The MAUP and UK Census Data Aggregation: Utilising Multi-level Models

201374125[a]

[a]Department of Geography and Planning, University of Liverpool, Liverpool, L69 7ZX

This version was compiled on March 6, 2019

The MAUP effects any spatial data that has some level of aggregation, using Multi-level models can help produce standard errors that more accurately represent the between-group variance in spatially aggregated data.

Modifiable Aerial Unit Problem | Multi-level Model | Property Price | Crime

1. Introduction

1.1. Property Price and Crime. Thaler (1978) conducted the first study that considered the effects of crime on property price through a pricing model, and found that people are willing to pay a premium for properties in areas where crime is lower. Hellman and Naroff (1979) similarly constructed a model from census data which utilised the inclusion of amenities to assess how the 'willingness to pay' for housing is affected by increases in the amenity level of a site, in addition to the level of crime. A study by Lynch and Rasmussen (2001) found conflicting results and state that simply counting crime distribution, due to the difference in reporting behaviour, provides a distorted view of how crime varies over space.

Gibbons and Machin (2008) considered the local variation in property prices in relation to crime and note that Macroeconomic and regional level models have highlighted what drives property prices at this scale, but localised variation in sub regional property prices are not explained by these frameworks.

Empirical findings from crime property price studies show that typically cost of housing and crime are negatively correlated, but sometimes do not fully consider the selection of their spatial scale. Table 1 gives an overview of the different spatial scales selected for the comparison between property price and crime, ordered from the largest to smallest.

Table 1. The Varying Spatial Scale of Property Price Research

Reference	Scale
Lynch and Rasmussen (2001)	City (Metropolitan)*
Thaler (1978)	City
Hellman and Naroff (1979)	City Centre
Gibbons and Machin (2008)	City (Sub Regions)

**A study of 11 separate metropolitan areas*

1.2. Multi Level Modelling and the MAUP. Multi-level modelling can be used to produce statistical analysis on data that has some level of aggregation, and therefore may have an inherent hierarchical structure (Dong et al., 2019). This differs to basic linear modelling, where all residuals (e_i) are assumed to have no correlation (Steele, 2019). With spatial data, the Modifiable Aerial Unit Problem (MAUP) arises with any level of point aggregation, meaning the shape and scale of the aggregation unit has a large influence on the resultant analysis (Fotheringham and Wong, 1991). In this case, individual properties are sold at a specific address, this is typically aggregated minimally at least to postcode, then

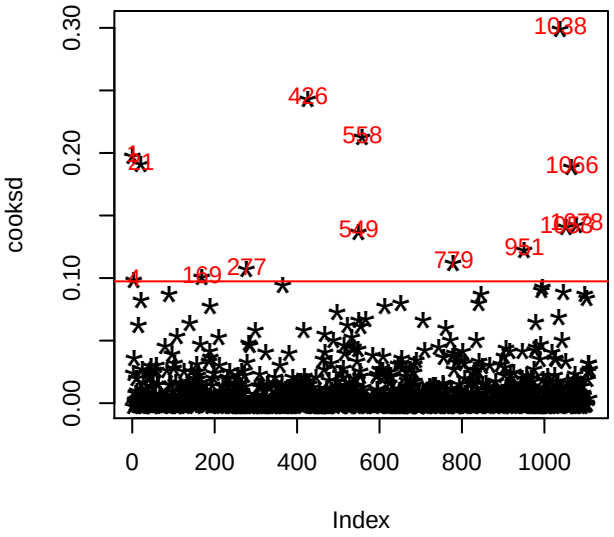

Fig. 1. Influential Obs by Cooks distance

in the examples above, into various sub-regional units, city level, metropolitan, and regional. Without implementing the use of a multi-level model, these hierarchical units are predetermined, typically due to government policy, which does not necessarily directly relate to the outcome variables being considered, and thus are considered to be used only to provide a usable level of aggregation and hierarchy, resulting in the MAUP (Jones et al., 2017). Multi-level modelling allows for an investigation of between group variability, by analysing macro and micro level models simultaneously. Meaning the effects of certain group level characteristics on individual outcomes can be assessed, which in turn reduces the effects of ecological fallacy (Steele, 2019).

