

A Guide to Research Methods in Environmental Justice Project Analysis

*Jeffrey Alan Johnson, Ph.D.
Principal, Johnson Analytical Research
Associate Faculty, Westminster College*

Identifying and remedying environmental injustices requires an effective base of information and tools for sound analysis. By relying on methods commonly used in social science and public health, those involved in environmental justice disputes can move toward fact-based debates and collaborative models for conflict resolution.

Potential Injustices as Hypotheses

Potential environmental injustices should be viewed as research hypotheses. This means that they must be:

1. Understood as a possible relationship between variables representing the environmental effects and affected populations, and
2. Tested by observation to determine whether the hypothesized relationship exists.

Generally speaking, hypothetical injustices can be categorized along two dimensions. An environmental project may have several different kinds of effects that constitute an injustice. The two most prominent kinds are projects that disproportionately harm (or deny benefits to) a group, and projects in which a group is denied participation in decision-making processes. Clarifying which type of effect is hypothesized is a critical first step, as the methods of analysis are very different for distributive and participatory injustices.

In either case, injustices may also be categorized along a causal dimension. Frequently cited causes for environmental injustice may include deliberate or unconscious discrimination, market forces (either as a cause or consequence of siting decisions), lifestyle and cultural factors, and social or political factors. Note that under many policies, including both Executive Order 12898 and Title VI of the Civil Rights Act, it is necessary only to show disproportionate effects from a project regardless of cause; hence the causal dimension may not be relevant to all analyses. Cause is most relevant when analysis aims at determining liability or identifying methods of remediation.

Operationalization and Measurement

Operationalizing a variable means defining the variable in an observationally specific way such that most observers will agree on the character of a case with respect to that variable. In operationalizing variables the definition must have both validity and reliability. Valid operationalizations will accurately reflect the concept that the researcher is attempting to measure. Reliable operationalizations will result in consistent classifications of across cases or studies.

Operationalizing affected groups is perhaps the most difficult task in environmental justice research. Depending on the hypothesized injustice, one can operationalize affected groups in three ways: geographically, demographically, or economically.

Spatial Level of Analysis. Operationalizing variables with regard to spatial level of analysis is a key task given that environmental effects will always be geographically limited. Choosing the proper level of analysis involves two considerations:

1. The overall scope of the study. This is typically the geographic extent of the environmental effects of the project. The major challenge here is finding a data source that corresponds as closely as possible with this scope. Geographic Information Systems can be used to define the study area and select smaller-scale units that fall within that area (e.g., all census tracts within 30 km of a facility).
2. The spatial unit of analysis. This is the geographic unit used to operationalize the affected populations for comparison. Common units used include census tracts and blocks, zip codes, counties, and distributions of specific populations or communities. It is often appropriate to use multiple units of analysis.

Demographic Measurement of Geographic Areas. Often the demographic characteristics of a spatial unit of analysis will be the main population variable. Binary nominal measurements (e.g., characterizing census blocks as either “white” or “minority”) are quite weak. The most effective measurement is a ratio variable (e.g., percentage of non-white population), as this allows for more subtle analysis.

Temporal Levels of Analysis. It is important to specify the time period in which a variable is analyzed. Environmental injustices may be limited to the present, in which case the focus is likely to be on the comparative analysis of affected groups. Where the hypothesized injustice occurred in the past but is still affecting present populations, changes in variables over time must be considered, as there may be time lags between causes and effects. Injustices that will take place in the future (usually the result of a project under consideration) require developing predictive models for changes in variables.

Health Effects. Effects of a project on health are most often operationalized in one of two ways:

1. Exposure to hazards. Exposure can be measured in multiple ways. The most direct is to measure the quantity of a hazard to which an average resident will be exposed. As this is not always possible, exposure may also be measured by proximity to the project or by the rate of occurrence of the exposure. Exposure analysis should include evaluation of a range of causal factors in exposure, including social and economic considerations.
2. Risk or incidence of adverse health outcomes. In cases of existing or past projects, incidence can be directly measured. Risk assessment attempts to predict the number of additional deaths from a project by estimating the amount of exposure to an environmental hazard caused by the project, the likelihood of that amount leading to adverse health effects, and the size of the population that will be exposed to the risk. The result is usually stated as a number of additional cancer and non-cancer deaths for a given population, usually 10,000 or 100,000.

