SOCIETY OF ACTUARIES

Data Visualization

2016 Call for Essays

I’m a visual learner. I find I’m often more effective at receiving and understanding ideas and concepts when they are communicated through images. If you’re giving me directions to a new location, a list of turns to make doesn’t help much … I’d much rather have a route map. When my wife and I discuss workshop projects or are considering some home remodeling, my typical opening remark is, “Let’s draw a picture …”.

The actuarial business world as well has always held visualization as an important tool. Just as soon as high-powered spreadsheets evolved, actuaries were among the first to ask how to graph, chart and visually display the results. And now, as high-powered analytics tools go even deeper into our work, we need to make sure our results are well understood and communicated through data visualization techniques.

We’ve put together some examples in this report from a recent call for articles that highlights several uses of data visualization. I think you’ll enjoy the creativity of this information. I hope it encourages us all to create new ways to visually communicate our actuarial work. Be sure to let us know your thoughts and share other examples of data visualization techniques at the SOA Engage Research Community, our new online forum for discussing ideas.

R. Dale Hall

R. Dale Hall, FSA, MAAA, CERA, CFA
Managing Director of Research
Society of Actuaries

Contents

3 A Time-Dependent Interactive Visualization on IFRS 4 Phase II General Approach
Jen Houng (Erik) Lie, FSA, CERA

9 Effective Communication of Stochastic Model Results
John Hegstrom, FSA, MAAA

12 Visualization of Social Network Data
Kailan Shang, FSA, CFA, PRM, SCJP

A Time-Dependent Interactive Visualization on IFRS 4 Phase II General Approach

Jen Houng (Erik) Lie, FSA, CERA

If I can't picture it, I can't understand it.— Albert Einstein

This article will discuss the consideration of a time-dependent interactive visualization on International Financial Reporting Standard (IFRS) 4 Phase II general approach.

The illustration website can be found at *http://ifrs4.getforge.io*.

Data

All data used in this illustration are fictitious, except the mortality table used (CSO 2001). A nonparticipating, single-pay policy with a five-year death benefit coverage is assumed. No surrender cash value will be paid. The policyholder is assumed to be age 45, male, nonsmoker. Ten thousand identical policies are assumed. Policy is not onerous at issue (2015). Discount rate is assumed to be a flat 3 percent in all years. Risk adjustment is assumed 0 for simplicity.

Visualization

Three staked column charts, each with two individual components presented in columns and one total value presented with a plotted line, are displayed. The three charts represent the following financials:

- IFRS Reserve
- Profit
- Total Comprehensive Income

The *x*-axis of each chart represents the actual or projection years, and the *y*-axis represents the dollar amount. The side panel on the left allows users to switch between these three charts.

A slider bar (with a play button below) allows users to switch across time (valuation year).

The control panel allows the user to change the following assumptions for interactive display:

- Actual Deaths
- Mortality Selection Factor
- Discount Rate

Some predefined scenarios are set for understanding the relationship between assumptions and financial outcomes.

The platform was constructed on the Web using JavaScript, and the chart plotting was done by Highcharts.

The Problem

The project is initiated with the intent to validate the IFRS 4 financial results. The new accounting standard introduces a lot of implicit logical relationships between the assumptions and the financial results; any mistake in the calculation may destroy these relationships. To solve the problem, an interactive visualization that can answer the "what-if" scenarios under different assumption setups can help our understanding and allow us to explore unexpected patterns under different conditions.

Considerations

The difficulties of formulating the visualization strategy are as follows:

- Existence of two time-related dimensions
- Multi-perspective

We will describe each in more detail in the sections that follow.

EXISTENCE OF TWO TIME-RELATED DIMENSIONS

The existence of two time-related dimensions—namely, the actual year and projection years—is a specific and unique feature that exists only in the actuarial industry. "Actual year" refers to the actual valuation year. For example, if this year is 2016, then we will run projection models to predict the future cash flows in 2017, 2018 and so on. As one year passes, we will again run projection models to predict 2018, 2019 and so on. In order to demonstrate the IFRS 4 calculation, it is essential to consider both the actual and projection years and to understand how the financials emerge in response to these two dimensions.

There are two options for mapping time: the mapping of time to space and the mapping of time to time (Aigner et al. 2011). One possible solution is to map

