6. The Effects of Aggregation on Multivariate Regression Parameters[1]

6.1. Summary

Several empirical studies of the Modifiable Area Unit Problem (MAUP) have been performed on census data, one of which has been about its effects on multivariate regression analysis. Recognizing that as much control as possible needs to be exerted in order to effectively study the MAUP, a spatial dataset generator was created that allows the user to construct sets of variables with various spatial and aspatial properties. The effect of aggregation on multivariate regression parameters, with special attention to the influence of spatial autocorrelation, is studied using a number of synthetic datasets created by the data generator. It is found that the effects depend on the combinations of autocorrelations of the unaggregated dependent and independent variables. It is also found that aggregation introduces collinearities between independent variables where none existed before. The patterns displayed provide hope that the effects of the MAUP on multivariate regression may not be as unpredictable as was once feared.

6.2. Introduction

The Modifiable Area Unit Problem (MAUP), a term introduced in Openshaw and Taylor's (1979) classic chapter, has long been recognized as a potentially troublesome feature of spatially aggregated data, such as census data. Aggregation of high-resolution (i.e. a large number of small spatial units) data to lower resolution (i.e. a smaller number of larger spatial units) areas is an almost unavoidable feature of large spatial datasets due to the requirements of privacy and/or data manageability. When the original data are aggregated, the values for the various univariate, bivariate, and multivariate parameters will more than likely change because of a loss of information. This phenomenon is called the *scale effect*. The N spatial units to which the higher-resolution data are aggregated, such as census enumeration areas or tracts, postal code districts, or political divisions of various levels, are arbitrarily created by some decision-making process and represent only one of an almost infinite number of ways to partition a region into N cells. Each partitioning will result in different values for the aggregated statistics; this variation in values is known as the *zoning effect*. The two effects are not independent, because the lower-resolution

[1] This chapter is based on Reynolds and Amrhein, 1998b, and was actually written before the other papers.

spatial structure may be built from contiguous higher-resolution units, such as census tracts from enumeration areas, and the resulting aggregate statistics will be different for each choice of aggregation.

Several studies (for example, Amrhein and Reynolds, 1996, 1997; Fotheringham and Wong, 1991; Amrhein and Flowerdew, 1993; Openshaw and Taylor, 1979) have been published that study the effects of the MAUP on a number of census datasets. Of these, only Fotheringham and Wong (1991) have examined the effects of the MAUP on multiple regression parameters, pessimistically concluding that its effects on multivariate analysis are essentially unpredictable. Amrhein (1993) presents the results of a statistical simulation of the MAUP by aggregating randomly-generated point data into square grids of various sizes, thus avoiding many of the problems associated with the use of census data. This chapter expands upon the ideas from both, using statistical simulations to study the effects of the MAUP on multivariate analysis. The fact that Steel and Holt's (1996) analytically derived rules for random aggregation agree with Amrhein's (1993) empirical rules corroborates that simulations are an effective tool for examining the effects of the MAUP.

6.3. The synthetic spatial dataset generator

The use of census data imposes a serious constraint upon those who seek to understand the mechanics of the MAUP simply because there is no control over the nature of a region's overall shape; the shapes, sizes and connectivities of its subregions; or the ranges, means, variances and covariances, frequency distributions, and spatial autocorrelations of the variables. The effects of aggregation on a given census variable can be determined readily enough, but few clues to underlying processes can be gleaned because the data cannot be systematically varied to test for the effects of changes. Other weaknesses of census data, such as random rounding and values missing due to the absence or suppression of data, only serve to make the drawing of any conclusions even more difficult. In order to study the MAUP, it is therefore advantageous to be able to construct synthetic spatial datasets over which a researcher can control and systematically vary all of the above features. This chapter employs the dataset generator described in detail in Chapter 3.

Figure 6.1 illustrates the region used for the experiments, which is divided into 400 subregions, along with three sample aggregations.

6.4. The experiments

Spatial autocorrelation is known to play a key role in the MAUP, as is illustrated in the following experiment. Consider a spatial dataset that contains negative spatial autocorrelation; that is, numbers that are dissimilar are located in adjoining regions. In the aggregation process, contiguous regions are joined and the individual variable values are (in this case) replaced by their average, hence creating a new dataset with a reduced variance. With some algebra, it is easy to show that the difference between the original variance and the aggregate variance (weighted by the number of units in each cell) is the sum (again weighted by the number of units) of the variance of the regions within each cell. For the negatively autocorrelated dataset, it is expected that the values in each cell will have a high variance, and hence the change in variance will be relatively large. As the spatial autocorrelation becomes more positive, the expected internal variance within each cell should decrease, since similar values will tend to become more likely to be adjacent, and hence the change in variance should become less. The influence of spatial autocorrelation on the behaviour of bivariate and multivariate statistics is more difficult to assess, however, as Chapter 5 demonstrates for the bivariate case, since each variable's MC and spatial pattern will cause it to respond to aggregation differently.