In either case, it is important to consider both the marginal effects of an individual project and the cumulative effects of all projects in assessing health effects.

Other effects. It is not out of the question to consider non-health effects of projects. The Bureau of Land Management has developed processes for evaluating visual impacts of projects, which could potentially be used to demonstrate injustices related to aesthetic considerations. Projects might also have significant effects on cultural resources or practices, present risks to property, or place financial burdens on local residents that are disproportionate to their share of the benefits received. Such effects require more specialized analysis that is addressed here.

Descriptive Statistics

Descriptive approaches are most useful for assessing conditions of changes over time. Such statistics measure primarily central tendencies or the degree of variation. The Excel Data Analysis Toolpack under Tools>Data Analysis (if it is installed) includes a descriptive statistics tool that will calculate all of the statistics below, except for a cross-tabulation.

Central Tendency. Measures of central tendency characterize a group of observations. There are three commonly used measures:

1. Mean or Average. The average is the sum of all cases divided by the number of cases. It is useful in characterizing the group either in itself or for comparison to other groups, but can be affected by unusually high or low values. To find the average using Excel, type:

=average(cell range for which you want the average)

2. Median. The median is the middle member of a group. It is especially valuable for variables with a few extreme cases, as these do not affect the median value. The average is more commonly used where the effects of extreme cases are important. To find the mean using Excel, type

=median(cell range for which you want the median)

3. Mode. The mode is the most common value for a variable. To find the mode using Excel, type:

=mode(cell range for which you want the median)

Degree of Variation. It is often helpful to know how much variation there is in the variable. This is most commonly measured by the standard deviation, which measures essentially the average difference between each value and the mean. A larger standard deviation indicates that the cases vary considerably. To find the standard deviation using Excel, type:

=stdev(cell range for which you want the median)

Cross-tabulation. Descriptive statistics for nominal and ordinal variables are usefully compared in a cross-tabulation, which shows the distribution of cases across two or more variables. It is usually shown in a table with one variable in rows and one in columns, as such:

	Minority Tracts	Non-minority Tracts	Total
Violations	53 (81.5%)	12 (18.5%)	65 (100%)
No Violations	174 (62.6%)	104 (37.4%)	278 (100%)
Total	227 (66.2%)	116 (33.8%)	343

There are two important aspects of interpreting a cross tabulation:

1. **Summation.** Whether percentages are based on the row, the column, or (rarely) the entire chart determines what can be concluded from the data. In the example above, one can see

that minority tracts are more likely to have violators than non-minority tracts because the percentages are in rows, showing that the minority tracts contain 66.2% of cases but 81.5% of violators. Summing the data in columns would show that sites in minority tracts are more likely to have violations. These two conclusions are independent from each other.

2. **Significance.** It is possible that the differences in categories are entirely random. In the absence of a test of statistical significance (described below), a general conclusion is quite weak.

Automatically performing cross-tabs using Excel can be performed using a rather complex set of array functions. In many cases where there are relatively few cases and data is unlikely to change, it may be easier to perform a cross-tab manually using the Data>Sort menu. Sort by the two variables, then highlight the cells with common values. The number of rows or columns highlighted indicates the count for that cell in the cross-tab.

In most cases, it will be necessary to use inferential statistical tests that allow for the drawing of inferences regarding the relationship between two variables in order to test hypotheses regarding environmental injustices.

Tests of Association

Tests of association are used to test the degree to which a change in one or more variables is associated with a change in another variable. This is commonly used to identify causal relationships, in which case the first variable (or, in the case of multivariate methods, variables) that are believed to cause the change are called “independent variables” while the last variable for which the independent variables cause change is referred to as the dependent variable. The terminology is often used even when the relationship in question is not necessarily causal, e.g., when identifying disproportionate effects on minority groups.

There are two main approaches to association, which are often used together to characterize a relationship. Both of these can be performed using the Data Analysis Toolpack in Excel as well as using the functions described below.