2. Methodology

2.1. Data Sources. This article analyses Price Paid Data (PPD)[1] for the Liverpool Local Authority District (LAD), provided by the HM Land Registry. Data was downloaded in .csv format and manipulated entirely through *R* (R Core Team, 2018) to ensure full reproducibility (See Appendix I). In addition, census data provided population and ethnicity estimates which were collected during the 2011 Census (27th March 2011), commissioned by the UK government for the Office of National Statistics (ONS, 2011). Crime data was downloaded from the Open Police Data website.

Shapefiles containing varying resolutions of census geography were provided by the Consumer Data Research Centre (CDRC), ranging from the smallest resolution 'Output Areas' (OA), 'Lower Super Output Areas' (LSOA), to 'Middle Super Output Areas (MSOA). These areas are defined by the ONS to provide statistics for smaller

[1]HM Land Registry data ©Crown copyright and database right 2017. Data is licensed under the Open Government License v3.0.

areas during the UK census (Dong et al., 2019). Output area classifications are determined for each census, primarily considering homogeneity within areas, suggesting heterogeneity between areas (Fotheringham and Wong, 1991).

Price paid data was provided at street level, but aggregated to postcode in *R* to ensure easy manipulation. Postcodes were checked for validity and their coordinates obtained from the free open source Postcodes API giving a total 4975 separate postcodes with price paid data attached. Crime data was provided with coordinate data so no postcodes were required, with a total of 14592 data points.

2.2. Predictor Variables. Proximity to the city centre as suggested in Cheshire and Sheppard (1995), is a simplified method that aggregates both the consideration of proximity to amenities (e.g. Gibbons and Machin, 2008; Hellman and Naroff, 1979), and the general effect of increased property price trending toward city centres due to high demand.

Minority groups and levels of immigration are shown to reduce property price (Sá, 2011), and Socially Rented properties are typically a lower than average property value.

2.3. Null Model. First a null MLM was run, exuding any predictor variables but allowing for intercept variance at MSOA level, expressed below:

$$\overline{y}_{i,j} = \beta_0 + \mu_{0,j} + \epsilon_{i,j} \tag{1}$$

Where:

$\overline{y}_{i,j}$: is the average log property sale price per i (OA) and j (MSOA)

β_0: is the intercept

$\mu_{0,j}$: is the random part (Level 2: MSOA)

$\epsilon_{i,j}$: is the random part (Level 1: OA)

2.4. Random Intercept Multi-level Model. Individual points that were determined to have a significant effect on the model were identified through the use of Cook's Distance (Loy and Hofmann, 2013).

Given the literature, and logical reasoning that crime is likely to directly reduce property price, and never increase it, a random intercept model was chosen. The second model (see Table 2) includes the proportion of minority groups, crime as a proportion of the OA population, and includes an MSOA scale predictor variable (`distCity`) representing the distance of the centre of each MSOA to the city centre.

$$\begin{aligned}\overline{\log(PropertyPrice)}_{i,j} =& \beta_0 + \beta_1 * minority_{i,j} \\ &+ \beta_2 * S\_Rent_{i,j} \\ &+ \beta_3 * \text{crime}_{i,j} \\ &+ \gamma * distCity_j + \mu_{0,j} + \epsilon_{i,j}\end{aligned} \tag{2}$$

Where:

γ: is the overall regression coefficient between $PropertyPrice$ and $distCity$

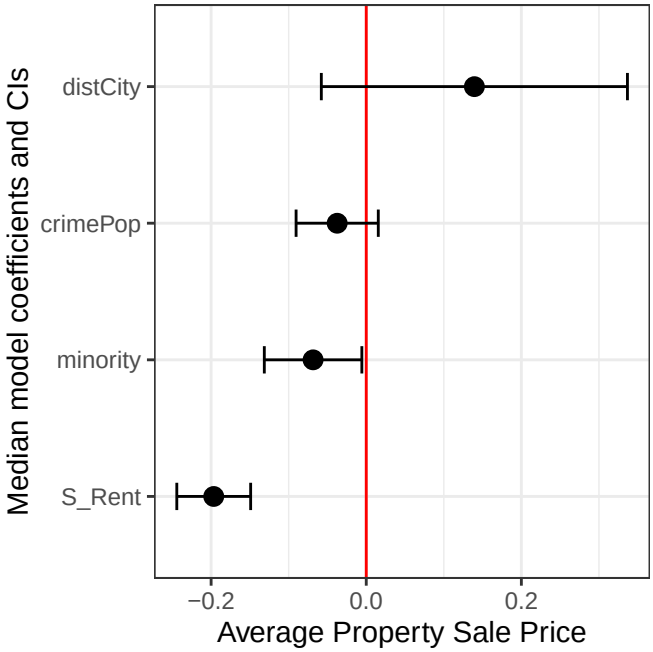

Fig. 2. Pattern of Fixed Intercepts (Model 1)

3. Data Descriptions

PPD were extracted from 2018 onwards giving a total of 9165 property transactions within the Liverpool region. Due to a very large standard deviation with this dataset (£112,087.41) relative to the mean (£163,145.44), to reduce homoskedasticity in the model estimate, prices were log transformed (Dong et al., 2019).

