Mathematical Statistics
Stockholm University

Factor-Augmented Modeling and Forecasting: regional animal abundance and dynamics

Ying Pang

Research Report 2016:10

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2016:10**,
http://www.math.su.se

Factor-Augmented Modeling and Forecasting: regional animal abundance and dynamics

Ying Pang

April 2016

Abstract

This paper analyzes and forecasts the regional abundance levels of several ungulates species in Kruger National Park, South Africa, which strengthens and extends the understanding about the local population density and dynamics. We employ a multi-level factor model to investigate regional population changes, in which covariation can be represented by common factors and idiosyncrasy can be captured by regional and/or species-specific factors. Additionally, the factors in various levels can be consistently estimated using principal components. Besides, we construct one-step-ahead forecast of populations and obtain the optimal forecasts using a handful of estimated factors as augmented predictors. Furthermore, we evaluate the forecasting accuracy, and conclude that using multi-level factors can lead to substantial improvement of predictive performance for most species of our interest. However, the extent of improvement differs widely across species and regions.

Keywords: Animal Population Dynamics, Factor-augmented Forecasting, Common Factor, Idiosyncratic Factor, Principal Component, Predictive Performance

Factor-Augmented Modeling and Forecasting regional animal abundance and dynamics

Ying Pang *

1 Introduction

In the past several decades, the animal population data sets have become increasingly available in ecological field. Many researchers have studied the source of bias and errors in the data collection strategies to obtain more accurate samples, such as Caughley (1974), McNaughton and Campbell (1991), Viljoen and Retiff (1994), and Redfern et al. (2002). Foremost, the purpose is to analyze the animal population abundance and dynamics. On one hand, some researchers apply the ecological or biological models and mechanisms for explaining population fluctuations under some hypothetical scenarios. For instance, these models can assess the viability of a reintroduced animal population after local extinction, and reverse a declining trend or growth rate in order to conservatively manage a species or community (Newman et al. (2014)). One the other hand, the statistical models and methodology can be employed and developed to meet the demand of wildlife managers or other concerned parties as well. In fact, it is more important to understand the reasons for variation abundance over space and time, which brings in a general question: what could cause the populations increase and/or decrease ?

Since the work of Elton (1924), the intrinsic influence on population changes has been widely investigated, and time series analysis has been extended to model animal population dependence. For instance, Post (2005) explain the spatial variation and dependence in herbivore population dynamics, and Månsson et al. (2007) combine time series models with biological mechanistic modeling, which enable the biological interpretability of statistical inference. Meanwhile, many studies suggest that the developmental processes of populations can be related to extrinsic influence and in-

*Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden, ypang@math.su.se.

teraction, such as vegetation dynamics and climatic changes; however, the potential environmental effects could be qualified and measured in different ways for various purpose of studies. For example, some adopt North Atlantic Oscillation as a measure of global weather phenomenon (Forchhammer et al. (2002); Post and Stenseth (1998)). And, some examine the direct, lagged and probably cumulative effects from atmospheric circulation, rainfall patterns and soil nutrients (Fritz and Duncan (1994); Mason (1996); Mason and Jury (1997); Ogutu and Owen-Smith (2003)). Moreover, a lot of comprehensive studies have demonstrated both intrinsic and extrinsic influence through key-factor analysis, in particular to analyze population dynamics of large-bodied and long-lived species, (Coulson et al. (1997); Forchhammer et al. (1998)).

In this paper, we investigate the regional populations of several ungulate species in Kruger National Park (KNP), South Africa. As aforementioned, the previous studies has shown clear evidence that environmental factors can significantly influence the animal population abundance and dynamics. Rather than directly measuring vegetation or climatic variables, we investigate underlying factors that can represent information concerning the population developments. Furthermore, we believe that the environmental changes can stimulate some similar patterns in population growth or decline over a surveyed area, which can be considered as the commonality and represented by a few factors in aggregate level. In the same time, these extrinsic effects can lead to the local dependence and divergence of species, which can be expressed by a small number of factors in disaggregate level, that are region- and species-specific. Moreover, the interaction within a population can be described by density dependence, which is the self-regulating effect over time in the population or species, and usually it can be modeled by an autoregression (AR).

We present a statistical model, which incorporates multi-level factor components and population density-dependence to represent the extrinsic and intrinsic influence, respectively. Moreover, we demonstrate that the forecasts using factors as augmented predictors can lead to substantial improvements of predictive performance in comparison to a benchmark. In addition, the factors in disaggregate level can make important contributions to explain and predict the population abundance.

The rest of this paper is structured as follows. Section 2 concisely describes KNP census data and presents the models along with estimation procedure. Section 3 introduces the experiment design, and illustrates the results for prediction and evaluation. At last, Section 4 concludes.

2 Materials and methodology

2.1 Data

In this paper, the empirical study uses census data of KNP (Ogutu and Owen-Smith (2003)). The survey area can be divided into Kruger South, Kruger Central and Kruger North (separated as north and far north in our analysis), and assume that the migration between adjacent districts to be negligible, which the display of regions is shown by Figure 1[1].

The KNP aerial census was conducted annually for all large ungulate species, excluding elephant, hippopotamus and buffalo, from 1977 to 1996. The counts were conducted for around four months between April and August in the dry season when the visibility conditions are best. More details and information about the survey can be found at Ogutu and Owen-Smith (2003) and Redfern et al. (2002). Notably, the population of seven species had decreased since 1987, and by 1995 six of them had declined to less than one-quarter of their peak levels. Therefore, our data covers the observations for eleven ungulates in four regions from 1977 to 1994. which tells that the sample size is smaller than the dimension of variables. Furthermore, the eleven species of our interest are: Burchell's zebra, blue wildebeest, waterbuck, warthog, sable antelope, greater kudu, impala, giraffe, tsessebe, eland and roan antelope. Note that the counts of three species (tsessebee, eland and roan antelope) in two regions

Figure 1: Regions display of KNP south, center, north and far north (The census area is green-colored)

[1] The source of image is from `http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0075-64582008000100017`.

(south and center) are not included in data, because the numbers are too few and merged with neighbors. In addition, the observation of impala in south at 1979 is considered to be biologically unrealistic, and thus treated as a missing value (Månsson et al. (2007)). And, the median value of proceeding and following two observations is used as a substitute for the original one. Moreover, as a general overview, Figure A.1 illustrates the developments of populations in logarithms.

2.2 Models

Let $P_{r,s,t}$ be a population of region r and species s at current time t. And we shall use P_t for the brevity in the following analysis, thus all the variables and coefficients will implicitly depend on s and r. Moreover, we start with a general population model for prediction,

$$P_{t+1} = P_t \cdot \exp\left\{ A' M_t + \sum_{i=1}^{q} \beta_i \ln P_{t-i+1} \right\} \ , \tag{1}$$

where M_t is a vector of factors in various levels, A is a coefficient vector for factors, β_i is a coefficient for lagged variable in logarithm, and q is a finite lag order. Furthermore, the factors can be obtained based on the regional population changes rates, and the estimation method will be discussed later.

