ORIGINAL ARTICLE

Model-Based Prediction of Nephropathia Epidemica Outbreaks Based on Climatological and Vegetation Data and Bank Vole Population Dynamics

S. Amirpour Haredasht[1], C. J. Taylor[2], P. Maes[3], W. W. Verstraeten[1,4,5], J. Clement[3], M. Barrios[1], K. Lagrou[6], M. Van Ranst[3], P. Coppin[1], D. Berckmans[1] and J.-M. Aerts[1]

[1] Measure, Model and Manage Bioresponses (M3-BIORES), Department of Biosystems, KU Leuven, Leuven, Belgium
[2] Department of Engineering, Lancaster University, Lancaster, UK
[3] National Reference Laboratory for Hantavirus Infections, Laboratory of Clinical Virology, Rega Institute, KU Leuven, Leuven, Belgium
[4] Royal Netherlands Meteorological Institute (KNMI), Climate Observations, De Bilt, The Netherlands
[5] Eindhoven University of Technology, Applied Physics, Eindhoven, The Netherlands
[6] Department of Experimental Laboratory Medicine, KU Leuven, Leuven, Belgium

Impacts

- Dynamic data-based model generates stochastic forecasts of the occurrence of nephropathia epidemica (NE) epidemics 3 months ahead.
- We have demonstrated that NE outbreaks can be accurately predicted using climatic and vegetation data or bank voles' population dynamics with a dynamic data-based model.
- Such a modelling approach can be used to develop new tools to help prevent future NE outbreaks.

Keywords:
Bank voles; nephropathia epidemica; hantaviruses; rodent-born diseases; population dynamics; model; satellite; prediction; time series; climate change; zoonoses; human disease

Correspondence:
J.-M. Aerts. Measure, Model and Manage Bioresponses (M3-BIORES), Department of Biosystems, KU Leuven, Kasteelpark Arenberg 30, B-3001 Leuven, Belgium.
Tel.: +32 16 321434 or +32 16 321436; Fax: +32 16 321480; E-mail: jean-marie.aerts@biw.kuleuven.be

The research was carried out at Katholieke Universiteit Leuven, Department of Biosystems, Measure, Model & Manage Bioresponses (M3-BIORES). To develop and test the prediction models, two datasets were used, namely a data set of NE cases in Finland and a dataset of NE cases in Belgium.

Received for publication February 6, 2012

doi: 10.1111/zph.12021

Summary

Wildlife-originated zoonotic diseases in general are a major contributor to emerging infectious diseases. Hantaviruses more specifically cause thousands of human disease cases annually worldwide, while understanding and predicting human hantavirus epidemics pose numerous unsolved challenges. Nephropathia epidemica (NE) is a human infection caused by *Puumala virus*, which is naturally carried and shed by bank voles (*Myodes glareolus*). The objective of this study was to develop a method that allows model-based predicting 3 months ahead of the occurrence of NE epidemics. Two data sets were utilized to develop and test the models. These data sets were concerned with NE cases in Finland and Belgium. In this study, we selected the most relevant inputs from all the available data for use in a dynamic linear regression (DLR) model. The number of NE cases in Finland were modelled using data from 1996 to 2008. The NE cases were predicted based on the time series data of average monthly air temperature (°C) and bank voles' trapping index using a DLR model. The bank voles' trapping index data were interpolated using a related dynamic harmonic regression model (DHR). Here, the DLR and DHR models used time-varying parameters. Both the DHR and DLR models were based on a unified state-space estimation framework. For the Belgium case, no time series of the bank voles' population dynamics were available. Several studies, however, have suggested that the population of bank voles is related to the variation in seed production of beech and oak trees in Northern Europe. Therefore, the NE occurrence pattern in Belgium was predicted based on a DLR model by using remotely sensed phenology parameters of broad-leaved forests, together with the oak and beech seed categories and average monthly air temperature (°C) using data from 2001 to 2009. Our results suggest that even without any knowledge about hantavirus dynamics in the host population, the time variation in NE outbreaks in Finland

could be predicted 3 months ahead with a 34% mean relative prediction error (MRPE). This took into account solely the population dynamics of the carrier species (bank voles). The time series analysis also revealed that climate change, as represented by the vegetation index, changes in forest phenology derived from satellite images and directly measured air temperature, may affect the mechanics of NE transmission. NE outbreaks in Belgium were predicted 3 months ahead with a 40% MRPE, based only on the climatological and vegetation data, in this case, without any knowledge of the bank vole's population dynamics. In this research, we demonstrated that NE outbreaks can be predicted using climate and vegetation data or the bank vole's population dynamics, by using dynamic data-based models with time-varying parameters. Such a predictive modelling approach might be used as a step towards the development of new tools for the prevention of future NE outbreaks.

Introduction

Hantaviruses are rodent- or insectivore-borne viruses, some of which are recognized as a cause of human haemorrhagic fever with renal syndrome. In Western and Central Europe and in Western Russia, one of the most important hantaviruses is the *Puumala virus*, which is transmitted to humans by infected red bank voles (*Myodes glareolus*). *Puumala virus* causes a mild form of haemorrhagic fever with renal syndrome called nephropathia epidemica (NE) (Clement et al., 2006).

Human hantavirus epidemics have often been explained by bank vole's abundance (Palo, 2009). Several studies showed the link between the abundance of the bank vole's food on the bank vole's population dynamics and NE incidence. Clear evidence for this can be found in the mast year phenomenon, which is defined as the abnormal abundance of seed production by oak or beech trees. It is generally believed that rodent peak populations occur after the mast years and cause NE outbreaks in Belgium (Tersago et al., 2009; Clement et al., 2010).

Several studies have been carried out to develop a tool for predicting the epidemic years of NE incidence in Belgium based on environmental factors. Clement et al. (2009) showed, through statistical analysis, that it was possible to predict the epidemic years based on the precipitation and temperature of the previous years. Tersago et al. (2009) showed a relation between annual numbers of NE incidence based on tree ecology and the average air temperature, together with summer and autumn precipitation by using a generalized linear model. Both studies related the incidence of NE cases to the mast year phenomenon that has a direct influence on the number of bank voles in the forest. The analysis of Clement et al. (2009) and Tersago et al. (2009) confirmed that peak years in NE cases are proceeded by a year with high seed production of at least native oak, beech or both. Their study aimed to forecast NE outbreaks, but the mast year theory was unable to predict the extreme NE outbreak in 2005. None of the above studies were able to quantify the prediction of the number of NE cases and the increasing trend in the number of NE cases in Belgium in recent years. In general, such disease forecasting is most useful to health services when NE cases can be predicted 2–6 months ahead, allowing tactical responses to be made when disease risk is expected to increase. For this reason, the present article focuses on predicting NE cases one season (or 3 months) ahead.

In our study, we hypothesize that, although the route of the transmission is indirect, the number of new cases is proportional to the square of the bank voles' abundance. Additionally, we hypothesize that the seed production of the beech and oak trees, together with the mixed forest phenological data, can be a useful indicator of bank vole population dynamics.

In this manner, the objective of this study was to develop mathematical models that allow prediction of the occurrence of NE cases 3 months ahead. To develop and test the models, two data sets were considered, one concerning NE cases in Finland and the other in Belgium. The second objective of the research is to answer the question 'Can we use our models to predict the number of NE cases over a longer time horizon, to help health services to optimize their disease control strategies?'.

Materials and Methods

Data

To develop and test the prediction models, two data sets were used, namely a data set of NE cases in Finland and a data set of NE cases in Belgium (Figs 1a and 2a). To predict the NE data, we selected several potential

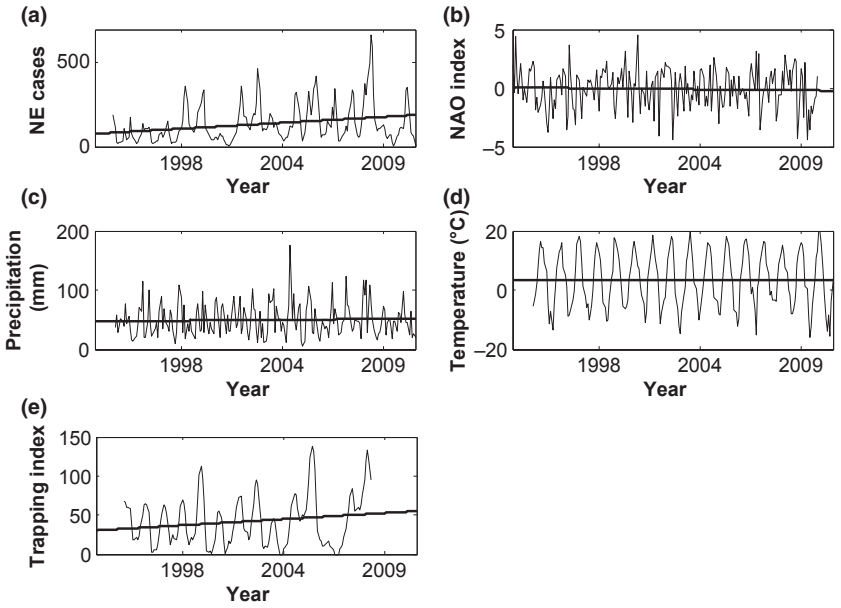

Fig. 1. Nephropathia epidemica (NE) cases in Finland (a) and climatological data [North Atlantic Oscillation (b), sum of monthly precipitation (mm) (c) and average monthly temperature (°C) (d)] together with interpolated bank vole trapping index (e). The data and trend lines are shown for clarity.

climatological input variables, namely air temperature (°C), precipitation (mm) and North Atlantic Oscillation as inputs for our predictive models. The bank voles' trapping index data in Finland were available from July 1995 to October 2008. The number of NE cases in Finland were modelled using data from 1996 to 2008. For the Belgium case, no time series of the bank voles' population dynamics were available. The NE occurrence pattern in Belgium was therefore predicted by using remotely sensed phenology parameters of broad-leaved forests, together with the oak and beech seed categories using data from 2001 to 2009.

