ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Forecasting Compositional Time Series with Exponential Smoothing Methods

Anne B. Koehler, Ralph D. Snyder, J. Keith Ord and Adrian Beaumont

November 2010

Working Paper 20/10

Forecasting Compositional Time Series with Exponential Smoothing Methods

Anne B. Koehler[1], Ralph D. Snyder[2], J. Keith Ord[3] and Adrian Beaumont[1]

Abstract

Compositional time series are formed from measurements of proportions that sum to one in each period of time. We might be interested in forecasting the proportion of home loans that have adjustable rates, the proportion of nonagricultural jobs in manufacturing, the proportion of a specific oxide in the geochemical composition of a rock, or the proportion of an election betting market choosing a particular candidate. A problem may involve many related time series of proportions. There could be several categories of nonagricultural jobs or several oxides in the geochemical composition of a rock that are of interest. In this paper we provide a statistical framework for forecasting these special kinds of time series. We build on the innovations state space framework underpinning the widely used methods of exponential smoothing. We couple this with a generalized logistic transformation to convert the measurements from the unit interval to the entire real line. The approach is illustrated with two applications: the proportion of new home loans in the U.S. that have adjustable rates; and four probabilities for specified candidates winning the 2008 democratic presidential nomination.

Keywords: compositional time series, innovations state space models, exponential smoothing, forecasting proportions

JEL classification: C22

[1] Department of Decision Sciences and MIS, Miami University, Oxford, OH 45056, USA. KOEHLEAB@muohio.edu
[2] Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia. ralph.snyder@monash.edu
[3] McDonough School of Business, Georgetown University, Washington, DC 20057, USA. ordk@msb.edu

1 Introduction

The need to forecast proportions arises in wide variety of areas, such as business, economics, geology, and political science. Specific examples include the proportion of home loans that have an adjustable rate, the proportion of nonagricultural jobs that are in manufacturing, the proportion of a rock's geochemical composition that is a specified oxide, and the proportion of an election betting market choosing a particular candidate. Special care must be taken when forecasting proportions. Standard forecasting methods allow the point forecast and/or the prediction intervals to range outside of the interval (0,1).

Whenever a time series y_t is a proportion, the time series $1 - y_t$ is also a proportion, and the sum of the two time series is, of course, 1. *Compositional Time Series* extend this property to include two or more positive time series whose values must sum to 1 at every time period. When the number of time series exceeds two, the forecasting procedure must guarantee that the forecasts of the individual time series lie within (0,1) and always sum to 1. Traditional multivariate forecasting methods do not have this important feature. The previous examples can obviously be extended to compositional time series with more than two time series by considering proportions for several categories of nonagricultural jobs, several oxides in the geochemical composition of a rock, and several candidates for the same office.

A key reference for the study of compositional time series is a book by Aitichison (1986). Transforming compositional time series by using log ratios is the main focus of that book and the approach used in our paper. Aitchison & Egozcue (2005) describe the development of compositional time series analysis over the previous twenty years and the importance of the log ratio method. A variety of methods have been applied to the analysis of compositional time series. A Bayesian approach can be found in work by Quintana & West (1988). It is not appropriate to consider a direct approach where the original (i.e. untransformed) time series is assumed to follow a Dirichlet distribution because that would require an assumptions of independence for these time series. However, Grunwald et al. (1993) have developed a new distribution for $y_{it}, i = 1, \ldots, m$, that is based on the Dirichlet distribution and does not have the restrictive independence assumption. We will be applying a vector exponential smoothing model of Hyndman et al. (2008) to a vector of log ratios of the time series. A related study has been done for vector ARIMA models in Brundson & Smith (1998).

In the next section of the paper, we describe the generalized logistic model that incorporates the vector exponential smoothing model. In the

third section, we examine an example of a forecasting a single proportion and an example of forecasting compositional time series. The paper ends with a section of conclusions.

