Geographically Weighted Regression - GWR

Joseph O'Malley

11 June, 2019

Contents

Using weighted regression to analyze health spending in the state of California.

```
#The dataset of employee satisfaction
g1 <- read.csv(filepath,sep=",",header = T)

## check columns
names(g1)
```

```
##  [1] "ZIPCODE"        "LAT"            "LNG"            "State"
##  [5] "Abbreviation"   "ANL_Rain"       "ANL_ave_temp"   "HURRICANE"
##  [9] "WEAT_Risk"      "HSHD_Exp_Appa"  "HSHD_Exp_Contr" "HSHD_Exp_Edua"
## [13] "HSHD_Exp_Ente"  "HSHD_Exp_Food"  "HSHD_Exp_Gift"  "HSHD_Exp_Heal"
## [17] "HSHD_Exp_Furn"  "HSHD_Exp_Hhop"  "HSHD_Exp_Misc"  "HSHD_Exp_Pers"
## [21] "HSHD_Exp_Read"  "HSHD_Exp_Toba"  "HSHD_Exp_Tran"  "HSHD_Exp_Util"
## [25] "AHIncome"       "EDU_Bach"
```

```
## dataset dimensions
dim(g1)
```

```
## [1] 29230    26
```

```
## subset data to focus on California
g2 <- g1[g1$Abbreviation %in% c("CA"), ]
dim(g2)
```

```
## [1] 1610   26
```

OLS Regression

```
library(spgwr)

# use OLS (regression) using annual income, tobacco spending, and bachelors degree to predict healthcar
global.lm <- lm(HSHD_Exp_Heal~AHIncome+HSHD_Exp_Toba+EDU_Bach+AHIncome, data=g2)
summary(global.lm)
```

```
##
## Call:
## lm(formula = HSHD_Exp_Heal ~ AHIncome + HSHD_Exp_Toba + EDU_Bach +
##     AHIncome, data = g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -214.22  -46.93   -5.74   38.93  910.11
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.007e+02  9.080e+00  11.091  < 2e-16 ***
## AHIncome      -3.777e-01  2.097e-01  -1.801   0.0719 .
## HSHD_Exp_Toba  6.950e+00  5.000e-02 138.989  < 2e-16 ***
## EDU_Bach       5.615e-03  7.648e-04   7.341 3.35e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.69 on 1606 degrees of freedom
## Multiple R-squared:  0.9952, Adjusted R-squared:  0.9952
## F-statistic: 1.116e+05 on 3 and 1606 DF,  p-value: < 2.2e-16
```

When trying to predict healthcare spending, tobacco spending had the highest impact (with a coefficient of 6.95). Having a bachelors degree also met the threshold for statistical significance, but had a very low positive weight. Income was slightly below the threshold and had a negetive weight.

find optimal bandwidth

```r
#find optimal bandwidth (indicates how the neighboring areas' characteristics will be counted in)
optimal.band<-gwr.sel(HSHD_Exp_Heal~HSHD_Exp_Toba, data=g2,
                      coords=cbind(g2$LNG, g2$LAT), longlat=TRUE, adapt=T)
optimal.band
```

Geographically Weighted Regression

```r
#run GWR
gwr.model <- gwr(HSHD_Exp_Heal~HSHD_Exp_Toba, data=g2, coords=cbind(g2$LNG, g2$LAT),
                 gweight=gwr.Gauss,
                 longlat=TRUE, adapt=optimal.band, hatmatrix=TRUE, se.fit=TRUE)

#We keep the local coefficients and standard error, and merge the Zip Codes
m<-as.data.frame(cbind((gwr.model$SDF)),g2$ZIPCODE)

# calculate the local t value by using coefficient/standard error
m$significance_toba<-m$HSHD_Exp_Toba/m$HSHD_Exp_Toba_se

# replace insignificant coefficient with 0
m$tobacco_coeff <- ifelse(abs(m$significance_toba) > 1.96, m$significance_toba, 0)

# check for outliers
```

```
q_toba<-quantile(m$tobacco_coeff, c(.05, .5, .95))
q_toba
```

```
##        5%       50%       95% 
##  20.58391  55.77901 143.43968
```

```
#We then set any number higher than the 95% "outlier" threshold to be the 95% value.
m$tobacco_coeff_smooth<- m$tobacco_coeff
#m$tobacco_coeff_smooth[m$tobacco_coeff_smooth > q[3]] <- q[3]

# check the quantile after smoothing the variable
quantile(m$tobacco_coeff_smooth, c(.05, .5, .95))
```

```
##        5%       50%       95% 
##  20.58391  55.77901 143.43968
```

