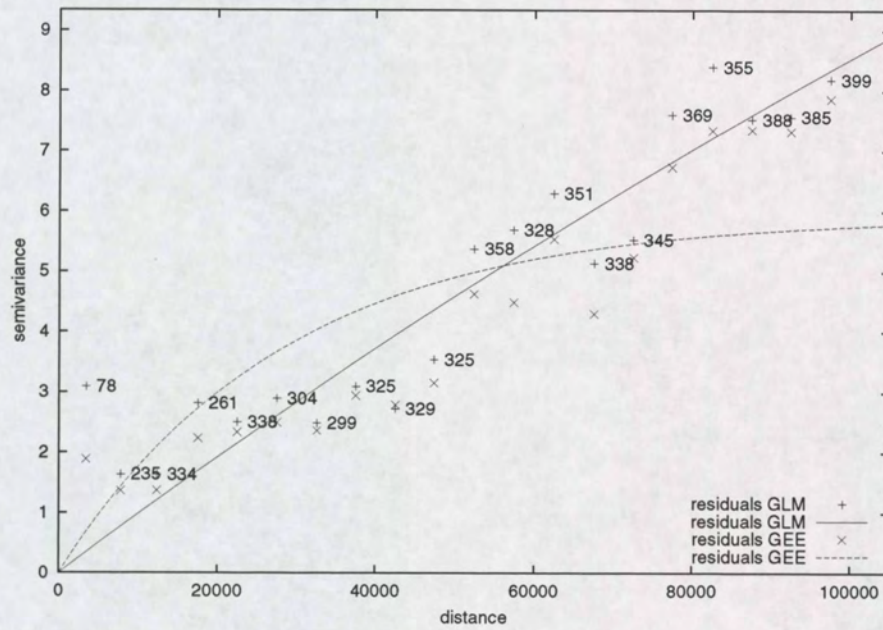


Figure 5.11: variogram of residuals for *Fulmarus glacialis*, period 5, obtained by GLM (+, solid line) and GEE (x, dashed line).

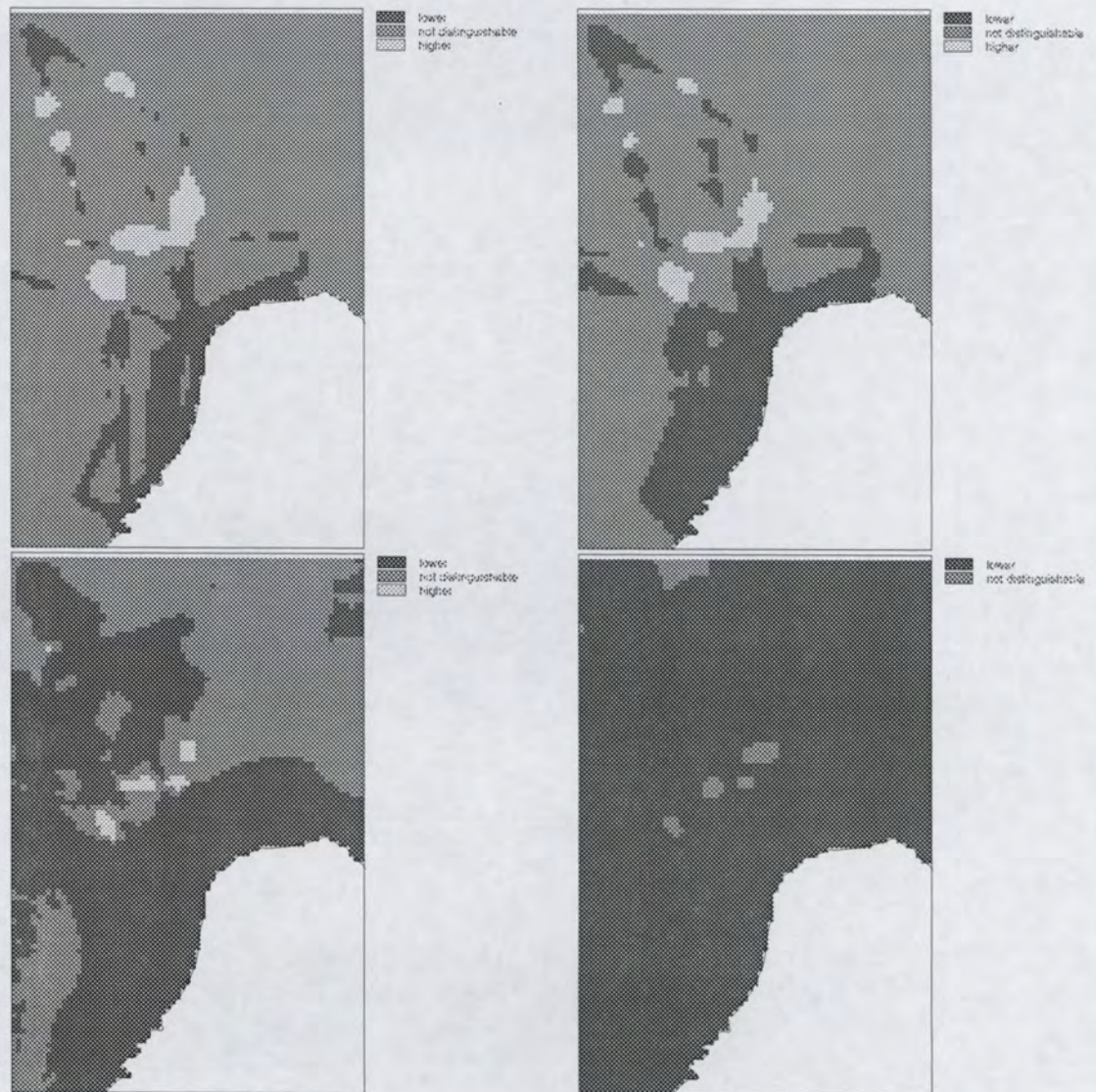


Figure 5.12: Map of *Fulmarus glacialis*, period 5, GEE. 95% Prediction intervals for 5 km \times 5 km block mean densities related to the density levels 1, 2, 5 and 10

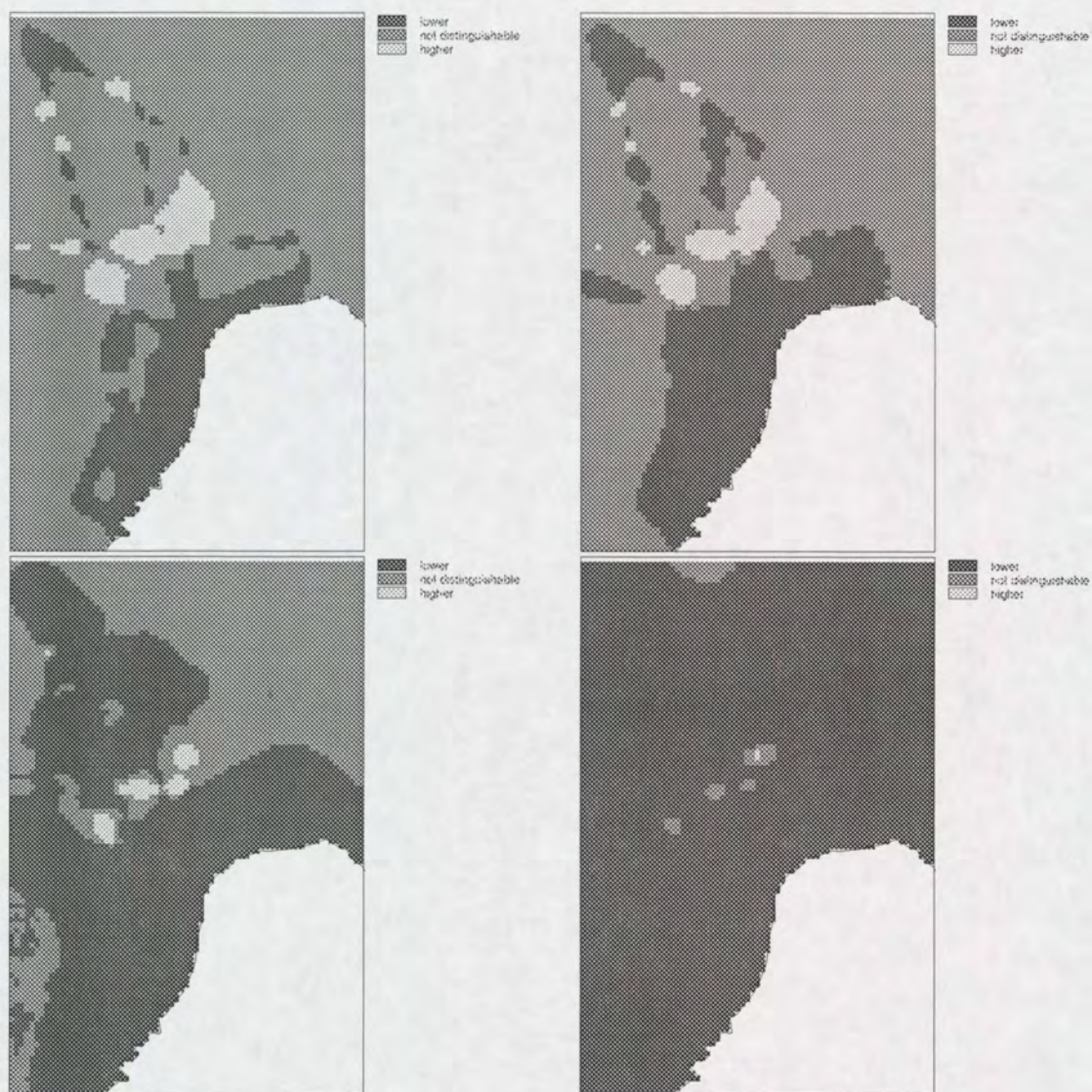


Figure 5.13: Map of *Fulmarus glacialis*, period 5, GLM. 95% Prediction intervals for 5 km \times 5 km block mean densities related to the density levels 1, 2, 5 and 10

6 Discussion and conclusions

The application of geostatistical methods in ecological research is in an early stage of development: as will be evident after reading the section on generalized linear models (2.3.1), many applications of spatial modelling in ecology are based on the unjustified assumption of spatially independent observations, and use unconditional prediction (sec. 2.2.4).

The combined approach of spatial modelling (geostatistics) and classical ecological modelling (using generalized linear or additive models) seems a promising step forward towards spatial statistical modelling, as applied to ecological data.

In the approach we followed, which is mostly based on the work of Gotway and Stroup (1997), a number of ad hoc decisions were taken, both with respect to the data handling as to the mathematical modelling. Unverified assumptions that were made with respect to the data are:

- differentiation in observation surfaces was ignored
- measurement error was assumed to be zero (in case of zero nugget variances), or assumed as having a zero-mean (in case of a positive nugget variance)
- no outlier analysis was performed

The following simplifications were made with respect to the statistical modelling:

- the family of trend models covered by the model selection procedure may have been too limited
- the model selection and estimation of trend parameter was based on the assumption of independent observations
- the calculation of prediction intervals (Eq. 3.4) was based on an oversimplistic model for the prediction error

- the spatial correlation of residuals was calculated as if they were *true* residuals, whereas in fact they were simple model residuals (working with predicted residuals, as obtained by leaving the i -th observation out when estimating the trend (Christensen, 1996, p. 343) may resolve this.)
- predictions are not guaranteed to be positive, because no constraint was set to the range for residual prediction
- the model for over-dispersion was fairly simple, and was not verified using sample data

We expect that the more elaborate estimating procedure ("method B"), based on generalized estimating equations or on generalized linear mixed models, combined with a suitable model for the prediction error will yield more realistic prediction intervals.

It should be noted here that the analysis of spatial correlation based on Pearson residuals of Eq. 3.1 may lead to a highly extreme value for a Pearson residual when positive densities occur at very unexpected circumstances. Such outlying residuals may completely dominate the variogram and may have to be removed before a sensible analysis of spatial correlation can take place. In this study, this step was omitted. An example of such an extreme may be present for *Sterna sandvicensis*, period 5 (Aug/Sep; section 5.2, page 45). One outlying observation (an occurrence of one single *Sterna sandvicensis* on open sea, see the corresponding figure on page 82) is probably the cause for the high, unstructured variation present in the variogram for this period (Fig. 5.7, page 45).

The trend models used here are admittedly very simple: given the two variables, depth of sea surface and distance to the (Dutch) coast, polynomials were fitted, and a first order interaction was allowed when main effects were present. A stepwise variable selection procedure was used to select one single model. To obtain more realistic results, especially in areas beyond the correlation distance from the measurements, models should be chosen that better reflect the current knowledge of the ecology of the species considered. First, the question *why* these birds prefer certain types of areas over others in the time period considered should be investigated. The criteria that make an area more or less preferred should be available as independent, known maps (like distance or depth), and then they can be used for predicting the trend. The present approach, based on stepwise variable selection, could be enriched using ecological knowledge to exclude those models that make no sense from an ecological perspective.

No quantitative comparisons of the current approach to the alternatives of (i) using a trend model only or (ii) using a strictly geostatistical approach (without trend function) only, were made. Although simple reasoning (given the maps of depth and distance to the coast, the sample maps and the result maps in chapter 5 and the maps in Appendix C) may be sufficient to support the choice for the combined approach, further research to support this choice quantitatively is needed. A rigorous validation of the prediction methods used (methods A and B) and of the nominal coverages of the confidence intervals obtained by them was not carried out. This should be done for a well-founded choice for (one of) our proposed methods. For point-wise prediction cross validation or bootstrap techniques could be used for this; for validating the interval predictions of block mean densities an approach based on simulated fields of bird densities may be necessary, because block mean data are not available. Both sampling errors for correlation parameters as well as trend parameters should be addressed in such a procedure.

For *Uria aalge*/*Alca torda*, an alternative, composite model was considered for the trend, consisting of a conditional Poisson distribution for the non-zero counts, and a binomial model for the presence (non-zero) or absence (zero counts) of birds (fairly similar to Welsh et al., 1996). The results are shown in appendix C. No prediction variances were obtained for this model.

For very rare species (*Sterna sandvicensis* in periods 1, 2 and 6; *Melanitta nigra* in period 3) the statistical methods used here are not suitable for spatial prediction. The techniques would be sufficient when all model inference (model structure and parameter estimates for trend and spatial correlation) were given, but exactly this information cannot be inferred from observed data. The current approach for observing (sampling) these species does not provide enough information for this. Estimating global means in the case of low observed bird densities (e.g., *Sterna sandvicensis*) may only be possible when information, external to the observed data, is added to facilitate the inference. For *Melanitta nigra* however, where the sparse positive densities show huge clusters, even estimating the global mean will remain a problem. Other, more species-oriented, monitoring strategies (combined with other estimating procedures) are required to enable accurate estimation of abundance for birds that tend to occur in compact, but huge clusters.

7 References

- Albert, P.S., McShane, L.S. (1995) A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. *Biometrics* 51, pp. 627-638.
- Augustin, N.H., Muggleston, M.A. and Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* 33: 339-347
- Augustin, N.H., Muggleston, M.A. and Buckland, S.T. (1998) The role of simulation in modeling spatially correlated data. *Environmetrics* 9, 175-196.
- Austin, M.P., Nicholls, A.O., Doherty, M.D. and Meyers, J.A. (1994) Determining species response functions to an environmental gradient by means of a beta-function. *Journal of Vegetation Science* 5: 215-228
- Austin, M.P., Nicholls, A.O. and Margules, C.R. (1990) Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. *Ecological Monographs* 60(2): 161-177
- Baptist, H.J.M., Wolf, P.A. (1993) Atlas van de vogels van het Nederlands Continentaal Plat. Rijkswaterstaat, Dienst Getijdewateren. Rapport DGW-98.013
- Bio, A.M.F., Alkemade, R. and Barendregt, A. (1998) Determining alternative models for vegetation response analysis: a non-parametric approach. *Journal of Vegetation Science* 9: 5-16
- Birks, H.J.B. (1996) Statistical approaches to interpreting diversity patterns in the Norwegian mountain flora. *Ecography* 19(3): 332-340
- Buckland, S.T. and Elston, D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology* 30: 478-495

- Carey, V.J. (1998) YAGS – yet another GEE solver. (Documentation and source code at <http://biosun1.harvard.edu/~carey/index.ssoft.html>)
- Chambers, J.M. and Hastie, T.J. (Eds.) (1993) Statistical models in S. Chapman and Hall, London.
- Christensen, R., 1991. Linear models for Multivariate, Time Series and Spatial Data. Springer Verlag, New York. 317 pp.
- Christensen, R., 1993. Quadratic Covariance Estimation and Equivalence of Predictions. *Mathematical Geology* 25 (5), pp. 541-558.
- Christensen, R., 1996. Plane Answers to Complex Questions: the Theory of Linear Models. Second Edition. Springer Verlag, New York. 452 pp.
- Cressie, N.A.C. (1993) Statistics for Spatial Data Revised Edition. Wiley, New York.
- Diggle, P.J., Liang, K-Y., Zeger, S.L. (1994) Analysis of Longitudinal Data. Oxford University Press, Oxford.
- Gotway, C.A., Stroup, W.W. (1997) A Generalized Linear Model Approach to Spatial Data Analysis and Prediction. *Journal of Agricultural, Biological and Environmental Statistics* 2(2), pp. 157-178.
- Gruijter, J.J. de, C.J.F. ter Braak, 1990. Model-Free Estimation from Spatial Samples: A Reappraisal of Classical Sampling Theory. *Mathematical Geology* 4 (22), pp. 407-415.
- Hansen, M.H., W.G. Madow, B.J. Tepping, 1983. An Evaluation of Model-Dependent Probability-Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, 78, pp. 776-793.
- Hastie, T.J., Tibshirani, R.J. (1990) Generalized Additive Models. Chapman and Hall, London.
- Huisman, J., Olff, H. and Fresco, L.F.M. (1993) A hierarchical set of models for species response analysis. *Journal of Vegetation Science* 4: 37-46
- James, F.C. and McCulloch, C.E. (1990) Multivariate analysis in ecology and systematics: Panacea or Pandora's box? *Ann. Rev. Ecol. Syst.* 21: 129-166
- Journel, A.G., Ch.J. Huijbregts, 1978. Mining Geostatistics. Academic Press, London, 600 pp.

- Journal, A.G., M. Rossi, 1989. When do we need a trend model in kriging? *Mathematical Geology* 21 (7), pp. 715-739.
- Journal, A.G., 1992. Geostatistics: Roadblocks and Challenges. in: A Soares (ed.), *Geostatistics Troia '92 Volume 1*, Kluwer, Dordrecht, pp. 213-224.
- Kitanidis, Peter K., 1986. Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis. *Water Resources Research*, 22 (4), pp. 499-507.
- Kitanidis, Peter K., 1991. Orthonormal Residuals in Geostatistics: Model Criticism and Parameter Selection. *Mathematical Geology* 23 (5), pp. 741-758.
- Kitanidis, P.K., 1993. Generalized Covariance Functions in Estimation. *Mathematical Geology* 25 (5), pp. 525-540.
- Liang, K-Y., Zeger, S.L. (1986) Longitudinal Data Analysis using Generalized Linear Models. *Biometrika* 73(1), pp. 13-22.
- McCullagh, P., Nelder, J.A. (1989) *Generalized Linear Models Second Edition*. Chapman and Hall, London.
- Nicholls, A.O. (1989) How to make biological surveys go further with Generalized Linear Models. *Biological Conservation* 48: 51-75
- Pebesma, E.J., De Kwaadsteniet, J.G. (1997) Mapping Groundwater Quality in the Netherlands. *Journal of Hydrology* 200, pp. 364-386.
- Pebesma, E.J., Wesseling, C.G. (1998) Gstat, a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences* 28(1), pp. 17-31. <http://www.geog.uu.nl/gstat/>
- Trexler, J.C. and Travis, J. (1993) Nontraditional regression analyses. *Ecology* 74(6): 1629-1637
- Welsh, A.H., Cunningham, R.B., Donnelly, C.F., Lindenmayer, D.B. (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological modelling* 88, pp. 297-308.
- Venables, W.N. and Ripley, B.D. (1994) *Modern applied statistics with S-Plus*. Springer-Verlag, New York

- Weseloh, R.M. (1996) Developing and validating a model for predicting gypsy moth (Lepidoptera: Lymantriidae) defoliation in Connecticut. *Journal of Economic Entomology* 89(6): 1546-1555
- Witte, R.H. (1995a) Noordzee tellingen februari/maart 1995. Rijkswaterstaat, Rijksinstituut voor Kust en Zee (RIKZ). Rapportnr NZ9412.
- Witte, R.H. (1995b) Noordzee tellingen april/mei 1995. Rijkswaterstaat, Rijksinstituut voor Kust en Zee (RIKZ). Rapportnr NZ9504.
- Witte, R.H. (1995c) Noordzee tellingen juni/juli 1995. Rijkswaterstaat, Rijksinstituut voor Kust en Zee (RIKZ). Rapportnr NZ9506.
- Witte, R.H. (1995d) Noordzee tellingen augustus/september 1995. Rijkswaterstaat, Rijksinstituut voor Kust en Zee (RIKZ). Rapportnr NZ9508.
- Witte, R.H. (1995e) Noordzee tellingen oktober/november 1995. Rijkswaterstaat, Rijksinstituut voor Kust en Zee (RIKZ).
- Witte, R.H. (1995f) Noordzee tellingen april/mei 1995. Rijkswaterstaat, Rijksinstituut voor Kust en Zee (RIKZ). Rapportnr NZ9504.
- Yee, T.W. and Mitchell, N.D. (1991) Generalized additive models in plant ecology. *Journal of Vegetation Science* 2: 587-602
- Zeger, S.L., Liang, K-Y. (1986) Longitudinal data analysis for discrete and Continuous Outcomes. *Biometrics* 42, pp. 121-130.