Crime data was extracted from 2018 onwards and filtered to include what were considered to be only crimes that pose a significant financial cost to victims, first considered in Cohen (1990) and reiterated in Lynch and Rasmussen (2001) as an important distinction when considering the relation of crime and property value (See Appendix I). All data was standardised due to the significant variation in scale to allow direct comparison (Eames, Ben-Shlomo and Marmot, 1993)

4. Results and Discussion

A general increase in correlation between two variables at higher level of aggregation results directly due to the way correlation is calculated, see Equation 3 (Fotheringham and Wong, 1991). This is reflected in these data; correlation between crime and property price at OA : −0.07, $LSOA$: −0.14, and $MSOA$: −0.24.

$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y} \tag{3}$$

Cooks distance (see Loy and Hofmann, 2013) reveals that several points have a very significant influence on the model (Figure 1), inclusion of these points led to a model which was unable to function (residual errors showed zero standard deviation). This is mainly due to some extreme outliers in the crime data. One data point, 23.87 standard deviations above the mean may be the result of issues with the data quality, so identified points were removed.

4.1. Null Model. The Null model estimate results reveal that MSOA-level variance is 0.647, and OA-level variance is 0.312, giving a Variance Partition Coefficient (VPC) of 0.65. Therefore around

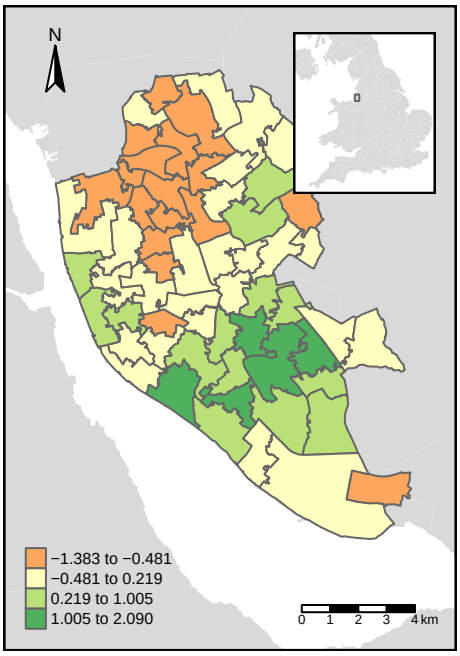

Fig. 3. Pattern of Random Intercepts (Model 2)

Table 2. ANOVA Comparison Between Models

Model	df	AIC	Statistic	P Value
Null Model	4	2151.40		
Model 1	6	2083.13	72.26	0.00
Model 2	7	2083.16	1.97	0.16

65% of the variation in property price is due to differences between MSOAs, suggesting that it is necessary to utilise a Multi-level model to simultaneously look at property price at the OA and MSOA scale.

Comparing the influence of Crime estimate on Property Price in a single-level linear model (Crime $= -0.12, \text{SE} = 0.04$) to the output of a multi-level model (Crime $= -0.047, \text{SE} = 0.028$) shows the extent to which an incorrect model without considering the MAUP can affect results.

4.2. Multi-level model with Predictor variables (Model 1). The VPC for the full MLM is -0.16, as expected this value is far lower with the inclusion of predictor variables, representing the factors that effect the between-MSOA variance, shown graphically on Figure 2.