Nephropathia epidemica

The NE cases in Finland were obtained from the Finnish National Institute for Health and Welfare, which maintains the National Infectious Disease Registry (http://www3.ktl.fi/), into which laboratory-confirmed diagnoses of Puumala virus infection have been reported since 1995. In our study, we used the monthly human NE cases reported in Finland in the period 1995–2008 (Fig. 1a).

The other data set was obtained from the Belgium Scientific Institute of Public Health (Brussels). We obtained data of NE cases, namely the weekly number of NE cases per postal code (a spatial entity smaller than the municipality) in the period 1994–2009 (Fig. 2a).

Bank vole's population data

The bank voles' trapping index data in Finland were derived from the work of Kallio et al. (2009). Bank voles were trapped in the Konnevesi area in Central Finland four times per year. In our study, the bank voles' trapping index data of the captured bank voles from July 1995 to October 2008 were used (Kallio et al., 2009) (Fig. 1e).

The first trapping session (spring) was carried out in early breeding season in May. The second (summer) was conducted typically during the first week of July, in the middle of the breeding season. The third (late summer) was typically in late August, and the fourth (late autumn) was carried out in late October–early November, after the breeding season typically just before the first snowfall.

Climate data

The objective of this study was to develop a predictive model of incidences of NE in Finland and Belgium by taking into account climatological data. Large-scale climate fluctuations, such as the North Atlantic Oscillation, influence the population structure of animals, their life history, as well as animal population dynamics (Stenseth et al., 2002; Loeuille and Ghil, 2004). Therefore, climate can influence host defence, vectors, pathogens and habitat

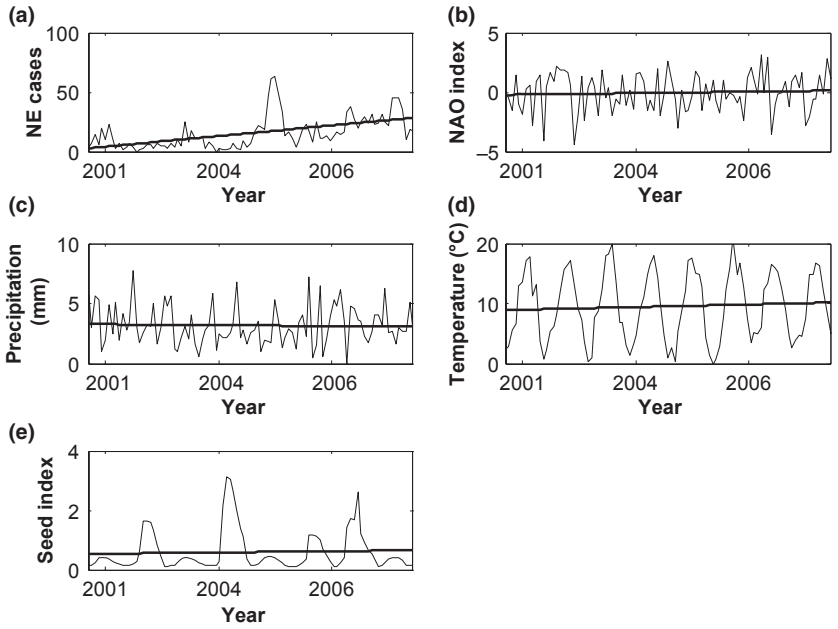

Fig. 2. Nephropathia epidemica (NE) cases in Belgium (a) and climatological data [North Atlantic Oscillation (b), sum of monthly precipitation (mm) (c) and average monthly temperature (°C) (d)] together with seed index (e). The data and trend lines are shown for clarity.

(Epstein, 2002). Several studies were carried out to develop a tool for predicting the epidemic years of NE incidence in Belgium based on climatological factors. Clement et al. (2009) showed, through statistical analysis, that it was possible to predict the epidemic years based on the precipitation and temperature of the previous years. In another study, a multiple-input, single-output (MISO) model was developed, describing the NE cases as a function of three inputs: average measured monthly precipitation (mm) and temperature (°C) in Belgium, as well as the estimated carrying capacity (voles ha^{-1}) from the SIR (susceptible, infective and remove with immunity) model of Sauvage et al. (2007) over an 11-year study period (1996–2008; Amirpour Haredasht et al., 2011). Hence, to predict the NE data, we selected several potential climatological input variables: air temperature (°C), precipitation (mm) and North Atlantic Oscillation as inputs for our predictive models.

Monthly indices of the North Atlantic Oscillation, defined as the difference in normalized sea level pressures between Ponta Delgada, Azores and Stykkisholmur–Reykjavik, Iceland, are available since 1865. These data were provided by the Climate Analysis Section, NCAR, Boulder, USA (Hurrell and Dickson, 2004; Fig. 1b).

The Royal Meteorological Institute of Belgium (Ukkel) provided daily data on air temperature (°C) and precipitation (mm) for 1996–2008. For modelling the dynamics of the NE cases, we calculated monthly average precipitation (mm) and average temperatures (°C) based on the daily reported climate data of Forges, which is the most endemic area in Belgium (and is located in the south of the country; Figs 2c and d).

The Finnish Meteorological Institute provided climatological data of monthly values for the Konnevesi area in Central Finland (again, the most endemic region of the country). For modelling the dynamics of the NE cases, we obtained the sum of precipitation (mm) and average air temperature (°C) from 1996 to 2008 (Figs 1c and d).

Enhanced vegetation index time series and phenological parameters

The enhanced vegetation index time series for the period 2001–2008 used in this study were derived from Barrios et al. (2010). In their study, the enhanced vegetation index time series were fitted to smooth curves from which a number of phenological parameters, such as the growing season period, were estimated.

The vegetation growing season was defined in terms of the energy reflectance in the red and infrared segment of the electromagnetic spectrum. Therefore, in temperate regions, the growing season can be conceived as the period of the year between green-up in spring when EVI values start to increase and senescence in fall when EVI values approach a value of zero, that is, growing season is the annual period of photosynthetical activity. The patterns of these parameters were examined for exploring possible connections between vegetation dynamics and

NE incidence. The EVI time series and phenological parameters used in this study are derived from (Barrios et al., 2010).

Tree seed production categories

The Tree Seed Centre of the Ministry of the Walloon Region in Belgium supplied categories of seed production of beech (*Fagus sylvatica*) and native oak species (*Quercus robur*, *Quercus petraea*). Tree seed production for each tree species was divided into four categories: 'very good years' (the species is fruiting throughout the Walloon territory, and practically all trees are bearing seed in high quantities), 'good years' (the species is fruiting throughout the territory, but the trees are bearing much less seed and some trees do not fruit), 'moderate years' (there is a reduced number of trees bearing seeds and sometimes only located in a portion of the territory) and 'low years' (years without fructification in significant quantities).

Two separate time series of seed production of beech ($B(t)$) and oak ($O(t)$) trees were developed. To quantify the seed production for each year, a value from 0 to 3 (0 = low, 1 = moderate, 2 = good, 3 = very good) was assigned for each category. The monthly time series were constructed as follows: for each year, the value of seed production was set equal to zero in the non-growing month and set equal to the seed production value (of that year) in the growing months. To decrease the complexity of the data-based model, we combined three inputs into one variable, that is, the composite time series of the seed index ($SI(t)$), calculated as follows:

$$SI(t) = EVI(t)B(t)O(t) + EVI(t) \quad (1)$$

where t represents discrete-time instants with a measurement interval of 1 month, $EVI(t)$ is the monthly enhanced vegetation index as estimated by Barrios et al. (2010), $B(t)$ and $O(t)$ are the monthly production of beech and oak nuts, respectively, as discussed previously, $SI(t)$ represents the availability of food for the bank vole's population in nature (Fig. 2e). Note that food is also available in the non-growing seasons, because otherwise the bank vole's population would die out, that is, the availability of food never becomes zero. As $B(t)$ and $O(t)$ are potentially equal to zero, we added $EVI(t)$ in equation (1) to avoid zero values for $SI(t)$ in the non-growing season.

The environmental carrying capacity $K(t)$ of the bank voles coincides with the variation in seed production in the northern part of Europe. $K(t)$ is one of the driving factors responsible for multi-annual fluctuations of the bank vole's population density (Sauvage et al., 2003). Amirpour Haredasht et al. (2011) estimated the carrying capacity by using the mechanistic model described in Sauvage et al. (2007) based on the number of NE cases in Belgium from 1996 until 2007. Our analysis suggested that seed index that is calculated by multiplying production of oak and beech nuts is better representing the estimated carrying capacity (Amirpour Haredasht et al., 2011) compared with a seed index calculated based on adding oak and beech nuts production (R^2 of 0.72 versus R^2 of 0.64).

Modelling

In the first part of our study, we used the mechanistic susceptible infected model described by Sauvage et al. (2003, 2007) to illustrate the dynamical mechanisms of NE and to formulate the interaction between human and bank vole populations. This model helped us in the second part of the study to build a model to predict the occurrence of NE.

In the second part, we used time series analyses (i) to interpolate bank voles' abundance for Finland [using a dynamic harmonic regression model (DHR)] and (ii) to predict the NE cases for both Finland and Belgium [using a dynamic linear regression model (DLR)]. The time series analyses were carried out using the Captain Toolbox (Taylor et al., 2007), as implemented in MATLAB® version 7.6 (R 2008a). Hereafter, we will describe the three different modelling approaches (mechanistic, DHR and DLR) in more detail.