The experiments in this chapter explore the effects of aggregation on the various parameters of the linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Three independent parameters are considered to be sufficient to capture enough of the complexities involved in multivariate linear regression without creating excessive computational and analytical overhead. Fotheringham and Wong (1991) use a four-variable regression model, in which the variables are all proportions; their results are compared to ours here.

Three different experiments are performed. In the first, y, x_1, x_2, and x_3 are all assigned the same level of spatial autocorrelation (as measured by the MC). Eight datasets are created in which all four variables have MCs of -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0 respectively, and have zero correlation between them. In the second experiment, x_1, x_2, and x_3 are assigned the same MC, while y is given a different one and again all variables are uncorrelated. Datasets are

created with MCs for dependent and independent variables chosen from -0.4, 0.0, 0.4, and 0.8, for a total of twelve combinations. The third experiment counts the number of statistically significant changes in correlations between variables for the datasets of the first experiment in order to estimate the potential for introduced collinearities. Obviously, having variables with no collinearity is an idealized case, since most variables will have some degree of correlation between them, but it is a good place to start.

The aggregation algorithm is described in detail in Chapter 3. For these experiments, as in Chapters 4 and 5, the regions are aggregated to M = 180, 160, 140, 120, 100, 80, 60, and 40 cells, representing from 45% to 10% of the original 400 regions, in order to assess the scale effect of the MAUP. All of these aggregations are performed independently in a run of the model, and each run is independent of the previous runs. To account for the variability of results introduced by the zoning effect, 1000 runs of the model are performed. After each aggregation, the data are fitted to the multiple linear regression model and the resulting parameters, plus the Moran Coefficient of the regression residuals (MC_{RR}), are saved.

Once all aggregations are completed, the maximum, minimum, mean and standard deviation of each parameter for each scale of aggregation are computed and saved for analysis. The analysis plots (see Figure 6.2b as an example, and Chapter 3 for a more detailed description) are arranged in groups of eight lines, one line for each scale of aggregation, with the labels for each line being listed in the plot's legend. Each group represents a set of initial conditions for an experiment, and is labeled on the plot with (MC_x, MC_y), where MC_x is the MC of the independent variables and MC_y that of the dependent variable. Each line represents the range of values of the parameter that are obtained for the scale over all the runs, and is also marked by the mean value (a heavy dot) and at the mean ± 1 standard deviation (a small horizontal line) to give a rough idea of the distribution of values.

6.5. Results

The results from the first experiment, in which the Moran Coefficients for the dependent and independent variables are the same, show that all of the multivariate regression parameters vary systematically with a change of scale and also with the level of spatial autocorrelation latent in the data. Figures 6.2 to 6.4 illustrate the variations in R^2, the MC of the residuals, and the val-

ues for β_0, β_1, and their standard errors; figures for β_2, and β_3 are similar to those of β_1, and are not shown. All of the figures show the same pattern, with the ranges for all scales decreasing with increasing spatial autocorrelation. This conforms to expectations, since we expect the scale effect to be less severe with greater positive autocorrelation due to more similar values tending to be aggregated. The figures also show that the variation of all parameter values increases with the magnitude of the scale effect over all levels of spatial autocorrelation. This again agrees with expectations, since more information is lost as the data values are aggregated into fewer cells, and with a larger number of regions going into each cell it is expected that there would be a greater degree of variation in results caused by the choice of partition, even for highly spatially autocorrelated data.

Since all the variables are generated randomly and are mutually uncorrelated, the values of R^2 for the unaggregated datasets are all close to zero. Figure 6.2a illustrates that aggregation can produce a model that can have, in extreme cases, from 20% to even 70% of the variation explained by the model, depending on the scale of aggregation and the spatial autocorrelation of the data. The distance of the maximum extreme values from the mean plus one standard deviation mark indicate they are all outliers in the frequency distributions, and as such they will tend to increase the mean value. But even with that in mind it is still apparent that aggregation tends to give models with better fits than the original data, with better fits being associated with greater aggregation. This agrees with expectations, since a reduction in the variability of the data values will tend to produce a better-fitting model (if covariance is also not reduced), but the loss of information caused by reducing the sample size offsets any apparent gain.