Given a biological reason about the multiplicative nature of population growth, we consider taking the nature logarithm on both sides of equation (1). Thus, letting $y_t = \ln P_t$, we obtain that

$$y_{t+1} = A' M_t + (1+\beta_1) y_t + \sum_{i=2}^{q} \beta_i y_{t-i+1} + \varepsilon_{t+1} \ , \tag{2}$$

where ε_{t+1} describes the remaining variance that cannot be included in the deterministic part of the model (2). More concisely, we rewrite model (2) by

$$y_{t+1} = A' M_t + B' W_t + \varepsilon_{t+1} \ , \tag{3}$$

where $W_t = (y_t, y_{t-1}, \ldots, y_{t-q+1})'$ and $B = (1+\beta_1, \beta_2, \ldots, \beta_q)'$. In addition, Figure 2 illustrates forecast model (3), which the future abundance level depends on the current effect of factor components and the direct and delayed density-dependence, plus a remaining variance.

The variable y_{t+1} can be considered as the one-step-ahead forecast. Additionally, in order to fit model (3), we substitute factor estimate $\widehat{M}_t$ for M_t, simply because true factors are latent and cannot be observed. Given that data available up to time T, the ordinary least squares (OLS) estimates $\widehat{A}$ and

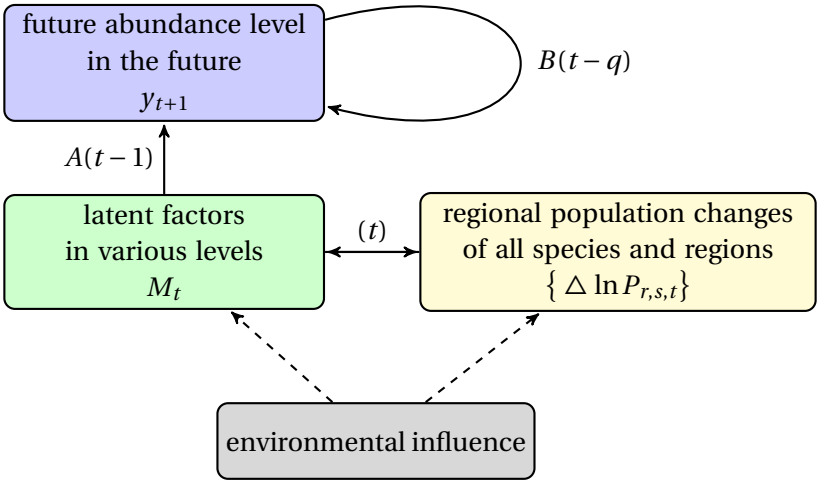


Figure 2: A model description. Coefficients B and A with arrows describe the intrinsic and extrinsic effect on the animal abundance level, respectively. The time interval given in brackets represents the primary time of interaction relative to t. The factors in various levels can be obtained by analyzing population change rates of all potential species and regions in the whole surveyed area. The dash arrows indicate the relations between which are not considered in our model.

$\widehat{B}$ can be produced by regressing y_{t+1} onto $\widehat{M}_t$ and W_t for all available t. Furthermore, the one-step-ahead forecast estimate $\hat{y}_{T+1|T}$ can be constructed straightforwardly by

$$\hat{y}_{T+1|T} = \widehat{A}'\widehat{M}_T + \widehat{B}'W_T \ . \tag{4}$$

2.3 Factor representation

2.3.1 Factor models

We define the population growth rates by

$$x_{r,s,t} = \ln\left(P_{r,s,t}/P_{r,s,t-1}\right) \ , \tag{5}$$

for $t = 1,\ldots,T$, $r = 1,\ldots,R$, $s = 1,\ldots,S_r$, where T is the sample size, R is the number of regions, and S_r is the number of species recorded in region r. In addition, let S be the number of species observed in the whole studied area, which indicates $S \geq S_r$. Then, we denote $X_t^r = (x_{r,1,t},\ldots,x_{r,s,t},\ldots,x_{r,S,t})'$, in which $x_{r,s,t} = 0$ if the species s is not observed in region r. Furthermore, we

obtain an N-dimensional vector $X_t = (X_t^{1\prime}, \ldots, X_t^{r\prime}, \ldots, X_t^{R\prime})'$ where $N = R \cdot S$. When the sample size T is smaller than the dimension of variables N, that is a case of high-dimensional framework.

A classic method to represent the covariation of a large number of observables is to extract a handful of factors that can capture common movement of data. Stock and Watson (2002) propose a factor model as follows,

$$X_t = \Lambda F_t + \mathbf{e}_t \ , \tag{6}$$

where F_t is an n-dimensional vector of factors common to all units and $\mathbf{e}_t$ is the idiosyncratic term. Note that "idiosyncratic" is a macroeconomic saying that refers to individual or characteristic features in this paper. Furthermore, Stock and Watson (2002) derive the consistent estimates of factors based on model (6), and successfully improve the predictive performance using these factor estimates as augmented predictors. In addition, model (6) is widely employed and developed for high-dimensional data, particularly in the macroeconomic application which often requires dimension reduction techniques.

We consider the situation that data X_t is available in disaggregate levels. For instance, Beck et al. (2009) and Beck et al. (2011) analyze the sectoral and/or regional inflation dynamics rather than in national level. Moreover, besides the common factor F_t, we also investigate factors in more specific levels. Therefore, we decompose the elements of $\mathbf{e}_t$ as follows,

$$e_{r,s,t} = \theta_{r,s} G_t^r + \gamma_{r,s} Q_t^s + u_{r,s,t} \ , \tag{7}$$

where G_t^r includes n_r regional factors that only influence variables in region r, Q_t^s includes n_s species-specific factors that only effect variables for species s, and $u_{r,s,t}$ is the remaining disturbance.

Combing the expression (7) and (6), we can represent a multi-level factor model,

$$x_{r,s,t} = \lambda_{r,s} F_t + \theta_{r,s} G_t^r + \gamma_{r,s} Q_t^s + u_{r,s,t} \ , \tag{8}$$

where $\lambda_{r,s}$, $\theta_{r,s}$ and $\gamma_{r,s}$ are values or row vectors of the factor loadings. Additionally, $\lambda_{r,s} F_t$ expresses the common or aggregate component, meanwhile, $\theta_{r,s} G_t^r$ and $\gamma_{r,s} Q_t^s$ represent regional and species-specific components, respectively. Moreover, we can rewrite (8) in a matrix form as

$$X_t = \Lambda F_t + \Theta G_t + \Gamma Q_t + \mathbf{u}_t \ , \tag{9}$$

where $G_t = (G_t^{1\prime}, \ldots, G_t^{R\prime})'$, $Q_t = (Q_t^{1\prime}, \ldots, Q_t^{S\prime})'$, and the matrices of factor loadings are with entries given by

$$\Lambda_{i,j} = \begin{cases} \lambda_{r,s} & i = r; j = s \ , \\ 0 & \text{otherwise} \ , \end{cases} \quad \Theta_{i,j} = \begin{cases} \theta_{r,s} & i = r; j = s \ , \\ 0 & \text{otherwise} \ , \end{cases} \quad \Gamma_{i,j} = \begin{cases} \gamma_{r,s} & i = r; j = s \ , \\ 0 & \text{otherwise} \ , \end{cases}$$

for all r and s.

