An Experiment in Epidemiological Forecasting:

A Comparison of Forecast Accuracies among Different Methods of Forecasting Deer Mouse Population Densities in Montana

By

Robert A. Yaffee, New York University, New York, NY, USA

Kostas Nikolopoulos, Manchester Business School, Manchester, UK

David P. Reilly, Automatic Forecasting Systems, Inc., Hatboro, PA, USA

Sven F. Crone, Lancaster University, Lancaster, UK

Kent D. Wagoner, Ithaca College, Ithaca, NY, USA

Richard J. Douglass, Montana Tech, University of Montana, Butte, MT, USA

Brian R. Amman, Thomas G. Ksiazek, and James N. Mills, Special Pathogens Branch, United States Centers for Disease Control and Prevention, Atlanta, GA, USA

July 7, 2008

Abstract

Hantavirus pulmonary syndrome (HPS) is a zoonotic disease associated with the deer mouse (*Peromyscus maniculatus*) in the United States. Prediction of increases in deer mouse population density could result in advance warning of high-risk conditions and help direct public health interventions. We used temperature, precipitation, and monthly counts of rodents captured at a Montana trapping site to develop two causal and 16 univariate forecasting models as well as to compare their efficacy in predicting deer mouse population densities. A naïve model, which carried forward the last observed population abundance value, was used as a basis of comparison. We generated 12 forecasts over a rolling origin fixed three-month window over three years. Only the univariate state-space local level models outperformed the naïve model. Failure of the causal models may be related to their dependence on accurate forecasts of temperature and precipitation variables well into the 3-month forecast horizon. Future models will incorporate remote measures of habitat quality including satellite-derived indices of greenness and primary productivity.

Introduction

In the southwestern United States, in 1993, there was a sudden outbreak of severe respiratory disease with a case fatality exceeding 50%. This disease, now called hantavirus pulmonary syndrome (HPS), was shown to be caused by a previously unrecognized virus, Sin Nombre hantavirus (SNV), which is carried by the deer mouse (*Peromyscus maniculatus*). Following the development of diagnostic reagents and the education of physicians to recognize the disease, HPS was detected throughout the range of the deer mouse in the US. As of March 2007, 465 cases had been confirmed in 32 states (Fig. 1). Further investigations have now shown that there are many more hantaviruses in the Americas. At least 37 hantavirus genotypes have been described in association with an approximately equal number of rodent species in North, South, and Central America.

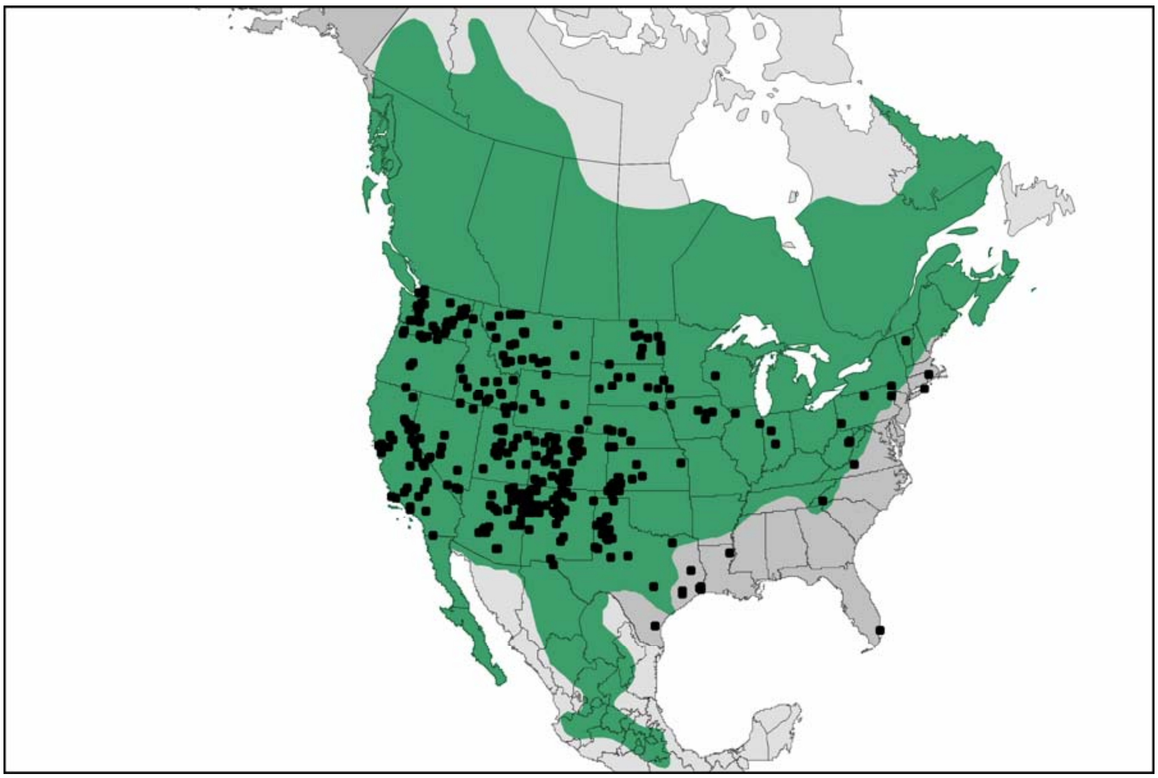


Figure 1: Distribution of the deer mouse in North America (shaded) and locations of 465 cases of hantavirus pulmonary syndrome in the USA as of April 1, 2007. Cases outside range of the deer mouse were associated with other hantaviruses and host species (Special Pathogens Branch, Centers for Disease Control and Prevention).

There is no vaccine and no specific treatment for HPS, so the best control measure is prevention. Because the development of the most effective intervention or prevention strategies depends on a thorough knowledge of the ecology of hantaviruses in nature, the U. S. Centers for Disease Control and Prevention (CDC) initiated ecological investigations of hantavirus hosts soon after SNV was discovered in 1993. An important part of these investigations has been a series of long-term mark-recapture studies of deer mouse populations in the southwestern United States (Mills et al., 1999), and Montana (Douglass et al., 2001). These studies have allowed investigators to follow deer mouse population densities and associated environmental variables (e.g., rainfall, temperature, habitat quality) over time.

Through analysis of the data from these studies we have been able to associate increases in rodent numbers at certain sites with increases in human cases of HPS (Yates et al. 2002). Knowledge of this association allowed successful prediction of periods of increased risk of human disease and the publication of advanced warning (CDC, 1998, 1999), possibly resulting in reduced morbidity and loss of life. Nevertheless, these predictions provided a short lead time, and they were only possible for localized areas where rodent populations were being monitored. In an effort to provide predictions with greater lead time over wider geographic areas, our attention recently has focused on the measurement of environmental factors associated with increases in rodent numbers (e.g., weather and habitat quality), and the development of mathematical models to predict rodent population abundance (Mills and Childs, 1998).

