DAILY PROGRAMMATIC INDEX

Daniel Karcz
Data Scientist
This presentation posits a hypothetical scenario between a Cost Estimator and Program Manager overseeing a cloud computation project at a company XYZ.

We will apply rigorous statistical methods on daily data, in order to provide proactive indications of cost over & under runs for a cloud computing project at corporation XYZ.

The proposed Daily Programmatic Index (DPI) value is an early warning system that will identify potentially significant cost swings.
Real time index pulses are a value add for cost estimators, risk managers, program managers, and earned value management practitioners.

Real time index pulses can be derived by:

1. Identification of programmatic cost drivers
2. Data dimensionality reduction via statistical techniques
3. Development of cost and risk indices
   - Overall supply chain index
   - Sector level supply chain indices
4. Holistic program management dashboards
   - Descriptive analytics - historical data
   - Predictive analytics - anticipated costs
WHY HIGH FREQUENCY DATA FOR COST ESTIMATION?

WHAT IS THE VALUE OF STATISTICAL METHODS?

1. Active monitoring of costs “vital signs”

2. Warning systems – cost problems

3. Data dimensionality reduction – more interpretable without loss of information
**END OF MONTH COST REPORTS**

- Monthly and quarterly reports
  - e.g. Contractor pay
  - e.g. Cloud computation usage
- Time consuming
- Time lag for identifying cost variances
- Not always standardized
- Subjective – error prone

**DAILY PROGRAMMATIC INDEX (DPI) COST PULSE**

*Warning!* cost shocks began on May 2\textsuperscript{nd} and continued

DPI would have provided an early warning system. Program manager and cost estimator could have taken action to manage accelerating cloud costs which commenced on May 2\textsuperscript{nd}. Instead of being surprised by end of month cost reports (i.e. May 31\textsuperscript{st}).
INDICES CAN BE SPECIFIC OR HOLISTIC FOR THE SUPPLY CHAIN

An index can be built upon one of the supply chain stages:

1. Raw materials index (input ex: production rates of mined copper, steel, silver, timber, corn, etc.)
2. Manufacturer index (input ex: production rates of widgets, prices, etc.)
3. Distributer index (input ex: cost of fuel, cost of labor, delivery time, etc.)

Or an index can be based upon the entire supply chain of one industry. Example: gold industry.
- Ore extracted from mine per week
- Grams of gold per ton of ore per week
- Cost of refining gold per gram per week
- Cost of securely storing and shipping gold per week
- Spot market price of gold per week
- Retail jewelry customer purchase pattern behavior per week
Federal Reserve Economic Data (FRED) is a database maintained by the Research division of the Federal Reserve Bank of St. Louis that has more than 765,000 economic time series from 96 sources.

- Banking & Interest Rates & Reserves
- Consumer Price Index (CPI)
- Gross Domestic Product (GDP)
- Monetary Aggregates
- Producer Prices Indexes
- U.S. Trade and International Transactions

HOW CAN THESE INDEX METHODS BE APPLIED TO PROGRAM MANAGEMENT?
The WEI is an index of real economic activity using timely and relevant high-frequency data. It represents the common component of ten different daily and weekly series covering consumer behavior, the labor market, and production.

1. Redbook same-store sales
2. Rasmussen Consumer Index
3. New claims for unemployment insurance
4. Continued unemployment insurance claims
5. Adjusted income/employment tax withholdings (from Booth Financial Consulting)
6. Railroad traffic originated (Association of American Railroads)
7. American Staffing Association Staffing Index
8. Steel production
9. Diesel, gasoline, and jet fuel
10. Average US electricity load
ST. LOUIS FED FINANCIAL STRESS INDEX

Measures the degree of financial stress in the markets and is constructed from 18 weekly data series, all of which are weekly averages of daily data series: seven interest rates, six yield spreads, and five other indicators. Each of these variables captures some aspect of financial stress.

