ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Forecasting *(evaluation)*

Andrea Guizzardi
Department of Statistical Sciences
University of Bologna

Sources:

Diebold (200X), Elements of forecasting, South Western ed.

- Cap.2 Six consideration basic to succesful forecasting
- Cap 4. Modelling and forecasting trend
- Cap 12 Units roots, Stocastic trends, ARIMA Forecasting Models and Smoothing

Guizzardi A. (2002) « La previsione Economica » Guaraldi Ed. Rimini

Capp.2,6,12

Forecasting

Despite frequent attempts to anticipate the future are often frustrated by the observed results, the forecasting exercises continue to be a subject of study for science and of interest to operators as:

- they provide information about the determinants of a phenomenon
- they guide the actions of practitioners being a tool to measure - and evaluate - the effectiveness of different strategies.

In social sciences prediction influence the behavior of the operators. The shift between forecasts and realizations reflects mainly the effectiveness of the forecast in influencing the action of the'' decision-maker'' .

Competitive advantages may result from a careful management and planning. In this a key role is played by expectations about the future formed on the basis of subjective or quantitative predictions.

To be defined:

1. availability of resources;
2. horizon, frequency, and areal detail;
3. availability and reliability of data sources;
4. timeliness and accuracy required.

Forecasting Evaluation

Regardless of the forecasting method used, there is the problem to evaluating the reliability of a forecast. In doing so the focus CAN NOT be the shift between the forecasts and realizations (the forecasting error size)

Example 1 (Granger 1989)

A financial advisor (fraudster) get the e-mails of 1024 financial newsletter subscribers.

He/she writes to 512 subscribers recommending to purchase a financial instrument because price will rise in the next month. At the same time, the remaining 512 are advised to take short positions, because the same instrument will lose its value.

Forecasting Evaluation

Whatever the result, the advisor repeats the strategy, writing to a half of the subscribers who have received the exact prediction that in the next month the price of the financial instrument will increase.The remaining 256 subscribers will be advised to open short positions.

After 3 additional rounds, using the same technique, there will be 32 people probably willing to pay to receive the forecast for next month performance.

Thus…

In evaluating the consultant's "forecasting model", it might be better to think that usually those who are able to advise with this degree of accuracy on price dynamics, do not sell their predictions; they use them!

Forecasting Evaluation

- In the definition of an objective criterion for assessing the forecasts is obvious to consider the comparison with what is going to accomplish.
- Unfortunately assess today whether or not a prediction approaches to reality is impossible by definition since the judgment would imply knowledge of the future.

- If you can not evaluate a forecast, it is possible to **evaluate the forecasting process** analyzing the properties of a sample of ex-post forecasts (that is in a context that is controlled experimentally - in the broadest sense).

Forecasting Evaluation

- It should be emphasized that, even ex-post, the distance between predicted (fitted) and observed values (actual) does not allow a unique evaluation of the forecast.
- There isn't an objective criterion to indicate how fitted and actual should be close since the evaluation depends on:
  - the loss function
  - the nature of the observed phenomenon, in relation to its degree of (un)predictability, conditionally to the availlable information (the weigth of the error component in the DGP that generated the observed time series)

- Thus **is not possible to evaluate a forecast by means of graphical comparison** between actual and fitted values.

Forecasting Evaluation

Example 2

If, for example, the observations are generated by the following MA process :

$$y_{t+1} = \theta_1 \varepsilon_t + \varepsilon_{t+1}$$

the optimal prediction one step ahead is a fraction (theta) of the realization of a white noise process.

The fitted time series is thus expected to follow a different trajectory respect to actual time series; thus in grafical analysis the observed data (actual) and their optimal forecast could appear distant or asynchronous.

Not being able to judge a predictor on the proximity of his ex-post predictions to actual data, do not exclude to assess "the process" of prediction, meaning the predictors properties

Forecasting Evaluation

(single model)

One step ahead forecast errors associated with an optimal forecaster

$$e_{t-H+h} = y_{t-H+h} - \hat{y}_{t-H+h} \qquad h = 1,2,\ldots,H$$

should comply with certain properties that can then be directly used as a yardstick. A good forecaster must indeed generate errors:

- unbiased (with zero-mean);
- random; ie which do not systematically present the same sign (prediction has not to over or under estimate systematically the ex-post values of the variable of interest).
- In the case of time series, the residuals should not be autocorrelated;

- These properties must hold for all t, and therefore for the H ex-post forecasts.

Forecasting Evaluation

(single model)

Testing unbiasdeness
is possible through an OLS regression of the prediction errors on the constant term.

$$e_h = c + \varepsilon_h \quad h = 1,2,...,H$$

If the estimated regression coefficient (c) differs significantly from 0, then the null hypothesis that the predictions are unbiased is rejected.