Furthermore, we clarify model (9) by giving a simple example. Suppose there are two regions and three species in a surveyed area, and only two species can be observed in the first region. which means $R = 2$, $S_1 = 2$ and $S_2 = S = 3$. Thus, model (9) is specified as

$$\begin{bmatrix} x_{1,1,t} \\ x_{1,2,t} \\ x_{1,3,t} \\ x_{2,1,t} \\ x_{2,2,t} \\ x_{2,3,t} \end{bmatrix} = \begin{bmatrix} \lambda_{1,1} \\ \lambda_{1,2} \\ \lambda_{1,3} \\ \lambda_{2,1} \\ \lambda_{2,2} \\ \lambda_{2,3} \end{bmatrix} F_t + \begin{bmatrix} \theta_{1,1} & 0 \\ \theta_{1,2} & 0 \\ \theta_{1,3} & 0 \\ 0 & \theta_{2,1} \\ 0 & \theta_{2,2} \\ 0 & \theta_{2,3} \end{bmatrix} \begin{bmatrix} {G_t}^1 \\ {G_t}^2 \end{bmatrix} + \begin{bmatrix} \gamma_{1,1} & 0 & 0 \\ 0 & \gamma_{1,2} & 0 \\ 0 & 0 & \gamma_{1,3} \\ \gamma_{2,1} & 0 & 0 \\ 0 & \gamma_{2,2} & 0 \\ 0 & 0 & \gamma_{2,3} \end{bmatrix} \begin{bmatrix} {Q_t}^1 \\ {Q_t}^2 \\ {Q_t}^3 \end{bmatrix} + \begin{bmatrix} u_{1,1,t} \\ u_{1,2,t} \\ u_{1,3,t} \\ u_{2,1,t} \\ u_{2,2,t} \\ u_{2,3,t} \end{bmatrix}$$

where $x_{1,3,t} = \lambda_{1,3} = \theta_{1,3} = \gamma_{1,3} = u_{1,3,t} = 0$.

2.3.2 Factor estimates

We adapt the model assumptions of Stock and Watson (2002) for our use. Therefore, assume that the error terms $\mathbf{u}_t$ are allowed to have limited serial correlation and weakly cross-sectional correlation in model (9). Besides, F_t, G_t, Q_t and $\mathbf{u}_t$ are assumed to be standard normally distributed and mutually independent.

We utilize the methodology developed by Beck et al. (2011) to deal with the complicated structure of multi-level factor model (9), and this employs principal component analysis (PCA) together with maximum-likelihood estimation. In addition, let k, k_r and k_s be the numbers of estimated factors for F_t, G_t^r and Q_t^s, respectively. Then, in order to obtain factor estimates, we consider minimizing the following objective function,

$$V(F,\Lambda) = (NT)^{-1} \sum_{t=1}^{T} (X_t - \Lambda F_t)'(X_t - \Lambda F_t) \ , \tag{10}$$

subject to the normalization $F'F/T = I_n$, where $F = (F_1, \ldots, F_T)'$ and I_n is an $n \times n$ identity matrix. By concentrating out $\widehat{\Lambda}$, minimizing objective function (10) is equivalent to maximizing $\text{Trace}\{F'[XX'/(NT)]F\}$ subject to $F'F/T = I_n$. Therefore, a theoretical factor estimate $\tilde{F}$ is given by $\tilde{F} = \sqrt{T}V$ where V consists of eigenvectors corresponding to n eigenvalues of covariance matrix $XX'/(NT)$ in decreasing order. Then, we reduce the dimension from n to k and obtain $\widehat{F}_t = (\tilde{F}_{1t}, \ldots, \tilde{F}_{kt})$. Moreover, the ith eigenvector in V is the ith principal component of data X_t, and thus $\widehat{F}_t$ is called principal-component-based factor estimate. In addition, $\widehat{F}_t$ is proved to be consistent when N grows much faster than T or $\sqrt{T}/N \to 0$ (Stock and Watson (2002), Bai and Ng (2002, 2013)).

During the estimation procedure for disaggregate factors, we do not want either regional or species-specific components mixed with the common factors, which means both G_t and Q_t are uncorrelated with F_t. Therefore, we use the estimated residuals instead of X_t as the foundation to form principal components. Firstly, by linearly regressing X_t onto $\widehat{F}_t$ for all t, we can obtain OLS estimate $\widehat{A}$ and resulting residuals $\{x_{r,s,t} - \hat{\lambda}_{r,s}\widehat{F}_t\}_{r=1,s=1,t=1}^{R,S,T}$. For each r, regional factor estimate $\widehat{G}_t^r$ can be produced using the first k_r principal components of the S_r residual variables $\{x_{r,s,t} - \hat{\lambda}_{r,s}\widehat{F}_t\}_{s=1,t=1}^{S,T}$. Secondly, by regressing X_t onto F_t and G_t for all t, we can have the OLS estimates $\widehat{A}$ and $\widehat{\Theta}$, and then calculate residuals $\{x_{r,s,t} - \hat{\lambda}_{r,s}\widehat{F}_t - \hat{\theta}_{r,s}\widehat{G}_t^r\}_{r=1,s=1,t=1}^{R,S,T}$. Regarding every species s, $\widehat{Q}_t$ can be estimated by the first k_s principal components of the $\sum_{r=1}^{R}\mathbf{1}(r_s)$ residual variables $\{x_{r,s,t} - \hat{\lambda}_{r,s}\widehat{F}_t - \hat{\theta}_{r,s}\widehat{G}_t^r\}_{r=1,t=1}^{R,T}$, where $\mathbf{1}(r_s)$ denotes a dummy variable that equals one if the species s can be observed in the region r, and equals zero otherwise.

This estimation method requires S_r to be large, otherwise $\widehat{G}_r$ and $\widehat{Q}_s$ are correlated. In order to get rid of the possible correlation between factors, we follow the approach of Beck et al. (2011) to update regional and species-specific factor estimates with the following iterative algorithm.

1. Use the estimates indicated in the previous paragraph to set initial values $\widehat{G}_t{}^{(0)}$ and $\widehat{Q}_t{}^{(0)}$.

2. Iterate for $i = 1, 2, \ldots$

   (a) Estimate the residuals $x_{r,s,t} - \hat{\lambda}_{r,s}\widehat{F}_t - \hat{\gamma}_{r,s}\widehat{Q}_t^{s(i-1)}$, where the OLS estimates for coefficients are obtained by regressing $x_{r,s,t}$ onto $\widehat{F}_t$ and $\widehat{Q}_t^{s(i-1)}$.

