**Use of Big Data and Weak-Signal Analysis to Counter Human Trafficking and Illegal, Unreported, Unregulated (IUU) Fishing.**

Novametrics LLC

World Bank Group Vendor Contract: #8005739

Contract: PR-17-41046

***Methodology w/***

***Conceptual Framework, Taxonomy & Algorithm***

Version November 2017

***Framework***

Human trafficking and IUU fishing are ubiquitous within the commercial seafood industry, as the vastness of the oceans, jurisdictional ambiguities, and enforcement inefficiencies limit effective policing. Current strategies to achieve United Nations Sustainable Development (SDG) goals related to human trafficking and IUU fishing have primarily involved inspections, surveys, reports, and the use of satellites, planes, and observers to identify violations.  These techniques are expensive, can only spot-check, and are vulnerable to evasion scenarios.

Human trafficking and IUU fishing are closely coupled.  The two phenomena exist within the same socioeconomic ecosystem. Overfishing reduces the profitability of the seafood industry at multiple levels, creating incentives for forced labor.  Furthermore, both activities typically occur on ships operating outside the regulatory process.

The victims of human trafficking are among the most vulnerable members of society. There is little incentive for them to come forward and report abuse. In addition, local customs and traditions may foster an environment in which certain forms of slavery are considered an accepted component of the social norm. Economic circumstances and the absence of enforced rule of law can create the opportunity to exploit those without access to alternative income-producing activities or social safety nets.

Although there is little direct data on human trafficking and IUU fishing, large amounts of indirect data exist.  Every community and enterprise has a complex mosaic of characteristics derived from the demographics, environmental resources, geographical location, economic activity, socio-cultural landscape, etc. that define a unique environment.  These characteristics are reflected in census data, transactions, survey data, remote imagery, and many other datasets.  While these datasets may be of varying quality and completeness, each has the potential of carrying information that reflects the crime, either by itself, or more commonly through combination with other datasets.

New methodologies in big data analytics hold promise for exploiting these existing diverse data sets to reveal human trafficking and IUU fishing.   In many areas, we are no longer data limited. We can go beyond the resolution of traditional data analysis, identifying diagnostic “signals” in what was previously considered “noise,” to improve estimates of prevalence, identify underlying causal relationships, and optimize return on investment (ROI) for interventions.

***Analytical Approach***

From an analytical perspective, eliminating human trafficking and IUU fishing can be classified as a “wicked problem”[[1]](#footnote-1) – the type of problem that defies a single solution and is characterized by a myriad of dynamically interconnected variables. Examples of “wicked problems” include reducing poverty, stopping outbreaks of civil violence, and mitigating natural disasters. Wicked problems occur in complex systems, where causal relationships are not direct, and where the circumstances that foster the problem vary from location to location.

While wicked problems may defy linear solutions, they can be characterized. More importantly, outcomes associated with wicked problems can be predicted. We predict outcomes through an “ecosystems” approach – a holistic approach that parameterizes the ecosystem in which the problem exists and allows for predictive modeling.

Consistent with criminal theory, the ecosystem includes three components: 1) an activity that can benefit from enslaved individuals, termed “potentially exploitive industries”, 2) individuals who can be enslaved, termed “population w/ low resiliency”, and 3) an environment in which abuse can occur, termed “limited societal safeguards”. The first two components constitute “motive”, the third one constitutes “opportunity”.

Efforts aimed at deterring human trafficking and IUU fishing focus on individual components of the ecosystem. Observers, declarations, and monitoring systems reduce the opportunity to commit the crime. The Port State Measures Agreement, ecolabeling schemes, and supply-chain certifications reduce motive by decreasing the value of illegal product.

While all efforts to stop human trafficking and IUU fishing are valuable components of an overall strategy, all have limitations. No single system operating in isolation is a silver bullet. In addition, it is difficult to evaluate the specific impact of any single effort. With over 1.8 million engine-powered fishing vessels operating across 300 million km2, monitoring the oceans is a formidable task. Differing international laws and jurisdictional ambiguities make prosecutions difficult, expensive, and time-consuming. Markets for fish and fish products are diverse, numerous, and globally distributed. Potential bad actors are omnipresent. As criminology theory recognizes, if an ecosystem can support a predator, a predator will appear; and if a predator is removed, another is likely to emerge.

Using big-data analytics and weak-signal analysis, we will apply an ecosystem approach and use the very characteristics that foster human trafficking and IUU fishing as a means for identifying the activity.

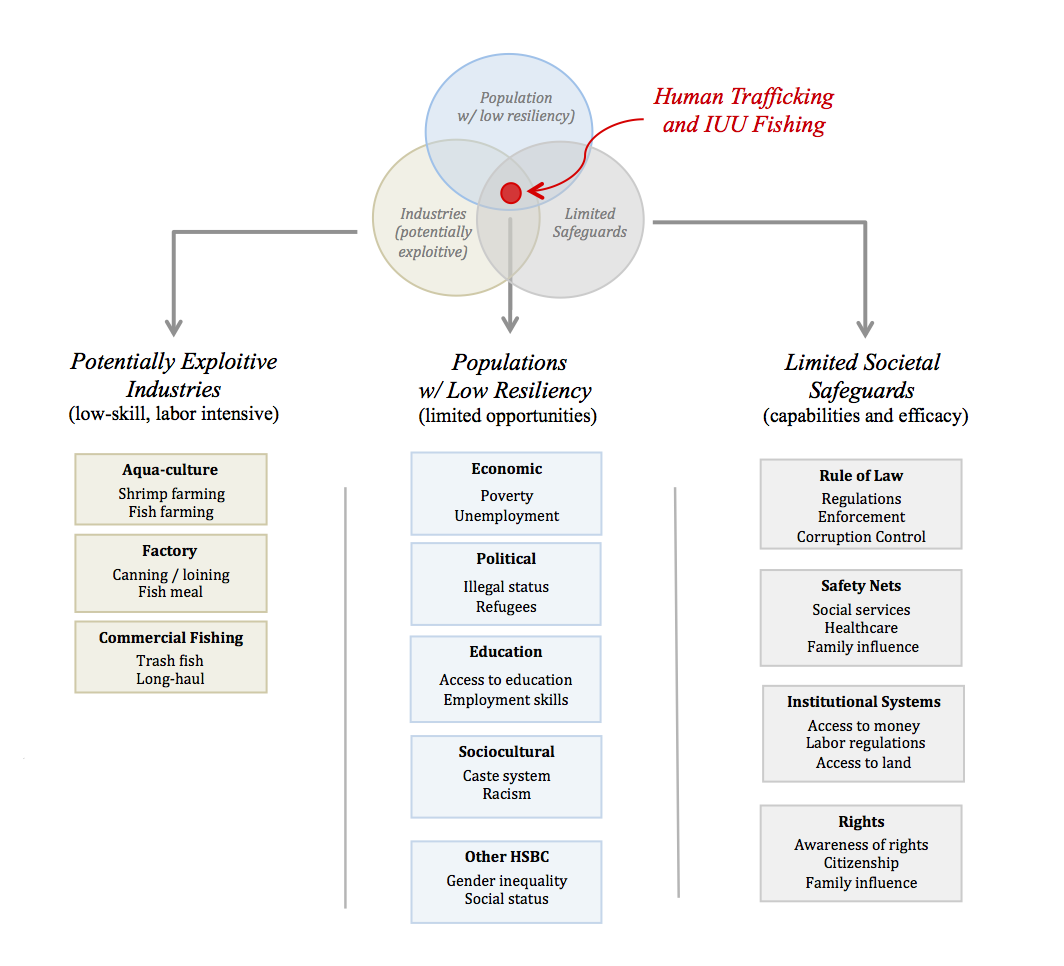


Figure 1: Schematic taxonomy of the ecosystem in which human trafficking and IUU fishing occurs.