Correlation. Correlation tests the degree to which a change in the independent variable is associated with a proportional change in the dependent variable. For instance, if an increase of one unit in the independent variable consistently results in an increase of three units in the dependent variable, the correlation coefficient, called r or Pearson's r , will be strong, but it will be no different than if the one-unit change in the independent variable consistently results in a one-unit change in the dependent variable. Correlation measures the consistency of change and not the amount of change.

To calculate the correlation coefficient using Excel, type

=CORREL(cell range for the first variable,cell range for the second variable)

The Correlation Coefficient for two variables is expressed as a value between -1 and 1. A coefficient of 1 indicates that a change in the independent variable always results in a directly proportional change in the dependent variable. A coefficient of -1 indicates an inversely proportional change. A coefficient of 0 indicates that there is no relationship between the variables. In social science, a coefficient of ± 0.25 is commonly thought of as a moderately close relationship, and a coefficient of ± 0.4 is a very strong relationship.

Exceptionally strong coefficients raise concerns of a spurious relationship, that is, one in which both variables are influenced by a third variable. For instance, there is a very strong correlation between

the size of a state's Jewish population and the incidence of HIV infection. This relationship is likely a consequence of the fact that both variables are strongly related to the size of a state's urban population rather than a consequence of a direct relationship between the variables.

The square of the correlation coefficient (r^2) is a measure of the percentage of the change in the dependent variable explained by the changes in the independent variables.

Regression. Regression is a tool for modeling relationships between two variables. This essentially shows how much change in the dependent variable results from a change in an independent variable, for example, how much an increase in minority population increases the probability of environmental violations. It calculates a linear relationship between the two variables, returning most importantly the slope of the resulting line. Regressions can be either univariate or multivariate, which estimates contributions of several independent variables to changes in dependent variables.

There are several functions related to regression in Excel. To simply find the slope and y-intercept of the linear relationship between two variables, highlight two adjacent cells in the same row, then type:

=LINEST(cell range for the dependent variable, cell range for the independent variable,1,0)

then press CTRL-SHIFT-ENTER instead of just ENTER to enter the formula as an array. The first number returned is the slope of the line; the second is the y-intercept.

Regression analysis can be used to forecast future changes in variables. For instance, one can create an estimate of the relationship between exposure to a chemical and cancer incidence, and then use that to forecast the additional incidences of cancer resulting from a proposed project.

Tests of Statistical Significance

Statistical significance refers to the likelihood that a relationship between variables is due to random chance rather than some underlying factor. It is important to note that a result can be statistically significant even where there is a relationship that is so weak as to be trivial in terms of substantive findings.

There are several such tests to be used according to the type of data analyzed. The Excel Data Analysis Toolpack includes tools for performing each of the analyses described below.

Student's t-test. The t-test is used to determine whether the differences between two groups are due to random chance. This can only be used to analyze cases with regard to a single variable broken down into two categories, e.g., the number of violations in minority vs. non-minority census tracts. To perform a t-test in Excel, first sort your data according to the variable that you are analyzing, then type:

=ttest(cell range for the first group, cell range for the second group,2,3)

The final two values in this function will be appropriate for most analyses in environmental justice, and correspond to the "t-test: Two Sample Assuming Unequal Variances" tool in the Data Analysis Toolpack.

Student's t-test can only be used with two groups of values. To analyze more groups, use the ANOVA (Analysis of Variance) test, which can be performed with Excel's Data Analysis Toolpack.

Chi-squared Test. This test is used to determine the statistical significance of a cross-tabulation by comparing the actual values of each cell to the values that would be expected if the distribution was

random. The expected value of each cell is the percentage of cases of the given value of the first variable, multiplied by the percentage for the given value of the second variable, multiplied by the total number of cases. For instance, using the table above, the expected number of minority tracts with violators would be

$$0.189 (\% \text{ of sites with violations}) \times 0.661 (\% \text{ of sites in minority tracts}) \times 343 (\text{total cases}) = 43.01$$

To perform a chi-squared test in Excel, prepare a cross-tab for the actual values and a separate one for the expected values. Then type:

$$=chitest(\text{cell range for the actual values}, \text{cell range for the expected values})$$

