

1. What is spatial autocorrelation?

The following image represents the residuals from a hedonic home price model. The homes are located around the university; the BAS building is shown as a large red cross. Positive residuals are shown in red—the darker the red, the larger the residual. Negative residuals are shown in blue, with darker blue representing a larger absolute value of the residual.

Homes clearly appear likely to have a residual of the same sign as their neighbors. We describe this as spatial autocorrelation of the residuals.

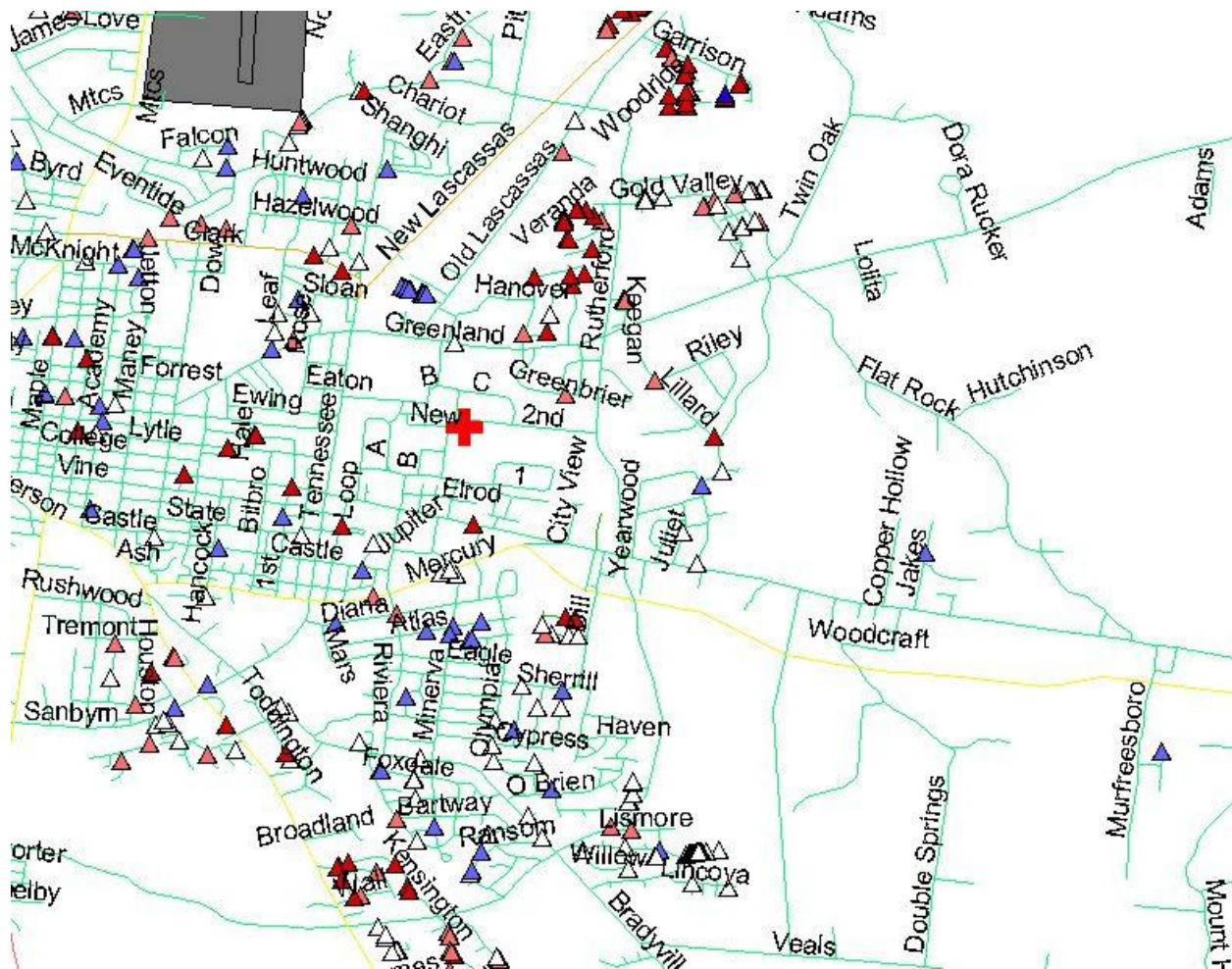


Figure 1: Residuals from a hedonic home price model. Red is positive, blue is negative.

