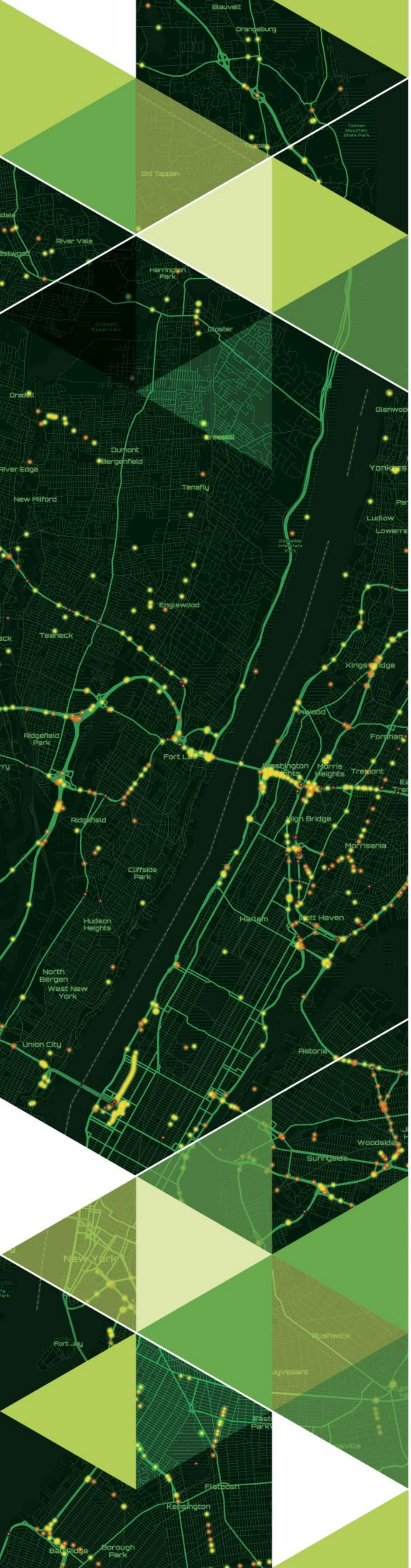


AN ESRI
TECHNICAL PAPER

May 2024

Creating Composite Indices Using ArcGIS: Best Practices

380 New York Street
Redlands, California 92373-8100 USA
909 793 2853
info@esri.com
esri.com

THE SCIENCE OF WHERE®

What's new?

The table below shows a version history of this technical paper.

Version	Date	Revisions
1	February 2023	- First release.
2	May 2024	- Added links to the ArcGIS Online Calculate Composite Index tool and the Social Equity Analysis solution. - New note highlighting additional resources on the topic of index creation. - Added more details to the *Special considerations for variable selection* section. - New section about *Scaling by classes.* - Added reference to PCA as an alternative combination method. - Added details about the popups created by the ArcGIS Pro Calculate Composite Index tool for exploring variable composition. - Added details on how to measure sensitivity to the methods by comparing rank. - Various minor edits for clarity and comprehension.

Contents

Creating Composite Indices Using ArcGIS: Best Practices

Composite indices use simple methods to combine multiple variables into a single indicator variable. Although the methods are not complex, creating an appropriate index is subjective and requires a detailed understanding of the consequences of each decision. This document details concepts and best practices for each step in the index creation workflow.

What is a composite index?

Composite indices combine multiple variables to create a single index variable, otherwise known as a composite indicator. Indices can provide an interpretable metric for subjects that are difficult to measure, such as social vulnerability or entrepreneurial innovation. Indices are widely used across social and environmental domains to help measure progress towards a goal and facilitate decisions.

Many organizations utilize indices in their operations, including:

- United Nations' Human Development Index
- Centers for Disease Control and Prevention's (CDC) Social Vulnerability Index
- European Commission's European Innovation Scoreboard

Indices can be created in different ways across ArcGIS, including using the Calculate Composite Index tool in ArcGIS Pro or ArcGIS Online, or using the Social Equity Analysis solution.

Why consider making your own index?

While it may be tempting to rely on existing indices to help make decisions, or as variables in an analysis, it is often more appropriate to create an index that was specifically designed for your purpose. There are three primary reasons that an existing index may not be applicable:

1. The existing index uses **variables** which are not suitable for the question you are answering. The results of a composite index are entirely dependent on the input variables and methods applied. These should be carefully selected based on the specific question the index is trying to answer, so it is unlikely that an existing index will be completely suitable for your specific question.

   For example, a business organization develops an index to measure economic opportunity in a region. An advocacy non-profit considers using this index to measure business opportunities for minorities but realizes that additional demographic variables are necessary to answer their specific question that focuses on social inequities.

2. The existing index uses different **study area boundaries** than you are interested in. The study area boundaries in a composite index have a strong influence on the results.

   Consider that the results of many preprocessing methods, such as percentiles or z-scores, depend on the values of the locations in the study area. For example, consider two counties in Texas: these may have similar rank when compared with other counties across the entire United States, but they may have significantly different rank when compared with only the counties in Texas. Therefore, these counties will yield very different results relative to each other for a national index versus a Texas index.

   Additionally, national indices are inherently not tailored to specific communities' needs. For example, while a nation-wide stress index might have *% without vehicles* as a variable, this variable may not be a source of stress in a highly walkable area, or in an area that is well-served by transit options.

3. The existing index may have been created using different **spatial units** than the spatial unit you are interested in. While aggregating or disaggregating the index to the desired spatial scale may be tempting, this will not necessarily yield the same results as an index that was calculated at the intended spatial scale, due to a phenomenon known as the Modifiable Areal Unit Problem (Norman, 2006; Openshaw, 1984).

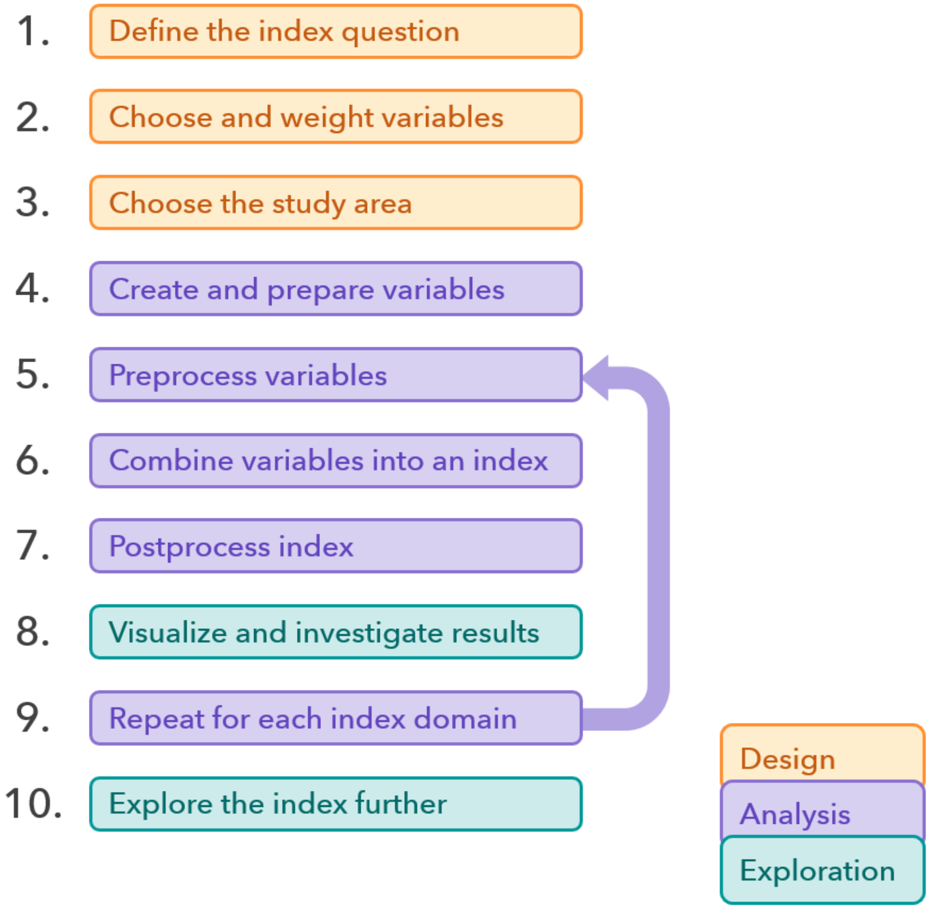
   For example, a local government wants to use an existing national hazard index to identify specific locations, such as neighborhoods, that need remediation against natural disasters. However, the existing index results are at a coarse spatial scale that does not capture the variation in hazard risk factors at the neighborhood level. The local government will get more appropriate results by creating an index using spatial units that match the spatial scale of the intended remediation.

These topics are considered in greater detail below.

Note: Due to the subjectivity and complexity of creating composite indices, we encourage you to read additional resources on the topic. In particular, we recommend the OECD (2008) *Handbook on Constructing Composite Indicators: Methodology and User Guide*, which informed many of the ideas presented in this paper.

The composite index workflow

Creating an appropriate composite index requires a detailed understanding of the workflow. The graphic provides an overview of a ten step process to create an effective index, and the subsequent sections document each step in detail.


1. Define the index question
2. Choose and weight variables
3. Choose the study area
4. Create and prepare variables
5. Preprocess variables
6. Combine variables into an index
7. Postprocess index
8. Visualize and investigate results
9. Repeat for each index domain
10. Explore the index further
Design
Analysis
Exploration

Step 1 Define the index question and dimensions

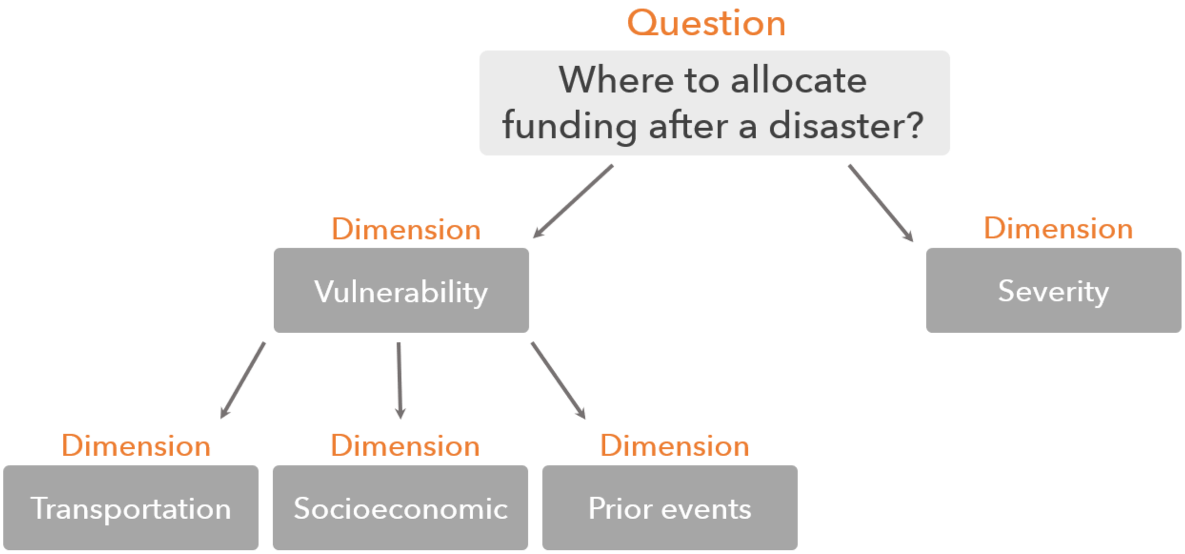
Before deciding which variables should be used in the index, it is important to work with stakeholders to clarify the **purpose** of the index. Identifying the purpose will help define the analysis **question.** Ask the stakeholders - what question will the index be used to answer? The following are some examples:

- The city planning department has received a federal grant for planting trees and asks - where should trees be planted?
- An advocacy group wants to create an index to answer – who is experiencing an unjust burden from environmental hazards?
- A disaster response organization wants to be able to quickly create an index following a natural disaster to help them allocate resources fairly and effectively. They ask - which neighborhoods should receive additional funding after a natural disaster?

Once the **question** has been defined, you should work with stakeholders to clarify their **priorities**. The priorities are the **dimensions** that are important for the index to capture in order to answer the analysis **question**. Ask your stakeholders: which factors matter to you? Some examples:

- The planning department will decide where to plant trees based on:
    a. the **need** of each community; for example, are there people with asthma or who have to wait in the sun for a bus or walk to work?
    b. the **feasibility** of planting trees; for example, how many trees can we plant?
- The advocacy group will identify the burden of environmental hazards in each neighborhood based on:
    a. the **exposure** to environmental hazards
    b. **vulnerability** to the hazards
- The disaster organization will provide funding to neighborhoods based on:
    a. the **severity** of the natural disaster
    b. **vulnerability** to the impacts of the natural disaster

Some dimensions may themselves be complex, consisting of multiple dimensions. For example, the disaster organization will provide funding to neighborhoods based on their vulnerability. However, vulnerability itself has multiple dimensions – people may be vulnerable based on income, but also based on available transportation. This can lead to a hierarchy of dimensions, as shown in the following graphic.



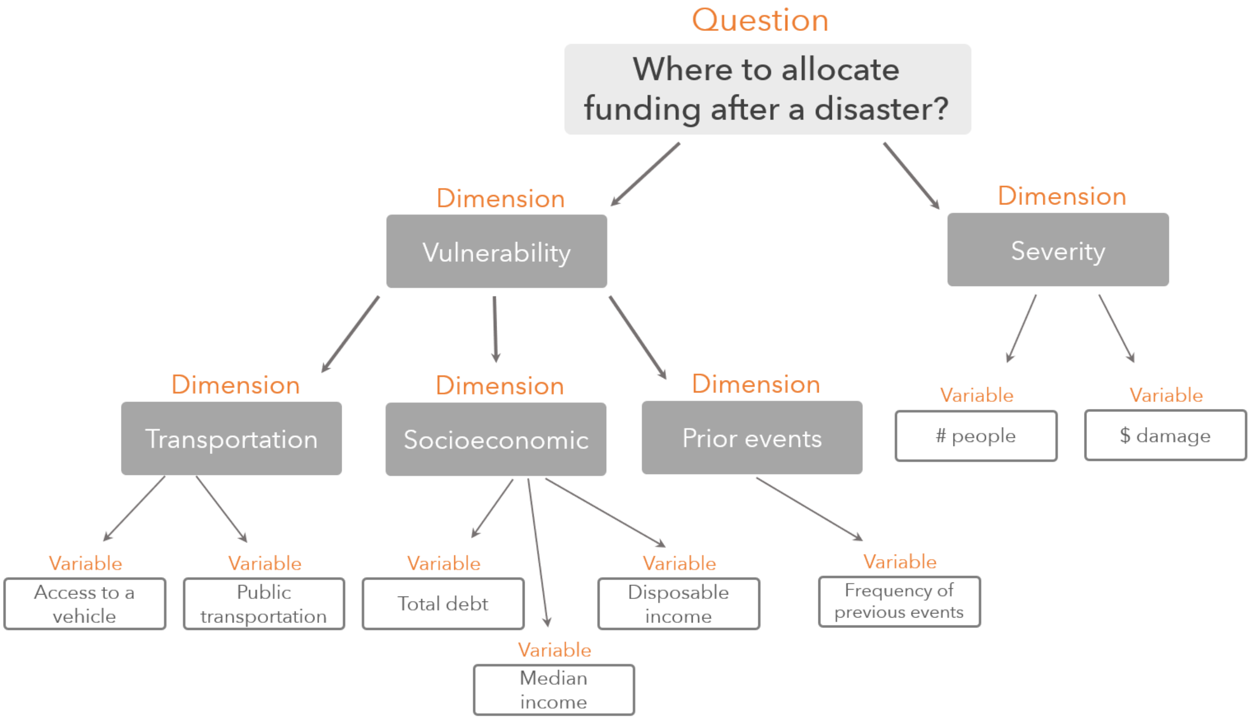
This is an illustrative example; it is not intended to represent a comprehensive index.