It has been hypothesized that increases in deer mouse population abundance are associated with environmental changes that follow a bottom-up, “trophic cascade” (Figure 1, Mills and Childs 1998, Yates et al. 2002). For example, the 1993 outbreak of HPS in the southwestern United States was preceded by an El Niño southern oscillation event that brought unusually high rainfall to the normally arid Southwest, resulting in improved vegetation quality, abundant food supplies and eventually, dramatic increases in rodent populations. The existence of such a cascade suggests that deer mouse population abundance (and thus human risk for HPS) might be forecast by quantifying the successive relationships among rainfall, food supply (as indexed by vegetation quality), and rodent population density and using weather variables as predictors.

Because of the universal availability of meteorological stations, weather (especially rainfall) is one of the easiest parameters in the trophic cascade to measure. Additionally, weather is the furthest “upstream” of any of the parameters in the cascade and thus, provides the greatest lead time for developing and implementing disease prevention interventions. Predictive models based on local weather data would use readily available data, be geographically specific, and provide the maximum advance warning. Nevertheless, we recognize that the same amount of rainfall can have totally different effects on vegetation (and thus on rodent population growth) depending on a multitude of factors including location, slope, topography, soil type, altitude, vegetation type, etc. Thus the stochastic effects of the numerous variables that intercede between weather and food supply might render weather variables as unreliable predictors. Because of their ready availability, long lead time, and potential utility, it was imperative to conduct a thorough and intensive forecasting experiment using weather variables and employ the newest and most powerful forecasting techniques available.

In the process of this test, we introduced a number of novel approaches, including the combination of causal and univariate methods, the comparison of state space methods to more traditional models, the use of four versions of the Theta method, the use of a rolling origin in the comparison of forecasting methods, and the use of measure of forecast accuracy that combined several standardized conventional criteria.

We hypothesized that we would be able to use variables derived from local weather conditions (temperature and precipitation) combined with monthly measurements of deer mouse population density to predict future deer mouse population density over a 3-month forecasting horizon. We invited four experts from three countries to use state-of-the-art forecasting tools that were in their area of expertise. Investigators were provided identical datasets including rodent abundance and weather variables and given 9 months to develop a predictive model. The approaches evaluated in this paper included 2 conditional and 16 univariate forecasting models. The accuracy of each model was assessed using identical criteria. We addressed three principal questions: (1) Can we develop an effective forecasting model using rodent abundance and local weather (temperature and precipitation) data? (2) Are causal models superior to univariate models? (3) Can we develop objective criteria for measuring forecast accuracy that allow us to rank our models in order of successful prediction?

This is one of the first attempts to bring together mathematicians, ecologists, and public health scientists to address an important public health problem. We hope that this preliminary attempt at combining our talents will result in improvements in the science in all three fields and lead to additional interdisciplinary collaborations to address problems related to the prevention of disease.

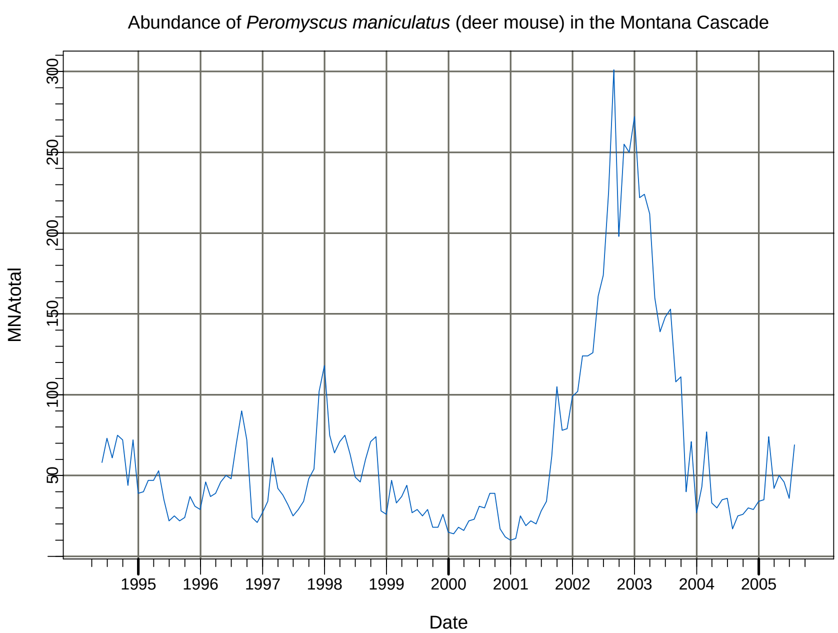
Methods

Measures and Data collection

Because of the completeness of the data, relative simplicity (low species diversity) of the rodent assemblage, and large numbers of deer mice, we chose the Cascade, Montana trapping site in Central Montana to test our forecasting tools.

Beginning in 1994, rodents were trapped on three 100 x 100m trapping grids consisting of 100 Sherman live-capture traps spaced at 10m intervals. The 3 grids were located 1 to 3 km apart in grassland habitat, at an elevation of 1250m, on a working cattle ranch. Trapping was conducted for 3 consecutive nights each month. Captured rodents were weighed and measured and affixed with a uniquely numbered ear tag so that they could be identified upon recapture in subsequent months. To measure population abundance, monthly captures were used to compute minimum number alive (MNA) from June 1994 through November 2005 (Fig. 2). MNA in month X is calculated as the number captured in month X plus those not captured in month X but captured in at least 1 previous month and at least 1 subsequent month (Chitty and Phipps, 1966).