2 Generalized logistic model for compositional time series

Compositional time series are time series $y_{1t}, y_{2t}, \ldots, y_{mt}$ with the restrictions that

$$0 < y_{it}$$

and

$$\sum_{i=1}^{m} y_{it} = 1$$

A model for compositional time series that combines the generalized logistic transformation and a vector exponential smoothing model is

$$\begin{aligned} y_{it} &= \begin{cases} \frac{\exp(z_{it})}{1+\sum_{k=1}^{m-1}\exp(z_{kt})} & i = 1, \ldots, m-1 \\ \frac{1}{1+\sum_{k=1}^{m-1}\exp(z_{kt})} & i = m \end{cases} && \text{(1a)} \\ \mathbf{z}_t &= \boldsymbol{\ell}_{t-1} + \boldsymbol{\varepsilon}_t && \text{(1b)} \\ \boldsymbol{\ell}_t &= \boldsymbol{\ell}_{\boldsymbol{t-1}} + \mathbf{A}\boldsymbol{\varepsilon}_t && \text{(1c)} \end{aligned}$$

where $\mathbf{z}_t = (z_{1t}, \ldots, z_{m-1,t})'$, $\boldsymbol{\ell}_t = (\ell_{1t}, \ldots, \ell_{m-1,t})'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{m-1,t})'$, and $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. We assume that $\{\boldsymbol{\varepsilon}_t\}$ and $\{\boldsymbol{\varepsilon}_{t+j}\}$ are independent for all $j \neq 0$. The matrix $\mathbf{A}$ is a diagonal matrix.

Equations 1b and 1c form the vector local level model in which each z_{it} has a univariate model that underlies simple exponential smoothing. These two equations can be replaced by other vector exponential smoothing models that allow for trend and seasonal patterns in the vector of time series. The general vector exponential smoothing model is

$$\begin{aligned} \mathbf{z}_t &= \mathbf{W}\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t \\ \mathbf{x}_t &= \mathbf{F}\mathbf{x}_{t-1} + \mathbf{A}\boldsymbol{\varepsilon}_t \end{aligned}$$

where $\mathbf{x}_{t-1}$ is a k-dimensional state vector of unobserved components, $\mathbf{W}$ is an $(m-1) \times k$ matrix, and $\mathbf{F}$ is an $(m-1) \times ((m-1))$ matrix. In the rest of this paper, we restrict our attention to the general logistic model with the vector local level model.

We now examine the process of forecasting compositional time series. Assume that we have observed values of the compositional time series $y_{i,t}, i = 1, \ldots, m$, from time period $t = 1$ to n. The first step is to transform the values at each time period t using log ratios as follows:

$$z_{i,t} = \ln\left(\frac{y_{it}}{y_{mt}}\right) \quad i = 1, \ldots, m-1.$$

This transformation alters the forecasting process from a restricted m-dim space to an unrestricted $(m-1)$-dim space. We will return later to a discussion of the invariance for the choice of the time series in the denominator. In the $(m-1)$-dim space, the point forecast of $\mathbf{z}_{n+h}$ and the covariance matrix of the h-step-ahead forecast error at time $t = n$ are

$$\hat{\mathbf{z}}_{n+h|n} = \boldsymbol{\ell}_n$$

and

$$\boldsymbol{\Sigma}_{n+h|n} = \boldsymbol{\Sigma} + (h-1)\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}.$$

The state vector $\boldsymbol{\ell}_t$ is calculated by exponential smoothing (where $\mathbf{e}_t$ denotes the estimated error) as follows:

$$\begin{aligned} \hat{\mathbf{z}}_{t|t-1} &= \boldsymbol{\ell}_{t-1} \\ \mathbf{e}_t &= \mathbf{z}_t - \hat{\mathbf{z}}_{t|t-1} \\ \boldsymbol{\ell}_t &= \boldsymbol{\ell}_{t-1} + \mathbf{A}\mathbf{e}_t. \end{aligned}$$