The schematic taxonomy (Fig. 1) can be translated into an algorithm and represented with a mathematical expression for vulnerability to human trafficking and IUU fishing (V) for each designated ecosystem α. Vulnerability Vα for ecosystem  is the product of the population with low resiliency (Pα) with exposure to potentially exploitive industries (Iα) and weak societal safeguards (Sα).

Vα  = Pα x Iα x Sα

The factors are multiplicative, as reduction of any factor to zero (e.g. removing the presence of potentially exploitive industries Iα) reduces vulnerability Vα to zero. Each factor contains multiple terms (examples of which are included within the columns of Figure 1) that sum together rather than multiply. For instance, in the column of societal safeguards, perfect labor regulations do not eliminate the vulnerability associated with poor law enforcement.

In recognizing the dynamic nature of the ecosystem and the presence of triggering events, the vulnerability estimate allows for time-dependent components with a delay factor w() that varies with a delay time . The total vulnerability V at time t2 is the time-interval since time t1 of past vulnerability, summed over all sectors of the ecosystem .

V(t2) = **∫** w(2) Σα Vα(t) dt = **∫** w(2) Σα Pα(t) x Iα(t) x Sα(t) dt

Where the integral over time is from t1 to t2 and:

Pα(t) = ƒP (wealth, education, …)

Iα(t) = ƒI (low-skill labor, isolated work, …)

Sα(t) = ƒS (rule of law, inspection, …).

In the schematic of our taxonomy, potential interventions can dismantle the ecosystem in which human trafficking and IUU fishing breeds by:

a) detaching the population with low resiliency from exposure to the risk landscape of potentially exploitive industries operating under limited societal safeguards,

b) insuring the robust and omnipresent societal safeguards, or

c) eliminating potentially-exploitive industries.

Depending on the scope of effort and the specific socio-cultural landscape in which the modern slavery exists, interventions could target increasing resiliency, reforming industries, or improving safeguards.

***Summary of Methodology***

*(see Appendix for detailed discussion of methodology)*

*1. Analytical Approach*

The initial process begins with the development of a conceptual framework. A taxonomy is built to that conceptual framework, and an algorithm is developed. The algorithm is represented with a mathematical expression to confirm logical validity.

*2. Data Collection and Review*

Data are collected and a database is developed in a raw persistent data storage layer. All data are reviewed for quality control purposes and checked for completeness, documented metadata, sampling process, definitions of indicators, units, etc. These data are referred to as Level 0 data.

*2.1 Data Organization*

The data are collated, along with source references and metadata. The data are reformatted into the appropriate file formats for use in statistical and geospatial analysis. A complete original version of the data is preserved.

*2.2. Initial Data Culling*

As a general rule, indicators are deleted if they are absent for a large majority of the data. Judgment is applied depending on the potential value of the indicator and the availability of alternative “proxy” indicators that might be reflecting comparable phenomena. Indicators are removed if they are judged to have inadequate coverage or are obviously redundant with other indicators that have greater distribution.

*3. Database Development*

The database consists of a raw persistent data storage layer and a virtualized data layer. The persistent layer preserves the raw data in its original form (Level 0 data). The virtualized layer contains the data products that have been developed for the analysis through various statistical and geospatial algorithms (Level 1 data).

*3.1 Standardization*

Because the analysis includes comparisons, we “standardize” the data. For the purposes of our analysis, we refer to standardizing data as providing a common reference. For example, instead of GDP, we may use GDP per capita, instead of Arable Land per km2, we may use Arable Land per capita, etc. In summary, whenever we have a measure, we determine how that measure should be standardized so that it can be compared from one location to another. The standardization process may include weightings.

*3.2 Vectorization of Data*

While the various indicator measures provide a “magnitude” they typically do not provide a “direction.” We therefore provide a direction to the value based on the how the indicators will be compiled. In most cases, this step requires adjusting as a reciprocal or an inverse. For instance, high infant mortality rates are considered a negative feature, while high immunization rates are considered a positive feature. We might therefore align the direction of the indicator measures by multiplying all values of one of the indicator measures by -1.

*3.3 Data Imputation*

Time-series data are interpolated. Interpolation is done using Piecewise Cubic Hermite Interpolation (PCHIP) or linear interpolation. If a region is missing so much data that interpolation is unreasonable (based on examining the time-series) and there have been significant changes in the region that could have an impact on the indicator, it is removed from the analysis.

# 3.4 Normalization

Because our goal is to perform statistical analysis that combines different types of data, we translate the units and scales of the indicators into a common form of measurement by normalizing. Novametrics employs a series of normalization techniques and applies the technique most appropriate for the data distribution. If the data present a normal or log-normal distribution, we normalize the data for each indicator using the deviations from the norm. If the data distribution does not follow a normal distribution or exhibits natural breaks, we use percentile ranking or assign values to each grouping.

*4. Analysis*

The analysis applies Novametrics’ weak signal analysis to big data to identify characteristics predictive of outcomes, and to reveal underlying causal relationships. The weak signal approach to big data finds inter-relationships among multiple variables. With many variables and many distinct populations, there may be multiple independent linkages. Multiple independent linkages indicate the problem has multiple causes; and the causes have different relative priority in different locations. Previously hypothesized relationships may be confirmed, but unexpected relationships are more common. It is the unexpected relationships that lead to a more sophisticated understanding, and in turn, offer opportunities for more nuanced and effective interventions.

*4.1 Novametrics Weak Signal Analysis*

Weak signal analysis allows us to go beyond the resolution of traditional data analysis, identifying diagnostic “signals” in what was previously considered “noise,” to improve estimates of prevalence, identify underlying causal relationships, and optimize return on investment (ROI) for interventions.

*4.2 Correlation Analysis*

In our holistic ecosystem-approach, correlations can indicate predictive relationships. The analysis remains agnostic about the form of the relationship and does not assume correlation implies causation.

*4.3 Factor Analysis*

Factor Analysis is used to identify the main factors that contribute to the variability of the data. Weightings for each of the factors are quantitatively determined through regression analysis. Composite measures are developed using indicators identified through the factor analysis with weightings derived from the regression analysis.

# 5. Evaluation, Groundtruthing, and Hindcasting

Novametrics validates its predictive analytics through ground truthing. Ground truthing is performed by applying the analysis to different areas and evaluating the predictive success. If suitable regions are not available, we use hindcasting. The analysis is adjusted for previous times and the predictions are compared to known outcomes.

*6. Visual Presentation*

A suite of presentation methods is used depending on the nature of the analysis and the decision-making needs of the user community. Methods include graphics, geospatial displays, and other methods for displaying quantitative information that allow for an intuitive understanding of the analytical basis for the recommendations.

