The distribution of (auto) insurance companies, by county, in Georgia, USA relative to population size and number of reported car accidents.

Team C3:

Corey Devin Anderson

Erin Zheng

William Won Jung

Chuck Youngman

Study Questions:

Q1) How are insurance companies distributed in the state?

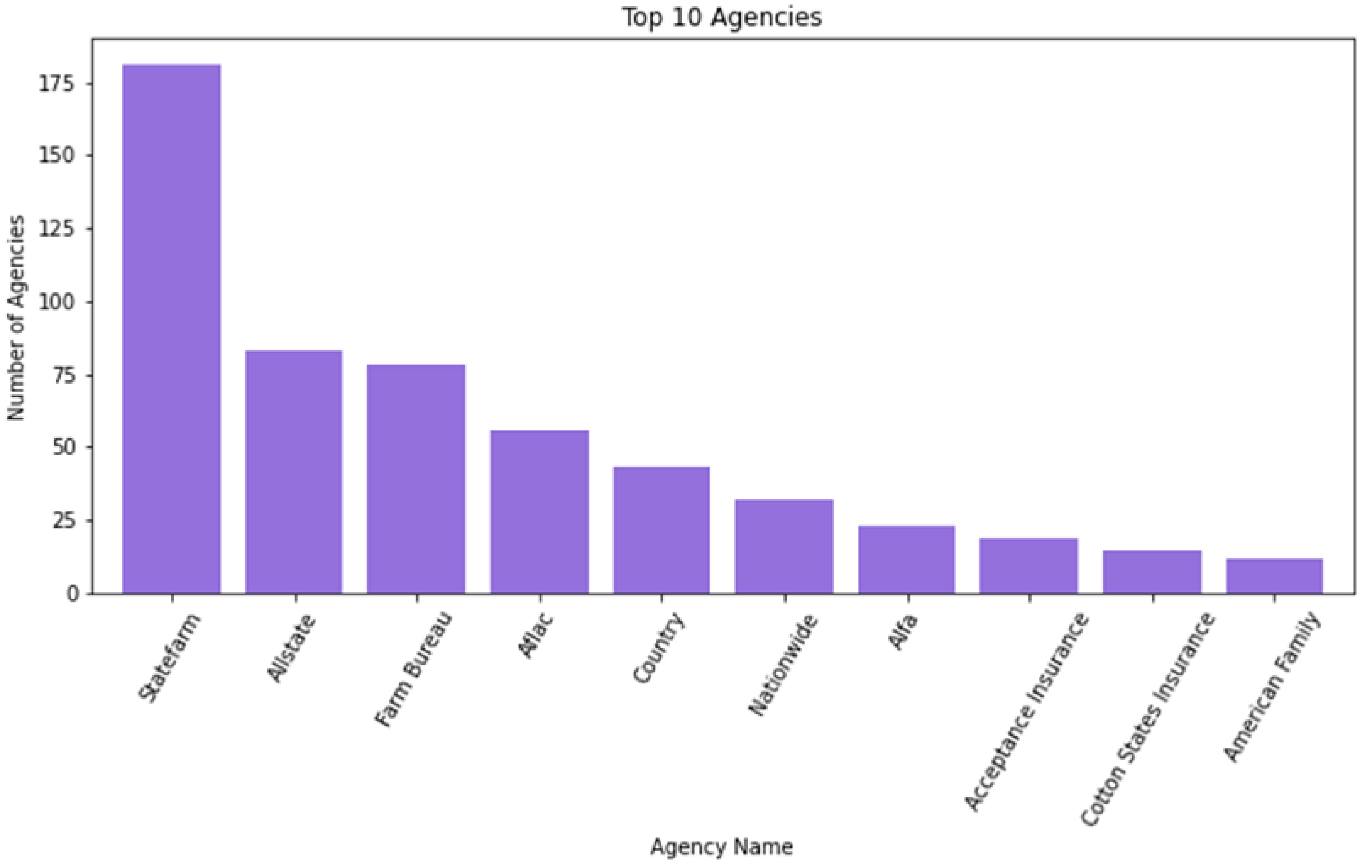
Q2) Which counties have higher than expected rates of insurance companies?