A GLM output

A.1 *Uria aalge/Alca torda*, period 2

| | Resid. Df | Resid. Dev | Test | Df | Deviance | Pr(Chi) |
|---|-----------|------------|---------------|----|----------|----------|
| 1 | 363 | 614.6 | | | | |
| 2 | 362 | 561.9 | + dis | 1 | 52.62 | 0.000000 |
| 3 | 360 | 537.1 | +dep+I(dep^2) | 2 | 24.80 | 0.000004 |
| 4 | 359 | 522.3 | +dis:dep | 1 | 14.86 | 0.000116 |
| 5 | 358 | 503.3 | +I(dis^2) | 1 | 19.01 | 0.000013 |
| 6 | 357 | 454.9 | +I(dep^3) | 1 | 48.40 | 0.000000 |
| 7 | 356 | 447.1 | +I(dis^3) | 1 | 7.78 | 0.005282 |
| 8 | 355 | 430.1 | +I(dep^4) | 1 | 16.97 | 0.000038 |
| 9 | 354 | 420.2 | +I(dep^5) | 1 | 9.92 | 0.001632 |

```
summary(az2.glm)
```

```
-----
```

```
formula = AZ2.km2 ~ dis + dep + I(dep^2) + I(dis^2) + I(dep^3) + I(dis^3) +
  I(dep^4) + I(dep^5) + dis:dep, family = quasi(link = "log", var = "mu"),
  maxit = 30)
```

Coefficients:

| | Value | Std. Error | t value |
|-------------|-------------|------------|---------|
| (Intercept) | -4.525e+001 | 1.523e+001 | -2.972 |
| dis | -1.657e-005 | 1.617e-005 | -1.025 |
| dep | 5.624e+000 | 2.035e+000 | 2.764 |
| I(dep^2) | -2.607e-001 | 1.023e-001 | -2.548 |
| I(dis^2) | 2.871e-010 | 1.242e-010 | 2.312 |
| I(dep^3) | 5.634e-003 | 2.419e-003 | 2.329 |
| I(dis^3) | -5.522e-016 | 2.876e-016 | -1.920 |
| I(dep^4) | -5.674e-005 | 2.699e-005 | -2.102 |


```

I(dep^5) 2.172e-007 1.139e-007 1.907
dis:dep -5.993e-007 2.640e-007 -2.270

```

(Dispersion Parameter for Quasi-likelihood family taken to be 2.104)

Null Deviance: 614.6 on 363 degrees of freedom

Residual Deviance: 420.2 on 354 degrees of freedom

```

Dispersion parameter: 2.104
Explained Deviance: 32%

```

A.2 *Fulmarus glacialis*, period 5

| | Resid. Df | Resid. Dev | Test | Df | Deviance | Pr(Chi) |
|---|-----------|------------|----------|----|----------|------------|
| 1 | 351 | 1354 | | | | |
| 2 | 350 | 1166 | +dis | 1 | 188.8 | 0.000e+000 |
| 3 | 349 | 1132 | +dep | 1 | 33.7 | 6.424e-009 |
| 4 | 348 | 526 | +dis:dep | 1 | 605.7 | 0.000e+000 |

```
summary(NS5.glm)
```

```

-----
formula = NS5.km2 ~ dis + dep + dis:dep,
family = quasi(link = "log", var = "mu"), maxit = 30)

```

Coefficients:

| | Value | Std. Error | t value |
|-------------|-------------|------------|---------|
| (Intercept) | -7.391e+000 | 6.911e-001 | -10.69 |
| dis | 3.298e-005 | 3.276e-006 | 10.07 |
| dep | 2.086e-001 | 1.707e-002 | 12.22 |
| dis:dep | -8.118e-007 | 7.924e-008 | -10.24 |

(Dispersion Parameter for Quasi-likelihood family taken to be 2.091)

Null Deviance: 1354 on 351 degrees of freedom

Residual Deviance: 526.3 on 348 degrees of freedom

Dispersion parameter: 2.091

Explained Deviance: 61%

A.3 *Sterna sandvicensis*, period 3

| | Resid.Df | Resid. Dev | Test | Df | Deviance | Pr(Chi) |
|---|----------|------------|------------|----|----------|---------|
| 1 | 461 | 226.9 | | | | |
| 2 | 460 | 204.9 | +ldis | 1 | 22.03 | 0.00000 |
| 3 | 459 | 187.7 | +I(ldis^2) | 1 | 17.21 | 0.00003 |

summary(GS3log.glm)

```
-----
formula = GS3.km2 ~ ldis + I(ldis^2),
family = quasi(link = "log", var = "mu"), maxit = 30)
```

Coefficients:

| | Value | Std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | -134.814 | 50.504 | -2.669 |
| ldis | 66.342 | 25.161 | 2.637 |
| I(ldis^2) | -8.257 | 3.126 | -2.642 |

(Dispersion Parameter for Quasi-likelihood family taken to be 1.269)

Null Deviance: 226.9 on 461 degrees of freedom

Residual Deviance: 187.7 on 459 degrees of freedom

Dispersion parameter: 1.269

Explained Deviance: 17%

A.4 *Sterna sandvicensis*, period 4

| | Resid.Df | Resid. Dev | Test | Df | Deviance | Pr(Chi) |
|---|----------|------------|------|----|----------|---------|
| 1 | 400 | 382.0 | | | | |

| | | | | | | |
|---|-----|-------|------------|---|-------|--------|
| 2 | 399 | 264.2 | +ldis | 1 | 117.8 | 0.0000 |
| 3 | 398 | 251.4 | +I(ldis^2) | 1 | 12.8 | 0.0004 |

```
summary(GS4log.glm)
```

```
-----
formula = GS4.km2 ~ ldis + I(ldis^2),
      family = quasi(link = "log", var = "mu"), maxit = 30)
```

Coefficients:

| | Value | Std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | -128.562 | 63.300 | -2.031 |
| ldis | 66.693 | 32.149 | 2.075 |
| I(ldis^2) | -8.687 | 4.073 | -2.133 |

(Dispersion Parameter for Quasi-likelihood family taken to be 1.449)

Null Deviance: 382 on 400 degrees of freedom

Residual Deviance: 251.4 on 398 degrees of freedom

Dispersion parameter: 1.449

Explained Deviance: 34%

A.5 *Sterna sandvicensis*, period 5

| | Resid.Df | Resid. Dev | Test | Df | Deviance | Pr(Chi) |
|---|----------|------------|------------|----|----------|---------|
| 1 | 351 | 529.4 | | | | |
| 2 | 350 | 360.3 | +ldis | 1 | 169.1 | 0.0000 |
| 3 | 349 | 328.8 | +dep | 1 | 31.5 | 0.0000 |
| 4 | 348 | 321.5 | +I(ldis^2) | 1 | 7.3 | 0.0068 |

```
summary(GS5log.glm)
```

```
-----
formula = GS5.km2 ~ ldis + dep + I(ldis^2),
      family = quasi(link = "log", var = "mu"), maxit = 30)
```


A.6. URIA AALGE/ALCA TORDA, PERIOD 2, ALTERNATIVE MODEL67

Coefficients:

| | Value | Std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | -31.1205 | 24.56281 | -1.267 |
| ldis | 16.0438 | 11.90214 | 1.348 |
| dep | -0.1236 | 0.03766 | -3.281 |
| I(ldis^2) | -1.9817 | 1.43627 | -1.380 |

(Dispersion Parameter for Quasi-likelihood family taken to be 2.854)

Null Deviance: 529.4 on 351 degrees of freedom

Residual Deviance: 321.5 on 348 degrees of freedom

Dispersion parameter: 2.854

Explained Deviance: 39%

A.6 *Uria aalge/Alca torda*, period 2, alternative model

AZ21.bin (Logistic regression with 0 (absent) and 1 (present) data)

| | Resid.Df | Resid. Dev | Test | Df | Deviance | Pr(Chi) |
|---|----------|------------|--------------------|----|----------|---------|
| 1 | 363 | 367.6 | | | | |
| 2 | 362 | 339.4 | +dep | 1 | 28.16 | 0.00000 |
| 3 | 360 | 322.9 | +I(dep^2)+I(dep^3) | 2 | 16.51 | 0.00026 |
| 4 | 359 | 311.3 | +dis | 1 | 11.62 | 0.00065 |
| 5 | 358 | 304.3 | +I(dis^2) | 1 | 7.04 | 0.00798 |

summary(AZ21.bin)

```
-----  
formula = alk.zeekoet ~ dep + dep^2 + dep^3 + dis + dis^2,  
family = quasi(link = "logit", var = "mu(1-mu)"),  
data = rikz2.na1, maxit = 30)
```

Coefficients:

| Value | Std. Error | t value |
|-------|------------|---------|
|-------|------------|---------|


```

(Intercept) -9.614e+000 2.011e+000 -4.780
           dis -1.528e-005 8.634e-006 -1.770
           I(dis^2) 8.314e-011 2.917e-011 2.851
           dep 6.484e-001 1.632e-001 3.973
           I(dep^2) -1.386e-002 3.733e-003 -3.712
           I(dep^3) 8.311e-005 2.535e-005 3.278

```

(Dispersion Parameter for Quasi-likelihood family taken to be 0.8126)

Null Deviance: 367.6 on 363 degrees of freedom
 Residual Deviance: 304.3 on 358 degrees of freedom

Dispersion parameter: 0.8126
 Explained Deviance: 17%

=====

AZ21.poi ((quasi-)Poisson regression on postive counts only)

| | Resid.Df | Resid. Dev | Test | Df | Deviance | Pr(Chi) |
|---|----------|------------|-----------|----|----------|----------|
| 1 | 73 | 113.4 | | | | |
| 2 | 72 | 96.5 | +dep | 1 | 16.97 | 0.000038 |
| 3 | 71 | 89.5 | +dis | 1 | 6.97 | 0.008282 |
| 4 | 70 | 80.8 | +I(dep^2) | 1 | 8.68 | 0.003215 |

summary(AZ21.poi)

```

glm(AZ21.km ~ dis + dep + dep^2, family = quasi(link = "log", var = "mu"),
     data = rikz2.na2, maxit = 30)

```

Coefficients:

| | Value | Std. Error | t value |
|-------------|-------------|------------|---------|
| (Intercept) | 2.924e+000 | 5.599e-001 | 5.223 |
| dis | 2.303e-006 | 1.089e-006 | 2.115 |
| dep | -1.033e-001 | 2.749e-002 | -3.759 |
| I(dep^2) | 7.543e-004 | 2.815e-004 | 2.680 |

A.6. URIA AALGE/ALCA TORDA, *PERIOD 2, ALTERNATIVE MODEL*69

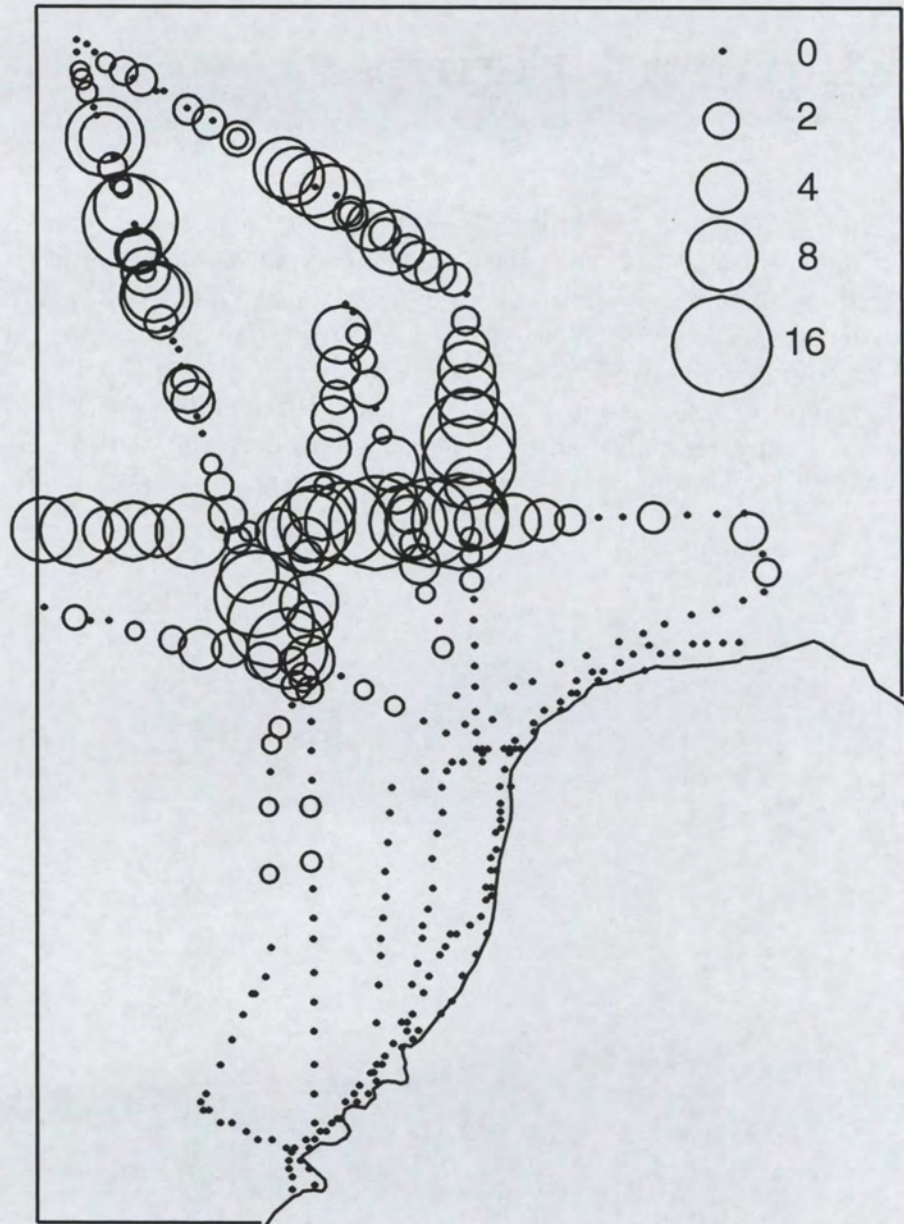
Null Deviance: 113.4 on 73 degrees of freedom
Residual Deviance: 80.8 on 70 degrees of freedom

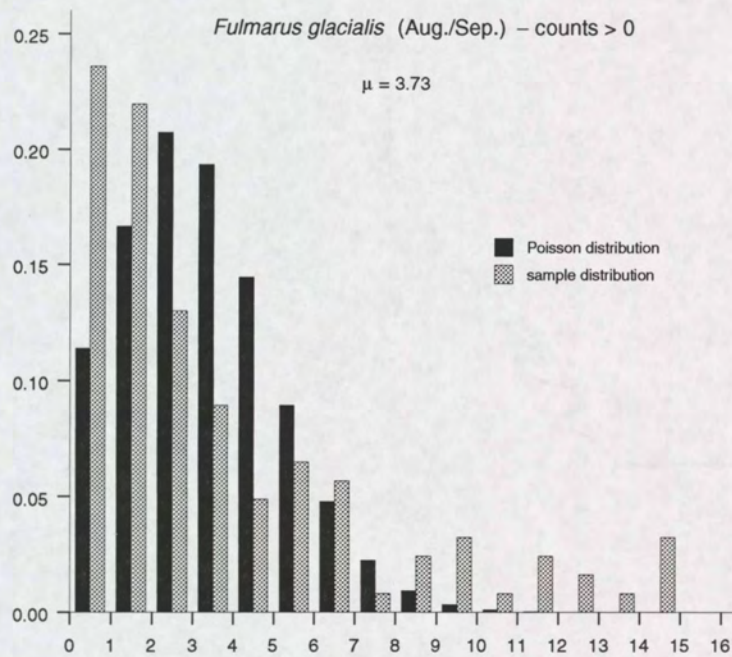
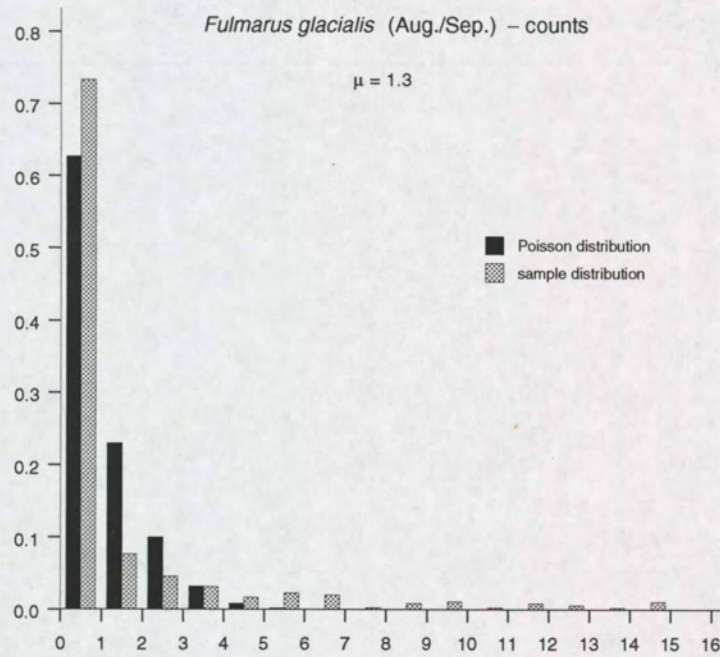
Dispersion parameter: 1.3
Explained Deviance: 36%

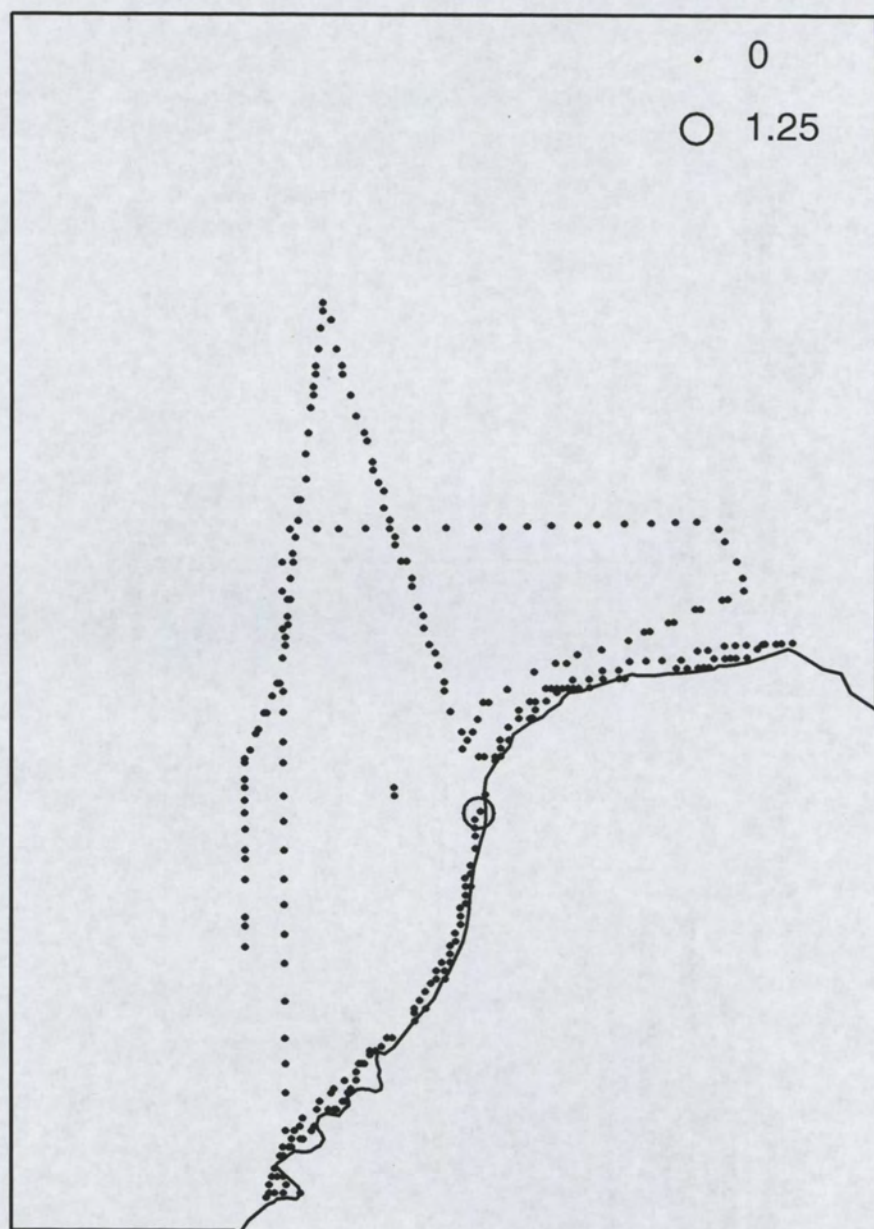
B Maps and Figures

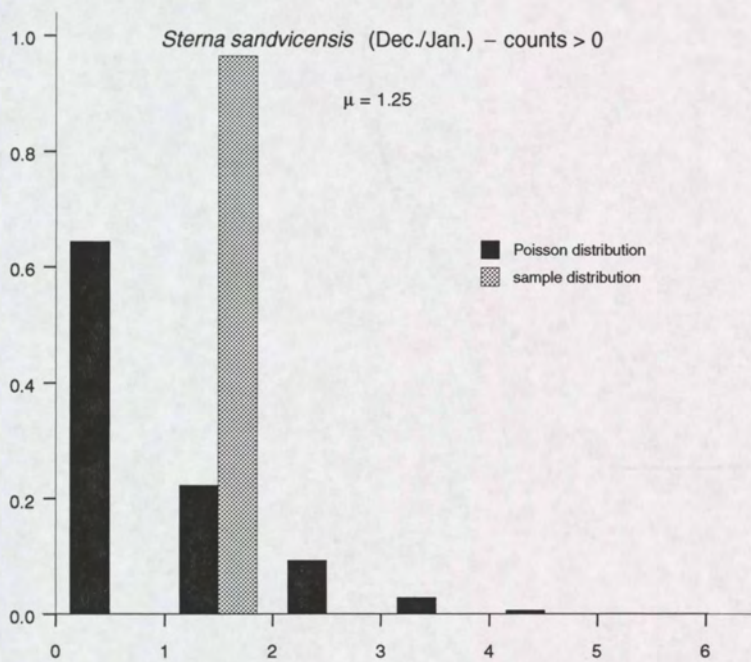
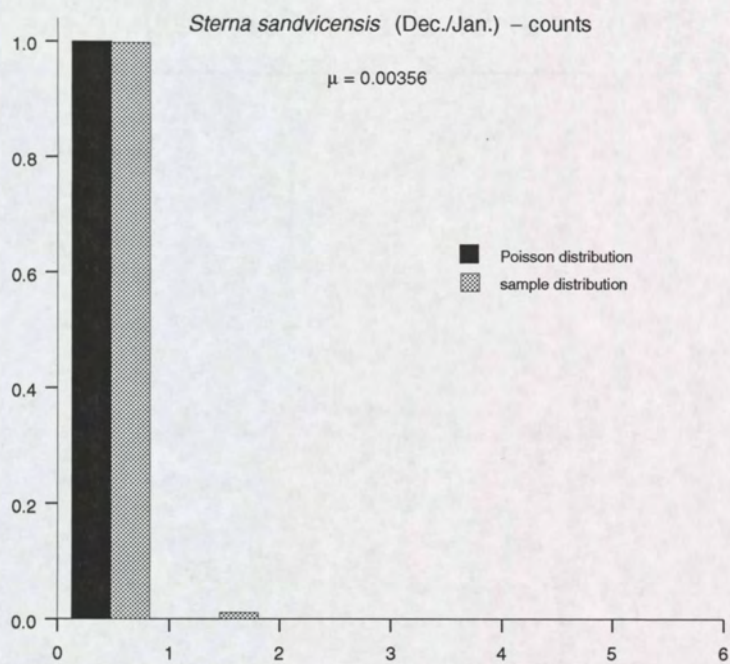
This appendix shows the maps with observed data for each species/period studied. Circle centres denote the centre of each observation area (a 150 m wide strip of approximately 6 km length, oriented in the flying direction), and circle size denotes the observed density (number of birds/km²; the *surface* of circles is taken proportional to the observed density).