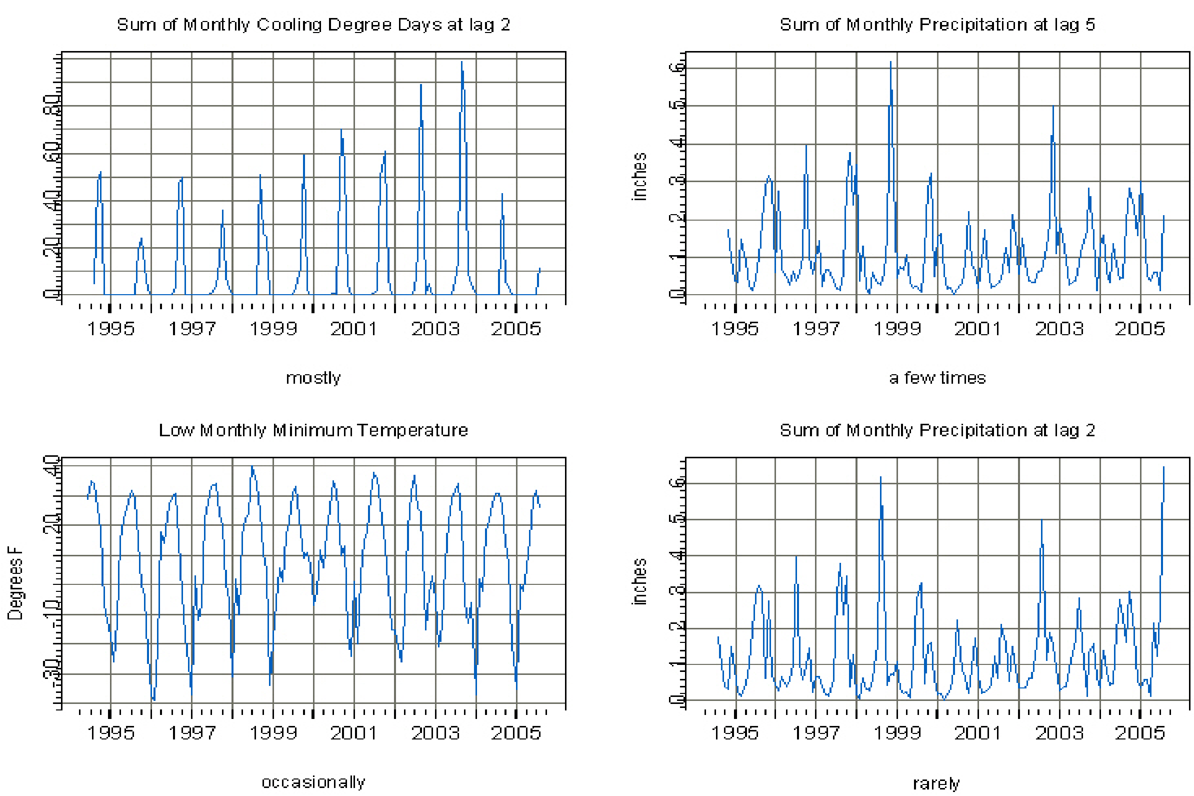
Figure 2: MNA time series for deer mice in Cascade, Montana



Possible predictor series included monthly weather data obtained from the National Oceanic and Atmospheric Administration weather station (Cascade 20SSE) 1 – 2 km from the trapping grids. The predictor series contained four statistically significant predictors. They included the sum of cooling degree days at lag 2 months, the sum of monthly precipitation at lag 5, the low minimum monthly temperature, and the sum of monthly precipitation at lag 2 (Fig. 3). The sum of cooling degree days per month is a measure of the amount of cooling needed for warm months. The number of cooling degree days is the difference obtained by subtracting 65° Fahrenheit from the average daily temperature, for each day that the average daily temperature surpasses that threshold (there are no negative cooling degree days). The sum of monthly cooling degree days is the sum of these differences (Source: National Oceanic and Atmospheric Administration World Wide Web site: http://answers.noaa.gov/noaa.answers/consumer/kbdetail.asp?kbid=348&p=t). Series that were not significant predictors were the sum of monthly snow (inches), maximum monthly temperature, minimum monthly temperature, average monthly temperature, sum of monthly heating degree days, and mean monthly minimum temperature (all temperatures in degrees F).

Because the abundance of deer mice is related to human risk, the level rather than the rate of change is of interest. In this form, this series was nonstationary, containing multiple level shifts, outliers, and end-effects (i.e., sudden, large level or slope shifts just before the point of forecast origin). Depending upon the model applied, there were at least 17 level shifts and 3 outliers within the series or forecast horizons.

Figure 3: Occasionally significant weather predictor series



Forecasting Models

In the search for a useful predictive model, 18 forecasting models were applied. Two conditional ("causal") forecasting models used readily available weather predictor variables. Those were 1) a transfer function model using Autobox version 6.0 and 2) a state space model using STAMP version 7.0, both of which could incorporate dynamic regressors. STAMP, in contrast to the single source of error state space method (Ord, K., Koehler, A. & Snyder, R, 1997), formulates the state space model with separate error terms for each random component. STAMP was used for state space modeling because it contained an augmented Kalman filter algorithm that could model and forecast nonstationary series. The augmented Kalman filter partitions the state vector into stationary and nonstationary components and then applies a diffuse prior distribution to initialize the weighting of the nonstationary components in the Bayesian sequential updating process. The regressor series observed to be statistically significant within some of these models were described above. Only the Autobox and Stamp “causal” models employed dynamic predictors.

Sixteen of the forecasting methods were univariate. These included 3) univariate Autobox, 4) Forecast Pro: expert system, which alternated between an ARIMA (0,1,0) or simple exponential smoothing model, 5) simple exponential smoothing (Stata was used because it performed exceptionally well compared to the National Institute of Standards and Technology Criterion datasets), 6) damped trend (SPSS v.15), 7) damped trend (Forecast Pro v 5), 8) Holt-Winters linear trend (Stata v. 9.2), 9) Holt-Winters additive seasonal (Stata), 10) Holt-Winters multiplicative seasonal (Stata), 11) state space local level model [STAMP-LLM], 12) state space local level model plus interventions (Stamp-LLM+interv), 13) original Theta AN (Assimakopoulos and Nikolopoulos) method, 14) Theta with optimal weighting, 15) generalized Theta, 16) Theta HB (Hyndman and Billah’s version in R), and 17) Neural Networks (Feedforward with an exhaustive grid search) (ANN FF). 18) A naïve model, in which the last actual MNA value was carried forward throughout the length of the three month forecast, is used as a basis for comparison.

The STAMP 7 and Autobox 6 causal models include dynamic regressors and interventions. The STAMP model includes a local level term, while the Autobox model includes a global intervention term. STAMP interventions include user-identified and modeled level shifts and additive outliers, while Autobox interventions include level shifts, additive outliers, and seasonal pulses. STAMP dynamic regressors, selected with PcGets version 1, generally included the sum of cooling degree days at lag 2 and occasionally the sum of precipitation at lag 5. Autobox dynamic regressors generally included the sum of cooling degree days at lag 2 and much less frequently the other three predictor series (low monthly minimum temperature, and sum of monthly precipitation at lags 2 and 5) displayed in Figure 3. Although the sum of cooling degree days at lag 2 was frequently significant, the size of its effect was very small. For example, an increase in one cooling degree day at lag 2 accounted for an increase in one more mouse tallied per month in the STAMP models.

The univariate models used different methods to forecast MNA. The state space models were run exclusively with STAMP because it uses an augmented Kalman filter, which can handle nonstationary time series. The state space models were based on a local level model, a random walk plus noise. When interventions were included in the univariate context, this reduced the model to a simple exponential smoother plus interventions. However, the interventions were identified and modeled by the user. Damped trend, without a seasonal component, forecasts were run with Forecast Pro and

SPSS. Forecast Pro was also used as an expert system, in which it alternated between an ARIMA (0, 1, 0) and a simple exponential smoother (SES) model. Four versions of the Theta model were tested: 1) The original Assimakopoulos and Nikolopoulos Theta model (Theta AN), 2) Hyndman and Billah's version of Theta in R (Theta HB), 3) an Optimal Weighting version which optimized weights on a three month out of sample segment of the data, and 4) a Generalized Theta, which combined 50% L(0) and 40% L(2) with additional Theta lines (L = -1,1,0,1,2,3) for the remaining 10% while optimizing on the last 12 months to obtain the forecast.

