Bike Share Propensity Index

Innovation Incubator – Spring 2023

Ulises Hernandez
Brynn Leopold
Drusilla van Hengel

Nelson\Nygaard

Contents

1 Introduction

2 Literature review

3 Methods

4 Results

Introduction

Bikeshare systems provide sustainable and resilient transportation options in densely populated urban areas.

Over the past decade, bikeshare programs experienced remarkable growth across the United States. For instance, New York City's Citi Bike, one of the most extensive bike share programs in the country, significantly expanded its network, going from around 6,000 bicycles at launch in May 2013 to over 20,000 bicycles by 2021.

As more systems hope to continue their expansion, it is important to distribute the benefits of bikeshare broadly. This project focuses on developing a better understanding of the attributes associated with bikeshare trip-making.

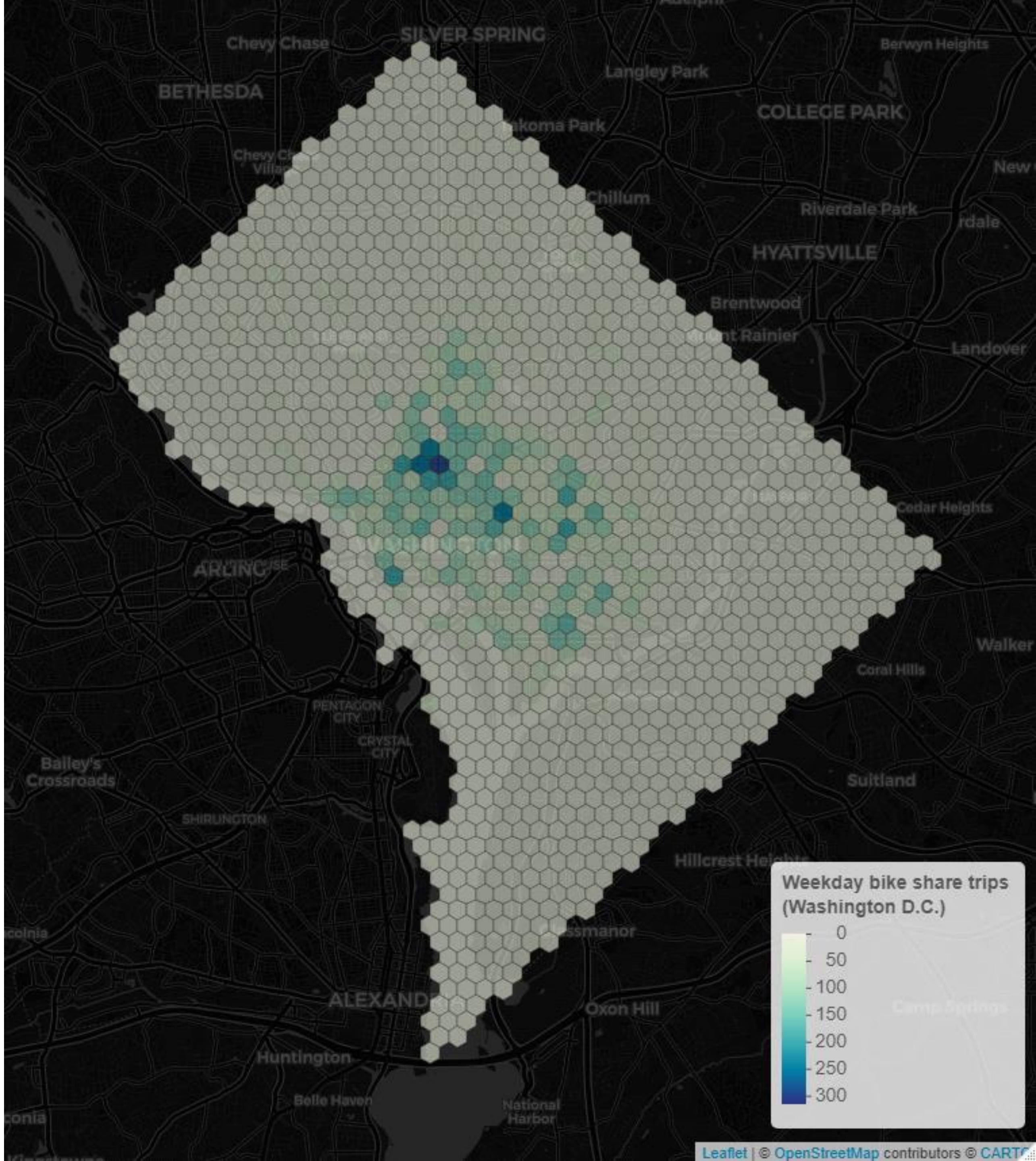
Insights into rider preferences, infrastructure requirements, pricing models, and network expansion strategies can enable cities to tailor their programs effectively, increase ridership, and reap the environmental and transportation benefits associated with bike sharing.

Project Description

This project developed a methodology for calculating a Bike Share Propensity Index (BSPI) to identify areas that exhibit characteristics favorable to bikeshare trip origin activity. The higher the BSPI value, the more likely it is for bikeshare to succeed in that area.

The results and insights gained from this Innovation Incubator project will enable us to better support cities in **determining when/where to expand networks and developing policy interventions to boost underperforming systems**. The BSPI research may also support station redistribution strategies.

Key to the creation of the BSPI is choosing a replicable and valid statistical modeling approach.