Using the generalized logistic transformation, the point forecasts of $y_{i,n+h}, i = 1, \ldots, m$, at time n are

$$\hat{y}_{i,n+h|n} = \begin{cases} \frac{\exp(\hat{z}_{i,n+h|n})}{1+\sum_{k=1}^{m-1}\exp(\hat{z}_{k,n+h|n})} & i = 1, \ldots, m-1 \\ \frac{1}{1+\sum_{k=1}^{m-1}\exp(\hat{z}_{k,n+h|n})} & i = m \end{cases}$$

In order for the forecasts to be invariant with respect to the choice for the time series y_{mt} in the log ratio transformation when there are more than two time series, we need to make restrictions to Model 1 and to the estimation procedure in the m-1-dim space. First, the diagonal values of $\mathbf{A}$ must be the same or, equivalently, replace $\mathbf{A}$ in the model by a constant α. Second, we used the mean of the first 10 values for $\mathbf{z}_t$ as the estimate for the initial state vector $\boldsymbol{\ell}_0$. Other estimates may be used, but the generalized logistic transformation of $\hat{\boldsymbol{\ell}}_0$ must produce the same m-dim vector no matter which

time series is chosen to be y_{mt}. Then, the maximum likelihood estimate for $\mathbf{A}$ can be found by minimizing $\log|\hat{\mathbf{\Sigma}}|$, where

$$\hat{\mathbf{\Sigma}} = n^{-1}\sum_{t=1}^{n}\mathbf{e}_t\mathbf{e}_t'.$$

The *prediction region* for $\mathbf{z}_{n+h}$ can be described as follows:

$$Q_{n+h|n} = (\mathbf{z}_{n+h} - \boldsymbol{\ell}_n)'\mathbf{\Sigma}_{n+h|n}^{-1}(\mathbf{z}_{t+h} - \boldsymbol{\ell}_n).$$

where

$$Q_{n+h|n} \sim \chi^2_{m-1}$$

The individual prediction intervals for $z_{i,n+h}$ can be found by using the following:

$$z_{i,n+h} \sim N(\ell_{i,n}, \sigma^2_{i,n+h|n})$$

where

$$\sigma^2_{i,n+h|n} = \sigma_{ii}(1 + (h-1)a^2_{ii}).$$

When there are two time series, the logistic transformation can be used to find prediction intervals for each $y_{i,n+h}, i = 1, 2$. However, when there are more than two time series, the generalized logistic transformation cannot be applied to the endpoints of these prediction intervals for $z_{i,n+h}, i = 1, \ldots, m-1$, to find prediction intervals for $y_{i,n+h}, i = 1, \ldots, m$, The reason for this situation is that the sum of the upper limits of the intervals for time series $y_{i,n+h}, i = 1, \ldots, m-1$, could exceed 1; making it impossible analytically to find a prediction interval for $y_{m,n+h}$. On the other hand, we can apply an exponential transformation to the prediction interval for $z_{i,n+h}$ to obtain a prediction interval for $y_{i,n+h}/y_{m,n+h}, i = 1, \ldots, m-1$.

3 Examples

3.1 Adjusted Rate Loans (ARL)

In this example, we examine forecasting the proportion of loans for new homes in the U.S.A. that are adjusted rate loans (ARL). The focus here is on a single time series y_t. The time period for this monthly time series is January 1985 to October 2008. A display of the time series can be seen in Figure 1.

