7. Summary of Conclusions

The results of this research clearly demonstrate why the Modifiable Area Unit Problem has been such a source of frustration for spatial analysts for so long. Even a relatively simple statistic like the weighted variance behaves in a complex manner, influenced by the spatial autocorrelation and arrangement of the unaggregated variable. More complex statistics, like the Moran Coefficient, correlation, covariance, and the bivariate regression slope parameters, are affected by the spatial arrangements of both variables, while the multivariate regression parameters are affected by those of all variables involved. Unfortunately, results reported in Chapter 4 amply indicate that the MC is not a sufficient measure of spatial organization for the purposes of prediction of results, since many different types of arrangement can have the same MC, and it is often the arrangement for the given MC that determines how a variable will behave under aggregation. Even so, it is still useful as a first approximation in most cases, and further research may be able to provide a summary statistic that can include pattern as well as spatial autocorrelation.

One of the common features to all the experiments is that the frequency distributions (which are a result of the zoning effect) of all of the aggregated statistics are either normally distributed or nearly so. The assumption of a normal distribution plays a pivotal role in most inferential statistical theory, so this empirical finding may help to further advance theoretical investigations of the MAUP. The finding is surprising, especially for something as complex as a MC of a regression residual, because due to Murphy's Law I would expect a distribution that would make the analysis of the MAUP with statistical theory even more difficult[1].

The relative change in variance shows a strong dependence on the spatial autocorrelation of the original variable, which of course is no surprise, but it also depends on the spatial arrangement of values. The aggregated Moran Coefficient depends not just on the initial spatial autocorrelation, but also on the spatial arrangement of the values, especially as the original MC increases and patterns become more distinct. Patterns with a large number of small clusters of similar values will show the greatest change in aggregate univariate statistics as the number of cells decreases because as the cell size increases, the likelihood of including regions with dissimilar values increases faster than it does when there are only a few large clusters. A more precise definition of

[1] OK, this is a bit cynical. Maybe I have been a post-graduate for too long.

the relationship must await a better way to describe the spatial arrangement of the data values, perhaps by using two or more spatial autocorrelation statistics in conjunction with each other.

The relative change in variance is strongly non-linearly correlated to the G statistic, which has been modified by dividing by the unweighted aggregate variance. This dependence does not appear to be because the unweighted aggregate variance is present on both sides of the regression equation, though what causes it and how it can be exploited are worth future research.

The covariance tends to behave in a similar way to the variance under aggregation, in spite of the possibility for it to increase or decrease. The range of the distributions of both statistics decreases with the decreasing number of aggregate cells for low values of spatial autocorrelation of variables, since increasing the cell size will not appreciably increase the (co)variation within each cell that can be lost by aggregation. As the MC increases, the within-cell variability will tend to increase with an increase in cell size as more dissimilar values are included, with the rate of this increase depending on the spatial arrangment (many small or fewer larger clusters).

When both variables have the same MC, the ranges of the covariance, correlation and regression slope parameter tend to increase as MC decreases, and to increase as the number of aggregate cells decreases. The MC of the regression residual (MC_{RR}) is not much affected by the initial correlation of the variables, but changes considerably with the increase in MC of the variables, showing a marked tendency to decrease as the number of aggregate cells decreases. This indicates that the statistical quality of regression results can actually be improved with aggregation, even though the values of the parameters are quite different from the original. This apparent improvement is offset by the loss of information caused by the reduction in sample size. When the variables have different MCs and the initial correlation is zero, the behaviour is still reasonably regular. The range of correlations tends to increase as the MC of the variables decreases, and the range of regression slope parameters is greatest when the MCs of the variables are the most different, and again tends to increase as either variable's MC decreases. The change in the MC_{RR} appears to depend primarily on the MC of the dependent variable. When the variables have different MCs and the initial correlation is non-zero, prediction of the statistics, and especially MC_{RR} and correlation, becomes difficult due to differences that are caused by the differences in spatial patterns of variables that have the same MC. Having a smaller number of initial zones in the aggregation increases the ranges of the aggregated statistics for variables with the same MC because

dissimilar values are closer together, increasing the chances of having aggregate cells with larger internal variations.

When the dependent and three independent variables in the multiple regression experiments have the same MCs, the variation of the statistics tends to decrease as MC increases. The mean of the distributions of the regression parameters remains essentially constant as the number of aggregate cells decreases. As with the bivariate case, the change of the MC_{RR} seems to be independent of the MC of the independent variables, but again this is caused by the initial correlations between variables being zero and so the initial MC_{RR} is the MC of the dependent variable. When the dependent variables have one MC and the independent variable has another, the MC of the independent variables tends to have more of an effect on the regression statistics than does that of the dependent variable. For a given MC of the dependent variable, the variability in the coefficients and their standard errors tends to decrease with increasing MC of the independent variables. However, for a given MC of the independent variables, the ranges of the statistics increase with an increase in the MC of the dependent variable. As the results from the bivariate analysis indicate, collinearities between variables are introduced when the initial correlations are zero. However, only 2 to 8 percent of the aggregations produce correlations that are statistically significantly different from zero.

The results of this research make it abundantly clear that those who use spatially referenced data should not try to extend any conclusions they draw to levels of spatial resolution that are different from the resolution of the data. As yet there is no way to estimate the value of a statistic computed at a finer scale of resolution (larger number of smaller regions) from aggregated data, applying results derived from a coarser spatial resolution will most likely lead to the drawing of erroneous conclusions.

8. Topics for Future Research

This research represents the first step in the systematic empirical exploration of the Modifiable Area Unit Problem, and much remains to be explored. All of the research work in this thesis is for variables that are averaged during aggregation, and it is suspected that variables that are summed or that are proportions (i.e. numerator and denominator aggregated separately) will not behave in the same way. Only a few of the possibilites have been explored for the multivariate re-

gression statistics, and more complex multivariate procedures such as factor analysis have not been tested at all. Before such analysis can properly proceed, however, a better way is required to numerically quantify spatial arrangements than the Moran Coefficient. A variogram certainly contains a complete description of the spatial structure, but then a way to describe the variogram would have to be concocted and we are no better off. The MC itself is not sufficient to describe the spatial arrangement, but perhaps using it in conjunction with other spatial autocorrelation statistics that describe the pattern differently will work.

It is hoped that my research will lead to further advances in the theoretical as well as empirical exploration of the MAUP, and that the knowledge that it is not totally intractable and chaotic might be enough to renew interest and research in this challenging statistical phenomenon.