To properly specify the variables that should be used in the index, work with stakeholders to define the question and dimensions, then visualize the design in a diagram as shown above. Step 2 will lay out considerations in selecting variables for each dimension.

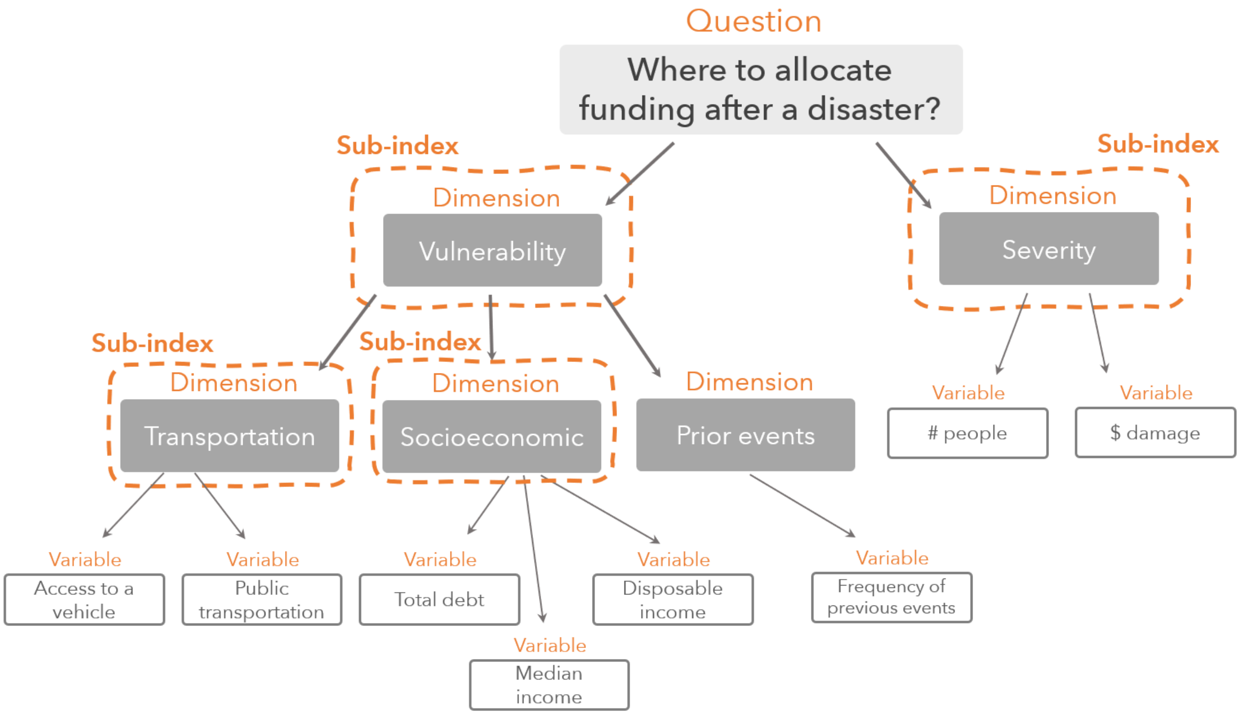
Note: Ideally, the question and dimensions are defined before selecting variables. However, discussions with stakeholders may begin with the variables already defined. In this case, it is still valuable to work backwards to define the dimensions and questions, as this process will help with analysis decisions, such as weighting of variables, which will be covered in subsequent steps.

Step 2 Choose and weight variables

In Step 1, we identified the index question and dimensions. The dimensions are the starting point for selecting variables. Identify the appropriate variable, or variables, that best represent each dimension.



If a dimension can't be represented by a single variable, or if a dimension includes other dimensions, a sub-index may be created to help organize the components of the index and appropriately weight each dimension. For example, continuing the example from above, the following figure shows the variables that were identified for each dimension. Each dimension that is outlined in orange will form a sub-index.



The following sections will explain important considerations in variable selection. Step 9 will explain the utility of sub-indices in more detail.

Selecting variables

It is a best practice to seek the advice of subject matter experts who understand the different factors that influence the dimension. Seek advice from subject matter experts who understand the different dimensions you've identified. Make every effort to consult:

- Domain-specific literature, such as academic journals, subject-matter books, and conference proceedings.
    - For example, search for research articles which have reviewed and recommended variables for your domain, such as Bigi et al. (2021) who do this for flood vulnerability indices.

- Domain experts, such as researchers, professors, and consultants.

- Members of the community, such as those who live in the places impacted by the index.

For example, for the index that aims to prioritize trees planted across a city's neighborhoods, you want to choose a variable (or variables) that represents the potential benefits dimension. To do this, you consult literature that associates shade from planted trees with a reduction in the risk of heat-related illnesses in the elderly. This leads you to include a variable for the total elderly population at each location. Consulting with members of the community further identifies that the adverse effects of lack of shade are felt most by those reliant on public transportation as a result of walking and standing in direct sunlight, leading you to require a second variable measuring the number of residents without access to vehicles. Since the potential benefits dimension relies on more than one variable, it can now form a sub-index.

Instinctually, it may seem appropriate to include all variables about a topic in the index, as each variable may be deemed to add new information. However, this is not the best approach.

*"The inclusion of any additional indicators should be underpinned by clear theoretical reasons, and adding more variables should not be an aim in itself, especially as there is no evidence that having more indicators per domain improves the measurement…" (*Allik et al. 2020*)*

Instead, follow a rationale for including variables that capture the different aspects of the problem in a better way than any individual variable (Allik et al. 2020).

Special considerations for variable selection

- It is important to consider the differences between counts and rates when building an index: a count captures the total measurement in an observed variable, while a rate (for example, incidents per 100,000 people) controls for differences that might influence the count, such as population or area. For example, a fatal crash count is influenced by the population count (e.g. it is expected to be higher in cities with large populations), while a fatal crash rate can highlight places where the portion of drivers experiencing a fatal crash is high, regardless of the population (e.g. emphasizing systemic problems regardless of traffic counts). Keep in mind that rates are not inherently better suited for an index; if the analysis question deals with understanding the total loss of life, for example, the traffic crash count may be more suitable than the traffic crash rate.

- Summary variables may be represented in different ways – for example, a temperature variable could be represented as the average high, or the maximum; an income variable could be represented as the mean or the median. There's no perfect choice – each variable may tell a different story, so the choice should be made carefully in order to fit the index purpose. Some of these options may be more skewed than others. In general, it is better to choose variables with a distribution that is closer to normal, however this is not always the case – steps 4 and 5 will go into more details about how to assess and deal with skewness. In some cases, it may also be an option to first create a sub-index that brings together the summary variables into a single value (see step 9 for more details).

- Be as specific as possible with your choice of variables. For example, if you are creating an index to measure the risk of extreme heat events, you might consider using a variable for the percentage of people who are in poverty. However this lacks specificity – there are other factors that more accurately reflect the true risk due to extreme heat, for example: not having air conditioning, or not owning a car. If the data is available, you should use variables about these specific factors instead. If you don't have access to data about these factors, you could use the poverty variable as a proxy, as long as you communicate this lack of specificity clearly as a potential limitation of the index.

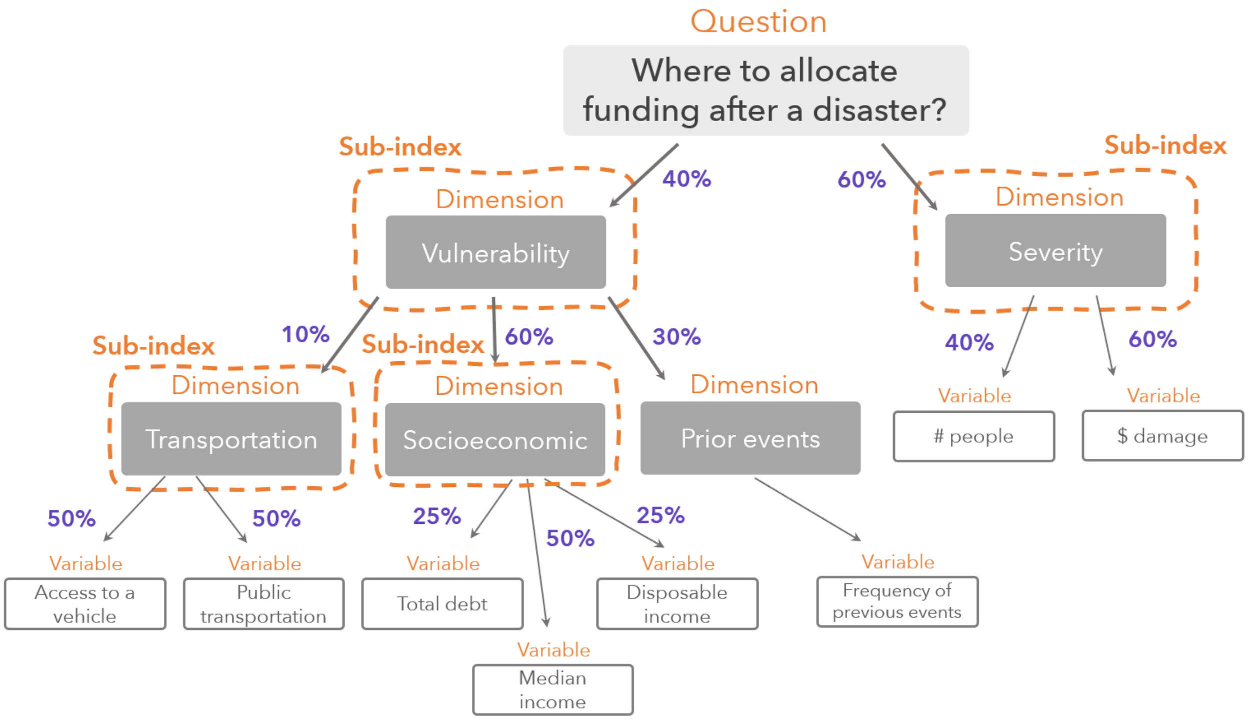
  The lack of specificity is particularly prevalent with variables about who people are, such as the percentage of minorities, or the percentage of people who are disabled. When considering these variables, you should ask whether there is a more appropriate variable – one that captures what people experience, rather than who they are. A variable like minority status may correlate with the phenomena the index is measuring due to inequities, but we should be cautious not to treat the minority status as a driver of the phenomenon (Townsend, 1984). Once we choose more appropriate variables about what people experience, we can still tie the final results back to the impacted people by analyzing the final index in conjunction with the outcome, for example, using regression to test for a relationship between minority status and social vulnerability.

Indeed, in some cases, you might intentionally include variables about who people are, instead of what they experience – this is perfectly acceptable when paired with a good justification. For example, if you were creating an index to help target vaccination clinics, it might be acceptable to use a race variable if the literature proves a valid link between a specific race and vaccine hesitancy.

Setting variable weights

Variable weights represent the relative importance of each variable as it contributes to the index. Weights have a significant impact on the resulting index. If an index does not apply any weights explicitly, the variables will be equally weighted. However, if some variables or sub-indices matter more than others, different weighting should be applied. Whether you choose to keep equal weights or alter weights to favor variables, the choice of weights is subjective and should be backed by a strong rationale.

As in variable selection, consult with subject matter experts for advice on the appropriate weights. Weights can also be defined using other methods, for example, a public opinion poll. Several methods, including more advanced quantitative methods, are covered in detail in OECD (2008), chapter 6.



Step 8 will describe some techniques you can use to evaluate the impact of the chosen weights on the resulting index.

Interactions between variables

Correlation among selected variables may lead to unintentional weighting. If your variables have been selected intentionally to capture different aspects of the problem, it is less likely that variables will be highly correlated. Further, grouping together variables into sub-indices representing each dimension often resolves this problem. Step 8 explains how to diagnose unintentional weighting, and step 9 explains how sub-indices can help remediate unintentional weighting.

Note: As you may have noted, sub-indices can serve several different purposes. Step 9 explains the purposes of sub-indices and methodological considerations in further detail.

Step 3 Choose the study area and the spatial units

As mentioned in **Why consider making your own index?**, it is important to ensure the study area and spatial units are reflective of how the index will be used. The spatial unit corresponds to each location in the index. The study area is the area covered by all of the study units. For example, the study unit may be Peruvian administrative subdivisions, and the study area is the country of Peru.

The choice of study area is important because several common preprocessing methods depend on summary statistics derived from the study units within the study area. For example, some preprocessing methods use the variable range, variable mean, or feature rank in their calculations, and these values will change based on the features that are included in the study area. To ensure the variables are preprocessed in an appropriate way, carefully select the study area which corresponds with the question the index will answer.

It’s also important to carefully consider the study unit used. Does each record correspond to a jurisdiction; a potential store location; a road segment? Identifying the study area will be simple if there is a clear study unit that corresponds to the index question identified in step 1. For example, for a United Nations index that aims to benchmark global development, it would be appropriate to use countries as the study unit because most of the factors that influence development vary at the country level.

However, the study unit may not be evident from the index question. For example, for an index that aims to identify areas to prioritize locations in need of improved internet infrastructure, the spatial unit is not clear. Since the index is about providing services to people, it would be appropriate to use a geographical unit, but there may be multiple options at different spatial scales. Generally, it is advised to use the smallest area (highest resolution) possible, to ensure each geographical unit has little variation within it (Allik et al., 2020). This maximizes the possibility that the variable values are reflective of the people that live in that area. It also reduces the impact of the Modifiable Areal Unit Problem (MAUP) whereby the aggregation of data into geographical units is biased by the aggregation scale (Openshaw, 1984).