Incubator Project Scope

Our team followed these concepts to narrow down the BSPI methodology

Focus on bike share trip activity at a specific locations

Studies on bike share systems can address various aspects of bike share activity. Some studies concentrate on predicting bike share demand, which encompasses trip origination and generation. They achieve this by utilizing demographic and infrastructure data related to station locations, as well as factors such as transit accessibility, proximity to points of interest, time of day, and weather conditions. Other research has delved into understanding the factors influencing bike share trip flows, often focusing on origin-destination trip pairs.

In our project, we specifically focus on trip origination at specific locations and aim to determine which demographic, infrastructure, and travel characteristics associated with these locations influence bike share trips.

Identify a statistical approach to assess propensity

Typically, propensity indexes are developed as composite scores. This usually involves selecting a set of variables commonly associated with bike share usage and then standardizing the range of those variables. For instance, normalizing all variables from 1 to 10 and then add them up to generate a final propensity index, where a higher score represents a greater propensity for usage.

While this approach is straightforward to convey and implement, our project focuses on identifying a statistical method that can assess the significance of explanatory variables. This method will help determine which variables should be included in the propensity index.

Use nationwide available datasets for explanatory variables

All statistical analyses are constrained by the availability and quality of the data. Developing a BSPI is no exception. While research has concentrated on similar types of explanatory variables for modeling bike share trips, such as demographics, land use, and transit infrastructure, the precise variables and data may differ considerably from one study to another.

Given that one of the project's objectives is replicability, we have given preference to using variables found in nationwide datasets, such as Census Data, or datasets that are routinely collected by local jurisdictions, such as bikeways infrastructure.

Key Methodology Takeaways from Literature Review

Independent Explanatory Variables

Bikeshare studies use two main categories of variables to predict bikeshare activity:

1. **Direct data from surveys on the characteristics of bike share riders and non-bike share riders**
2. **Indirect data related to trip origins/destinations characteristics**
    - Demographic and socioeconomic variables
    - Built environment/Urban design variables
    - Transportation/Travel characteristics variables

Dependent Variables

- Studies often try to predict **trip volume (continuous variable), origin-to-destination flows (continuous variable), or bikeshare membership (discrete variable).**
- Most recent studies, particularly those interested on trip volume, highlight the presence of spatial autocorrelation in the dependent variable.
- New methods are needed to reduce the impact of spatial and temporal autocorrelation (of things like bikeshare stations) on ridership variability.

Statistical methods to address spatial autocorrelation

Neighborhoods, block groups, or stations close to each other exhibit similar characteristics and usage patterns. Three key approaches for reducing the statistical impacts of proximity are summarized below.

Geographically Weighted Regression (GWR): GWR recognizes that the relationships between independent and dependent variables may vary across different locations within a study area. GWR estimates separate regression coefficients for each location and reveals how the relationships change spatially.

A critical component of GWR models involves a **moving window or kernel that downweights peripheral observations**. The kernel moves through the study area, and at each location, it computes a local model. It utilizes data under the kernel to construct a local model at that location, with data points farther away from the kernel center receiving lower weights in the solution

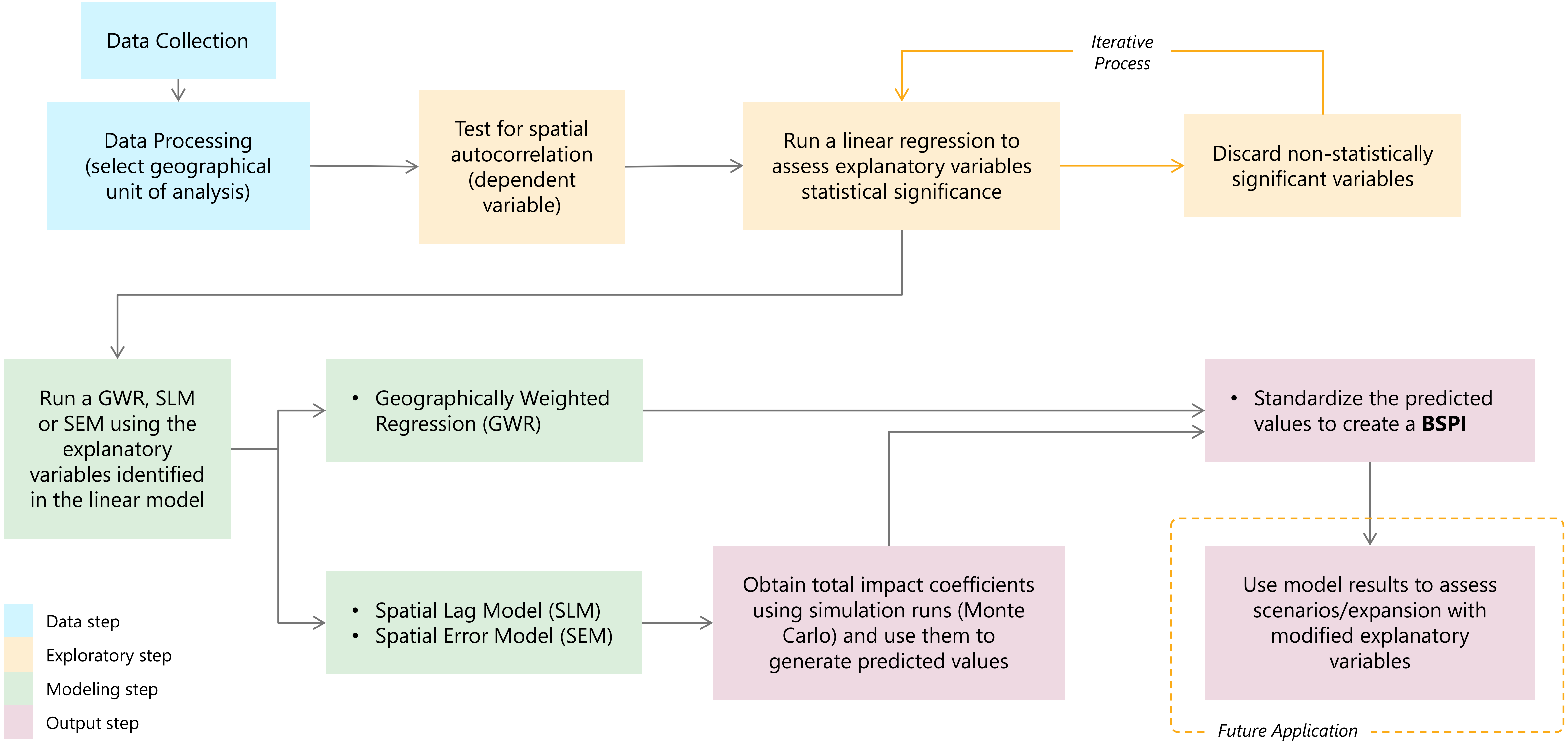
Spatial Lag Model: A Spatial Lag Model captures the idea that a location's value is influenced by the values of nearby locations. This approach is useful when there is a spatial pattern in the dependent variable that can be explained by the values of neighboring observations.