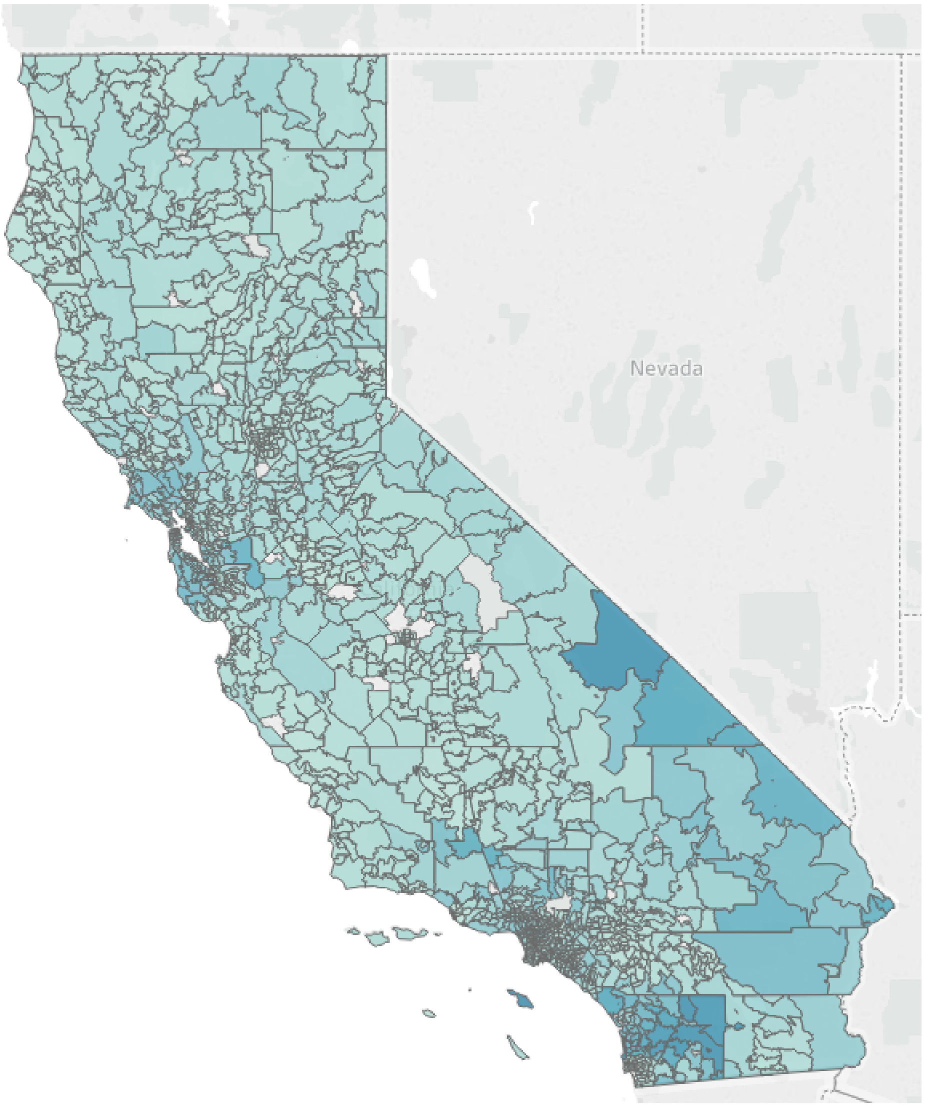
```
# create rank variable
m$tobacco_coeff_rank<-rank(m$tobacco_coeff)

# output to csv for Tableau viz
# write.csv(m, "filepath/gwr_results.csv")
```

