Evaluating Density Forecasts with Applications to ESPF

Kanemi Ban, Masaaki Kawagoe and Hideaki Matsuoka

September 2013

The views expressed in “ESRI Discussion Papers” are those of the authors and not those of the Economic and Social Research Institute, the Cabinet Office, or the Government of Japan. (Contact us: https://form.cao.go.jp/esri/en_opinion-0002.html)
Evaluating Density Forecasts with Applications to ESPF

Kanemi Ban*
Masaaki Kawagoe†
Hideaki Matsuoka¶

September 2013

Abstract

This paper evaluates density forecasts using micro data from the ESP forecast (ESPF), a monthly survey of Japanese professional forecasters. The ESPF has collected individual density forecasts since June 2008. We employ two approaches: Probability Integral Transform (PIT) and Ranked Probability Score (RPS). First, we apply Berkowitz's (2001) test to individual density forecasts produced every June. We fail to reject the independency in FY 2010 and 2011 real GDP growth rates. As for CPI inflation rates, we reject the independency in all the samples during FY 2008 to 2011, but fail to reject it if the sample is limited to a half with better forecast performance. The result may ensure that individual densities coincide with the unobserved true data generation process of the actual outcomes. Second, we calculate RPS, following Kenny, Kostka, and Masera (2012), and compare the Mean Probability Distribution (MPD), the average of individual densities, with three benchmarks -- Uniform, Normal and Naïve distributions -- and individual density forecasts. The MPD turns out to be a “good” density: it beats the benchmarks in most cases and ranks about fifth out of around 35 participants every year. Subjective judgments added to the MPD are likely to deteriorate the performance in the case of CPI inflation rate, but to improve in the case of real GDP growth rate.

JEL classification:C22,C53

Keywords: Density Forecast, Probability Integral Transform, Scoring the Densities

*: Osaka University.
†: Economic and Social Research Institute, Cabinet Office, Japanese Government.
¶: Japan Center for Economic Research.

The authors owe gratitude to participants at the ESRI-JCER conference held on February 21 and the 30th meeting of Society of Applied Economic Time Series Analysis held on July 6, especially Kazumi Asako (Hitotsubashi University) and Yoshinori Kawasaki (Institute of Statistical Mathematics). The views expressed herein are our own and in no way represent those of the organizations we belong to.
1. Introduction

The ESP Forecast (ESPF, hereafter), a monthly survey of Japanese professional forecasters, was launched in 2004 and has been established as a valuable information source. The ESPF’s consensus forecasts (CF) are covered by the media and widely quoted in government and central bank documents. Its quality is confirmed by annual performance reviews. The annual review announces the top-five forecasters out of roughly 30 on average every year. Table 1 shows how many times a forecaster has been selected for the best five. Only 6 out of 50 in total were awarded three times or more. In fact, the CF belongs to this group: the CF would have been awarded three times if it had been evaluated as a participant in the past reviews from FY 2004 to 2011.

<table>
<thead>
<tr>
<th>Record of Best 5</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of forecasters</td>
<td>30</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>50</td>
</tr>
</tbody>
</table>

Note that CF’s good performance is relative when compared with individual forecasts. That is, it still may be subject to a substantial forecast error. In fact, as for real GDP growth rate, the actual outcomes were often seen as “outliers” in terms of distribution of point forecasts (Komime, et al. 2009). However, also note dispersion of point forecasts across survey respondents may underestimate overall uncertainties surrounding the actual outcomes because each point forecast itself represents an underlying subjective distribution each forecaster has in his mind, as Zarnowitz and Lambros (1987) pointed out. Therefore, it is essential to collect data on individual subjective distributions to correctly measure the overall uncertainties.

Thus, the ESPF has asked participants to submit their density forecasts of real GDP growth and CPI inflation rates for both the current and the next fiscal year every month

---

1 The first four-year experiences were surveyed by Komine, Ban, Kawagoe and Yoshida (2009).
since June 2008. A density forecast is an estimate of the complete probability distribution of the possible future values of a given target variable, which provides a description of forecast uncertainty. The ESPF publishes the Mean Probability Distribution (MPD), the average density forecast across individuals’, which is intended to indicate the uncertainty of the CF. Such risk information should be valuable to ESPF users, and it is, therefore, important to examine the performance of MPD.