Figure 1 Three-Dimensional Plot

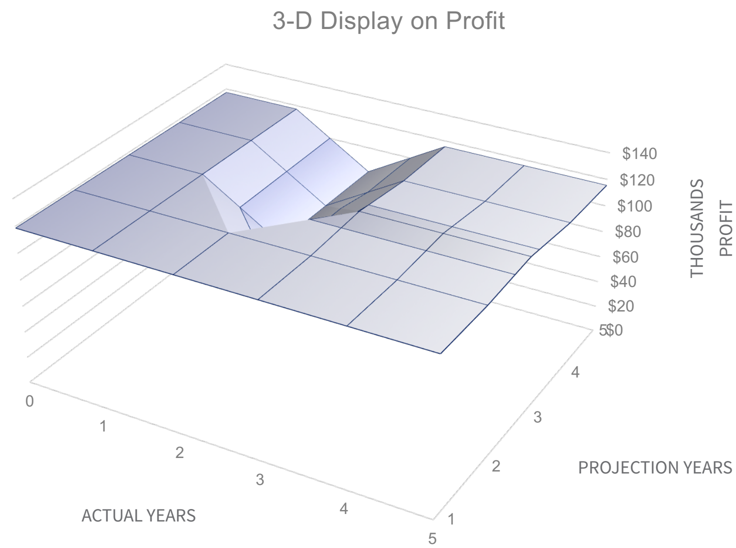
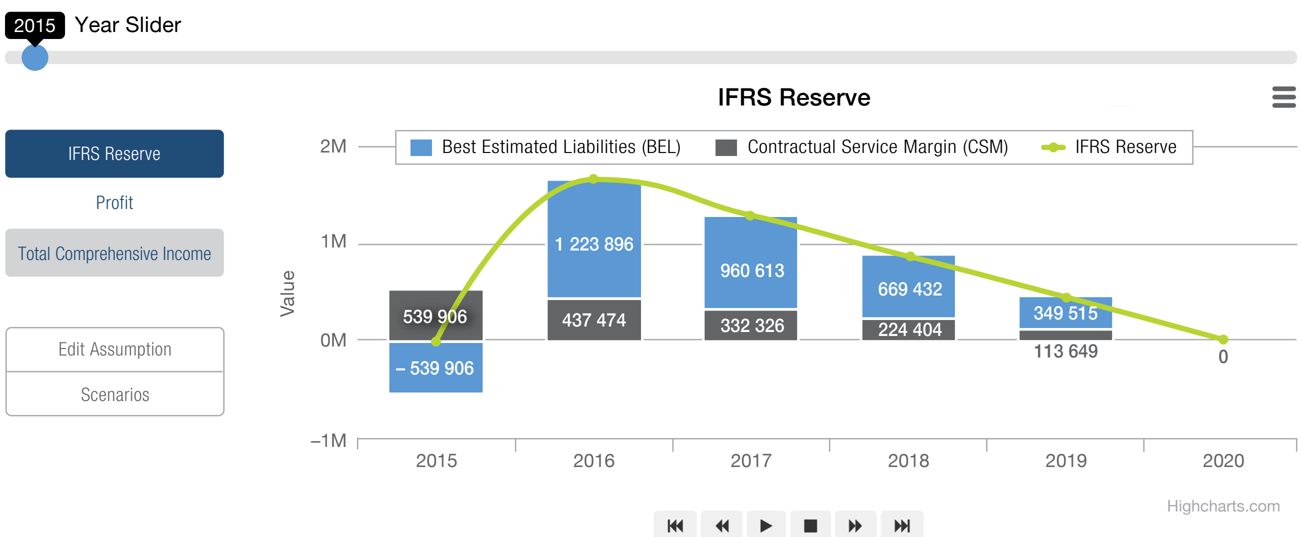


Figure 2 Slider Bar for the Time Dimension



both time dimensions to space by using a three-dimensional plot (Figure 1). Using this method, the *x*-axis and the *y*-axis represent the actual year and the projection year, respectively, and the *z*-axis represents the dollar amount of profit.

The outcome of the three-dimensional plot is not ideal. Attempts to show abstract, nonspatial data using three-dimensional views have not been successful, mostly because the information is not displayed clearly and because getting around in three dimensions is much harder than navigation in two dimensions (Ware 2008). In addition, further breakdowns on profit are required; this creates more information, which is hard to encode in 3-D graphs.

An alternative solution is to map one time dimension to space and another one to time. This solution, by nature, is more intuitive. A slider bar with a play button is created, which introduces the extra dimension and gives users clear, direct control of the time axis (Figure 2).

MULTIPERSPECTIVE

Unlike many visualization projects in which only one final item or message is intended to be delivered, the IFRS 4 income statement creates the following three financials:

- IFRS reserve
- Profit
- Total comprehensive income

In addition, there are three approaches to present the profit (Vlaminckx et al. 2013):

- Traditional approach
- Summarized margin approach
- IFRS approach

The three approaches are mathematically equivalent.[1] For the ease of understanding, the summarized margin approach is used as the default presentation format. Navigational tabs and corresponding pages are created to allow users to switch between different approaches (Figure 3). This enables users to explore the similarities and differences between these presentation formats in the visualized charts.

Demonstration

The following demonstration is equal to "Scenario 2" in the illustration on the website, which includes the following assumptions:

- Mortality selection factor in 2017 increases from 100% to 110% and returns to 100% in 2018 and onward.

- Actual deaths in 2018 are not affected by the change in the 2017 assumption.

We are interested in looking at how profit emerges over time. At the end of years 2015 and 2016, there are no changes in the profit pattern, since everything is equal to expected. In 2017, the mortality selection factor increases, so there are larger-than-expected deaths in 2018–2020, larger claims, and lower profit. However, the profit in 2017, as already passed, is not affected at all. In 2018, since we did not modify the actual deaths, the actual death is equal to the inception expectation (not the current best estimated), creating experience variance in that year. However, the total profit in that year (adding up the expected revenue and experience variance) should be similar to that of the inception expectation. The total profit projection for the end of each "actual year" should be similar to that of the inception expectation (e.g., 121,062 vs 118,982 in year 2018); the tiny difference is created by the distortion of the contractual service margin (CSM) amortization schedule with the time value of money.

All the previous statements, while highly abstract and difficult to understand, are illustrated clearly in Figure 4 (and even better on the website with the help of animation).

Besides the scenario described here, the reader may want to think about the following:

1. How might IFRS reserve and profit change if the number of actual deaths deviates from expected?

2. How might IFRS reserve and profit change if mortality selection is increased *after* a particular year? Will the total profit (compared to inception) increase or decrease?
3. How might total comprehensive income change if the discount rate becomes 5 percent *in* or *after* a particular year?