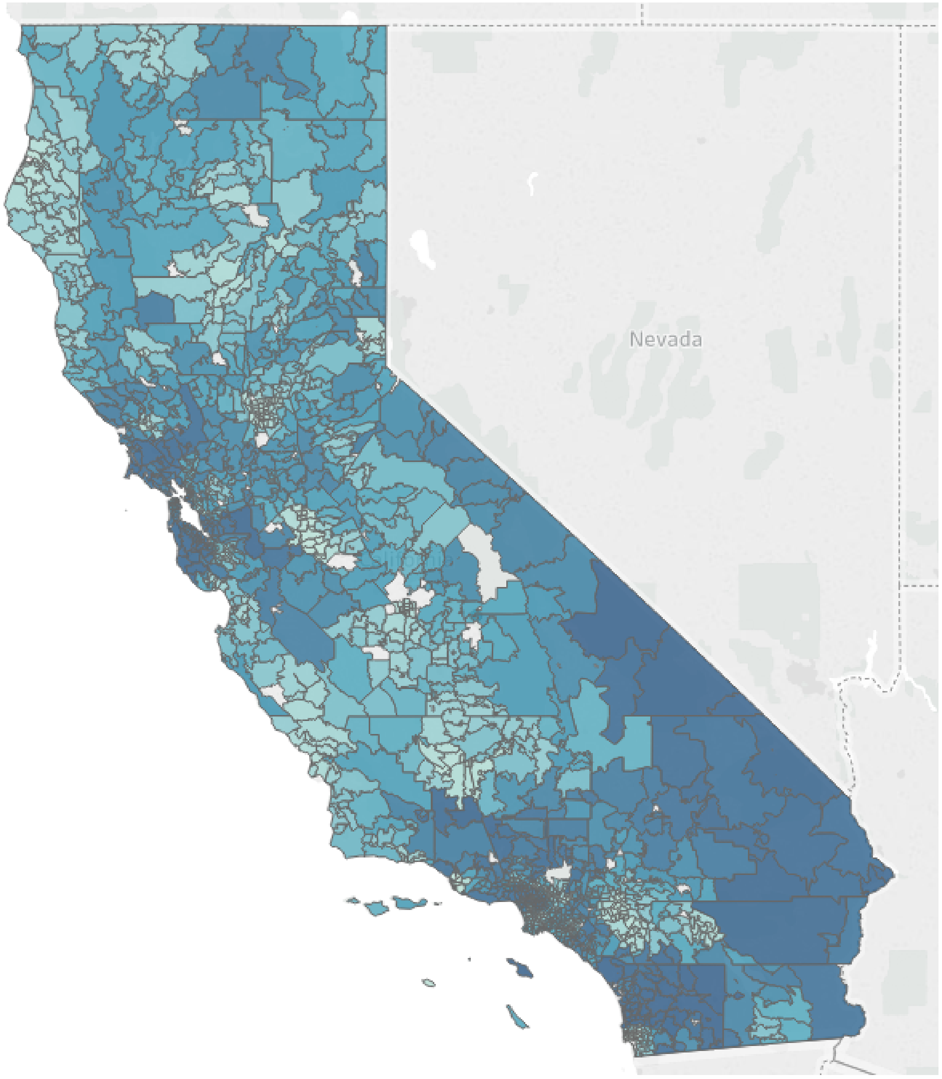
Results

My analysis focused on the state of California, as it is a very large state with several distinct regions. I wanted to know more about what leads to higher expenditures on healthcare (by household).
I started this process by using OLS regression to look at several features including: income, tobacco spending, and education. Income was slightly below the threshold for significance, while Bachelors degree met this threshold but had a low magnitude impact. Tobacco spending, however, had a large positive relationship with healthcare spending with a coefficient of 6.95. This seems logical that people probably smoke/chew the tobacco they purchase and this leads to conditions that need additional medical attention and spend. This is fairly straight forward, but may help inform which specific zip codes to direct attention to (whether that be CMS, insurance, hospital systems, etc). That being said, I moved forward with analysis using GWR on tobacco spending.