This paper employs two approaches to evaluate density forecasts, Probability Integral Transform (PIT) and Ranked Probability Score (RPS). PIT is used to examine whether our observed individual density forecasts can be regarded to coincide with its true data generating process. Because our small sample disallows Diebold, Gunther and Tay’s (1998) graphical methods, we will resort to Berkowitz’s (2001) proposal. After confirming the consistency between the two, we will go on to calculate RPS following the procedures Boero, Smith and Wallis (2011) and Kenny, Kostka, and Masera (2012) proposed. We will run horse races between MPD and three benchmarks as well as individual density forecasts.

The rest of the paper is organized as follows. Section 2 explains the two approaches to evaluate the densities, PIT and RPS. In Section 3, we will examine the density forecast data the ESPF has accumulated over the past four years. Section 4 shows empirical results and, finally, Section 5 concludes with our assessment of density forecasts.
2. Methodology

There are two approaches to the density forecasts, PIT (Probability Integral Transform) and RPS (Ranked Probability Score), which will be briefly explained in this section.

2.1 PIT (Probability Integral Transform)

The probability integral transform proposed by Diebold, Gunther and Tay (1998) is based on the relationship between the data generating process, \( f(y_t) \), and the sequence of density forecasts, \( p(y_t) \). The probability integral transform \( z_t \) is the cumulative density function corresponding to the density \( p_t(y_t) \) evaluated at \( y_t \).

\[
\int_{-\infty}^{y_t} p(u) du = P(y_t)
\]

Let \( q_t(z_t) \) denote the density of \( z_t \). Assuming that \( \partial P^{-1}(z_t)/\partial z_t \) is continuous and nonzero over \( y_t, z_t \in (0,1) \) with density

\[
q(z_t) = \frac{\partial P^{-1}(z_t)}{\partial z_t} \frac{f(P^{-1}(z_t))}{p(P^{-1}(z_t))}.
\]

If \( p(y_t) = f(y_t) \), \( q(z_t) \) is a density of uniform distribution, \( U(0,1) \).

Diebold, Gunther and Tay (1998) characterise both the density and dependence structure of the entire \( z_t \) sequence when \( p(y_t) = f(y_t) \). Suppose that \( \{y_t\}_{t=1}^m \) is generated from \( \{f(y_t|\Omega_t)\}_{t=1}^m \) where \( \Omega_t = \{y_{t-1}, y_{t-2}, \ldots\} \). If a sequence of density forecasts \( \{p(y_t)\}_{t=1}^m \) coincides with \( \{f(y_t|\Omega_t)\}_{t=1}^m \), the sequence of probability integral transforms of \( \{y_t\}_{t=1}^m \) with respect to \( \{p(y_t)\}_{t=1}^m \) is identical, independent uniform distribution with \( U(0,1) \). Thus we have

\[
q(z_m, z_{m-1}, \ldots, z_1) = \frac{f(P^{-1}(z_m)|\Omega_m)}{p(P^{-1}(z_m))} \frac{f(P^{-1}(z_{m-1})|\Omega_{m-1})}{p(P^{-1}(z_{m-1}))} \ldots \frac{f(P^{-1}(z_1)|\Omega_1)}{p(P^{-1}(z_1))}.
\]

---

2 Here we suppress for simplicity a subscript indicating when the forecast is formulated, say, t-\( \tau \), where \( \tau \) is forecast horizon, because \( \tau \) is assumed to be fixed.
If \( p(y_t) = f(y_t) \), each of ratio is a \( U(0,1) \) and the product of a m-variate \( U(0,1) \) for \( \{z_{t,h},h=1,\ldots,m\} \) is distributed iid \( U(0,1) \).