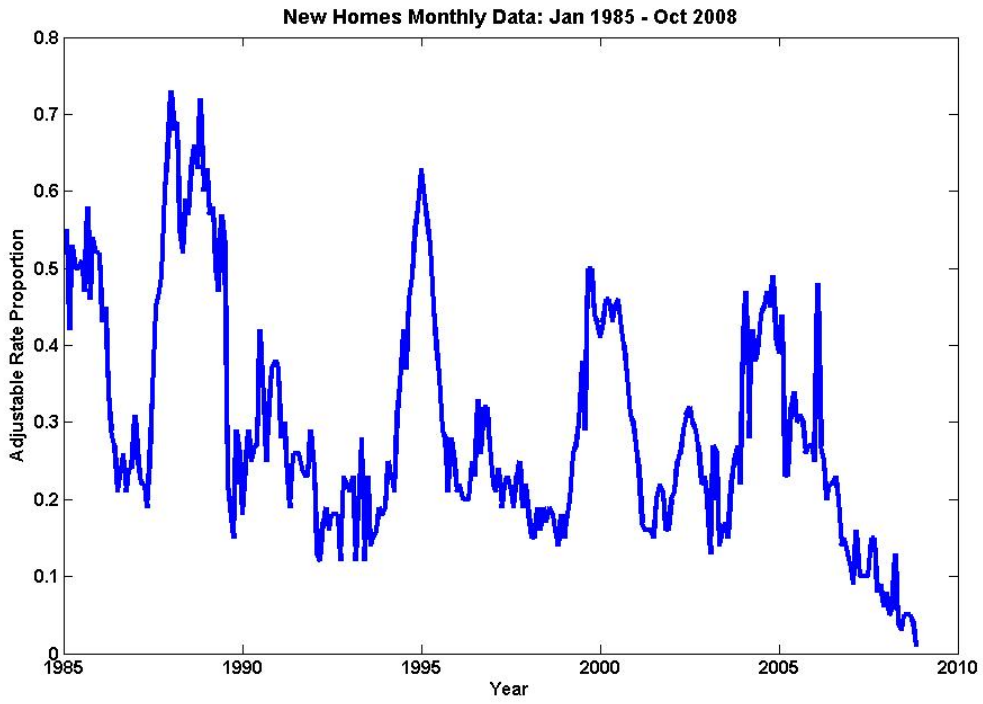

Figure 1: Proportion of loans for new homes in the U.S.A. that are adjusted rate loans, monthly from January 1985 to October 2008.

If the values were not restricted to lying between 0 and 1, an appropriate model for this time series would be the local level model that underlies simple exponential smoothing. The local level model is

$$y_t = \ell_{t-1} + \varepsilon_t \quad (2a)$$
$$\ell_t = \ell_{t-1} + \alpha\varepsilon_t \quad (2b)$$

The point forecast of y_{n+h} is ℓ_n and the prediction intervals can be found using $y_{n+h} \sim N(\ell_n, \sigma^2(1 + (h-1)\alpha^2)$. The point forecasts for Model 2 would always remain within the restricted range, but prediction intervals would allow impossible values.

Forecasting a single proportion is, of course, the same as forecasting a compositional time series with $y_{1t} = y_t$ and $y_{2t} = 1 - y_t$. Thus, using Model 1 to forecast the proportion of new home loans that have an adjusted rate will resolve this potential problem for the prediction intervals. To illustrate the problem with Model 2 and the correction in the prediction interval using Model 1, we withheld the last 12 months of the ARL data and forecasted these months with both models. Figure 2 includes the actual values, the model fits to the historical data, and the forecasts and prediction interval from both Models 1 and 2 . The point forecasts from the two models happen to be the same for these time periods, but clearly the prediction intervals using the standard local level model are wider than those using the logistic transformation. Moreover, the lower limits for the standard local level model drop below 0. In this example, the estimates for both the ℓ_0 and α are maximum likelihood estimates that are found simultaneously.

It is important to check that the accuracy of the forecasts is not reduced when we use the logistic transformation. We would not normally expect the point forecasts from the two models to be the same as they are in the example for Figure 2. The two measures of forecasting accuracy in our study are defined in the next section.