Note that not necessarily if the errors have 0 mean they are randomly distributed. the previous test never reject unbiasdeness if the residual sum is zero.
This could also happen if all the errors have the same sign except one of magnitude equal to: minus the sum of remaining errors.

Thus is also useful to support previous analysis testing for the randomness in the numbers of positive (or negative) signs.

Forecasting Evaluation

(single model)

To test the null of absence of a sistematic tendency to over or under estimate actual values, is possible to calculate the following $\chi 2$ statistic :

$$\chi^2 = \frac{2 \cdot \left(O_+ - E_+\right)^2}{E_+}$$

where O+ and E+ denote the number of Observed and Expected overestimation errors (positive errors).
The statistic is distributed as a Chi-square with a degree of freedom and can be used to reject the hypothesis of equal relative frequency between overestimation (+) and underestimation (-) errors.

It should be noted that the test does not reject the null hypotesis even if same sign prediction errors' are clustered in only two sequences (runs) of cardinality H/2.
A further test the signs sequence is therefore necessary

Forecasting Evaluation

(single model)

The runs test

Runs tests can be used to test the randomness of a residual sequence by marking the positive residuals with + and the other with − (zero are omitted)

A "run" of a sequence is a maximal non-empty segment of the sequence consisting of adjacent equal elements. For example, the sequence

"+++−−+−−++++++"

consists of **five runs**, three of which consist of +'s and the others of −'s.

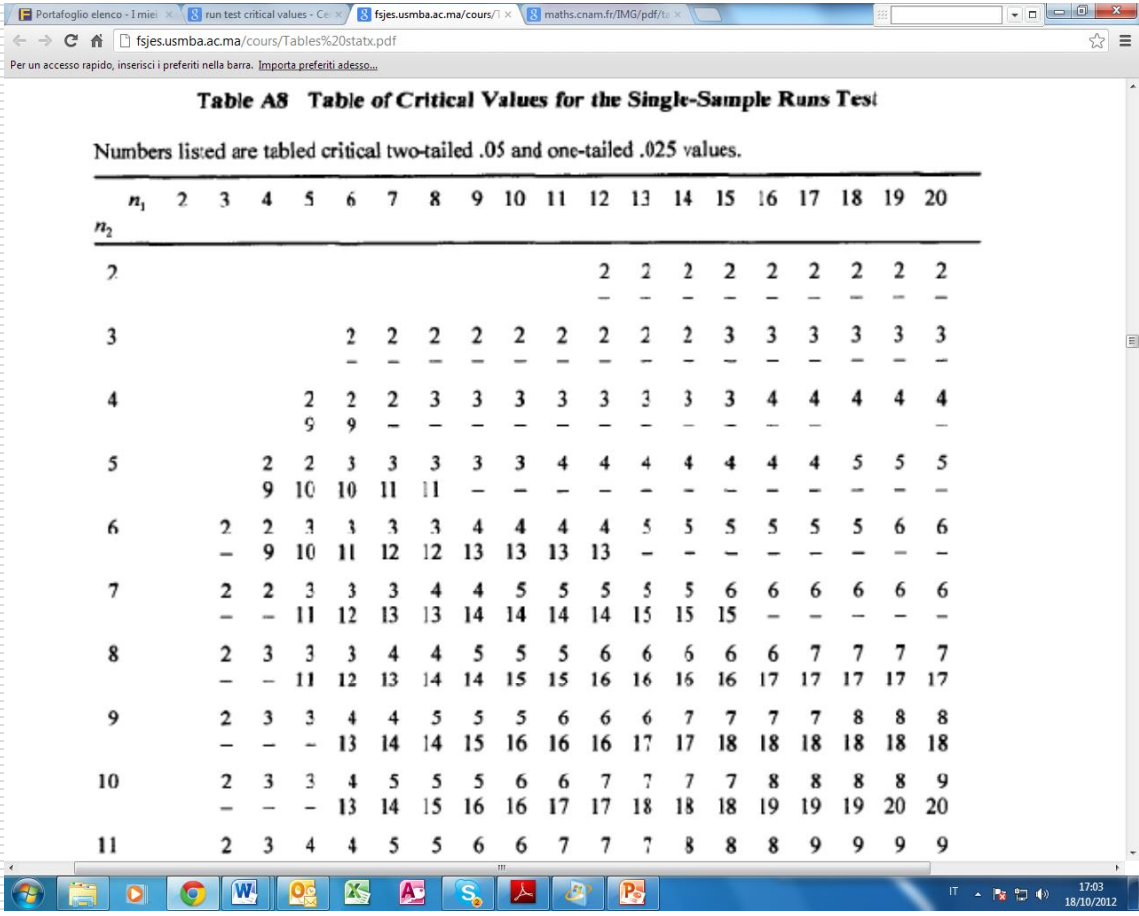
If +'s and −'s have equal probabilities under the null hypothesis that the two elements (+ and -) follow a random sequence, the number of runs in a sequence of length **n** is approximatively (for **n**>20) a Normal random variable with:

- Mean $\mu = 1 + \frac{2 \cdot n_1 \cdot n_2}{n}$; n1 / n2 = # positive / negative values
- Stand. Dev. $\sigma = \sqrt{\frac{(\mu-1)(\mu-2)}{n-1}}$; n=n1+n2

Forecasting Evaluation

(single model)

- If the number of runs is significantly higher or lower than expected, the hypothesis of statistical randomness of the sequence in the residual signs may be rejected.
- If **n**<20 use the critical values shown beside.



fsjes.usmba.ac.ma/cours/Tables%20statx.pdf

Per un accesso rapido, inserisci i preferiti nella barra. Importa preferiti adesso...

Table A8 Table of Critical Values for the Single-Sample Runs Test

Numbers listed are tabled critical two-tailed .05 and one-tailed .025 values.

n_2 \ n_1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
											–	–	–	–	–	–	–	–	–
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
					–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
				9	9	–	–	–	–	–	–	–	–	–	–	–			–
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
			9	10	10	11	11	–	–	–	–	–	–	–	–	–	–	–	–
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
		–	9	10	11	12	12	13	13	13	13	–	–	–	–	–	–	–	–
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
		–	–	11	12	13	13	14	14	14	14	15	15	15	–	–	–	–	–
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
		–	–	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
		–	–	–	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
		–	–	–	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9

Forecasting Evaluation

(rival models; descriptive framework)

Defined a loss function, rival model forecasting performance can be ranked (descriptive ranking) considering the cost associated to the error of each model.

In case that the operational consequences of a prediction error can be summarized by means of a quadratic loss function:

$$FMSE = 1/H \sum_{h=1}^{H} \left(y_{t+h} - \hat{y}_{t+h} \right)^2$$

different specifications are ranked considering FSME of the H ex-post forecasting errors.

The forecasting model with the minimum FMSE win the competition.

This (descriptive) way to evaluate forecasting performance can be extended easily to setting with other loss functions

Forecasting Evaluation

(rival models; inferential framework)

Descriptive comparison does not take into account the role played by sampling variability in determine the loss function values associated to competing models. Some inferential evaluation strategies are available upon all of part of the following hypotesis:

- (1) Loss is quadratic,
- (2) the forecast errors are:
  - (2a) zero mean
  - (2b) Gaussian
  - (2c) serially uncorrelated
  - (2d) contemporaneously uncorrelated

Hypotesis 2a) can be tested with usual unbiasdeness tests (see previous slides) Many normality test are available for hypothesis 2b); serial (un)correlation tests are suitable for hypothesis 2c).

Forecasting Evaluation

(rival models; inferential framework)

If all previous hypothesis holds the simple F Test can (the ratio of sample variances) can used to test the equality in the performance of two rival models (say M1 and M2). Infact hypothesis 2a) imply that FMSE is the forecasting error variance.

$$H_0 : \sigma^2_{M1} = \sigma^2_{M2}$$

A test for H0 is then given by the ratio of two sample variances s2

$$TU = \frac{s^2_{M1}}{s^2_{M2}} = \frac{\sum_{h=1}^{H} e^2_{M1,h}}{\sum_{h=1}^{H} e^2_{M2,h}} \sim \mathrm{F}^{\mathrm{H}}_{\mathrm{H}}$$

where the greatest variance is placed at numerator. The statistic is distributed as an F with H degrees of freedom in the numerator and the denominator, the critical values are tabulated.

If Random Walk errors are placed at denominator, the statistics is known as Theil-U statistic

Forecasting Evaluation

(rival models; inferential framework)

- However, the TU statistic is of little use in practice because the conditions required are too restrictive.
- Assumption (2d) is particularly unpalatable. Its violation produces correlation between the numerator and denominator of TU, which will not then have the F distribution.

- In the case of non-zero correlation between the two errors series (e1 and e2), the TU statistical is distributed as a F with a fat right tail. Thus is conservative, in the sense that rarely leads to reject the null hypothesis at the usual confidence levels.