- **relative to population size**
- **relative to accident counts**

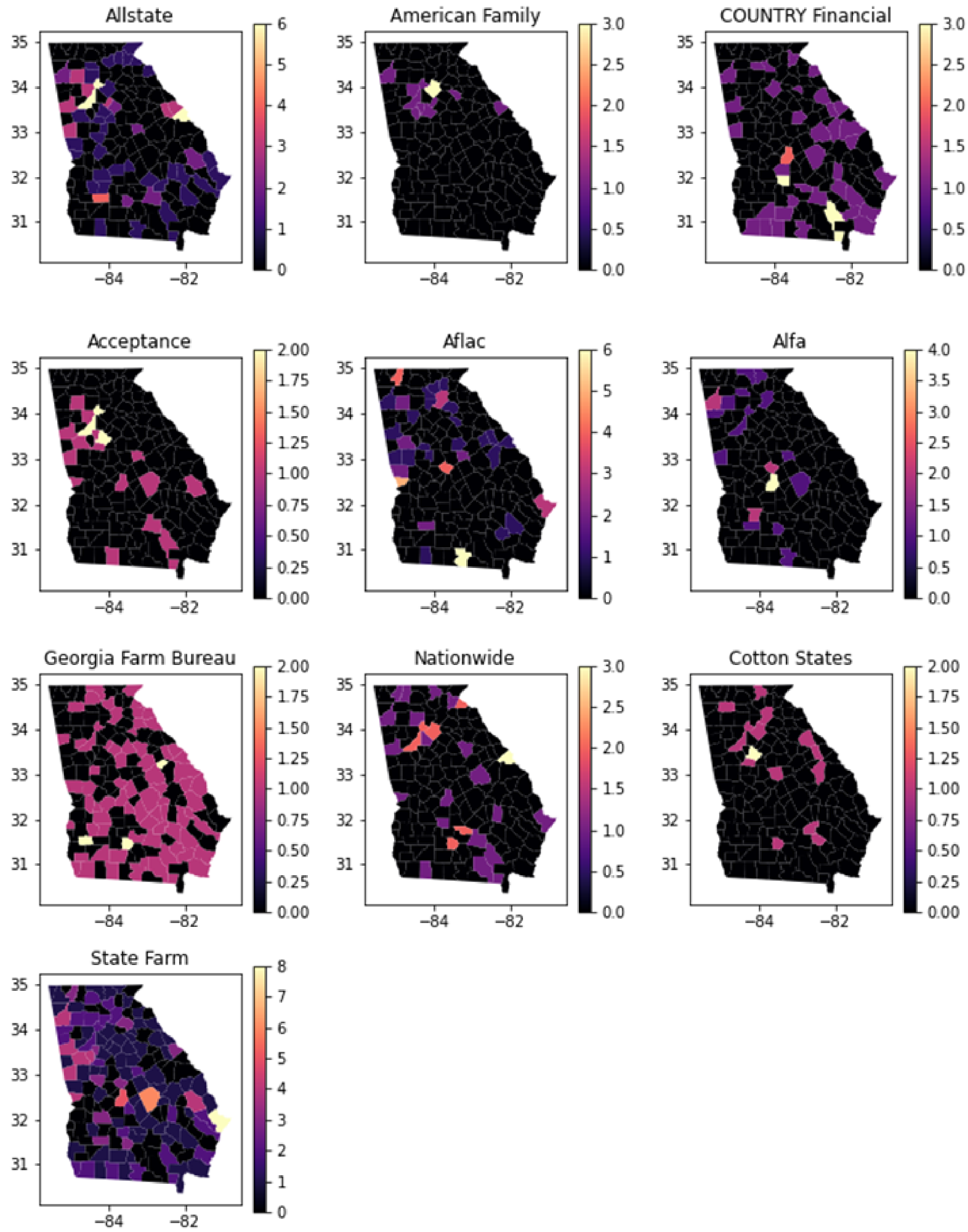
Q3) Does population size or frequency affect where offices are located?

