## Notes on Testing Predictive Validity of a Risk Terrain Model

There are several reasonable ways to empirically test the predictive validity of a risk terrain model, including OLS, logistic, or geographically weighted (GW) regression models. Some of these methods are shown in the table below. ArcGIS currently provides tools for OLS and Geographically Weighted Regression modeling; there is an informative training video for these tools available here: http://training.esri.com/acb2000/showdetl.cfm?did=6&Product_id=942. The most appropriate method for testing predictive validity will depend on a variety of things, including the nature of your outcome events data, cell size, and the way you prepare your data (e.g., binary or interval values). You are encouraged to be innovative and explore different empirical methods that suit your needs the best.

| **Possible Statistical Methods for Testing Predictive Validity** | | | |
|---|---|---|---|
| | | (Dependent Variable)<br>**Outcome event values for each cell are:** | |
| | | **Binary**<br>(e.g. Present? Yes/No) | **Interval**<br>(e.g. Counts) |
| (Independent Variable)<br>**Risk values for each cell are:** | **Binary**<br>(e.g. The RTM yielded only two possible risk values, such as 0-1: Highest risk/Not highest risk) | Chi-Squared; Fisher's Exact | OLS Regression; T-test |
| | **Ordinal/Interval**<br>(e.g. The RTM yielded two or more possible risk values, such as 0-9) | Logistic Regression | OLS Regression; Geographically Weighted Regression |

You might conceptualize the composite risk values of your risk terrain map as either ordinal or interval. An ordinal variable is similar to a categorical variable (e.g. Male or Female) except there is a clear ordering of ordinal variables (e.g., Low, Medium, High). Risk is often conceptualized as an ordinal variable in RTM because, while it is possible for two places to have risk values of 5 and 10, for example, it is not possible to say that the place with a risk value of 10 is twice as risky as the place with a risk value of 5. The difference between an ordinal and interval variable is that the interval between the values of the interval variable are equally spaced (which permits one to say, for example, that "4" is twice as great as "2"). Although risk is usually conceptualized as an ordinal variable with regard to risk assessments, risk values (as the product of RTM) can be treated as interval values for the purpose of testing the predictive validity of a risk terrain model.

The decision to use ordinal or interval risk values for testing the predictive validity of a RTM depends, in part, on the conversation you want to have about the results. With interval risk values, regression results can be used to explain the likelihood of an outcome event occurring at places when the risk at a place increases by a unit of 1. With ordinal risk values (which require dummy coding and a reference variable omitted), regression results can be used to explain the magnitude of risk between risk values at places.

When doing a t-test or ANOVA, the assumption is that the sample means are normally distributed. This is not generally the case with the locations of crime data (i.e., crimes cluster). However, if the distribution of the individual outcome event locations is not normal, the distribution of the sample means will be normally distributed if your sample size is about 30 or larger. This is due to the "Central Limit Theorem." Given that your unit of analysis in RTM is very small cells within a grid of your study area (e.g., 100ftX100ft cells), then your sample size will most likely be well over 30 cells.

## Spatial Autocorrelation Basics

Distributions among geographical units, such as risk terrain cells, are usually not independent, meaning that values found in a particular cell are likely to be influenced by corresponding values in nearby cells. Imagine, for instance, that a home's value will be more affected by the value of neighboring houses compared to houses that are farther away: Houses that are on one side of town tend to have less effect on the value of a house far away on the opposite side of town. Stated another way: things tend to be influenced by other things that are closer to them, compared to other things that are farther away. This phenomenon is referred to as spatial autocorrelation.

When testing the predictive validity of a risk terrain model, you may have to control for spatial autocorrelation. Again, this refers to feature similarity based on both feature locations and feature values simultaneously. Or stated another way, spatial autocorrelation refers to the degree to which cells within a risk terrain (i.e. spatial features) and their associated values (e.g., the dependent variable: crime counts) tend to be clustered together throughout the landscape (i.e., positive spatial autocorrelation) or dispersed (i.e., negative spatial autocorrelation). Note that spatial autocorrelation--for the purposes of testing predictive validity--is measured using the dependent variable (i.e., the outcome events) and not the independent variable (e.g., risk values). Controlling for spatial autocorrelation allows you to measure the effect of a place's risky environment on the attraction of outcome events without the similar patterns of outcome events in neighboring places confounding your results.

Moran's I is an area-wide analysis used to measure spatial autocorrelation. ArcGIS has a "Spatial Autocorrelation (Moran's I)" tool in ArcToolbox that can be used to measure spatial autocorrelation, with values approaching 1 when geographical units are situated near other similar geographical units, and approaching –1 when geographical units are situated near dissimilar geographical units. A Moran's I value of 0 indicates the absence of autocorrelation, or independence, among geographical units. If spatial autocorrelation exists, then you should create a spatial lag control variable to include in your regression model when testing the predictive validity of your risk terrain model. This can be done with the "Generate Spatial Weights Matrix" tool. If spatial autocorrelation does not exist, then a spatial lag control variable is not needed.

## Notes on Selecting Risk Map Layers

There are several reasonable ways to empirically validate and select risk map layers using traditional statistical methods. Some of these methods are shown in the table below. ArcGIS currently provides tools for OLS and Geographically Weighted Regression modeling; there is an informative training video for these tools available here: Esri Training: Regression Analysis Basics in ArcGIS 9.3. Chi-Squared, Fisher's Exact, and T-tests will have to be performed outside of ArcGIS with a separate statistical analysis software package, such as EpiInfo (It's Free). The most appropriate method will depend on a variety of things, including the operationalization schema of your risk map layers (e.g., binary valued or ordinal scale).

| **Possible Statistical Methods for Selecting/Validating Risk Map Layers** | | | |
| --- | --- | --- | --- |
| | | (Dependent Variable)<br>**Outcome event values for each cell are:** | |
| | | **Binary**<br>(e.g. Present? Yes/No) | **Interval**<br>(e.g. Counts) |
| (Independent Variable)<br>**Risk factor spatial influence values for each cell are:** | **Binary**<br>(e.g. Spatial influence of risk factor is present? Yes/No) | Chi-Squared; Fisher's Exact | OLS Regression; T-test |
| | **Ordinal**<br>(e.g. Spatial influence of risk factor is operationalized as High, Medium, or Low) | Chi-Squared; Fisher's Exact | ANOVA; OLS Regression |
| | **Interval**<br>(e.g. Intensity value of risk) | Logistic Regression | OLS Regression; Geographically Weighted Regression |

With a binary valued independent variable in an OLS Regression model, the zero (0) value becomes the reference category. So, for example, if "0=no risk" and "1=risk" and the results are statistically significant, you could say that "Risky places had an X increase in outcome events compared to non-risky places."

Remember to dummy code an ordinal level independent variable for regression modeling. For example, if the composite risk of a place is operationalized as "high", "medium", or "low", (which might be the case if you symbolize and reclassify your final risk terrain map by standard deviational break--e.g., "high"=greater than +2SD; "medium"=between the mean and +2SD; "low"=less than the mean), then create three binary-valued variables: "High? (Yes/No)", "Medium? (Yes/No)", "Low? (Yes/No)." Include all but the reference variable (usually the lowest value, e.g., "Low?") in the regression model. If the results are significant, you could talk about the magnitude of risk between each dummy variable and the reference variable.