Forecasting Evaluation

(rival models; inferential framework)

- To test whether the differences between the FMSEs of rival forecasts can be attributed to sampling variability, is possible to use the Harvey version of the (Morgan) Granger and Newbold test - which remains consistent in the presence of non-normally distributed heteroskedastic forecasting errors:

$$e_{+,t} = \beta e_{-,t} + \varepsilon_t$$

where $\quad e_{+,t} = e_{M1,t} + e_{M2,t} \qquad e_{-,t} = e_{M1t} - e_{M2t}$

and tests whether $\beta = 0$, in order to verify the null hypothesis of equivalent forecasting performances of the two predictors being considered. The test statistic is:

$$S_2^* = \frac{\hat{\beta}}{\sqrt{\frac{\sum e_{-,t}^2 \hat{\varepsilon}_t^2}{\left(\sum e_{-,t}^2\right)^2}}} \qquad \text{where:} \qquad \hat{\beta} = \frac{\sum e_{-,t} e_{+,t}}{\sum e_{-,t}^2}$$

which is distributed approximately as a Student'T, with (n-1) df.

Forecasting Evaluation

(rival models; inferential framework)

- A second parametric test is developed from the most general formulation used for encompassing tests

$$y_t = \alpha_0 + \alpha_1 y_{M1,t} + \alpha_2 y_{M2,t} + \varepsilon_t$$

- Where $y_{Mj,t}$ $j = 1,2$ denotes the one-step-ahead forecast

- Instead of testing for $\alpha 2=0$ (to assess whether the first forecast encapsulates all of the useful predictive information contained in the second one), we test for $\alpha_1 = \alpha_2$, to determine whether both forecasts contain an equal amount of predictive information.

Forecasting Evaluation

(rival models; inferential framework)

- To reduce the contemporaneous correlation problem (multicollinearity)is possible to orthogonalize the rival forecasts

$$y_t = \alpha_0 + \beta_1 y_{-,t} + \beta_2 y_{+,t} + \varepsilon_t$$

- By construction, $\beta_1 = 0.5(\alpha_1 - \alpha_2)$ and ; $\beta_2 = 0.5(\alpha_1 + \alpha_2)$ therefore, if the null hypothesis $\beta = 0$ is rejected, then $\alpha_1 \neq \alpha_2$.

- Moreover, if >0, M1 is the best forecasting model, provided that M1 is the model with the largest number of variables.

Forecasting Evaluation

(rival models; inference in NONPARAMETRIC framework)

Rank test

An exact finite-sample non-parametric test could be performed starting from the Wilcoxon signed-ranks test for paired samples, implemented on the differences between the absolute errors of two rival forecasts (say, M1 and M2):

$$TW = \sum_{h=1}^{H} I_{+} \cdot rank\left(\left|e_{M1,t}\right| - \left|e_{M2,t}\right|\right)$$

where I+ is an indicator function, equal to 1 when the absolute error of M1 is larger than that of M2 and equal to 0 otherwise.

In the TW statistic the differences between the paired MAPEs are computed, and the ranks associated with positive differences are summed.

Forecasting Evaluation

(rival models; inference in NONPARAMETRIC framework)

- The null hypothesis being tested is that the rank sum for the positive differences is equal to the rank sum for the negative differences, in which case the forecasting performance of the two models being compared is similar.
- Critical values are tabulated
- In the asynptotic case:

$$\frac{TW - \frac{H \cdot (H+1)}{4}}{\sqrt{\frac{H \cdot (H+1) \cdot (2H+1)}{24}}} \overset{A}{\sim} N(0,1)$$

- If the one-step-ahead forecasting errors are zero-mean and the assumption that they are normally distributed holds, then it is possible to perform more powerful parametric tests.

Data Snooping

- Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection. When such data reuse occurs, there is always the possibility that any satisfactory results obtained may simply be due to chance rather than to any merit inherent in the method yielding the results. (White, 2000)

- Every time a good forecasting model is found by an extensive model specification search there is always a danger that the forecasting ability of the model is given by luck. Even when there is no exploitable relation between the dependent and independent variable one can find it by trying a high number of specification on a given data-set. The model will look good but it will be useless in practice.

- One among the first which treat the problem was Halbert White who introduced the Reality Check test (RC) (2000)

- Here will be presented the Superior Predictive Ability (SPA) test introduced by Hansen (2005)

- The SPA test is more powerful and less sensitive to poor and irrelevant alternatives. These achievements are gained by the employ of a studentized test statistic that reduce the influence of erratic forecasts. Indeed the RC can be manipulated by the inclusion of poor and irrelevant forecast in the set of alternative forecasts.

Superior Predictive Ability (SPA) test

Hansen (2005)

- The SPA test allows for the simultaneous comparison of m series of forecasts
- The drawback of SPA test is that it does not aim explicitly at the identification of the models which overperform the benchmark
- Hansen defines the relative performance variable as:

$$d_{k,t} \equiv L\left(\xi_t, \delta_{0,t-h}\right) - L\left(\xi_t, \delta_{k,t-h}\right)$$

Where

$d_{k,t}$ is the performance of model k compared to the benchmark $(k = 0)$ at time t

$L(\cdot)$ is a loss function like MSE or Directional Accuracy (DA)

ξ_t is the variable of interest

and $\delta_{k,t-h}$ is the k-th decision rule (e.g. the h-step-ahead forecast of ξ_t).