Forecasting Protocol

There were 138 months (11.5 years) in the MNA series, from June 1994 through November 2005, inclusive. We began forecasting three months ahead in November 2002 [month 102]. After each 3-month forecast, we rolled the forecast origin ahead for one season (three months). For example, the origin of the forecast was rolled ahead one (three month) season for each forecast [102, 105, . . . ,135] so that the next forecast extended over the next season. Seasons were defined as winter: December -- February; spring: March -- May; summer: June -- August; and fall: September -- November. The process was reiterated until there were no more forecasts to be generated. In sum, 12 three month *ex ante* forecasts were generated.

The comparative analysis was not done as a blind experiment. The decision to perform this comparative analysis was made after several of the forecasts had already been generated. Forecasters were not blind after generating their first complete set of forecasts and most forecasters submitted multiple sets of forecasts.

Principal Measures of Forecast Accuracy

We assessed the accuracy of forecasts with multiple measures. We used mean absolute error (MAE, eq. 1) to obtain an absolute measure of error, mean absolute percentage error (MAPE, eq. 2) to place MAE on a comparable scale, median absolute percentage error (MedAPE, eq. 3) to avoid outlier distortion, and percent worse than naïve (last value carried forward, eq. 4) to obtain a good relative measure.

$$Mean\ Absolute\ Error\ =\ \sum_{t=1}^{h}\left|\frac{A_t - F_t}{h}\right| \qquad (Eq.1)$$

$$\begin{aligned} &where \\ &A = actual \\ &F = forecast \\ &h = \ periods\ in\ forecast\ horizon \end{aligned}$$

$$Mean\ Absolute\ Percentage\ Error\ =\ \frac{100}{h}\sum_{t}^{h}\left|\frac{A_t - F_t}{A_t}\right| \qquad (Eq.2)$$

MedAPE: Midpoint of sorted MAPE for 3 month forecast horizon. (Eq. 3)

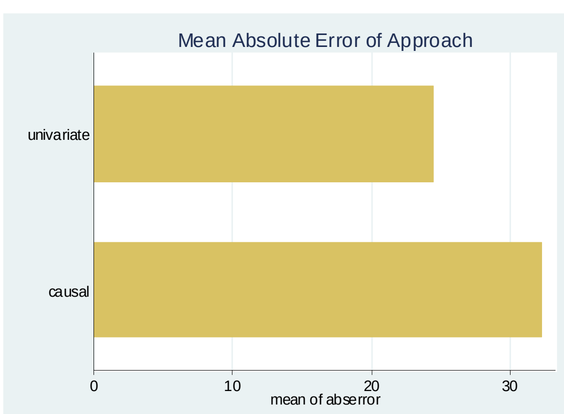
$$Percent\ worse\ =\ \frac{100}{N_f}\sum_{f=1}^{f} k_s \qquad \text{(Eq. 4)}$$

$$\begin{aligned} &where \\ &N_f\ = number\ of\ \ forecasts \\ &k_s = \begin{cases} 1\ if\ \ |F - A| < |F - A_{naive}| \\ 0\ otherwise \end{cases} \\ &A_{naive} = last\ value\, carried\ foreward \end{aligned}$$

Armstrong (2001) commented on the relative advantages and disadvantages of each of these measures. He notes that MAE has good face validity. It can provide a good baseline when forecasts are of very small numbers, but it contains no scale and is vulnerable to outlier distortion. By comparison, MAPE has a comparable scale, but it is vulnerable to distortion from outliers or forecasts of small numbers. For example, if there is but one mouse, a forecast of two mice can have a MAPE of 100%. MedAPE has good outlier protection, yet no control for difficulty in forecasting. Percent worse is the complement of percent better. As a percentage of counts of whether forecast is better than the absolute error of the naïve forecast, it should provide good outlier protection, good reliability, and maintain the same direction of accuracy (with more accurate measures having lower scores) as other measures. When the accuracies of models or methods are similar, percent worse may have difficulty distinguishing among them.

Results

We developed 18 forecasting models and compared their forecast accuracy using several measures. The 16 "univariate" models had a smaller mean absolute error and mean absolute percentage error than the 2 "causal" models (Figs. 4, 5).



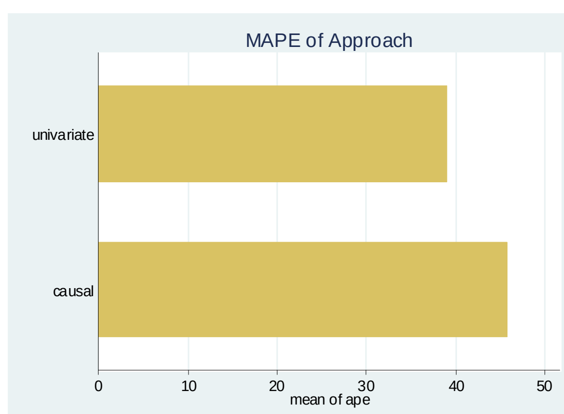


Figure 4: MAE by type

Figure 5: MAPE by type

However, outliers may have distorted the mean measures in the horizontal bar charts. Box-plots showing medians for the two classes of forecasts suggest that the causal models might slightly outperform the univariate models (Fig 6).

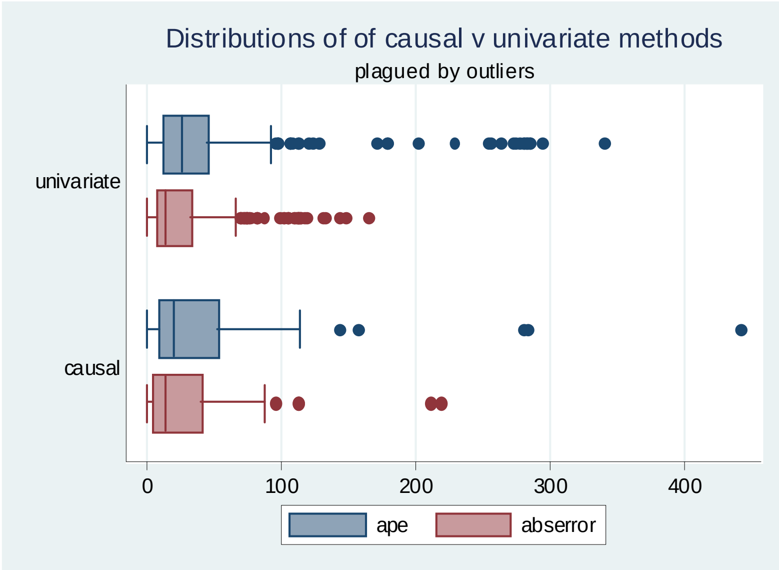


Figure 6: Box-plots of absolute percentage error and absolute error by type

We graphed forecast accuracy in terms of MAPE with error bars representing +/- 2 standard deviations (Fig 7). Because the whole population of forecasts is contained in our dataset, we did not require hypothesis testing with standard errors. Although these distributions may not be normal, the overlap of the error bars, suggests that there is no clear way to identify a single best forecast using MAPE.

