NCCC-134
APPLIED COMMODITY PRICE ANALYSIS, FORECASTING AND MARKET RISK MANAGEMENT

Price Density Forecasts in the U.S. Hog Market: Composite Procedures

by

Andres Trujillo-Barrera, Philip Garcia,
and Mindy Mallory

Price Density Forecasts in the U.S. Hog Market: Composite Procedures

Andres Trujillo-Barrera

Philip Garcia

Mindy Mallory

Paper presented at the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management
St. Louis, Missouri, April 22-23, 2013

Andres Trujillo-Barrera is assistant professor in the Marketing and Consumer Behaviour Group at Wageningen University, Philip Garcia is the T. A. Hieronymus Distinguished Chair in Futures Markets, and Mindy Mallory is Assistant Professor in the Department of Agricultural and Consumer Economics at the University of Illinois at Urbana-Champaign.

Price Density Forecasts in the U.S. Hog Market: Composite Procedures

Abstract

We develop and evaluate quarterly out-of-sample individual and composite density forecasts for U.S. hog prices using data from 1975.I to 2010.IV. Individual forecasts are generated from time series models and the implied distribution of USDA outlook forecasts. Composite density forecasts are constructed using linear and logarithmic combinations, and several straightforward weighting schemes. Density forecasts are evaluated on goodness of fit (calibration) and predictive accuracy (sharpness). Logarithmic combinations using equal and mean square error weights outperform all individual density forecasts and all linear combinations. Comparison of the USDA outlook forecasts to the best logarithmic composite demonstrates the consistent superiority of the composite procedure, and identifies the potential to provide hog producers and market participants with accurate expected price probability distributions that can facilitate decision making.

Keywords: Density Forecast Combination, Commodity Price Analysis.

Introduction

In a variety of settings agricultural economists have been involved in developing better price forecasts to assist decision makers (Leuthold et al., 1970; Brandt and Bessler, 1981; Bessler and Kling, 1986; Zapata and Garcia, 1990; Wang and Bessler, 2004; Colino et al., 2012). Often the research has provided innovative techniques for generating and combining point forecasts, and considerable evidence exists that combining forecasts dominates the best individual forecasts. While these studies have been informative, they have focused primarily on the mean of the expected price distribution and have often used an out-of-sample root mean squared error measure calculated ex post to assess the degree of risk in prices. An attractive alternative, which potentially can provide decision makers with a more precise assessment of the expected price distribution, involves direct density forecasts and the use of density composite methods. The importance of density forecasts for agricultural commodity prices was recognized as early as the mid-1960s by Bottum (1966) and Timm (1966), but their subsequent development and use have been scarce. Research has mainly focused on estimation procedures (Fackler and King, 1990; Sherrick, Garcia, and Tirupattur, 1996; Egelkraut, Garcia, and Sherrick, 2007) with less attention on forecast evaluations and comparisons.

To date, no studies exist that examine the usefulness of composite density forecast procedures in the agricultural commodity price literature.

Research in economics on forecasting and combining densities is emerging (Elliott and Timmermann, 2008). Kascha and Ravazzolo (2010) contend that "our knowledge of when and why predictive density combinations work is still very limited," recognizing the need to increase our understanding of the circumstances when procedures are most effective. Timmermann (2006) identifies a number of viable alternatives to formulate density composites, however, there is no consensus as to the most effective method. The search for effective methods for density forecasts and their combinations is empirically motivated by the need to develop robustness to unknown instabilities, and a reduction of idiosyncratic biases (Aastveit et al., 2012).

Density forecasts offer information about the uncertainty of the predictions, and a more precise description of risk and its dynamics to decision makers (Corradi and Swanson, 2006; Hall and Mitchell, 2007; Geweke and Amisano, 2010). Clements (2004) argues the desire for more than traditional point forecasts is in line with the increasing recognition that assessing uncertainty is an indispensable part of the forecasting and decision making. Recently Colino et al. (2012) investigated the predictive accuracy of quarterly point forecasts of hog prices from outlook programs, futures prices, and a variety of time series models in hopes of providing more precise information to decision makers. These forecasts traditionally have been an important source of information for hog industry producers, meat packers, and retailers. Their findings confirm the difficulty in generating individual accurate forecasts and the value of composite procedures in improving the predictive performance. Density forecasts and their composites may further help improve the information available to decision makers that emerges from market and outlook forecasts.

In this study we investigate the usefulness of newly developed density forecast, evaluation, and combination procedures for generating ex-ante distributions of quarterly hog prices. The density forecasts are developed for one and two quarters ahead using time series models and data for the period 1975 to 2010 (Colino et al., 2012). We also construct implied distributions of USDA outlook hog price forecasts based on the distribution of historical forecast errors (Isengildina-Massa et al., 2011). To evaluate and compare density forecasts we use sharpness and calibration (Mitchell and Wallis, 2011). Sharpness refers to the precision of a forecast and is measured by a forecast's log score which reflects the extent a forecast attains a high density value at the actual observation. Calibration refers to the correct distributional specification, i.e., whether the density of the forecast resembles the true distribution, and it is measured using the probability integral transform (PIT). To combine density forecasts we compare linear and logarithmic pooling, and several weighting schemes. The relative predictive accuracy of the forecasts is also assessed based on differences of log scores using a variant of the Diebold and Mariano test. The aim is to provide a more complete description of hog price forecast distributions over time, and to evaluate, compare, and combine predictive distributions that may help improve USDA hog price forecasting performance and the information provided to decision makers.

Our findings indicate that the out-of-sample performance of USDA forecast is poor compared to individual time series and composite forecasts. We find that forecasting performance improves by combining forecast densities, but the performance strongly depends on the distribution and weighting scheme used. Logarithmic pooling using either mean square error weights (MSEW) or equal weights (EW) significantly improves the sharpness over individual forecasts; however, neither linear combinations nor recursive log score weights improve over the best individual forecasts.

We also observe that the performance evaluation criteria of the individual forecasts differ between point forecasts evaluated using root mean squared error (RMSE), and density forecasts evaluated with average log scores. In fact, the model with the smallest RMSE does not provide the best log score measures. Evaluation of the combination results shows that logarithmic pooling outperforms linear pooling both in terms of sharpness and calibration. The MSE weighted composite consistently outperforms the recursive and adaptive combination procedure, but is virtually indistinguishable from the simple equal weighted composite. The difficulty in outperforming the equally weighted procedure corresponds to the well-known combination puzzle identified in point forecasting (Smith and Wallis, 2009), and points to the added error that more sophisticated procedures may add to the forecast process.

Forecasting Models

Individual Forecasting Models and Data

Economic density forecasts like their point and interval counterparts can arise from subjective expert opinions, time series models, forward looking markets, and simulations. Without doubt densities based on time series models have been the most commonly used procedures in the literature, but forecasts based on expert opinions are also prevalent in macroeconomics. For agricultural prices, time series models are the primary method to generate price forecasts. Outlook programs also provide expert opinions of future prices, but they have focused primarily on point forecasts (see Colino, Irwin, and Garcia (2011) for a review). For instance, USDA provides interval forecasts of several commodity prices including hog prices, but does not offer a distribution over the interval. However, indirect procedures can be used to obtain the implied distribution of the experts' forecast. For instance, Isengildina-Massa et al. (2011) show that implied densities can be obtained by estimating the distribution of forecast error.

Quarterly data spanning the period 1975.I-2010.IV are used to generate and evaluate hog price density forecasts. Univariate time series models use prices obtained from USDA; a Vector Autoregression (VAR) model incorporates pork production, sows farrowing, and beef prices obtained from USDA, and a quarterly average corn price assembled from Barchart (Figure 1). For USDA expert forecasts for one quarter (h=1) and two quarters ahead (h=2), are taken from USDA Livestock, Dairy, and Poultry Outlook. Data sources are provided in Table 1.

Colino, Irwin, and Garcia (2011) provide an in-depth analysis of the specification of ARMA and VAR models applied to this dataset. After testing for the stationarity, identifying the lag structure, and examining residuals, they chose an autoregression of order 5 (AR(5)) and an unrestricted vector autoregression of order 5 VAR(5). The VAR (5) specification is consistent with previous hog price forecasting literature, and is the result of a thorough process of examining potential variables, structural changes, and preliminary estimations of reduced VARs. Further, these procedures performed well in the point forecast composite analysis conducted by Colino et al. (2012) for the same period. As a result, we adopt and re-estimate the AR(5) and VAR(5) models for our analysis. Implied densities are generated from forecast errors of the models under normality which we could not reject in tests of the residuals.[1] Another forecast used by Colino et al. (2012) is No change or naive, this forecast takes the value of the current observation, and the variance of the forecast error is developed using the last 40 observations.

The U.S. hog industry has undergone structural change during the last two decades. Looking at figure1, it is evident there is a positive trend in pork production, while sow farrowing has remained relatively constant. Colino, Irwin, and Garcia (2011) argue that the significant production expansion is the result of new breeding technologies and capital concentration, particularly in the 1990s. As a result of this structural change, we also include an exponential smoothing model to allow added flexibility as more recent information may have added importance in generating forecasts than more distant past observations.