Superior Predictive Ability (SPA) test

Hansen (2005)

$\boldsymbol{d}_t$, t = 1,..., n, - where $\boldsymbol{d}_t$ is the vector of relative performance - is viewed as our data, and we therefore state all assumptions in terms $\boldsymbol{d}_t$

Given $\boldsymbol{\mu} \equiv E(\boldsymbol{d}_t)$ the null hypothesis can be formulated as

$$H_0: \boldsymbol{\mu} \leq \mathbf{0}$$

It can be stated as: **none of the employed models is more accurate than the benchmark δ_0**

The studentized test statistic is: $$T_n^{SPA} \equiv \max\left[\max_{k=1,\dots,m} \frac{n^{1/2}\overline{d}_k}{\widehat{\omega}_k}, 0\right]$$

$\widehat{\omega}_k^2$ is a consistent estimator of $\omega_k^2 \equiv var\left(n^{1/2}\overline{d}_k\right)$ and $\overline{d}_k = \frac{1}{n}\sum_n d_{k,t}$

The p-values are calculated with bootstrap resamples.

Directional forecasting

Directional forecast

→ the ability to predict market movements.

Directional Forecasting Evaluation

(see *Blaskowitz, O; Herwartz, H (2008) Testing directional forecast value in the presence of serial correlation, SFB 649 Discussion Paper 2008-073*

- The Problem:

  Suppose that upon examination a variable has been found historically to have increased by 0.5 units in every period. Further to this, suppose that two sets of forecasts were generated for this variable (denoted as A and B) and these predicted a decrease of 0.5 units in every period (forecast A), and an increase of 100.5 units in every period (forecast B). Which is then the more accurate set of forecasts?

  Obviously the errors associated with the forecasts are vastly different with A having an (absolute) error of 1 unit and B having an error of 100 units → typical forecast evaluation statistics will lead to A being viewed as 'better' than B.

  However, suppose that an investor should decide whether to sell (go short on) or buy (go long on) a futures contract. In this case (where market movement matters) B is the preferred forecast as it predicts correctly the direction of change. Although this is a very simplified and extreme example, it does illustrate that accuracy in terms of predicting direction and returning 'small' errors will not necessarily coincide.

  Leitch and Tanner (1991) show that profit has a stronger and more significant relationship with directional accuracy than accuracy as measured by conventional forecast evaluation statistics.

Leitch, G. and Tanner, E. (1991) 'Forecast evaluation: Profits versus the conventional error measures', American Economic Review, 81, 580-590.

(FROM http://www.economicsnetwork.ac.uk/showcase/cook_directional)

Serial correlation should be considered.

Pearson's test in multiway table

- Gleser and Moore (1985) demonstrated that under positive dependence between successive observations, the Pearson chi-squared test has a **null distribution** that is asymptotically larger than that obtained under serial independence.
- Porteous (1987) extended Tavare's results to multiway tables and showed that Pearson's test will be valid when **all but one** of the variables under consideration are serially independent.

	OBSERVED			
	Positive outcome	No change out.	Negative Out.	Total
Positive forecast	10	9	7	26
no change forec.	15	12	6	33
Negative forecast	5	7	11	23
Total	30	28	24	**82**

	EXPECTED			
	Positive outcome	No change out.	Negative Out.	Total
Positive forecast	9,51	8,88	7,61	26,00
no change	12,07	11,27	9,66	33,00
Negative forecast	8,41	7,85	6,73	23,00
Total	30	28	24	**82**

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} = 6{,}4$$

p=0,171 → not rejec

Pesaran-Timmermann test (1992)

- The Pesaran-Timermann test (PT) is a **distribution-free** procedure which test the correct prediction of the direction of change in the variable under consideration.
- It is of particular interest when the underlying probability distribution of forecast is difficult to derive analytically or when the forecast are made available only in form of qualitative data.

The 2 x 2 case

Let $x_t = E(y_t|\Omega_{t-1})$ be the predictor of y_t formed respect to the information available at time $t-1$ and suppose that there are n observations on (y_t, x_t). The PT test is based on the proportion of times that the direction of change in y_t is correctly predicted in the sample.