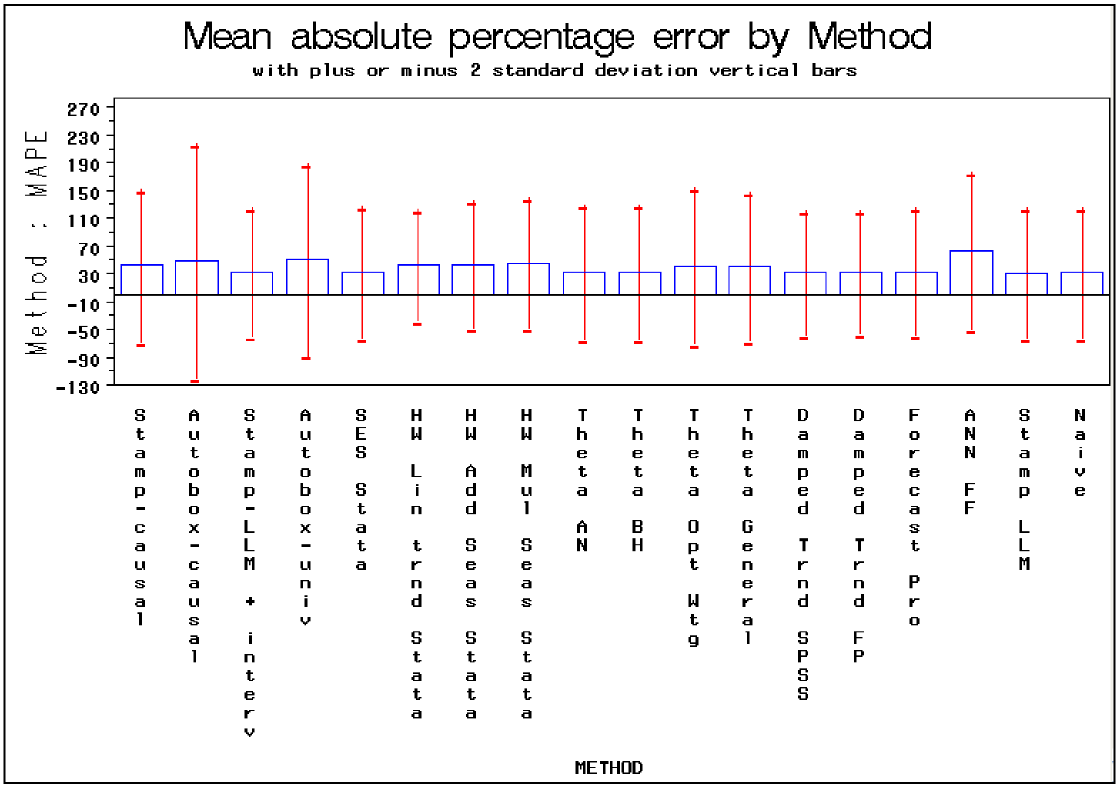
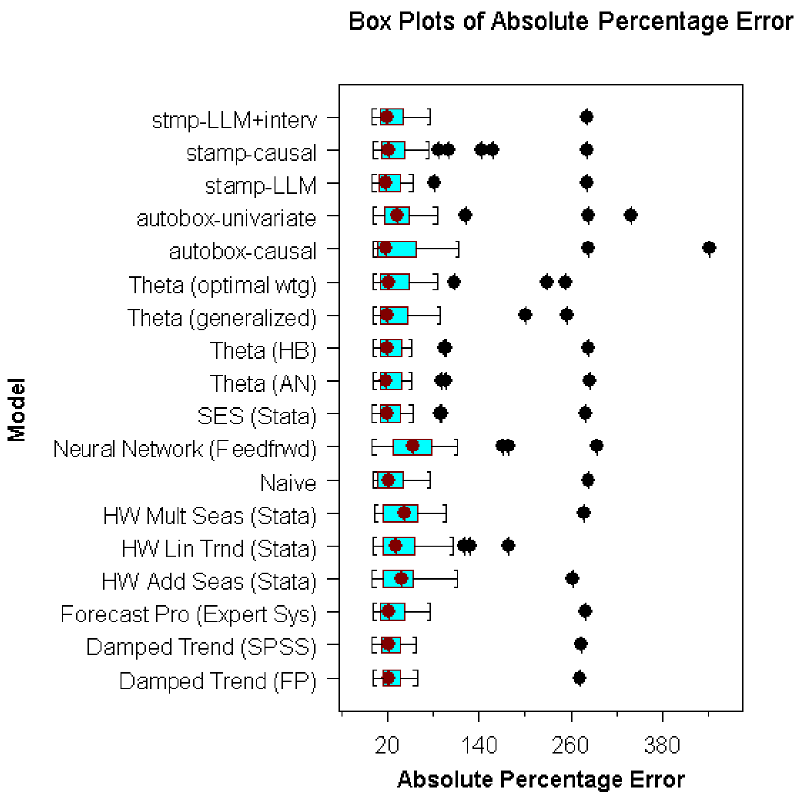


Figure 7: MAPE with plus or minus 2 standard deviation bars by method

To determine if MAPE is distorted by outliers we graphed the medians of the different methods, which are much more resistant to outlier distortion than the mean (Fig 8). The more extreme observations, on the right-hand sides of the distributions, were probably distorting the means, so we could not rely only on average measures. However, the interquartile ranges greatly overlapped, indicating that many of these methods had similar accuracy.

Figure 8: Box –Plot of MedAPE



We sorted our methods by forecast accuracy, as measured by MAE, MAPE, MedAPE, and Percent worse (Tables 1 – 4). Methods are sorted in order of decreasing accuracy for MAE (Table 1), MAPE (Table 2), MedAPE (Table 3), and percent worse (Table 4).

Table 1: Forecasting models sorted by MAE

Mean absolute error

Obs	METHOD	MAE	mae_rank
1	Stamp LLM	18.2729	1
2	Naive	18.2778	2
3	Stamp-LLM + interv	18.9329	3
4	SES Stata	19.7811	4
5	Theta AN	19.9575	5
6	Theta BH	20.0399	6
7	Forecast Pro	20.0679	7
8	Damped Trnd SPSS	20.2755	8
9	Damped Trnd FP	20.6913	9
10	Theta General	21.9275	10
11	Theta Opt Wtg	22.7314	11
12	HW Lin trnd Stata	23.3750	12
13	HW Add Seas Stata	27.2713	13
14	Autobox-causal	30.8313	14
15	HW Mul Seas Stata	32.2643	15
16	Stamp-causal	33.5151	16
17	Autobox-univ	33.9283	17
18	ANN FF	46.2160	18