In a hedonic home price regression, spatial autocorrelation of the residuals is usually considered to be the result of “neighborhood effects.” It occurs because variables omitted from the model are spatially autocorrelated. To illustrate, the next map shows the occurrence of brick homes throughout this same area: the blue triangles represent brick homes, the red triangles are non-brick. Since brick homes are more expensive, the regression residuals for brick homes will be higher if the model fails to include brick as an independent variable. Since brick is autocorrelated (a home is more likely to be brick if its neighbors are brick), then the residuals will be autocorrelated.

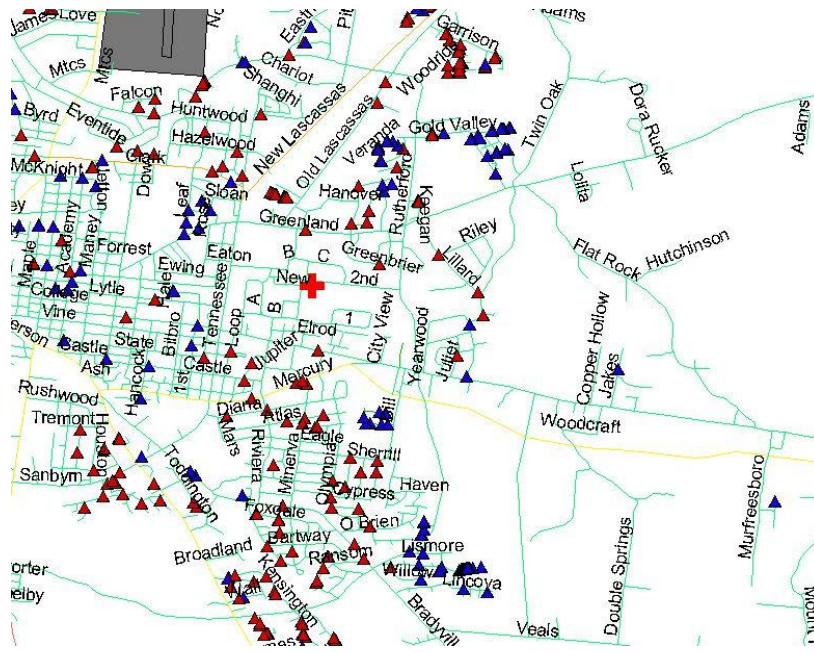


Figure 2: Occurrence of brick homes: red is non-brick, blue is brick.

Many variables exhibit some kind of spatial autocorrelation, which means that when they are omitted from a model the residuals will be spatially autocorrelated. Figure 3 presents an example of per capita protein consumption, viewed on a world map. Physically adjacent countries clearly have similar levels of protein consumption. But it is also true that culturally similar countries (e.g., Australia and the U.S.) have similar levels. Spatial autocorrelation is generalizable beyond physical distance to any kind of meaningful distance. There is a tradition within cross-cultural studies of considering both physical distance and cultural distance when examining autocorrelation of variables or regression residuals.

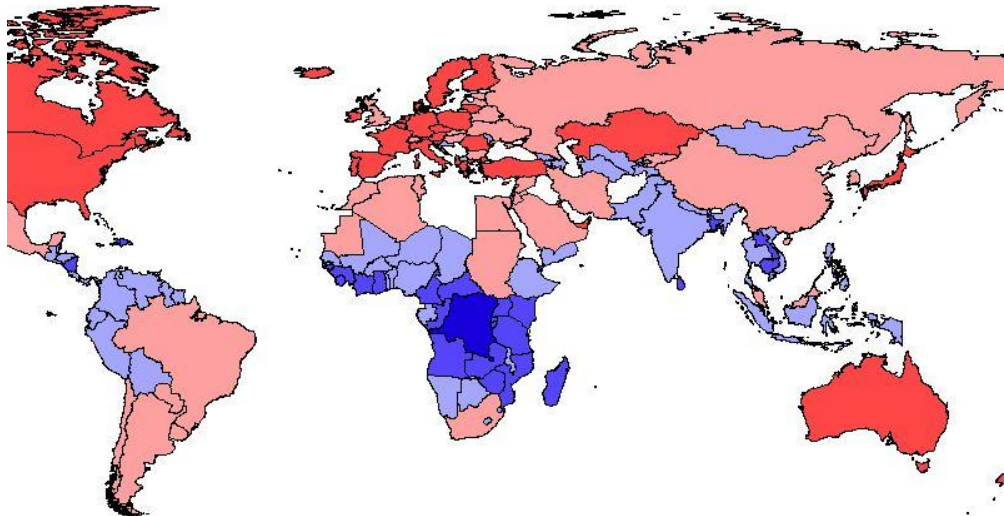


Figure 3: Per capita protein consumption (WHO, 2004). Blue is low, red is high.