**Appendix: *Summary of Analytical Products***

|  |  |  |
| --- | --- | --- |
| ***Task*** | ***Analytical Step*** | ***Product*** |
|  | Analytical approach. In consultation with subject matter experts, a conceptual framework, taxonomy and algorithm are developed. The algorithm is expressed in mathematical terms. | Dynamic Taxonomy & Iterative Algorithm |
|  | Data collection, referencing, reformatting, collating, and review for quality and completeness. This step includes the development of appropriate database architecture, transfer of all relevant metadata, and construction of the Database Dictionary. | Data Dictionary with sources and citations |
| Macintosh HD:Users:gvdv:Desktop:Normal_distribution_and_scales.gif | Data standardization, normalization, vectorization, and data imputation (interpolation and extrapolation). This step uses a range of statistical methodologies based on the data distributions to provide a complete, quality-controlled database for analysis. | Database with documented data standards |
|  | Statistical analysis, regression analysis, culling, identification of proxy measures, identification of primary indicators, weightings, and composite indicator formation. | Statistically-derived Composite Measures for algorithm |
|  | Integration of tabular data products, survey data, and modeling data with geospatial data and geospatial analysis. This step includes assigning composite measures to specific locations and/or populations. | Geospatial data products |
| Macintosh HD:Users:gvdv:Desktop:Screen Shot 2015-12-28 at 1.07.27 PM.png | Weak signal analysis to identify underlying linkages, both systemic and causal that are predictive of outcomes. | Identification of underlying linkages |
| Macintosh HD:Users:gvdv:Desktop:Screen Shot 2015-12-28 at 1.09.53 PM.png | Identification of linkages and expected return on investment for various strategies. | ROI estimates for strategies |
|  | Implementation heat maps. Optimization with geospatial data to evaluate locations where each action will have the greatest impact | Heat Maps to optimize results |
|  | Dashboard. Data visualization to provide evaluation through spatial (map), temporal (time-series), and impact evaluation. Customized to decision-making requirements. | Data visualization to inform decision-making. |

**Appendix: Detailed Discussion of Methodology**

***1. Development of Analytical Approach***

* See main document for the development of the analytical approach.

***2. Data Collection and Review***

* Every population has a complex mosaic of characteristics derived from their demographics, environmental resources, geographical location, ethnic history, wealth, dialects, income-producing activities, religious sects, access to markets, educational levels, etc. These characteristics are reflected in population data, survey data, remote imagery, social media, infrastructure, market transactions, and many other datasets. While these datasets may be of varying quality and completeness, each has the potential of carrying information that reflects a characteristic of the population, either by itself, or more commonly through combination with other datasets.
* Our methodologies in big data analytics allow us to go beyond the traditional analysis of “big data” by integrating non-traditional data through advanced statistical and signal processing algorithms. In many cases, we are no longer data limited. Our data fusion methodologies allow us to identify diagnostic “signals” predictive of outcomes in what was previously considered “noise.”
* Not all forms of “Big Data” are large volumes of data. And not all large volumes of data are “Big Data”. Using an example presented in one discussion of “Big Data:

An airplane on a regular one-hour flight has thousands of sensors covering everything from the speed of air over the airframe to the amount of carbon dioxide in each section of the cabin. Each sensor is effectively an independent data-collection device. The real interest is usually in combinations of sensor readings (such as speed of air combined with air pressure). The combinations are complex and vary with the error tolerance and characteristics of each device. A holistic analysis of the airplane’s performance is “big data” analysis, although the size of the dataset is not large. Even a hundred thousand sensors, each producing an eight byte reading every second creates less than 3GB of data in an hour of flying (100,000 sensors x 60 minutes x 60 seconds x 8 bytes). Conversely, media and telecommunications stream generate petabytes of simple and structured data. Internet search engines (e.g. Google) and relational databases can parse this well-structured data quickly. While the data volume is large, it isn’t “big” in the same way as the data coming from the myriad of sensors in the earlier example.

* Early discussions of “big data” recognized that data could be “big” based on such dimensions as volume, variety and velocity (e.g. Gartner / Meta Group, 2001). More recently, the dimensions of “big data” have expanded to “5 Vs”: volume, variety, velocity, variability, and value (Jain, IBM, 2017).
* Novametrics does not use any particular definition of “big data.” We absorb all variants of “big data” in our data fusion methodologies. The goal is to develop compiled indicators that can serve as surrogate measures for the sociocultural characteristics of the population. We do not assume that any of these indicators by themselves are predictive. The premise is that by combining indicators, we are able to discern characteristics that are predictive of outcomes.

*2.1 Data Organization*

* As a means for organizing the data, evaluating data quality and identifying the appropriate indicators to use, we group the data into various categories or “buckets”. Because we collect all available data, in many cases we will have 500-1,000 measures that can be used to evaluate a location, market, or population. Some indicators may appear in more then one bucket. That does not affect the analysis. When the relationship between the indicators and the phenomena that is trying to be measured is determined, the indicators are used collectively without regard to their buckets. The buckets are primarily for two purposes: 1) to organize the data and thus allow for consistent normalization and standardization procedures, and 2) to identify either redundant data or data that can be used as proxies through high correlations. Sample data bucket categories include:
* Demographics (population density, age distribution, etc.)
* Governance (aid, democracy indices, % military expenditures, openness of press, history of conflict, presence of violence/terrorism, rule of law, regulatory quality failed state index, control of corruption, male/female leadership (gender disparity), etc.)
* Economics (GDP per capita, GINI, unemployment, inflation, GDP growth, etc.)
* Education (literacy rate adult, literacy rate youth, literacy rate males/females (gender disparity), trained teachers, proximity to schools, etc.)
* Health (immunization, disease prevalence, infant mortality, access to healthcare facilities, sanitation, improved water access, physicians, teenage mothers, HIV/AIDS, malaria, under-5 mortality rate, orphans, etc.)
* Infrastructure (urbanization, road access (RAI = within 2 miles of a year-round road), paved roads, communication technology, mobile cellular subscriptions, internet access, transportation, etc.)
* Natural Resources (freshwater access & withdrawals, agriculture (% land, % subsistence, % GDP), fishing, rubber, timber, oil, gas, minerals, etc.)
* Wealth Potential (land-use, land-ownership, land-productivity, water, livestock, agriculture income, economic freedom indices, purchasing power, access to markets, exports, manufacturing, inflation, employment, bureaucracy, etc.)
* Social (dependency ratios, women in wage-earning jobs, women in parliament, etc.)
* Cultural (ethnic groups, dialects, refugees, etc.)
* Religiosity (# of religions, participation, extremist organizations, etc.)
* Transparency (freedom of press, corruption, democracy, etc.)
* As we use various methods for reducing the number of indicators for analysis, it often becomes useful to reduce the number of indicator categories as well. The categories are assigned numbers so they can be sorted and mapped into each other.
* All data are reviewed for quality control purposes. Novametrics checks for completeness, documented metadata, sampling process, definitions of indicators, units, etc. The data are collated, along with source references and metadata. The data are reformatted into the appropriate file formats for use in statistical and geospatial analysis.
* Copies of the original data, along with the metadata, are maintained and cloud hosted. The original data are maintained for full traceability and reproducibility.