*"**spatial dependence observed in our data does not reflect a truly spatial process, but merely the geographical clustering of the sources of the behaviour of interest**. For example, citizens in adjoining neighbourhoods may favour the same (political) candidate not because they talk to their neighbors, but because citizens with similar incomes tend to cluster geographically, and income also predicts vote choice. Such spatial dependence can be termed attributional dependence"* (Darmofal, 2015: 4)

Spatial Error Model: The Spatial Error Model accounts for spatial autocorrelation by allowing for spatially correlated error terms. It assumes that the residuals from the model are spatially autocorrelated and models this correlation explicitly. It's suitable when spatial dependencies are present in the unexplained variation of the dependent variable.

*"If so the behaviour is likely to be highly social in nature, and **understanding the interactions between interdependent units is critical to understanding the behaviour in question**. For example, citizens may discuss politics across adjoining neighbours such that an increase in support for a candidate in one neighbourhood directly leads to an increase in support for the candidate in adjoining neighbourhoods"* (Darmofal, 2015: 4)

BSPI methodology overview



Data Collection

Data Processing (select geographical unit of analysis)

Test for spatial autocorrelation (dependent variable)

Run a linear regression to assess explanatory variables statistical significance

Iterative Process

Discard non-statistically significant variables

Run a GWR, SLM or SEM using the explanatory variables identified in the linear model

- Geographically Weighted Regression (GWR)

- Spatial Lag Model (SLM)
- Spatial Error Model (SEM)

Obtain total impact coefficients using simulation runs (Monte Carlo) and use them to generate predicted values

- Standardize the predicted values to create a **BSPI**

Use model results to assess scenarios/expansion with modified explanatory variables

Future Application

🟦 Data step

🟨 Exploratory step

🟩 Modeling step

🟪 Output step

Data Inventory

For this project, we gathered more than 50 variables from various sources for **Washington D.C.** and classified them into three types. While this categorization does not significantly impact the model specification, it proves helpful in ensuring that the modeling exercise incorporates variables identified as relevant in recent research. We actively cleaned and transformed the data to prepare it for use in the statistical model. A brief description of key steps is provided below.

Data Gathering: Collect relevant data from identified sources. Most datasets were directly downloaded as shapefiles from their respective sources, except for U.S. Census data, which we obtained directly in R using the Census API.

Handling Missing Data: Identify and address missing data. This step primarily involved ensuring that all modeled areas had a non-NA value. For most variables, this meant imputing a value of zero.

Removing Duplicates: Check for and eliminate any duplicate entries in the datasets.

Detecting Outliers: Identify and address outliers that might skew the analysis. This step is particularly crucial for the dependent variable (bike share trips); we removed trips with a duration of less than one minute and trips lasting three or more hours.

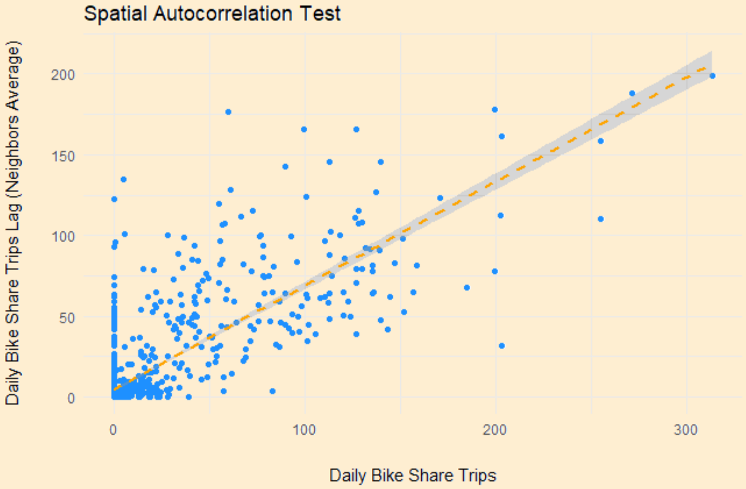
Converting Data Types: Ensure that variables are in the correct format (numeric, categorical, date) for analysis. This also involved preparing the data to be used as an absolute metric or as a percentage. For instance, household vehicle ownership was initially obtained as households per census tract. However, we transformed and tested this metric as a percentage of households per tract.