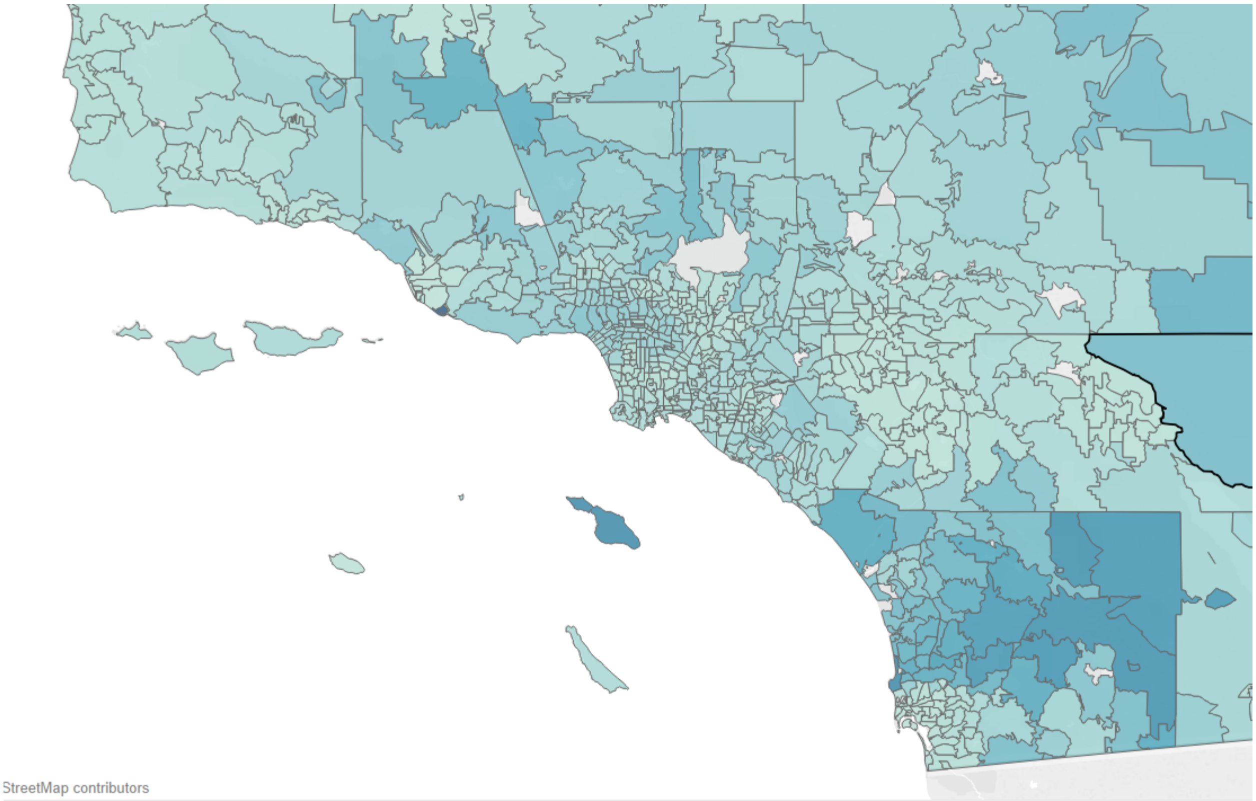
map 1 - California Tobacco Spending by zip (coefficients)



