Spatial_Regression_Pilot_Test

Ta Yun Yang

06/17/2020

This document is used to generate a pilot test report for property value evaluation in Los Angeles County incorporating the spatial regression adjustments.

Importing Data

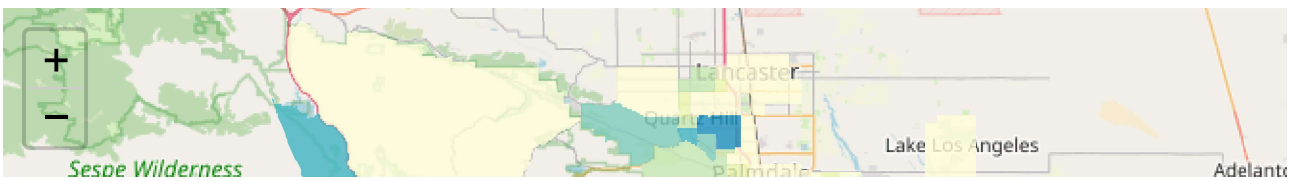
```
parcel <- st_read("D:\\Project Data\\Data_Viz project data\\plot_geo.geojson")
```

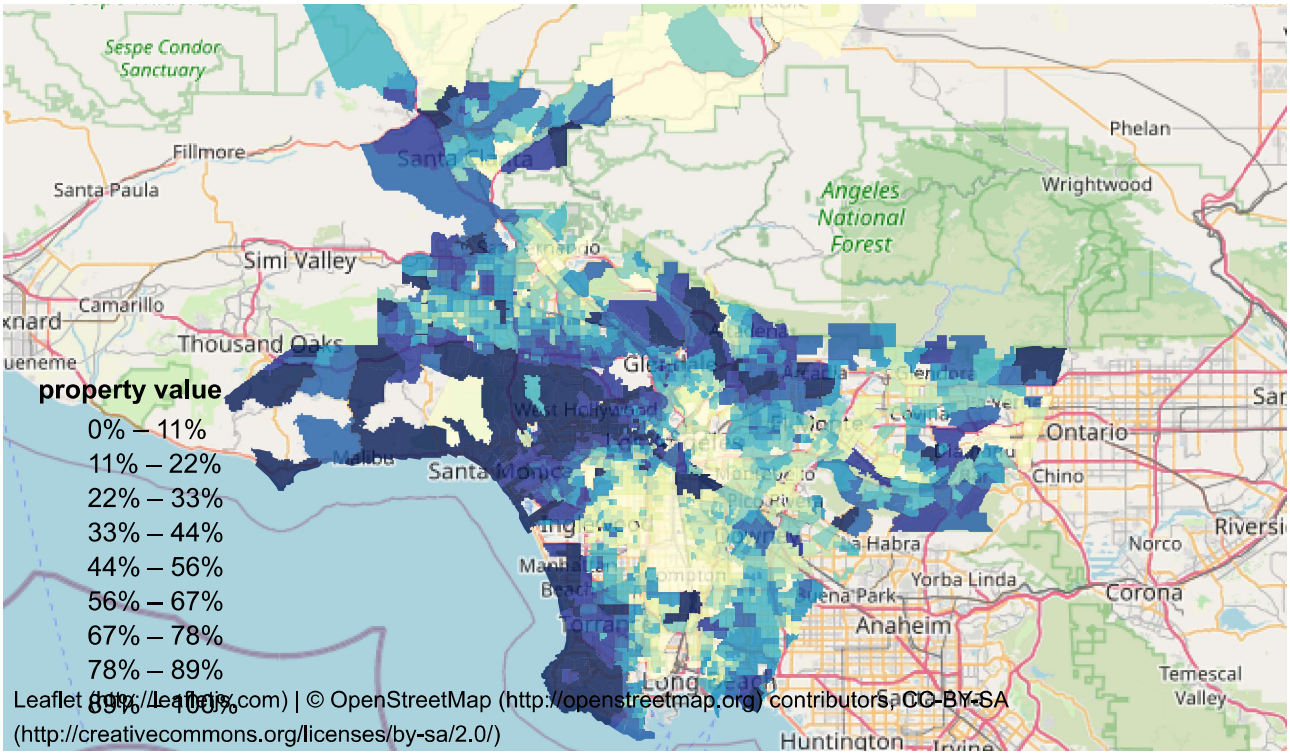
```
## Reading layer `plot_geo' from data source `D:\Project Data\Data_Viz project data\plot_geo.geojs
on' using driver `GeoJSON'
## Simple feature collection with 2216 features and 28 fields
## geometry type:  POLYGON
## dimension:      XY
## bbox:           xmin: -118.9452 ymin: 33.69692 xmax: -117.6552 ymax: 34.8233
## geographic CRS: WGS 84
```

```
selected_cols <- c("CT", "ZIPcode5", "TotalValue", "EffectiveYearBuilt", "SQFTmain",
                   "Bedrooms", "Bathrooms", "geometry")
parcel <- parcel[, selected_cols]

unemployment <- read.csv("D:\\Project Data\\Data_Viz project data\\ACS\\ACS_17_5YR_S2301_with_ann.
csv")
unemployment <- unemployment[2:nrow(unemployment), c("GEO.display.label", "HC04_EST_VC01")]
colnames(unemployment) <- c("GEO.display.label", "unemployment")

unemployment$CT <- ""
for (i in seq(1, nrow(unemployment), 1)){
    if (substr(unemployment$GEO.display.label[i], 18, 18) != ","){
        unemployment[i, "CT"] <- paste0(substr(unemployment$GEO.display.label[i], 14, 17),