Exponential smoothing techniques are a class of forecasting methods with the property that forecasts are weighted combinations of past observations, with recent observations given relatively more weight than more distant observations. The fact that weights decrease exponentially is reflected in its name (Hyndman and Khandakar, 2008). The classical time series decomposition of the systematic (non random) part of an observation into three elements, trend, seasonality, and level, is used by exponential smoothing methods. Several functional forms may arise depending on whether the interaction of the three elements is additive or multiplicative, and also if trend, and/or level are included. We may have models that include both trend and seasonality, and others that only include one element. In our forecast estimation using exponential smoothing, the optimal values of the smoothing parameters were calculated for each period.

While the previous time series models have been tested in other venues and worked well, they may not be optimal. Rather, they represent reasonable models that have been applied frequently to forecast hog prices. Since our primary focus is not on finding the best possible specification for each individual model, but rather to investigate the additional information that density forecasts and their composites offer, they should provide a reasonable structure for analysis.

In our analysis, we are interested on the performance of experts' forecasts from USDA, and the potential gains that can be achieved by forecast combination. However, USDA does not provide density forecasts, only intervals. Therefore, we construct the USDA implied density forecasts by estimating the distribution of its forecast errors, following a procedure used by Isengildina-Massa et al. (2011) to build confidence intervals for USDA corn, soybean and wheat price forecasts.

To identify and estimate the forecast error distribution of the USDA hog price forecasts we follow several steps. We start by plotting a histogram of forecast errors, and find that the distribution resembles a bell shaped curve (figure 2). Next, we use a Cullen and Frey graph (Cullen and Frey, 1999). As seen in figure 3 this graph allows us to identify the choice of distributions given the skewness-kurtosis of the whole sample. Based on figure 3, the distribution of forecast errors has little skewness and kurtosis close to three. As a result, among other competing distributions such as

[1]Test of residuals failed to identify ARCH effects or autocorrelation.

logistic, uniform, exponential, gamma, and beta, the normal is the closest theoretical distribution to the USDA forecast errors.[2]

Using the normal for the forecast errors distribution, we fit the data using maximum likelihood estimation. As seen in figure 2, the QQ-plot, empirical and theoretical cdf, and PP-plot appear to confirm the goodness of fit of the normal distribution. In addition, to evaluate goodness of fit Anderson-Darling statistics are calculated. Both the normal and the logistic provide good fits, but AIC and BIC criteria both favor the normal distributions (e.g., normal AIC=913.63 $\leq$ logistic AIC= 916.65).

We are interested in generating forecasts from 1985.I until 2010.IV to match the forecasts generated from the time series models. To generate the empirical density for 1985.I we use the forecast errors from 1975.I until 1984.IV. For the next period in 1985.II, we use the forecast errors from 1975.I until 1985.I, and so on. We use the same evaluation of the errors distribution as previously described. The Anderson-Darling statistic never rejects normality, and as a result we use a parametric fit of the forecast errors with a normal distribution.

Evaluation of Density Forecasts

Accurate density forecasts should fulfill two requirements, producing probability estimates that are correct, and generating estimates that give a high density value at the observation. The first requirement is called calibration, while the second is called sharpness (Mitchell and Wallis, 2011).

Evaluating density forecasts is complicated because the true density is not observed, not even after the realization of the forecasted variable. Nevertheless, several approaches have been developed to test the calibration of density forecasts. Diebold, Gunther, and Tay (1998) proposed the use of the probability integral transform (PIT) as a measure of goodness of fit across all forms of probabilistic forecasts. The PIT is defined as:

$$PIT_t = \int_{-\infty}^{Y_{t+n}} f(y_t)dy \equiv F(Y_{t+n}), \tag{1}$$

where $f(y_t)$ and $F(Y_t)$ are the probability and cumulative density functions of variable y_t, and Y_{t+n} is the realized value at the forecast horizon n. As shown in Rosenblatt (1952), given the true data generating process, the cumulative densities at the realizations will be uniform. Similarly, if a density forecast is correctly specified, then the probability integral transform of the series of realizations is uniformly distributed, and in the case of one-step-ahead forecast, also independently and identically distributed (iid). Therefore, testing the uniformity and iid characteristics of the PIT series serves as a test of correct specification of the distribution, independent of a particular loss function, and overcoming the problem of not observing the true distribution.

However, tests of uniformity have low power, and as an alternative, Berkowitz (2001) proposed a transformation of the series from uniformly to normally distributed, arguing that tests for normal

[2]As a robustness check 5000 bootstraps are included for the skewness-kurtosis location of the sample distribution.

distribution have more power. We evaluate individual and combined densities' goodness of fit (calibration) by means of the Berkowitz test where the original PIT series is transformed. Let ϕ^{-1} be the inverse of the standard normal distribution. If a sequence of PIT_t is iid and U(0,1), then $z_t = \phi^{-1}(PIT_t)$ is iid and N(0,1). By using a likelihood ratio, independence and normality can be jointly tested (Berkowitz, 2001).

In addition to calibration we are also interested in measuring sharpness. For that we use scoring rules, defined as functions of predictive distributions and realized outcomes used to evaluate predictive densities (Gneiting and Raftery, 2007; Bjornland et al., 2011). The log score is the logarithm of the probability density function of the forecast evaluated at the realized value. Scoring rules serve as a way to compare forecasts by measuring the distance between the true distance and the possibly misspecified model. This is based on the link between logarithmic scores with the Kullback-Leibler information criterion (KLIC). KLIC measures the expected divergence of the model's density with respect to the true density and is defined as:

$$\begin{aligned} KLIC_i \int h(y_t) ln \frac{h(y_t)}{f_i(y_t)} dy \\ = E(log(h(y_t))) - E(log(f_i(y_t))), \end{aligned} \tag{2}$$

where $h(y_t)$ is the true density and $fi(y_t)$ is the predictive density of model i. KLIC is non-negative and attains its lower bound only when $h(y_t)$ equals $fi(y_t)$. Although $h(y_t)$ is unobserved, notice that a comparison between competing models $KLIC_i$ and $KLIC_j$ only requires evaluation of $E(log(fi(y_t)))$, since the expected true density $E(log(h(y_t)))$ can be treated as a constant across the models and cancels out. This implies that minimizing KLIC involves maximizing the $E(log(fi(y_t)))$. This term, known as the average logarithmic score or log score, can be estimated by:

$$E(log(f_i(y_t))) = \frac{1}{n} \sum_{t=0}^{n-1} log(f_i(y_t)). \tag{3}$$

The log score rewards models that on average allocate higher probability to events that actually occurred. For example, in figure 4 consider density forecast functions $f_1(y)$ and $f_2(y)$ evaluated ex-post in time t at the realized price y_t. Density $f_1(y)$ is preferred to $f_2(y)$ since it assigns a higher probability to the realized price; this also applies to the monotonic transformation comprising the log scores, log($f_1(y_t)$) and log($f_2(y_t)$).[3] In our empirical analysis we calculate log scores to compare density forecasts for the out-of-sample observations, and set the weights of one of the combinations schemes following Bjornland et al. (2011) and Kascha and Ravazzolo (2010).

To further assess the predictive accuracy of alternative forecasts, we test the significance of the differences in log scores between two forecasts using a technique described by Mitchell and Hall

[3]Since the density function is a value between zero and one, the codomain of the log score is (-∞,0) where the less negative values (closer to zero) are preferred to the more negative values.

(2005) and Mcdonald and Thorsrud (2011). We regress the differences in the log scores of competing forecasts on a constant and use heteroskedasticity and autocorrelation consistent estimators (HAC) robust standard errors to determine its significance.

$$log(f_1(y_t)) - log(f_2(y_t)) = c \tag{4}$$

where c is the constant. If the difference is positive and significantly different than zero, then $f_1(y_t)$ is considered a superior density forecast than $f_2(y_t)$. The HAC covariance matrix is estimated using Newey and West (1994) non parametric bandwidth selection procedure (Zeileis, 2004). Using sharpness and calibration as forecast evaluation criteria is a departure from procedures used in point forecasting. Point forecasts are generally evaluated by root mean square error (RMSE), defined as the square root of the sum of squares of forecast errors:

$$RMSE = \sqrt{\frac{1}{n}\sum(p_t - y_t)^2} \tag{5}$$

where p_t is the point forecast that corresponds to the mean of the density forecast, and y_t is the realized price. While a strong correlation exists between log scores and RMSE, the relationship is not one to one (Kascha and Ravazzolo, 2010). We argue forecast evaluation criteria that are focused on features of the probability distribution are especially relevant when prices are highly variable. This is particularly true if a decision maker's loss function is not quadratic and depends on higher moments of a possible outcome (Bjornland et al., 2011). Mcdonald and Thorsrud (2011) argue that the recent financial crisis has highlighted the importance of having not only good point forecasts, but also a good assessment of the whole range of possible outcomes.

To illustrate consider figure 5. Two competing density forecasts $f_1(y)$ and $f_2(y)$ have the same mean, p_t. If y_t is the realized outcome, f_1 and f_2 will have the same RMSE for this observation, so it cannot inform the choice of f_1 versus f_2. However, f_2 is clearly the superior forecast distribution, at least for this observation, because f_2 places a higher probability on the observed outcome than does f_1. For precisely this reason, $f_2(y)$ obtains a higher log score than f_1 at this observation. If this occurred repeatedly $f_2(y)$ correctly would be chosen over f_1 using log scores, even though both would deemed equal by RMSE.

