CAPCO

THE CAPCO INSTITUTE JOURNAL OF FINANCIAL TRANSFORMATION

DATA MANAGEMENT

Synthetic financial data: An application to regulatory compliance for broker-dealers

J. B. HEATON | JAN HENDRIK

DATA ANALYTICS

50TH EDITION | NOVEMBER 2019

THE CAPCO INSTITUTE

JOURNAL OF FINANCIAL TRANSFORMATION

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

CONTENTS

DATA MANAGEMENT

DATA ANALYTICS

DATA INTELLIGENCE

DEAR READER,

Welcome to the milestone 50th edition of the Capco Institute Journal of Financial Transformation.

Launched in 2001, the Journal has covered topics which have charted the evolution of the financial services sector and recorded the fundamental transformation of the industry. Its pages have been filled with invaluable insights covering everything from risk, wealth, and pricing, to digitization, design thinking, automation, and much more.

The Journal has also been privileged to include contributions from some of the world's foremost thinkers from academia and the industry, including 20 Nobel Laureates, and over 200 senior financial executives and regulators, and has been co-published with some of the most prestigious business schools from around the world.

I am proud to celebrate reaching 50 editions of the Journal, and today, the underlying principle of the Journal remains unchanged: to deliver thinking to advance the field of applied finance, looking forward to how we can meet the important challenges of the future.

Data is playing a crucial role in informing decision-making to drive financial institutions forward, and organizations are unlocking hidden value through harvesting, analyzing and managing their data. The papers in this edition demonstrate a growing emphasis on this field, examining such topics as machine learning and AI, regulatory compliance, program implementation, and strategy.

As ever, you can expect the highest caliber of research and practical guidance from our distinguished contributors, and I trust that this will prove useful to your own thinking and decision making. I look forward to sharing future editions of the Journal with you.

Lance Levy, **Capco CEO**

FOREWORD

Since the launch of the Journal of Financial Transformation nearly 20 years ago, we have witnessed a global financial crisis, the re-emergence of regulation as a dominant engine of change, a monumental increase in computer processing power, the emergence of the cloud and other disruptive technologies, and a significant shift in consumer habits and expectations.

Throughout, there has been one constant: the immense volume of data that financial services institutions accumulate through their interactions with their clients and risk management activities. Today, the scale, processing power and opportunities to gather, analyze and deploy that data has grown beyond all recognition.

That is why we are dedicating the 50th issue of the Journal of Financial Transformation to the topic of data, which has the power to change the financial industry just as profoundly over the coming 20 years and 50 issues. The articles gathered in this issue cover a broad spectrum of data-related topics, ranging from the opportunities presented by data analytics to enhance business performance to the challenges inherent in wrestling with legacy information architectures. In many cases, achieving the former is held back by shortcomings around the quality of, and access to, data arising from the latter.

It is these twin pillars of opportunity and challenge that inform the current inflection point at which the financial industry now stands. Whilst there is opportunity to improve user experiences through better customer segmentation or artificial intelligence, for example, there are also fundamental challenges around how organizations achieve this – and if they can, whether they should.

The expanding field of data ethics will consume a great deal of senior executive time as organizations find their feet as they slowly progress forward into this new territory. In my view, it is critical that organizations use this time wisely, and do not just focus on short-term opportunities but rather ground themselves in the practical challenges they face. Financial institutions must invest in the core building blocks of data architecture and management, so that as they innovate, they are not held back, but set up for long-term success.

I hope that you enjoy reading this edition of the Journal and that it helps you in your endeavours to tackle the challenges of today's data environment.

Guest Editor
Chris Probert, **Partner, Capco**

SYNTHETIC FINANCIAL DATA: AN APPLICATION TO REGULATORY COMPLIANCE FOR BROKER-DEALERS

J. B. HEATON | One Hat Research LLC
JAN HENDRIK WITTE | Honorary Research Associate in Mathematics, University College London

ABSTRACT

The hype of "big data" has not escaped the investment management industry, although the reality is that price data from U.S. financial markets are not really big data; price data is small data. The fact that sellers and advisors in financial markets use small data to generate and test investment strategies creates two major problems. First, the economic mechanisms that generate prices (and, therefore, returns) may change through time, so that historical data from an earlier time may tell us little or nothing about future prices and returns. Second, even if data-generating-mechanisms are somewhat stable through time, inferences about the profitability of investment strategies may be sensitive to a handful of outliers in the data that get picked up again and again in different strategies mined from the same small data set. In this article, we present an answer to the financial small data problem: using machine-learning (ML) methods to generate "synthetic" financial data. The essential part of our approach to developing synthetic data is the use of ML methods to generate data that might have been generated by financial markets but was not. Synthetic price and return data have numerous uses, including testing new investment strategies and helping investors plan for retirement and other personal investment goals with more realistic future return scenarios. In this article, we focus on a particularly important use of synthetic data: meeting legal and regulatory requirements such as best interest and fiduciary requirements.