It is often the case that a necessary variable is not available at the chosen spatial scale. Consider the possible impacts of the MAUP when it is necessary to aggregate or downscale variables to a different spatial unit (i.e. counties to tracts or vice-versa).

Step 4 Create and prepare variables

The goal of this step is to ensure the variables identified in step 2 are available in a single dataset, and that the values of each variable are well understood and appropriate for the index. This involves several data engineering steps: gathering available data or creating your own data, exploring the variable values, and applying techniques to prepare the variables for analysis.

Numerous organizations publish data that can be used in your index, such as the United States Census Bureau, and the United Nations. The ArcGIS Living Atlas of the World and open data portals like ArcGIS Hub can also be used to find spatial data.

Tools such as Network Analyst and Spatial Analyst in ArcGIS Pro can help you create more effective variables representing spatial phenomena. For example, when creating a healthcare access index, instead of using the count of hospitals in a county, you could use Network Analyst to calculate the average distance to the nearest hospital. Instead of using air pollution station data, you may consider using Spatial Analyst to estimate average pollution levels using an interpolated surface.

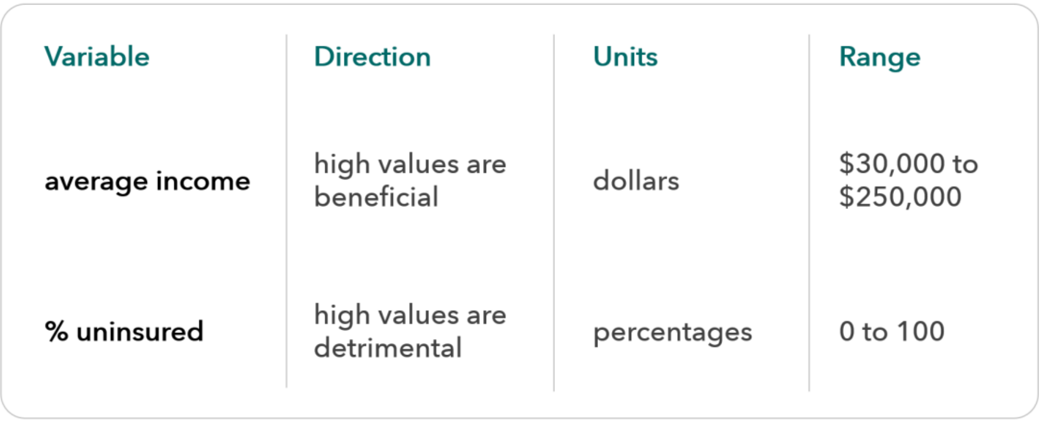
Note: If the study unit does not correspond to the necessary unit, i.e. you have data for tracts but need data for counties, you may have to change the geographic unit by aggregating or downscaling to the desired unit (for example, using Apportion Polygon, or Areal Interpolation tools in ArcGIS Pro). Consider using the Enrich tool in ArcGIS Pro or ArcGIS Online which provides high quality demographic data at various geographies.

If data is to be converted from one geography to another, the conversion should be applied to the raw data, rather than derived data such as rates, percentages, or indices (Norman, 2010).

Once variables have been created and collated, it is important to explore the values within each variable. In ArcGIS Pro, you can use Data Engineering to launch histograms to show the distribution, create maps to display the spatial variation, and calculate summary statistics to understand the values. Note variables with large counts of nulls, skewed distributions, outliers, and any unexpected spatial patterns. If a variable has any missing values, impute values if appropriate (for example, using the Fill Missing Values tool), or find supplemental data if not. Take note of any skewed variables – you may choose to take action on these in the next step.

Step 5 Preprocess variables

Preprocessing refers to the various data preparation steps that ensure variables are compatible and can be properly combined into an index. It is often the case that an analysis starts with incompatible variables. For example, a social vulnerability index with average income and percent uninsured variables may have the following properties that make the variables incompatible:

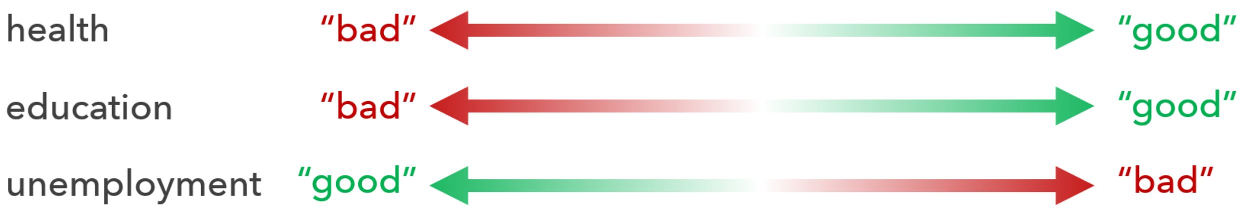


Variable	Direction	Units	Range
average income	high values are beneficial	dollars	$30,000 to $250,000
% uninsured	high values are detrimental	percentages	0 to 100

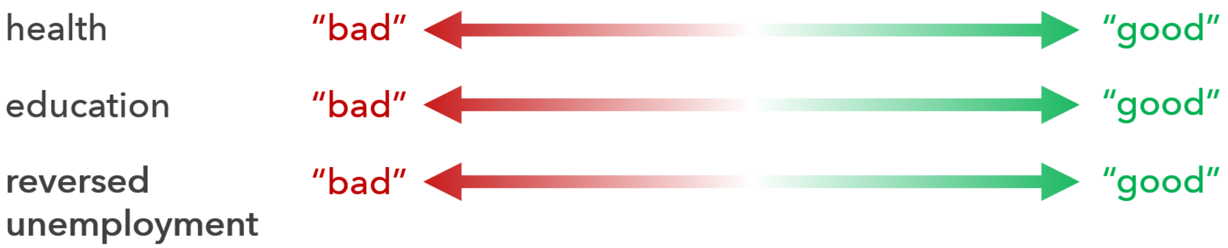
There are various ways to preprocess variables to make them compatible, and this section provides guidance on two key concepts: **reversing variables** to achieve a consistent direction among variables and **scaling variables** to achieve a consistent unit and range among variables.

Reverse variables

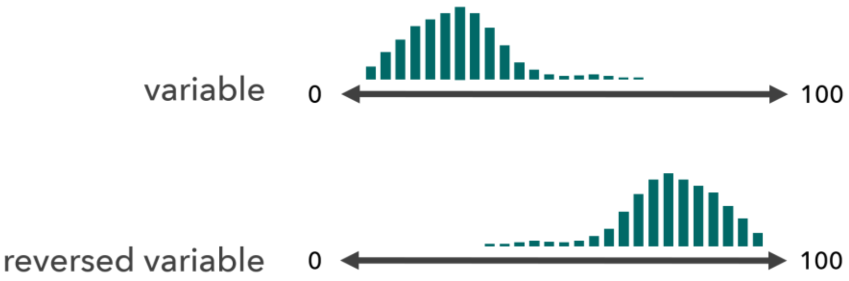
Consider the objective of the index and how the index should change as each variable increases. For example, an index measuring human development in different countries favors countries that **increase** health and education variables, while favoring countries that **decrease** unemployment variables.



Consistency in direction among variables is important so that locations with high values represent a common meaning that can be properly aggregated and represented in the index. The unemployment variable can be reversed so that locations with high values across all variables represent high measures of human development.

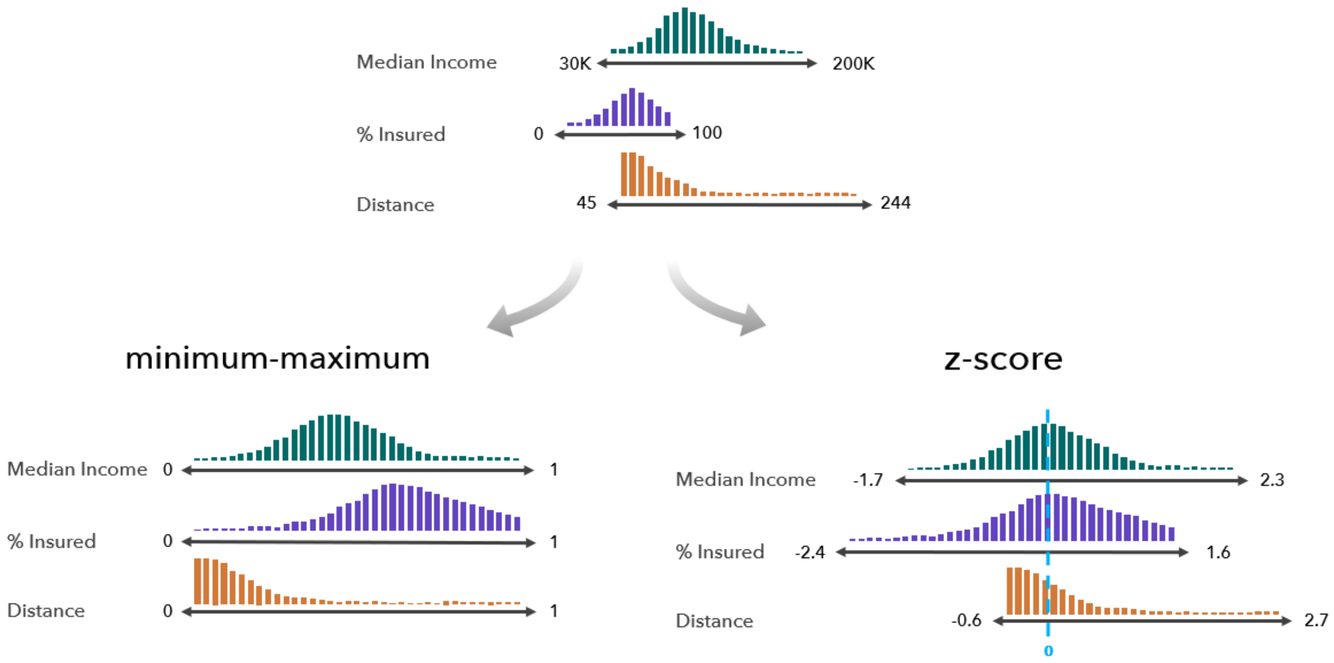


You can reverse a variable by multiplying each value by -1 and scaling the field between the original range of the variable. The result is a variable with the opposite direction that has a mirrored distribution, which preserves the original difference in values.



Scale variables

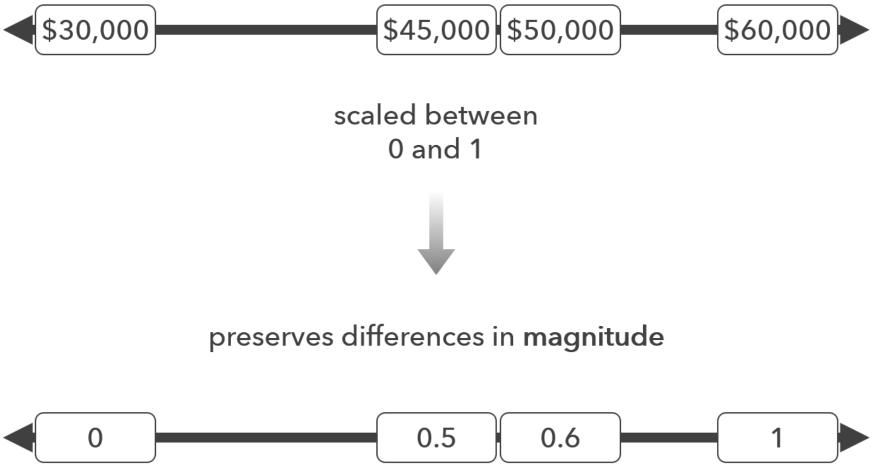
Various methods are available when scaling variables to a common unit and range. For example, the Calculate Composite Index tool in ArcGIS Pro offers seven methods, including scaling to a common minimum and maximum, or standardizing using z-scores.



The various scaling methods have important advantages and considerations. Rather than list them all, this section provides key concepts to help you make the right decisions as you evaluate any scaling option for your index.

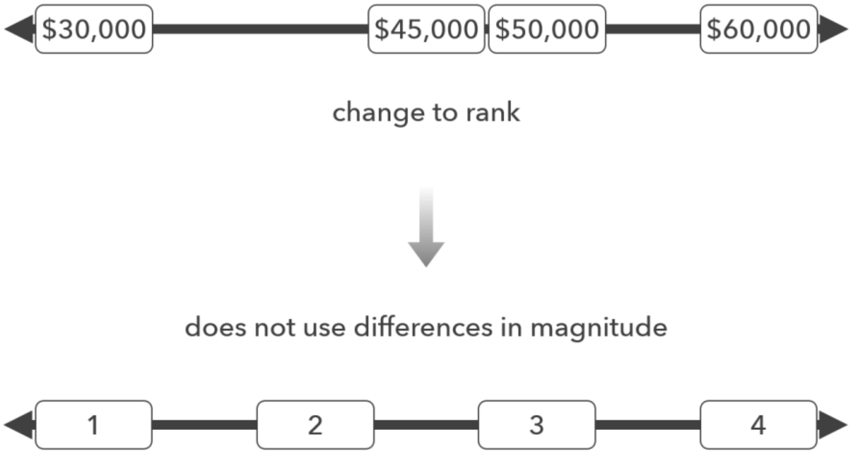
Do you care about differences in magnitude or rank?

As variables are converted to a new consistent unit and range, some scaling methods preserve differences in the magnitude of values within a variable and others use the rank of the values within the variable. Consider a variable with four values representing average income. One option is to scale the variable between 0 and 1, preserving the differences in magnitude.



In other words, the difference between $30,000 and $45,000 is exactly half of the total range in values, and similarly, the difference between 0 and 0.5 is exactly half of the range.

An alternate option is to use the rank. This removes the differences in magnitude and only informs on the position of values within the variable.