A simple nonparametric test such as a Kolmogorov-Smirnov test is available to assess whether the probability integral transform series is iid \( U(0,1) \). As Diebold, Gunther and Tay (1998) point out, such test is not so informative. Rather than testing, they suggest visual assessment using the graphical tools and the correlogram supplemented with the usual confidence interval. They demonstrate that histograms of transformed forecast data can reveal useful information about model failure.

Berkowitz (2001) suggests that it is more fruitful to take the inverse normal CDF transformation of \( z_t \) to \( z_t^* \). If \( p(y_t) = f(y_t) \), a sequence of \( z_t^* \) becomes iid standard normal. Berkowitz argues the more powerful tools can be applied to testing iid N(0,1). Here we will use a one-degree of freedom test of independence against a first-order autoregressive structure represented in the equation,

\[
x_{i,t} - \mu_i = \rho(x_{i,t-1} - \mu_i) + \varepsilon_{i,t}.
\]

Although Berkowitz (2001) also proposes a three-degree of freedom test of zero-mean, unit variance and independency, we will try the simple test.

A problem in this paper is that the ESPF has short history of density forecasts. If we use a sequence of mean average density forecasts, the number of sample is not more than four, too small to test independency and normality. Fortunately, individual density forecasts are available in the case of the ESPF, which allows us to resort to dynamic panel techniques.

Thus, we will follow the below procedures for testing independency, where \( i \) and \( t \) denote individual forecaster and time, respectively.

1. Probability integral transform with respect to the individual density forecasts
   \[
z_{i,t} = \int_{-\infty}^{x_{i,t}} p_{i,t}(u)du
\]
2. If \( z_{i,t} = 0 \) or \( z_{i,t} = 1 \), then those data are dropped for testing
3. The inverse normal CDF transformation of \( z_{i,t} \) to \( x_{i,t} \),
   \[
x_{i,t} = \Phi^{-1}(z_{i,t})
\]
4. Test of \( H_0: \rho = 0 \) in the autoregressive model
\[ x_{i,t} - \mu_t = \rho(x_{i,t-1} - \mu_t) + \epsilon_{i,t} \]  

(5) If \( H_0: \rho = 0 \) is rejected, the same procedure is undertaken by using subgroup of individual density forecasts.

Note the second step is added to deal with “outliers,” which will be examined in Section 3. If actual outcomes are outside individual density forecasts, \( z_{i,t} \) is equal to zero or one, i.e. \( x_{i,t} = -\infty \) or \( x_{i,t} = +\infty \), thereby causing a computational difficulty. We decide to proceed by just dropping such failure data.

### 2.2 Scoring the Densities

RPS was proposed by Boero, Smith and Wallis (2011) and was applied to the ECB SPF densities by Kenny, Kostka, and Masera (2012).

RPS is a kind of squared errors of cumulative distribution functions (CDF) for a density forecast and a binary outcome variable. The density forecast \( p_{i,t-\tau}(B_{jt}) \) is produced by a forecaster \( i \) at time \( t-\tau \) (\( t = 1, \ldots T, \tau = 1, \ldots D \)) for the outcome available at time \( t \), and \( B_{jt}(j=1,\ldots K) \) is a bin for which probability is placed. Therefore,

\[ \sum_{j=1}^{K} p_{i,t-\tau}(B_{jt}) = 1. \]

In the case of outcome, \( y_t \) is the binary variable taking a value of 1 in \( B_{it} \) covering the actual value, and zero otherwise. The CDFs of \( p_{i,t-\tau}(B_{it}) \) and \( y_t \) are \( P_i(t) \) and \( Y_t \), respectively. Kenny, Kostka, and Masera (2012) calculates the following,

\[ RPS_{it} = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{K} \left( p_{i,t-\tau}(B_{jt}) - Y_t \right)^2. \]  

(4)