   (b) Obtain $\widehat{G}_t{}^{(i)}$, in which $\widehat{G}_t^{r(i)}$ consists of the first k_r principal components of residual variables from last step for all t and s.

   (c) Calculate the residuals $x_{r,s,t} - \hat{\lambda}_{r,s}\widehat{F}_t - \hat{\theta}_{r,s}\widehat{G}_t^{r(i)}$, where the OLS estimates for coefficients are obtained by regressing $x_{r,s,t}$ onto $\widehat{F}_t$ and $\widehat{G}_t^{r(i)}$.

   (d) Construct $\widehat{Q}_t{}^{(i)}$, in which $\widehat{Q}_t^{s(i)}$ includes the first k_s principal components of residual variables from last step for all t and r.

   (e) Check conditions for all t

   - $\max_r\{\max_t|\widehat{G}_t^{r(i)} - \widehat{G}_t^{r(i-1)}|\} < CL$;
   - $\max_s\{\max_t|\widehat{Q}_t^{s(i)} - \widehat{Q}_t^{s(i-1)}|\} < CL$,

   where CL is a criteria level. If both conditions are satisfied, break the loop and jump to step 3, otherwise let $i = i + 1$ and continue.

3. Obtain the modified estimations $\widehat{G}_t = \widehat{G}_t{}^{(i)}$ and $\widehat{Q}_t = \widehat{Q}_t{}^{(i)}$.

3 Results

3.1 Simulation studies

3.1.1 Experimental design

We consider a Monte Carlo experiment to assess predictive performance of our forecast model. The data of predictors can be simulated based on multilevel factor model (8), recalling that

$$x_{r,s,t} = \lambda_{r,s} F_t + \theta_{r,s} G_t^r + \gamma_{r,s} Q_t^s + u_{r,s,t} \ ,$$

where

$$\begin{cases} u_{r,s,t} = a u_{r,s,t-1} + (1+b^2)\nu_{r,s,t} + b\nu_{r,s-1,t} + b\nu_{r,s+1,t} \\ \nu_{r,s,t} = \sigma_{r,s,t}\eta_{r,s,t} \\ {\sigma_{r,s,t}}^2 = \rho_0 + \rho_1 {\nu_{r,s,t-1}}^2 + \rho_2 {\sigma_{r,s,t-1}}^2. \end{cases} \tag{11}$$

Factor loadings $\lambda_{r,s}$, $\theta_{r,s}$, and $\gamma_{r,s}$ are drawn from a uniform distribution on $[0.1, 0.8]$. And, the processes ζ_t^F, ζ_t^G and ζ_t^Q are all from a standard normal family and mutually independent. Furthermore, the disturbance $\eta_{r,s,t}$ is a standard normal deviate, and independent of factors. In addition, in the equation system (11), errors $u_{r,s,t}$ are serially correlated with an AR coefficient a and cross-sectionally correlated with a moving average coefficient b. Moreover, the innovation $\nu_{r,s,t}$ where the series $\nu_{r,s,t}$ follows an autoregressive conditional heteroscedastic process. Note that we can have a simple case $u_{r,s,t} \overset{iid}{\sim} N(0,1)$, given $a = b = \rho_1 = \rho_2 = 0$ and $\rho_0 = 1$.

The scalar variable to be forecast is simulated by

$$y_{t+1} = A' M_t + \varepsilon_{t+1} \ ,$$

where $M_t = (F_t', G_t^{r\prime}, Q_t^{s\prime})'$. The coefficient matrix A has entries drawn from a uniform distribution on $[0.1, 0.8]$ and errors $\varepsilon_t \overset{iid}{\sim} 0.1 \cdot N(0,1)$. Furthermore, there is no lagged variable involved in simulation procedure, which results in the disappearance of W_t in the model (3).

3.1.2 Simulation evaluation

The evaluation to factor and forecast estimation are summarized by three statistics, in which the first two are suggested by Stock and Watson (2002). Firstly, $R^2_{\hat{F},F}$ is to examine the estimation $\hat{F}$, which is computed by

$$R^2_{\hat{F},F} = \frac{E\left[\text{Trace}\left(\widehat{F}' P_F \widehat{F}\right)\right]}{E\left[\text{Trace}\left(\widehat{F}'\widehat{F}\right)\right]} \ , \tag{12}$$

where $P_F = F(F'F)^{-1}F'$, and the expectation averages the relevant statistic over the Monte Carlo repetitions. When the factor estimates converge to true factors, it indicates that $R^2_{\widehat{F},F}$ converges to one in probability. Moreover, $R^2_{\widehat{G},G}$ and $R^2_{\widehat{Q},Q}$ are defined in a similar way.

The second one is to examine predictive performance, assuming true factors are known. The statistic $S^2_{\hat{y},\tilde{y}}$ measures how close the forecast estimate $\hat{y}_{t+1|t}$ is to theoretical forecast $\tilde{y}_{t+1|t}$, which is calculated by

$$S^2_{\hat{y},\tilde{y}} = 1 - \frac{E\left(\hat{y}_{t+1|t} - \tilde{y}_{t+1|t}\right)^2}{E\,\hat{y}_{t+1|t}{}^2} \quad . \tag{13}$$

Given the historic information available up to time T, the forecast estimate $\hat{y}_{T+1|T}$ is obtained by equation (4), recalling that

$$\hat{y}_{T+1|T} = \widehat{A}'\widehat{M}_T + \widehat{B}'W_T \quad ,$$

where the coefficient estimates are obtained by fitting forecast model (3) in which M_t is replaced by $\widehat{M}_t = (\widehat{F}_t', \widehat{G}_t^{r\prime}, \widehat{Q}_t^{s\prime})'$ for $t = 1, \ldots, T$. Furthermore, the theoretical forecast is constructed by

$$\tilde{y}_{T+1|T} = \widetilde{A}'M_T + \widetilde{B}'W_T \quad , \tag{14}$$

where $\widetilde{A}$ and $\widetilde{B}$ are the OLS estimates obtained by fitting model (3) with generated factors M_t for $t = 1, \ldots, T$. When numbers of estimated factors are equal to numbers of true factors, it indicates that $\hat{y}_{T+1|T} - \tilde{y}_{T+1|T} \xrightarrow{p} 0$, and thus $S^2_{\hat{y},\tilde{y}}$ should be close to one when T and N are large jointly (Stock and Watson (2002)).

The third statistic is the mean squares forecasting error (MSFE) which is also used to examine the forecast performance. It compares the forecast estimate $\hat{y}_{T+1|T}$ with real observation y_{T+1} over replications,

$$\text{MSFE} = \frac{1}{MC}\sum_{j=1}^{MC}\left(\hat{y}^{[j]}_{T+1|T} - y^{[j]}_{T+1}\right)^2 \quad , \tag{15}$$

where MC is the number of Monte Carlo replications. And, the smaller value of MSFE is, the better forecast model performs.