## *2.2 Initial Data Culling*

* Indicators are removed if they are judged to have inadequate coverage or are obviously redundant with other indicators that have greater distribution. The general approach is to plot the indicators available for the region of interest, identify a natural break, and determine the location of the break based on quartiles and distribution of the data.
* In one analysis, the decision was made to remove indicators that had less than 70% of the data for the regions of the study. The value of 70% was chosen as the threshold because of a natural break in the distribution of the data points per indicator. The break was based on quartiles and distribution of data.
* Judgment is applied depending on the potential value of the indicator and the availability of alternative “proxy” indicators that might be reflecting comparable phenomena about the subject.

***3. Database Development***

* The database consists of a raw persistent data storage layer and a virtualized data layer. The persistent layer preserves the raw data in its original form (Level 0 data). The virtualized layer contains the data products that have been developed for the analysis through various statistical and geospatial algorithms (Level 1 data).

*3.1 Data Standardization*

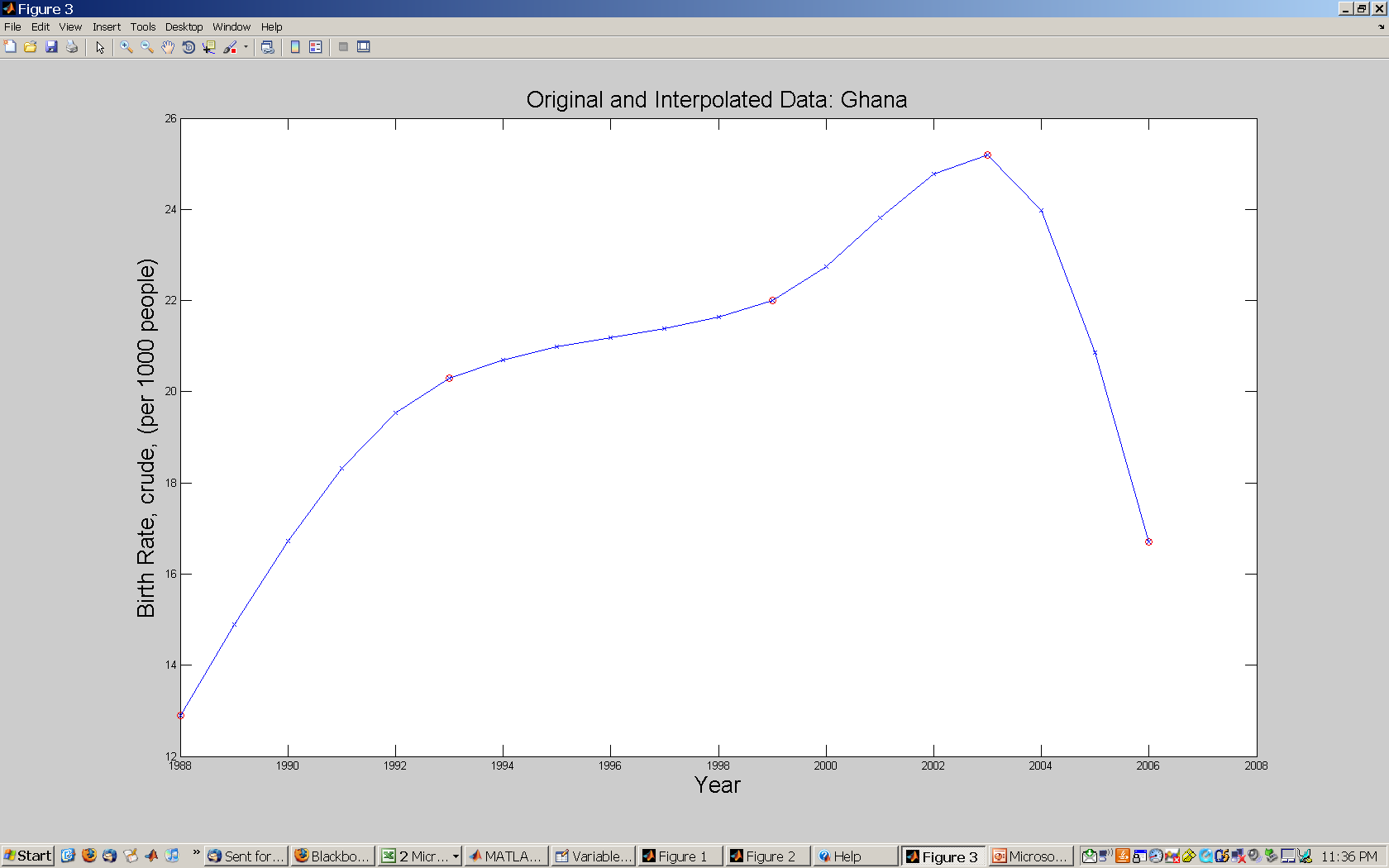
* Because the analysis includes comparisons, it is important to “standardize” the data. For the purposes of our analysis, we refer to standardizing data as providing a common reference.
* For example, instead of GDP, we may use GDP per capita; instead of Arable Land per km2, we might use Arable Land per capita; instead of “number of telephones,” we may use “number of telephones per capita”; etc. In summary, whenever we have a measure, we consider how that measure should be standardized so that it can be compared from one region to another. If it is related to population, it is standardized against population (per capita). If it is related to area, it is standardized against area (per km2). If it is related to population density, it is standardized to population density (per person/km2)? In some cases, the standardizations can be combined. For example, if we measure telephone lines per person/km2, the combination of population and area allows us have a measure that provides a separation in access to telephone lines based on urban areas (high population density) compared to rural areas (low population density).
* Although we keep all original data in the database, once we begin the analysis the duplicate indicators that are not standardized are generally not used. For example, if one indicator is “Number of Telephones” and another indicator is “Number of Telephones per capita”, we remove from the analysis the indicator that is not standardized.
* *Standardizing with weightings:* When standardizing data across regions, it may be important to consider weighting the standardization. As an illustration of the process, consider Africa. If the Seychelles’ population is 1/50th of Africa’s total population, its weighting factor would be 0.02 (2%). To find the average of a measure for the whole of Africa, we cannot simply average the measures for each country. Instead, we must find the average in relation to the weighting factors. To conceptualize why we must do this, consider Seychelles again: if it has a literacy rate of 100% with only 2% of Africa’s total population in its borders while Nigeria has a literacy rate of 70% for 15% of Africa’s population, then clearly Nigeria’s literacy rate should weigh more heavily than the Seychelles in calculating the average literacy rate for the whole of Africa. Thus, in calculating the average of a measure for the whole of Africa, we multiply each measure by its country’s population weighting factor and then *add* each of these weighted components to find the average of this measure for the whole of Africa as weighted for population.

*3.2 Vectorization of Data*

* While the various indicator measures provide a “magnitude” they do not provide a “direction.” We therefore provide a direction to the value based on the how the indicators will be compiled. In most cases, this step requires adjusting the value to its reciprocal or inverse.
* For example, the indicator “Political Rights” is ranked on a scale as follows: 1.0 - 2.5 = Free, 3.0 – 5.0 = Partly Free, and 5.5 – 7.0 = Not Free. We may reverse the scale so that a higher value indicates a freer rather than more oppressed nation. As another example: high immunization rates are “good”, but high infant morality rates are "bad," so we might multiply the infant mortality rates by -1 so increases in rates imply improvement for both measures.