Q1 & Q2:

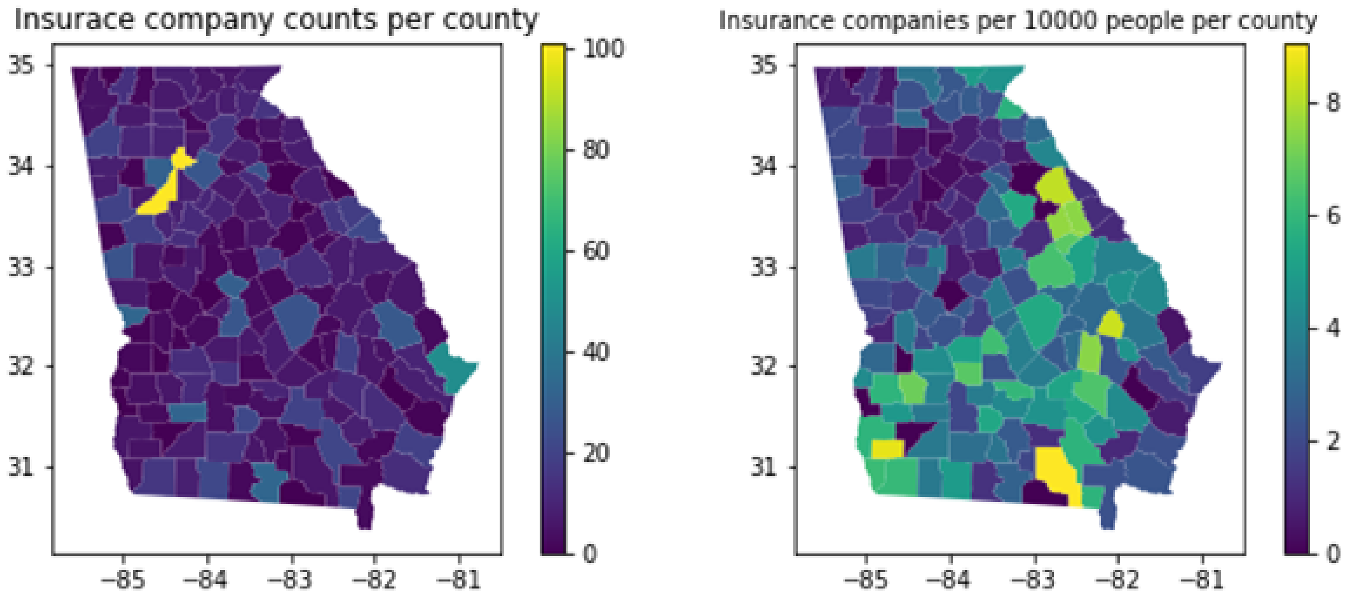
The most frequently occurring insurance agency was State Farm, with over 170 locations. Next were Allstate and Farm Bureau, which had over 70 locations. There were a total of 910 agencies, but after the first 10 agencies, there was a natural break in the distribution, with most of the remaining distribution having less than 5 locations.