Category	Variable	Unit	Data Source
Transit	Bus Stops	count	Open Data DC
	Bus Ridership	daily average	Open Data DC
	Metrorail stops	count	Open Data DC
Built environment, Land Use	Bike Share Station Count	count	Capital Bikeshare
	Length of bikeways (class I, II and IV)	feet	Open Data DC
	Average commute time to work	minutes	U.S. DOT Equitable Transportation Community Explorer
	Estimated Average Walk Time to Points of Interest	minutes	
	Dwelling units - single family	units	Replica
	Dwelling units - multifamily	units	
	Dwelling units - mixed use	units	
	Total building area	square feet	
	Single family building area	square feet	
	Multifamily building area	square feet	
	Retail building area	square feet	
	Office building area	square feet	
	Attractions building area	square feet	
	Industrial building area	square feet	
	Health building area	square feet	
	Education building area	square feet	
	Civic building area	square feet	
	Transportation building area	square feet	
	Open space building area	square feet	

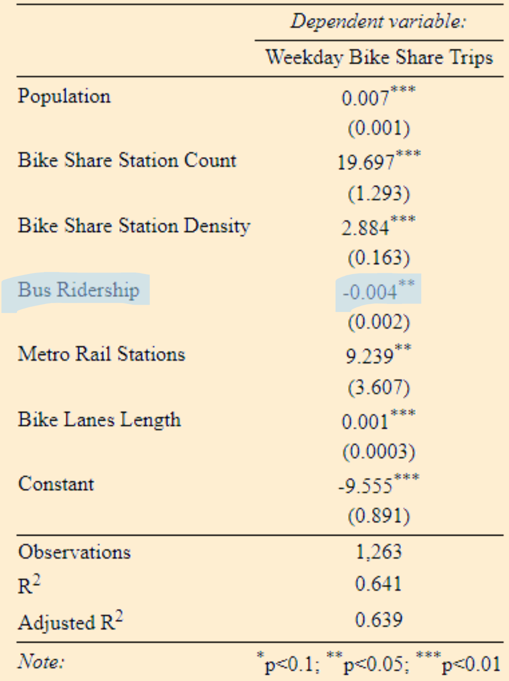
Data Inventory cont.

Category	Variable	Unit	Data Source
Demographic and socioeconomic	Total population	persons	ACS 5-year estimates 2017-2021
	White population (not Hispanic)	persons	
	Black or African American population (not Hispanic)	persons	
	American Indian and Alaska Native population (not Hispanic)	persons	
	Asian population (not Hispanic)	persons	
	Native Hawaiian and Other Pacific Islander population (not Hispanic)	persons	
	Other race population (not Hispanic)	persons	
	Two or more races population (not Hispanic)	persons	
	Hispanic (any race)	persons	
	Total households	household units	
	Households with annual income below $25,000	household units	
	Households with annual income between $25,000 - $50,000	household units	
	Households with annual income between $50,000 - $75,000	household units	
	Households with annual income between $75,000 - $100,000	household units	
	Households with annual income above $100,000	household units	
	Population 18 years old and younger	persons	
	Population between 18 - 24 years old	persons	
	Population between 25 - 39 years old	persons	
	Population between 40- 54 years old	persons	
	Population between 55 - 64 years old	persons	
	Population 65 years old and older	persons	
	Total female population	persons	
	Total male population	persons	
	Households with zero vehicles	household units	
	Households with one vehicle	household units	
	Households with two or more vehicles	household units	
	Households with one or two vehicles	household units	
	Population with Income below 200% of poverty level	percent from total population	

