vignette-geospatial

Liuqian Bao, Xiaofeng Cai, Jiajia Feng, Sophie Shi, Jiahui He

2024-12-12

Spatial Analysis of Travel Distances Across Different Modes Using Geographically Weighted Regression (GWR)

Goal

The goal is to understand how travel behavior (e.g., distances traveled for different modes of transportation) varies across different counties through a spatial regression analysis.

Data Sources

This vignette uses a database of 26,095 sample households residing in California, containing detailed information on their travel behavior on one assigned day for each household, from April 19, 2016 through April 25, 2017, provided by the Transportation Secure Data Center (TSDC). [1] Specifically, we obtained travel mode data from `PersonData.Rds`, `HHData.Rds`and geographical data, i.e. the physical location and shape of each county, from`counties.shp`, files given in the dataset.

[1]"Transportation Secure Data Center." (2019). National Renewable Energy Laboratory. Accessed Jan. 15, 2019: www.nrel.gov/tsdc.

Methodology

Geographically Weighted Regression (GWR):

Geographically Weighted Regression (GWR) is a spatial analysis technique that extends traditional regression by allowing the relationships between dependent and independent variables to vary spatially. Unlike ordinary least squares (OLS) regression, which assumes global stationarity of the coefficients, GWR incorporates geographic context into the model. This approach accounts for spatial heterogeneity, a common characteristic in spatial datasets, where relationships can change over space due to localized factors.

In GWR, the regression is performed repeatedly for each location in the dataset, weighting observations according to their spatial proximity to the focal location. The weighting is determined using a kernel function, which can be fixed or adaptive, depending on the data's spatial distribution.[2]

Bandwidth Selection: The selection of an appropriate bandwidth is a crucial step for the GWR model. Bandwidth is a parameter that governs the spatial extent, over which neighboring observations influence the estimation of local parameters. The bandwidth serves as a key filter determining the degree of localization in the analysis.

A bandwidth that is too narrow may lead to oversensitivity to local variations, potentially capturing noise in the data. On the other hand, too broad bandwidths can result in over smoothed representations , masking subtle spatial patterns. With a proper bandwidth value, we are able to achieve the balance to ensure the GWR model accurately captures the true spatial heterogeneity without being unduly influenced by distant observations.

Adaptive bandwidths offer an effective solution, as they can vary based on the size of each geographical area and that of its neighbors. Thus, the model can select a narrower bandwidth in dense areas, and a larger one for suburban areas. [3]

[2] Charlton, M., & Fotheringham, A. S. (2009). Geographically weighted regression. [White Paper].

[3] Kiani et al.(2024, February 29). *Mastering geographically weighted regression: Key considerations for building a robust model: Geospatial Health.* Mastering geographically weighted regression: key considerations for building a robust model | Geospatial Health. https://www.geospatialhealth.net/gh/article/view/1271/1365

Data Pre-Processing

Load data

```r
library(sf)
library(sp)
library(dplyr)
library(GWmodel)
library(ggplot2)
library(mapview)
library(leafpop)

person_data <- readRDS("./data/PersonData.Rds")
hh_data <- readRDS("./data/HHData.Rds")
bg_density <- readRDS("./data/hh_bgDensity.Rds")
shapefile <- st_read("./data/counties/counties.shp")
boundaries <- st_read("./data/ca_state_boundaries/CA_State.shp")
```

```r
head(person_data)
```

Person dataset includes basic demographics, employment/student status, and travel behavior variables.

```
## # A tibble: 6 x 50
##      hhid  pnum  Male   Age persHisp persWhite persAfricanAm persNativeAm
##     <dbl> <dbl> <dbl> <dbl>    <dbl>     <dbl>         <dbl>        <dbl>
## 1 1031985     1     1    74        0         1             0            0
## 2 1031985     2     0    73        0         1             0            0
## 3 1032036     1     1    46        0         1             0            0
## 4 1032036     2     0    47        0         1             0            0
## 5 1032036     3     1    15        0         1             0            0
## 6 1032036     4     1    14        0         1             0            0
## # i 42 more variables: persAsian <dbl>, persPacIsl <dbl>, persOthr <dbl>,
## #   persDKrace <dbl>, persRFrace <dbl>, bornUSA <dbl>, DriverLic <dbl>,
## #   TransitPass <dbl>, Employed <dbl>, WorkFixedLoc <dbl>, WorkHome <dbl>,
## #   WorkNonfixed <dbl>, WorkDaysWk <dbl>, TypicalHoursWk <dbl>,
## #   FlexSched <dbl>, FlexPrograms <dbl>, Disability <dbl>, DisLicensePlt <dbl>,
## #   TransitTripsWk <dbl>, WalkTripsWk <dbl>, BikeTripsWk <dbl>, Student <dbl>,
## #   WorkMode <chr>, SchoolMode <chr>, EducationCompl <chr>, workday <dbl>, ...
```

```r
head(hh_data)
```