Mechanistic epidemiological model

Many authors have linked the fluctuations of bank vole population with variability in numbers of NE cases (Heyman et al., 2002). Davis et al. (2005) used a density-dependent mass action principle to explain the transmission of infection from rodents to humans. The mass action principle explains the direct transmission of viruses through direct contact between infected bank voles and susceptible humans. In practice, human infection mainly occurs indirectly via the contaminated environment (McCaughey and Hart, 2000; Sauvage et al., 2007). Sauvage et al. (2007) described NE dynamics using the susceptible infected model. In the susceptible infected model, the bank voles are represented as contaminating the environment, through which the virus is spread into the human population (Sauvage et al., 2007). The modelling process of Sauvage et al. (2007) was similar to that used by Berthier et al. (2000). If we use the notation of Berthier et al. (2000), we can write the approximate number of new primary cases based on the susceptible infected model of Sauvage et al. (2007) as:

$$C(t) \propto H_S(t)\varepsilon G(t) \quad (2)$$

where $C(t)$ is the number of new primary NE cases, $H_S(t)$ is the susceptible number of humans, ε is a constant rate (ha^{-1} $year^{-1}$) at which each contaminated faeces leads to transmission of the disease when a human inhales it and $G(t)$ is the site contamination/decontamination dynamics describing the spread of infection from voles to humans.

As is to be expected with density-dependent pathogen transmission, the higher host densities increase both the number of infected individuals and the infection prevalence. Published field studies of bank voles are available where abundance and prevalence of infection voles have been estimated (Escutenaire et al., 2000; Olsson et al., 2002). In the model of Sauvage et al. (2007), the bank voles contaminated the environment which spread the virus into the human population. Site contamination/decontamination dynamics ($G(\boldsymbol{t})$) allowed the spread of infection from voles to humans. The modelling process was similar to that used by Berthier et al. (2000). In the Sauvage et al. (2007) model, the virus survival outside the host, for example, in the ground, was considered to be about 12 days (Kallio, 2003). Elasticity analyses of Sauvage et al. (2007) model showed that the virus survival outside the host has very little influence of the dynamic of the $G(t)$. Therefore; the site contamination/decontamination dynamics $G(t)$ is proportional to the number of infected bank voles $I(t)$ as follows:

$$G(t) \propto I(t) \quad (3)$$

$$I(t) = N(t)\rho(t) \quad (4)$$

where $I(t)$ is the number of infected bank voles, $N(\boldsymbol{t})$ is the total population of bank voles and $\rho(t)$ is the prevalence of the infection in the bank voles population (i.e. the proportion of bank voles population that is seropositive).

The literature shows a positive correlation between the prevalence of hantavirus infection and the abundance of bank voles (Olsson et al., 2005). To emulate 3-year cyclic variations in bank voles' population dynamics, Sauvage et al. (2003, 2007) used 3-year periodic carrying capacities. The environmental carrying capacity $K(t)$ of the bank voles coincides with the variation in seed production in the northern part of Europe. $K(t)$ is one of the driving factors responsible for multi-annual fluctuations of the bank vole's population density (Sauvage et al., 2003). The fully mechanistic model of Sauvage et al. (2003, 2007), takes into account the infection dynamics and bank vole's population fluctuation, showed that in the cyclic population, the infected voles fluctuated with the same pattern as that of vole's abundance. In other words, prevalence should increase with abundance of bank voles, that is, $\rho(t) = \rho_1 N(t)$ where ρ_1 is a constant (Davis et al., 2005). This will lead to a corresponding and amplified variation in the number of new primary cases in humans (Davis et al., 2005). This can be seen by substituting $\rho(t) = \rho_1 N(t)$ into equation (3) and subsequently equation (2), which yields:

$$C(t) \propto H_S(t)N^2(t). \quad (5)$$

However, this relationship only holds for some limited range of N as prevalence $\rho(t) = I(t)\ N(t)$ is constrained to be between 0 ($I(t) = 0$) and 1 ($I(t) = N(t)$), and there must necessarily be a saturation effect as N increases. The saturation effect may be seen mathematically if we defined the prevalence with generalized linear model with logit link function as $\rho(t) = 1 - e^{-\rho_1 N(t)}$. From this equation, it is clear that the prevalence has a positive correlation with the bank vole's population ($N(t)$). In this formula, the saturation point can be seen as if ($N(t)$) increases, the prevalence will tend to one. The maximum prevalence ($\rho(t)$) found in the published field studies of bank voles (Escutenaire et al., 2000; Olsson et al., 2002; Sauvage et al., 2002) was a prevalence of 29% (Sauvage et al., 2002). Hence, we can conclude that $-\rho_1 N(t) \ll 1$ and in that case the term $e^{-\rho_1 N(t)}$tends to be approximately equal to $1 + \rho_1 N(t)$ so $\rho(t) = \rho_1 N(t)$.

Equation (5) seeks to represent a complex set of interactions between human and bank voles' populations. These state that: (i) although the route of the transmission is indirect, the number of new cases is proportional to the abundance of bank voles; (ii) the spatial and temporal fluctuation in bank voles can lead to the spatial seasonal and cyclic variation in human cases; and (iii) the number of new NE cases is proportional to the number of susceptible humans who are at risk (e.g. people who have occupational activities related to forest work or farmers).

In this study, we assume that number of susceptible human who are exposed to contact with voles over the study period is fixed. The study of Olsson et al. (2009) showed a strong positive correlation between the square of the bank vole's abundance and the NE cases in Sweden.

From this, we conclude that new primary NE cases in humans are proportional to the square of the bank voles' abundance. This formula helped us in the second part of the study to build a DLR model to predict the occurrence of NE by using the square of the bank voles' abundance as an input.

Interpolation of vole abundance for Finland (DHR model)

The bank voles' trapping index data for Finland were derived from the work of Kallio et al. (2009) and are based on

captured bank voles from July 1995 to October 2008. As discussed below, these data show clear evidence of periodicity; hence, the DHR model (Young et al., 1999) is selected for the analysis. One of the desirable features of the DHR model when estimated within the state-space framework is that it can function well even if there are large gaps in the data and thus can be used to interpolate the bank voles' trapping index time series. The DHR model is defined here as follows:

$$R(t) = L(t) + Q(t) + e(t) \tag{6}$$

where $R(t)$ is the observed time series (i.e. bank voles trapping index), $L(t)$ is a trend or low-frequency component, $e(t)$ is an 'irregular' component, normally defined for analytical convenience as a normally distributed Gaussian sequence with zero mean value and variance σ^2 (i.e. discrete-time white noise) and $Q(t)$ is the following cyclical term based on six components:

$$Q(t) = \sum_{i=1}^{6} \{\alpha_i(t)\cos(f_i t) + \beta_i(t)\sin(f_i t)\}. \tag{7}$$

In this model, $\alpha_i(t)$ and $\beta_i(t)$ are stochastic time-variable parameters, and f_i, $i = 1, 2,\ldots, 6$ are the frequencies associated with the cyclical component. Here, the number of periodic components and the numerical values of the frequencies were obtained by reference to the spectral properties of the series, as defined by the autoregressive spectrum. This was estimated by fitting an autoregression model to the data, for which the model order was first identified by the Akaike's Information Criteria (Akaike, 1974).

The trend components $L(t)$ are represented using a generalized random walk model. The generalized random walk family of models includes the well-known random walk and integrated random walk models as special cases (Taylor et al., 2007), and the present research is limited to these two examples:

Each of the time-variable parameters in this analysis ($\alpha_i(t)$ and $\beta_i(t)$, $i = 1, 2, \ldots, 6$) was represented as a random walk process of the form:

$$\alpha(t) = \alpha(t-1) + \eta_i(t) \tag{8}$$

where $\eta_i(t)$ is a zero mean, white noise input. The trend component $L(t)$ in equation (6) is modelled as an integrated random walk of the form:

$$\begin{aligned} L(t) &= L(t-1) + S(t-1) \\ S(t) &= S(t-1) + \eta_7(t) \end{aligned} \tag{9}$$

where $S(t)$ represents the 'slope' of the trend, and $\eta_7(t)$ is a zero mean, white noise input.

In this manner, the estimation algorithm is instructed that each parameter is a stochastic variable that is likely to change by an unknown but small amount over each sampling interval, within the stochastic limits imposed by the covariance matrix P for the set of random walk processes. An associated noise variance ratio matrix was estimated using frequency domain optimization (Young et al., 1999).

In particular, the variances associated with each $\eta_i(t)$ in the random walk and integrated random walk models are critical in estimating the time-variable parameters and need to be optimized against the data. The ratio of these variances to the variance σ^2 of the residual $e(t)$ is the noise variance ratio referred to previously. The optimum noise variance ratios (or 'hyperparameters') are exploited by the recursive DHR algorithm to yield estimates of the time-variable parameters and, hence, estimates of each unobserved component, which collectively make up the total DHR output (Taylor et al., 2007).

Such a state-space model is particularly well suited for estimation based on optimal time-variable parameter recursive estimation, in which the time-variable parameters (acting as surrogate 'states') are estimated sequentially by the Kalman Filter (Kalman, 1960) while working through the data in temporal order. In the offline situation, where all the time series data are available for analysis, this filtering operation may be accompanied by optimal fixed-interval smoothing (Bryson and Ho, 1969).