*3.3 Data Imputation*

* We refer to data interpolation as “imputation of missing data”. This can get quite elaborate if we are trying to correct for data that might be missing for non-random reasons. For example, high-income households might be less likely to report their income and therefore we are correcting for such missing data. For new data sets, we begin by graphing the data and considering what the data represents. Based on the type of data and what we know about the region during the time of interpolation is critical for informed interpolation. Absent compelling reasons, the default is to use either Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) or linear regression. Linear regression assumes a linear relationship. Accordingly PCHIP is preferable in most cases, although it makes little difference over short intervals.
* If a region is missing so much data that interpolation is unreasonable (based on examining the time-series) and there have been significant changes in the region that could have an impact on the indicator, it is removed from the analysis. Judgment is applied depending on the potential value of the indicator and the availability of alternative “proxy” indicators that might be reflecting comparable phenomena about the population.
* There is a risk in having too much “interpolated data” relative to “raw data”. That said, the majority of our data is interpolated. Even in the United States, census data is only taken once every 10 years. As a result, data interpolation is a common practice. Ultimately, the amount of raw data required for interpolation is a judgment call. It depends on the type of data and what has occurred in the area of interest. In may be possible to interpolate with confidence using just two data points, if for example, we are measuring population growth in a region that has been stable over the time period in question. Using the same example of population growth, it may also be possible to interpolate (extrapolate) from a single data point, if we know the characteristics of the region and are confident that it has developed in a manner consistent with another region. In this case, we are essentially using the growth rate of another region as a proxy for the growth rate of our region. For example, if the average population growth in an area is 3%/year, we might assume that the population growth in our area of interest is 3%/year. If such an approach is taken, it is documented in the metadata, both with the basis for the assumptions and the selection of the proxy indicator.
* When using linear regression or PCHIP, we plot the data. We may also use more than one calculation on the time-series. For example, in a technology time series shifting from periods when telephone lines were prevalent to when cellular phones overtook landlines may provide a better estimate of technology growth.
* Below is an example of PCHIP interpolation applied to a time-series.

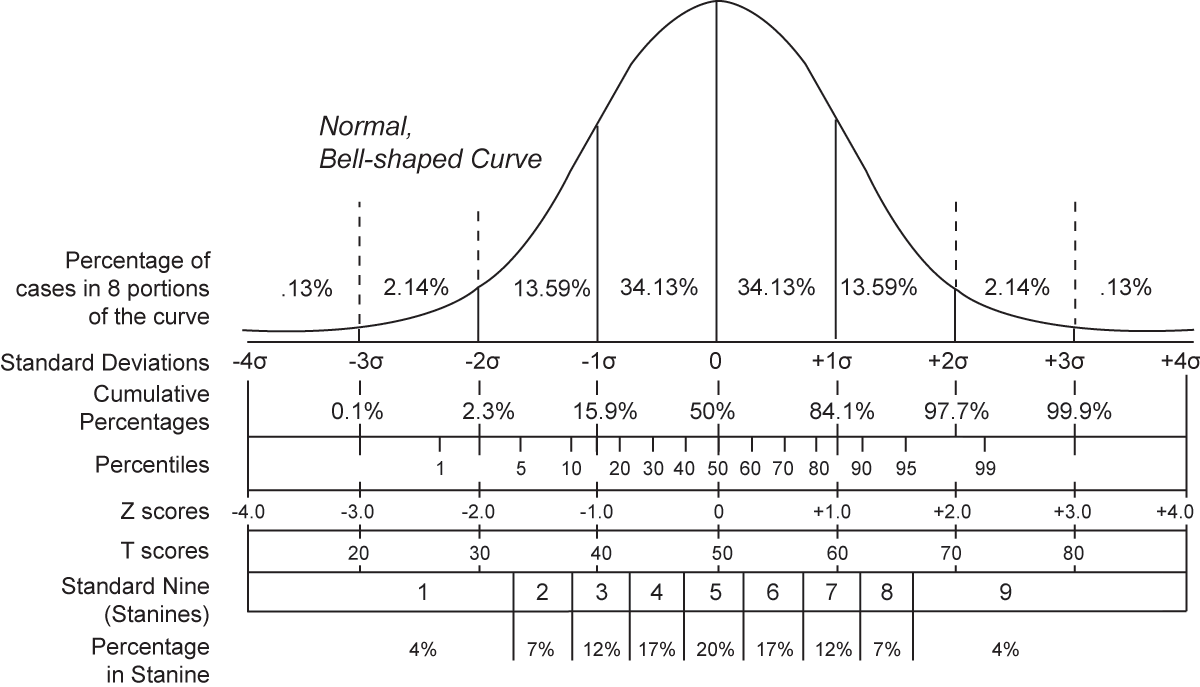


# 3.4 Normalization

* Because our goal is to perform statistical analysis that combines different types of data, we translate the units and scales of the indicators into a common form of measurement by normalizing. There are several ways of normalizing data. The first step is always the same: we examine the data.
* If the data present a normal distribution, we normalize the data for each indicator using the deviations from the norm – also called the “Z Score” because the normal distribution is some times referred to as the “Z” distribution:

[(value – mean) / (standard deviation)] = normalized value.

* If the data distribution does not follow a normal distribution or exhibits natural breaks, we use percentile ranking or assign values to each grouping. This type of normalization is referred to as “ranking”. It has the advantage of recognizing different groupings, minimizing the affect of outliers, and allowing for assigning values to groupings according to their relative values.



***4. Analysis***

* The analysis applies Novametrics’ weak signal analysis to big data to identify characteristics predictive of outcomes, and to reveal underlying causal relationships. Weak signal analysis, along with correlation analysis and factor analysis, allows us to empirically develop composite measures customized for each sociocultural and economic landscape.

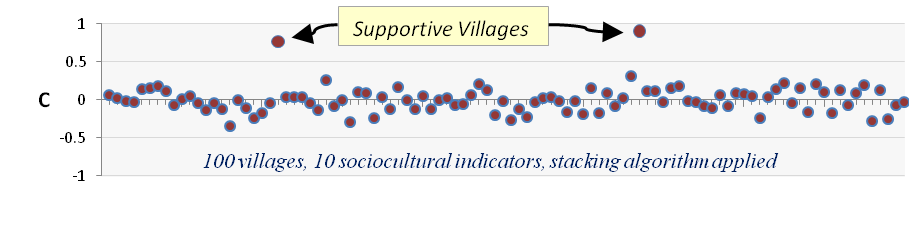
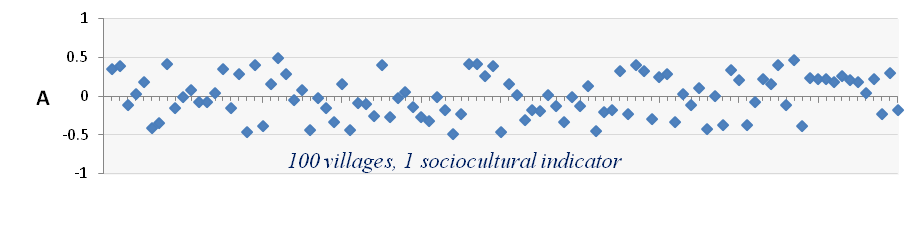
# 4.1 Novametrics Weak Signal Analysis