Figure 6.2b illustrates the change of the MC_{RR} with aggregation. One of the basic assumptions of a linear regression model is that the residuals are independent, and it is clear that this assumption is being violated since spatially autocorrelated residuals are not independent[2]. Since the initial correlations between the variables are all zero, all of the regression slope parameters are also initially zero so that the initial MC_{RR} will simply be the MC of the deviation of y about its mean, which equals the MC of y. The diagram illustrates the tendency for the regression residuals to become more randomly autocorrelated, with that for the initially negative residuals tending to

[2] Since each observation can be partly predicted from its neighbours, the information content of observations is reduced. See Section 5.3, Griffith (1988, pp. 82-83), and Cliff and Ord (1981, p. 199) for details.

increase, while that for the initially positive ones tending to decrease. The change in residuals for the MC of 1.0 does not follow the pattern of the rest of them, but still does tend to decrease slightly. As with the findings of Chapter 5, it appears that aggregation tends to improve the statistical quality of linear regression, even though it changes all of the parameter values.

Figures 6.3 and 6.4 show that the regression coefficients and their standard errors behave similarly under aggregation. The mean values of the β_0 and β_1 estimates b_0 and b_1 remain close to their unaggregated values over all levels of spatial autocorrelation and all scales. In contrast, the average value of the standard error for all coefficients shows a definite increase with the scale effect. This is not unexpected, as Fotheringham and Wong (1991) point out, since the standard error depends partly on the number of aggregated units. Interestingly, even though the range of variation of the standard error due to the zoning effect decreases with increasing spatial autocorrelation, the mean value for a given scale remains essentially constant. The β_2 and β_3 coefficient estimates b_2 and b_3 and their standard errors behave similarly and are not shown.

The results of the second experiment, in which the independent variables x_1, x_2 and x_3 contain the same level of spatial autocorrelation, while y has a different one, are presented in Figures 6.5 to 6.7. Each plot consists of 12 groups of lines, with each group representing a combination of MCs for the dependent and independent variables. The groups are organized in four sets of three, with each set's dependent variable having the same Moran Coefficient.

As before, the range of variation of the various parameters increases as the scale decreases. Figure 6.5a shows that the range of R^2 decreases as the MC of both the independent and dependent variables increases, though it appears to decrease faster with the increase in the independent variables' MC than with the dependent variable's. This is consistent with the results shown in Figure 6.2a and indicates that, as before, less information is lost when the variables are highly autocorrelated, resulting in smaller variations of the aggregated statistic values.

By examining Figure 6.5b and comparing it to Figure 6.2b, it is apparent that the behaviour of the MC of the residuals depends more on the spatial autocorrelation of the dependent variable than that of the independent variables, since the distributions do not change significantly with the MC of the independent variables. As explained above, this is due to the initial values of the slope parameters being zero, resulting in the initial MC_{RR} being the MC of the dependent variable. As before, the behaviour will depend on the spatial pattern of the variables, not just on their MCs.

As with the first experiment, the regression coefficients and their standard errors each behave in roughly the same way for each combination of spatial autocorrelations. There are three clearly visible patterns, aside from the usual increase in variability with decreasing aggregation scale. First, the mean values of the distributions for the regression coefficients tend to remain fairly stable as the number of aggregate cells decreases, while the means of the standard errors tend to increase. Second, for a given MC of the independent variable, the variability of the ranges increases with increasing MC of the dependent variable, though this effect becomes much less dramatic as the MC of the independent variables increases. The size of some of the ranges is interesting, especially with the intercept parameter b_0 which can be almost 80 above or below the mean of 20 for the 40-cell case in the third from last group in Figure 6.6a. Third, for a given MC of the dependent variable, the range decreases with increasing MC of the independent variables. The patterns are reflected in the those for the standard errors, as shown in Figures 6 and 7 for b_0 and b_1 (those for b_2 and b_3 are similar and not shown). Since the multivariate linear regression model parameter estimates are of the form $\mathbf{b}=(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$, it is expected that variations in the spatial autocorrelation of the independent variables **X** will influence the outcome more than those of the dependent variable **Y**. These figures should serve as a clear warning to those who would blindly use multivariate regression methods on aggregated georeferenced data and then expect the results to apply to a higher resolution!

Comparison of these results with those of Fotheringham and Wong (1991) is difficult because the dependent and each of their four independent variables had a different MC, ranging from almost 0.9 for their P^{black} to about 0.25 for P^{eld}. Even from the very simple second experiment, it is clear that having the dependent and independent variables with different MCs increases the complexity of the response of the regression parameters to aggregation. Differences in the spatial patterns of the variables, as shown above, can also hamper comparisons, as results may be very different for variables with the same MCs.

