Dr. Richard Cross
Assistant Director, Data Services

Richard is an assistant director responsible for numerous analytical productivity and data quality initiatives. He designs, implements, and operates systems that apply lean manufacturing principles to data production. Prior to Moody's Analytics, he was a consultant with McKinsey. He has a PhD and an MS in aerospace engineering from Georgia Tech, and an SB in aeronautics and astronautics from MIT.

Dr. Cristian deRitis
Senior Director, Consumer Credit Analytics

Cristian is a senior director who develops credit models for a variety of asset classes. His regular analysis and commentary on consumer credit, housing, mortgage markets, securitization, and financial regulatory reform appear on Economy.com and in publications such as *The Wall Street Journal* and *The New York Times*. Cristian has a PhD in economics from Johns Hopkins University and is named on two US patents for credit modeling techniques.

View all articles of this edition at
MoodysAnalytics.com/RiskPerspectives

When Good Data Happen to Good People: Boosting Productivity with High-Quality Data

BY DR. RICHARD CROSS AND DR. CRISTIAN DERITIS

With ever-increasing requirements for a higher quantity and quality of analytical output, the need to boost productivity in risk management has become more acute. In pursuing these productivity gains, we have observed that investments in data quality can offer dramatic improvements and typically pay for themselves. In this article, we aim to enable readers to make pragmatic upgrades by showing the mechanisms through which data quality and productivity interact, drawing on a useful analogy to lean manufacturing principles. From this discussion, we are able to define data quality as it pertains to risk analytics. We provide a quantitative and qualitative discussion of the benefits that can be realized with better data quality. Finally, we conclude with case studies that provide real examples of data quality in practice.

Introduction

Post-crisis regulations such as the Dodd-Frank Act have dramatically increased the consequences to financial institutions of unsound risk analytics. The analytical complexity and massive downside risks of unacceptable regulatory submissions lead firms to maintain large headcounts of high-cost analytical employees who transform data into projections of their institutions' financial performance. In our experience, this "get it right at any cost" situation results in material inefficiency, waste, and delay.

While there are many sources of waste, we frequently observe data quality to be a root cause. Dealing with some "garbage in" when there is no room for even a little "garbage out" is expensive. Bank risk managers and stress testing teams felt this acutely in the 2017 Comprehensive Capital Analysis and Review (CCAR) stress tests, when the Federal Reserve released a data error and revised it a week later. The incident prompted rework, compressed schedules, and created uncertainty.

Obviously, increasing data quality improves the productivity of these analytical risk management processes. What may not be obvious is how best to invest in data quality and what return on investment may be possible. Even defining data quality in the context of risk analytics is not straightforward.

In this article, we quantify the impact of data quality improvements on analytical productivity. We describe the key mechanisms of waste caused by data that we have observed in our work and provide examples of how to address them. These mechanisms lead to a functional definition of data quality. We conclude with several examples of the impact of improving data quality on efficiency in analytical tasks.

What's the Bottom Line?

Investing in data quality can provide a range of substantial cost savings. In research[1] and in our own experience, data quality investments consistently lead to a 30% to 50% reduction in direct costs – expenses such as payroll that are necessary for the actual production of analytical results and supporting documentation.

Improved data quality can also provide substantial indirect gains, cutting the costs that arise from quality problems and uncertainty. Although quantification of indirect costs remains elusive, we find two broad sources of such costs: model development time and confidence levels.

Model development time: Accurate data is a necessary – though not sufficient – condition for constructing a predictive model. If historical performance data for a loan portfolio are incorrect, a model developed on such a dataset will fail to capture the true underlying relationships between performance and economic factors. Noisy data will either provide weak signals at best or spurious correlations at worst.

From our experience developing a wide variety of econometric forecasting models, we find that poor data quality is the main reason for increasing the cycle time for model development. Having analytical modelers spend time addressing fundamental data issues during the model development process is wasteful for two reasons. First, being downstream consumers of data, modelers will waste time locating and communicating with the appropriate data experts within the organization. Second, the data corrections that modelers ultimately develop for their specific projects will not be captured at the source. The latter issue is particularly costly, as it implies that an institution may end up paying to address a single data error multiple times.

Confidence: High data quality creates confidence. It reduces noise, which in turn reduces model uncertainty. More broadly, model users who have low confidence in reported data are inclined to add an "uncertainty premium" to model results.

In the specific case of loss forecasting and allowance calculation, bad data may lead managers to assume conservative estimates ultimately leading to higher-than-necessary capital allocation. In this case, the cost of poor data quality directly translates into higher-than-required capital buffers and loss allowance provisions. While this may be prudent, overly conservative projections can price lenders out of the market, disappointing shareholders and ceding opportunities to more nimble competitors.