Table 2: Forecasting models sorted by MAPE

Mean absolute percentage error

Obs	METHOD	MAPE	mape_ rank
1	Stamp LLM	30.8445	1
2	Damped Trnd FP	31.4732	2
3	Damped Trnd SPSS	31.5207	3
4	Naive	31.5959	4
5	Stamp-LLM + interv	32.1105	5
6	SES Stata	32.3526	6
7	Theta AN	32.5539	7
8	Forecast Pro	32.6062	8
9	Theta BH	32.7424	9
10	Theta General	39.8989	10
11	Theta Opt Wtg	41.4875	11
12	Stamp-causal	42.1822	12
13	HW Lin trnd Stata	43.3061	13
14	HW Add Seas Stata	43.5801	14
15	HW Mul Seas Stata	45.6779	15
16	Autobox-causal	49.2464	16
17	Autobox-univ	50.4155	17
18	ANN FF	63.6246	18

Table 3: Forecasting Models sorted by MedAPE

Median Absolute Percentage Error

Obs	METHOD	MedAPE	Med_rank
1	Stamp LLM	18.5468	1
2	Autobox-causal	19.0708	2
3	Theta AN	19.0981	3
4	SES Stata	20.0127	4
5	Theta BH	20.1417	5
6	Stamp-LLM + interv	20.1553	6
7	Theta General	20.4098	7
8	Damped Trnd SPSS	22.0576	8
9	Naive	22.1014	9
10	Stamp-causal	22.4163	10
11	Forecast Pro	22.7373	11
12	Theta Opt Wtg	22.8761	12
13	Damped Trnd FP	23.0143	13
14	HW Lin trnd Stata	32.0758	14
15	Autobox-univ	34.0471	15
16	HW Add Seas Stata	38.9371	16
17	HW Mul Seas Stata	42.3424	17
18	ANN FF	53.0188	18

Table 4: Forecasting models sorted by Percent Worse

Method sorted by PctWorse

Obs	METHOD	pctworse	pw_rank
1	Theta AN	47.222	2.0
2	Theta BH	47.222	2.0
3	Stamp LLM	47.222	2.0
4	Stamp-LLM + interv	52.778	5.5
5	SES Stata	52.778	5.5
6	Damped Trnd SPSS	52.778	5.5
7	Damped Trnd FP	52.778	5.5
8	Autobox-causal	55.556	8.0
9	Stamp-causal	58.333	10.0
10	Theta Opt Wtg	58.333	10.0
11	Theta General	58.333	10.0
12	Autobox-univ	75.000	13.0
13	HW Lin trnd Stata	75.000	13.0
14	HW Add Seas Stata	75.000	13.0
15	HW Mul Seas Stata	77.778	15.5
16	Forecast Pro	77.778	15.5
17	ANN FF	86.111	17.0
18	Naive	100.000	18.0

We also sought to distinguish one group of forecast methods from the others in accordance with measures of forecast accuracy—absolute error, absolute percentage error, and error squared. We used Classification and Regression Trees (CART 6.0), with a twoing splitting criterion, to split the methods according to levels of the forecast accuracy criteria. Because cross-validation revealed only 11.6% correct classification, these results were discarded.

Discussion

Our primary research question addressed the development of an effective forecasting model using locally available weather variables. Although two causal models were developed, they had lower accuracy than the naïve model that simply brought forward the previous MNA value across the 3-month forecast horizon. With a three month forecast horizon, we were not able to develop an effective forecast based on a causal model using weather (temperature and precipitation) predictor variables. To answer our second research question about the relative accuracy of causal versus univariate models, by most measures, the causal models were inferior to most univariate models. We will discuss possible reasons for the failure of the causal models below. Finally, we employed four criteria of forecast error to indicate forecast accuracy.

It was difficult to determine which models were better because the measures of forecast accuracy for so many of these models were almost identical. A Classification and Regression Trees (CART 6.0) program provided an overall correct classification of only 11.6%. Thus we were not able to use CART to define cut-points to construct an ordered typology of the preferred methods. Bagging (bootstrap aggregation) resulted in overall correct percentages of classification of less than 10 percent. The accuracies of these methods were very similar, given the size of the series and the number of forecasts generated.

The most accurate model for our difficult series depended upon the forecast accuracy criterion is used. If we employ the MAE, MAPE, and MedAPE, the state- space local level model (STAMP LLM) appears to be the most accurate. If we employ the percent worse measure, this method is tied with the original Theta developed by Assimakopoulos and Nikolopoulos along with the Billah and Hyndman versions of Theta method implemented in R. The top six to nine models, regardless of the forecast accuracy criteria employed, seem almost indistinguishable in forecast accuracy. Were we to look for those methods that appear among the top six ranked in terms of accuracy for at least three of these forecast accuracy criteria, we discover the state space (STAMP) local level model, state space local level model plus interventions (STAMP-LLM + interv), Theta

(AN) , Theta (BH), and the simple exponential smoother (Stata) methods. If we were to examine the top nine ranks for models that appeared in at least three criteria, we could add to the foregoing set damped trend (SPSS), damped trend (Forecast Pro), as well as the naïve model. Moreover, there is no simple, easily made, hard and fast cut-point among the accuracies indicated in the above tables.

Given the numerous level shifts and end-effects that characterize this series, it would be expected that models containing local level adjustment would provide the best forecasts and, indeed, the top performers in the first two groups included local smoothing. Two outstanding performers were state space models with the augmented Kalman filter, which allows modeling nonstationary series. The best forecasting models were the state space local level univariate model and the state space local level model with interventions. The state space models adjusted to the unstable level of the series by sequentially updating a state vector with a one-step-ahead autoregressive forecast of a state vector plus a regression on the innovation. This sequential updating in the state space model corrects for difference between the observed and the estimated value after a lag of one month and helps adjust to intervention effects. The local level model can be formulated as a simple exponential smoother (Harvey, 1992). The local smoothing adjustment in the Theta method, the Stata simple exponential smoother, and the damped trend models perform a similar function. A local smoothing capability was common to these outstanding models.

Several factors may help explain the failure of the causal models using weather variables. The weather variable that commonly turned out to be statistically significant in these models (sum of cooling degree days at lag 2) was associated with a very small increase in the mouse population. The forecasting models with weather predictors required preliminary prediction of some of the weather variables well into the 3-month forecasting horizon. However, accurate prediction of the weather is not currently possible beyond an approximately two week horizon (Trenberth, K., 2007; NCAR & UCAR, 2007). Attempting to predict for longer periods is to build error into the forecasts of the MNA. The aggregated error compounds the difficulty of making accurate forecasts of the MNAtotal.

Limitations

Our primary goal was to identify one or more methods that would provide useful forecasts of deer mouse population abundance. Because we only tested these methods on a single, particularly difficult series, we cannot assess the general capability of each method. We had neither a large sample of different series nor a large sample of forecasts. A sample of 12 clustered forecasts or 36 iterated forecasts was not large enough to distinguish one from the other in cases where the forecast accuracies of several of these methods were almost identical.