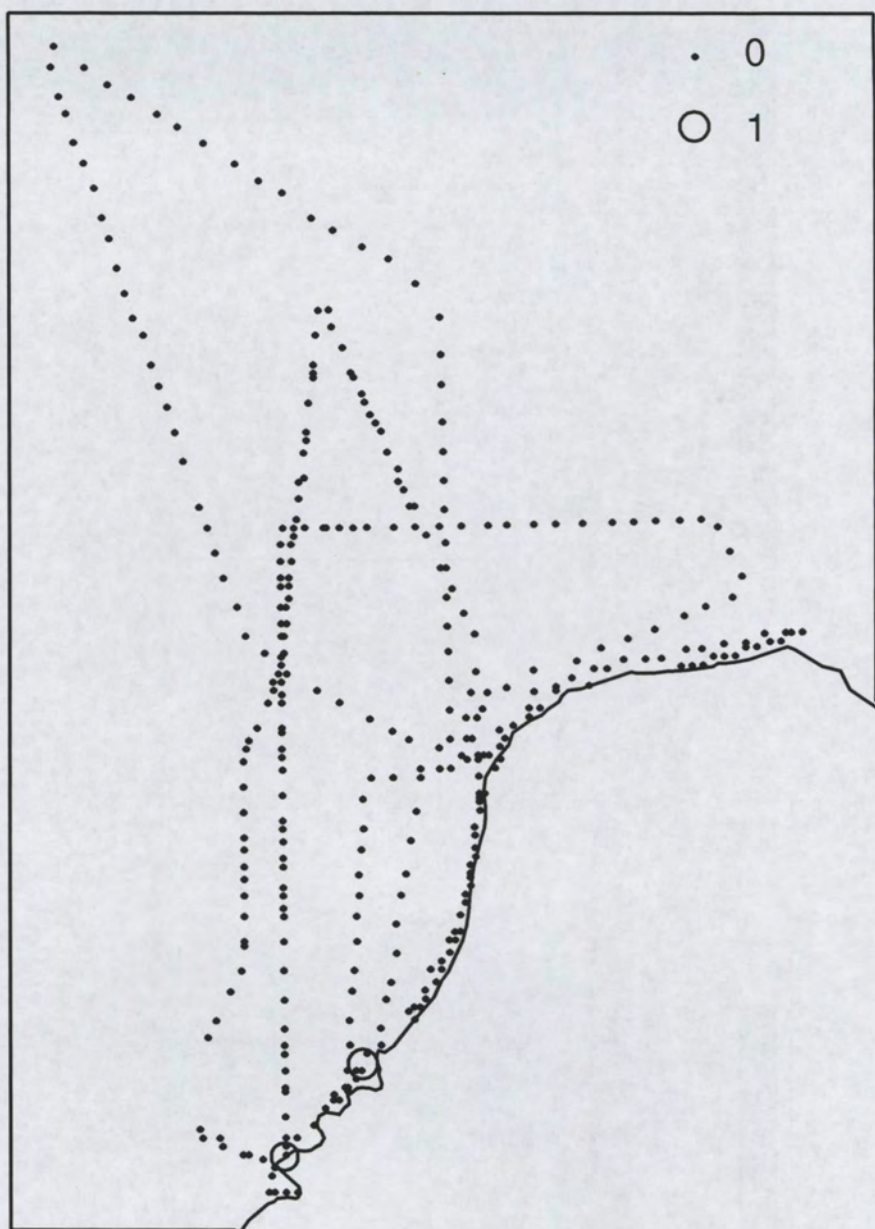
For each species/period studied, a histograms of the sample distribution of all observed densities and of all positive (non-zero) observed densities is shown along with a Poisson distribution having the same mean as the sample shown.

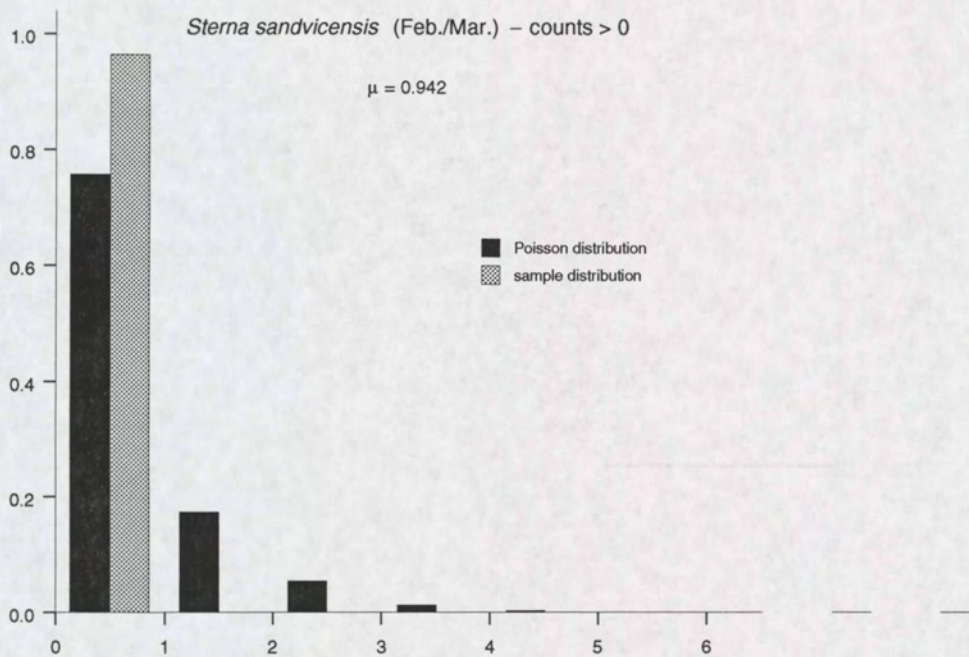
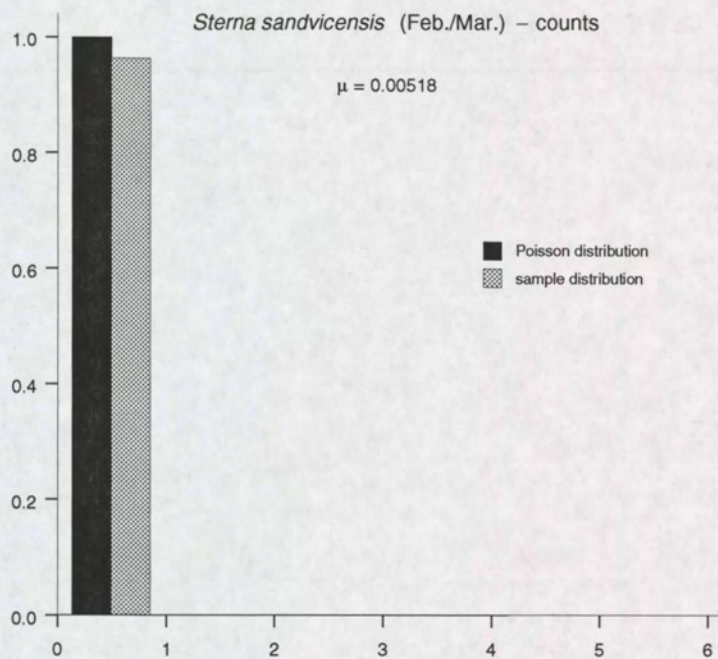
Fulmarus glacialis (Aug/Sep)

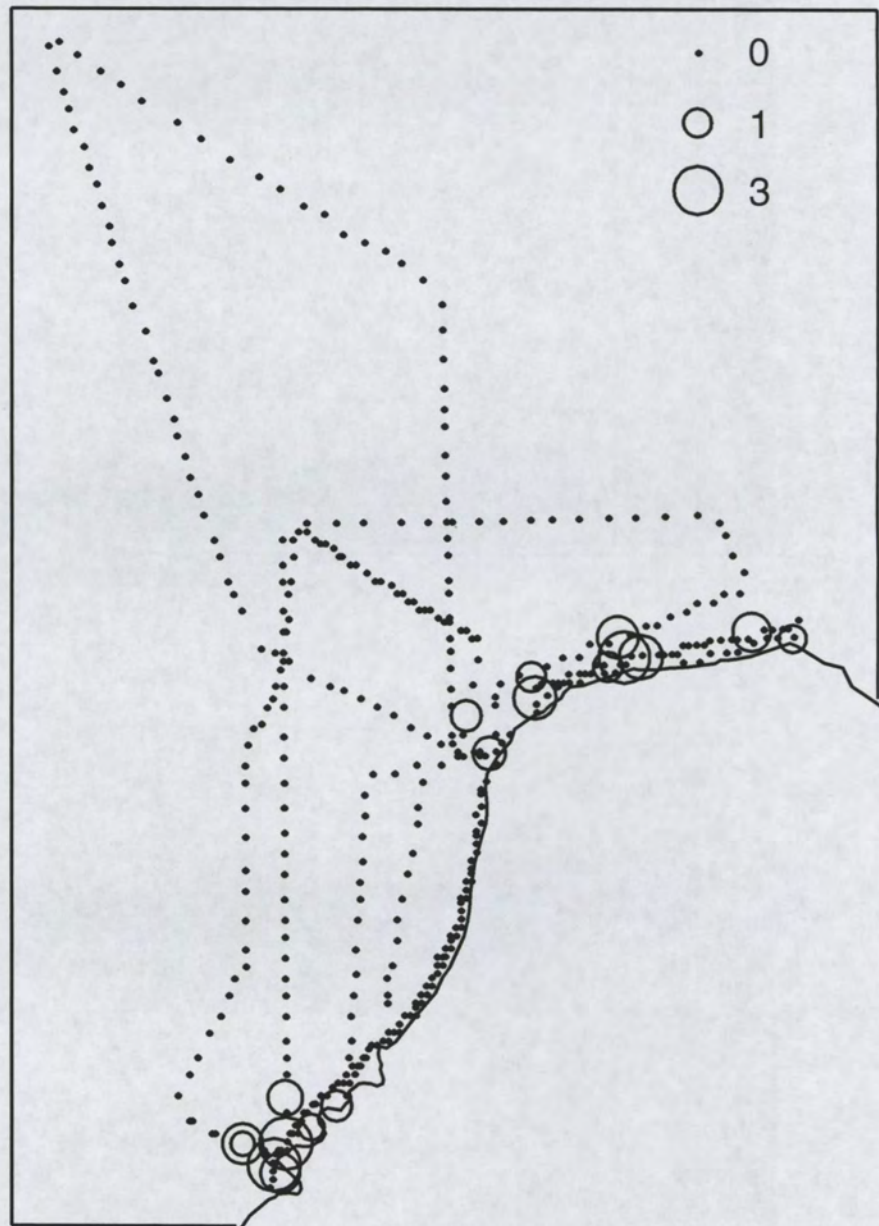


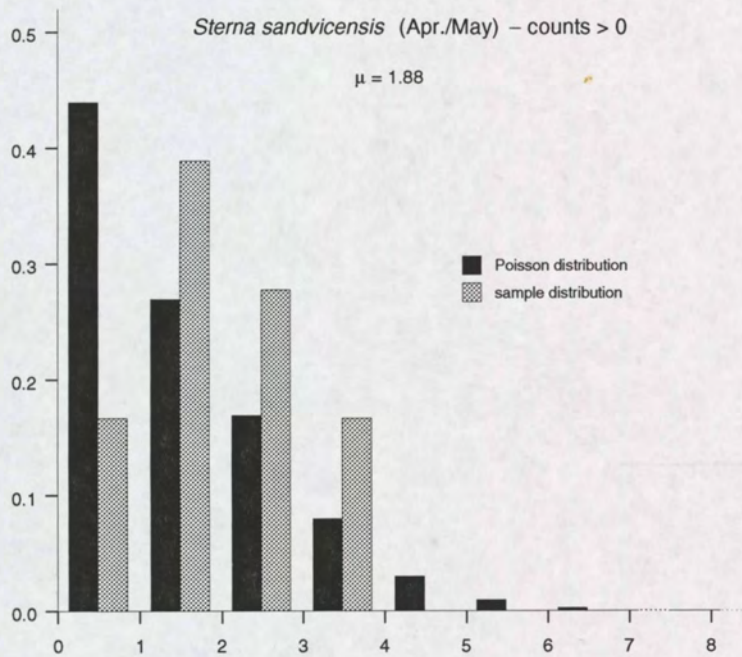
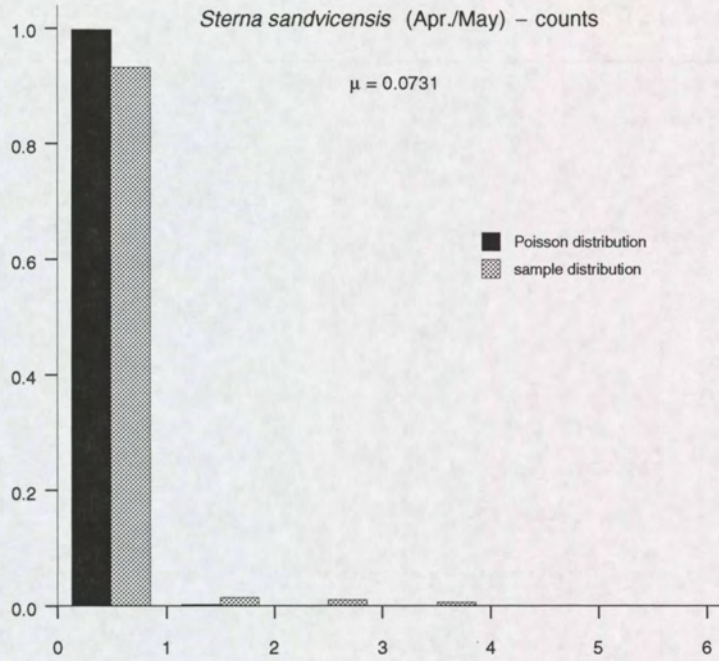
Sterna sandvicensis (Dec/Jan)

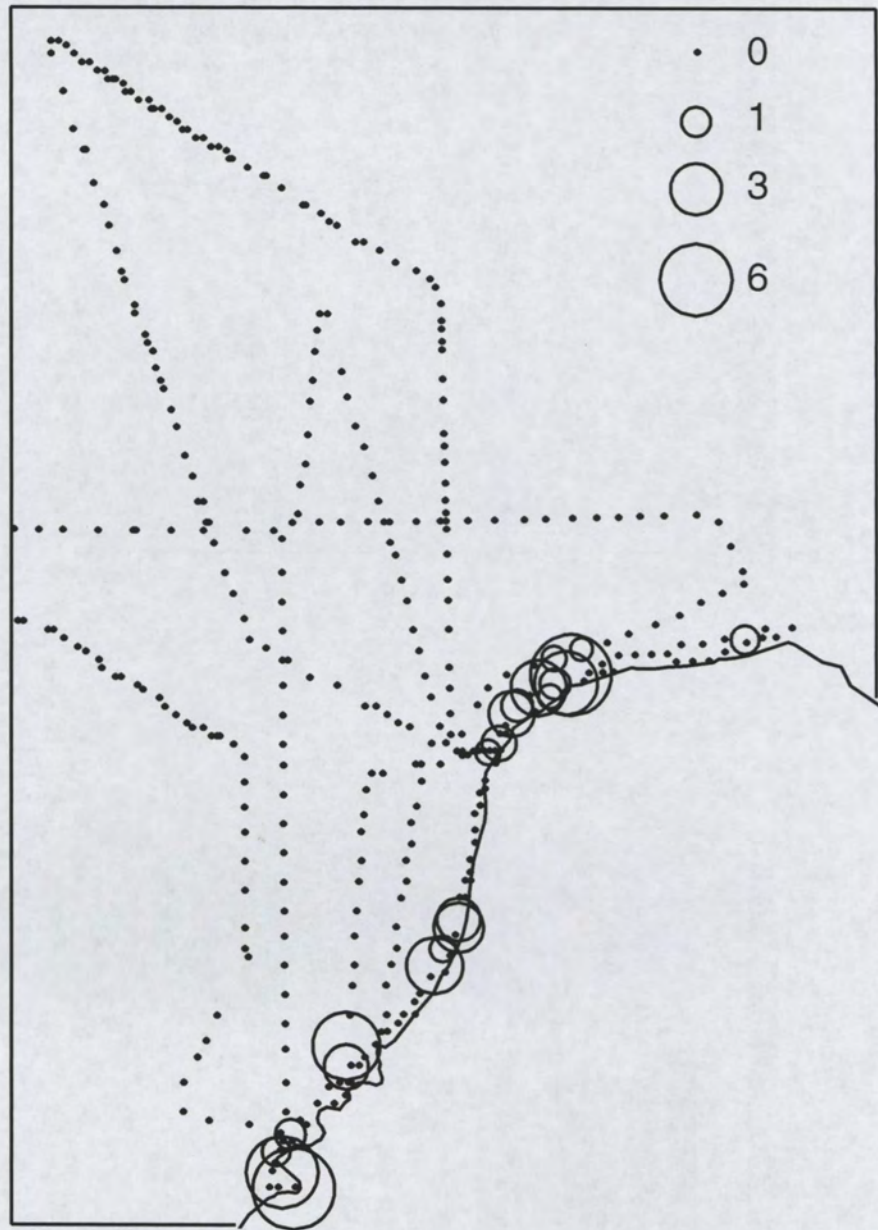


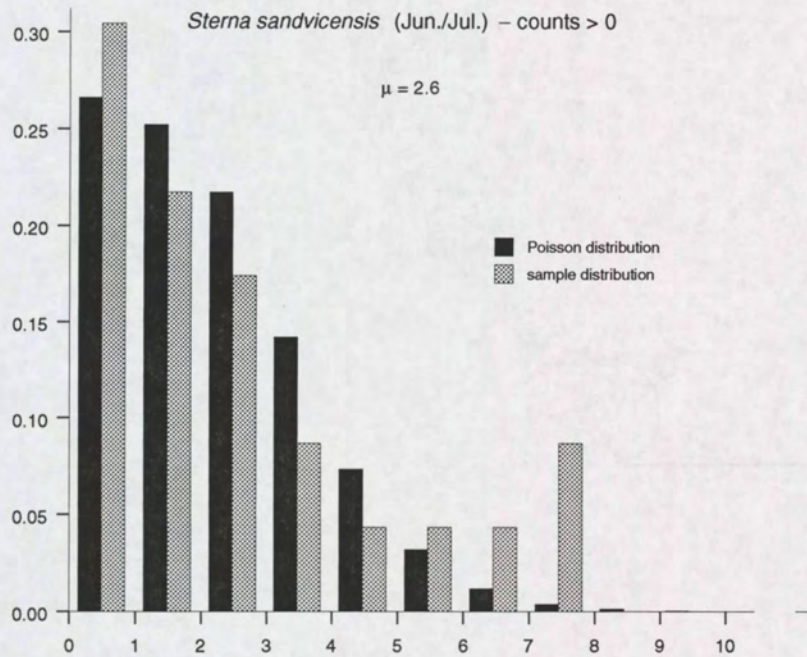
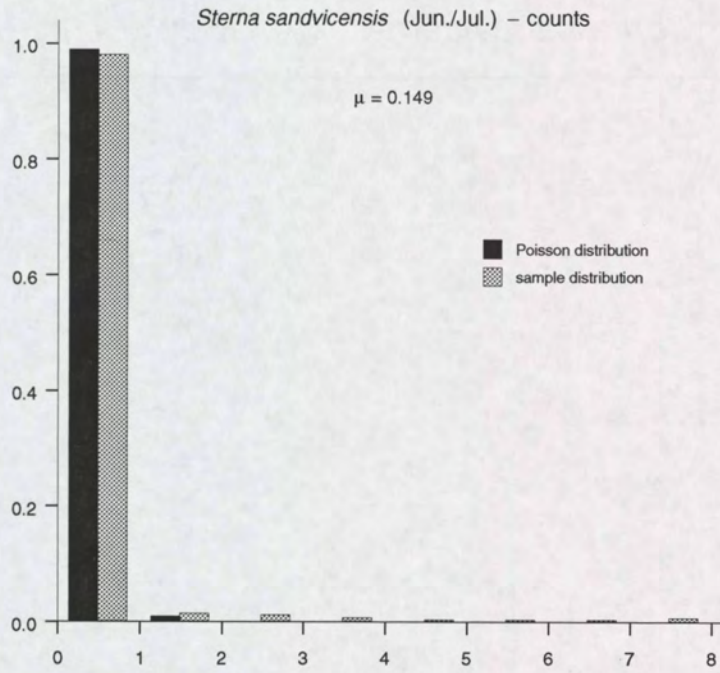
Sterna sandvicensis (Feb/Mar)