Predictive models of NE cases using climate and vole/phenological data (DLR model)

Dynamic linear regression or DLR analysis is a generic tool for modelling non-stationary data (Young, 1999; Taylor et al., 2007). Here, the stochastic evolution of each parameter is assumed to be described by the general random walk model process mentioned previously. Represented as a variation of equation (6), the DLR model usually consists of trend and input function components which, in the Captain Toolbox (Taylor et al., 2007), are implemented in the state-space form. For the NE data, we investigated several model structures that related the output (number of NE cases) to several potential input variables. The models that yielded the best forecasting results differ for each data set (Belgium and Finland). However, the general form is shown below.

$$\begin{aligned} C(t) &= \sum_{i=1}^{i=n} m_i(t)u_i(t - nt_i) + e(t) \\ e(t) &\sim N\{0, \epsilon^2\} \quad t = 1, 2, \ldots, N \end{aligned} \tag{10}$$

where $m_i(t)$, $i = 1, 2,...,n$ are either constant parameters (the normal regression model) or they may vary over the observation interval to reflect possible changes in the regression relationship; $u_i(t)$, $i = 1, 2, \ldots, n$ are the regression (input or exogenous) variables that are assumed to affect the 'dependent' variable; $nt_i(t)$ is the time delay between each input i and their first effects on the output;

and $e(t)$ is an 'irregular' component, normally defined for analytical convenience as a normally distributed Gaussian sequence with zero mean value and variance (i.e. discrete-time white noise).

The parametric time variability was not known prior to the analysis, and so, each time-variable parameter was defined as a non-stationary stochastic variable. This adds a statistical degree of freedom to the estimation problem, so allowing for the estimation of parameter variations. Such variations may result from physical changes in the process or from some form of non-linearity in the data. In this manner, the models obtained are all inherently self-adaptive, namely they change their parameters automatically in an optimal manner to reflect changes in the nature of the time series.

The cross-correlation function was utilized to help clarify which of the available inputs had the strongest linear relationship with the output (NE cases). The cross-correlation function is a representation of the linear correlation between an assumed input variable and the output at different lags, plotted against this lag.

Statistical analysis

The goodness of fit of the NE prediction was calculated by using a mean relative prediction error (MRPE) defined as (Oltjen and Owens, 1987):

$$\text{MRPE } \% = \frac{1}{n}\sum_{t=1}^{n}\sqrt{\left[\frac{C(t)-\hat{C}(t)}{C(t)}\right]^2} \times 100 \qquad (11)$$

where $C(t)$, $i = 1, 2, 3, \ldots, n$ are the number of reported NE cases, $\hat{C}(t)$ are the last predicted NE cases at the time to which forecasts were made (depending on the forecasting horizon) and t represents discrete-time instants with a measurement interval of 1 month. MRPE values were calculated for all samples and then averaged for the whole data set. The 'best' model was considered to be the one with the lowest MRPE value for a given time series. MRPE has been used by several authors to quantify the prediction performance of models (Oltjen and Owens, 1987; Talpaz et al., 1991; Aerts et al., 2003).

To quantify how close forecasts or predictions were to the eventual NE cases, we also calculated the mean absolute deviation error (MAD) for the selected model as:

$$\text{MAD} = \frac{1}{n}\sum_{t=1}^{n}|C(t)-\hat{C}(t)|. \qquad (12)$$

To find the time delays (nt_i) in equation 10 between each input and their first effects on the NE cases, the time delay for each input was varied from 3 to 12 months. Hence, for a model with two inputs, in total 100 models with different time delays were generated, and each of these models was used to predict the NE cases 3 months ahead. For each sample of the data set (i.e. every month from 2001 onwards for the Finland case and every month from 2003 onwards for the Belgium case), we estimated the model parameters and generated the 3-month-ahead prediction.

To calculate the MRPE for each predictive horizon (one up to 6 months ahead), we followed the same processes mentioned previously, again varying the time delay for each input to avoid the problem of having to predict the inputs.

Results

The DHR model for interpolating bank voles' trapping index

The bank voles' trapping index data for Finland showed some evidence of yearly fluctuations. The DHR model estimates the frequencies associated with these periodic components by reference to the spectral properties of the series, as defined by the autoregressive spectrum. This is estimated by initial autoregressive analysis of the series to identify the number and values of the fundamental and harmonic frequencies associated with the periodicity.

The Akaike's Information Criteria (Akaike, 1974) yielded a 31st order autoregressive model for the bank vole population time series, and the associated autoregressive spectrum (31), presented in Fig. 3, suggests that the bank vole population series had a strong 36-month periodicity. This 3-year cycle is probably caused by predators that generate the cyclic population fluctuations of voles observed in Northern Europe (Korpimäki et al., 2002). In addition to the 36-month periodicity, five of the harmonics (i.e. 18, 12, 9, 6, 4 and 3 months) also had pronounced peaks. The annual cycle is indicated by the pronounced peaks at periods of 12, 6, 4 and 3. (months/cycle) (Fig. 3). In fact, as the other harmonics demonstrated in Fig. 3 were relatively small, they were ignored.

Subsequent Kalman filter/fixed-interval smoothing estimations based on the state-space version of the model yielded the estimates of the time-variable parameters and, hence, the estimates of each unobserved component, which collectively make up the total DHR model output.

Figure 4a shows the resulting interpolated series, together with the original measured bank voles' trapping index time series for Finland. Based on this analysis, it was possible to resample the data of the bank voles' trapping index at the same frequency as the other variables, that is, at a time interval of 1 month.

To quantify the performance of the DHR's interpolation, we used the interpolated time series to predict the

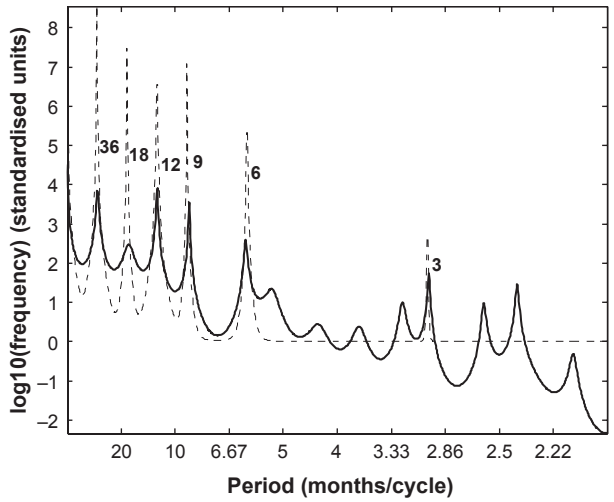

Fig. 3. Autoregressive spectrum for the bank voles trapping index in Finland. Solid is the autoregressive spectrum of the bank voles trapping index (based on the estimated autoregressive model for these data) and the dashed trace is the autoregressive spectrum of the DHR model (based on the estimated autoregressive of the DHR response) using frequency domain Noise Variance Ratio optimization.

bank vole's population one step ahead with a MRPE of 14%. As can be seen in Fig. 4b, all the model predictions fall within the two times standard error bounds, indicating an acceptable prediction performance.

Selection of the DLR model inputs for forecasting NE cases

The first criterion used in this study to help selecting the relevant inputs for the DLR model was a trend line. For this preliminary analysis, we generated a straight line trend by selecting an integrated random walk model with noise variance ratio = 0 for each time series. As illustrated in Figs 1 and 2, the bank voles' trapping index and the number of NE cases were the only variables in Finland that indicated an increasing trend. For Belgium, the seed index, average air temperature (°C) and NE cases all indicated an increasing trend (Fig. 2).

We performed the cross-correlation function analysis between the time series of the number of human NE cases reported in Finland and the associated climatic variables, that is, North Atlantic Oscillation, average monthly temperature (°C) and sum of the precipitation (mm), at time lags of 0–24 months. In this regard, the analysis revealed with a 95% confidence interval that the highest correlation coefficient was obtained for NE incidence in Finland and the average monthly air temperature (°C) variable with a value of 0.43 and 0.50 for time lags of 4 and 8 months, respectively, between the monthly air temperature and NE cases.

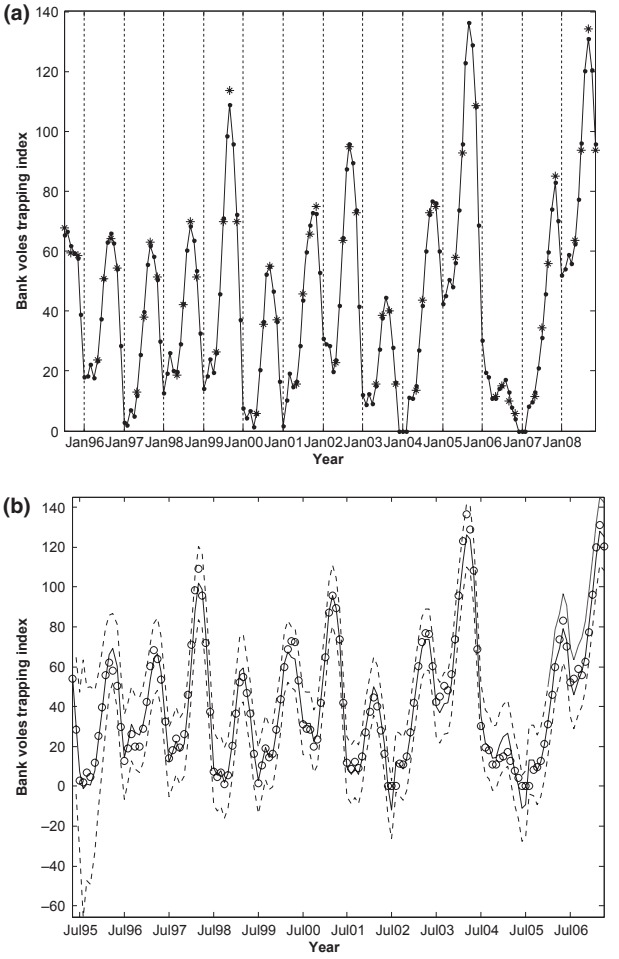

Fig. 4. Interpolated bank voles trapping index in Finland per month (•) compared with the original measured values (*) four times per year (top). Comparison of the 1-month-ahead DHR model predictions (solid lines), the interpolated bank voles' population index (circles) and two times standard error bounds (dashed lines) in Finland (bottom).