The indirect benefits of confidence may go beyond the business users of models. High-quality data are necessary to gain the confidence of model validators, senior managers, regulators, auditors, and other interested parties. Even if a model is well-constructed and estimated using state-of-the-art techniques, data anomalies can distract and call into question the integrity of model results – adding to compliance and other costs.

Quality is Productivity is Quality

The inseparable relationship between quality and productivity has been known in the manufacturing world for years,[2,3] and research on the topic of data quality has made effective use of the analogy.[4] Experience with serving our risk management clients and our own internal data processes has shown that the analogy also applies to quantitative analytical work. We have found, time and again, that there is a virtuous cycle between increasing quality and increasing productivity. Better-quality data boost analytical productivity by reducing wasted effort, idle resources, process bloat, and the number of required judgment calls. Likewise, higher productivity increases quality by automating error-prone tasks, reducing haste, and leaving time to evaluate results.

1   Hansen, Mark David. "Zero Defect Data: Tackling the Corporate Data Quality Problem." Massachusetts Institute of Technology. January 1991.

2   Shewhart, Walter A. *Economic Control of Quality of Manufactured Product*. New York: McGraw-Hill, 1931.

3   Deming, W. Edwards. *Out of the Crisis*. Cambridge, Mass.: MIT Press, 1986.

4   See Ref 1.

We have identified four major buckets of waste: rework, questioning, process deviations, and peak load.

Rework: The most obvious source of waste is having to discard a result and do the same task again. The serial nature of analytical processes makes rework issues especially costly when upstream steps are error-prone and weak error detection results in late identification of problems. Such situations may require even more of the process to be repeated.

Questioning: Evaluating anomalous results incurs additional costs, both in time spent and in the quality of the resources applied. Questioning an analytical result usually entails several higher-end tasks including troubleshooting, trade-off analysis, and ultimately making judgment calls. Judgment calls frequently require escalation of an issue to a person with the authority to make a decision. Improved data quality should reduce questioning time by reducing the frequency of anomalous results, caused by actual errors or poor models, and facilitating the troubleshooting process.

Process deviations: Unacceptable input data, such as model drivers with values outside required bounds, may not be addressable by rework and could require a change to the process itself. Changes must be developed, documented, and often validated. Furthermore, process deviations increase the probability of repeated errors, should the amended analytical step be repeated.

Peak load: All of the above reasons may delay the completion of upstream tasks, leaving less time for downstream tasks. When the available time for downstream tasks gets compressed, the organization may be forced to apply more resources to meet deadlines. This creates a spike in the demand for analytical capacity, which may require carrying excess staff or contracting for temporary help. When upstream tasks are completed efficiently, there is a decreased probability of downstream tasks being compressed.

The experience of the 2017 CCAR stress tests has elements of all four of these types of waste. The data quality concern was that the Federal Reserve issued incorrect data for BBB bond yields in its supervisory scenarios. A week later, it issued corrected scenario data. The rework this caused is obvious: Analysts set to work upon the initial release of data and were required to repeat some or all

of this work in response to the revised data. Additional questioning occurred, with institutions seeking to determine what this change meant for them and how best to proceed. Analytical results received further questioning after recalculations to evaluate the impact of the revised guidance. The unanticipated correction in scenario guidance certainly created process deviations, since doubtless few, if any, institutions anticipated this occurrence. Finally, the rework and additional week of delay in receiving definitive figures from the Federal Reserve compressed schedules and created higher peak loads.

Defining Data Quality

Identifying the channels where data problems can impair productivity enables us to propose functional requirements for data quality in risk analytics. High-quality data should be (1) verifiably correct, (2) fit for use, and (3) documented. These are thematically similar to data requirements in regulatory guidance,[5] but in this section we tailor their meaning and add specifics in terms of how they relate to the productivity of risk analytics processes.

Verifiably correct: Numerical correctness is clearly the minimum threshold for data quality. Running an analysis using incorrect data will likely incur rework or process deviation waste. However, the ability to independently verify numerical correctness further increases the quality of the data. This is especially important when using third-party data such as economic or market risk variables. The ability to independently verify data accelerates troubleshooting and avoids communication iterations with vendors or internal parties that would add delay and downstream peak load issues. Verifiability can come in several forms, such as backlinks to primary sources, quality declarations, unique identifiers, and accessible quality logs.

Fit for use: For data to produce sound analytical results, they must accurately quantify the concept they intend to measure. Modelers should consider both the definition of the data and their objective properties, such as time series length, frequency, timeliness, and consistency. Data that are too aggregated or nonspecific may provide weak or wrong fits, such as if national data on house prices were used when one could use state-, metro-, or ZIP code-level data. Using true high-frequency data should almost always be superior to interpolating lower-frequency data. Dealing with

---

5   See Basel Committee on Banking Supervision, "Principles for effective risk data aggregation and risk reporting," BCBS 239, January 2013; and Office of the Comptroller of the Currency, "Supervisory Guidance on Model Risk Management," OCC 2011-12, April 4, 2011.

outliers or definitional breaks reduces degrees of freedom in model estimation. Data that are fit for use should produce better-functioning models with more trusted results. This not only speeds up model development, but also reduces the expected questioning time and probability of process deviations.