Density forecast combination

While point forecast combinations are often based on root mean square error (RMSE), combining density forecasts is less straightforward. In particular, a combined density may have different characteristics than the individual densities from which it is constructed. For instance, a linear combination of normal distributions with different means and variances will be a mixture normal (Hall and Mitchell 2007), and for many distributions analytical solutions of the combinations are not feasible requiring the use of simulation techniques (Bjornland et al., 2011). Combining density forecasts imposes new requirements beyond those for combining point forecasts, namely the combination must result in a distribution. This implies that the combination must be convex with weights confined to the zero-one interval so that the probability forecast never is negative and probabilities always sum to one.

Several approaches to combine probability distributions have been considered in the literature including linear pools, with equal weights, (Wallis, 2005), combination of weights based on KLIC (Amisano and Giacomini, 2007; Hall and Mitchell, 2007; Jore, Mitchell, and Vahey, 2010), a Bayesian framework (Geweke and Amisano, 2010; Eklund and Karlsson, 2007), and recalibration of linear combinations (Ranjan and Gneiting, 2010; Gneiting and Ranjan, 2011). Here, we employ the most widely used procedures, linear and logarithmic pooling, which have been shown to work well with quarterly data that demonstrate limited deviations from normality and no ARCH effects i.e.Kascha and Ravazzolo (2010)) which is consistent with the structure the data and models used.[4] Isengildina-Massa2011 The first approach is the convex linear combinations ("linear opinion pool")

$$\bar{F}_c = \sum_{i=1}^{N} \omega_{t+h,i} F_{t+h,i} \tag{6}$$

where the linear pool $\bar{F}_c$ is made up of N competitive forecast densities $F_{t+h,i}$, where i represents each forecast model, and ω its weight with $0 \leq \omega_{t+h,i} \leq 1$ and $\sum_{i=1}^{N} \omega_{t+h,i} = 1$.
The second approach is the logarithmic opinion pool where densities are expressed by:

$$\bar{f}^l = \frac{\prod_{i=1}^{N} f_{t+h}^{\omega_t}}{\int \prod_{i=1}^{N} f_{t+h} \omega_t} \tag{7}$$

where ω_{t+h}, are weights chosen such that the integral in the denominator is finite.
Winkler (1968), shows that the logarithmic combination retains the symmetry of the individual forecasts for the case of normal densities. For instance, consider a set of normal densities with means and variances $\mu_i, \sigma_i, i = 1 \ldots N$, and denote transformed weights by $\alpha_i = \frac{\omega_i}{\sigma_i^2}$. The logarithmic pool is a normal density, $N(\mu_c, \sigma_c^2)$, with mean and variance given by $\mu_c = \sum_{i=1}^{N} \alpha_i \mu_i / \sum_{i=1}^{N} \alpha_i$, and $\sigma_c^2 = (\sum_{i=1}^{N} \alpha_i)^{-1}$.

The logarithmic combination offers certain advantages, Hora (2004) showed that application of linearly combined forecasts can produce suboptimal density forecasts, as a linear pool tends to be over dispersed and gives prediction intervals that are too wide on average. Meanwhile, logarithmic pool is less dispersed than the linear combination and is also unimodal (Genest and Zidek, 1986), as shown in figure 6.

Choice of weights

To aggregate the predictive densities, weights for each model need to be identified. We use four procedures and divide the data into three periods to allow for initial estimation, training, and out-of sample evaluation. In the initial estimation, we develop the structure of the forecast models. Then in the training period, we recursively generate out-of-sample forecasts which are used to train or develop the composite weights. Finally in the evaluation period, we continue to recursively re-estimate the models and their composite weights and forecast.

[4]The hog price series is also stationary

The first procedure uses recursive log score weights (RLSW), following Amisano and Giacomini (2007), Hall and Mitchell (2007) and Jore, Mitchell, and Vahey (2010). The recursive weights of RLSW for forecast h-steps ahead are:

$$\omega_{t+h,i} = \frac{exp \sum_{\tau=t_s}^{t=h} ln(f_i(y_{t+h}))}{\sum_{k=1}^{N} exp(\sum_{\tau=t_s}^{t=h} ln(f(y_{t+h}))} \tag{8}$$

where y_{t+h} is the actual observation, $\tau = t_s$ is the beginning of a training period used to initialize the weights, i corresponds to the individual forecast model, and N is the number of models in total. RLSW are based on the log score of the out-of-sample performance of density forecast models, the weight for each model is the ratio of its log score performance over the sum of the log score performance of all models. Since the weights are updated recursively through time, Jore, Mitchell, and Vahey (2010) recommend it's use in the presence of structural changes or uncertain instabilities. The second procedure is based on mean squared error weights (MSEW) and is often used in point forecasting. It is calculated as:

$$\omega_{t+h,i} = \frac{1/MSE_{t+h,i}}{\sum_{k=1}^{N} 1/(MSE_{t+h,k})}, \tag{9}$$

where $MSE_{t+h,i} = \frac{1}{(t-h)} \sum_{\tau=t_s}^{t=h} (y_{t+h} - \mu_{t+h,i})^2$, where y_{t+h} is the actual observation and $\mu_{t+h,i}$ is the mean forecast of model i. The weights in this procedure also change in the evaluation period to allow for a model's improving forecast performance.

The last procedures are more direct. The third procedure, equal weights (EW), gives the same weight to all models, Wallis (2005), The final procedure, selects the model with the best average log score up to the time the forecast is generated, giving all the weight to that model. It is identified as (SELECT).

For point forecasting, numerous papers find that simple equal weighted combinations often outperform more sophisticated adaptive and weighting methods (Bjornland et al., 2011; Colino et al., 2012). However, it is not clear if such result can be generalized to density combinations. Jore, Mitchell, and Vahey (2010) find recursive log score weights give more accurate forecasts than other weight schemes when analyzing U.S. macroeconomic data. They claim that RLSW takes into account shifting variance and structural breaks. Kascha and Ravazzolo (2010) also find the same result for the US inflation, but for other countries considered (UK, New Zealand, and Norway) RLSW weights yields worse forecasts than alternative schemes. Bjornland et al. (2011) examined Norwegian GDP and inflation using a suite of models, finding that logarithmic RSLW outperformed forecasts from other schemes. Results from the literature appear to be mixed, and evidence suggests that a solution for an optimal forecast combination procedure is uncertain. Different combining rules may be suitable in different situations, as a result, there is not a single all-purpose optimal combining procedure (Winkler, 1986).

1 Results

For the analysis the data are divided into three periods for estimation, training, and out-of-sample forecast evaluation. The individual econometric models are first estimated in the period 1975.I-1984.IV. Then we generate the composite forecasts recursively, starting with the period 1985.I - 1993.IV to obtain the optimal weights for forecasting the observation of 1994.I. For the following period 1994.II we use as a training period for the composite structure the period 1985.I-1994.I, therefore adding the last observation to the estimation window. The same procedure is followed for the next forecasts until 2010.IV. Finally, out-of-sample forecasts are recursively generated for the 1994.I-2010.IV evaluation period.

The out-of-sample forecasting performance is evaluated in terms of sharpness, predictive accuracy, and calibration. Table 2 provides the sharpness results for the evaluation period. Along with the average log score (lnS), the root mean square error (RMSE) is presented for comparison because it is a traditional measure of point forecast performance.[5] Table 3 presents selected pair-wise tests of competing models in which individual models are compared against the best composite model to determine if the differences are significantly different.
Since in Table 2 the logarithmic pooling with MSEW model is the best overall forecast in terms of sharpness (highest log score), all pair wise comparisons are made against this model. Table 4 provides the calibration (goodness of fit) of individual and combined density forecast measured using the Berkowitz test on the normalized probability integral transform (PIT).

Results from Table 2 show for the individual models, the exponential smoothing procedure has the best performance at h=1 and h=2, with the highest (less negative) lnS. Exponential smoothing gives more weight to recent observations, incorporating instabilities, and cyclical, components in a more flexible way. The USDA's forecast performed the worst at h=1, and only outperform the No Change forecast at h=2. Nevertheless, at the second horizon, its performance deteriorates less rapidly than the other individual forecasts.

Although a close relationship between average log score lnS and RMSE exists, the correspondence is not one-to-one. For instance, VAR(5) exhibits the lowest RMSE among individual models across different horizons, however its out-of-sample density forecast performance ranks comparatively low in terms of sharpness. Also note that exponential smoothing is the best forecast in terms of sharpness in both h=1, and h=2, but it ranks second in terms of RMSE at h=1, and only third at h=2. The heterogeneity in the results at different horizons, and the difference in ranking between point and density forecast criteria provide further motivation to consider density composites since they can provide insurance against selecting an innefective procedure.

The results for the density combinations in Table 2 show that pooling either linearly or logarithmically improves sharpness over the individual models. The composites based on equal and MSE weights always outperform the individual forecasts. Logarithmic pooling based on MSE weights (Log MSEW) provides the best combination followed closely the equally weighted logarithmic composite (Log EW).

[5]The findings for RMSE of AR(5), VAR(5), and USDA are the same as the results in Colino et al. (2012) table 5 except for small differences due to rounding.