2. Why is spatial autocorrelation a problem?

Ordinary least squares requires that the residuals in the estimated model not be correlated with each other. Violation of this property causes the estimated standard errors of the coefficients to be biased, so that one cannot trust the t-statistics, and one therefore cannot make hypothesis tests regarding the estimated coefficients (Kennedy 1998). Even worse, the

presence of autocorrelation is often a sign of omitted variables, so that even the estimated coefficients may be biased. For example, in a hedonic housing model a spatial autocorrelation test compares the residual for a particular house to the residuals of houses in the same neighborhood. Typically, these models show significant autocorrelation of the residuals, since the factors leading a house to have a particularly high (low) residual are often causing neighboring houses to have a particularly high (low) residual. These “neighborhood effects” are a sign of omitted variables, a problem that can usually most easily be cured by incorporating a spatially lagged dependent variable in the model.

In regressions on cross-national datasets, proximity of nations could be either spatial or cultural, and autocorrelation among spatially or culturally proximate residuals would indicate “neighborhood effects” in which variables similarly affecting related nations have been omitted.

3. How does one test for spatial autocorrelation?

There are several methods. The most common is a modified version of the Moran’s I test¹ that is used for regression residuals. The null hypothesis is that there is no autocorrelation; a z-score with a high enough absolute value allows one to reject the null hypothesis. In most cases, only positive autocorrelation will have a valid meaning.

Note that any spatial autocorrelation test will require a weight matrix, defining the spatial relationship between each of the observations. Each element w_{ij} in weight matrix \mathbf{W} can be interpreted as the *proximity* between the observation in row i and the observation in row j : the higher w_{ij} , the closer together are i and j .

The R function `lm.morantest`, found in package `spdep`, will test for autocorrelation of regression residuals, returning a p-value.

4. If one has spatial autocorrelation, what does one do about it?

The finding of autocorrelation indicates that there are likely to be important omitted variables. The most common way to handle this problem is to create a spatially lagged dependent variable $\hat{y} = \mathbf{W}y$, where y is an $n \times 1$ vector of the dependent variable, and \mathbf{W} is an $n \times n$ spatial weight matrix, with the rows standardized to sum to one. The spatially lagged variable \hat{y} will, however, be endogenous: if observation i ’s value depends on the values of proximate observations, then the values of proximate observations will in turn depend on the value of observation i . The simplest way around this problem is to substitute an instrument for y .

Introducing \hat{y} as an independent variable in the model usually eliminates autocorrelation. The coefficient of \hat{y} can be interpreted as the effect on the dependent variable of transmission across space. This interpretation, however, masks the specific traits which determine the dependent variable—it tells us that spatial mechanisms are a determinant, but doesn’t tell us what those spatial mechanisms are. Some econometricians would therefore argue that it is preferable to use spatially lagged variables sparingly, after first attempting to introduce other independent variables and trying various functional forms.

HOMEWORK

You will use the Standard Cross-Cultural Sample in this homework. Pick an interesting variable and create a model explaining that variable. Use the link below to find programs and a “primer” that will guide you in building a model with spatial lag terms (one for distance, one for language) within the context of multiple imputation.²

<http://escholarship.org/uc/item/7cm1f10b>

Turn in an *ex ante* discussion, tables for descriptive statistics (for all variables in the unrestricted model) and regression results (for your final model), and an *ex post* discussion. Be sure to test for correct functional form, normality of residuals, multicollinearity, heteroskedasticity, and spatial autocorrelation (using both weight matrices). Remove irrelevant independent variables, testing for the propriety of your action. Report all of these test statistics in the notes section of your regression results table.

¹ Moran’s I is commonly used to test for autocorrelation in variables. Be careful to use the correct, modified, form when testing autocorrelation in regression residuals.

² More advanced R scripts can be found under “supporting materials” here: <http://www.escholarship.org/uc/item/5rh6z6z6>