                                        substr(unemployment$GEO.display.label[i], 19, 20))}
    else{unemployment[i, "CT"] <- paste0(substr(unemployment$GEO.display.label[i], 14, 17), paste(
"00"))}
}

median_income = read.csv("D:\\Project Data\\Data_Viz project data\\ACS\\ACS_17_5YR_S1903_with_ann.
csv")
median_income <- median_income[2:nrow(median_income), c("GEO.display.label", "HC03_EST_VC02")]
colnames(median_income) <- c("GEO.display.label", "median_income")

median_income$CT <- ""
for (i in seq(1, nrow(median_income), 1)){
  if (substr(median_income$GEO.display.label[i], 18, 18) != ","){
    median_income[i, "CT"] <- paste0(substr(median_income$GEO.display.label[i], 14, 17),
                                     substr(median_income$GEO.display.label[i], 19, 20))}
  else{
    median_income[i, "CT"] <- paste0(substr(median_income$GEO.display.label[i], 14, 17), paste("0
0"))}
}

parcel <- merge(parcel, unemployment, by.x="CT", by.y="CT")
parcel <- merge(parcel, median_income, by.x="CT", by.y="CT")

parcel$ZIPcode5 <- NULL
parcel <- parcel[(parcel$EffectiveYearBuilt != 0) & (parcel$TotalValue != 0), ]
parcel$age <- 2018 - parcel$EffectiveYearBuilt
parcel$logvalue <- log(parcel$TotalValue)
parcel$median_income <- as.numeric(parcel$median_income)
parcel$unemployment <- as.numeric(parcel$unemployment)
```

