A Small Data Approach to Risk Analysis: Applying the Pareto Principle to Risk Metrics

Andrew Sheves
andrew@dcdr.io
https://research.dcdr.io

Abstract

Organizations have become increasingly complex as systems become more sophisticated and intertwined. This complexity makes threat identification and risk analysis increasingly difficult. Machine learning (ML) and artificial intelligence (AI) would seem to offer a solution to problems of scale and complexity but these systems cannot access much of the dialogue or thought that contributes to a decision, meaning the data needed to create effective models is limited. Moreover, months or even years must pass before a strategic decision can be assessed as 'right' or not, meaning that decades may be required for ML or AI to create useful models. However, a 'small data' approach utilizing the Pareto Principle could focus on the 20% of factors that have the biggest impact on organizations. Tracking +/- a dozen critical metrics and placing these into a historic context alongside a directional trend should allow risk managers and decision-makers to focus their attention on the events that are likely to have the biggest impact on their organizations and respond accordingly. A small data approach will speed up and simplify the risk analysis process, free up resources, and avoid creating the impression of greater understanding than there actually is. This paper explains how such a system could work.

The problem: Things are Complex and Noisy

In 1984, Charles Perrow introduced the idea of tightly coupled systems: situations where the complexity and interoperability of systems were so great, that humans would be unable to prevent many failures[1]. In the intervening period, systems have become even more complex, making risk analysis increasingly difficult. The speed with which data is created increases the amount of information that may be relevant considerably and adds significant noise to any analysis. Moreover, as organizations become more sophisticated, there is a tendency to try to track more and more data points with the idea that more is better. Unfortunately, this increases both the amount of data to process and the amount of signal to track. The outcome is that vital signal is not only drowned out by background noise, but also masked by other signals[2].

[1] *Normal Accident*, Charles Perrow
[2] *The Signal and the Noise*, Nate Silver

Why AI isn't The Answer (Yet)

Unfortunately, despite their benefits in handling large data sets, current machine learning (ML) algorithms and artificial intelligence (AI) systems seem ill-suited to risk analysis and decision-making for two reasons.

> First, a high degree of nuance is required to understand complex decisions. This requires more extensive data sets compared to, for example, those needed to differentiate between cats and dogs. Moreover, as most human decision-making is conducted mentally or in discussions, the way decisions are reached is not in a format easily accessible by machines. Technically, this is an easily solvable problem, but it will require people to change how they record their decision-making and time to compile and analyze a sufficiently large amount of data.
>
> Second, the results of many business decisions are not apparent for months or years, meaning that the feedback loop for any decision-making system is painfully long[3]. This will further extend the delay before AI can supplement decision-making effectively.

However, even when these problems are solved, a big-data solution to risk analysis will always suffer from a GIGO problem: in this case, the quality of the answer relies upon the quality of the question. Therefore, efficiency and friction will always be a factor until we get better at asking the right question and can spend the necessary time asking many variations on the same question until we get a sense of the 'best' answer.

Therefore, while ML and AI will be able to assist our decision-making in the future, these technologies will not be satisfactory for some time. However, there is an opportunity to use specific indicators to track risk metrics which, in turn, can be used for risk analysis and decision-making. We can call this the small data approach and use the Pareto Principle as a foundation for the model.

Applying the Pareto Principle to Risk Analysis

In the late 1800s, Italian economist Vilfredo Pareto identified that 80% of Italy's land was owned by 20% of the population, giving rise to the Pareto principle or 80/20 rule[4]. Although the exact split might not be 80/20, this type of power law distribution can be found everywhere, from the Gini index of wealth distribution to book sales.

Applying this concept to risk analysis would mean that instead of trying to keep track of a vast array of data points, with all the associated issues outlined above, monitoring the most significant 20% should help identify the most significant changes in the threat landscape.

This approach isn't perfect: there would be a long tail of lesser threats that are not tracked as carefully. Moreover, genuine Black Swans (evens for which we have no historic context[5]) can still catch us off guard. However, the effectiveness of a properly-applied 80/20 distribution coupled with the efficiency of tracking 1/5 of the amount of data should still produce a net-positive effect as:

- Appropriate attention is being paid to the most consequential factors.
- The time spent on routine risk analysis is reduced significantly, freeing up resources to examine long-tail threats and plan for unforeseen events.

