Validating Default Probabilities on Short Time Series

–Working Paper –

Stefan Blochwitz, Stefan Hohl, Dirk Tasche, Carsten S. Wehn*

May 7, 2004

We present two approaches to examine the accuracy of default probability forecasts for different rating grades. In particular, we analyze the respective advantages and disadvantages of the two methods. Also, the effect of independence assumptions is taken into account by modelling latent variables like the asset correlation and dependency in time. Both tests, the Extended Traffic Light Approach as well as an ad hoc normal test work on time-varying default probability forecasts. They are considered with respect to their practical use and potential application in validating default forecasts in credit institutions.

Key words: Basel II, Internal Ratings Based Approach, Validation, Estimation of Default Probabilities, Time Dependency

*Stefan Blochwitz, Dirk Tasche and Carsten S. Wehn work in the Banking and Financial Supervision department at the Deutsche Bundesbank. Stefan Blochwitz is head of group for the on-site inspection of IRB models in the Supervisory Review Process, Dirk Tasche is involved as senior analyst in the current negotiations on Basel II and Carsten Wehn is senior examiner for Internal Market Risk Models. Stefan Hohl is a Senior Economist (Supervision) at the Representative Office of the Bank for International Settlements for Asia and the Pacific in Hong Kong. Nevertheless, all statements made in the present article are the authors' own opinions and should not be cited as being those of the Deutsche Bundesbank or of the Bank for International Settlements.
Corresponding author: Dirk Tasche, E-mail: tasche@ma.tum.de
The authors are grateful to numerous colleagues from regulatory bodies and commercial banks for helpful discussions.

1 Introduction

Due to the ground breaking changes in risk management of credit portfolios, banks are facing more and more complex challenges in determining appropriate default probabilities for assets held in certain portfolios and associated with rating grades. Incentives are currently set towards sophisticated risk measurement methods by the new Basel Capital Accord, see BCBS (2003), that intends regulatory capital requirements to be calculated in a more risk adjusted way.

Thus, risk managers and developers for banks' credit risk estimation models as well as supervisors are confronted with the issue of validating those risk estimates. This is a rather problematic task as the data is by far not that frequently available as in other risk areas. In credit risk, defaults are recorded most commonly only once per year, and hence a comparison between the forecasts and the respective realizations can only be made rarely. Most credit risk models also include several latent variables that determine the overall behavior of the credit portfolio (e.g. asset correlation).

Despite these issues, approaches to the validation have to be made that should be understandable by a bank's practitioners as well as by examiners who are responsible for auditing the appropriateness and adequacy of the estimation and modelling procedures. A recent example of such an approach is given by Balthazar (2004), relying heavily on simulation methods. Tasche (2003) presents a method avoiding simulations but requiring explicit specification of asset correlations.

In the present analysis, we investigate for two validation methods for PD estimates the error that occurs when neglecting correlation with respect to time and asset correlation. We therefore briefly sketch the Extended Traffic Light Approach (ETLA) that is based on a multinomial model and an ad hoc normal test as a well understood alternative to the ETLA. We run numerical simulations for several situations to incorporate the timely dependency of a systematic variable and also the correlation of each obligor with respect to this global variable.

For the notation, we use the following conventions: We denote in the following for the years $t = 1, \ldots, T$ the forecasts for the default probabilities (PDs) by $\hat{p}_t$ and the respective observed default rates as $\tilde{p}_t$. The number of obligors in the respective rating grade is denoted by N_t, $D_{i,t}$ is the indicator of the default event for the i-th obligor at time t and $d_{i,t}$ is the respective realization. Thus the observed default rate reads

$$\tilde{p}_t = \frac{\sum_{i=1}^{N_t} d_{i,t}}{N_t}.$$

Further on, $D_t = \sum_{i=1}^{N_t} D_{i,t}$ is the total number of defaults in the rating grade. Whenever

appropriate, p_t is the "real" (but unobservable) PD. Further notation will be introduced where it occurs first.