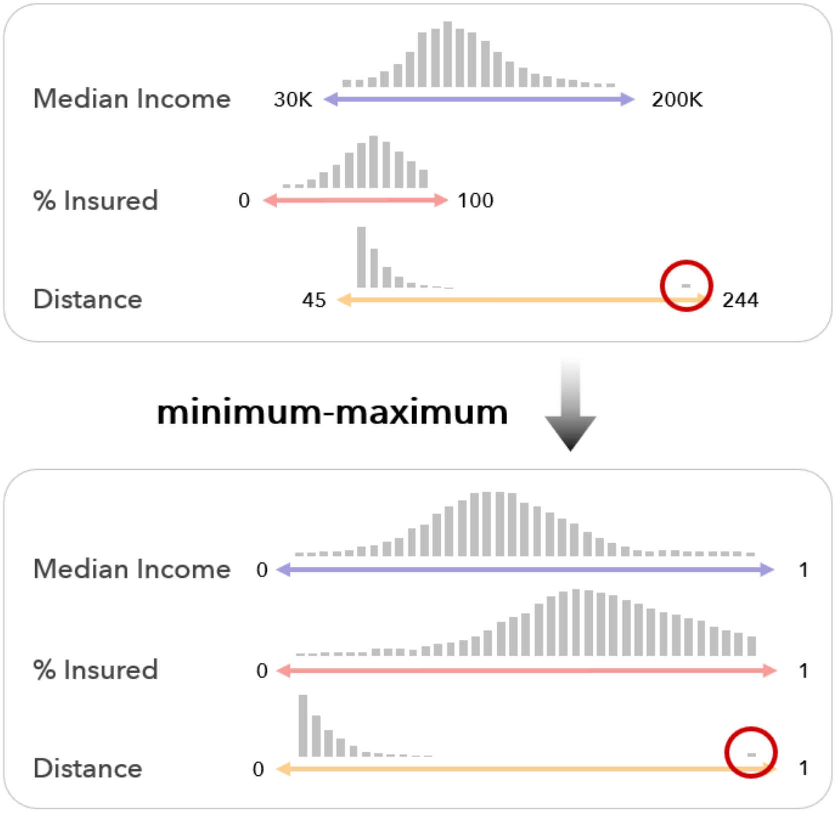


In other words, the magnitude of the difference between $30,000 to $45,000, which is $15,000, does not matter; the only information that matters is that $45,000 is the next value that is higher than $30,000.

Neither approach is inherently better than the other. The right approach depends on the analysis question and design of the index. The following are considerations to reach a decision on whether to preserve magnitude differences or use ranking methods.

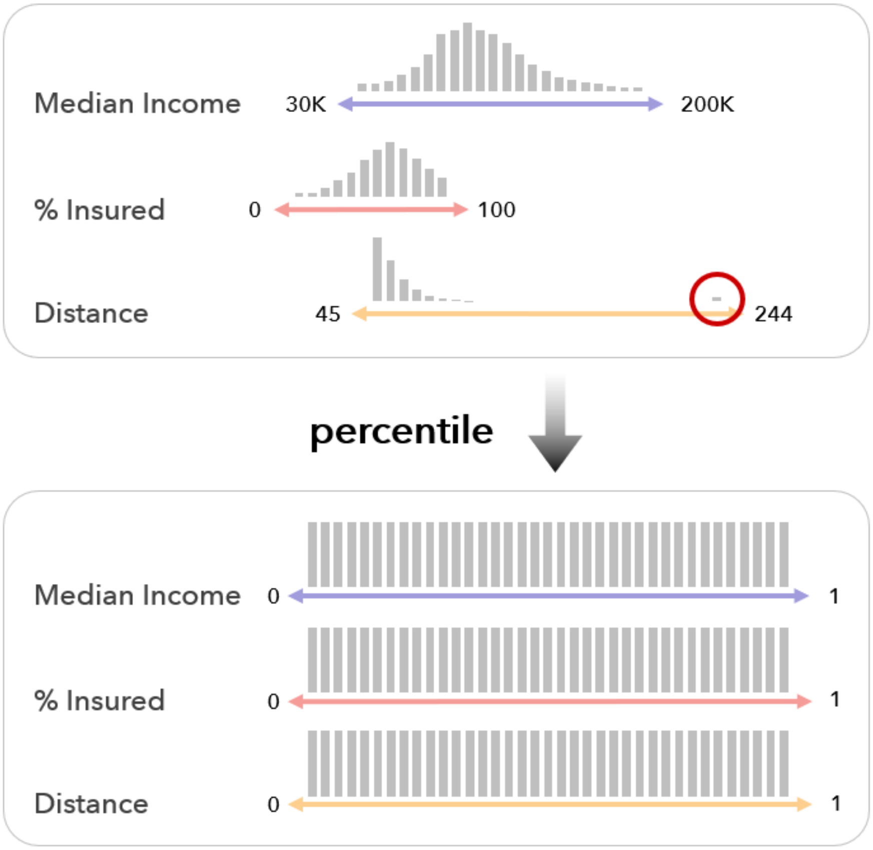
First, consider if the question being answered by the index requires you to preserve the difference in magnitude. In some cases, the magnitude difference is not necessary. For example, a resource allocation index used to find the most deprived locations in a city may use the percentile method, which uses the ranked differences, because it's more important to quantify for each variable whether locations are better or worse than other locations, rather than how much better or worse they are. But an air pollution index may scale using the minimum-maximum method, which preserves differences in magnitude, because the aim is to identify locations that are high in terms of absolute values, not rank.

Second, determine how extreme values should be treated in the index, and examine if any of the variables contain outliers or heavily skewed distributions. Methods that preserve the difference in magnitude will reward (or punish) indicators with extreme values. However, this also makes these methods more susceptible to outliers; when a variable has a single, very large value, most values in the resulting 0 to 1 distribution may be close to 0, which may result in a less meaningful contribution to the index if other variables do not have outliers and skewness in the same direction.



After applying min-max scaling, most Distance values are close to zero due to an outlier

Methods that use the differences in rank remove the impact of extreme values and are robust to outliers, as the variables are changed to a uniform distribution.



After using percentiles, the outlier does not compress Distance values

There are additional options to remediate issues with skewness and outliers, such as variable transformation and standardizing using z-scores. These options are later in this section.

Do you care about a reference value?

Several scaling methods offer ways to compare each variable to a reference value, such as the variable's average, a specified threshold, or an aspirational value. For example:

- A business wants to evaluate the performance of its retail stores against the average in the district. A business analyst standardizes sales indicators using the z-score, resulting in negative values for stores that underperform against the average, and positive values for stores that outperform compared to the average.

- An economist establishes that patients with a travel time greater 110 minutes to their nearest healthcare facility are subject to life-threatening outcomes. A scaling method sets a threshold at 110 minutes, and locations with a higher travel time are "flagged" as 1, while all other locations are given a value of 0.

- A human development index intends to measure progress towards improving life expectancy to 100 years and uses each country's current life expectancy as a variable, containing a range between 45 and 90 years. The life expectancy indicator is scaled using the minimum-maximum method between 0 and 1, but a custom range between 45 and 100 is used to include the aspirational value.

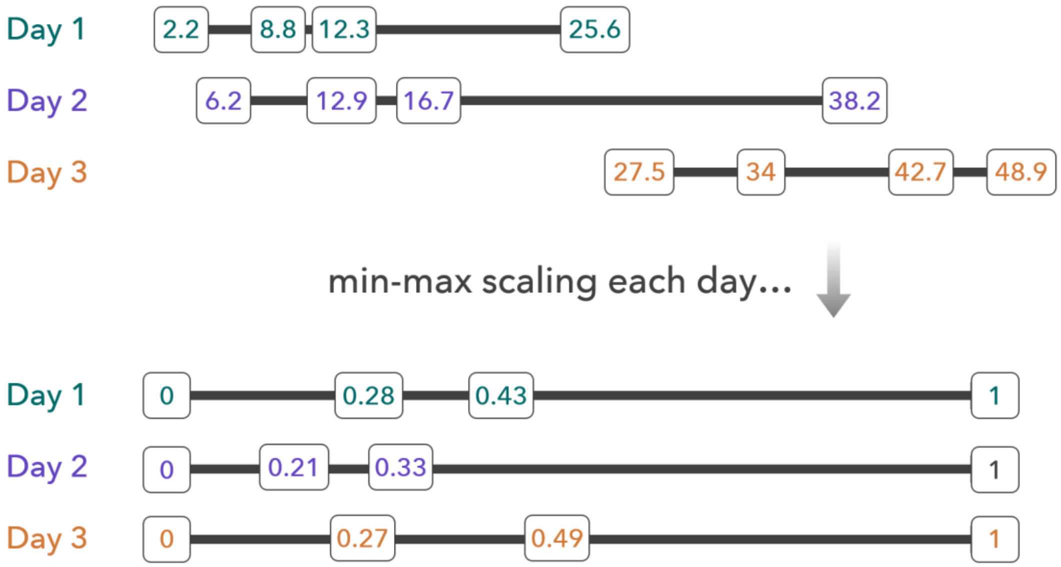
> **Note**: Consider that in this example, a regular min-max scaling of the variable between 0 and 1 would assign the country with the highest life expectancy a value of 1. Using a custom data range, the country with the current highest life expectancy is assigned a value less than 1, as it still has not reached the aspirational goal of 100 years.

Converting variables to the same scale using reference values driven by domain experts can help make an index that more appropriately answers the intended question.

Do you want to compare the index over time?

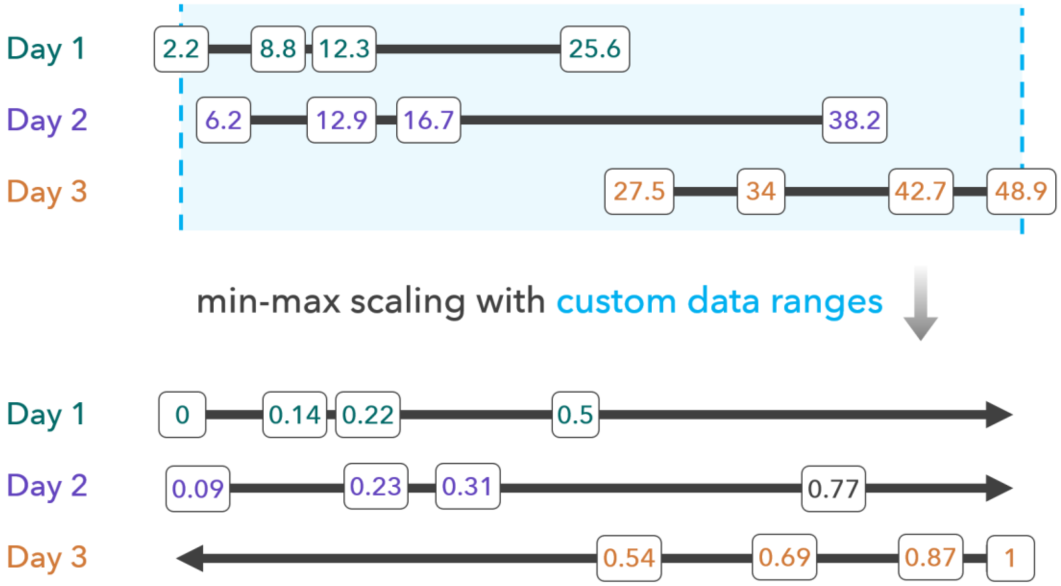
When the goal of an index is to compare changes in index values across time, it is important to ensure that the index calculated in each time step uses the same preprocessing method, and that the possible data ranges across all time steps are consistent.

For example, an air quality index uses an ozone parts per million (ppm) variable. On the first day, the ozone ppm range across all locations is between 2.2 and 25.6. On the second day, the range changes to 6.2 to 38.2. On the third day, the range changes to 27.5 to 48.9. If this variable is preprocessed using min-max scaling between 0 and 1, different ozone ppm values will correspond to the maximum index value each day, and the air quality index result will be incomparable across time.



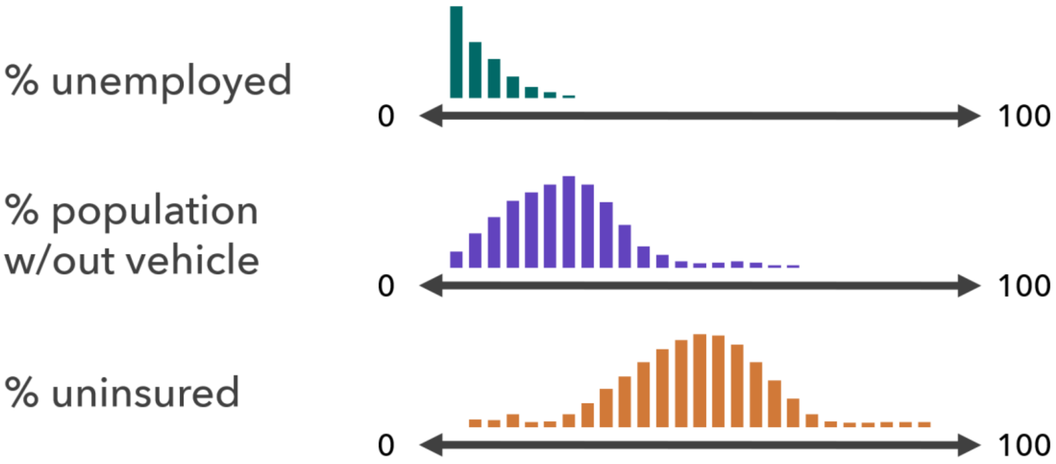
The data ranges across all time steps must be consistent to make the index result comparable across time steps. A useful method to apply a consistent data range in this example is minimum-maximum scaling with a custom data range.

Determine an appropriate custom data range by considering the following two options. First, if the data for all time steps is available, use the **range across all time steps**. Alternatively, if the data for all time steps is not yet available, and you are creating an index that must be compatible with future time steps, consult with domain expertise to determine a **possible range** of ozone ppm values, regardless of the available data.



When is it acceptable to keep variables in their original form?

When variables are already consistent in terms of unit and range, it may be possible to forgo preprocessing the variables. This is often the case when using rate variables, such as percentages. For example, three demographic variables measuring percentages of the population may already be compatible in terms of units.
However, there may still be cases where variables with consistency in unit may not have consistency in range. For example, an index using the percent of the population that is unemployed, the percent of population that does not own a vehicle, and the percent of population without insurance has compatible units (percentages) and directions (high values are detrimental). However, the ranges may differ.

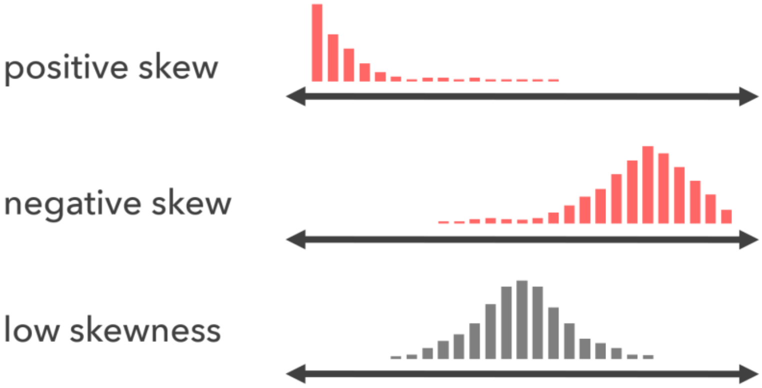


The different ranges may cause unintended weighting, where contributions by variables with higher ranges become more significant to the index. To prevent unintended weighting, scale the percent variables to a new scale (for example, using min-max scaling) so that the ranges are brought to a common unit and range. There may also be cases where different ranges among variables are acceptable. For example, in a national index of disaster risk, there may be variables indicating the percentage of homes at risk of flooding, winter storms, tornadoes, and each of the other hazards. In this case, where each location lies between 0% and 100% for each hazard may be more important than where each location lies relative to the other locations.