However, here is an assumption that the uncertainty the forecasters face is of the same magnitude if the forecast horizon, \( \tau \), is the same. This assumption may fail to hold in our data: the uncertainty is much greater in forecasting FY 2008 outcomes in the midst of the Great Recession than in forecasting FY 2011 counterparts in its recovery phase,
for example. Average deviation values are derived to cope with this difficulty\(^3\): first, \(RPS_{i,t-\tau}\) are calculated for each \(t\) and \(\tau\); second, their deviation values \((DV_{i,t-\tau})\) are obtained for each \(t\) so that the mean \(\mu_{i,\tau}\) and the standard deviation \(\sigma_{i,\tau}\) are normalized to 50 and 10, respectively; and finally, we take an average of deviation values across \(t\) and \(\tau\).

\[
RPS_{i,t-\tau} = \sum_{\tau=1}^{K} \left( P_{i,t-\tau}(B_{j_{\tau}}) - Y_{t} \right)^2,
\]

\[
DV_{i,t-\tau} = 50 + 10 \times (RPS_{i,t-\tau} - \mu_{t-\tau}) / \sigma_{t-\tau}
\]

\[
ADV_i = \frac{1}{T} \sum_{\tau=1}^{T} \frac{1}{D_i} \sum_{\tau=1}^{D_i} DV_{i,t-\tau}
\]

\(^3\) Another reason for using deviation values is that we need to add up scores across different forecast horizons. The longer the forecast horizon is, the more difficult it is to forecast, which implies that it is necessary to use some weights to adjust for the difference to add up the scores for each individual forecaster. Because the weights are unavailable, we decide to use the deviation values.
3. Data

The ESPF has collected density forecasts of real GDP growth and CPI inflation rates of both the current and the next fiscal year\(^4\) made by professional forecasters every month since June 2008. Each forecaster is requested to attach probabilities that growth and inflation rates fall to each bin. MPD (Mean Probabilities Distribution), the average of individual densities, are published every month.

Developments of the MPDs for GDP growth and CPI inflation rates are respectively shown as box plots by target and period (Figure 1). The boxes represent their 50% inter-quartile ranges, and the tails represent the upper and lower quartiles. The horizontal dot lines indicate realisations. Note the actual real GDP growth rates in FY 2008 and 2009 turned out to be unexpectedly low, locating below many MPDs, especially in long forecast horizons. This suggests many forecasters tended to underestimate negative effects on the output growth of the Great Recession. However, the bias is less visible in case of CPI inflation rate, especially in FY 2008, thanks to the sticky nature of price movements.

June forecasts will be used for the independence test because they are the first forecasts based on a set of actual outcomes in the previous fiscal year\(^5\). A series of June forecasts represent an unbroken sample of one-year density forecasts with no overlapping information. Fan charts in Figure 2 illustrate every June MPD for the current and the next fiscal year, providing visually a set of prediction intervals covering 10, 25, 50, 75 and 90 percentiles.

\(^4\) The Japanese fiscal year starts in April and ends in March.
\(^5\) The actual outcomes available at the end of April for CPI inflation and in early May for GDP growth rate, respectively.
Figure 1 MPDs and Realised Values

(1) Real GDP Growth Rates

(2) CPI Inflation Rates
Figure 2 Fan Charts
(1) Real GDP Growth Rates

June 2008 GDP Forecast

June 2009 GDP Forecast

June 2010 GDP Forecast

June 2011 GDP Forecast

(2) CPI Inflation Rates

June 2008 CPI Forecast

June 2009 CPI Forecast

June 2010 CPI Forecast

June 2011 CPI Forecast
We now turn to individual density forecasts. Figure 3 shows how many bins each forecaster actually uses. A few bold forecasters put 100 percent in a bin, while very cautious ones put in more than ten! The average is about 4.5 for both GDP and CPI. Moreover, both bin distributions look very similar, although the width of each bin is a half percentage point for GDP, twice as large as that for CPI.

