Paper SAS3014-2019

Generating Financial Synthetic Instruments for Stress Testing: An Application of Machine Learning and Statistics in SAS® Viya®

Christian Macaro and Rocco Cannizzaro, SAS Institute Inc.

ABSTRACT

In stress testing, an analysis of financial risk exposure needs to be carried out with respect to two key components: the current portfolio positions and the future portfolio positions. The risk analysis of current portfolio positions, properly projected in the future, is a well-known topic of discussion. The risk analysis of future portfolio positions, (that is, financial positions that might originate in the future) is less studied and it poses analytical challenges. In this paper, a random forest regression is used to predict the overall future portfolio volume. Then, conditionally on the future volume, a Copula model is used to generate synthetic positions. Finally, a K-Nearest Neighbor classification method is used to fill-in non-numerical attributes of the newly generated positions.

INTRODUCTION

Analyzing financial risk exposure is fundamental for financial institutions. On one hand, these analyses are often required by institutional regulators to adhere to financial best practices. On the other hand, being able to assess whether they can operate under any circumstances should be implemented by every financial entity that aims at a successfully thriving during good times and surviving during adverse circumstances.

In this context, some institutional regulators (the International Accounting Standards Board, IASB, and the Financial Accounting Standards Board in the USA (FASB)) have made huge progress by determining that any methodology that assesses risk exposure should not focus on incurred losses, but on expected losses. This concept is crucial as it requires financial institutions to perform a detailed analysis of their financial position conditionally on alternative scenarios.

The objective of a stress testing analysis is to measure the resilience of a financial institution to hypothetical adverse scenarios, including economic downturns or prolonged recessions. The analysis must be forward looking (typically 3 to 5 years) with respect to the expected evolution of the business over the forecast period.

A business evolution plan includes assumptions about the growth of certain segments of the portfolio (that is, by product, line of business, geography, and so on). Such assumptions result in the requirement to simulate the creation of future financial positions. This requirement becomes particularly interesting when the amount of historical information available is large and data are possibly distributed.

In this paper, we provide a general guideline on how to leverage SAS Viya analytical capabilities to do the following:

1. Help with the definition of the business evolution plan.
2. Generate synthetic positions that match the assumptions of the business evolution plan.

In doing so, we will take advantage of the possibility of storing the data and performing the computation in the cloud. The availability of cloud storage and cloud computing is not required. A similar analysis can be performed using SAS procedures that pre-date cloud-based technology.

STRESS TESTING: AN EXAMPLE

Stress testing for financial institutions can be characterized by the process of assessing the risk exposure conditionally on the following two aspects:

1. Future projections of the risk factors.
2. Future assumptions about the business evolution plan.

The latter point is obviously related the first one, and it typically involves a combination of qualitative and quantitative analysis of the company's future strategy.

For simplicity, we consider a retail bank whose core business consists of mortgages and auto loans only. The portfolio data are a representation of a typical portfolio from the retail banking industry. As previously mentioned, assumptions about the future business evolution plan are typically the result of qualitative and quantitative assessment of external factors. In this case, we assume that the risk factors affecting the bank are exclusively represented by the following three macroeconomic variables:

1. Real Gross Domestic Product change rate (RGDP)
2. Unemployment change rate (UR)
3. House price index (HPI)

Figure 1 shows that the relationship between the total exposure of the portfolio and HPI is strong. The relationship between RGDP and UR with the total exposure is not obvious, although intuition based on economic theory suggests it.