We used readily available weather variables to form our causal models. Success of one or more of these models would have allowed prediction of local deer mouse population abundance and, by inference, human disease risk with the greatest lead time, using reliable, widely available, and inexpensive data. However, identical rainfall and temperature values may have different effects on habitat quality depending upon location, topography, and vegetative cover. Lacking the success of weather based models, the next best alternative is likely to be to avoid the problem of varying effects of weather on vegetation by basing predictive models directly on measures of vegetation quality. Although they are more expensive than weather data, several satellite-derived indices (e.g., indices of greenness and plant primary productivity) provide relative measures of habitat quality and potential food availability. We plan to experiment with forecasting models based on these indices. Still, there are other factors besides weather and vegetation quality that affect monthly deer mouse populations. These include, but are not limited to predator-prey relationships, disease, interspecific competition, and intrinsic population processes that affect fecundity, territoriality, or migration. Thus, even more proximal models based on vegetation characteristics are not a guarantee of successful prediction.

Implications

In our models, the best predictor of future deer mouse population density was current deer mouse population density. This suggests that uninterrupted, regular sampling at frequent intervals of deer mouse populations at this time may provide useful data for local predictive models of deer mouse population density and associated risk of human disease due to infection with Sin Nombre virus.

Forecasts with shorter time horizons may provide increased accuracy, but such models would have limited utility because they provide little time for preparation of public health interventions. More research into time series models with mixed frequency of regressors might improve forecasting accuracy. Weather variables, especially precipitation, might more successfully be used as a predictor for rodent populations in areas, such as the southwestern United States, where populations are more likely to be moisture dependent. The comparative analysis performed on the data from Cascade, Montana should be tested on data from other sites to determine the applicability of our results to other deer mouse population datasets.

Eleven years represents a tremendous trapping effort, but is nevertheless a very short time series on which to base a forecasting model. Surges in rodent population density (which are often associated with increased risk of human disease; Yates et al. 2002) are infrequent -- our series contained only one large surge. Thus our 11-year data series distills to N=1 target events for epidemiological purposes. Although others might argue that there are two smaller surges, there appear to be too few such events to allow meaningful and reliable statistical modeling of associated ecological patterns and processes. Development of accurate predictive models for relatively rare public health emergencies, including outbreaks of HPS, may require data series measured in decades rather than a few years.

With the publication of these results we describe and analyze an intensive effort to address an important public health forecasting problem. We also provide a careful consideration of possible explanations for the limited success of our models and offer very important guidance for follow-up studies. We hope that the publication of these results will stimulate others to experiment with alternative approaches to forecasting human risk due to zoonotic agents and provide a background for considering which predictor variables and which modeling techniques are likely to be useful.

Acknowledgments

We thank Drs. Pierre Rollin, and C.J. Peters for CDC support of this research. We also thank our field collaborators for essential assistance, encouragement, and intellectual stimulation: Dr. Terry Yates, Cheryl Parmenter, Dr. Bob Parmenter, Dr. Chris Hice, Dr. Charlie Calisher, Dr. Jeff Root, Jeff Doty, Dr. Ken Abbott, Rachelle Macintosh, Dr. Amy J. Kuenzi, Dr. Andy Hopkins, and Kevin Hughes. We would like to thank David H. Hendry, Siem Jan Koopman, Hans-Martin Krolzig, Andrew C. Harvey, Neil Ericsson, Jurgen Doornik, Eric Zivot, Ralph Snyder, and Rob Hyndman. We would like to thank several consultants from SAS, Stata, and Insightful for their helpful suggestions. From SAS, Inc. we would also like to thank Donna Woodward, Leonardo Auslender, and Phil Gibbs for helpful suggestions concerning programming. From Stata, we would like to thank Brian Poi, Kerry Kammire, and Dave Drukker for helpful comments. From Insightful, we would like to thank Rachel Flodin, Brian Russell.

References:

Armstrong, J.S. and Collopy, F (1992) Error Measures for generalizing about forecasting methods: empirical comparisons, IJF (4), 69-90.

Armstrong, J.S. (2001). “Evaluating Forecasting Methods,” in Armstrong, J.S. (Ed). Principles of Forecasting, Kluwer Academic Publishers: Norwalk, MA. 444-472.

Assimakopoulos, V. and Nikolopoulos, K. (2000). The Theta Model: A Decomposition Approach to Forecasting, International Journal of Forecasting, Vol. 16. No. 4, 521-530.

Auslender, L.E. (1998). Alacart: Poor Man’s Classification Trees. SUGI 1998. SAS Institute. Cary: NC.

Autobox User’s Guide v.6. (2006). Automatic Forecasting Systems, Inc., Hatboro. Pa.

Brieman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees. Chapman and Hall/CRC: New York.

CART (Classification and Regression Trees) User's Guide (1997). Salford Systems, San Diego, CA.

CDC. (1998). Hantavirus Pulmonary Syndrome - Colorado and New Mexico, 1998. Morbidity and Mortality Weekly Report 47:449-452,

CDC. (1999). Update: hantavirus pulmonary syndrome -- United States, 1999. Morbidity and Mortality Weekly Report 48 (24):521-525,

Chitty D, and Phipps E, 1966. Seasonal changes in survival in mixed populations of two species of Voles. J. Animal Ecol. 35:313-331.

Clements, M. and Hendry, D.F. (eds). (2004). A Companion to Economic Forecasting. Blackwell Publishing: Malden, MA.

Douglass, R. J., Wilson, T., Semmens, W. J., Zanto, S. N., Bond, C. W., Van Horn, R. C., and Mills. J. N. (2001). Longitudinal studies of Sin Nombre virus in deer mouse dominated ecosystems of Montana. American Journal of Tropical Medicine and Hygiene. 65 (1):33-41.

Durbin, J. and Koopman, S.J. (2000) Time Series Analysis by State space Methods. Oxford, UK: Oxford University Press, pp. 1-47, 52-53, 65-81, 115-120.

Gardiner, Jr., E.S. (1999). Rule-Based Forecasting vs. Damped-Trend Exponential Smoothing. Management Science, Vol. 45. No. 8. 1169-1176.

Goodwin, P. & Lawton, R (1999). On the asymmetry of symmetric MAPE. International Journal of Forecasting (15) 405-408.

Hamilton, J.D. (1994). Time Series Analysis. Princeton University Press: Princeton, NJ., 372-408.

Harvey, A. (1992). Forecasting Structural Time Series Models and the Kalman Filter, Cambridge University Press: New York, Chapter 2.

Harvey, A. (1993) Time Series Models. MIT Press: Cambridge, MA, Chapters 1 - 4.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning. Springer: New York, 266-296.

Hendry, D. F. and Krolzig, H.-M. (2001). Automatic Econometric Model Selection using PcGets. Version 1.0. User's Guide, Timberlake Consultancy, Ltd: London, UK.

Koopman, Harvey, Shepard and Ericsson. (2006) Stamp 7. Structural Time Series Analyser, Modeler and Predictor. Timberlake Consultancy, Ltd.: London.