- The test requires that the probability of changes in the direction of y_t is time-invariant and does not take the extreme value of 0 and 1.
- The **null hypothesis** under test is that y_t and x_t are independent

Implementation of Pesaran-Timmermann test

The formula of PT is set out below. This version of the test statistic is provided by Granger and Pesaran (2000)

$$PT = \frac{\sqrt{N}\left(\frac{N_{pp}}{N_{pp}+N_{np}} - \frac{N_{pn}}{N_{pn}+N_{nn}}\right)}{\left(\frac{\hat{\pi}_f(1-\hat{\pi}_f)}{\hat{\pi}_o(1-\hat{\pi}_o)}\right)^{1/2}} \sim N\,(0,1)$$

The subscripts p and n indicate "positive" and "negative" respectively so that N_{pn} is the number of times the actual outcome is "negative" when the forecasted value is "positive". $\hat{\pi}_o = \frac{N_{pp}+N_{np}}{N}$ is the probability that observed outputs are up and $\hat{\pi}_f = \frac{N_{pp}+N_{pn}}{N}$. is the probability that outcomes are forecast to be up.

An example....
Neural network vs Logit

- we compare L1, $N1_{16}$ and the Naïve specifications with both PT (Pesaran-Timmerman) and SPA tests. The Naïve model is defined as the model which forecasts always a negative return.
  - The PT test highlights that the market timing ability is different across the models
  - The superior predictive ability (SPA) consider each model as benchmark in turn. The p-values are calculated with 1000 bootstrap resamples assuming the negative DA (directional accuracy) as loss function.
  - A low p-values indicate that the benchmark predictive accuracy is not inferior to the predictive accuracy of at least one of the alternative specifications.

Results confirm that $N1_{16}$ is the only model providing systematic and sizeable improvements in predictive accuracy respect to the competitor

	PT	SPA
Naïve	/	0.003
L1 (Logistic regression)	1.69 (p<0,1)	0.011
$N1_{16}$	5.19 (p<0,01)	1

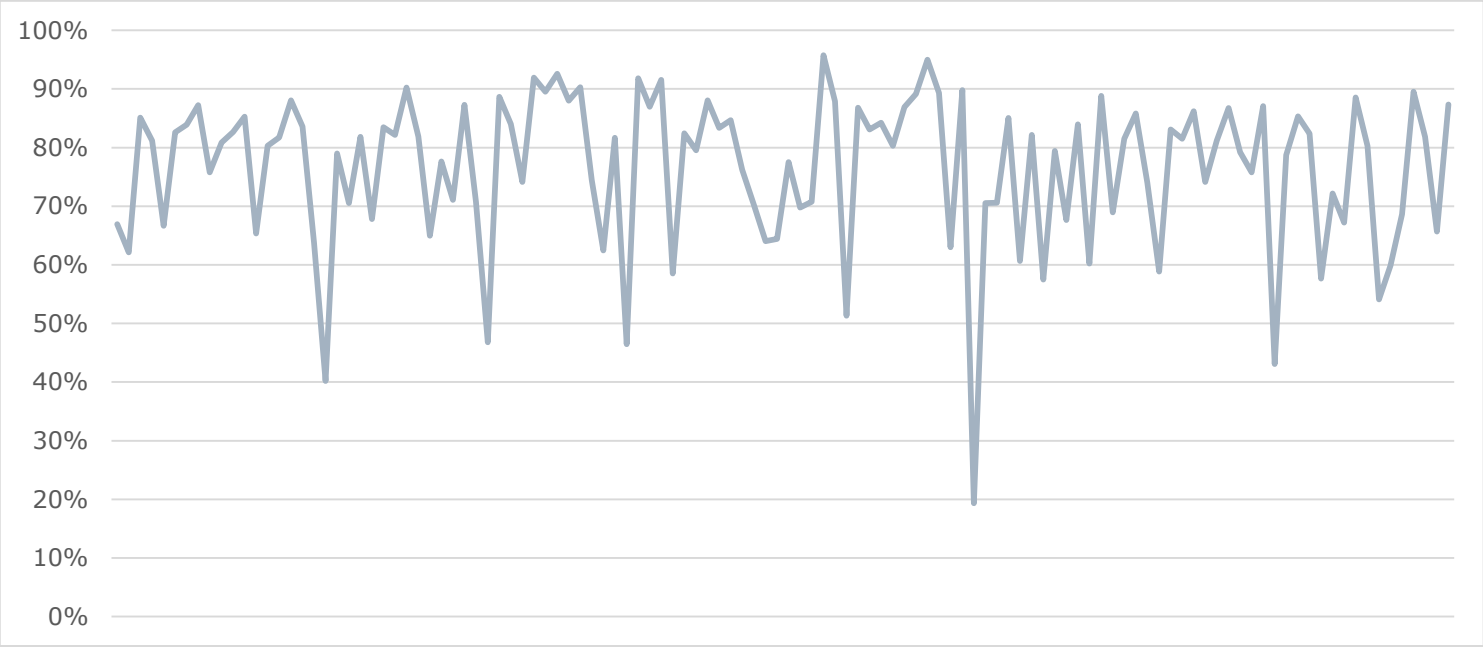
An example: Volatility forecast

Problem -> forecast the VIX index o its futures?