Household dataset includes household-level demographics, survey date, and home county.

```
## # A tibble: 6 x 19
##      hhid CTFIP County        MPO    City  DOW   HH_size HH_nTrips HH_nEmployees
##     <dbl> <dbl> <chr>         <chr>  <chr> <chr>   <dbl>     <dbl>         <dbl>
## 1 1031985  6095 Solano        MTC    Vall~ Tues~       2         4             0
## 2 1032036  6073 San Diego     SANDAG San ~ Satu~       5        31             1
## 3 1032053  6047 Merced        Merced Merc~ Thur~       6        46             1
## 4 1032425  6083 Santa Barbara Santa~ Gole~ Mond~       2         0             2
## 5 1032558  6037 Los Angeles   SCAG   Los ~ Frid~       1         6             0
## 6 1033586  6061 Placer        SACOG  Linc~ Frid~       3        10             1
## # i 10 more variables: HH_nStudents <dbl>, HH_nLicenses <dbl>, HH_nCars <dbl>,
## #   HH_nBikes <dbl>, HH_income <dbl>, HH_anyTransitRider <dbl>,
## #   HH_homeType <dbl>, HH_homeowner <dbl>, HH_isHispanic <dbl>,
## #   HH_intEnglish <dbl>
```

```r
head(bg_density)
```

Block group density dataset contains how urban are the areas around CHTS respondents' homes.

```
## # A tibble: 6 x 3
##      hhid bg_density bg_group
##     <int>      <dbl> <fct>
## 1 1449245       818. Suburban
## 2 3007304       818. Suburban
## 3 3008384       818. Suburban
## 4 3007273       818. Suburban
## 5 1452535       818. Suburban
## 6 3007437      2843. Urban
```

```r
head(shapefile)
```

Counties file contains the county names and their corresponding latitudes and longitudes.

```
## Simple feature collection with 6 features and 17 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -123.5363 ymin: 34.89747 xmax: -117.9808 ymax: 38.85292
## Geodetic CRS:  WGS 84
##   STATEFP COUNTYFP COUNTYNS CTFIP            NAME               NAMELSAD LSAD
## 1      06      107 00277318  6107          Tulare          Tulare County   06
## 2      06      009 01675885  6009       Calaveras       Calaveras County   06
## 3      06      047 00277288  6047          Merced          Merced County   06
## 4      06      079 00277304  6079 San Luis Obispo San Luis Obispo County   06
## 5      06      097 01657246  6097          Sonoma          Sonoma County   06
## 6      06      041 00277285  6041           Marin           Marin County   06
##   CLASSFP MTFCC CSAFP CBSAFP METDIVFP FUNCSTAT       ALAND    AWATER
## 1      H1 G4020  <NA>  47300     <NA>        A 12494707314  37391604
## 2      H1 G4020  <NA>   <NA>     <NA>        A  2641820029  43810423
## 3      H1 G4020  <NA>  32900     <NA>        A  5011554680 112760479
## 4      H1 G4020  <NA>  42020     <NA>        A  8543230300 820974619
## 5      H1 G4020   488  42220     <NA>        A  4081430061 497530414
## 6      H1 G4020   488  41860    41884        A  1347585499 797420416
##      INTPTLAT     INTPTLON                       geometry
```

```
## 1 +36.2288317 -118.7810618 MULTIPOLYGON (((-118.3606 3...
## 2 +38.1846184 -120.5593996 MULTIPOLYGON (((-120.02 38....
## 3 +37.1948063 -120.7228019 MULTIPOLYGON (((-120.0521 3...
## 4 +35.3852268 -120.4475409 MULTIPOLYGON (((-120.214 35...
## 5 +38.5250258 -122.9376050 MULTIPOLYGON (((-122.513 38...
## 6 +38.0518169 -122.7459738 MULTIPOLYGON (((-123.0233 3...
```

Merge Data

Since we are using data from two separate Rds files `PersonData.Rds` and `HHData.Rds`, we combine these two datasets by using a left_join function on `hhid` which is a unique identifier for each household.

```
# Merge Data
combine_data <- left_join(person_data, hh_data) %>% left_join(bg_density)
head(combine_data)
```

```
## # A tibble: 6 x 70
##      hhid  pnum  Male   Age persHisp persWhite persAfricanAm persNativeAm
##     <dbl> <dbl> <dbl> <dbl>    <dbl>     <dbl>         <dbl>        <dbl>
## 1 1031985     1     1    74        0         1             0            0
## 2 1031985     2     0    73        0         1             0            0
## 3 1032036     1     1    46        0         1             0            0
## 4 1032036     2     0    47        0         1             0            0
## 5 1032036     3     1    15        0         1             0            0
## 6 1032036     4     1    14        0         1             0            0
## # i 62 more variables: persAsian <dbl>, persPacIsl <dbl>, persOthr <dbl>,
## #   persDKrace <dbl>, persRFrace <dbl>, bornUSA <dbl>, DriverLic <dbl>,
## #   TransitPass <dbl>, Employed <dbl>, WorkFixedLoc <dbl>, WorkHome <dbl>,
## #   WorkNonfixed <dbl>, WorkDaysWk <dbl>, TypicalHoursWk <dbl>,
## #   FlexSched <dbl>, FlexPrograms <dbl>, Disability <dbl>, DisLicensePlt <dbl>,
## #   TransitTripsWk <dbl>, WalkTripsWk <dbl>, BikeTripsWk <dbl>, Student <dbl>,
## #   WorkMode <chr>, SchoolMode <chr>, EducationCompl <chr>, workday <dbl>, ...
```