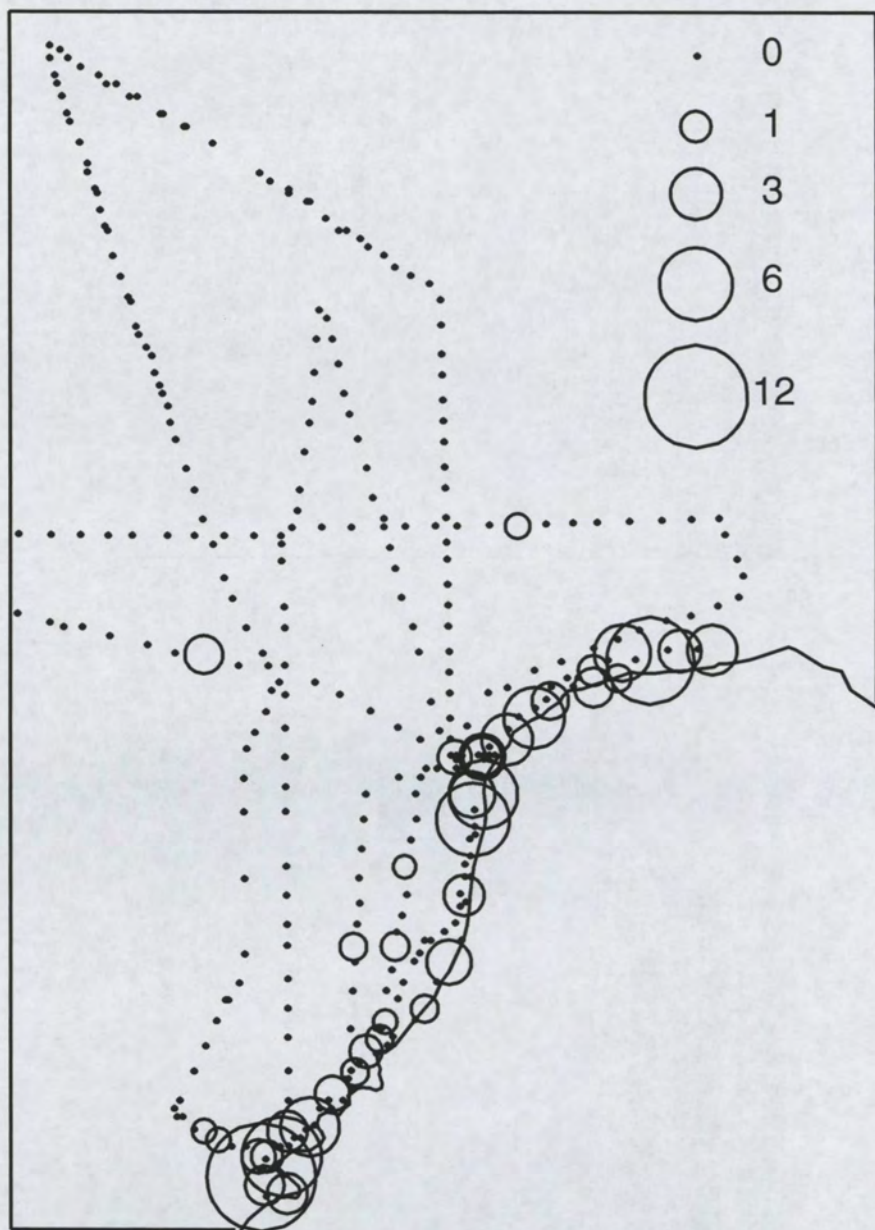


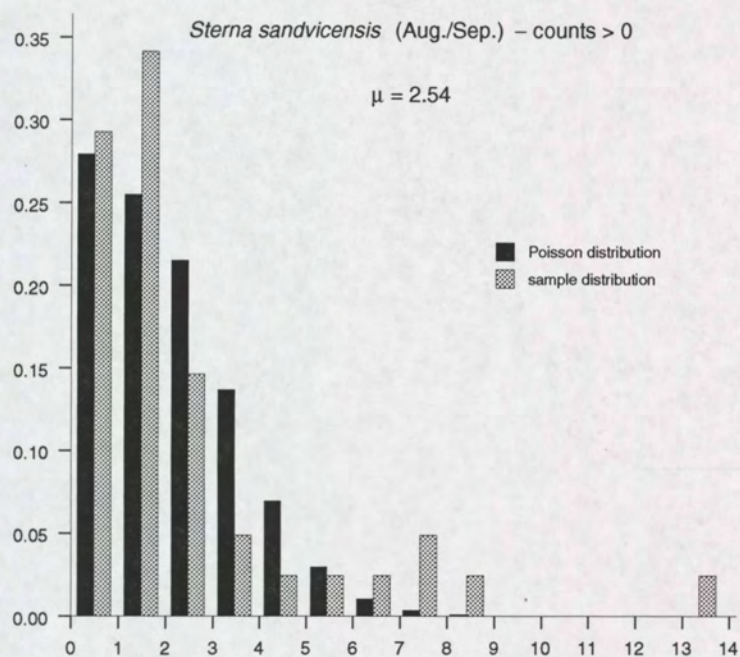
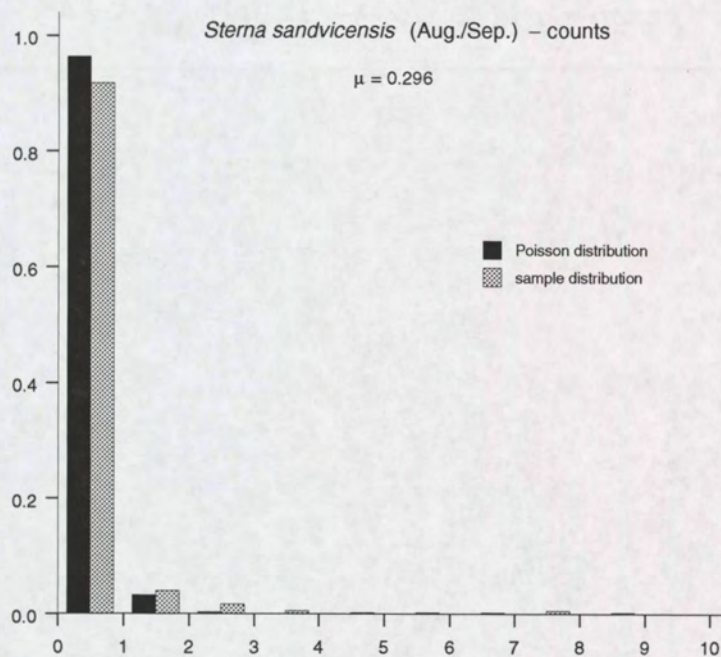
Sterna sandvicensis (Apr/May)

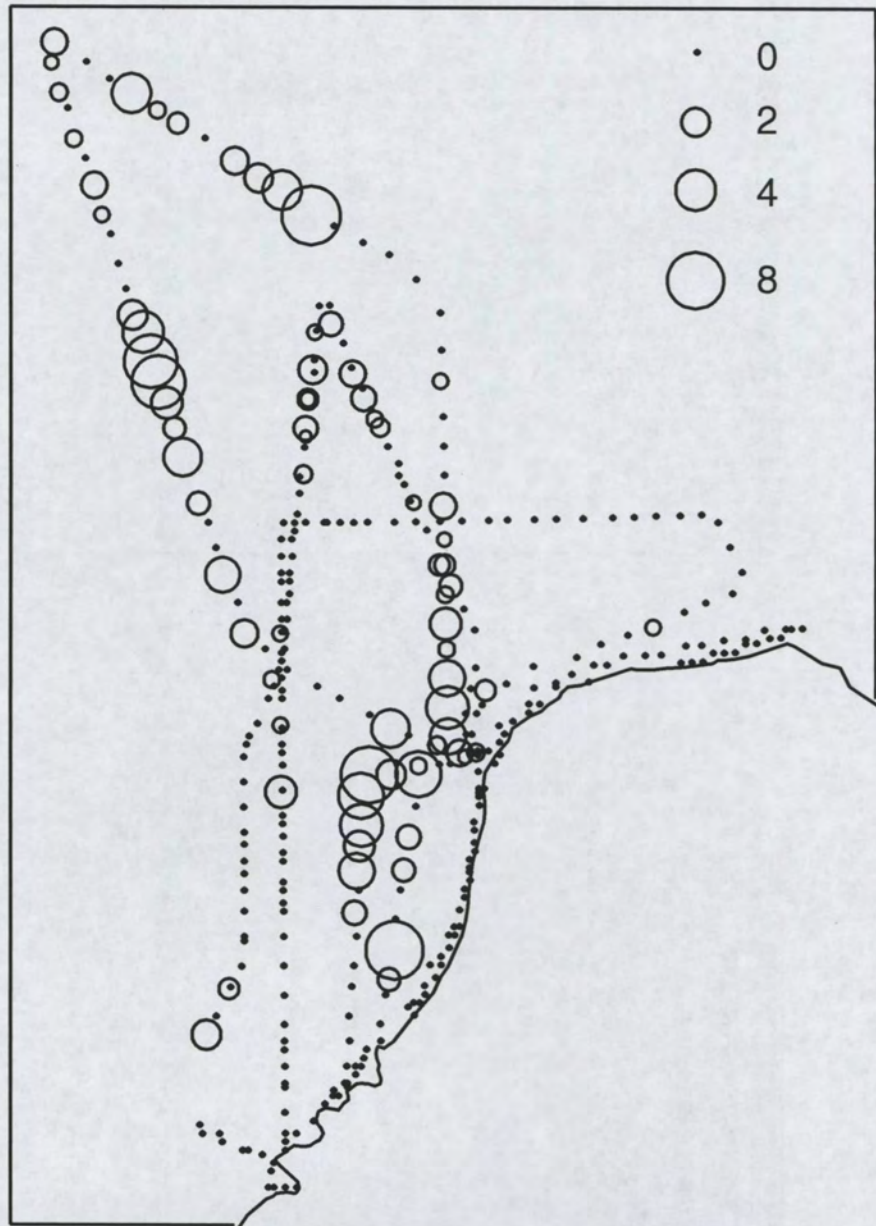


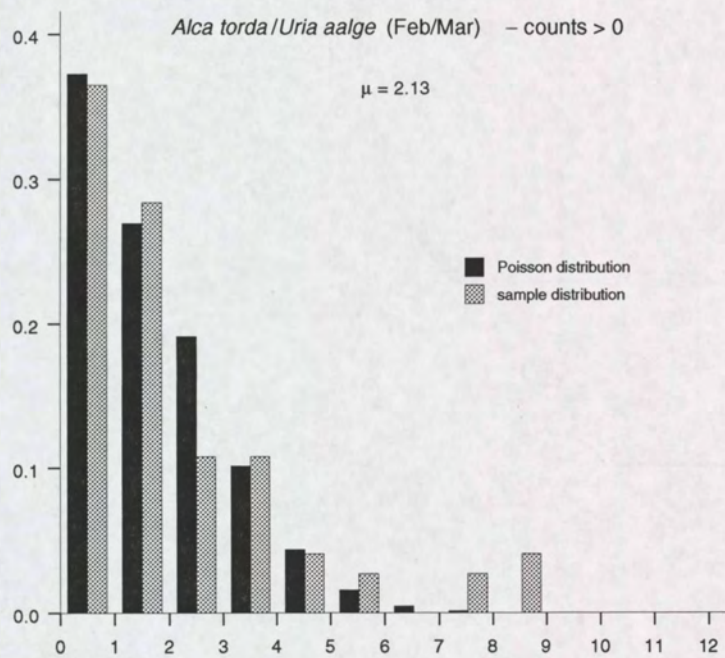
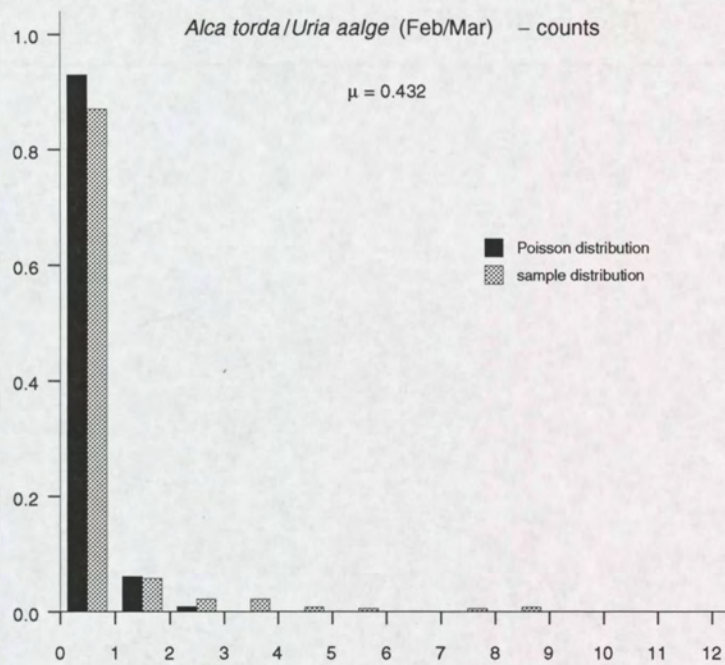
Sterna sandvicensis (Jun/Jul)

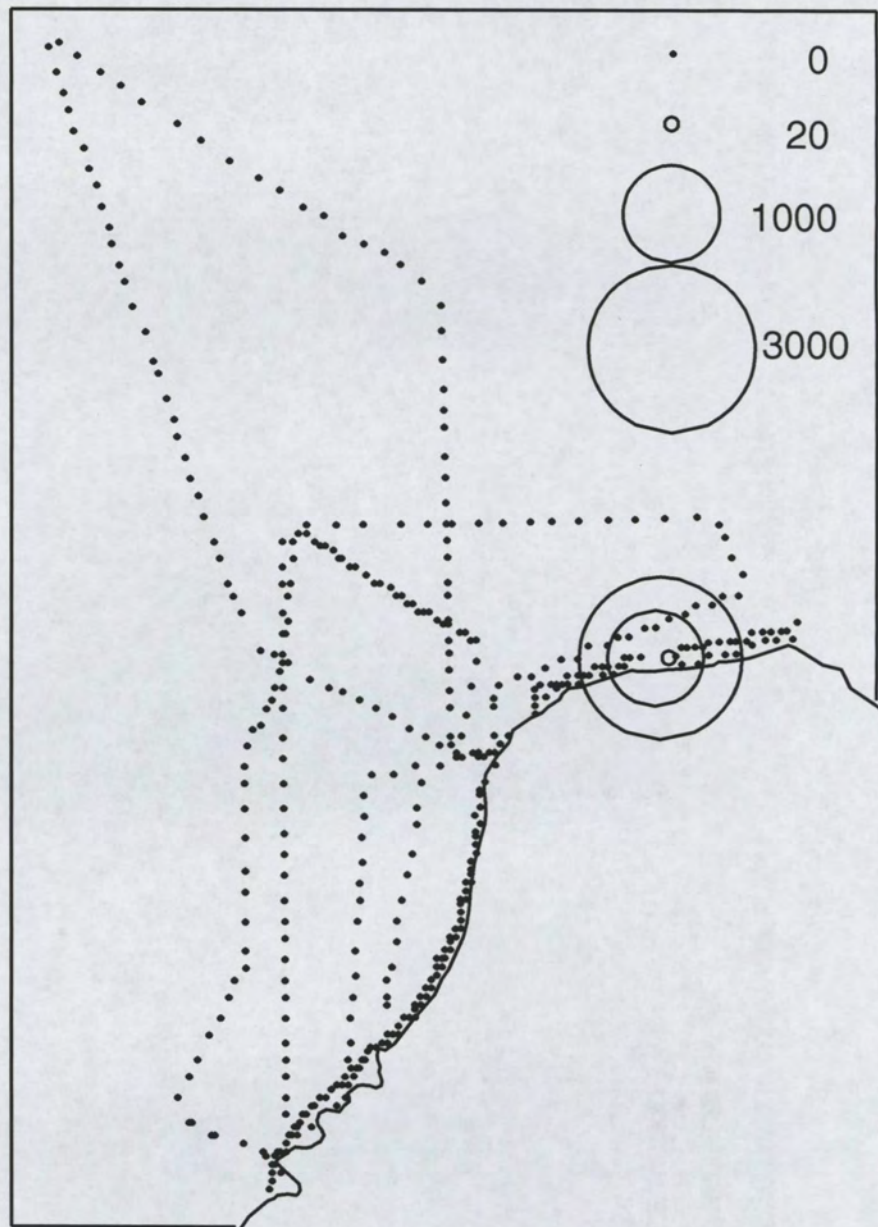


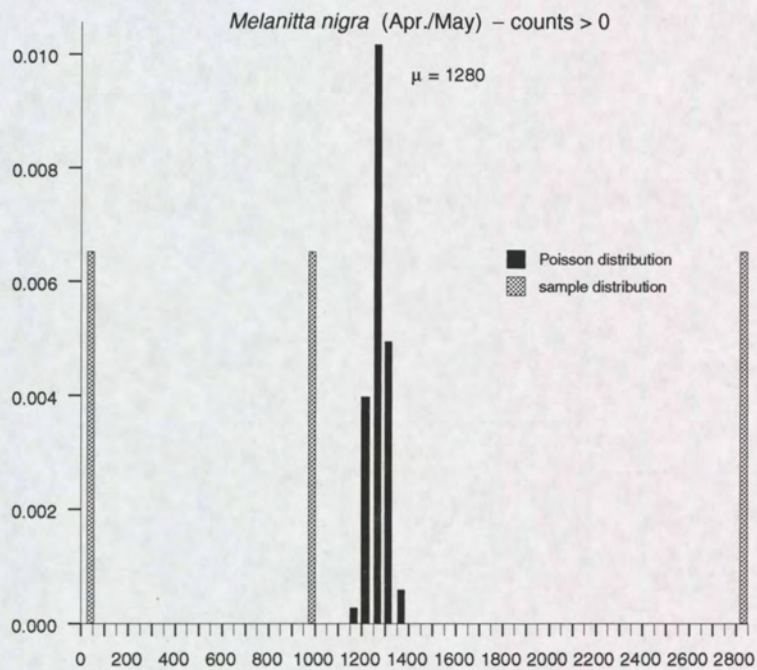
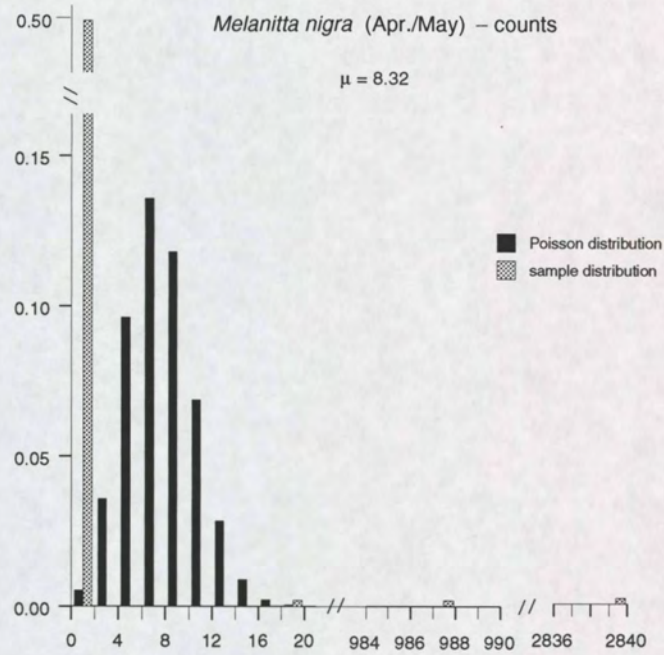
Sterna sandvicensis (Aug/Sep)



Alca torda/Uria aalge (Feb/Mar)



Melanitta nigra (Apr/May)



C the *Uria aalge*/*Alca torda* alternative

The logistic model obtained for the densities converted to a 0 (for zero density) and 1 (for non-zero density) variable, is as follows:

| | |
|----------------------|--|
| Terms in final model | dis + dis ² + dep + dep ² + dep ³ |
| Null Deviance | 367.6 on 363 degrees of freedom |
| Residual Deviance | 304.3 on 358 degrees of freedom |
| Explained Deviance | 17% |
| Dispersion parameter | 0.8126 |

The variogram for residuals from this model is shown in Fig. C.1 (top).