3.2 Measures of forecasting accuracy

The measures of forecasting accuracy that are employed in this paper are the absolute scaled error and the continuous ranked probability score. The absolute scaled error (ASE) (Hyndman & Koehler, 2006) is

$$\text{ASE} = \frac{|e_{t+h|t}|}{MAE_t}$$

where the h-period-ahead forecast error at time t is

$$e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$$

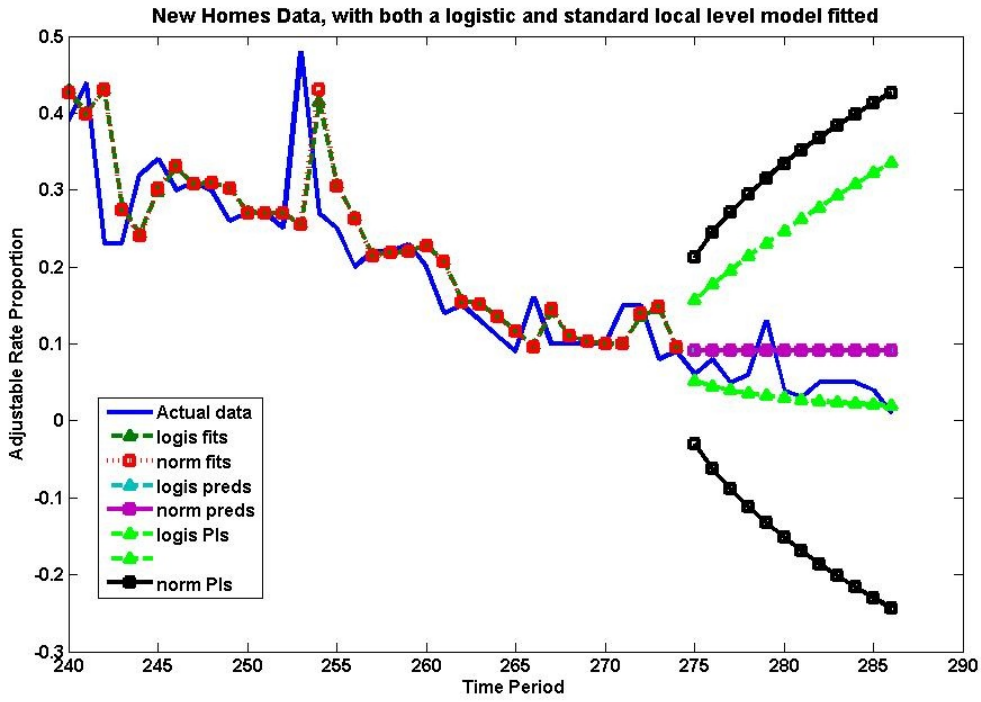

Figure 2: Point forecasts and prediction intervals for the Adjusted Rate Loan time series using the logistic model (1) and the standard local level model (2).

and

$$\text{MAE}_t = \frac{1}{t-1}\sum_{i=2}^{t}|y_i - y_{i-1}|$$

The continuous ranked probability score (CRPS) was chosen in order to compare entire prediction distributions on which both the point forecasts and prediction intervals would depend.The CRPS measures the closeness of two distributions in terms of the squared difference between their cumulative distribution functions and is defined by

$$\text{CRPS}(F, y_{obs}) = \int_{-\infty}^{\infty}(\delta(y \geq y_{obs}) - F(y))^2 dy. \tag{3}$$

$F(y)$ is the cumulative distribution function for y_{t+h} and is based on the model and estimates of the initial value and parameters. Also, y_{obs} is the the observed value of y_{t+h}, and $\delta(y \geq y_{obs})$ equals 1, if argument is true and 0 otherwise. Hence, Equation 3 may be re-written as

$$\text{CRPS}(F, y_{obs}) = \int_{-\infty}^{y_{obs}} F^2(y) + \int_{y_{obs}}^{\infty}(1 - F(y))^2 dy \tag{4}$$