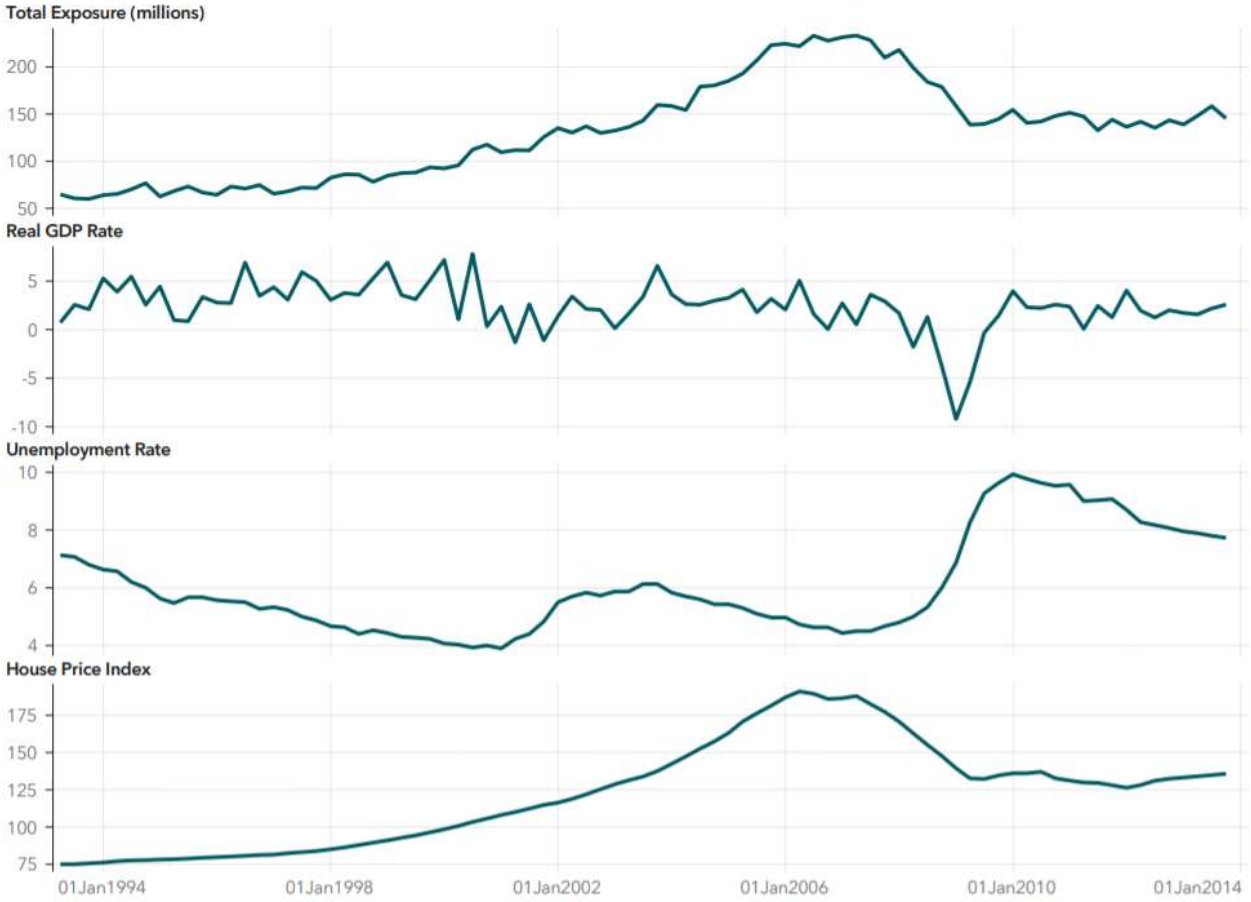

Figure 1. Graphs of Aggregated Portfolio, Real GDP Rate, Unemployment Rate, and House Price Index Over Time

Figure 2 confirms the strong linear dependence between total exposure of the portfolio and HPI. Figure 2 also emphasizes non-linear relationships between UR, HPI, and total exposure. At this point, assumptions about the business evolution plan should be made based on both analytical results as well as the company's business strategy.