Correlation coefficients of 0.70, 0.54 and 0.30 were found between the NE outbreak in Finland and the bank vole's trapping index for time lags of 3, 4 and 5 months, respectively. Therefore, the number of NE cases in Finland was modelled using average monthly air temperature (°C) and bank vole's trapping index as inputs for the DLR model.

Similarly, the cross-correlation function analysis revealed, with a 95% confidence interval, that the NE incidence in Belgium has no significant correlation coefficient with average monthly precipitation (mm) for neither the south of Belgium nor the monthly North Atlantic Oscillation. Hence, only the average monthly air temperature was considered in the model with a correlation coefficient of around 0.30 for time lags of 4, 5 and 10 months between the monthly air temperature and NE cases. The correlation coefficient between the NE outbreaks in Belgium and seed

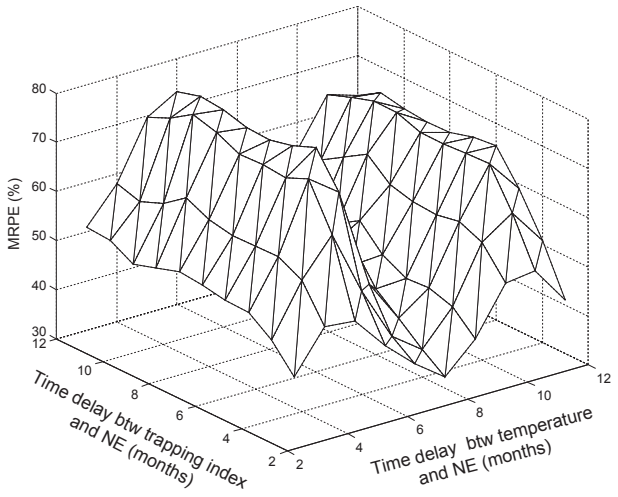

Fig. 5. Mean relative prediction error (MRPE) as a function of time delay between temperature and nephropathia epidemica (NE) cases and of time delay bank voles' trapping index and the NE cases for the Finland case for the 3 month a head prediction. The smaller MRPE indicates a closer fit of the model to the data.

index with a time lag of 12 months amounted 0.70. Therefore, the number of NE cases in Belgium was modelled using average monthly air temperature (°C) and seed index as inputs for the DLR model.

Prediction model for Finland

In the Finland case, we predicted the occurrence of NE cases three months (or one season) ahead based on the squared bank vole's trapping index and average monthly air temperature (°C; see equation 10) as Inputs for the following DLR model:

$$C(t) = L(t) + b_{R^2}(t)R^2(t-3) + b_T(t)T(t-8) + e(t)$$
$$e(t) \sim N\{0, \epsilon^2\} \quad t = 1, 2, \ldots, N \tag{13}$$

where $C(t)$ is the number of NE cases reported in Finland per month; t represents discrete-time instants with a measurement interval of 1 month, $T(t)$ and $R(t)$ represent the two inputs of the model, namely average monthly air temperature (°C) and the bank vole population index, respectively. The mechanistic model of Sauvage et al. (2007) was utilized to explain the use of squared bank voles' trapping index as an input for the DLR model (equation 5); $L(t)$ is a trend or low-frequency component; and $b_{R^2}(t)$and $b_T(t)$ are either constant parameters or they may vary over the observation interval to reflect possible changes in the regression relationship. nt_i is the time

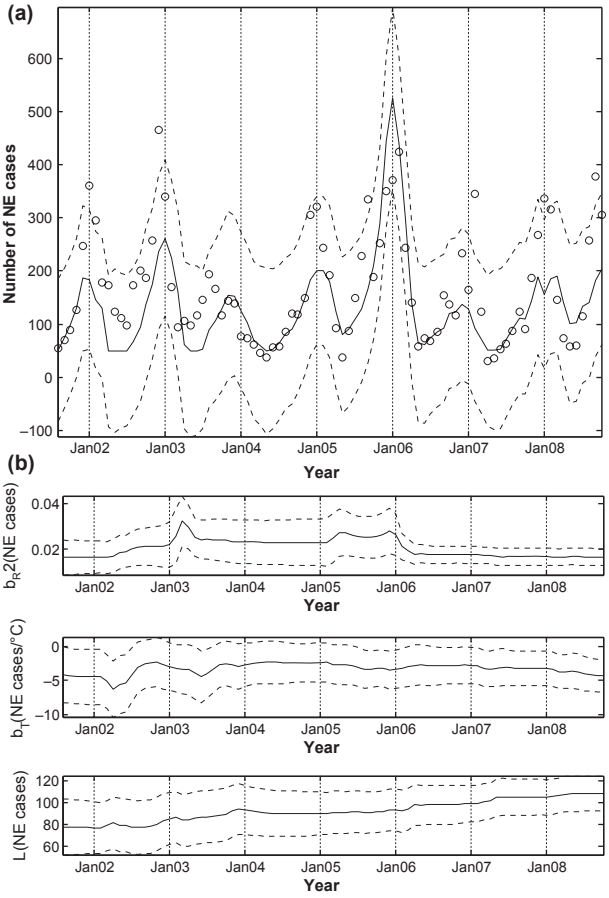

Fig. 6. Comparison of the 3-month-ahead DLR model predictions (solid lines), the reported nephropathia epidemica (NE) cases (circles) and two times standard error bounds (dashed lines) in Finland using as inputs temperature (°C) and bank voles' trapping Index (top). Estimated time-varying parameters (solid lines) and two times standard error bounds (dashed lines) (bottom).

delay between each input i and their first effects on the output; and $e(t)$ is an 'irregular' component, defined for analytical convenience as a normally distributed Gaussian sequence with zero mean value and variance σ^2 (i.e. discrete-time white noise).

These parameters were estimated under the assumption that the trend $L(t)$ and $R(t)$ evolve as random walk processes (equation 8), while $T(t)$ evolves as an integrated random walk process (equation 9). The associated noise variance ratio coefficients were calculated by maximum likelihood optimization based on the sum of squares of the 3-step-ahead forecasting errors. In this case, the associated noise variance ratio coefficients for each repressor were estimated to be close to zero, again implying relatively time-invariant parameters. Indeed, for the results considered below, the noise variance ratio coefficients were assumed to be zero for simplicity. In this regard, it should be noted that future research, which might take

into account spatial characteristics of the climatological data and/or be based on a longer data set, should further investigate optimal values for these noise variance ratio hyperparameters.

This selected model was found to predict NE incidence in Finland 3 months (or one season) ahead with a 34% MRPE (Fig. 5). Note from equation (13) and Fig. 5 that the model is based on a 3 months delayed bank vole trapping index and an 8 months delay for average air temperature. The developed model for Finland could predict the NE cases with a MAD = 20 NE cases per month (Fig. 6a).

Two times standard errors of the estimated parameters are demonstrated in Fig. 6b. The fact that all the model parameters fall within the two times standard error bounds is indicating an acceptable prediction performance of the model.

Prediction model for Belgium

To predict the occurrence of NE cases in Belgium 3 months (or one season) ahead, we used the squared seed index and climatological data (average monthly temperature) as inputs for DLR model (equation 10). The particular DLR model developed for the Belgium case from equation 10 was as follows:

$$C(t) = L(t) + b_{\mathrm{SI}^2}(t)\mathrm{SI}^2(t-12) + b_T(t)T(t-10) + e(t)$$
$$e(t) \sim N\{0, \epsilon^2\} \quad t = 1, 2, \ldots, N \tag{14}$$

where $C(t)$ is the number of NE cases reported in Belgium per month; t represents discrete-time instants with a measurement interval of 1 month; $SI(t)$ and $T(t)$ represent the two inputs for this model, namely the monthly seed index and the average monthly air temperature (°C), respectively; $L(t)$ is a trend or low-frequency component; and, following a similar approach to equation (10), $b_T(t)$ and $b_{SI^2}(t)$are either constant parameters (the normal regression model) or they may vary over the observation interval to reflect possible changes in the regression relationship, as is typical for DLR models.

It has been demonstrated that increasing numbers of beech and oaks fruit in the so-called 'mast years' increase the number of bank voles that are associated with outbreak of NE in North-West Europe (Clement et al., 2009, 2011). The variation in seed production in the northern part of Europe, represented in equation (10) by $SI(t)$, is one of the driving factors responsible for multi-annual fluctuations of the bank vole's population density and coincides with the carrying capacity (Sauvage et al., 2003). The carrying capacity of a biological species in an environment more specifically is the population size of the species that the environment can sustain, given the food, habitat, water and other necessities available in the environment (Sayre, 2008). In our study, the mast year theory was utilized to explain the use of squared seed index as an input to the DLR model for Belgium.

The parameters of equation 10 were estimated under the assumption that $T(t)$ evolves as an integrated random walk process (equation 9), while $SI^2(t)$ and $L(t)$ vary as random walk processes (equation 8). The associated noise variance ratio coefficients were calculated using maximum likelihood optimization, based on the sum of squares of the 3-step-ahead forecasting errors. However, the model with the lowest MRPE (Fig. 5) was the one with the noise variance ratio coefficients of zero for each regressor, implying time-invariant parameters in this case.

