AGILE: GIScience Series, 3, 30, 2022. https://doi.org/10.5194/agile-giss-3-30-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

Handling the MAUP: methods for identifying appropriate scales of aggregation based on measures on spatial and non-spatial variance

Alexis Comber [1, 2], Paul Harris [3], Kristina Bratkova [1,2], Hoang Huu Phe [4], Minh Kieu [5], Quang Thanh Bui [6], Thi Thuy Hang Nguyen [7], Eric Wanjau[2], and Nick Malleson [1,2]

[1] School of Geography, University of Leeds, Leeds, UK
[2] Leeds Institute for Data Analytics, University of Leeds, UK
[3] Sustainable Agriculture Sciences, Rothamsted Research, North Wyke, UK
[4] RD Consultants, Hanoi City, Vietnam
[5] Faculty of Engineering, University of Auckland, New Zealand
[6] Faculty of Geography, VNU University of Science, Hanoi, Vietnam
[7] VNU Vietnam Japan University, Vietnam National University, Hanoi, Vietnam

Correspondence: Alexis Comber (a.comber@leeds.ac.uk)

Abstract. The Modifiable Areal Unit Problem or MAUP is frequently alluded to but rarely addressed directly. The MAUP posits that statistical distributions, relationships and trends can exhibit very different properties when the same data are aggregated or combined over different reporting units or scales. This paper explores a number of approaches for determining appropriate scales of spatial aggregation. It examines a travel survey, undertaken in Ha Noi, Vietnam, that captures attitudes towards a potential ban of motorised transport in the city centre. The data are rich, capturing travel destinations, purposes, modes and frequencies, as well as respondent demographics (age, occupation, housing etc) including home locations. The dataset is highly dimensional, with a large n (26339 records) and a large m (142 fields). When the raw individual level data are used to analyse the factors associated with travel ban attitudes, the resultant models are weak and inconclusive - the data are too noisy. Aggregating the data can overcome this, but this raises the question of appropriate aggregation scales. This paper demonstrates how aggregation scales can be evaluated using a range of different metrics related to spatial and non-spatial variances. In so doing it demonstrates *how* the MAUP can be directly addressed in analyses of spatial data.

Keywords. Aggregation, Scale, Spatial Support, Ecological Fallacy

1 Introduction

The city of Ha Noi in Vietnam, has increasing levels of air pollution and congestion as the city expands, placing pressure on existing transport infrastructures. Ongoing research has undertaken a survey of travel behaviours in Ha Noi (Bratkova et al., 2022; Malleson et al., 2022; Wanjau et al., 2022). The survey captured answers to 142 questions and had 26339 respondents, whose home locations are shown in Figure 1. One of the aims of the survey was to understand public perceptions of a potential ban on motorbikes from parts of the city centre, asking respondents whether agreed or disagreed with the proposed ban.However, attempts to construct binomial models of the respondent attitudes to the proposed travel ban resulted only in very weak models, despite extensive experimentation, data manipulation and transformation.

One option to overcome this is to spatially aggregate the data, but this introduces the Modifiable Areal Unit Problem or MAUP (Openshaw, 1984b, a; Dungan et al., 2002) and raises the question of determining appropriate scales of aggregation. The MAUP, in brief, is the variation in statistical distributions, relationships and trends when the same raw data are aggregated over spatial units at different spatial scales. Essentially the MAUP describes a process of distortion. It results in different model outcomes, such as coefficient estimates (Brunsdon and Comber, 2021), and importantly different process understandings. It is common to find research papers saying things like "the MAUP should always be tested for" (Hint: this is an easy point to make if you are ever reviewing a spatial analysis paper: "Did you examine the MAUP?"). But rarely is a full examination of the effects of the MAUP undertaken. In a recent paper Comber and Harris (2022) recommended that sensitivity to the MAUP can be investigated by identifying the aggregation scales at which the processes under investigation are considered to be stable, where stability is in respect to "variances, covariances and higher moments, in context of the subsequent data analyses" (p15), as well as measures indicators of spatial association (Anselin, 1995; Hui, 2009) and local spatial covariances (variograms) (Harris et al., 2010).