In the case of the standard local level model, y_{t+h} would have a normal distribution with mean $\hat{y}_{t|t+h}$ and standard deviation $\sigma_{t|t+h}$. When $F(y)$ is the cumulative distribution function for the normal distribution, the CRPS has the following form (Gneiting & Raftery, 2007):

$$\begin{aligned}\text{CRPS}(F, y_{obs}) &= \sigma_{t|t+h}\left[2\phi\left(\frac{y_{obs} - \hat{y}_{t|t+h}}{\sigma_{t|t+h}}\right) - \frac{1}{\sqrt{\pi}}\right. \\ &\quad \left. + \left(\frac{y_{obs} - \hat{y}_{t|t+h}}{\sigma_{t|t+h}}\right)\left(2\Phi\left(\frac{y_{obs} - \hat{y}_{t|t+h}}{\sigma_{t|t+h}}\right) - 1\right)\right] \end{aligned} \tag{5}$$

where ϕ is the standard normal density function, and Φ is the cumulative normal distribution function. In order to have positive values for the CRPS, we are using the negative of the CRPS in Gneiting & Raftery (2007).

The distribution of y_{t+h} is unknown for the logistic model. For this model, an approximate CRPS using quantiles in the transformed space can be found as follows:

- For large R, find R ordered quantiles, $z^{(i)}, i = 1, \ldots, R$, in the normal distribution for z_{t+h} which has mean $\hat{z}_{t+h|t}$ and standard deviation $\sigma_{t+h|t}$.
- Use the logistic transformation to find R values, $y^{(i)}, i = 1, \ldots, R$.

- Identify the value, $y^{(k)}$ that is immediately below the observed value of y_{t+h}.
- Based on Equation 4, calculate

$$\begin{aligned} \text{CRPS}(F, y_{obs}) &= \sum_{i=1}^{k-1} \left(\frac{i}{R}\right)^2 d_i + \left(\frac{k}{R}\right)^2 (y_{obs} - y^{(k)}) \\ &\quad + \left(1 - \frac{k}{R}\right)^2 (y^{(k+1)} - y_{obs}) \\ &\quad + \sum_{i=k+1}^{R-1} \left(1 - \frac{i}{R}\right)^2 d_i \end{aligned} \tag{6}$$

  where $F(y^i) = i/R$ and $d_i = y^{(i+1)} - y^{(i)}$.

3.3 Comparison for forecasts for ARL time series

In this section, we compare Forecasts for the ARL times series from Models 1 and 2. We would hope that the gain in meeting the (0,1) restriction for point forecasts and prediction intervals will not result in a lack of forecasting accuracy. The comparisons will be done with both the ASE and the RPS. Thirty-six observations were withheld from the end of the time series, and rolling forecasts were found for 1 to 12 periods ahead. Thus, there are 24 forecasts for each forecasting horizon, 1 to 12. At each time period, the parameters were re-estimated for each model.

In Table 1, the mean and median ASE are presented for each forecast horizon. The results in Table 1 show that there is no loss of accuracy in the point forecasts by using a logistic transformation. In Table 2 , the mean and median CRPS are shown for each of the 12 forecast horizons. The distribution of y_{t+h} in the standard local level model is a normal distribution, and hence, CRPS in Equation 5 can be used. However, the distribution of y_{t+h} in the logistic model does not have a normal distribution. Thus, for logistic model, the approximate CRPS in Equation 6 must be used. The CRPS is always smaller for the logistic model than for the standard local level model in Table 2 .

3.4 Iowa Election Market

As an example of compositional time series with more than two time series, we look at the Iowa Election Market for the U.S.A. democratic presidential

Horizion	Median		Mean	
	logistic	standard	logistic	standard
1	0.507	0.496	0.870	0.886
2	0.922	0.943	1.092	1.119
3	0.934	0.985	1.176	1.201
4	0.953	0.998	1.112	1.133
6	1.149	1.147	1.407	1.416
12	1.720	1.781	2.129	2.147

Table 1: ARL: mean and median ASE for the logistic model (Model 1) and the standard local level model (Model (2)

Horizion	Median		Mean	
	logistic	standard	logistic	standard
1	0.016	0.018	0.031	0.034
2	0.026	0.028	0.036	0.039
3	0.028	0.030	0.039	0.043
4	0.028	0.032	0.037	0.040
6	0.034	0.038	0.044	0.048
12	0.052	0.056	0.065	0.069

Table 2: ARL: mean and median CRPS for the logistic model (Model 1) and the standard local level model (Model 2).

nominee in 2008 (http://tippie.uiowa.edu/iem/markets). Four time series are selected that are the probabilities of winning the democratic nomination for each of the following candidates: Clinton, Obama, Edwards, and Other Candidates. Each time series is composed of seventy-seven weekly observations (closing price on Sunday night; scaled to add to 1) from March 4, 2007 to August 17, 2008. These time series are displayed in Figure 3.

