Validierung des In-house Credit Assessment Systems der OeNB

Manuel Mayer,
Christoph Leitner[1]

Das In-house Credit Assessment System (ICAS) der OeNB dient der Bonitätsbewertung nichtfinanzieller Unternehmen in Österreich. Es basiert auf dem Common Credit Assessment System (CoCAS), das von der Deutschen Bundesbank und der OeNB in Zusammenarbeit mit der Wirtschaftsuniversität Wien entwickelt wurde. Seit seiner Einführung im Jahr 2011 haben sich drei weitere Zentralbanken (Banco de España, Banque Nationale de Belgique und Banco de Portugal) dem gemeinsamen Projekt angeschlossen. CoCAS ist eine Plattform für die Bewertung von Unternehmen, deren Kredite von inländischen Kreditinstituten als notenbankfähige Sicherheiten für Refinanzierungs- und Innertageskredite vom Eurosystem verwendet werden können. Somit unterliegt CoCAS neben der Validierung durch die OeNB auch der jährlichen Validierung durch das Eurosystem im Rahmen des Eurosystem Credit Assessment Framework (ECAF) Monitoring and Performance Reports. Dieser Beitrag gibt einen Überblick über die verschiedenen Methoden zur Validierung des In-house Credit Assessment Systems der OeNB, insbesondere das grundlegende theoretische Rahmenwerk und statistische Tests zur Validierung der Kalibrierungsgüte der CoCAS-Ratings durch die OeNB wie auch über die ECAF Performance Monitoring Methodology, die die Basis der Validierung durch das Eurosystem bildet.

Im Eurosystem haben nationale Zentralbanken die Möglichkeit, Kreditinstituten liquiditätszuführende Refinanzierungs- und Innertageskredite zu gewähren. Diese Kreditgeschäfte sind mit ausreichenden notenbankfähigen Sicherheiten zu unterlegen, die hohe Bonitätsanforderungen erfüllen müssen. Die Erfüllung dieser Bonitätsanforderungen kann neben einem Rating durch eine externe Ratingagentur (ECAI) oder einem dafür zugelassenem IRB-Institut bzw. einem Rating Tool auch mit Hilfe des In-house Credit Assessment Systems (ICAS) der jeweiligen Zentralbank überprüft werden, das im Fall der OeNB auf dem Common Credit Assessment System (CoCAS)[2] basiert.

Der Ratingprozess von CoCAS besteht aus zwei Stufen und wird in der OeNB von der Abteilung Statistik – Aufsicht, Modelle und Bonitätsanalysen durchgeführt. In einer ersten Stufe wird anhand eines statistischen Modells ein so genanntes quantitatives Rating erstellt. In einer zweiten Stufe wird es mittels eines OeNB-Expertenmodells von einem Analysten bestätigt oder abgeändert. CoCAS wurde in Zusammenarbeit der OeNB mit der Deutschen Bundesbank und der Wirtschaftsuniversität Wien entwickelt. Mittlerweile haben sich drei weitere Zentralbanken im Eurosystem (Banco de España, Banque Nationale de Belgique, Banco de Portugal) dem gemeinsamen Projekt angeschlossen.

In Übereinstimmung mit den regulatorischen Rahmenbedingungen im Eurosystem und mit der Standardliteratur im Kreditrisikobereich wird der Ausfall von Unternehmen in CoCAS als stochastisches Ereignis abgebildet, dessen Eintrittswahrscheinlichkeit nicht direkt beobachtbar ist, sondern geschätzt werden muss. Die Überprüfung der Genauigkeit der von CoCAS geschätzten Ausfallwahrscheinlichkeiten erfolgt einerseits über die regelmäßige interne Kontrolle durch die OeNB und andererseits über die jährliche Validierung durch das Eurosystem im Rah-

[1] *Oesterreichische Nationalbank, Abteilung Statistik – Aufsicht, Modelle und Bonitätsanalysen, manuel.mayer@oenb.at; christoph.leitner@oenb.at.*

[2] *Siehe Deutsche Bundesbank (2015) und Oesterreichische Nationalbank (2015).*

men des Eurosystem Credit Assessment Framework (ECAF) Monitoring and Performance Report (siehe EU, 2015). Dieser Beitrag gibt einen Überblick über die verschiedenen statistischen Verfahren, die zur Validierung der CoCAS-Ratings verwendet werden.