To find the answers to these questions, go to the website, try the interactive illustration, and find out!

Conclusion

Communicating with data visualizations is not just about being more effective by replacing text and numbers. A thoughtfully crafted visualization increases our understanding of the data (Rodriguez and Kaczmarek 2016).

Apart from the example illustrated here, there are many other actuarial domains that can benefit from utilizing data visualization techniques, such as predictive underwriting, strategic asset allocation and stochastic modeling. There is a huge opportunity for actuaries to think about how data visualization may create an impact for our work.

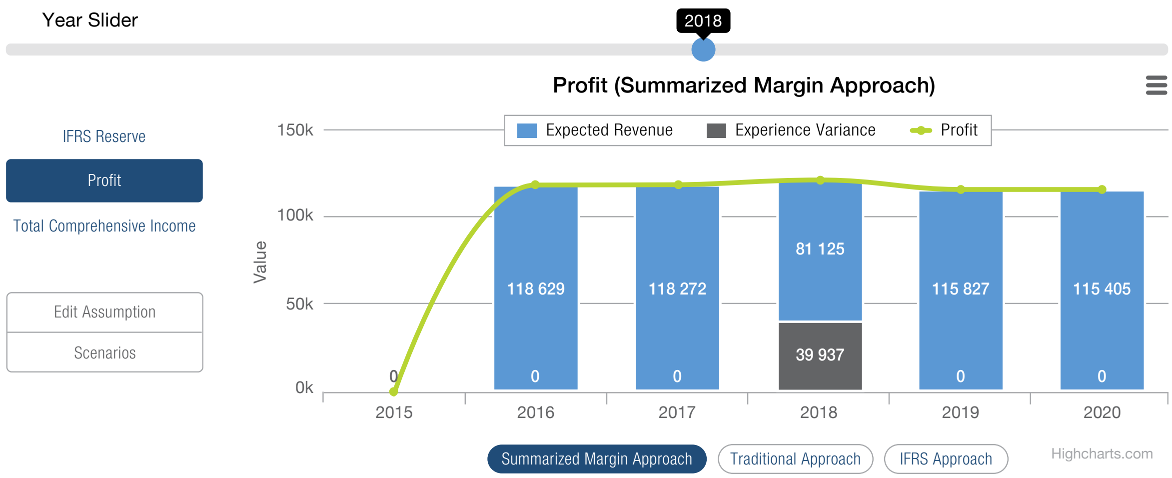
1 Proof is provided in the appendix.

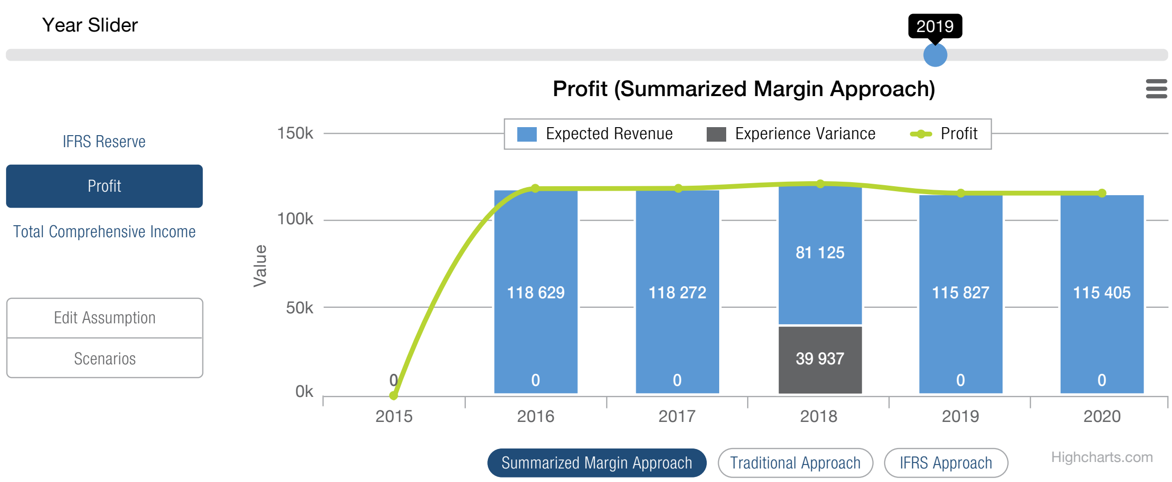
Figure 3 Different Approaches to Present Profit