1. INTRODUCTION

In the age of "big data", those in the investment management industry have a "small data" problem. While it is tempting to think of financial market data as voluminous, financial-market data is tiny by comparison to many big data collections. Companies like Walmart, Amazon, PayPal, Facebook, and Google collect petabytes (one petabyte equals a million gigabytes) of data every hour. Their daily data collections dwarf the data that financial market transactions generate. While the hype of big data has not escaped the investment management industry, the reality is that price data from U.S. financial markets are in fact "small data".

Nevertheless, financial market participants often use the small data of financial markets to generate and test investment strategies. This comes with two major problems. First, the economic mechanisms that generate prices (and, therefore, returns) may change through time. That is, price-data-generating mechanisms may be nonstationary, so that historical data from an earlier time may tell us little or nothing about future prices and returns. Second, even if data-generating-mechanisms are somewhat stable through time, inferences about the profitability of investment strategies may be sensitive to a handful of outliers in the data that get picked up again and again in different strategies mined from the same small data.

It now appears that the outlier problem may be far more serious than previously recognized. For decades, investment advisors and broker-dealers have assumed that the historical premium of equities over risk-free securities implied (1) that stocks are a generally superior investment strategy for the "long term", and (2) that the superiority of the overall stock-market returns implied that professional money managers could earn even higher returns by actively seeking out stocks with the best risk-return characteristics among the total set of stock market offerings. It turns out, however, that the first implication is highly fragile because the overall historical superiority of equities over risk-free securities rests on the superior performance of a handful of securities. That is, while it is easy to form an intuition that the returns to individual stocks will be roughly bell-shaped and centered around the market return, the return distribution is often highly skewed over time, with a handful of stocks that earn above the total index return and a majority that earn below it.

Recent research has demonstrated that the superiority of equities as a whole (that is, the entire stock market) over risk-free securities in the last century or so does not reflect a reliable tendency for smaller groups of equities to outperform risk-free securities. In a pathbreaking work, Bessembinder (2018) finds that the majority of U.S. listed common stocks have returned (inclusive of dividends) less than the risk-free rate (that is, the one-month Treasury bill) over their lives as listed companies, so that just 4% of listed U.S. companies account for all of the gains of the U.S. stock market from 1926 to 2016. Bessembinder et al. (2019) find similar results for the period 1990 to 2018: a majority of both U.S. and non-U.S. stocks underperform the one-month U.S. treasury bill rate over this period.

Researchers are now realizing that the power of passive indexing to beat active managers year-after-year may rest on this empirical fact, since large indexes tend to catch the handful of extreme winners that stock-pickers and other active managers may miss [Ikenberry et al. (1992), Heaton et al. (2017))][1] Historical price and return data that contains a handful of outliers that drive investment performance is an unreliable basis, on its own, for generating and testing investment strategies.

In this article, we present an answer to the financial small data problem: using machine-learning (ML) methods to generate "synthetic" financial data. Outside of financial services, synthetic data has been used to allow analysis of otherwise confidential data by making modest changes that protect privacy but leave statistical inferences intact [Little (1993), Rubin (1993)]. Methods of generating synthetic data have also been used in a number of other contexts where actual data was lacking in sufficient quantities, such as in training image classification systems [Krizhevsky et al. (2012), Tremblay et al. (2018), Wang et al. (2019)], training systems to read Indic handwriting [Roy et al. (2018)], generating synthetic mobile payment transactions to train fraud-detection algorithms [Lopez-Rojas et al. (2016)], and augmenting data from wearable health sensors [Taewoong Um et al. (2017)], among others.

> *"In the age of "big data", those in the investment management industry have a "small data" problem."*

In these examples, the goal in generating synthetic data for problems like these involves generating data that is different in ways that does not alter the fundamentally stability of the relationship to be learned.