Choropleth for the Property Values in Los Angeles

```
color_Q <- colorQuantile("YlGnBu", domain=parcel$TotalValue, n=9)
leaflet(parcel) %>%
  addPolygons(stroke = FALSE, fillOpacity = .8, smoothFactor = 0.2, color = ~ color_Q(TotalValue)
  ) %>%
  addTiles() %>%
  addLegend("bottomleft", pal=color_Q, values=~TotalValue,
            title = "property value",
            opacity = 1
  )
```






```
for (i in seq(0, 1, 0.11)){
    print(paste0(paste(i), "th quantile of property value is $", paste(quantile(parcel$TotalValue,
i))))
}
```

```
## [1] "0th quantile of property value is $48036"
## [1] "0.11th quantile of property value is $218459.465"
## [1] "0.22th quantile of property value is $249886.36"
## [1] "0.33th quantile of property value is $277168.295"
## [1] "0.44th quantile of property value is $307175.38"
## [1] "0.55th quantile of property value is $344193.1"
## [1] "0.66th quantile of property value is $392150.18"
## [1] "0.77th quantile of property value is $474087.3"
## [1] "0.88th quantile of property value is $614214.24"
## [1] "0.99th quantile of property value is $1394058.43"
```

OLS Regression for Property Value Evaluation

$$PropertyValue_i = \beta_0 + \beta_1 MedianIncome_i + \beta_2 Unemployment_i + \beta_3 SquareFootage_i$$

$$+\beta_4 Age_i + \beta_5 Bedrooms_i + \beta_6 Bathrooms_i + u_i$$

Where we use median income, unemployment rate, median house square footage, median house age, median number of bedrooms and median number of bathrooms in census tract i to predict the median property value in census tract i

```
LA_pred.ols <- lm(TotalValue ~ median_income + unemployment + SQFTmain + age + Bedrooms + Bathroom
s, data=parcel, na.action=na.exclude)
summary(LA_pred.ols)
```

```
##
## Call:
## lm(formula = TotalValue ~ median_income + unemployment + SQFTmain +
##     age + Bedrooms + Bathrooms, data = parcel, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3933467   -77969   -28171    46882  7574016
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   384868.055  34790.672  11.062  < 2e-16 ***
## median_income    -22.897      9.620  -2.380   0.0174 *
## unemployment     111.106    108.032   1.028   0.3038
## SQFTmain         156.959      2.697  58.196  < 2e-16 ***
## age            -2915.140    378.093  -7.710 1.90e-14 ***
## Bedrooms      -43025.456   9355.921  -4.599 4.49e-06 ***
## Bathrooms      23104.075  10808.143   2.138   0.0327 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279100 on 2173 degrees of freedom
## Multiple R-squared:  0.6333, Adjusted R-squared:  0.6323
## F-statistic: 625.4 on 6 and 2173 DF,  p-value: < 2.2e-16
```

In the above results, housing characteristics like house square footage, house age, number of bedrooms and number of bathrooms are very effective (statistically significant) in predicting property values. However, the high magnitudes and different signs for the coefficients of #bathrooms and #bedrooms are abnormal. This might be attributed to multicolinearity with house square footage. Therefore, we simplify the model into

$$PropertyValue_i = \beta_0 + \beta_1 MedianIncome_i + \beta_2 Unemployment_i + \beta_3 SquareFootage_i + \beta_4 Age_i + u_i$$

```
LA_pred.ols <- lm(TotalValue ~ median_income + unemployment + SQFTmain + age, data=parcel, na.acti
on=na.exclude)
summary(LA_pred.ols)
```

```
##
## Call:
## lm(formula = TotalValue ~ median_income + unemployment + SQFTmain +
##     age, data = parcel, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3882508   -78923   -32213    46169  7625693
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   330710.699  30220.184  10.943   <2e-16 ***
## median_income    -24.147      9.593  -2.517   0.0119 *
## unemployment     110.589    108.567   1.019   0.3085
## SQFTmain         157.640      2.643  59.635   <2e-16 ***
## age            -3276.875    355.412  -9.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 280800 on 2175 degrees of freedom
## Multiple R-squared:  0.6287, Adjusted R-squared:  0.628
## F-statistic: 920.7 on 4 and 2175 DF,  p-value: < 2.2e-16
```

Where we can observe that house characteristics are still effective in predicting property values that properties with less year of usage and larger space have higher property values. In contrast, we also see the negative effect in median income which might against our intuition. In order to correct the violation of OLS assumption due to the existance of spatial auto-correlation, spatial regression analysis is done in the following sector.