Group the data by county (CTFIP), and calculates the average total number of miles traveled by a person for each travel mode (Drive Alone, Drive with Others, Passenger, Walk, and Total) on the survey day.

```
summarized_data <- combine_data %>%
  select(hhid, pnum, DriveAlone_Dist, Driveothers_Dist, Passenger_Dist, Walk_Dist, Bike_Dist, CTFIP, Su
  group_by(CTFIP) %>%
  summarise(
    avg_DriveAlone_Dist = mean(DriveAlone_Dist, na.rm = T),
    avg_Driveothers_Dist = mean(Driveothers_Dist, na.rm = T),
    avg_Passenger_Dist = mean(Passenger_Dist, na.rm = T),
    avg_Walk_Dist = mean(Walk_Dist, na.rm = T),
    avg_Bike_Dist = mean(Bike_Dist, na.rm = T),
    avg_Sum_Pmt = mean(Sum_PMT, na.rm = T))
```

Visualizing County Data on a Map

The interactive plot below shows the average number of miles the person traveled on survey day, with yellowish color representing high values and purplish color representing low values by counties in California. By pointing at each of the country, we are able to observe the corresponding values for `avg_DriveAlone_Dist`,`avg_Driveothers_Dist`, `avg_Passenger_Dist`, `avg_Walk_Dist`, `avg_Bike_Dist`, with their representations shown below.

`avg_DriveAlone_Dist`: average number of miles the person drove alone on survey day `avg_Driveothers_Dist`: average number of miles the person drove with others on survey day

`avg_Passenger_Dist`: average number of miles the person rode in a car as a passenger on survey day

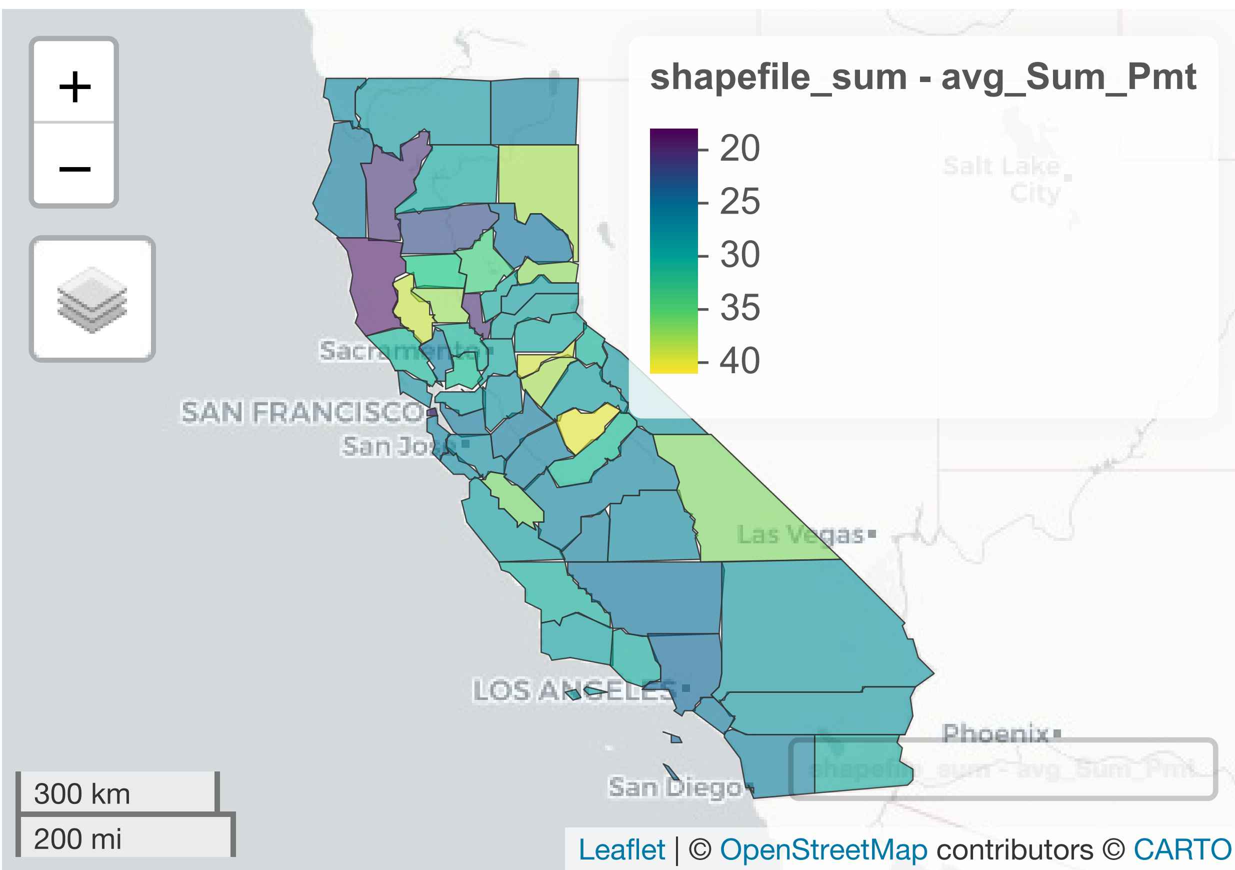
`avg_Walk_Dist`: average number of trips the person made on by walking on survey day

