Transit Usage in Seattle: A Spatial Investigation

Peter Silverstein

2024-12-10

Final Project

GIS and Spatial Analysis

QMSS5070

Introduction

Research Question:

1. How does transit usage percentage (percent of trips using mass transit / total trips per census tract) vary spatially in and around Seattle and Tacoma, Washington?
2. How does this variation relate to population density, median income, and median age at the census tract level?

Purpose of Study:

There are essentially two purposes to this study. The first is to better understand where there are concentrations of high and low transit usage around the region. If there is clustering and we see hotspots and coldspots, further policy-focused questions can be asked. For example: given clustering, what characteristics of a census tract makes in more or less likely to be in one of these hot or cold zones? How might we allocate resources across hot zones, cold zones, and those in-between to increase the adoption of transit by commuters? Is the dispersion of transit availability closely related to the demand and does the dispersion favor certain demographic groups over others?

The second research question is a very basic attempt at answering one of these follow-up questions. By understanding how the three variables (population density, median income, and median age) are related to the outcome of interest (percentage of commuter trips taken using public transit), we can begin to fill in the knowledge gaps demonstrated by the questions above.

Hypotheses

1. I believe we will see transit hotspots close to urban centers (e.g., Seattle and Tacoma, the two biggest cities in the region of interest). Further, I believe the opposite will be true for coldspots–they should exist further outside urban centers. These ideas are based on the fact that transit lines themselves tend to be clustered in high-density, urban areas, meaning opportunities for mass transit travel are more convenient and plentiful in more central urban areas.
2. I expect that transit use percentage is positively associated with population density and median income and negatively associated with age. I make this hypothesis about population density based on the reasoning above. I expect younger people to (a) be more likely to live in highly urban areas and (b) be less likely to own a personal vehicle (such as a car). Of the three variables, I am the least confident about median income, because I think the relationship could be pulled in both positive and negative directions. On one hand, urban areas tend to be more expensive and thus have a higher requirement for income to live there. On the other hand, lower income should be associated with lower rates of car ownership and thus lower income would be associated with higher transit ridership.

Data and Methodology:

Data Sources:

1. The **Puget Sound Regional Association (PSRC) Household Travel Survey (HTS) 2017-23** is a biennial survey of commuters done in the King, Kitsap, Pierce, and Snohomish counties of Washington state (the counties surrounding Seattle and Tacoma). The present analysis uses the Trips dataset from the HTS. Each observation in the dataset represents a single trip taken by a respondent and includes a variety of variables. Most important for my analysis are origin/destination tract and mode of travel, although the dataset also includes date, time, distance, speed, etc.

2. All **census tract-level ACS 2022 5-year estimates for demographic data and the associated geometries** were accessed via the R `tidycensus` package, which leverages an API connection to the US Census Bureau to provide US Census data for a specified geographic area.
3. Finally, **Stanford's Cities and Towns of the United States, 2014** dataset provided point data to allow me to add city labels to my maps for reference.

Data Preparation/Spatial Data Management:

Data Cleaning

The output from `tidycensus` is already quite clean, so the majority of data cleaning steps were conducted on the PSRC HTS dataset. After loading the dataset, I selected my columns of interest and converted their types where appropriate and useful (e.g., string to factor). The next step was to collapse the mode of travel column from around 50 unique responses (an artifact of (a) a very detailed survey and (b) some option changes over the years of the survey) to just 4 useful categories, outlined below:

1. *Mass Transit:* included in this category trips that used a metro bus, private bus or shuttle, urban rail/light rail, school bus, ferry, paratransit, and commuter rail. Essentially I included any multi-occupancy transit vehicle.
2. *Personal Vehicle:* trips including all single-occupancy motor vehicles. This includes personal cars, ride-shares, taxis, motorcycles, and car-share services.
3. *Active Transit:* included walking, running, biking, and skateboarding.
4. *Other:* included helicopter/plane, "other" responses.