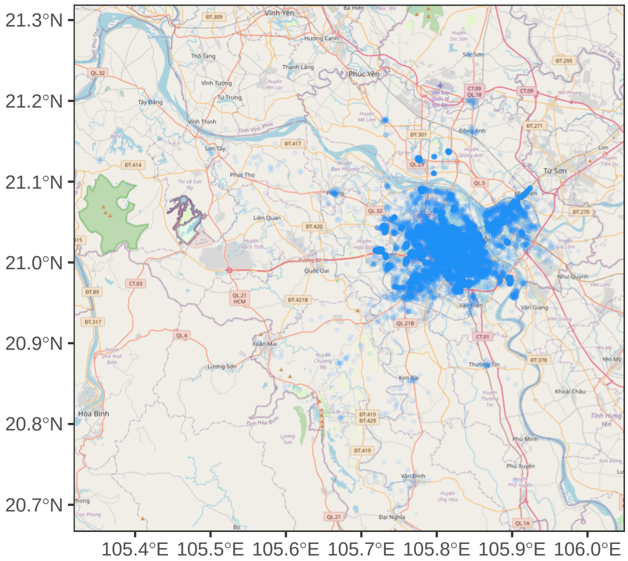


Figure 1. The spatial extent and home locations of the Ha Noi travel survey respondents, with a transparency term and an OSM backdrop (copyright OpenStreetMap contributors).

This paper explores the impact of different spatial of aggregation using the home locations indicated in the travel survey described above. It undertakes an evaluation of the impacts of the MAUP using 6 exploratory measures and then uses these to determine an appropriate aggregation scale before undertaking an analysis of the factors associated with respondent attitudes to the proposed travel ban in Ha Noi.

2 Methods

The survey data were aggregated over spatial aggregation units at different scales and a number of metrics were calculated before a final analysis of behaviours was undertaken at a selected scale.

2.1 Data

A survey of travel behaviours and attitudes has been undertaken as part of the Urban Transport Modelling for Sustainable Well-Being in Ha Noi project[1] in order to support evidence-based policy making around the Ha Noi transport system. The survey asks respondents for basic demographic information as well as details about their travel behaviour (e.g. their common journeys) and their transport related aspirations (e.g. ownership of different types of vehicle). It also asks questions specifically related to a possible ban on motorbikes from the city centre. The aim of this analysis is to examine the response to this question with respect to a mixture of demographic and travel related variables. Specifically the analysis sought to understand how variations in age, gender, occupation, type of residence, typical trip purpose, mode and distance were related to whether the respondent agreed with the ban or not. Aside from distance, all of the variables were categorical (see below). In the aggregations the explanatory variables were converted to counts of each response for each spatial unit, and then rates were calculated based on the number of observations in each unit. For each group of compositional data listed below, one of the categories was dropped as is the usual practical for compositional data in regression:

- Age: less_18, 18_25, 26_35, 36_55, 56_75, more_75; dropped more_75;
- Gender: female, male; dropped male;
- Occupation: retired, student, private, fdi, state; dropped retired;
- Home type: high_rise, private_house, old_building, social_house, private_new, resettlement; dropped old_building;
- Home ownership: owner, rent, parent_house, morgate; dropped parent_house;
- Trip purpose: visit, education, work, shopping, caring, leisure; dropped work;
- Trip mode: taxi, moto, walk, bus, ebike, bike, car, tram; dropped tram;
- Opinion on the proposal ban: agree, disagree, neutral, strongdisagree, strongagree.

The mean travel distance for the main journey made by each respondent was calculated for the respondents in each cell. The response variable (Opinion on the proposal ban) was collapsed such into a binary variable of the proportion of respondents in each cell that agreed with the proposed ban of motorised transport in the city centre in some way – i.e. composed of agree and strongagree – to create a ban agreement response variable.

[1] https://urban-analytics.github.io/UTM-Hanoi/

2.2 Scales

A series of different sized hexagonal grid cells were constructed over which the observations in Figure 1 were aggregated. These ranged from an approximate 10 × 10 coverage of the area containing 102 grid cells, each with an area of 503 km^2, to 100 × 100 grid cells with 7305 grid cells with an area of 5 km^2, as sample of which are shown in Figure 2.

2.3 Metrics

labelmetrics

The survey data were combined over the cells in the aggregation layer in the manner described above and a series of metrics were used to explore the effect of aggregation scale. These were:

- **Variance** of the target variable, ban agreement proportion;
- **Filtered Variance** of the target variable, with the data filtered for cells with more than 5 respondents;
- The residual variogram's **Nugget** effect arising from a linear regression model fitted with a spatially autocorrelated error term, where the variogram is specified using an exponential function decaying with respect to the Euclidean distance separating the cells, and all parameters are estimated using restricted maximum likelihood (REML) (Lark et al., 2006). The Nugget ranges from 0 to 1, where it is the proportion of small-scale random variation to the total variation (both random and structural);
- The residual variogram's correlation **Range** from the REML estimation above, noting that this should always be evaluated relative to scale of aggregation. To achieve this the logged Range is divided by the cell parameter value;
- The number of **PCA Components** that explain 80% of variation in the aggregated data;
- **Moran's** I of the arising from the residuals of the REML model.