1 Theoretisches Rahmenwerk für die Validierung der CoCAS-Ratings durch die OeNB

Im Folgenden wird das grundlegende theoretische Rahmenwerk erläutert, das der Validierung der auf Basis von CoCAS erstellten Ratings durch die OeNB zugrunde liegt. Im Einklang mit der Standardliteratur zur Validierung von Kreditrisikosystemen wird ein Ausfall als stochastisches Ereignis modelliert, dargestellt durch eine Bernoulli-Zufallsvariable $X_{i,j}$, die den Wert 1 annimmt, wenn Schuldner i in Ratingklasse j in einer bestimmten zukünftigen Periode (typischerweise 1 Jahr) ausfällt und den Wert 0 annimmt, falls dies nicht geschieht. Die latente Ausfallwahrscheinlichkeit eines Schuldners in Ratingklasse j wird mit $p_{i,j}=P(X_{i,j}=1)$ bezeichnet. Unter der Annahme, dass die einzelnen Ausfallereignisse voneinander unabhängig sind und dass Schuldner innerhalb einer Ratingklasse dieselbe Ausfallwahrscheinlichkeit aufweisen $(p_{i,j}=p_j)$, folgt die Anzahl von Ausfällen D_j innerhalb einer Ratingklasse einer Binomialverteilung:

$$D_j = \sum_{i=1}^{n_j} X_{i,j} \sim Bin(n_j, p_j),$$

wobei n_j die Anzahl von Schuldnern innerhalb einer Ratingklasse bezeichnet. Eine Ratingklasse wird als richtig kalibriert angesehen, wenn das Kreditrisiko innerhalb der Ratingklasse weder über- noch unterschätzt wird, also die geschätzten Ausfallwahrscheinlichkeiten den wahren Ausfallwahrscheinlichkeiten entsprechen. Ein gesamtes Ratingsystem, bestehend aus $j=1,\ldots,K$ Ratingklassen, wird als richtig kalibriert bewertet, wenn jede einzelne Ratingklasse richtig kalibriert ist. Der Test eines Ratingsystems kann somit anhand folgender Null- und Alternativhypothese formuliert werden:

$$\boldsymbol{H}_0: \quad \forall j \in K: \quad p_j = pd_j,$$

$$\boldsymbol{H}_1: \quad \exists j \in K: \quad p_j \neq pd_j,$$

wobei pd_j die vom zu validierenden Ratingsystem geschätzte Ausfallwahrscheinlichkeit für alle Schuldner in Ratingklasse j bezeichnet. Statistische Tests für das oben genannte Testproblem können grundsätzlich in Tests unterteilt werden, bei denen alle Ratingklassen einzeln getestet werden (einfache Tests) und jene, bei denen die Ratingklassen des Ratingsystems gemeinsam getestet werden (gemeinsame Tests).[3]

2 Testverfahren der Validierung der CoCAS-Ratings durch die OeNB

Zur Gruppe der einfachen Tests gehört der Sterne-Test, der in Aussenegg et al. (2011) beschrieben wird und auf eine Ratingklasse j angewendet werden kann. Der p-Wert des Tests ist als die Wahrscheinlichkeit zu interpretieren, mit der in dieser Ratingklasse d_j Ausfälle oder Ausfälle, die noch mehr gegen die Nullhypothese sprechen, eintreten. Dabei sprechen all die Ausfälle noch mehr gegen die Nullhypothese, die eine gleich hohe oder geringere Eintrittswahrscheinlichkeit als d_j haben.