The Poisson model applied to non-zero observations only, has:

| | |
|----------------------|--------------------------------|
| Terms in final model | dis + dep + dep ² |
| Null Deviance | 113.4 on 73 degrees of freedom |
| Residual Deviance | 80.8 on 70 degrees of freedom |
| Explained Deviance | 36% |
| Dispersion parameter | 1.3 |

The variogram for residuals from this model is shown in Fig. C.1 (bottom).

Variances for the residuals of the logistic (1/0) regression were taken as $\hat{\mu}(s_i)(1 - \hat{\mu}(s_i))$. Predicted values and prediction variances for the trend and trend+residual of the 0/1 data are shown in Fig. C.2. Predicted values and prediction variances for the trend and for trend+residual for the non-zero densities are shown in Fig. C.3.

Final predictions for this model were obtained by multiplying the predicted value from the 1/0 model, $\hat{Y}_1(s_i)$, with the predicted value for positive densities, $\hat{Y}_{>1}(s_i)$:

$$\hat{Y}_{\text{alt}}(s_i) = \hat{Y}_1(s_i) \times \hat{Y}_{>1}(s_i)$$

and $\hat{Y}_{\text{alt}}(s)$ is shown in Fig. C.4 (top, right). No prediction variances were obtained for this model.

The conditional Poisson distribution was assumed for positive counts. A slightly better approach would be to use the truncated Poisson distribution, suggested by Welsh et al. (1996).

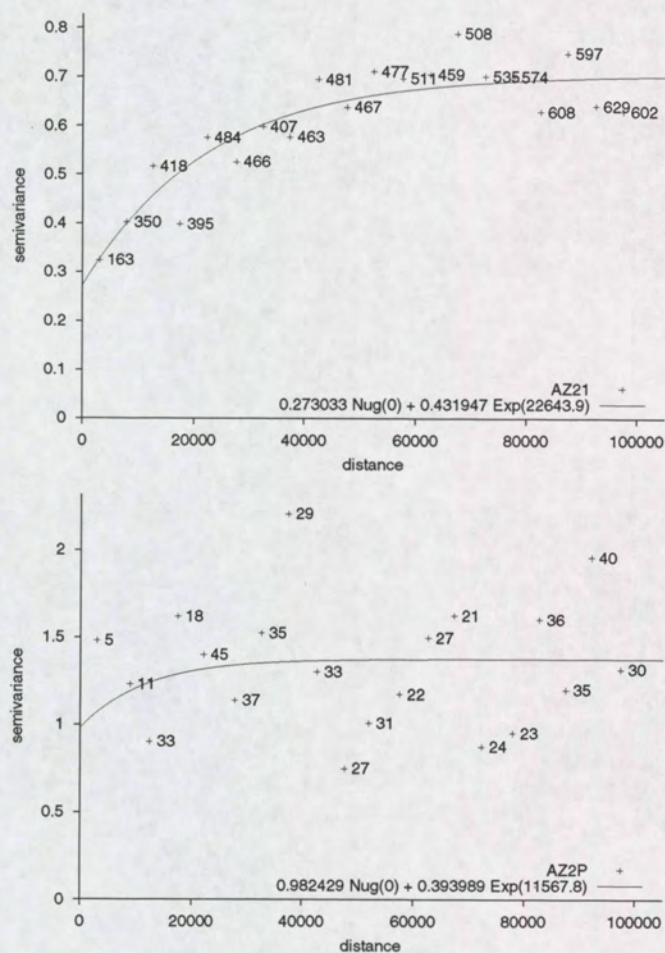


Figure C.1: Sample semivariograms (+) and fitted semivariogram models (—) for *Uria aalge/Alca torda*, period 2, under the alternative model. Top: semivariogram for 0/1 absence/presence Pearson residuals; bottom semivariogram for Pearson residuals based on the model for non-zero densities: semivariogram for Numbers indicate N_j , the number of residual pairs

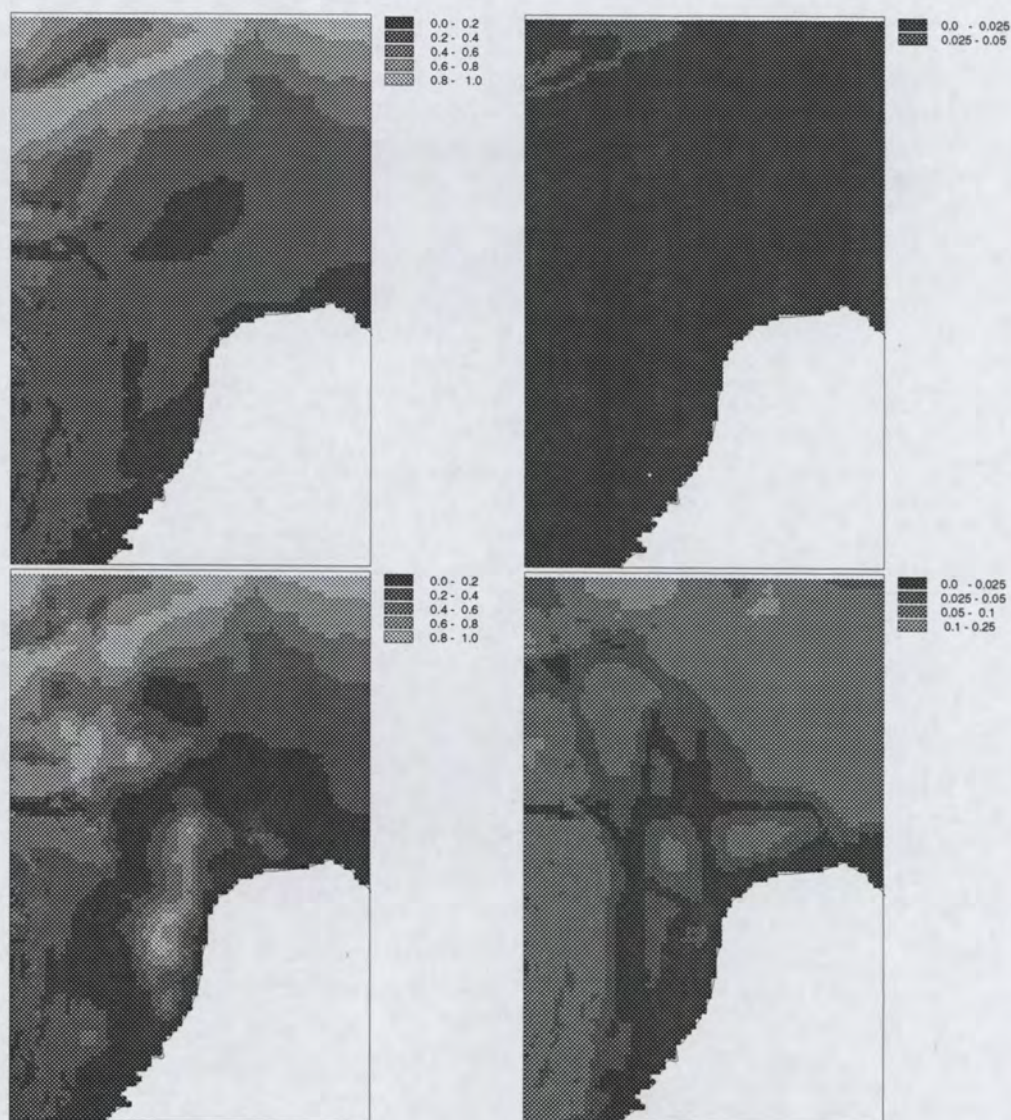


Figure C.2: Map of *Uria aalge/Alca torda*, period 2. Predicted value and prediction variance for the trend of 1/0 data (top: left and right) and for the sum of trend and residual (bottom: sum of predicted value left; sum of prediction variances right)

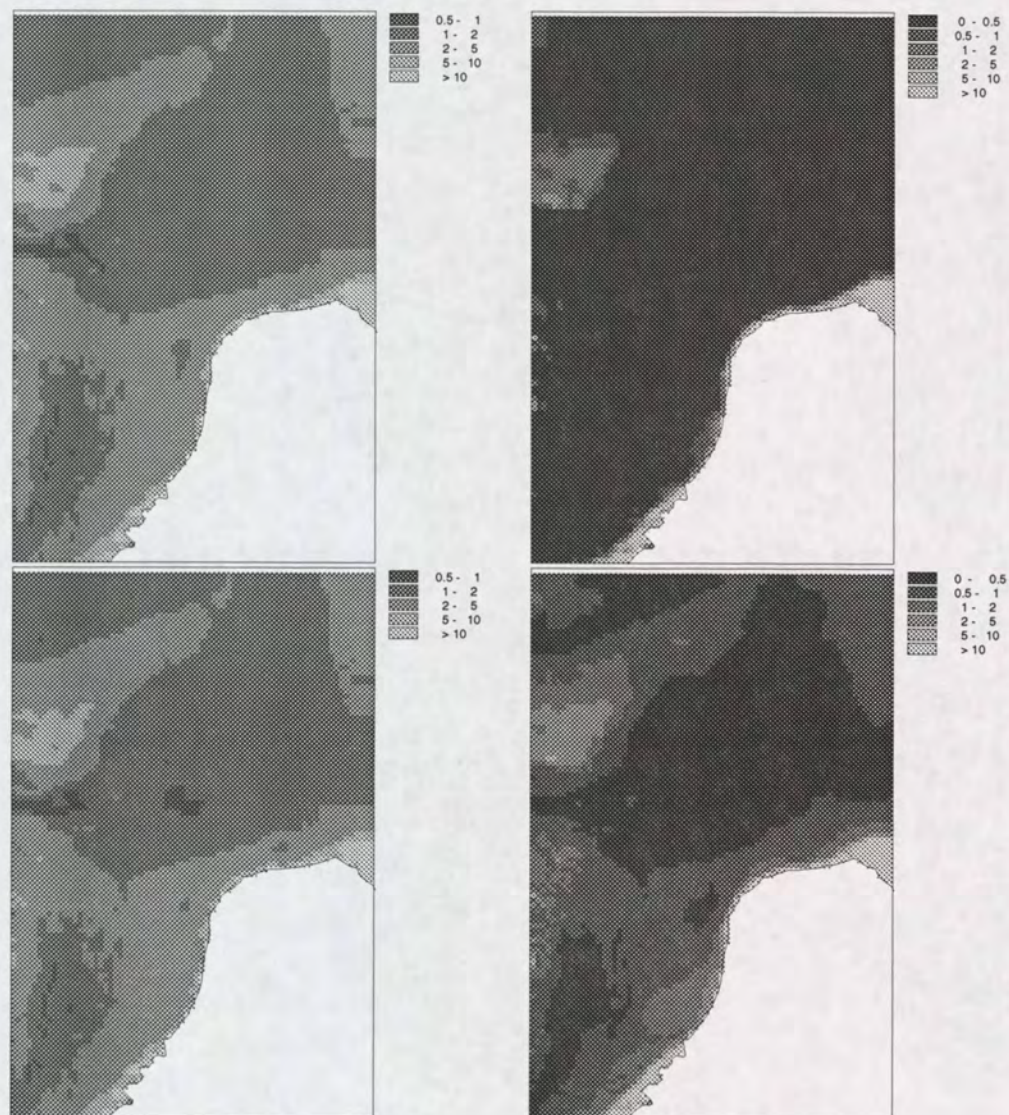


Figure C.3: Map of *Uria aalge/Alca torda*, period 2. Predicted value and prediction variance for the trend of non-zero densities (top: left and right) and for the sum of trend and residual (bottom: sum of predicted value left; sum of prediction variances right)

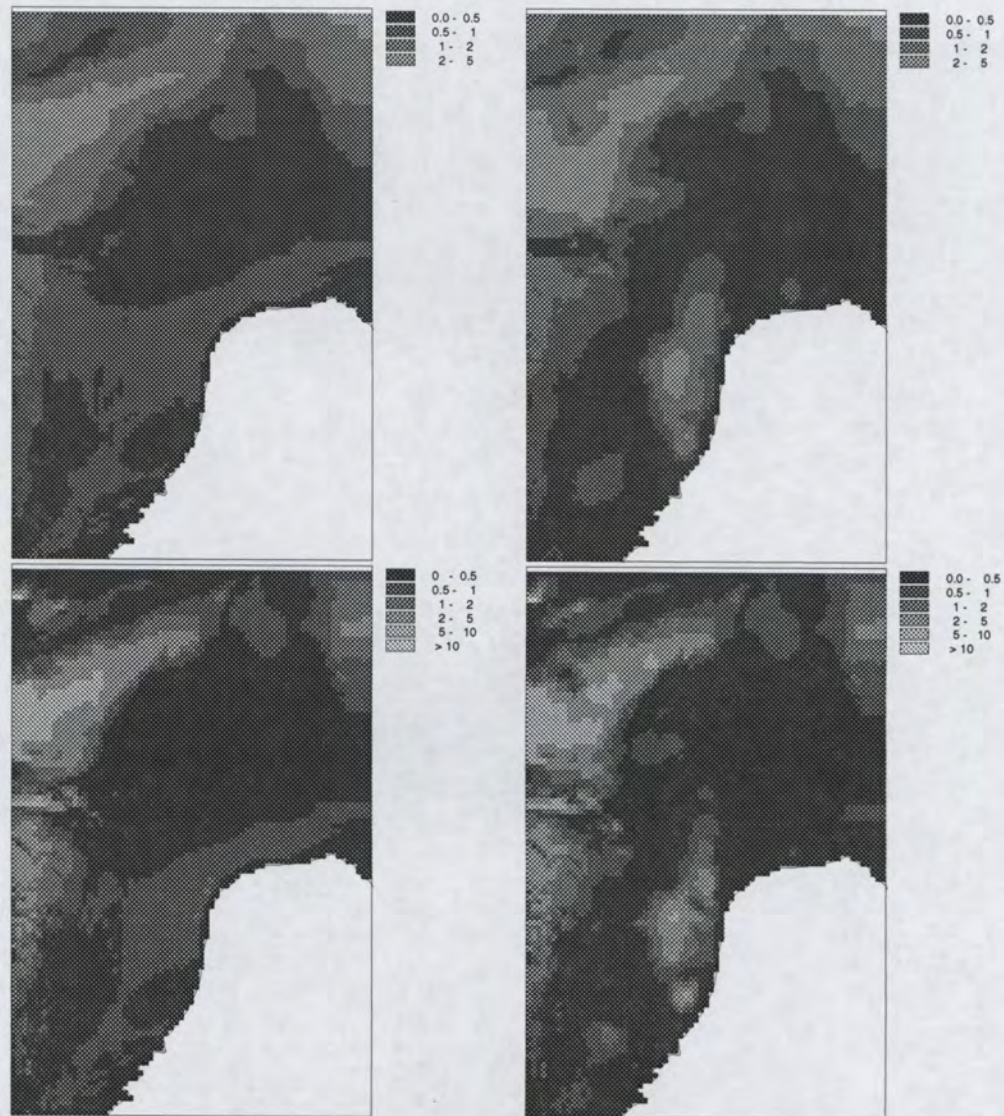


Figure C.4: Map of *Uria aalge*/*Alca torda*, period 2. Predicted value and prediction variance for the combined prediction of trend (top, left) and for the combined prediction of sum of trend and residual (top, right). For reference: the prediction map of the trend (bottom, left) and sum of trend and residual (bottom, right) as obtained from the standard approach, section 3.1 and 5.3

D Block averages

D.1 Theory

Theory of prediction of block averages is found Subsection 2.2.5 and Section 3.1.3.

D.2 Implementation

One implementation for predicting block averages (block kriging) is found in gstat (Pebesma and Wesseling, 1998). An example command file for block kriging with regular blocks on a grid mask map is:

```
data(NS5): 'NS5dat.txt', x=3, y=4, v=8, beta=[0,1],
  X=10, ave, VarFunction = "mu";
variogram(NS5): 0.944782 Nug(0) + 1.7139 Exp(34491.3);
mask: 'NS5.pr'; # map with mean value at prediction locations
predictions(NS5): 'NS5.rpr';
variances (NS5): 'NS5.rvr';
block: dx=10000, dy=10000;
```

Here, block size is 10000×10000 . The beta, X, and VarFunction key words are explained in appendix H.

For irregular shaped blocks, the points that discretise the block should be defined with the area statement, which has the same syntax as data statements:

```
data(NS5): 'NS5dat.txt', x=3, y=4, v=8, beta=[1],
  X=-1&10, ave, VarFunction = "mu";
variogram(NS5): 0.944782 Nug(0) + 1.7139 Exp(34491.3);
area: 'NS5coast.eas', x=1, y=2, v=3, X=3;
output: 'dlt10km.out';
```


Here, the block discretizing data NS5coast.eas were obtained as follows:

```
pcrcalc NS5coast.pr=mif(afst5km.csf lt 20000, NS5.pr)
(cuts the part from NS5.pr where 'afst5km', distance to the coast, is less then
20000 metres), and
map2col -g NS5coast.pr NS5coast.eas
(converts the grid map to an GeoEAS (ascii) column file)
```


E GEE: S-Plus output

This section lists some results for *Fulmarus glacialis*, period 5, as S-Plus output. The preliminary operations that were done to obtain rikz5 (period 5) and NS (densities of *Fulmarus glacialis*) are:

```
rikz <- read.table("/home/edzer/pc/rikz/ana/data/rikzdat.txt",header=T)

period <-(floor(rikz$dag/45)+1)      # get period index
period[period>4]<-(period[period>4]-1)
rikz$period<-period
rm(period)

# three species, as density:
rikz$AZ<- rikz$alk.zeekoet/rikz$km2 # Uria aalge/Alca torda
rikz$NS<- rikz$noordse.stormvogel/rikz$km2 # Fulmarus glacialis
rikz$GS<- rikz$grote.stern/rikz$km2 # Sterna sandvicensis

rikz1 <- na.omit(rikz[rikz$period == 1,])
rikz2 <- na.omit(rikz[rikz$period == 2,])
rikz3 <- na.omit(rikz[rikz$period == 3,])
rikz4 <- na.omit(rikz[rikz$period == 4,])
rikz5 <- na.omit(rikz[rikz$period == 5,])
rikz6 <- na.omit(rikz[rikz$period == 6,])
```

The following is the output of a number of YAGS runs with simple models for *Fulmarus glacialis*, period 5.

S-PLUS : Copyright (c) 1988, 1998 MathSoft, Inc.

S : Copyright Lucent Technologies, Inc.