Based on the individual performance of the Log MSEW composite, we test its predictive accuracy against the best individual forecast (exponential smoothing), the best linear pooling forecast (Lin MSEW), the USDA implied forecast, and the other logarithmic pooling forecasts based on different weighting schemes. We also compare the best linear pooling model (LIN MSEW) to the best individual forecast model to further assess the effect of pooling. The results indicate the Log MSEW composite dominates the best of the individual models (exponential smoothing), the best linear composite (LIN MSEW), but not the logarithmic equally weighted composite (Log EW). Interestingly, while the average log scores for the linear equally weight and MSE are superior to the individual forecasts, the best linear composite (Lin MSEW) does not provide statistically superior accuracy compared to the best individual forecast, suggesting that functional form of the combination may influence the accuracy. This last finding also supports the notion that it is difficult to improve on equally weighted forecasts. The USDA forecast does not perform well relative to Log MSEW, suggesting that its forecasts can be substantially improved using composite methods. The RLSW weights deliver substantially worse forecasts than alternative combination schemes, even worse that some of the individual models. The test of equal predictive accuracy finds statistically significant differences between Log RLSW and the best weighting procedures.

Examination of how the methods perform over time can also be informative for decision makers. We compare forecast behavior over time by calculating cumulative difference of the log scores for alternate forecasts:

$$ClnS = \sum_{\tau t_s}^{t} ln(f_1(y_{t+h})) - ln(f_2(y_{t+h})), \tag{10}$$

where f_1 and f_2 correspond to the competing forecasts. ClnS increases over time when f_1 is more accurate than f_2. Using this measure, we examine the cumulative differences of log scores between the best density combination and the best individual model (Log MSEW and exponential smoothing), the best logarithmic pool and the best linear pool (Log MSEW and Linear MSEW), and the best density combination and the USDA (Log MSEW and USDA). We provide the ClnS for h=1, and h=2 in figure 7 and figure 8. The ClnSs between the best logarithmic and linear combinations show only a modest increase from 1994 to 2010 at both h=1 and h=2. The difference between the best logarithmic pool and the exponential smoothing although significant (Table 3) is small for h=1, but becomes increasingly larger at h=2. This highlights the effectiveness of the exponential smoothing model as it captures shorter-term characteristics but does less well at longer horizons. USDA forecasts behave similarly for both h=1 and h=2; it is consistently outperformed by the best combination.

Table 4 provides the calibration (goodness of fit) of individual and combined density forecast at h = 1 using the Berkowitz test on the normalized probability integral transform (PIT). Recall that a density forecast is calibrated (correctly specified) if the joint null hypothesis of normality and no autocorrelation of the normalized PITS is not rejected. Results show that univariate models AR(5), Exponential smoothing, and No Change are calibrated (Berkowitz test fails to reject the null at 5% level), while the VAR(5) model is not well calibrated. Notice while the VAR(5) is the best individual model in terms of RMSE (Table 2), it is actually the worst model in terms calibration and ranks relatively low in sharpness.

Results in Table 4 show that all the logarithmic pooling combinations are calibrated, while the linear pooling combinations are not because the null hypotheses are rejected. The results demonstrate a clear dominance of logarithmic over linear pooling in terms of calibration, which is consistent with small but significant difference in sharpness and predictive accuracy. Note that the linear pooled forecasts do not improve over the best individual density forecasts as established by the test of equal predictive accuracy. On balance, logarithmic pooling with MSE (Log MSEW) and equally weighted combinations (Log EW) are superior to individual densities, linear combinations, and to combinations derived from recursive weights. figure9 provides a depiction of the forecast densities (2000.I) consistent with the statistical findings, supporting the attractiveness of the logarithmic MSE composite.

1.1 Summary and Conclusions

Useful forecasts should be probabilistic, taking the form of predictive distributions over future quantities and events (Gneiting and Ranjan, 2011). In this paper we investigate the density forecasting performance of different individual and composite measures using hog prices and different time series models and forecasts provided by the USDA. In theory, composite forecasts can provide information from additional sources and protect against particularly bad forecasts when using a particular model. These advantages can lead to superior density forecast accuracy.

The evaluation of the density and composite forecasts was based on sharpness and calibration. In terms of sharpness, we find that the performance of the individual forecasts varied substantially although exponential smoothing appeared to be the best individual model across different horizons. Consistent with the emerging density forecast literature, density forecast combinations improved the sharpness over individual forecasts, but the form of the combination and the weighting procedures influenced the performance. Linear composites improved over individual forecasts, but the test of equal predictive accuracy indicated that the difference was not always statistically significant. In contrast, logarithmic composites based on MSE and equal weights were always better than the individual forecasts and the linear composites in terms of sharpness. The best logarithmic composite (Log MSEW) also was statistically more accurate than the other forecasts investigated except for logarithmic equal weights forecasts. Somewhat unexpectedly, the sharpness performance of the linear and logarithmic recursive log score weight forecasts (RSLW) was quite poor. Previous research has also identified that recursive log score weighting scheme has performed poorly in other situations. The exact reason for the poor performance is uncertain, but may point to the complexity of the weighting scheme. Regardless, this finding clearly identifies the need for careful empirical assessment of alternate weighting schemes as different combining rules appear to be more suitable in different situations.

Most of the individual forecast models were well calibrated, except for the VAR(5) which exhibited the best individual point forecasts based on the RMSE. Interestingly, the VAR(5) poor performance also emerged in the sharpness comparisons and point to the importance of using log score measures when evaluating densities and their relative accuracy. All of the logarithmic pooled forecasts are well calibrated, while the linear pooled forecasts generally reject the null hypothesis. The only linear pooled forecast that fails to reject the null hypothesis of calibration is the linear RLSW composite that performs poorly in terms of sharpness. The relatively poor performance of the linear composites here is consistent with research that has shown that linear composites can be over dispersed and give prediction intervals that too wide on average. Logarithmic composites in contrast tend to be less dispersed and unimodal which facilitates the calibration.

Similar to Colino et al. (2012) we also find that the use of composite forecasting methods can reduce the outlook forecast errors. Here, in terms of sharpness, all of the composite density forecasts are superior, and the best logarithmic forecast is statistically much more accurate. Detailed comparison of the relative performance of the USDA forecasts and the best logarithmic forecast over time demonstrates the consistent superiority (which appears to be increasing in magnitude through time) of the composite density forecasts at both one and two quarter ahead horizons. The clear superiority of these logarithmic density forecasts identifies the potential to provide hog producers and market participants with more accurate expected price probability distributions which can facilitate decision making. Importantly, the difference between the best logarithmic composite forecasts and the easier to generate logarithmic equal weighted forecasts is small and statistically not significant. This should reduce the complexity and costs of developing these density forecasts and perhaps provide a more readily available alternative. Finally, here we examined the ability of USDA to characterize out-of-sample forecast densities, identifying large improvements which can be made using composite procedures. Further work to assess the ability of other outlook programs seems warranted and potentially useful in providing information to decision makers.

References

Aastveit, K., K. Gerdrup, A. Jore, and F. Ravazzolo. 2012. "Short-term Forecasting: Norges Bank's Density Combination Approach." Unpublished.

Amisano, G., and R. Giacomini. 2007. "Comparing Density Forecasts via Weighted Likelihood Ratio Tests." *Journal of Business & Economic Statistics* 25:177–190.

Berkowitz, J. 2001. "Testing Density Forecasts, With Applications to Risk Management." *Journal of Business & Economic Statistics* 19:465–474.

Bessler, D.A., and J.L. Kling. 1986. "Forecasting Vector Autoregressions with Bayesian Priors." *American Journal of Agricultural Economics* 68:144.

Bjornland, H.C., K. Gerdrup, A.S. Jore, C. Smith, and L.A. Thorsrud. 2011. "Weights and Pools for a Norwegian Density Combination." *The North American Journal of Economics and Finance* 22:61–76.

Bottum, J.C. 1966. "Changing Functions of Outlook in the U.S." *Journal of Farm Economics* 48:1154.

Brandt, J.A., and D.A. Bessler. 1981. "Composite Forecasting: An Application with U.S. Hog Prices." *American Journal of Agricultural Economics* 63:135.

Clements, M.P. 2004. "Evaluating the Bank of England Density Forecasts of Inflation." *The Economic Journal* 114:844–866.

Colino, E., S. Irwin, and P. Garcia. 2011. "Improving the Accuracy of Outlook Price Forecasts." *Agricultural Economics* 42:357–371.

Colino, E., S. Irwin, P. Garcia, and X. Etienne. 2012. "Do Composite Procedures Really Improve the Accuracy of Outlook Forecasts?" *Journal of Agricultural and Resource Economics* 37.

Corradi, V., and N.R. Swanson. 2006. "Predictive Density Evaluation." *Handbook of Economic Forecasting* 1:197–284.

Cullen, A.C., and H.C. Frey. 1999. *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer.

Diebold, F.X., T.A. Gunther, and A.S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39:863–883.

Egelkraut, T.M., P. Garcia, and B.J. Sherrick. 2007. "The Term Structure of Implied Forward Volatility: Recovery and Informational Content in the Corn Options Market." *American Journal of Agricultural Economics* 89:1–11.

Eklund, J., and S. Karlsson. 2007. "Forecast Combination and Model Averaging Using Predictive Measures." *Econometric Reviews* 26:329–363.

Elliott, G., and A. Timmermann. 2008. “Economic Forecasting.” *Journal of Economic Literature* 46:3–56.

Fackler, P.L., and R.P. King. 1990. “Calibration of Option-Based Probability Assessments in Agricultural Commodity Markets.” *American Journal of Agricultural Economics* 72:73.

Genest, C., and J.V. Zidek. 1986. “Combining Probability Distributions: A Critique and an Annotated Bibliography.” *Statistical Science* 1:pp. 114–135.