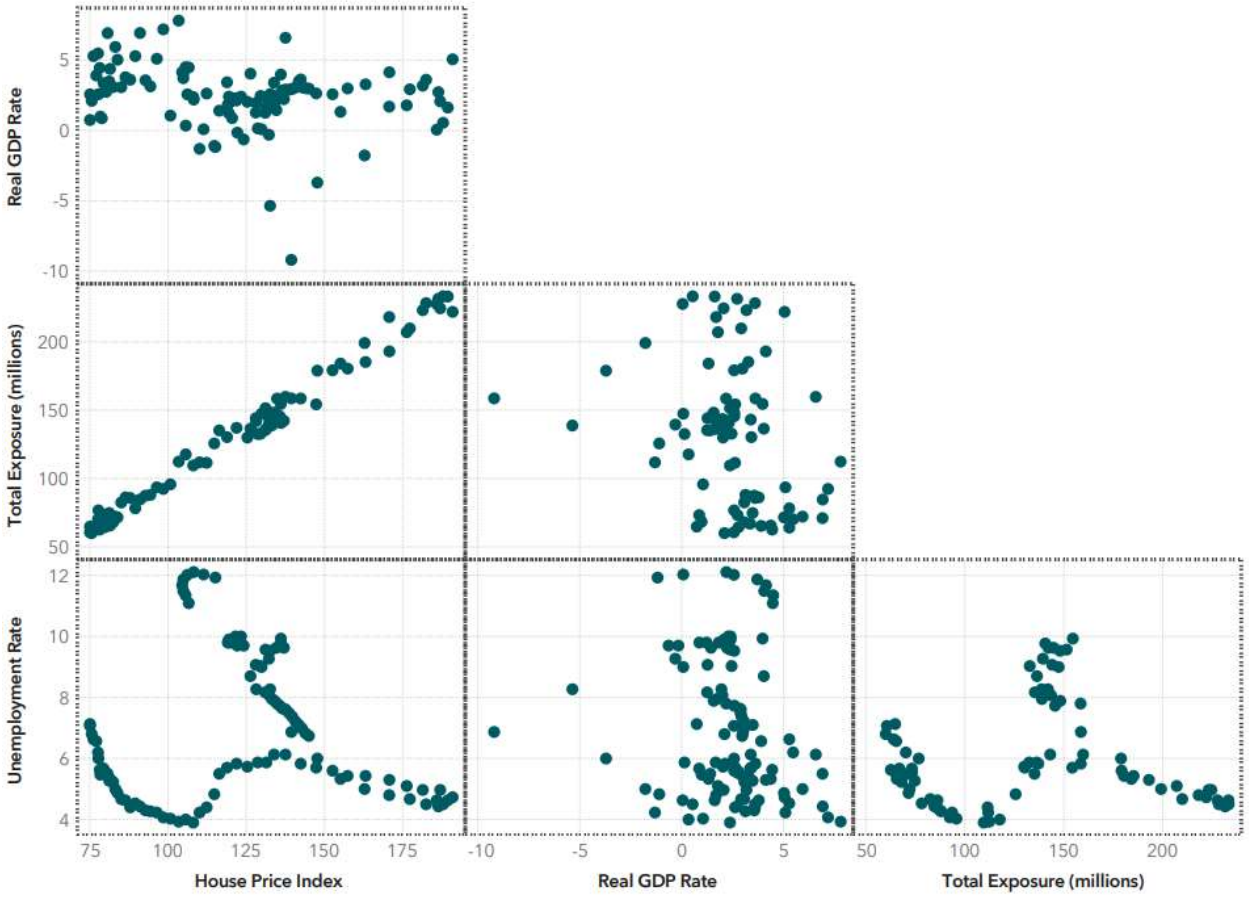

Figure 2. Scatterplot Matrix of Real GDP Rate, Total Exposure, Unemployment Rate, and House Price Index

SCENARIO FORECASTING: PROC FOREST

There are a lot of analytical models that can be used to assist a company in determining their own business evolution plan. The simplest and most common tool is the linear regression model. More sophisticated models derived from macroeconomic theory consist of vector autoregressive models (Henry and Kok 2013). Other models focus on the uncertainty of selecting the relevant risk factors by considering a Bayesian probabilistic model of all possible combinations (Macaro and Sanford 2015). Recent advances in machine learning theory suggest a new group of models based on the random forest theory. A random forest model is essentially a random collection of tree models where the behavior of the variable of interest is explained with a series of simple classification rules. The advantage of using these models is that their methodology has been conceived by keeping in mind the following two aspects:

1. Out-of-sample predicting capability
2. Nonlinear relationship between the main variable of interests and a collection, possibly large, of external factors

Both features of these models seem to be suitable for determining a business evolution plan for a company.