In order to simplify the model an *ANOVA* test was run comparing the model with, and without the `distCity` MSOA level predictor variable (Table 2). Given the Median model regression coefficients (fixed effects) suggested that this variable had no significant effect on the model (Figure 2; Table 2). The result of the *ANOVA* revealed a non-significant difference between the two models ($pchisq = 0.160$). Therefore this variable was removed giving the new model:

$$\begin{aligned}\overline{\log(PropertyPrice)}_{i,j} =& \beta_0 + \beta_1 * minority_{i,j} \\ &+ \beta_2 * S\_Rent_{i,j} \\ &+ \beta_3 * crime_{i,j} \\ &+ \mu_{0,j} + \epsilon_{i,j}\end{aligned} \tag{4}$$

Table 3. Multi-level Model Summary

Term	Estimate	Std Error	Statistic	Group
Property Price	0.03	0.09	0.29	fixed
Crime	-0.04	0.03	-1.55	fixed
Minority	-0.08	0.03	-2.43	fixed
Social Rented	-0.20	0.02	-8.31	fixed
MSOA: Std Dev	0.73			MSOA_CD
Residual Std Dev	0.57			Residual

4.3. Final Model (Model 2). The Group level random effects were re-calculated after removal of `distCity`, and used to produce Figure 3. This figure reveals the group level variance (MSOA level), influenced by between group heterogeneity. The VPC for this model is essentially 0, suggesting explanatory variables chosen are sufficient to explain the between MSOA variation in property price. This then reveals that Figure 3 provides an estimate for the variation in random effects between the property price at MSOA scale.

The variability (standard deviation) of the MSOA level random effect is 0.73, and the residual variability is 0.57 (Table 3). This suggests that there is still a large amount of random deviation outside the scope of the chosen predictor variables.

Fixed effects show that despite the high residual variability, chosen predictor variables do reduce the property prices, with the highest estimate given by Social Rented (Table 3). Suggesting that above all, the proportion of socially rented properties in an area most significantly reduces the average property price. The low effect due to crime could be due to differences in reporting between areas (Lynch and Rasmussen, 2001), or due to high property prices focusing towards city centres, where crime is high (Cheshire and Sheppard, 1995).

References

Cheshire, Paul and Stephen Sheppard. 1995. "On the Price of Land and the Value of Amenities." *Economica* 62(246):247.

Cohen, Mark A. 1990. "A Note on the Cost of Crime to Victims." *Urban Studies* 27(1):139–146.

Dong, Guanpeng, Levi Wolf, Alekos Alexiou and Dani Arribas-Bel. 2019. "Inferring Neighbourhood Quality with Property Transaction Records by Using a Locally Adaptive Spatial Multi-Level Model." *Computers, Environment and Urban Systems* 73:118–125.

Eames, M, Y Ben-Shlomo and M G Marmot. 1993. "Social Deprivation and Premature Mortality: Regional Comparison across England." *BMJ* 307(6912):1097–1102.

Fotheringham, A S and D W S Wong. 1991. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and Planning A: Economy and Space* 23(7):1025–1044.

Gibbons, Stephen and Stephen Machin. 2008. "Valuing School Quality, Better Transport, and Lower Crime: Evidence from House Prices." *Oxford Review of Economic Policy* 24(1):99–119.

Hellman, Daryl A and Joel L Naroff. 1979. "The Impact of Crime on Urban Residential Property Values." p. 8.

Jones, Kelvyn, David Manley, Ron Johnston, Dewi Owen and Chris Charlton. 2017. "Modelling Residential Segregation as Unevenness and Clustering: A Multilevel Modelling Approach Incorporating Spatial Dependence and Tackling the MAUP- Includes Code in MLwiN, Stata and R. Revised Version January 2018." *Submitted* .

Loy, Adam and Heike Hofmann. 2013. "Diagnostic Tools for Hierarchical Linear Models." *Wiley Interdisciplinary Reviews: Computational Statistics* 5(1):48–61.

Lynch, Allen K. and David W. Rasmussen. 2001. "Measuring the Impact of Crime on House Prices." *Applied Economics* 33(15):1981–1989.

ONS. 2011. "2011 Census - Office for National Statistics." https://www.ons.gov.uk/census/2011census.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sá, Filipa. 2011. "Immigration and House Prices in the UK." p. 31.

Steele, Fiona. 2019. "LEMMA: 5.0 Introduction to Multilevel Modelling: C 5-0." https://www.cmm.bris.ac.uk/lemma/mod/lesson/view.php?id=274.

Thaler, Richard. 1978. "A Note on the Value of Crime Control: Evidence from the Property Market." *Journal of Urban Economics* 5(1):137–145.

