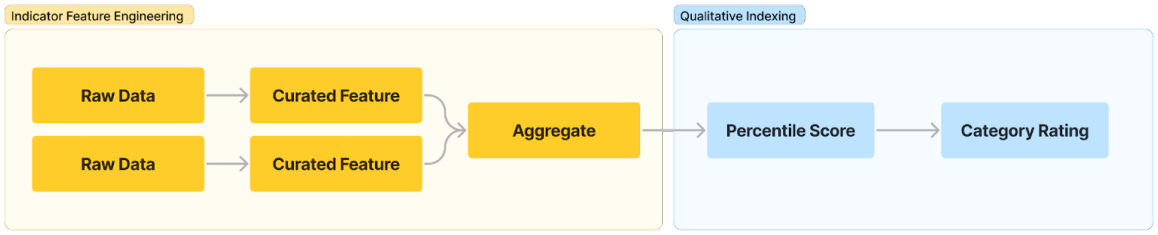
CA Indicator Processing Manual

Last Updated: July 31th, 2023

By Gigi Sung

This document serves as an internal guide designed to assist researchers in the development of an indicator and its corresponding index. To ensure consistency, we will first establish relevant terminology and the overall methodology of indexing before delving into the process of individual signal construction.



Each CA indicator serves two purposes. The first is to extract valuable information from raw data into different curated features which enables additional downstream models such as Climate Price. The second is to convert the features into percentile ranking and a category rating for qualitative understanding. The above diagram explains the flow of data in a typical indicator processing pipeline.

Raw Data: Raw data are data obtained from the sources with little or no transformation. These raw data are extracted from their sources into the CA data lake and act as input to the overall process.

Curated Feature: The curated features are specifically engineered from raw data. They might be obtained through simple processing of the raw data (e.g., population with outliers removed), or are created from an aggregate of different raw data streams (e.g., population density which is calculated using population divided by area). These features are the most statistically significant outputs of each indicator and will be used in our downstream ML models. Therefore, it is crucial to make sure that the output features are error and outlier free.

Aggregate: Aggregate values are a single feature used to describe all underlying features. This value is calculated from a specifically designed formula for each indicator. It captures the aggregated distribution of the underlying raw features and thus can also be used for downstream machine learning and statistical analysis. The most used technique will be to add each underlying feature together after standardizing them to the same scale. It is important to note that while summing the features, the distribution of each feature should be as normally distributed as possible by applying the appropriate transformation technique.

Percentile score: The percentile scores are relative measures, representing a location's comparative ranking among all other locations for a specific Aggregate Value. All scores are confined within a range of 0 (lowest potential value) to 100 (highest potential value). The scores for each component are determined based on the given value's percentile ranking in the national distribution (rounded to the nearest hundredth).

Category Rating: With a percentile score, we can ascribe certain breaks in the scale to create a qualitative rating of each indicator (e.g., Low, Medium, High). This describes the comparative nature of a location's value in relation to all other locations at the same geographical level. Determining the breaks requires domain-knowledge, and the exact breaks for indicators should be determined with domain knowledge.

Percentile score calculation

While the **Curated Features** and **Aggregate Value** have unique calculations for each indicator, the **Percentile Score** and **Category Rating** can be standardized across indicators.

We recommend using **pd.Series.rank()** on the aggregate value series to compute the rank across all scenarios. Then, a 5 category rating, Very Low, Low, Medium, High, Very High, can be calculated based on a break of 20, from 0 to 100.

Appendix 1: Risk Indicator Aggregation Overview

The following is a list of features that aggregate into various climate risk indicators. In general, the aggregated value should reflect a combination of the distribution from the absolute risk and the relative change from baseline risk (i.e. 50% weight goes into actual risk and another 50% goes into the change in risk between baseline and future).

When aggregating features in the final step, all features need to be scaled to 0 and 1 to make sure all distributions are added in the same scale.

Risk	***Feature List***	***Aggregated Value Methodology***
Heat	- ***Local Threshold*** = historic 98th percentile (7 days a year) - ***Heat Wave Frequency*** = 2 and more consecutive days above local threshold - ***Heat Wave Intensity*** = Avg degrees above threshold * length of heat wave - ***Heat Risk*** = ***Frequency * Intensity*** ***- Change in Risk*** = Future Heat Risk/Baseline Heat Risk - ***Cooling degree days*** = degree days required to cool to 18 degrees Celsius	***Future Heat Risk * 0.5*** ***+*** ***Change in risk * 0.5*** Measures the conditional damage of heatwaves and the expected increase in power consumption under different climate change scenarios
Inland Flooding	***Local Threshold*** = 98th percentile ***Storm Frequency*** = Number of days in a year with precipitation more than local threshold ***Storm Intensity*** = Avg volume above threshold for each storm event ***Storm Risk***= Storm Frequency * Storm Intensity ***Change in Storm Risk =*** Future Storm Risk/Baseline Storm Risk ***Flood Frequency*** = Historic Flood Frequency ***Flood Zone Proximity*** = Distance to FEMA flood zone	***Flood Zone Proximity * 0.25*** ***+*** ***Flood Frequency * 0.25*** ***+*** ***Change in Storm Risk * 0.5*** Measures the conditional damage of an inland flooding event at the location under different climate change scenarios
Coastal Flooding	***Frequency =*** 1/return period (100 or 250) ***Inundation***= max inundation level for the specified return period in a specified year ***Sea level rise coef*** = Scenario Sea level rise/RCP45 Sea level rise ***Baseline 100 Year Risk*** =1/100 * Baseline 100 Year Inundation ***Baseline 250 Year Risk*** =1/250* Baseline 250 Year Inundation ***Future 100 Year Risk*** =1/100 * Future 100 Year Inundation * Sea Level Rise coef ***Future 250 Year Risk*** =1/250* Future 250 Year Inundation * Sea Level Rise coef ***Change in risk =*** mean(Future 100 Year Risk/Baseline 100 Year Risk) + (Future 250 Year Risk /Baseline 250 Year Risk) - 1	***Future 100 Year Risk * 0.25*** ***+*** ***Future 250 Year Risk * 0.25*** ***+*** ***Change in risk * 0.5*** Measures the conditional damage of coastal flooding at the location under different climate change scenarios

Drought	***Fresh Water Quality*** = Water Quality Index ***Baseline Water Stress*** = Water Supply/Water Demand ***Change in Water Stress*** = Change in water stress under different climate change scenarios ***(NEW) Frequency*** = drought frequency ***(NEW) Intensity*** = drought intensity	***Fresh Water Quality * 0.25*** ***+*** ***Baseline Water Stress * 0.25*** ***+*** ***Delta Water Stress *0.5*** Drought measures the availability and quality of fresh water in a watershed.
Wind	***Frequency*** = Annual Frequency of hurricane ***Severity*** = Maximum Speed of hurricane as a proxy of the expected damage ***Wind risk*** = ***Frequency * Severity*** ***Change in wind risk =*** Any way to estimate increase or decrease in Hurricane risk based on climate change?	***Frequency * Severity*** Measures the conditional damage of a hurricane event at the location
Fire	***Burn Probability =*** likelihood of a given location on your landscape burning ***Severity =*** Conditional Risk to Potential Structure ***Building Coverage =*** Pct. Cover of buildings of an approximate area ***Climate coef = Change in Humidity, Windspeed from 2050 to 2020 (Change in weather patterns)***	***Burn Probability * Severity*** Measures the conditional damage of a fire event at the location under different climate change scenarios