How to Interpret the Index:
The average value of the index, which begins in late 1993, is designed to be zero. Thus, zero is viewed as representing normal financial market conditions. Values below zero suggest below-average financial market stress, while values above zero suggest above-average financial market stress.
<table>
<thead>
<tr>
<th>Date</th>
<th>Type</th>
<th>Function</th>
<th>Unit</th>
<th>Price</th>
<th>Usage</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/3/2019</td>
<td>computation</td>
<td>active storage</td>
<td>gigabyte</td>
<td>$0.02</td>
<td>2982</td>
<td>$59.64</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>computation</td>
<td>BigQuery Storage Write API</td>
<td>gigabyte</td>
<td>$0.03</td>
<td>3323</td>
<td>$88.08</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>computation</td>
<td>Long-term storage</td>
<td>gigabyte</td>
<td>$0.01</td>
<td>37140</td>
<td>$371.40</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>computation</td>
<td>Streaming inserts</td>
<td>megabyte</td>
<td>$0.01</td>
<td>8318</td>
<td>$83.18</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>computation</td>
<td>BigQuery Storage Read API</td>
<td>terabyte</td>
<td>$1.10</td>
<td>3459</td>
<td>$3,804.50</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>computation</td>
<td>queries</td>
<td>terabyte</td>
<td>$5.00</td>
<td>26.8</td>
<td>$144.00</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Data Scientists</td>
<td>hourly rate</td>
<td>$57.30</td>
<td>20</td>
<td>$1,146.00</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Database Administrators</td>
<td>hourly rate</td>
<td>$47.12</td>
<td>20</td>
<td>$942.40</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Developers</td>
<td>hourly rate</td>
<td>$64.50</td>
<td>20</td>
<td>$1,290.00</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Engineer</td>
<td>hourly rate</td>
<td>$95.80</td>
<td>20</td>
<td>$1,916.00</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Data Scientists OT</td>
<td>overtime hourly rate</td>
<td>$85.95</td>
<td>0</td>
<td>$ -</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Database Administrators OT</td>
<td>overtime hourly rate</td>
<td>$70.68</td>
<td>0</td>
<td>$ -</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Developers OT</td>
<td>overtime hourly rate</td>
<td>$96.75</td>
<td>10</td>
<td>$967.50</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>labor</td>
<td>Engineer OT</td>
<td>overtime hourly rate</td>
<td>$143.70</td>
<td>0</td>
<td>$ -</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>Training</td>
<td>personal development session</td>
<td>rate per day per employee</td>
<td>$250.00</td>
<td>0</td>
<td>$ -</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>utility</td>
<td>Electricity</td>
<td>KwH</td>
<td>$0.84</td>
<td>6940</td>
<td>$5,829.60</td>
</tr>
<tr>
<td>2/3/2019</td>
<td>utility</td>
<td>Space Rent</td>
<td>sq footage</td>
<td>$18.00</td>
<td>1</td>
<td>$18.00</td>
</tr>
</tbody>
</table>
1. Control Charts
   • Visualize whether data collected over time is stationary (i.e. “stable”) or if there are signals that could be out of the ordinary (i.e. “shocks”).

2. Principal Components Analysis (PCA)
   • With increasing data availability data sets are often large and difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability while minimizing information loss.
METHOD 1: CONTROL CHARTS — VISUAL EXAMPLE

- Outlier (i.e. “shocks”)
- Average (i.e. “stable”)
Query Storage Read API costs exploded in late February and early March due to subsidiary corporate XYZ and ABC mergers, and aggressive cloud testing.

Control charts would successfully alert a Program Manager or Cost Estimator that a specific cost is burning hot, and a cost valve must be turned to cool large deviations in cost.

Notice the peak costs are above 3 standard deviations away from the average.
CONTROL CHARTS: ELECTRICITY & CLOUD COMPUTING

- The Average Electricity Rate in the U.S. is **10.42 cents per kilowatt-hour** (2021)

- **ENRON** – The 2000 California electricity crisis (Western U.S. energy crisis) was a situation in which the state of California had a shortage of electricity supply caused by market manipulations and capped retail electricity prices. Delays in approval of new power plants, and market manipulation decreased supply. This caused an **800% increase** in wholesale prices from April 2000 to December 2000.

DPI monitoring would have served as an early warning system!

WHAT CAUSED THE COST SPIKES IN APRIL/MAY OF 2021?

WARNING SIGNS?

“XYZ” MERGED ALL SUBSIDIARY COMPANIES, WHICH INCREASED CLOUD USER BASE BY 7X.

![Computational Costs (actuals) graph]

- Active Storage
- Queries
- BigQuery Storage Read API
- BigQuery Storage Write API
- Long-term storage
- Streaming inserts
Principal Components Analysis

Step 1: Standardize all data inputs to have same scale.

Step 2: Covariance Matrix Computation.

Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.
Step 3 continued: PCA puts maximum possible information in the first component, then maximum remaining information in the second and shown on (scree plot).

Trade off dimension reduction vs. interpretability

Organizing information in principal components this way, will allow you to reduce dimensionality without losing much information. Important! Principal components are less interpretable, when they are constructed as linear combinations of the initial variables.

Step 4: Recast the data along the Principal Component Axis.
Step 1: Monitor DPI principal component

• Start from 40,000 foot view (PCA component 1 shock).

• Then observe each feature that was integrated into PCA. Did one cost driver change or a combination of cost drivers? Use control charts to observe cost shock deviations.
Step 2: if DPI identifies potentially significant cost swing during the day, then visualize control charts for each feature to identify cost driver’s initiating swings.
Step 3: Program manager and cost estimator take proactive action to get ahead of budgetary contingency plans. Hypothetical examples below:

1) Overtime: employee biometric work sign in enabled, so overtime hours are tracked in source system of record at end of day.
   ✓ Action: temporarily suspend overtime pay until new cost estimation is developed and approved.

2) Cloud usage 7X increase
   ✓ Action: implement IT chargeback model to hold internal organizations accountable for computational costs.

3) Cloud computation increase
   ✓ Action: limited cloud usage per organization on a weekly basis.
Academic and Public Sources