Fotheringham and Wong's (1991) (hereafter referred to as FW for brevity) analysis of the change in Moran Coefficients of the variables can be compared with experimental results, however, using the diagrams of Chapter 4. Even though the change in the MC depends on the spatial arrangement of the variable, Figures 4.2b, 4.4a, and 4.8 show that the distributions widen as the number of aggregate cells decreases (also shown in FW's Figure 6), and that the mean value ei-

ther decreases or increases monotonically, unlike most of the examples in their Figure 6 which increase and then decrease. These differences could be the result of FW's performing only 20 random aggregations for each spatial scale (20 being not nearly enough to approximate the true distribution of aggregate values), having more than twice the number of base units as we used, and using proportional variables (i.e. numerator and denominator are aggregated separately and the results divided) rather than variables that are simply summed or averaged, or perhaps to unknown violations of the regression model assumptions. Further research needs to be done to study the effects of the MAUP on proportion-type variables.

Also of interest in a study of multivariate linear regression are conditions that violate the assumptions of the model. The easiest one to study is collinearity, the presence of correlation between the independent variables[3]. For this experiment, the datasets used in the first experiment, which all have zero correlation between the variables, are aggregated in the model as before and the number of correlations that are statistically significantly different from zero are counted for each level of aggregation. Table 6.1 summarizes the results for the sets that have MCs of -0.4, 0.2, and 0.8 for the aggregation levels of 180, 100, and 40 cells, while Figure 6.8 illustrates the variation of correlation with MC for the datasets whose variables have the MCs of -0.4 and 0.8. Note that the values in the row labelled *Any* will be less than the sum of the values in the columns if more than one of the correlations is significant at the same time, which occurs frequently for the -0.4 MC case at all levels of aggregation, but less so for the other datasets.

Figure 6.8 and Table 6.1 demonstrate that the ranges of the introduced correlations decrease as the MCs of the variables increase, while as usual the ranges increase with decreasing numbers of cells. The reduction in the range is caused by the decreasing amount of variability lost as the variables become more positively spatially autocorrelated, so as the range decreases fewer values in the distribution cross over into the critical range. As illustrated in Chapter 5, predicting how a pre-existing non-zero correlation between two of the variables will be affected by aggregation is not simple, as the change will depend on the interaction between the spatial distributions of

---

[3] Note that the paper which forms this chapter was initially written before my more detailed analysis of bivariate statistics in Chapter 5. Since the counting of significant changes in r was not a topic discussed in Chapter 5, I decided to leave this in as is.

the variables. The fact that there can be significant changes in the collinearities reinforces the need for caution when using multivariate regression techniques on aggregated data.

6.6. Conclusions

In order to systematically examine the role of spatial autocorrelation in the data on the response of multivariate regression parameters to aggregation, a multiple linear regression model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ was employed, as three independent variables are sufficient to capture much of the complexity of multivariate regression while minimizing the computational and analytical overhead. The first two of the three experiments performed were designed to test the effect of various spatial autocorrelation levels in the independent and dependent variables on the variation of the regression parameters with aggregation. The third experiment tests to see how much collinearity is introduced between independent variables with increasing aggregation, when there was none in the unaggregated data.

When all variables have the same spatial autocorrelation, as measured by the Moran Coefficient, the variation of the parameters tends to decrease as the Moran Coefficient increases, as expected, indicating that more positively autocorrelated data are less affected by the MAUP. For all values of MC tested, the mean values of the coefficient estimates b_0, b_1, b_2 and b_3 are found to be essentially constant over all levels of resolution, even as the range of the distributions increases. Change in the variability is reflected in the standard errors for the coefficients, whose mean values and ranges tend to increase with decreasing spatial resolution. The mean value of R^2 shows a very large variability for negatively autocorrelated data that tends to decrease with increasing values of the Moran Coefficient. The change of the MC of the residuals depends on the MC of the dependent variable more than that of the independent variable, since the initial values of the β coefficients are zero and hence the initial MC_{RR} is that of the dependent variable.