Other preprocessing considerations

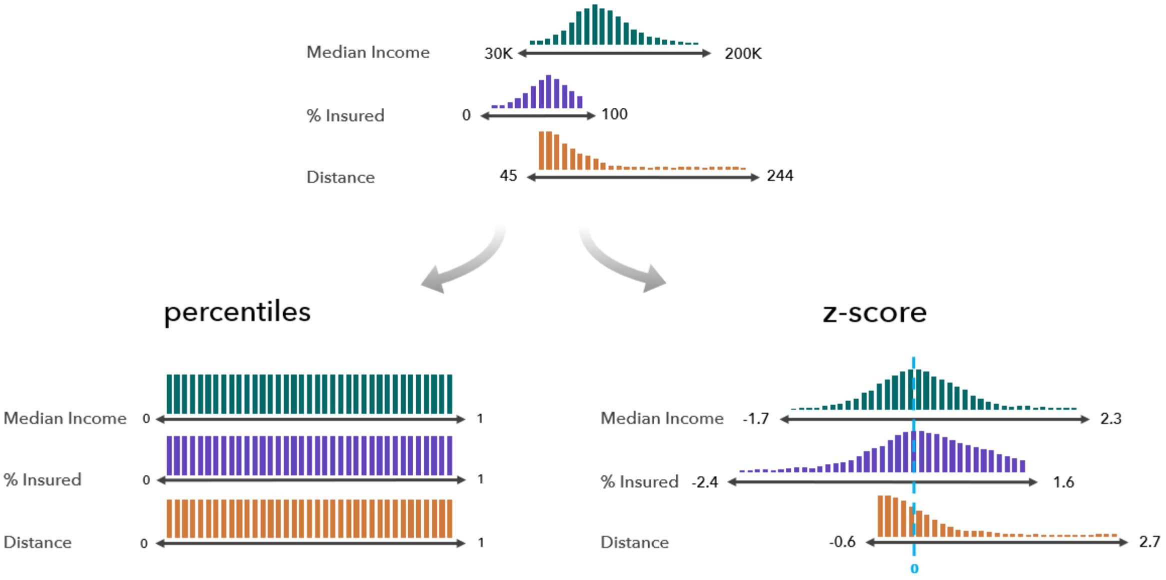
Skewness

Skewness measures the symmetry of the distribution in a variable. When the distribution is symmetrical on both sides, as seen in a normal distribution, the skewness will be close to zero. When distributions have longer tails on one side, skewness will be positive (long tail on the right) or negative (long tail on the left).



High skewness in a variable has the potential to alter the impact of the variable on the index. If one variable is highly skewed and all other variables are not, the skewed variable may have a lesser impact on the resulting index when using methods that preserve magnitude and distribution, such as minimum-maximum scaling. This is because most of the values are compressed to a relatively small portion of the distribution.

There are multiple options to help remediate issues with skewness and data compression. Methods that standardize using the z-score shift the distribution to be centered around its mean, reducing the impact of the skewness, but not eliminating it completely. Scaling methods that use rank, such as the percentile method, convert variables to the uniform distribution, eliminating issues with the shape of the distribution at the cost of losing information about differences in magnitude.



To preserve the magnitude in differences in variables while resolving issues with skewness, consider transforming the variable. A variable transformation changes the shape of the distribution by applying a mathematical operation to each value, such as the logarithm. The resulting shape may have reduced skewness and therefore lead to more consistent results as the variable contributes to the index.

Note: Use exploratory tools such as Data Engineering in ArcGIS Pro to identify skewness, then consider transforming variables with skewed distributions to alter the shape of the distribution.

Outliers

Outliers are variable values that significantly differ from the rest of the values in the variable. Their effect can be similar to the effect of skewness, leading to data compression and diminishing the impact of a variable on the index, particularly when no other features have similar outlier patterns. However, unlike skewness, outliers cannot be remediated by variable transformation, as outliers are often preserved even after transforming the variable.

Scaling methods that use rank eliminate the effect of the outlier in the variable at the cost of losing information about differences in magnitude.

It is important to investigate outliers, as they often correspond to valid data that should be regarded in the measurements of the index. In the cases when an outlier represents data collection errors, efforts should be taken to correct the data, or remove the location from analysis if possible.

Scaling by classes

Variables may be scaled to a consistent unit and range by binning the continuous values into ordinal classes. For example, income in dollars may be classified into five equal classes: 1 - Lowest, 2 - Low, 3 - Medium, 4 - High, 5 - Highest. In ArcGIS Pro, the Reclassify Field tool can be used to create these classes.

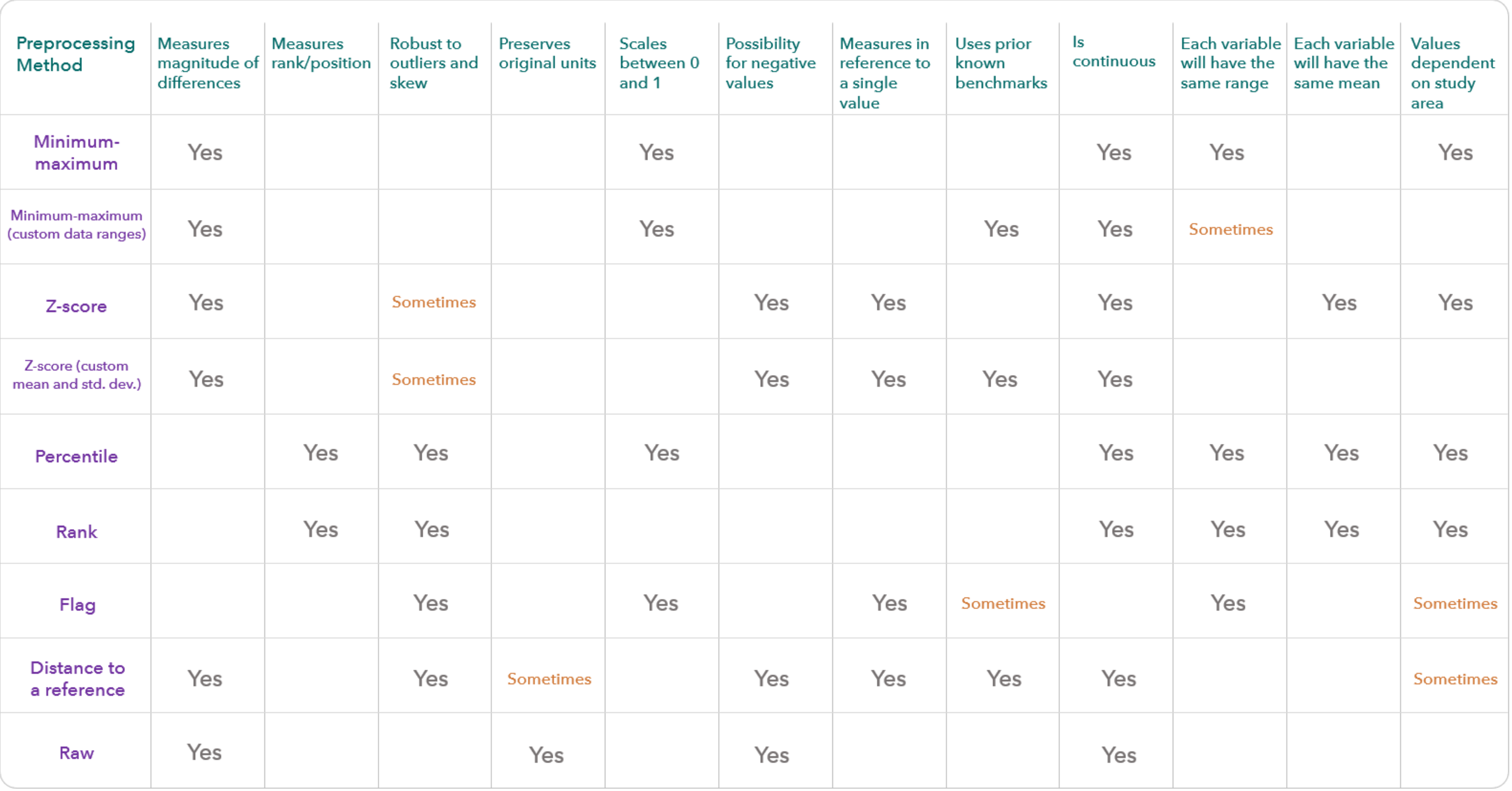
It is generally recommended to avoid scaling by classes, as it removes much of the variability in the data. The arbitrary cut-off points can introduce significant differences between locations with similar values - for example, two features with a one dollar difference in income can be assigned to a different class due to the choice of the cut-off point. Maintaining the variability by using a continuous scaling method (such as minimum-maximum scaling) would more accurately reflect the similarity between the locations.

In practice, however, some input variables may simply not be available as continuous values. For example, an index might include a variable about neighborhood condition, which is a qualitative ordinal variable between 1 and 4. In this case you should either treat this variable as continuous (i.e. apply the continuous scaling method to this variable) so you can keep all other input variables continuous, or consider scaling all other variables in the index into the same number of classes.

Method comparison

With the previous considerations, you may compare the characteristics of each method as you select a preprocessing option for your index. The figure in the following page summarizes the characteristics of common preprocessing methods, many of which are available in ArcGIS Pro via the Calculate Composite Index tool.

Note: For other methods and to learn more, see OECD (2008).



Preprocessing Method	Measures magnitude of differences	Measures rank/position	Robust to outliers and skew	Preserves original units	Scales between 0 and 1	Possibility for negative values	Measures in reference to a single value	Uses prior known benchmarks	Is continuous	Each variable will have the same range	Each variable will have the same mean	Values dependent on study area
Minimum-maximum	Yes				Yes				Yes	Yes		Yes
Minimum-maximum (custom data ranges)	Yes				Yes			Yes	Yes	Sometimes		
Z-score	Yes		Sometimes			Yes	Yes		Yes		Yes	Yes
Z-score (custom mean and std. dev.)	Yes		Sometimes			Yes	Yes	Yes	Yes			
Percentile		Yes	Yes		Yes				Yes	Yes	Yes	Yes
Rank		Yes	Yes						Yes	Yes	Yes	Yes
Flag			Yes		Yes		Yes	Sometimes		Yes		Sometimes
Distance to a reference	Yes		Yes	Sometimes		Yes	Yes	Yes	Yes			Sometimes
Raw	Yes			Yes		Yes			Yes			

Note: For more information on these methods, consult the Calculate Composite Index tool documentation.

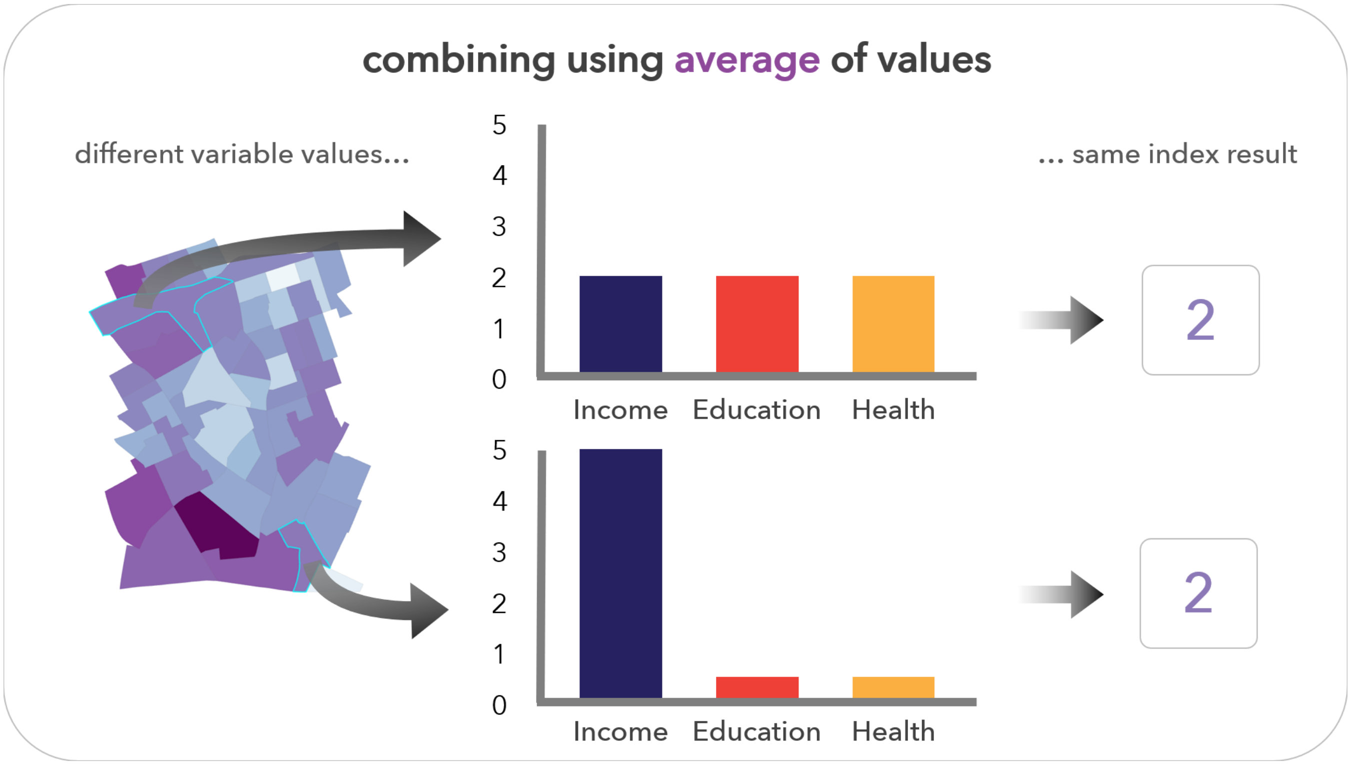
Step 6 Combine variables

Once variables are compatible (discussed in Step 5 Preprocessing Variables), the next step is to combine them into an index. There are various methods to combine variables, such as the sum, mean, multiplication, or geometric mean. The most common methods are categorized into two groups: **additive** (including sum and mean), and **multiplicative** (including multiply and geometric mean).

One of the most important concepts to consider when selecting an aggregation method is whether high values in a variable should compensate for low values in other variables, a concept known as compensability.