F-test. The F-test is used to test the statistical significance of variance between two variables, and is used with correlations as regressions. To perform an F-test in Excel, type:

$$=fetest(\text{cell range for the first variable}, \text{cell range for the second variable})$$

Interpretation. Statistical significance is usually expressed as a level that represents the likelihood that the relationship between the values exists by chance. To say a relationship is “significant at the .05 level” indicates that the relationship would occur by chance less than 5% of the time. Usually, one chooses a threshold significance level and accepts results as valid if they exceed this level. Commonly used significance levels in scientific research are the .10, .05, and .01 levels.

Two problems with this approach present themselves in environmental justice research. The approach of using threshold levels rather than specific values dates predates the use of computers for calculations. Modern statistical software and spreadsheets allow the precise calculation of a significance level. In addition, the high threshold levels used in scientific research reflect the importance of avoiding false positives in such research, which prevents scientists from making a claim in the absence of near-certainty. This may not be appropriate for environmental justice research, where false negatives present serious health risks to populations. These considerations suggest reporting of the precise level of significance rather than simply identifying tests that meet the threshold level.

The Excel tests described above all return values between 0 and 1 that indicate the likelihood that a relationship is NOT due to chance, opposite the threshold level approach. To calculate the significance level, subtract the results of the test from 1.

Qualitative Methods

Whether due to the nature of the problem or the availability of data, qualitative research is often necessary in evaluating environmental justice cases. Qualitative research involves a wide range of analytical approaches, far more than can be suggested here. These approaches tend to require far more intensive efforts on the part of the researcher, come with relatively little guidance on specific techniques and practices, and raise serious concerns regarding reliability, validity, and bias. But they can produce information that cannot be gathered through quantitative research, especially with regard to the history of a case. This is especially true with regard to participatory injustices and in developing bases for negotiated responses to identified injustices.

Documentary Analysis. As environmental justice controversies are commonly in the public realm, there is often extensive documentation of both substantive and procedural issues. Often, however, parties may be reluctant to make such documentation available. Open records laws such as the

federal Freedom of Information Act and the state GRAMMA law can force agencies to turn over records as needed. If litigation is involved, court records are likely to be accessible and may include documents from private actors that would otherwise be unavailable. Most Environmental Impact Statements are available online from the EPA or the agency that prepared the statement. Among qualitative methods, documentary analysis has the potential to be the most objective, however it is often limited to data that participants are comfortable placing in the public realm.

Interviews. Individual interviews and focus groups are exceptionally effective tools for uncovering the understandings, values, and perspectives of participants. Depending on the aim, interviews can vary from highly structured interviews that do not deviate from a set list of questions (often useful when that responses will be interpreted for use in quantitative methods) to very loosely structured discussions that give ethnographic kinds of insights (insights into culture, values, and worldviews, for instance). Interviews present less significant problems of access and empathetic biases on the part of the researcher than participant observation, described below.

Participant Observation. Researchers can gain much insight by observing participants in environmental justice processes. Participant observation can refer to two methods of research: the observation of participants by a researcher who is herself outside of the process, and the direct participation of the researcher in the case being studied. These can generate exceptionally useful insights, especially in documenting issues that would not be documented otherwise and in going beyond the participants conscious understandings of the process. However, these processes are highly constrained by the access that participants are willing to offer, the amount of trust that participants have in the researcher, and the extent to which the researcher can keep empathy with participants from becoming a source of bias. Participant observation also raises the possibility that the researcher's presence may alter the behavior of the actors involved, though this is more likely to be a concern in academic research than in analyses aimed at understanding the concerns raised by specific projects.

T r i a n g u l a t i o n

The most effective research designs will not be limited to a specific model of analysis. The combination of multiple statistical and qualitative methods, referred to as “triangulation,” should be part of any good research plan. A combination of methods that present consistent findings is much more powerful than an individual test, as it reduces the likelihood that the findings of the study are due to errors or biases in the methods used. Where the findings are inconsistent, differences in method can clarify complexities and expose previously unknown issues in the case being analyzed.