Documented: Documentation is essential to the interpretation of data. To do their job effectively, modelers and analysts need to know specifics on how the data are defined and constructed: Exactly which loans are included in this sample? Where estimation is used, one should know the uncertainty associated with the estimate: Is the volatility in this time series due to volatility in real life or uncertainty in estimation? Anomalous data points should be notated and explained: Does this temporary spike in charge-offs represent a policy change, an unexplained but actual rise, or erroneous data? This knowledge gives modelers the tools to decide proper treatment of the data when creating models and increases the confidence in their choices. Questioning time for analysts and validation teams should be reduced when tracing the sources of model results.

Quality in Practice

Several guiding principles underlie the data quality initiatives we have implemented in our analytical processes:

» Prioritize upstream data inputs, especially in model development.

» Implement high-throughput quality checks to verify the thousands of variables being forecast.

» Maximize use of objective pass/fail tests with low rates of false positives.

» Log judiciously and store intermediate results.

» Press data vendors to improve their delivered data quality.

» Catalog data centrally and insure all users have access to the latest catalog.

The following case studies illustrate the application of these principles in real situations.

Case 1: Consumer Credit Loss Modeling

A large regional bank needed to create expected credit loss models for its auto loan portfolio to complete a regulatory submission for the Federal Reserve's annual CCAR and Dodd-Frank Act stress testing programs. Two broad data quality issues – cataloging and consistency – impacted the cost, quality, and timing of the models that were produced for this project.

A member of the IT department provided the project managers with a data dictionary and a high-level summary of the bank's historical data. At first blush, the information indicated that the bank had monthly, historical performance data available from 2003 to 2015 for more than 3 million auto loan and lease originations. The data included a number of commonly used borrower and loan characteristics that were captured at the time the accounts were originated.

Based on this wealth of information, the data were sufficient to build a robust competing risk model of performance at the loan level. We interviewed the model's intended users, who identified a set of desirable features and variables they wanted the model to consider – such as the debt-to-income ratio of borrowers – when predicting default probabilities.

After defining the scope and success criteria for the models, the first task was for bank staff to inspect the bank's system of records and construct a loan-level dataset containing monthly performance observations for all of the loans in its historical portfolio.

The data quality issue of cataloging was immediately apparent at this stage of the project, as the modeling team discovered the bank's loan data were stored on a variety of disparate systems – the result of years of mergers and acquisitions. Not only was the provided data dictionary incomplete and outdated, but it also failed to indicate that variables were defined and labeled inconsistently across multiple databases.

This second data quality issue, inconsistency of data across time, resulted in significant direct and indirect costs to the bank. The modeling project was halted for four months while the bank commandeered resources to merge and normalize the data. Five full-time, highly skilled database programmers were deployed for this effort at a direct cost in excess of $250,000. In addition, the bank incurred indirect costs such as loss of revenue and competitiveness due to delaying other scheduled activities and projects. The loss of modeling and business team time was highly disruptive, and the competing business priorities created a tumultuous work environment.

The compressed timeline that resulted translated into a direct cost, as the bank had to hire contractors to expedite model development and validation. Despite the additional expense,

the quality of the models ultimately suffered; portions of the model development process that are typically sequential, such as experimentation and revision, had to be done in tandem to save time. The revised timeline did not allow for the full battery of tests that had been originally planned, necessitating some compromises on business user requests.

High data quality is fundamental to sound risk management and analytical productivity.

Lack of confidence in the data introduced real costs as well. Senior leadership grew nervous with the delays and developed a contingency plan in the event that data cleaning and model development were not completed in time. This effort increased the overall cost of the project by about a third and produced models that were both inferior and unusable for other applications. In the end, the main modeling effort was completed in time and the contingency plan was not needed – but not before the additional expense was incurred.

An observer may have attributed delays in the project to the modeling team's efforts, as the only tangible project deliverable – model documentation – was delayed relative to the projected delivery date. However, looking at the project through a wider lens – as the bank did in a subsequent debrief – it was clear that the root cause of the delay was traceable to poor data quality.

Case 2: Macroeconomic Scenarios

Each month, Moody's Analytics produces macroeconomic forecasts and alternative economic scenarios for 58 countries. The process involves about 45 economists, many of whom are senior, and usually takes three weeks per month. Data for the process come from more than 100 sources around the world.