Exploratory Step

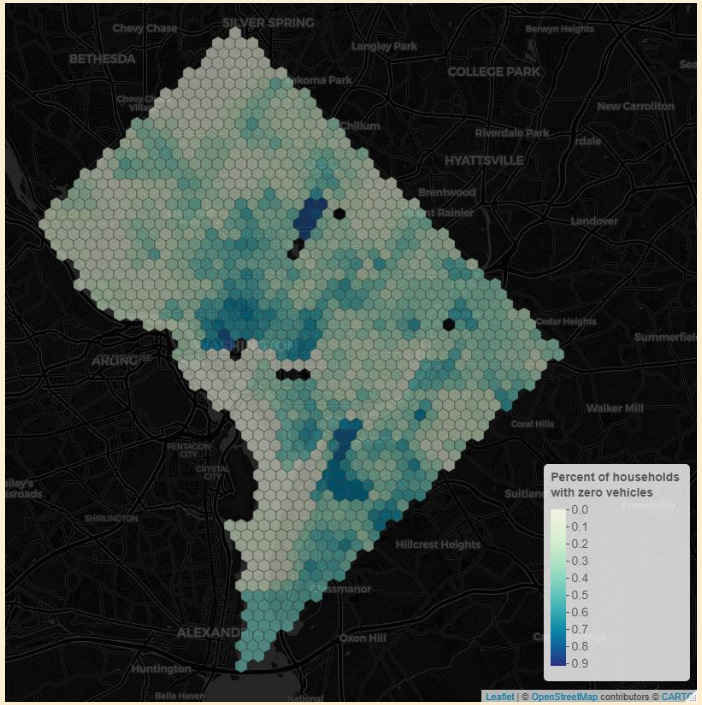


After collecting and processing the data, **the first exploratory step involved testing the dependent variable for spatial autocorrelation**. The scatterplot above illustrates the relationship between daily bikeshare trips by hex bin and the average daily bikeshare trips of the neighboring hex bins. The **positive correlation suggests that hex bins with high bikeshare activity are likely neighboring hex bins with high bikeshare activity**. Furthermore, a more formal test (Moran's I test) substantiated the spatial autocorrelation in bikeshare trips. Therefore, a spatial model was deemed appropriate for evaluating variables that explain bikeshare ridership.



	Dependent variable:
	Weekday Bike Share Trips
Population	0.007^{***}
	(0.001)
Bike Share Station Count	19.697^{***}
	(1.293)
Bike Share Station Density	2.884^{***}
	(0.163)
Bus Ridership	-0.004^{**}
	(0.002)
Metro Rail Stations	9.239^{**}
	(3.607)
Bike Lanes Length	0.001^{***}
	(0.0003)
Constant	-9.555^{***}
	(0.891)
Observations	1,263
R^2	0.641
Adjusted R^2	0.639
Note:	*p<0.1; **p<0.05; ***p<0.01

We analyzed the correlation of multiple transportation, built environment, and socioeconomic variables with bikeshare trip volume. This iterative process involved running dozens of linear regressions to identify the variables with the highest significance. We conducted this assessment initially by variable type, and then we mixed variables to find a robust set for use in the spatial models. This process is discretionary and requires judgment to assess the statistical significance of the explanatory variable and whether the coefficient has the expected direction.



In one of our initial correlation analyses, we observed that all transportation variables were statistically significant and had the anticipated impact, except for Bus Ridership. **We opted to exclude Bus Ridership** from the selected independent variables since we concluded that this negative correlation could be due to the high bus ridership and low bikeshare activity in the southeast of D.C. However, characterizing this zone was still possible through other variables such as income.

RESULTS

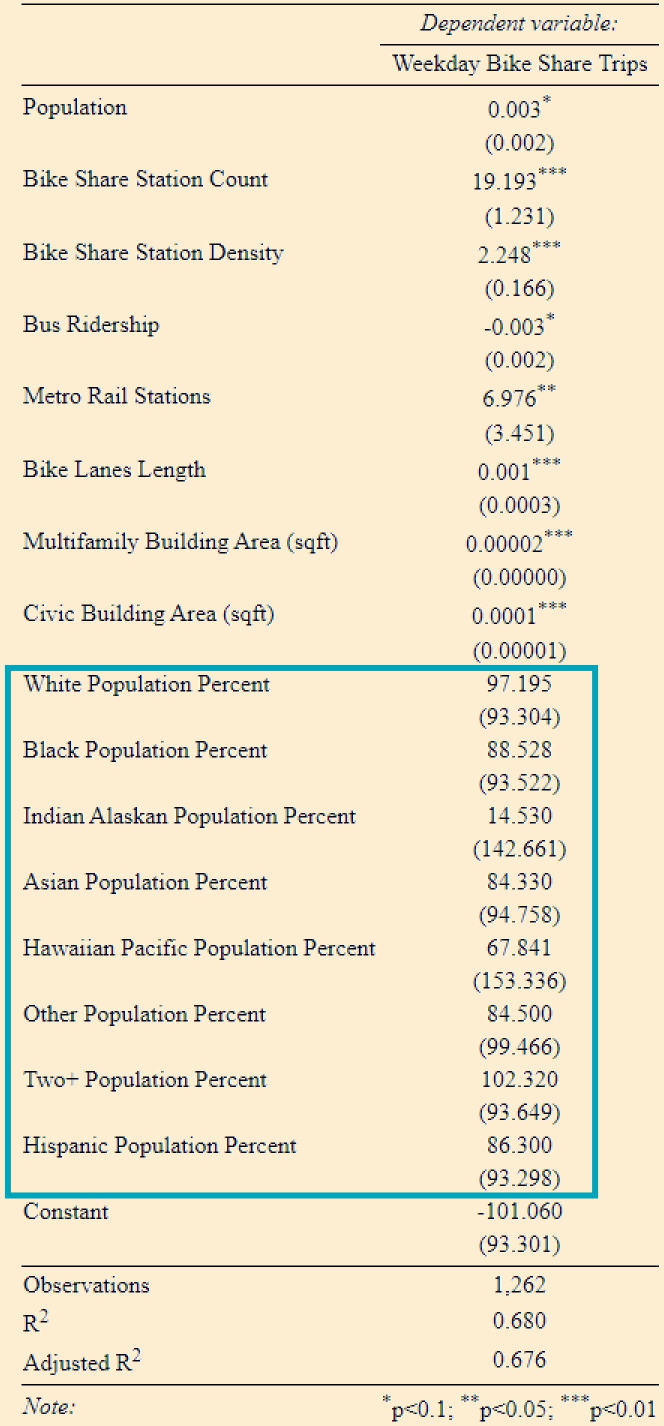
Exploratory Step

One of the most interesting findings is that race composition is not a meaningful characteristic for explaining bikeshare ridership—specifically when referring to the race of individuals near bike stations, not the actual people using bikeshare.