Apendix I: Code

201374125[a]

[a]Department of Geography and Planning, University of Liverpool, Liverpool, L69 7ZX

This version was compiled on March 6, 2019

1. Data Cleaning

```
library(rgdal) # spatial tools
library(maptools) # more spatial tools
library(classInt) # spatial tools
library(lme4) # lmer multi level models
library(merTools) # modelling functions
library(ggplot2) # visualisations
library(sf) # for processing geopackage files/shapefiles
library(tmap) # for plotting maps/choropleths
library(RCurl) # api querying
library(jsonlite) # api querying
library(httr) # api querying
library(dplyr) # tidy data tools
library(data.table) #alternative data structures
library(kableExtra) # extra table functions
library(scales) # convert numbers to currency
library(broom) # convert summaries into data.frames
```

```
## Read in all spatial data ##
liv <- st_read("./data/Liverpool/shapefiles/OA.shp")

OA <- st_read("./data/Liverpool/shapefiles/Liverpool_oa11.shp")
LSOA <- st_read("./data/Liverpool/shapefiles/Liverpool_lsoa11.shp")
MSOA <- st_read("./data/Liverpool/shapefiles/Liverpool_msoa11.shp")
####
## Read and tidy census csv data
livPop <- read.csv("./data/Liverpool/tables/CT0010_oa11.csv")
livPop$minority <- rowSums(livPop[,4:96]) / livPop[,2]
livPop <- livPop[,c(1,2,97)]
colnames(livPop)[1] <- "OA_CD"

# spatial join census data to OA
liv <- merge(liv, livPop, by = "OA_CD")

## Read name points, choose only Liverpool centre ##
centre <- st_read("./data/names/NamedPlace.shp")

centre <- centre[centre$distname == "LIVERPOOL",]
centre <- st_as_sf(centre)
centre <- st_transform(centre, 27700) # transform to BNG
```

```
# find centre of each MSOA
livCentroid <- st_centroid(MSOA)

# find distance from city centre to each MSOA centre
MSOA$distCity <- st_distance(x = livCentroid, y = centre)
MSOA$distCity <- as.numeric(MSOA$distCity)
temp <- as.data.frame(MSOA)
```

```
temp <- subset(temp, select = -geometry)
liv <- merge(liv, temp, by.x = "MSOA_CD", by.y = "MSOA11CD") #spatial join
```

```
## Read in price paid data ##
pp <- read.csv("./data/pp-2018.csv", header = FALSE)

pp <- pp[pp$V5 != "O",] # exclude 'other'
pp <- pp[pp$V15 != "B",] # filter out non market sales
pp <- pp[c("V2", "V3","V4","V12")]

colnames(pp) <- c("price", "date", "postcode", "V12")

## Change date to a proper date format, remove time
pp$date <- lapply(pp$date, as.character)
pp$date <- substr(pp$date,1,nchar(pp$date))
pp$date <- as.Date(pp$date, format= "%Y-%m-%d")

# remove outlier dates
pp <- subset(pp, date > "2018-01-01")

# only properties within Liverpool
pp <- pp[pp$V12 == "LIVERPOOL",]
pp <- pp[pp$postcode != "",] # remove empty postcodes

# write to csv as this takes a long time to run
write.csv(pp, file = "./data/pp.csv")
```

```
# read in all crime csv, contained in seperate files for each month
csv_files <- dir("./data/crime/", pattern='*.csv$', recursive = T)
csv_files <- paste0("./data/crime/", csv_files) # list all directories
crime <- bind_rows(lapply((csv_files), read.csv)) # bind all csv

# subset to certain crime defined in main paper
crime <- crime[crime$Crime.type == c("Vehicle crime",
                                     "Criminal damage and arson",
                                     "Violence and sexual offences",
                                     "Burglary", "Robbery",
                                     "Theft from the person"),]

crime <- crime[,c(2,5:7,8)]

# again write to csv as takes a long time
write.csv(crime, file = "./data/crime.csv")
```

```
# read in new price paid csv
pp <- read.table("./data/pp.csv", sep = ",", header = TRUE)