Version 5.0 Release 3 for Linux 2.0.32 : 1998

Working data will be in .Data

Warning messages:

1: The functions and datasets in library section yags are not supported by


```

StatSci. in: library(yags, lib.loc = ".")
2: function "dyn.load" is obsolete; will assume ./yags/yags_1.o has been
    linked automatically in: dyn.load(paste(library, "/", section, "/",
    section, "_1.o", sep = ""))
sel <- unique.loc(rikz5$x,rikz5$y)
NS5 <- rikz5$NS[sel]
x <- rikz5$x[sel]
y <- rikz5$y[sel]
# dis <- poly(rikz5$afst5km[sel])
# dep <- poly(rikz5$diep5km[sel])
dis <- rikz5$afst5km[sel]
dep <- rikz5$diep5km[sel]
day <- rikz5$dag[sel]
cor.met <- dist2d(x,y)
vgm.params <- setup.vgm(200000, 5000, "exp", NULL, 4, 26000)

print(summary.yags(spat.yags(
  NS5 ~ dep,
  id=day,
  cor.met=cor.met,
  family=poisson,
  alpfun=vgm.alpfun,
  scalefun=vgm.scalefun,
  wcorigen=vgm.corigen)))
[1] "Starting iteration 1 with coefs"
[1] -0.74665740  0.02837257
      c1      a
5.974038 32753.85
[1] "Starting iteration 2 with coefs"
[1] -0.88385624  0.03260472
      c1      a
5.775868 32132.83
[1] "Starting iteration 3 with coefs"
[1] -0.89034315  0.03276027
      c1      a
5.777542 32113.84
[1] "Starting iteration 4 with coefs"
[1] -0.89087979  0.03277667
      c1      a
5.776877 32111.59
      c1      a

```



```
5.776898 32111.41
[1] "Scale parameters are really naive"
```

Call:

```
spat.yags(formula = NS5 ~ dep, id = day, cor.met = cor.met, family = poisson,
  alpfun = vgm.alpfun, scalefun = vgm.scalefun, wcorigen = vgm.corigen)
```

Summary of raw residuals:

| Min | 1Q | Median | 3Q | Max |
|-----------|------------|------------|-----------|----------|
| -6.890239 | -0.9462653 | -0.6806069 | -0.413399 | 13.28201 |

Coefficients:

| | Estimate | Naive S.E. | Naive z | Robust S.E. | Robust z |
|-------------|------------|-------------|-----------|-------------|-----------|
| (Intercept) | -0.8909345 | 0.604221358 | -1.474517 | 0.8133425 | -1.095399 |
| dep | 0.0327784 | 0.009757592 | 3.359271 | 0.0153620 | 2.133733 |

Estimated Scale Parameter: 4.79508

Number of Iterations: 4

Working Correlation Parameter(s)

```
[1] 5.776898 32111.408225
```

Warning messages:

```
No formal definition of old-style inheritance; consider
  setOldClass(c("pmat", "blo.... in: checkOldClass(c("pmat", "block",
    "vstack"))
```

```
print(summary.yags(spat.yags(
  NS5 ~ dis,
  id=day,
  cor.met=cor.met,
  family=poisson,
  alpfun=vgm.alpfun,
  scalefun=vgm.scalefun,
  wcorigen=vgm.corigen)))
[1] "Starting iteration 1 with coefs"
[1] -3.865205e-01 5.271024e-06
      c1      a
7.420227 41984.6
[1] "Starting iteration 2 with coefs"
[1] -3.025795e-01 4.752669e-06
```



```

      c1      a
7.248421 41544.9
[1] "Starting iteration 3 with coefs"
[1] -3.054164e-01 4.785897e-06
      c1      a
7.240712 41572.64
      c1      a
7.240621 41572.06
[1] "Scale parameters are really naive"

```

Call:

```

spat.yags(formula = NS5 ~ dis, id = day, cor.met = cor.met, family = poisson,
          alpfun = vgm.alpfun, scalefun = vgm.scalefun, wcorigen = vgm.corigen)

```

Summary of raw residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|------------|----------|------------|----------|
| -4.00739 | -0.9156712 | -0.77299 | -0.4719219 | 13.51688 |

Coefficients:

| | Estimate | Naive S.E. | Naive z | Robust S.E. | Robust z |
|-------------|---------------|--------------|-----------|--------------|------------|
| (Intercept) | -3.053254e-01 | 6.722387e-01 | -0.454192 | 5.150038e-01 | -0.5928605 |
| dis | 4.783625e-06 | 3.391263e-06 | 1.410573 | 2.295206e-06 | 2.0841809 |

Estimated Scale Parameter: 5.455069

Number of Iterations: 3

Working Correlation Parameter(s)

```

[1] 7.240621 41572.060950

```

```

print(summary.yags(spat.yags(
  NS5 ~ dep + dis,
  id=day,
  cor.met=cor.met,
  family=poisson,
  alpfun=vgm.alpfun,
  scalefun=vgm.scalefun,
  wcorigen=vgm.corigen)))
[1] "Starting iteration 1 with coefs"
[1] -6.954250e-01 2.072487e-02 1.799401e-06
      c1      a

```



```

6.175877 35782.44
[1] "Starting iteration 2 with coefs"
[1] -7.908760e-01  3.785916e-02 -2.318783e-06
      c1      a
5.944197 28995.84
[1] "Starting iteration 3 with coefs"
[1] -8.603621e-01  4.009353e-02 -2.383314e-06
      c1      a
5.90397 28666.35
[1] "Starting iteration 4 with coefs"
[1] -8.672941e-01  4.059542e-02 -2.491057e-06
      c1      a
5.92734 28501.44
[1] "Starting iteration 5 with coefs"
[1] -8.680779e-01  4.062998e-02 -2.488314e-06
      c1      a
5.921818 28499.23
[1] "Starting iteration 6 with coefs"
[1] -8.682494e-01  4.064158e-02 -2.490816e-06
      c1      a
5.922443 28495.44
      c1      a
5.922459 28495.33
[1] "Scale parameters are really naive"

```

Call:

```

spat.yags(formula = NS5 ~ dep + dis, id = day, cor.met = cor.met, family =
  poisson, alpfun = vgm.alpfun, scalefun = vgm.scalefun, wcorigen =
  vgm.corigen)

```

Summary of raw residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|-----------|------------|------------|----------|
| -5.92477 | -1.024913 | -0.7362392 | -0.3071871 | 13.68878 |

Coefficients:

| | Estimate | Naive S.E. | Naive z | Robust S.E. | Robust z |
|-------------|---------------|--------------|------------|--------------|------------|
| (Intercept) | -8.682668e-01 | 5.971063e-01 | -1.4541243 | 7.983055e-01 | -1.0876372 |
| dep | 4.064235e-02 | 1.668531e-02 | 2.4358168 | 2.407680e-02 | 1.6880299 |
| dis | -2.490747e-06 | 4.291654e-06 | -0.5803701 | 3.486807e-06 | -0.7143348 |

Estimated Scale Parameter: 5.189705

Number of Iterations: 6

Working Correlation Parameter(s)

[1] 5.922459 28495.333042

add a measurement error of 1.0 on the 'Pearson residual' scale:

vgm.params <- setup.vgm(200000, 5000, "exp", NULL, 4, 26000, meas.error = 0.5)

print(summary.yags(spat.yags(

NS5 ~ dep + dis,

id=day,

cor.met=cor.met,

family=poisson,

alpfun=vgm.alpfun,

scalefun=vgm.scalefun,

wcorigen=vgm.corigen)))

[1] "Starting iteration 1 with coefs"

[1] -6.954250e-01 2.072487e-02 1.799401e-06

c1 a

5.68847 38230.64

[1] "Starting iteration 2 with coefs"

[1] -7.596402e-01 3.606815e-02 -1.875636e-06

c1 a

5.288886 31483.43

[1] "Starting iteration 3 with coefs"

[1] -7.935018e-01 3.725001e-02 -1.873805e-06

c1 a

5.217543 31333.65

[1] "Starting iteration 4 with coefs"

[1] -7.938233e-01 3.744167e-02 -1.929181e-06

c1 a

5.222869 31230.71

c1 a

5.223126 31228.76

[1] "Scale parameters are really naive"

Call:

spat.yags(formula = NS5 ~ dep + dis, id = day, cor.met = cor.met, family =
poisson, alpfun = vgm.alpfun, scalefun = vgm.scalefun, wcorigen =
vgm.corigen)

Summary of raw residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|-----------|------------|-----------|----------|
| -5.8417 | -1.047695 | -0.7601036 | -0.355378 | 13.57978 |

Coefficients:

| | Estimate | Naive S.E. | Naive z | Robust S.E. | Robust z |
|-------------|---------------|--------------|------------|--------------|------------|
| (Intercept) | -7.937541e-01 | 6.378208e-01 | -1.2444782 | 8.021069e-01 | -0.9895864 |
| dep | 3.742622e-02 | 1.966682e-02 | 1.9030135 | 2.460351e-02 | 1.5211740 |
| dis | -1.920701e-06 | 4.839876e-06 | -0.3968493 | 3.221195e-06 | -0.5962698 |

Estimated Scale Parameter: 4.946718

Number of Iterations: 4

Working Correlation Parameter(s)

[1] 5.223126 31228.762421

F S-Plus source code

S-Plus source code Copyright (C) 1999 Edzer J. Pebesma. All source files listed are covered by the GPL ¹ (General Public Licence) version 2 or later. The functions `spat.yags` and `spat.yags.pmat.fit` were modified from the YAGS library, (C) V.J. Carey. YAGS is also covered by the GPL.

Note that the functions are not guaranteed to work in general. They have only been tested for a limited number of cases. Also, they have only been tested with S-Plus version 5.0 release 3 for Linux 2.0.32, on a Linux computer running Red Hat 6.0. (Run time for the models in the previous section was a few minutes on a 300 MHz Pentium II.)

S-Plus sources for functions and the data sets can be downloaded from
<http://www.geog.uu.nl/~pebesma/rikz/>

¹See <http://www.gnu.org/copyleft/gpl.html>

G Problems with estimating α

It may for certain combination of regression models and variogram models be difficult or even impossible to find good estimates. This will lead to rather cryptical error messages from somewhere in the estimating routines. Three cases will be discussed. For clarity, the option `trace=T` was added to the `fit.vgm` call in `vgm.alpfun`.

divergence This happens when the variogram parameters tend to diverge.

For instance, when we try to fit an exponential semivariogram model to an apparent linear sample semivariogram, then the range and sill will both run to infinity, in order to reach a linear form:

```
> vgm.params <- setup.vgm(100000, 5000, "exp", 1, 4, 26000)
> print(summary.yags(spat.yags(NS5 ~ dis, id = day, cor.met = cor.met,
family = poisson, alpfun = vgm.alpfun, scalefun = vgm.scalefun,
wcorigen = vgm.corigen)))
[1] "Starting iteration 1 with coefs"
[1] -3.865205e-01  5.271024e-06
24287.9 : 1 4 26000
20592.3 : 3.64209 5.94946 122213
12818.2 : 2.55434 22.0598 505152
12333.9 : 2.22897 99.8575 2325650
12266.5 : 2.16544 441.457 10311200
12255.7 : 2.15042 2047.86 47862300
12255.5 : 2.14671 10605.9 247912000
```

```
Problem in nls( ~ sqrt(npairs) * (semivariance - (c0...
step factor reduced below minimum
Use traceback() to see the call stack
```

This would suggest the use of a linear semivariogram model. And indeed:


```

vgm.params <- setup.vgm(100000, 5000, "lin", 1, 1e-4, NULL)
print(summary.yags(spat.yags(
NS5 ~ dis,
id=day,
cor.met=cor.met,
family=poisson,
alpfun=vgm.alpfun,
scalefun=vgm.scalefun,
wcorigen=vgm.corigen)))

```

```
[1] "Starting iteration 1 with coefs"
```

```
[1] -3.865205e-01  5.271024e-06
```

```
      c0      c1
```

```
0.2189661 8.583332e-05
```

```
[1] "Starting iteration 2 with coefs"
```

```
[1] -6.525330e-02 -2.084776e-06
```

```
      c0      c1
```

```
1.367178 0.0001605816
```

```
[1] "Starting iteration 3 with coefs"
```

```
[1] -1.247553e-01 -2.195701e-06
```

```
      c0      c1
```

```
1.497853 0.0001730579
```

```
[1] "Starting iteration 4 with coefs"
```

```
[1] -1.400878e-01 -2.105377e-06
```

```
      c0      c1
```

```
1.48342 0.0001735114
```

```
<...>
```

```
[1] "Starting iteration 24 with coefs"
```

```
[1] -1.349033e-01 -2.143132e-06
```

```
      c0      c1
```

```
1.491289 0.0001735326
```

```
[1] "Starting iteration 25 with coefs"
```

```
[1] -1.337002e-01 -2.151913e-06
```

```
      c0      c1
```

```
1.493128 0.0001735381
```

```
      c0      c1
```

```
1.493292 0.0001735327
```

```
Call:
```

```
spat.yags(formula = NS5 ~ dis, id = day, cor.met = cor.met, family = pois
```



```
alpfun = vgm.alpfun, scalefun = vgm.scalefun, wcorigen = vgm.corigen)
```

```
"Error code was niter==maxiter"
```

```
...
```

this run almost converges to a stable solution. Increasing the maxiter argument to `yags` may be sufficient to find it.

Other options would be (i) to increase the distance up to which semi-variances are calculated to 200000 in order to include the range where semivariance do not increase anymore, or (ii) to fix the nugget variance (defining a known measurement error) of for instance 1, in order to stabilise the non-linear fitting algorithm. Both options are applied in the following run:

```
vgm.params <- setup.vgm(200000, 5000, "exp", NULL, 4, 26000, 1)
print(summary.yags(spat.yags(
NS5 ~ dis,
id=day,
cor.met=cor.met,
family=poisson,
alpfun=vgm.alpfun,
scalefun=vgm.scalefun,
wcorigen=vgm.corigen)))
[1] "Starting iteration 1 with coefs"
[1] -3.865205e-01  5.271024e-06
      c1      a
6.476162 47932.73
[1] "Starting iteration 2 with coefs"
[1] -3.298177e-01  4.849199e-06
      c1      a
6.418728 47531.4
[1] "Starting iteration 3 with coefs"
[1] -3.331056e-01  4.882647e-06
      c1      a
6.414231 47571.51
[1] "Starting iteration 4 with coefs"
[1] -3.328747e-01  4.879262e-06
      c1      a
6.415432 47566.74
      c1      a
```



```
6.415448 47566.86
[1] "Scale parameters are really naive"
```

```
Call:
spat.yags(formula = NS5 ~ dis, id = day, cor.met = cor.met,
family = poisson, alpfun = vgm.alpfun, scalefun = vgm.scalefun,
wcorigen = vgm.corigen)
```

```
Summary of raw residuals:
```

| Min | 1Q | Median | 3Q | Max |
|-----------|------------|------------|------------|----------|
| -4.033156 | -0.8946614 | -0.7526912 | -0.4520617 | 13.53812 |

```
Coefficients:
```

| | Estimate | Naive S.E. | Naive z | Robust S.E. | Robust z |
|-------------|---------------|--------------|-----------|-------------|------------|
| (Intercept) | -3.328963e-01 | 7.711126e-01 | -0.431709 | 5.20710e-01 | -0.6393123 |
| dis | 4.879611e-06 | 3.907880e-06 | 1.248659 | 2.58003e-06 | 1.8913000 |

```
Estimated Scale Parameter: 5.540637
```

```
Number of Iterations: 4
```

```
Working Correlation Parameter(s)
```

```
[1] 6.415448 47566.855327
```

no convergence In the following example, the `nls` function does not converge in the maximum number of iterations allowed. Still, convergence seems likely when this maximum would be increased.

```
> vgm.params <- setup.vgm(200000, 5000, "exp", 1, 4, 26000)
> print(summary.yags(spat.yags(NS5 ~ dis, id = day, cor.met = cor.met,
family = poisson, alpfun = vgm.alpfun, scalefun = vgm.scalefun,
wcorigen = vgm.corigen)))
[1] "Starting iteration 1 with coefs"
[1] -3.865205e-01 5.271024e-06
61844.9 : 1 4 26000
27405.6 : 0.350084 7.11415 62114.5
27256.9 : -0.941789 7.52947 21686.1
26181.6 : 1.9305 5.47429 47511.5
25402.2 : -0.100484 7.21236 27106.4
21258.8 : 0.0411863 7.4257 45094.9
20560.9 : -1.01024 8.22089 33373.6
```



```

20214.2 : -1.17568 8.57386 37797.4
20158.2 : -1.41506 8.75144 35595.1
20157.4 : -1.44539 8.78863 35586.4
      c0      c1      a
-1.445395 8.78863 35586.41
[1] "Starting iteration 2 with coefs"
[1] -2.943658e-01 4.935112e-06
47446.6 : 1 4 26000
23707.7 : 0.323105 6.75955 59823.1
22923 : -0.897907 7.23781 23482.8
21117.6 : 1.12144 5.93896 46244
20221.2 : -0.433798 7.2286 30212.8
18511.4 : -0.63419 7.70097 41125.5
18215.6 : -1.21181 8.13155 34919.8
18184.7 : -1.30436 8.28505 35921.6
18184.4 : -1.40768 8.3617 34979.5
18183.9 : -1.31085 8.29002 35841
18183.7 : -1.39979 8.35606 35055.6
18183.3 : -1.31905 8.29629 35773.8
18183.2 : -1.39316 8.35129 35118.9
18182.9 : -1.32582 8.30143 35717.6
18182.8 : -1.38757 8.34725 35171.6
18182.6 : -1.33143 8.30567 35670.6
18182.6 : -1.38287 8.34384 35215.4
18182.5 : -1.33606 8.30917 35631.3
18182.4 : -1.37893 8.34096 35251.9
18182.3 : -1.3399 8.31206 35598.6
18182.3 : -1.37562 8.33854 35282.3
18182.2 : -1.34309 8.31445 35571.2
18182.2 : -1.37285 8.33651 35307.6
18182.2 : -1.34574 8.31643 35548.3
18182.1 : -1.37053 8.33481 35328.7
18182.1 : -1.34793 8.31807 35529.3
18182.1 : -1.36859 8.33339 35346.2
18182.1 : -1.34976 8.31943 35513.4
18182.1 : -1.36697 8.3322 35360.8
18182.1 : -1.35128 8.32057 35500.2
18182.1 : -1.36562 8.3312 35373
18182.1 : -1.35254 8.32151 35489.1
18182.1 : -1.36449 8.33037 35383.2
18182 : -1.35359 8.32229 35479.9

```



```

18182 : -1.36355 8.32967 35391.6
18182 : -1.35446 8.32294 35472.2
18182 : -1.36276 8.32909 35398.6
18182 : -1.35519 8.32348 35465.8
18182 : -1.3621 8.3286 35404.5
18182 : -1.35579 8.32393 35460.5
18182 : -1.36156 8.3282 35409.4
18182 : -1.3563 8.3243 35456
18182 : -1.3611 8.32786 35413.4
18182 : -1.35672 8.32461 35452.3
18182 : -1.36072 8.32758 35416.8
18182 : -1.35707 8.32487 35449.2
18182 : -1.3604 8.32734 35419.7
18182 : -1.35736 8.32509 35446.7
18182 : -1.36014 8.32715 35422
18182 : -1.3576 8.32527 35444.5
Problem in nls( ~ sqrt(npairs) * (semivariance - (c0...: maximum number of
Use traceback() to see the call stack

```

singular regression parameter covariances This problem occurs when the regression parameters (β) are not estimable anymore under the given semivariogram model. The solution would be to choose another regression model, or another semivariogram model.

```

vgm.params <- setup.vgm(100000, 5000, "lin", 1, 1e-4, NULL)

print(summary.yags(spat.yags(
NS5 ~ dis + dep + dis:dep,
id=day,
cor.met=cor.met,
family=poisson,
alpfun=vgm.alpfun,
scalefun=vgm.scalefun,
wcorigen=vgm.corigen)))
[1] "Starting iteration 1 with coefs"
[1] -8.192145e+00  3.790883e-05  2.300028e-01 -9.469364e-07
      c0      c1
1.001955 1.638228e-05

```

Problem in solve.qr(a): apparently singular matrix
Use traceback() to see the call stack


```
traceback()
15: eval(action, sys.parent())
14: doErrorAction("Problem in solve.qr(a): apparently singular matrix", 1000)
13: stop("apparently singular matrix")
12: solve.qr(a)
11: solve.default(S2)
10: solve(S2)
9: spat.yags.pmat.fit(X, Y, id, weights, cor.met, family, alpfun, scalefun,
8: spat.yags(NS5 ~ dis + dep + dis:dep, id = day, cor.met = cor.met, family =
7: summary.yags(spat.yags(NS5 ~ dis + dep + dis:dep, id = day, cor.met = cor.met,
6: length(names(sig))
5: length(names(sig)) > 0
4: hasMethod("show", el(class(x), 1))
3: (nargs() == 1 || hasArg("prefix")) && hasMethod("show", el(class(x), 1))
2: print(summary.yags(spat.yags(NS5 ~ dis + dep + dis:dep, id = day, cor.met =
1:
Message: Problem in solve.qr(a): apparently singular matrix
```


H Mean-dependent covariances in gstat

In generalized linear models, the assumption of constant (homoscedastic) variances is replaced (generalized) by an assumption of mean-dependent variances. For instance, for binomial variables the variance of an observation y_i can be taken as $\mu_i(1 - \mu_i)$, and for Poisson variables it can be taken as μ_i .

In geostatistics, the variances and covariances are typically chosen to be stationary, meaning that they do not depend on mean values. Recently, this assumption has been loosened (e.g. Gotway and Stroup, 1997).

Gstat now implements two options for prediction with mean dependent covariances – the binomial case where $\text{Var}(y_i) = \mu_i(1 - \mu_i)$ and the Poisson case where $\text{Var}(y_i) = \mu_i$. It is implemented as an extension of simple kriging with a known, non-constant mean, and we will explain this option first. This functionality will be available in the upcoming release (2.1) of gstat.

H.1 Simple kriging with known, non-constant mean

In gstat, non-constant mean functions are implemented as linear functions, e.g.

$$Z(s_i) = \sum_{j=1}^p f_j(s_i)\beta_j + e(s_i)$$

By default, only an intercept is defined ($p = 1$, $f_1(s_i) = 1$ for all s_i , and β_1 is an unknown constant, resulting in ordinary kriging). Adding X column numbers can be used to add other functions, and this would result in universal kriging. When, however, the mean parameters are *known*, simple kriging is used. Consider the following example:

```
data(AZ2): 'AZ2n.txt', x=3, y=4, v=11, beta=[0,1], X=7;  
variogram(AZ2): 1.40457 Exp(9342.37);
```



```
mask: '/home/edzer/rikz/trend/AZ2.pr';
predictions(AZ2): 'AZ2pr.map';
variances(AZ2): 'AZ2var.map';
```

Here, $p = 2$, $f_1(s_i) = 1$ for all s_i , $f_2(s_i)$ is given in column 7 for the data locations, and in the mask map for prediction locations. The vector β is now defined: $\beta_1 = 0$ and $\beta_2 = 1$, resulting in the known mean function

$$1 \times 0 + f_2(s_i) \times 1 = f_2(s_i)$$

In other words, this value for **beta** results in a mean value that *equals* the value equal of $f_2(s_i)$. Prediction at s_0 is now obtained by

$$\hat{Z}(s_0) = f_2(s_0) + \hat{e}(s_0)$$

with $\hat{e}(s_0)$ the simple kriging prediction of the residual. Note that the data (column 11) are observed values of $Z(s_i)$, they are not residuals.