2 Extended Traffic Light Approach (ETLA)

In Blochwitz, Hohl & Wehn (2003), a traffic light approach is presented as a tool to select examination samples and to identify suspicious rating grades. This approach is rather a graphical visualization of the observed default rate in relation to the forecasted default probability than a statistical test. The proposal is based on the asymptotic assumption of no correlation in time and on the observation that if the default events in year t are independent and if all the obligors in the portfolio have the same probability of default p_t the number D_t of defaults in year t is binomially distributed with probability parameter p_t and size parameter N_t:

$$D_t \sim \mathcal{B}(N_t, p_t).$$

As a consequence, by the central limit theorem, the distribution of the standardized default rate

$$\bar{p}_t \quad = \quad \frac{D_t - N_t p_t}{\sqrt{N_t p_t (1-p_t)}} = \frac{\tilde{p}_t - p_t}{\sqrt{\frac{p_t(1-p_t)}{N_t}}}$$

can be approximately described by the standard normal distribution as long as $N_t p_t$ is not too small. Blochwitz et al. (2003) analyze in their paper the effect of the incorporation of asset correlation and conclude that in most relevant cases this effect is rather small. They do so by comparing first and second error levels of an independent and high granular portfolio and also by numerical simulations. The present analysis aims to shed further light on the behavior of the proposed ETLA.

An interpretation as a statistical test might be that if default events are assumed to be independent and, additionally, independence in time is taken as given, under the Null hypothesis of correct forecasts a multinomial distribution with well-defined probabilities of the outcomes (identified with the traffic light colours) turns out to be the distribution of the test statistic. For this statistic, probabilities (corresponding to the colors green, yellow, orange, and red) with $\pi_g + \pi_y + \pi_o + \pi_r = 1$ and a color mapping $C(x)$ are defined by

$$C(x) = \begin{cases} g, & x \leq \Phi^{-1}(\pi_g), \\ y, & \Phi^{-1}(\pi_g) < x \leq \Phi^{-1}(\pi_y), \\ o, & \Phi^{-1}(\pi_y) < x \leq \Phi^{-1}(\pi_o), \\ r, & \Phi^{-1}(\pi_o) < x, \end{cases}$$

where Φ^{-1} denotes the inverse function of the standard normal distribution function. In the present paper, the probabilities were chosen as $\pi_g = 0.5$, $\pi_y = 0.3$, $\pi_o = 0.15$ and $\pi_r = 0.05$. With this definition, under the assumption of independence of the annual numbers of default, the vector (L_g, L_y, L_o, L_r) with L_c counting the appearances of colour $c \in \{g, y, o, r\}$ in the sequence $C(\bar{p}_1), \ldots, C(\bar{p}_T)$ will be approximately multinomially distributed with

$$P\left[A = (L_g, L_y, L_o, L_r)\right] = \frac{T!}{L_g! L_y! L_o! L_r!} \pi_g^{L_g} \pi_y^{L_y} \pi_o^{L_o} \pi_r^{L_r},$$

for every quadruple (L_g, L_y, L_o, L_r) of non-negative integers such that $L_g + L_y + L_o + L_r = T$. In order to construct critical regions for tests of the underlying probabilities of defaults, for the case of $T \leq 9$ the order function

$$\Lambda = \Lambda(L_g, L_y, L_o, L_r) = 1000 L_g + 100 L_y + 10 L_o + L_r$$

turned out to be appropriate. With this notation, the traffic lights test of the hypothesis "All true probabilities of default in the years $t = 1, \ldots, T$ are smaller than their corresponding forecasts $\hat{p}_t$" can be specified as follows:

Reject the hypothesis at confidence level β if

$$\Lambda \leq \nu_\beta,$$

where ν_β is calculated as the greatest number ν with the property that $P[\Lambda \leq \nu] < 1 - \beta$. If the critical value ν_β is exceeded by the test statistic, do not reject the hypothesis at level β.

3 Normal Test

The construction of a normal test of PDs is based on the following observation: If $X_1, X_2, X_3 \ldots$ are independent random variables with (not necessarily equal) means $\mu_1, \mu_2, \mu_3, \ldots$ and common variance $\sigma^2 > 0$ then by the central limit theorem the distribution of the standardized sum

$$\frac{\sum_{t=1}^{T} (X_t - \mu_t)}{\sqrt{T}\sigma}$$

will converge to the standard normal distribution for T tending towards ∞. In most cases of practical interest, the rate of convergence is quite high. Therefore, even for small values of T (e.g. $T = 5$) approximating the standardized sum with the standard normal distribution seems reasonable.