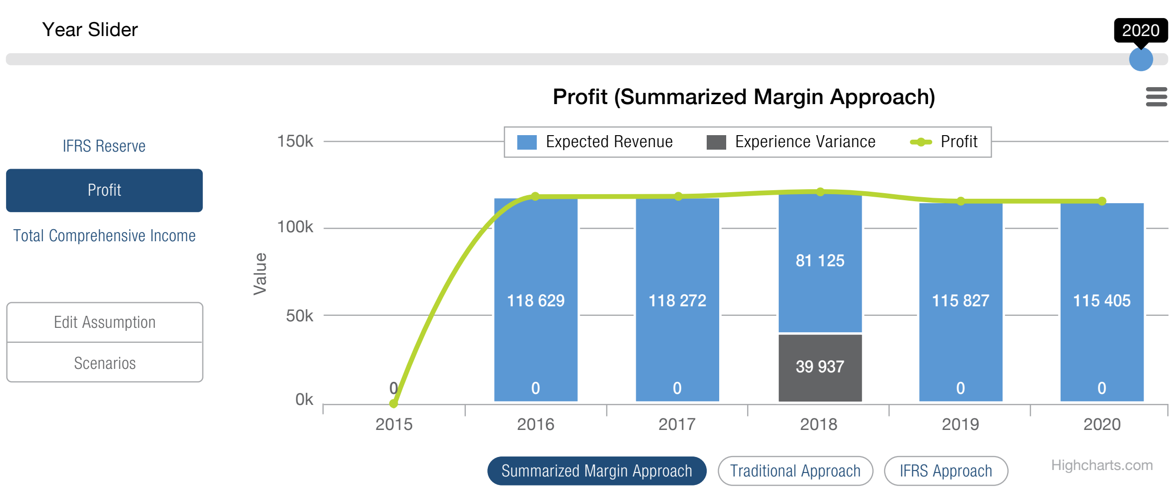
Figure 4 Profit over Time

(Continued on page 8)

(Continued)


Year Slider
2018
Profit (Summarized Margin Approach)
IFRS Reserve
Profit
Total Comprehensive Income
Edit Assumption
Scenarios
Expected Revenue
Experience Variance
Profit
150k
100k
50k
0k
Value
0
118 629
0
118 272
0
81 125
39 937
115 827
0
115 405
0
2015
2016
2017
2018
2019
2020
Summarized Margin Approach
Traditional Approach
IFRS Approach
Highcharts.com


Year Slider
2019
Profit (Summarized Margin Approach)
IFRS Reserve
Profit
Total Comprehensive Income
Edit Assumption
Scenarios
Expected Revenue
Experience Variance
Profit
150k
100k
50k
0k
Value
0
118 629
0
118 272
0
81 125
39 937
115 827
0
115 405
0
2015
2016
2017
2018
2019
2020
Summarized Margin Approach
Traditional Approach
IFRS Approach
Highcharts.com


Year Slider
2020
Profit (Summarized Margin Approach)
IFRS Reserve
Profit
Total Comprehensive Income
Edit Assumption
Scenarios
Expected Revenue
Experience Variance
Profit
150k
100k
50k
0k
Value
0
118 629
0
118 272
0
81 125
39 937
115 827
0
115 405
0
2015
2016
2017
2018
2019
2020
Summarized Margin Approach
Traditional Approach
IFRS Approach
Highcharts.com

Appendix

PROOF OF EQUIVALENCE

SUMMARIZED MARGIN APPROACH

Profit = Expected Revenue + Experience Variance

= (Expected CSM Amortization) − (Delta on Claims + Delta on CSM Amortization)

= (Expected CSM Amortization) − (Actual Death Claims − Expected Death Claims) − (Expected CSM Amortization − Actual CSM Amortization)

= Expected Death Claims + Actual CSM Amortization − Actual Death Claims

IFRS APPROACH

Profit = IFRS Revenue + IFRS Expense

= (Expected Death Claims + Actual CSM Amortization) − (Actual Death Claims)

= Expected Death Claims + Actual CSM Amortization − Actual Death Claims

TRADITIONAL APPROACH

Profit = Net Cash Flow − Increase in IFRS Reserve + Interest on IFRS Reserve

= Premium(t) − Actual Death Claims(t) − (IFRS Reserve(t) − IFRS Reserve(t − 1)) + (IFRS Reserve(t − 1) + Premium(t − 1)) × Discount Rate

= Premium(t) − Actual Death Claims − (BEL(t) + CSM(t)) − (BEL(t − 1) + CSM(t − 1)) × (1 + Discount Rate)

= Premium(t) − Actual Death Claims − (BEL(t) − BEL(t − 1) × (1 + Discount Rate)) − (CSM(t) − CSM(t − 1) × (1 + Discount Rate))

= Premium(t) − Actual Death Claims − (BEL(t) − BEL(t − 1) × (1 + Discount Rate)) − (CSM(t − 1) − Actual CSM Amortization + Interest on CSM − ΔBEL − CSM(t − 1) × (1 + Discount Rate))

= Premium(t) − Actual Death Claims − (BEL(t) − (Actual BEL(t) − Expected BEL(t)) − BEL(t − 1) × (1 + Discount Rate)) + Actual CSM Amortization

= Premium(t) − Actual Death Claims − (Premium(t) − Expected Death Claims) + Actual CSM Amortization

= Expected Death Claims + Actual CSM Amortization − Actual Death Claims

---

References

Aigner, Wolfgang, Silvia Miskch, Heidrun Schumann, and Christian Tominski. 2011. *Visualization of Time-Oriented Data.* London: Springer.

Rodriguez, Julie, and Piotr Kaczmarek. 2016. *Visualizing Financial Data.* Indianapolis, IN: John Wiley & Sons.

Vlaminckx, Dirk, Christophe Vandeweghe, and Bert Lievens. 2013. "IFRS 4 (Phase II) for Insurance Contracts." Presentation at Deloitte seminar, October 9, 2013.

Ware, Colin. 2008. *Visual Thinking for Design.* Burlington, MA: Morgan Kaufmann Publishers.

Jen Houng (Erik) Lie is an actuarial consultant for Deloitte in Hong Kong. He can be reached at *erlie@deloitte.com.hk*.