3.1.3 Results of simulation

The free parameters of experiment are T, R, S, n, n_r, n_s, a, b, ρ_0, ρ_1 and ρ_2. For each Monte Carlo replication, we construct the factor estimates $\widehat{F}$, $\widehat{G}$, $\widehat{Q}$, and forecast estimate $\hat{y}_{T+1|T}$ and $\tilde{y}_{T+1|T}$, specified k, k_r, k_s and q.

The statistics $R^2_{\widehat{F},F}$, $R^2_{\widehat{G},G}$, $R^2_{\widehat{Q},Q}$ and $S^2_{\hat{y},\tilde{y}}$ are computed given the numbers of estimated factors equals true factors', i.e. $k = n$, $k_r = n_r$ and $k_s = n_s$. In addition, the values of MSFE are calculated using factor estimates obtained under two scenarios: (i) let $k = n$, $k_r = n_r$ and $k_s = n_s$; (ii) select k, k_r and k_s over $0 \le k \le n$, $0 \le k_r \le n_r$ and $0 \le k_s \le n_s$. Furthermore, the values of MSFE produced by AR forecasts are reported as benchmark for comparison, where the order of lagged variables q can be determined by the Bayes information criterion.

The results over 1000 Monte Carlo replications are shown in Table 1. We consider the scenario for independent factors and errors, and examine how the predictive performance changes with respect to the sample size, the dimension of variables and the number of factors. Panel A and B contain results of small and large samples, respectively. On one hand, there are two facts shared for both small and large samples. Firstly, the values of MSFE.i and MSFE.ii are much smaller than the benchmark (MSFE.ar), which indicates that the forecasts using multi-level factors can perform strikingly better than the AR forecasts. Secondly, we find that all the values of MSFE.ii are smaller than MSFE.i, which implies that it is not necessary to use as many factors as possible for predictions. In other words, the predictive performance can be improved by using a relatively small number of estimated factors as augmented predictors. On the other hand, there are lots of difference between the behavior of small and large samples. From panel A, the deterioration of both factor and forecast estimates can be detected when N and T are smaller than 50 . Additionally, panel B suggests that our forecasts perform quite well, which results from the fact that most of $S^2_{\hat{y},\tilde{y}}$ are over 0.95. Furthermore, as T and N jointly grow, the statistics for factor examination become closer to 1, in which most of them exceed 0.9; however, a slight deterioration can be found when the numbers of generated factors are relatively large, particularly in the examination for common factors.

Besides, we investigate how the dependence of errors can influence the predictive performance. Pre-fixed the sample size and dimension ($T = 100$ and $N = 500$), panel C shows the results when errors are assumed to be dependent, which errors can be simulated by model (11). After introducing serial and cross-sectional correlation in the errors, there is little influence to the factor estimates, although the common factor estimates become less accurate. In addition, the values of MSFE.ii are slightly smaller than the ones of the independent case, and thus the predictive performance has not changed much.

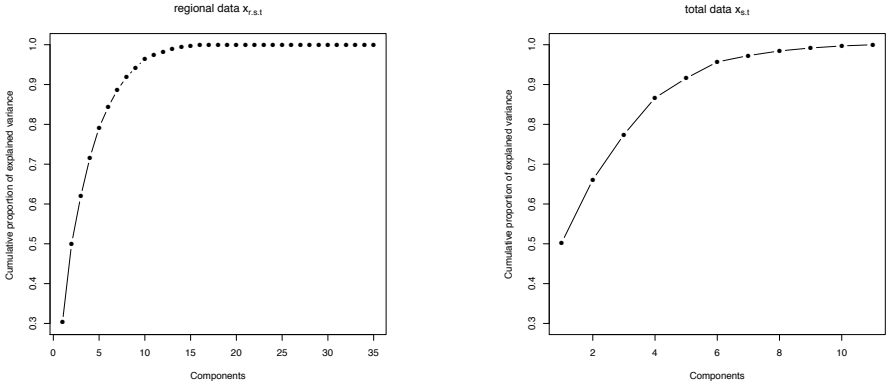
Table 1: Simulation results over 1000 Monte Carlo replications

Free Parameters											Factor Examination			Forecast Evaluation			
T	R	S	n	n_r	n_s	a	b	ρ_0	ρ_1	ρ_2	$R^2_{\hat{F},F}$	$R^2_{\hat{G},G}$	$R^2_{\hat{Q},Q}$	$S^2_{\hat{y},\bar{y}}$	MSFE.i	MSFE.ii	MSFE.ar
A. INDEPENDENT FACTORS AND ERRORS: SMALL T AND N ($T, N \leq 50$)																	
25	5	5	2	2	2	0	0	1	0	0	0·6106	0·6960	0·7027	0·7195	0·3496	0·0972	1·0347
25	5	10	5	2	2	0	0	1	0	0	0·6790	0·7232	0·7491	0·7363	0·3867	0·0984	1·0094
50	5	10	5	2	2	0	0	1	0	0	0·6967	0·7453	0·8034	0·8666	0·2887	0·0979	1·0454
B. INDEPENDENT FACTORS AND ERRORS: LARGE T AND N ($T, N \geq 100$)																	
100	10	10	5	2	2	0	0	1	0	0	0·8651	0·8126	0·8178	0·9394	0·2102	0·0670	0·9586
100	10	25	5	2	2	0	0	1	0	0	0·9312	0·8777	0·9185	0·9491	0·2061	0·0675	0·9741
100	20	25	5	2	2	0	0	1	0	0	0·9630	0·9262	0·9383	0·9653	0·2035	0·0688	1·0199
100	20	25	5	5	5	0	0	1	0	0	0·8242	1·0000	0·8023	0·9075	0·2399	0·0541	0·9743
250	10	10	5	2	2	0	0	1	0	0	0·9570	0·8456	0·8480	0·9524	0·2253	0·0748	1·0206
250	10	25	5	2	2	0	0	1	0	0	0·9514	0·9158	0·9121	0·9727	0·1857	0·0661	1·0408
250	20	25	5	2	2	0	0	1	0	0	0·9757	0·9406	0·9416	0·9781	0·1803	0·0633	0·9415
250	20	25	5	5	5	0	0	1	0	0	0·9006	0·8868	0·9087	0·9592	0·1946	0·0617	1·0448
500	10	10	5	2	2	0	0	1	0	0	0·9602	0·8596	0·8617	0·9624	0·2112	0·0657	1·0159
500	10	25	5	2	2	0	0	1	0	0	0·9765	0·9315	0·9172	0·9761	0·2029	0·0693	1·0432
500	20	25	5	2	2	0	0	1	0	0	0·9793	0·9521	0·9498	0·9853	0·1842	0·0683	1·0714
500	20	25	5	5	5	0	0	1	0	0	0·9233	0·8779	0·8879	0·9653	0·1808	0·0681	0·9932
C. DEPENDENT FACTORS AND ERRORS: LARGE T AND N ($T = 100$ AND $N = 500$)																	
100	20	25	5	5	5	0·5	0	1·00	0	0	0·8252	1·0000	0·9391	0·9016	0·2216	0·0448	0·9495
100	20	25	5	5	5	0·9	0	1·00	0	0	0·8212	1·0000	0·9873	0·9078	0·2388	0·0551	1·0287
100	20	25	5	5	5	0·9	0	0·05	0·05	0·9	0·8255	1·0000	0·9682	0·9099	0·2264	0·0535	1·0158
100	20	25	5	5	5	0	1	1·00	0	0	0·8261	1·0000	0·8691	0·9063	0·2229	0·0453	0·9689
100	20	25	5	5	5	0	1	0·05	0·05	0·9	0·8217	1·0000	0·9339	0·9076	0·2205	0·0525	1·0052
100	20	25	5	5	5	0·9	1	0·05	0·05	0·9	0·8254	1·0000	0·8910	0·8992	0·2258	0·0533	0·9106