In order to apply the normal approximation to the case of PD forecasts $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_T$ and observed percentage default rates $\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_T$ an estimator τ^2 of the assumed common variance σ^2 must be specified. The obvious choice of τ^2 is

$$\tau_0^2 = \frac{1}{T-1} \sum_{t=1}^{T} (\tilde{p}_t - \hat{p}_t)^2.$$

This estimator will be unbiased if the forecast PDs exactly match the true PDs. However, τ_0^2 will be upwardly biased as soon as some of the forecasts differ from the corresponding true PDs. The bias can be considerably reduced by choosing

$$\tau^2 = \frac{1}{T-1} \left(\sum_{t=1}^{T} (\tilde{p}_t - \hat{p}_t)^2 - \frac{1}{T} \left(\sum_{t=1}^{T} (\tilde{p}_t - \hat{p}_t) \right)^2 \right).$$

Under the hypothesis of exact forecasts, τ^2 is unbiased. In case of mismatches, it is also upwardly biased, but to a less extent than τ_0^2. The normal test of the hypothesis "All true probabilities of default in the years $t = 1, \ldots, T$ are smaller than their corresponding forecasts $\hat{p}_t$" goes as follows:

Reject the hypothesis at confidence level β if

$$\frac{\sum_{t=1}^{T} (\tilde{p}_t - \hat{p}_t)}{\sqrt{T}\tau} > z_\beta,$$

where z_β is calculated as the standard normal β-quantile (e.g. $z_{0.99} \approx 2.33$).
If the critical value z_β is not exceeded by the test statistic, accept the hypothesis at level β.

4 Model for the simulation study

Goal of the study is to generate close-to-reality time series of annual default rates and to apply both the normal and the traffic lights test methodologies to them. Closeness to reality in this case means that the rates in different years can be stochastically dependent and that the same holds for the default events within one year. Essentially, the model can be considered as an extension of the Vasicek model which was used in deriving the Basel II risk weight functions into the time dimension (cf. Gordy (2003)).

Assume that a fixed portfolio is being observed in years $t = 1, \ldots, T$. At time t the number of obligors in the portfolio is the a priori known deterministic number N_t. The

change in the general economic conditions from year $t-1$ to year t is expressed by the random variable S_t. Small values of S_t reflect poor economic conditions, large values stand for good conditions. The joint distribution of S is normal with standardized marginal distributions and correlation matrix

$$\Sigma = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1T} \\ r_{21} & 1 & r_{23} & \cdots & r_{2T} \\ \vdots & & \ddots & & \vdots \\ r_{T1} & \cdots & \cdots & r_{TT-1} & 1. \end{pmatrix}.$$

Defining $r_{st} = \vartheta^{|s-t|}$ for some appropriately chosen $\vartheta \in [0,1]$ is common practice in panel analysis and is also the approach which is followed in this simulation study. This is in line with assuming an autoregressive process of first order for the global variable. The unconditional default probability in year t is p_t. Similar to Gordy (2003), we assume that, conditional on S, the numbers of default are independent and binomially distributed with sizes N_t and conditional default probabilities

$$p_t(S) = \Phi\left(\frac{p_t - \sqrt{\rho_t}S_t}{\sqrt{1-\rho_t}}\right).$$

The ρ_t are interpreted as the correlations of the changes in the obligors' asset values from year $t-1$ to year t. The annual percentage default rates $\tilde{p}_t$ will be calculated as $\tilde{p}_t = \frac{D_t}{N_t}$, where D_t denotes the number of defaults in year t.

5 Subject and Results of the Simulation Study

Both the normal test as well as the traffic light test were derived by asymptotic considerations – with regard to the length of the time series of the observed default rates in case of the normal test and with regard to the portfolio size in the case of the traffic light test. As a consequence, even in the case of complete independence in time and in the portfolio it is not clear that the type I errors[1] observed with the tests will be dominated by the nominal error levels. Of course, the compliance with the nominal error level is much more an issue in the case of dependencies of the annual default rates or of the default events in the portfolio. When compliance with the nominal error level for the type I error is confirmed, the question has to be examined which test is the more powerful, i.e. for which test the type II errors[2] are lower.

In order to clarify these points, simulation scenarios for $T = 5$ years of default experience as described in Tables 1 (for type I errors) and 2 (for type II errors) were generated.

[1] I.e. the probabilities of errorneously rejecting the hypothesis.
[2] I.e. the probabilities of not rejecting the hypothesis if specific alternatives are true.

Tables 3 and 4 list the parameter settings that were used for the implementation[3] of the scenarios. For all simulation runs, a constant over time portfolio size of 1,000 obligors was fixed. Finally, Tables 5 and 6 report the observed error rates in the cases of the type I error and the type II error respectively.

With regard to the type I errors, according to Table 5 both test methodologies seem to be essentially in compliance with the nominal error levels. At high error levels (10% and 5%), the normal test fits the levels better than the traffic light test does. For low error levels (2.5% and less), the order of compliance is just reversed with the traffic light test performing better. Both test methodologies face in some scenarios relatively bad performance at the very low levels. Serious outliers are observed at 5% level for the traffic light test as, in the dependence scenarios with larger PDs (DC-LC and DV-LV), type I errors of more than 10% occur.