```

```
# find unique postcodes and turn to vector
pcd_list <- pp[!duplicated(pp$postcode), ]
pcd_list <- as.vector(pcd_list$postcode)

# empty list
datalist <- list()

## Loop to check all postcodes and find coordinates
for( i in 1:length(pcd_list))
  tryCatch({ # try catch makes loop continue on error
  print(i) # ensure loop is working
```

```
  pcd <- pcd_list[i] # iterate list
  postcodes_io <- "https://api.postcodes.io/postcodes/"
  postcodes_io_results <- fromJSON(paste0(postcodes_io,pcd)) # api for each pc
  pcd <- data.frame(pcd,
                    postcodes_io_results$result$longitude,
                    postcodes_io_results$result$latitude) # get long and lat

  datalist[[i]] <- pcd # save to list

   if(i %% 50==0) { # api restricts large searches so delay every 50 for 2s
    date_time<-Sys.time()
    while((as.numeric(Sys.time()) - as.numeric(date_time))<2){}
      cat(paste0("block: ", i/10, "\n"))
   }
}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")}) # list errors

# fast list into data frame
postcodeTxt <- data.table::rbindlist(datalist, fill = TRUE)

# remove duplicates
postcodeTxt <- postcodeTxt[!duplicated(postcodeTxt), ]
# write to csv as this takes a long time
write.csv(postcodeTxt, file = "./data/ppPostcodes.csv")
```

```
#read in pp postcodes
postcodes <- read.table("./data/ppPostcodes.csv", sep = ",", header = TRUE)
postcodes[1] <- NULL
colnames(postcodes)[1] <- 'postcode'

#join to price paid data
ppPostcodes <- merge(x = postcodes, y = pp, by = "postcode", all.x = TRUE)

# group by postcode and provide mean property value
ppPostcodes <- ppPostcodes %>%
   group_by(postcode) %>%
   summarise(price=mean(price, na.rm = TRUE))

# join to postcodes list again
ppPostcodes <- merge(x = ppPostcodes, y = postcodes, by = "postcode",
                     all.x = TRUE)
# set coordinate columns
coordinates(ppPostcodes)= ~postcodes_io_results.result.longitude +
                           postcodes_io_results.result.latitude
# crs is WSG
proj4string(ppPostcodes)=CRS("+init=epsg:4326") # set it to lat-long

# change to BSG
ppPostcodes <- st_as_sf(ppPostcodes)
ppPostcodes <- st_transform(ppPostcodes, 27700)
```

```
# read in crime data
crime <- read.csv("./data/crime.csv")
crime$crime <- 1 # count of crime point
coordinates(crime) <- ~Longitude + Latitude
proj4string(crime)=CRS("+init=epsg:4326") # set it to lat-long

crime <- st_as_sf(crime)
crime <- st_transform(crime, 27700)
```

```
# Spatial join crime points to OA and group by sum of points in each
crimeOA <- st_join(crime,OA) %>%
  group_by(OA11CD) %>%
  summarize(crime = sum(crime))

# remove geometry turn to df
crimeOA <- as_tibble(crimeOA)
crimeOA <- subset(crimeOA, select=-geometry)

colnames(crimeOA)[1] <- "OA_CD"

#####
# Spatial join price paid pc to OA, group by OA and show mean price
priceOA <- st_join(ppPostcodes, OA) %>%
  group_by(OA11CD) %>%
  summarize(price = mean(price))

priceOA <- as_data_frame(priceOA)
priceOA <- subset(priceOA, select=-geometry)

colnames(priceOA)[1] <- "OA_CD"

# merge both to liv spatial data frame
liv <- merge(liv, crimeOA, by = "OA_CD")
liv <- merge(liv, priceOA, by = "OA_CD")
```

```
# turn into normal dataframe and remove geometry
livDf <- as_data_frame(liv)
livDf <- subset(livDf, select=-geometry)
# log house prices due to large scale and range (reference)
livDf$logPrice <- log(livDf$price)
# create crime per population
livDf$crimePop <- livDf$crime / livDf$pop
## Scale the data (normalise/standardise) due to different scales
livDf[,5:26] <- scale(livDf[,5:26])
```