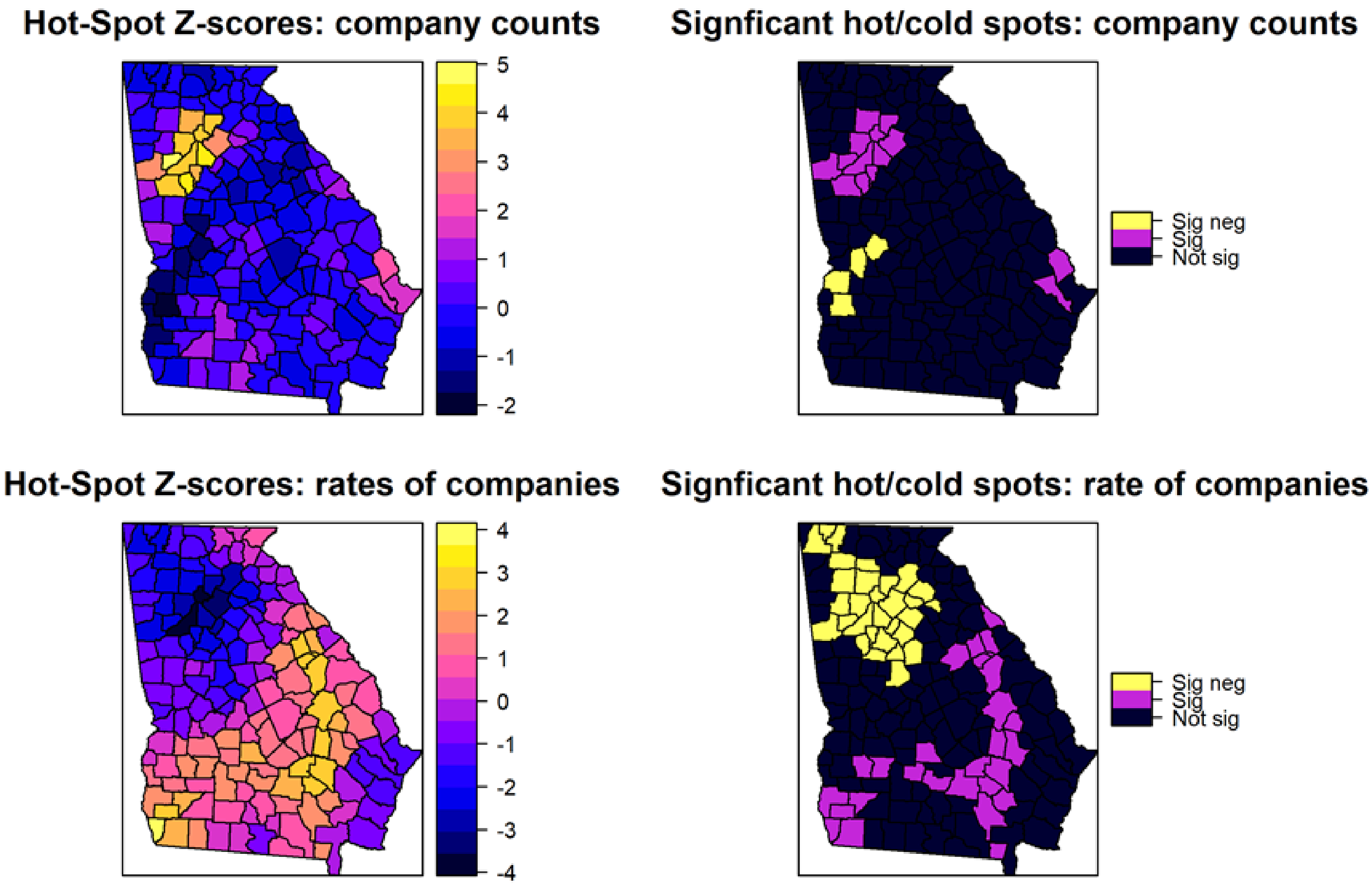


Overall, the most common agencies displayed substantial variation among counties in the state. However, there were agencies such as Farm Bureau that were more evenly distributed among rural counties. Ownership between agencies is not always obvious and could affect agency distributions. For example, Cotton States insurance is affiliated with COUNTRY Financial insurance; which could explain what looks like a negative spatial correlation between Cotton States counts and COUNTRY Financial counts.



Counts of insurance companies were highest, by far, in Fulton County (= 101) followed by Chatham county (49). Overall, in terms of counts of agencies, the Atlanta metropolitan area was a significant hot spot: a higher than expected proportion of the total count of agencies was associated with that county and its surrounding counties (figure below, left). The third highest was a tie between Lowndes County and Muscogee County (both with a count of 33), those counties the relative frequency distribution tailed out more slowly. Nine of the 159 counties that had zero insurance agencies.


Insurace company counts per county
35
34
33
32
31
−85
−84
−83
−82
−81
100
80
60
40
20
0
Insurance companies per 10000 people per county
35
34
33
32
31
−85
−84
−83
−82
−81
8
6
4
2
0



Hot-Spot Z-scores: company counts

Signficant hot/cold spots: company counts

Hot-Spot Z-scores: rates of companies

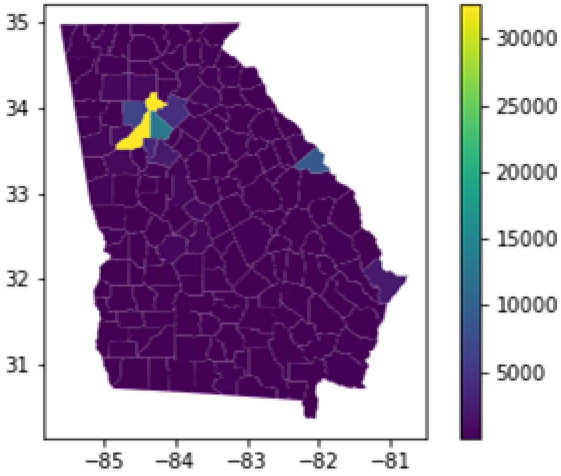
Signficant hot/cold spots: rate of companies

When population was used as a baseline for insurance agency counts, resultant rates of insurance agencies showed the inverse pattern: there was a significant hot spot over a belt of rural counties in southeastern Georgia, and a significant coldspot in the Atlanta metro area. The highest baseline rates were in Clinch, Miller and Candler counties.