The individual densities are also presented in terms of actual outcomes. Figure 4 shows to what extent the individual densities cover the outcomes, i.e. whether the realized values are considered possible by each forecaster. Lower probabilities for GDP than for CPI suggest the former is subject to more risks that actual outcomes may be “outliers” in terms of individual density forecasts. Note there still remain sizeable outlier risks even in a relatively short forecast horizon, as observed in FY 2009 and 2010 GDP growth rates. Here the MPD has a clear advantage because the realized values are located within the range of the MPD (Figure 1(1)).

There seems to be a puzzle: differences in outlier risks between the two variables observed in Figure 4 have no visible counterparts in bin distributions in Figure 3. A possible interpretation may be “over-confidence” on the part of GDP forecast and/or “under-confidence” on the part of CPI forecast.
Figure 3 Number of Bins Used

Table:

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
<th>FY2010</th>
<th>FY2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>3.88</td>
<td>4.66</td>
<td>4.35</td>
<td>4.59</td>
</tr>
<tr>
<td>STD</td>
<td>1.34</td>
<td>1.98</td>
<td>1.74</td>
<td>1.79</td>
</tr>
<tr>
<td>#</td>
<td>428</td>
<td>631</td>
<td>677</td>
<td>679</td>
</tr>
</tbody>
</table>

Figure 4 Hitting Ratio of Density Forecast

Table:

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
<th>FY2010</th>
<th>FY2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>3.89</td>
<td>4.51</td>
<td>4.53</td>
<td>4.53</td>
</tr>
<tr>
<td>STD</td>
<td>1.32</td>
<td>1.64</td>
<td>1.78</td>
<td>1.66</td>
</tr>
<tr>
<td>#</td>
<td>346</td>
<td>551</td>
<td>599</td>
<td>623</td>
</tr>
</tbody>
</table>
4. Results

4.1 Independence Test

Following the procedures explained in Section 3.1, it is necessary to test whether transformed \( z \) is a failure or not. The failures are counted in Table 2 and dropped when testing. The actual growth rates turn out to be outliers for all of the individual forecasts in June 2008 and for two-thirds in June 2009. As a result, we will use the June forecasts in 2010 and 2011 only for testing independence. OLS will be applied to these samples, where individual effects are assumed to be the same, because of two-period time series.

<table>
<thead>
<tr>
<th>period</th>
<th>Real GDP growth rate</th>
<th>CPI inflation rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of sample</td>
<td># of failure</td>
</tr>
<tr>
<td>June 2008</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>June 2009</td>
<td>27</td>
<td>20</td>
</tr>
<tr>
<td>June 2010</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>June 2011</td>
<td>40</td>
<td>1</td>
</tr>
</tbody>
</table>

As for CPI inflation rate, the outcome is not covered by about half of density forecasts in June 2011. Fortunately, limiting individuals to those with consecutive time series forecasts still keeps sufficiently large samples that we will use Arellano-Bond method for dynamic panel.

Table 3 summarizes the estimation results. According to the Column Eq.1, the null hypothesis, \( H_0: \rho = 0 \), is not rejected in 34 real GDP growth rate samples. Thus, a sequence of density forecasts of GDP growth rate \( \{p(y_t)\}_{t=1}^m \) may coincide with date generating process, \( \{f(y_t | \Omega_t)\}_{t=1}^m \).
Table 3 Estimation Results

<table>
<thead>
<tr>
<th></th>
<th>Eq.1</th>
<th>Eq.2</th>
<th>Eq.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent var.</td>
<td>GDP growth rate</td>
<td>CPI inflation rate</td>
<td></td>
</tr>
<tr>
<td>$\rho_1$ estimate</td>
<td>0.055</td>
<td>-0.587</td>
<td>-0.605</td>
</tr>
<tr>
<td>standard error</td>
<td>0.255</td>
<td>0.236</td>
<td>0.565</td>
</tr>
<tr>
<td>p-value</td>
<td>0.830</td>
<td>0.013</td>
<td>0.285</td>
</tr>
<tr>
<td>Sample period</td>
<td>FY 2010 to 2011</td>
<td>FY 2008 to 2011</td>
<td>FY 2008 to 2011</td>
</tr>
<tr>
<td># of observations</td>
<td>34</td>
<td>29</td>
<td>14</td>
</tr>
<tr>
<td># of individuals</td>
<td>20</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td># of instruments</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

However, as for CPI inflation rate, the null is rejected in the full sample with 20 forecasters (see Column Eq. 2). Then, we select 10 out of 20 individuals, who are superior to the rest in terms of forecast performance, and fail to reject the null in this subsample (see Column Eq. 3).