Compensability

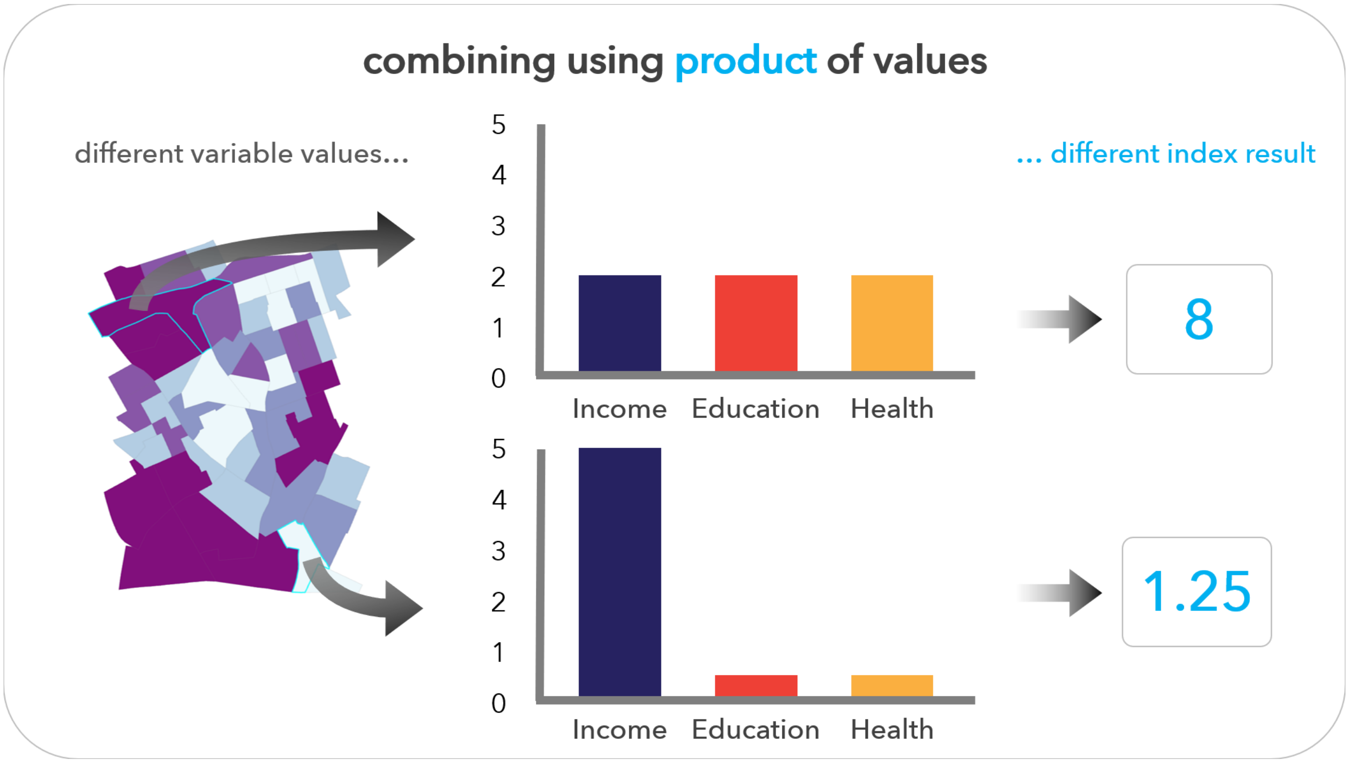
Compensability refers to “the possibility of offsetting a disadvantage on some indicators by a sufficiently large advantage on other indicators” OECD (2008). Additive methods are compensatory, allowing for a sufficiently high variable to compensate for low values in another variable.



Additive methods allow high variable values to compensate for low variable values

Multiplication and geometric mean are partially non-compensatory methods. Multiplicative methods have two important characteristics:

1. **Raising input values together is more highly rewarded:** All variable values must have high values to receive a high index value, and even a small number of variables with low values will greatly lower the resulting index.



combining using product of values

different variable values...

... different index result

Multiplicative methods do not allow high variable values to compensate for low variable values

2. **Raising lower values is more highly rewarded:** Consider the following two examples using geometric mean combination:

Case 1	Variable 1	Variable 2	Variable 3	Variable 4	Result
Start	50	50	50	40	47.29
Changed Variable 4	50	50	50	50	50

Case 2	Variable 1	Variable 2	Variable 3	Variable 4	Result
Start	50	50	50	10	33.44
Changed Variable 4	50	50	50	20	39.76

Case 1 has an uplift of 2.71 when variable 4 increases 10 units, while Case 2 has an uplift of 6.33 with the same unit increase. The difference being that Case 2 had a lower starting value for the changed variable.

Consider that the effects of using a partially non-compensatory method are dependent on the direction of the variables. In addition to ensuring that the variables all have the same direction, also ensure that the direction is such that the non-compensatory method has the intended effect. For instance, since raising lower values is more highly rewarded, confirm that low index values are indeed, intended to be more highly rewarded. If they are not, you should reverse the direction of all input variables before combining variables.

When to use a non-compensatory method?

To decide if a **non-compensatory** method is needed, consider the following questions:

- **Is it acceptable if a high value in one variable and a low value in another variable "cancel out"?** For example, in an influenza risk index, should low overall health values in a location be offset by high vaccination rates? If so, you may use a compensatory method, such as sum or mean. If not, use a non-compensatory method such as geometric mean.

- **Is it acceptable if the index is high when only one variable is sufficiently high**? For example, should a country with an extremely high gross domestic product and low values in all other metrics have a high human development index score? If so, you may use a compensable method, such as sum or mean. If not, use a non-compensatory method such as geometric mean.

If you answered no to both questions, you might be considering using a multiplicative method instead of the more common additive methods. However, multiplicative methods have some drawbacks, details of which are included in the following section.

Note: Compensability does not mean correlation. Correlation is a measure of whether variables tend to increase or decrease in tandem. Whereas compensability is a characteristic of the method used to combine variable values, specifically whether variables can offset the values of each other.

Additional considerations

- Aggregation methods that use the mean, such as the arithmetic mean or the geometric mean, often result in more predictable values that fall within the range of the input variables. For example, the mean or geometric mean of percentage variables will be within 0 and 100, while the sum or multiplication will have a range that depends on the number of variables and magnitude of values.

- When variables contain negative values, for example, when the variable is scaled using z-scores, use an additive method, such as the sum or mean. Multiplicative methods can result in unexpected results when variables have negative values, as the sign of the result depends on how many variables with a negative sign are included in the calculation. Variables which contain zeroes can cause similar unexpected results with multiplicative methods – their result will always be zero, even if all of the other variable values are high.

- Variable weights inform on the relative importance of each variable. For additive methods, weights should be used as a multiplier for each variable value. For multiplicative methods, weights should be used as exponents, and each variable value is raised to the power of the weight.

- Other aggregation methods exist beyond the commonly used sum, mean, multiplication, and geometric mean:
    - The MPI and AMPI methods detailed in Maziotta and Paretto (2016) and Maziotta and Paretto (2018) are non-compensatory approaches that maintain the characteristics of an additive approach.
    - The dimension reduction technique Principal Component Analysis (PCA) is sometimes utilized to create indices, for example in Kolak et al. (2020). PCA can be run in ArcGIS Pro by using the Dimension Reduction tool. However, this method has also gained some critique, for example, Spielman et al. (2020) analyzed a social vulnerability index created using PCA and found it to be lacking internal and theoretical consistency.
    - OECD (2008) also details several other methods that can be used to combine variables.

- If your index includes sub-indices, there are additional considerations in your choice of method to combine variables. Step 9 contains more details.

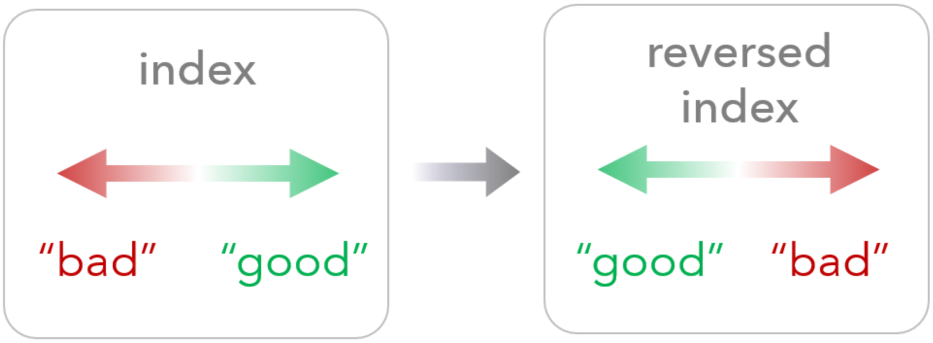
Step 7 Postprocess the index

Once combined, the index may need to be postprocessed to ensure it can be readily interpreted and used. This section outlines considerations for postprocessing workflows that can be applied.

Index direction

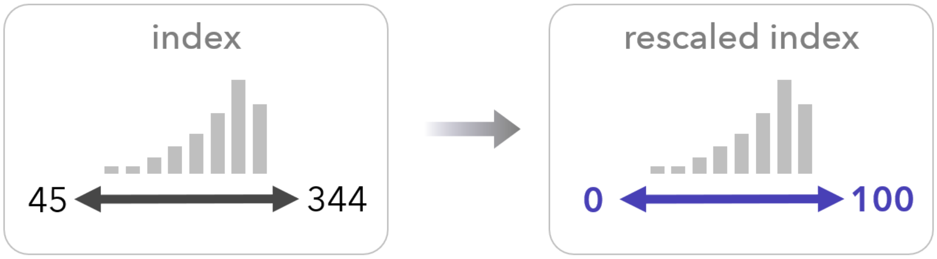
High values in indices may represent beneficial or detrimental conditions. It is important to confirm that high values align with the purpose of the index. For example, a social vulnerability index may have high values represent high vulnerability to adverse circumstances (**high values** represent places where more remediation needed); while a human development index may have high values represent highly developed regions (**low values** represent places where more remediation needed).

If the direction of the index does not align with its intended purpose and title, the index should be reversed.



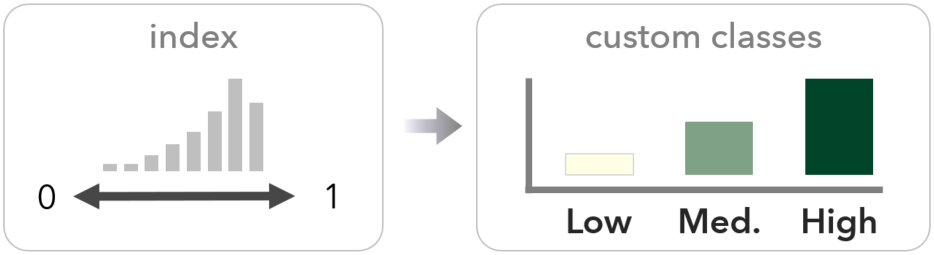
Index scale

Index values for a particular location are only interpretable when the minimum and maximum possible values are known. Consider that an index value of 3 only gains meaning if you know that the index ranges between 0 and 10. However, the combination step may result in index ranges that make it difficult to interpret results for a particular location. For example, interpreting the performance of a location with a value of 412 when the range is between 262 and 635 is not immediately intuitive. In this case, you can use minimum-maximum scaling to scale between 0 and 100, resulting in the location having an interpretable index value of 40.

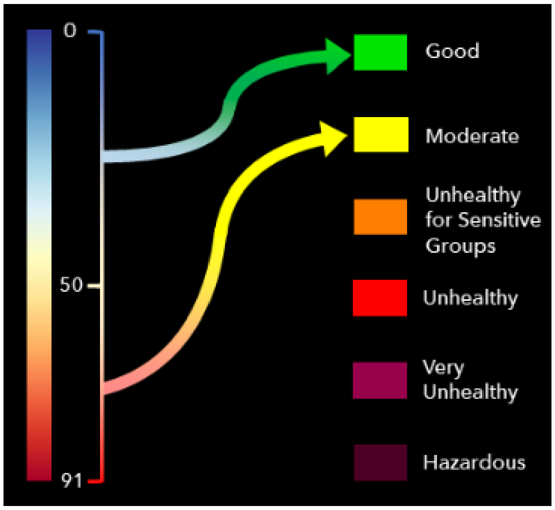


Index classification

While indices are typically numeric representations of complex phenomena, it is often easier for an audience to interpret an index that has been categorized into classes. These classes could be based on statistical characteristics of the data, for example, standard deviations, or quantiles. Alternatively, custom classes could be defined in order to categorize results into recognizable classes, like “Low”, “Medium”, and “High”.



Custom classification of the index can be particularly useful when the categories have direct ties to specific thresholds or intended policy. For example, an air quality index may be more impactful when the combined pollutant indicator range is classified into categories that inform the public on specific risks.



An air quality index classifies pollutant ranges for public interpretation

Step 8 Visualize and investigate results

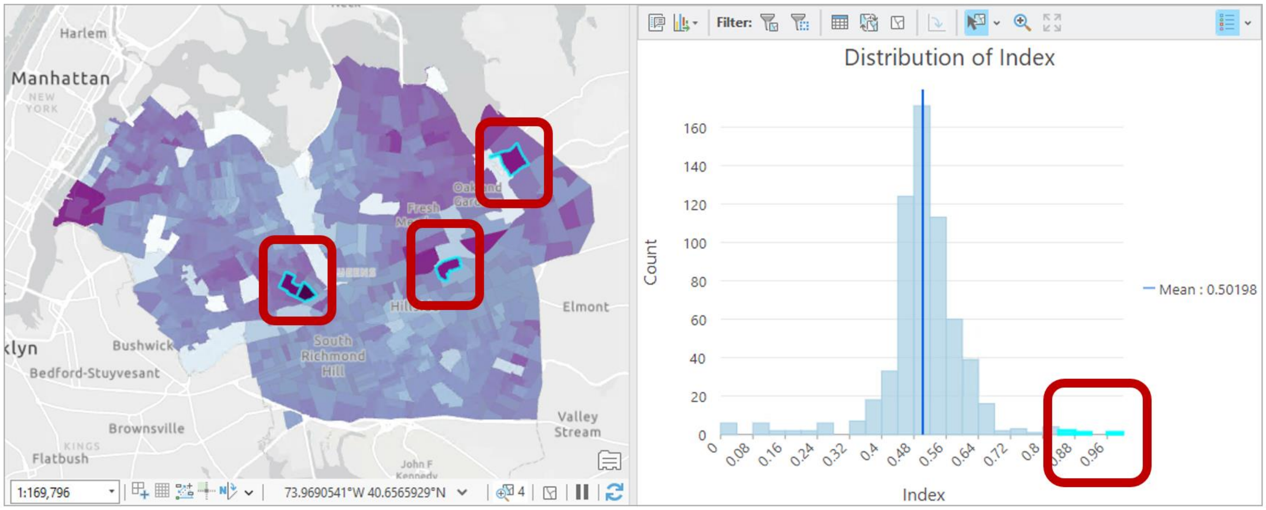
The development of the index involves a series of subjective steps: from the selection and weighting of variables to the choice of preprocessing, combination, and postprocessing methods. A thorough evaluation of the index is necessary to determine how these choices impacted the results.
Consider the following questions as you evaluate the index:

- What is the distribution of the resulting index?
- What are the spatial patterns of the resulting index?
- What is the composition of the index in different locations?
- How important was each variable to the index?