In practice, a random forest model can be used to forecast the total exposure of a portfolio, given the projections of the macroeconomic variables (RGDP, UR, and HPI). The FOREST procedure that is available in SAS® Visual Data Mining and Machine Learning can be used for this purpose.

```
proc forest
   data     = sascas1.data
   outmodel = sascas1.fit
   seed     = 12345
   ;

   target  TotalExposure      / level = interval;
   input   RGDP UR HPI        / level = interval;
   id      Date RGDP UR HPI Scenario;
   output
      out = sascas1.Forecasts
   ;
run;
```

Figure 3 confirms that a major driver in the determination of the portfolio exposure is the House Price Index.

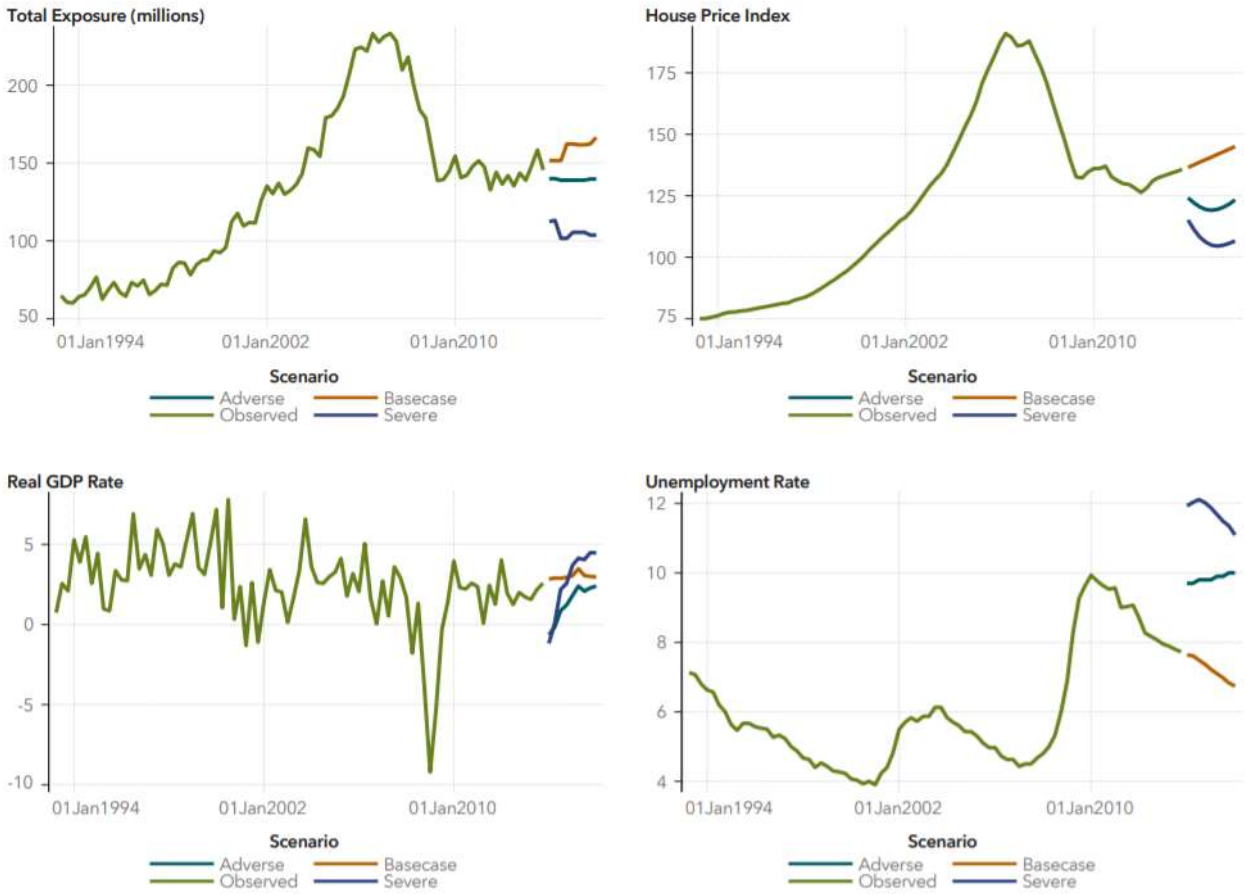

Figure 3. Projection Graphs of Aggregated Portfolio, Real GDP Rate, Unemployment Rate, and House Price Index Over Time

SYNTHETIC POSITIONS GENERATION: PROC CCOPULA

After the conditional forecasts of the aggregated portfolio values have been obtained for each scenario, the synthetic positions need to be generated. A simple and straightforward approach is to resample existing portfolio positions until the targets defined in the business evolution have been reached. The problem with this approach is that no new position is used. Samples of the existing portfolio might determine a portfolio exposure that matches the current risk exposure. To introduce true randomness in the generation of new portfolio position, a more sophisticated approach is proposed.

A COPULA model seems to be best for generating synthetic positions. These models are characterized by the capability of modeling the dependence across many variables by considering their marginal empirical cumulative probability distributions and a merging function called the “COPULA” function. Here is an example of the CCOPULA procedure that is available in SAS® Econometrics:

```
/* Number of instruments to generate */
%let Ninst = 10000;

proc ccopula data = sascas1.portfolio;