Effective Communication of Stochastic Model Results

John Hegstrom, FSA, MAAA

One of the challenges facing financial modelers is how best to display a distribution of results. The goal is to convey the necessary information without making the recipient work too hard to understand it. Also, the viewer should be able to come to his or her own conclusion without being led to a specific outcome. Blending art with science helps develop an effective graphical presentation of a distribution of results.

A fixed deferred annuity product with a market value adjustment was chosen as a test case for developing graphical displays of results. The results distribution consists of the present value of distributable earnings at issue from a group of new contracts. One thousand economic scenarios were run to produce the distribution of results. In addition, four product variations were compared.

There are many options for presentation of the results. At one extreme, a point estimate of the mean of the distribution could be communicated as a single number. At the other extreme, a complete table of results by economic scenario could be provided. Clearly, neither of these extremes would be appropriate for decision making. Several alternatives will be explored. The intention is to convey the nature of complex results in a simple-to-understand format.

Understand Your Audience

Understanding your audience is the first step in effective visual communication. Busy executives do not want to be overwhelmed with too much data in a graph or chart. On the other hand, they need enough information to make effective decisions. Technical audiences will want the ability to extract more details. Avoid jargon or acronyms that your audience may not understand. Know in advance if any of your audience is visually impaired with color blindness or other conditions, and make adjustments as necessary.

Types of Graphs

There are a number of approaches to take when presenting graphical results. Following are some of these options, along with a brief description and example of each.

HISTOGRAM

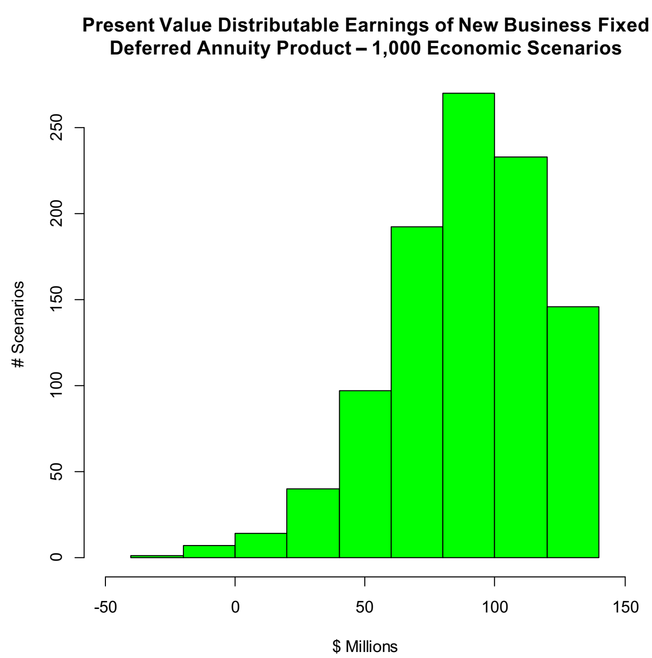
One of the most common approaches to display a distribution of results is a histogram. Figure 1 shows a basic histogram of the results from the base annuity product run.

Note that there are choices of bin size, coloring, scale and annotations, among others. The choice of bin size is important. If the bins are too large, important details about the distribution will be left out. If the bins are too small, the big picture will be obscured by the noise. Colors can have meaning or just aesthetic appeal. Annotations can provide valuable supplemental information to aid the viewer. One of the drawbacks of a histogram is that it is difficult to display multiple distributions side by side effectively.

DENSITY PLOT

Another graph related to the histogram is the density plot. Kernel density estimation (KDE) estimates the probability

Figure 1 Histogram



density function based on sample data. The curve will have the same shape as a histogram but will be a smooth representation of the data. Figure 2 shows a density plot (green line) overlaid on the previous histogram.

The density plot provides a good representation of the data. However, there are some parameters that need to be set, such as the smoothing kernel distribution and bandwidth. Most statistical packages do a good job of selecting optimal parameters, but the data should be compared to a histogram for reasonability. Several density plots can be overlaid using different colors to compare different sets of results. However, this can be confusing if different plots are somewhat similar.

STRIP CHART WITH JITTER

A strip chart can be created showing the distribution of data points. Normally, a strip chart shows data in one dimension. However, by adding jitter, which is controlled random noise, it becomes easier to see the distributional pattern. In addition, color can be added to the data points to provide valuable information, as in Figure 3.

In this strip chart, the data points are colored based on the relative level of interest rates over the modeling period. This provides valuable information to the viewer on an important factor that is driving results. The jittering takes place along the *x*-axis, providing some space between points. Strip charts can be shown with comparable data sets side by side, if desired.

VIOLIN PLOT

Violin plots are basically density plots shown in a symmetric fashion, allowing side-by-side visualization of comparable data. In addition, they typically show the quartiles and median of the data. The modified violin plot in Figure 4 shows the results of several variations of the product and the corresponding results distributions.

This violin plot has been modified from the standard violin plot template in a few ways. First, the mean is shown instead of the median, as the mean of the distribution has special significance as a probability-weighted expected value, assuming that the economic scenarios are equally probable. Second, only the middle 50 percent of the data is indicated by the dark blue shapes; the standard whiskers have been removed for the sake of clarity. Third, the 99th percentile of the data is indicated by the red symbol and a label. This risk measure could easily be set to other measures, such as a conditional tail expectation (CTE) level or a value at risk (VaR) amount.