These questions can be addressed by creating data visualizations such as maps and charts. You can also coordinate with stakeholders to review the results to make sure they are as expected.

Index distribution and spatial patterns

A useful first step is to evaluate the index distribution. Use a histogram and a choropleth map to evaluate the range, average, variation, and general patterns of the index. Explore high and low values on the map by making selections and identifying these locations in the map, paying special attention to outliers. Where possible, compare the results with any available domain information to confirm that the index follows expected patterns.



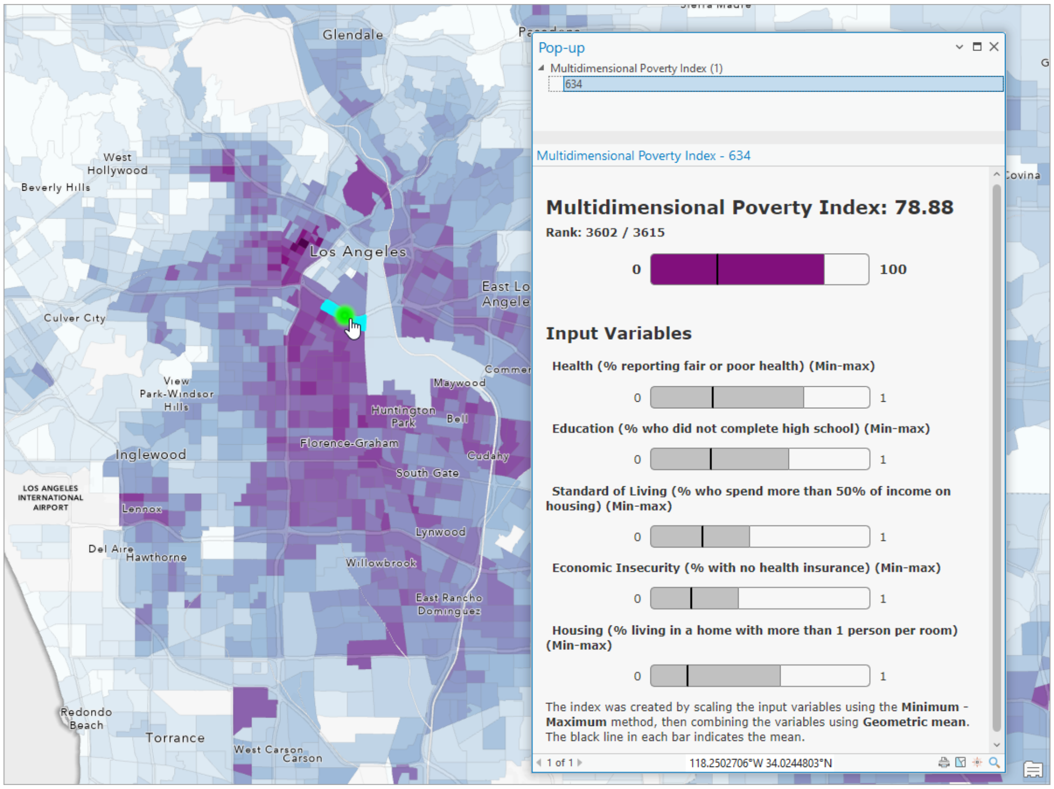
extremely high index values correspond to known high-cost neighborhoods

Outliers in the index may represent important locations to highlight and remediate or may represent a different process driving unexpected index values.

To evaluate local outliers (locations whose index value differs significantly from its neighbors), consider running a Cluster and Outlier Analysis.

Variable composition

A useful next step after gaining a general understanding of the index is to identify the composition of the index in different locations of interest: consider locations you're familiar with and check if the variable values that compose the index make sense. Interrogate locations with high index values and verify which variables are driving high index results. Repeat your exploration for locations with low index values. After running the Calculate Composite Index tool in ArcGIS Pro, click features on the map to explore the variable composition in the popups.



Pop-up

Multidimensional Poverty Index (1)

634

Multidimensional Poverty Index - 634

Multidimensional Poverty Index: 78.88

Rank: 3602 / 3615

0 **100**

Input Variables

Health (% reporting fair or poor health) (Min-max)

0 1

Education (% who did not complete high school) (Min-max)

0 1

Standard of Living (% who spend more than 50% of income on housing) (Min-max)

0 1

Economic Insecurity (% with no health insurance) (Min-max)

0 1

Housing (% living in a home with more than 1 person per room) (Min-max)

0 1

The index was created by scaling the input variables using the **Minimum - Maximum** method, then combining the variables using **Geometric mean.** The black line in each bar indicates the mean.

1 of 1

118.2502706°W 34.0244803°N

An example of a popup created by the Calculate Composite Index tool in ArcGIS Pro

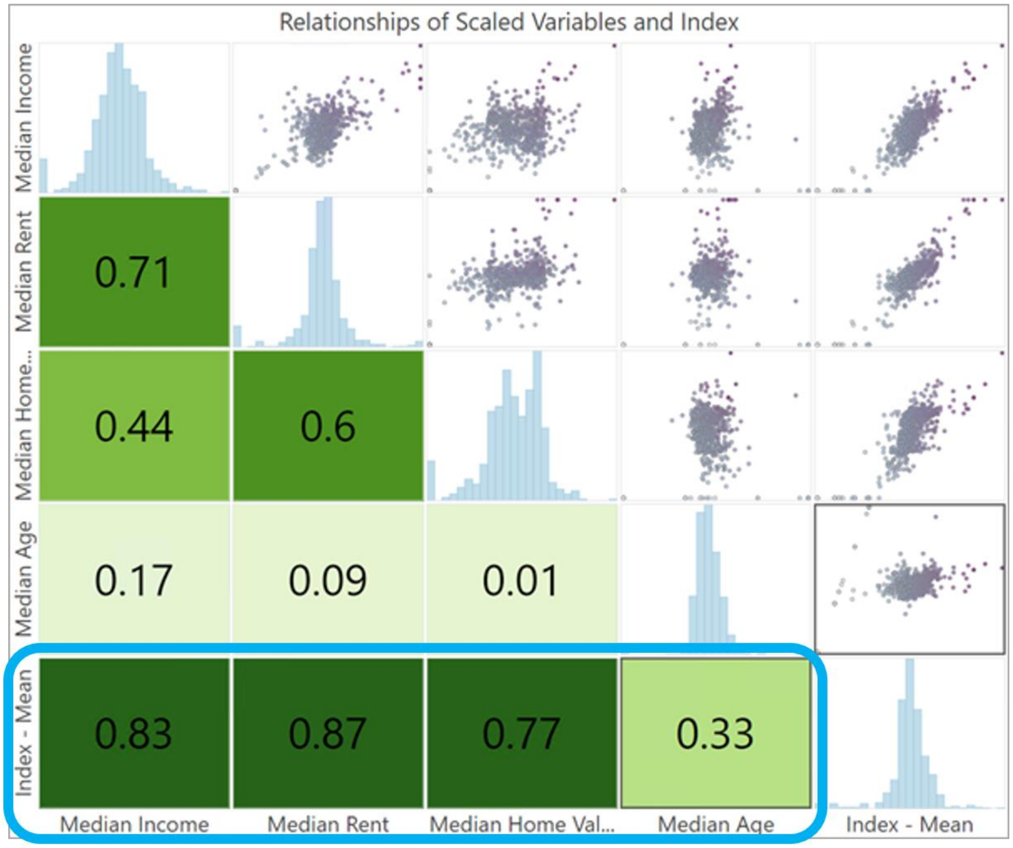
Consider whether the aggregation methods allowed locations with vastly different variable values to have similar index results, and whether this is reasonable for the purpose of the index. Consult with step 6 and the “compensability” section to make alternate aggregation choices if necessary.

You could also consider mapping the coefficient of variation (the standard deviation divided by the mean) of the variables at each location. Where the coefficient of variation is large, there is high variation between the variable values. Study these locations to assess whether the index results make sense.

Variable relationships

It is also important to evaluate which variables are most important to the resulting index and whether any variables are not contributing meaningfully or whether any single variable is driving most of the variation in the index.

Use a scatterplot matrix of the index and its preprocessed variables to identify relationships and visualize measures of correlation, such as Pearson’s R.



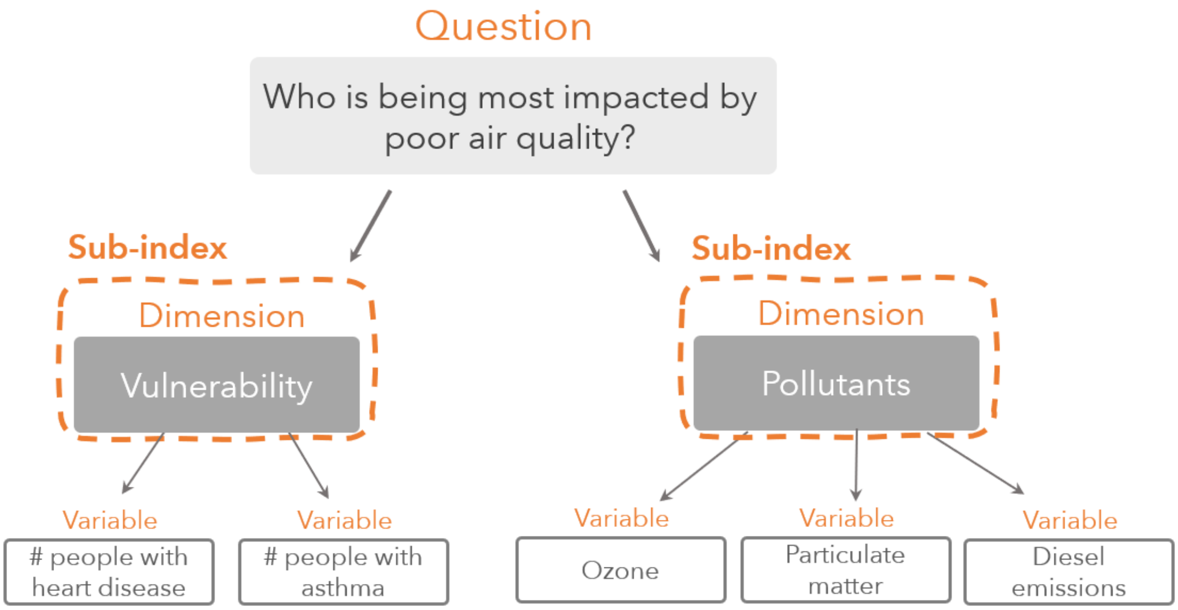
In this example, the median income, median rent, and median home value variables all have high correlation with the index (and amongst each another). The median age variable does not have high correlation with the index. This indicates that index results have been driven predominantly by the income, rent, and home value variables.

Variables with high or low correlation are not necessarily problematic, but it is important to consider whether highly correlated variables are leading to unintentional weighting. In spite of using equal weights, each of the three income related variables mattered much more to the index than the age variable due to their correlation with each other. This might indicate that two of the three variables should be removed, or that the income variables are representing a single dimension so these should be a sub-index. Step 9 will explain how sub-indices can help avoid unintentional weighting. If you did intend to weight some variables more than others, the scatterplot matrix can give you a sense of whether these had the intended effect, however this interpretation is limited if the variables are correlated.

The index is a measure of variance – index values change as the variables change. Therefore, variables with low variance contribute less to the index. An extreme example of this is a variable where every value is the same: it would not have any contribution to the index. On a scatterplot, variables with low variance may surface due to low correlation with the index. To confirm this, create box plots of the preprocessed variables. You may consider removing these variables since they are contributing very little to the results, or you could switch the preprocessing method to z-score standardization for all variables, as this equalizes variance across the variables.

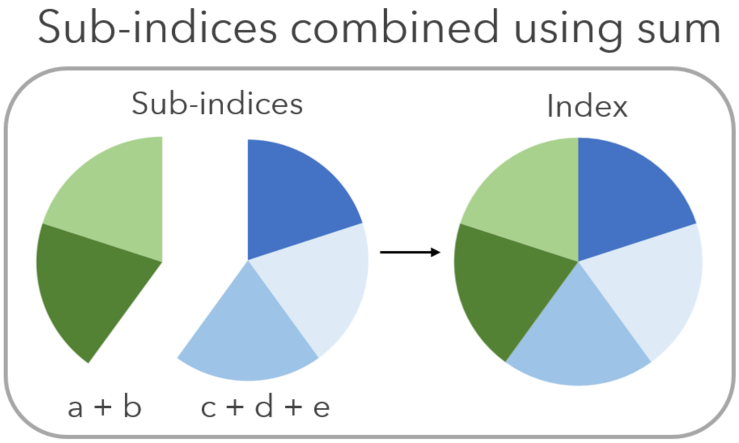
Step 9 Repeat for sub-indices and combined index

So far you have considered the process of creating a simple composite index, selecting variables, preprocessing them, combining them, and postprocessing the index. In some cases, your index may itself become an input variable to create a broader index, as a sub-index representing a dimension. For example, the index example in the image is defined by two dimensions, which each comprise a sub-index.

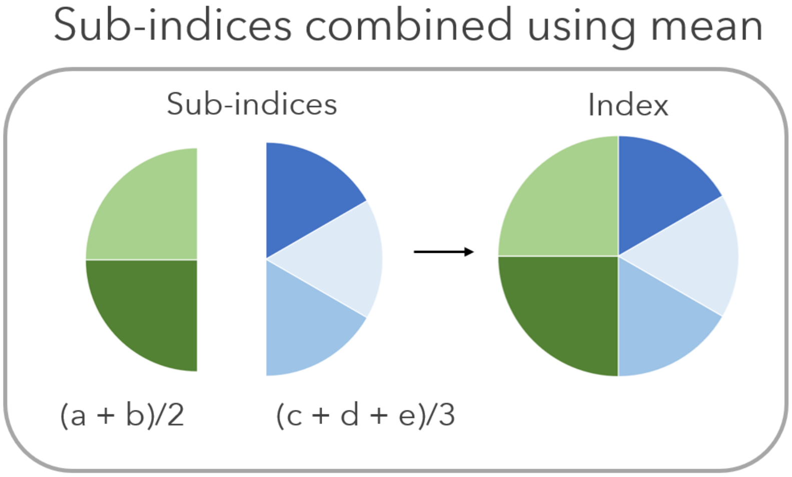


When creating an index which contains sub-indices, most of the methodological considerations are the same as before. However, there are some additional important considerations which can impact the choice of preprocessing and combination methods. To understand these, it is important to understand the utility and benefits of sub-indices.