Geweke, J., and G. Amisano. 2010. “Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns.” *International Journal of Forecasting* 26:216–230.

Gneiting, T., and A.E. Raftery. 2007. “Strictly Proper Scoring Rules, Prediction, and Estimation.” *Journal of the American Statistical Association* 102:359–378.

Gneiting, T., and R. Ranjan. 2011. “Combining Predictive Distributions.” *Arxiv*, Jun., pp. .

Hall, S.G., and J. Mitchell. 2007. “Combining Density Forecasts.” *International Journal of Forecasting* 23:1–13.

Hora, S.C. 2004. “Probability Judgments for Continuous Quantities: Linear Combinations and Calibration.” *Management Science* 50:597–604.

Hyndman, R., and Y. Khandakar. 2008. “Automatic Time Series Forecasting: The forecast Package for R.” *Journal of Statistical Software* 27:1–22.

Isengildina-Massa, O., S. Irwin, D.L. Good, and L. Massa. 2011. “Empirical confidence intervals for USDA commodity price forecasts.” *Applied Economics* 43:3789–3803.

Jore, A., J. Mitchell, and S.P. Vahey. 2010. “Combining Forecast Densities from VARs with Uncertain Instabilities.” *J. Appl. Econ.* 25:621–634.

Kascha, C., and F. Ravazzolo. 2010. “Combining Inflation Density Forecasts.” *J. Forecast.* 29:231–250.

Leuthold, R.M., A.J.A. MacCormick, A. Schmitz, and D.G. Watts. 1970. “Forecasting Daily Hog Prices and Quantities: A Study of Alternative Forecasting Techniques.” *Journal of the American Statistical Association*, pp. 90–107.

Mcdonald, C., and L.A. Thorsrud. 2011. “Evaluating density forecasts: model combination strategies versus the RBNZ.” *Reserve Bank of New Zealand Discussion . . .* , pp. .

Mitchell, J., and S.G. Hall. 2005. “Evaluating, Comparing and Combining density forecasts using the KLIC with an application to the Bank of England and NIESR ’fan’ charts of Inflation.” *Oxford Bulletin of Economics and Statistics* 67:995–1033.

Mitchell, J., and K.F. Wallis. 2011. “Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness.” *Journal of Applied Econometrics* 26 , pages:1023–1040.

Newey, W.K., and K.D. West. 1994. “Automatic Lag Selection in Covariance Matrix Estimation.” *The Review of Economic Studies* 61:631–653.

Ranjan, R., and T. Gneiting. 2010. “Combining Probability Forecasts.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72:71–91.

Rosenblatt, M. 1952. “Remarks on a multivariate transformation.” *The Annals of Mathematical Statistics* 23:470–472.

Sherrick, B.J., P. Garcia, and V. Tirupattur. 1996. “Recovering Probabilistic Information From Option Markets: Tests Of Distributional Assumptions.” *Journal of Futures Markets* 16:545–560.

Smith, J., and K.F. Wallis. 2009. “A Simple Explanation of the Forecast Combination Puzzle.” *Oxford Bulletin of Economics and Statistics* 71:331–355.

Timm, T.R. 1966. “Proposals for Improvement of the Agricultural Outlook Program of the United States.” *Journal of Farm Economics* 48:1179–1184.

Timmermann, A. 2006. “Forecast Combinations.” In G. Elliott, C. W. J. Granger, and A. Timmermann, eds. *Handbook of Economic Forecasting*. Amsterdam, NL: North Holland, vol. 1, chap. 4, i ed., pp. 135–196.

Wallis, K.F. 2005. “Combining Density and Interval Forecasts: A Modest Proposal*.” *Oxford Bulletin of Economics and Statistics* 67:983–994.

Wang, Z., and D.A. Bessler. 2004. “Forecasting Performance of Multivariate Time Series Models with Full and Reduced Rank: an Empirical Examination.” *International Journal of Forecasting* 20:683–695.

Winkler, R.L. 1986. “[Combining Probability Distributions: A Critique and an Annotated Bibliography]: Comment.” *Statistical Science*, pp. 138–140.

—. 1968. “The Consensus of Subjective Probability Distributions.” *Management Science* 15:B—-61.

Zapata, H.O., and P. Garcia. 1990. “Price Forecasting with Time Series Methods and Nonstationary Data: An Application to Monthly U.S. Cattle Prices.” *Western Journal of Agricultural Economics* 15:123–132.

Zeileis, A. 2004. “Econometric Computing with HC and HAC Covariance Matrix Estimators.” *Journal of Statistical Software* 11:1–17.

2 Tables and Figures

Table 1: Sources of Quarterly Data, 1975.I - 2010.IV

Hog Price: Livestock, Dairy, and Poultry Outlook, Barrows & gilts, n. base, i.e. $/cwt http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do? documentID=1350 and Red Meat Yearbook before 2005
Sows Farrowings: Number of Sows from USDA NASS hogs and pigs http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do? documentID=1086
Commercial Pork Production: NASS Livestock slaughter http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do? documentID=1096
Beef price: Livestock, Dairy, and Poultry Outlook, Choice steers, 5-area Direct, $/cwt http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do? documentID=1350 and Red Meat Yearbook before 2005
Corn prices: Quarterly average of daily prices, $/bushel AgMAS, barchart.com
USDA Hog Price Forecast: Livestock, Dairy, and Poultry Outlook, Barrows & gilts, n. base, i.e. $/cwt http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do? documentID=1350

Table 2: Log Score and RMSE of Individual and CombinedForecasting Models, 1994.I - 2010.IV

	h=1		h=2	
Individual Model	lnS	RMSE	lnS	RMSE
AR(5)	-3.1084	5.2208	-3.4593	7.0601
Exp Smooth	-3.0240	4.9617	-3.4136	7.2277
No Change	-3.2980	6.3042	-3.6572	9.0587
VAR(5)	-3.1405	4.6605	-3.4595	6.0505
USDA	-3.3206	6.5738	-3.5462	8.1549
Combinations				
Select	-3.2147	5.2976	-3.522	6.7799
Linear Pooling				
LIN EW	-3.0098	4.7816	-3.3138	6.4489
LIN RLSW	-3.1009	5.0610	-3.4564	6.7057
LIN MSEW	-2.9873	4.6446	-3.2845	6.1635
Logarithmic Pooling				
LOG EW	-2.9454	4.5674	-3.2252	6.0439
LOG RLSW	-3.0916	5.0505	-3.4993	6.6099
LOG MSEW	-2.9386	4.5130	-3.2237	5.9426

Table 3: Test of Equal Predictive Accuracy

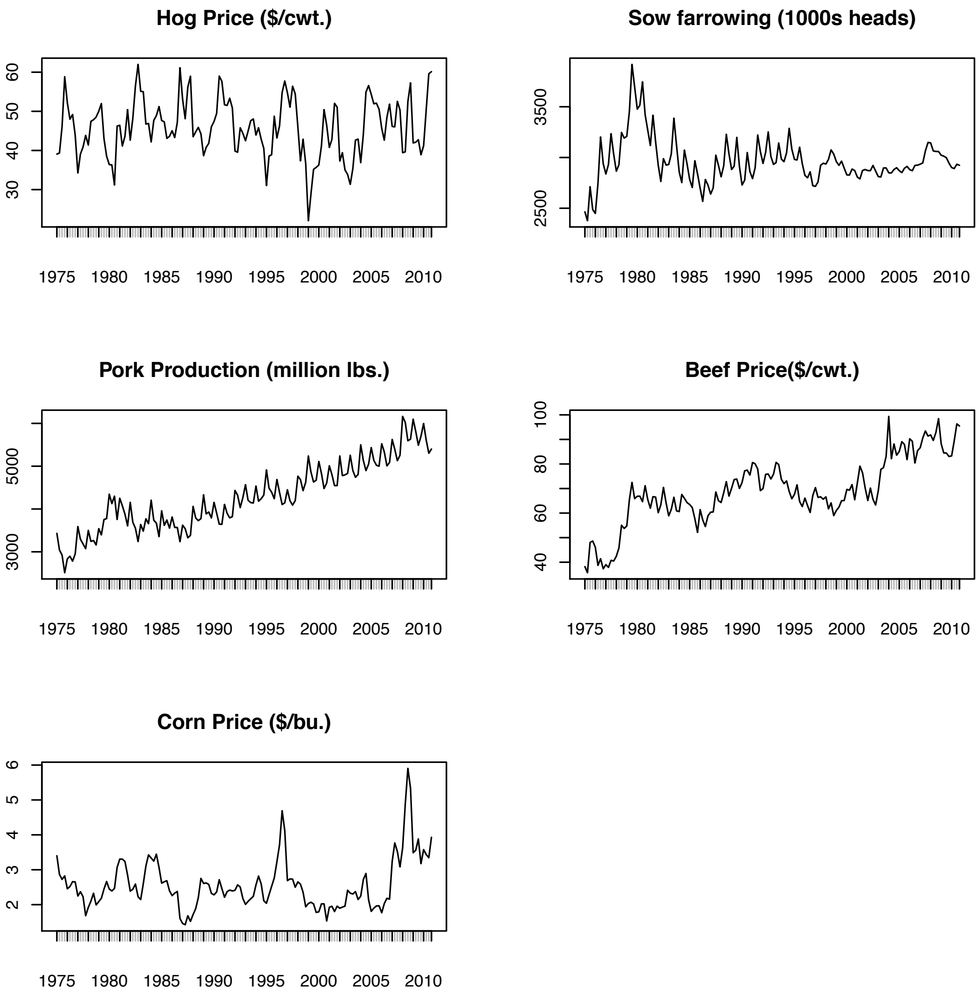
	h=1			h=2		
Log Score Comparison	Diff.	SE	P value	Diff.	SE	P value
Log MSEW vs Exp Smooth	0.0854	0.0412	0.0420**	0.1899	0.0762	0.0152**
Log MSEW vs USDA	0.3821	0.0788	0.0000**	0.3225	0.1265	0.0131**
Log MSEW vs Lin MSEW	0.0488	0.0252	0.0569*	0.0607	0.0360	0.0962*
Log MSEW vs Log RLSW	0.1530	0.0433	0.0007**	0.2756	0.0774	0.0007**
Log MSEW vs Log EW	0.0068	0.0108	0.5286	0.0015	0.0239	0.9511
Lin MSEW vs Exp Smooth	0.0366	0.0340	0.2864	0.1291	0.0942	0.1754