The idea was to use these metrics to identify the aggregation scales at which the aggregation process stabilises - i.e. levels out in some way.

3 Results

The results of using different scales of aggregation are shown in in Figure 3. The Variance, Filtered Variance, and PCA Components all provide measures of (a-spatial) variability in the data and the the Nugget, Range and Moran's I, all provide measures of spatial variability. Low values for the variogram's Range indicate weak residual spatial autocorrelation, while high variogram Nugget values similarly indicate weak residual spatial autocorrelation. The weakest residual spatial autocorrelation is when both occur in tandem, and vice versa for the strongest residual spatial autocorrelation. Theoretically, as data are aggregated to coarser scales then the Nugget will approach 1 and Range will approach 0. Unlike the variogram, Moran's I only provides a single metric of residual spatial autocorrelation and as such, provides a less detailed summary.

Examining Figure 3, some trends are evident. The Nugget, Variance, Filtered Variance and PCA Components stabilise around the same range - between 50 and 70. They have plateaus in their trend (before continuing to increase). For the Range, if the elbow is considered, it stabilises between 40 and 70 before continuing to decrease. The spatial autocorrelation in Moran's I is moderately positive across scales and to a degrees shows some levelling off at this range also. Together these metrics, although calculated in different ways and capturing different but related process, suggest that an aggregation scale in the range of 50 × 50 to 70 × 70 grid cells may be appropriate.

The final stage of the analysis was to fit a model, using a sampling grid constructed with approximately 50 × 50 grid cells as shown Figure 4. The approach was as follows: aggregate the data as before, filter the data for cells with more than 5 observations (resulting in 173 grid cells), fit an initial model and then undertake model selection through a stepwise AIC evaluation to determine a final parsimonious model. The coefficient estimates and their significance are shown in Figure 4.

The coefficient estimates in Figure 4 show a number of things:

- the effect of average (mean) trip distance (in kilometres) has only a marginal effect on the proportion of people who agree with the proposed ban;
- in terms of demographics, the proportion of females and people aged 26 to 75 are negatively associated with the proportion agreeing with the ban, and the proportion of people living in resettlement accommodation highly so;
- the proportions of people employed by the state in some way or Foreign Direct Investment (`fdi`) were positively associated with agreement with the proposed ban, as were those living in their own homes;
- in terms of the nature and mode of the main journey or trip, the proportion of people going shopping for their main was negatively associated with the ban and unsurprisingly the proportion of people using a motorbike for that trip was also negatively associated with agreement to the proposed ban.

To emphasis the importance of this, a comparative analysis was undertaken with 25 × 25 grid cells (514 cells) and the

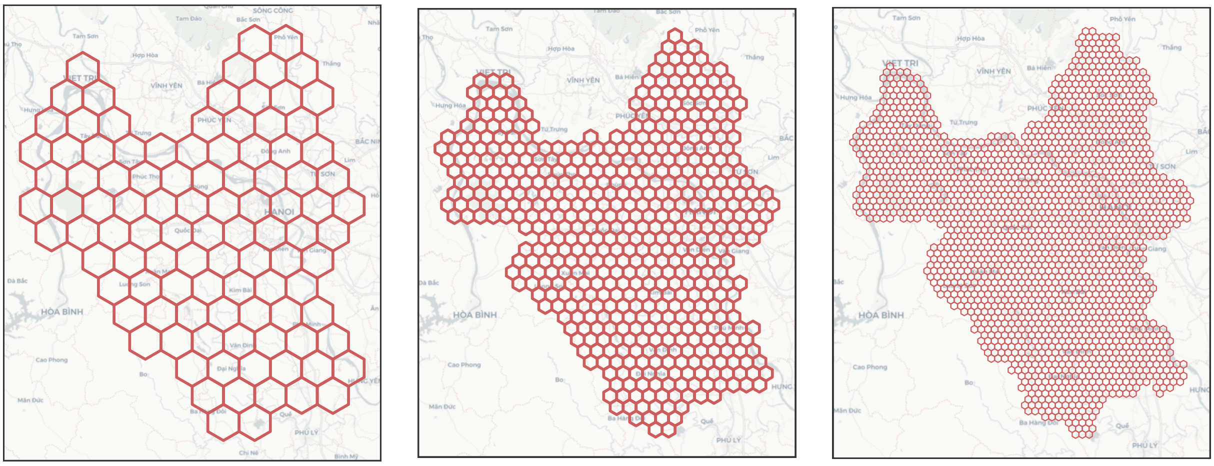


Figure 2. Example scales of aggregation: left 10 by 10 cells centre 25 by 25, right 50 by 50.