4.2 Results of RPS calculations

It may be desirable to only use the June samples because their data generation process is likely to coincide with the unobserved true one, as shown in Section 4.1. However, focusing on the June samples is costly, since most of the collected data gets dumped. Thus, we will use not only the June samples, but also the samples of 17 months for GDP and those of 16 months for CPI, following the ESPF’s annual performance reviews\(^6\).

A look at Figure 1 may pose another problem: how should the FY2008 samples be handled? Their forecast errors are quite large, due to worldwide financial shocks, compared with the other years’ counterparts, thereby heavily affecting our calculation results.

Based on the above consideration, we will calculate RPSs for four cases, 2 (June samples only, and 17 or 16 month samples) by 2 (samples with and without FY 2008

---

\(^6\) See Komine et al. (2009) for the details of the annual performance review.
data), for each of real GDP growth and CPI inflation rates.

We will compare a RPS of the MPD with those of three benchmarks, and also with those of individual density forecasts, following Kenny, Kostka, and Masera (2012). The first benchmark is a uniform distribution forecast, assigning equal probabilities to all bins covered by the MPD. The second is a normal distribution, the mean and the variance of which are set equal to the ESPF’s CF and that of the corresponding forecast errors, respectively. The third is a naïve forecast arbitrarily set equal to the MPD in the previous month. Figure 5 provides benchmark examples of real GDP growth and CPI inflation rates in June 2011.

The RPS and ADV calculation results are reported in Table 4, where smaller scores mean better performance, as is obvious from Eq.(4) and (7). Table 4 also reports rank orders of the MPD and benchmarks, compared with individual. The MPD ranks about 5th out of about 35 forecasters and, therefore, is a “good” density forecast.

The MPD always beat two benchmarks, Uniform and Naïve. The performances of the former are rather poor, ranking near the end of the list in seven out of eight cases. The performance difference between the MPD and the Naïve may indicate how valuable newly arrived information is after the previous forecast was finalized.

The performance of the Normal benchmark depends on a target variable. The Normal ranks in the top 10 percent of the participants in the case of the CPI inflation rate, while belonging to a bottom half in the case of output growth rate. As is observed from Figure 5, the normal benchmark has smaller variance than the MPD in CPI, and a larger one in GDP. Hence, as for CPI, subjective judgement contained in the MPD spreads out risks concentrated in a few bins, thereby worsening the score. But the situation may be just the opposite in the output growth rate: the judgement concentrate probabilities in fewer bins and succeed to raise the score. This may mean that different characteristics are required to be a “good” forecaster: prudence for CPI inflation and overconfidence for real GDP growth rate.
Figure 5 Benchmark Density Forecasts
(1) Real GDP Growth Rate in June 2011