Financial markets present a far different problem. Financial-market data is likely to be generated by mechanisms (interactions of traders using information) that are not stable through time and, more importantly, are unstable ("nonstationary", in statistical parlance) in unpredictable ways. While dogs generally look the same over a period of decades (even allowing for new hybrid breeds), allowing for successful image recognition algorithms, we know very little about the mechanisms that generate prices and how those mechanisms change through time. Even if a researcher finds a good model of price behavior in a particular period of time, there is little

[1] To illustrate the problem with outliers, consider an (equally weighted) index of five securities, four of which (although it is unknown which) will return 10% over the relevant period, and one of which will return 50%. Suppose that active managers choose portfolios of one or two securities and that they equally weight each investment. There are 15 possible one- or two-security "portfolios". Of these 15, 10 will earn returns of 10%, because they will include only the 10% securities. Just five of the 15 portfolios will include the 50% winner, earning 30% if part of a two-security portfolio and 50% if it is the single security in a one-security portfolio. The mean average return for all possible actively managed portfolios will be 18%, while the median portfolio of all possible one- and two-stock portfolios will earn 10%. The equally weighted index of all five securities will earn 18%. Thus, in this example, the average active management return will be the same as the index, but two-thirds of the actively managed portfolios will underperform the index because they will omit the 50% winner.

reason to believe that prices will behave today as they did 10 or 20, or even five years, ago. Synthetic data that only mimics historical data is unhelpful, since the primary danger presented by the use of historical data is that past performance of a given investment product or strategy may depend on data outliers that do not repeat; outliers that we now know are driving returns.

It is important to note that these "outliers" are, in general, not necessarily apparent to the naked eye. While some outliers are easy to see in the data as individual stocks that have extreme returns, other outliers are outliers in relationships among stocks, rare occurrences in the high-dimensional relationships that exist among a huge number of possible return combinations.

We have developed a method for generating synthetic financial data. The essential part of our approach to developing synthetic data is the use of ML methods to generate data that might have been generated by financial markets but was not. This is an important contrast between our application and previous methods to generate synthetic data. While much synthetic data seeks to introduce randomizations in existing data that do not influence the learning task or statistical inferences, our goal is to generate synthetic data that introduces randomizations that matter for inferences. Our approach is to assume that the features of the past data that are relatively common are more likely to repeat than the features of the past data that are relatively rare. By identifying the features of the past data that are relatively rare relative to the other features of the data, we can generate data that does not assume that those rarities will appear in the future as they did in the past.

Synthetic price and return data have numerous uses, including testing new investment strategies and helping investors plan for retirement and other personal investment goals with more realistic future return scenarios. In this article, however, we focus on a particularly important use of synthetic data: meeting legal and regulatory requirements such as best interest and fiduciary requirements.

2. THE COMPLIANCE PROBLEM: SMALL DATA AND CHANGING RULES

Recent research casting doubt on the superiority of many equity strategies could not come at a worse time for broker-dealers. In June 2019, the U.S. Securities and Exchange Commission (SEC) adopted Regulation Best Interest (RBI) to regulate the conduct of broker-dealers who make recommendations to retail customer of a securities transaction or investment strategy involving securities. Among its many requirements, the regulation requires broker-dealers to exercise reasonable diligence, care, and skill in making a recommendation to a retail customer. This is known as the "Care Obligation."

The SEC's Final Rule states that "whether a broker-dealer's recommendation satisfies the Care Obligation will be an objective evaluation turning on the facts and circumstances of the particular recommendation and the particular retail customer" and further states that the care obligation requires that a broker-dealer understands "potential risks, rewards, and costs associated with the recommendation." The SEC further states that "[s]cienter [bad intent] will not be required to establish a violation of Regulation Best Interest". This suggests that negligence or recklessness will be sufficient to state a claim against broker-dealers for violations of RBI, including the failure to adequately discharge the Care Obligation.

RBI places substantial new compliance burdens on broker-dealers, but it is not the only law or regulation governing the recommendation and sale of investment strategies. Investment advisors governed by the U.S. Investment Advisors Act of 1940 are fiduciaries to their clients, as are those who oversee pension plans governed by the U.S. Employee Retirement Income Security Act (ERISA).

Even those without these statutory and regulatory duties are liable under common law to those they mislead negligently (in certain circumstances) or recklessly (in many circumstances). This may apply, for example, to otherwise lightly regulated hedge fund managers.