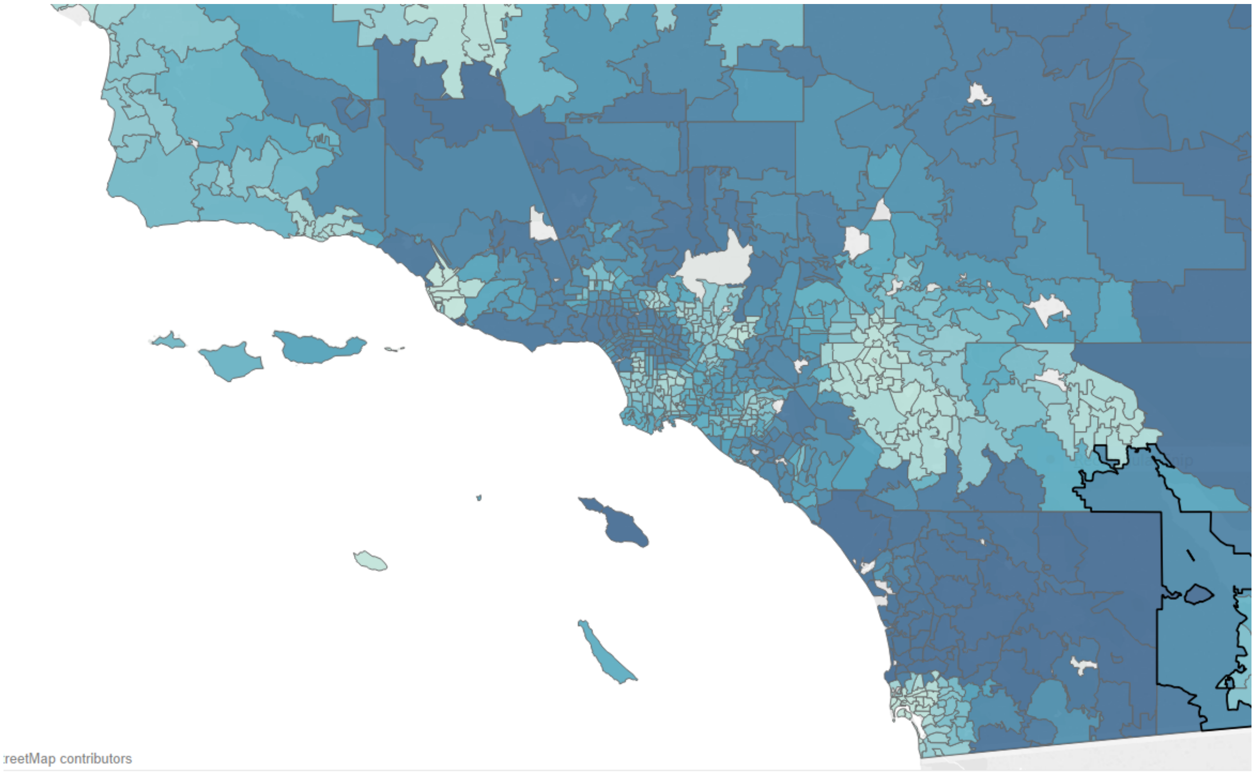
map 2 - California Tobacco Spending by zip (Ranked)



map 3 - Southern California Tobacco Spending by zip (coefficients)



map 4 - Southern California Tobacco Spending by zip (Ranked)



I created 4 separate maps using the output of my GWR in Tableau. The first two (one for coefficients, one ranked) are the entire state of California where you can see the differences in Tobacco spending with the darker counties being higher spending. As you can see the rural areas appear to be darker in general (map 1), but as you look at the ranked version this relationship becomes more obvious (map 2). I then wanted to take a closer look at the very compact zip codes in the southern part of the state (maps 3 & 4). Once again you can see the cities generally have lower rates of smoking aside from certain pockets of neighborhoods in LA. The ranked map (map 4) also shows the drastic difference in the zip codes in and around San Diego.

Conclusion

Overall, there are so many use cases for this data. The most obvious are CMS ad campaigns against smoking and its effects in select zip codes in an effort to reduce the development of chronic conditions caused by smoking that greatly increase spend on healthcare.