Clearly applying a vector local level model directly to the four time series would not be reasonable because once values are known for three of the time series, the fourth value is known also. A standard modification that might be applied to the vector local level model is the following:

$$y_{mt} = 1 - \sum_{i=1}^{m-1} y_{it} \tag{7a}$$

$$\mathbf{z}_t = \boldsymbol{\ell}_{t-1} + \boldsymbol{\varepsilon}_t \tag{7b}$$

$$\boldsymbol{\ell}_t = \boldsymbol{\ell}_{\boldsymbol{t-1}} + \mathbf{A}\boldsymbol{\varepsilon}_t \tag{7c}$$

where $\mathbf{z}_t = (y_{1t}, \ldots, y_{m-1,t})'$. Other definitions and assumptions are the same as in Model 1. We will refer to this model as the "standard modified vector local level model."

In order for the forecasts from this model to be invariant to the choice of time series for y_{mt}, the requirements that we imposed for Model 1 are necessary here. Hence, $\mathbf{A}$ must be replaced by a single parameter α, and we again estimated $\boldsymbol{\ell}_0$ by the mean of the first 10 observations of $\mathbf{z}_t$. For this model, we need an estimate $\hat{\boldsymbol{\ell}}_0$ for $\boldsymbol{\ell}_0$ that has the property that the m values of $\hat{\ell}_{i,0}, i = 1, \ldots, m-1$, and $1 - \sum_{i=1}^{m-1} \hat{\ell}_{i,0}$ are independent of the choice for y_{mt}.

Model 7 still has problems for compositional time series because it could generate values for the time series y_{mt} that would be negative or greater than 1. However, we used forecasts from Model 7 in a comparison with forecasts from the Model 1 to examine the effects of the generalized logistic transformation on the point forecasts.

For this comparison, the last thirty observations were withheld, and rolling point forecasts were found for 1 to 6 periods ahead. Thus, there were 25 point forecasts for each horizon from 1 to 6 periods ahead. In both models, the time series for Obama was selected as time series $y_{4,t}$, and the parameters and initial values were re-estimated at each time period. The mean and median ASE are used to compare the forecasts from the modified vector local level model and the generalized logistic model (Model 1) for each of the four time series. The results for 1 to 4 periods ahead can be seen in Table 3, and they indicate that the generalized logistic model produce