2. Introduction

2.1. Table 1.

```
\begin{table}[ht]
    \caption{The Varying Spatial Scale of Property Price Resarch}\label{t1}
\centering
\begin{tabular}{ll}
  \hline
Reference & Scale \\
  \hline
\cite{Lynch2001} & City (Metropolitan)* \\
\cite{Thaler1978} & City \\
\cite{Hellman1979} & City Centre \\
\cite{Gibbons2008} & City (Sub Regions) \\
   \hline
\multicolumn{1}{l}{\footnotesize\textit{*A study of 11 separate \\
                                        metropolitan areas}}
\end{tabular}
\end{table}
```

3. Methodology

3.1. Null Model.

3.1.1. Equation 1.

```
\begin{equation}\label{eq:null}
\overline{y} _ { i , j } = \beta _ { 0 } + \mu _ { 0 , j } +
                           \epsilon _ { i , j }
\end{equation}

Where:
\begin{itemize}[label=]
\itemsep0em
    \item $\overline{Y} _ { i , j }$: is the average log house sale price per
                                      ${i}$ (OA) and ${j}$ (MSOA)
    \item $\beta _ { 0 }$: is the intercept
    \item $\mu _ { 0 , j }$: is the random part (Level 2: MSOA)
    \item $\epsilon _ { i , j }$: is the random part (Level 1: OA)
\end{itemize}
```

3.2. Random Intercept Multi-level Model.

3.2.1. Cooks Distance (Figure 1).

```
model.1 <- lmer(logPrice ~ distCity + minority + S_Rent + crimePop +
                  (1|MSOA_CD), data = livDf)
summary(model.1)

cooksd <- cooks.distance(model.1) # find cooks distance (influential obs)

plot(cooksd, pch="*", cex=2)  # plot cook's distance
abline(h = 8*mean(cooksd, na.rm=T), col="red")  # add cutoff line
text(x=1:length(cooksd)+1, y=cooksd,
     labels=ifelse(cooksd>8*mean(cooksd, na.rm=T),names(cooksd),""), col="red")

influential <-  cooksd > 8*mean(cooksd, na.rm=T)# influential row numbers

extremeOutliers <- livDf[influential,]
extremeOutliers <- extremeOutliers[order(-extremeOutliers$crimePop),]

livDf <- livDf[!influential, ] # exclude very high crime areas (24 std dev)
```

3.2.2. Figure 2.

```
\begin{figure*}
  \begin{center}
    \includegraphics{a1_files/figure-latex/fig2-1}
  \end{center}
  \caption{(a) MSOA-scale random effects (Model 1).
           (b) Fixed effects and statistical uncertainty measures (Model 1).}
\label{fig2}
\end{figure*}
```

3.2.3. Equation 2.

```
\begin{equation}\label{eq:mlm}
  \begin{aligned}
    \overline{\log ( {HousePrice} )} _ { i , j }
    = & \beta _ { 0 } + \beta _ { 1 } * minority _ { i , j } \\
    + & \beta _ { 2 } * S _ { - } R e n t _ { i , j } \\
    + & \beta _ { 3 } * \text {crime} _ { i , j } \\
    + & \gamma * d i s t C i t y _ { j } + \mu _ { 0 , j } +
        \epsilon _ { i , j }
  \end{aligned}
\end{equation}
```

```
Where:
\begin{itemize}[label=]
\itemsep0em
    \item $\gamma$: is the overall regression coefficient between
                    $HousePrice$ and $distCity$
\end{itemize}
```

4. Data Descriptions

```
# convert means and std to GBP format
sd <- paste0('£',formatC(sd(pp$price), big.mark=',',
                         format = 'f', digits = 2))
mean <- paste0('£',formatC(mean(pp$price), big.mark=',',
                           format = 'f', digits = 2))
```

5. Results and Discussion

```
# random sample of msoa
sample.20 <- sample(unique(livDf$MSOA_CD), 20, replace = FALSE)

#plot correlations between crime and price
ggplot(livDf[livDf$MSOA_CD %in% sample.20,], aes(x = crimePop, y = logPrice)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
 facet_wrap(~ MSOA_CD) +
  xlab("Crimes per Population") +
  ylab("Log Property Price") +
  theme_bw()
```

```
# find variation in correlation between spatial scales
livOA <- data.table(liv)
livLS <- data.table(livOA[,c(3,5:24)])
livLS$crime <- livLS[ ,sum(crime), by = LSOA_CD][,2]
livLS$price <- livLS[ ,mean(price), by = LSOA_CD][,2]

livMS <- data.table(liv[c(2,5:24)])
livMS$crime <- livMS[ ,sum(crime), by = MSOA_CD][,2]
livMS$price <- livMS[ ,mean(price), by = MSOA_CD][,2]
```

```
\begin{equation}\label{eq:cor}
r _ { x y } = \frac { \operatorname { cov } ( x , y ) } { s _ { x } s _ { y } }
\end{equation}
```

5.1. Null Model.

```
model.null <- lmer(logPrice ~ 1 + (1|MSOA_CD), data = livDf)
summary(model.null)

VPC <- 0.6466 / (0.3412 + 0.6466)
VPC
```

```
model.1 <- lmer(logPrice ~ minority + S_Rent + distCity + crimePop +
                  (1|MSOA_CD), data = livDf)
summary(model.1)

VPC <- 0.02362 / (-0.04036 + 0.13062 - 0.19526 - 0.06893 + 0.02362)
VPC
```