`avg_Bike_Dist`: average number of trips the person made by bike on survey day

```
shapefile_sum <- shapefile %>%
  left_join(summarized_data, by = "CTFIP")

mapviewOptions(fgb = F)

mapview(shapefile_sum,
        zcol = "avg_Sum_Pmt", # assigned color based on sum distance
        legend = TRUE,
        label = as.character(shapefile_sum$CTFIP),
        popup = popupTable(shapefile_sum,
                           zcol = c("avg_DriveAlone_Dist",
                                    "avg_Driveothers_Dist",
                                    "avg_Passenger_Dist",
                                    "avg_Walk_Dist",
                                    "avg_Bike_Dist",
                                    "avg_Sum_Pmt")))
```



Perform Geographically Weighted Regression

Convert shape file into spatial data

Convert the `county.shp` file into spatial data.

```
coords <- st_coordinates(st_centroid(st_geometry(shapefile_sum)))

gwr_data <- shapefile_sum %>%
  select(avg_DriveAlone_Dist, avg_Driveothers_Dist, avg_Passenger_Dist, avg_Walk_Dist, avg_Bike_Dist, a
  st_drop_geometry()

gwr_data <- cbind(gwr_data, coords)



# convert to spatial data
coordinates(gwr_data) <- ~X + Y
proj4string(gwr_data) <- CRS("+proj=longlat +datum=WGS84")
```

Construct Regression formula:

We use the `gwr_formula` function to construct our regression formula. We took `avg_Sum_Pmt` as the response variable which is the average number of miles the person traveled on survey day, and `avg_DriveAlone_Dist`,`avg_Driveothers_Dist`, `avg_Passenger_Dist`, `avg_Walk_Dist`, `avg_Bike_Dist` as predictors.

```
gwr_formula <- avg_Sum_Pmt ~ avg_DriveAlone_Dist + avg_Driveothers_Dist + avg_Passenger_Dist + avg_Walk_
```

Select GWR Bandwidth

Use `bw.gwr()` to find the optimal bandwidth for GWR analysis.

```
# Perform bandwidth selection for GWR
bw <- bw.gwr(
  formula = gwr_formula,
  data = gwr_data,
  adaptive = T)
bw
```

In our analysis, we determined the optimal bandwidth here, which is 35.

Results

Here is a summary of the geographically weighted regression model fit:

```
gwr_model <- gwr.basic(
  formula = gwr_formula,
  data = gwr_data,
  bw = bw,
  adaptive = T)

gwr_result <- gwr_model$SDF
summary(gwr_result)
```

```
## Object of class SpatialPointsDataFrame
## Coordinates:
##          min        max
```

```
## X -123.89432 -115.36552
## Y   33.03547   41.74308
## Is projected: FALSE
## proj4string : [+proj=longlat +datum=WGS84 +no_defs]
## Number of points: 58
## Data attributes:
##    Intercept       avg_DriveAlone_Dist avg_Driveothers_Dist avg_Passenger_Dist
##  Min.   :-0.4934   Min.   :0.9145      Min.   :0.8048       Min.   :0.8516
##  1st Qu.: 0.4961   1st Qu.:0.9518      1st Qu.:0.9750       1st Qu.:0.8755
##  Median : 0.8308   Median :0.9979      Median :1.0536       Median :0.9178
##  Mean   : 0.9485   Mean   :1.0118      Mean   :1.0237       Mean   :0.9157
##  3rd Qu.: 1.4073   3rd Qu.:1.0629      3rd Qu.:1.0721       3rd Qu.:0.9444
##  Max.   : 2.5491   Max.   :1.1466      Max.   :1.1642       Max.   :1.0113
##  avg_Walk_Dist     avg_Bike_Dist         y              yhat
##  Min.   :-0.9423   Min.   :1.616   Min.   :18.13   Min.   :19.12
##  1st Qu.: 1.6040   1st Qu.:2.339   1st Qu.:25.74   1st Qu.:25.73
##  Median : 4.0746   Median :2.797   Median :28.15   Median :28.10
##  Mean   : 3.5652   Mean   :2.692   Mean   :28.95   Mean   :28.98
##  3rd Qu.: 5.6299   3rd Qu.:3.087   3rd Qu.:31.11   3rd Qu.:31.31
##  Max.   : 6.9179   Max.   :3.916   Max.   :40.88   Max.   :40.47
##     residual            CV_Score Stud_residual       Intercept_SE
##  Min.   :-1.035504   Min.   :0   Min.   :-2.66169   Min.   :0.6972
##  1st Qu.:-0.321663   1st Qu.:0   1st Qu.:-0.75237   1st Qu.:0.8132
##  Median : 0.005585   Median :0   Median : 0.01299   Median :0.9091
##  Mean   :-0.029618   Mean   :0   Mean   :-0.11165   Mean   :0.9201
##  3rd Qu.: 0.244794   3rd Qu.:0   3rd Qu.: 0.49366   3rd Qu.:1.0281
##  Max.   : 1.141008   Max.   :0   Max.   : 3.06768   Max.   :1.2451
##  avg_DriveAlone_Dist_SE avg_Driveothers_Dist_SE avg_Passenger_Dist_SE
##  Min.   :0.04545        Min.   :0.07797         Min.   :0.07302
##  1st Qu.:0.04917        1st Qu.:0.08661         1st Qu.:0.08168
##  Median :0.05427        Median :0.09965         Median :0.09066
##  Mean   :0.05475        Mean   :0.09926         Mean   :0.09005
##  3rd Qu.:0.05991        3rd Qu.:0.10985         3rd Qu.:0.09689
##  Max.   :0.07354        Max.   :0.12553         Max.   :0.11771
##  avg_Walk_Dist_SE avg_Bike_Dist_SE  Intercept_TV     avg_DriveAlone_Dist_TV
##  Min.   :0.8861   Min.   :1.050    Min.   :-0.6075   Min.   :12.72
##  1st Qu.:1.0182   1st Qu.:1.122    1st Qu.: 0.5059   1st Qu.:17.83
##  Median :1.1827   Median :1.170    Median : 0.9482   Median :19.25
##  Mean   :1.4982   Mean   :1.234    Mean   : 0.9975   Mean   :18.72
##  3rd Qu.:1.5748   3rd Qu.:1.337    3rd Qu.: 1.6638   3rd Qu.:20.28
##  Max.   :3.3180   Max.   :1.605    Max.   : 2.6042   Max.   :21.74
##  avg_Driveothers_Dist_TV avg_Passenger_Dist_TV avg_Walk_Dist_TV
##  Min.   : 6.411          Min.   : 7.903        Min.   :-0.6299
##  1st Qu.: 9.098          1st Qu.: 9.168        1st Qu.: 1.0900
##  Median :10.650          Median :10.414        Median : 2.0566
##  Mean   :10.565          Mean   :10.306        Mean   : 2.9079
##  3rd Qu.:12.364          3rd Qu.:11.350        3rd Qu.: 5.5485
##  Max.   :13.466          Max.   :12.864        Max.   : 6.4995
##  avg_Bike_Dist_TV    Local_R2
##  Min.   :1.107    Min.   :0.9879
##  1st Qu.:1.860    1st Qu.:0.9922
##  Median :2.414    Median :0.9937
##  Mean   :2.236    Mean   :0.9937
##  3rd Qu.:2.723    3rd Qu.:0.9952
```

```
##  Max.   :3.160    Max.   :0.9978
```

```r
head(gwr_result)
```