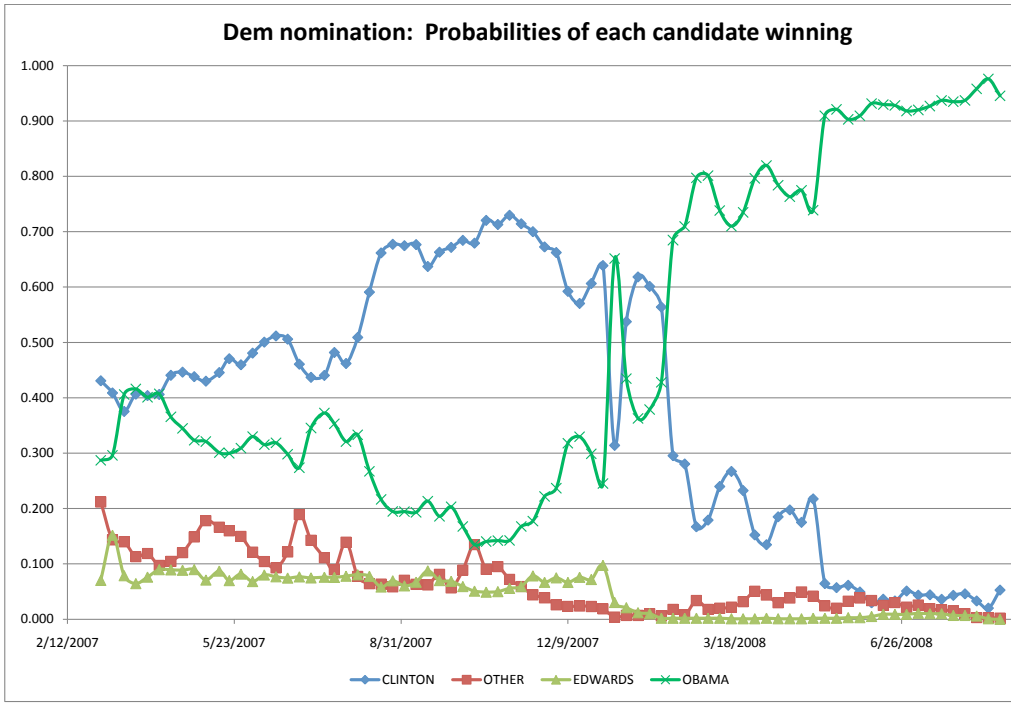

Figure 3: Weekly Iowa Election Market probabilities of candidates winning the democratic nomination for president of the U.S.A from March 4, 2007 to August 10, 2008

Candidate	Horizon	Median		Mean	
		logistic	standard	logistic	standard
Clinton	1	0.528	0.789	1.123	1.297
	2	0.774	0.810	1.839	1.880
	3	1.216	1.340	2.428	2.447
	4	1.158	1.170	2.449	2.482
Obama	1	0.640	0.659	1.027	1.175
	2	0.679	0.697	1.718	1.775
	3	1.137	1.238	2.236	2.273
	4	0.884	0.836	2.255	2.301
Edwards	1	0.036	0.068	0.100	0.158
	2	0.088	0.100	0.166	0.223
	3	0.099	0.122	0.219	0.275
	4	0.117	0.150	0.251	0.308
Other	1	0.430	0.453	0.483	0.451
	2	0.499	0.418	0.579	0.559
	3	0.582	0.522	0.618	0.590
	4	0.472	0.542	0.576	0.578

Table 3: Mean and median ASE for the generalized logistic model (Model 1) and the standard modified vector local level model (Model 7).

more accurate forecasts for the single candidates over these short forecast horizons.

4 Conclusions

Standard exponential smoothing models may not be appropriate for single time series that must lie within the unit interval (0,1) unless the values are near the middle of the interval. The logistic model is a reasonable alternative to the local level model because the point forecasts seem to be as accurate as the those from the local level model, and point forecasts and prediction intervals will always be contained within the unit interval (0,1). A vector exponential smoothing model, even the modified version, is not appropriate for compositional time series with more than two time series. The generalized logistic model is a reasonable model for compositional time series in this case.

References

Aitchison, J. & Egozcue, J. J. (2005). Compositional data analysis: where are we and where are we should we be heading? *Mathematical Geology*, *37*, 829–850.

Aitichison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Brundson, T. M. & Smith, T. M. F. (1998). The time series analysis of compositional data. *Journal of Official Statistics*, *14*, 237–253.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction,and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Grunwald, G. K., Raftery, A. E., & Guttorp, P. (1993). Time series of continuous proportions. *Journal of the Royal Statistical Society Series B*, *55*, 103–116.

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688.

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with Exponential Smoothing*. Berlin: Springer.

Quintana, J. M. & West, M. (1988). Time series analysis of time series data. In: Bernado, J. M., de Groot, M. H., Lindley, D. V., & Smith, A. F. M. (eds.), *Bayesian Statistics 3*, pp. 747–756, New York: Oxford University Press.