* Our analysis assumes that all data have value. While the datasets may be of varying quality and completeness, each has the potential of carrying an indicator that reflects a characteristic of a population, either by itself or through combination with other datasets. The analysis methodology is therefore designed with an open architecture that can incorporate all available data. Poor quality or erroneous data gets filtered out of the analysis.
* “Weak Signal” analysis reveals characteristics of a population or a region that may not be evident in any individual set of data. Adapting methodologies originally developed for signal processing, we integrate social, economic, political, cultural, and environmental data sets to develop quantitative measures of a given population or market. We do not assume that any of these indicators by themselves are predictive.
* The concept of data stacking is illustrated by the two astronomical images in the figure following. The picture on the left is a single image. The picture on the right is the combination of 27 images of the same object – mathematically stacked one on top of the other. The composite of 27 images provides greater resolution and therefore a more accurate picture than any of the single images. In data processing terminology, we are able to amplify the “weak signals” and suppress the “noise”. The weak signals (small-scale physical features of the planet) are present in the individual images but they cannot be seen because of the blurriness of the picture. The blurriness, or noise, is caused by atmospheric refraction, vibrations of the telescope, etc. and varies randomly in each of the images. When 27 of the images are stacked upon each other, the random noise cancels itself out and the signals co-add among the images. The result is an enhancement of the signal relative to the noise.



Left is the image created by a single astronomical dataset. Right is the same image created by stacking multiple datasets so “weak signals” (subtle physical features of the planet) that were previously hidden in the noise of the data can now be recognized.

* A corresponding example of the data-stacking concept as applied to our model is measuring (developing an image of) a population’s resiliency to a potential agricultural disruption such as crop failure. In our methodology, a single image might be, for example, a dataset consisting of the percentage of the region’s population devoted to agricultural activity, or the amount of land in each geographical area devoted to agriculture. Viewed alone, this single data set gives us a hazy, incomplete picture of how sensitive a population might be to a crop failure by telling us the percentage of land devoted to agriculture. It does not reveal other characteristics that might also be important for determining how resilient the population might be to a potential agricultural disruption. For example, it does not include information on the diversity of the local economy, the population’s ability to engage in alternative income producing activities, their access to roads and markets, or even the sensitivity of the crops being cultivated. If we could combine all of these individual characteristics we would have a clearer picture of the population’s resiliency.
* We can apply our weak signal analysis to identify populations of concern. In the figure below, we use the example of a region that contains 100 villages. Our goal is to identify the villages in the region that are most likely to be supportive of an economic activity. In this example, plot “A” may be a measure of the literacy rate in each of the 100 villages (normalized and standardized so they can be compared). Literacy rates by themselves are not indicators of support for economic development, but they may be partial surrogate indicators for the population’s exposure to other cultures and potential for a diversity of income-generating activities.
* Plot “B” shows ten measures for the same 100 villages. It includes not only the literacy rates shown in plot “A”, but also nine other potential surrogate indicators such as infant mortality, market access, % electricity, governance, religiosity, arable land, gender disparity, etc. Again, none of these independent measures by themselves are direct indicators of the population’s support for a particular development operation, but they are all potentially reflective of underlying sociocultural phenomena.
* The bottom plot “C” uses the concept of weak signal analysis to identify the two villages that would be the most supportive of economic development. In plot “B” none of the villages appear to be better candidates than any of the others. When the data are combined according to our algorithm, however, the random fluctuations cancel out and the weak signals that would otherwise be hidden in the noise become additive and appear above the noise level.



*4.2 Correlation Analysis*

* In holistic ecosystem-approach, correlations can indicate predictive relationships. The analysis remains agnostic about the form of the relationship and does not assume correlation implies causation (“*cum hoc ergo propter hoc*”). As an example, consider two population characteristics “A” and “B” that correlate with significant statistical confidence. There are at least five options:

*Option 1:* The correlation is the result of random coincidence and does not reveal any causal relationships between A and B.

*Option 2:* A is “causing” B, with the independent variable A causing the change in the dependent variable B.

*Option 3*: B is “causing” A, with the independent variable B causing the change in the dependent variable A.

*Option 4:* A and B are both dependent variables, both following an independent population characteristic C that has not been measured.

*Option 5:* A and B are part of a larger correlated system with no unique causal factor, that is, no independent variable.

Option 5 is characteristic of “coupled systems” in physics, in which “causality” resides in the linkages between variables. In a fully coupled “holistic” system, no variable is truly independent. Such systems are common in the natural environment. For example, in atmosphere-ocean interactions that lead to the El Niño and La Niña climate events, there are no dependent vs. independent variables. Atmospheric pressure highs and lows induce winds that push surface seawater, and warm and cool patches of the sea surface induce variations in atmospheric pressure. Neither the atmosphere nor the ocean operates independently of the other. Neither can be taken as the independent variable in a causal relationship. Yet the relationship is unambiguous and allows us to predict both the atmospheric and oceanic effects with high degrees of certainty.

* A holistic approach to wicked problems treats problems as coupled systems that lack true independent variables, but that, nevertheless, offer situations where we can predict outcomes and intervene to effect change. The big-data systems approach finds inter-relationships among many variables, not only two. With many variables and many distinct populations at risk, there may be multiple independent correlation patterns. This indicates the problem has multiple causes, and the causes do not have equal relative priority in different places. Previously hypothesized relationships may be confirmed, but unexpected relationships are just as common. And it is the unexpected relationships that lead to a more sophisticated understanding, and in turn, offers opportunities for more nuanced and effective interventions.
* In a holistic approach, the correlations among population attributes are treated as a coupled system that can be influenced at several points, rather than a cause-effect process that can be modified only through its dependent variable.
* If several variables connected with modern slavery are interdependent, whatever the exact form of the causal relationship, an intervention designed to decrease a particular attribute “A” should decrease all variables that correlate positively with “A”, should increase all variables that correlate negatively with “A”, and should leave unchanged the variables that are uncorrelated with “A”. The advantage of a holistic systems approach is that it allows us to achieve our objectives by identifying the characteristics to be modified, therefore allowing us to identify the interventions that will provide the greatest return on investment.
* X is positively (negatively) correlated to Y if Y tends to be large (small) when X is large and Y is small (large) when X is small (large). Although the concept of correlation can be related to that of regression, these are not totally equivalent as they are derived from different sets of assumptions. Usually, when one talks of correlation, one refers to the parametric Pearson product-moment correlation “”, which is practically estimated from a finite data set as the value “r”. In statistical jargon, “r” is the ratio of the covariance of X and Y to the square root of the product of the individual variances of X and Y. Mathematically, this estimator takes the following form, which can be used for computation:



where are the arithmetic means of the N data points: . Numerically, “r” varies between -1 and 1 and we note that when r=1 or r=-1, X=Y or X=-Y. Note that this expression is commutative, that is one can change the order of X and Y and the value of “r” will remain the same. This implies that neither of the variables X or Y is considered the dependent and the other the independent variable. This is different than regression analysis, which implies a functional relationship between X and Y.