- The majority of literature focuses on the prediction of VIX index, this is acceptable from a theoretical point of view but origin **a "trading bias"**
- On one hand practitioners can approximate the VIX index by a straddle of options. In this case the "bias" arises from (no) Delta-neutrality (*the rate of change of the option value with respect to changes in the S&P 500 price is not equal to zero during the trading day*)
- On the other hand, it is possible to trade future contracts on VIX. In this case practitioners who aim to use econometric forecasts of VIX dynamics are exposed to the "poor" correlation between VIX and VXF returns (Degiannakis, 2008).
- Finally, it should be considered that when returns are calculated as the log closing price difference (between two consecutive days) other bias arises from the jumps present in the points where the series rolls to the new contract.

Correlation between VIX and VXF

- Correlation between VIX and its future (calculated on OTCR) takes value **0.748** on the whole sample (2454 observations), rising to **0.765** if we consider an average of the 116 correlations computed on monthly sub-samples. The correlation's distribution it is quite erratic with many spikes around 0.4 and one minimum near 0.2
- It is noteworthy to say that the gap on OTCR is more serious than that measured on daily returns as in this latter case correlation reaches **0.8248** on the whole sample and **0.7838** when calculated as an average of monthly sub-samples



The idea of coincident indicator

- We consider both lagged endogenous and exogenous variables selected among the Asian world stock indices that closes right before the US market open in the hypothesis that their movements already account for possible market sentiment on latest economic news or response to progress in major world affairs (Shen et al., 2012)

- The autocorrelation analysis suggest to keep lag 0 and 1 of the Nikkei 225 (N225), Hang Seng (HSI), Australian ASX200 and SENSEX (BSESN) returns

Exchange	Open	Close
Tokyo	01:00	07:00
Hong Kong	02:30	09:00
Sydney	01:00	07:00
Bombay	04:45	10:30
New York	15:30	22:00

Logit

- The logistic models were estimated with maximum likelihood on a sample of T_2= 2086 observations ranging from 27 March 2007 to 08 July 2015
- One model was estimated with both lagged and coincident variables (L10) and not all variable turn out to be significant
- After a backward stepwise selection process only the current BSESN is retained being highly significant variable and with the expected negative sign (L1)

- L10 shows a better goodness-of-fit but a worst BIC criterion, so model with the current BSESN index return as single explicative turn out to be our best choice.
- The BSESN has the expected sign

- **It is worth to note that BSESN dynamic is the most recent information available in our data set (represents the market that close last).**

	L10	L1
Intercept	-0.2570***	-0.2509***
$OTCR_{t-1}$	-0.011	
$OTCR_{t-2}$	-0.006	
$ASX200_t$	-0.0882	
$ASX200_{t-1}$	-0.0773	
$BSESN_t$	-0.1053***	-0.0942***
$BSESN_{t-1}$	0.0462	
HSI_t	0.0385	
HSI_{t-1}	0.0297	
$N225_t$	0.0306	
$N225_{t-1}$	-0.0064	
BIC	2.9008×10^3	2.8635×10^3
Pseudo R^2	0.0112	0.0068

Neural Network

- In order to account for a more general form of non-linearity in the relation, we also estimate a feed-forward single hidden layer neural network. As in the logistic specifications we consider two sets of inputs (i.e. the whole 10 variables and the current BSESN). Models are labelled respectively as $N10_j$ and $N1_j$ where **j** stands for the number of neurons

- The hidden layer (activated by a hyperbolic tangent function); **j** ranges from 2 to 20. We consider 2 neurons in the output layer activated by a soft-max function as they represent complementary probabilities

- Each network is trained updating weights and biases with scaled conjugate gradient.

- The **estimation sample** is constituted by T_1=1718 observations ranging from 27 March 2007 to 21 January 2014.

- To avoid overfitting we early stop the training algorithm measuring network performance on a **validation sample** of T_2-T_1=368 observation (from 22 January 2014 to 8 July 2015).

- The loss function is the Cross-Entropy as it provides similar relative errors for both big and small difference between targets and forecasted values (unlike the MSE).

- The maximum iterations is capped at 1000.

The choice of best network

- In order to choose the best network architecture (the number of neurons in the input and hidden layer) between $N1_j$ and $N10_j$ all the possible 38 architectures are compared on a **test set** of T_3-T_2 = 368 observations from 09 July 2015 to 20 December 2016.