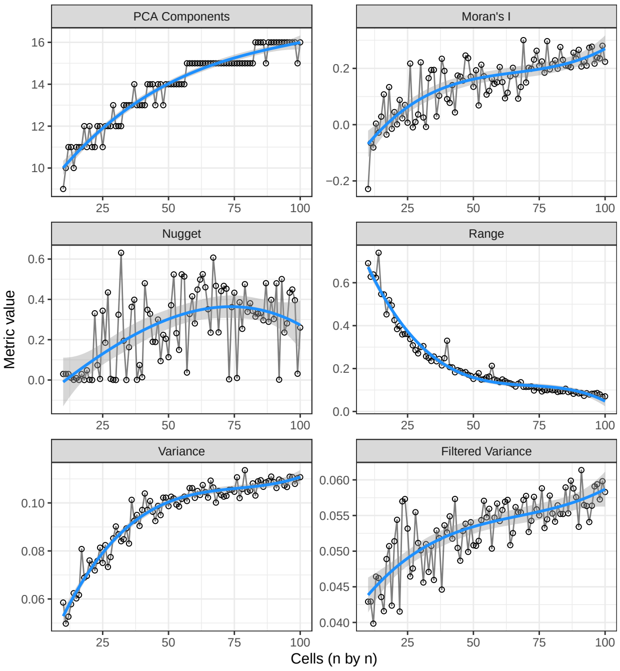


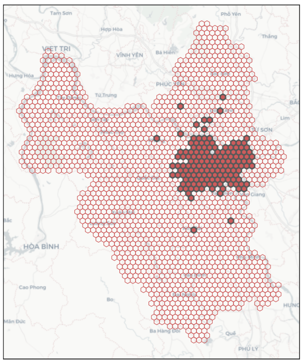
Figure 3. The metric scores under different scales of aggregation (grid cell number), with a 3rd order polynomial trend line fitted.

results are shown in Figure 5. There are number of things to note:

1. The AIC selected parsimonious model has a different composition. If the the names of the covariates are examined, this spatially coarser model has retained different age, occupation, ownership, trip purpose and vehicle variables.

2. The strength of the relationship between the covariates and target variable (ban agreement percentage) has changed in some cases (`own_owner`, `purp_shopping`, `vehic_moto`).

3. The nature of the relationship between the covariates and the target variable has changed. The signs are reversed for the age covariates, indicating a different relationship with the outcome.



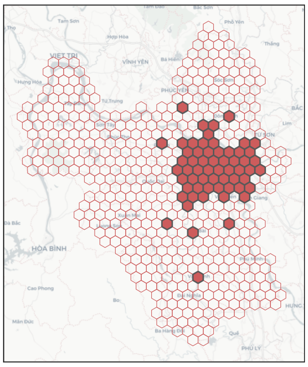
	Estimate	Std. Error	Pr(>|t|)
(Intercept)	0.561	0.092	0.000
distance	0.007	0.002	0.003
gender_female	−0.196	0.061	0.002
age_26_35	−0.454	0.115	0.000
age_36_55	−0.634	0.113	0.000
age_56_75	−0.470	0.144	0.001
occup_state	0.584	0.086	0.000
occup_fdi	0.512	0.090	0.000
type_resettlement	−3.108	1.320	0.020
own_owner	0.238	0.062	0.000
purp_shopping	−0.285	0.136	0.038
vehic_moto	−0.163	0.062	0.009
vehic_bike	−0.273	0.185	0.143

Figure 4. The grid cells from a 50 by 50 model, with the filtered grid cells (shaded) used in the final analysis, and the summary of the coefficient estimates, standard errors and significance, from the final fitted model.

Together these suggest a very different understanding of the processes and factors relating to the degree of ban agreement.