For prediction at arbitrarily spaced point locations one proceeds as follows. The last three statements are replaced by something like:

```
data(): 'pt1', x=1, y=2, X=3;
output: 'AZ2n.out';
```

indicating that column 3 in pt1 holds the values of $f_2(s_0)$.

H.2 Mean-dependent covariances

Mean-dependent covariances are obtained when in addition to a varying, known mean value, a variance function is defined. This is done by

```
data(AZ2): 'AZ2n.txt', x=3, y=4, v=11, beta=[0,1], X=7,
  VarFunction = 'mu';
```

for Poisson data, or, alternatively `VarFunction = 'mu(1-mu)'` for binomial data. In the first case, $\sigma^2(s_i) = \mu_i$, in the second case $\sigma^2(s_i) = \text{Var}(Z(s_i)) = \mu_i(1 - \mu_i)$.

The covariance for data pair $Z(s_i), Z(s_j)$ is now obtained by

$$\text{Cov}(Z(s_i), Z(s_j)) = \sigma(s_i)\sigma(s_j)\phi\rho(h)$$

The variogram defined in the command file for the observed data should be that of the standardised (Pearson) residuals, which is related to the correlogram $\rho(h)$ by

$$\gamma_p(h) = \phi(1 - \rho(h))$$

H.3 Differences from previous approach

In the previous approach (Pebesma et al, 1999), we estimated the trend and predicted the residual separately, and added them afterwards. Now, the simple kriging prediction is directly the sum of the trend (which is defined for data and prediction locations) and the predicted residual, so these two do not have to be added afterwards. The prediction variance from the simple kriging is however still the residual prediction variance. To obtain the same variances as in the previous approach, we will have to add the estimation error of the mean (trend).

I Geostats 2000 conference paper

MAPPING SEA BIRD DENSITIES ON THE NORTH SEA: COMBINING GEOSTATISTICS AND GENERALISED LINEAR MODELS

Edzer J. Pebesma¹, Ana F. Bio², Richard N.M. Duin³.

(Conference paper, presented at Geostats 2000, held April 8-11, 2000, Kaapstad, South Africa)

Abstract. Using airborne strip-transect monitoring data, maps of bird densities were estimated for several bird species. Because the densities are transformed count data with many genuine zeros, we combined a generalised linear modelling approach with geostatistics for the spatial interpolation. Water depth and distance to coast were used as covariates to model the trend. A generalised estimating equations-like approach was used to estimate the trend, the spatial correlation function and the over-dispersion parameter. The residual variance was taken proportional to the (varying) mean, and non-stationary residual variances and covariances were obtained from known means, a stationary correlogram and the over-dispersion parameter. The results for one species (*Fulmarus glacialis*) are shown as approximate 95% prediction intervals of 5 km \times 5 km block mean densities.

I.1 Introduction

Since 1984, the Dutch National Institute for Coastal and Marine Management (RIKZ) monitors sea birds on the Dutch part of the North Sea (NCP)

¹Dept. of Physical Geography, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands, e.pebesma@geog.uu.nl

²Dept. of Environmental Sciences, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands

³National Institute for Coastal and Marine Management, P.O. Box 20907, 2500 EX The Hague, The Netherlands

using an airborne observation technique (Baptist and Wolf, 1993). Since 1989 this monitoring is carried out systematically on a bi-monthly basis. Each monitoring round consists of three days of flying, following a fixed flight schedule. The reason for monitoring is to get insight (i) in the spatial distribution of sea birds (and, to a lesser extent, marine mammals) over the NCP and (ii) in temporal changes in the spatial distribution of sea birds, and (iii) gaining knowledge about the sea bird species involved, and the fish or shell species they eat. This paper will address the first goal only. To avoid misinterpretation of inaccurately predicted values, and given the fact that the spatial mapping involves large areas with no measurements nearby, we have to present interpolated results as interval estimates.

For a given bird species, the data collected are numbers of birds observed within consecutive strips of approximately 150 m wide and 6 km long. Division through strip area transforms the numbers to bird densities.

Because large parts of the mapping area have no observations nearby, simple interpolation methods such as inverse distance weighted interpolation or ordinary kriging would fail to give sensible predictions at locations beyond correlation distances from observations. As a first approach, two known covariates that may be important to birds, water depth and distance to the (Dutch) coast, were used as external information to model the trend, in a regression-like manner.

Count data are usually modelled with log-linear models, by using the Generalised Linear Models (GLM) framework (McCullagh and Nelder, 1989). Mostly for applications to longitudinal data (temporally correlated data), this framework has been extended to that of the Generalised Estimating Equations (GEE) by Liang and Zeger (1986) and Zeger and Liang (1986; see also Diggle, Liang and Zeger, 1994), but their emphasis remained on testing hypotheses and estimating regression parameters. For spatial interpolation we need to use these models in a predictive manner, and Gotway and Stroup (1997) did this by addressing point kriging prediction. Here, we follow their approach, and we attempt to extend the method with block kriging and (approximate) prediction interval estimation.

The next section deals with the monitoring data. Then, a section follows on the spatial interpolation method, and a the results section is concluded by a discussion.

I.2 The Monitoring Data

The bird data are collected as strip transect counts. During the flight, depending on light conditions either on one or on both sides of the plane, all

visible birds within a strip of 150 m wide are registered for fixed periods of approximately 2 minutes. This corresponds to a strip length of approximately 6 km, and a "support" of measurements of approximately 1 km². For a given species a single "observation" therefore does not correspond to the occurrence of individual birds, but to the number of birds observed within a strip that has a known location and size (derived from observation time, flying speed and flight plan coordinates). An example of observed densities for *Fulmarus glacialis* (Fulmar) is shown in Fig. I.1.

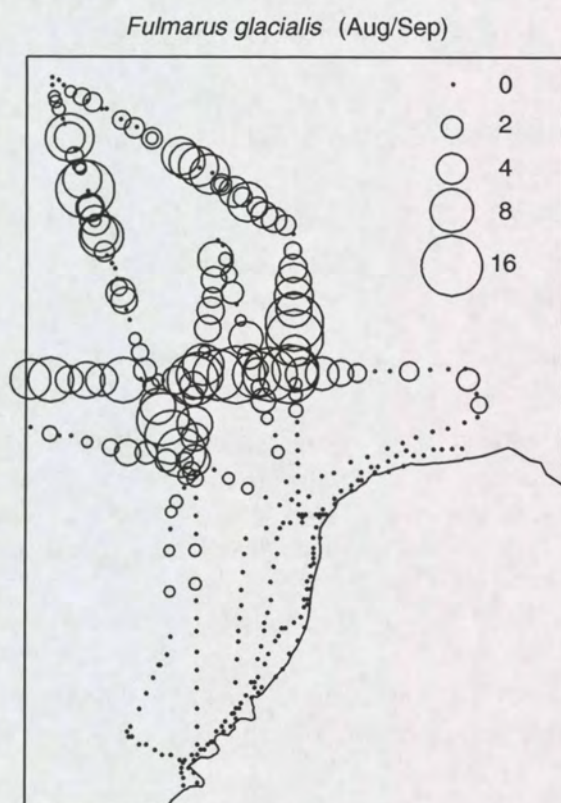


Figure I.1: Observed counts of *Fulmarus glacialis* during the Aug/Sep monitoring of 1995. Circle size denotes bird density (km⁻²) and circle centres denote the observation location centers; the curved line is the (generalised) Dutch coast line. The study area is 355 × 545 km.

I.3 Spatial Interpolation

I.3.1 Generalised Linear Models

Given observations $Z(s_i), i = 1, \dots, n$ taken at n spatial locations s_i , and covariates $X_j(s)$, the standard linear model for the data is

$Z(s) = \sum_{j=0}^p X_j(s)\beta_j + e(s)$, with $Z(s) = (Z(s_1), \dots, Z(s_n))'$ the observations, $X_j(s) = (X_j(s_1), \dots, X_j(s_n))'$ the known covariates, and $p+1$ the number of unknown regression coefficients β_j , where β_0 is usually an intercept. ANOVA and regression models are obtained when $\text{Cov}(e(s))$, the covariance of $e(s)$, is a diagonal matrix (i.e., independent residuals). Universal kriging is the prediction of $Z(s)$ when the residuals are spatially correlated (e.g., Christensen, 1991; Cressie, 1993).

Generalised linear models (McCullagh and Nelder, 1989) extend the standard linear model with (i) a non-linear link function and (ii) a variance function according:

- (i) the function $g(\cdot)$ that relates the expectation of $Z(s)$ to the linear model, $g(\mu(s)) = \eta(s) = \sum_{j=0}^p X_j(s)\beta_j$ [with $\mu(s) = E(Z(s))$], is called the link function, and
- (ii) the function $V(\cdot)$ that relates the variance of $Z(s)$ to the mean value $\mu(s)$ is called the variance function: $\text{Var}(Z(s)) = V(\mu(s))$.

An identity link function and a constant variance function brings one back to standard linear models.

One model often used for countdata is the log-linear model (McCullagh and Nelder, 1989), which we will use for density our observations $Z(s)$:

$$Z(s) = \mu(s) + e(s), \quad E(Z(s)) = \mu(s), \quad \log(\mu(s)) = \eta(s) = \sum_{j=0}^p X_j(s)\beta_j \quad (\text{I.1})$$

with $\mu(s)$ the expected density at s (birds per km^2). For the variance of the residuals we assumed it is proportional to the mean value,

$$\text{Var}(e(s)) = \phi\mu(s) \quad (\text{I.2})$$

with ϕ the over-dispersion parameter.

As the count data often consist of many zeros, we cannot fit model (1) directly to transformed data as this would require taking the log of zeros. Therefore an iterative method that works with re-estimation of residuals is used for this. Details are found in McCullagh and Nelder (1989, sec. 2.5). Generalised linear models were originally developed for independent data, but recently extensions that allow for temporal or spatial correlation have been developed.

I.3.2 Spatial correlation

Equation (I.2) shows that when $\mu(s)$ is not a constant, the variance of $e(s)$ will be non-constant and as a consequence we cannot model $e(s)$ as a stationary random function. Therefore, calculating the variogram from $e(s)$ is of no use, but we could use Pearson residuals instead,

$$p(s) = \frac{Z(s) - \hat{\mu}(s)}{\sqrt{V(\hat{\mu}(s))}}$$

which will have a constant variance. Here, we will work with the standardised residuals

$$r(s) = \frac{Z(s) - \hat{\mu}(s)}{\sqrt{\hat{\mu}(s)}}$$

that will, under the assumed model (2), have a constant variance ϕ . Assuming stationarity of the spatial correlation, spatial correlation of these residuals can then be modelled for instance with the residual semivariogram, which is a scaled and mirrored version of the autocorrelogram $\rho(h)$ of $e(s)$:

$$\gamma_r(h) = \frac{1}{2}E(r(s) - r(s+h))^2 = \phi(1 - \rho(h)).$$

The advantage of using $r(s)$ over $p(s)$ is that the estimation of the dispersion factor (the sill of $\gamma_r(h)$) is done in the variogram modelling step, along with the parameters that determine the shape of the autocorrelation (or the shape of the variogram: the range and relative nugget).

The variance function (I.2) can then be exploited to obtain variances and covariances of the field $Z(s)$ from $\mu(\cdot)$, ψ and $\rho(h)$:

$$\text{Var}(Z(s)) = \phi\mu(s), \quad \text{Cov}(Z(s), Z(t)) = \phi\sqrt{\mu(s)\mu(t)}\rho(s-t) \quad (\text{I.3})$$

by substituting estimated means $\hat{\mu}(s)$ for $\mu(s)$.

I.3.3 Parameter estimation

Regression coefficients $\beta = (\beta_0, \dots, \beta_p)'$ and correlation parameters $\alpha = (\alpha_1, \dots, \alpha_q)'$ were fitted using a Generalised Estimating Equation (GEE)-like approach (Liang and Zeger, 1986; Diggle et al., 1994). The iterative procedure starts at $i = 1$, and proceeds as follows:

1. estimate α^i given β^{i-1}
2. estimate β^i given α^i

3. on convergence, stop, else go to step 1.

The iteration starts with β^0 obtained from a standard generalised linear model (assuming independent residuals). Convergence was reached if the largest element of $|\beta^i - \beta^{i-1}|$ was smaller than 0.0001, and our model needed 6 iterations to converge.

The actual fitting was done by using the S-Plus library YAGS (Carey, 1998), which we extended with spatial correlation function parameter estimators. Spatial correlation parameters (α) were estimated by fitting permitted models to the standardised residual sample variograms using methods of moments estimators and least squares fitting (Cressie, 1993). Details on estimating β are found in Carey (1998) or Liang and Zeger (1986).

I.3.4 Predicting $Z(s_0)$

Spatial prediction (interpolation) of $Z(s_0)$ at an unobserved location s_0 was done by adding the estimated trend to the predicted residual: $\hat{Z}(s_0) = \hat{\mu}(s_0) + \hat{e}(s_0)$. Given the covariate vector $x(s_0) = (x_0(s_0), \dots, x_p(s_0))$, the trend is estimated as $\hat{\mu} = g^{-1}(\hat{\eta}(s_0)) = \exp(\hat{\eta}(s_0))$, where $\hat{\eta}(s_0) = \sum_{j=0}^p x_j(s_0)\hat{\beta}_j$ and where $\hat{\beta}$ was obtained from the GEE step of Section I.3.3. Next, $e(s_0)$ was predicted by using simple kriging assuming a "known" (i.e., the estimated), non-constant mean function $\hat{\mu}$:

$$\hat{e}(s_0) = \lambda'(Z(s) - \hat{\mu}(s))$$

where the weight vector $\lambda = (\lambda_1, \dots, \lambda_n)'$ was obtained by solving $\Sigma\lambda = c_0$, where element (i, j) of Σ is $\text{Cov}(e(s_i), e(s_j))$, and where $c_0 = (\text{Cov}(e(s_0), e(s_1)), \dots, \text{Cov}(e(s_0), e(s_n)))'$, and with all covariances obtained from the stationary autocorrelogram and dispersion factor through (I.3). For block kriging, the elements in c_0 were replaced by the appropriate integrated covariogram values.

For the simple kriging part of this study, we used gstat (Pebesma and Wesseling, 1998), which was extended for this purpose with a variance function mechanism that allowed simple kriging with non-constant means and mean-dependent covariances.

I.3.5 Interval estimation

Since the predicted values can hardly be believed to accurately predict real densities due to the large areas without observations (Fig. I.1), we need to be able to present the accuracy information by ways of prediction intervals. Because

accuracy information for the trend is available on the log-scale, we used a two-step approach to construct prediction intervals. First, approximate 95% confidence intervals for $\eta(s_0)$ were obtained on the log-scale as

$$[\eta_L(s_0), \eta_U(s_0)] = [\hat{\eta}(s_0) - 2\sigma_\eta(s_0), \hat{\eta}(s_0) + 2\sigma_\eta(s_0)]$$

with $\sigma_\eta(s_0)$ the estimation variance for $\sum_{j=0}^p x_j(s_0)\hat{\beta}_j$. Note that this confidence interval only addresses the error regarding the estimation of β .

When transforming this interval back to the working scale by taking the exponent of both sides, an approximate 95% confidence interval for the mean $\mu(s_0)$ is obtained, denoted as

$$[\mu_L(s_0), \mu_U(s_0)].$$

Starting from this latter interval, to account for the predicted residual $\hat{e}(s_0)$ and its prediction error, $\sigma_e(s_0)$ we

- shifted the interval towards the predicted value $\hat{Z}(s_0)$ by adding $\hat{e}(s_0)$
- stretched the width of the interval with (four times) the kriging prediction error.

Assuming the kriging prediction error to be symmetrically distributed, we approximated 95% prediction intervals by

$$[\mu_L(s_0) + \hat{e}(s_0) - 2\sigma_e(s_0), \mu_U(s_0) + \hat{e}(s_0) + 2\sigma_e(s_0)]. \quad (\text{I.4})$$

I.4 Results

Depth and distance to coast were entered as covariates in (1). The fitted semivariogram for the standardised residuals $r(s)$ for *Fulmarus glacialis* as well as the sea depth map is shown in Fig. I.2. The sample semivariogram was obtained by pooling the variograms of each of the three observation days that completed the Sep/Aug monitoring round of Fig. I.1. It should be noted here that for a number of alternative variogram and regression models, the parameter estimation procedure gave no convergence. The variogram of Fig. I.2 was fitted by fixing the nugget variance to 1, in order to stabilize the estimation procedure.

The approximate 95% prediction intervals for 5 km \times 5 km block average sea bird density obtained by (I.4) are shown in Fig. I.3. Following one of the presentation methods for displaying confidence interval maps by Pebesma

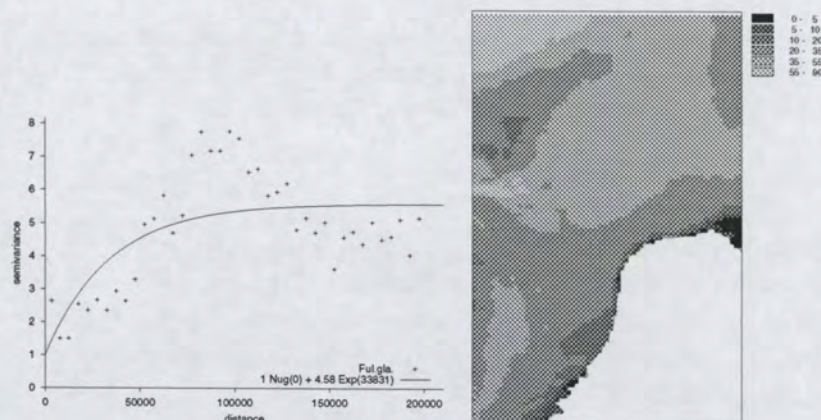


Figure I.2: Left: sample semivariogram of standardised residuals for *Fulmarus glacialis*: + denotes semivariogram estimates, — the fitted model. Right: sea depth map (units: m) of the study area

and De Kwaadsteniet (1997), the position of these intervals is shown relative to four reference density levels. This position is *lower* when the complete confidence interval is below the reference level, *higher* when the interval is completely above the level, or the confidence level can straddle the reference level, in which case the predicted value is *not distinguishable* from the reference level, based on available monitoring data and covariates.

I.5 Discussion

In the current study we interpolated sea bird densities (bird counts divided through the counted surface) from airborne transect monitoring data to predict $5 \text{ km} \times 5 \text{ km}$ block average sea bird densities. In order to prevent end-users from ignoring possibly large inaccuracies, predicted densities were presented in maps as approximate 95% prediction intervals (Fig. I.3).

In the approach used here a method for estimating the trend based on log-linear modelling (1), residual variance that is proportional to the mean response (2) and spatially correlated residuals, is combined with a method for predicting the residuals based on covariances that are proportional to the mean response, a second order stationary autocorrelogram (3) and simple block kriging. These conditions seem to be fairly good in accordance with the observations we had and the processes considered (block mean densities):

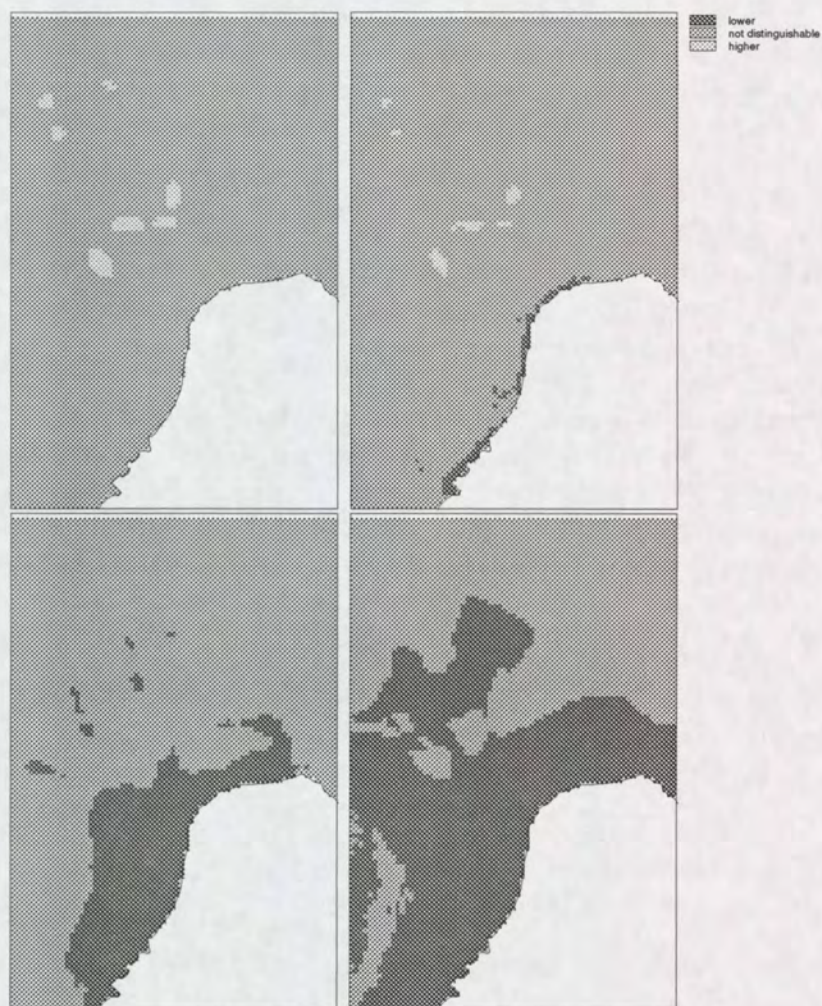


Figure I.3: Map of *Fulmarus glacialis*, period 5, on the Dutch part of the North Sea. 95% prediction intervals for 5 km \times 5 km block mean densities related to the density levels 1 (upper left), 2 (upper right), 4 (lower left) and 7 (lower right) birds/km²

it ensures positive estimates of the trend, and a fairly flexible modelling of the trend function is allowed. The approach results in reasonable estimates (i.e., the trend function) in regions where data are sparse, and in location-specific, data-driven predictions of deviations from the trend function in areas where observations suggest such deviations.