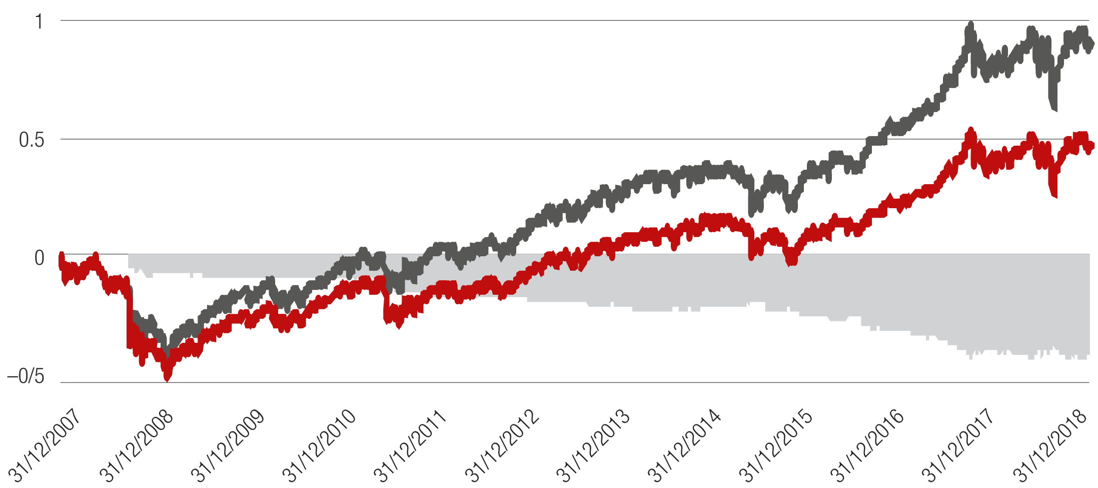
3. A SYNTHETIC DATA SOLUTION

Given the known limitations of historical data, how can a broker or fiduciary gain confidence that an investment strategy will not result in future regulatory action or litigation? Put differently, what work would a broker-dealer or fiduciary want to show was done to support its recommendations and actions if accused of basing advice on bad inferences from historical data?

Our proposal is to use synthetic data to test investment products for investor best interest and the requirements of fiduciary duty. Inexpensively generated synthetic data can supplement more suspect historical data to better screen products for a client's best interest and the satisfaction of fiduciary requirements.

Our specific approach uses ML fraud-detection algorithms [Bolton and Hand (2002), Kou et al. (2004), Abdallah et al. (2016), Porwal and Mukund (2019)] in a novel way. Fraud-detection algorithms are outlier detectors, designed to detect fraudulent transactions that make up a very small proportion of all financial transactions. For example, the Kaggle Credit Card Fraud Detection Dataset, a dataset of "anonymized credit card transactions labeled as fraudulent or genuine" contains only 0.172% fraudulent transactions; the remaining 99.828% of the transactions are genuine.

Figure 1: Correction to compensate for unsuitable data



Source: Bloomberg and the authors

We use a fraud-detection approach to identify high-dimensional outliers in the historical dataset and replace them with a larger alternative dataset that reflects the different ways in which the joint prices might alternatively have been realized in the past. We build on the ML sub-specialty of deep learning that has recently provided a successful set of solutions [Sangeetha et al. (2017), Choi and Lee (2018), Fu et al. (2016), Phua et al. (2018), Zhang et al. (2018)]. We apply the fraud-detection algorithm approach to generate synthetic data that is less dependent on historical outliers. The resulting synthetic datasets have little to no dependence on historical anomalies while maintaining all other characteristics with a high degree of accuracy. This method is far superior to simplistic Monte Carlo-based approaches.

> *"A recommendation that depends on such a small number of non-repetitive historical anomalies is unlikely to pass muster under regulatory requirements. Synthetic data would have shown why."*

4. AN EXAMPLE: THE DOW JONES INDUSTRIAL AVERAGE AS AN INVESTMENT STRATEGY

As an example, consider the Dow Jones Industrial Average (DJIA) as if it was a marketed investment strategy. Starting with a dataset of daily closing prices of all DJIA constituents for the period January 2, 2008 to May 22, 2019, we recreate the index and deploy our modified fraud-detection algorithm to identify days for which the price change in the 30 considered stocks exhibit highly unusual activity. These are changes that typically are unnoticeable to the naked eye, occurring as they do in the complex interrelationships among the stocks. While the historical DJIA has an annual return of 6.2% over the period (see dark grey line in Figure 1), synthetic data that is not as sensitive to outliers suggests an average annual return of 3.4% (the red line in the figure). The difference (the light grey area in the figure), 2.8% per year, demonstrates the importance of a very small number of non-repetitive historical anomalies.

A broker-dealer that recommended the DJIA, but did not consider the dependence of a DJIA investment strategy on a handful of outliers unlikely to repeat, would open itself to legal and regulatory liability easily demonstrated by proof of the undisclosed (and perhaps even unanalyzed) importance of those outliers.

A recommendation that depends on such a small number of non-repetitive historical anomalies is unlikely to pass muster under regulatory requirements. Synthetic data would have shown why.