(2) CPI Inflation in June 2011
Table 4: Results of RPS Calculations

<table>
<thead>
<tr>
<th></th>
<th>period evaluation</th>
<th>GDP</th>
<th>Uniform</th>
<th>Normal</th>
<th>Naïve</th>
<th>number of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MPD</td>
<td>score</td>
<td>ranking</td>
<td>MPD</td>
<td>score</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>June samples (1)</strong></td>
<td>FY2009 to 2011</td>
<td>RPS 1.29</td>
<td>1.34</td>
<td>1.56</td>
<td>1.37</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADV 46.26</td>
<td>56.47</td>
<td>57.00</td>
<td>48.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Including FY2008</td>
<td>RPS 3.10</td>
<td>2.81</td>
<td>3.15</td>
<td>23</td>
<td>nan.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADV 46.22</td>
<td>47.08</td>
<td>51.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CPI</strong></td>
<td>FY2009 to 2011</td>
<td>RPS 0.45</td>
<td>1.06</td>
<td>0.39</td>
<td>0.51</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADV 46.35</td>
<td>57.86</td>
<td>45.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Including FY2008</td>
<td>RPS 0.42</td>
<td>1.01</td>
<td>0.33</td>
<td>2</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADV 46.19</td>
<td>57.58</td>
<td>44.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>17 or 16 month samples for each FY (2)</strong></td>
<td>FY2009 to 2011</td>
<td>RPS 0.70</td>
<td>1.08</td>
<td>0.39</td>
<td>0.51</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADV 46.78</td>
<td>55.34</td>
<td>51.36</td>
<td>48.94</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Incl. FY2008</td>
<td>ADV 46.80</td>
<td>52.87</td>
<td>49.98</td>
<td>51.27</td>
<td>19</td>
</tr>
<tr>
<td><strong>CPI</strong></td>
<td>FY2009 to 2011</td>
<td>RPS 0.40</td>
<td>1.01</td>
<td>0.33</td>
<td>0.51</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADV 46.06</td>
<td>61.08</td>
<td>44.36</td>
<td>48.48</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Incl. FY2008</td>
<td>ADV 46.13</td>
<td>59.24</td>
<td>45.14</td>
<td>48.88</td>
<td>17</td>
</tr>
</tbody>
</table>

Note: (1) Calculated from individuals without missing responses.  
(2) Calculated from individuals with more than 70 per cent responses.
5. Conclusion

The density forecasts the ESP Forecasts (or ESPF) has collected every month since June 2008, are evaluated by two approaches: Probability Integral Transform (PIT) and Ranked Probability Score (RPS).

First, we apply Berkowitz’s (2001) test to individuals’ density forecasts because the small sample of the ESPF disallows Diebold, Gunther and Tay’s (1998) graphical or nonparametric approach. June forecasts are examined because they are the first ones formulated using a set of actual outcomes of the previous fiscal year, i.e. without overlapping information. As for real GDP growth rates, we drop FY 2008 and 2009 data, which are greatly affected by the Great Recession, and fail to reject the null of independence in FY 2010 and 2011. As for CPI inflation rates, we reject the null in the period of FY 2008 to 2011 in the full sample of 20 individuals, but fail to reject it in its subsample of 10 individuals with better forecast performance.

Second, we compare the Mean Probability Distribution (MPD), the average of individual densities, with three benchmarks and individuals’ density forecasts by calculating RPS. MPD turned out to be a “good” density forecast, ranking around fifth during the sample period and beating the benchmarks in most cases. The good performance of the MPD is robust enough to stand up to changes in sample periods and variables. Subjective judgments added to the MPD are likely to deteriorate the performance in the case of CPI, but to improve in the case of GDP.
References


Kawagoe, Masaaki (2012) “What Has Been Learned in Eight Years since the Launch of the ESPF ?: Reviews and Prospects,” Nakahara Encouragement Award Lecture given at the annual meeting of the Japan Association of Business Cycle studies held on November 17. (in Japanese)


## Appendix

The table below shows the samples used in the independence test in Table 3.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Eq.1</th>
<th>Eq.2</th>
<th>Eq.3</th>
<th>Sample ID</th>
<th>Eq.1</th>
<th>Eq.2</th>
<th>Eq.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>36</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>37</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>38</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td></td>
<td></td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>41</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td></td>
<td></td>
<td>42</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>43</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>X</td>
<td>X</td>
<td></td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td></td>
<td></td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>46</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>48</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>49</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>X</td>
<td></td>
<td></td>
<td>51</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td>52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>X</td>
<td></td>
<td></td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td>54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>X</td>
<td>X</td>
<td></td>
<td>55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td>57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>X</td>
<td>X</td>
<td></td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>61</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>X</td>
<td>X</td>
<td></td>
<td>62</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td>63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td>65</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td>68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td>69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>X</td>
<td>X</td>
<td></td>
<td>70</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>