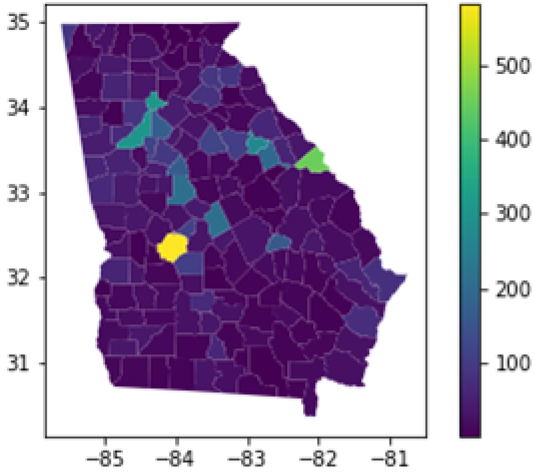
At the present time, we can only speculate on why there are more insurance companies per 10000 people in rural southeastern Georgia. There appear to be many small offices spread around many of these small towns, perhaps because there is a local market for basic services. It might also be worthwhile to examine demographic factors (such as age-class structure) that could influence choices about insurance agents.

In terms of accidents, there was a significant hot spot in the Atlanta metro area. However, accident rates were highest in Macon ("Macon-Bibb") county, followed by Richmond County (in the Augusta metropolitan region). Together with the Atlanta metropolitan region, these areas constituted statistically significant hot-spots for accidents.

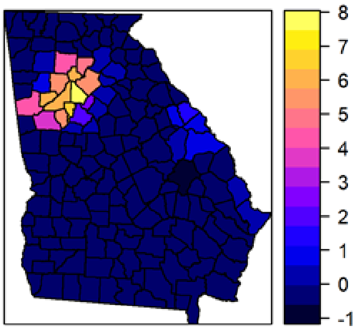
Automobile accident counts per county



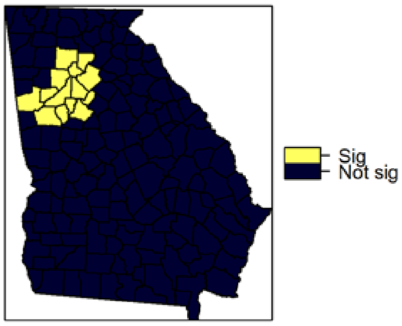
Automobile accidents per 10000 people per county



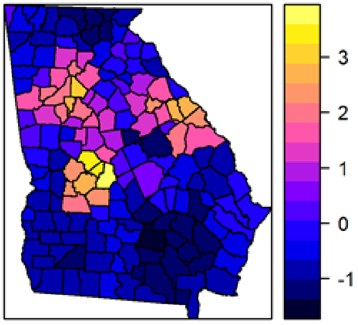
Hot-Spot Z-scores: accident_counts



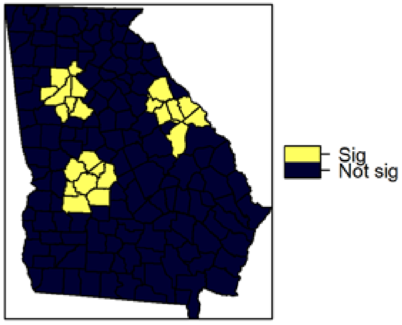
Signficant hot/cold spots: accident counts