The high complexity and large resource requirement of this forecast process create monthly peak loads which impose bounds on analytical throughput and scheduling. Consequently, we strategically invested in reducing both the size and duration of this peak load while improving quality, with the goal of materially increasing analytical output and shortening process time. To that end, we redesigned the forecasting process with an emphasis on implementing the concepts previously discussed. Several key investments were made, notably in a new infrastructure for updating historical data, a new initial forecast system, standardized forecast quality-check programs, and mistake-

proofing the forecast editing process.

The central features of the new historical data infrastructure are improved archiving, logging, transparency, and high-throughput quality checks. The system takes as inputs a mapping spreadsheet with definitions, equations, and metadata and produces the complete dataset needed for the forecast. The system also runs a barrage of quality tests on all data and results. Along the way, it archives the mapping information, raw data, processed data, and the output of all quality checks run. Archiving intermediate results improves our ability to diagnose problems quickly and explain unexpected results. Furthermore, we log each calculation to accelerate troubleshooting. This not only makes the actual update process go faster, but also facilitates answering questions that may come from analysts downstream.

The quality checks run by the historical data infrastructure efficiently surface potential problems and focus attention on what is important. These checks include:

» Data with large or deep revisions relative to the previous month's forecast run

» Metadata changes

» Discontinued data

» Highly lagged data

» Sparse data

Next, the initial forecasting system merges the new historical data, model equations, and carried-over forecast assumptions to produce an initial forecast. If it fails to produce the intended results, forecasters need to perform a substantial amount of cleanup work downstream. We substantially improved the initial forecasting system by the addition of integrity checks, which verify that intended steps did, in fact, occur as intended. These checks verify the following:

» Every variable in the model has updated history.

» All intended exogenous overrides are applied to the forecast.

» Exogenous add-factor variables created by re-endogenizing variables produce exactly the desired result.

Additionally, we archive a copy of the initial forecasting package – data, equations, add-factors, and quality findings – to facilitate troubleshooting and trace the causes of specific results downstream.

After that, we enhanced the process by which economists impart judgment into the forecast by adding universal quality checks and redesigning support programs to prevent forecasters from taking actions that would impair quality.

The universal quality checks inspect for objectively implausible

---

Our support program redesign implemented *poka-yoke*, or "inadvertent error prevention," to protect data quality.

---

results such as values going negative that should not, long-run seasonality in a forecast of a deseasonalized variable, and unprecedented growth being forecast in the first period of forecast. The support programs display these quality findings prominently each time an economist runs the model.

The support program redesign implemented *poka-yoke*, or "inadvertent error prevention,"[6] in several ways. If an economist attempts to edit the forecast of a variable that he or she should not edit, the solve program halts and displays an appropriate message. The program also firewalls actual historical figures from editing and overrules any attempted edits where there is reported history. The same protection of numerical values also extends to shared assumption data that flow into all models in use.

Overhauling the process with a focus on data quality achieved telling results. Economists reported their time before and after the operational redesign and showed a 30% reduction in direct time spent forecasting. Quality improved measurably: Delivery schedules moved forward by several days, and forecast restatements went from common to rare. More than one economist remarked that the improvement in upstream data quality reduced cleanup time and let them spend more time adding value to the forecast.

Conclusion

Data quality is fundamental to sound risk management and analytical productivity. As our case studies have illustrated, problems with data quality can trickle through and affect every subsequent analysis, model, and decision made by an institution. These problems incur substantial direct and indirect costs. Conversely, high data quality creates a virtuous cycle

of productivity and quality, reducing labor costs, accelerating schedules, and bolstering confidence in analytical output.

Despite the common refrain from every business executive that information is central to his or her ability to make proper decisions, data quality often doesn't get the attention – or the funding – that revenue-generating initiatives do. However, our experience and research show that the benefits of investing in data quality are widespread and pay for themselves quickly.

The indirect costs of not investing in data quality are significantly higher than the direct costs of implementing quality programs. The costs of data assurance are low relative to the benefits. Before dismissing data quality initiatives as expensive, businesses need to consider their return on investment through the lens of cost avoidance.

---

6   Shingo, Shigeo. *Zero Quality Control: Source Inspection and the Poka-Yoke System*. Portland, Ore.: Productivity Press, 1986.

RISK PERSPECTIVES

Find additional articles, interviews, and multimedia content at MoodysAnalytics.com/RiskPerspectives

CONTACT US
Visit us at **moodysanalytics.com** or contact us at a location below:

AMERICAS	EMEA	Asia (excluding Japan)	JAPAN
+1.212.553.1653	+44.20.7772.5454	+852.3551.3077	+813.5408.4100
clientservices@moodys.com	clientservices.emea@moodys.com	clientservices.asia@moodys.com	clientservices.japan@moodys.com

MOODY'S
ANALYTICS