In general, according to Table 6 the traffic light test appears to be more powerful than the normal test. In case of low PD forecasts to be checked, compared to the case of larger PDs, for both test methodologies power is very low. However, whereas in most scenarios differences in power are not dramatic, the traffic light test sees a heavy collapse of performance in the "independence with varying larger PDs" (I-LV) scenario where at levels 2.5% and 1% the normal test is more than 10% better.

To sum up, both test methodologies seem to be reliable with respect to compliance with the nominal error levels, even in case of dependencies that were not taken into account in their designs. The traffic light test is more powerful than the normal, and should therefore be preferred to the normal test. However, the normal test appears to be slightly more robust than the traffic light test with respect to violations of the assumptions underlying its design. This observation might favor simultaneous applications of the tests.

6 Conclusion

With the Extended Traffic Light Approach (ETLA) by Blochwitz et al. (2003), a flexible tool for monitoring the probability of default (PD) forecasts of rating grades was provided. As the design of the ETLA is based on an assumption of cross-sectional and inter-temporal independence of default events, in the paper at hand we checked its robustness with respect to violations of this assumption. For this purpose, the ETLA was used as a statistical test of the hypothesis of adequate PD forecasts, and its performance was compared to the performance of an ad hoc normal test. A common feature of both these tests is the suitability for being applied to hypotheses of non-constant PD fore-

[3] Every scenario was investigated with 25,000 simulation runs.

Scenario	Description
I-SC	Independence of default events and annual default rates, small and constant unconditional PDs.
I-LC	Independence of default events and annual default rates, larger and constant unconditional PDs.
DC-SC	Time dependence, constant asset correlations, small and constant unconditional PDs.
DC-LC	Time dependence, constant asset correlations, larger and constant unconditional PDs.
I-SV	Independence of default events and annual default rates, small and varying unconditional PDs.
I-LV	Independence of default events and annual default rates, larger and varying unconditional PDs.
DV-SV	Time dependence, varying asset correlations, small and varying unconditional PDs.
DV-LV	Time dependence, varying asset correlations, larger and varying unconditional PDs.

Table 1: Scenarios for type I error simulations.

Scenario	Description
I-SV	Independence of default events and annual default rates, small and varying unconditional PDs.
I-LV	Independence of default events and annual default rates, larger and varying unconditional PDs.
DV-SV	Time dependence, varying asset correlations, small and varying unconditional PDs.
DV-LV	Time dependence, varying asset correlations, larger and varying unconditional PDs.

Table 2: Scenarios for type II error simulations.

Scenario	Correlation in time (ϑ)	Asset correlations	PD forecasts (in %)	True PDs (in %)
I-SC	0	0; 0; 0; 0; 0	0.3; 0.3; 0.3; 0.3; 0.3	0.3; 0.3; 0.3; 0.3; 0.3
I-LC	0	0; 0; 0; 0; 0	3.0; 3.0; 3.0; 3.0; 3.0	3.0; 3.0; 3.0; 3.0; 3.0
DC-SC	0.2	0.05; 0.05; 0.05; 0.05; 0.05	0.3; 0.3; 0.3; 0.3; 0.3	0.3; 0.3; 0.3; 0.3; 0.3
DC-LC	0.2	0.05; 0.05; 0.05; 0.05; 0.05	3.0; 3.0; 3.0; 3.0; 3.0	3.0; 3.0; 3.0; 3.0; 3.0
I-SV	0	0; 0; 0; 0; 0	0.1; 0.2; 0.3; 0.4; 0.6	0.1; 0.2; 0.3; 0.4; 0.6
I-LV	0	0; 0; 0; 0; 0	1.0; 2.0; 3.0; 4.0; 6.0	1.0; 2.0; 3.0; 4.0; 6.0
DV-SV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	0.1; 0.2; 0.3; 0.4; 0.6	0.1; 0.2; 0.3; 0.4; 0.6
DV-LV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	1.0; 2.0; 3.0; 4.0; 6.0	1.0;2.0;3.0;4.0;6.0

Table 3: Parameter settings for type I error simulations.