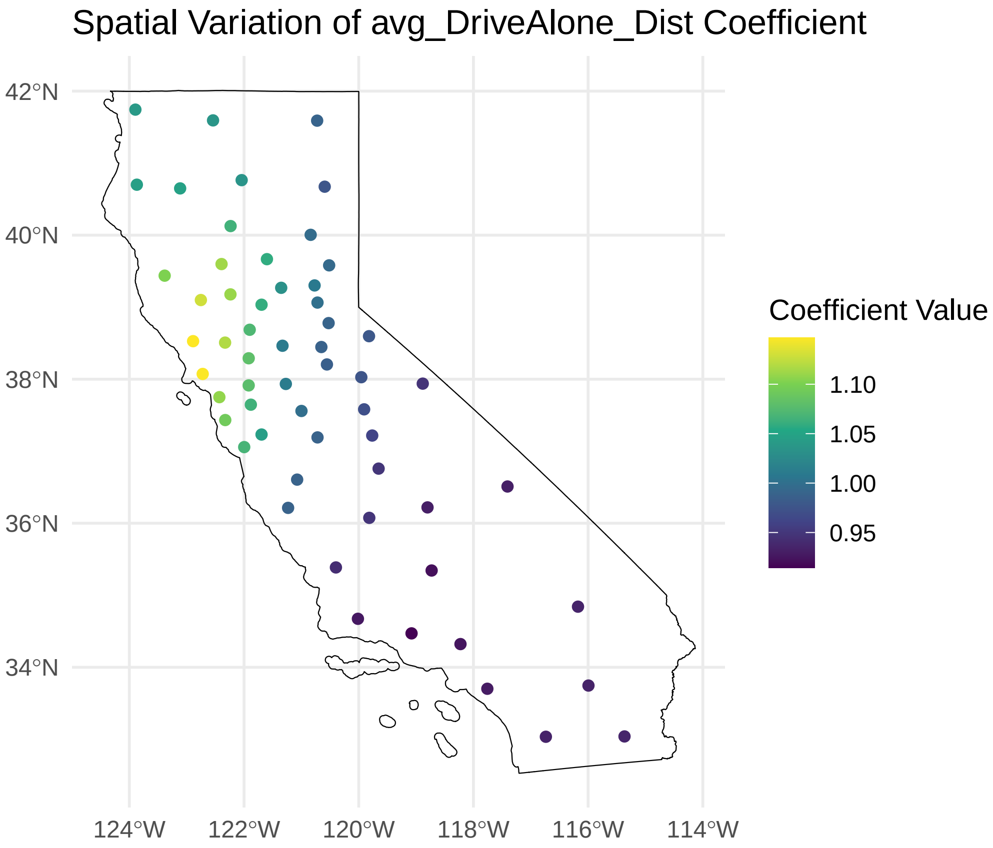
```
## class       : SpatialPointsDataFrame
## features    : 6
## extent      : -122.8884, -118.8004, 35.38639, 38.52794  (xmin, xmax, ymin, ymax)
## crs         : +proj=longlat +datum=WGS84 +no_defs
## variables   : 24
## names       :          Intercept, avg_DriveAlone_Dist, avg_Driveothers_Dist, avg_Passenger_Dist,    a
## min values  : -0.493407547943922,   0.931195113605572,    0.804772664961579,  0.882009855991374, 2.07
## max values  :   2.15676300222343,    1.14663162009081,     1.07245452213885,    1.0113345223337, 6.82
```

```r
gwr_sf <- st_as_sf(gwr_result)
```