3.2 Empirical studies

3.2.1 Preliminary analysis

Let south, center, north and far north be numbered by $r = 1,2,3,4$, respectively. For our data, $S_r = 8$ for $r = 1,2$, $S_r = 11$ for $r = 3,4$ and time span from 1977 to 1994, which indicates $R = 4$, $S = 11$, $N = 44$ and $T = 18$. Next, we consider two sets of X_t: one is regional data $\{x_{r,s,t}\}$ and the other is the total data $\{x_{s,t}\}$, where $x_{s,t} = \ln\left(\sum_r P_{r,s,t} \big/ \sum_r P_{r,s,t-1}\right)$. After scaling variables to have zero mean and unit standard deviation, we perform PCA for a overview. In addition, Figure 3 illustrates the cumulative proportion of variation explained by each principal component regarding regional and total data. And, we can see that the first two principal components can explain approximately 50.4 % and 66.1 % of the variation of regional and total data, respectively.



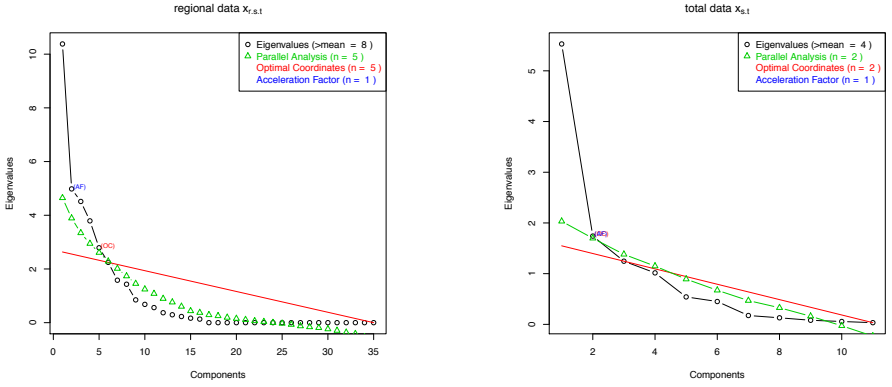
(a) regional population growth rates

(b) total population growth rates

Figure 3: Cumulative proportion of variation explained by the principal components

In theory, we lose the least variance by dropping the last component, simply because the last principal component explains the smallest part of the variance, which has the smallest eigenvalue of the covariance matrix. And, that is how PCA works for dimension reduction. The important question is how many of the last components should be dropped without losing too much information. In other words, how many principal components need to remain in order to resemble the commonality appropriately. There are many different methods to determine the number of n. Moreover, Figure 4 gives the answers to appropriate values of n provided by the Kaiser rule,

parallel analysis[2], and the Cattell subjective scree test which both optimal coordinate and acceleration factor in details can be found in Cattell (1966). Based on the regional data, we obtain that $n_{Kaiser} = 8$, $n_{parallel} = 5$, meanwhile the scree test of the optimal coordinate and acceleration factor gives $n = 5$ and $n = 1$ respectively. Therefore, we determine $k = 5$ as the number of estimated common factors at most in the following model fit. Besides, using total data can provide a smaller number of common factors. Actually, it is not surprising that a relatively small number of principal components can capture a large amount of common movement for the total population change rates.



(a) regional population growth rates

(b) total population growth rates

Figure 4: Solutions to the number of principal components to retain by the Kaiser rule (black), parallel analysis (green), Cattell subjective scree test: optimal coordinate (red) and acceleration factor (blue)

3.2.2 Empirical evaluation

The computation of MSFE does not require any knowledge of true factors, thus it can be modified and adopted in empirical study. And, MSFE is an average of forecasting errors over a time period from T_0 to T_1,

$$\text{MSFE} = \frac{1}{T_1 - T_0 + 1} \sum_{T=T_0-1}^{T_1-1} \left(\hat{y}_{T+1|t} - y_{T+1}\right)^2 \ . \tag{16}$$

[2] The Kaiser rule and parallel analysis are performed as classical methods to determine n. Let $n_{Kaiser} = \sum_i \mathbf{1}_{\{\lambda_i \geq \bar{\lambda}\}}$ and $n_{parallel} = \sum_i \mathbf{1}_{\{\lambda_i \geq LS\}}$, where λ_i is the ith eigenvalue of covariance matrix, and $\bar{\lambda}$ is the mean of eigenvalues and LS is a location statistics, for instance the 95th percentile.

And, relative MSFE (RMSFE) is calculated by

$$\text{RMSFE} = \text{MSFE}/\ \text{MSFE}_{[0]}\ , \tag{17}$$

where the benchmark $\text{MSFE}_{[0]}$ is an AR forecast as benchmark. Additionally, a value of RMSFE smaller than one indicates that our forecast model performs better than AR(q) model.

Specified the numbers of k, k_r, k_s and q, we estimate factors based on regional data $\{x_{r,s,t}\}$, and then, using multi-level factor estimates, construct one-year-ahead forecasts for both regional and total populations in logarithms, which are $y_{t+1}^{reg} = \ln P_{r,s,t+1}$ and $y_{t+1}^{tot} = \ln\left(\sum_r P_{r,s,t+1}\right)$. For each populations, we can construct the forecasts by equation (4) over a forecasting time period, and calculate MSFE using formula (16). Moreover, with different combinations of cross section $k = 1,\ldots,6$, $k_r = 0,1,2$, $k_s = 0,1,2$, and $q = 0,\ldots,4$, we can obtain the minimum MSFE that results from the optimal forecast, and report its RMSFE along with the selected numbers of factors and lag order. Furthermore, we report the minimum

3.2.3 Empirical results

At first, the length of forecasting period need to be determined. In other words, suppose the forecast ends at 1994, and then the choice of T_0 should be addressed. Our data has a time span of 18 years, which is a quite small sample, and it results in the predictive performance being sensitive to the choice of T_0. In theory, a larger initial sample includes more historical information and thus produces the better factor and forecast estimates. However, it shortens the forecasting time period, which results in fewer forecast errors to be averaged and thus the results cannot be trustworthy. Consequently, regarding various T_0, Table B.1 summarizes the values of MSFE from the optimal forecasts for total populations. It is clear that, for most species, the ranges of MSFE values are relatively large, which indicates the results are sensitive to T_0. Additionally, it is seen that the values of MSFE become smaller by the increase in T_0, however, the trends of decrease in MSFE values are not monotonous. Therefore, it is a trade-off in compromised handling to forecast the populations for last eight years and use the previous observations as initial samples, which means the forecasting time period starts at 1988 and ends at 1994.