4 Discussion

This paper presents work from a project where the aim is inform policy makers of the travel patterns and nature of the journeys people make as part of their every day routines in Ha Noi. Data collection was delayed due to COVID-19 and has only just been completed (Spring of 2022). The dataset is extremely rich both in terms of the number and geographical spread of observations but also in its thematic content. It also incredibly noisy. In this context, spatial aggregation provides a method to summarise the data into a more manageable format from which understandings and trends can be extracted, in order to identify potential strategies to mitigate the effects on pollution and congestion in the city, which are increasing with



	Estimate	Std. Error	Pr(>|t|)
(Intercept)	−1.667	0.599	0.007
distance	0.007	0.004	0.051
age_18_25	2.274	0.628	0.001
age_56_75	1.911	0.602	0.002
age_36_55	2.273	0.627	0.001
age_26_35	2.224	0.635	0.001
occup_private	−0.354	0.116	0.003
own_owner	0.394	0.109	0.001
own_morgate	1.479	0.570	0.012
purp_shopping	−0.752	0.150	0.000
purp_education	0.225	0.114	0.054
purp_leisure	1.451	0.640	0.027
purp_caring	−1.030	0.298	0.001
vehic_moto	−0.371	0.096	0.000
vehic_bus	−1.563	0.562	0.007

Figure 5. The grid cells from a 25 by 25 model, with the filtered grid cells (shaded) used in the final analysis, and the summary of the coefficient estimates, standard errors and significance from the model.

rapid urbanisation. However, spatial aggregation introduces well-known and well-recognised distortions through the MAUP. As yet there is no toolkit to mitigate or understand the effects of the MUAP. This paper provides an initial investigation and demonstrates a set of approaches for doing this, through a simplified case study. It this methodological suggestion which the main contribution of this research, rather than the coefficient estimates as presented in Figure 4 and interpreted above. Future work will examine the relationships with the many other factors captured by the survey that are of interest, and alternative regression models, especially non-linear spatial ones such a spatial GAM splines, to develop more detailed analyses.

Acknowledgements. This work has received funding from the British Academy under the Urban Infrastructures of Well-Being programme [grant number UWB190190]. The data and code will be made available if issues around respondent privacy can be overcome.

References

Anselin, L.: Local indicators of spatial association—LISA, Geographical analysis, 27, 93–115, 1995.

Bratkova, K., Comber, A., Hoang Huu, P., Kieu, M., Malleson, N., Bui Quang, T., Nguyen Thi Thuy, H., and Wanjau, E.: Let the Data Speak for Itself: Developing a New Data Dashboard for a Hanoi Transport Survey, Zenodo, https://doi.org/10.5281/zenodo.6408078, 2022.

Brunsdon, C. and Comber, A.: Opening practice: supporting reproducibility and critical spatial data science, Journal of Geographical Systems, 23, 477–496, 2021.

Comber, A. and Harris, P.: The Importance of Scale and the MAUP for Robust Ecosystem Service Evaluations and Landscape Decisions, Land, 11, 399, 2022.

Dungan, J. L., Perry, J., Dale, M., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti, M., and Rosenberg, M.: A balanced view of scale in spatial statistical analysis, Ecography, 25, 626–640, 2002.

Harris, P., Charlton, M., and Fotheringham, A. S.: Moving window kriging with geographically weighted variograms, Stochastic Environmental Research and Risk Assessment, 24, 1193–1209, 2010.

Hui, C.: A Bayesian solution to the modifiable areal unit problem, in: Foundations of Computational Intelligence Volume 2, pp. 175–196, Springer, 2009.

Lark, R., Cullis, B., and Welham, S.: On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML, European Journal of Soil Science, 57, 787–799, 2006.

Malleson, N., Bratkova, K., Comber, A., Huu, P. H., Kieu, M., Quang, T. B., Nguyen Thuy, H., and Wanju, E.: Upscaling a Spatial Survey with Propensity Score Matching: Implications of a Motorbike Ban in Ha Noi, Zenodo, https://doi.org/10.5281/zenodo.6410193, 2022.

Openshaw, S.: Ecological fallacies and the analysis of areal census data, Environment and planning A, 16, 17–31, 1984a.

Openshaw, S.: The modifiable areal unit problem, CATMOG 38, Geo Abstracts, Norwich, 1984b.

Wanjau, E., Bratkova, K., Comber, A., Huu Hoang, P., Kieu, M., Malleson, N., Quang Bui, T., and Nguyen Thi Thuy, H.: Spatial Interaction Modelling by Transport Mode: A glimpse into the impacts of a Motorbike Ban in Hanoi, Zenodo, https://doi.org/10.5281/zenodo.6408064, 2022.