---

[3] *Superforcasting*, Philip Tetlock
[4] Vilfredo Pareto biography, Wikipedia
[5] *The Black Swan*, Nicholas Nassem Taleb

- There's a recognition that there are 'known unknowns' which might affect the organization. Conversely, a system that gives the appearance of tracking everything creates the false impression that there are no unknowns.

Therefore, the benefits of an 80/20 approach should outweigh the downsides while also overcoming many of the previously mentioned problems of trying to use big data for risk analysis and decision-making. This approach would be particularly beneficial for smaller organizations that could not afford an ML / AI solution, even if one were available.

The Intent

The remainder of this paper describes a system that uses a small data approach for risk analysis and decision-making. The specific intent is to develop a simple system for tracking risk metrics, specifically threat indicators, to speed up and simplify risk-based decision-making. To be assessed as successful, the system should:

- Place results in an ordered scale where the relative severity of each factor is clear.
- Improve focus by reducing the number of factors to track.
- Narrowing the range of considerations.
- Produce updated results for the relevant factors within the decision-making timeframe.
- Share information in a clear, easy-to-understand format.

What are the Characteristics of a Key Metric?

As with much in life, the devil lies in the details: specifically, what are the key metrics? There is wide latitude for these to become highly subjective and tailored to a particular industry or sector, but this can be mitigated by adding some strict constraints. The initial constraints imposed are as follows. To be useful, the metric has to be:

- Broad, not narrow (meaning the metric has widespread effects)
- Publicly available
- Easily understandable
- Updated frequently
- Commonly used

The first point is most important as we are looking for macro indicators: measurements that have widespread effects. A good example is the price of crude oil which, for example, affects transportation, heating, the cost of shipped goods, and even the velocity of transition to renewables. Similarly, the price of grain has a direct effect on the cost of food and the number of people who have a calorie deficiency, but it also affects the balance of payments of countries that rely on grain imports and can lead to instability where bread is subsidized.

Additionally, there are single measurements that can act as analogs for a range of factors. Vaclav Smil cites childhood mortality as a broad reflection of the quality of life in a country as factors like availability of healthcare and income levels which contribute to, or detract from, infant survival rates, also contribute to the quality of life[6].

[6] *Numbers Don't Lie*, Vaclav Smil

What Are The Key Metrics?

Based on the criteria above, an initial set of 12 metrics have been identified that should provide an effective foundation for this model. These have been separated into three categories: market, economic, and social.

The initial set of metrics and sources is shown below. This list may be adapted over time based on utility and user feedback.

Market

- Shipping (Freightos FBX container index)
- Wheat (Chicago Board of Trade (CBOT))
- Global Uncertainty (St Louis Fed (FRED))
- Iron and Steel (Google Finance Iron and Steel index)
- Oil (Brent Crude future)

Economic

- GDP growth (International Monetary Fund (IMF))
- Uncertainty (FRED)
- Inflation (World Bank)

Social

- Freedom and democracy (Freedom House)
- Conflict (Heiselberg)
- Development (UN Sustainable Development Goals)
- Connectivity (Internet usage - The World Bank)

Tracking the Metrics

One immediate problem is that the metrics are measured on very different scales and the reporting frequency of each differs significantly. However, because the requirement is for a system to indicate what has changed and by how much, we can say that the system has to reflect:

- What 'normal' looks like
- If a metric deviates from normal
- If it deviates, by how much and in which direction? (i.e., A threat vector.)
- Any broad trends reflected in the data.

Therefore, absolute values are less important than showing relative change for each metric, which can be demonstrated by a percentage change. So, for each metric, we need to collect historic data for the relative time interval and know the current value. From there, we can compare the current value relative to historic measurements and identify any short-term trends.