Scenario	Correlation in time (ϑ)	Asset correlations	PD forecasts (in %)	True PDs (in %)
I-SV	0	0; 0; 0; 0; 0	0.1; 0.2; 0.3; 0.4; 0.6	0.15; 0.25; 0.35; 0.45; 0.65
I-LV	0	0; 0; 0; 0; 0	1.0; 2.0; 3.0; 4.0; 6.0	1.5; 2.5; 3.5; 4.5; 6.5
DV-SV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	0.1; 0.2; 0.3; 0.4; 0.6	0.15; 0.25; 0.35; 0.45; 0.65
DV-LV	0.2	0.05; 0.06; 0.07; 0.08; 0.09	1.0; 2.0; 3.0; 4.0; 6.0	1.5; 2.5; 3.5; 4.5; 6.5

Table 4: Parameter settings for type II error simulations.

Nominal level	0.1	0.05	0.025	0.01	0.005	0.001
I-SC, normal	0.109	0.059	0.045	0.027	0.020	0.014
I-SC, traffic	0.135	0.085	0.043	0.011	0.007	0.001
I-LC, normal	0.130	0.081	0.055	0.037	0.028	0.016
I-LC, traffic	0.104	0.062	0.030	0.013	0.005	0.001
DC-SC, normal	0.092	0.049	0.030	0.017	0.013	0.007
DC-SC, traffic	0.124	0.076	0.029	0.018	0.016	0.008
DC-LC, normal	0.116	0.070	0.044	0.026	0.019	0.010
DC-LC, traffic	0.136	0.113	0.026	0.024	0.023	0.018
I-SV, normal	0.111	0.059	0.043	0.024	0.017	0.012
I-SV, traffic	0.132	0.088	0.043	0.013	0.005	0.001
I-LV, normal	0.128	0.077	0.051	0.032	0.024	0.014
I-LV, traffic	0.096	0.060	0.029	0.012	0.004	0.001
DV-SV, normal	0.083	0.037	0.021	0.010	0.007	0.003
DV-SV, traffic	0.115	0.071	0.027	0.017	0.015	0.007
DV-LV, normal	0.113	0.062	0.036	0.019	0.013	0.005
DV-LV, traffic	0.126	0.108	0.023	0.022	0.022	0.017

Table 5: Type I errors (normal = with normal test, traffic = with traffic light test).

Nominal level	0.1	0.05	0.025	0.01	0.005	0.001
I-SV, normal	0.736	0.836	0.875	0.922	0.944	0.964
I-SV, traffic	0.685	0.782	0.874	0.946	0.972	0.990
I-LV, normal	0.252	0.366	0.467	0.575	0.643	0.754
I-LV, traffic	0.259	0.374	0.600	0.688	0.760	0.871
DV-SV, normal	0.862	0.927	0.956	0.977	0.984	0.992
DV-SV, traffic	0.811	0.868	0.950	0.965	0.969	0.983
DV-LV, normal	0.775	0.858	0.908	0.946	0.961	0.979
DV-LV, traffic	0.733	0.760	0.933	0.935	0.936	0.955

Table 6: Type II errors (normal = with normal test, traffic = with traffic light test).

casts. This situation is likely to occur in practice since many rating systems are hybrids that combine point-in-time and through-the-cycle features.

The comparison of ETLA and normal test was carried out by means of a simulation study. Reliability with respect to type I error levels as well as power measured by type II error sizes were examined. Overall, the performance of ETLA and normal test is broadly equal. However, in general the ETLA appears to be slightly more powerful while the normal test is slightly more robust with respect to correlation of default events and in time. In consideration of the strong conceptual differences, the observation of comparable performance of the tests indicates that further developments in the field of PD validation might not reach much improvement. Nevertheless, this is only a conjecture so that further research for its verification is needed.

The extension of the ETLA to simultaneous monitoring of several rating grades represents another direction for further research since for rating systems with many grades a purely random rejection of appropriate estimation for one or two grades becomes likely. Therefore, the simultaneous ETLA could reflect the adequacy of a rating system as a whole, in contrast to the simultaneous application of ETLA to several rating grades which gives only local pictures of the system (e.g., in order to select samples for a deeper examination).

References

Balthazar, L. (2004), 'PD estimates for Basel II', *RISK Magazine* **17**(4), 84–85.

BCBS (2003), 'The New Basel Capital Accord – Third Consultative Document', http://www.bis.org/bcbs/cp3full.pdf. Basel Committee on Banking Supervision.

Blochwitz, S., Hohl, S. & Wehn, C. S. (2003), Reconsidering Ratings, Working paper, submitted for publication.

Gordy, M. B. (2003), 'A Risk-Factor Model Foundation for Ratings-Based Bank Capital Rules', *Journal of Financial Intermediation* **12**(3), 199–232.

Tasche, D. (2003), A traffic lights approach to PD validation, Working Paper.

Back to 'INSIGHTS