Figure 2 Plot

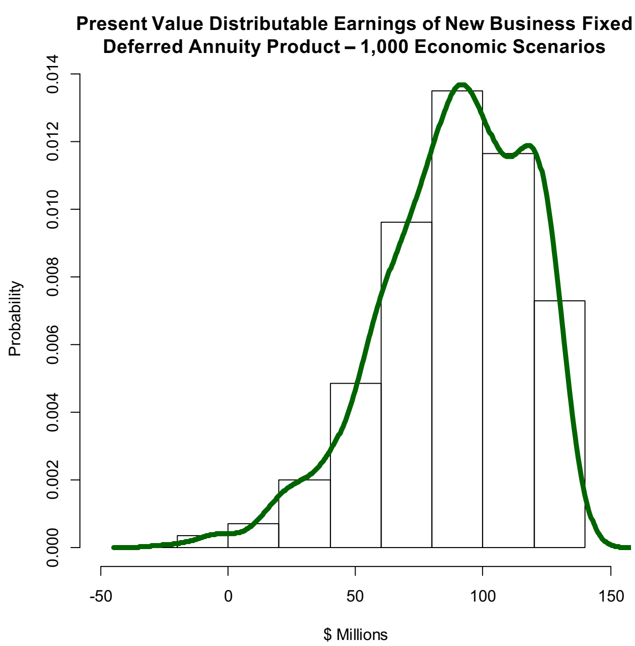


Figure 3 Strip Chart with Jitter

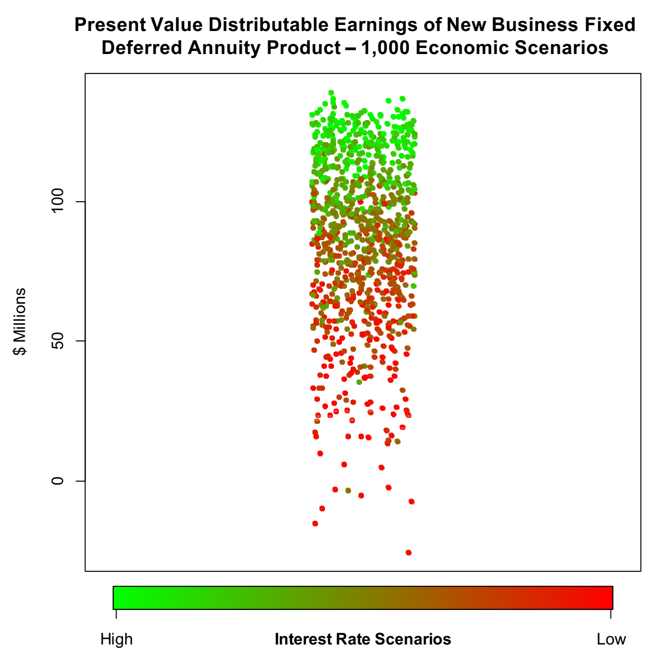
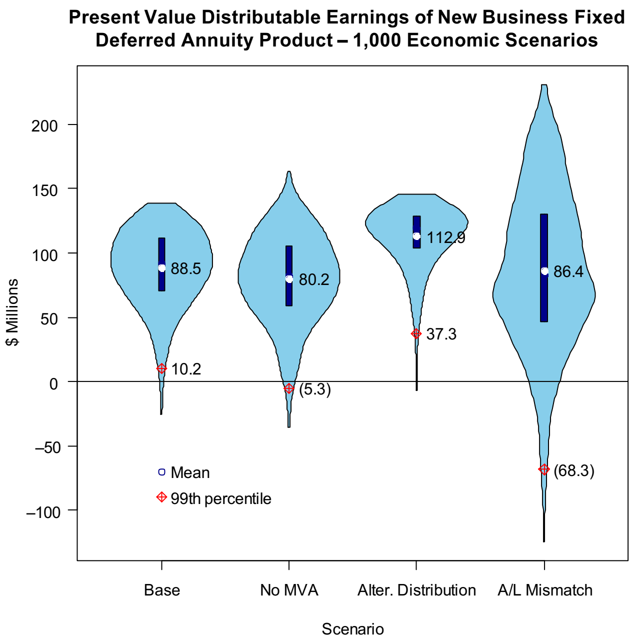


Figure 4 Violin Plot



This modified violin plot has the advantage of being able to easily display results side by side, show the shape of the results distribution and show the key metrics as well. It is a valuable chart for presenting results to decision makers. Figure 4 leads to the conclusion that the alternative distribution strategy (lower excess lapse rates) offers a superior risk/reward profile and should be pursued if viable. Another conclusion is that stretching the asset duration to pick up yield does not pay off in this instance and increases risk.

ASOP 41

Graphs can be considered a part of actuarial communications and a component of an actuarial report under Actuarial Standard of Practice (ASOP) 41. It is necessary to follow ASOP 41 when creating, presenting and distributing graphs, if it applies. ASOP 41 lays down requirements as to form, content and disclosures. Proper disclosure of methodology and assumptions is required. It is important to disclose that the distributions shown are just estimates and don't encompass all possibilities.

Conclusion

With today's computing power, large numbers of scenarios can be run. The resulting avalanche of results data needs to be analyzed and communicated in an efficient and effective manner. There are many powerful tools, such as violin plots and strip charts (with jitter), that can be successfully used to analyze and communicate results.

The visual displays in this article were created using the freely available R software package. The annuity product used in the examples resembles real-world products but is fictitious.