I then implemented a number of filters to filter data that didn't pass muster for realism/were outside the scope of this question. This included the following operations:

1. Filtered non-complete survey responses
2. Filtered distance to the range of 0 to 150 miles
3. Filtered trip duration to only include those greater than 0 minutes
4. Filtered speed to exclude speeds greater than 150 mph
5. Filtered to remove observations with missing value for travel mode

Spatial Joins

The next step was to count the number of trips and join these values to the geometries/ACS data from tidyverse. I used the `summarize()` function to count the number of total trips and number of trips per category for each census tracts. I then joined these counts to my geometry table via the `left_join()` function and matched on GEOID. Further, I removed any tracts from the analysis with 0 total trips. I acknowledge that this is a bit of a simplistic solution to missingness and will discuss it further in my analysis and conclusions. Finally, I calculated the percentage of trips in each tract that were made by mass transit mode (count of mass transit / count of total trips).

GIS Methodology Overview

Manipulation/Analytical Methods Used

As mentioned above, I counted the number of trips inside each census tract, then converted the total trip and mass transit trip counts to percentage. Finally, I used a simple join function to associate the counts with their respective geometries.

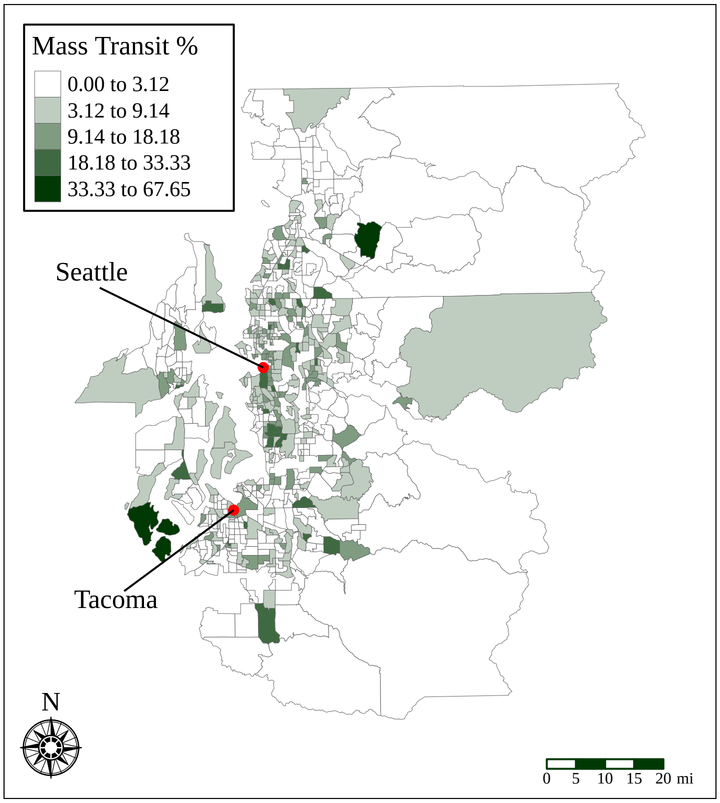
For the analysis in this project, I will perform a global cluster analysis (using Moran's I) and visualize any hot and cold zones using Getis-Ord Gi*. These approaches should help me understand whether transit use is clustered and visualize where it is clustered. I will then run a Spatial Error Model (SEM) and a Spatial Lag Model (SLM) regressing mass transit percentage on population density, median income, and median age. I believe that the SEM is the better conceptual choice - I believe that it is likely that unmeasured factors (e.g., land use, transit quality/reliability, parking availability) are spatially correlated, while the idea that ridership in one tract influences another is a bit harder to intuit. That said, I will run both models in order to compare the results.

Software Used

All data loading, cleaning, and manipulation, mapping (both choropleth and Getis-Ord Gi*), table-creation, regression modeling, and write-up were performed with R and R-Studio.

Results and Analysis

Mass Transit Trips as a % of Total Trips (classification = Jenks)


Mass Transit %
0.00 to 3.12
3.12 to 9.14
9.14 to 18.18
18.18 to 33.33
33.33 to 67.65
Seattle
Tacoma
N
0 5 10 15 20 mi