For assessing propensity through the bikeshare station's environment, it was better to exclude race as an explanatory variable. This is because race might merely express other characteristics that effectively impact bikeshare ridership. In urban planning, it is well-documented that minority neighborhoods have historically received minimal or no investment in walk and bike facilities. Bikeshare systems are typically located in the urban core, where white male workers might constitute a larger proportion of the total workers in the area.

This result underscores the importance of understanding the needs of disadvantaged communities to ensure that bikeshare functions for them (or identifying alternatives if it does not).

It also emphasizes the need to focus propensity assessments on built environment characteristics more likely to have a causal relationship with cycling.



	Dependent variable:
	Weekday Bike Share Trips
Population	0.003*
	(0.002)
Bike Share Station Count	19.193***
	(1.231)
Bike Share Station Density	2.248***
	(0.166)
Bus Ridership	-0.003*
	(0.002)
Metro Rail Stations	6.976**
	(3.451)
Bike Lanes Length	0.001***
	(0.0003)
Multifamily Building Area (sqft)	0.00002***
	(0.00000)
Civic Building Area (sqft)	0.0001***
	(0.00001)
White Population Percent	97.195
	(93.304)
Black Population Percent	88.528
	(93.522)
Indian Alaskan Population Percent	14.530
	(142.661)
Asian Population Percent	84.330
	(94.758)
Hawaiian Pacific Population Percent	67.841
	(153.336)
Other Population Percent	84.500
	(99.466)
Two+ Population Percent	102.320
	(93.649)
Hispanic Population Percent	86.300
	(93.298)
Constant	-101.060
	(93.301)
Observations	1,262
R^2	0.680
Adjusted R^2	0.676
Note:	*p<0.1; **p<0.05; ***p<0.01

Selected variables for Spatial Modeling

Variable	Coefficient	Statistical significance
Population	0.004	95%
Bike Share Station Count	18.7	99%
Bike Share Station Density	2.18	99%
Metrorail Stations	6.01	90%
Bike Lanes Length (ft.)	0.001	99%
Single family Building Area (sq. ft.)	-0.000016	99%
Multifamily Building Area (sq. ft.)	0.000019	99%
Civic Building Area (sq. ft.)	0.00008	99%
Average Commute Time (mins.)	-0.11	99%
Percent of population below 200% poverty	-18.7	99%