Measurement Intervals

In this system, 'historic' and 'short-term' are relative to the tracked metric. However, differences in reporting timescales can be overcome by the general assumption that the frequency of updates correlates with how quickly changes are felt. For example, daily changes in the price of crude oil are

felt by consumers at the pump within days. Meanwhile, societal changes aren't apparent for years, roughly corresponding to the interval between major surveys. This is not a perfect correlation but the fundamental assumption that reporting frequency and the timeline over which a change is felt appears sound. This assumption will be examined over time and timescales will be adjusted as necessary.

This results in the following measurement intervals.

Market

- Measured - Weekly (averages)
- Trend interval - 21 days
- Comparison interval - 90 days

Economic

- Measured - Monthly or quarterly
- Trend interval - Three months or three consecutive quarters
- Comparison interval - One year

Social

- Measured - Annually
- Trend interval - 3 years
- Comparison interval - Decade-to-decade

Year-on-year comparisons may also be included in longer-term analysis to help identify seasonal patterns or deviations.

Reflecting the Relative Value

Relative value shows the current price or value for a metric compared to the low and high values for the comparison interval. This measurement allows us to put a value into context and say whether it is relatively high or low for the period.

Calculations are based on the current value expressed as a percentage of the difference between the highest and lowest values for the comparison period. Percentage bands are used to describe the current value, relative to the comparison interval.

- 80 - 100 Very High
- 60 - 79 High
- 40 - 59 Mid-range
- 20 - 39 Low
- 1 - 19 Very low

This framework allows us to place a value into context and describe it in natural language consistently, e.g., "*wheat prices are very high at the moment*". This removes subjectivity from discussions and allows decision-makers to use standardized relative terms to describe the metrics relevant to their organizations during risk analysis and decision-making discussions.

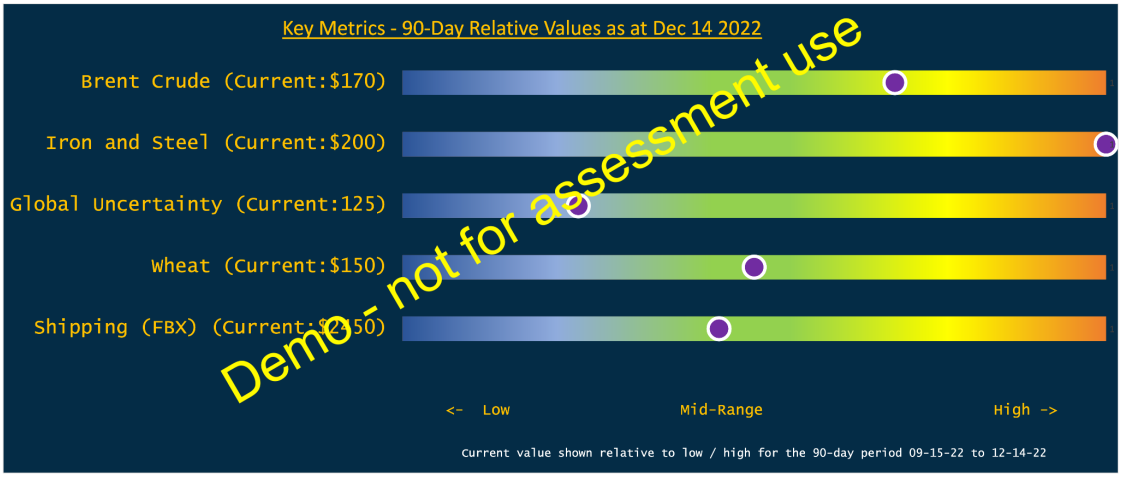

Fig 1 - Example of a relative values graph

Reflecting Trends

The change vector for each metric helps decision-makers see the direction and magnitude of charge for a factor over the trend interval. This will allow users to broadly determine if the resultant risk(s) is / are increasing or decreasing and plan accordingly.

The trend calculation is based on the percentage change from the start to the end of the trend interval for each metric. Bands are used to describe the current value, relative to the comparison interval.