Take Queen Weight for neighbored geometries

```
LA_W.queen <- poly2nb(parcel, queen=TRUE)
LA_W <- nb2listw(LA_W.queen , style="W", zero.policy=TRUE)
```

Spatial Regression (Adjusting OLS using the weight derived from geographical adjacencies)

```
LA_W.sar <- lagsarlm(TotalValue ~ median_income + unemployment + SQFTmain + age, data=parcel,
                     LA_W, zero.policy=TRUE)
summary(LA_W.sar)
```

```
##
## Call:lagsarlm(formula = TotalValue ~ median_income + unemployment +
##     SQFTmain + age, data = parcel, listw = LA_W, zero.policy = TRUE)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3860187.1   -51055.8    -8215.8    39365.4  7766350.1
##
## Type: lag
## Coefficients: (numerical Hessian approximate standard errors)
##                  Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)   171902.0823  27261.6617  6.3056 2.870e-10
## median_income    -11.8236      7.5957 -1.5566    0.1196
## unemployment      11.4402          NA      NA        NA
## SQFTmain         152.9393      2.5155 60.7979 < 2.2e-16
## age            -2707.2726    334.7542 -8.0873 6.661e-16
##
## Rho: 0.32118, LR test value: 200.14, p-value: < 2.22e-16
## Approximate (numerical Hessian) standard error: 0.021488
##     z-value: 14.947, p-value: < 2.22e-16
## Wald statistic: 223.42, p-value: < 2.22e-16
##
## Log likelihood: -30339.38 for lag model
## ML residual variance (sigma squared): 7.0354e+10, (sigma: 265240)
## Number of observations: 2180
## Number of parameters estimated: 7
## AIC: 60693, (AIC for lm: 60891)
```

After the correction using the weight of spatial auto-correlation, coefficient of median income become insignificant and the magnitudes of coefficients of house characteristics slightly decrease. To conclude, the property values are basically driven by house properties, and the effects from demographic attributes are more ambiguous.