Die Nullhypothese, dass das gesamte Ratingsystem richtig kalibriert

[3] *Die im Folgenden diskutierten Tests sind zweiseitige Tests. Zu einseitigen gemeinsamen Tests siehe auch Coppens et al. (2016).*

ist, wird verworfen, wenn die oben genannte Nullhypothese für mindestens eine Ratingklasse verworfen wird. Bei einem solchen Testproblem (multiples Testproblem) tritt das Phänomen der so genannten Alphafehler-Kumulierung auf. Dies bezeichnet den Umstand, dass der Alpha-Fehler[4] für den Test der Nullhypothese, dass das gesamte Ratingsystem richtig kalibriert ist, größer ist als der Alpha-Fehler für die Tests der einzelnen Ratingklassen. Um den Alpha-Fehler für den Test des gesamten Ratingsystems zu beschränken, steht in der Literatur eine Reihe von Methoden zur Verfügung. Zur Validierung der CoCAS-Ratings kommt die min-P-Methode zur Anwendung, die von Westfall und Wolfinger (1997) vorgeschlagen wurde. Westfall und Wolfinger (1997) argumentieren, dass damit für diskrete Verteilungen die Teststärke gegenüber der klassischen Bonferroni-Methode verbessert werden kann (der Beta-Fehler[5] des statistischen Tests reduziert werden kann).

Neben Tests, bei denen jede Ratingklasse einzeln getestet wird, stehen auch Verfahren zur Verfügung, mit denen alle Ratingklassen gemeinsam getestet werden. Zu Letzteren zählen beispielsweise der Hosmer-Lemeshow-Test[6], der Spiegelhalter-Test[7] sowie der multivariate Sterne-Test.[8] Die Teststatistik des Hosmer-Lemeshow-Tests ist die Summe der quadrierten Abweichungen zwischen der Anzahl der prognostizierten Ausfälle und der Anzahl der beobachteten Ausfälle in allen Ratingklassen, gewichtet mit der Inversen der theoretischen Varianz der Anzahl von Ausfällen:

$$HSLS = \sum_{j=1}^{K} \frac{(n_j pd_j - d_j)^2}{n_j pd_j (1 - pd_j)}.$$

Unter der Nullhypothese, dass die geschätzten Ausfallwahrscheinlichkeiten den tatsächlichen, latenten Ausfallwahrscheinlichkeiten entsprechen, und unter der Annahme, dass der Stichprobenumfang groß genug ist, um die Normalverteilungsannahme zu treffen, ist die Teststatistik $HSLS_{\chi^2}$-verteilt mit K Freiheitsgraden.[9]

Der Spiegelhalter-Test basiert auf dem „Mean Square Error“ (MSE) zwischen dem Ausfallindikator y_i und der geschätzten Ausfallwahrscheinlichkeit pd_i (auch als Brier Score[10] bekannt), $MSE = \frac{1}{N}\Sigma_{i=1}^{N}(y_i - pd_i)^2$, wobei N die Gesamtzahl der Schuldner (in allen Ratingklassen) bezeichnet. Unter der Nullhypothese, dass die geschätzten Ausfallwahrscheinlichkeiten mit den latenten Ausfallwahrscheinlichkeiten übereinstimmen, ergeben sich der Erwartungswert und die Varianz von MSE als: $E[MSE] = \frac{1}{N}\Sigma_{i=1}^{N} pd_i(1 - pd_i)$ und $Var[MSE] = \frac{1}{N^2}\Sigma_{i=1}^{N} pd_i(1 - pd_i)(1 - 2pd_i)^2$.