This selected model was found to predict NE incidence in Belgium 3 months (or one season) ahead with a 40% MRPE (Fig. 8a). Note from equation (14) and Fig. 7 that the model with the lowest MRPE was obtained using the average air temperature lagged by 10 months and the seed index of the previous year (12 month). The developed model for Belgium could predict the NE cases with a MAD = 7 NE cases per month.

Figure 8b shows that all the model parameters fall within the two times standard errors of the estimated parameters, which indicate an acceptable prediction performance of the model.

Discussion

At present, there are no practical tools for forecasting the number of NE cases based on observed climatological and

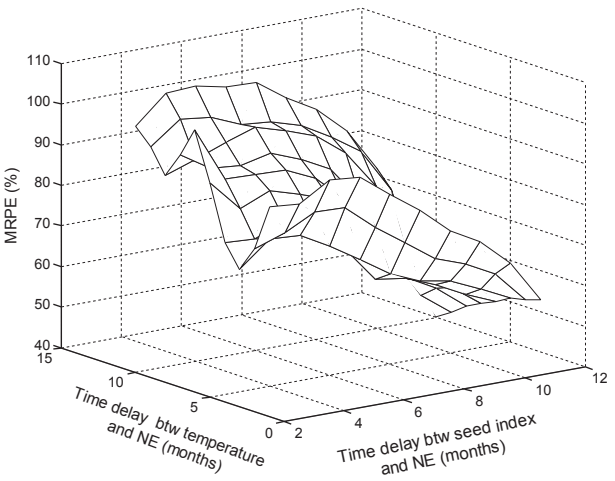

Fig. 7. Mean relative prediction error (MRPE) as a function of time delay between temperature and nephropathia epidemica (NE) cases and of time delay between seed index and the NE cases for the Belgium case for the 3 month a head prediction. The smaller MRPE indicates a closer fit of the model to the data.

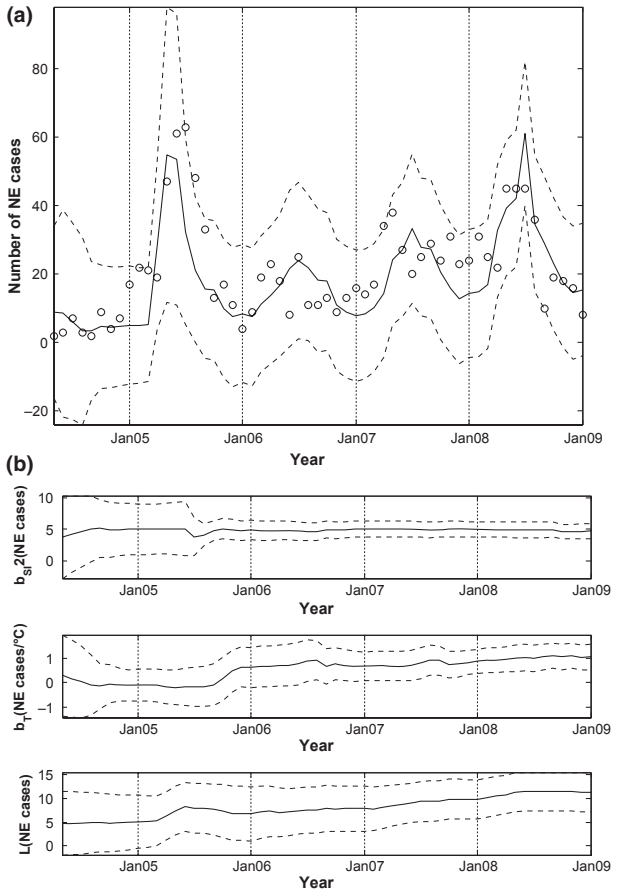

Fig. 8. Comparison of the 3 months ahead DLR model predictions (solid lines), the reported nephropathia epidemica (NE) cases (circles) and two times standard error bounds (dashed lines) in Belgium using as inputs temperature (°C) and seed index (top). Estimated time-varying parameters (solid lines) and two times standard error bounds (dashed lines) (bottom).

vegetation data. Recent work focuses on NE warning systems (Clement et al., 2009), in which flags are raised when epidemics are expected. Setting the threshold for what is an epidemic (defined as a number of cases substantially exceeding that expected based on recent experience or what is thought normal) is subjective. The term epidemic does not combine well with the term prediction (if the expected number is predicted based on recent experience, the prediction can never be 'epidemic' according to the above definition). In general, disease forecasting is most useful to health services when it predicts case numbers 2–6 months ahead, allowing tactical responses to be made when disease risk is predicted to increase. For this reason, this study avoids the problem of setting epidemic thresholds and focuses on predicting NE cases.

In this study, we answered two general epidemiological questions. First, how well the selected models perform in predicting the number of NE cases one season ahead? Secondly, can we use our models to predict the number of NE cases over a wider horizon to help health services to optimize their disease control strategies? Figures 6 and 8 illustrate that the model was predicting the NE cases quite well, but when looking at the results in more detail, it is clear that the model showed particularly good results in the epidemic months where the data were more dynamic. Figures 9 and 10 illustrate the model performance for a range of prediction horizons. NE cases can be predicted up to 6 months ahead with a MRPE of 38% for Finland cases and with a MRPE of 46% for Belgium cases.

The number of NE cases in Finland were modelled based on data from 1996 to 2008. The output of the DLR model was the number of NE cases in Finland, while the inputs were the average measured monthly temperature (°C) and interpolated squared time series of bank voles' trapping index in Finland. The latter input variable was obtained using a DHR model.

Using spectral analysis, the component signals of the bank voles' trapping index time series in Finland were first extracted. The resulting DHR model successfully described the temporal characteristics of the bank voles' trapping index time series and its different dynamical mechanisms with a MRPE of 14% (Fig. 4b). The analysis confirmed that the bank vole's population series contained a seasonal and cyclic components, indicated by the pronounced autoregressive spectrum peaks at periods of 36, 18, 12, 9, 6, 4 and 3 (months/cycle; Fig. 3).

In this study we showed that, although the route of transmission is indirect, new NE cases are proportional to the square of the bank voles' abundance (represented in our model using the DHR interpolated tracking index). Finding this relationship helped us to predict the occurrence of NE cases, by using the square of the bank voles' abundance as

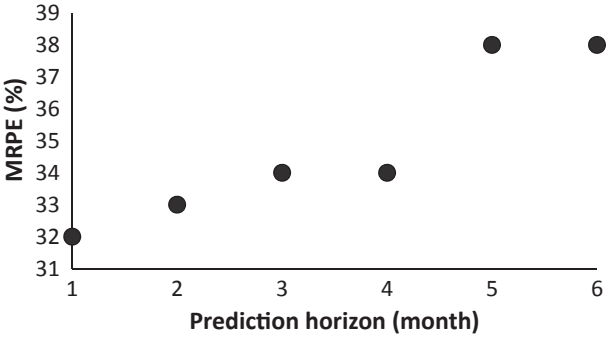

Fig. 9. Mean relative prediction error (MRPE) versus prediction horizon for the Finland case. The smaller MRPE indicates a closer fit of the model to the data.

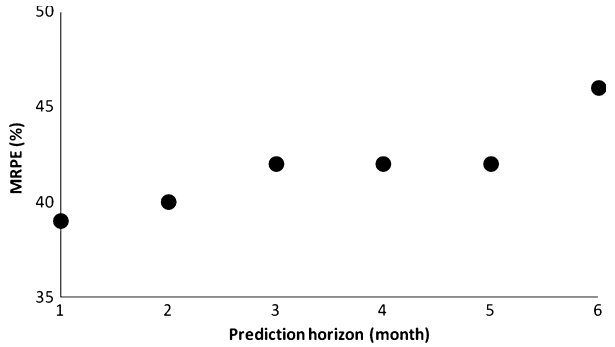

Fig. 10. Mean relative prediction error (MRPE) versus prediction horizon for the Belgium case. The smaller MRPE indicates a closer fit of the model to the data.

an input to the subsequent DLR model. Previously developed mechanistic susceptible infected models (Sauvage et al., 2007) were utilized to explain this use of the square of the bank voles' trapping index.

Our results showed that the NE cases in Finland may be predicted 3 months ahead with a 34% MRPE, solely by using the population dynamics of the carrier species (bank voles) and air temperature, without any knowledge about hantavirus dynamics in the host population (bank voles). Figure 6 shows an illustrative forecast with associated two times standard error bounds. It is clear from Fig. 6 that 94% of the data points lie within the two times standard error bounds, revealing that, despite the fact that we had a short time series without any measurement about the vegetation dynamics, our model could predict the NE cases. Although the cyclical dynamics are captured quite well, the size of some epidemics is not, as indicated by some significant underestimates (e.g. year 2003) and overestimates (year 2006). It should be noticed that the bank voles' trapping index used as inputs in the DLR model were measured in the Konnevesi area in Central Finland. As we used these bank voles' trapping index data to model the NE cases for the whole country, it might be expected that taking into account spatial variation of the bank voles' population would potentially improve the modelling results. The results of our study showed the potential prediction ability of the data-based modelling approach in predicting the NE cases. In future work, the data-based modelling approach may be improved by integration of estimated bank vole population dynamics measured in the field with factors such as vegetation coverage and abundance of food for bank voles. This can provide us with an expert tool to predict (and aid prevention of) the incidence of NE cases.

As described earlier, we investigated 100 DLR models with time delays ranging from 3 months to 1 year, and we selected the best models based on the MRPE. The best model had a 4 months lag between the NE cases in Finland and the bank voles' trapping index. The 3-month lag may readily be interpreted biologically: (i) the signs of symptoms and the antibody responses in humans take 2–6 weeks from the actual infection time (Vapalahti et al., 2003; Kramski et al., 2009); (ii) the transmission of Puumala virus to other bank voles was reported by Yanagihara et al. (1985) to occur 2 weeks after infection.