Plot the coefficients

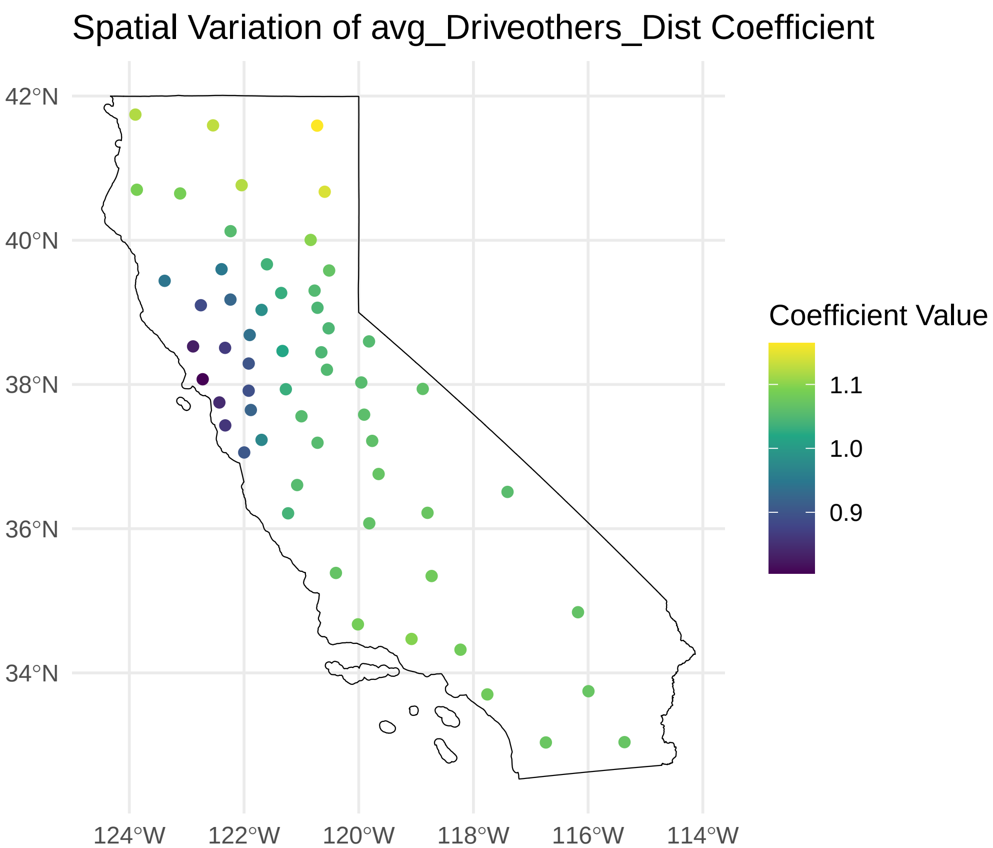
Driving Alone Around 38°N, 123°W which is the Bay Area, the higher coefficients indicate that the average drive-alone distance contributes more significantly to changes in total traveling distance.

```r
ggplot(data = gwr_sf) +
  geom_sf(aes(color = avg_DriveAlone_Dist)) +
  geom_sf(data = boundaries, fill = NA, color = "black", linewidth = 0.2) +
  scale_color_viridis_c() +
  theme_minimal() +
  labs(title = "Spatial Variation of avg_DriveAlone_Dist Coefficient",
       color = "Coefficient Value")
```