5.2. Multi-level model with Predictor variables (Model 1).

```
### For valid likelihood-based model comparisons, the MLM needs to be estimated
### using the ML method rather than the Restricted ML

anova(model.null, model.1)

## Model without dist city effects

model.2 <- lmer(logPrice ~ crimePop + minority + S_Rent +
                  (1|MSOA_CD), data = livDf)

summary(model.2)

anova(model.null,model.2,model.1)

pchisq(1.970, df = 1, lower.tail = FALSE)

## inclusion of city dist has no significant benefit to the model
```

5.2.1. Table 2.

```
table <- tidy(anova(model.null, model.1, model.2))
table <- table[c("term", "df", "AIC", "statistic", "p.value")]
table <- table[order(table$term, decreasing = TRUE),]
table$term <- c("Null Model", "Model 1", "Model 2")

# tell Kable to not show anything for NA values
options(knitr.kable.NA = '')
# Table 2
kable(table, digits = 2, caption = "ANOVA Comparison Between Models",
      longtable = FALSE, booktabs = TRUE,
      linesep = '', format = "latex",
      row.names = FALSE,
      col.names = c("Model",
                    "df",
                    "AIC",
                    "Statistic",
                    "P Value")
      )
```

5.3. Final Model (Model 2).

```
model.2 <- lmer(logPrice ~ crimePop + minority + S_Rent + (1|MSOA_CD),
                data = livDf)

summary(model.2)

VPC <- 0.02793 / (-0.04162 -07608 - 0.19547 + 0.02793)
VPC
```

```
MSOA_re <- REsim(model.2)
MSOA <- merge(MSOA, MSOA_re, by.x = "MSOA11CD", by.y = "groupID", all.x = TRUE)
```

```
# read in geopackage for gbr
uk <- st_read("./data/gbr.gpkg", layer="gadm36_GBR_3")

# set to bng
uk <- st_as_sf(uk)
uk <- st_transform(uk, 27700)
```

5.3.1. Figure 3.

```
# find extent of postcodes (slightly larger than liv region)
livRegion <- st_bbox(ppPostcodes) %>%
  st_as_sfc()

# select only polygons from England and Wales (same as census data)
inset <- subset(uk, NAME_1 == "England" | NAME_1 == "Wales")

# seet inset extent to just england and wales
insetEx <- st_bbox(inset) %>%
  st_as_sfc()

# base map grey
m1 <- tm_shape(uk, bbox = liv) + tm_fill()

# show mean random effects at msoa level
m2 <- m1 + tm_shape(MSOA) +
  tm_polygons(col = "mean",
              title = "",
              palette = "RdYlGn",
              midpoint = 0,
          n = 4, # number of bins
          style='jenks') +
  # thicker frame, legend, stop overlap
  tm_layout(frame.lwd = 2, legend.position = c("left", "bottom"),
            inner.margins = 0.1) +
  tm_compass(type = "arrow", position = c("left", "top")) +
  tm_scale_bar()

ukMap <- tm_shape(uk, bbox = insetEx) + tm_fill() +
  tm_shape(livRegion) + tm_borders(lwd = 1) +
  tm_layout(frame.lwd = 2)

library(grid) # just use for printing inset on top
m2
print(ukMap, vp = viewport(.7, .82, width = 0.25, height = 0.25))
```

5.3.2. Table 3.

```
table <- tidy(model.2)
table$term <- c("Property Price", "Crime", "Minority", "Social Rented",
                "MSOA: Std Dev", "Residual Std Dev")

# tell Kable to not show anything for NA values
options(knitr.kable.NA = '')
# Table 2
kable(table, digits = 2, caption = "Multi-level Model Summary",
      longtable = FALSE, booktabs = TRUE,
      linesep = '', format = "latex",
      row.names = FALSE,
      col.names = c("Term",
                    "Estimate",
                    "Std Error",
                    "Statistic",
                    "Group")
      ) %>%
  # add line after row 5
  row_spec(4, hline_after = T)
```

```
model.2 <- lmer(logPrice ~ crimePop + minority + S_Rent + (1|MSOA_CD),
                data = livDf, REML = FALSE)
```

```
VPC <- 0.02793 / (-0.04162 -07608 - 0.19547 + 0.02793)
VPC
```