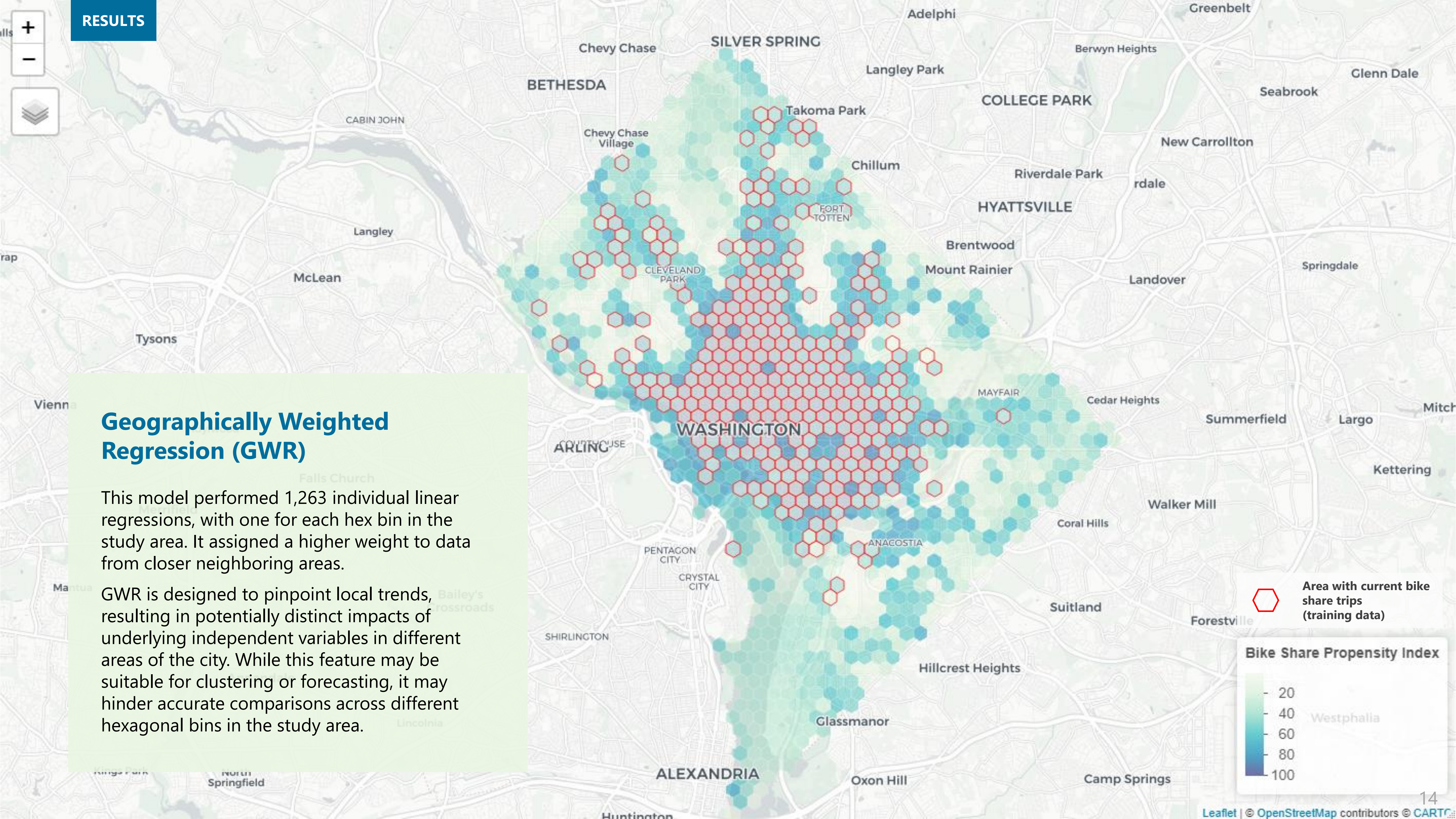
Adjusted R^2 0.72

RESULTS

Geographically Weighted Regression (GWR)

This model performed 1,263 individual linear regressions, with one for each hex bin in the study area. It assigned a higher weight to data from closer neighboring areas.

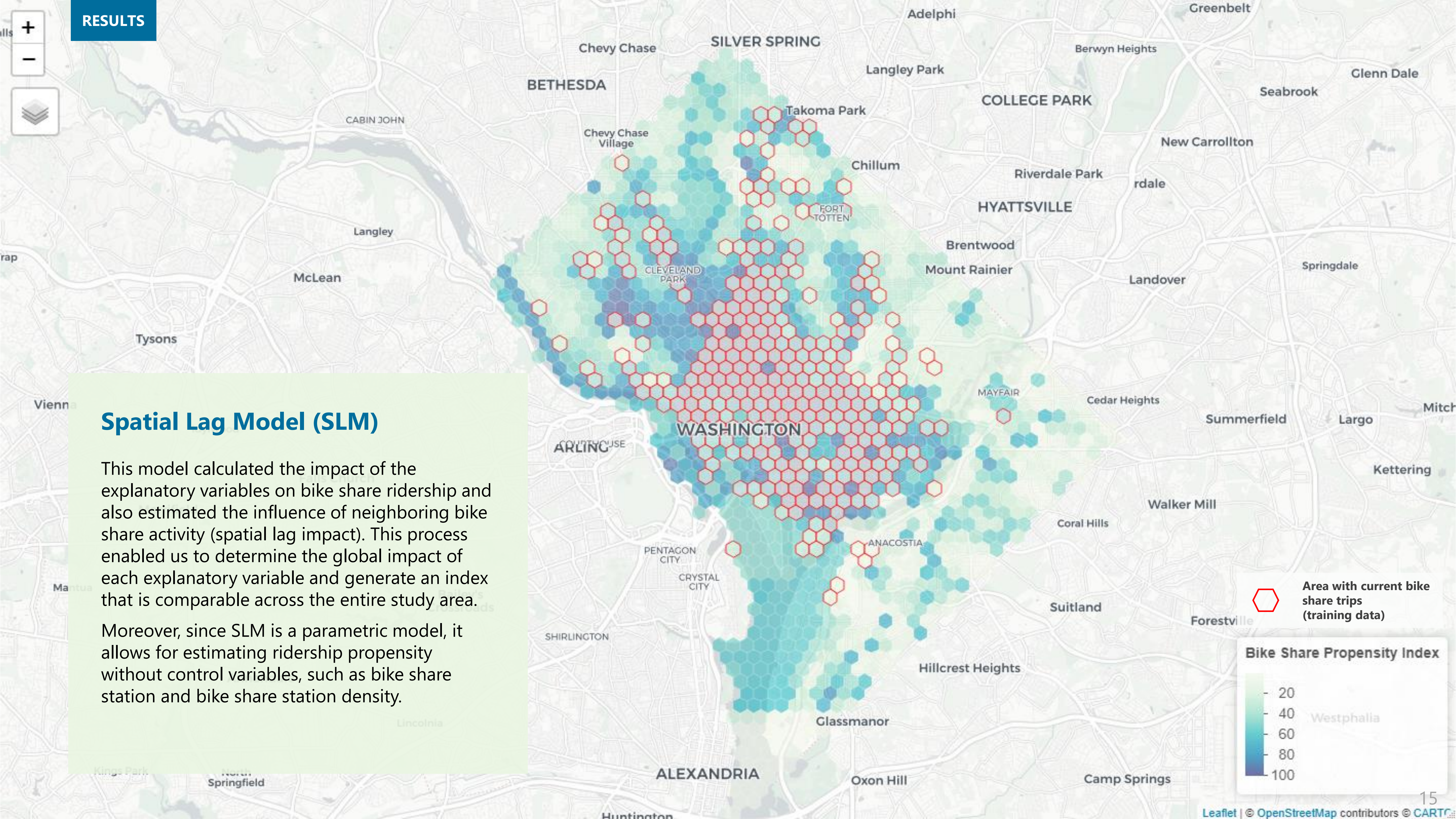
GWR is designed to pinpoint local trends, resulting in potentially distinct impacts of underlying independent variables in different areas of the city. While this feature may be suitable for clustering or forecasting, it may hinder accurate comparisons across different hexagonal bins in the study area.