* The numerator of “r” is called the covariance between X and Y. “Covariance” simply says that when two variables covary, if one is above its expected value, the other one also tends to be above its expected value. The expected value is the value that the variable would have if the world was perfect (i.e. if there was no error) and we denote it by “E(…)”. Consider the product computed for a single sample. Since X and Y are drawn from distributions (i.e. from the real world), single values of X and Y likely differ from E(X) or E(Y), which are practically estimated by . Looking at this expression, if when X is greater than E(X) Y is greater than E(Y), the product will be positive. Similarly, if when X is smaller than E(X) Y is smaller than E(Y), the covariance will also be positive. If X and Y are out of phase, the product will be negative. Since X and Y are taken from probability distributions, if we compute this product for an infinite number of samples, we will have a distribution of positive and negative values. The question is, given the distribution of the data sets of X and Y, what is the expected value of that product averaged over all the samples? In other words, what is the expected value of the covariance? Mathematically, this becomes



.



* Because N is not infinite, however, there is a chance that one picks sets of X and Y that produce high values of “r”, even when no correlation actually exist. This is equivalent to answering the question “if X and Y have nothing to do with each other, what is the probability, given N samples, that they will produce a value of “r” that is greater or smaller than r, a critical value of r?”
* First, this implies one must decide of an acceptable probability of being wrong or to conclude that a correlation exist when actually there is none; typically this is 5%, or =0.05. Then, one computes the number of “degrees of freedom (df)”, this is typically the number of data points (N) minus the number of parameters you must estimate to compute the statistics. In this case, one uses the data to compute two parameters (), so df=N-2.Finally, we look up in critical “r” values the value r that corresponds to the  chosen and df. If the absolute value of “r” computed from the data is larger than the absolute value of r, the probability that we found a correlation when none exist is smaller than  (i.e. 5%), so we have high confidence that our results are not due to chance.



* Strictly speaking, Pearson correlation can only be applied if X and Y follow a “bivariate normal distribution” and if the level of measurement of X and Y is on an interval or ratio scale. On interval measurement scales, one unit on the scale represents the same magnitude across the whole range of the scale. If the measurements of X or Y are not drawn from a bivariate normal distribution or if they are made on an ordinal scale, Pearson correlation cannot be used. In that case, other alternative measures of correlation are recommended, based on the ranks of the observations relative to one another: Spearman’s rs or Kendall’s . Note that either of these will work on any data that can be used for Pearson, but these non-parametric tests are not as powerful: more samples are needed to achieve a similar degree of confidence.
* To compute the *Spearman c*orrelation coefficient (rs), one must first rank the observations of X and Y from smallest to largest and we replace the data values with the respective ranks of each sample in the ordered set: [Xk,Yk] 🡪 [Lk,Mk]. In case of ties, the value of the rank will be the average of the tied ranks. We then compute the value , that is the difference between the ranks of Y and the ranks of X. Spearman’s coefficient is then computed by:



.



* *Kendall’s*  is a little more complicated to compute. Instead of taking the difference of the ranks, one must choose one of the variable and order the data set of ranks [Lk,Mk] according to either Lk or Mk. Since the ranks of Lk and Mk are not necessarily identical, one of the vectors will simply be an ordered list from 1 to N, but the other will likely be a scrambled list of ranks. From this scrambled vector, we construct a new vector Q. To make Q, we proceed from the first pair of rank, look at the rank corresponding to the entry of the scrambled vector and count all the ranks smaller than itself for every point that has not yet been considered, that is from point 2 to N. We then proceed to the second rank value and compare it to the ranks of the scrambled vector for point 3 to N, etc. Note that if both the L and M vectors order similarly, Q will be a vector full of 0 as there will never be a rank greater than it’s own. Kendall’s  is then computed as:

.