Table 2 presents the results of optimal forecasts regarding total populations (in logarithms), together with the AR forecasts as benchmark. There are nine species of which factor-augmented forecasting performs better than benchmark because of RMSFE smaller than one; however, the improvement

Table 2: The optimal one-year-ahead out-of-sample forecasts: total populations

	The optimal forecasts					Benchmark
	k	k_r	k_s	q	RMSFE	root $\text{MSFE}_{[0]}$
zebra	2	-	-	-	1.3925	0.29
wildebeest	4	-	1	-	0.5658	0.28
waterbuck	2	-	1	2	0.4007	0.62
warthog	4	-	1	3	0.3352	0.68
sable	2	-	-	2	0.4314	0.72
kudu	1	-	-	2	0.6745	0.46
impala	2	-	-	-	0.9376	0.35
giraffe	5	-	-	-	1.9369	0.66
tsessebee	2	-	-	-	0.3953	0.66
eland	6	-	1	-	0.4473	0.72
roan	5	-	-	2	0.5877	0.66

The symbol - presents the entry of zero, which indicates no factor or lagged variable used for prediction.

differs among these species. Additionally, the largest improvement is found in warthogs, for which MSFE is reduced to nearly 34% of benchmark's. And, the least gain is given by impala, which there is a decrease of 6% MSFE for benchmark. Furthermore, it seems that the gains come from using one, or two, common factor(s) regarding sable, kudu, impala and tsessebee. Not surprisingly, regional factors are not helpful to predict total population, which results from $k_r = 0$ for all the species. However, we have seen that forecasting with one species-specific factor has achieved substantial improvement for four species: wildebeest, waterbuck, warthog and eland. Furthermore, some gains can be accomplished by including lagged variables for the waterbuck and warthog. Therefore, when forecasting total populations, we can successfully improve the predictive performance using factors as augmented predictors, which factors in various levels are estimated based on regional population changes.

Table 3 contains the results of optimal forecasts for regional population (in logarithms). It is seen that, in south and center, the predictive performance for all of the eight species are similar, and we find that factor-augmented forecasting performs better than benchmark for six species. Furthermore, forecasting using regional and/or species-specific factors shows considerable improvement regarding warthog, sable and kudu in south. And, it seems that forecasting with one, or two, common factor(s) can achieve improvement in most cases. Moreover, the predictive performance are much alike in north and far north, because two regions are adjacent. Particularly for zebra, the AR forecasts, as benchmark, can perform slightly better than

Table 3: The optimal one-year-ahead out-of-sample forecasts: regional populations

	k	k_r	k_s	q	RMSFE		k	k_r	k_s	q	RMSFE
SOUTH						CENTER					
zebraS	2	-	-	-	0.5971	zebraC	2	-	-	-	0.5934
wildebeestS	2	-	-	-	0.9108	wildebeestC	1	-	-	1	1.4379
waterbuckS	4	2	2	1	0.8900	waterbuckC	3	-	2	2	0.4907
warthogS	2	2	1	-	0.2160	warthogC	3	1	2	-	0.3575
sableS	2	1	1	-	0.1459	sableC	4	1	-	-	0.5301
kuduS	2	1	-	2	0.1703	kuduC	1	-	2	1	0.6059
impalaS	1	1	-	-	1.1322	impalaC	2	-	-	-	0.8462
giraffeS	1	-	1	3	1.0142	giraffeC	1	-	-	3	1.2161
NORTH						FAR NORTH					
zebraN	2	-	-	-	1.1310	zebraFN	3	-	-	3	1.1739
wildebeestN	4	2	-	2	0.3455	wildebeestFN	4	1	1	-	0.3843
waterbuckN	4	-	-	-	0.5852	waterbuckFN	4	1	2	1	0.1802
warthogN	3	1	2	-	0.3722	warthogFN	2	-	-	1	0.4739
sableN	4	-	-	2	0.5945	sableFN	2	1	1	2	0.2827
kuduN	2	1	1	2	0.4364	kuduFN	1	-	-	2	0.7901
impalaN	2	1	-	3	0.6044	impalaFN	2	2	-	-	0.3040
giraffeN	2	-	1	3	0.8372	giraffeFN	1	1	1	-	0.2482
tsessebeeN	4	-	2	-	0.5546	tsessebeeFN	3	1	2	2	0.4200
elandN	5	2	1	-	0.2359	elandFN	4	-	-	-	0.4198
roanN	2	-	1	-	0.6108	roanFN	3	1	1	2	0.3275

The symbol - presents the entry of zero, which indicates no factor or lagged variable used for prediction.

the forecast estimates constructed by our model. However, for the rest ten species, predictive improvement can be accomplished by involving multi-level factors as augmented predictors.

4 Conclusions

This paper focuses on analyzing and forecasting ungulate population abundance levels with the aerial census in KNP. We show that factor components, in various levels, play an important role in representing environmental effects to animal population fluctuations. And, we also demonstrate influence of density-dependence within populations to some extent. Moreover, we estimate factors using regional population growth rates, which the estimation method builds on the work of Stock and Watson (2002) and Beck et al. (2011). With multi-level factor estimates as augmented predictors, the resulting forecasts can make substantial improvement in comparison to AR

forecasts. In addition, the extent of predictive improvement differs widely across regions and species.