- +10% or more — Sharp increase
- +2 to 9.9% — Moderate Increase
- -1.9 to +1.9% — Relatively stable
- -9.9 to -2% — Moderate Decreasing / dropped
- -10% or more — Sharp decrease

Similar to the relative value, this measurement indicates a movement trend for that metric over the interval period (e.g., 21 days). Therefore, it is a relative measurement and should not be used for detailed calculations or forecasting.

For example, if the price of steel increases by 10% in an interval, a reasonable use of the data is to assess that the cost of raw materials will likely increase in the short term so unfinished or planned construction projects will require additional capitalization. However, it would be wrong to project a 10% increase in steel costs each month and to use that for budget planning.

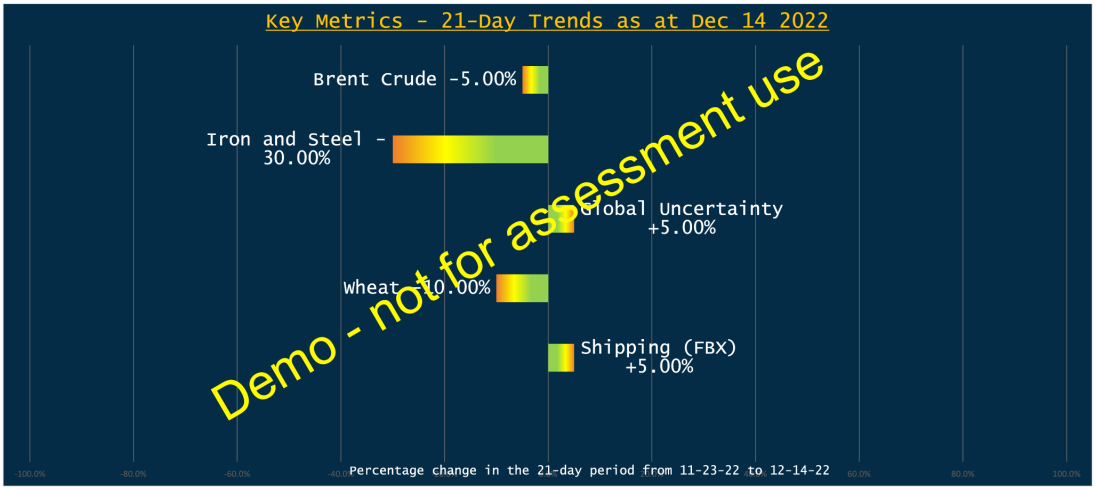

Fig 2 - Example of a trends graph

Possible Improvements to the Trend Vector

There is room for improvement with the trend vector, but this will not become clear until several iterations of the model are run and feedback is received. However, possible changes to the trend vector are:

- Adding additional bands to further 'shade' the increase / decreasing descriptions.
- Reflecting the percentage change as a value between +90 and -90 degrees so the change can be shown as a directional arrow. This could be achieved using the method for converting the degree of a slope into a % grade (Degrees = Tan-1(.50 percentage)), but this measurement produces an exponential result which may be misleading.
- Instead of assessing the change based on the start and end measurements for the interval, there may be more effective measurements that reflect change. For example, comparing the interval change to the average change in the two preceding intervals.

Roll Out Roadmap

The value of these metrics is contextual: how does the current value relate to recent highs and lows and in what direction is it trending? This context will improve as more comparative data is available. Feedback can be accelerated by creating retrospective reports but the true value of the model will only be apparent once its projections can be used in real-world situations and the outcomes evaluated.

The shorter intervals for market measurements mean that there will be significantly more opportunities to assess the model's effectiveness using this data. Therefore, the roll-out of the model will be as follows:

Phase 1 - By Dec 31, 2022

- Semi-weekly market reports (Monday and Thursday)
- Monthly market reviews

Phase 2 - By Jan 30, 2023

- Inclusion of economic data once per week
- Economic reviews added to monthly reports

Phase 3 - By Feb 28, 2023

- Inclusion of societal data in monthly reports

Phase 4 - By Mar 31, 2023

- Quarterly reviews begin
- Sensitivity indexing beta (create a 'live' risk profile for a user's operations based on the threat level and relative impact)