Assignment Two: The MAUP and Multilevel Modelling

Contents

1 Demonstrating the MAUP

1.1 Background

Areal units in zoning systems amalgamate into objects that constitute the basic units for the observation and analysis of spatial phenomena (Openshaw 2015). Yet, no gold standard for guiding the spatial aggregation process exists, with the validity of zonal objects subject to the arbitrary and modifiable decision-making of quantitative geographers. Problematically, the analysis of socioeconomic data involving areal units is encumbered by the modifiable areal unit problem (MAUP): "the sensitivity of analytical results to the definition of units for which data are collected." According to the literature, the MAUP constrains the reliability of analyses for aggregated spatial data, as findings have shown varying results with the scale of aggregation and configuration of zoning systems (Avery and Clark 2015).

In practice, the MAUP is condensed into two issues of scale and zoning sensitivity which this paper will attempt to demonstrate in Section 1.2. The first issue, described as the *scale problem*, is the variation in findings when data for zonal units are progressively aggregated. This has been demonstrated empirically by Avery and Clark (2015) who found that whilst correlation coefficients did not increase monotonically with aggregation[1], a general increase in data aggregation corresponds to an increase in correlation coefficients.

The second issue, the *zoning problem*, pertains to the variation in findings when alternative combinations of zonal units are analysed with the scale or number of units held constant (Openshaw 2015). Zoning sensitivity in multivariate analysis has been demonstrated empirically in Fotheringham and Wong (2015) who simulated the aggregation of 871 block groups into 218 zones in 150 different iterations. They highlight the severity of the zoning problem by demonstrating the possibility of concluding with one iteration of zones no association between the percentage of blue-collar workers and mean family income, with another iteration finding a unit increase in blue-collar worker percentages as reducing mean family income by $20,000. In all, ignoring scale and zoning sensitivity in model calibration can lead to inferential conclusions that a researcher's areal data is applicable to the constituents who form the zones under study - the ecological fallacy problem (Openshaw 2015).

[1] At higher levels of aggregation, there is smaller adjacency of zonal units meaning groupings are more heterogenous leading to lower correlation coefficients.

1.2 MAUP analysis

1.2.1 Data

To demonstrate the scaling and zoning sensitivities of the MAUP, we calculate the bivariate strength of association between two open data variables. For our first variable, *crime_count*, we submit a HTTP request using the POST verb to send a custom polygon for retrieving all street-level crimes occuring in 2012.

```r
# download geojson
u <- "http://statistics.data.gov.uk/boundaries/E08000012.json"
# store in temporary directory
downloader::download(url = u, destfile = "/tmp/lpool.geojson")
lpool <- readOGR(dsn = "/tmp/lpool.geojson", layer = "OGRGeoJSON")
# access coords slot
lpool <- lpool@polygons[[1]]@Polygons[[1]]@coords
# build lat/lon + date string to send with postrequest
curl.string <- paste0('poly=',paste0(sprintf('%s,%s',lpool[,2], lpool[,1])
                                     , collapse = ':'))

# build dates list for loop
dates = c("2012-01", "2012-02", "2012-03", "2012-04", "2012-05", "2012-06",
          "2012-07", "2012-08", "2012-09", "2012-10", "2012-11", "2012-12")

# dates = c("2012-01", "2012-02")

document <- lapply(dates, function(month) {
  # format acceptable packet for http request
  curl.string <- list(poly=c(curl.string), date=c(month))
  # post custom polygon to police api
  r <- httr::POST("https://data.police.uk/api/crimes-street/all-crime",
                  body = curl.string, encode="multipart", verbose())
  json <- content(r, "text", encoding = "ISO-8859-1")
  # return as data.frame
  jsonlite::fromJSON(txt=json)
})
```

```r
# cast lat/lon columns to numeric data type
document$lat <- as.numeric(document$lat)
document$lon <- as.numeric(document$lon)

# convert data.frame to shapefile in rgdal
coordinates(document) <- ~lon+lat
proj4string(document) <- CRS("+init=epsg:4326") # WGS 84

# write once
writeOGR(document, "/Users/samcomber/Documents/spatial_analysis/shp/crimes",
         "crimes", driver = "ESRI Shapefile")
```

Regarding our second variable, *tweet_count*, we aggregate geo-referenced tweets containing timestamps relating to Twitter postings within the municipality of Liverpool for 2012[2].

[2] This dataset was data mined by Guy Lansley from UCL, and processed by Dani Arribas-Bel.

```
tweets <- readOGR(dsn = "/Users/samcomber/Documents/spatial_analysis/shp/tweets", layer = "tweets_liverp
tweets_sample <- as.data.frame(tweets[sample(nrow(tweets), 1000),])
```

To demonstrate the MAUP, we aggregate a count of each crime and tweet into separate variables for each region in the shapefile, before computing correlations between the two vectors of values in the dataframe. Specifically, we investigate the scaling problem using a hex binned lattice which **. Finally, we demonstrate the zoning problem by constraining cardinality - i.e. the number of regions in the polygon shapefile - to 100 regions, but generating 200 iterations of randomly .

1.2.2 MAUP findings

** CORR TABLE **

2 Multilevel Modelling

In this section, we implement a two-level multilevel model that accounts for spatial heterogeneity effects between the levels.

2.1 Model specification

```
# msoa and lsoa geographies were joined by location in QGIS
# as R lacks any robust library for polygon.in.polygon joins
lsoa.join <- readOGR(dsn = "/Users/samcomber/Documents/spatial_analysis/shp/
                      spatial_join", layer = "lsoa_join", stringsAsFactors = FALSE)
```

** lsoa and oa group variables **

** avg housing price over areal unit **

2.2 Analysis

2.3 Interpretation

3 Bibliography

4 Appendix

Avery and Clark. 2015. "R: A Language and Environment for Statistical Computing." Journal Article. http://www.R-project.org.

Fotheringham and Wong. 2015. "R: A Language and Environment for Statistical Computing." Journal Article. http://www.R-project.org.

Openshaw. 2015. "R: A Language and Environment for Statistical Computing." Journal Article. http://www.R-project.org.