Hyndman, R. and Billah, H. (2003). Unmasking the Theta Method, International Journal of Forecasting, Vol. 19. Number 2, 287-290.

Kuenzi, A., Douglass, R., and Bond, C. (2000). Sin Nombre Virus in Deer Mice Captured inside Homes, Southwestern Montana. Emerging Infectious Diseases, 6(4):.386-388.

Krebs, C.J. (2001). Ecology: The Experimental Analysis of Distribution and Abundance. 5th ed. Benjamin Cummings, Menlo Park, California.

Lewellen, R. H. & Vessey, S. H. (1998). The Effect of Density Dependence and Weather on Population Size of a Polyvoltine Species. Ecological Monographs, 68(4):571-594.

Mills, J. N. et al, (1995) Guidelines for working with Rodents potentially infected with Hantavirus. Journal of Mammology. (76) 716-712.

Mills, J. N. and Childs, J.E. (1998). Ecologic Studies of Rodent Reservoirs: Their Relevance for Human Health. Emerging Infectious Diseases. 4(4):529-537

Mills, J. N., Yates, T. L., Ksiazek, T. G., Peters, C. J., and Childs, J. E. (1999). Long-term studies of hantavirus reservoir populations in the southwestern United States: rationale, potential, and methods. Emerging Infectious Diseases 5 (1):95-101.

Merritt, J.F., Maricio, L. & Bozinovic, F. (2001). Seasonal Regulation in Fluctuating Small Mammal Populations. Oikos, (94), pp. 505-514.

Maindonald, J. and Braun J. (2003). Data Analysis and Graphics Using R. Cambridge University Press: New York: NY., 242-252.

Mathews, B.P., & Diamontopoulos, A. (1994). Toward a taxonomy of forecast error measures: A factor-comparative investigation of forecast error dimensions, JOF (13), 409-416.

National Center for Atmospheric Research and the University Corporation for Atmospheric Research Office of Programs (NCAR & UCAR). "The Edge of Predictability: Long Range Weather Forecasting," http://www.ucar.edu/research/prediction/prediction.jsp, June 30,2007.

Ord, J.K., Koehler, A.B., and Snyder, R.D. (2007). "Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models." *Journal of the American Statistical Association.*, 92, 1621-1629.

Snyder, R.D. (2006) personal communication. June 23, 2006.

Stata 9 (2006) Time Series Reference Manual, StataCorp. College Station, TX.

Stellwagon, E. and Goodrich's. (2006). Forecast Pro User's Guide, Business Forecast Systems, Belmont, MA.

Steinberg, D. (2001) .CART: Tree Structured Non-Parametric Data Analysis. Manual. Salford Systems, Inc. San Diego, CA., 23-27.

Stroble, C., Zeileis, A. Boulesteix, AL, and Hothorn, T. (2005) Bias in Variable Importance Measures of Ensemble Methods Based on Classification Trees. World Wide Web:
http://www.r-project.org/useR-2006/Abstracts/Strobl+Zeileis+Boulesteix+Hothorn.pdf.

Taylor, J.W. (2003). Exponential Smoothing with a Damped Multiplicative Trend. International Journal of Forecasting (19).715-725.

Trenberth, K.E. (2007). Issues and Prospects for prediction of climate on multiple time scales. International Symposium on Forecasting Keynote Address: New York, NY. June 26, 2007.

Yates, T. L., Mills, J. N., Parmenter, C. A., Ksiazek, T. G., Parmenter, R. R., Vande Castle, J. R., Calisher, C. H., Nichol, S. T., Abbott, K. D., Young, J. C., Morrison, M. L., Beaty, B. J., Dunnum, J. L., Baker, R. J., Salazar-Bravo, J., and Peters. C. J. 2002. The ecology and evolutionary history of an emergent disease: hantavirus pulmonary syndrome. Bioscience 52 (11):989-998,

Ye, J. (1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. Journal of the American Statistical Assn. Vol.. 93. No. 441. (Mar. 1998), 120-131.

Zivot, E. and Wang, J. (2006). Modeling Financial Time Series with S-Plus. 2nd ed. Springer: New York., 519-569.

Appendix I

Principal Formulae of five of the more accurate methods on the CDC Montana Deer Mouse Series: June 1994 to November 2005. The lowest number in ranking is most accurate.

1.

state space local level model

an exponentially weighted moving average :

$y_t = ln(MNAtotal)$

$measurement\ equation: \ y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma^2_{\varepsilon_t})$

$transition\ equation: \ \mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim NID(0, \sigma^2_{\eta_t})$

$initialization: \ \mu_1 \sim N(\mu_1, P_1)$

$estimation: augmented\ Kalman\ filter\ sequential\ updating\ of\ mean\ and\ variance$

$where\ \mu_t = level\ and\ \varepsilon_t = irregular\ are\ unobserved\ components$

$of\ state\ vector, \alpha. \ \alpha_1 = initial\ mean\ of\ state\ vector. \ P_1 = initial\ state\ variance.$

Source: Durbin, J. and Koopman, S.J(2001) . Time Series by State Space Methods, Oxford University Press : Oxford,UK,chapter 1.

2.

state space local level model with interventions

an exponentially weighted moving average :

$y_t = \ln(mnatotal)$

$$measurement\ equation: \ y_t = \mu_t + \sum_{\mathbf{t=1}}^{\mathbf{k}} \mathrm{w}_t \mathrm{I}_{t\text{-}\tau} + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma^2_{\varepsilon_t})$$

$$transition\ equation: \ \mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim NID(0, \sigma^2_{\eta_t})$$

$$initialization: \ \mu_1 \sim N(\mu_1, P_1)$$

estimation : augmented Kalman filter sequential updating of mean and variance

where $\mu_t = level$ *and* $\varepsilon_t = irregular$ *are unobserved components*

of state vector, α.

$\alpha_1 =$ *initial mean of state vector.*

$P_1 =$ *initial state variance.*

$w_t =$ *Intervention (outlier or level shift) parameter estimate.*

$I_{t\text{-}\tau} =$ *Intervention at time t with lag* τ.

Source : Durbin, J. and Koopman, S.J(2001) . Time Series

by State Space Methods, Oxford University Press : Oxford,UK,chapter 1.

3.

Naïve Model: Last value of MNAtotal carried forward over three month forecast horizon.

4.

Theta AN model: The original version of the Assimakopoulos, V. and Nikolopoulos, K. Theta model, in which the forecast is average of simple exponential smoother and linear regression line derived from MNAtotal series data.

5.

Stata Simple Exponential Smoother:

$$S_t = \hat{\alpha} x_t - (1-\hat{\alpha}) S_{t-1} \quad for\ \ t = 1, \ldots, T$$

$S_t =$ *forecast of series* S_{t-1}

$\hat{\alpha} =$ *estimated smoothing parameter*

$x_t =$ *current observation value*

$S_0 =$ *initial value*

Source: Stata Time Series Reference Manuel, Release 9 (2005), StataCorp, College Station, Tx., 247.