A number of important aspects may have been missed however. From an ecological point of view, the variables that were taken as covariates may not be the preferred ones. Better variables would address the behaviour of the animals more directly, and should for instance be related to feeding (e.g. availability of fish or shell species) and breeding (e.g. distance to breeding colonies). Only very general information (coast aversion, relation of food to water depth) is carried by the covariates currently used (sea depth and distance to coast). Still, non-linear transformation or even more complex functions (higher order polynomials or interactions) might be considered for the two covariates used here.

Issues with respect to measurement error in the data have not been addressed here. These may include systematic errors (birds are more easily missed when small waves are present, when they occur in many small groups, or when light conditions are suboptimal), or random errors (large counts are estimates rather than exact counts). Also, for the interpolation step we assumed that the spatial field did not change between the three flight dates.

On the mathematical side of the problem, several simplifying assumptions were made in this study, of which we need to mention:

- during the estimation of spatial correlation (Sec. 3.3) the residuals are treated as regular observations (correlation resulting from subtracting a common, estimated trend was ignored)
- no other variance functions than (2) have been considered (e.g., the negative binomial would have been a viable alternative)
- the assumptions regarding the distribution of the errors in $\hat{\beta}$ as well as in $\hat{e}(\cdot)$ being normal (or, at least symmetrical)
- the correlation of the predicted residuals with the estimation error in $\hat{\beta}$ has been ignored, and the procedure would therefore not reproduce (predict) the observed values at observation locations
- using $\mu(s_0) + \hat{e}(s_0)$ to predict $Z(s_0)$ may result in negative predicted densities; likewise the interval (4) may straddle zero.

All these issues lead to a potential bias and underestimation of uncertainties. Therefore, the prediction intervals should be interpreted with some care.

No full comparison between regression and spatial correlation model alternatives has been done. We found however that the GEE algorithm did not converge for many models that were slightly more elaborate than the ones we used (for example, fitting an unknown nugget effect or an interaction term of the covariates).

Acknowledgements

The authors acknowledge the helpful discussions with Richard Witte, Cor Berrevoets and Hans Hartholt. Vincent Carey helped with some issues regarding the extension of YAGS. Part of this paper was written while the first author was a visiting scholar at the Department of Geological and Environmental Sciences at Stanford University. The Netherlands Organisation for Scientific Research (NWO) supported this visit with a travel stipend.

References

- Albert, P.S., McShane, L.S. (1995) A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. *Biometrics* 51, pp. 627–638.
- Baptist, H.J.M., Wolf, P.A. (1993) Atlas van de vogels van het Nederlands Continentaal Plat. Rijkswaterstaat, Dienst Getijdewateren. Rapport DGW-98.013 (*in Dutch*)
- V.J. Carey (1998) YAGS – yet another GEE solver.
<http://biosun1.harvard.edu/~carey/index.ssoft.html>)
- Christensen, R., 1991. Linear models for Multivariate, Time Series and Spatial Data. Springer Verlag, New York. 317 pp.
- Cressie, N.A.C. (1993) Statistics for Spatial Data, Revised Edition. Wiley, New York.
- Diggle, P.J., Liang, K-Y., Zeger, S.L. (1994) Analysis of Longitudinal Data. Oxford University Press, Oxford.
- Gotway, C.A., Stroup, W.W. (1997) A Generalized Linear Model Approach to Spatial Data Analysis and Prediction. *Journal of Agricultural, Biological and Environmental Statistics* 2(2), pp. 157–178.

- Liang, K-Y., Zeger, S.L. (1986) Longitudinal Data Analysis using Generalized Linear Models. *Biometrika* 73(1), pp. 13-22.
- McCullagh, P., Nelder, J.A. (1989) *Generalized Linear Models*, Second Edition. Chapman and Hall, London.
- Pebesma, E.J. and De Kwaadsteniet, J.W. (1997), Mapping groundwater quality in the Netherlands. *Journal of Hydrology* 200, 364-386.
- Pebesma, E.J. and Wesseling, C.G. (1998), Gstat, a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences* 24(1), 17-31. <http://www.geog.uu.nl/gstat/>
- Zeger, S.L., Liang, K-Y. (1986) Longitudinal data analysis for discrete and Continuous Outcomes. *Biometrics* 42, pp. 121-130.



INTERUNIVERSITAIR CENTRUM VOOR GEO-ECOLOGISCH ONDERZOEK

Het ICG onderzoekprogramma richt zich op de volgende thema's gebundeld in het onderzoekprogramma "Patterns and Processes in Changing Environments":¹

1. "Dynamics and Palaeorecords of Depositional Environments"
 - a. "Dynamics and evolution of river and coastal systems"
 - b. "Palaeoclimatology and paleoecology of the Quaternary"
2. "Functioning of Landscape Ecosystems"
 - a. "Processes and change in terrestrial ecosystems"
 - b. "Processes and changes in geomorphological systems"

Bij vele onderzoeken wordt gebruik gemaakt van technieken van "Methods, data analysis and modelling". Een deel van de ICG medewerkers houdt zich bezig met ontwikkelen en uitvoeren van deze technieken.

De rapporten die in de ICG-reeks verschijnen worden in zeer beperkte oplage gedrukt en worden verspreid onder deelnemende onderzoeksgroepen en andere belangstellenden. Overname van gegevens en/of citeren is alleen toegestaan na overleg met de auteur(s) en/of leider(s) van het onderzoek. Losse afleveringen zijn verkrijgbaar bij onderstaande personen:

Fysische Geografie en Milieukunde
Universiteit Utrecht

Dr. J.H. van den Berg
(030-2532752)

IBED/Fysische Geografie en Bodemkunde
Universiteit van Amsterdam

Dr. Ir G.B.M. Heuvelink
(020-5257448)

IBED/Palynologie en Paleo/Actuo-ecologie
Universiteit van Amsterdam

Prof dr. H. Hooghiemstra
(020-5257857)

Kwartairgeologie en Geomorfologie
Vrije Universiteit Amsterdam

Dr. C. Kasse
(020-4447381)

Centrum voor Isotopenonderzoek
Rijksuniversiteit Groningen

Afdeling Geografie en Geologie,
Fysische Geografie en Regionale Geografie
Katholieke Universiteit Leuven, België

TOT NU TOE VERSCHENEN:

- 95/1 Zeeberg, J.J., Universiteit Utrecht, Vakgroep Fysische Geografie: *The nature and distribution of Late Pleistocene dunes in the European lowlands and on the Russian platform*
- 95/2 Dinter, M. van, Sorber, A.M. en H.J.A. Berendsen, Universiteit Utrecht, Vakgroep Fysische Geografie: *Inventarisatie van de sedimentatie van zand op de oeverwallen van de Waal en de Gelderse IJssel tijdens het hoogwater van januari en februari 1995*
- 95/3 Sorber, A.M. en G. de Vaan, Universiteit Utrecht, Vakgroep Fysische Geografie: *Ruimtelijke variatie van de sedimentaire structuur en textuur van de bedding van de Grensmaas (stuw Borgharen, km. 15.5 - Maaseik, km. 52.7)*
- 95/4 Hessel, R., Universiteit Utrecht, Vakgroep Fysische Geografie: *Investigation of drought patterns - a case study in Southern Germany*

¹ Research Programme 1999-2003

- 95/5 Meursing, L., Universiteit Utrecht, Vakgroep Fysische Geografie: *De hydraulische ruwheid van doorstroomde vegetatie. Analyse van gepubliceerde model- en prototype metingen.*

1996

- 96/1 Asselman, N.E.M., Universiteit Utrecht, Vakgroep Fysische Geografie: *Suspended sediment concentrations during high discharge events in the river Rhine.*
- 96/2 Makaske, B. en M. Terlien, Universiteit Utrecht, Vakgroep Fysische Geografie: *Le développement géomorphologique de la partie méridionale du Delta intérieur du Niger.*
- 96/3 Van der Wateren-de Hoog, B., Universiteit Utrecht, Vakgroep Fysische Geografie: *Adaptation of a daily weather generator to represent long term precipitation persistence.*
- 96/4 Fehse, J. Universiteit van Amsterdam, Sectie Palynologie en Paleo/Actuo-ecologie: *The Chocó biogeographic region - high levels of biodiversity and endemism threatened in Colombia's Pacific lowland.*
- 96/5 Asselman, N.E.M., Universiteit Utrecht, Vakgroep Fysische Geografie: *Grainsize characteristics used to identify sediment transport pathways on fine grained aggrading floodplains.*
- 96/6 Boer, A. de, Universiteit Utrecht, Vakgroep Fysische Geografie: *(Semi)automatische meetsystemen voor het sedimenttransport in rivieren - literatuurstudie.*
- 96/7 Kleinhans, M.G., Universiteit Utrecht, Vakgroep Fysische Geografie: *Sediment transport in de Nederlandse rijntakken -verwerking metingen 1988-1995 en toetsing transportvergelijkingen.*
- 96/8 Hesselink, A.W., Universiteit Utrecht, Vakgroep Fysische Geografie: *Eilanden en zandbanken in de Rijntakken rond 1850*
- 96/9 Kleinhans, M.G., Universiteit Utrecht, Vakgroep Fysische Geografie: *Sediment transport in the dutch Rhine branches - annual transport and interim sediment budget.*

1997

- 97/1 Asselman, N., van Deursen, W., Kwadijk, J., Middelkoop, H. en C. Wesseling (Universiteit Utrecht, Vakgroep Fysische Geografie) van Dijk, P. en F. Kwaad. Universiteit van Amsterdam, Vakgroep Fysische Geografie en Bodemkunde: *Environmental change and the river Rhine, implications for discharge, sediment supply and water management - progress report 1*
- 97/2 Kabout, J.A.H., Hesselink, A.W. en H.J.A. Berendsen, Universiteit Utrecht, Vakgroep Fysische Geografie: *Inventarisatie van de sedimentatie van zand op de oeverwallen van de Waal tijdens het hoog water van februari en maart 1997.*
- 97/3 Kleinhans, M.G., Universiteit Utrecht, Vakgroep Fysische Geografie: *Sedimenttransport in de Waal: betrouwbaarheidsanalyse en meetstrategie.*
- 97/4 Bruinsma, M. en J.C.J. Kwadijk, Universiteit Utrecht, Vakgroep Fysische Geografie: *Uitbreiding Rhine Flow Model 1902 - 1980*
- 97/5 Storms, J. en J.C.J. Kwadijk, Universiteit Utrecht, Vakgroep Fysische Geografie: *Verandering van de kans op extreme afvoeren 1990 - 2100 voor het UKHI klimaat scenario.*
- 97/6 Hoek, W.Z., Vrije Universiteit, Sectie Kwartairgeologie en Laaglandgenese: *Reference list of Late glacial and Early Holocene pollen diagrams from The Netherlands and adjacent parts of Belgium and Germany.*
- 97/7 Kleinhans, M.G., Universiteit Utrecht, Vakgroep Fysische Geografie: *Sedimenttransport in de Waal: hoogwater Maart 1997.*
- 97/8 Wilbers, A., Universiteit Utrecht, Vakgroep Fysische Geografie: *Duinkarakteristieken en dune tracking tijdens een hoogwater in de Rijntakken.*
- 97/9 Ancker, J.A.M. van den en Jungerius, P.D., Arens Bureau voor Strand- en Duinonderzoek, Universiteit van Amsterdam, Vakgroep Fysische Geografie en Bodemkunde: *Eolische processen langs de Waal, zomer 1997.*

1998

- 98/1 Wateren, B. van der, Universiteit Utrecht, Vakgroep Fysische Geografie: *Een afvoer model gebaseerd op het probability distributed principe.*
- 98/2 Asselman, N.E.M., Universiteit Utrecht, Vakgroep Fysische Geografie: *The impact of climatic change on suspended sediment transport in the river Rhine*
- 98/3 Buma, J., Universiteit Utrecht, Vakgroep Fysische Geografie: *Finding the most suitable slope stability model for the assessment of the impact of climate change on a landslide in South East France*
- 98/4 Buma, J., Universiteit Utrecht, Vakgroep Fysische Geografie: *The impact of climate change on a landslide in South East France, simulated using different GCM-scenarios and downscaling methods for local precipitation.*
- 98/5 Buma, J., Universiteit Utrecht, Vakgroep Fysische Geografie: *Modelling the impact of climate change on a landslide in the Italian Dolomites*
- 98/6 Dijk, P.M. van and Kwaad, F.J.P.M., Universiteit van Amsterdam, Vakgroep Fysische Geografie en Bodemkunde: *Estimation of suspended sediment supply to the stream network of the river Rhine under present-day climate and land use*
- 98/7 Hesselink, A.W., Universiteit Utrecht, Vakgroep Fysische Geografie: *Ontwikkeling van de uiterwaarden langs de Lek. Vanaf de 16e eeuw tot heden*
- 98/8 Imeson, A.C., Cammeraat, L.H. and Bergkamp, G., Universiteit van Amsterdam, Vakgroep Fysische Geografie en Bodemkunde: *Mediterranean Desertification and Land Use. Annual report for 1996*
- 98/9 Blom, J. van, Coppus, R., Dekker, L.C. and Sevink, J., Universiteit van Amsterdam, Vakgroep Fysische Geografie en Bodemkunde: *De bodems van de loofbossen op de oudere duinen en strandwallen van de Nederlandse kust. Profielontwikkeling en bodemverzuring*
- 98/10 Lenders, R., Maren, B. van and Mol, J.-W., Universiteit Utrecht, Vakgroep Fysische Geografie: *Wind-, golf-, en stromingsgeïnduceerd sedimenttransport in kribvakken langs de Waal*
- 98/11 Lev, T., M. van der Perk, A. Gillett, J.P. Absalom, N.M.J. Crout and G. Voigt, Universiteit Utrecht, Vakgroep Fysische Geografie: *GIS-based modelling of radiocaesium transfer to agricultural food products in the Chernobyl region, Ukraine.*
- 98/12 Wilbers, A., Universiteit Utrecht, Vakgroep Fysische Geografie: *Bodemtransport en duinontwikkeling tijdens afvoergolven in de Rijn en Waal.*
- 98/13 Kleinhans, M.G., Universiteit Utrecht, Vakgroep Fysische Geografie: *Kalibratie van de Valbuis Fysische Geografie Utrecht.*
- 98/14 Cohen, K.M., S. Quartel en H.J.A. Berendsen, Universiteit Utrecht, Vakgroep Fysische Geografie: *Zanddikte op de oeverwallen van de Waal (km 900 - 910) een jaar na het hoogwater van 1997.*
- 98/15 Schans, H., Universiteit Utrecht, Vakgroep Fysische Geografie: *Representativiteit van kribvakmetingen uit 1996 en 1997 ten opzichte van de hele Waal.*
- 98/16 Schans, H., Universiteit Utrecht, Vakgroep Fysische Geografie: *Bed level development in Bovenrijn, Pannerdensch Kanaal and the upstream part of the Waal.*
- 98/17 Asselman, N.E.M., Universiteit Utrecht, Vakgroep Fysische Geografie: *Estimation of the sediment load in the lower Rhine basin using sediment rating curves.*
- 98/18 Asselman, N.E.M., Universiteit Utrecht, Vakgroep Fysische Geografie: *The concept of a suite of linked models to simulate sediment transport in the Rhine basin.*
- 98/19 Wilbers, A.W.E., Universiteit Utrecht, Vakgroep Fysische Geografie: *Ruimtelijke variabiliteit van duinkarakteristieken in de Waal tijdens een afvoergolf in 1997.*
- 98/20 Dijk, P.J.M. van, Kwaad F.J.P.M. Universiteit van Amsterdam, Fysische Geografie en Bodemkunde: *The Rhine basin sediment supply model: the quality of morphometric input parameters and snowmelt modelling.*

- 98/21 Hesselink, A.W. Universiteit Utrecht ,Disciplinegroep Geomorfologie en klimaat: *Beschrijving van steekboringen in twee uiterwaarden langs de IJssel en de Waal, Nederland. Data rapport.*

1999

- 99/1 Cohen K.M., H.J.A. Berendsen, Universiteit Utrecht, Faculteit Ruimtelijke Wetenschappen, Fysische Geografie: *Inventarisatie van zand op oeverwallen van de Waal (km 900-917) na het hoogwater van November 1998.*
- 99/2 Tietema A., Universiteit van Amsterdam, Faculteit der Ruimtelijke Wetenschappen, Fysische Geografie en Bodemkunde: *Nitraatuitspoeling in een intrekgebied bestudeerd met een dynamisch GIS.*
- 99/3 Hiemstra J.F. & J.J.M. van der Meer, Universiteit van Amsterdam, Faculteit der Ruimtelijke Wetenschappen, Fysische Geografie en Bodemkunde. *Neogene Glacial History at the Allan Hills, Antarctica – Section Logs*
- 99/4 Moor J.J.W. , Vrije Universiteit Amsterdam, Faculteit der Aardwetenschappen, Kwartairgeologie en Geomorfologie: *Sub-arctic Rivers in Northern Russia. The influence of vegetation, landscape, climate and hydrology on the river morphology of two catchments areas in the Usa Basin in Northeast-European Russia.*
- 99/5 Dijk P.M. van & F.J.P.M. Kwaad, Universiteit van Amsterdam, Faculteit der Ruimtelijke Wetenschappen, Fysische Geografie en Bodemkunde: *The supply of sediment to the river Rhine drainage network*
- 99/6 Kleinhans M.G., Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Sedimenttransport in de Waal: hoogwater november 1998*
- 99/7 Hesselink A.W., Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Opzet digitale inventaris oude rivierkaarten: combineren van verschillende inventarissen van oude rivierkaarten.*
- 99/8 Wilbers A.W.E., M.G. Kleinhans, Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Gevoeligheidsanalyse dune tracking in 2 dimensies.*
- 99/9 Middelkoop, H. & A. Kroon: Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Analyse historische waterstanden Maas -Benedenrivierengebied.*
- 99/10 Wilbers, A., Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Bodemtransport en duinontwikkeling in de Rijntakken: bodempeilingen hoogwater november 1998.*
- 99/11 Linden, S. van der, Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Modelling runoff in arctic river systems: the impact of climate change.*

2000

- 00/1 Tietema, A. & J. Kros, Universiteit van Amsterdam, Faculteit der Natuurwetenschappen, Wiskunde en Informatica, Fysische Geografie: *Modelling critical nitrogen loads and nitrate leaching in Dutch forest ecosystems.*
- 00/2 Tietema, A., Universiteit van Amsterdam, Faculteit der Natuurwetenschappen, Wiskunde en Informatica, Fysische Geografie: *Validatie van gemodelleerde nitraatuitspoeling in bossen.*
- 00/3 Dankers, R., Universiteit Utrecht, Faculteit Ruimtelijke Wetenschappen, Fysische Geografie: *Application of remote sensing in hydrological modelling of sub-arctic environments, a literature review.*
- 00/4 Middelkoop, H., N.E.M. Asselman, H. Buitenveld, M. Haasnoot, J.C.J. Kwadijk, J.A.P.H. Vermulst, W.P.A. van Deursen, P.M. van Dijk, and C. Wesseling, Universiteit Utrecht, Faculteit Ruimtelijke Wetenschappen, Fysische Geografie: *The impact of climate change on the river Rhine and the implications for water management in the Netherlands.*
- 00/5 Kroon, A., S. Vermeer, Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Eolische processen Millinger Duin.*

- 00/6 Lloyd Davies M.T., J.J.M. van der Meer, Universiteit van Amsterdam, Faculteit der Natuurwetenschappen, Wiskunde en Informatica, Fysische Geografie: *Field work carried out in the Allan Hills and Taylor glacier, Taylor Valley, south Victoria Land, Antarctica, 1999-2000.*
- 00/7 Veer J.A. , Vrije Universiteit Amsterdam, Faculteit Aardwetenschappen, Afdeling Kwartairgeologie en Geomorfologie: *Geochemical comparison of Lateglacial lacustrine deposits in the Weerterbos area (southern Netherlands)*
- 00/8 Middelkoop H., B.G. Ruessink, Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Analyse historische waterstanden Maas - Benedenrivierengebied II.*
- 00/9 Putten, M.J. van, Vrije Universiteit Amsterdam, Faculteit Aardwetenschappen, Afdeling Kwartairgeologie en Geomorfologie : *Fluvial response to climatic fluctuations in north eastern European Russia.*
- 00/10 Pebesma, E.J., R.N.M. Duin & A.M.F. Bio, Universiteit Utrecht, Ruimtelijke Wetenschappen, Fysische Geografie: *Spatial interpolation of sea bird densities on the Dutch part of the North Sea.*