The first benefit of sub-indices is to **thematically group variables** for each dimension. This can help you communicate to stakeholders and end-users how the index was calculated, and which dimensions are driving the results, since each sub-index can be interrogated individually at each location. The CDC's Social Vulnerability Index takes this approach – each of their four sub-indices is a thematic group, leveraged so that the different dimensions of the index can be mapped and understood in isolation. The CDC's methodology is to sum together each variable within each sub-index, then sum together the sub-indices, using equal weights at each stage. Therefore, the final index value for each location is no different than if all of the original variables were summed together. The image shows this concept – each of the five variables has an equal contribution to the index.



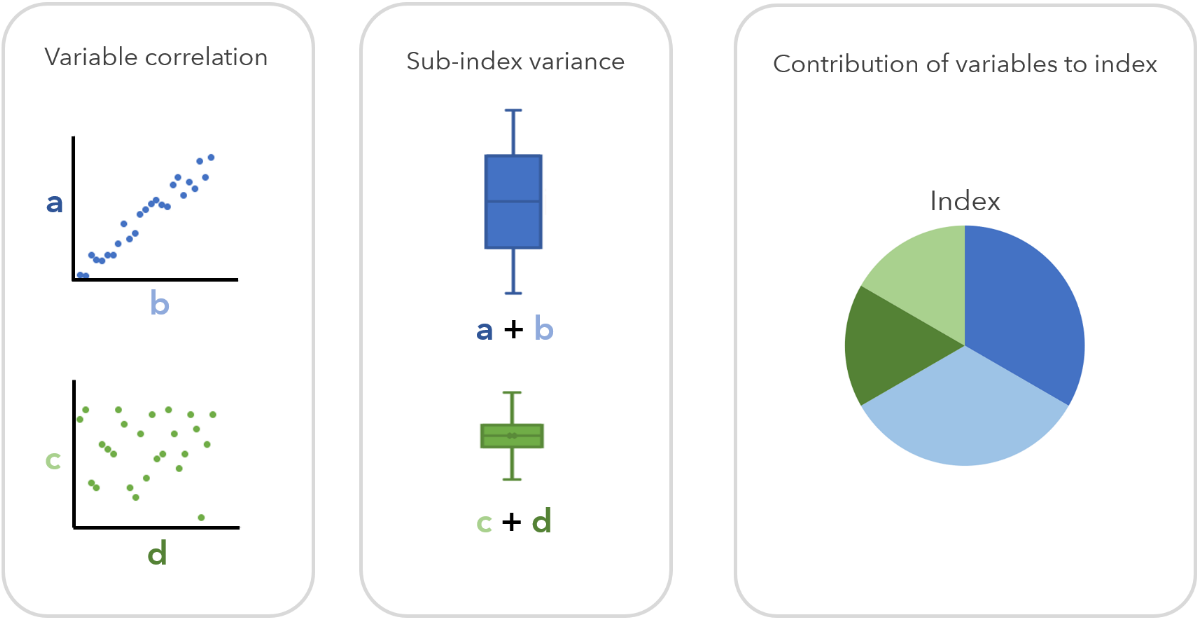
Combining variables using the mean is also an additive method to combine variables. However, sum and mean have subtle but important differences for the creation of sub-indices. As you saw above, when the variables in a sub-index are combined using a sum, each variable has the same contribution to the index. This means that since the dimensions have different numbers of variables, one dimension contributes more than the other. However, as the image below shows, if variables in sub-indices are combined using the mean, since the mean divides by the number of variables, each variable no longer has the same contribution to the index – instead each sub-index has the same contribution.



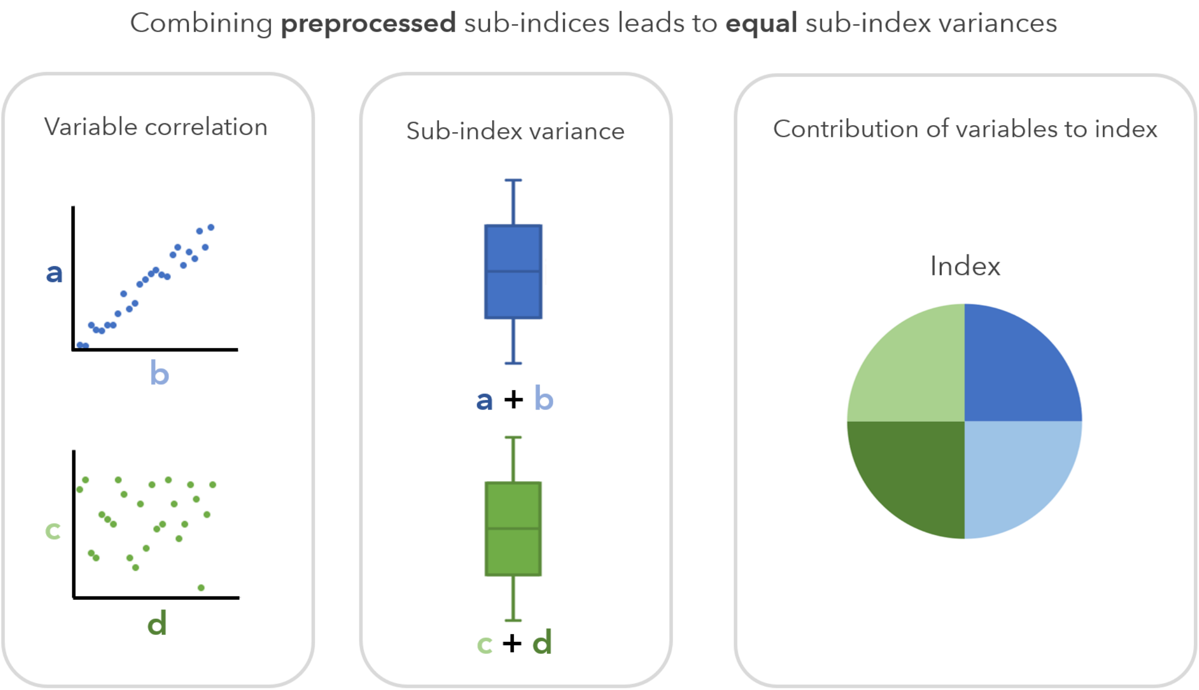
We can use this characteristic to our advantage. This allows us to apply sub-indices to correct for the **different number of variables** within each dimension, achieving balanced contributions of each dimension. It still allows us to interrogate each sub-index individually, but also gives us the opportunity to correct the index value if we intended for each dimension to have equal contribution. Even if we did not intend for the dimensions to have equal weights, adjusting the sub-indices so that they have equal contribution before applying the weights helps to ensure that the weights have their intended effect. Note that for multiplicative methods, multiply does not correct for the number of variables (similar to sum), and geometric mean does (similar to mean).

The third benefit of sub-indices is to help account for unintentional weighting of variables due to **differences in correlation** between variables. When some variables are highly correlated, and others have low correlation, the highly correlated variables will contribute more to the final index. This happens because for both additive and multiplicative methods the variance increases with increasing correlation (illustrated for additive methods by the variance sum law). This means that when the variables values are combined, high variance dominates – as illustrated in the image below. As an example of how correlation can impact results, imagine an index with variables **a** and **b** that are correlated, and variables **c** and **d** that are uncorrelated. Consider a location where variable **a** is high – variable **b** is also likely to be high due to the correlation. Variables **c** and **d** may be high or low – they have no relationship with variable **a** or with each other. In this location, the index is likely to be high, predominantly due to the influence of the correlated variables **a** and **b**.

Combining **raw** sub-indices leads to **unequal** sub-index variances



Sub-indices can be used to resolve the problem with correlated variables, by providing a way to equalize the variance. This is done by combining the variables into sub-indices, then scaling the sub-index results to achieve equal variance prior to combining into the index, as illustrated in the image below. The z-score preprocessing method is the most effective scaling method to ensure equal variance, however other methods can be used as long as you verify that the variances become similar. Note that correcting for differences in correlation only relies on the sub-indices being scaled – any method can be used to combine variables. Use scatterplot matrices to assess whether unintentional weighting due to correlation is a problem for your index, then use box plots to evaluate the variance of the preprocessed sub-indices (see step 8 for more details).


Combining **preprocessed** sub-indices leads to **equal** sub-index variances
Variable correlation
a
b
c
d
Sub-index variance
a + b
c + d
Contribution of variables to index
Index

Step 10 Explore the index further

The final step of the index creation process involves evaluation, consulting with stakeholders, and refinement. Consider applying some of the following to make your index even more effective:

- Use different variations of preprocessing and combination methods to learn how each combination impacts your resulting index. You can quantify how the change in methods changes the results by comparing the rank of the output index (for example, by calculating the change in the index_rank field in the output of the Calculate Composite Index tool in ArcGIS Pro). If you see large changes in rank, this indicates the index is very sensitive to the methods; conversely, if the ranks do not change a considerable amount, you can infer that the results are robust between methods. It's always important to justify the methods you choose, but assessing the sensitivity can help to contextualize how much these decisions matter.

- Consider a different spatial scale and compare the results with your current index. For example, run the index at the Census tract level in addition to the county level. Do these tell a different story?

- Investigate spatial clustering in the index results, for example, using the Hot Spot Analysis and Cluster and Outlier Analysis tools in ArcGIS Pro, or the Find Hot Spots and Find Outliers tools in ArcGIS Online. These can help you identify regions with statistically significant clustering of high index values, as well as find spatial outliers where the index may differ from the values of its immediate neighbors. These methods can reveal results that hint at new questions or problems, for example, examination of spatial outliers could reveal that a particular variable has driven these outliers, which could either be a genuine finding, or could indicate that the preprocessing or combination steps may need to be modified.

- Use the Multivariate Clustering tool in ArcGIS Pro to find features with similar input variable values. This may help you reveal common patterns that drive the index results across the study area. For example, a region with high human development index values may be characterized as having two multivariate clusters: one with high index values as a result of high education, and another with high index values as a result of high health outcomes.

- If the index is being used to answer a question which has a measurable outcome, build a regression analysis to test how well the index can predict the outcome. For example, the Virginia Health Opportunity Index creators tested how well their health index predicted health outcomes within the state. You could use the Generalized Linear Regression tool in ArcGIS Pro to do this. The diagnostics from this result may help justify the design and demonstrate the effectiveness of the index.

- Ensure that those who are impacted by the index are considered in the design and included in the process where possible. For example, use Survey123 to solicit community feedback, or create dynamic applications to encourage stakeholders to engage with the results and improve transparency. The index creation process is subjective, and it is crucial to include documentation which communicates the assumptions, intended uses, and limitations of the index.

References

Allik, M., Leyland, A., Ichihara, M. Y. T., & Dundas, R. (2020). “Creating small-area deprivation indices: a guide for stages and options.” *J Epidemiol Community Health* 74: 20–25.

Bigi ,V., Comino, E., Fontana, M., Pezzoli, A., & Rosso, M. (2021). “Flood Vulnerability Analysis in Urban Context: A Socioeconomic Sub-Indicators Overview.” *Climate* 9(1):12.

Mazziotta, M., & Pareto, A. (2016). “On a generalized non-compensatory composite index for measuring socio-economic phenomena”. *Social Indicators Research*, 127, 983–1003.

Mazziotta, M., & Pareto, A. (2018). “Measuring Well-Being Over Time: The Adjusted Mazziotta–Pareto Index Versus Other Non-compensatory Indices”. *Social Indicators Research,* 136, 967–976.

Norman, P. (2006). “Sociodemographic spatial change in the UK: data and computational issues and solutions. *GIS Development special issue Maps and Census* 10(12): 30-34.

Norman, P. (2010) “Identifying change over time in small area socio-economic deprivation”. *Applied Spatial Analysis and Policy* 3(2-3) 107-138.

OECD/European Union/EC-JRC (2008). “Handbook on Constructing Composite Indicators: Methodology and User Guide” OECD Publishing, Paris.

Openshaw, S. (1984). “The Modifiable Areal Unit Problem: Concepts and Techniques in Modern Geography” Geobooks, Norwich.

Townsend, P. (1987). “Deprivation” *Journal of Social Policy*, 16(2), 125-146.

Kolak M, Bhatt J, Park YH, Padrón NA, Molefe A. (2020) “Quantification of Neighborhood-Level Social Determinants of Health in the Continental United States” JAMA Netw Open.

Spielman, S.E., Tuccillo, J., Folch, D.C. et al. (2020) “Evaluating social vulnerability indicators: criteria and their application to the Social Vulnerability Index.” Nat Hazards 100, 417–436.

About the Authors

Lynne Buie
Lynne is a senior product engineer on the spatial statistics team at Esri. She has a bachelors in geography and a masters in GIS, both from the University of Edinburgh. She started her career as a GIS analyst and developer in the public sector, then moved to Esri where she has been building spatial statistics and data engineering capabilities since 2019.

Catherine McSorley
Catherine McSorley is a Senior Product Engineer on the Spatial Statistics team at Esri. She has a background in mathematics, statistics, and data science from the University of Wisconsin-Madison and North Carolina State University, as well as several years of consulting experience. Catherine is passionate about using spatial data to answer important community questions and enjoys being on the cutting edge of researching the newest analytical methods.

Alberto Nieto
Alberto Nieto is a senior product engineer on the spatial statistics team at Esri. In his role, he helps research, build, and maintain spatial data science capabilities in ArcGIS and works with organizations to learn about the problems our software can help solve. Alberto’s background includes fourteen years of experience, including previous roles as a GIS Developer at Capital One and NOAA's Climate Prediction Center, and as a GIS Analyst at the United States Census Bureau and the Alachua County Environmental Protection Department.

THE SCIENCE OF WHERE®

Esri, the global market leader in geographic information system (GIS) software, offers the most powerful mapping and spatial analytics technology available.

Since 1969, Esri has helped customers unlock the full potential of data to improve operational and business results. Today, Esri software is deployed in more than 350,000 organizations including the world's largest cities, most national governments, 75 percent of Fortune 500 companies, and more than 7,000 colleges and universities. Esri engineers the most advanced solutions for digital transformation, the Internet of Things (IoT), and location analytics to inform the most authoritative maps in the world.

Visit us at esri.com.

Contact Esri

380 New York Street
Redlands, California 92373-8100 USA

1 800 447 9778
T 909 793 2853
F 909 793 5953
info@esri.com
esri.com

Offices worldwide
esri.com/locations

For more information, visit **esri.com**.