Die Teststatistik des Spiegelhalter-Tests folgt unter der Annahme eines

[4] *Auch Fehler 1. Art genannt. Ein Fehler 1. Art liegt vor, wenn die Nullhypothese zurückgewiesen wird, obwohl sie in Wirklichkeit zutrifft.*

[5] *Auch Fehler 2. Art genannt. Ein Fehler 2. Art liegt vor, wenn die Nullhypothese fälschlicherweise bestätigt wird, obwohl die Alternativhypothese korrekt ist.*

[6] *Siehe Hosmer und Lemeshow (1980).*

[7] *Siehe Spiegelhalter (1986).*

[8] *Siehe Aussenegg et al. (2011).*

[9] *Für den Fall, dass für die Validierung ein In-Sample-Datensatz verwendet wird, der auch für die Schätzung des Modells verwendet wurde, sind K-2 Freiheitsgrade zu verwenden.*

[10] *Siehe Brier (1950).*

hinreichend großen Stichprobenumfangs näherungsweise einer Standardnormalverteilung:

$$Z = \frac{MSE - E[MSE]}{\sqrt{Var[MSE]}}.$$

Der Hosmer-Lemeshow-Test wie auch der Spiegelhalter-Test basieren auf der Annahme eines hinreichend großen Stichprobenumfangs. Aussenegg et al. (2011) argumentieren, dass diese Annahme in vielen Fällen zu verzerrten Testergebnissen führt und schlagen als Alternative den multivariaten Sterne-Test vor, dem keine Normalverteilungsannahme zugrunde liegt. Die Methodik des multivariaten Sterne-Tests wird im Detail in Aussenegg et al. (2011) beschrieben, wobei der p-Wert des multivariaten Sterne-Tests analog zum beschriebenen einfachen Sterne-Test zu interpretieren ist.

Tabelle 1 zeigt die Testergebnisse des Hosmer-Lemeshow-Tests, des Spiegelhalter-Tests, des multivariaten Sterne-Tests sowie des einfachen Sterne-Tests für das statistische Modell von CoCAS zur Bewertung von Firmen, die nach IFRS bilanzieren.[11] Die p-Werte zeigen, dass bei einem üblichen Signifikanzniveau von 5% für keinen der vier Tests die Nullhypothese der korrekten Kalibrierung des Ratingsystems verworfen werden kann.

Tabelle 1

Validierungstests für CoCAS

	p-Wert
Homser-Lemeshow-Test	0,6948
Spiegelhalter-Test	0,2392
Multivariater Sterne-Test	0,7406
Einfacher Sterne-Test (min-P)	0,7395

Quelle: OeNB.

Neben der Kalibrierungsgüte wird von der OeNB zur Validierung der CoCAS-Ratings auch die so genannte Trennschärfe („discriminatory power“)[12] zwischen Schuldnern, die im Beobachtungszeitraum ausfallen, und jenen, bei denen dies nicht geschieht, gemessen. Ein häufig verwendetes Maß für die Trennschärfe ist der AUROC („Area Under the Receiver Operating Characteristic Curve“), der ein ordinales Maß für die Qualität eines Ratingsystems darstellt. Der AUROC[13] und der zugehörige – durch eine lineare Transformation des AUROCs ableitbare – „Accuracy Ratio“[14] (AR) zeigen für das statistische Modell von CoCAS (zur Bewertung von Unternehmen, die nach IFRS bilanzieren) eine sehr gute Trennschärfe an: 0,9118 bzw. 0,8237.

3 Validierung von ICAS durch das Eurosystem

Da das Hauptanwendungsgebiet des ICAS die Bewertung von Unternehmen ist, deren Kredite von Kreditinstituten als Sicherheiten für Refinanzierungs- und Innertageskredite vom Eurosystem verwendet werden können, unterliegt das ICAS auch der Validierung durch das Eurosystem im Rahmen des Eurosystem Credit Assessment Framework (ECAF) Monitoring and Performance Report. Dies geschieht auf Basis der so genannten ECAF Performance Monitoring Methodology.[15] Jedes Jahr wird für jede relevante Ratingklasse auf der Ra-