For Belgium, time series of the bank voles' population dynamics were not available. However, several studies have suggested that the population of bank voles is related with the variation in seed production in Northern Europe. The NE occurrence pattern in Belgium was therefore predicted by using remotely sensed phenology parameters of broad-leaved forests, together with the oak and beech seed categories and average monthly air temperature (°C). This approach can be useful in the areas with a lack of, or very few, bank voles' trapping data. However, this may be specific to Belgium (or other places) where rodent population dynamics are affected mainly by food resources and not other mechanisms. In the boreal zones (e.g. Finland), primarily coniferous forests do not show masting, so the vole cycles are determined predominantly by interactions between voles and their specialist mammalian predators and winter food resources (Hanski et al., 1991; Korpimäki et al., 2005; Huitu et al., 2007).

The DLR analysis revealed that changes in forest phenology and temperature fluctuations, considered as an effect of climate change, may affect the mechanics of NE transmission. NE cases in Belgium may be accurately predicted 3 months ahead with a 40% MRPE, based only on the average monthly temperature and monthly values of the seed index, in this case without any knowledge of the bank vole's population dynamics. Figure 8 shows an illustrative forecast with associated two times standard error bounds. It is clear from Fig. 8 that 93% of the data points lie within the two times standard error bounds of the forecast, revealing that despite the fact that we had a short time series without any measurement about the bank vole's population, our model could predict the NE cases.

Puumala virus is the most common cause of haemorrhagic fever with renal syndrome in Europe with an average of 1000 serology-verified cases annually in Finland (incidence 19 per 100 000; Vapalahti et al., 2003). In Belgium, about 300 cases, at most, of NE have been diagnosed during an epidemic year. In Finland, bank vole populations have 3-year cycles, which cause the 3-year cycles in the NE outbreaks, whereas NE outbreaks in Belgium have 2-year cycles. In contrast to the high risk to humans in Finland in late autumn–winter, the highest numbers of NE cases in Western Europe (Belgium) occur during summers in years when bank voles' abundance is high, and where high epidemic years are predicted by

climate, tree seed production and masting (Clement et al., 2009; Tersago et al., 2009). A high summer temperature induces the formation of flower buds of deciduous forest trees (oak and beech), resulting in the next year in an abundance of seed production. Acorn and beech mast fall down in autumn, contributing to the high winter survival and early breeding of voles in a mast year. This leads to high rodent densities and more human NE cases in summer (Clement et al., 2009; Tersago et al., 2009). In the boreal zone, like in Finland, primarily coniferous forests do not provide significant mast production, so the vole cycles are determined predominantly by interactions between voles and their specialist mammalian predators and winter food resources (Hanski et al., 1991; Korpimäki et al., 2005; Huitu et al., 2007). Our study shows that, despite all the differences between the nature of the NE epidemic in Finland and Belgium, the use of DLR modelling techniques is a valuable approach to predict the NE outbreaks in both countries.

The modelling approach that was introduced in our study is a hybrid between data-based modelling in one hand and a mechanistic susceptible infected model (Sauvage et al., 2007) based on mast theory on the other hand. In this way, the proposed data-based mechanistic modelling approach creates added value. Such data-based mechanistic models can take advantage of the fact that they combine mechanistic process knowledge with measured information (e.g. vegetation indexes, climatological data, etc.). This makes them understandable from a biological/ecological point of view, while simultaneously accounting for real-time measured information to predict future NE outbreaks. These kinds of models have the advantage that they are relatively easy to implement without taking into consideration too many complex biological processes and their parameters.

In this study, we quantified the predictive accuracy of the model over different forecasting horizons (from 1 to 6 months). NE cases were predicted up to 6 months ahead with a MRPE of 46% for Belgium cases and with a MRPE of 38% for Finland cases. Although the MRPE values in our study were not very low, they were considerably lower than the values found by other researchers for similar modelling approaches in similar applications (Briët et al., 2008). In the study of Briët et al. (2008), malaria cases were predicted based on the rainfall in Sri Lanka. They used an autoregressive integrated moving average model to predict the malaria cases in all the districts of Sri Lanka. The prediction error ranged from minimum 22% (for one of the districts) for a 1-month-ahead prediction horizon up to minimum 51% (for one district) for a 4-month-ahead prediction. Furthermore, it should be noticed that in our study, the climatological values used as inputs in the data-based model were measured at only one weather station in Finland or Belgium, which was located close to the endemic regions. As we used these climatological data to model the NE cases for the whole country, it might be expected that taking into account spatial characteristics of the climatological data would potentially further improve the modelling results, although the dynamics and the trend in the NE cases could already be predicted in a satisfactory way in both countries. The epidemiological models can be estimated based on variables and parameters believed to be important for the dynamics of the disease. A data set covering a longer period might also improve the modelling results. The results of our study showed the potential prediction ability of the data-based modelling approach in predicting the NE cases.

To evaluate the prediction ability of the DLR modelling approach furthermore, we applied the method of persistence forecasting. Persistence forecasting is a forecasting method assuming that the future condition will be the same as the present condition. It is often used as a standard of comparison in measuring the degree of skill of forecasts obtained using other methods (Anbarci et al., 2010). In our study, the persistence forecast represented the degree of month-to-month variation of NE cases. The one-step-ahead persistence forecast was calculated for the NE cases for both countries. The persistence forecasting approach had a MRPE of 68% in the case of Belgium (1996–2011) and a MRPE of 55% in the case of Finland (1995–2011). Comparing these values with the results of our study (namely MRPE of 39% for Belgium and 32% for Finland cases for 1-month-ahead prediction), we can conclude that the data-based modelling approach clearly outperformed the method of persistence forecasting.

To quantify how close these forecasts or predictions were to the actual NE cases, we finally calculated the MAD error for both countries. The developed model for Belgium could predict the NE cases with a MAD = 7 NE cases per month, and the developed model for Finland could predict NE cases with a MAD = 20 NE cases per month. These were considered to be acceptable prediction accuracies for early warning of future outbreaks because the MAD values were relatively small compared with the large fluctuations (up to 465 NE cases per month for Finland and up to 63 NE cases per month for Belgium) we want to predict. Based on the above-mentioned findings, we concluded that our model performed satisfactory over the entire 3- to 6-month predictive horizons. As a result, the predictive performance should give the authorities enough time to develop disease control strategies.

Conclusions

In this research, we demonstrated that NE outbreaks can be predicted based on climate and vegetation data or

bank vole's dynamics using dynamic data-based models. The time variation in NE outbreaks in Finland could be predicted 3 months ahead with a 34% MRPE, by taking into account the population dynamics of the carrier species (bank voles), even without any knowledge about hantavirus dynamics in the host population. The time series analysis revealed that the vegetation index, changes in forest phenology (which can be derived from satellite images) and climate fluctuation, which are all considered as an effect of climate change, affect the mechanics of NE transmission. NE outbreaks in Belgium were predicted 3 months ahead with a 40% MRPE, based only on the climatological data and vegetation data, and without any knowledge of the bank vole's population dynamics. Such a modelling approach could be used in a next step to develop tools for prevention of NE outbreaks.

Acknowledgements

This research was supported by the KU Leuven (project IDO/07/005). Piet Maes was supported by a postdoctoral grant from the 'Fonds voor Wetenschappelijk Onderzoek (FWO)-Vlaanderen'. The Captain Toolbox for time series analysis and forecasting (Taylor et al., 2007) was utilized for DHR/DLR estimation.

Nomenclature

Bank vole dynamics	
I	Infected vole density (vole ha^{-1})
N	Bank vole density (vole ha^{-1})
ρ	Prevalence of the infection in the bank voles' population
C	New primary cases in humans (cases)
H_s	Susceptible number of humans (cases)
G	Contaminated proportion of the soil litter surface
ε	Indirect contamination rate (ground to susceptible human, ha^{-1} year^{-1})
Vegetation dynamics	
EVI	The enhanced vegetation index
SI	seed index
O	Seed production of oak
B	Seed production of beech
$K(t)$	Environmental carrying capacity (vole ha^{-1})
Interpolation of vole abundance for Finland (DHR model)	
$R(t)$	Bank voles trapping index
$Q(t)$	Sustained cyclical or quasi-cyclical component
$L(t)$	Trend or low-frequency component
$\alpha_i(t)$	Stochastic time-variable parameters
$\beta_i(t)$	Stochastic time-variable parameters
f_i	The frequencies associated with the (normally longer period) cyclical component (cycles month^{-1})
$\eta_i(t)$	The zero mean, white noise input
$S(t)$	Second state variable (generally known as the 'slope'
Predictive models of NE cases using climate and vole/ phenological data (DLR model)	
t	Discrete-time instants with a measurement interval of 1 month
nt_i	The time delay between each input i and their first effects on the output (month)
$b_i(t)$	The model parameters to be estimated for each input i (cases °C^{-1}, cases depending on the input i)
T	Average monthly air temperature (°C)
$\hat{C}(t)$	The predicted NE cases (Cases month^{-1})
$C(t)$	The reported NE cases (Cases month^{-1})

Reference

Aerts, J. M., M. Lippens, G. De Groote, J. Buyse, E. Decuypere, E. Vranken, and D. Berckmans, 2003: Recursive prediction of broiler growth response to feed intake by using a time-variant parameter estimation method. *Poult. Sci.* 82, 40–49.

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19, 716–723.