   /* Model the dependency of the portfolio variables */
   var
      TotalExposure  /* Financial Exposure */
      PD             /* Probability of Default */
      LGD            /* Loss Given Default */
      LTV            /* Loan-To-Value */
   ;
   /* Fit Normal-copula */
   fit Normal /
               marginals = empirical
               method    = mle
               store     = sascas1.fit
   ;
   /* Simulate new instruments */
   simulate /
               ndraws     = &Ninst.
               outuniform = sascas1.uniform_sim
               seed = 12345
   ;
run;
```

Notice that in the CCOPULA procedure code example, three additional variables are considered:

1. Probability of default (PD)
2. Loss-Given-Default (LGD)
3. Loan-to-Value (LTV)

These additional variables, together with the portfolio exposure, are the variables of interest. These are the variables for which the multivariate dependence should be maintained.

Table 1 shows the list of synthetic instruments that are generated by the CCOPULA procedure. In addition to the variables of interests (PD, LGD, LTV, and Exposure), the table shows three additional variables: Scenario, Horizon, and Short_Flag. These additional variables are generated in a data step and map the newly generated synthetic positions to the assumptions provided in the business evolution plan. The Scenario variable indicates the scenario under which the portfolio synthetic instrument becomes active. The Horizon variable indicates the horizon of activation of the portfolio synthetic instrument. Finally, the Short_Flag variable indicates whether the synthetic instrument will be added or deleted to or from the portfolio. Notice that the Short_Flag variable is always set to FALSE when the

business evolution plan assumes growth. It is worth emphasizing that the aggregate exposure of the synthetic instruments needs to match the assumptions made in the business evolution plan regarding the total exposure.