John Hegstrom, FSA, MAA, is a consulting actuary with Actuarial Resources Corporation. He can be reached at *john.hegstrom@arcval.com.*

Visualization of Social Network Data

Kailan Shang, FSA, CFA, PRM, SCJP

Traditional data such as numerical data, categorical data and ordinal data are all well-organized data that can be measured on the same standard. The characteristics of traditional data can be summarized by descriptive statistics such as subgroup size, mean, median, standard deviation and correlation. They can also be presented using graphs such as a line chart, bar chart, pie chart, histogram and scatter plot. On the other hand, nontraditional data such as social network data are difficult to measure and summarize. The useful information of each data record varies case by case and usually cannot be represented by descriptive statistics. Nontraditional methods of data visualization are needed to effectively communicate the key information of the data. Some nonnumerical data-sets such as social network data are of large volume. Effective data visualization can also help us explore the data set for pattern recognition, explanation and better understanding.

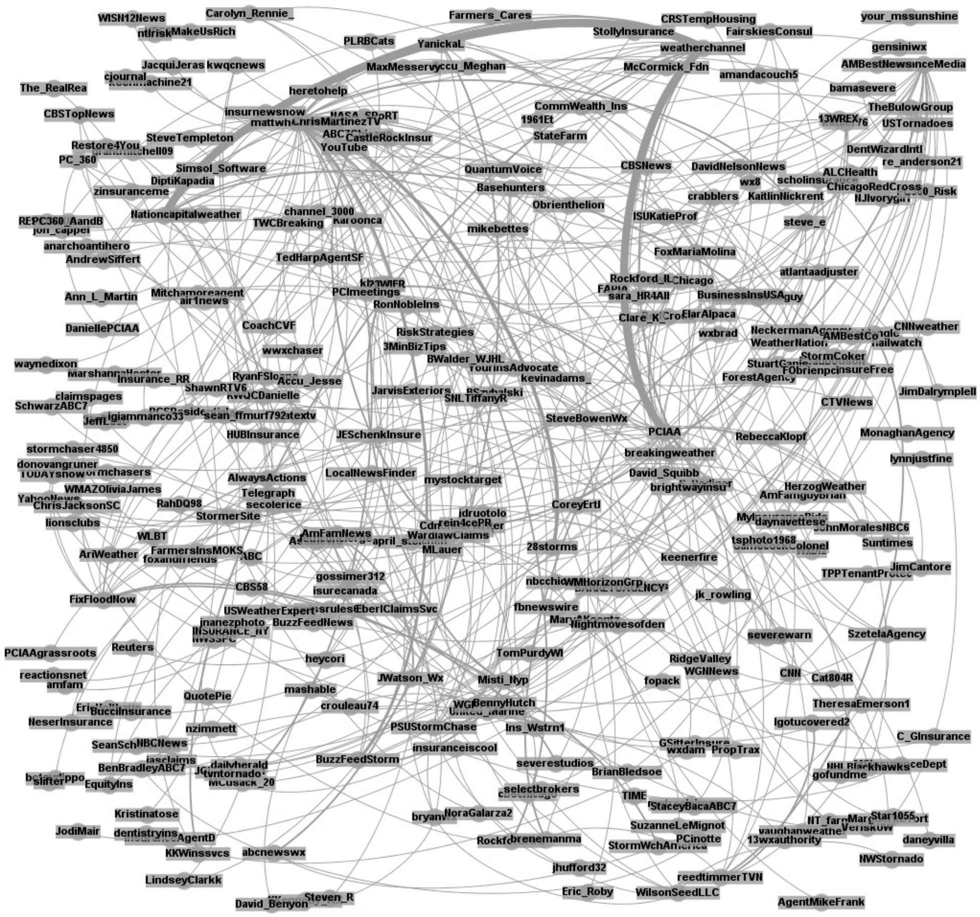
Data

In a recent Society of Actuaries (SOA) research project, Twitter data of several severe tornado events were studied to measure the social media response to catastrophic events. This type of data provides a unique perspective of monitoring and managing catastrophic risk. Social networks can help disseminate warnings and tips to reduce losses and expedite after-event recovery. With fast and wide spread of information, social networks can be used for effective loss control.

Each Twitter data record contains the content of the tweet and accompanying metadata such as user name, user bio, count of friends, count of followers, count of favorites and post time. The relevant Twitter data search for the Rochelle tornado outbreak that happened on April 9, 2015, contains about 95,000 records. Given the diversity of the information in the tweets, traditional data analysis and visualization methods are not very helpful for exploring the data and discovering patterns. Text-mining models and different data visualization methods are needed to summarize and present the information in social network data.