Amirpour Haredasht, S., J. M. Barrios González, P. Maes, W. W. Verstraeten, J. Clement, G. Ducoffre, K. Lagrou, M. Van Ranst, P. Coppin, D. Berckmans, and J. Aerts, 2011: A dynamic data-based model describing nephropathia epidemica in Belgium. *Biosyst. Eng.* 109, 77–89.

Anbarci, N., J. Boyd III, E. Floehr, J. Lee, and J. Jin Song, 2010: Population and income sensitivity of private and public weather forecasting. *Reg. Sci. Urban Econ.* 41, 124–133.

Barrios, J. M., W. W. Verstraeten, P. Maes, J. Clement, J.-M. Aerts, S. Amirpour Haredasht, J. Wambacq, K. Lagrou, G. Ducoffre, M. Van Ranst, D. Berckmans, and P. Coppin, 2010: Satellite derived forest phenology and its relation with nephropathia epidemica in Belgium. *Int. J. Environ. Res. Public Health*, 7, 2486–2500.

Berthier, K., M. Langlais, P. Auger, and D. Pontier, 2000: Dynamics of a feline virus with two transmission modes within exponentially growing host populations. *Proc. R. Soc. Lond. B: Biol. Sci.* 267, 2049–2056.

Briët, O. J., P. Vounatsou, D. M. Gunawardena, G. N. Galappaththy, and P. H. Amerasinghe, 2008: Models for short term malaria prediction in Sri Lanka. *Malaria J.* 7, 76–86.

Bryson, A. E., and Y. C. Ho, 1969: Applied Optimal Control: Optimization, Estimation, and Control. Blaisdell, Waltham.

Clement, J., P. Maes, and M. Van Ranst, 2006: Hantaviruses in the old and new world. *Perspect. Med. Virol.* 16, 161–177.

Clement, J., J. Vercauteren, W. W. Verstraeten, G. Ducoffre, J. M. Barrios, A. M. Vandamme, P. Maes, and M. Van Ranst, 2009: Relating increasing hantavirus incidences to the changing climate: the mast connection. *Int. J. Health Geogr.* 8, 1–11.

Clement, J., P. Maes, C. van Ypersele de Strihou, G. van der Groen, J. M. Barrios, W. W. Verstraeten, and M. van Ranst, 2010: Beechnuts and outbreaks of nephropathia epidemica (NE): of mast, mice and men. *Nephrol. Dial. Transplant.* 25, 1740–1746.

Clement, J., P. Maes, J. M. Barrios, W. W. Verstraeten, S. Amirpour Haredasht, G. Ducoffre, J.-M. Aerts, and M. Van Ranst, 2011: Global Warming and epidemic Trends of an emerging viral Disease in Western-Europe: the Nephropathia epidemica Case. In: Casalegno, S. (ed.), Case Studies on the Economy, Human Health, and on Urban and Natural Environments, pp. 39–52. InTech: Rijeka, Croatia.

Davis, S., E. Calvet, and H. Leirs, 2005: Fluctuating rodent populations and risk to humans from rodent-borne zoonoses. *Vector Borne Zoonotic Dis.* 5, 305–314.

Epstein, P. R., 2002: Climate change and infectious disease: stormy weather ahead? *Epidemiology*, 13, 373–375.

Escutenaire, S., P. Chalon, R. Verhagen, P. Heyman, I. Thomas, L. Karelle-Bui, T. Avsic-Zupanc, A. Lundkvist, A. Plyusnin, and P. P. Pastoret, 2000: Spatial and temporal dynamics of Puumala hantavirus infection in red bank vole (*Clethrionomys glareolus*) populations in Belgium. *Virus Res.* 67, 91–107.

Hanski, I., L. Hansson, and H. Henttonen, 1991: Specialist predators, generalist predators, and the microtine rodent cycle. *J. Anim. Ecol.* 60, 353–367.

Heyman, P., R. Van Mele, F. De Jaegere, J. Klingström, C. Vandenvelde, A. Lundkvist, F. Rozenfeld, and M. Zizi, 2002: Distribution of hantavirus foci in Belgium. *Acta Trop.* 84, 183–188.

Huitu, O., I. Jokinen, E. Korpimäki, E. Koskela, and T. Mappes, 2007: Phase dependence in winter physiological condition of cyclic voles. *Oikos*, 116, 565–577.

Hurrell, J. W., and R. R. Dickson, 2004: Climate variability over the North Atlantic. Marine Ecosystems and Climate Variation, pp. 15–32. Oxford University Press, Oxford.

Kallio, E., 2003: Stability of Puumala-virus outside the, 2nd European meeting on viral zoonoses, St Raphae.

Kallio, E., M. Begon, H. Henttonen, P. Koskela, and T. Mappes, 2009: Cyclic hantavirus epidemics in humans – predicted by rodent host dynamics. *Epidemics*, 1, 101–107.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.* 83, 95–108.

Korpimäki, E., K. Norrdahl, T. Klemola, T. Pettersen, and N. C. Stenseth, 2002: Dynamic effects of predators on cyclic voles: field experimentation and model extrapolation. *Proc. R. Soc. Lond. B: Biol. Sci.* 269, 991–997.

Korpimäki, E., K. Norrdahl, O. Huitu, and T. Klemola, 2005: Predator–induced synchrony in population oscillations of coexisting small mammal species. *Proc. R. Soc. Lond. B: Biol. Sci.* 272, 193–202.

Kramski, M., K. Achazi, B. Klempa, and D. H. Krüger, 2009: Nephropathia epidemica with a 6-week incubation period after occupational exposure to Puumala hantavirus. *J. Clin. Virol.* 44, 99–101.

Loeuille, N., and M. Ghil, 2004: Intrinsic and climatic factors in North-American animal population dynamics. *BMC Ecol.* 4, 6–17. Mathworks, Matlab, R2008a, Tutorial.

McCaughey, C., and C. A. Hart, 2000: Hantaviruses. *Med. Microbiol.* 49, 587–599.

Olsson, G. E., M. Hjertqvist, A. Lundkvist, and B. Hörnfeldt, 2009: Predicting high risk for human hantavirus infections, Sweden. *Emerg. Infect. Dis.* 15, 104–106.

Olsson, G. E., N. White, C. Ahlm, F. Elgh, A.-C. Verlemyr, P. Juto, and T. R. Palo, 2002: Demographic factors associated with Hantavirus infection in bank voles (*Clethrionomys glareolus*). *Emerg. Infect. Dis.* 8, 924–929.

Olsson, G., N. White, J. Hjalten, and C. Ahlm, 2005: Habitat factors associated with bank voles (*Clethrionomys glareolus*) and concomitant hantavirus in northern Sweden. *Vector Borne Zoonotic Dis.* 5, 315–323.

Oltjen, J. W., and F. N. Owens, 1987: Beef cattle feed intake and growth: empirical Bayes derivation of the Kalman filter applied to a nonlinear dynamic model. *J. Anim. Sci.* 65, 1362–1370.

Palo, R. T., 2009: Time series analysis performed on nephropathia epidemica in humans of northern Sweden in relation to bank vole population dynamic and the NAO index. *Zoonoses Public Health*, 56, 150–156.

Sauvage, F., C. Penalba, P. Vuillaume, F. Boue, D. Coudrier, D. Pontier, and M. Artois, 2002: Puumala hantavirus Infection in Humans and in the Reservoir Host, Ardennes Region, France. *Emerg. Infect. Dis.* 8, 1509–1511.

Sauvage, F., M. Langlais, N. G. Yoccoz, and D. Pontier, 2003: Modelling hantavirus in fluctuating populations of bank voles: the role of indirect transmission on virus persistence. *J. Animal Ecol.* 72, 1–13.

Sauvage, F., M. Langlais, and D. Pontier, 2007: Predicting the emergence of human hantavirus disease using a combination of viral dynamics and rodent demographic patterns. *Epidemiol. Infect.* 135, 46–56.

Sayre, N. F., 2008: The Genesis, History, and Limits of Carrying. *Ann. Assoc. Am. Geogr.* 98, 120–134.

Stenseth, N. C., A. Mysterud, G. Ottersen, J. W. Hurrell, K.-S. Chan, and M. Lima, 2002: Ecological effects of climate fluctuations. *Science*, 297, 1292–1296.

Talpaz, H., P. J. Sharpe, H. I. Wu, I. Plavnik, and S. Hurwitz, 1991: Modeling of the dynamics of accelerated growth following feed restriction in chicks. *Agric. Syst.* 36, 125–135.

Taylor, C. J., D. J. Pedregal, P. C. Young, and W. Tych, 2007: Environmental time series analysis and forecasting with the Captain toolbox. *Environ. Model. Software*, 22, 797–814.

Tersago, K., A. Servais, P. Heyman, G. Ducoffre, and H. Leirs, 2009: Hantavirus disease (nephropathia epidemica) in Belgium: effects of tree seed production and climate. *Epidemiol. Infect.* 137, 250–256.

Vapalahti, O., J. Mustonen, A. Lundkvist, H. Henttonen, A. Plyusnin, and A. Vaheri, 2003: Hantavirus infections in Europe. *Lancet. Infect. Dis.* 3, 653–661.

Yanagihara, R., L. Amyx, and D. C. Gajdusek, 1985: Experimental infection with Puumala virus, the etiologic agent of nephropathia epidemica, in bank voles (*Clethrionomys glareolus*). *J. Virol.* 55, 34–38.

Young, P., 1999: Data-based mechanistic modelling, generalised sensitivity and dominant mode analysis. *Comput. Phys. Commun.* 117, 113–129.

Young, P., D. Pedregal, and W. Tych, 1999: Dynamic harmonic regression. *J. Forecasting*, 18, 369–394.