Appendices

A Figures

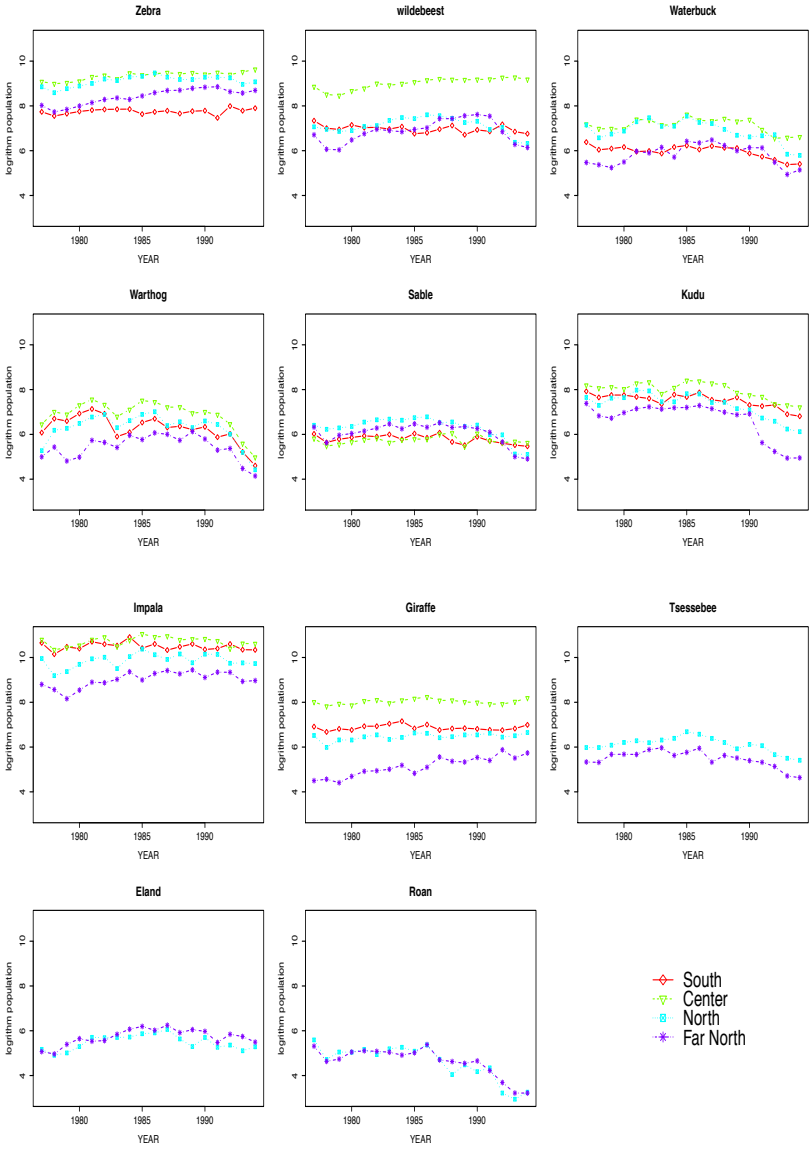


Figure A.1: The development of regional populations (in logarithms) for individual species from 1977 to 1994

B Tables

Table B.1: MSFE values of optimal forecasts for total populations given various T_0

	Year T_0									
	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
zebra	0.18	0.13	0.15	0.12	0.13	0.09	0.11	0.12	0.16	0.14
wildebeest	0.08	0.08	0.07	0.04	0.03	0.03	0.03	0.01	0.01	0.01
waterbuck	0.27	0.24	0.17	0.16	0.12	0.11	0.10	0.11	0.05	0.01
warthog	0.23	0.22	0.21	0.15	0.14	0.16	0.17	0.03	0.03	0.03
sable	0.23	0.22	0.22	0.22	0.25	0.22	0.14	0.02	0.02	0.02
kudu	0.21	0.21	0.15	0.14	0.15	0.16	0.19	0.02	0.02	0.02
impala	0.31	0.16	0.10	0.11	0.06	0.06	0.06	0.05	0.05	0.01
giraffe	0.81	0.69	0.74	0.83	0.69	0.57	0.67	0.83	0.41	0.61
tsessebee	0.27	0.29	0.18	0.17	0.04	0.05	0.03	0.02	0.01	0.01
eland	0.37	0.32	0.24	0.23	0.22	0.23	0.24	0.21	0.02	0.00
roan	0.24	0.26	0.28	0.25	0.22	0.08	0.09	0.11	0.14	0.02

References

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29.

Beck, G. W., Hubrich, K., and Marcellino, M. (2009). Regional inflation dynamics within and across euro area countries and a comparison with the united states. *Economic Policy*, 24(57):142–184.

Beck, G. W., Hubrich, K., and Marcellino, M. G. (2011). On the importance of sectoral and regional shocks for price-setting.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Caughley, G. (1974). Bias in aerial survey. *The Journal of Wildlife Management*, pages 921–933.

Coulson, T., Albon, S., Guinness, F., Pemberton, J., and Clutton-Brock, T. (1997). Population substructure, local density, and calf winter survival in red deer (cervus elaphus). *Ecology*, 78(3):852–863.

Elton, C. (1924). Periodic fluctuations in the numbers of animals: their causes and effects. *British Journal of Experimental Biology 2*, pages 119–163.

Forchhammer, M. C., Post, E., Stenseth, N. C., and Boertmann, D. M. (2002). Long-term responses in arctic ungulate dynamics to changes in climatic and trophic processes. *Population Ecology*, 44(2):113–120.

Forchhammer, M. C., Stenseth, N. C., Post, E., and Landvatn, R. (1998). Population dynamics of norwegian red deer: density–dependence and climatic variation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1393):341–350.

Fritz, H. and Duncan, P. (1994). On the carrying capacity for large ungulates of african savanna ecosystems. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 256(1345):77–82.

Månsson, L., Ripa, J., and Lundberg, P. (2007). Time series modelling and trophic interactions: rainfall, vegetation and ungulate dynamics. *Population ecology*, 49(4):287–296.

Mason, S. (1996). Rainfall trends over the lowveld of south africa. *Climatic Change*, 32:35–54.

Mason, S. and Jury, M. (1997). Climatic variability and change over southern africa: a reflection on underlying processes. *Progress in Physical Geography*, 21(1):23–50.

McNaughton, S. and Campbell, K. (1991). Long-term ecological research in african ecosystems. *Long-term research. Wiley, Chichester*, pages 173–189.

Newman, K., Buckland, S., Morgan, B., King, R., Borchers, D., Cole, D., Besbeas, P., Gimenez, O., and Thomas, L. (2014). *Modelling Population Dynamics: Model Formulation, Fitting and Assessment using State-Space Methods*. Methods in Statistical Ecology, Springer.

Ogutu, J. O. and Owen-Smith, N. (2003). Enso, rainfall and temperature influences on extreme population declines among african savanna ungulates. *Ecology Letters*, 6(5):412–419.

Post, E. (2005). Large-scale spatial gradients in herbivore population dynamics. *Ecology*, 86(9):2320–2328.

Post, E. and Stenseth, N. C. (1998). Large-scale climatic fluctuation and population dynamics of moose and white-tailed deer. *Journal of animal ecology*, 67(4):537–543.

Redfern, J., Viljoen, P., Kruger, J., and Getz, W. (2002). Biases in estimating population size from an aerial census: a case study in the kruger national park, south africa: Starfield festschrift. *South African Journal of Science*, 98(9 & 10):p–455.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

Viljoen, P. and Retiff, P. (1994). The use of the global positioning system for real-time data collecting during ecological aerial surveys in the kruger national park, south africa. *Koedoe*, 37(1):149–157.