When all of the independent variables have a particular Moran Coefficient, and the dependent variable has a different one, it appears that the MC of the independent variables tends to play a larger role in the variation of the regression coefficients, R^2, and the MC_{RR}, than does the MC of the dependent variable. For a given MC of the dependent variable, the variability in the coefficients and their standard errors tend to decrease with increasing MC of the independent variables. However, for a given MC of the independent variables, the variability tends to *increase*

with increasing MC of the dependent variable. The range of R^2 decreases as the MC of either the dependent or independent variables increase. It appears that the change in MC_{RR} depends on the MC of the dependent variable for initially uncorrelated variables.

Results from the third experiment reveal that collinearities between independent variables can be introduced by aggregation. The mean values of the ranges of correlations remain at or very near 0.0 for all resolutions and MCs of the variables. As one would expect, the ranges of the aggregate correlations are much greater for the variables with low or moderate MC than for those that are more highly autocorrelated, resulting in more statistically significant changes of correlations, many of which will occur simultaneously. Of course few datasets have no correlations between the variables, but it will be difficult to predict the change in a non-zero correlation until a way to incorporate the spatial patterns of the variables into the analysis is found.

The results of the experiments in this chapter only scratch the surface of the behaviour of multivariate regression parameters when data are aggregated from one level of spatial resolution to another. It is clear that the spatial autocorrelation of each of the variables involved influences the behaviour, and that if each variable has a different autocorrelation it will be difficult to predict ahead of time what the behaviour of the regression parameters will be. Exploration of the effect of the MAUP on multivariate regression using variously autocorrelated variables and various degrees of collinearity is a focus for future research.

The variables used in these experiments are all variables that were averaged during the aggregation process. The behaviour of variables that are proportions, in which numerator and denominator are aggregated individually, and variables that are summed in aggregation, also needs to be examined. Comparison of FW's results to ours indicates that multivariate models constructed with variable other than averaged variables may behave differently under aggregation from the model described in this chapter. Models that involve combinations of different variable types may behave even more differently. All of these require further research.

The ultimate goal of the research is, of course, to see if it is possible to empirically estimate error in a spatial dataset that has been introduced by aggregation, and the presence of recognizable patterns indicates that the prospects are perhaps not as gloomy as FW first believed.

Table 6.1: Total number of statistically significant correlations between the variables created by the aggregation process. The number of instances when any of the combinations produced a significant correlation is recorded in the row labelled Any.

	MC = -0.4			MC = 0.2			MC = 0.8		
Cells	180	100	40	180	100	40	180	100	40
y, x_1	20	52	64	1	2	12	0	0	3
y, x_2	13	60	65	0	1	6	0	0	2
y, x_3	27	42	79	0	2	10	0	0	2
x_1, x_2	20	57	60	0	2	1	0	0	0
x_1, x_3	33	45	61	0	0	10	0	0	4
x_2, x_3	14	54	71	0	0	6	0	0	1
Any	120	279	337	1	7	33	0	0	12

6.7. References

Amrhein, C. G., 1993: Searching for the elusive aggregation effect: Evidence from statistical simulations. *Environment and Planning A*, **27**, 105-119.

Amrhein, C. G., and H. Reynolds, 1996: Using spatial statistics to assess aggregation effects. *Geographical Systems*, **2**, 83-101.

Amrhein, C. G., and H. Reynolds, 1997: Using the Getis statistic to explore aggregation effects in Metropolitan Toronto Census data. *The Canadian Geographer*, **41(2)**, 137-149.

Amrhein, C. G., and R. Flowerdew, 1993: Searching for the elusive aggregation effect: Evidence from British census data. Unpublished manuscript available from the authors.

Griffith, D. A., 1988: *Advanced Spatial Statistics*. (Dordrecht: Kluwer).

Fotheringham, A. S., and D. W. S. Wong, 1991: The modifiable area unit problem in multivariate analysis. *Environment and Planning A*, **23**, 1025-1044.

Reynolds, H., and C. Amrhein, 1998a: Using a spatial dataset generator in an empirical analysis of aggregation effects on univariate statistics. *Geog. and Env. Modelling*, **1(2)**, 199-219.

Reynolds, H., and C. G. Amrhein, 1998b: Some effects of spatial aggregation on multivariate regression parameters. *Econometric Advances in Spatial Modelling and Methodology: Essays in Honour of Jean Paelinck*, D. Griffith, C. Amrhein and J-M. Huriot (eds.). Dordrecht: Kluwer.

Openshaw, S., and P. Taylor, 1979: A million or so correlation coefficients: Three experiments on the modifiable area unit problem, in *Statistical Applications in the Spatial Sciences*, ed. N. Wrigley, (London: Pion), 127-144.

Steel, D. G., and D. Holt, 1996: Rules for random aggregation. *Env. and Planning A*, **28**, 957-978.