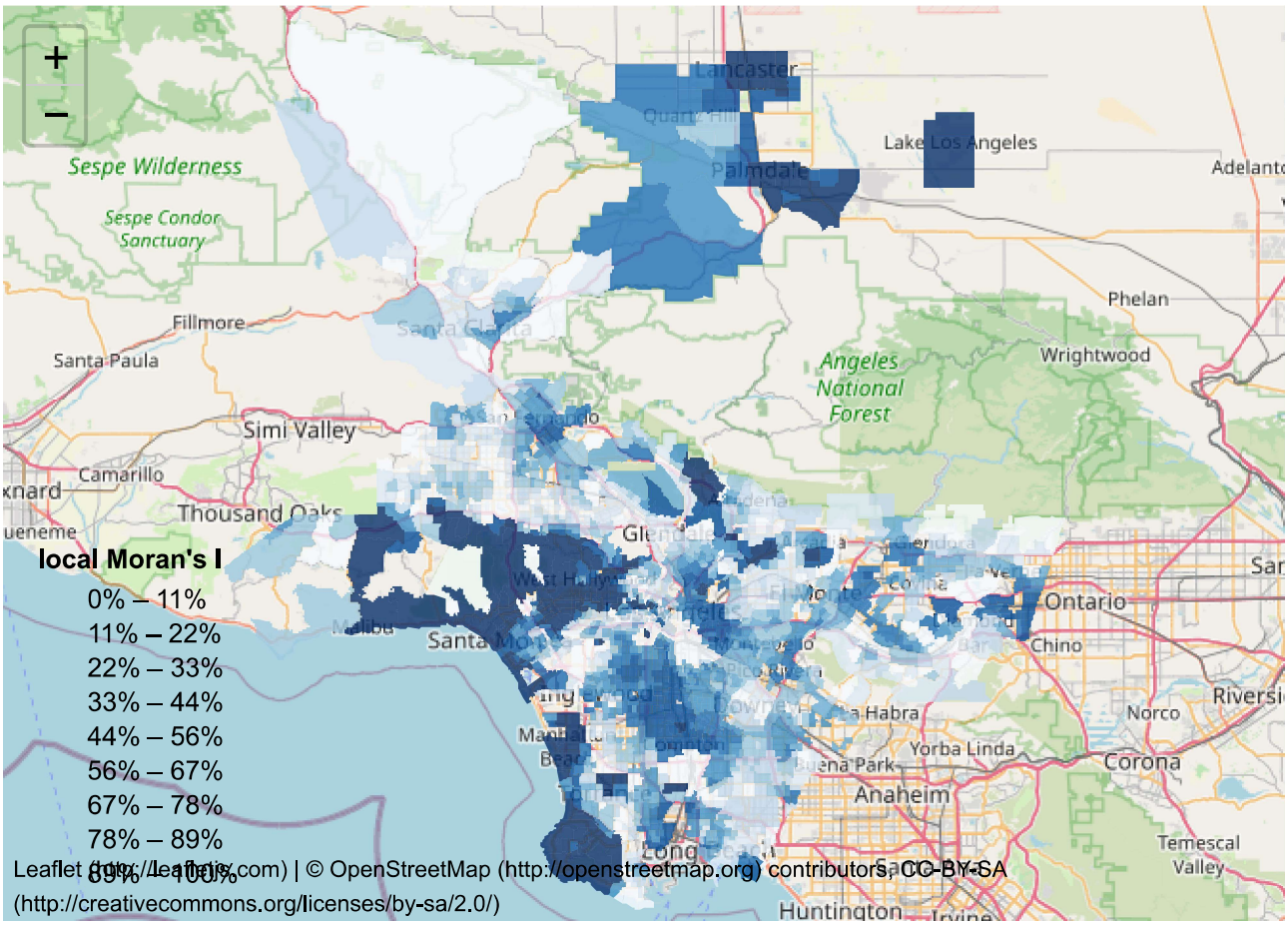
Global Moran's I

```
lm.morantest(LA_pred.ols, LA_W, alternative="two.sided", zero.policy=TRUE)
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = TotalValue ~ median_income + unemployment +
## SQFTmain + age, data = parcel, na.action = na.exclude)
## weights: LA_W
##
## Moran I statistic standard deviate = 21.453, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## Observed Moran I      Expectation         Variance
##     0.2743221730    -0.0009394768     0.0001646315
```

Local Mora's I

```
localm <- localmoran(parcel$TotalValue, LA_W)
temp_mi <- data.frame(id=parcel$CT, MI=data.frame(localm)$Ii)
plot_mi <- merge(parcel, temp_mi, by.x="CT", by.y="id")

local_mi <- colorQuantile("Blues", plot_mi$MI, n=9)
leaflet(plot_mi) %>%
  addPolygons(stroke = FALSE, fillOpacity = .8, smoothFactor = 0.2, color = ~ local_mi(MI)
  ) %>%
  addTiles() %>%
  addLegend("bottomleft", pal=local_mi, values=~MI,
            title = "local Moran's I",
            opacity = 1
  )
```



The above choropleth shows the distribution of clustering effect for the property values. Regions with darker colors have higher property value correlations with their neighbored regions.

Reference

PPHA38520 Multilevel GIS Applications in the Social Sciences Course Materials from Ned English, The University of Chicago

American Community Survey (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml))

Assessor Parcels Data from Los Angeles County Open Data Portal (https://data.lacounty.gov/Parcel-/Assessor-Parcels-Data-2006-thru-2019/9trm-uz8i (https://data.lacounty.gov/Parcel-/Assessor-Parcels-Data-2006-thru-2019/9trm-uz8i))