Instid	Scenario	Horizon	Short	Exposure	LGD	LTV	PD
synth_inst_0000000021	Severe	5	FALSE	$5,503,497	67%	15%	6%
synth_inst_0000000020	Severe	5	FALSE	$249,790	67%	79%	5%
synth_inst_0000000019	Severe	5	FALSE	$2,861,376	61%	71%	8%
synth_inst_0000000018	Severe	2	FALSE	$2,499,328	22%	37%	5%
synth_inst_0000000016	Severe	2	FALSE	$331,030	12%	68%	4%
synth_inst_0000000015	Basecase	9	FALSE	$8,335,864	20%	32%	1%
synth_inst_0000000014	Basecase	9	FALSE	$67,839	41%	0%	8%
synth_inst_0000000013	Basecase	8	FALSE	$8,694,373	41%	75%	1%
synth_inst_0000000012	Basecase	4	FALSE	$3,701,503	20%	67%	38%
synth_inst_0000000011	Basecase	4	FALSE	$74,217	10%	18%	3%
synth_inst_0000000010	Basecase	4	FALSE	$74,217	10%	20%	13%
synth_inst_0000000009	Basecase	4	FALSE	$3,980,549	43%	68%	1%
synth_inst_0000000008	Basecase	4	FALSE	$73,921	23%	63%	6%
synth_inst_0000000007	Basecase	4	FALSE	$5,503,497	18%	20%	6%
synth_inst_0000000006	Basecase	1	FALSE	$8,335,864	40%	32%	2%
synth_inst_0000000005	Basecase	1	FALSE	$1,733,701	12%	83%	8%
synth_inst_0000000004	Basecase	1	FALSE	$3,980,549	10%	53%	7%
synth_inst_0000000003	Adverse	8	FALSE	$7,534,308	44%	34%	32%
synth_inst_0000000002	Adverse	8	FALSE	$74,234	14%	68%	2%
synth_inst_0000000001	Adverse	6	FALSE	$330,575	30%	58%	9%

Table 1. Copula Simulation of Synthetic Instruments

SYNTHETIC POSITIONS GENERATION FOR CLASSES: PROC FASTKNN

One problem with generating a synthetic position with a COPULA model is that you can only capture the multivariate dependence across numeric continuous variables. Two possible solutions to this problem are the following:

1. Random sampling of the class variables from similar pre-existing portfolio positions.
2. Model the relationship between the variables of the COPULA model and the class variables of interest.

The first solution is computationally fast and efficient; however, it does not guarantee consistency of the results. For example, the COPULA model might generate a synthetic instrument with a probability of default close to one (PD≈1), while the random sampling component could assign a AAA rating grade.

The second solution requires additional complex computing, but it can prevent the inconsistencies affecting the simple random sampling method. For example, the K-Nearest Neighbors (KNN) machine learning algorithm can be used to properly match the rating grade given the probability of default. The FASTKNN procedure that is available in SAS Visual Data Mining and Machine Learning can be used for this purpose.

SYNTHETIC POSITIONS ELIMINATION

A closer look at the random forest analysis shows that not all horizons require an increment of the portfolio to be covered by new synthetic positions. In fact, at some horizons (especially in the adverse and severe scenarios), the total exposure of the portfolio shrinks. This can be accomplished with a random elimination without replacement of the original portfolio positions.

Table 2 shows the required random elimination of portfolio instruments as indicated by the business evolution plan. In this case, the Short_Flag variable is set to TRUE. This indicates that the position is available only up to the designated horizon.