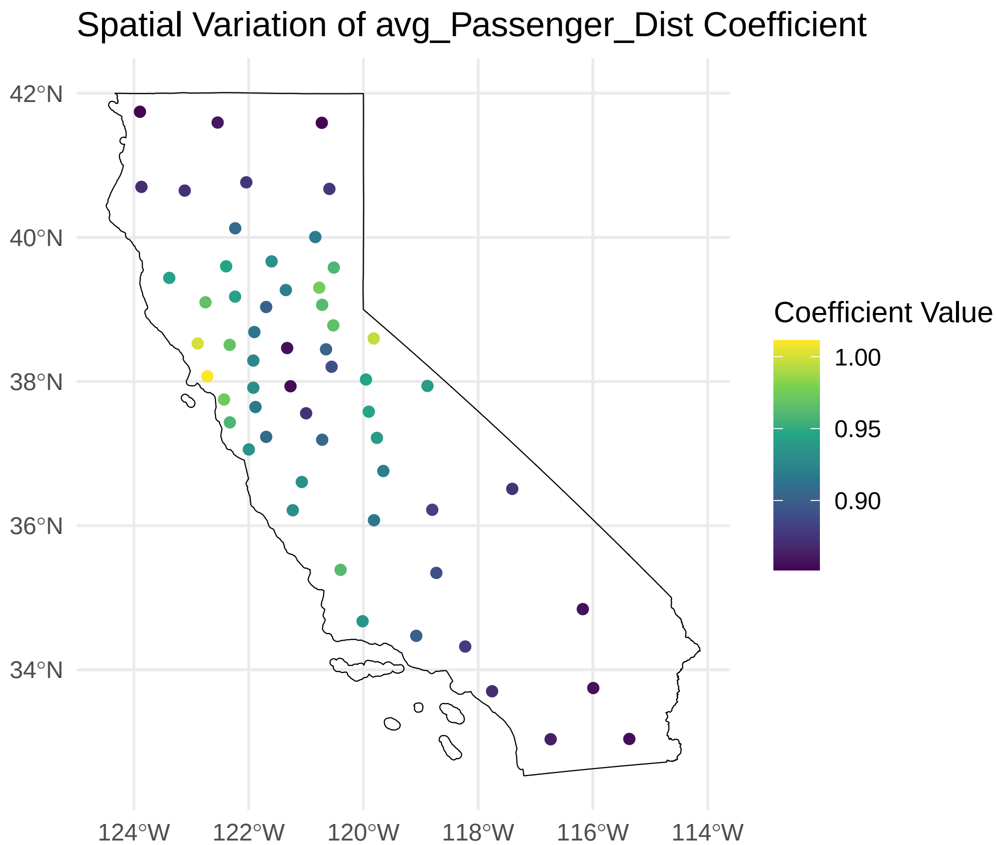
Driving with Others Coefficients are higher in southern and northern regions (green regions), while some coastal areas near 38°N have lower coefficients (dark blue), which suggest a converse relationship to the drive-alone distance.

```
ggplot(data = gwr_sf) +
  geom_sf(aes(color = avg_Driveothers_Dist)) +
  scale_color_viridis_c() +
  geom_sf(data = boundaries, fill = NA, color = "black", linewidth = 0.2) +
  theme_minimal() +
  labs(title = "Spatial Variation of avg_Driveothers_Dist Coefficient",
       color = "Coefficient Value")
```



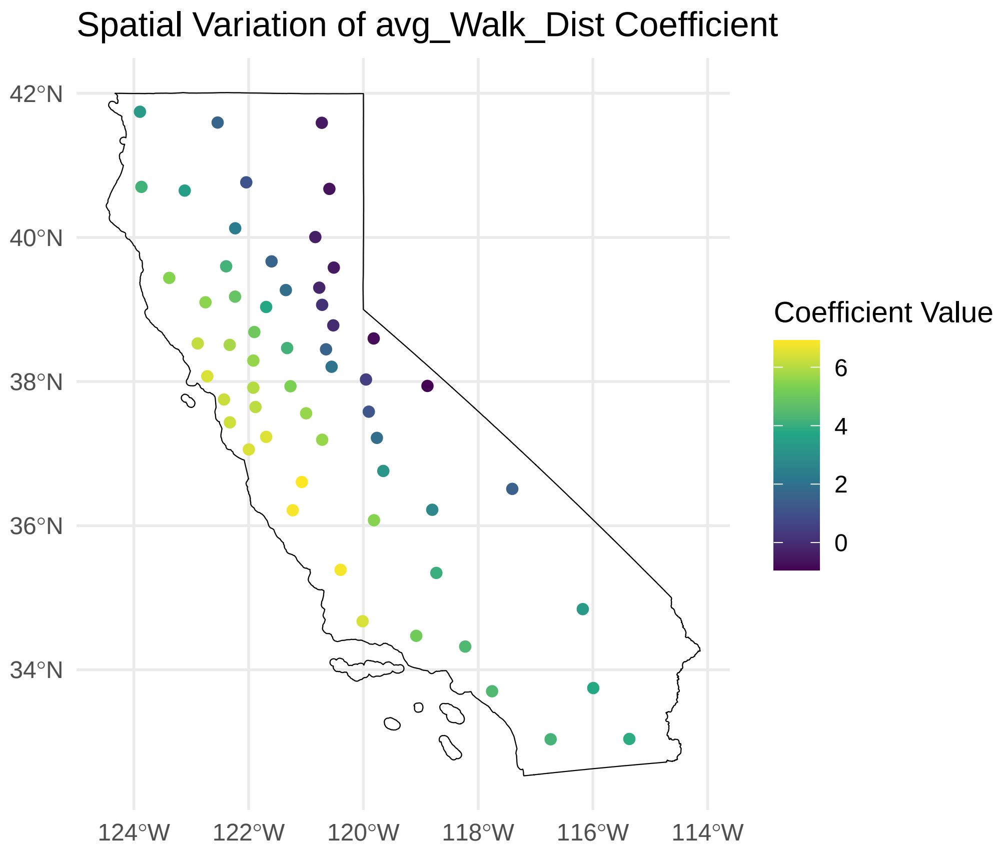
Riding as Passenger The coefficients are generally low (mostly dark blue) across all regions, particularly in southern areas, suggesting that passenger distance is less influential.

```
ggplot(data = gwr_sf) +
  geom_sf(aes(color = avg_Passenger_Dist)) +
  geom_sf(data = boundaries, fill = NA, color = "black", linewidth = 0.2) +
  scale_color_viridis_c() +
  theme_minimal() +
  labs(title = "Spatial Variation of avg_Passenger_Dist Coefficient",
       color = "Coefficient Value")
```