- We use the expected performance index (EP) defined as

$$EP(N\cdot_j) = \frac{1}{K}\sum_k MDA_k(N\cdot_j)$$

Where

- $k = 1,2,\dots,30000$ are the number of simulations (each simulation starts from different random initial weights. The best $EP(N\cdot_j)$

- MDA is the Mean Directional Accuracy

- The best $EP(N\cdot_j)$ value for both N1 and N10 specification is obtained with j=16 neuron in the hidden layer which is identified as best complexity for the two models

- The expected performance of a network with only the $BSESN_t$ as input, 16 hidden nodes and 2 output neurons is $EP(N1_{16}) = 62{,}8$, while $EP(N10_{16}) = 59\%$.

Neural network vs Logit

- Both neural network specification gain a maximum MDA of 65,8%
- The logistic regression shows a much lower MDA (59,2%)

The accuracy gap between the two model comes from the greater ability of the networks to grasp the upward movements (which are less frequent in the sample).

- If we focus on upward returns the neural network MDSs are 43% and 47% (respectively for 1 and 10 inputs)
- The logistic regression performs only 2 true positive in 152 positive observations.

In order to have robust results we compare L1, $N1_{16}$ and the Naïve specifications with both PT and SPA tests. The Naïve model is defined as the model which forecasts always negative.

- The PT test highlights that the market timing ability is different across the models
- The superior predictive ability (SPA) consider each model as benchmark in turn. The p-values are calculated with 1000 bootstrap resamples assuming the negative DA as loss function.
- A low p-values indicate that the benchmark predictive accuracy is not inferior to the predictive accuracy of at least one of the alternative specifications.

Results confirm that $N1_{16}$ is the only model providing systematic and sizeable improvements in predictive accuracy respect to the competitor

	PT	SPA
Naïve	/	0.003
L1 (Logistic regression)	1.69 (p<0,1)	0.011
$N1_{16}$	5.19 (p<0,01)	1

Optimizing a trading strategy

- The accuracy does not necessarily imply profitability
- In order to check economic profits we simulate a simple trading strategy that suggest to open a long position if the forecasted probability is above a certain threshold, or to open a short one if it is under 1 minus the threshold. We remain flat when neither of the two situations is realized
- We try to maximize the trading strategy considering six different thresholds allowing to avoid "false signals" leaving out the weakest signals.
- Profits (losses) are calculated as simple sum of OTCR since the strategy open and close the position within the same day.
- Bid-ask spread are neglected, commission are set to 50 cents each contract

		$N1_{16}$	L1
No filter	/	539.6 (100%)	192.4 (100%)
Filter I	49.5% < p < 50.5%	510.7 % (95.3%)	191.4 % (99.7%)
Filter II	49% < p < 51%	518.3 % (92.7%)	232 % (99.2%)
Filter III	47.5% < p < 52.5%	490.1 % (85.9%)	281.6 % (96%)
Filter IV	45% < p < 55%	508.4 % (73%)	353.6 % (77.2%)
Filter V	40% < p < 60%	375.9 % (43.5%)	46 % (43.5%)

Trading simulation

- Finally, the profitability of the previous three optimal strategies (i.e. probability thresholds) are evaluated on a new trading set made of T_4-T_3=188 observations (ranging from 02/01/2017 to 29/09/2017) to avoid sample overlaps with the previous phase in which we set the optimal thresholds
- The goal is to test if there are significant differences between the economic performances of the best strategies that can be implemented starting from Neural network, logistic and Naïve predictions.

Rival models	Cumulative returns	# false signals (1-DA)	Returns standard dev.	maximum # consec. losses	Max drawdown	SPA test
Naïve	95.8 % (100%)	39.4%	3.2%	5	19.8%	0.110
L1	120 % (85.6%)	42.6%	2.8%	3	16.3%	0.232
$N1_{16}$	145.8% (100%)	34.0%	3,1%	4	17.6%	1

- The neural network outperforms both the benchmark and logistic regression in terms of cumulative return. It also displays the lowest number of losses, confirming the ability of the neural network to anticipate even upward movements in the VXF
- If we consider the risk the strategy built with logistic regression outputs is the best. The result depends on the applied probability filter: limiting the percentage of market entries to 86% the strategy produces the most favourable number of consecutive losses and the smallest maximum drawdown

Trading simulation 2

- It should be considered that the strategy related to neural network outputs has the highest Sharpe ratio.
- The SPA test confirms, on an inferential frame, that $N1_{16}$ provides systematic improvements in economic performance respect to competitors.

Considerations:

- High returns mean high risks, no exceptions
- The margin request for selling one VIX future start from 13500. It depend on the broker used. In moments of panic it spread widely