5. CONCLUSION

Sellers of investment strategies face considerable legal and regulatory hurdles in marketing their products. Synthetic data may provide the only defensible basis for testing investment strategies for compliance. It is becoming better understood that historical data may not support the sale of many investment strategies because the strategies are too highly dependent on outliers that are unlikely to repeat in the future. Sellers and investment advisors who have not taken steps to test "potential risks, rewards, and costs associated with [a] recommendation" beyond looking at historical performance will likely find themselves in an indefensible position with regulators and litigants.

Synthetic data generation methods can identify rare (often very high-dimensional) outliers in data and replace them systematically to capture what data might have been generated instead but was not. Our method uses fraud-detection algorithms "inverted" to identify and replace outliers, creating an enormous number of synthetic datasets that can be used to test investment strategies. Here, we illustrate the approach with the Dow Jones Industrial Average, an easy-to-understand "investment strategy" proxy, but our method applies equally to the most complex quantitative strategies, including those that operate at very high frequency. Our approach provides a way for financial services firms to use advances in ML to solve a new and pressing compliance problem: avoiding regulatory violations in investment advice and oversight.

REFERENCES

Abdallah, A., M. A. Maarof, and A. Zainal, 2016, "Fraud detection system: a survey," Journal of Network and Computer Applications 68, 90-113

Bessembinder, H., 2018, "Do stocks outperform treasury bills?" Journal of Financial Economics 129, 440-457

Bessembinder, H., T.-F. Chen, G. Choi, and K. C. J. Wei, 2019, "Do global stocks outperform US treasury bills," working paper, https://bit.ly/2lFYAzz

Bolton, R. J., and D. J. Hand, 2002, "Statistical fraud detection: a review," Statistical Science 17:3, 235-255

Choi, D., and K. Lee, 2018, "An artificial intelligence approach to financial fraud detection under IoT environment: a survey and implementation," Security and Communication Networks, https://bit.ly/2mmlsVb

Fu, K., D. Cheng, Y. Tu, and L. Zhang, 2016, "Credit card fraud detection using convolutional neural networks," International Conference on Neural Information Processing, https://bit.ly/2nl0fed

Heaton, J. B., N. Polson, and J. H. Witte, 2017, "Why indexing works," Applied Stochastic Models in Business and Industry 33, 690-693

Ikenberry, D., R. Shockley, and K. Womack, 1992, "Why active managers often underperform the S&P500: the impact of size and skewness," Journal of Private Portfolio Management 1, 13-26

Kou, Y., C. T. Lu, S. Sinvongwattana, and Y. P. Huang, 2004, "Survey of fraud detection techniques," proceedings of the 2004 IEEE International Conference on Networking

Krizhevsky, A., I. Sutskever, G. E. Hinton, 2012, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, https://bit.ly/2nTNOGL

Little, R. J., 1993, "Statistical analysis of masked data," Journal of Official Statistics 9, 407-426

Lopez-Rojas, E. A., A. Elmir, and S. Axelsson, 2016, "PaySim: a financial mobile money simulator for fraud detection," available at https://bit.ly/2MJdNfN

Phua, C., V. Lee, K. Smith, and R. Gayler, 2018, "A comprehensive survey of data mining-based fraud detection research," arXiv:1009.6119

Porwal, U., and S. Mukund, 2019, "Credit card fraud detection in e-commerce: an outlier detection approach," arXiv: 1811.02196

Roy, P. P., A. Mohta, and B. B. Chaudhuri, 2018, "Synthetic data generation for Indic handwritten text recognition," arXiv:1804.06254v1

Rubin, D. B., 1993, "Discussion: statistical disclosure limitation," Journal of Official Statistics 9:2, 461-468

Sangeetha, K. N., Veenadevi, and B. A. Usha, 2017, "Benefits of SVM and deep learning in credit card fraud detection: a survey," International Conference on Signal, Image Processing Communication & Automation (ICSIPCA), 233-236

Taewoong Um, T., F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulic, 2017, "Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks," arXiv:1706.00527v2

Tremblay, J., A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, 2018, "Training deep networks with synthetic data: bridging the reality gap by domain randomization," arXiv:1804.06516

Wang, K., F. Shi, W. Wang, Y. Nan, and S. Lian, 2019, "Synthetic data generation and adaptation for object detection in smart vending machines," arXiv: 1904.12294

Zhang, R., F. Zheng, and W. Min, 2018, "Sequential behavioral data processing using deep learning and the Markov transition field in online fraud detection," arxiv:1808.05329

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward.

Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and asset management and insurance. We also have an energy consulting practice in the US. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Hong Kong
Kuala Lumpur
Pune
Singapore

EUROPE
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

FSC
www.fsc.org
MIX
Paper from responsible sources
FSC® C001648