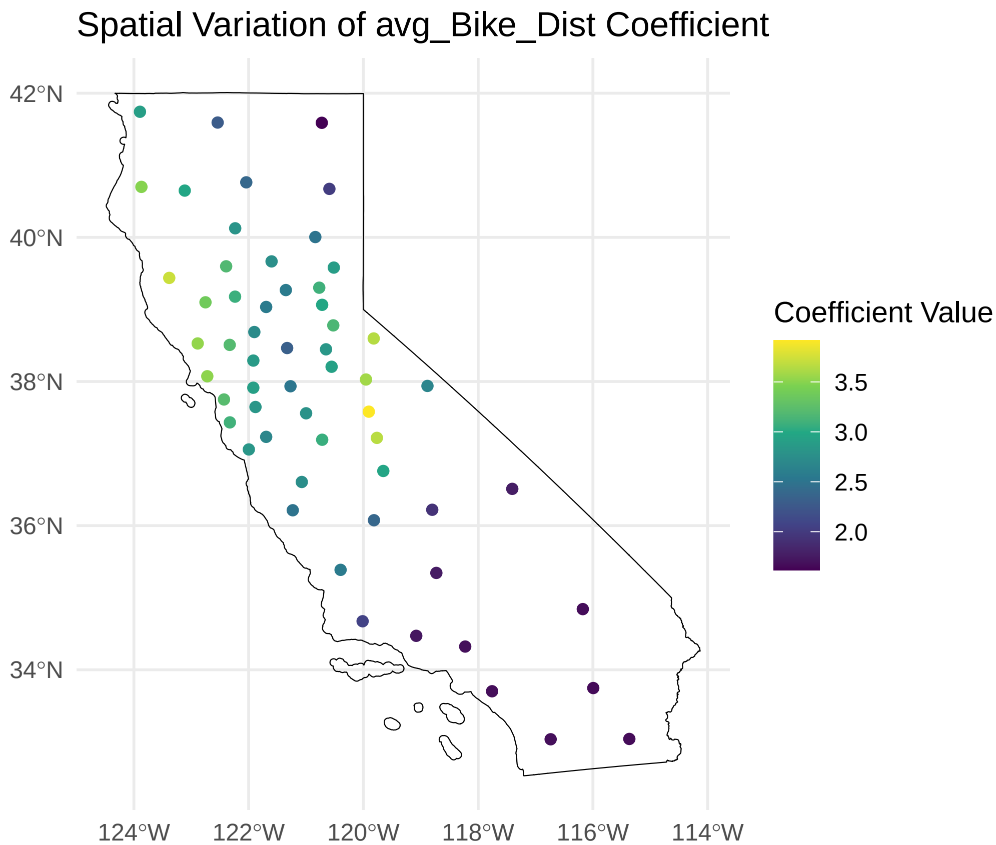
Walking Coastal areas (especially below 38°N) have significantly higher coefficients (yellow to green), suggesting that walking distance contributes more significantly to changes in total traveling distance.

```r
ggplot(data = gwr_sf) +
  geom_sf(aes(color = avg_Walk_Dist)) +
  geom_sf(data = boundaries, fill = NA, color = "black", linewidth = 0.2) +
  scale_color_viridis_c() +
  theme_minimal() +
  labs(title = "Spatial Variation of avg_Walk_Dist Coefficient",
       color = "Coefficient Value")
```



Biking Coefficients are lower in southern regions (dark purple regions), suggesting that bike distance is less influential.

```
ggplot(data = gwr_sf) +
  geom_sf(aes(color = avg_Bike_Dist)) +
  geom_sf(data = boundaries, fill = NA, color = "black", linewidth = 0.2) +
  scale_color_viridis_c() +
  theme_minimal() +
  labs(title = "Spatial Variation of avg_Bike_Dist Coefficient",
       color = "Coefficient Value")
```



Summary **Driving Alone**: Higher influence in the northern regions.

Driving with Others: Higher influence in central and southern regions.

Walking: Stronger impacts in coastal regions.

Cycling: Stronger impacts in northern and central regions.

Passenger: Riding as a passenger is relatively consistent across regions.

Future Study

This is a very simple example of using the geographically weighted regression, where we only use the model to investigate how each travelling mode weight in the total travelling distance across the state of California. Future research could perform more analysis and/or apply to other forms of data:

1. **Incorporating Additional Variables:** Include expainatory variables such as income, population density, and land-use characteristics to study how they relate to travel models and how the relationships vary geographically.
2. **Enhancing GWR Adjustments:**
   - **Bandwidth Optimization:** Experiment with fixed and adaptive bandwidths to determine the best spatial scale for analyzing the data, potentially improving model accuracy.
   - **Multiscale GWR (MGWR)**: Explore multiscale GWR to account for variables that may operate at different spatial scales, providing more nuanced insights into local and regional variations.
   - **Kernel Function Selection**: Investigate the impact of different kernel functions (e.g., Gaussian, bisquare) on the model's results to ensure the most appropriate spatial weighting is applied.