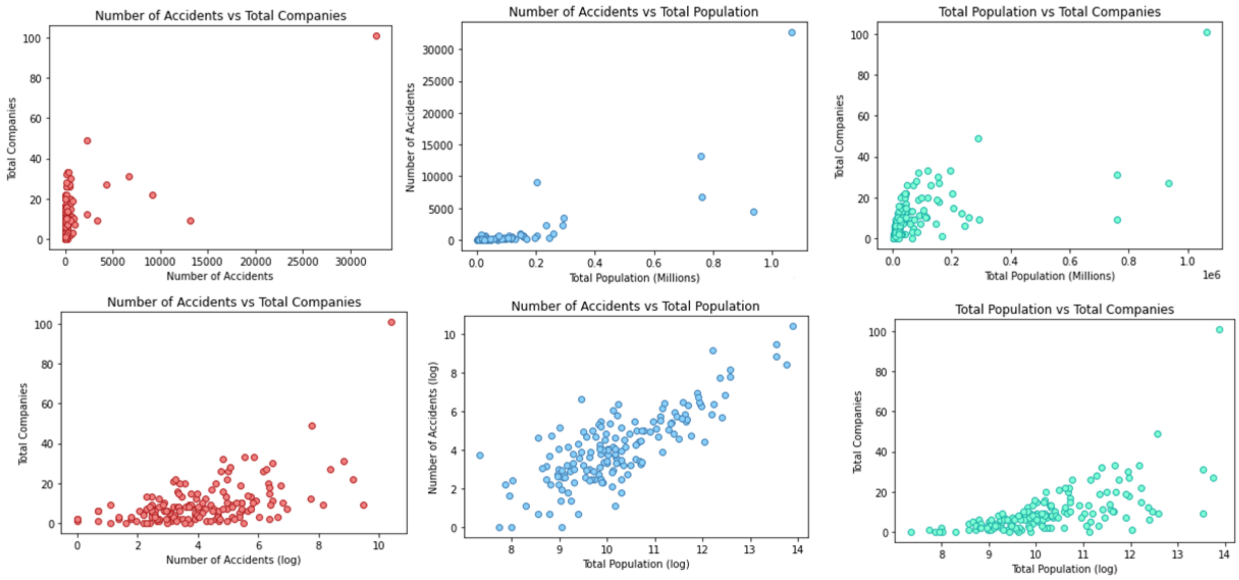
Hot-Spot Z-scores: accident_rates



Signficant hot/cold spots: accident rates



Scatterplots



Example scatter plots for agency counts, accident counts, and county census size:

Total companies in the county against accident frequency and log accident frequency (**left panels)**. Accident frequency and log accident frequency against total population in the county and log total population in the county (middle panels). Total companies in the county against total population in the county and log total in the county population (**right panels**).

The counts for all three variables are overdispersed. Both agency and accident count data were right skewed and over-dispersed. For example, the variance to mean ratio for agency count per county equaled 0.08 (mean = 9.50 agencies; var = 123.57). When plotted against total population size, or against each other, the error terms are heteroskedastic (i.e., the variance in company and accident counts is more variable in the counties with the largest census sizes). All three variables show evidence of spatial structure, and even with log transformation, error terms in these could be autocorrelated.

Q3 (Does population size or frequency affect where offices are located?)

Many factors likely contribute to where agencies are located. Population size is an obvious baseline for agency counts and, when corrected for, different patterns emerged for rates of insurance agencies. We picked accident frequency as our initial predictor variable because these data are readily available and it seemed like an obvious starting point for an auto-insurance company, but there are likely many other important predictors that were not included, that would be interesting to include in further studies as information about age-class distributions.

The regression coefficient for accident count was statistically significant at α = 0.05 in the Quasi-Poisson GLM (**insurance_data/insurance_GLM_preliminary.R**). Since many of the assumptions of regression are stretched by these data, we are hesitant to report the results formally here; for heuristic purposes, we have made the code and regression table summary available in the .R files.

Further analyses might consider a spatial regression or geographically-weighted regression, but these models can become unwieldy and more limited in model evaluation options for a tricky GLM. While building highly predictive models of agency distributions cannot be based on population size and accident counts alone, additional exploratory spatial analysis with methods suitable for inhomogeneous count data, as in point pattern analysis might be the most fruitful approach.

Conclusions

Obtaining complete information about all insurance agencies and reported accidents will likely require the use of many sources. While gigantic (4.2 million rows), even the Kaggle dataset may have biases associated with where and how accidents are reported. Our results provide valuable basic information about the intensity of auto-insurance agencies and accidents in different counties in Georgia, including where there are significant hot spots and cold spots in the frequency of auto insurance agencies and accidents, as well as detailed maps showing the frequency of different companies in different Georgia counties. In the process of analyzing these data we developed useful code for places searches with the Google Maps API, that includes a slick recursive call to the next_page_token, and an additional search strategy that improves the search process by using an optimized grid of locations for the places search, followed by a spatial join with the county polygons (to avoid reverse geocoding). These methods can applied to other regions containing insurance agencies, or other types of places easily accessed through the places API.