Figure 1 Word Cloud Example

Figure 2 Social Network



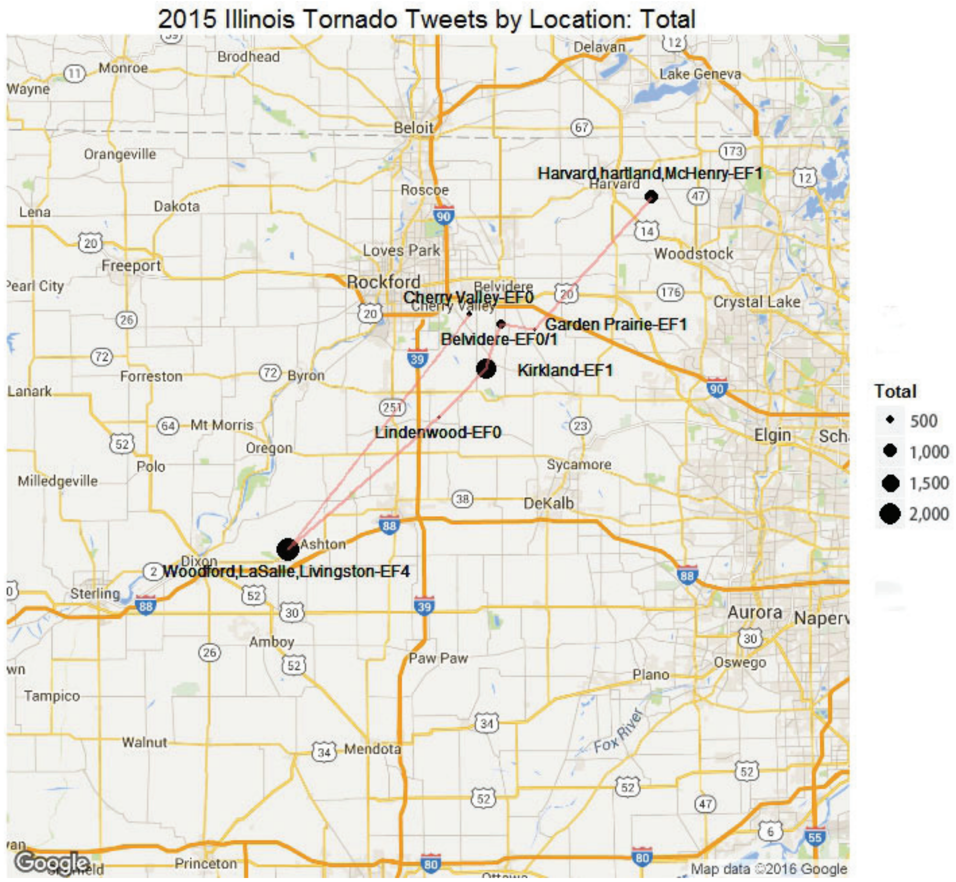
Word Cloud

Unlike numerical data, the key information of social network data cannot be presented using the mean, the median, a histogram and so on. One way to present key information is through the use of a word cloud. A word cloud contains the most popular words based on the frequency of appearance in the Twitter data. It can clearly show the most important topics in the data set. Figure 1 shows an example of a word cloud using the Twitter tornado data. The size of each word is determined by the frequency of the word in the data set—the higher the frequency, the larger the size of the word. The keywords include users mentioned in the tweets, the location of the tornado, the time of the tornado outbreak, the impact of the tornado and so on. There are several tools available to create word clouds, including open source statistical software R with package "wordcloud" and online tool Wordle (http://www.wordle.net).

When constructing a word cloud, consider the following helpful points for effective communication:

- The keywords cannot be too many, so that most words are legible.
- The word size is usually set according to the frequency of the word in the data set. However, sometimes one word is dominating, which makes other popular words too small to recognize. Other scaling methods, such as taking the logarithm of the word frequency, can be used.
- Some common words such as "is," "at," "who" and "that" are not useful for presenting the key information of the data. These stop words need to be removed from the word cloud.
- Some words have similar meaning. For example, the words "rains," "raining," and "rained" are

Figure 3 Twitter Data Geolocation Analysis



Note: EF is the Enhanced Fujita scale of a tornado.

all derived from the root word "rain." These morphological affixes need to be removed before counting the frequency of words so that the final list contains words with different meanings. Words with similar meanings can be aggregated to avoid redundancy.

Social Network

Network analysis is important to understand the key influencers in the network and how information is spread over the network. Figure 2 shows the network derived from insurance-related Twitter tornado data. It includes all the users having six or more interactions with others in the data set. Visualization of the network can help us identify the centers of the network (nodes with many edges connected to other users) and the most frequent interactions between users (thick edges). Figure 2 was made using Gephi, an open source graph visualization platform.

It is important that network visualization clearly highlight the key information that readers are looking for. Following are some guidelines:

- The size of the node or edge needs to vary by its significance in the network.
- Proximity usually indicates the closeness of users. Instead of using the thickness of edges to show the degree of interaction, users with frequent interactions can be presented by the proximity of users in the graph. Using proximity, it is easier to show subgroups with high connection within their groups but low connection with other groups.

With the identified key influencers and most frequent interactions, the information flow of the network can be understood. It can also help narrow down the scope of the data that need further detailed study.

Geolocation

Geolocation analysis of social network data can help measure the impact of catastrophic events. A high data volume at a location means a high-impact event caused by high severity, high population density, or both. Adding data volume and event severity on a geolocation map can indicate whether a high correlation between the data volume and the event severity exists or not. Figure 3 shows the volume of Twitter tornado data and severity of tornado by location. Each dot represents a recorded tornado event, with the location and severity information in the data label. The dot size is proportional to the data volume. In general, high severity leads to high Twitter data volume. Figure 3 was made using open source statistical software R with package "ggmap," which allows users to load certain online maps and add data points at specified locations defined by longitude and latitude.

Some nontraditional data visualization methods such as network graphs and geolocation graphs can be combined together to show the network connectivity among locations.

Conclusion

Social network data are difficult to analyze using traditional statistical measures. Well-designed data visualization can not only help deliver the key information of social network data effectively but also facilitate the discovery of patterns in the data. The visualization of social network data needs to be intuitively easy to understand because the delivered information is usually more qualitative than quantitative. The size of the objects in the graph can be used to reflect the weight of the data.

Kailan Shang, FSA, CFA, PRM, SCJP, is Managing Director of Swin Solutions, Inc. He can be reached at *kailan.shang@swinsolutions.com.*