** Significant at 5%, * Significant at 10%

Table 4: Berkowitz Test for Calibration

Individual Model	Likelihood Ratio	Significance
AR5	6.2187	0.1014
Exp. Smoothing	4.1360	0.2472
Random Walk	2.1301	0.5458
VAR5	28.5000	0.0000**
USDA	4.1281	0.2479
Combinations		
Selec	4.4249	0.2191
Linear Pooling		
Lin EW	8.4421	0.0377**
Lin RLSW	7.2431	0.0645*
Lin MSEW	10.3329	0.0159**
Logarithmic Pooling		
Log EW	4.8225	0.1856
Log RLSW	4.7642	0.1898
Log MSEW	5.5753	0.1342

** Significant at 5%, * Significant at 10%



graphicx

package

Figure 1: Time-series plots of variables used for U.S. hog model, 1975.I - 2010.IV.

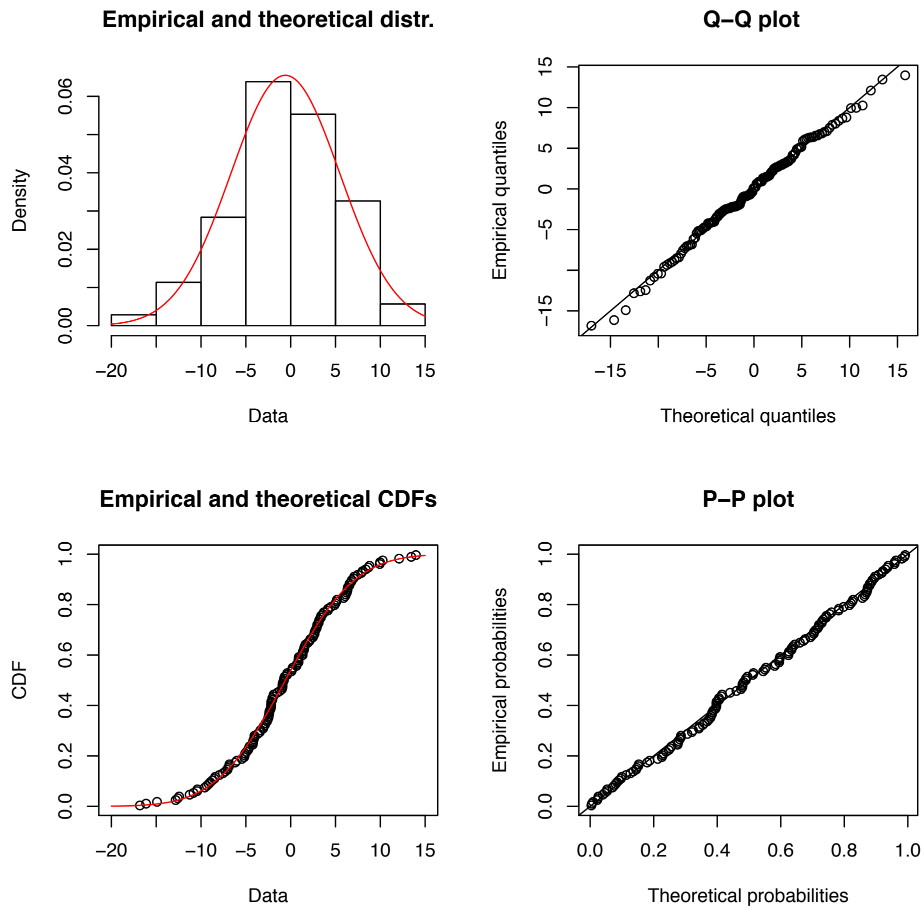


Figure 2: Empirical and Theoretical PDF, CDF, QQ-Plot, and PP-plot of the Forecast Error Distribution