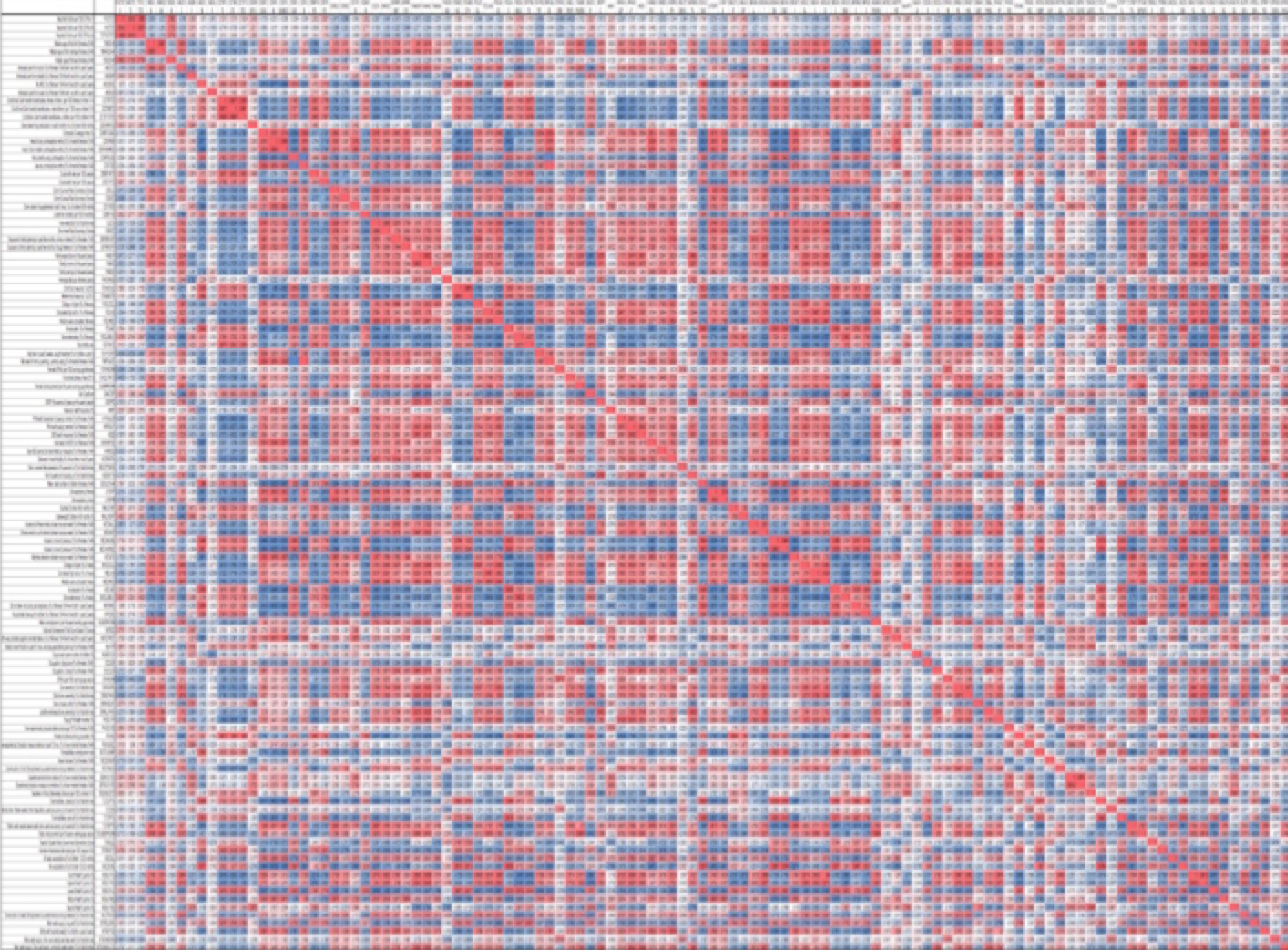


* Significance testing of the Spearman or Kendall coefficient works as for Pearson, but one must use different tables, that is a table for the critical Spearman coefficient and a different table for the critical Kendall’s .
* Correlation analysis is never used blindly, for reasons illustrated below with Anscombe’s quartet. The Anscombe quartet is a set of four different data sets, each comprised of 11 samples, giving exactly the same statistics for the mean and variance of X and Y as well as the same Pearson correlation coefficient. The numbers above the plots in bold are the correlation coefficients computed with the Pearson (r), Spearman (rs) and Kendall’s () method. The corresponding values below the graphs show the p-values, or the probability that the observed correlation coefficient is due to chance. Only the first data set is actually suitable for a Pearson correlation analysis in spite of the large r-values and small p-values. Kendall’s  and Spearman’s rs are giving better results, especially in the last (bottom right) case. This last case also illustrates the effect of outliers on the Pearson correlation results.



* When investigating a large data set, one useful method to understand the relationship between all the parameters is to compute a so-called correlation matrix. This is nothing else than a table showing the correlation results for all the possible parameter pairs. The correlation matrix is symmetric because the correlation between X and X is 1 and the actual number of pairs to consider is only . Below is a graphical representation of a correlation matrix, where strong positive correlations are shown in red and strong negative correlations are shown in blue. Note the symmetry around the diagonal.





*4.3 Factor Analysis*

* The role of factor analysis is to identify the main factors that contribute to the variability of the data. While there may be hundreds of indicators, perhaps only about a dozen control the variability. The others are not important. The problem with Factor Analysis is that it is a blunt instrument. As we saw previously in the discussion of weak signal analysis, often there are significant signals within the indicators that do not appear to have major variability.
* There are two types of factor analysis: *confirmatory* (which tests a hypothesis that the data can be organized into a number of factors, and *exploratory* (which derives factors without knowing how many “buckets” they might fall into). We are using *exploratory* factor analysis, because we do not have a preconceived notion as to how many factors will be most salient in the dataset. *Exploratory* factor analysis makes no initial assumptions about existing relationships of the factors.
* If the objective is to represent a large dataset with a just a few indicators, Principal Component Analysis (PCA) can be used to identify the variables that are responsible for the greatest variation in the data. As with the correlation matrix, PCA is in some sense, removing the data that are redundant. PCA assumes that much of the data are correlated. In fact, if the indicators being used are not correlated, the analysis will not be useful.
* Factors are based on covariance. A factor is not an indicator. It is based on running covariances on all of the indicators and coming up with the factors that bundle together certain indicators at certain weights. The factors make up the total variance of all the indicators.
* The number of factors is NOT the number of indicators from the data set. In the context of factor analysis, factors are sets of indicators. We will discover the actual indicators that define these factors in a following step.
* In deciding the number of factors, the most common method is to generate a scree plot. The term “scree” refers to the rubble of rock at the bottom of a cliff or hillside.
* The scree plot is a two dimensional graph with factors on the x-axis and *eigenvalues* on the y-axis. Eigenvalues represent the variance accounted for by each underlying factor. They are not represented by percentages but scores that total to the number of items. For example, a 74 indicator dataset will theoretically have 74 possible underlying factors, each factor will have an eigenvalue that indicates the amount of variation in the items accounted for by each factor.

:::::Desktop:screeplot_week10.pdf

* From the scree plot you can see that the first couple of factors account for most of the variance, then the remaining factors all have small eigenvalues. A researcher might examine this plot and decide there are 2 underlying factors and the remainder of factors are just “scree”. Another might choose 9. There is a “Kaiser” Stopping rule that states only the number of factors with eigenvalues over 1.00 should be considered for an analysis. Another might choose to keep all of the factors. Novametrics selects the factors based on visual inspection of the scree plots and consideration of any natural breaks.
* The next step is to extract the indicators from the factors. From the factors, we determine which of the indicators are actually responsible for those factors. In statistical subroutine modules (e.g. Stata), the rotation command (‘rotate’) will determine the “factor loadings” or weights of each indicator. ‘Method’ refers to how the rotation is to be done. Our default is almost always the “varimax” criterion.
* With the final set of indicators, we can determine weightings through regression. The p-value can be used to determine a weight that represents each indicator's correlation with the outcome.
* As an example, if we were looking at the district level for the relationship of various indicators to conflict, we run a regression of conflicts per capita on the indicators. The conflicts are the outcome variable, and the indicators are independent variables.
  1. To calculate the index score weighting, we multiply the normalized value of a given indicator from a given district in a given year by the reciprocal of the p-value (1/p-value).
  2. Thus, for each normalized indicator Xn, our result is a final value Xn\*:

Xn\*District i Year z = X District i Year z \* *w*Xn1

* 1. In this example, this equation can be used to multiply the weighting by the normalized indicator value for each district, each year, in the entire dataset.
* Composite measures (for each indicator, for each year, for each location) = ((Weighting)\*(Sign-corrected normalized value)). The sum of each value **Xn\*** across all indicators for a given location and year forms the composite measure.
* Composite measured man be ranked based on their distribution. For example, we might assign values of Low, Medium, and High (and fractions thereof), based on natural breaks. In each case, we plot the values, observe natural breaks in the distribution, and assign corresponding low-medium-high categories. For example, based on one distribution, we decided that (-infinity to -0.1)=LOW, (-0.1 to 0.25)=MEDIUM, (0.25 to +infinity)=HIGH. When translated into a geospatial display the values are assigned color codes so that high areas may be coded green and low areas coded yellow.

# **5. Evaluation, Groundtruthing, and Hindcasting**

* Novametrics validates its predictive analytics through ground truthing. Ground truthing is performed by applying the analysis to different areas and evaluating the predictive success. If suitable regions are not available, we use hindcasting. The analysis is adjusted for previous times and the predictions are compared to known outcomes.
* The objective in ground-truthing is to assign probabilities, so that recommendations are provided in quantitative probabilistic terms with established confidence levels. When recommendations are presented quantitatively and in probabilistic terms, risk mitigation strategies can be applied and expected return on investment (ROI) can be calculated.

***6. Visual Presentation***

* A suite of presentation methods is used depending on the nature of the analysis and the decision-making needs of the user community. Methods include graphics, geospatial displays, and other methods for displaying quantitative information that allow for an intuitive understanding of the analytical basis for the recommendations.
* In general, the visual presentations are customized to the decision-making needs of the end user and the nature of the various data streams.

1. The original use of the term ‘wicked problem” is attributed to design theorist Horst Rittel. [↑](#footnote-ref-1)