[11] *Die Ergebnisse beziehen sich auf die Beobachtungsperiode Jänner 2010 bis Juni 2015.*

[12] *Siehe z. B. Basel Committee on Banking Supervision (2005) und Lingo und Winkler (2008).*

[13] *Der Wertebereich von AUROC geht von 0 bis 1, wobei 0,5 für eine zufällige und 1 für eine perfekte Ratingklasseneinteilung der Ausfälle steht.*

[14] *Der Wertebereich von AR geht von -1 bis 1, wobei 0 für eine zufällige und 1 für eine perfekte Ratingklasseneinteilung der Ausfälle steht.*

tingskala des Eurosystems die Anzahl der Ausfälle gezählt und die Ausfallrate berechnet. Anschließend werden auf Basis statistischer Tests für jede Ratingklasse kritische Werte berechnet, mit denen die beobachteten Ausfallraten verglichen werden. Diese Methode wird für einen einjährigen wie auch für einen mehrjährigen Beobachtungszeitraum angewandt. Im Fall eines Überschreitens der kritischen Werte werden weitere Maßnahmen zur Überprüfung der Kalibrierungsgüte des Kreditrisikosystems und gegebenenfalls Korrekturmaßnahmen getroffen.

Literaturverzeichnis

Aussenegg, W., F. Resch und G. Winkler. 2011. Pitfalls and Remedies in Testing the Calibration Quality of Rating Systems. Journal of Banking and Finance 35(3). 698–708.

Basel Committee on Banking Supervision. 2005. Studies on the Validation of Internal Rating Systems. Working Paper 14 (http://www.bis.org/publ/bcbs_wp14.pdf).

Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review 78(1). 1–3.

Coppens F., F. Gonzales und G. Winkler. 2007. The Performance of Credit Rating Systems in the Assessment of Collateral used in Eurosystem Monetary Policy Operations. ECB Occasional Paper Series, 65.

Coppens F., M. Mayer, L. Millischer, F. Resch, S. Sauer und K. Schulze. 2016. Advances in Multivariate Back-testing for Credit Risk Underestimation. ECB Working Paper Series, 1885.

Döhler, S. 2010. Validation of Credit Default Probabilities using Multiple-Testing Procedures. The Journal of Risk Model Validation 4(4). 59–92.

EU. 2015. Guideline (EU) 2015/510 of the ECB of 19 December 2014 on the Implementation of the Eurosystem Monetary Policy Framework (ECB/2014/60), OJ L 91, 2.4.2015. 3–135.

European Central Bank. 2011. The Implementation of Monetary Policy in the Euro Area. General Documentation on Eurosystem Monetary Policy Instruments and Procedures.

Hosmer, D. W. und S. Lemeshow. 1980. A Goodness-of-fit Test for the Multiple Logistic Regression Model. Communication in Statistics – Theory and Methods 10. 1043–1069.

Lehmann, E. L. und J. P. Romano. 2006. Testing Statistical Hypotheses. Springer.

Lingo, M. und G. Winkler. 2008. Discriminatory Power: An Obsolete Validation Criterion? The Journal of Risk Model Validation 2(1). 1–27.

Spiegelhalter, D. J. 1986. Probabilistic Prediction in Patient Management and Clinical Trails. Statistics in Medicine 5(5). 421–433.

Westfall, P. H. und R. D. Wolfinger. 1997. Multiple Tests with Discrete Distributions. The American Statistician 51(1). 3–8.

Oesterreichische Nationalbank. 2015. Common Credit Assessment System zur Bonitätsbeurteilung von nichtfinanziellen Unternehmen – das statistische Ratingmodell. Statistiken – Daten & Analysen 4Q/15.

Deutsche Bundesbank. 2015. Das Common Credit Assessment System zur Prüfung der Notenbankfähigkeit von Wirtschaftsunternehmen. Monatsbericht Januar 2015.

[15] *Siehe European Central Bank (2011).*