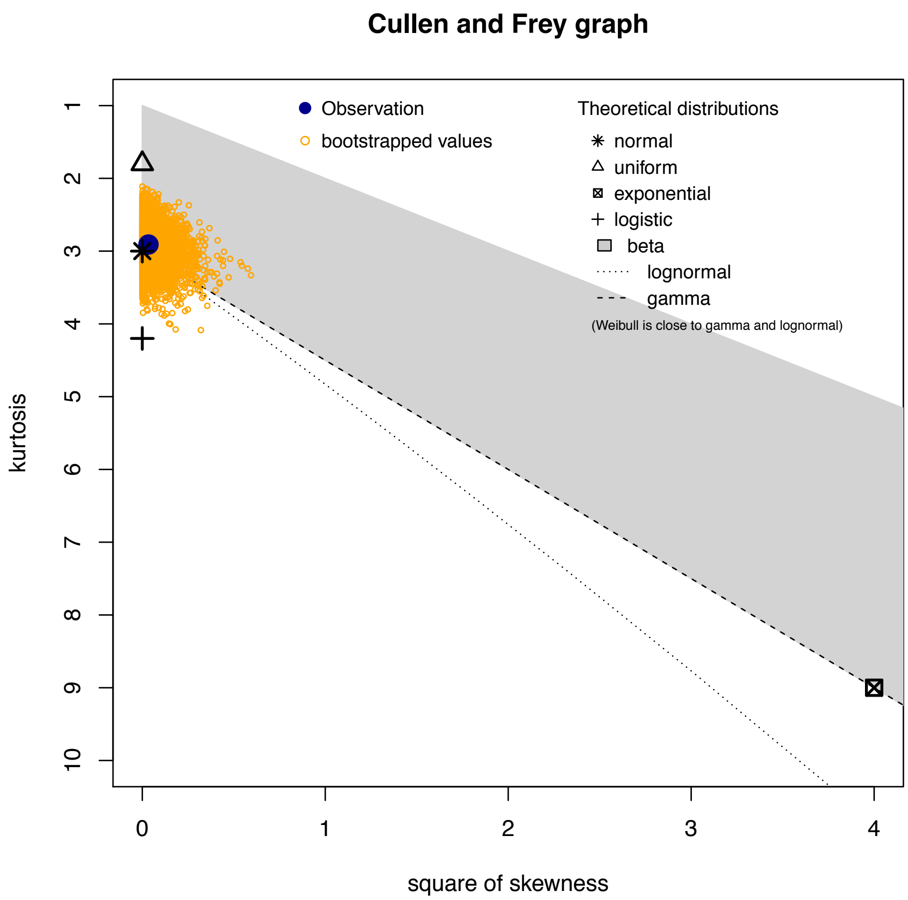


Figure 3: Cullen and Frey, Skewness-Kurtosis Graph

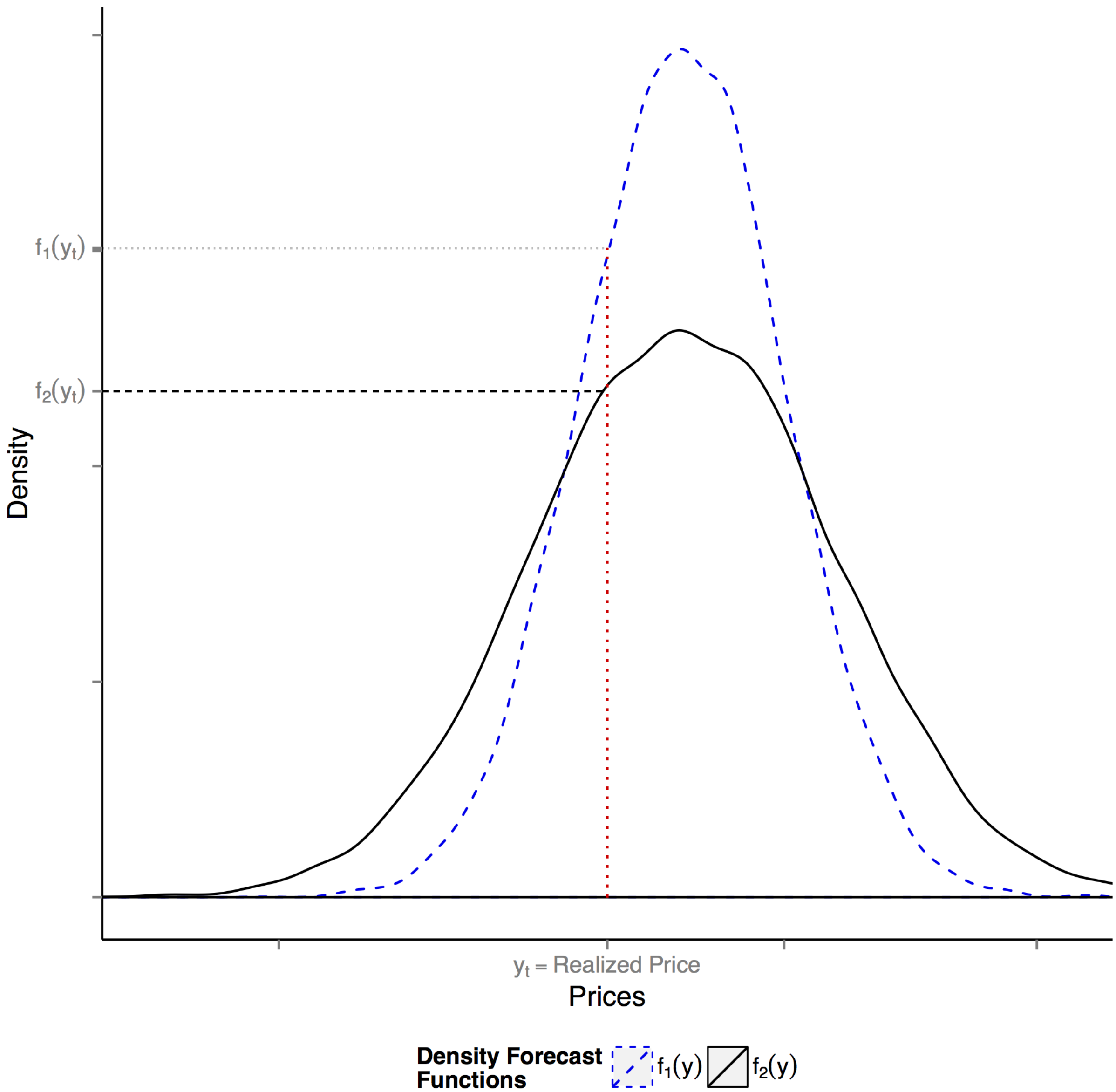


Figure 4: Density forecast functions at realized price, $f_i(y_t)$

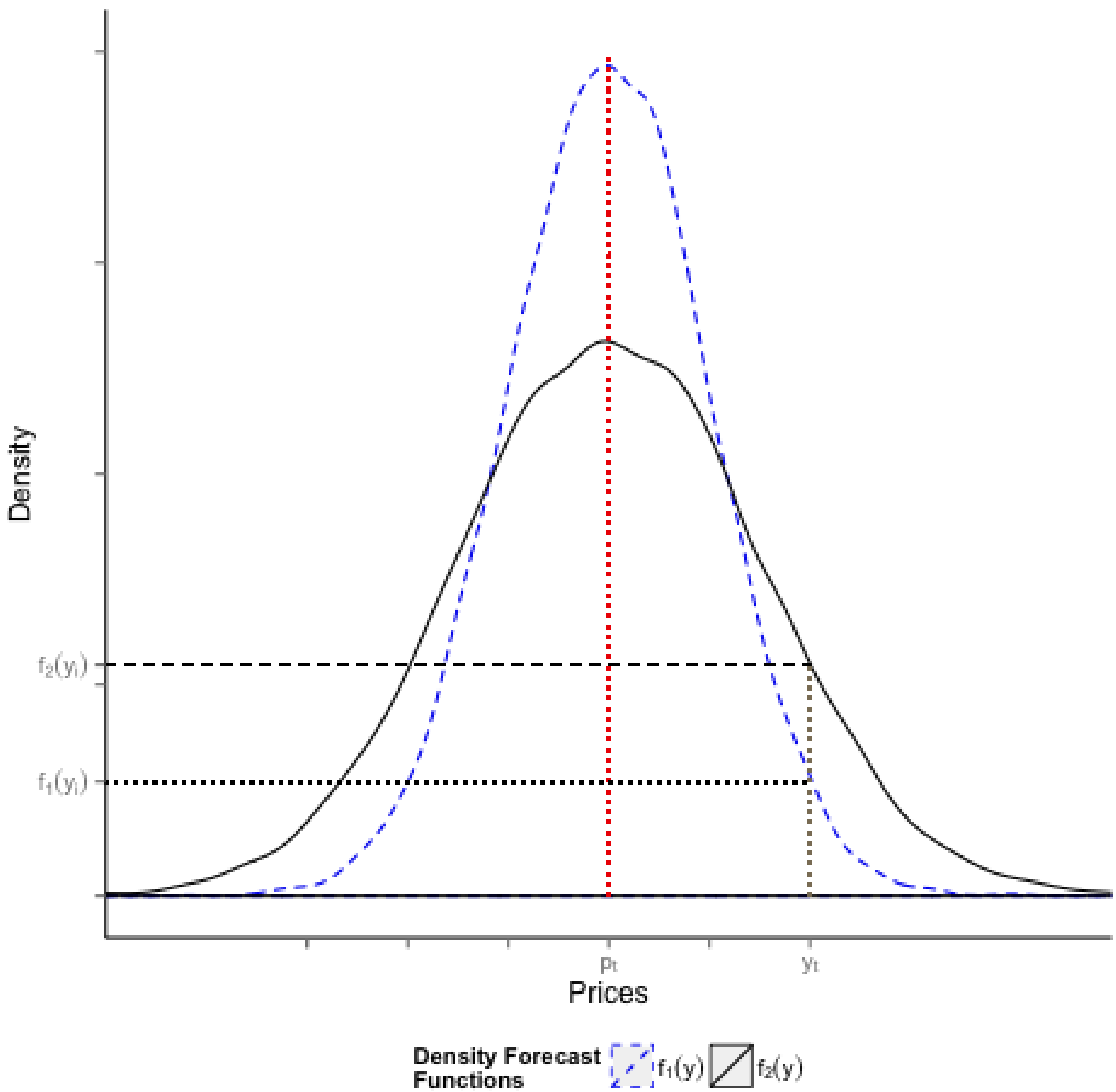
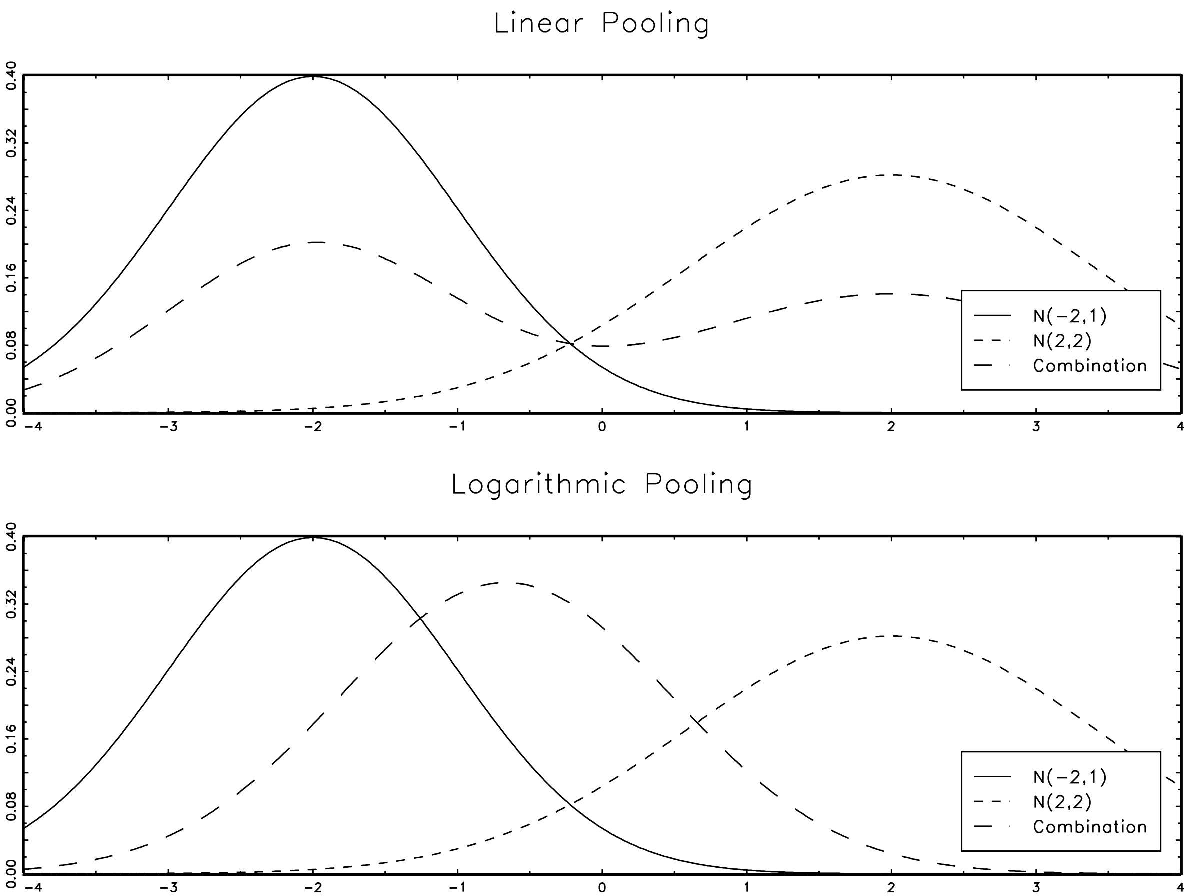