Instid	Scenario	Horizon	Short	Exposure	PD	LTV	LGD
Inst_0000000952	Adverse	1	TRUE	$331,030	5%	11%	12%
Inst_0000000934	Adverse	1	TRUE	$72,126	2%	67%	1%
Inst_0000000925	Severe	1	TRUE	$360,358	2%	157%	55%
Inst_0000000925	Adverse	1	TRUE	$180,179	1%	79%	28%
Inst_0000000898	Severe	1	TRUE	$7,125,049	6%	0%	26%
Inst_0000000886	Severe	1	TRUE	$4,960,407	5%	63%	18%
Inst_0000000886	Adverse	1	TRUE	$4,960,407	5%	63%	18%
Inst_0000000853	Adverse	1	TRUE	$44,055	10%	10%	15%
Inst_0000000853	Basecase	6	TRUE	$44,055	10%	10%	15%
Inst_0000000841	Severe	1	TRUE	$1,971,221	4%	21%	35%
Inst_0000000826	Severe	1	TRUE	$279,482	5%	68%	5%
Inst_0000000769	Severe	1	TRUE	$7,390,223	8%	9%	6%
Inst_0000000745	Severe	1	TRUE	$2,782,037	2%	93%	41%
Inst_0000000745	Adverse	3	TRUE	$2,782,037	2%	93%	41%
Inst_0000000643	Adverse	1	TRUE	$123,004	1%	52%	17%
Inst_0000000628	Severe	3	TRUE	$367,279	1%	52%	36%
Inst_0000000628	Severe	1	TRUE	$367,279	1%	52%	36%
Inst_0000000598	Severe	1	TRUE	$97,844	15%	57%	162%
Inst_0000000562	Severe	1	TRUE	$73,921	1%	78%	9%
Inst_0000000550	Basecase	6	TRUE	$2,980,000	25%	57%	15%
Inst_0000000505	Severe	3	TRUE	$8,335,864	6%	26%	53%
Inst_0000000487	Severe	1	TRUE	$5,503,497	1%	3%	12%
Inst_0000000418	Adverse	1	TRUE	$48,290	32%	11%	11%
Inst_0000000403	Adverse	3	TRUE	$70,486	5%	41%	31%
Inst_0000000358	Severe	1	TRUE	$79,825	1%	60%	4%
Inst_0000000325	Severe	1	TRUE	$3,701,503	6%	1%	33%
Inst_0000000283	Severe	8	TRUE	$4,867,433	8%	67%	45%
Inst_0000000283	Severe	3	TRUE	$4,867,433	8%	67%	45%

Table 2. Random Elimination of Instruments

CONCLUSION

Stress testing analysis is fundamental for financial entities operating in an ever-changing environment. The emphasis that institutional regulators have posed on assessing expected loss conditionally on the scenarios has inevitably pushed the financial entities to define clear business evolution plans conditionally on alternative scenarios. In turns, this has created a demand for a methodology that can produce synthetic positions.

REFERENCES

Henry, Jérôme and Christoffer Kok, eds. 2013. *A Macro Stress Testing Framework for Assessing Systemic Risks in the Banking Sector*. Occasional Paper Series No. 152. Available https://www.ecb.europa.eu/pub/pdf/scpops/ecbocp152.pdf.

Macaro, Christian and Kenneth Sanford. 2015. "Incorporating External Economic Scenarios into Your CCAR Stress Testing Routines." *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1756-2015.pdf.

SAS Institute Inc. 2018. "The FOREST Procedure." In *SAS® Visual Data Mining and Machine Learning 8.3: Procedures*. Cary, NC: SAS Institute Inc. Available http://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=casml&docsetTarget=casml_forest_toc.htm&locale=en (accessed January 3, 2019).

SAS Institute Inc. 2018. "The FASTKNN Procedure." In *SAS® Visual Data Mining and Machine Learning 8.3: Procedures*. Cary, NC: SAS Institute Inc. Available

http://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=casml&docsetTarget=casml_fastknn_toc.htm&locale=en (accessed January 3, 2019).

SAS Institute Inc. 2018. “The CCOPULA Procedure.” *In SAS® Econometrics 8.3: Econometrics Procedures*. Cary, NC: SAS Institute Inc. Available http://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=casecon&docsetTarget=casecon_ccopula_toc.htm&locale=en (accessed January 3, 2019).

SAS Institute Inc. 2018. SAS® Viya 3.4. Cary, NC: SAS Institute Inc. Available http://support.sas.com/documentation/onlinedoc/viya/index.html.

ACKNOWLEDGMENT

The authors are grateful to Linda Roberts, a Senior Technical Writer in the Publications Division at SAS Institute, and to Brad Kellam, a Principal Technical Editor in the Publications Division at SAS Institute, for their valuable assistance in the preparation of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Christian Macaro
SAS Institute Inc.
Christian.Macaro@sas.com

Rocco Cannizzaro
SAS Institute Inc.
Rocco.Cannizzaro@sas.com