⬡ Area with current bike share trips (training data)

Bike Share Propensity Index

20
40
60
80
100

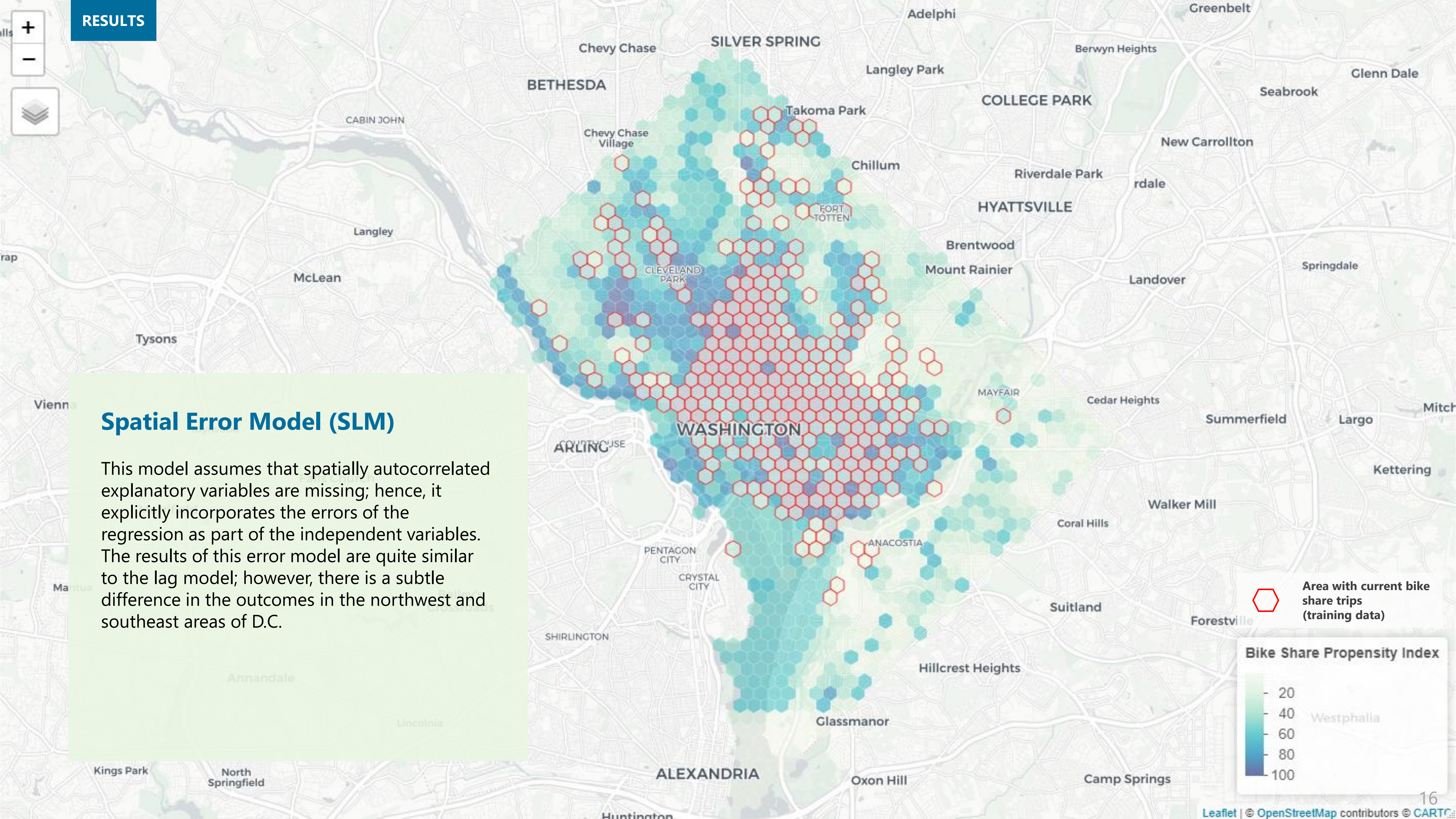


Spatial Lag Model (SLM)

This model calculated the impact of the explanatory variables on bike share ridership and also estimated the influence of neighboring bike share activity (spatial lag impact). This process enabled us to determine the global impact of each explanatory variable and generate an index that is comparable across the entire study area.

Moreover, since SLM is a parametric model, it allows for estimating ridership propensity without control variables, such as bike share station and bike share station density.

RESULTS



Spatial Error Model (SLM)

This model assumes that spatially autocorrelated explanatory variables are missing; hence, it explicitly incorporates the errors of the regression as part of the independent variables. The results of this error model are quite similar to the lag model; however, there is a subtle difference in the outcomes in the northwest and southeast areas of D.C.

Next Steps

Bike Share Propensity Index Refinement

- Test the Spatial Lag Model and Spatial Error Model on additional cities
- Refine model to replace income variable with built form/infrastructure variable

Further Research

- Integrate BSPI methodology into bike share expansion analysis
  - *BSPI: Which areas of the city have the built environment characteristics to support bike share?*
  - Mobility deserts: which areas of the city lack access to transportation options?
  - Equity assessment: where do cities want to prioritize investments to provide high quality options?
- Explore methods that account for temporal autocorrelation