Figure 5: Log Scores and RMSE



graphicx package

Figure 6: Linear Pooling and Log Pooling, Source:Kascha and Ravazzolo (2010)

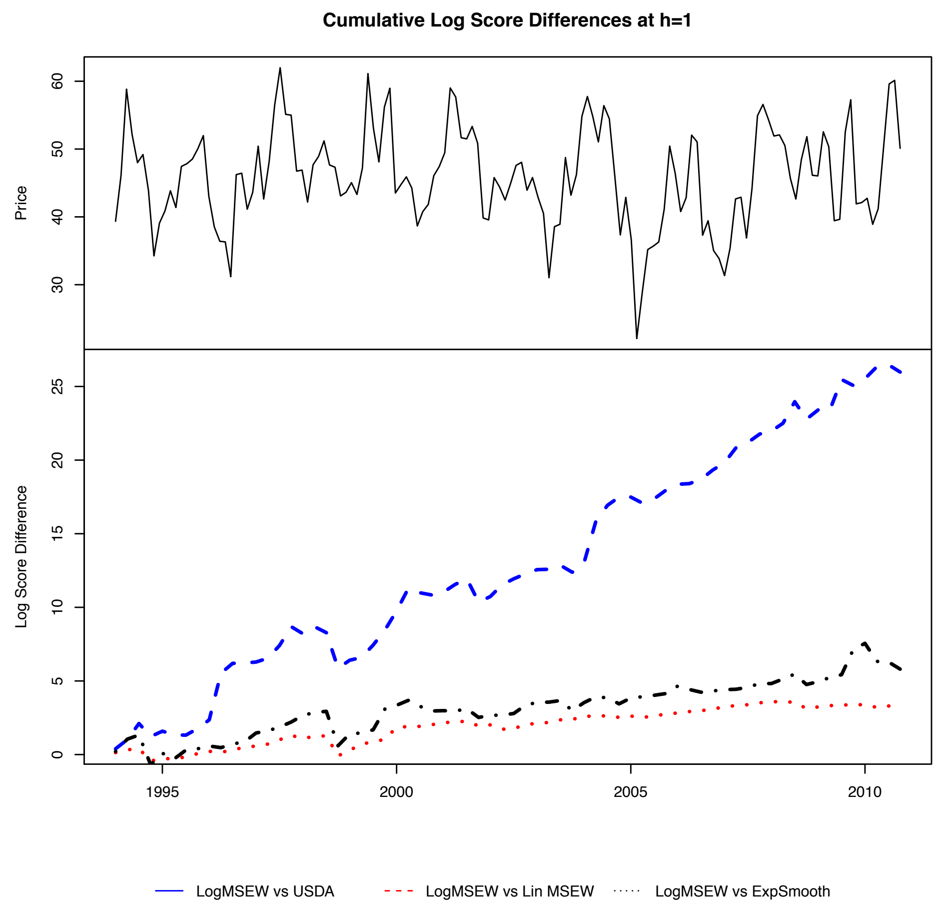


Figure 7: Cumulative Difference of Log Scores at h=1

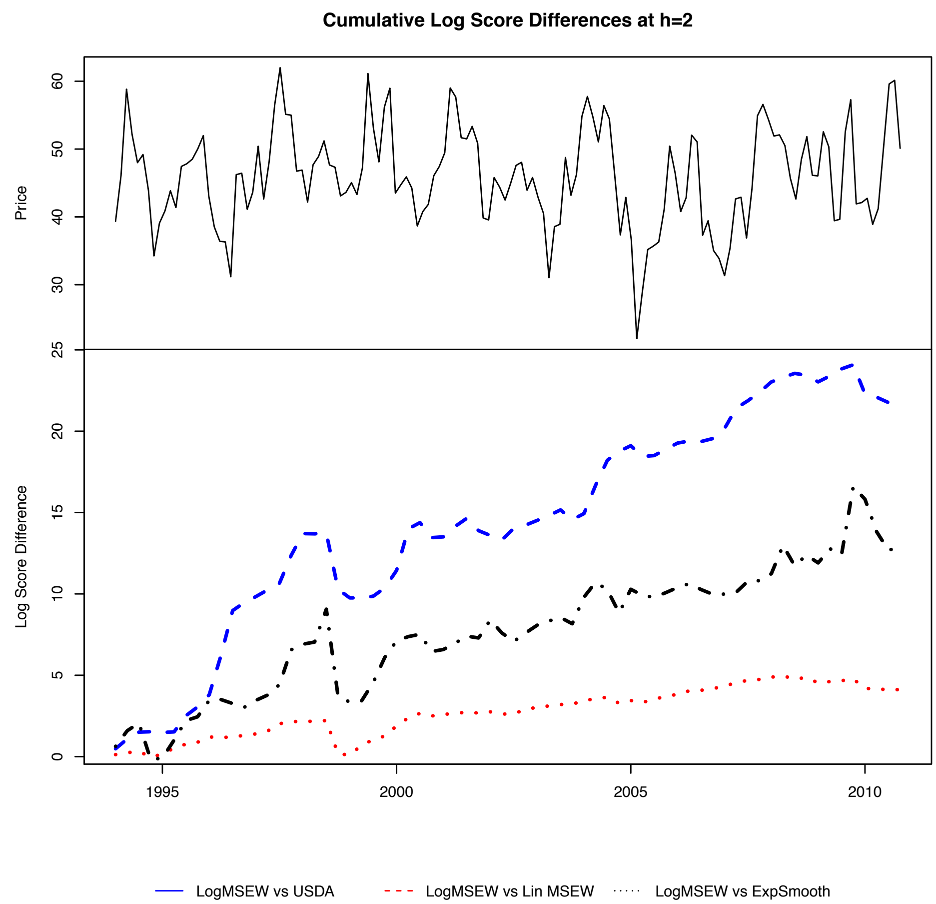


Figure 8: Cumulative Difference of Log Scores at h=2

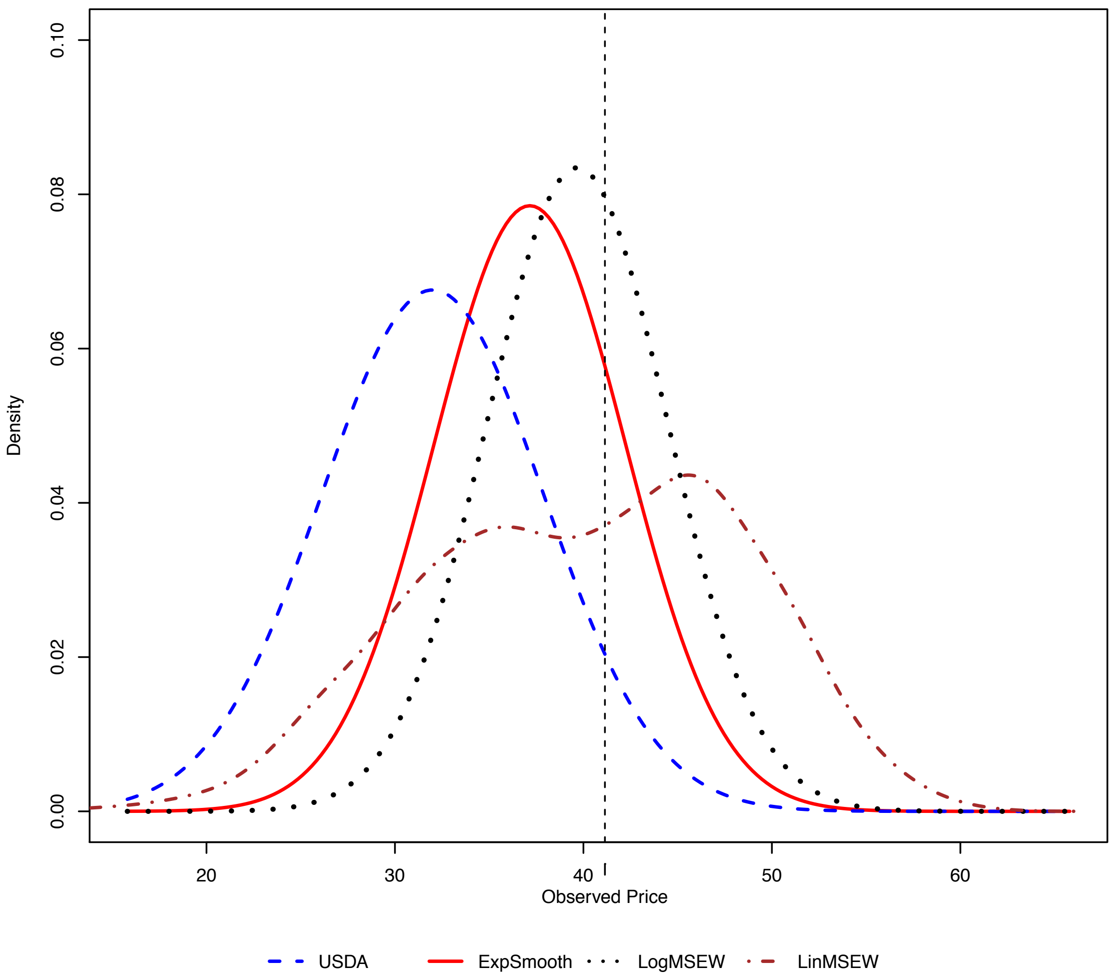


Figure 9: Density Forecasts in 2000.I